MOLECULAR ECOLOGY

# Population structure in Arabidopsis thaliana - Implications for detecting local adaptation



| | |
|---|---|
| Journal: | *Molecular Ecology* |
| Manuscript ID: | draft |
| Manuscript Type: | Original Article |
| Date Submitted by the Author: | |
| Complete List of Authors: | Kronholm, Ilkka; Max-Planck-Institute for Plant Breeding Research, Dept. of Plant Breeding and Genetics<br>Loudet, Olivier; INRA, Génétique et amelioration des plantes<br>De Meaux, Juliette; Max Planck Institute for Plant Breeding Research, Plant Breeding and Genetics |
| Keywords: | local adaptation, Fst, computer simulation, microsatellite, SNP, Population Genetics - Empirical |
| | |

# Population structure in *Arabidopsis thaliana* – Implications for detecting local adaptation

**Ilkka Kronholm[1], Olivier Loudet[2], Juliette de Meaux[1*],**

[1]Department of Plant Breeding and Genetics, Max-Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany, [2]INRA, Genetics and Plant Breeding SGAP UR254, F-78026 Versailles, France.

[*]Corresponding author

Keywords: local adaptation, $F_{ST}$, population genetics, computer simulation, microsatellite, SNP

Adress of corresponding author:

Juliette de Meaux
Max Planck Institute for Plant Breeding Research
Carl-von-Linné-Weg 10
50829 Cologne, Germany

Tel: +49 (0) 221 50 62 465
Fax: +49 (0) 221 50 62 413
e-mail: demeaux@mpiz-koeln.mpg.de

Running title: Population structure in *A. thaliana*

Email addresses:

IK: kronholm@mpiz-koeln.mpg.de
OL: loudet@versailles.inra.fr
JdM: demeaux@mpiz-koeln.mpg.de

# Abstract

Local adaptation is often invoked as an explanation for maintenance of genetic diversity. Yet, local adaptation is difficult to detect, because patterns of genetic diversity produced by selection can be confounded by demographic effects. Characterisation of population structure is therefore of primary importance for disentangling the effects of demography from selection. In this study we describe the population structure of *Arabidopsis thaliana* at multiple spatial scales. Genetic differentiation between regions is low but differentiation between populations within regions is high. This suggests that statistical power to detect local adaptation is greatest at a regional scale in *A. thaliana*, where even weak selection may be detected. We also found that gene diversity was correlated with differentiation. This prompted us to investigate the relationship between mutation and migration rates for various estimators of genetic differentiation using computer simulations. From these studies we found that $\Phi_{ST}$ type estimator, is the only estimator that is independent from mutation rate. However, it assumes single step mutation model and displays high sampling variance. We discuss the implications of our results for studies of local adaptation and offer some suggestions for future studies.

48

# 49 Introduction

50

51 The striking match between phenotype and their local environment observed in many species
52 has motivated many evolutionary studies on local adaptation, which we define as the outcome
53 of geographically variable selection. Advances in molecular genetics have made it possible to
54 finally answer questions about the genetic basis of adaptation (reviewed in Orr, 2005). Yet,
55 demonstrating the impact of natural selection on geographic patterns of phenotypic or
56 molecular variation is not straightforward. The effect of natural selection can be confounded
57 by demographic factors, such as population structure, population growth or bottlenecks (Storz,
58 2005). However, such demographic factors should impact the whole genome whereas
59 selection should only alter genetic variation in and around the genes controlling the adaptive
60 trait. Neutral expectations can be derived from empirical distributions of diversity statistics
61 across the whole genome. In particular, the action geographically variable selection on genes
62 controlling putatively adaptive phenotypes can be inferred from the discrepancies between
63 gene specific and genome-wide patterns of population differentiation. Such genome-wide
64 patterns reflect the species population structure.

65

66 One way to quantify population structure is to use the summary statistic $F_{ST}$, which measures
67 population differentiation. Basically $F_{ST}$ and its hierarchical extensions, quantify how genetic
68 diversity is partitioned within and between populations or groups of populations (see
69 Excoffier, 2007). By using many presumably neutral markers one can build a distribution of
70 expected $F_{ST}$ values and then compare these to $F_{ST}$ values of genes that are hypothesised to be
71 subject to selection (Beaumont, 2005). This can also be done for phenotypes using $Q_{ST}$, a
72 measure of genetic differentiation in quantitative traits (reviewed in Merilä, Crnokrak, 2001).

Some recent studies have raised concerns about the reliability of $F_{ST}$ for characterisation of population structure using markers with high mutation rates, such as microsatellites (Balloux *et al.*, 2000; Hedrick, 1999; Hedrick, 2005; Jost, 2008). High levels of within population diversity bias the classical Wright's $F_{ST}$. The problem is rooted in the mathematics of $F_{ST}$. If a locus has multiple alleles, classical $F_{ST}$ can be low even if populations share no alleles (Hedrick, 2005; Jost, 2008). Alternative estimators have been proposed to solve the problem. An analogous estimator to $F_{ST}$, $\Phi_{ST}$, takes into account the distances between alleles thereby correcting for mutation rate (Excoffier, 2007; Slatkin, 1995). 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, true differentiation or $D$, measures differentiation between populations that is not bound by levels of within population variation (Jost, 2008). These estimators provide improved measures of differentiation between populations, yet their usefulness for detecting signatures of local adaptation is not clearly established.

Our study system, *Arabidopsis thaliana* (L.) Heyhn. (Brassicaceae), is a model organism for many aspects of plant molecular biology. In recent years, it has become a model in plant population genetics, because the molecular details of adaptation can be elucidated (Mitchell-Olds, Schmitt, 2006). *A. thaliana* is an annual weed that colonises disturbed habitats. It is distributed from northern Europe to northern Africa and from western Europe to central Asia (Hoffmann, 2002). *A. thaliana* therefore grows in a wide range of habitats. Recent studies have also revealed that it has considerable genetic variation in its genome, despite its high rates of self-fertilisation (Clark *et al.*, 2007; Nordborg *et al.*, 2005; Schmid *et al.*, 2005). In addition, *A. thaliana* displays abundant phenotypic variation for most traits, such as flowering time, seed dormancy, seed size, tolerance to several abiotic stresses including drought,

(reviewed in Alonso-Blanco, Koornneef, 2000). Interestingly, there is evidence that flowering time variation is adaptive in *A. thaliana* (Caicedo *et al.*, 2004; Le Corre, 2005; Toomajian *et al.*, 2006). However, most studies of local adaptation in *A. thaliana* have focused on investigating adaptation either at a small geographical scale (Kuittinen *et al.*, 1997; Le Corre, 2005; Stenoien *et al.*, 2005) or at a species wide scale (Caicedo *et al.*, 2004; Hopkins *et al.*, 2008; Samis *et al.*, 2008; Stinchcombe *et al.*, 2004) and no study so far has combined these two scales. Recent studies population genetics in *A. thaliana* have started to put more emphasis on within population sampling but different markers have been used, making comparisons between datasets difficult (He *et al.*, 2007; Le Corre, 2005; Pico *et al.*, 2008; Stenoien *et al.*, 2005).

