Genetics of local adaptation in *Arabidopsis thaliana* – seed dormancy as a case study

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Ilkka Kronholm

aus Turku, Finnland

Köln, 2010

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln in der Abteilung für Pflanzenzüchtung und Genetik (Prof. Dr. M. Koornneef) in der Gruppe Dr. J. de Meaux angefertigt.

Berichterstatter:	Prof. Dr. Maarten Koornneef
	Prof. Dr. Thomas Wiehe
Prüfungsvorsitzender:	Prof. Dr. Siegfried Roth
Tag der Disputation:	30.06. 2010

Publications

Kronholm, I., Loudet, O. & de Meaux, J. 2010. Influence of mutation rate on estimators of genetic differentiation – lessons from *Arabidopsis thaliana*. *BMC Genetics* 11: 33

Abstract

Local adaptation occurs when natural selection favours different phenotypes in different locations. Here, I studied the genetics of adaptation using local adaptation for seed dormancy in *Arabidopsis thaliana* as a model system. I asked, is there local adaptation for seed dormancy and what environmental factors drive it? What is the genetic basis of adaptation and what is the molecular nature of adaptive changes? To answer these questions I conducted a population genetic study, comparing neutral markers, candidate genes and traits. Some QTL-mapping experiments were also performed. The results obtained indicate that there is local adaptation in seed dormancy and this is mediated by the amount of precipitation received during the summer months. Local adaptation seems to occur at a regional geographic scale. Based on genetic mapping and studies, the large effect gene *DOG1* is mainly responsible for adaptation, together with several other loci with minor effects. A population genetic study of *DOG1* revealed that there is a signature of local selection on *DOG1*. Several functional alleles of *DOG1* are segregating in natural populations. Mutations that increase or decrease dormancy seem to have occurred several times independently. This likely happens because of a low migration rate, new mutations occur in separate populations but they cannot migrate efficiently to other populations and thus no single mutation becomes fixed. The molecular basis of adaptive changes could not be determined, yet some candidate mutations for functional changes were identified. In addition, some of the results raised concerns about the proper way to estimate genetic differentiation. Therefore, the statistical properties of some estimators of genetic differentiation were studied using computer simulations. An estimator that takes mutation model into account can be used to compare different types of markers.

Zusammenfassung

Lokale Anpassung erfolgt, wenn durch natürliche Selektion in verschiedenen Regionen, unterschiedliche Phänotypen favorisiert werden. In der vorliegenden Arbeit wurden die genetischen Grundlagen der evolutionären Anpassung am Beispiel der Samenruhe im Modellsystem *Arabidopsis thaliana* untersucht. Hierbei habe ich ermittelt, ob lokale Anpassung der Samenruhe vorliegt und welche ökologischen Faktoren diese möglicherweise vorantreiben. Was ist die genetische Grundlage der Adaptation und was sind die molekularen Mechanismen, die zur Anpassung führen? Um diese Fragen zu beantworten, führte ich eine populationsgenetische Studie durch, in der neutrale Markergene, Kandidatengene und Merkmalsausprägungen verglichen wurden. Zudem wurden Experimente zur Ermittlung von QTLs durchgeführt. Die Ergebnisse weisen darauf hin, dass lokale Anpassung der Samenruhe auftritt und diese durch die Niederschlagsmenge in den Sommermonaten beeinflusst wird. Lokale Anpassung scheint auf regionaler Ebene statt zu finden. Basierend auf Genkartierung und anderen Experimenten ergab sich, dass das Haupteinflussgen *DOG1* größtenteils für die Adaptation verantwortlich ist, zusammen mit mehreren anderen Genorten mit geringerem Einfluss. Eine populationsgenetische Untersuchung von *DOG1* zeigte, dass es Anzeichen lokaler Selektion von *DOG1* gibt. Einige funktionale Allele von *DOG1* mendeln in natürlichen Populationen. Mutationen, welche die Samenruhe vermindern oder erhöhen sind vermutlich mehrfach unabhängig voneinander aufgetreten. Dies resultiert wahrscheinlich aus einer geringen Durchmischungsrate, da neue Mutationen in separaten Populationen entstehen, sich diese allerdings nicht effektiv ausbreiten können und daher kein einzelnes Allel in allen Populationen fixiert wird. Die exakten, molekularen Grundlagen der Anpassung konnten nicht ermittelt werden, aber einige Mutationen wurden identifiziert, die vermutlich funktionale Veränderungen bewirken. Darüber hinaus geben einige Ergebnisse Anlass, Bedenken über den richtigen Weg der Ermittlung genetischer Differenzierung zu äußern. Daher wurden die statistischen Eigenschaften einiger Schätzfunktionen genetischer Differenzierung mittels Computersimulationen untersucht. Eine Schätzfunktion, welche mögliche Arten von Mutationen einbezieht erwies sich beim Vergleich verschiedener Markertypen als sinnvoll.

Contents

List of Figures

List of Tables

Glossary

This is a list of abbreviations used in this study. Gene names are given in *italics* and are capitalised, mutants are in lowercase.

ABA	Abscisic acid
AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
AR	Allelic richness
CI	Confidence interval
D25	Time required to reach germination of 25 %
D50	Time required to reach germination of 50 %
D75	Time required to reach germination of 75 %
Df	Degrees of freedom
DNA	Deoxyribonucleic acid
DOG1	*DELAY OF GERMINATION 1*
DTB	Days to bolting
FRI	*FRIGIDA*
GA	Gibberellic acid
LD	Linkage disequilibrium
LOD	Logarithm of odds
mRNA	Messenger ribonucleic acid
MSD	Mean squared deviations
MTIME	Maturation time
PCA	Principle component analysis
PCR	Polymerase chain reaction
QTL	Quantitative trait locus
r	Correlation coefficient (Pearson)
REML	Restricted maximum likelihood
SNP	Single nucleotide polymorphism
WODS	Weeks of dry storage

This is a list of mathematical notation used. Lowercase subscripts mean indexes for a particular group, e. g. N_g means the number of individuals in group g and P_g means number of populations in group g.

σ_b^2	Genetic variance between populations (from molecular markers)
σ_w^2	Genetic variance within populations (from molecular markers)
σ_a^2	Genetic variance between regions (from molecular markers)
σ_G^2	Genetic variance (from a quantitative trait)
σ_E^2	Environmental variance (from a quantitative trait)
σ_{GB}^2	Genetic variance between populations (from a quantitative trait)
σ_{GW}^2	Genetic variance within populations (from a quantitative trait)
σ_m^2	Genetic variance from new mutations (for a quantitative trait)
σ_α^2	Variance of allelic effects of new mutations (for a QTL)
F_{ST}	Summary statistic, measures genetic differentiation between populations
F_{IS}	Summary statistic, measures departure from Hardy-Weinberg expectation within populations
F_{CT}	Summary statistic, measures genetic differentiation between regions
Φ_{ST}	Summary statistic, analogous to F_{ST}
F'_{ST}	Summary statistic, standardised F_{ST}
D	Summary statistic, measures allelic differentiation
Q_{ST}	Summary statistic, analogous to F_{ST} (for a quantitative trait)
H^2	Broad sense heritability
h_m^2	Mutational heritability
Θ	Population mutation rate (for microsatellite markers)
H_S	Subpopulation heterozygosity, also called gene diversity
H_T	Total heterozygosity
N	Number of individuals (in a population)
P	Number of populations
G	Number of regions (groups of populations)
L	Number of loci
A	Number of alleles
m	Migration rate
μ	Mutation rate
p	Allele frequency
a	Allelic effect

Chapter 1

Introduction

Organisms appear as if designed for a purpose, as many features in organisms match strikingly the environment they live in. The fundamental insight of Darwin was that organisms are designed by natural selection for the purpose of achieving reproductive success (fitness) (Gardner 2009). If organisms in a population differ with respect of some trait and this trait affects the reproductive success of the said organims, then certain types will leave more descendants than others. If the differences in the trait are hereditary, the composition of the population with respect to this trait will change over time. Darwin called this process natural selection (Darwin 1859). Later, Darwin's ideas were formulated in the terms of population genetics. A central result derived from population genetic theory concerning the action of natural selection is called the fundamental theorem of natural selection. The fundamental theorem of natural selection states that natural selection always acts to increase the mean fitness of a population (Fisher 2003). As there are other evolutionary mechanisms than selection, it is important to realise that the fundamental theorem describes only the part of change in fitness that is due to the action of natural selection and not the whole change in fitness (Crow 2002). The fundamental theorem therefore isolates the part of evolutionary change that constitutes adaptation.

While there are other evolutionary mechanisms than natural selection, selection is ultimately responsible for all the functional features organisms have. Adaption therefore, is at the core of evolutionary theory. In this work I study the genetics adaptation using local adaptation for seed dormancy in *Arabidopsis thaliana* as a model system.

1.1 Local Adaptation

As most adaptations were brought about by past natural selection, studying adaptation would require comparing the ancestral and the derived, adapted populations (Kawecki & Ebert 2004). This is often impossible for natural populations and for traits that are unconditionally adaptive. Therefore, local adaptation is interesting to study because it presents a situation where adaptation can be studied in progress.

When discussing what circumstances are favourable to adaptation[1], Darwin (1859) wrote that

> *"But if the area be large, its several districts will almost certainly present different conditions of life; and then if natural selection be modifying and improving a species in the several districts, there will be intercrossing with the other individuals of the same species on the confines of each. And in this case the effects of intercrossing can hardly be counterbalanced by natural selection always tending to modify all the individuals in each district in exactly same manner to the conditions of each;. . . "*

Then Darwin wrote that factors that prevent intercrossing should facilitate adaptation.

> *"In hermaphrodite organisms which cross only occasionally, and likewise in animals which unite for each birth, but which wander little and which can increase at a very rapid rate, a new and improved variety might be quickly formed on any one spot, and might there maintain itself in a body, so whatever intercrossing took place would be chiefly between the individuals of the same new variety. A local variety when thus formed might subsequently slowly spread to other districts."*

It is clear that Darwin understood that local selection would make populations better adapted to local conditions, while intercrossing (gene flow and recombination in modern terms) would act against natural selection to prevent adaptation to local conditions. Local adaptation is thus hardly a new idea in evolutionary biology.

Local adaptation means a situation where natural selection favours different phenotypic values or alleles in different populations. The reason for this is that environmental conditions are different (Kawecki & Ebert 2004). Usually, local adaptation is taken to imply that there is a trade-off; adapting to one population will

[1]The original title of Darwin's chapter reads "circumstances favourable to natural selection", thus he seems to equate natural selection and adaptation here. In modern terms he is talking about adaptation.

cause fitness to be low in the other populations. To express the previous in mathematical terms, the fitness function has different optima in different populations. The canonical equation for stabilising selection is

$$W = \mathrm{e}^{\left(-\frac{(P-Z_{Opt})^2}{2\omega^2}\right)} \quad \text{(Turelli 1984)} \tag{1.1}$$

where W is fitness, P is trait value, Z_{Opt} is the optimal trait value and ω is selection intesity. Thus local adaptation can be defined as a case where, the optimal trait value, Z_{Opt} is different for different populations $Z_{Opt1} \neq Z_{Opt2} \neq Z_{Opt3}$. This definition of local adaptation assumes the presence of trade-offs, since fitness follows the same exponential function with different optima in different populations. In this study, this is the definition used for local adaptation. In order for local adaptation to be possible there has to be at least some limits to gene flow and constraints on phenotypic plasticity (Kawecki & Ebert 2004).

Since plants are sessile organisms and cannot move once established, they have to tolerate any changes that occur in the environment. This can impose strong selection pressures for the populations to adapt to local conditions. Plants are often relatively easy to transplant to different locations and this has inspired a long history of research into local adaptation in plants, starting from the classical experiments of Turesson (1922, 1925) who coined the term *ecotype* to describe locally adapted genotypes. Clausen et al. (1941) transplanted *Potentilla glandulosa* originating from different altitudes to three different common gardens along an altitude gradient. They observed that differences between different genotypes persisted in a common garden and genotypes originating from a similar environments tended to do better at altitudes corresponding to their origin than genotypes from different altitudes. Local adaptation was mediated by differences in flowering time and the plants responsiveness to frost.

Flowering time is a well studied trait (Komeda 2004; Koornneef et al. 2004) and is involved in local adaptation in many different species (Schemske 1984; Le Corre 2005; Hall & Willis 2006). The study of Hall & Willis (2006) on *Mimulus guttatus* is a nice demonstration that natural selection favours different optima for flowering time in two different populations. Another example of a trait that has been intensively studied is heavy metal tolerance (Macnair 1987). Evolution of heavy metal tolerance has been obseved in many species, like in *Anthoxanthum odoratum* for which Antonovics & Bradshaw (1970) observed steep a cline in zinc tolerance across a transect of contamined soil near a mine and uncontaminated pasture.

A recent meta-analysis of reciprocal transplant experiments in plants revealed that local adaptation is rather common but not universal, with strong evidence for local adaptation found in 45 % of the cases (Leimu & Fischer 2008). Perhaps surprisingly, geographic distance did not have an effect on whether local adaptation was more common. However, population size did have a strong effect, with large

populations being more often locally adapted than smaller ones (52 % vs. 9 %) (Leimu & Fischer 2008). This is in agreement with population genetic theory, since selection is more effective and mutational supply is higher in large populations.

1.2 Genetic basis of adaptation

How does adaptation occur at the genetic level? The first theory of adaptation was put forth by R. A. Fisher in 1930 (Fisher 2003). Fisher argued that adaptation proceeds through many small changes, as small substitutions have the highest probability of being beneficial, because they are unlikely to cause deleterious pleiotropic effects. If this model were correct, only the methods of quantitative genetics would be applicable for studying adaptive traits, since the individual genes would have too small effects to be studied empirically. However, it turns out that this model lacks an important component: it does not consider the probability of fixation in addition to the allele being beneficial (Orr 2005). Orr showed that in Fishers model, the distribution of allelic effects fixed during the whole process of adaptation, also called an adaptive walk, follows an exponential distribution (Orr 1998). With few genetic changes of large effect that occur at the start of the adaptive walk, and then smaller changes later on. Why this happens is easy to understand intuitively. Every new mutation is initially present as a single copy in the population and is at great risk of being lost by genetic drift. Mutations that cause a large increase in fitness can more easily escape this loss as selection acts stronger on them. Later, building on the work of Gillespie (1983), Orr showed that essentially the same conclusions apply to a mutational landscape model of DNA sequences (Orr 2002, 2003).

This theory has been tested empirically with microbial systems and most of the results obtained so far are in qualitative agreement with the theory (Betancourt & Bollback 2006; Eyre-Walker & Keightley 2007). The problem is that since the theory is probabilistic, a large number of beneficial mutations are needed to test its predictions and these are difficult to obtain empirically. Also, in some cases the theory does not accurately predict the distribution of fixed allelic effects, such as in the computer simulation study of Cowperthwaite et al. (2005), where there was an excess of beneficial mutations of small effect. Moreover, if environmental change is gradual rather than sudden, fixed allelic effects are of smaller magnitude depending on the rate of environmental change (Collins et al. 2007; Kopp & Hermisson 2007; Collins & de Meaux 2009; Kopp & Hermisson 2009).

In the context of local adaptation, Griswold (2006) using computer simulations has shown that major effect alleles are again expected, especially if there is a balance between selection and migration. The distribution of allelic effects did not follow the exponential distribution in this case. Migration between the

populations tended to increase the effect of segregating alleles, and alleles of moderate magnitude explained most of the phenotypic differences observed between the populations.

Taking the above into consideration, population genetic theory predicts that in QTL-mapping experiments for adaptive traits, it is expected that there will be few QTLs of large effect and a larger number of QTLs of small effect. This prediction seems also to hold, at least qualitatively, for many cases. Most of the phenotypic differences between domesticated maize and its progenitor teosinte seem to be attributable to few QTL of large effect (Doebley 2004). The same pattern appears to emerge for sunflower domestication (Burke et al. 2002). For different *Mimulus* species for which the genetic basis of different pollinator syndromes was investigated (Bradshaw et al. 1998; Bradshaw & Schemske 2003) and for different stickleback morphs, where 1 major and 4 minor QTLs control differences in pelvic spines (Shapiro et al. 2004) which are an adaptive response to different environments.

1.2.1 Molecular basis of adaptive changes

Trivially, natural selection does not care about the molecular basis of the mutations that increase fitness. However, it has been suggested that phenotypes could be changed more easily by *cis*-regulatory changes than protein coding changes. The argument is that *cis*-regulatory elements are modular and thus they potentially have less deleterious pleiotropic effects, than changing a function of a protein (Stern 2000; Carroll 2005). This view has been challenged on several grounds, first it is not clear if mutations in *cis*-regulatory elements would truly have less pleiotropic effects than mutations in coding regions (Hoekstra & Coyne 2007). Transcription factors themselves can be modular (Lynch & Wagner 2008) and more importantly, there is plenty of evidence that protein coding changes and gene duplications have contributed to adaptation (Hoekstra & Coyne 2007). Carroll (2005, 2008) has argued that only (animal) form is expected to evolve predominantly by *cis*-regulatory changes and other kinds of phenotypic change might not follow this pattern. However, this argument is not very persuasive. First of all, the distinction between morphology and physiology is often less than clear, secondly, both morphology and physiology are built up by genetic networks during development (Hoekstra & Coyne 2007).

As different authors disagree over theoretical arguments, only empirical data will likely solve the debate. Stern & Orgogozo (2008) reviewed the data of mutations known to cause phenotypic differences within and between species. Review of the data revealed that in 22 % of cases a *cis*-regulatory mutation was involved, if only morphology was considered the figure was 40 %. Whether the mutations had been actually shown to be adaptive was not considered. There might be different biases in this data due to experimental reasons. However, currently there is

no evidence for the predominance of *cis*-regulatory changes. Nonetheless, the molecular basis of adaptive changes remains an interesting question. There was some indication that different types of mutations could be involved in within vs. between species differences (Stern & Orgogozo 2008). Whether this is true, remains to be seen.

1.3 How to detect local adaptation?

This study uses the methods of population genetics to detect local adaptation. While reciprocal transplant experiments are very powerful, only a limited number of populations can be studied due to logistical reasons. Moreover, seed dormancy is a difficult trait to study in the field. Thus, a population genetic approach was chosen for this study to permit studying seed dormancy variation on a large geographic scale.

Natural selection is not the only evolutionary force that changes allele frequencies. Real populations are finite in size and thus allele frequencies change over time due to random sampling, this is called genetic drift. Sudden changes in population size (population bottle-necks or founder events) can also change allele frequencies. Allele frequencies within a population can also change due to migration of individuals from other populations or due to mutations (Wright 1931). In order to study the effects of natural selection these demographic effects must be separated from the effects of selection. A reasonable assumption is that demographic events will affect all loci in the genome, but due to independent segregation, selection will only affect loci or closely linked regions that have an effect on fitness.

Selection within a population will reduce genetic diversity, as selection will remove poorly adapted genotypes. In the context of local adaptation, selection will favour different genotypes in different populations, so the populations will become genetically differentiated.

1.3.1 Genetics of populations

Allele frequency differences between populations can be quantified using the summary statistic F_{ST}. When allele frequencies at a locus are very different in different populations F_{ST} will be high and when allele frequencies are similar F_{ST} will be low (Wright 1951; Holsinger & Weir 2009). Originally Wright (1951) defined F_{ST} as the correlation between gametes (alleles) chosen randomly from a single population relative to the entire set of populations. Later it has been shown that F_{ST} can also be thought as an intraclass correlation coefficient, thus F_{ST} can be interpreted as a measure how genetic variation is partitioned between populations Holsinger & Weir (2009). F_{ST} can be estimated from genetic marker data using

the variances of allele frequencies, and has a general form

$$F_{ST} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{1.2}$$

Where σ_b^2 is the genetic variance between populations and σ_w^2 is the genetic variance within populations. F_{ST} can also be defined in terms of coalescence times as

$$F_{ST} = \frac{\bar{t}_1 - \bar{t}_0}{\bar{t}_1} \qquad \text{(Excoffier 2007)} \tag{1.3}$$

where $\bar{t}_1$ equals the average coalescence time of two alleles from different populations and $\bar{t}_0$ equals the average coalescence time of two alleles from the same population.

If many loci are compared, a neutral expection for F_{ST} can be obtained from the distribution of locus specific F_{ST} values. It is important to realise that F_{ST} (and many other summary statistics in population genetics) have two kinds of variance, sampling variance and evolutionary variance (Cockerham 1969). Sampling variance is due to the fact that not every individual within a population can be sampled. Sampling variance can be decreased by sampling more individuals. Evolutionary variance is caused by genetic drift. If we could run the process of evolution several times, we would get different allele frequencies for a locus each time due to the stochastic effect of drift. Evolutionary variance is not affected by the number of sampled individuals. Therefore, in order to detect a value of F_{ST} that is somehow different from the expected value, it is necessary to compare this value to a distribution of values derived from multiple loci. By comparing the locus of interest to a neutral distribution, it can be inferred whether selection is acting on this locus (Beaumont & Balding 2004; Beaumont 2005; Storz 2005). Local adaptation will increase F_{ST} relative to neutral expection, while balancing selection will maintain low F_{ST} between populations (Charlesworth et al. 1997).

An F_{ST} analog, Q_{ST} can also be calculated for quantitative traits (Spitze 1993). It has been shown that $Q_{ST} = F_{ST}$ when the trait is neutral (Rogers & Harpending 1983; Lande 1992) and this result holds for many different demographic scenarios (Lande 1992; Whitlock 1999). If $Q_{ST} > F_{ST}$ this has been taken to indicate local adaptation and $Q_{ST} < F_{ST}$ has been suggested to indicate selection for the same optima in different populations. Porcher et al. (2004) showed experimentally that $Q_{ST} > F_{ST}$ if there is local selection. In most studies Q_{ST} was often found to be higher than F_{ST} for many different phenotypic traits, reviewed in Merilä & Crnokrak (2001) and again in Leinonen et al. (2008).

1.3.2 Population genetics of local adaptation

In order to illustrate how natural selection influences the measures of genetic differentiation described in section 1.3.1, I did a computer simulation of local adaptation. Let there be four populations of size 500 of outcrossing diploid hermaphrodites which exchange migrants at a rate of $m = 0.1$. There is a single quantitative trait which is under stabilising selection following equation 1.1. Let there be 20 loci that can potentially have an effect on the trait if there is a mutation in these loci, and there are also 30 loci which do not affect the trait. The loci are unlinked. The phenotypic trait is neutral for the first 6000 generations, after which the environment changes such that the populations will have different optima for the phenotype afterwards. Output of the simulation is shown in Figure 1.1, when selection starts the populations diverge rapidly in the phenotypic trait, as shown by Q_{ST} and allele frequencies at the underlying QTL, as shown by F_{ST} at QTL. Migration keeps F_{ST} at the other markers low.

The F_{ST} distributions of the different loci are shown in Figure 1.2, all of the neutral loci have very low F_{ST} but some of the QTLs have higher values of F_{ST}. The populations have differentiated phenotypically. There is a correlation between F_{ST} at the QTL and the amount of phenotypic variance explained with loci having larger phenotypic effects are more differentiated. Only some of the loci that could affect the trait have mutations in them, with only few that have large effects and several that have smaller effect. This was suggested by theoretical arguments in section 1.2.

In the above example, parameter values were obviously chosen so that large effects could be seen and it is not intended to be an accurate description of a particular situation that happens in nature. However, it serves to illustrate the population genetic reasoning which is later applied to an empirical setting used in this study.

1.3.3 Estimation of genetic differentiation

Some recent studies have raised the concerns about the reliability of F_{ST} for characterisation of population structure using markers with high mutation rates, such as microsatellites (Hedrick 1999; Balloux et al. 2000; Hedrick 2005; Jost 2008). High levels of within population diversity bias F_{ST} downwards, because F_{ST} is estimated using heterozygosities or genetic variances, see equation 1.2. If a locus has multiple alleles, classical F_{ST} can be low, even if populations share no alleles (Kalinowski 2002; Hedrick 2005; Jost 2008). In addition to classical F_{ST}, there are other estimators that have been proposed over the years. An analogous estimator to F_{ST}, Φ_{ST}, takes into account the distances between alleles thereby correcting for mutation rate (Slatkin 1995; Excoffier 2007). Classical F_{ST}, estimated in the

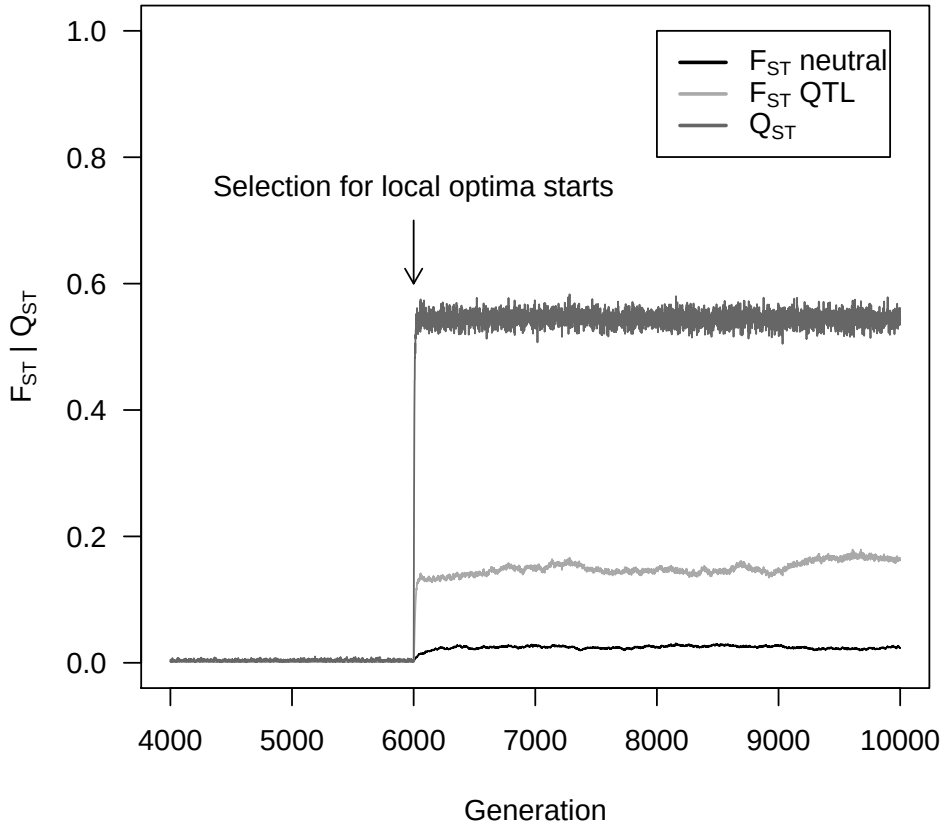

Figure 1.1: Computer simulation illustrating how selection causes high genetic differentiation in the phenotypic trait and the underlying QTL, while differentiation at neutral markers remains low. For simplicity, mean F_{ST} across loci is shown. Values are the means of 5 different replicate simulations. Different coloured lines correspond to genetic differentiation at different markers as indicated by the legend. For the first 6000 generations the phenotypic otimum is the same for all populations, while afterwards each population has its own optimum.

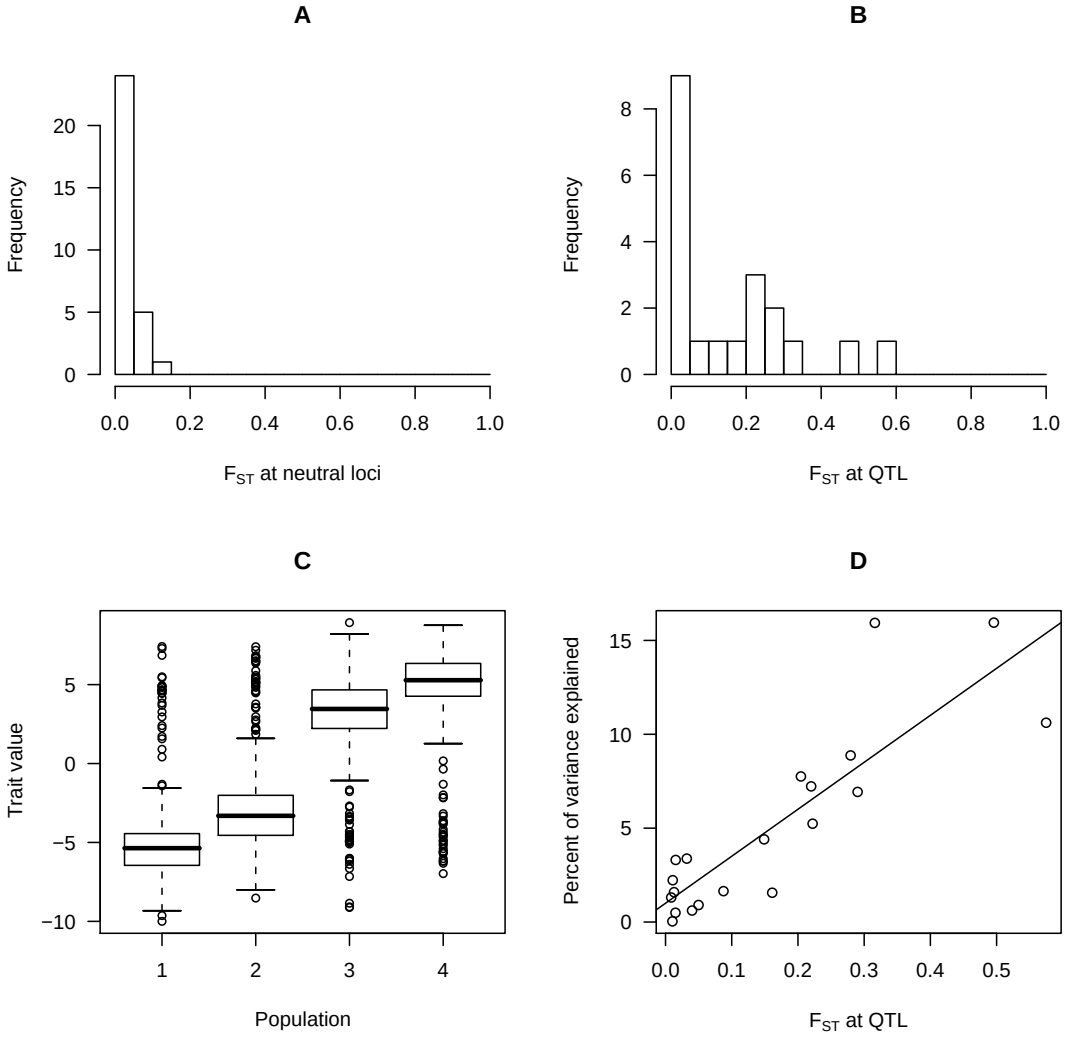

Figure 1.2: State of the computer simulation at the final generation. A) Distribution of F_{ST} at the neutral markers. B) Distribution of F_{ST} at QTL. C) Distribution of phenotypes for the different populations, means coincide with the different optima of the fitness function. D) Relationship between F_{ST} at QTL and the amount of genetic variance explained by that locus.

framework of Weir & Cockerham (1984) considers only allele identity while Φ_{ST} considers distances between the alleles, be it differences in repeat number (e.g. in the case of microsatellite) or number of pairwise differences between DNA haplotypes. Another measure, F'_{ST}, standardises the observed F_{ST} value with the maximum possible value that F_{ST} could attain given the amount of observed diversity (Hedrick 2005). Finally, Jost (2008) derived recently a new measure of genetic differentiation, D, to replace G_{ST} (or its equivalent for empirical studies F_{ST}). D measures allelic differentiation by partitioning heterozygosity into within and between population components, see appendix C for definitions of D and F'_{ST}.

1.4 Seed dormancy

Seed dormancy is defined as the inability of the seed to germinate in favourable conditions (Finch-Savage & Leubner-Metzger 2006). Normally, seeds depend on some environmental cues for germination to be induced, such as moisture and light. However, if the seeds are dormant, these cues are not sufficient to induce germination, and seeds need some other extrinsic or intrinsic signals that first break seed dormancy. Non-dormant seeds of *Arabidopsis thaliana* need light and water to be able to germinate (Bentsink & Koornneef 2008). Germination starts with the uptake of water by the dry seed, this is followed by rupture of the seed coat and endosperm, then the radicle can emerge from the seed.

1.4.1 Physiological and molecular mechanisms of dormancy

Seed dormancy has evolved multiple times during the history of plants, thus it is not surprising that different species have different physiological mechanisms for dormancy (Finch-Savage & Leubner-Metzger 2006). I will restrict my attention to a type of dormancy (physiological dormancy) that is present in the Brassicaceae and thus in *A. thaliana*. Seeds can have two types of dormancy, primary dormancy which is induced when the seeds mature on the mother plant and secondary dormancy which can be induced after seed dispersal.

There are two phases in plant seed development, embryo and endosperm development and seed maturation. First the embryo develops and grows, then growth is arrested and the seed enters a maturation phase. Dormancy starts developing during the early maturation phase and increases as maturation proceeds (Raz et al. 2001). During seed maturation, seeds also accumulate storage compounds and develop desiccation tolerance.

There is ample evidence that on the physiological level, seed dormancy is induced and maintained by the plant hormone Abscisic acid (ABA). ABA deficiency is associated with absence of primary dormancy in the mature seed while increased

ABA content in the mature seed can increase seed dormancy (Finch-Savage & Leubner-Metzger 2006). Another hormone group, gibberellins, is associated with germination. There is some evidence that in *A. thaliana* dormancy depends on the ratio of these two hormones, and gibberellins can only promote germination when ABA content is low (Ali-Rachedi et al. 2004).

In addition to intrinsic physiological factors, dormancy is also regulated by many environmental factors. During seed maturation temperature and photoperiod can influence the depth of dormancy in mature seeds (Roach & Wulff 1987; Munir et al. 2001; Donohue et al. 2005a). At least in *A. thaliana*, the effects of temperature seem to be larger than the effects of photoperiod (Donohue et al. 2005a), with colder temperatures during seed maturation inducing stronger dormancy. Temperature experienced by the seeds after dispersal is additional important factor. Germination of seeds of *A. thaliana* is promoted by colder temperatures and also increased light levels (Kugler 1951). Germination can also be affected by other environmental factors such as nitrate levels (Alboresi et al. 2005).

Some of the genes known to be involved in seed maturation and establisment of dormancy are known from mutant screens. Many of these have roles in hormone metabolism or transduction of these signals (Holdsworth et al. 2008). Genes such as *ABI3, FUS3, LEC1* and *LEC2* are involved in seed maturation and hormone signaling (Holdsworth et al. 2008). In addition some genes have been found that affect dormancy, but are not involved in hormone signaling, such as *HUB1* and *HUB2* (Leon-Kloosterziel et al. 1996), which are involved in chromatin remodeling (Liu et al. 2007), and *DOG1* (see section 1.4.2). Phytochromes are involved in germination responses to light and photoperiod (Shinomura 1997; Casal & Sanchez Rodolfo 1998). There is also some evidence that temperature can indirectly affect the light sensitivity of phytochrome mediated germination responses (Donohue et al. 2007b), with different phytochrome genes having different effects depending on the environment (Donohue et al. 2007a).

Dormancy can be released by a process called after-ripening. After-ripening happens during dry storage of dormant seeds at ambient temperatures. During after ripening the seeds become able to germinate at higher temperatures, their ABA content is decreased and they become more responsive to light (Finch-Savage & Leubner-Metzger 2006). After a certain period of after-ripening the seeds will lose their dormancy completely and will be able to germinate if conditions become permissive. In *A. thaliana*, there is genetic variation for the duration of after-ripening requirement (Evans & Ratcliffe 1972; Alonso-Blanco et al. 2003). The precise molecular mechanims of after-ripening are unknown. However, transcriptional changes are associated with different states of dormancy in the seeds (Cadman et al. 2006; Carrera et al. 2007). *DOG1* mRNA levels also seem to decrease during after-ripening (Finch-Savage et al. 2007).