We ask the following questions in this study: 1) What is population structure of *A. thaliana* at different spatial scales? 2) Given the population structure of *A. thaliana*, what is the best scale to look for adaptation, i. e. where statistical power is the greatest? 3) What is the best estimator of genetic differentiation in context of detecting local adaptation? To answer these questions, we characterised population structure of *A. thaliana* in a hierarchical sample and quantify how genetic variation is distributed at different spatial scales. We use a common set of markers to genotype 41 *A. thaliana* populations located in four geographic regions (Spain, France, Norway and central Asia). We also used computer simulations to examine the performance of different estimators of genetic differentiation and discuss which of them is best suited for studies of local adaptation.

# Methods

## Population samples

In total 289 individuals from 41 populations were genotyped. Detailed information about the populations can be found in the supplementary material (Table S1). A map showing the locations of the sampled populations is shown in figure 1. We analysed 7, 15, 13 and 6 populations from Spain, France, Norway and Central Asia, respectively. Number of sampled individuals from each population ranges from 3 to 11 with a mean of 7. Three regions in Western Europe: Spain, France, Norway create a South – North cline. The fourth region is composed of populations from Kyrgyzstan and Tajikistan, and will be referred to as the Central Asian region throughout the paper. The Spanish populations are described in Pico *et al.* (2008). French populations were collected by Valerie Le Corre and some of them are described in Le Corre (2005). The Norwegian populations were obtained from O. A. Rognli through NARC (Norway). Populations from Central Asia were collected by OL and described at http://www.inra.fr/vast/collections.htm. Field collected plants were subjected to one or two generations of self-fertilisation in the greenhouse before DNA extraction.

## Genotyping

DNA was extracted from young leaves using BioSprint 96 robot and BioSprint 96 DNA Plant Kit (Qiagen) according to manufacturer's instructions.

Plants were genotyped at 24 microsatellite loci, 20 of which are located in the nuclear genome and 4 in the chloroplast genome. Details of the microsatellite loci used and genotyping procedures can be found in the supplementary material (Table S2). Microsatellites were amplified using standard methods and allele sizes were determined using capillary

electrophoresis. To determine the actual number of repeats in each allele, the accession Col-0 was also genotyped for each locus and using its PCR product size and the genome sequence the number of repeats was deduced for each allele. The Spanish accessions had already been genotyped previously for some of the loci used here, as described in Pico *et al.* (2008). We verified that our allele sizes corresponded to the allele sizes reported previously by re-genotyping a subsample at selected alleles.

The plants were also genotyped for a set of 149 single nucleotide polymorphism (SNP) markers (developed by Warthmann *et al.*, 2007) by Sequenom, inc. (San Diego, CA). Detailed description of the SNP markers is found in supplementary material (Table S4). Out of the 149 SNP markers, 137 had good quality data and were polymorphic in the whole sample.

**Data analysis – Genetic diversity and population structure**

All statistical analyses were done using the statistical environment R (R project core team, 2006) unless otherwise stated. Methods not implemented by R-packages were implemented via R-scripts written by IK and are available upon request.

For analysis of genetic diversity, the microsatellite data were used. Only the 20 nuclear microsatellites were used to analyse of genetic diversity and F-statistics. Measures of genetic diversity: Nei's gene diversity (Hs) and allelic richness, allelic richness is a measure of the number of alleles independent of sample size, were calculated using FSTAT 2.9.3 (Goudet, 2001). Differences in measures of genetic diversity between groups of populations were tested using permutation tests, permuting populations among regions, implemented in FSTAT. The microsatellite population mutation rate, $\theta$, is the product of effective population size and

mutation rate at a locus was calculated following equation 15 of Kimmel *et al.* (1998). The performance of this summary statistic based method has been shown to be comparable to likelihood-based methods (RoyChoudhury, Stephens, 2007). $\theta$ was calculated for each locus within each region. For SNP data the minor allele frequency was calculated for each locus in each region.

$F_{ST}$ was estimated according to Weir & Cockerham (1984) for microsatellites and SNP markers, using the R-package "hierfstat" (Goudet, 2005). All other genetic differentiation methods were implemented via R-scripts written by IK. For microsatellites the standardised genetic differentiation measure, $F'_{ST}$ (Hedrick, 2005), was estimated using the maximised variance component method of Meirmans (2006). In order to take the distance between the microsatellite alleles into account (Slatkin, 1995) we estimated $\Phi_{ST}$ in an AMOVA framework (Michalakis, Excoffier, 1996). 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.

To check correlations between different marker types and methods we calculated pairwise $F_{ST}$ values between populations using different marker types and analysis methods. Here the matrices of pairwise $F_{ST}$ values are not fully independent because both matrices being compared include the same set of populations. Therefore Mantel-tests were used to assess the statistical significance of the correlations. Mantel-tests were done using the R-package "vegan" (Oksanen *et al.*, 2007).

To investigate population structure in our sample independently of our sampling we used the program STRUCTURE v2.1 (Falush *et al.*, 2003; Pritchard *et al.*, 2000). STRUCTURE uses a

196 model based clustering algorithm that assigns individuals probabilistically to genetic clusters,
197 where the number of clusters (*K*) is initially not known. This analysis allows us to answer the
198 question: to how many genetic clusters (populations) can the data be partitioned into, given
199 the observed data? Given that *A. thaliana* is highly inbreeding, the data was analysed in
200 haploid format. In the analysis we used a model with linked loci, both nuclear and chloroplast
201 microsatellite markers and SNPs were included in the analysis. Genetic distances between the
202 markers were based on *Arabidopsis* physical and genetic maps. Allele frequencies were
203 assumed to be correlated. In the simulations we set length of burnin to 100000, admixture
204 burnin to 50000 and the number of MCMC replicates to 200000. The probability of the data
205 under the model was estimated. Other options were left as default. Simulations were run 5
206 times for each value of *K*. The optimal value of *K* was evaluated using the $\Delta K$ method
207 (Evanno *et al.*, 2005). Since the $\Delta K$ method has been shown to detect higher order population
208 structure (Evanno *et al.*, 2005), we investigated population structure of our data using a
209 hierarchical approach. After initial round of analysis the resulting clusters, and in some case
210 regions, were taken and clustering was performed again for a subset of the data. Plots of
211 individual assignments were drawn using the program DISTRUCT (Rosenberg, 2004).
212
213 Since *A. thaliana* is naturally highly inbred, some of the assumptions made by STRUCTURE
214 are not met. We thus also analysed our data using principal component analysis as
215 implemented in the R package "adegenet" (Jombart, 2008). Unlike STRUCTURE, which
216 implements a model based clustering, principal component analysis permits us to analyse
217 genetic structuring of the populations without making assumptions of Hardy-Weinberg
218 equilibrium or linkage equilibrium. We used both microsatellite and SNP markers in the
219 principle component analysis.
220

221 **Data analysis – Computer simulations**

222 In order to investigate the behaviour of $F_{ST}$, $F'_{ST}$, $\Phi_{ST}$ and $D$ under high mutation rates,
223 computer simulations using EasyPop 1.8 (Balloux, 2001) were performed. The simulation
224 scheme was set to 10 populations with 500 individuals each, 20 freely recombining loci and
225 random mating hermaphrodites. Populations followed an island model of migration.
226 Migration rates (probability that a given individual will migrate in each generation), $m$ ranged
227 from 0.1 to 0.00001 and mutation rates (probability that a given allele will mutate in each
228 generation), $\mu$ from 0.00001 to 0.01. In order to simulate microsatellite loci we first examined
229 a pure single step mutation model. Then we relaxed this assumption by using a mixed
230 mutation model in which the loci followed a single step mutation model but with the
231 probability of 0.2 to mutate to any state. The number of possible allelic states was set to 30.
232 The effect of self-fertilisation was examined by doing simulations with proportion of self-
233 fertilisation set to 0.9. Simulations were run for 2000 generations. In the end to simulate
234 realistic sampling situation, 30 individuals were sampled from each population for parameter
235 estimation. Each simulation was repeated 5 times for a given set of parameter values. For
236 each simulated dataset we calculated genetic differentiation statistics, gene diversity (*Hs*), and
237 microsatellite population mutation rate ($\theta$) as described earlier.
238
239 We also performed coalescent simulations to investigate the effect of different marker types
240 on $F_{ST}$ calculations. We investigated sequence haplotypes (these would be derived by
241 sequencing a number of loci from many individuals), independent single SNP markers and
242 microsatellite markers following a single step mutation model. All coalescent simulations
243 were performed using the program ms (Hudson, 2002). We simulated an island model of
244 population structure with 10 populations, 20 individuals were sampled from each population.
245 For sequence haplotypes and microsatellites 30 independent loci were simulated, for SNP
246 markers we 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. We set up the simulations so that the effective population size was 1000 for each population and then parameters $m$ and $\mu$ ranged from 0.0001 to 0.1 for $m$ and 0.00001 to 0.001 for $\mu$. Each simulation with given a parameter set was repeated 5 times.

# Results

### Genetic diversity in *Arabidopsis thaliana*

Genetic diversity was measured using microsatellite markers. Plots of microsatellite allele sizes illustrate some variation between regions and among loci (supplementary material, Figure S1). On the basis of visual inspection, there appears to be no gross departure from the continuous distribution of allele sizes predicted by single step mutation model. To examine whether genetic diversity is the same in each of the regions, indices of genetic diversity were calculated for each region (Table 1) In Norway and Asia the least number of alleles were observed. Concordantly, allelic richness was highest in Spain (AR = 2.269), intermediate in France (AR = 1.720), lowest in Norway and Asia (1.245 and 1.383, respectively, Table 1.). Differences in allelic richness were significant ($p < 0.05$; 1000 permutations) in all comparisons except when comparing the Central Asian populations to those in Norway or France. Similar trend could be observed for gene diversity (Table 1).

For the 137 polymorphic SNP markers in the total sample, 135 were polymorphic in Spain, all 137 loci were polymorphic in France, 119 were polymorphic in Norway and only 67 loci were polymorphic within Central Asia. Minor allele frequency plots for each region are shown in Figure 2. SNPs used in this study are biased toward high frequency. See supplementary

material for description of ascertainment scheme, ascertainment bias cannot be corrected in our sample (see supplementary material and Figure S3). Therefore, we examined the effect of ascertainment bias on $F_{ST}$ in the dataset in which they were selected (Nordborg *et al.*, 2005), hereafter we call this the Nordborg dataset. When we used the 137 SNPs used in our study to calculate $F_{ST}$ between genetic clusters defined by Nordborg *et al.* (2005) in the Nordborg dataset, we obtained $F_{ST}$ = -0.0018. Then we sampled 137 SNPs at random from the Nordborg dataset 1000 times and calculated $F_{ST}$ between the genetic clusters for each sampled dataset, the 95 % quantile for $F_{ST}$ in the sampled datasets was -0.0051 – 0.0271.

**Population structure of *Arabidopsis thaliana***

We tested whether our geographically structured sample of A. thaliana exhibited hierarchical population structure. We therefore analysed our data using the program STRUCTURE. To determine the number of clusters, plots of likelihood and $\Delta K$ values of different values of $K$ were analysed (supplementary material, Figure S4).

When all populations were included in the analysis the highest $\Delta K$ value was observed for $K$ = 3 (Figure S4), this partition groups Spanish and French populations together, leaving Norwegian populations as one group and Central Asian populations as a third group (Figure 3). We then analysed each of the geographic regions separately. Higher order clustering was observed within most regions. At the lowest hierarchical level, genetic clusters mostly correspond to sampled populations, with the exception of some Norwegian populations (Figure S5). There also seems to be some migration events and admixture detectable especially in the Spanish and French populations (see supplementary material for details).

Principal component analysis mostly corroborated results obtained with STRUCTURE. When all populations are included in the analysis, the first principal component (PC) separates the Central Asian populations from the rest and the second PC separates the Norwegian populations from the Spanish and French (Figure 4). PC 1 and 2 explain 10.4 % and 8.5 % of the genetic variation, respectively. The remaining components explain considerably less variation. When analysing only the Spanish or Central Asian populations PCA results are similar to STRUCTURE results. In France and Norway STRUCTURE detects several more clusters than the first components of PCA (Figure S6).

**Genetic differentiation in *Arabidopsis thaliana***

We observed that genetic differentiation ($F_{ST}$) for microsatellite loci correlates with gene diversity (*Hs*) and the population mutation rate ($\theta$) (Figure 5). For instance, in the Spanish populations the correlation between *Hs* and $F_{ST}$ was r = -0.862 (95 % CI = -0.944 – -0.678) with $p < 0.001$ (Table 2). We further examined the correlation between diversity and various alternative estimators of differentiation. There was positive albeit non-linear relationship between *Hs* and $F'_{ST}$, ( r = 0.479, [95 % CI = 0.076 – 0.760], p = 0.033). $\Phi_{ST}$ was not correlated with *Hs*, ( r = -0.294, [95 % CI = -0.652 – 0.170], p = 0.208). A similar pattern was observed when $\theta$ was used instead of Hs (data not shown). $\Phi_{ST}$ instead, is independent from genetic diversity in our data, except in Central Asian populations (Table 2, Figure 5).

Next we calculated measures of genetic differentiation for microsatellites and SNP markers between populations within regions and between regions (Figure 6). For microsatellites genetic differentiation between populations was the 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 (Figure 6). However, it should be noted that the confidence intervals are sometimes broad (Figure 6), especially in Central Asia. Genetic differentiation between geographic regions was smaller than between populations within regions (Figure 6). Differentiation measured by $F'_{ST}$ from microsatellites was generally higher than with other methods (Figure 6).