1.4.2 *DOG1* – Seed dormancy QTL

Some QTL mapping experiments for seed dormancy have been done in *A. thaliana* and several QTLs have been identified (van Der Schaar et al. 1997; Alonso-Blanco et al. 2003; Clerkx et al. 2004; Bentsink et al. 2010). These were named the Delay Of Germinaton (*DOG*) loci. One of these QTLs, *DOG1*, has been identified by map based cloning from a cross between the accessions Landsberg (Ler) and Cape Verde Islands (Cvi) (Bentsink et al. 2006). *DOG1* was found to encode a protein of unknown molecular function. The Cvi allele of *DOG1* was shown to have a large effect on seed dormancy and increase it relative to the Ler allele. Since these two alleles differ from each other at the sequence level by several mutations the causal mutation could not be identified. There was some evidence that the two alleles were differentially expressed due to changes in *cis*-regulation, but there were also amino acid differences between the two alleles (Bentsink et al. 2006). Mutations that completely abolish the function of *DOG1* remove seed dormancy completely in both Ler and Cvi genetic backgrounds (Bentsink et al. 2006; Schwab 2008).

The genetic function of *DOG1* is known, but biochemical function is not. *DOG1* is expressed only at the seed stage in the embryo, and the peak of *DOG1* expression occurs during seed maturation. The level of gene expression (or protein) seems to be correlated with the degree of seed dormancy, as colder temparatures experienced by the mother plant increase dormancy as well as *DOG1* expression (Schwab 2008).

1.4.3 Ecological relevance of seed dormancy

Why should plants have a mechanism like seed dormancy that can prevent germination even if the present conditions would be permissive for growth? The answer is that seedling establishment must be considered in its ecological context. Plants have to time their germination to the proper season. If a plant flowers in the spring and sets seeds, often there could be a short period of time in the summer when there is enough moisture to induce germination in the absence of dormancy. However, these conditions could be very transient, such that, in a short period of time the drought is coming. Germinating at this time could be fatal, as young seedlings are very vulnerable and could not tolerate such drought.

Indeed, there is evidence that the timing of germination is under strong natural selection in many different species (Marks & Prince 1981; Kalisz 1986; Biere 1991; Gross & Smith 1991), including *A. thaliana* (Griffith et al. 2004; Donohue et al. 2005b). In addition Donohue et al. (2005b) showed that there was local adaptation for germination timing, as the optimal timing of germination differed in two different locations.

Germination timing can have also consequences on the subsequent life-history

expressed by the plant, as germinating at different times can bring the organism to different environments and this could influence selection on traits other than germination timing (Evans & Cabin 1995). This has been demonstrated in *A. thaliana*, where manipulating the timing of germination influenced the strength of natural selection on other traits (Donohue 2002). Moreover, recent evidence from *A. thaliana* indicates that germination timing can determine the subsequent life-history expressed by a plant (Wilczek et al. 2009), indicating that the timing of germination could be an even more important decision in determining plant life-history than the timing of flowering.

In addition of timing germination to the proper season within a year, seed dormancy can also postpone germination to subsequent years, creating a seed bank in the soil. Seed banks have some ecological and population genetic consequences. Plants can escape bad years by persisting in the soil as seeds (Venable & Lawlor 1980). It has been suggested that seed dormancy could be an adaptation to variable environments (Venable & Brown 1988), a kind of evolutionary bet-hedging (Slatkin 1974). Although seed dormancy as mechanism of bet-hedging is an attractive hypothesis, empirical evidence supporting it remains inconclusive (Evans & Dennehy 2005), yet there are some studies that suggests this the case (Evans et al. 2007; Simons 2009). It has been also suggested that seed dormancy could evolve as a response to avoid competition between siblings (Ellner 1986). However, some newer theoretical models have shown that dormancy will not eliminate competition if there is some kind of clonal reproduction, such as selfing (Kobayashi & Yamamura 2000). As selfing is rather common in plants, sib-competition seems an unlikely reason for the evolution of dormancy. Yet, in clonal populations, competition for habitat space can favour the evolution of dormancy, even in the absence of environmental variation (Satterthwaite 2010). This assumes that at least some seeds are retained in the parental neighbourhood and survival in the seed bank is high. In addition to plant ecology, having a seed bank has an influence on population genetics of plant populations, as a seed bank increases the effective population size of a plant population thus increasing the amount of genetic variation that can be maintained. If large seed banks can persist in the soil, effective population size can even exceed the census size (Vitalis et al. 2004). Genetic differentiation among populations is then also decreased.

Since germination timing has been shown to be under selection, the obvious question to ask is: are germination phenology in the field and physiological seed dormancy somehow related? Does measuring dormancy in the laboratory tell us something about the germination behaviour of that genotype in field? In a follow up experiment to Donohue et al. (2005b), where natural selection on germination timing was found, Huang et al. (2010) genotyped the RIL population used earlier and found that the QTLs for germination timing on the field and by extention

fitness colocalised on QTLs for seed dormancy measured in the laboratory. Of the dormancy QTLs, *DOG1* and *DOG6* localised at the same positions as QTLs for fitness. This experiment justifies the approach used in this study, in that measuring physiological seed dormancy is a relevant ecological trait.

It should be noted that other organims than plants can also have traits that ecologically resemble seed dormancy, called collectively germ banking (reviewed in Evans & Dennehy 2005). Many animals have resting or dormant stages, such as many insects and crustaceans that can have egg dormancy or diapause states. Thus many of the ecological consequences of seed dormancy apply to some non-plant organisms as well. There is also evidence that germination timing is also under strong selection in animals as well, for instance Koeller et al. (2009) detected local adaptation in shrimp hatching phenology in the Atlantic Ocean.

1.5 The study system – *Arabidopsis thaliana*

In this study, I studied the population genetics of an annual flowering plant, *Arabidopsis thaliana* (L.) Heyhn. (Brassicaceae). *Arabidopsis thaliana* is a well established model system for plant molecular biology. Its genome has been sequenced and there are many genetic tools available for it (Arabidopsis Genome Initiative 2000). It has a small genome, only 120–134 Mb with 5 chromosomes. Current estimate for the number of protein coding genes is 27 379 (TAIR9 release, www.arabidopsis.org). Generation time of *A. thaliana* can be short in the greenhouse, with some genotypes completing their life cycle within six weeks, although there is extensive variation within the species.

Arabidopsis thaliana is self-compatible and self-fertilisation is the predominant form of reproduction. Nevertheless, some outcrossing does occur in nature (Abbott & Gomes 1989; Hoffmann et al. 2003). Estimates of outcrossing rate from molecular markers range from 1 to 10 % (Abbott & Gomes 1989; Bergelson et al. 1998; Bakker et al. 2006; Pico et al. 2008; Bomblies et al. 2010; Platt et al. 2010) with some variation among populations. While this may seem rather low, there is sufficient outcrossing in *A. thaliana* to limit linkage disequilibrium (LD) to only 10 kb on large geographic scale (Kim et al. 2007). While LD is variable across the genome and extends longer within local populations, the population recombination rate in *A. thaliana* is higher than in humans (Kim et al. 2007).

Within species *A. thaliana* displays abudant natural genetic variation in several traits, such as flowering time, seed dormancy, seed size, tolerance to different abiotic factors and resistance to drought (Alonso-Blanco et al. 2009).

1.5.1 Ecology of *Arabidopsis thaliana*

Until recently there was not very much interest in the ecology of *A. thaliana*, this is now changing since the power of *A. thaliana* genetics can be combined with evolutionary studies (Mitchell-Olds & Schmitt 2006). *Arabidopsis thaliana* is a small weedy plant that is mostly found in disturbed or early successional habitats. It mostly prefers to grow on sandy soil and is absent from limestone-derived soils (Hoffmann 2002). In Western and Central Europe it is often found within habitats disturbed by humans, such as railroads, roadsides and fields. It cannot compete with other grasses very well (personal observations).

Arabidopsis thaliana has a wide range distribution being most abundant in Western and Central Europe and its continuous range extends all the way to Central Asia. Latitudinally, *A. thaliana* occurs from Southern-Europe up to Scandinavia (Hoffmann 2002). Some observations have been made also from Northern Africa, Japan and Korea, while the plant seems to have been introduced to North America (Hoffmann 2002). Since the last glaciation, *A. thaliana* has spread from its glacial refugia, it is now clear that one af these was in the Iberian Peninsula (Sharbel et al. 2000; Beck et al. 2008; Pico et al. 2008), while other refugia may have been in the Apennine Peninsula or the Balkans (Beck et al. 2008). Yet another possibility is a refugia in Central Asia (Sharbel et al. 2000) or perhaps China (He et al. 2007). However, sampling in these regions is very limited so final conclusions cannot be made.

Several large scale surveys of genetic diversity have been conducted in *A. thaliana* (Bergelson et al. 1998; Miyashita et al. 1999; Sharbel et al. 2000; Nordborg et al. 2005; Bakker et al. 2006; Pico et al. 2008; Platt et al. 2010). The main findings are that there is extensive genetic variation within *A. thaliana*. Population structure is strong, with some 35 – 38.5 % of variation exists between local populations within larger geographical regions and 33 – 56.7 % within populations (Bakker et al. 2006; Nordborg et al. 2005). Recently, Bomblies et al. (2010) showed that considerable genetic variation can be present within small geographical regions. Despite strong population structure most of the variation is shared among genotypes. Nordborg et al. (2005) found that the allele frequency spectrum did not fit to the standard neutral models of population genetics and that simple models of population growth did not improve the model fit. Thus, the accurate population genetic model describing the demographic history of *A. thaliana* is likely to be complex. There is isolation by distance that can be detected with a large number of markers (Pico et al. 2008; Platt et al. 2010).

There appear to be two major life-history strategies within *A. thaliana*, the winter annuals and the summer annuals (Laibach 1951; Nordborg & Bergelson 1999). The difference between the two types is that the winter annuals need to experience a period of cold temperatures before flowering can be effectively

induced. The difference in cold requrement is in many cases mediated by a large effect locus. Loss of function mutations in the gene *FRIGIDA* confer the plants the ability to flower early in the absence of cold treatment (Johanson et al. 2000). Generally, the winter annuals germinate in the autumn, overwinter as vegetative rosettes, flower and set seed in the spring. In the summer their seeds remain dormant (Baskin & Baskin 1972, 1983). Even though most plants germinate in the autumn, in some populations there have been reports of plants that germinate later in the following spring (Griffith et al. 2004; Montesinos et al. 2009). Most populations of *A. thaliana* seem to be winter annuals in Scandinavia (Nordborg & Bergelson 1999), Spain (Montesinos et al. 2009) and north America (Baskin & Baskin 1972). However, in central Europe populations with summer annuals are found (Le Corre 2005) and some populations in North America may have this life-history as well, as autumn flowering populations have been observed (Griffith et al. 2004). In addition, the distinction between winter annuals and summer annuals is not clear cut, since genotypes with the *fri*-mutation can also exhibit the winter annual life-cycle depending on the environmental conditions (Wilczek et al. 2009).

In *A. thaliana* there is some evidence that a seed bank is formed, as *A. thaliana* can be germinated from soil samples (Ratcliffe 1976; Lundemo et al. 2009). As *A. thaliana* nees light for germination, seeds that get buried too deep, will likely form a seed bank in the soil. Using genetic markers Lundemo et al. (2009) estimated that effective population size in *A. thaliana* is in fact larger than observed census sizes due to the fact that large numbers of individuals can lie dormant in the seed bank.

1.6 Study questions

Seed dormancy can have many ecological and genetic consequences. This study focuses on local adaptation brought about by predictable environmental differences among different populations. Some of the results raised new questions during the course of the study. I observed that genetic diversity (and thus mutation rate) was correlated with genetic differentiation between populations for the microsatellite markers, see section 3.1.1. This prompted me to study the properties of different estimators of genetic differentiation.

In this study I asked the following questions:

1. Which of the many different estimators of genetic differentiation are suited for studies of local adaptation?
    - As D and F'_{ST} were derived rather recently, their properties have not been studied extensively.
    - Can different types of markers be compared to each other?

- Which is the estimator that allows to compare different types of markers?

2. Is there local adaptation in seed dormancy in *A. thaliana*?

   - Before this study, little was known about the amount genetic variation present for seed dormancy in *A. thaliana*, especially within populations.
   - If local adaptation is present, at what geographical scale does it exist?
   - What are the environmental variables that mediate selection on seed dormancy?

3. If there is local adaptation what is the genetic basis of adaptation?

   - How many genes are involved and what kind of distribution of effects they have?
   - Are seed dormancy QTLs identified from one cross, such as *DOG1*, involved in local adaptation throughtout the distribution of *A. thaliana*?

4. What is the molecular basis of adaptation?

   - Are adaptive mutations affecting seed dormancy genes under local adaptation *cis*-regulatory changes or protein coding changes.

To answers these questions I undertook a population genetic study of seed dormancy variation in *A. thaliana*. Where I combined neutral markers, a candidate gene for seed dormancy and the phenotypic trait. To capture potentially locally adapted populations, the sampling spanned several geographic regions, where environmental conditions were different. To address genetic basis of adaptation I performed crosses between individuals to determine the number of QTL causing the differences between populations that are potentially locally adapted. The properties of different estimators of genetic differentiation were studied using computer simulations. I also discuss the results in the context of how population subdivision will affect adaptation.

Chapter 2

Methods

2.1 Sampling

For the population genetic study of local adaptation in seed dormancy 289 individuals from 41 populations were genotyped and phenotyped. A summary of the sampled populations is presented in Table 2.1. This sample will be subsequently referred to as the population sample. The sampling was hierarchical in that there were 4 larger geographical regions, Spain, France, Norway and Central Asia with 7, 15, 13 and 6 populations within them respectively. Detailed information about the sampled populations is found in the supplementary Table B.1. The three regions in Western Europe create a south-north cline. The Central Asian region is composed of populations from Kyrgyzstan and Tajikistan. Figure 2.1 shows a map with the sampled populations. The Spanish populations are described in Pico et al. (2008). The French populations were collected by and described by Le Corre (2005). The Norwegian populations were obtained from Odd-Arne Rognli through NARC (Norway). Populations from Central Asia were collected by Olivier Loudet and are described at http://www.inra.fr/vast/collections.htm.

2.2 Crosses

I constructed several mapping populations based on interesting *DOG1* alleles (see section 3.4). Crosses were made using standard methods. F_1-individuals were selfed to produce F_2-seeds, from these F_2-populations were grown. Leaves were collected for DNA extraction from F_2-individuals after bolting. Note that for seed dormancy, phenotyping is done on F_3-seeds, thus for mapping F_2-genotypes and F_3-phenotypes were used. The populations generated are given in Table 2.2.

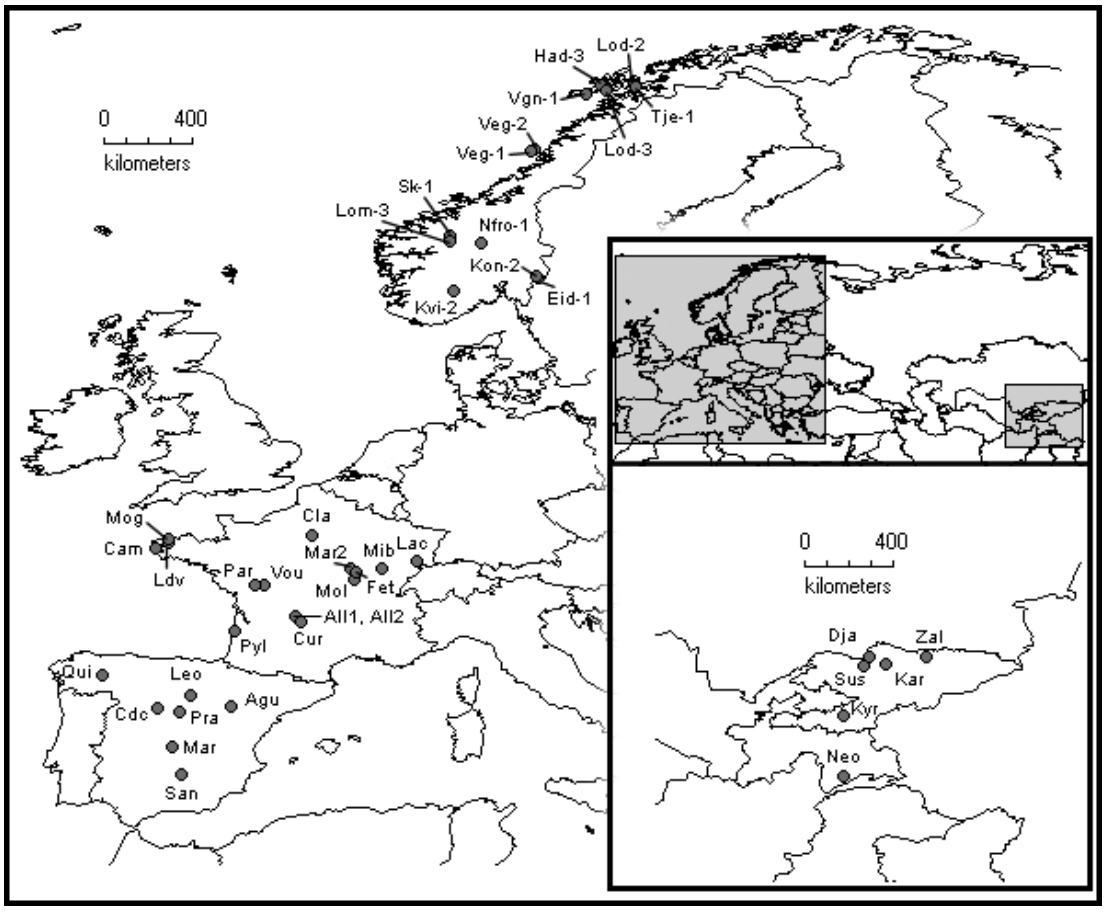

Figure 2.1: Map of the populations used in this study. Inset shows the Central Asian populations.

Table 2.1: Summary of the sampled populations

Region	Populations	Genotypes
Spain	7	70
France	15	109
Norway	13	64
Central Asia	6	46
4	41	289
Wageningen, the Netherlands	–	54
Total	–	343

2.3 Genotyping

Field collected plants were subjected to one or two generations of self-fertilisation in the greenhouse before DNA extraction. DNA was extracted from young leaves using a BioSprint 96 robot and the BioSprint 96 DNA Plant Kit (QIAGEN) according to manufacturer's instructions. The population samples (see section 2.1) were genotyped for 24 microsatellite markers and 149 SNP markers, in addition *DOG1* was genotyped in all these plants. The mapping populations (see section 2.2) were genotyped with the same 149 SNP markers and a SNP marker that distinguished between different *DOG1* alleles.

2.3.1 Microsatellites

The population samples were genotyped for 24 microsatellite loci, 20 of which are located in the nuclear genome and 4 in the chloroloplast genome. Primer sequences and a general description of the loci are given in supplementary Table B.2. Forward primers were labelled with a fluorescent dye and a PIG-tail sequence was added to reverse primers to avoid problems with +A activity of *Taq*-polymerase (Brownstein et al. 1996). Microsatellites were amplified using standard PCR methods and allele sizes were determined using capillary electrophoresis on a ABI3130*xl* Genetic Analyzer (Applied Biosystems). The size standard used was GeneScan 600 LIZ. Genotype calls and PCR product sizes were determined with the GeneMapper 3.7 software (Applied Biosystems). To determine the actual number of repeats in each allele, the accession Col-0 was also genotyped for each locus for reference. Using the Col-0 PCR product size and the Col-0 genome sequence the number of repeats was deduced for each allele. The Spanish accessions had been previously genotyped for some of the loci used here, as described in Pico et al. (2008). I verified that

Table 2.2: F_2-populations generated for this study. Parents of the crosses with their *DOG1* haplotypes are shown. For each cross the size of the F_2-population is given.

Parent	Region	*DOG1* haplotype
Trs-1	Wageningen	21
Trs-2	Wageningen	4
Cam-4	France	15
Fet-6	France	5
All2-1	France	1
Kon-2-2	Norway	19
Nfro-1-4	Norway	18
Cross	N	
Trs-1 x Trs-2	127	
Cam-4 x Fet-6	126	
Kon-2-2 x Fet-6	121	
All2-1 x Cam-4	145	
All2-1 x Fet-6	133	
Kon-2-2 x Nfro-1-4	122	

the allele sizes corresponded to the previously reported sizes by re-genotyping a subsample at selected alleles.

2.3.2 SNP markers

The population samples were also genotyped for 149 SNP markers developed by Warthmann et al. (2007), by Sequenom, inc. (San Diego, CA.). Out of 149 SNP markers, 137 had good quality data and were polymorphic in the whole sample. Thus only these markers were used in the population genetic analyses. The 149 SNP markers were also used to genotype the F_2-populations at the University of Chicago DNA sequencing facility (Chicago, IL.).

To genotype SNP markers around *DOG1*, I first designed primers to amplify 9 loci around *DOG1* to discover SNP markers. Primer sequences and positions are presented in supplementary Table B.3. The SNP discovery panel was composed of 16 different genotypes from different regions, 2 from Wageningen, 4 from Spain, 3 from France, 3 from Norway and 4 from Central Asia. DNA sequencing was performed as in section 2.3.3.

From each of the 9 sequenced fragments, 4 to 1 SNPs were chosen for genotyping (Supplementary Table B.3) and pyrosequencing assays were designed for these SNPs with the Assay Design Software 1.0.6 (Qiagen). The SNPs were genotyped using pyrosequencing (Fakhrai-Rad et al. 2002), with the PSQ 96MA Pyrosequencing system (Qiagen) according to manufacturer's instructions. Briefly, the target sequence was amplified by PCR with a biotin labelled primer. The PCR products were immobilised to streptavidin coated beads and the DNA strands were separated. An internal sequencing primer was added to the single stranded PCR product. In the pyrosequencing reaction nucleotides are added iteratively. When a nucleotide is incorporated by DNA polymerase, pyrophosphate is released, which leads via a series of enzymatic reactions to the emission of light. The light signal is then detected by a camera. Because the added nucleotides are known, the sequence of the template can be determined.

Primer sequences for the SNP markers are given in supplementary Table B.4. To limit the number of biotinylated primers needed, a universal primer method was used (Aydin et al. 2006). A universal sequence was added to the 5' end of the specific primers. In the PCR reaction four primers were used, the specific primers and the universal primers with appropriate universal primer labelled with biotin. For some assays the four primer reaction did not work efficiently, so two separate PCR reactions had to be performed.

To genotype *DOG1* in the QTL-analyses pyrosequencing assays were designed to genotype SNP markers in *DOG1* that distinguished between the different haplotypes (see section 3.4). Primers for these assays are given in supplementary Table B.4. Assay D_a1 distinguished between groups of haplotypes (15, 5), 1 and 4; assay D_a2 distinguished between groups of haplotypes 15 and (1, 5, 4); assay D_a3 between groups of haplotypes (1, 5, 15) and 4; and assay D_a4 between haplotypes 5, 18 and 19.

2.3.3 DNA sequencing

Based on preliminary results, the first exon of *DOG1* is the most polymorphic region of the gene (M. Debieu, unpublished results). Thus, I designed primers to amplify the first exon of *DOG1* and sequenced it from the population sample. The primers used were D1E1 - 5'-AAA CAC AAA CAC GCA AAC CA and i1re - 5'-GCC GCA CCG TAC TGA CTA CC. PCR and Sanger sequencing was performed using standard protocols. Only the primer i1re was used in the sequencing since there is a poly-A stretch at the end of the promoter of *DOG1*. For some genotypes PCR products were cloned to a pCR®4-TOPO Vector using TOPO TA Cloning Kit (Invitrogen). Standard protocols for cloning were used. Electrophenograms were inspected for errors and sequences could be aligned unambiguously using BioEdit 7.0.5.3 (Hall 1999).

While sequencing *DOG1* from the population sample, it became apparent that some genotypes had a large length polymorphism in the first intron. The primers ee1f - 5'-CGA CGG CTA CGA ATC TTC AG and i1re were used to amplify this large length polymorphism. The PCR products were separated using a 1 % agarose gel and a band was cut away from the gel and purified using QIAquick Gel Purification Kit (QIAGEN) according to manufacturer's instructions. Purified PCR products were cloned into the vector described above and sequenced.

2.3.4 Other markers

It was revealed that there was an insertion in the first intron of some genotypes (see section 3.4.1). The presence or absence of this insertion was genotyped using primers ee1f and i1re, the length polymorphism could be revealed by resolving the PCR products on a 3 % agarose gel. This assay was also used to genotype the *DOG1* haplotype in the cross between Trs-1 and Trs-2.

2.4 Phenotyping

For the quantitative genetic experiments all lines were first multiplied in the greenhouse under the same environmental condititions to remove or equalise any possible maternal effects. All plants were grown in the same climatised greenhouse with a temperature +20 °C during the day and +18 °C during the night. Natural light was supplemented with lamps to reach a photoperiod of 16h of light when necessary. For the common garden experiment, the plants were grown in standard soil for *Arabidopsis* in round pots with a diameter of 6 cm with one plant in each pot.

2.4.1 Common garden experiment

The common garden experiment was started in the fall of 2007. Since *A. thaliana* is mostly self fertilising, genetically identical replicates for each genotype are obtained from selfed progeny of that genotype. In the experiment each genotype was replicated three times. Replicates were randomised within blocks which corresponded to different positions in the greenhouse. From a pilot experiment it was known that there were large differences in flowering time between the genotypes. Since I was interested in measuring seed dormancy I wanted the seeds for all genotypes to mature in similar environmental conditions. Therefore, in order to synchronise flowering time I planted the genotypes in three different groups. The seeds were stratified (cold treatment +4 °C, of wet seeds) in the dark for four days to induce germination. After this they were planted on soil and moved to the greenhouse.

After 14 days of growing the plants were moved to climate chamber for vernalisation (cold treatment of the rosette), to make flowering possible for all genotypes, with a temperature of +4 °C and short day (8h of light) photoperiod for 28 days. Subsequently the plants were moved back to the greenhouse. As a consequence of planting at different times and rosette vernalisation I was able to synchronise flowering and the seeds matured during March – April 2008 for most genotypes.

Seed dormancy was measured for each replicate as described in section 2.4.3. Flowering time was also recorded using three different measurements, time (days) from sowing to bolting (DTB), plants were considered to bolt when flowering stem reached height of 5 cm, the time when first flower was opened (DTF) and the number of rosette leaves leaves at bolting. These three measures are highly correlated. I calculated how long was the development time for the seeds in the experiment from the date of first flower opened to the date of silique harvest, referred from here on as maturation time (MTIME). Due to the fact that all plants could not be harvested at the same time, this permitted me to investigate whether slight differences in maturation time had an effect on seed dormancy.

2.4.2 QTL mapping experiments

The mapping populations for QTL mapping experiments were grown in the same greenhouse as the common garden experiment, with the same vernalisation treatment. Except for crosses Trs-1 x Trs-2 and All2-1 x Cam-4 which were not vernalised. The mapping populations were grown in 96-hole trays. The genotyping was done on F_2-lines, while the phenotyping was done for F_3-lines as dormancy can only be measured from the seeds of the mother plant.

For the cross Trs-1 x Trs-2, the parental lines behaved in an unexpected way in the first experiment with F_2 lines (there were no differences between the parent, even though in previous experiments the parents were different). Therefore, a second experiment was done, for which F_2-lines that were homozygous in the region of *DOG1* were selected. In the second experiment 19 F_3-lines of both genotypes, 38 lines in total, were used. In this kind of experiment the two different *DOG1* genotypes are compared to each other, while rest of the background is randomised. The F_3-lines were grown in two environments, in the greenhouse and in a growth bank with temperature set to +16 °C and long day (16h light) photoperiod. This is referred to as the cold environment. Thus dormancy was measured from F_4-seeds. In the second experiment for the greenhouse conditions, the dormancy of the parents returned to a level seen in previous experiments.

2.4.3 Seed dormancy measurements

Ripening of the siliques (structures in which seeds mature in Brassicaceae) was assessed visually by observing a colour change to brown. *A. thaliana* produces siliques over a long period of time, and siliques were harvested when there were enough ripened siliques on the plant (usually siliques were harvested from the main stem). On the day the seeds were harvested from a given replicate, germination experiment was immediately started for those seeds.

To measure the physiological seed dormancy I performed a germination experiment for each seed batch (replicate) following Alonso-Blanco et al. (2003). I measured the ability of the seeds to germinate in a time course experiment, for each time point, a sample of approximately 50 - 100 seeds was taken and put on a small Petri dish with filter paper and 700 µl of water was added. Then the Petri dishes were transferred to a growth cabinet with a +25 °C during the day and +20 °C during the night. The photoperiod was 12h of light. After one week, the number of germinated and dormant seeds was counted using a preparation microscope; seeds were counted as germinated when the root tip had protruded the seed coat. After-ripening occurred at room temperature and seeds were stored in paper bags. For each seed batch germination tests were performed immediately after harvest (0 weeks) and then subsequently 1, 2, 4, 8, 16, 24, 32, 40, and 52 weeks after harvest for the common garden experiment (see 2.4.1). For the QTL mapping experiments germination test were performed in the same manner, except a data point was added on 12 weeks, and the 1 week data point was dropped for some crosses and a 3 week data point added for the cross Kon-2-2 x Nfro-1-4. For the QTL mapping experiments germination test were stopped after 16 weeks as nearly complete germination was observed for all lines. When a seed batch was germinating at 100 % in two consecutive tests it was considered to have lost dormancy and subsequent data points were imputed as 100 % for that batch.

For the common garden experiment the dormancy measurements were stopped after 52 weeks. A small number of seed batches had not reached 100 % germination at this time, therefore I performed a viability test for these seed batches following Cadman et al. (2006). The procedure was the same as in the germination test but instead of pure water, a solution with 100 µmol/l Gibberellin ($GA_{4/7}$) (Duchefa, Haarlem) and 38 µmol/l Fluridone (Sigma-Aldrich, Seelze) was added. The chemicals were initially dissolved in a small volume of ethanol. Gibberellin is a plant hormone that promotes germination and Fluridone is a drug that blocks the synthesis of abscisic acid, a hormone that maintains seed dormancy. After this treatment virtually all seed batches germinated to 100 %, there was only one replicate that germinated only to 90 %, therefore seed viability does not influence the results in any way.

2.5 Statistical analysis

All statistical analyses were done using the R statistical package (R Development Core Team 2006) unless otherwise stated. Methods not implemented in R-packages were implemented via scripts that are available upon request.

2.5.1 Statistical analysis of seed dormancy data

If seed dormancy would be measured as proportion of germination at a given time point, this would yield a strongly bimodal distribution that is constrained between 0 and 1, and such a situation would be highly unsuitable for many statistical models. Thus, to obtain a measure of seed dormancy for a given replicate I fitted a binomial regression through the germination data for each replicate using a logit link function, $\ell(\rho) = \ln(\rho/(1-\rho))$ (Venables & Ripley 2002). The response variable was the number of seeds germinated over the total number of seeds in one germination test. From the fitted function I calculated the time, $\mathrm{D}(\rho)$, for which the probability of germination, ρ, was 0.25, 0.5 or 0.75. This can be obtained from

$$\mathrm{D}(\rho) = \frac{\ell(\rho) - \beta_0}{\beta_1} \tag{2.1}$$

where $\ell(\rho)$ is the value of the link function at ρ, β_0 is the slope of the regression function and β_1 the intercept. This is a measure of the time of dry storage required to reach a given probability of germination, the unit is weeks of dry storage (WODS). A similar measure was used before by Alonso-Blanco et al. (2003) to measure seed dormancy. Using a probit link function produces nearly identical results. Figure 2.2 illustrates the seed dormancy data and how a measure of seed dormancy is obtained. Note that for some replicates this method can give slightly negative values for dormancy if the germination in the first data point is over ρ. While this is not ideal, I did not set the negative values to zero, as this would truncate the distribution of dormancy values. Negative values for some replicates can be seen to reflect very low dormancy of some genotypes. Moreover, setting negative values to zero does not affect the biological conclusions obtained.

For some replicates the fit to a binomial regression model was not ideal, therefore I also calculated dormancy with an iterative method which searches the data for points that define the interval where probability of germination reaches 0.25, 0.5 or 0.75. Then, assuming that release of dormancy is linear in this interval, the algorithm fits a line through these points and the time required to reach a given germination fraction is solved from this equation.

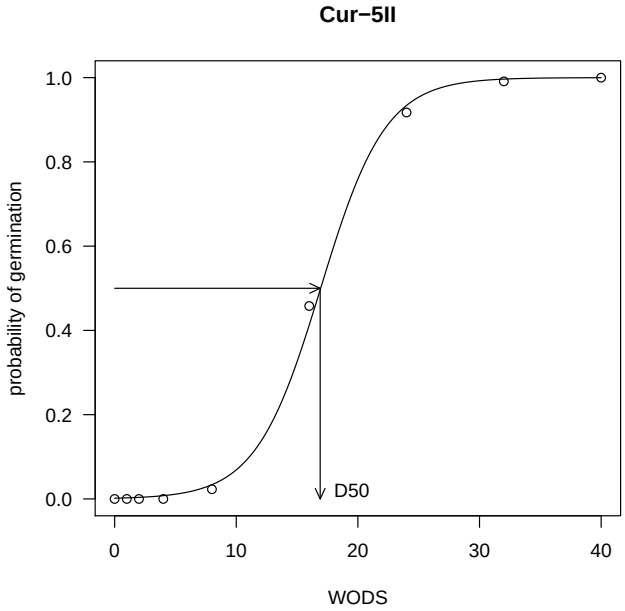

Figure 2.2: Illustration of seed dormancy data for replicate II of genotype Cur-5. The arrows illustrate how a measure of seed dormancy (D50) is obtained.

2.5.2 Statistical analysis of the common garden experiment

Measures of seed dormancy were calculated for each replicate, referred to as D25, D50 and D75 for time required to reach 0.25, 0.5 and 0.75 germination respectively, using both binomial regression and linear methods (see section 2.5.1). The three different time points were calculated, to investigate if they reflect different aspects of dormancy. To calculate genotype means I used a linear model

$$y_{ijk} = \mu + g_i + b_j + e_{ijk} \tag{2.2}$$

where y_{ijk} is the phenotypic observation of the kth replicate of the ith genotype in block j, μ is the overall mean, g_i is the genotypic effect of the ith genotype, b_j is the block effect of the jth block and e_{ijk} is the residual. Genotypic means were obtained from the term $\mu + g_i$. In this way possible block effects are taken into account when calculating genotypic means. In general, block effects were non-existent or very small. For flowering time only one block was different from the rest after adjusting for multiple comparisons (TukeyHSD), for dormancy also only one block (albeit a different one) remained significantly different from the rest. In general, I conclude that positional effects in the greenhouse are minimal and do not affect any biological conclusions of this study.

To investigate differences between populations and regions I used a linear model

$$y_{ijk} = \mu + r_i + p_{ij} + e_{ijk} \tag{2.3}$$

where y_{ijk} is the mean phenotype of the kth genotype in the jth population within the ith region, μ is overall mean, r_i is the effect of ith region, p_{ij} is the effect of the jth population nested within the ith region and e_{ijk} is the residual.

2.5.3 Analysis of climate data

To find possible causes for selection, I examined if the trait values of the populations are related to any environmental variables. I used the program DIVA-GIS 5.2.0.2 (Hijmans et al. 2001) (available at www.diva-gis.org) in combination with the 2.5 arc-minute resolution current global climate environment data (Hijmans et al. 2005) (available at www.worldclim.org) to extract climatic data for our populations. This data is an average of the conditions in the past 50 years. Then I built a linear model that explains variation in plant traits by climatic conditions, population means were used in this analysis.