We also examined whether different types of markers and methods give consistent results regarding population structure. We calculated pairwise $F_{ST}$ values between populations within each of the geographical regions using microsatellite and SNP markers. If different markers and methods give consistent results pairwise $F_{ST}$ values should be correlated. Then we used Mantel-tests to determine the significance of correlations. Correlations are presented in table 3. In Spain, the correlation between the $F_{ST}$ in microsatellites and SNP markers is not significant ($r = 0.414$, $p = 0.135$). The correlation between microsatellite $\Phi_{ST}$ and SNP $F_{ST}$ is $r = 0.584$, $p = 0.067$. In French populations correlation between microsatellite $F_{ST}$ and SNP $F_{ST}$ is $r = 0.977$, $p < 0.001$. Correlation between microsatellite markers and SNP markers are moderate in Norwegian and central Asian populations (Table 3). Correcting for high microsatellite mutation rates using $\Phi_{ST}$ decreases correlations between microsatellite and SNP markers in all regions except Spain (Table 3). We also examined the correlation between pairwise microsatellite differentiation using different analysis methods. Correlations between microsatellite $F_{ST}$ and $F'_{ST}$ were generally high. However, correlations between microsatellite $F_{ST}$ and $\Phi_{ST}$ seemed to be lower (Table 3); in the Central Asian populations correlation between microsatellite $F_{ST}$ and $\Phi_{ST}$ was not significant, $r = 0.408$, $p = 0.113$.

## Computer simulations

We used forward population genetic simulations to investigate the behaviour of different estimators with varying migration and mutation rates. The best estimator, in the context of local adaptation, should be robust to mutation rate to allow comparisons between different marker types. Results of forward population genetic simulations show that FST tends to zero when mutation rate increases (Figure 7). Replicate simulations cluster very well showing that there is little variance among replicates. This results follows the analytical expectation presented in Hedrick (2005). If mutations follow a pure single step model $\Phi_{ST}$ is essentially independent from mutation rate (Figure 7). $F'_{ST}$ and $D$ are not independent from mutation rate. In our simulations it can be observed that unexpectedly, when migration rate is very low, increasing mutation rate up to 0.01 causes also $F'_{ST}$ and $D$ to go downward (Figure 7, panels C and D). If the assumptions of single step mutation model are relaxed $\Phi_{ST}$ has the same trend as $F_{ST}$ although the effect is somewhat slower (Figure 8.). In our mixed mutation model there is a probability of 0.2 that a mutation generates an allele of any size. Results are the same if rate of self-fertilisation was set to 0.9 (Figure S7).

Next we examined the effect of mutation rate on different marker types. We simulated DNA haplotypes (derived by re-sequencing short fragments from multiple individuals), microsatellite markers and single SNP markers. Results from the simulations are presented in figure 9. We calculated $\Phi_{ST}$ that takes into account distance between different haplotypes or microsatellite alleles. By applying this method to both haplotypes and microsatellites gives essentially the same results (Figure 9) and $\Phi_{ST}$ is independent from mutation rate for both marker types. Single SNP markers also give $F_{ST}$ values that are nearly identical to the ones obtained with other types of markers. Mean SNP $F_{ST}$ = 0.028, 0.040, 0.206 and 0.695 for migration rates $m$ = 0.1, 0.01, 0.001 and 0.0001 respectively. This is in accordance with haplotype and microsatellite markers (Figure 9). Therefore, $\Phi_{ST}$ for haplotype data,

microsatellites (following single step mutation model) and $F_{ST}$ for single SNPs (free of ascertainment bias) give comparable estimates of differentiation.

# Discussion

In this study we have characterised the population structure of *A. thaliana* at multiple spatial scales and discuss its implications for detecting local adaptation. We found that the traditional $F_{ST}$ estimator is not suitable for comparing different marker types; instead $\Phi_{ST}$ which takes mutation rate into account is theoretically preferable. However, in practise $\Phi_{ST}$ assumes a mutation model, which may not be correct in all cases. We document extensive population structure in *A. thaliana*. Geographic regions are weakly differentiated from one another but there is strong differentiation between populations within regions. When differentiation in neutral markers is high, it will be difficult to detect outlier loci or higher $Q_{ST}$ than $F_{ST}$ even if selection is strong. This is a consequence of the non-linear relationship of $F_{ST}$ to population migration rate, $F_{ST} = 1 / (4Nm + 1)$. Therefore detecting local adaptation in our sample of *A. thaliana* using $F_{ST}$ vs. $Q_{ST}$ comparisons has the highest statistical power between regions. On the other hand, selection for maintaining diversity may be easier to detect within regions where $F_{ST}$ is high (Beaumont, Balding, 2004).

## Implications for detecting local adaptation

We found that genetic diversity is related to genetic differentiation in our sample, indicating that microsatellite mutation rates are high relative to migration rate (Figure 7, Figure 5 and Table 2). A relationship between genetic diversity and differentiation has also been observed in other studies and organisms. Carreras-Carbonell *et al.* (2006) studied two subspecies found in the triplefin fish, *Tripterygion delaisi*. They found that $F_{ST}$ was negatively correlated with

expected heterozygosity (r = -0.9) and that genetic differentiation between the two subspecies was low using traditional $F_{ST}$. However, clustering methods clearly differentiated the two subspecies and using an analogous measure of $F'_{ST}$, genetic differentiation was higher. Similarly O'Reilly *et al.* (2004) found a relationship between heterozygosity and $F_{ST}$ in the fish walleye pollock, *Theragra chalcogramma*, which was erroneously attributed to homoplasy. This relationship has also been found in *Arabidopsis lyrata*, a relative of *A. thaliana* exhibiting a markedly different life-history and more genetic diversity than *A. thaliana* (Clauss, Mitchell-Olds, 2006; Muller *et al.*, 2007). This shows that a wide variety of organisms are in the parameter space where variation in $F_{ST}$ reflects variation in both migration and mutation rates.

We used computer simulations to examine the behaviour of different estimators of genetic differentiation in response to changing migration and mutation rates. Our results show that, because it takes distances between different alleles into account, $\Phi_{ST}$ is the only estimator that is completely independent from mutation rate if its assumptions are met (Figure 7) (Slatkin, 1995). However, deviations from pure single step mutation model make this type of estimator dependent on mutation rate (Figure 8). We also showed that if the assumptions of $\Phi_{ST}$ are met, both DNA haplotypes and microsatellites give comparable estimates to single SNP $F_{ST}$ (Figure 9). In *A. thaliana* as in many other species, practise, microsatellite loci were often shown to deviate from single step mutation model (Calabrese, Sainudiin, 2005; Ellegren, 2004; Symonds, Lloyd, 2003). In *A. thaliana* many microsatellite loci were shown to deviate from the single step mutation model. Hence, caution is needed when using $\Phi_{ST}$. Our simulations further demonstrate that $D$ or $F'_{ST}$ both depend on mutation rate (Jost, 2008) (Figure 7 and supplementary material). 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. This means that $F'_{ST}$ or $D$ are useful for studies where the

amount of genetic differentiation is of interest *per se*, such as in conservation studies (Hedrick, 2005; Jost, 2008). Instead, they can be misleading for studies of local adaptation, where the goal is to compare markers with different mutation rates.