2.5.4 Quantitative genetics

Heritability, which measures what proportion of observed variance is genetic variation, was estimated as

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} \tag{2.4}$$

Q_{ST} measures how quantitative genetic variation is partitioned between populations and is estimated as

$$Q_{ST} = \frac{(1 + F_{IS})\sigma_{GB}^2}{(1 + F_{IS})\sigma_{GB}^2 + 2\sigma_{GW}^2} \quad \text{(Bonnin et al. 1996).} \tag{2.5}$$

Assuming complete selfing (i.e. $F_{IS} = 1$), this reduces to

$$Q_{ST} = \frac{\sigma_{GB}^2}{\sigma_{GB}^2 + \sigma_{GW}^2} \tag{2.6}$$

where σ_{GB}^2 is genetic variation between populations and σ_{GW}^2 is genetic variation within populations. To estimate Q_{ST} the variance components from model 2.8 were substituted to equation 2.6.

Since *A. thaliana* is mostly self fertilising, genetic variance components can be estimated in a straight forward manner from the common garden experiment (see section 2.4.1). Assuming complete selfing, variation between replicates within genotypes allows estimating σ_E^2, the environmental variance component, and variation

between genotypes allows estimating σ^2_G, the genetic variance component. Dominance variation is not defined, because all lines were assumed to be homozygous. I used two different methods to estimate the variance components, a linear mixed effects model, from which variance components were estimated using REML (Venables & Ripley 2002) or a Bayesian method of estimating variance components implemented in WinBUGS 1.4.3 (Lunn et al. 2000).

Variance components were estimated from a model

$$y_{ijk} = \mu + b_i + g_j + e_{ijk} \tag{2.7}$$

where y_{ijk} is the phenotypic observation of the kth replicate of the jth genotype in the ith block, b_i is the block effect for the ith block, g_j is the genotypic effect for jth genotype. Blocks were included as fixed effects and genotypes as random effects. For Q_{ST} the model was extended such that

$$y_{ijkl} = \mu + b_i + p_j + g_{jk} + e_{ijkl} \tag{2.8}$$

where p_j is the population effect, other terms are as above, for block i, population j, genotype k nested within population and replicate l nested within genotype. Blocks are included as fixed effects and population and genotype as random effects. Specification of the WinBUGS models followed the same logic and was done following O'Hara & Merilä (2005) (see appendix D for details of the implementation).

2.5.5 Genetic diversity and population structure

Measures of genetic diversity, Nei's gene diversity (H_S) and allelic richness (AR) were calculated using FSTAT 2.9.3 (Goudet 2001). Allelic richness is a measure of the number of alleles independent of sample size. To compare genetic diversity between groups of populations a permutation test was used, permuting populations within regions as implemented in FSTAT. The microsatellite population mutation rate, Θ, is the product of effective population size and mutation rate at a locus. It was calculated following Kimmel et al. (1998). Θ was estimated from

$$\widehat{\Theta} = (1/\widehat{P_0}^2 - 1)/2 \ , \tag{2.9}$$

where

$$\widehat{P_0} = \frac{2N \sum_{i=1}^{A} p_i^2 - 1}{2N - 1} \ , \tag{2.10}$$

A is the number of alleles at a locus, p_i is the frequency of allele i and N is the number of individuals. The performance of this summary statistic based method

has been shown to be comparable to likelihood-based methods (RoyChoudhury & Stephens 2007). Θ was calculated for each locus within each region.

To check whether $\Phi_{ST} > F_{ST}$ for the microsatellite loci, I used a permutation test (Hardy et al. 2003) to assess the difference between Φ_{ST} and F_{ST}. The test was implemented in the program SPAGeDi 1.2 (Hardy & Vekemans 2002). The test works by permuting microsatellite allele sizes among the allelic states to test if stepwise mutations contribute to genetic differentiation.

To investigate population structure in the sample without regard to the sampling scheme, the data was analysed using a principle component analysis (PCA). PCA for genetic markers was implemented in the R-package *adegenet* (Jombart 2008). PCA does not make any assumptions of Hardy-Weinberg equilibrium or linkage equilibrium. Both microsatellite and SNP markers were used in the analysis. The PCA results were also used in correcting for population structure in the candidate gene association analysis (see section 2.5.8). In a sample with all regions and some accessions from Wageningen, 4 components were selected for population structure correction. For samples within Spain, France, Norway and Central Asia, 7, 7, 10 and 6 components respectively, were selected for the correction.

2.5.6 F-statistics

F_{ST} was estimated according to Weir & Cockerham (1984) using the R-package *hierfstat* (Goudet 2005). All other genetic differentiation methods were implemented via R-scripts. The standardised genetic differentiation measure, F'_{ST} (Hedrick 2005), was estimated using the maximised variance component method of Meirmans (2006). In order for the distance between microsatellite alleles or sequence haplotypes to be taken into account (Slatkin 1995), I estimated Φ_{ST} following Michalakis & Excoffier (1996). In some instances differentiation indices between regions were calculated in a hierarchical setting, taking into account the partition of variation between populations within regions (Excoffier 2007). Confidence intervals for different measures of genetic differentiation were generated by bootstrapping over loci. The measure of allelic differentiation, D (Jost 2008), was also estimated. Equations for calculating the different estimators are presented in the appendix C.

To compare genetic differentiation in *DOG1* and phenotypic traits to neutral markers I calculated F_{ST} for SNP markers using the method of Weir & Cockerham (1984), for microsatellite markers and sequence haplotypes I used Φ_{ST}, this way the different mutation rate of different markers can be taken into account and different types of markers can be compared to each other (see section 3.1.2). To compare F_{ST} of *DOG1* to neutral markers I used the empirical distribution of neutral markers and compared *DOG1* to the quantiles of this distribution. The reason for using empirical distribution was that the mean neutral F_{ST} was high

in this dataset. The validity of the Lewontin-Krakauer method for approximating the distribution of F_{ST} is not known, when F_{ST} is high (Whitlock 2008).

To measure possible isolation by distance, a matrix of pairwise F_{ST} values between populations was correlated to a matrix of pairwise geographic distances[1]. The matrices are not fully independent, since they include the same set of populations. Therefore, Mantel-tests were used to assess the statistical significance of the correlations, as implemented in the R-package *vegan* (Oksanen et al. 2007).

For the computer simulations (see section 2.6), I calculated the expected F_{ST} to which the estimated values can be compared. In an island model, the time to coalescence of a pair of alleles within a population is $\bar{t}_0 = 2NP$ and time to coalescence of a pair of alleles from different populations is $\bar{t}_1 = 2NP + (P - 1)/2m$ (Slatkin 1991), where m is the migration rate and N is the population size. Substituting these into equation 1.3 yields

$$F_{ST} = \frac{1}{1 + 4NmP/(P-1)} \tag{2.11}$$

now the parameter values used in the simulations can be substituted to this equation to obtain expected F_{ST}.

2.5.7 Sequence analysis of *DOG1*

I constructed a haplotype network of the first exon and an insertion polymorphism in the first intron of *DOG1* using the program TCS v 1.21 (Clement et al. 2000). TCS implements a maximum parsimony method to construct the evolutionary relationships between the haplotypes. For all analyses that required an outgroup the *Arabidopsis lyrata* sequence of the first exon of *DOG1* was used. For estimation of genetic differentiation Φ_{ST} was used (see section 2.5.6), pairwise distances between the haplotypes (these are equivalent to Hamming distances) were used in the calculation. Sequence diversity indices were calculated using DnaSP v. 4.10.4 (Rozas et al. 2003).

As the first intron is sometimes involved in regulation of gene expression (Rose et al. 2008), I used the bioinformatic tool IMETER (Rose et al. 2008) to analyse is the insertion in intron 1 predicted to influence gene expression. IMETER gives a score for each intron based on characteristics that have been observed to affect gene expression, such as sequence composition and intron length (Rose et al. 2008). The different intron 1 sequences for *DOG1* were obtained from the GeneBank, these were the ones deposited by Bentsink et al. (2006).

[1]R-script to calculate geographic distances between the populations was kindly provided by Sylvain Antoniazza

2.5.8 Candidate gene association

I tested whether genetic variation in *DOG1* is associated with genetic variation in seed dormancy. Population structure in our sample is strong, F_{ST} between populations is usually high (Supplementary Table B.6). In order to avoid spurious marker-phenotype associations that arise due to the fact that some alleles can be associated with certain populations, population structure has to be taken into account. I performed an association test using mixed model association following Yu et al. (2006) with the PK_T method of Stich et al. (2008). The model takes into account population structure and kinship of individuals within populations. The model for association was

$$y = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e} \tag{2.12}$$

where y is a vector of genotypic means, $\mathbf{X}$ is an incidence matrix for fixed effects, that includes marker effect and linear independent columns describing population structure derived from principle component analysis of marker data (see page 31). β is a vector of fixed effects, $\mathbf{Z}$ is an incidence matrix for random effects, u is a vector of genetic effects and $\mathbf{e}$ is the residual. Variance of genetic effects is assumed to be $\mathrm{Var}(u) = 2\mathbf{K_T}\sigma^2_G$, where $\mathbf{K_T}$ is a $N \times N$ matrix of kinship coefficients that defines the genetic covariance between all entries. Variance of residuals is assumed to be $\mathrm{Var}(\mathbf{e}) = \sigma^2_E\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This type of model, called the animal model, is very common in quantitative genetics and has many different applications (Lynch & Walsh 1998).

Kinship coefficients were calculated as in Stich et al. (2008), following Bernardo (1993) and Lynch (1988a). The matrix $\mathbf{K_T}$ is calculated as

$$K_{Tij} = \frac{S_{ij} - 1}{1 - T} + 1 \tag{2.13}$$

where S_{ij} is the proportion of marker loci that share alleles between genotypes i and j and T is the probability that an allele from genotype i and an allele from genotype j are indentical by state, given that they are not identical by descent. Negative kinship values were set to 0. I used a single value of T for all genotypes. Because the value of T is unknown, I calculated the kinship matrix for $T = \{0, 0.025, ..., 0.975\}$ to obtain an estimate of T that most effectively corrects for population structure, following Stich et al. (2008).

Assuming that the neutral markers used in this study are not causally linked to the phenotype, it is expected that the distribution of p-values is uniform from an association test using the neutral markers given that the method adheres to the nominal α-level. In this case the expected p-values can be calculated from $r(x_i)/L$, where $r(x_i)$ is the rank of the observed p-value for the ith marker and L is the number of loci. If p-values are uniformly distributed, a diagonal line is

observed in a plot of observed vs. expected p-values. I calculated mean squared difference (MSD) between observed and expected p-values for the neutral markers as measure of deviation from the uniform distribution. I used the MSD value as a criterion for selecting the value for T, following Stich et al. (2008). Correcting for population structure is important in our sample, without a correction many markers would be associated with the phenotype, even though they are likely not causally associated (Figure 2.3).

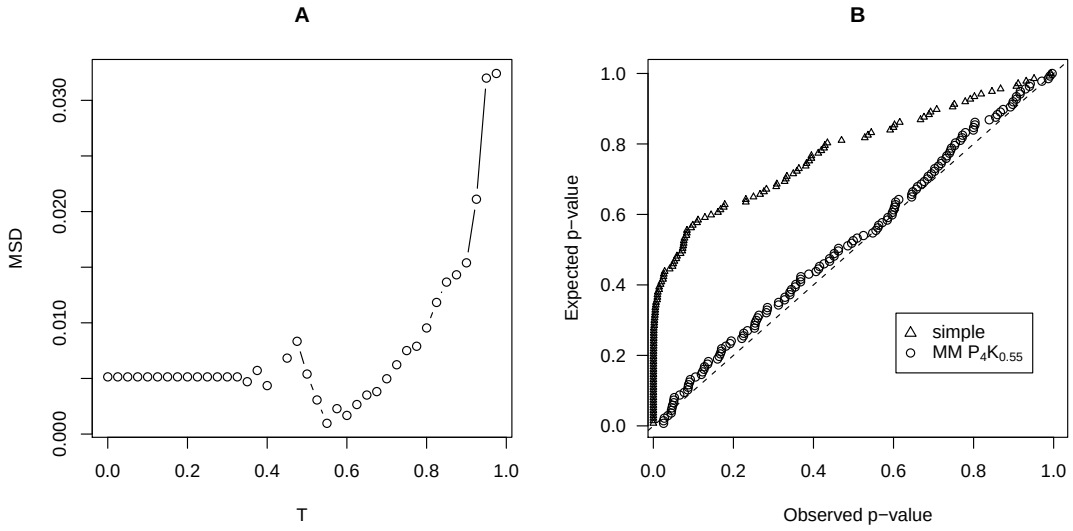

Figure 2.3: An example of population structure correction. In panel A) MSD values calculated for different $\mathbf{K_T}$ matrices. The value of T that gave the lowest MSD value was used in the actual association study. Panel B) a plot of expected vs. observed p-values, triangles are p-values for simply testing association of each SNP marker to the phenotype without any correction and circles are p-values using the mixed model correction. If p-values follow the uniform distribution they should lie on the dashed line.

In order to increase statistical power, I included a sample of accessions from Wageningen in addition to the population sample when performing the association test for *DOG1* over all genotypes. This increased the sample to 343 genotypes (Table 2.1). I also tested for association within geographic regions of the population sample. For each new sample, T was estimated again. The value of T that gave the matrix $\mathbf{K_T}$ with the lowest MSD value was selected for the association test of *DOG1* (Figure 2.3). At first, the SNP markers were used to determine the optimal value of T for each trait and sample (Supplementary Table B.8). When using the SNP markers, the model 2.12 was implemented using the R-package *EMMA*

(Kang et al. 2008). Second, after determining the optimal T value, the association test for *DOG1* using the model 2.12 was done using the program TASSEL 2.0.1 (Bradbury et al. 2007). Switching between the different programs was done because *EMMA* did not support multiple alleles. Sequence haplotypes of the first exon of *DOG1* were used as different alleles in the association study. Since there were multiple tests done due to multiple alleles, I corrected for multiple testing using the Bonferroni-Holm correction (Holm 1979).

2.5.9 QTL mapping

To check co-segregation of *DOG1* with dormancy in the F_2-populations, the populations were analysed using a linear model

$$y_{ij} = \mu + g_i + e_{ij} \tag{2.14}$$

where y_{ij} is the phenotypic observation of the jth line in genotypic class i, g_i is the effect of ith *DOG1* genotypic class and e_{ij} is the residual. Following Lynch & Walsh (1998), I denote the genotypic values of the genotypes D_1D_1, D_1D_2 and D_2D_2 as 0, $(1+k)a$ and $2a$, respectively. Taking the estimates of the different genotypes from the linear model, the effect of allele D_2 is obtained from $a = (D_2D_2 - D_1D_1)/2$ and the dominance coefficient from $k = ((D_1D_2 - D_1D_1)/a) - 1$. Interval estimates for a and k were obtained by fitting the model 2.14 in a Bayesian context (see appendix D for details).

All genetic mapping and QTL-analyses were implemented using the R package *R/qtl* (Broman et al. 2003). For the QTL-mapping, genetic maps were constructed for the F_2-populations based on the markers that were polymorphic between the two parents, out of the 149 genotyped SNPs. Marker order was based on the *A. thaliana* physical map and genetic distances were calculated using the methods outlined in Lander & Green (1987), the Kosambi mapping function was used.

For QTL analyses, I used the Haley-Knott regression (Haley & Knott 1992) and its extended version (Feenstra et al. 2006) to scan for QTLs in the genome. Evidence for the precence of a QTL was evaluated using a LOD (logarithm of odds) score, which compares the likelihood of model with and without a QTL. Significance thresholds for LOD scores were obtained using the permutation method of Churchill & Doerge (1994), which permutes trait values among the genotypes, 1000 permutations were performed. First the genome was scanned for QTL, and the found large effect QTLs were selected as covariates and the genome was scanned again for additional QTL. After this a multiple QTL model was fitted to the data to estimate QTL effects and refine their map positions. Only those QTLs that exceeded the LOD threshold in the multiple QTL model were retained.

2.6 Computer simulations

2.6.1 Effect of mutation rate on estimators of genetic differentiation

In order to investigate the behaviour of F_{ST}, F'_{ST}, Φ_{ST} and D under different mutation rates, computer simulations using EasyPop 1.8 (Balloux 2001) were performed. The simulation scheme was set to 10 populations with 500 individuals each, 20 freely recombining loci and random mating hermaphrodites. Populations followed an island model of migration. Migration rates (probability that a given individual will migrate in each generation), m ranged from 0.1 to 0.00001 and mutation rates (probability that a given allele will mutate in each generation), μ from 0.00001 to 0.01. In order to simulate microsatellite loci, I first examined a pure single step mutation model. Then I relaxed this assumption by using a mixed mutation model in which the loci followed a single step mutation model but with the probability of 0.2 to mutate to any state. The number of possible allelic states was set to 30. The effect of self-fertilisation was examined by doing simulations with proportion of self-fertilisation set to 0.9. Simulations were run for 2000 generations. In the end to simulate realistic sampling situation, 30 individuals were sampled from each population for parameter estimation. Each simulation was repeated 5 times for a given set of parameter values. For each simulated dataset I calculated genetic differentiation statistics, gene diversity (H_S) and microsatellite population mutation rate (Θ).

2.6.2 Effect of mutation rate on estimator of quantitative genetic differentiation

I examined how the mutation rate at underlying QTL affects Q_{ST}. I used quantiNEMO (Neuenschwander et al. 2008), with the same settings as described in 2.6.1 for neutral markers with the following exceptions: the number of QTL underlying the variation in the quantitative trait was 10 and there were 21 possible allelic states for each QTL. I used the random mutation model in quantiNEMO for the QTL alleles, in this model, allelic effects are drawn from a normal distribution. I also ran the simulations using the incremental mutation model, where the allelic effect of a new mutation resembles its ancestor. Variance of allelic effects was set to 0.1. The quantitative trait simulated was neutral and did not have any effect on fitness. The simulation was started at a state where all loci were monomorphic, the number of generations was 4000. The time to reach equilibrium was longer for low migration and mutation rates and in these cases number of generations was 6000. This is longer than for neutral markers because simulations had to be started from monomorphic state; otherwise distribution of allelic effects becomes

unrealistic. Variance components for Q_{ST} were estimated from genotypic values, which are returned by quantiNEMO as output, using R-scripts that I wrote. The statistical model was a mixed-effect model with populations as random factors; REML-estimates of variance components were used. This was done in order to calculate Q_{ST} also in the presence of self-fertilisation, which quantiNEMO does not calculate as a standard output. Q_{ST} was estimated from the equation 2.5. This method of estimating Q_{ST} gives the same results as the standard output of quantiNEMO when mating is random (results not shown). For each parameter set, 50 replicates of simulations with a single quantitative trait were run.

2.6.3 Comparing different marker types

I also performed coalescent simulations to investigate the effect of different marker types on F_{ST} calculations. I investigated sequence haplotypes (these would be derived by sequencing a number of loci from many individuals), independent single SNP markers and microsatellite markers following a single step mutation model. All coalescent simulations were performed using the program ms (Hudson 2002). I simulated an island model of population structure with 10 populations, 20 individuals were sampled from each population. For sequence haplotypes and microsatellites 30 independent loci were simulated, for SNP markers I simulated 100 independent SNPs. For single SNPs and haplotypes, multiple hits were not permitted. The microsatellite mutation model was implemented via R-script. In the program ms migration and mutation rate are expressed in terms of effective population size, $4Nm$ and $4N\mu$ respectively. I set up the simulations so that the effective population size was 1000 for each population and then parameters m and μ ranged from 0.0001 to 0.1 for m and 0.00001 to 0.001 for μ. Each simulation was repeated 5 times for each parameter combination.

Chapter 3

Results

3.1 Influence of mutation rate on different estimators of F_{ST}

3.1.1 Mutation rate is important in *A. thaliana*

In the *A. thaliana* dataset, microsatellite genetic diversity, and by extension mutation rate, is negatively correlated with F_{ST} (Figure 3.1). For instance, in the Spanish populations the correlation between H_S and F_{ST} was r = -0.862 (95 % CI = -0.944 - -0.678) with p < 0.001. For F'_{ST}, there was a positive relationship between H_S and F'_{ST}, (r = 0.479, p = 0.033), this was also true for D (Figure 3.1). Φ_{ST} was not correlated with H_S, (r = -0.294, p = 0.208). A similar pattern was obseved when the population mutation rate (Θ) was used instead of H_S (Table B.5). For Θ and F_{ST} r = -0.682, p < 0.001, for Θ and F'_{ST} r = 0.500, p = 0.025 and for Θ and Φ_{ST} r = -0.301, p = 0.197. Φ_{ST} is independent from genetic diversity and mutation rate in the data, except for H_S in the Central Asian populations (Table B.5).

3.1.2 Computer simulations - F_{ST}

The forward simulation results for the single step mutation model are presented in Figure 3.2. If mutation rate was high relative to migration rate, increasing mutation rate caused F_{ST} to decrease despite very limited migration. Φ_{ST} is not affected by mutation rate (Figure 3.2). These results are similar to those obtained by Hedrick (2005) and Balloux & Goudet (2002). Yet, for F'_{ST} and D I observed that increasing mutation rate leads to increased differentiation (Figure 3.2). This is in contrast to what was claimed in Hedrick (2005) but is compatible with Jost (2008). When mutation rate was increased up to 0.01, all estimators went down.

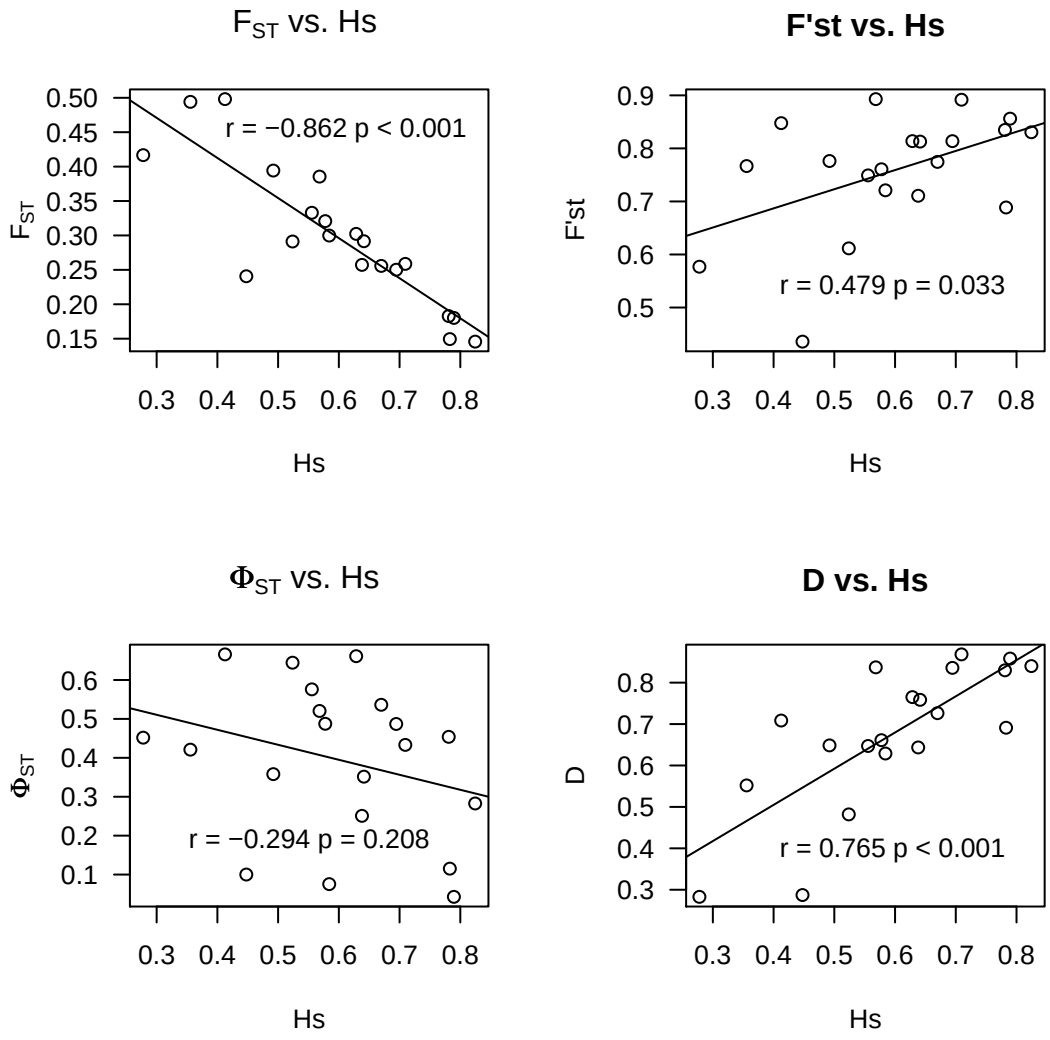

Figure 3.1: Correlation between Hs and different estimators of genetic differentiation for microsatellite markers the in Spanish populations. Gene diversity, Hs, was calculated for each locus and plotted against different estimators of genetic differentiation.

Table 3.1: Expected F_{ST} values were calculated as described in the methods. Observed values are simulation means for the different estimators or marker types. Simulation values are shown for a single mutation rate, $\mu = 0.00001$. MSAT stands for microsatellite loci and DNA for DNA sequence haplotypes.

Forward simulations					
m	E(F_{ST})	O(F_{ST})	O(Φ_{ST})	O(F'_{ST})	O(D)
0.1	0.0045	0.0040	0.0049	0.0190	0.0148
0.01	0.0431	0.0440	0.0414	0.1920	0.1549
0.001	0.3104	0.3104	0.3067	0.7581	0.6492
0.0001	0.8182	0.7570	0.7667	0.9831	0.9307
0.00001	0.9783	0.8339	0.8500	0.9931	0.9580
Coalescent simulations					
m	E(F_{ST})	O($F_{ST(SNP)}$)	O($\Phi_{ST(MSAT)}$)	O($\Phi_{ST(DNA)}$)	
0.1	0.0023	0.0277	0.0287	0.0299	
0.01	0.0220	0.0400	0.0420	0.0488	
0.001	0.1837	0.2057	0.2003	0.2127	
0.0001	0.6923	0.6950	0.7124	0.6933	

This likely reflects the fact that the number of alleles was restricted in a single locus, so some homoplasy could occur. Since mutation rates this high are biologically unrealistic it is not a source of concern. When the assumptions of the single step mutation model were relaxed, the effect of mutation rate was apparent also for Φ_{ST} (Figure 3.3).

Moreover, I observed that F'_{ST} and D were consistantly higher than the expected value for F_{ST} (Table 3.1). Together with their dependence on mutation rate, this suggests that F'_{ST} and D cannot be related to coalescence times the same way as F_{ST} can.

Next I examined how to compare different marker types. I simulated DNA sequence haplotypes, microsatellite markers (following SSM) and SNP markers. The results of the simulations show that by using Φ_{ST} different marker types can be compared regardless of mutation rate (Figure 3.4 and Table 3.1). Φ_{ST} is independent of mutation rate for both microsatellites and DNA sequences (Figure 3.4). Single SNP markers gave also comparable estimates to microsatellites and DNA sequences (Table 3.1). Therefore different types of markers can be compared if Φ_{ST} is used.

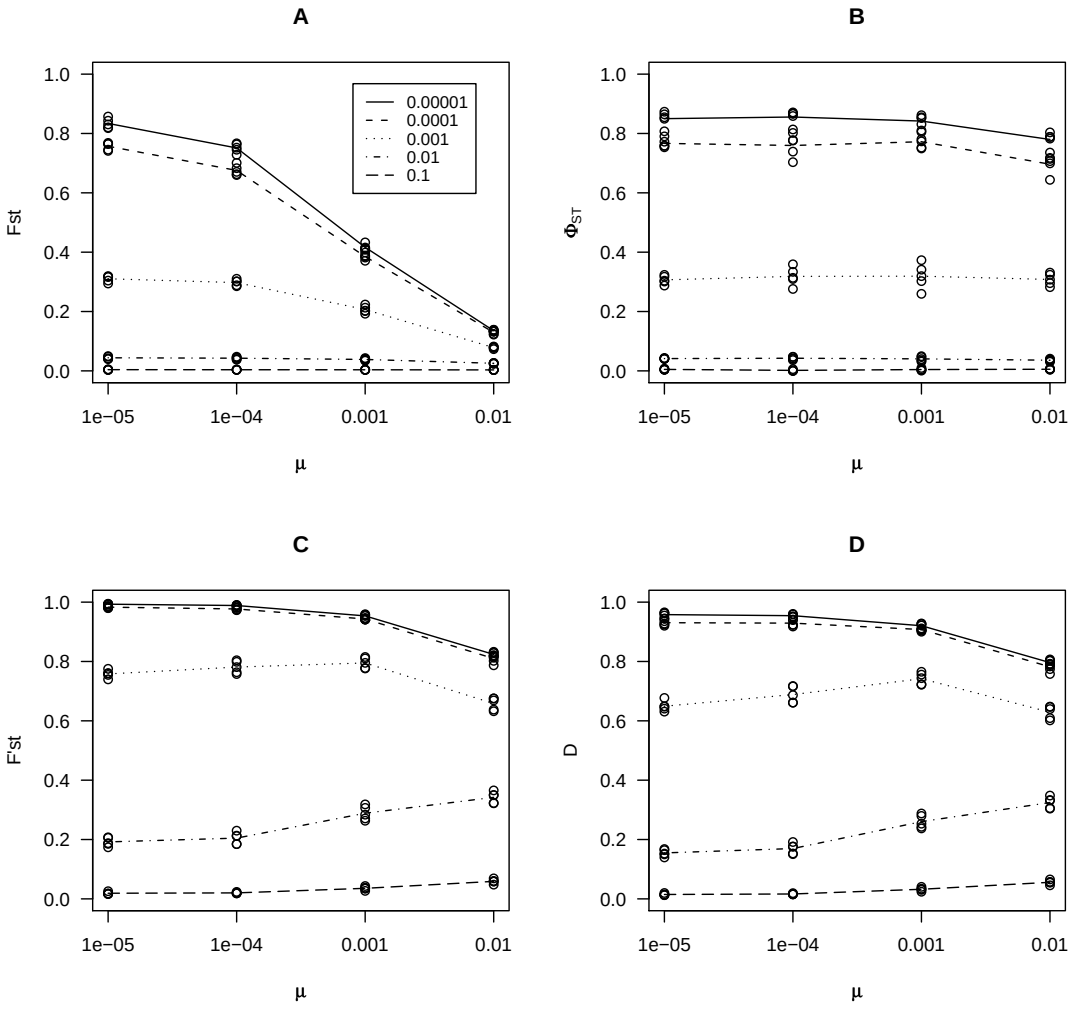

Figure 3.2: Different estimators of genetic differentiation were plotted against mutation rate. Different line types represent different migration rates as indicated by the legend. Different estimators are F_{ST}, Φ_{ST}, F'_{ST} and D in panels A, B, C and D respectively.

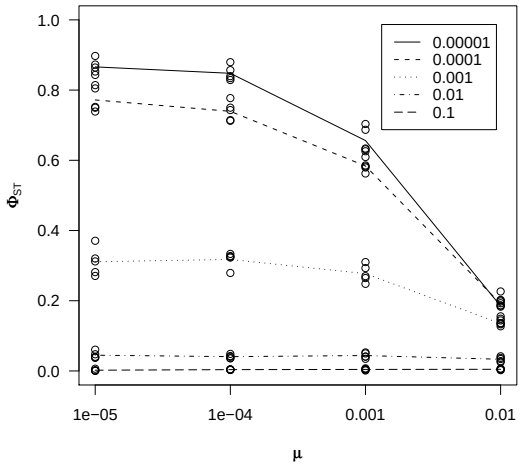

Figure 3.3: The effect of mutation rate on genetic differentiation calculated from Φ_{ST} using mixed mutation model. In this model, there was a probability of 0.2 that when a mutation occurs the allele will mutate to any state. Different lines represent different migration rates as indicated by the legend.

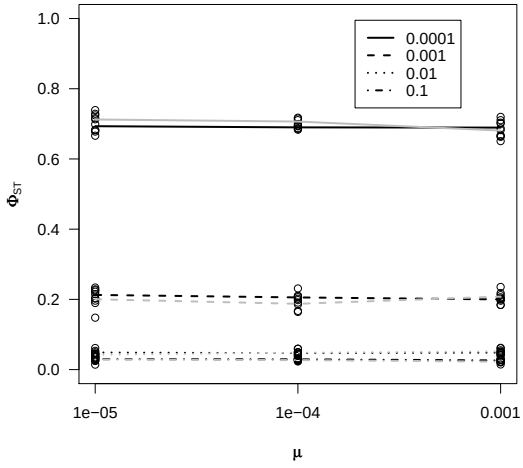

Figure 3.4: The effect of mutation rate on genetic differentiation, calculated from Φ_{ST}. Black lines represent DNA sequences, grey lines are microsatellites. Different types of lines represent different migration rates as indicated by the legend.

3.1.3 Computer simulations - Q_{ST}

I further examined the effect of mutation rate on Q_{ST}, estimator of genetic differentiation from quantitative traits. Q_{ST} had similar properties as F_{ST} in that if migration rate was low and mutation rate high it decreased Q_{ST} (Figure 3.5, panel A). I also performed the simulations using an incremental mutation model for the QTL, where each new mutation has an allelic value close to its ancestor. In this case Q_{ST} has similar properties as Φ_{ST} and is not affected by mutation rate (Figure 3.5, panel B).

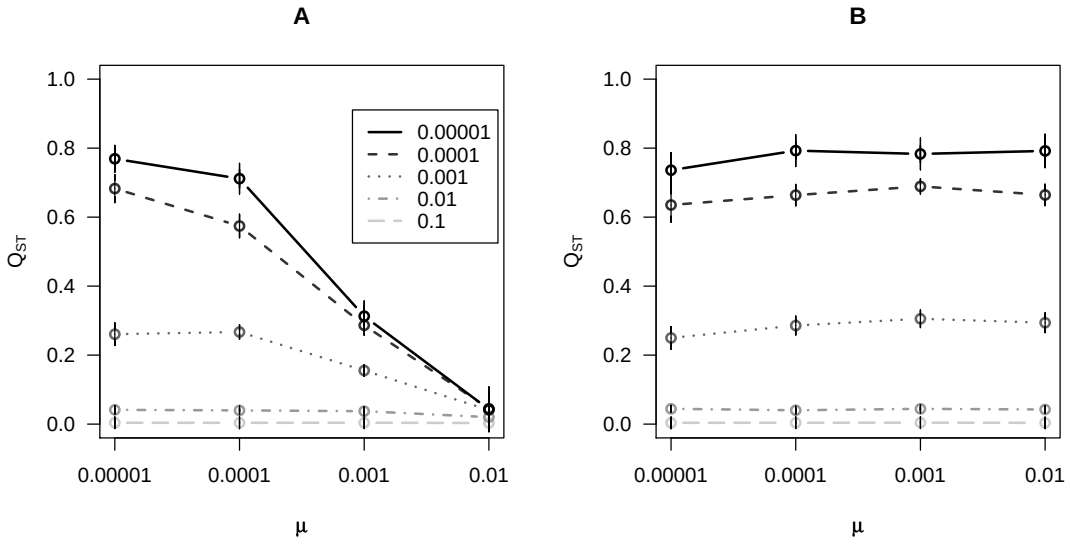

Figure 3.5: The effect on mutation rate on Q_{ST}. Different lines represent different migration rate as indicated by the legend. Simulations were done using two different mutation models, in panel A) random mutation model B) incremental mutation model.

3.2 Population genetics of *A. thaliana*

3.2.1 Genetic diversity

Genetic diversity was measured using microsatellite markers. To examine whether genetic diversity is the same for each region, indices of genetic diversity were calculated for the different regions. The allelic richness was highest in Spain (AR = 2.269), intermediate in France (AR = 1.720), lowest in Norway and Central Asia (1.245 and 1.383, respectively). Differences in allelic richness were significant (p

Table 3.2: Indices of genetic diversity were calculated for the 20 nuclear microsatellite markers. AR = Allelic richness and H_S = gene diversity

Region	AR	H_S
Spain	2.269	0.598
France	1.720	0.392
Norway	1.245	0.144
Asia	1.383	0.228

< 0.05, 1000 permutations) in all comparisons except when comparing the Central Asian populations to those of Norway or France. A similar trend was observed for Hs (Table 3.2).