Since microsatellites raise some concerns, using SNP markers seems preferable since they usually have two alleles and low mutation rate and thus avoid the problems associated with $F_{ST}$ (Jost, 2008). The SNP loci in our study show an ascertainment bias (Figure 2 and supplementary material). Yet, when we examined the effect of ascertainment bias in the dataset where the SNP markers were ascertained, the effect was minor. Although it is not certain that this behaviour would be the same in our dataset, estimates of differentiation based on SNPs are broadly concordant with $\Phi_{ST}$ estimates based on microsatellites (Figure 6). Different markers and methods are also often correlated with each other (Table 3). In Norway and central Asia however, microsatellite $\Phi_{ST}$ seems to be lower than $F_{ST}$ for SNP markers (Figure 6). There are two possible explanations for this: either microsatellites do not follow single step mutation model very accurately, or ascertainment bias is affecting the SNP markers differently in some regions, or both. In many cases, confidence intervals for $\Phi_{ST}$ are broad in our data, a likely consequence of the high sampling variance displayed by $\Phi_{ST}$ type estimators Balloux & Lugon-Moulin (2002).

It is known that there is great deal of variation in mutation rates between different genes due to evolutionary constraints (Clark *et al.*, 2007). In order to directly compare differentiation across genes mutation rate has to be taken into account. However, the implications of high mutation rate for using $F_{ST}$ to detect loci under selection are smaller than for $F_{ST}$ vs. $Q_{ST}$ comparisons. The relationship between diversity and differentiation in studies of local adaptation was considered by Beaumont & Nichols (1996) in their method to detect outlier loci which jointly considers heterozygosity and $F_{ST}$ (Beaumont, Balding, 2004; Beaumont,

Nichols, 1996). This method was shown to be robust to mutation rate variation among loci (Beaumont, Nichols, 1996).

Our results have greater implications for studies of local adaptation based on $Q_{ST}$ vs. $F_{ST}$ comparisons. A recent review and meta-analysis of $F_{ST}$ vs. $Q_{ST}$ studies (Leinonen *et al.*, 2008) noted that using $F'_{ST}$ would generally change the conclusions of $F_{ST}$ vs. $Q_{ST}$ studies. However, our study shows that using $F'_{ST}$ or $D$ in $Q_{ST}$ studies is not appropriate, because true measures of genetic differentiation are not independent from the high mutation rate of microsatellites. This may cause the true measure to be overly conservative. If microsatellites are used it should be first examined whether or not there is a relationship between diversity and differentiation in the data. If this is the case then $\Phi_{ST}$ type estimator should be used. Nonetheless, our simulation also shows that in the case of high mutation rate and departure from single step mutation model, $\Phi_{ST}$ is also misleading. SNP markers do not suffer from these issues in $F_{ST}$ estimation because they are bi-allelic. This consideration suggests that SNPs or DNA haplotypes generated by re-sequencing should be preferably used, yet it may not be feasible in non-model organisms. Moreover, SNP markers may often have some ascertainment bias.

**Population structure in *Arabidopsis thaliana***

At a fine geographical scale, sampled populations of *A. thaliana* often correspond with genetic clusters determined by clustering methods (Supplementary data). Therefore treating the sampled populations as separate genetic populations is justified. Our results show that *A. thaliana* populations from different regions are differentiated. In a cluster analysis, the Spanish and French populations are grouped together, whereas the Norwegian and Central

Asian regions each form a separate cluster (Figure 3 and Figure 4). This is consistent with previous results (Beck *et al.*, 2008; Nordborg *et al.*, 2005; Schmid *et al.*, 2006). Our results further confirm that Iberian Peninsula has the highest genetic diversity and was a glacial refugia for *A. thaliana* (Beck *et al.*, 2008; Pico *et al.*, 2008; Sharbel *et al.*, 2000). Differentiation between Spain and France appears to be very small (Figure 6) as expected if French populations are derived from Iberian populations (Pico *et al.*, 2008). Colonisation may have occurred from multiple sources after last glaciation in France and Norway, as suggested by higher order population structure observed in these regions (Supplementary material, see also Nordborg *et al.*, 2005; Stenoien *et al.*, 2005.). French populations may include individuals that come from the Apennine Peninsula, Balkans or even central Asia (Beck *et al.*, 2008; Sharbel *et al.*, 2000). Genetic differentiation between populations within regions is high (Figure 6) indicating that *A. thaliana* is geographically structured on a small scale throughout its range. For microsatellites $F'_{ST}$ is higher than other estimates. This indicates that true genetic differentiation also between regions, due to input of new mutations, can be substantial. Differentiation where effects of mutations are not considered is lower (Figure 6).

Geographic structuring seems to vary between regions, since $F_{ST}$ estimates tend to follow an opposite trend as genetic diversity. This is not just an effect of lowered genetic diversity on $F_{ST}$ estimates since the effect is consistent across different methods and marker type (Figure 6.). It may instead reflect lower effective population size outside Iberian Peninsula caused by bottlenecks that occurred during colonisation of Western Europe. Reduced level of genetic differentiation between regions despite high population differentiation suggest that effective population size is much smaller within populations and so genetic drift is faster within populations than over whole regions.

Low differentiation between regions suggests that in A. thaliana, local adaptation is most likely to be detected across broad regions. Therefore we may be able to detect local adaptation that has occurred when *A. thaliana* migrated from Iberian Peninsula to France and Northern Europe. Instead, we may lack the power to detect adaptation that has occurred on a fine geographical scale. From South of Spain to Northern Norway, our population sample spans widely different climates and photoperiods. There are many phenotypes and candidate genes whose adaptive relevance can be studied in this context, like genes regulating flowering time e. g. *PHYTOCROME C* studied in Balasubramanian et al. (2006) and Samis et al. (2008). Other examples are seed dormancy, cold tolerance, for which some candidate genes controlling its natural variation have been recently found (Alonso-Blanco *et al.*, 2005). The comparison between Western European and Central Asian populations can be used also to study adaptive relevance of traits found in the accessions Shahdara, Kondara and Kashmir-1 and 2 that come from this region. These lines are parents of many publicly available QTL-mapping populations, e. g. Loudet *et al.* (2002), see also the website http://www.inra.fr/vast/RILs.htm. A number of interesting genes contributing to natural variation have been cloned in these mapping populations, such as loci causing genetic incompatibility, variation in growth or sulfate content (Alcázar *et al.*, 2009; Loudet *et al.*, 2008; Loudet *et al.*, 2007), to cite but a few examples.

Since *A. thaliana* is primarily a self fertilising plant, different genotypes can be propagated essentially indefinitely. The lines described here will be made available to the community along with the genotypic data permitting investigators to test whether their phenotypes or genes of interest may be locally adapted.