For the 137 SNP markers used, 135 were polymorphic within Spain, all 137 were polymorphic within France, 119 within Norway and only 67 loci were polymorphic within Central Asia. SNPs used in this study are biased towards high frequency. However, when the SNPs were selected they were selected from an alignment that included genotypes from several different geographic locations and so are not geographically biased.

3.2.2 Population structure

I calculated measures of genetic differentiation for microsatellites and SNP markers between populations within regions and between regions. For microsatellites, genetic differentiation between populations was lowest in Spain ($F_{ST} = 0.2900$, $\Phi_{ST} = 0.3556$), intermediate for France ($F_{ST} = 0.4937$, $\Phi_{ST} = 0.6818$) and for Asia ($F_{ST} = 0.6026$, $\Phi_{ST} = 0.3101$) and the highest in Norway ($F_{ST} = 0.8004$, $\Phi_{ST} = 0.8128$). A similar trend was observed for both microsatellites and SNP markers (Supplementry Table B.6). However, the confidence intervals were sometimes quite broad, especially in Central Asia. Genetic differentiation between geographic regions was smaller then between populations within regions (Supplementary Table B.6). Differentiation measured by F'_{ST} from microsatellites was $F'_{ST} = 0.7208$ for Spain. 0.8115 for France, 0.9436 for Norway and 0.8413 for Central Asia. Values for D were 0.6393, 0.6509, 0.7241 and 0.6334 for the Spanish, French, Norwegian and Central Asian populations respectively.

I tested whether Φ_{ST} was higher than F_{ST} for the microsatelllite loci by using a permutation test that permutes allele sizes between different alleles (Hardy et al. 2003). If $\Phi_{ST} > F_{ST}$, one possibility is that stepwise mutations contribute to differentiation. Within Spanish populations, the difference is suggestive albeit not significant, 2-sided test p = 0.0629. Within the French populations, the

difference was significant p = 0.0210. In the Norwegian and Central Asian populations, differences were not significant(p = 0.1009 and p = 0.8561 respectively). This suggests that stepwise mutations may contribute to genetic differentiation in Spain and France. In the Norwegian and Central Asian populations instead, the microsatellite loci may exhibit some departure from the stepwise mutation model.

The PCA analysis grouped the populations into three groups, with one cluster including the Spanish and French populations, one cluster represented the Norwegian populations and the Central Asian populations formed another cluster (Figure 3.6). The first component separates the Central Asian populations from the other ones and explains 10.4 % of the variation and the second component separates the Norwegian populations from the Spanish and French and explains 8.5 % of the variation. When the analysis was done within each of the regions, it was evident that within Spain and France there is more admixture than within Norway or Central Asia (Supplementary Figure A.1). The Norwegian and Central Asian populations form groups that are mostly well separated, in contrast there are many individuals within Spain that cluster near another population. These individuals are possible migrants.

3.3 Seed dormancy variation in *A. thaliana*

I checked whether the different methods (see 2.5.1) used to calculate seed dormancy gave similar results. The correlation between binomial regression and the linear method was high r = 0.96, 0.96 and 0.80 for D25, D50 and D75 respectively, for all correlations p < 2.2E-16. The binomial regression seems theoretically a better way to measure dormancy, therefore all analyses were done using dormancy values calculated with it. Generally, biological conclusions remain the same, regardless of the method used. Next I checked whether maturation time in the greenhouse had any effect on dormancy measurments. Maturation time had no practical effect on the dormancy measurements, for instance, including maturation time in the model used to estimate heritability changed the estimate by 0.0018. Therefore I did not include maturation time in the models.

There are significant differences in seed dormancy between regions and populations within regions (Figure 3.7, Table 3.3). Seed dormancy was strongest in the Central Asian populations. There were some genotypes that were still dormant even after a year of after-ripening. For the European regions, the Spanish populations were the most dormant, the French populations had a lower dormancy than the Spanish populations and finally the Norwegian populations had the lowest dormancy. Among the European regions, seed dormancy decreases when going from Southern to Northern Europe (Figure 3.7).

However, there was a lot of variation within each of the regions, with differences

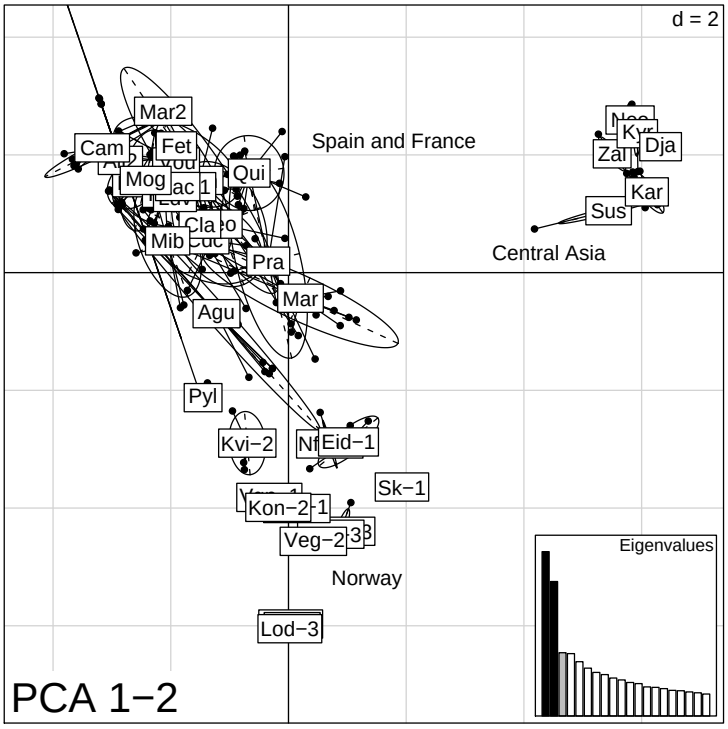

Figure 3.6: PCA of the population sample. X-axis is the first principle component and Y-axis the second component. Inset shows a barplot of eigenvalues (principal components) displaying the relative contribution of each component to the total genetic variance.

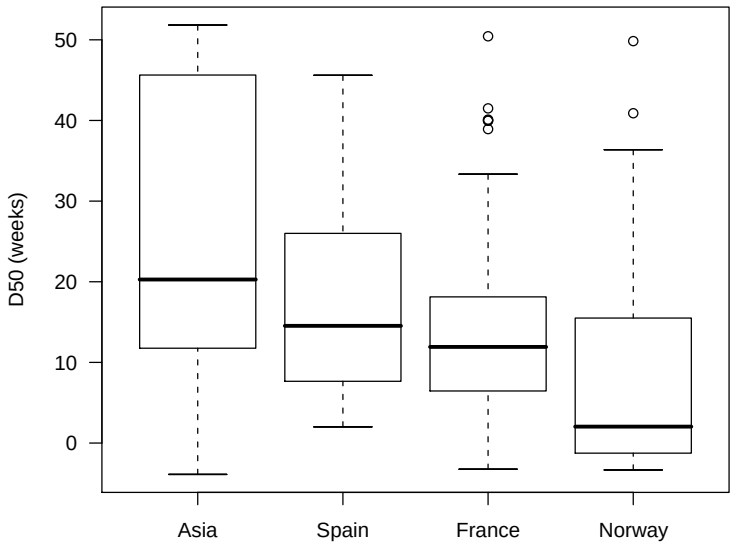

Figure 3.7: Seed dormancy (D50) plotted for each of the regions. Data are genotype means

in population means often greater than differences in region means (Supplementary Table B.9). In Spain the population Leo had a lower dormancy than the rest of the Spanish populations and low genetic variation within a population when Leo was compared to the rest of the Spanish populations (Supplementary Table B.9). In contrast some of the other Spanish populations, like Mar, Pra and San had comparably much higher genetic variation.

In the French populations some variation was again observed. However, since not all of the French populations have the same number of sampled individuals comparing the genetic variances is difficult. Nevertheless it seems that there are some populations that have relatively low dormancy and genetic variation, such as Lac, Mar2, Ldv and Vou while other populations are more dormant (Supplementary Table B.9).

Within Norway the situation is clearer, since most of the populations are completely non-dormant or have very low dormancy. There is one population however, Sk-1, that has strong dormancy and no genetic variation. In addition there is a group of three populations, Lod-2, Lod-3 and Tje-1 that are geographically close to each other and also have moderate to strong dormancy. These populations also have the most genetic variation of the Norwegian populations (Supplementary

Table 3.3: Analysis of variance table for seed dormancy in different regions. Data are genotype means

	Df	F-value	p-value
Region	3	42.872	< 2.2E-16
Population within region	37	12.830	< 2.2E-16
Residuals	245		

Table B.9).

Differences within a region were the greatest in the Central Asian populations. Populations Sus and Kar were very stongly dormant, while the populations Zal, Dja and Kyr had moderate dormancy and the population Neo was almost completely non-dormant. The populations Kar, Kyr and Zal have some genetic variation, while the other populations do not have any variation (Supplementary Table B.9).

The heritability values for seed dormancy are presented in table 3.4. The heritability, calculated over all genotypes in population sample, was high, around 0.8. Heritability remained high when calculated over genotypes within each of the regions (Table 3.4). The heritability estimates calculated using the REML or Bayesian method were nearly identical. This shows that the differences observed in seed dormancy between the different genotypes are mostly due to genetic variation.

3.3.1 Local adaptation in seed dormancy

To test if the observed differences in seed dormancy are adaptive, Q_{ST} of seed dormancy was compared to F_{ST} values from neutral markers. Q_{ST} values for dormancy are presented in table 3.5. Q_{ST} for dormancy was high, 0.71 – 0.75, when calculated over all populations or the European populations. Q_{ST} was also high within France, Norway and Central Asia (0.72, 0.92 and 0.79, repectively). Within Spain Q_{ST} for dormancy was only 0.38 (Table 3.5). Despite that some of the observed Q_{ST} values were high, they were never outside the distribution of neutral markers (Table 3.5). Within Norway and Central Asia, there are some markers that have an F_{ST} value of 1, thus within these regions it is not possible to detect selection using this method. There is uncertainty in estimating Q_{ST}, indicated by the large interval observed for Q_{ST} estimates.

Irrespective of the statistical issues in estimating Q_{ST}, it can still be interesting to use it as an exploratory tool. If some populations have high pairwise Q_{ST} values, they can be good candidates for further study. Such an example is the population Leo in Spain. It had a much lower dormancy and genetic variation than the other Spanish populations. When this population was compared to the other

Table 3.4: Heritabilities, H^2, for seed dormancy in different regions, 2.5 % and 97.5 % denote the limits of the 95 % highest posterior density interval.

Region	Trait	H^2	2.5 %	97.5 %
All	D25	0.7829	0.7431	0.8191
	D50	0.8058	0.7694	0.8387
	D75	0.7859	0.7464	0.8218
Spain	D25	0.6000	0.4709	0.7159
	D50	0.6961	0.5877	0.7889
	D75	0.7423	0.6463	0.8229
France	D25	0.6926	0.6002	0.7726
	D50	0.7507	0.6730	0.8174
	D75	0.7129	0.6271	0.7880
Norway	D25	0.5844	0.4476	0.7072
	D50	0.7370	0.6345	0.8238
	D75	0.7922	0.7055	0.8641
Central Asia	D25	0.8348	0.7508	0.9002
	D50	0.8468	0.7677	0.9080
	D75	0.7785	0.6710	0.8648

Spanish populations, for instance to population Agu, pairwise Q_{ST} was higher than expected from neutral markers. Pairwise Q_{ST} between Agu and Leo was 0.8770 for D50 and 0.9308 for D25, when compared to neutral markers the probability of observing equal or greater F_{ST} values was 0.034 for D50 and 0.013 for D25. Even though a pairwise comparison for Q_{ST} would never be significant taking into account the large uncertainly in its estimation, population Leo can still be flagged as a case warranting further study.

Even though Q_{ST} for dormancy is not higher than expected, variation in seed dormancy was related to the environment. Summer precipitation explains variation in seed dormancy (Figure 3.8, Table 3.6). Populations that received more precipitation in the summer were less dormant. The R^2 of the model was 0.31. There were some outlier populations that were quite dormant but received a fair amount of precipitation like Mog and Sk-1, or were non-dormant but received considerably more precipitation than the other populations like Veg-1 and Veg-2. These outliers did not drive the relationship, as excluding them increased the R^2 of the model to 0.41. Setting the small negative values for some the non-dormant populations to zero had almost no effect. The climate of the Central Asian populations

Table 3.5: Q_{ST} values for seed dormancy in different regions. Q_{ST} 2.5 % and 97.5 % denote the limits of the 95 % highest posterior density interval for Q_{ST}. 95 % F_{ST} indicates the value for the 95 % quantile of neutral marker F_{ST}.

Region	Q_{ST} D50	Q_{ST} 2.5 %	Q_{ST} 97.5 %	95 % F_{ST}
All	0.7523	0.6478	0.8421	0.7973
Europe	0.7053	0.5746	0.8184	0.7674
Spain	0.3815	0.1084	0.7301	0.6471
France	0.7237	0.5246	0.8785	0.7857
Norway	0.9237	0.8025	0.9911	1.0000
Central Asia	0.7912	0.5494	0.9523	1.0000

is quite different from Western Europe, therefore only the European populations were used, but the relationship remained significant when the Central Asian populations were included. The effect of precipitation was the strongest for D25, but remains significant for D50, p = 0.005. Summer precipitation has an effect even when it was included in a model with region already entered, p = 0.044. Other variables such as latitude, temperatures or other precipitation variables always explained less of the variation in seed dormancy.

Furthermore, when the French and Norwegian populations are used, pairwise Q_{ST} values for D25 were weakly correlated to differences in precipitation between the populations, r = 0.2057, p = 0.029 (Mantel-test). This correlation was only suggestive for D50, r = 0.1888, p = 0.066. In other regions this relationship was not significant. Neutral markers are not correlated with precipitation in these populations (Table 3.9). This further suggests that summer precipitation could drive local adaptation in seed dormancy.

Table 3.6: Linear model for D25 and summer precipitation, defined as precipitation in the warmest quarter of the year. Only the European populations are included.

	Df	F-value	p-value
Summer precipitation	1	16.16	0.0003
Residuals	33		

Taken together, these results suggest that seed dormancy is locally adapted in *A. thaliana*, since genetic variation is correlated to the environment. The environmental factor excerting the selection pressure is most likely the amount of precipitation received in the summer months. However, precipitation does not explain all of the variation, so there must be other factors that influence variation

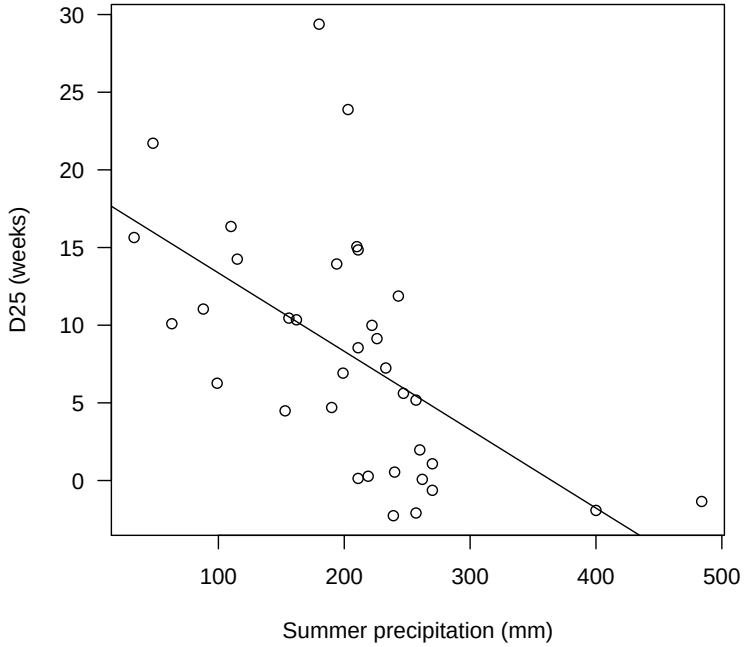

Figure 3.8: Seed dormancy is related to summer precipitation defined as precipitation in the warmest quarter of the year. The data are population means. D25 is shown in the figure as it has the strongest relationship. Only the European populations are included.

in dormancy as well.

3.4 Population genetics of *DOG1*

22 haplotypes could be defined for *DOG1* on the basis of exon 1 sequence and a large insertion in the first intron (Supplementary Table B.7). A summary of haplotype frequencies by region is presented in Table 3.7. Different haplotypes are at high frequency in different regions. In spain haplotypes 5, 9, 10 and 14 are at moderate frequencies, other haplotypes present in Spain are at low frequencies. In France there are two predominant haplotypes, 1 and 15. In Norway three haplotypes are at high frequencies, 2, 18 and 19. Finally, in the Central Asian populations there are only three different haplotypes 4 and 21 at nearly equal frequencies and 22 also at moderate frequency.

Table 3.7: Summary of *DOG1* haplotype frequencies in different regions

Region	Haplotype																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Spain	0.06	-	-	-	0.22	0.06	0.01	0.04	0.20	0.13	0.06	-	-	0.16	0.04	-	0.01	-	-	-	-	-
France	0.52	-	0.06	0.04	0.04	-	-	-	-	-	-	-	-	-	0.31	0.01	-	-	-	-	0.03	-
Norway	-	0.23	-	-	-	-	-	-	-	-	-	0.08	0.09	-	-	-	-	0.25	0.33	0.02	-	-
Central Asia	-	-	-	0.39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.43	0.17

In total there are 11 haplotypes segregating in the Spanish populations, 6 in the French and Norwegian populations and 3 in the Central Asian populations. Haplotype diversity, (Hd), is 0.87 in Spain, 0.62 in France, 0.77 in Norway and 0.64 in Central Asia. This is in contrast to genetic diversity estimates from neutral markers (Table 3.2), for which the Norwegian populations have the lowest genetic diversity.

The haplotype network of *DOG1* is presented in Figure 3.9. The *A. lyrata* outgroup cannot be joined to the network with 95 % confidence, however, haplotype 5 is likely to be the ancestral haplotype in *A. thaliana*. It occupies the central part of the network, with other branches of the network radiating from it (Figure 3.9). The most common haplotype in a population is likely to be the ancestral haplotype (Posada & Crandall 2001). Haplotype 5 is the most common haplotype in Spain. The Spanish haplotypes are mostly found in the central part of the network, while haplotypes from other regions occupy the peripheral parts of the network. Interestingly, the closely related haplotypes 18 and 19 that are found only in Norway at high frequency are connected to haplotype 5 by a long branch. Haplotypes 15 and 1 that are common in France are not closely related to each other, unlike haplotypes 4, 21 and 22 which are common in the Central Asian populations. The common haplotypes in France are present in Spain at low frequencies.

3.4.1 Sequence analysis of *DOG1*

Sequence analysis of the length polymorphism in the first intron (see section 2.3.3), revealed the 756 bp insertion to be a transposon that had inserted 14 bp after the start of the first intron. The three characteristics of DNA transposons were identified: 1) A target site duplication, the sequence 5'-GGTTTGGAC was found to be duplicated flanking the insertion. 2) Terminal inverted repeats characteristic for transposons were identified, when the insertion sequence was reverse complemented and aligned with itself. 3) Homology to other transposons was identified, as BLAST search againts the databases revealed homology to other transposons. The insertion could be identified as a *sTag1* element (Shankar et al. 2001). This is an non-autonomous DNA transposon, derived from the *Tag1* transposon present in *A. thaliana*.

I then analysed the predicted intron mediated enhancement of gene expression in different intron 1 alleles using the bionformatic tool IMETER (Rose et al. 2008). The results are shown in Table 3.8. While the other intron alleles have similar positive values, an intron with the insertation has a negative value which is quite different from the rest. Although, it is difficult to assess whether there differences between the intron alleles have a biological meaning, the allele 21 is predicted (bioinformatically) to lower gene expression relative to the other introns.

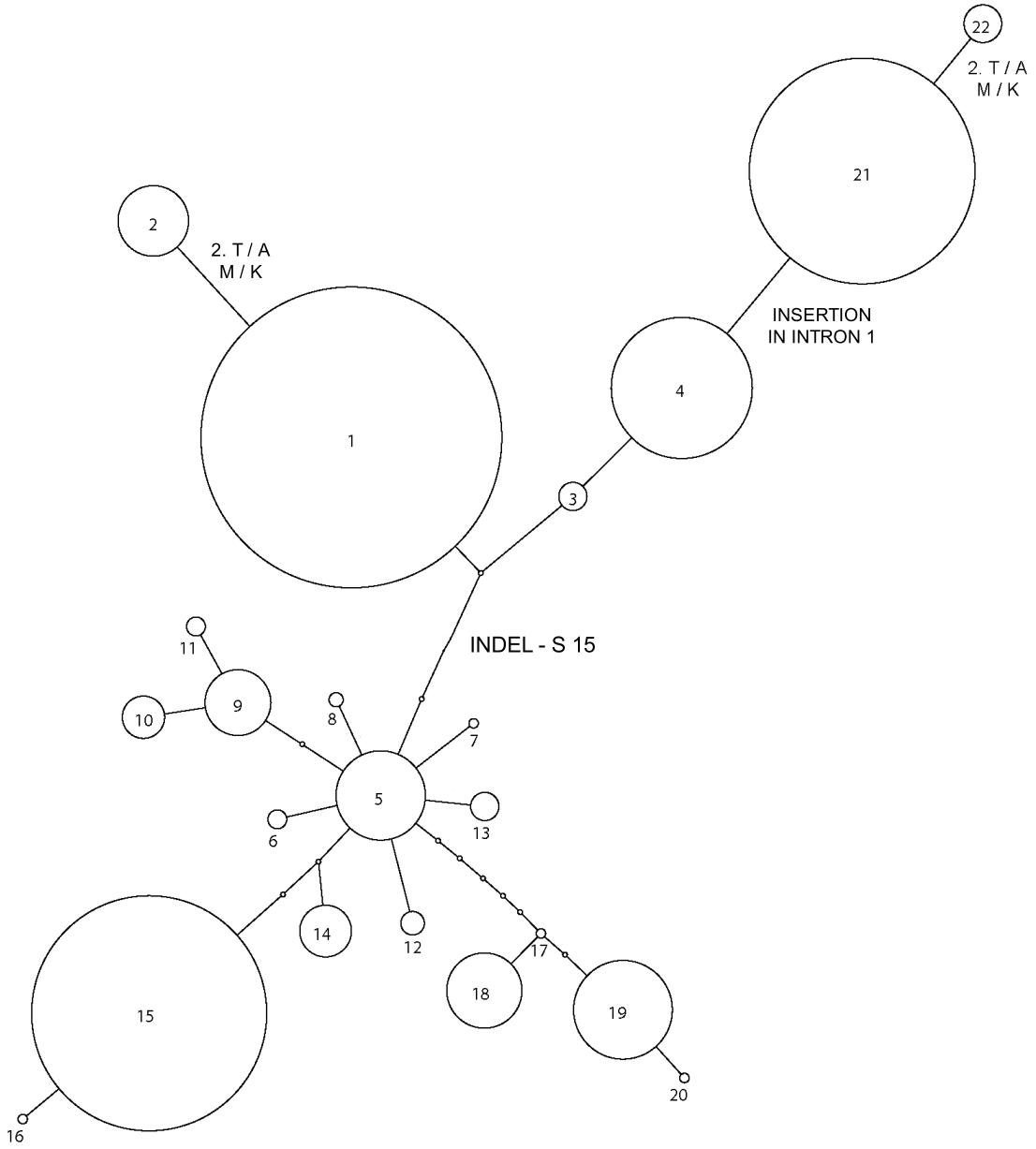

Figure 3.9: Haplotype network of *DOG1*. Each node represents a single mutation, the radius of the circle is proportional to the frequency of that haplotype. The population sample and the accessions from Wageningen were combined for the network.

Table 3.8: IMETER values for different intron 1 alleles. Positive values are predicted to increase gene expression, while negative values are predicted to decrease it.

Accession (haplotype)	score
Sha (4)	18.1
Ler	16.8
Fei-0	15.3
Cvi (14)	15.7
An-1	17.7
Trs-1 (21)	−23.2

For amino acid substitutions 18 non-synonymous changes and 7 synonymous changes were observed in a total of 131 codons. This gives a ratio 2.57, which could indicate selection on amino acid changes. However, the McDonald-Kreitman test was not significant with 10 synonymous substitutions between species and 16 non-synonymous substitutions between species when compared to *A. lyrata*. This gives a neutrality index of 1.61 which is not significantly different from 1, p = 0.555 (Fisher's exact test).

3.4.2 Selection on *DOG1*

While the restricted geographic distribution of haplotypes in *DOG1* reveals the possibility that there is local adaptation, this could also be a result of restricted migration and drift. Therefore I tested whether genetic differentiation in *DOG1* is higher than expected by chance alone. I compared Φ_{ST} of *DOG1* to the F_{ST} distribution obtained from 137 SNPs and 20 microsatellites. *DOG1* seems to be an outlier (Figure 3.10). When only the European populations are considered, there is only one marker that has a higher F_{ST} than *DOG1*. This gives the probability of observing equal or higher values, P($\geq$ Obs. value) = 0.0064. When the Central Asian populations are included, there are two markers with higher F_{ST}, giving P($\geq$ Obs. value) = 0.0127. This suggests that genetic differentiation in *DOG1* is higher than expected if *DOG1* is evolving neutrally.

I also tested for selection on *DOG1* by using another approach. If genetic differentiation between populations increases as an environmental variable increases, and this happens faster for *DOG1* than for neutral genetic markers, this suggests that selection is operating. Pairwise F_{ST} between populations was correlated to geographic distance or absolute differences in summer precipitation between populations. There was isolation by distance at a regional scale in neutral markers,

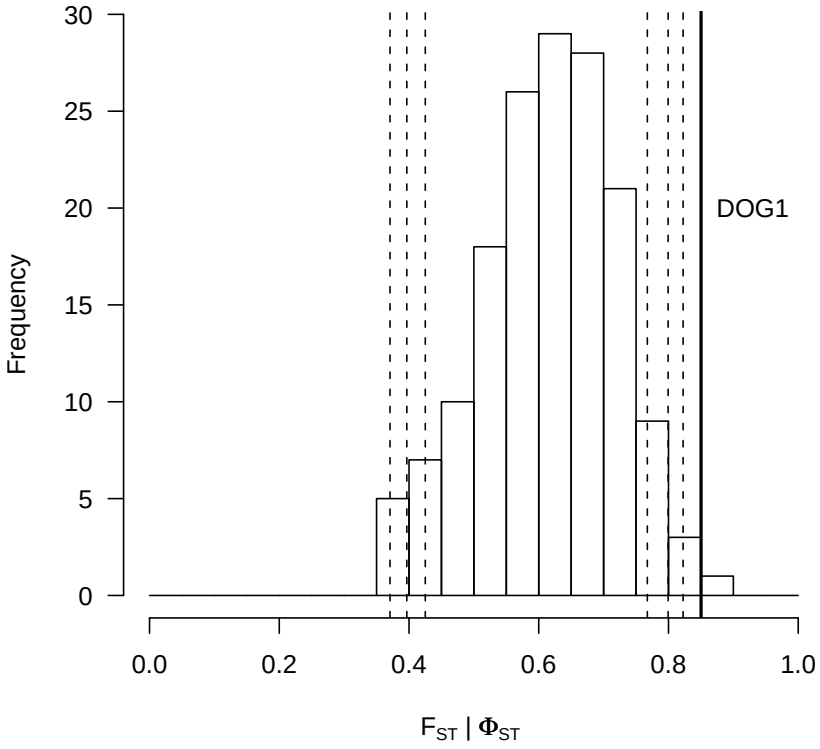

Figure 3.10: Distribution of F_{ST} values in the European populations. The histogram shows the distribution for the neutral markers. Solid line if the Φ_{ST} value for *DOG1*. Dashed lines are the 1%, 2.5%, 5%, 95%, 97.5% and 99% quantiles of the neutral distribution.

but isolation by distance was always higher for *DOG1* (Table 3.9). Between Spain and France isolation by distance was not significant for neutral markers, but for *DOG1* there was weak isolation by distance. Neutral differentiation never increases with increasing precipitation differences between populations, but for *DOG1* there seemed to be a slight increase in genetic differentiation. Although, this is only suggestive for all European regions together and for Spanish and French populations. However, when Norwegian and French populations are compared the correlation was weak but significant. (Table 3.9). This further suggests that *DOG1* variation in these populations is not neutral.

Table 3.9: Correlations between genetic differentiation and geography. Pairwise F_{ST} between populations, for SNPs or *DOG1*, correlated either to geographic distance or absolute differences in summer precipitation. Significance of correlations tested with the Mantel-test, 1000 permutations.

Region	Distance vs. SNP F_{ST}	Distance vs. *DOG1* Φ_{ST}	Precipitation vs. SNP F_{ST}	Precipitation vs. *DOG1* Φ_{ST}
European regions	r = 0.1774, p = 0.003	r = 0.3662, p < 0.001	r = -0.0626, p = 0.666	r = 0.1229, p = 0.055
Spain and France	r = 0.0147, p = 0.467	r = 0.1791, p = 0.021	r = -0.0785, p = 0.704	r = 0.1385, p = 0.053
France and Norway	r = 0.1958, p = 0.001	r = 0.4185, p < 0.001	r = 0.1299, p = 0.256	r = 0.1715, p = 0.002

Finally, if *DOG1* is under selection, population genetic theory predicts that there should be a peak of F_{ST} at the position of *DOG1* when genetic differentiation is viewed along the chromosome (Charlesworth et al. 1997). I tested this by genotyping SNP markers near *DOG1*. The results show that Φ_{ST} indeed peaks at the position of *DOG1* (Figure 3.11). Charlesworth et al. (1997) also suggested calculating between population heterozygosity ($H_T - H_S$), as if there is different amount of crossing over in different chromosomal regions this could cause problems, since within population genetic variance is included in F_{ST} measurements. The results show that there is also a peak for between population heterozygosity at the position of *DOG1* (Figure 3.11). This indicates that the high genetic differentiation in *DOG1* is caused by local selection for different alleles.

3.5 Natural genetic variation in *DOG1*

To investigate whether natural genetic variation in *DOG1* segregating in natural populations is also functional variation I performed a candidate gene association study with *DOG1*. Some QTL-mapping experiments were also done to confirm

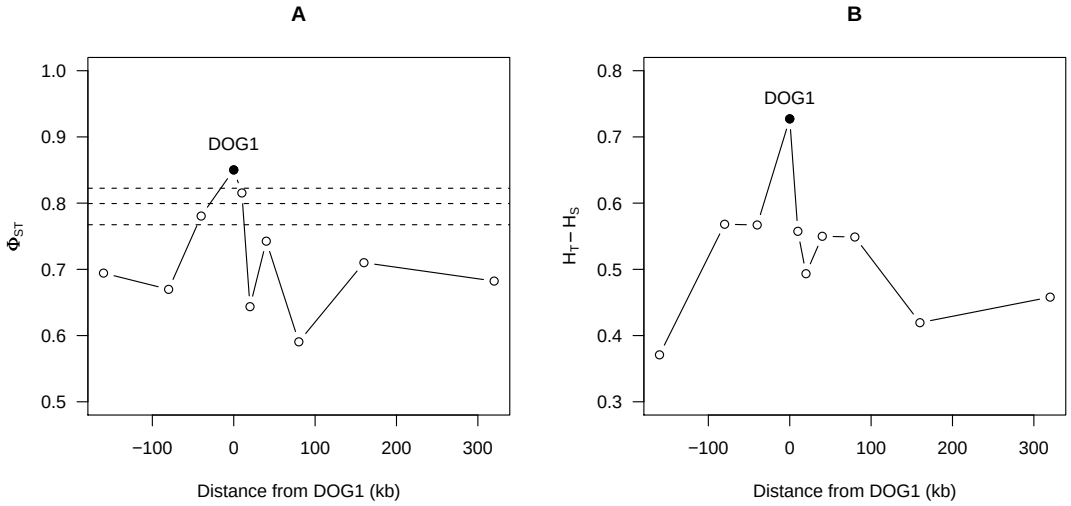

Figure 3.11: Genetic differentiation viewed along the chromosome near *DOG1*. Each locus consists of a set of closely linked SNP markers, *DOG1* is the filled circle. A) Φ_{ST} along the chromosome. The dashed lines are the upper quantiles of the neutral distribution. B) Between population heterozygosity ($H_T - H_S$) along the chromosome.

the associations and to investigate the genetic architecture of seed dormancy in *A. thaliana*.

3.5.1 Candidate gene association

First I tested each *DOG1* haplotype for association with seed dormancy (three different time points) in the whole sample and within each of the regions. The results are shown in Table 3.10, some of the alleles that are segregating in natural populations are associated with seed dormancy. Haplotype 4, which is present in the French, the Wageningen and the Central Asian populations was most strongly associated allele. It was associated with increased dormancy. Haplotype 4 has the highest marker R^2 values explaining up to 9 % of the variance in the French populations. Haplotypes 6, 9 and 10 are weakly associated with dormancy when only the Spanish populations are considered, although they are not significant after correcting for multiple testing (Table 3.10). Haplotype 13 is weakly associated with an increase of dormancy in the whole sample. Haplotype 15 is associated with decreased dormancy in the French populations. Interestingly, within France there are two predominant haplotypes segregating, 15 and 1 (Table 3.7). Although the effect of haplotype 15 is seen only for D25, it explains a comparatively large amount

of the variance, 5 % (Table 3.10). Haplotypes 18 and 19 are also weakly associated with decreased dormancy in the whole sample, these haplotypes segregating are at high frequency in the Norwegian populations (Table 3.7) and are connected to the other haplotypes by a long branch (Figure 3.9). Haplotypes 21 and 22 are both associated with decreased dormancy in the Central Asian populations. As there are only three different haplotypes segregating in Central Asia (4, 21 and 22) it implies that haplotype 4 is associated with decreased dormancy also in Central Asia.

In general, variation in *DOG1* explains only a small amount of the variance in seed dormancy, mostly the haplotypes explain around 2–3 % of the variance (Table 3.10). However, it is likely that the effects of the haplotypes are not estimated correctly in many cases, since the strong population structure in the sample translates into a rather limited power to detect associations. There are several haplotypes that are unique to certain populations. Because population structure has to be accounted for, this will decrease the power to detect associations. Moreover, there are many different haplotypes and many of them at rather low frequencies, this will also decrease power.

3.5.2 QTL mapping

I performed crosses to check cosegregation of several *DOG1* alleles with seed dormancy. The results are shown in Table 3.11. For the cross between the parents Trs-1 and Trs-2, the F_3-lines were used in the analysis (see section 2.5.9). The difference in D50 between haplotypes 21 and 4 was -2.54 WODS (t-test, $p = 0.006$), which translates into an allelic effect of $a = 1.27$. The difference between the lines was the same in the colder environment, even though the lines were more dormant in general.

To reveal additional loci that contribute to seed dormancy in the F_2-populations, QTL mapping was performed. The populations Cam-4 x Fet-6 (CF), Kon-2-2 x Fet-6 (KF), All2-1 x Fet-6 (AF) and All2-1 x Cam-4 (AC) were analysed by QTL mapping. In addition to *DOG1* other QTLs were revealed. The LOD thresholds obtained by permutation and that correspond to an $\alpha = 0.05$ were 3.04, 2.98, 3.03, 2.97 for populations CF, KF, AF and AC respectively. Results of the QTl-mapping are shown in Table 3.12. In addition to *DOG1* some additional QTL were detected, in a cross between Kon-2-2 and Fet-6, which is a cross between a Norwegian and a French genotype there is an additional large effect QTL on chromosome 4. In this cross there were indications of additional small effect QTLs on top of chromosome 5 and on chromosome 3. However, these were not significant in the multiple QTL model. A picture of the LOD profile of this cross in shown in Figure 3.12. In most crosses only a single QTL in addition to *DOG1* could be detected and no other QTL than *DOG1* was constantly detected in all of the crosses. *DOG1* was always

Table 3.10: Significant assocations for *DOG1* and seed dormancy. Associations have been tested for all three time points for the whole sample and within each region. For multiple testing corrections the Bonferroni-Holm method was used.

Haplotype	Sample	Trait	p-value	p-adjusted	Direction	Marker R^2
2	All	D50	0.014	0.245	increase	0.006
		D75	0.002	0.042	increase	0.010
4	All	D50	1.10E-07	2.42E-06	increase	0.028
		D25	2.55E-08	5.61E-07	increase	0.027
		D75	4.69E-06	1.03E-04	increase	0.021
	France	D50	7.05E-04	0.005	increase	0.049
		D25	2.35E-05	1.65E-04	increase	0.088
		D75	0.005	0.038	increase	0.032
6	Spain	D25	0.050	0.450	increase	0.017
9	Spain	D50	0.045	0.451	decrease	0.016
		D25	0.023	0.253	decrease	0.022
10	Spain	D50	0.023	0.255	decrease	0.020
		D25	0.023	0.253	decrease	0.022
		D75	0.025	0.270	decrease	0.019
13	All	D50	0.008	0.160	increase	0.007
		D25	0.016	0.304	increase	0.005
		D75	0.010	0.187	increase	0.007
15	France	D25	0.002	0.011	decrease	0.050
18	All	D50	0.010	0.184	decrease	0.007
		D25	0.014	0.288	decrease	0.005
		D75	0.007	0.124	decrease	0.008
19	All	D50	0.015	0.252	decrease	0.006
		D25	0.016	0.304	decrease	0.005
		D75	0.039	0.620	decrease	0.005
21	Central Asia	D50	0.001	0.003	decrease	0.025
		D25	0.005	0.009	decrease	0.019
		D75	1.84E-04	5.52E-04	decrease	0.039
22	All	D50	1.63E-05	3.42E-04	decrease	0.019
		D25	7.43E-06	1.56E-04	decrease	0.018
		D75	4.35E-04	0.009	decrease	0.013
	Central Asia	D50	0.001	0.003	decrease	0.025
		D25	0.003	0.008	decrease	0.021
		D75	2.60E-04	5.52E-04	decrease	0.037

Table 3.11: Co-segregation of different *DOG1* alleles with seed dormancy in F_2-populations. D50 difference is the difference in the mean homozygote values for the different haplotypes. The significance of this difference was tested with a *post hoc* test (TukeyHSD), corrected for multiple testing. Haplotype on the right in the first column is always the more dormant haplotype. For a and k the numbers in parenthesis are the 95 % highest posterior density intervals.

Haplotypes	Difference D50	p-adjusted	R^2	allelic effect, a	dominance coefficient, k
1 and 5	−1.4	5.69E-13	0.35	0.7 (0.51 – 0.85)	−0.3 (−0.69 – 0.03)
15 and 5	−4.37	7.55E-15	0.54	2.19 (1.81 – 2.53)	−0.15 (−0.46 – 0.03)
15 and 1	−1.56	3.49E-05	0.12	0.78 (0.38 – 1.09)	0.02 (−0.64 – 0.76)
19 and 5	−2.27	4.66E-15	0.52	1.13 (0.92 – 1.32)	−0.12 (−0.38 – 0.12)
18 and 19	−0.06	0.12	–	–	–

the QTL with the largest effect except in cross All2-1 x Cam-4 , where it had a modest efect.

The other QTLs correponds most likely to previously detected *DOG* loci (Alonso-Blanco et al. 1999; Clerkx et al. 2004; Bentsink et al. 2010). QTL on top of chromosome 5 is likely to be *DOG4*, QTL at position 46 on chromosome 4 could be *DOG5*, QTL at the top of chromosome 1 is at the same position as *DOG2*. On chromosome 2 QTL at position 64.3 is near the position of *DOG20*, QTL at chromosome 3, position 70.1 is *DOG6* and QTL at position 22.0 in chromosome 4 is likely to be *DOG18*. Since there are some gaps in the genetic maps and recombination is limited in F_2 populations of this size the map position of the QTL can be displaced of those of the recombinant inbred line populations. In some cases the QTL at *DOG1* was slightly displaced from the *DOG1* marker to the flanking, tightly linked, marker. This is due to the low recombination between these markers. It is also likely that the effects of the large effect QTLs are overestimated, as the mapping populations investigated here are rather small (Beavis 1998).

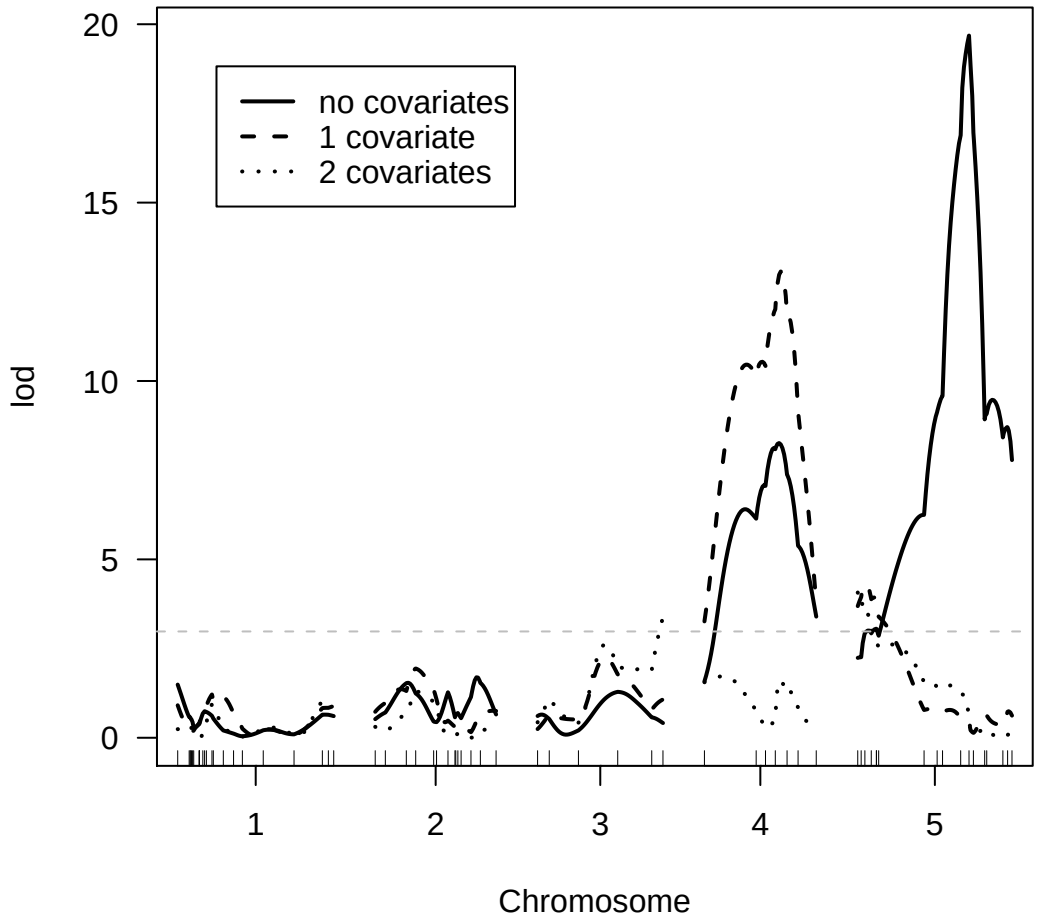

Figure 3.12: LOD profile for QTL mapping in cross Kon-2-2 x Fet-6. The different lines correspond to genome scans with different covariates. The first covariate is the QTL on chromosome 5 and the second covariate is the QTL on chromosome 4. The dashed grey line indicates the LOD significance threshold.

Table 3.12: Results of QTL mapping in F_2-populations. CHR indicates chromosome. Pos. indicates the position of the QTL on the genetic map, units are in centimorgans. Marker indicates the closest marker to the QTL postion. The column % var. indicates the % of variance explained by this QTL in the mapping population, the column direction indicates the parental allele that increases dormancy.

Cross	CHR	Pos.	Marker	LOD	% var.	Direction
Cam-4 x Fet-6	5	75.0	DOG1 / AtMSQTsnp392	28.275	53.4	Fet-6
	5	3.0	AtMSQTsnp325	6.539	8.1	Fet-6
Kon-2-2 x Fet-6	5	66.0	DOG1	28.566	44.4	Fet-6
	4	46.0	AtMSQTsnp281	15.337	17.8	Fet-6
All2-1 x Fet-6	5	61.0	DOG1	15.208	33.2	Fet-6
	1	1.0	AtMSQTsnp8	8.607	16.6	All2-1
All2-1 x Cam-4	2	64.3	AtMSQTsnp184	3.439	3.8	All2-1
	3	70.1	AtMSQTsnp235	5.565	6.4	Cam-4
	4	22.0	AtMSQTsnp263	22.207	34.0	All2-1
	5	57.0	DOG1 / AtMSQTsnp388	10.350	13.0	All2-1

Chapter 4

Discussion

4.1 Mutation rate and estimation of genetic differentiation

4.1.1 F_{ST} and mutation rate

The simulations show that D and F'_{ST} both depend on mutation rate (Figure 3.2). For D this is shown by Jost (2008), but the fact that F'_{ST} is dependent on mutation rate is not made clear by Hedrick (2005). New mutations increase differentiation between populations, especially if migration rate is low. D and F'_{ST} measure allelic differentiation, regardless of the process that generates these differences (Jost 2009). This means that D or F'_{ST} are usefull for studies where the amount of genetic differentiation is of interest *per se*. Instead, studies where genetic differentiation needs to be compared across loci, such as in this study, these estimators become problematic.

Φ_{ST} is the only estimator that is completely independent from mutation rate (Figure 3.2). The estimator was derived for this purpose (Slatkin 1995), see also Balloux & Goudet (2002). I also showed that if the assumptions of Φ_{ST} are met, both DNA haplotypes and microsatellites give estimates comparable to F_{ST} calculated for bi-allelic SNPs (Figure 3.4). In a large empirical dataset in humans, where microsatellites analysed using a stepwise mutation model and SNPs gave comparable results, a similar conclusion was reached (Sun et al. 2009).

In *A. thaliana* and in many other species, microsatellite loci were often shown to deviate from pure single step mutation model (Symonds & Lloyd 2003; Ellegren 2004; Calabrese & Sainudiin 2005; Bhargava & Fuentes 2010). The results show further that Φ_{ST} became dependent on mutation rate, if there was some deviation from the single step mutation model (Figure 3.3). This is also seen in our dataset in some geographic regions, as we find that Φ_{ST} is not different from F_{ST}. Yet,

it could be argued that Φ_{ST} is still preferred over F_{ST}, even if there are some deviations from the SMM model Φ_{ST} still perform better than F_{ST}.

Correlating estimates of population differentiation to levels of diversity has highlighted the effect of mutation rate on estimates of differentiation in *A. thaliana*. We observe this effect in the four distinct geographical regions (Figure 3.1 and Supplementary Table B.5). Therefore, mutation rates seem to significantly impact estimates of population differentiation in this species. This relationship has also been found in *Arabidopsis lyrata* (Muller et al. 2007), a relative of *A. thaliana* exhibiting a markedly different life-history and more genetic diversity than *A. thaliana* (Clauss & Mitchell-Olds 2006). This relationship has also been observed in some fish species (O'Reilly et al. 2004; Carreras-Carbonell et al. 2006). As shown by simulations the problem is most severe in systems where diversity is high and migration between populations is low. It seems that populations of many organisms have the potential to be in the region of parameter space where F_{ST} reflects variation in both migration and mutation rates. Therefore, studies of population structure should systematically investigate this effect.

In conclusion, genetic differentiation in different types of markers: SNPs, microsatellites and DNA sequences, can be compared if they are analysed using Φ_{ST}. Then, different mutation rate of the different markers is taken into account.

4.1.2 Q_{ST} and mutation rate

For Q_{ST} the simulations show that it is affected by the mutation rate of the underlying QTL (Figure 3.5). As Q_{ST} is equivalent to F_{ST} under neutrality (Lande 1992) this result is perhaps not surprising. However, it is interesting to note that this bias is observed in a parameter range that seems to be empirically relevant. In the simulations the parameter μ, the genic mutation rate was varied. The overall polygenic mutation rate, or variance contributed mutation in each generation (assuming additive effects), of a quantitative trait is $\sigma_m^2 = 2\mu L\sigma_\alpha^2$ (Lynch 1988b), where L is the number of QTL, 2 accounts for diploidy and σ_α^2 is the variance of allelic effects. In the simulations the parameters $L = 10$ and $\sigma_\alpha^2 = 0.1$ were held constant in all simulations. This translates into σ_m^2 of twice as large as genic mutation rate, so that for $\mu = 0.0001$ $\sigma_m^2 = 0.0002$. The environmental variance, σ_E^2, was set to 1 so mutational heritability, $h_m^2 = \sigma_m^2/\sigma_E^2$ (Lynch 1988b; Lynch & Walsh 1998) was equal to σ_m^2. When genic mutation rate of QTL was $\mu = 0.001$, effects on Q_{ST} were rather large when migration rates were low (Figure 3.5). This corresponded to $h_m^2 = 0.002$. It is interesting to note that empirical estimates of mutational heritabilities frequently fall around this value. For instance, h_m^2 for the well studied trait of *Drosophila* bristle number seems to be around 0.0035 – 0.0043 (Lynch & Walsh 1998). Schultz et al. (1999) estimated h_m^2 for few life-history traits in *A. thaliana* and found that they fall around 0.003. Estimating mutational her-

itability empirically is not easy, but empirical data from several sources suggests that for many traits it is around 0.002 – 0.003 (Lynch & Walsh 1998).

Interestingly, if new alleles display incremental changes of function, Q_{ST} becomes independent of the mutation rate. However, alleles of large effect contributing to quantitative trait variation have been observed frequently in natural populations (Orr 2005). A prominent example is provided by the *FRIGIDA* locus, a gene contributing to quantitative variation in flowering time in *A. thaliana*. Several loss-of-function mutations were reported to segregate in natural populations (Le Corre 2005). Thus QTL are likely to frequently deviate from a purely incremental mutation model.

Traditionally it was thought that $Q_{ST} < F_{ST}$ would indicate selection for the same phenotypic optima in different populations (Merilä & Crnokrak 2001). Intuitively, systems where F_{ST} is large, because migration is low, offer potentially the greatest statistical power to detect $Q_{ST} < F_{ST}$. Results presented here instead suggest that the bias in Q_{ST} may be the greatest in such cases and that the utility of $Q_{ST} - F_{ST}$ comparisons may be limited to detecting diversifying selection. High polygenic mutation rates will bias Q_{ST} downwards, if local adaptation ($Q_{ST} > F_{ST}$) is of interest, the test will remain consevative.

A recent meta-analysis of studies comparing $F_{ST} - Q_{ST}$ noted that using F'_{ST} would generally change the conclusions of the studies (Leinonen et al. 2008). However, as discussed above using F'_{ST} or D in such studies is not appropriate, because these measures of genetic differentiation are not independent from the high mutation rate of microsatellites. Yet, this is a concern only for a subset of cases where migration rate is low.

4.1.3 Effects of mating system

Simulations were performed under random-mating as well as with a self-fertilizing rate of 0.9 and this yielded essentially the same results. From a population genetics perspective, self-fertilisation should reduce effective population size and thus coalescence times of alleles within populations (Nordborg & Donnelly 1997). This is precisely what was observed in the simulations: absolute F_{ST} values were higher, but the relationship of the statistics to mutation rate remains qualitatively the same (Supplementary Figure A.3). This is also true for Q_{ST} (Supplementary Figure A.4). Thus the effect of self-fertilisation is mainly to increase F_{ST} values, but it does not alter the effect of mutation rates on the estimates.

4.2 Seed dormancy is locally adapted in *A. thaliana*

4.2.1 Seed dormancy variation

I observed extensive genetic variation in seed dormancy in the sampled populations. When calculated over all genotypes in the sample, seed dormancy had a high heritability, around 0.8. Within the individual regions this was lower, but still remained quite high around 0.6 – 0.8 (Table 3.4). There were differences in seed dormancy between and within regions (Figure 3.7, Table 3.3 and Supplementary Table B.9). In Europe there was a tendency for dormancy to decrease when going from south to north. The Spanish populations were the most dormant, then the French and the Norwegian populations were mostly non-dormant.

There was extensive variation within each of the regions. Particularly interesting was that many populations had a lot of genetic variation (Supplementary Table B.9). As the amount of genetic variation within a population should be related to the intensity of selection, this could indicate that selection on seed dormancy is not present or is very mild in these particular populations. Alternatively, there could be some mechanism that maintains genetic variation in dormancy, perhaps temporally variable natural selection. However, theoretical arguments would suggests that this is unlikely, since temporal variation favours the evolution of generalist phenotypes (Kisdi 2002). If some kind of bet-hedging stategy due to temporal variation would be favoured in these populations one would expect that phenotypic plasticity would evolve. On the other hand, in this study seed dormancy was measured only in a single environment due to logistic limitations. Therefore, it is possible that the large genetic variation observed in some populations would not be expressed under different environmental conditions.

Nevertheless, the fact that some populations within regions were very different from their neighbours hints that these populations may be locally adapted. For instance the population Leo within Spain and the population Neo within Central Asia both have less dormancy than other populations within their regions and very low genetic variation within population. This is also true for the population Sk-1 in Norway, except that it is very dormant in contrast to its neighbours.

It is difficult to compare these results to other studies. Most other studies that investigated population differentiation in seed dormancy only measured germination % at a few time points, so dormancy was not estimated accurately or the studies did not measure genetic variances in a proper design. Evans & Ratcliffe (1972) measured germination in some British *A. thaliana* populations, and observed there was some variation within and between populations, but proper genetic variances were not estimated. This was also true for Napp-Zinn (1976) and

Ratcliffe (1976). Naylor & Jana (1976) studied seed dormancy variation in *Avena fatua*. The authors observed some differences between populations and there was a correlation between parents and offspring, although heritability was not estimated. The authors conclude that there was local adaptation, yet no evidence for this was presented. Andersson & Milberg (1998) studied dormancy of four different species, again observing some variation within and between populations, this time concluding that there was no local adaptation. Nevertheless, it is clear that genetic variation in seed dormancy exists in many different species. In this respect the situation somewhat similar to results obtained here, since extensive variation between and within populations was observed. Extensive variation within *A. thaliana* in seed dormancy was also observed by M. Debieu in a worldwide accession sample (unpublished results).

4.2.2 Genetic differentiation in seed dormancy

To examine if the observed differences in seed dormancy were higher than differences expected based on neutral markers, I compared Q_{ST} for dormancy to F_{ST} calculated from neutral markers.

Recently there has been some criticisms of using Q_{ST} to infer the action of selection. The core of the problem is that Q_{ST} has both high sampling variance (O'Hara & Merilä 2005; Goudet & Büchi 2006) and evolutionary variance (Miller et al. 2008; Whitlock 2008). Also, it is not enough to compare Q_{ST} simply to the mean F_{ST}, but it must be compared to the distribution of neutral F_{ST} values (Whitlock 2008). Moreover, non-additive gene action and dominance may bias Q_{ST} downwards (Whitlock 1999; Goudet & Büchi 2006; Goudet & Martin 2007), additionally for some demographic scenarios Q_{ST} can be smaller than F_{ST} (Miller et al. 2008). The case where $Q_{ST} > F_{ST}$ still remains as an indication of local adaptation (Goudet & Büchi 2006; Goudet & Martin 2007). In our study dominance is not an issue, since the lines are homozygous, however more serious concern is the fact that neutral F_{ST} is very high in our populations (Supplementary Table B.6). Given the high variance of Q_{ST} estimates it is nearly impossible to obtain Q_{ST} estimate that is statistically significantly different from the neutral markers. Instead, in this study Q_{ST} was used as an exploratory tool (Whitlock 2008) to investigate overall patterns and to identify interesting populations for further study.

Q_{ST} for seed dormancy was high in many instances but never outside the distribution of neutral markers. There were some cases where pairwise Q_{ST} between populations was suggestive of local adaptation, for instance in Spain for populations Agu and Leo (see section 3.3). Although, this does not conclusively prove local adaptation, it suggests that these populations could be investigated further.

Most studies comparing Q_{ST} and F_{ST} have found Q_{ST} to be higher than F_{ST}

for most traits (Merilä & Crnokrak 2001; Leinonen et al. 2008), indicating that local adaptation is common. Notably, the exception to this pattern has been the studies in *A. thaliana* (Kuittinen et al. 1997; Stenøien et al. 2005), where Q_{ST} for flowering time and some other traits was observed to be in line with neutral markers. Yet, Le Corre (2005) observed higher Q_{ST} than F_{ST} for flowering time in populations from France (some these populations were also used here). Banta et al. (2007) also found higher Q_{ST} for flowering time in *A. thaliana*. Although, these previous studies could be criticised on the grounds that their analysis of $Q_{ST} - F_{ST}$ differences is not correct, since they compared Q_{ST} to the mean F_{ST}. However, the conclusion that flowering time is locally adapted is very likely to be true as other lines of evidence for local adaptation on flowering time have been found. Evidence for this comes from clines (Stinchcombe et al. 2004) and sequence variation suggests that *FRI* is not evolving neutrally (Le Corre et al. 2002; Toomajian et al. 2006). It should be also noted that in this study, it was observed that the point estimate of Q_{ST} for flowering time (DTB) was outside the distribution of the neutral markers (data not shown).

It is likely that, high neutral differentiation, mediated partly by high selfing rate in *A. thaliana* obscures the detection of higher Q_{ST} than F_{ST}. This has also been suggested by other studies. In a highly selfing snail species, Chapuis et al. (2007) detected higher Q_{ST} than F_{ST} only when populations were grouped together by habitat type.

4.2.3 Ecological factors influencing dormancy

It could be argued that the observed differences in seed dormancy could also be explained by the strong population structure and thus genetic drift. It is well known that *A. thaliana* does have strong population structure (Nordborg et al. 2005; Pico et al. 2008; Platt et al. 2010). However, this is unlikely because seed dormancy was correlated with summer precipitation (and this correlation was stronger than for latitude). Populations receiving more precipitation in the summer were less dormant. Genetic differentiation at neutral markers was not correlated with precipitation, which supports the view that differences in dormancy between populations do not reflect the action of genetic drift. This is reinforced by the result that Q_{ST} for seed dormancy was weakly correlated with precipitation differences between populations in some comparisons (see section 3.3).

The results fit the ecological predictions about the role of seed dormancy in natural populations, plants can avoid summer drought by not germinating in the spring (Baskin & Baskin 1972, 1983). Populations with shorter summer droughts can probably benefit by germinating in the late summer or early autumn, as early autumn germinants had higher fitness in some studies than late autumn germinants (Donohue 2002). In the Spanish populations that are late flowering and likely

have the winter annual life cycle, the observed strong seed dormancy is likely a way to escape drought. Yet, Montesinos et al. (2009) observed that in some Spanish populations (although these were different than the ones here) few plants germinated also during late winter and early spring in addition to the normal autumn germination. Suggesting that alternative life-histories are possible also in some populations in Spain. Although, Griffith et al. (2004) have suggested that spring germination could be a failure of the seed to experience dormancy breaking conditions in the autumn. Most of the Norwegian populations were non-dormant and these populations also have the winter annual life cycle. Since these populations also receive more precipitation in the summer, the short growing season probably selects for non-dormant genotypes. The French populations exhibited intermediate levels of dormancy, which could reflect the fact that there are both summer and winter annuals in the populations and the requirements for dormancy can be different. Interestingly, Evans & Ratcliffe (1972) speculated that differences in seed dormancy between some *A. thaliana* populations were an adaptation to diffrences in summer precipitation, although no evidence of this was presented. As the relationship between precipitation and dormancy was the strongest for D25, it could indicate that precipitation is important in determining the time when first seeds can start to germinate.

Studies in other plant species have also found a correlation with environmental variables and seed dormancy. This relationship was found in *Digitaria milanjiana*, where the amount of dormancy was related to precipitation as in this study (Hacker 1984; Hacker et al. 1984). Meyer & Monsen (1991) found that germination patterns in *Artemisia tridentata* were correlated with mean January temperature after short chilling treatments of seeds. A relationship between germination patterns and the environment was found for *Linum perenne* (Meyer & Kitchen 1994), where populations from high altitudes responded to chilling treatments, where as this response was different in populations from lower altitudes. This was also observed for several species of *Penstemon* (Meyer et al. 1995). However, correlation with dormancy and environment has not been always found (Schütz & Milberg 1997) or it was not clear what were the conditions driving seed dormancy differences (Petru & Tielbörger 2008).

In our study, we did not observe a significant relationship with January temperature but summer precipitation. This can be explained by the fact the ecology of *A. thaliana* is different from the species for which January temperatures were found to be important for determining variation in germination. As in the big sagebrush studied by Meyer & Monsen (1991), which germinates during the winter or spring in contrast to autumn germinating *A. thaliana*.

4.2.4 Selection at *DOG1*

Further support for local adaptation in seed dormancy comes from the fact that a signal of local selection was observed at *DOG1*. First, it was observed that Φ_{ST} for *DOG1* was higher than expected (Figure 3.10). Again, it could be argued that this could also happen by chance alone, as neutral F_{ST} is quite high and F_{ST} has rather high variance. However, it was also observed that genetic differentiation at *DOG1* was related to geographic distances and precipitation and these correlations were stronger in some instances than for neutral markers (Table 3.9). Another argument against selection at *DOG1* that could be made is that there could be ascertainment bias in the SNP markers (Clark et al. 2005), which could lower the neutral F_{ST}. But ascertainment bias is not likely to be a major problem, because the SNP markers used were selected from a sample that included genotypes from many different locations, even though high frequency SNPs were selected, there was no geographical bias in SNP ascertainment. Moreover, the microsatellite markers used do not suffer from such a bias and the mean Φ_{ST} of the microsatellite markers is nearly equal to the mean F_{ST} of the SNP markers. For the European populations microsatellite $\Phi_{ST} = 0.660$ and SNP $F_{ST} = 0.621$, such a small difference suggests that SNPs are not biased (see also Supplementary Table B.6, generally SNP F_{ST} is equal or greater than microsatellite Φ_{ST}).

Nevertheless, I further investigated whether *DOG1* is under local selection by examining Φ_{ST} along the chromosome at the position of *DOG1*. There was a clear peak in Φ_{ST} and in between population heterozygosity at the position of *DOG1* (Figure 3.11), indicating that the high Φ_{ST} of *DOG1* is likely to have been caused by selection and not by a lower recombination rate in this part of the chromosome, for example. This is reinforced by the result that between population heterozygosity also peaks at the position of *DOG1* (Charlesworth et al. 1997). Together these results strongly suggest that local selection is responsible for the high genetic differentiation observed at *DOG1*.

4.2.5 *DOG1* is associated with dormancy variation

In order for a gene to be under selection it has to affect the phenotype. Since *DOG1* is a cloned dormancy QTL, one could assume that all variation in this gene is functional. However, the cross from where *DOG1* was cloned and several subsequent analyses used the accession Ler as the other parent. This accession has low dormancy and a *DOG1* allele that has not been found in any other accession so far (M. Debieu, unpublished results). Also, for a population genetic study, alleles that are at high frequencies are often the most interesting ones, and these were not necessarely the ones that were present in the previous QTL mapping populations. Therefore, I did an association analysis in order to confirm that the

variation segregating in our sample is also functional varition.

Perhaps not surprisingly, natural variation in *DOG1* was associated with dormancy (Table 3.10). Several alleles were associated with dormancy, including the haplotypes 18 and 19, common in Norway, which are associated with lowered dormancy. The allele 4 was associated with increased dormancy and it is present both in the French populations and in the Central Asian populations. In the Central Asian populations allele 21 is associated with lower dormancy, as mainly alleles 21 and 4 are segregating in these populations. The allele 15, common in the French populations was associated with dormancy, although only for D25.

To confirm if these associations were real, I did several crosses and checked whether *DOG1* co-segregates with dormancy. Co-segregation could be confirmed for crosses where the alleles 1 and 5; 15 and 5; 1 and 15; 19 and 5; 21 and 4 segregated in the F_2-population (Table 3.11). In the cross, where the parental alleles were 18 and 19, seed dormancy did not segregate in the F_2-population, suggesting that these alleles are functionally identical. I could confirm all the associations that were tested and some functionally different alleles that were not found in the association were found by genetic analysis. Likely, the association lacks statistical power, since some of the alleles are at low frequencies and strong population structure limits segregation. This is also reflected in that Aranzana et al. (2005) observed the known association between *FRIGIDA* and flowering time could only be found if the many different loss-of-function alleles of *FRIGIDA* were grouped together, underscoring the low power to detect the effects of alleles at low frequency. Theoretically, the co-segregation of *DOG1* and dormancy could be caused by nearby linked genes as LD extends quite far in such small F_2-populations. However, this seems unlikely, since in the sample used for association LD is surely much more limited. Since no false positives were found based on confirmed results, it seems likely that some of the alleles with weaker evidence for association are also causal variants.

Combining genetic evidence and population genetic patterns for *DOG1* suggests that there has been selection for lower dormancy in Norway, where the alleles 18 and 19 are at high frequency. Curiously, these haplotypes are linked to haplotype 5 by a long branch (Figure 3.9). It appears that haplotype 5 has spread from Spain to France, and there, at least two alleles, 15 and 1, have been derived independently from haplotype 5. Both of these alleles lower dormancy. Haplotype 4 appears to increase dormancy and from again independently has been derived haplotype 21 which decreases dormancy relative to haplotype 4. There are also indications that alleles that increase dormancy have been derived multiple times, such as allele 6 in Spain and allele 13 in Norway. It is likely that allele 14, which is the same allele that the accession Cvi has, is a strong allele (Bentsink et al. 2006). Evolution in *DOG1* is charecterised by multiple independent mutations that either

increase or decrease dormancy. The structure of the haplotype network or phylogenetic relationships of the alleles are not correlated with their function.

Based on information from neutral markers, evolutionary relationships between the *DOG1* haplotypes and their function, a hypothetical scenario for the evolutionary history of *DOG1* can be proposed. The results obtained from neutral markers (see section 3.2.2 show that the Spanish populations have the most genetic variation and that the French populations are very closely related to the Spanish. On the other hand the Norwegian populations have low genetic diversity and a are separated from the Spanish and the French populations. This is in line with what is known about the phylogeography of *A, thaliana* (Bergelson et al. 1998; Sharbel et al. 2000; Beck et al. 2008; Pico et al. 2008). After the last glaciation *A. thaliana* spread from the Iberian Peninsula to Northern Europe. It appears that only some haplotypes (1, 5 and 15) of *DOG1* have been able to successfully colonise the French populations from the Iberian Peninsula. The common alleles in the French populations are at low frequencies in the Spanish populations. Then mostly likely haplotypes 5 and 1 have colonised Scandinavia, where new mutations that lower dormancy have occurred and have been swept to high frequency by selection.

Recently, Atwell et al. (2010) performed a genome-wide association study in *A. thaliana* in which several phenotypes were measured, seed dormancy among them. This study found that *DOG1* was very weakly associated with seed dormancy, and it was not among the top associations. This contrast with the results presented here. A possible explanation is that in the study of Atwell et al. (2010) only 200 lines were used (and the dormancy phenotype was not available for all of these) versus 343 lines used in this study. Additionally, the lines used by Atwell et al. (2010) were mainly single accessions and not population samples like here. However, in the study of Atwell et al. (2010) *DOG1* was associated with flowering time. However, I could not confirm this, in crosses *DOG1* always segregated seed dormancy but never with flowering time (data not shown). Moreover, the *dog1* mutant has no flowering time phenotype and *DOG1* is expressed only in the seed (Schwab 2008). Thus, the association with *DOG1* and flowering time is a false positive. Since, flowering time and dormancy are both traits which are adaptive, variation within these traits is not randomly distributed but is geographically structured. Indeed, false positive rate for flowering time is much higher than for other traits (Aranzana et al. 2005; Atwell et al. 2010). As dormancy and flowering time can co-vary, for instance, the Norwegian populations are all late flowering and mostly non-dormant. Thus, alleles of *DOG1* that cause low dormancy are completely associated with a late flowering genetic background in Norway.

4.3 Genetic basis of adaptation

DOG1 was shown to be involved in adaptation. In all of crosses performed except one, *DOG1* was found to be segregating. As the parents were chosen on the basis of their *DOG1* alleles this is not surprising. However, in a cross between the genotypes Kon-2-2 and Fet-6 only two large effect QTL were detected. As Kon-2-2 is from Norway and has very low dormancy, while Fet-6 has moderate dormancy and is from France, these genotypes are locally adapted. Yet, only two QTL were detected in this cross, there are likely to be other smaller effect loci, but these two loci are likely to responsible for the majority of adaption. It was perhaps surprising that the other QTL was not *DOG6* as this QTL has been found to have a large effect in many different crosses (Bentsink et al. 2010; Huang et al. 2010). Overall, *DOG6* was found segregating as a small effect QTL only in a single cross in this study.

The results are compatible with the expectation that when there is local adaptation, there should be few QTLs of large effect and larger number of QTLs of small effect (Griswold 2006). However, the low resolution of the QTL mapping probably causes many small effect QTLs to go undetected. Additionally, the QTL effects are probably exaggerated due to Beavis effect (Beavis 1998), so the results must be interpreted with caution. Because of the low number of QTLs detected, distributions of QTL effects cannot be built, hence the data cannot be formally compared to the neutral expectation. These results are in line with QTL effects detected for flowering time in *A. thaliana* (Koornneef et al. 2004), but different from those obtained by Buckler et al. (2009) for maize flowering time, where no large effect QTLs were detected. However, flowering time in maize may be a special case as maize is an obligate outcrosser. Thus flowering time in maize cannot evolve by mutations with large effect as such individuals could not reproduce with others.

4.3.1 Genetic architecture of dormancy

It could be argued that fixed adaptive mutations are expected often to be dominant in diploid organisms, as dominant mutations have much smaller chance of being lost by genetic drift when rare. Here I observed that for *DOG1* there was slight evidence for dominance in crosses between alleles 1 and 5, and likewise between alleles 15 and 5. Yet, this effect was not significant as the 95 % highest posterior density interval for the dominance coefficient barely included zero (Table 3.11). In other crosses there were no indications of dominance. This was also observed by Bentsink et al. (2006) for the Ler and Cvi alleles. Possibly the selfing nature *A. thaliana* renders dominance irrelevant as new mutations can be made homozygous immediately in the next generation.

Why is it that some loci are detected consistantly in crosses, while other are detected only occasionally? It may be that some genes in the genetic network that determines the development of dormancy, occupy parts of the network which can change without causing any pleiotropic effects on other traits. All variation in such genes would be additive, and since additive genetic varition is available to selection, natural selection would use only variation in such genes for local adaptation. This suggests that there might be only a small number of genes that can be responsible for response to selection on seed dormancy. Similar argument has been made for flowering time by Roux et al. (2006). Bentsink et al. (2010) combined several mapping populations to analyse seed dormancy and found that QTLs for seed dormancy contributed only additive variation and no epistatic varition was found. Although, there is a theoretical argument based on allele frequency distributions, that genetic variance is expected to be additive, even if gene action would not be (Hill et al. 2008).

4.3.2 What is the molecular basis of adaptive changes?

There were many different mutations observed between the different *DOG1* alleles (see section 3.4.1) and many of them were amino acid substitutions. Therefore it is not easy to pinpoint the functional mutations between the different alleles. One strong candidate mutation for a functional change is the transposon insertion in haplotype 21. Introns that are close to the promoter are predicted to enhance gene expression and this effect seems to be dependent on the physical size of the intron and its sequence composition (Rose et al. 2008). Haplotype 21 is predicted to lower gene expression *in silico* and this is at least consistent with the fact that haplotype 21 segregates with lower dormancy in a cross between haplotypes 21 and 4. The haplotypes 21 and 4 are very similar in sequence except the insertion. However, further experiments need to be done to show that the insertation is the causal mutation. At present this must remain a hypothesis.