# References


Alcázar R, García AV, Parker JE, Reymond M (2009) Incremental steps toward incompatibility revealed by Arabidopsis epistatic interactions modulating salicylic acid pathway activation. *Proceedings of the National Academy of Sciences, USA* **106**, 334-339.

Alonso-Blanco C, Gomez-Mena C, Llorente F, *et al.* (2005) Genetic and molecular analyses of natural variation indicate CBF2 as a candidate gene for underlying a freezing tolerance quantitative trait locus in Arabidopsis. *Plant Physiology* **139**, 1304-1312.

Alonso-Blanco C, Koornneef M (2000) Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. *Trends in Plant Science* **5**, 22-29.

Balasubramanian S, Sureshkumar S, Agrawal M, *et al.* (2006) The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana. *Nature Genetics* **38**, 711-715.

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**, 1414-1422.

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**, 155-165.

Beaumont MA (2005) Adaptation and speciation: what can $F_{st}$ tell us. *Trends in Ecology and Evolution* **20**, 435-440.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B Biological Sciences* **263**, 1619-1626.

Beck JB, Schmuths H, Schaal BA (2008) Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology* **17**, 902-915.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences, USA* **101**, 15670-15675.

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

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.

Clark RM, Schweikert G, Toomajian C, *et al.* (2007) Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. *Science* **317**, 338-342.

Clauss MJ, Mitchell-Olds T (2006) Population genetic structure of Arabidopsis lyrata in Europe. *Molecular Ecology* **15**, 2753-2766.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetetics* **5**, 435-445.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

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

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.
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/izea/softwares/fstat.html.
Goudet J (2005) hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184-186.
He F, Kang D, Ren Y, *et al.* (2007) Genetic diversity of the natural populations of Arabidopsis thaliana in China. *Heredity* **99**, 423-431.
Hedrick PW (1999) Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**, 313-318.
Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* **59**, 1633-1638.
Hoffmann MH (2002) Biogeography of Arabidopsis thaliana (L.) Heynh. (Brassicaceae). *Journal of Biogeography* **29**, 125-134.
Hopkins R, Schmitt J, Stinchcombe JR (2008) A latitudinal cline and response to vernalization in leaf angle and morphology in Arabidopsis thaliana (Brassicaceae). *New Phytologist* **179**, 155-164.
Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338.
Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.
Jost L (2008) Gst and its relatives do not measure differentiation. *Molecular Ecology* **17**, 4015-4026.
Kimmel M, Chakraborty R, King JP, *et al.* (1998) Signatures of Population Expansion in Microsatellite Repeat Data. *Genetics* **148**, 1921-1930.
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)**, 144-152.
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**, 4181-4192.
Leinonen T, O'Hara RB, Cano JM, Merila J (2008) Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology* **21**, 1-17.
Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F (2002) Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics* **104**, 1173-1184.
Loudet O, Michael TP, Burger BT, *et al.* (2008) A zinc knuckle protein that negatively controls morning-specific growth in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences, USA* **105**, 17193-17198.
Loudet O, Saliba-Colombani V, Camilleri C, *et al.* (2007) Natural variation for sulfate content in Arabidopsis thaliana is highly controlled by APR2. *Nature Genetics* **39**, 896-900.
Meirmans PG (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.
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.
Mitchell-Olds T, Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* **441**, 947-952.

620 Muller MH, Leppälä J, Savolainen O (2007) Genome-wide effects of postglacial colonization
621 in Arabidopsis lyrata. *Heredity* **100**, 47-58.
622 Nordborg M, Hu TT, Ishino Y*, et al.* (2005) The pattern of polymorphism in Arabidopsis
623 thaliana. *PLoS Biology* **3**, e196.
624 O'Reilly P, Canino M, Bailey K, Bentzen P (2004) Inverse relationship between FST and
625 microsatellite polymorphism in the marine fish walleye pollock (Theragra
626 chalcogramma): implications for resolving weak population structure. *Molecular*
627 *Ecology* **13**, 1799-1814.
628 Oksanen J, Kindt R, Legendre P*, et al.* (2007) vegan: Community Ecology Package. R
629 package version 1.8-7.
630 Orr HA (2005) The genetic theory of adaptation: a brief history. *Nat Rev Genet* **6**, 119-127.
631 Pico FX, Mendez-Vigo B, Martinez-Zapater JM, Alonso-Blanco C (2008) Natural Genetic
632 Variation of Arabidopsis thaliana is Geographically Structured in the Iberian
633 Peninsula. *Genetics* **180**, 1009-1021.
634 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using
635 multilocus genotype data. *Genetics* **155**, 945-959.
636 Rosenberg NA (2004) DISTRUCT: a program for graphical display of population structure.
637 *Molecular Ecology Notes* **4**, 137-138.
638 RoyChoudhury A, Stephens M (2007) Fast and Accurate Estimation of the Population-Scaled
639 Mutation Rate, {theta}, From Microsatellite Genotype Data. *Genetics* **176**, 1363-1366.
640 Samis KE, Heath KD, Stinchcombe JR (2008) Longitudinal Clines in Flowering Time and
641 Phytochrome C in Arabidopsis Thaliana. *Evolution* **26**, 26.
642 Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A
643 multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure
644 from a neutral model of DNA sequence polymorphism. *Genetics* **169**, 1601-1615.
645 Schmid KJ, Torjek O, Meyer R*, et al.* (2006) Evidence for a large-scale population structure
646 of Arabidopsis thaliana from genome-wide single nucleotide polymorphism markers.
647 *Theoretical and Applied Genetics* **112**, 1104-1114.
648 Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in Arabidopsis
649 thaliana: biogeography and postglacial colonization of Europe. *Molecular Ecology* **9**,
650 2109-2118.
651 Slatkin M (1995) A Measure of Population Subdivision Based on Microsatellite Allele
652 Frequencies. *Genetics* **139**, 457-462.
653 Stenoien HK, Fenster CB, Tonteri A, Savolainen O (2005) Genetic variability in natural
654 populations of Arabidopsis thaliana in northern Europe. *Molecular Ecology* **14**, 137-
655 148.
656 Stinchcombe JR, Weinig C, Ungerer M*, et al.* (2004) A latitudinal cline in flowering time in
657 Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proceedings of*
658 *the National Academy of Sciences, USA* **101**, 4712-4717.
659 Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population
660 divergence. *Molecular Ecology* **14**, 671-688.
661 Symonds VV, Lloyd AM (2003) An analysis of microsatellite loci in Arabidopsis thaliana:
662 mutational dynamics and application. *Genetics* **165**, 1475-1488.
663 team RDc (2006) R: A language and environment for statistical computing. http://www.R-
664 project.org. R Foundation for Statistical Computing, Vienna, Austria.
665 Toomajian C, Hu TT, Aranzana MJ*, et al.* (2006) A nonparametric test reveals selection for
666 rapid flowering in the Arabidopsis genome. *PLoS Biology* **4**, e137.
667 Warthmann N, Fitz J, Detlef W (2007) MSQT for choosing SNP assays from multiple DNA
668 alignments. *Bioinformatics* **23**, 2784-2787.
669 Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population
670 structure. *Evolution* **38**, 1358-1370.