The fact that many non-synonymous substitutions were observed in the first exon of *DOG1* by itself could suggest that these substitutions are not neutral. In particular there were three sites 13, 14 and 15 which have two, four and four amino acid variants segregating respectively (Supplementary Figure A.2). There were also other substitutions that seemed quite interesting. The first methionine has been substituted to lysine in haplotypes 2 and 22 (Supplementary Figure A.2). This event seems to have occurred two times independently (Figure 3.9). However, the haplotypes harbouring this substitution were not null alleles, since some of the genotypes with haplotypes 2 or 22 had some dormancy. These haplotypes were clearly different from the *dog1* mutant. There is another methionine at position 19 that preserves the intact reading frame, but the start of translation in these two alleles is not known. This much amino acid variation seems quite striking, yet

the formal neutrality (McDonald-Kreitman) test was not significant due to many amino acid substitutions when compared to *A. lyrata* (see section 3.4.1). However, it also could be argued that the high amino acid variation in *A. thaliana* could be the result of relaxed selection against slightly deleterious variants due to selfing (Bustamante et al. 2002).

In conclusion no single causal mutation could be identified in this study. In fact there were some hints that both *cis*-regulatory variation and amino acid changes contribute to functional variation in *DOG1*.

4.4 Adaptation in subdivided populations

Generally, when there is local adaptation, genetic differentiation for the trait and QTL is expected to be higher than for neutral markers (Le Corre & Kremer 2003), see also section 1.3. However, these predictions have not always been supported. Using simulations Latta (1998) showed that sometimes allele frequencies at QTL controlling the trait under local selection do not diverge greatly even though phenotypic differences developed between the populations. This is due to the fact that selection build up covariance (between population linkage disequilibrium) between alleles at different QTL (Latta 1998). Hall et al. (2007) argued that this has happened in the European aspen, where there are genetically based phenotypic differences in several phenological traits ($Q_{ST} > F_{ST}$), but low differentiation at genes associated with variation in these traits. Extensive simulations have supported the fact that the build up of allelic covariance can cause this phenomenon in some circumstances, especially if migration rate is high and intensity of selection is from low to moderate (Le Corre & Kremer 2003), but by no means in all cases.

The results presented here contrast with those of Latta (1998) and Hall et al. (2007), as higher differentiation at QTL influencing the trait under local adaptation was observed. There is a simple explanation for this, as low differentiation at QTL is likely to be observed only when migration is high and selection intensity is low or moderate (Le Corre & Kremer 2003). In the *A. thaliana* populations studied here, migration is likely to be rather low. When migration is low, high differentiation at QTLs is also observed often in simulations (Le Corre & Kremer 2003).

Another phenomenon observed often in simulations is that local adaptation also increases genetic differentiation at neutral markers, even in the absence of linkage with QTLs. This is observed especially when migration is low and there is selfing (Le Corre & Kremer 2003), like in the system examined in this study. In fact, Porcher et al. (2006) observed that local selection increased neutral F_{ST} in experimental populations of *A. thaliana*. Selection tends to create linkage disequilibrium between neutral markers and QTL as mating is no longer random (the mating partner of an individual is likely to be an individual with also a high fitness) and

this effect is enhanced in the presence of self-fertilisation. This effect also partly explains why higher Q_{ST} than F_{ST} was not detected, but higher differentiation at QTL was still observed (Porcher et al. 2006).

Recently, by using simulations Yeaman & Guillaume (2009) showed that adopting a continuum-of-alleles model (Crow & Kimura 1964) for the QTL permitted local adaptation in the presence of stronger gene flow than a model with biallelic loci or a Gaussian approximation. As QTL effects can be larger this also permits larger selection coefficients for individual loci. The results presented here suggests that models that allow multiple alleles for QTLs may be more realistic when describing natural populations. This is important, since many models in quantitative genetics have used a biallelic genetic architecture when examining questions such as, how much genetic variation can be maintained by a balance between migration and selection (Phillips 1996; Spichtig & Kawecki 2004).

Additionally, in species where migration is low or occurs at a small scale relative to the range of the species, such as in *A. thaliana* (Platt et al. 2010), recurrent beneficial mutation likely influences patterns of genetic diversity (Pennings & Hermisson 2006). Let $2N_e\mu$ be the population mutation rate into the beneficial allele, then if $2N_e\mu > 0.01$ recurrent introductions of the beneficial mutation start to play a role in adaptation. Additionally if migation between populations is low, such that $2N_e\mu > 2N_em$, where m is the migration rate. It is likely that adaptation in each population then happens from its own mutational origin (Pennings & Hermisson 2006). Formally the results of Pennings & Hermisson (2006) were derived for the haploid situation, but diploidy is unlikely to drastically affect the results. This happens because gene flow is too weak force relative to mutation rate for the same allele to spread to all populations where it would be beneficial. Therefore, in *A. thaliana* and other species where population structure is strong it is not surprising to see that multiple alleles of independent origin give the same or related phenotype. This is seen for the results presented in this study, as functionally similar *DOG1* alleles have been derived multiple times. Moreover, other studies in *A. thaliana* have found similar patterns, there are multiple independent loss of function mutations segregating for *FRI* (Johanson et al. 2000; Le Corre et al. 2002; Le Corre 2005; Toomajian et al. 2006).

4.5 Conclusions

Next I consider the answers to the questions posed in section 1.6. This study found evidence for local adaptation in seed dormancy in *A. thaliana*. Although, this was perhaps not so surprising considering what was already known about seed dormancy and its ecological importance. Strongest evidence for local adaptation was found on a large geographic scale, that is, between regions. However, some

indications of local adaptation also within regions were found, but the evidence for this was not as strong. Local adaptation for seed dormancy seems to be mediated by the amount of precipitation received in the summer months. This again, is in line with the ecological function that seed dormancy has. However, summer precipitation explained only a moderate portion of the total variance, which implies that there must be other unaccounted environmental factors that influence the distribution of seed dormancy.

As for the genetic basis of adaptation, this study demonstrated that *DOG1*, a seed dormancy QTL, is under local selection. Based on effects observed in crosses, *DOG1* is likely to be a major QTL in natural populations, and seems to be involved in local adaptation throughout the distribution of *A. thaliana*. Nevertheless, other QTLs were detected. The number of QTLs detected was not large enough fit their effects to any particular distribution, yet they are compatible with the expectation that adaptation should involve few QTLs of large effect and more loci of smaller effects. The molecular basis of adaptation could not be yet revealed and further experiments are required to discover the causal mutations. There are some hints that both *cis*-regulatory and amino acid variation are involved.

I also investigated what is the best way to compare different markers that may have different mutation rates when estimating genetic differentiation. Methods that take into account the distance between the different alleles were found to be suitable for this task, and by using them different types of markers can be compared.

Bibliography

Abbott, R.J. & Gomes, M.F. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity* 62: 411–418.

Alboresi, A., Gestin, C., Leydecker, M.T., Bedu, M., Meyer, C. & Truong, H.N. 2005. Nitrate, a signal relieving seed dormancy in *Arabidopsis*. *Plant, Cell & Environment* 28: 500–12.

Ali-Rachedi, S., Bouinot, D., Wagner, M.H., Bonnet, M., Sotta, B., Grappin, P. & Jullien, M. 2004. Changes in endogenous abscisic acid levels during dormancy release and maintenance of mature seeds: studies with the Cape Verde Islands ecotype, the dormant model of *Arabidopsis thaliana*. *Planta* 219: 479–88.

Alonso-Blanco, C., Aarts, M.G.M., Bentsink, L., Keurentjes, J.J.B., Reymond, M., Vreugdenhil, D. & Koornneef, M. 2009. What has natural variation taught us about plant development, physiology, and adaptation? *Plant Cell* 21: 1877–1896.

Alonso-Blanco, C., Bentsink, L., Hanhart, C.J., Blankestijn-de Vries, H. & Koornneef, M. 2003. Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics* 164: 711–29.

Alonso-Blanco, C., Blankestijn-de Vries, H., Hanhart, C.J. & Koornneef, M. 1999. Natural allelic variation at seed size loci in relation to other life history traits of *arabidopsis thaliana*. *Proc Natl Acad Sci U S A* 96: 4710–7. Journal Article.

Andersson, L. & Milberg, P. 1998. Variation in seed dormancy among mother plants, populations and years of seed collection. *Seed Science Research* 8: 29–38.

Antonovics, J. & Bradshaw, A.D. 1970. Evolution in closely adjacent plant populations. VIII. Clinal patterns at a mine boundary. *Heredity* 25: 349–362.

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P. & Nordborg, M. 2005. Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics* 1: e60.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, D., M. B.and Weigel, Marjoram, P., Borevitz, J.O., Bergelson, J. & Nordborg, M. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465: 627–631.

Aydin, A., Toliat, M.R., Bahring, S., Becker, C. & Nurnberg, P. 2006. New universal primers facilitate Pyrosequencing. *Electrophoresis* 27: 394–7.

Bakker, E.G., Stahl, E.A., Toomajian, C., Nordborg, M., Kreitman, M. & Bergelson, J. 2006. Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Molecular Ecology* 15: 1405–18.

Balloux, F. 2001. A computer program for the simulation of population genetics. *Journal of Heredity* 92: 301–302.

Balloux, F., Brunner, H., Lugon-Moulin, N., Hausser, J. & Goudet, J. 2000. Microsatellites can be misleading: an empirical and simulation study. *Evolution* 54: 1414–22.

Balloux, F. & Goudet, J. 2002. Statistical properties of population differentiation estimators under step mutation in a finite island model. *Molecular Ecology* 11: 771–783.

Banta, J.A., Dole, J., Cruzan, M.B. & Pigliucci, M. 2007. Evidence of local adaptation to coarse-grained environmental variation in *Arabidopsis thaliana*. *Evolution* 61: 2419–2432.

Baskin, J.M. & Baskin, C.C. 1972. Ecological life cycle and physiological ecology of seed germination of *Arabidopsis thaliana*. *Canadian Journal of Botany* 50: 353–360.

Baskin, J.M. & Baskin, C.C. 1983. Seasonal changes in the germination responses of buried seeds of *Arabidopsis thaliana* and ecological interpretation. *Botanical Gazette* 144: 540–543.

Beaumont, M.A. 2005. Adaptation and speciation: what can F_{ST} tell us. *Trends in Ecology and Evolution* 20: 435–440.

Beaumont, M.A. & Balding, D.J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* 13: 969–80.

Beavis, W.D. 1998. QTL analyses: power, precision, and accuracy. In: *Molecular Dissection of Complex Traits* (A.H. Paterson, ed), pp. 145–162. CRC Press, New York.

Beck, J.B., Schmuths, H. & Schaal, B.A. 2008. Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology* 17: 902–915.

Bentsink, L., Hanson, J., Hanhart, C.J., Blankestijn-de Vries, H., Coltrane, C., Keizer, P., El-Lithy, M.E., Alonso-Blanco, C., de Andrés, M.T., Reymond, M., van Eeuwijk, F., Smeekens, S. & Koornneef, M. 2010. Natural varitation for seed dormancy in *Arabidopsis* is regulated by additive genetic and molecular pathways. *Proceedings of the National Academy of Sciences USA* 107: 4264–4269.

Bentsink, L., Jowett, J., Hanhart, C.J. & Koornneef, M. 2006. Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. *Proceedings of the National Academy of Sciences USA* 25: 25.

Bentsink, L. & Koornneef, M. 2008. Seed dormancy and germination. In: *The Arabidopsis Book* (C.R. Somerville & E.M. Meyerowitz, eds). American Society for Plant Biologists, Rockville, MD.

Bergelson, J., Stahl, E., Dudek, S. & Kreitman, M. 1998. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* 148: 1311–1323.

Bernardo, R. 1993. Estimation of coefficient of coancestry using molecular markers in maize. *Theoretical and Applied Genetics* 85: 1055–1062.

Betancourt, A.J. & Bollback, J.P. 2006. Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. *Current Opinion in Genetics & Development* 16: 618–623.

Bhargava, A. & Fuentes, F.F. 2010. Mutational dynamics of microsatellites. *Molecular Biotechnology* 44: 250–266.

Biere, A. 1991. Parental effects in *Lychnis flos-cuculi*. II: Selection on the time of emergence and seedling performance in the field. *Journal of Evolutionary Biology* 4: 467–486.

Bomblies, K., Yant, L., Laitinen, R.A., Kim, S.T., Hollister, J.D., Warthmann, N., Fitz, J. & Weigel, D. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genetics* 6: e1000890.

Bonnin, I., Prosperi, J.M. & Olivieri, I. 1996. Genetic markers and quantitative genetic variation in *Medicago truncatula* (Leguminosae): a comparative analysis of population structure. *Genetics* 143: 1795–805.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. & Buckler, E.S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633–2635.

Bradshaw, H.D., Otto, K.G., Frewen, B.E., McKay, J.K. & Schemske, D.W. 1998. Quantitative trait loci affecting differences in floral morphology between two species of monkeyflower (Mimulus). *Genetics* 149: 367–382.

Bradshaw, H.D. & Schemske, D.W. 2003. Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature* 426: 176–178.

Broman, K.W., Wu, H., Saunak, S. & Churchill, G.A. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19: 889–890.

Brownstein, M.J., Carpten, M.D. & Smith, J.R. 1996. Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques* 20: 1004–1010.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., Goodman, M.M., Harjes, C., Guill, K., Kroon, D.E., Larsson, S., Lepak, N.K., Li, H., Mitchell, S.E., Pressoir, G., Peiffer, J.A., Rosas, M.O., Rocheford, T.R., Romay, M.C., Romero, S., Salvo, S., Villeda, H.S., Sofia da Silva, H., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S. & McMullen, M.D. 2009. The genetic architecture of maize flowering time. *Science* 325: 714–718.

Burke, J.M., Tang, S., Knapp, S.J. & Rieseberg, L.H. 2002. Genetic analysis of sunflower domestication. *Genetics* 161: 1257–1267.

Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D. & Hartl, D.L. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416: 531–4.

Cadman, C.S., Toorop, P.E., Hilhorst, H.W. & Finch-Savage, W.E. 2006. Gene expression profiles of *Arabidopsis* Cvi seeds during dormancy cycling indicate a common underlying dormancy control mechanism. *Plant Journal* 46: 805–22.

Calabrese, P. & Sainudiin, R. 2005. Models of microsatellite evolution. In: *Statistical methods in Molecular Evolution* (R. Nielsen, ed), pp. 289–305. Springer, New York.

Carrera, E., Holman, T., Medhurst, A., Dietrich, D., Footitt, S., Theodoulou, F.L. & Holdsworth, M.J. 2007. Seed after-ripening is a discrete developmental pathway associated with specific gene networks in *Arabidopsis*. *Plant Journal* 53: 214–224.

Carreras-Carbonell, J., Macpherson, E. & Pascual, M. 2006. Population structure within and between subspecies of the Mediterranean triplefin fish *Tripterygion delaisi* revealed by highly polymorphic microsatellite loci. *Molecular Ecology* 15: 3527–3539.

Carroll, S.B. 2005. Evolution at two levels: On genes and form. *PLoS Biology* 3: e245.

Carroll, S.B. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134: 25–36.

Casal, J.J. & Sanchez Rodolfo, A. 1998. Phytochromes and seed germination. *Seed Science Research* 8: 317–329.

Chapuis, E., Trouve, S., Facon, B., Degen, L. & Goudet, J. 2007. High quantitative and no molecular differentiation of a freshwater snail (*Galba truncatula*) between temporary and permanent water habitats. *Molecular Ecology* 16: 3484–3496.

Charlesworth, B., Nordborg, M. & Charlesworth, D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of gene diversity in subdivided populations. *Genetical Research* 70: 155–174.

Churchill, G.A. & Doerge, R.W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963–971.

Clark, A.G., Hubisz, M.J., Bustamante, C., Williamson, S. & Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15: 1496–1502.

Clausen, J., Keck, D.D. & Hiesey, W.M. 1941. Regional differentiation in plant species. *American Naturalist* 75: 231–250.

Clauss, M.J. & Mitchell-Olds, T. 2006. Population genetic structure of *Arabidopsis lyrata* in Europe. *Molecular Ecology* 15: 2753–66.

Clement, M., Posada, D. & Crandall, K.A. 2000. TCS: a computer program to estimate gene genealogies. *Molecular Ecology* 9: 1657–9.

Clerkx, E.J., El-Lithy, M.E., Vierling, E., Ruys, G.J., Blankestijn-De Vries, H., Groot, S.P., Vreugdenhil, D. & Koornneef, M. 2004. Analysis of natural allelic variation of *Arabidopsis* seed germination and seed longevity traits between the accessions Landsberg erecta and Shakdara, using a new recombinant inbred line population. *Plant Physiology* 135: 432–43.

Cockerham, C.C. 1969. Variance of gene frequencies. *Evolution* 23: 72–84.

Collins, S. & de Meaux, J. 2009. Adaptation to different rates of environmental change in *Chlamydomonas*. *Evolution* 63: 2952–2965.

Collins, S., de Meaux, J. & Acquisti, C. 2007. Adaptive walks towards a moving optimum. *Genetics* 176: 1101–1118.

Cowperthwaite, M.C., Bull, J. & Meyers, L.A. 2005. Distributions of beneficial fitness effects in RNA. *Genetics* 170: 1449–1457.

Crow, J.F. 2002. Here's to Fisher, additive genetic variance, and the fundamental theorem of natural selection. *Evolution* 56: 1313–1316.

Crow, J.F. & Kimura, M. 1964. The theory of genetic loads. In: *Proceedings of the XIth International Congress of Genetics*, pp. 495–505.

Darwin, C. 1859. *The Origin of Species*. John Murray, London.

Doebley, J. 2004. The genetics of maize evolution. *Annual Review of Genetics* 38: 37–59.

Donohue, K. 2002. Germination timing influences natural selection on life-history characters in *Arabidopsis thaliana*. *Ecology* 83: 1006–1016.

Donohue, K., Dorn, L., Griffith, C., Kim, E., Aguilera, A., Polisetty, C.R. & Schmitt, J. 2005a. Environmental and genetic influences on the germination of *Arabidopsis thaliana* in the field. *Evolution* 59: 740–757.

Donohue, K., Dorn, L., Griffith, C., Kim, E., Aguilera, A., Polisetty, C.R. & Schmitt, J. 2005b. The evolutionary ecology of seed germination of *Arabidopsis thaliana*: Variable natural selection on germination timing. *Evolution* 59: 758–770.

Donohue, K., Heschel, M.S., Butler, C.M., Barua, D., Sharrock, R.A., Whitelam, G.C. & Chiang, G.C.K. 2007a. Diversification of phytochrome contributions to germination as a function of seed-maturation environment. *New Phytologist* 177: 367–379.

Donohue, K., Heschel, M.S., Chiang, G.C.K., Butler, C. & Barua, D. 2007b. Phytochrome mediates germination responses to multiple seasonal cues. *Plant, Cell & Environment* 30: 202–212.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5: 435–445.

Ellner, S. 1986. Germination dimorphisms and parent-offspring conflict in seed germination. *Journal of Theoretical Biology* 123: 173–185.

Evans, A. & Cabin, R.J. 1995. Can dormancy affect the evolution of post-germination traits? The case of *Lesquerella fendleri*. *Ecology* 76: 344–356.

Evans, J. & Ratcliffe, D. 1972. Variation in 'after-ripening' of seeds of *Arabidopsis thaliana* and its ecological significance. *Arabidopsis Information Service* 9: 3–5.

Evans, M.E.K. & Dennehy, J.J. 2005. Germ banking: bet-hedging and variable release from egg and seed dormancy. *Quarterly Review of Biology* 80: 431–451.

Evans, M.E.K., Ferrière, R., Kane, J.M. & Venable, D.L. 2007. Bet hedging via seed banking in desert evening primroses (Oenothera, Onagraceae): Demographic evidence from natural populations. *American Naturalist* 169: 184–194.

Excoffier, L. 2007. Analysis of population subdivision. In: *Handbook of statistical genetics* (D.J. Balding, M. Bishop & C. Cannings, eds). Wiley.

Excoffier, L., Smouse, P.E. & Quattro, J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479–91.

Eyre-Walker, A. & Keightley, P.D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* 8: 610–618.

Fakhrai-Rad, H., Pourmand, N. & Ronaghi, M. 2002. Pyrosequencing^{TM}: An accurate detection platform for single nucleotide polymorphisms. *Human Mutation* 19: 479–485.

Feenstra, B., Skovgaard, I.M. & Broman, K.W. 2006. Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. *Genetics* 173: 2269–2282.

Finch-Savage, W.E., Cadman, C.S.C., Toorop, P.E., Lynn, J.R. & Hilhorst, H.W.M. 2007. Seed dormancy release in *Arabidopsis* Cvi by dry after-ripening, low temperature, nitrate and light shows common quantitative patterns of gene expression directed by environmentally specific sensing. *Plant Journal* 51: 60–78.

Finch-Savage, W.E. & Leubner-Metzger, G. 2006. Seed dormancy and the control of germination. *New Phytologist* 171: 501–523.

Fisher, R.A. 2003. *The Genetical Theory of Natural Selection. A Complete Variorum Edition*. Oxford University Press, Inc., Oxford.

Gardner, A. 2009. Adaptation as organism design. *Biology Letters* 5: 861–864.

Gillespie, J.H. 1983. Molecular evolution over the mutational landscape. *Evolution* 38: 1116–1129.

Goudet, J. 2001. *FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). - Available on the web at http://www2.unil.ch/popgen/softwares/fstat.htm.*

Goudet, J. 2005. hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* 5: 184–186.

Goudet, J. & Büchi, L. 2006. The effects of dominance, regular inbreeding and sampling design on Q_{ST}, an estimator of population differentiation for quantitative traits. *Genetics* 172: 1337–47.

Goudet, J. & Martin, G. 2007. Under neutrality, $Q_{ST} \leq F_{ST}$ when there is dominance in an island model. *Genetics* 176: 1371–1374.

Griffith, C., Kim, E. & Donohue, K. 2004. Life-history variation and adaptation in the historically mobile plant *Arabidopsis thaliana* (Brassicaceae) in North America. *American Journal of Botany* 91: 837–849.

Griswold, C.K. 2006. Gene flow's effect on the genetic architecture of a local adaptation and its consequences for QTL analyses. *Heredity* 96: 445–453.

Gross, K.L. & Smith, A.D. 1991. Seed mass and emergence time effects on performance of *Panicum dichotomiflorum* Michx. across environments. *Oecologia* 87: 270–278.

Hacker, J.B. 1984. Genetic variation in seed dormancy in *Digitaria milanjiana* and its correlation with rainfall at the collection site. *Journal of Applied Ecology* 21: 947–959.

Hacker, J.B., Andrew, M.H., McIvor, J.G. & Mott, J.J. 1984. Evaluation in contrasting climates of dormancy characteristics of seed of *Digitaria milanjiana*. *Journal of Applied Ecology* 21: 961–969.

Haley, C.S. & Knott, S.A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315–324.

Hall, D., Luquez, V., Garcia, V.M., St Onge, K.R., Jansson, S. & Ingvarsson, P.K. 2007. Adaptive population differentiation in phenology across a latitudinal gradient in european aspen (*Populus tremula*, L.): A comparison of neutral markers, candidate genes and phenotypic traits. *Evolution* 61: 2849–2860.

Hall, M.C. & Willis, J.H. 2006. Divergent selection on flowering time contributes to local adaptation in *Mimulus guttatus* populations. *Evolution* 60: 2466–2477.

Hall, T.A. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucl Acids Symp Ser* 41: 95–98. URL `http://www.mbio.ncsu.edu/BioEdit/page2.html`.

Hardy, O.J., Charbonnel, N., Freville, H. & Heuertz, M. 2003. Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. *Genetics* 163: 1467–82.

Hardy, O.J. & Vekemans, X. 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* 2: 618–620.

He, F., Kang, D., Ren, Y., Qu, L.J., Zhen, Y. & Gu, H. 2007. Genetic diversity of the natural populations of *Arabidopsis thaliana* in China. *Heredity* 99: 423–431.

Hedrick, P.W. 1999. Highly variable loci and their interpretation in evolution and conservation. *Evolution* 53: 313–318.

Hedrick, P.W. 2005. A standardized genetic differentiation measure. *Evolution* 59: 1633–1638.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.

Hijmans, R.J., Guarino, L., Cruz, M. & Rojas, E. 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter* 127: 15–19.

Hill, W.G., Goddard, M.E. & Visscher, P.M. 2008. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* 4: e1000008.

Hoekstra, H.E. & Coyne, J.A. 2007. The locus of evolution: Evo devo and the genetics of adaptation. *Evolution* 61: 995–1016.

Hoffmann, M.H. 2002. Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae). *Journal of Biogeography* 29: 125–134.

Hoffmann, M.H., Bremer, M., Schneider, K., Burger, F., Stolle, E. & Moritz, G. 2003. Flower visitors in a natural population of *Arabidopsis thaliana*. *Plant Biology* 5: 491–494.

Holdsworth, M.J., Bentsink, L. & Soppe, W.J.J. 2008. Molecular networks regulating *Arabidopsis* seed maturation, after-ripening, dormancy and germination. *New Phytologist* 179: 33–54.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65–70.

Holsinger, K.E. & Weir, B.S. 2009. Genetics in georaphically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* 10: 639–650.

Huang, X., Schmitt, J., Dorn, L., Griffith, C., Effgen, S., Takao, S., Koornneef, M. & Donohue, K. 2010. The earliest stages of adaptation in an experimental plant population: strong selection on QTLs for seed dormancy. *Molecular Ecology* 19: 1335–1351.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337–8.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R. & Dean, C. 2000. Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* 290: 344–347.

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

Jost, L. 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* 17: 4015–4026.

Jost, L. 2009. D vs. G_{ST}: Response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Molecular Ecology* 18: 2088–2091.

Kalinowski, S.T. 2002. Evolutionary and statistical properties of three genetic distances. *Molecular Ecology* 11: 1263–73.

Kalisz, S. 1986. Variable selection on the timing of germination in *Collinsia verna* (Scrophulariaceae). *Evolution* 40: 479–491.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. & Eskin, E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.

Kawecki, T.J. & Ebert, D. 2004. Conceptual issues in local adaptation. *Ecology Letters* 7: 1225–1241.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. & Nordborg, M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 39: 1151–1155.

Kimmel, M., Chakraborty, R., King, J.P., Bamshad, M., Watkins, W.S. & Jorde, L.B. 1998. Signatures of population expansion in microsatellite repeat data. *Genetics* 148: 1921–1930.

Kisdi, E. 2002. Dispersal: risk spreading versus local adaptation. *American Naturalist* 159: 579–596.

Kobayashi, Y. & Yamamura, N. 2000. Evolution of seed dormancy due to sib competition: Effect of dispersal and inbreeding. *Journal of Theoretical Biology* 202: 11–24.

Koeller, P., Fuentes-Yaco, C., Platt, T., Sathyendranath, S., Richards, A., Ouellet, P., Orr, D., Skuladottir, U., Wieland, K., Savard, L. & Aschan, M. 2009. Basin-scale coherence in phenology of shrimps and phytoplankton in the north Atlantic Ocean. *Science* 324: 791–793.

Komeda, Y. 2004. Genetic regulation of time to flower in *Arabidopsis thaliana*. *Annual Review of Plant Biology* 55: 521–535.

Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. 2004. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annual Review of Plant Biology* 55: 141–72.

Kopp, M. & Hermisson, J. 2007. Adaptation of a quantitative trait to a moving optimum. *Genetics* 176: 715–719.

Kopp, M. & Hermisson, J. 2009. The genetic gasis of phenotypic adaptation I: Fixation of beneficial mutations in the moving optimum model. *Genetics* 182: 233–249.

Kugler, I. 1951. Untersuchungen über das Keimverhalten einiger Rassen von *Arabidopsis thaliana* (L.) Heynh. Ein Beitrag zum Problem der Lichtkeimung. *Beiträge zur Biologie der Plfanzen* 28: 211–243.

Kuittinen, H., Mattila, A. & Savolainen, O. 1997. Genetic variation at marker loci and in quantitative traits in natural populations of *Arabidopsis thaliana*. *Heredity* 79 (Pt 2): 144–52.

Laibach, F. 1951. Über sommer- und winterannuelle Rassen von *Arabidopsis thaliana* (L.) Heynh. Ein Beitrag zum Ätiologie der Blütenbildung. *Beiträge zur Biologie der Plfanzen* 28: 173–210.

Lande, R. 1992. Neutral theory of quantitative genetic variance in an island model with local extinction and colonization. *Evolution* 46: 381–389.

Lander, E.S. & Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences USA* 84: 2363–2367.

Latta, R.G. 1998. Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *American Naturalist* 151: 283–292.

Le Corre, V. 2005. Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits. *Molecular Ecology* 14: 4181–92.

Le Corre, V. & Kremer, A. 2003. Genetic variability at neutral markers, quantitative trait land trait in a subdivided population under selection. *Genetics* 164: 1205–19.

Le Corre, V., Roux, F. & Reboud, X. 2002. DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: Extensive nonsynonymous variation is consistent with local selection for flowering time. *Molecular Biology and Evolution* 19: 1261–1271.

Leimu, R. & Fischer, M. 2008. A meta-analysis of local adaptation in plants. *PLoS ONE* 3: e4010.

Leinonen, T., O'Hara, R.B., Cano, J.M. & Merilä, J. 2008. Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology* 21: 1–17.

Leon-Kloosterziel, K.M., van de Bunt, G.A., Zeevaart, J.A. & Koornneef, M. 1996. *Arabidopsis* mutants with a reduced seed dormancy. *Plant Physiology* 110: 233–40.

Liu, Y., Koornneef, M. & Soppe, W.J.J. 2007. The absence of histone H2B monoubiquitination in the *Arabidopsis hub1* (*rdo4*) mutant reveals a role for chromatin remodeling in seed dormancy. *Plant Cell* 19: 433–444.

Lundemo, S., Falahati-Anbaran, M. & Stenøien, H.K. 2009. Seed banks cause elevated generation times and effective population sizes of *Arabidopsis thaliana* in northern Europe. *Molecular Ecology* 18: 2798–2811.

Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 10: 325–337.

Lynch, M. 1988a. Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution* 5: 584–599.

Lynch, M. 1988b. The rate of polygenic mutation. *Genetical Research* 51: 137–148.

Lynch, M. & Walsh, B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland.

Lynch, V.J. & Wagner, G. 2008. Resurrecting the role of transcription factor change in the evolution of development. *Evolution* 62: 2131–2154.

Macnair, M.R. 1987. Heavy metal tolerance in plans – a model evolutionary system. *Trends in Ecology and Evolution* 2: 354–359.

Marks, M. & Prince, S. 1981. Influence of germination date on survival and fecundity in wild lettuce *Lactuca serriola*. *Oikos* 36: 326–330.

Meirmans, P.G. 2006. Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* 60: 2399–2402.

Merilä, J. & Crnokrak, P. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology* 14: 892–903.

Meyer, S.E. & Kitchen, S.G. 1994. Life history variation in blue flax (*Linum perenne*: Linaceae): Seed germination phenology. *American Journal of Botany* 81: 528–535.

Meyer, S.E., Kitchen, S.G. & Carlson, S.L. 1995. Seed germination timing patterns in intermountain *Penstemon* (Scrophulariaceae). *American Journal of Botany* 82: 377–389.

Meyer, S.E. & Monsen, S.B. 1991. Habitat correlated variation in mountain big sagebrush (*Artemisia tridentata* spp. Vaseyana) seed germination patterns. *Ecology* 72: 739–742.

Michalakis, Y. & Excoffier, L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142: 1061–1064.

Miller, J.R., Wood, B.P. & Hamilton, M.B. 2008. F_{ST} and Q_{ST} under neutrality. *Genetics* 180: 1023–1037.

Mitchell-Olds, T. & Schmitt, J. 2006. Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. *Nature* 441: 947–952.

Miyashita, N.T., Kawabe, A. & Innan, H. 1999. DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. *Genetics* 152: 1723–1731.

Montesinos, A., Tonsor, S.J., Alonso-Blanco, C. & Pico, F.X. 2009. Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from contrasting native environments. *PLoS ONE* 4: e7213.

Muller, M.H., Leppälä, J. & Savolainen, O. 2007. Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. *Heredity* 100: 47–58.

Munir, J., Dorn, L.A., Donohue, K. & Schmitt, J. 2001. The effect of maternal photoperiod on seasonal dormancy in *Arabidopsis thaliana* (Brassicaceae). *American Journal of Botany* 88: 1240–1249.

Napp-Zinn, K. 1976. Population genetical and gene geographical aspects of germination and flowering in *Arabidopsis thaliana*. *Arabidopsis Information Service* 13: 30.

Naylor, J.M. & Jana, S. 1976. Genetic adaptation for seed dormancy in *Avena fatua*. *Canadian Journal of Botany* 54: 306–312.

Nei, M. & Chesser, R.K. 1983. Estimation of fixation indices and gene diversities. *Annals of Human Genetics* 47: 253–259.

Neuenschwander, S., Hospital, F., Guillaume, F. & Goudet, J. 2008. quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics* 24: 1552–1553.

Nordborg, M. & Bergelson, J. 1999. The effect of seed and rosette cold treatment on germination and flowering time in some *Arabidopsis thaliana* (Brassicaceae) ecotypes. *American Journal of Botany* 86: 470–475.

Nordborg, M. & Donnelly, P. 1997. The coalescent process with selfing. *Genetics* 146: 1185–95.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J.D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M. & Bergelson, J. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* 3: e196.

O'Hara, R.B. & Merilä, J. 2005. Bias and precision in Q_{ST} estimates: problems and some solutions. *Genetics* 171: 1331–1339.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, R.B., Henry, M. & Stevens, M.H.H. 2007. *vegan: Community Ecology Package. R package version 1.8-7.* URL `http://cran.r-project.org, http://cc.oulu.fi/~jarioksa/`.

O'Reilly, P., Canino, M., Bailey, K. & Bentzen, P. 2004. Inverse relationship between F_{ST} and microsatellite polymorphism in the marine fish walleye pollock (*Theragra chalcogramma*): implications for resolving weak population structure. *Molecular Ecology* 13: 1799–1814.

Orr, H.A. 1998. The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. *Evolution* 52: 935–949.

Orr, H.A. 2002. The population genetics of adaptation: The adaptation of DNA sequences. *Evolution* 56: 1317–1330.

Orr, H.A. 2003. The distribution of fitness effects among beneficial mutations. *Genetics* 163: 1519–1526.

Orr, H.A. 2005. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* 6: 119–127.

Pennings, P.S. & Hermisson, J. 2006. Soft sweeps II-molecular population genetics of adaptation from recurrent mutation or migration. *Molecular Biology and Evolution* 23: 1076–1084.

Petru, M. & Tielbörger, K. 2008. Germination behaviour of annual plants under changing climatic conditions: separating local and regional environmental effects. *Oecologia* 155: 717–728.

Phillips, P.C. 1996. Maintenance of polygenic variation via a migration-selection balance under uniform selection. *Evolution* 50: 1334–1339.

Pico, F.X., Mendez-Vigo, B., Martinez-Zapater, J.M. & Alonso-Blanco, C. 2008. Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the iberian peninsula. *Genetics* 180: 1009–1021.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K., M., D., Holub, E.B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J. & Borevitz, J.O. 2010. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics* 6: e1000843.

Porcher, E., Giraud, T., Goldringer, I. & Lavigne, C. 2004. Experimental demonstration of a causal relationship between heterogeneity of selection and genetic differentiation in quantitative traits. *Evolution* 58: 1434–1445.

Porcher, E., Giraud, T. & Lavigne, C. 2006. Genetic differentiation of neutral markers and quantitative traits in predominantly selfing metapopulations: confronting theory and experiments with *Arabidopsis thaliana*. *Genetical Research* 87: 1–12.

Posada, D. & Crandall, K.A. 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution* 16: 37–45.

R Development Core Team 2006. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Ratcliffe, D. 1976. Germination characteristics and their inter- and intrapopulation variability in *Arabidopsis*. *Arabidopsis Information Service* 13: 34.

Raz, V., Bergervoet, J.H. & Koornneef, M. 2001. Sequential steps for developmental arrest in *Arabidopsis* seeds. *Development* 128: 243–252.

Roach, D.A. & Wulff, R.D. 1987. Maternal effects in plants. *Annual Review of Ecology and Systematics* 18: 209–235.

Rogers, A.R. & Harpending, H.C. 1983. Population structure and quantitative characters. *Genetics* 105: 985–1002.

Rose, A.B., Elfersi, T., Parra, G. & Korf, I. 2008. Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. *Plant Cell* 20: 543–551.

Roux, F., Touzet, P., Cuguen, J. & Le Corre, V. 2006. How to be early flowering: an evolutionary perspective. *Trends in Plant Science* 11: 375–81.

RoyChoudhury, A. & Stephens, M. 2007. Fast and accurate estimation of the population-scaled mutation rate, Θ, from microsatellite genotype data. *Genetics* 176: 1363–1366.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496–2497.

Satterthwaite, W.H. 2010. Competition for space can drive the evolution of dormancy in a temporally invariant environment. *Plant Ecology* online early: DOI:10.1007/s11258–009–9696–y.

Schemske, D.W. 1984. Population structure and local selection in *Impatiens pallida* (Balsaminaceae), a selfing annual. *Evolution* 38: 817–832.

Schultz, S.T., Lynch, M. & Willis, J.H. 1999. Spontaneous deleterious mutation in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences USA* 96: 11393–11398.

Schütz, W. & Milberg, P. 1997. Seed dormancy in *Carex canescens*: regional differences and ecological consequences. *Oikos* 78: 420–428.

Schwab, M. 2008. *Identification of novel seed dormancy mutants in Arabidopsis thaliana and molecular and biochemical characterization of the seed dormancy gene DOG1.* Ph.D. thesis, University of Cologne, Germany.

Shankar, P.C., Ito, S., Kato, M., Matsui, M., Kodaira, R., Hayashida, N. & Okazaki, M. 2001. Analysis of Tag1-like elements in *Arabidopsis thaliana* and their distribution in other plants. *DNA Research* 8: 107–113.

Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jonsson, B., Schluter, D. & Kingsley, D.M. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428: 717–23.

Sharbel, T.F., Haubold, B. & Mitchell-Olds, T. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology* 9: 2109–2118.

Shinomura, T. 1997. Phytochrome regulation of seed germination. *Journal of Plant Research* 110: 151–161.

Simons, A.M. 2009. Fluctuating natural selection accounts for the evolution of diversification bet-hedging. *Proceedings of the Royal Society B: Biological Sciences* 276: 1987–1992.

Slatkin, M. 1974. Hedging one's evolutionary bets. *Nature* 250: 704–705.

Slatkin, M. 1991. Inbreeding coefficients and coalescence times. *Genet Res* 58: 167–175.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462.

Spichtig, M. & Kawecki, T. 2004. The maintenance (or not) of polygenic variation by soft selection in heterogeneous environments. *American Naturalist* 164: 70–84.

Spitze, K. 1993. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics* 135: 367–74.

Stenøien, H.K., Fenster, C.B., Tonteri, A. & Savolainen, O. 2005. Genetic variability in natural populations of *Arabidopsis thaliana* in northern Europe. *Molecular Ecology* 14: 137–48.

Stern, D.L. 2000. Evolutionary developmental biology and the problem of variation. *Evolution* 54: 1079–1091.

Stern, D.L. & Orgogozo, V. 2008. The loci of evolution: How predictable is genetic evolution? *Evolution* 62: 2155–2177.

Stich, B., Mohring, J., Piepho, H.P., Heckenberger, M., Buckler, E.S. & Melchinger, A.E. 2008. Comparison of mixed-model approaches for association mapping. *Genetics* 178: 1745–1754.

Stinchcombe, J.R., Weinig, C., Ungerer, M., Olsen, K.M., Mays, C., Halldorsdottir, S.S., Purugganan, M.D. & Schmitt, J. 2004. A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. *Proceedings of the National Academy of Sciences USA* 101: 4712–4717.

Storz, J.F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14: 671–688.

Sun, J.X., Mullikin, J.C., Patterson, N. & Reich, D.E. 2009. Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution* 26: 1017–1027.

Symonds, V.V. & Lloyd, A.M. 2003. An analysis of microsatellite loci in *Arabidopsis thaliana*: mutational dynamics and application. *Genetics* 165: 1475–88.

Toomajian, C., Hu, T.T., Aranzana, M.J., Lister, C., Tang, C., Zheng, H., Zhao, K., Calabrese, P., Dean, C. & Nordborg, M. 2006. A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biology* 4: e137.

Turelli, M. 1984. Heritable genetic variation via mutation–selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology* 25: 138–193.

Turesson, G. 1922. The genotypical response of the plant species to the habitat. *Hereditas* 3: 211–350.

Turesson, G. 1925. The plant species in relation to habitat and climate. *Hereditas* 6: 147–236.

van Der Schaar, W., Alonso-Blanco, C., Leon-Kloosterziel, K.M., Jansen, R.C., van Ooijen, J.W. & Koornneef, M. 1997. QTL analysis of seed dormancy in *Arabidopsis* using recombinant inbred lines and MQM mapping. *Heredity* 79 (Pt 2): 190–200.

Venable, D.L. & Brown, J.S. 1988. The selective interactions of dispersal, dormancy and seed size as adaptations for reducing risk in variable environments. *American Naturalist* 131: 360–384.

Venable, D.L. & Lawlor, L. 1980. Delayed germination and dispersal in desert annuals: Escape in time and space. *Oecologia* 46: 272–282.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*, 4th edn. Springer, New York.

Vitalis, R., Glémin, S. & Olivieri, I. 2004. When genes go to sleep: the population genetic consequences of seed dormancy and monocarpic perenniality. *American Naturalist* 163: 295–311.

Warthmann, N., Fitz, J. & Detlef, W. 2007. MSQT for choosing SNP assays from multiple DNA alignments. *Bioinformatics* 23: 2784–2787.

Weir, B.S. & Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Whitlock, M.C. 1999. Neutral additive genetic variance in a metapopulation. *Genet Res* 74: 215–21.

Whitlock, M.C. 2008. Evolutionary inference from Q_{ST}. *Molecular Ecology* 17: 1885–1896.

Wilczek, A.M., Roe, J.L., Knapp, M.C., Cooper, M.D., Lopez-Gallego, C., Martin, L.J., Muir, C.D., Sim, S., Walker, A., Anderson, J., Egan, J.F., Moyers, B.T., Petipas, R., Giakountis, A., Charbit, E., Coupland, G., Welch, S.M. & Schmitt, J. 2009. Effects of genetic perturbation on seasonal life history plasticity. *Science* 323: 930–934.

Wright, S. 1931. Evolution in mendelian populations. *Evolution* 16: 97–159.

Wright, S. 1951. The genetical structure of populations. *Annals of Eugenics* 15: 323–354.

Yeaman, S. & Guillaume, F. 2009. Predicting adaptation under migration load: The role of genetic skew. *Evolution* 63: 2926–2938.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. & Buckler, E.S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38: 203–208.

Appendix A

Supplementary figures

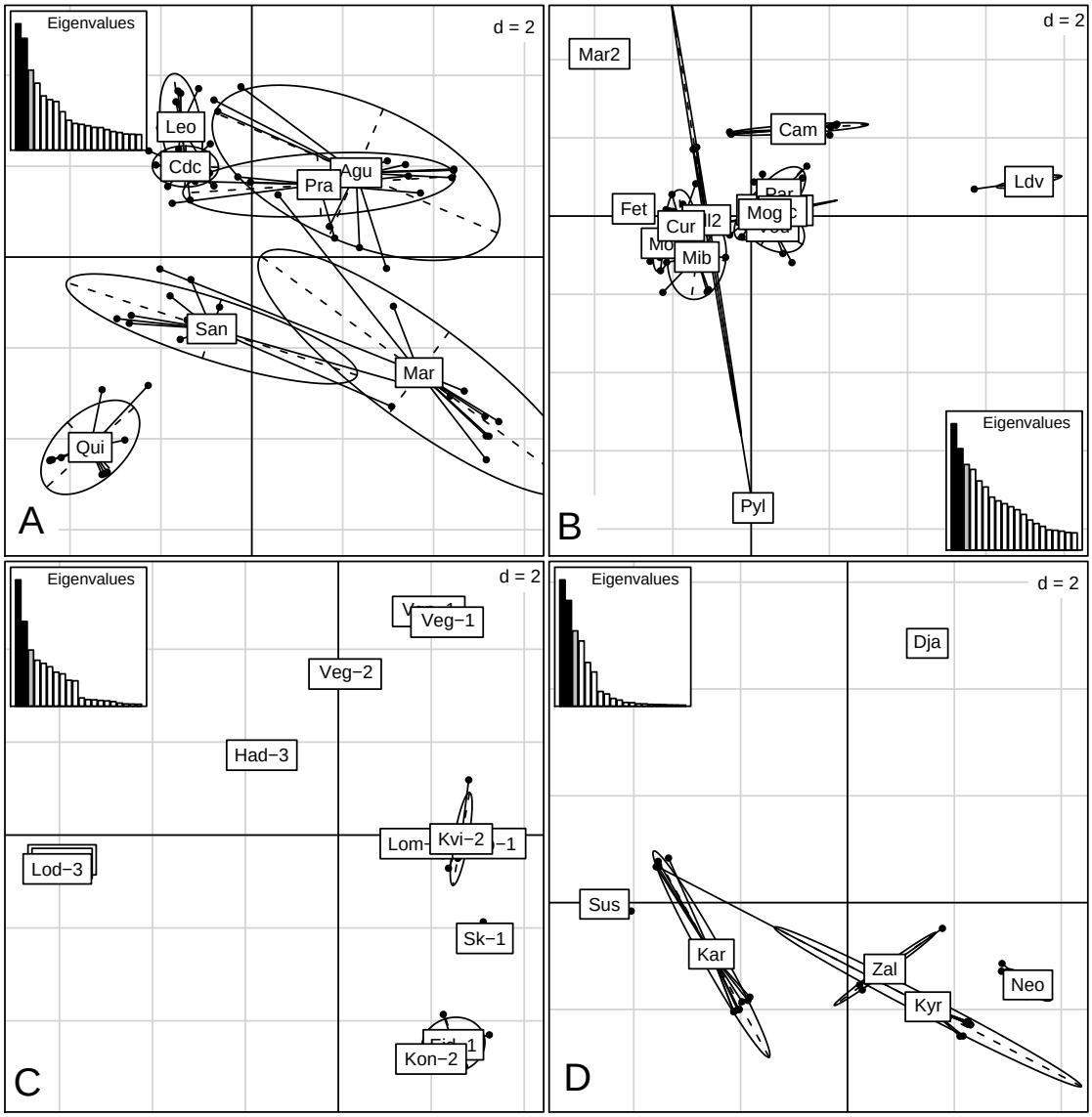

Figure A.1: PCA within each of the regions. The different regions, Spain, France, Norway and Central Asia are in panels A, B, C and repectively. X-axis is the first principal component and Y-axis the second principal component. Inset shows a barplot of eigenvalues (principal components) displaying the relative contribution of each component to the total genetic variance.

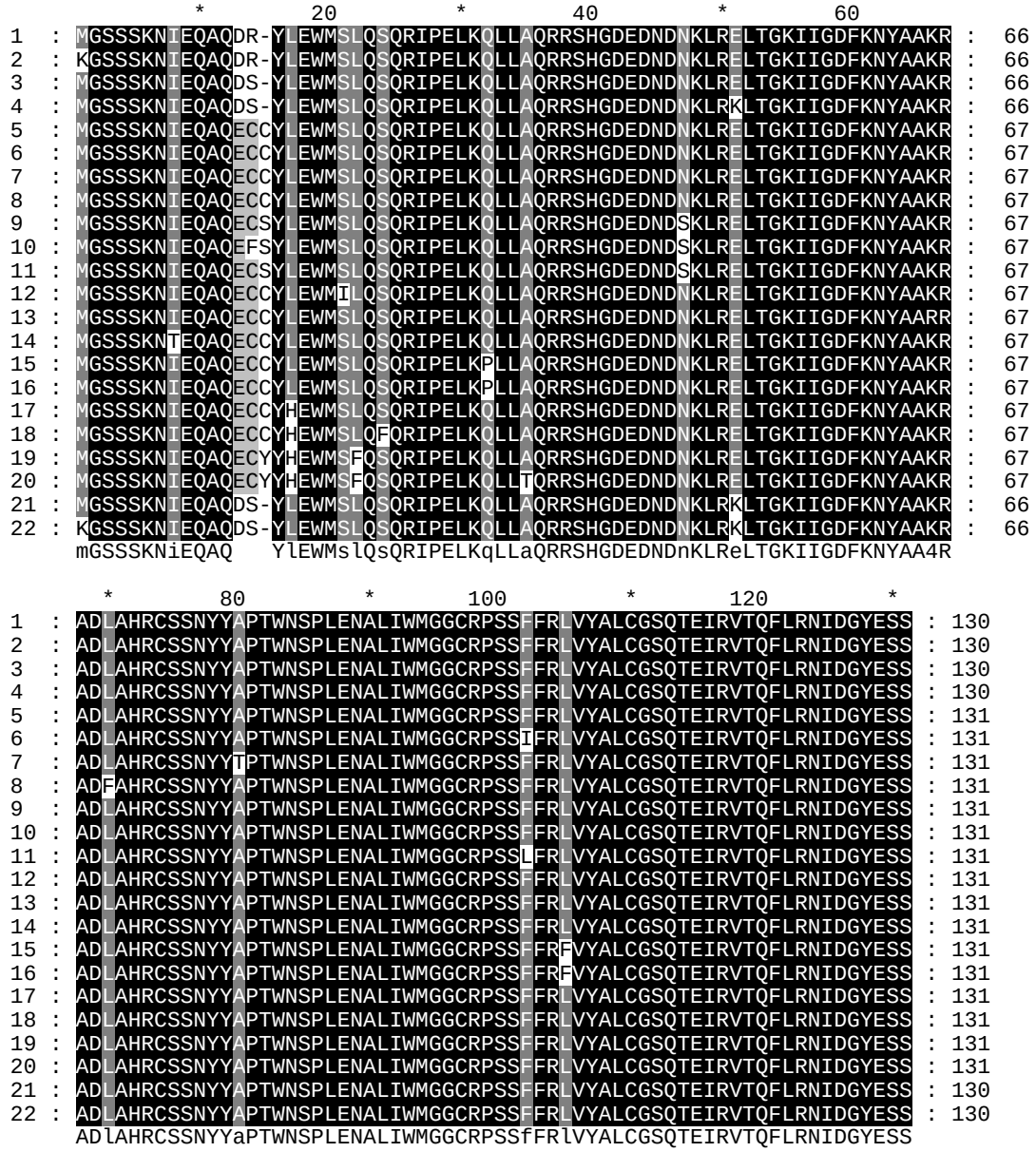

Figure A.2: Amino acid alignment of exon 1 of *DOG1*. Numbers on the left indicate different haplotypes.

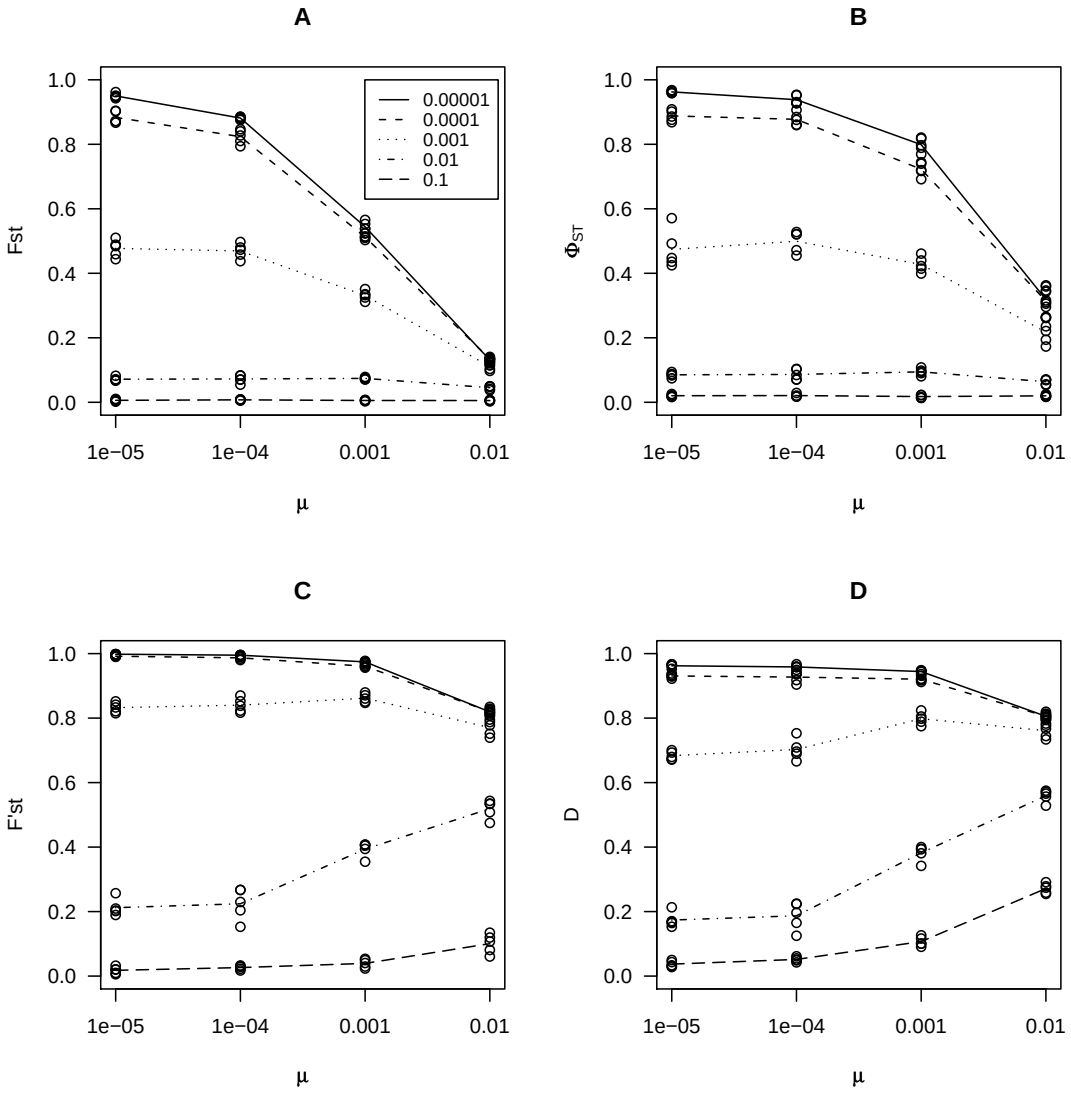

Figure A.3: Results of forward simulations with selfing rate of 0.9. The mixed mutation model was used. The panels show the relationships with different estimators of genetic differentiation to different migration and mutation rates. X-axis is the mutation rate; different types of lines correspond to different migration rates as in the legend in panel A. Different estimators are F_{ST}, Φ_{ST}, F'_{ST} and D in panels A, B, C, and D respectively.

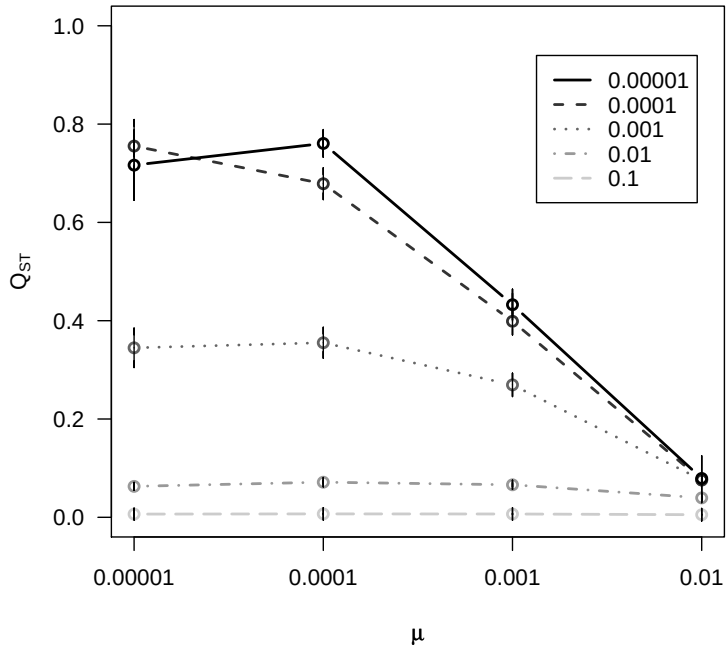

Figure A.4: Results of computer simulations for Q_{ST} with selfing rate of 0.9. Mutation rate is on the x-axis, different migration rates correspond to different lines as indicated by the legend. Points are means of 50 replicate simulations and vertical lines correspond to 95 % confidence intervals.

Appendix B

Supplementary tables

Table B.1: Detailed information on sampled *A. thaliana* populations. Coordinates are given in decimal degree format, N is the number of individuals sampled from the population.

Abbreviation	Population	Region	Latitude (N)	Longitude (E)	Description	N
Agu	Aguaron	Spain	41.321141	−1.340049	Mediterranean forest of evergreen *Quercus* species	10
Cdc	Ciruelos de Coca	Spain	41.214709	−4.544289	Sandy open sites and edges of a pine forest	10
Leo	San Leonardo de Yagüe	Spain	41.797002	−3.11332	Rocky slopes and sandy riversides	10
Mar	Marjaliza	Spain	39.58253	−3.93033	Mediterranean forest of evergreen *Quercus* species	10
Pra	Prádena del Rincón	Spain	41.05598	−3.54306	Trackway and riverside along a grassland	10
Qui	Quintela	Spain	42.692971	−6.92869	Roadsides in the surroundings of a village	10
San	Santa Elena	Spain	38.332574	−3.5076	Roadside and rocky slopes along a Mediterranean forest of evergreen *Quercus* species	10
Mol	Molnet	France	46.92444	4.103333	Edge of a path in rural area with pastures and forest patches	11
All1	Allassac	France	45.260833	1.478611	Cultivated field (winter wheat)	7
All2	Allassac	France	45.260833	1.478611	Cultivated field (maize)	5
Cur	Curemonte	France	45.001667	1.743056	Meadow along a path	9
Cla	Clamart	France	48.801111	2.264167	Waste ground	5
Pyl	Le Pyla	France	44.659444	−1.170000	Sandy soil beside a road	6
Vou	Vouillé	France	46.641389	0.166389	Fallow field	9
Par	Parthenay	France	46.649167	−0.244722	Garden	10
Ldv	Landivisiau	France	48.511389	−4.068056	Garden in an urban area	9

Continued on next page...

Table B.1 – Continued

Abbreviation	Population	Region	Latitude (N)	Longitude (E)	Description	N
Mar2	Marigny l'Eglise	France	47.356389	3.936389	Roadside	7
Fet	Fétigny	France	47.200278	4.179167	Stone wall	4
Lac	Lachapelle-sous-Chaux	France	47.705278	6.821944	Roadside	4
Mog	Sibiril-Mogueriec	France	48.665278	−4.062778	Roadside on sandy soil, rural area close to seaside	3
Cam	Camaret-sur-Mer	France	48.276667	−4.595556	Public garden on sandy soil, urban area close to seaside	10
Mib	Mirebeau sur Bèze	France	47.333056	5.319167	Edge of a field cultivated with mustard	10
Nfro-1	Espe, Kvikne, Nord-Fron	Norway	61.577450	9.661783	South-West slope of mountain pasture landscape	5
Sk-1	Harsheimlii, Skjåk	Norway	61.898966	8.253983	By old cart road on South-West soil cliff in old cultivated landscape	5
Lom-3	Storlii, Lom	Norway	61.681333	8.231000	Sheepwalk in cultivated landscape, dry ground	5
Vgn-1	Festvåg, Vågan	Norway	68.170000	14.220833	South slope between road and the sea, 15 meters from the sea	5
Had-3	Hesthumpen, Hadsen	Norway	68.507616	14.892350	South abrupt rocks in small dry patches of cultivated landscape and sheep pasture	5
Lod-2	Ulvikfjellet, Lødingen	Norway	68.565716	16.318666	On a precipitous slope	5
Tje-1	Kilstadsfjellet, Tjeldsund	Norway	68.424966	16.369383	On a small shelf and dry ground under south cliff	5
Lod-3	Holandstinden, Lødingen	Norway	68.304750	15.114416	South cliff, small shelf on the top of rocky ground	5
Veg-1	Nes, Vega	Norway	65.706827	11.934589	Headland, dry ground with ground cover	5

Continued on next page...

Table B.1 – Continued

Abbreviation	Population	Region	Latitude (N)	Longitude (E)	Description	*N*
Veg-2	Strandberg, Vega	Norway	65.612733	11.838610	Beach rock	4
Eid-1	Åbogen, Eidskog	Norway	60.100000	12.133333	South rock, block ground under precipitous, partly overhanging shale cliff	5
Kon-2	Granlivarden, Kongsvinger	Norway	60.150000	12.083333	South rock, rocky ground, scree, surrounded by residential settlement	5
Kvi-2	Sanåker, Håtveitbygdi, Morgedal, Kviteseid	Norway	59.483333	8.433333	Along a route, sand gravel and rocky moss	5
Zal	Tchong-Kemin valley, Djachyl-Kul lake	Central Asia	42.796400	76.349600	Eroded slope	4
Sus	Otmok village, Suusamyr river	Central Asia	42.188033	73.406933	River bank	6
Neo	Jawshangoz village, Shokhdara river	Central Asia	37.357333	72.467866	Flat rocky meadow	8
Kyr	Kyrgyz-Ata reserve, Kyrgyz-Ata river	Central Asia	40.010787	72.457581	Grassy/rocky slope	8
Kar	Susamyr village, West Karakol river	Central Asia	42.300200	74.369116	River Bank	11
Dja	Djarly-Kayindy reserve, Tchong-Tcholok canyon	Central Asia	42.589766	73.629716	Rocky slope	9

Table B.2: Primer sequences and description of the microsatellite loci used. Primer sequences are given from 5' to 3', the reverse primers had a PIG-tail sequence, GTTTCTT, attached to their 5' end and the forward primers had a fluorescent label. CHR = chromosome, CLH = chloroplast, pos. = physical position in the genome in basepairs.

		Primer sequences		Genome position		
Locus	Label	Forward (core)	Reverse (core)	CHR	pos.(bp)	Repeat type
nga248n	VIC	TACCGAACCAAAACACAAAGC	CTCTGTATCTCGGTGAATTCTCC	1	9887614	CT
nga1145	FAM	CCTTCACATCCAAAACCCAC	GCACATACCCACAACCAGAA	2	682624	CT
nga172	PET	AGCTGCTTCCTTATAGCGTCC	CATCCGAATGCCATTGTTC	3	786450	GA
nga1139	NED	TAGCCGGATGAGTTGGTACC	TTTTTCCTTGTGTTGCATTCC	4	15408735	CT
AthGENEA	VIC	ACCATGCATAGCTTAAACTTCTT	ACATAACCACAAATAGGGGTG	1	21608507	A
nga6	FAM	TGGATTTCTTCCTCTCTTCAC	ATGGAGAAGCTTACACTGATC	3	22908202	GA
nga1111	PET	TGGTTCGGTTACAATCGTGT	AGTTACCAGATTGAGCTTTGAGC	4	5054189	TC
nga106	NED	GTTATGGAGTTTCTAGGGCACG	TGCCCCTTTTGTTCTTCTC	5	5320244	GA
nga111	VIC	CTCCAGTTGGAAGCTAAAGGG	TGTTTTTTAGGACAAATGGCG	1	26564737	GA
nga361	FAM	AAAGAGATGAGAATTTGGAC	ACATATCAATATATTAAAGTAGC	2	13170956	GA
msat3.18	PET	TCATACCTACATATTGCCCT	TACCTCAAAAGAGCAAACA	3	21254251	AT
AthS0262	NED	TTGCTTTTTGGTTATATTCGGA	ATCATCTTCCCATGGTTTTT	5	9611233	CTT
nga1126	VIC	CGCTACGCTTTTCGGTAAAG	GCACAGTCCAAGTCACAACC	2	11645028	GA
nga8	FAM	GAGGGCAAATCTTTATTTCGG	TGGCTTTCGTTTATAAACATCC	4	4593289	GA
nga1107	PET	GCGAAAAAACAAAAAAATCCA	CGACGAATCGACAGAATTAGG	4	17060748	GA
nga129	NED	TCAGGAGGAACTAAAGTGAGGG	AACACTGAAGATGGTCTTGAGG	5	19409716	GA
nga59	VIC	TTAATACATTAGCCCAGACCCG	GCATCTGTGTTCACTCGCC	1	8601	CT
nga63	FAM	ACCCAAGTGATCGCCACC	AACCAAGGCACAGAAGCG	1	3224501	GA
nga168	PET	GAGGACATGTATAGGAGCCTCG	TCGTCTACTGCACTGCCG	2	16298919	CT
nga162	NED	CTCTGTCACTCTTTTCCTCTGG	CATGCAATTTGCATCTGAGG	3	4608284	GA
ATCP46615	VIC	AATTTTTTTCCATTGCACATTG	TCAGAAATAGTCGAACGGTCG	CLH		A
ATCP7905	FAM	CGAACCCTCGGTACGATTAA	TGGAGAAGGTTCTTTTTCAAGC	CLH		A
ATCP70189	PET	CGGGTTGATGGATCATTACC	GCAATGCACAAAAAAAGCCT	CLH		A
ATCP28673	NED	GCGTTCCTTTCATTTAAGACG	TGCACTCTTCATTCTCGTTCC	CLH		T

Table B.3: Primers used to amplify and sequence the SNP discovery fragments. Primer sequences are given from 5' to 3'. Gene model indicates the gene that is closest or overlaps to the assayed fragment. Distance indicates the physical distance to *DOG1*, positive values indicate that the fragment is downstream and negative values upstream from *DOG1*. The last column shows how many SNPs were eventually genotyped from the corresponding fragment.

Locus	Forward	Reverse	Gene model	Distance	SNPs
s1	CTTAAAACGGGCTGCTTGAC	GGAACCATCGTCAGTTCCTT	At5G45850	10 kb	4
s2	ATGGTTCCGTTTGTTCTGCT	TCCAATTCGTTACCCGTCTC	At5G45880	20 kb	2
s3	CAGCTGAAGATGTTGGTATTGTT	GCGTAATCTTAACCAAACGATCTA	At5G45930	40 kb	3
s4	TCCAGCCTAGACTCATTCAAGTT	GGCTTTCCCAATCCATTCTA	At5G46040	80 kb	4
s5	TCCGTTTCTGATTTGTGTCC	TAAACAGCTGCTTGGGTCCT	At5G46250	160 kb	2
s6	AATCCCTTGCGTAATGTTGC	GGAAAGGAAGATCCGAGGAG	At5G46600	320 kb	2
s7	CATGTGGCTTGCATAGGTTG	ATGGATCCTCTTCTCCCTCA	At5G45730	−40 kb	3
s8	TTCATTGTCCTCGTTGGTCA	GTTGTTCCAGGCGTTTCAAG	At5G45640	−80 kb	3
s9	TGGATACTTCCACCGTGATCT	CGCACTAGCATATGGCATCA	At5G45430	−160 kb	1

Table B.4: Primer sequences used in the pyrosequencing assays. Primer sequences are given from 5' to 3'. For the SNP assays core sequences are given, the forward and reverse primers had a the corresponding universal sequence attached to their 5' end. In the PCR reaction either the forward or reverse biotin labelled universal primer was used as indicated by the column bio. Sequence to analyse is the template sequence to be analysed in the pyrosequencing reaction. The last four assays were used for genotyping *DOG1* in the QTL-analyses. For these assays the universal primers were not used.

	Universal forward primer GTGACGTACTAGCAACG	Universal reverse primer TAGCAGGATACGACTATC			
SNP(s)	Forward primer	Reverse primer	Sequencing primer	Bio	Sequence to analyse
s1_545	TGGCAAGAACCCTAACTGATC	TGGCAATTAATCAAGATTTGAACA	AATACCATCCACATGATT	R	ATGATCAT/CAATC/GC
s1_803	CATTACTCGATCTGTCGCCATCT	ATGGTTTCTTCAACCGCAAGTTT	TCTCGAACTCACTTACTCC	R	AC/ATTGAA/GCT
s2_128	CGCCTGCACAAACACAATTAA	ACCGAAACAAATTCTTGAGGTC	ATGTGGATCAAGATCACG	R	GAA/TCAAC
s2_483	TATGACAGTGCTTGCATCCTCTAG	GTGTGCATGGCTAATTTACAACG	TGTATGGCCATCTCAAA	R	ATGAT/CAAG
s3_390	CTGCGTAAAAACCCGACAG	TGTGCCATTAGAGATGTTCCTCAA	TTAATCTTTAATGGCTC	F	TC/GTCTG
s3_520	TCTCTTCCACTTCAGCTGGATTT	CACCAGTTATCGGTATTCTACACG	CTACACGATAAGAAAGCG	F	TTTAAA/TCA/CGTTA
s4_66	TTCCCGACTGCCATACTTGTG	CGATCAAGCACCAGAGAGTATGAA	GCCATACTTGTGCTCGT	R	ATAC/TGATGTCCCG/AAGG
s4_339	TTCGGTTATGGACGCGATGAT	TCCTAGGCTTCACAACATTCTCAA	CAATGTGATACCTCTTGG	R	G/ATTTCCG/AGTTA
s5_155	TTTGGTAAAGCTGGAAGGTAGG	CAACGTCTCTTACTCTGGTTCTGA	GATCACATATGGATACACTT	F	TTAATA/GCTG
s5_360	GTTACATGCCTTTGTGGAGTATGA	CGACTTCCTTCCTAGCTGGTCTT	CTTTTTCAGCTGCCTC	F	CACCG/TTTTCA
s6_70	CAACCAAGTTTTATGAAAATACCA	CGGCTATGTTATATGAAAAAACGA	TTGAAAAATATCTAGC	F	A/CATGG
s6_628	CAAGAACCACAACGACGGTCAT	GGAGTTGCCTTAACGCTTGTGTCT	AGGTGTTGGAAAAAATG	F	CTCTA/TTGGGC
s7_165	GTTTTGTAGAAGCCCACTTGTTGA	AACCAACCCACATGCGTTAC	TATATATGAGCAAAAGAGC	R	CACAA/GCCA
s7_204	TGTAGAAGCCCACTTGTTGATCTC	GCAGCATGAAATCACATTCAAGAC	AACCCACATGCGTTA	F	CAAG/TTGAA
s7_358	TGCTGCACAAAGGTTGTATCAA	CTTTTGGCGACAAACTCCACA	AAACTCCACATTCCCA	F	ATCA/GCCG
s8_170	GAGGGCAGATCTTATAGCAGCAC	ATTCGATAGACCGACGAGTTTTAG	GCACCAGCAACATGA	R	GGA/GCAAG
s8_365	TCGTCGGTCTATCGAATCTTTG	TTTTTCTAGTAGAGGCCCAAATTG	AGATGCTAACATTTTAAAGG	F	TAAAC/TTCA/TTTTG
s9_490	TGCCATATCTTTTCCTTATTCACC	AGAGAGGGCGAAGAGACAACA	CCATTGTTAAGCACCG	F	CACTAC/GACT
D_a1	ATGGGATCTTCATCAAAGAACAT	TGAGATCGTCGTTGAGCTAAGA	CATCGAACAAGCTCAAG	R	AA/TT/CG/CT[TGT]TATCTCGAGT
D_a2	GCAATCTCAACGCATCCC	GCGTTCTCTAAAGGACTGTTCCA	CGTCGTTGAGCTAAGAG	F	TG/TGTTTGA
D_a3	same as in D_a1	TGCGGCGTAATTTTTGAAA	GATAATGATAACAAGCTTCG	R	TG/AAGTTAAC
D_a4	same as in D_a1	same as in D_a1	same as in D_a1	R	AATGTTA/GTTATCA/TCGAA/GTG

Table B.5: Correlation between genetic diversity and genetic differentiation between populations in different regions. Correlation coefficients are given with 95 % confidence intervals are in parenthesis. H_S is subpopulation heterozygosity and Θ is an estimate of microsatellite mutation rate

	H_S		Θ	
Spanish populations	r(95 % CI)	p	r(95 % CI)	p
F_{ST}	−0.862 (−0.944 – −0.678)	< 0.001	−0.682 (−0.864 – −0.342)	< 0.001
F'_{ST}	0.479 (0.046 – 0.760)	0.033	0.500 (0.074 – 0.771)	0.025
Φ_{ST}	−0.294 (−0.652 – 0.170)	0.208	−0.301 (−0.656 – 0.163)	0.197
D	0.765 (0.488 – 0.902)	< 0.001	0.655 (0.300 – 0.851)	0.002
French populations				
F_{ST}	−0.867 (−0.948 – −0.681)	< 0.001	−0.625 (−0.840 – −0.238)	0.004
F'_{ST}	0.645 (0.270 – 0.850)	0.003	0.705 (0.370 – 0.878)	< 0.001
Φ_{ST}	−0.260 (−0.639 – 0.220)	0.282	−0.144 (−0.561 – 0.332)	0.557
D	0.876 (0.700 – 0.952)	< 0.001	0.756 (0.460 – 0.901)	< 0.001
Norwegian populations				
F_{ST}	−0.916 (−0.968 – −0.791)	< 0.001	−0.599 (−0.828 – −0.198)	0.007
F'_{ST}	0.199 (−0.280 – −0.599)	0.413	0.446 (−0.011 – 0.748)	0.056
Φ_{ST}	−0.109 (−0.536 – 0.364)	0.658	−0.021 (−0.471 – 0.437)	0.931
D	0.631 (0.248 – 0.843)	0.004	0.717 (0.390 – 0.884)	< 0.001
Central Asian populations				
F_{ST}	−0.801 (−0.928 – −0.506)	< 0.001	−0.494 (−0.795 – 0.002)	0.052
F'_{ST}	−0.116 (−0.578 – 0.403)	0.669	0.125 (−0.395 – 0.584)	0.645
Φ_{ST}	−0.628 (−0.857 – −0.192)	0.009	−0.132 (−0.589 – 0.389)	0.625
D	0.565 (−0.013 – 0.860)	0.055	0.484 (−0.124 – 0.828)	0.111

Table B.6: Genetic differentiation as measured by F_{ST} between populations within regions and between regions, using F_{CT}. Estimate of F_{ST} is given over all loci and its 95 % CI, obtained by bootstrapping over loci, is in parenthesis. Between region estimated were not calculated for D, since hierarchical estimation has not been derived for it.