## Acknowledgements

We thank Valerie Le Corre for providing the French populations, Odd-Arne Rognli and Anna Monika Lewandowska for providing the Norwegian populations and Carlos Alonso-Blanco and Xavier Pico for providing the Spanish populations. We thank Sinead Collins, Maarten Koornneef, Maria Clauss and Marie-Hélène Muller for comments on the manuscript. Funding was provided by Max Planck Gesellschaft and by a grant from DFG to JdM within the collaborative research network SFB680.

This work belongs to a project investigating local adaptation in *Arabidopsis thaliana*, focusing on seed dormancy. Ilkka Kronholm is a PhD student, whose thesis this study is a part of. He is interested in the genetic basis and population genetics of adaptation. Juliette de Meaux is interested in the molecular basis of adaptive innovations in plants. Her research focuses on various phenotypes of adaptive relevance, such as seed dormancy and innate immunity. Olivier Loudet's main interest is in the quantitative genetics of natural variation and genotype x environment interactions.

# Figure legends

**Figure 1 - Map of sampled populations**

Geographical locations of populations used in this study. Inset shows the Central Asian region and overview.

**Figure 2 - SNP minor allele frequency distributions for each region**

X-axis shows the minor allele frequency. Note that y-axis has a relative frequency density scale; the area under the histogram is equal to one.

**Figure 3 - STRUCTURE plot of all populations analysed together**

STRUCTURE plot with K = 3, all populations used. Labels above the figure designate the regions for populations. Labels below the figure are the populations. Each vertical column is an individual and height of the coloured bars is proportional to the probability of belonging to one of three clusters.

**Figure 4 - PCA analysis of all populations**

Principal component analysis of all populations. First and second principal components are plotted. Labels designate populations, different regions are indicated in the plot.

**Figure 5 - Correlations between genetic diversity and genetic differentiation in Spanish populations**

Gene diversity, Hs, was calculated for each locus and is plotted against different estimators of genetic differentiation. A) $F_{ST}$ B) $F'_{ST}$ C) $\Phi_{ST}$ D) $D$

**Figure 6 - Genetic differentiation within and between regions**

Genetic differentiation between populations within regions and between regions. Differentiation was estimated from microsatellite markers using three different methods (for explanation see methods). Monomorphic loci were removed from the analysis when appropriate. Points are estimates of differentiation over all loci, bars represent its 95 % CI, obtained by bootstrapping over loci. For SNP markers $F_{ST}$ was used.

**Figure 7 - Results of computer simulations for single step mutation model**

Different estimators of genetic differentiation are plotted against mutation rate. Different lines represent different migration rates. Migration rates 0.1, 0.01, 0.001, 0.0001 and 0.00001 correspond to different lines as indicated by the legend in panel A. Different estimators are $F_{ST}$, $\Phi_{ST}$, $F'_{ST}$ and $D$ in panels A, B, C and D respectively.

**Figure 8 - Results of computer simulations for mixed mutation model**

The effect of mutation rate on genetic differentiation calculated from $\Phi_{ST}$ using mixed mutation model. In this model, there is a probability of 0.2 that when a mutation occurs the allele will mutate to any state. Different lines represent different migration rates. Migration rates are 0.1, 0.01, 0.001, 0.0001 and 0.00001 correspond to different lines as indicated by the legend.

**Figure 9 - Results of coalescent simulations for different marker types**

The effect of mutation rate on genetic differentiation, calculated from $\Phi_{ST}$. Black lines represent estimates from DNA haplotypes, grey lines are estimates from microsatellite alleles. Different line types represent different migration rates. Migration rates are 0.1, 0.01, 0.001 and 0.0001 correspond to different lines as indicated by the legend.

# Tables

**Table 1 - Indices of genetic diversity for each region**

| | Allelic richness | Gene diversity, (Hs) |
|---|---|---|
| Spain | 2.269 | 0.598 |
| France | 1.720 | 0.392 |
| Norway | 1.245 | 0.144 |
| Asia | 1.383 | 0.228 |
| | | |
| Comparison by permutation test | | |
| Spain vs. France | p = 0.017 | p = 0.068 |
| Spain vs. Norway | p < 0.001 | p < 0.001 |
| Spain vs. Asia | p = 0.002 | p = 0.007 |
| France vs. Norway | p = 0.011 | p = 0.002 |
| France vs. Asia | p = 0.163 | p = 0.167 |
| Norway vs. Asia | p = 0.591 | p = 0.487 |

Indices of genetic diversity for each region based on 20 nuclear microsatellite markers. Significance of differences was assessed by permutation tests, 1000 permutations.

**Table 2 - Correlations between genetic diversity and genetic differentiation**

| | *Hs* | |
|---|---|---|
| Spanish populations | r (95 % CI) | p |
| $F_{ST}$ | -0.862 (-0.944 – -0.678) | <0.001 |
| $F'_{ST}$ | 0.479 (0.046 – 0.760) | 0.033 |
| $\Phi_{ST}$ | -0.294 (-0.652 – 0.170) | 0.208 |
| *D* | 0.765 (0.488 – 0.902) | <0.001 |
| French populations | | |