Region or comparison	Marker type				
	Microsatellite				SNP
Within regions	F_{ST}	Φ_{ST}	F'_{ST}	D	F_{ST}
Spain	0.2900 (0.2531 – 0.3345)	0.3556 (0.2500 – 0.4646)	0.7208 (0.6591 – 0.7833)	0.6393 (0.5455 – 0.7275)	0.3476 (0.3199 – 0.3753)
France	0.4937 (0.4598 – 0.5292)	0.6818 (0.4612 – 0.7989)	0.8115 (0.7441 – 0.8720)	0.6509 (0.5202 – 0.7660)	0.5456 (0.5257 – 0.5636)
Norway	0.8004 (0.7501 – 0.8499)	0.8128 (0.7581 – 0.8879)	0.9436 (0.9209 – 0.9630)	0.7241 (0.6151 – 0.8175)	0.9274 (0.9162 – 0.9387)
Central Asia	0.6026 (0.4864 – 0.7253)	0.3101 (0.1070 – 0.8259)	0.8413 (0.7639 – 0.8984)	0.6334 (0.5251 – 0.7336)	0.7806 (0.7343 – 0.8206)
Between regions	F_{CT}	Φ_{CT}	F'_{CT}	–	F_{CT}
Spain vs. France	0.0267 (0.0035 – 0.0631)	−0.0064 (−0.0647 – 0.0941)	0.1383 (0.0112 – 0.2830)	–	0.0553 (0.0331 – 0.0784)
Spain vs. Norway	0.0870 (0.0309 – 0.1548)	0.0921 (0.0205 – 0.1729)	0.4071 (0.1980 – 0.6104)	–	0.1791 (0.1354 – 0.2247)
France vs. Norway	0.1319 (0.0649 – 0.2016)	0.0753 (−0.0219 – 0.2363)	0.5255 (0.3195 – 0.7228)	–	0.2149 (0.1759 – 0.2558)
European regions	0.0792 (0.0405 – 0.1250)	0.0469 (−0.0260 – 0.1476)	0.3637 (0.2307 – 0.4975)	–	0.1448 (0.1177 – 0.1726)
All regions	0.1094 (0.0679 – 0.1585)	0.1159 (0.0696 – 0.2070)	0.4304 (0.2982 – 0.5634)	–	0.2257 (0.2017 – 0.2530)

Table B.7: Polymorphic positions in the observed *DOG1* haplotypes. ID = haplotype number, position starts at the A in ATG = 1, the insert column indicates the presence of the insertion in intron 1.

	Position																									
ID	2	6	23	33	39	40	41	43-45	50	54	62	66	69	71	95	103	140	147	151	168	197	208	238	243	304	313	Insert
1	T	A	T	T	T	C	G	InDel	T	G	G	G	A	C	A	G	A	T	G	C	A	C	G	C	T	C	-
2	A	A	T	T	T	C	G	InDel	T	G	G	G	A	C	A	G	A	T	G	C	A	C	G	C	T	C	-
3	T	A	T	T	T	T	C	InDel	T	G	G	G	A	C	A	G	A	T	G	C	A	C	G	C	T	C	-
4	T	A	T	T	T	T	C	InDel	T	G	G	G	A	C	A	G	A	T	A	C	A	C	G	C	T	C	-
5	T	A	T	T	A	T	G	TGT	T	G	G	G	A	C	A	G	A	T	G	C	A	C	G	C	T	C	-
6	T	A	T	T	A	T	G	TGT	T	G	G	G	A	C	A	G	A	T	G	C	A	C	G	C	A	C	-
7	T	A	T	T	A	T	G	TGT	T	G	G	G	A	C	A	G	A	T	G	C	A	C	A	C	T	C	-
8	T	A	T	T	A	T	G	TGT	T	G	G	G	A	C	A	G	A	T	G	C	A	T	G	C	T	C	-
9	T	A	T	T	A	T	G	TCT	T	G	G	G	A	C	A	G	G	T	G	C	A	C	G	C	T	C	-
10	T	A	T	T	A	T	T	TCT	T	G	G	G	A	C	A	G	G	T	G	C	A	C	G	C	T	C	-
11	T	A	T	T	A	T	G	TCT	T	G	G	G	A	C	A	G	G	T	G	C	A	C	G	C	C	C	-
12	T	A	T	T	A	T	G	TGT	T	G	T	G	A	C	A	G	A	T	G	C	A	C	G	C	T	C	-
13	T	A	T	T	A	T	G	TGT	T	G	G	G	A	C	A	G	A	T	G	C	G	C	G	C	T	C	-
14	T	A	C	T	A	T	G	TGT	T	G	G	G	A	C	A	G	A	T	G	T	A	C	G	C	T	C	-
15	T	A	T	T	A	T	G	TGT	T	G	G	G	A	C	C	G	A	T	G	T	A	C	G	C	T	T	-
16	T	T	T	T	A	T	G	TGT	T	G	G	G	A	C	C	G	A	T	G	T	A	C	G	C	T	T	-
17	T	A	T	A	A	T	G	TGT	A	A	G	G	G	C	A	G	A	C	G	C	A	C	G	G	T	C	-
18	T	A	T	A	A	T	G	TGT	A	A	G	G	G	T	A	G	A	C	G	C	A	C	G	G	T	C	-
19	T	A	T	A	A	T	G	TAT	A	A	G	T	G	C	A	G	A	C	G	C	A	C	G	G	T	C	-
20	T	A	T	A	A	T	G	TAT	A	A	G	T	G	C	A	A	A	C	G	C	A	C	G	G	T	C	-
21	T	A	T	T	T	T	C	InDel	T	G	G	G	A	C	A	G	A	T	A	C	A	C	G	C	T	C	present
22	A	A	T	T	T	T	C	InDel	T	G	G	G	A	C	A	G	A	T	A	C	A	C	G	C	T	C	present

Table B.8: Minimum MSD values and their corresponding T values for different traits and samples.

Trait	D25		D50		D75	
	Min(MSD)	T	Min(MSD)	T	Min(MSD)	T
Sample						
All regions + W	1.46E-03	0.55	9.68E-04	0.65	7.29E-04	0.55
Spain	2.75E-04	0.75	3.15E-04	0.725	3.05E-04	0.725
France	1.71E-04	0.45	5.37E-04	0.6	2.83E-04	0.6
Norway	2.52E-03	0.975	7.00E-04	0.95	1.15E-03	0.875
Central Asia	1.71E-02	0.85	1.23E-02	0.85	8.17E-03	0.85

Table B.9: Population means and genetic variation within populations for seed dormancy. σ_G^2 is the amount of genetic variation within a population and H^2 is heritability within a population.

Region	Population	D50	σ_G^2 D50	H^2 D50
Spain	Agu	19.383	15.7346	0.0959
Spain	Cdc	11.730	10.2359	0.5356
Spain	Leo	6.864	5.9245	0.3313
Spain	Mar	27.218	226.4866	0.8512
Spain	Pra	15.685	110.209	0.8035
Spain	Qui	20.377	52.4344	0.5675
Spain	San	19.980	116.1429	0.7299
France	All1	20.689	141.9779	0.7177
France	All2	13.518	4.3987	0.149
France	Cam	15.428	17.4596	0.3661
France	Cla	13.575	2.5233	0.0761
France	Cur	17.584	53.5376	0.7289
France	Fet	11.458	0	0
France	Lac	−0.911	5.956	0.8746
France	Ldv	6.121	10.3624	0.544
France	Mar2	2.510	2.1499	0.2646
France	Mib	21.632	102.5252	0.611
France	Mog	43.968	0	0

Continued on next page...

Table B.9 – Continued

Region	Population	D50	σ_G^2 D50	H^2 D50
France	Mol	14.617	11.0574	0.2551
France	Par	16.655	27.4916	0.2156
France	Pyl	9.353	3.2899	0.3819
France	Vou	6.047	18.5708	0.5024
Norway	Eid-1	3.746	8.8888	0.1356
Norway	Nfro-1	−1.575	0.0358	0.032
Norway	Vgn-1	0.110	0.1557	0.1792
Norway	Sk-1	30.077	0	0
Norway	Had-3	5.497	0	0
Norway	Lod-3	14.462	19.4329	0.2664
Norway	Tje-1	33.271	80.5805	0.2462
Norway	Kon-2	−1.855	0.0168	0.0581
Norway	Kvi-2	9.542	0	0
Norway	Lom-3	1.245	0	0
Norway	Lod-2	13.058	42.6187	0.6305
Norway	Veg-1	−1.280	0	0
Norway	Veg-2	−2.052	0.0974	0.0966
Central Asia	Dja	20.922	0	0
Central Asia	Kar	37.742	62.0237	0.2857
Central Asia	Kyr	17.087	176.7456	0.9343
Central Asia	Neo	0.169	3.0268	0.2145
Central Asia	Sus	48.997	1.0253	0.3425
Central Asia	Zal	24.245	226.505	0.9054

Appendix C

Equations

Here I give the equations for different F_{ST} estimators and other calculations that I implemented for this study in R. Φ-statistics were estimated using a analysis of variance framework for molecular markers (AMOVA) (Excoffier et al. 1992; Michalakis & Excoffier 1996). Total variance is partitioned into covariance[1] components and then they are used to calculate Φ-statistics.

Since most of the genotypes were homozygous, the within individual level was omitted from the analysis.

For microsatellite markers the pairwise allelelic distance is defined as $\delta_{ij}^2 = (a_i - a_j)^2$, where δ_{ij}^2 is simply the squared difference in repeat numbers between allele i and j. For DNA sequences this becomes the number of pairwise differences.

For one group of populations the AMOVA becomes as in table C.1. Where sums of squares are defined as

$$SSD(T) = \frac{1}{2N}\sum_{i=1}^{2N}\sum_{j=1}^{i}\delta_{ij}^2 \ , \ SSD(WP) = \sum_{p=1}^{P}\frac{1}{2N_p}\sum_{i=1}^{2N_p}\sum_{j=1}^{i}\delta_{ij}^2 \text{ and} \qquad \text{(C.1)}$$

[1] see Excoffier (2007) for why these are covariance rather than variance components

Table C.1: AMOVA for one group of populations

Source of variation	Degrees of freedom	Sum of squares (SSD)	Expected mean squares
Among populations	$P-1$	$SSD(AP)$	$n^{'}\sigma_b^2 + \sigma_w^2$
Within populations	$2N-P$	$SSD(WP)$	σ_w^2
Total	$2N-1$	$SSD(T)$	

Table C.2: AMOVA for several groups of populations

Source of variation	Degrees of freedom	Sum of squares (SSD)	Expected mean squares
Among groups	$G-1$	$SSD(AG)$	$\sigma_w^2 + n^{''}\sigma_b^2 + n^{'''}\sigma_a^2$
Among populations within groups	$P-G$	$SSD(AP)$	$\sigma_w^2 + n^{'}\sigma_b^2$
Within populations	$2N-P$	$SSD(WP)$	σ_w^2
Total	$2N-1$	$SSD(T)$	

$$SSD(AP) = SSD(T) - SSD(WP)$$

and where

$$n^{'} = \frac{1}{P-1}\left(2N - \sum_{p=1}^{P}\frac{(2N_p)^2}{2N}\right)$$

The covariance components are

$$\sigma_w^2 = SSD(WP)/(2N-P)\ ,\ \sigma_b^2 = (SSD(T)/(2N-1) - \sigma_w^2)/n^{'} \tag{C.2}$$

Then a single locus estimate of Φ_{ST} is defined as

$$\Phi_{ST} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

To obtain multilocus estimates, the covariance components are summed over all loci, such that

$$\Phi_{ST} = \frac{\sum_{l=1}^{L}\sigma_{bl}^2}{\sum_{l=1}^{L}\sigma_{bl}^2 + \sum_{l=1}^{L}\sigma_{wl}^2} \tag{C.3}$$

If there are multiple groups of populations this adds one additional level of hierarchy. Then the AMOVA becomes as in table C.2. The sums of squares are defined as

$$SSD(AG) = SSD(T) - \sum_{g=1}^{G}\frac{1}{2N_g}\sum_{i=1}^{2N_g}\sum_{j=1}^{i}\delta_{ij}^2\ ,\ SSD(AP) = \sum_{g=1}^{G}\frac{1}{2N_g}\sum_{i=1}^{2N_g}\sum_{j=1}^{i}\delta_{ij}^2 - SSD(WP)$$

and

$$SSD(WP) = \sum_{g=1}^{G}\sum_{p}^{P_g}\frac{1}{2N_{gp}}\sum_{i=1}^{2N_{gp}}\sum_{j=1}^{i}\delta_{ij}^2 \tag{C.4}$$

$SSD(T)$ is as in equation C.1. The n primes are defined as

$$S_G = \sum_g^G \sum_{p=1}^{P_g} \frac{2N_{gp}^2}{N_g}\,, n^{'} = \frac{2N - S_G}{P - G}\,, n^{''} = \frac{S_G - \sum_{p=1}^{P} \frac{2N_p^2}{N}}{G-1} \text{ and } n^{'''} = \frac{2N - \sum_{g=1}^{G} \frac{2N_g^2}{N}}{G-1}$$

The covariance components are calculated as

$$\sigma_b^2 = (SSD(AP)/(P-G) - \sigma_w^2)/n^{'}\,,\ \sigma_a^2 = (SSD(AG)/(G-1) - \sigma_w^2 - n^{''}\sigma_b^2)/n^{'''} \tag{C.5}$$

where σ_w^2 is the same as in equation C.2

The Φ-statistics are

$$\Phi_{CT} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_w^2}\ ,\Phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \text{ and } \Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_w^2}$$

where Φ_{CT} is the genetic differentiation between regions, Φ_{SC} is the genetic differentiation between populations within regions and Φ_{ST} is the genetic differentiation between populations. The previous can be reduced to an equivalent of F_{ST} estimated by Weir & Cockerham (1984) by substituting δ_{ij}^2 with a function that gives $\delta_{ij}^2 = 0$ when $a_i = a_j$ and $\delta_{ij}^2 = 1$ when $a_i \neq a_j$.

The calculation of the standardised measure of genetic differentiation F'_{ST} was done using the AMOVA approach outlined above following Meirmans (2006). The standardised measures divides F_{ST} by the maximum value that could be obtained given given the observed amount of diversity (Hedrick 2005). F'_{ST} is defined as

$$F'_{ST} = \frac{F_{ST}}{F_{ST(max)}} \tag{C.6}$$

$F_{ST(max)}$ is obtained from

$$F_{ST(max)} = \frac{\sigma_{b(max)}^2}{\sigma_{b(max)}^2 + \sigma_w^2}$$

where $\sigma_{b(max)}^2$ is estimated as in equation C.2, except that sums of squares are redefined as

$$SSD(AP)_{max} = SSD(T)_{max} - SSD(WP) \text{ and}$$

$$SSD(T)_{max} = \frac{1}{2N} \sum_{p=1}^{P} \left(\sum_{i=1}^{2N_p} \sum_{j=1}^{i} \delta_{ij}^2 + \sum_{p \neq p2}^{P} N_p N_{p2} \right)$$

When there are several groups of populations F_{CT} and F_{SC} can be maximised as

$$F_{CT(max)} = \frac{\sigma_{a(max)}^2}{\sigma_{a(max)}^2 + \sigma_b^2 + \sigma_w^2} \text{ and } F_{SC(max)} = \frac{\sigma_{b(max)}^2}{\sigma_{b(max)}^2 + \sigma_w^2} \tag{C.7}$$

These maximisations have to performed separately, since $F_{CT(max)}$ has to be relative to the actual $SSD(AP)$ not to $SSD(AP)_{max}$. The maximised sums of squares for between groups are

$$SSD(AG)_{max} = SSD(T)_{max} - SSD(WP) \text{ and} \tag{C.8}$$

$$SSD(T)_{max} = \frac{1}{2N} \sum_{g=1}^{G} \left(\sum_{i=1}^{2N_g} \sum_{j=1}^{i} \delta_{ij}^2 + \sum_{g \neq g2}^{G} N_g N_{g2} \right) \tag{C.9}$$

$SSD(WP)$ is calculated like in equation C.4 and covariance components are likewise calculated like in equation C.5. For between populations within groups the maximised sums of squares are

$$SSD(AP)_{max} = SSD(WG)_{max} - SSD(WP) \tag{C.10}$$

$$SSD(WG)_{max} = \sum_{g=1}^{G} \left[\frac{1}{2N_g} \sum_{p=1}^{P_g} \left(\sum_{i=1}^{2N_p} \sum_{j=1}^{i} \delta_{ij}^2 + \sum_{p \neq p2}^{P_g} N_p N_{p2} \right) \right] \tag{C.11}$$

where again $SSD(WP)$ is given by equation C.4.

Recently Jost (2008) derived a new estimator of genetic differentiation, D, which was intented to replace F_{ST}. D is defined as

$$D = \left(\frac{H_T - H_S}{1 - H_S} \right) \left(\frac{P}{P-1} \right) \tag{C.12}$$

H_T and H_S were estimated following Nei & Chesser (1983). Let $\bar{N}$ be the harmonic mean of population sample sizes and p'_{ij} be the frequency of allele i in the sample from the jth population. Then

$$H'_j = 1 - \sum_{i=1}^{A} (p'_{ij})^2 \ , \ H'_S = (1/P) \sum_{j=1}^{P} H'_j \text{ and } H'_T = 1 - \left[\sum_{i=1}^{A} \left((1/P) \sum_{j=1}^{P} p'_{ij} \right)^2 \right] \tag{C.13}$$

and the estimators are

$$\hat{H}_S = \left(2\bar{N}/(2\bar{N}-1) \right) H'_S \text{ and } \hat{H}_T = H'_T + \hat{H}_S / 2\bar{N}P \tag{C.14}$$

Substituting $\hat{H}_T$ and $\hat{H}_S$ from equation C.14 in place of H_T and H_S in equation C.12 yields the estimator for D.

Appendix D

Bayesian models

Here I give the specifics of the Bayesian models used in estimating H^2 and Q_{ST} from the common garden experiment, and estimation of allelic effects and dominance coefficients for *DOG1* in the F_2-populations. The Bayesian models corresponded to model 2.7 for H^2, model 2.8 for Q_{ST} and model 2.14 for the analysis of F_2-populations. The BUGS code to implement these analyses is given below.

The estimation of H^2 and Q_{ST} followed the model of O'Hara & Merilä (2005). In Bayesian models the choice of priors is important as this can sometimes influence the result if not enough information is contained in the data. In the model for H^2 there are two variance components σ^2_G and σ^2_E for genetic and environmental effects and σ^2_B for block effects (see section 2.5.4). For Q_{ST} the variance components are σ^2_{GB}, σ^2_{GW}, σ^2_E and σ^2_B. In all cases the prior for the overall mean was normally distributed with mean of 0 and variance 10^6. I compared several different priors for the variance components, a standard gamma prior on precisions

$$\sigma^{-2}_{GB}, \sigma^{-2}_{GW}, \sigma^{-2}_E \sim \Gamma(10^{-4}, 10^{-4})$$

A uniform prior on standard deviations

$$\sigma_{GB}, \sigma_{GW}, \sigma_E \sim \text{Unif}(0, 10^4)$$

or alternatively a uniform prior on variances

$$\sigma^2_{GB}, \sigma^2_{GW}, \sigma^2_E \sim \text{Unif}(0, 10^5)$$

Another possibility is to place a uniform prior on Q_{ST} (or heritability) and then place another uniform prior on the sum of standard deviations

$$Q_{ST} \sim \text{Unif}(0, 1), \quad \sigma_{GB} + \sigma_{GW} \sim \text{Unif}(0, 10^5) \quad \text{and} \quad \sigma_E \sim \text{Unif}(0, 10^4)$$

As Q_{ST} is the statistic of interest this seems better on theoretical grounds, similarly in the previous variances can be used instead of standard deviations. In the

actual model the variance components have to be calculated from sum of standard deviations, using equation 2.6 the ratio of standard deviations is

$$\frac{\sigma_{GW}}{\sigma_{GB}} = \sqrt{1/Q_{ST} - 1}$$

then precisions can be expressed in terms of this ratio and and the sum of standard deviations

$$\sigma_{GB}^{-2} = \left(\left(1 + \frac{\sigma_{GW}}{\sigma_{GB}}\right)\left(\frac{1}{\sigma_{GB} + \sigma_{GW}}\right)\right)^2$$
$$\sigma_{GW}^{-2} = \left(\left(1 + \frac{\sigma_{GB}}{\sigma_{GW}}\right)\left(\frac{1}{\sigma_{GB} + \sigma_{GW}}\right)\right)^2$$

First I compared the results obtained using different priors, for heritability the choice of prior did not affect the results and the point estimates for H^2 and variance components were essentially identical using REML. The standard gamma prior on precisions was chosen to analyse the data. For Q_{ST} the gamma prior on precisions and the uniform prior on Q_{ST} performed best, as the other priors had sometimes convergence problems. Both priors gave similar point estimates as REML, there were few cases where the uniform prior on Q_{ST} seemed to be more in line with the REML estimates. Thus, the data was subsequently analysed with uniform prior on Q_{ST} and the sum of standard deviations. In all cases gamma prior was used on the precision for block effects, but the results were essentially the same whether block effects were included or not. When estimating allelic effects and dominance coefficients for *DOG1*, uniform prior on standard deviations gave the same point estimates as the standard linear model and this prior was used in the subsequent analyses.

For the Markov Chain Monte Carlo (MCMC) simulation, two independent chains were run. First there was a burn-in period of 10000 iterations and then, for H^2, 40000 iterations giving a sample of 80000 iterations for parameter estimation. For Q_{ST} there was first a burn-in of 10000 and then 50000 iterations with thinning set to 5 yielding a sample of 100000 for parameter estimation. For the analysis of the F_2-populations a burn-in of 10000 with a sampling of 50000 was used.

BUGS code

```
#Calculating heritability
Model
{
for(i in 1:N) {
```

```
Mu[i] <- BetaGeno[genotype[i]] + BetaBlock[block[i]]
Trait[i] ~ dnorm(Mu[i], tau.err) #Data is in variable Trait
}
for(g in 1:gen){BetaGeno[g] ~ dnorm(theta, tau.gen) }
for(b in 1:blo){BetaBlock[b] ~ dnorm(0, tau.block) }

theta ~ dnorm(0, 1.0E-6) # theta is the overall mean
sigma2.gen <- 1 / tau.gen # sigma2.gen is the genetic variance
sigma2.err <- 1 / tau.err # sigma2.err is the environmental variance

Heritability <- sigma2.gen / (sigma2.gen + sigma2.err)

#Gamma priors for precisions
tau.gen ~ dgamma(0.0001, 0.0001)
tau.err ~ dgamma(0.0001, 0.0001)
#Prior for block variation
tau.block ~ dgamma(0.0001, 0.0001)

}

Inits #Initial values for the two chains
list(theta = 1, tau.gen = 0.1, tau.err = 0.1, tau.block = 0.1)
list(theta = 10, tau.gen = 0.1, tau.err = 0.1, tau.block = 0.1)

Data #This section describes the data

#Where N = number of replicates, gen = number of genotypes,
#blo = number of blocks
#Trait = vector of phenotypic measurements for each replicate
#genotype = vector giving the genotype of each replicate,
#block = vector giving the block for each replicate
list(N = xxx, gen = xxx, blo = xxx, Trait = c(n1, n2, n3, ..., ni),
genotype = c(g1, g2, g3, ..., gi), block = c(b1, b2, b3, ..., bi))

#Calculating Qst

Model
{
for(i in 1:N) {
Mu[i] <- BetaPop[population[i]] + BetaGeno[genotype[i]] + BetaBlock[block[i]]
Trait[i] ~ dnorm(Mu[i], tau.err)
}
for(p in 1:pop) { BetaPop[p] ~ dnorm(theta, tau.pop) }
```

```
for(g in 1:gen) { BetaGeno[g] ~ dnorm(0, tau.gen) }
for(b in 1:blo) { BetaBlock[b] ~ dnorm(0, tau.block) }

theta ~ dnorm(0, 1.0E-6)

#Prior for block variation
tau.block ~ dgamma(0.0001, 0.0001)

#Uniform prior on Qst and standard deviations
Qst ~ dunif(0, 1)
sd.popgen ~ dunif(0, 10000) #sd.pop + sd.gen
ratio.genpop <- sqrt( (1 - (1 / Qst)) * (-1)) #sd.gen / sd.pop
tau.pop <- pow( (1 + ratio.genpop) * (1 / sd.popgen), 2)
sigma2.pop <- 1 / tau.pop #precisions are calculated using the sd.ratio
tau.gen <- pow( (1 + (1 / ratio.genpop)) * (1 / sd.popgen), 2)
sigma2.gen <- 1 / tau.gen
sd.err ~ dunif(0, 1000); tau.err <- pow(1 / sd.err, 2)
sigma2.err <- 1 / tau.err

}

Inits #Inital values for the two chains
list(theta = 20, Qst = 0.5, sd.popgen = 100, sd.err = 10, tau.block = 0.1)
list(theta = 10, Qst = 0.5, sd.popgen = 10, sd.err = 10, tau.block = 0.1)

Data #This section describes the data
#This section is identical with the heritability model,
#except pop = xxx gives the number of populations
#and the vector population = c(p1, p2, p3, ..., pi)
#giving the population for each replicate

#Calculating allelic effects and dominance coefficents in an F2-population

Model
{
for(i in 1:N) {
Mu[i] <- BetaGeno[genotype[i]]
Trait[i] ~ dnorm(Mu[i], tau.err)
}
for(g in 1:gen) {BetaGeno[g] ~ dnorm(theta, tau.gen) }

theta ~ dnorm(0, 1.0E-6)
```

```
#Allelic effect and dominance coefficient
allelic <- (BetaGeno[1] - BetaGeno[2]) / 2
dominance <- ((BetaGeno[3] - BetaGeno[2]) / allelic) - 1
#Note that index of BetaGeno has to be changed according to how genotypic
#classes are coded in the genotype vector
#in this example heterozygotes are coded as 3,
#while the homozygotes are 1 and 2

#Uniform priors on SD
sigma.gen ~ dunif(0,100)
sigma.err ~ dunif(0,100)
tau.gen <- 1 / (sigma.gen*sigma.gen)
tau.err <- 1 / (sigma.err*sigma.err)

}

Inits
list(theta = 1, sigma.gen = 1, sigma.err = 1)
list(theta = 10, sigma.gen = 1, sigma.err = 1)

Data
#This the same as in previous models
#N = xxx, gen = 3, now gen in the number of genotypic classes
#Trait = c(n1, n2, ..., ni)
#genotype = c(g1, g2, ..., gi) a vector giving the genotypic class
#for each replicate
```

Abstract in Finnish

Tiivistelmä

Paikallisella adaptaatiolla tarkoitetaan tilannetta, jossa luonnonvalinta suosii eri fenotyyppejä eri populaatioissa. Tässä tutkimuksessa on tutkittu lituruohon, *Arabidopsis thaliana*, paikallisten adaptaatioiden genetiikkaa käyttäen siementen lepotilaa mallitapauksena. Tutkimuksessa on tarkasteltu ovatko erot siementen lepotilan kestossa adaptiivisia vertailtaessa eri lituruohopopulaatioita ja mitkä ovat ne ympäristötekijät, jotka luovat tämän valintapaineen? Mikä on paikallisen adaptaation geneettinen ja molekylaarinen tausta? Vastatakseni näihin kysymyksiin lituruohopopulaatioita tutkittiin populaatiogenetiikan menetelmin vertaillen fenotyyppiä, neutraaleita markkereita sekä kandidaattimarkkereja. Myös geenikartoituskokeita tehtiin. Tulokset osoittavat, että erot siementen lepotilan kestossa ovat adaptiivisia ja että kesäkuukausien sademäärä todennäköisesti aiheuttaa valintapaineen. Paikallinen adaptaatio on ilmeistä vertailtaessa isoja maantieteellisiä alueita. Geenikartoituskokeiden perusteella *DOG1*-geeni vastaa suuressa osin adaptaatiosta yhdessä muiden lokusten kanssa, joilla on pienempi vaikutus. Populaatiogeneettisen tutkimuksen perusteella paikallisen luonnonvalinnan vaikutus voidaan havaita *DOG1*-geenissä. Useita toiminnallisesti erilaisia *DOG1*-geenin alleeleja voidaan havaita lituruohopopulaatioissa. Siementen lepotilan kestoa pidentäviä tai lyhentäviä mutaatiota on tapahtunut useita kertoja toisistaan riippumatta. Tämä selittyy sillä, että populaatioiden kesken tapahtuva muuttoliike eli migraatio on alhainen, joten mutaatiovauhti on verrattavissa migraatiovauhtiin. Adaptiivisten mutaatioden molekylaarista perustaa ei voitu selvittää, mutta joitakin kandidaattimutaatioita havaittiin. Lisäksi jotkin tulokset herättivät kysymyksen mikä on tilastollisesti paras tapa estimoida geneettistä erilaistumista. Niinpä eräiden geneettisen erilaistumisen estimaattorien tilastollisia ominaisuuksia tarkasteltiin tietokonesimulaatioiden avulla. Mutaatiomallin huomioonottavaa estimaattoria voidaan käyttää erityyppisten markkerien vertailuun.

Acknowledgements

First, I'd like thank my supervisor Juliette for her support during these years, it was pretty tough sometimes but I learned a lot in the end. Second, I'd like to thank Maarten for giving me the opportunity to work in this project and also for sharing many stories about *Arabidopsis*. I'd also like to thank the current and former members of group de Meaux: Marilyne, Madlen, Sinéad, Fei, Jinyong, Li and the special guest star George. I learned a lot from you and it was fun. Ute provided excellent technical help in the lab and Adriana helped me with some germination experiments. People from the Soppe group shared their expertise on seed dormancy, particularly Wim, Melanie and Kazumi. The splendid technical staff at the MPI work very hard the make life easier for others, so thanks to the people at ADIS, the technicians and the gardeners. I'd especially like to thank Kurt Stüber who organised a course that introduced me to computer programming. That is probably the most useful skill I learned during my studies, Ulrieke Göbel also gave me some bioinformatic support.

I'd like to thank our collaborators and co-authors, Olivier Loudet, Xavier Pico, Carlos Alonso-Blanco. The following people also contributed *Arabidopsis* seeds and information: Valerie Le Corre, Odd-Arne Rognli and Anna Lewandowska.

During my studies I have had the opportunity to interact with many scientists who have given me thoughts on seed dormancy evolution, so thanks to Annie Schmitt and Kathleen Donohue. I'm grateful to Jérôme Goudet who let me visit his lab for a short time, and taught me a lot about F_{ST} and Q_{ST}. Thanks also to Sylvain, Alex and Sam for interesting discussions and quantiNEMO help. I'd also like to thank Bob O'Hara for advice on the Baysian models and Q_{ST}, Benjamin Stich for advice on the mixed model association and Nabil Elrouby for advice on transposons.

I would like to thank my thesis committee, Prof. Dr. Maarten Koornneef, Prof. Dr. Thomas Wiehe, Prof. Dr. Siegfried Roth and Dr. Wim Soppe for taking their time to review my work.

Terkkuja myös vanhemmille, jotka ovat aina kannustaneet eteenpäin opin tiellä. Kiitos myös lituruohoille siitä, että ne ovat vallan mainio tutkimuskohde!

"I love fools' experiments. I am always making them"
– Charles Darwin

This thesis was written with LaTeX.

Erklärung

"Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von den auf seite II angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Maarten Koornneef betreut worden."

Köln, im September 2010

Ilkka Kronholm

Lebenslauf

Persönliche Daten

Name	Ilkka Kronholm
Geburtsdatum	07.12.1983
Geburtsort	Turku, Finnland

Schulbildung

01.06.2002	Gymnasium am Turun Suomalaisen Yhteiskoulun lukio und Abitur

Studium

02.11.2006	Master of Science (Magister) an der Universität in Turku, Finnland. Mein Hauptthema war die Genetik, mein zweites Thema waren: Ökologie, Biochemie und tieresche Physiologie. Das Thema der Magisterarbeit war: Genetische Differenzierung zwischen des Blattkäfers *Galerucella sagittariae* in verschiedenen Habitaten. (auf Finnisch)
2006–heute	Doktorarbeit am Max-Planck-Institute für Pflanzenzüchtungsforschung in Köln, Deutschland. Das Thema meiner Arbeit lautet: Genetics of local adaptation in *Arabidopsis thaliana* – Seed dormancy as a case study.

Stipendien

08.06.2006	Universität von Turku, Sahlberg Entomologische Stipendium, 400 Euros für die Magisterarbeit.
05.05.2006	Kuopio Naturforschungsgesellschaft, 1000 Euros für die Magisterarbeit.
06.04.2006	Finnische Gesellschaft für Biologie Vanamo, 800 Euros für dir Magisterarbeit.

Praktika

01.06.06 – 31.08.06	Universität von Turku, Labor von Genetik: Praktikant.
11.04.05 – 31.12.05	Universität von Turku, Labor von Genetik: Teilzeit Technischer Assistent.

Publikationen

Kronholm, I., Loudet, O. & de Meaux, J. 2010. Influence of mutation rate on estimators of genetic differentiation – lessons from *Arabidopsis thaliana*. *BMC Genetics* 11: 33

Präsentationen

25.08.09	European Society for Evolutionary Biology meeting. Turin, Italien. Mündliche Präsentation: Local adaptation in *Arabidopsis thaliana* – seed dormancy as a case study.