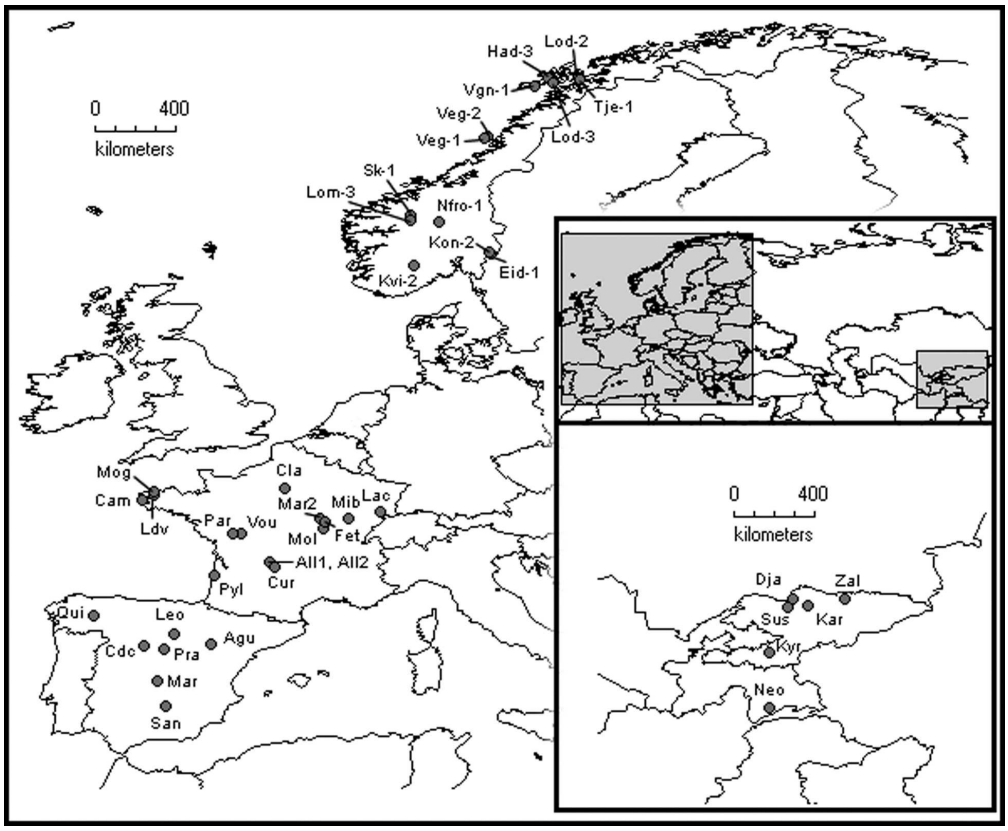
| | | |
|---|---|---|
| $F_{ST}$ | -0.867 (-0.948 – -0.681) | <0.001 |
| $F'_{ST}$ | 0.645 (0.270 – 0.850) | 0.003 |
| $\Phi_{ST}$ | -0.260 (-0.639 – 0.220) | 0.282 |
| $D$ | 0.876 ( 0.700 – 0.952) | <0.001 |
| Norwegian populations | | |
| $F_{ST}$ | -0.916 (-0.968 – -0.791) | <0.001 |
| $F'_{ST}$ | 0.199 (-0.280 – -0.599) | 0.413 |
| $\Phi_{ST}$ | -0.109 (-0.536 – 0.364) | 0.658 |
| $D$ | 0.631 (0.248 – 0.843) | 0.004 |
| Central Asian populations | | |
| $F_{ST}$ | -0.801 (-0.928 – -0.506) | <0.001 |
| $F'_{ST}$ | -0.116 (-0.578 – 0.403) | 0.669 |
| $\Phi_{ST}$ | -0.628 (-0.857 – -0.192) | 0.009 |
| $D$ | 0.565 (-0.013 – 0.860) | 0.055 |

Correlations between genetic diversity in markers and genetic differentiation between populations in different regions. Correlation coefficients are given with 95 % confidence intervals are in parenthesis. *Hs* is subpopulation heterozygosity.

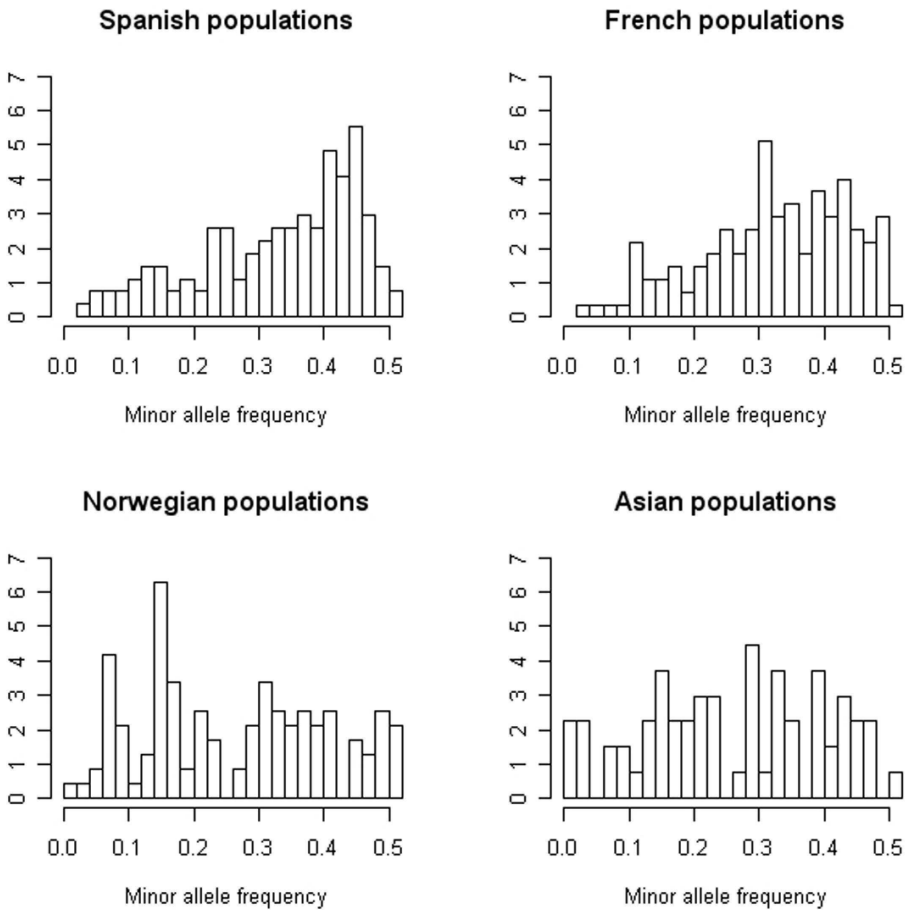
**Table 3 - Correlations between pairwise differentiation and different marker types and analysis methods**

| | SNP $F_{ST}$ | Microsatellite $F_{ST}$ | Microsatellite $F'_{ST}$ |
|---|---|---|---|
| Within Spain | | | |
| Microsatellite $F_{ST}$ | r = 0.412, p = 0.135 | – | – |
| Microsatellite $F'_{ST}$ | r = 0.486, p = 0.092 | r = 0.904, p < 0.001 | – |
| Microsatellite $\Phi_{ST}$ | r = 0.584, p = 0.067 | r = 0.672, p = 0.034 | r = 0.502, p = 0.083 |
| Within France | | | |
| Microsatellite $F_{ST}$ | r = 0.977, p < 0.001 | – | – |
| Microsatellite $F'_{ST}$ | r = 0.899, p< 0.001 | r = 0.929, p < 0.001 | – |
| Microsatellite $\Phi_{ST}$ | r = 0.720, p < 0.001 | r = 0.723, p < 0.001 | r = 0.672, p < 0.001 |
| Within Norway | | | |
| Microsatellite $F_{ST}$ | r = 0.444, p = 0.002 | – | – |
| Microsatellite $F'_{ST}$ | r = 0.590, p < 0.001 | r = 0.910, p < 0.001 | – |
| Microsatellite $\Phi_{ST}$ | r = 0.242, p = 0.061 | r = 0.530, p = 0.03 | r = 0.529, p = 0.01 |
| Within Central Asia | | | |
| Microsatellite $F_{ST}$ | r = 0.768, p = 0.022 | – | – |
| Microsatellite $F'_{ST}$ | r = 0.773, p = 0.013 | r = 0.947, p = 0.004 | – |
| Microsatellite $\Phi_{ST}$ | r = –0.097, p = 0.61 | r = 0.408, p = 0.113 | r = 0.317, p = 0.19 |

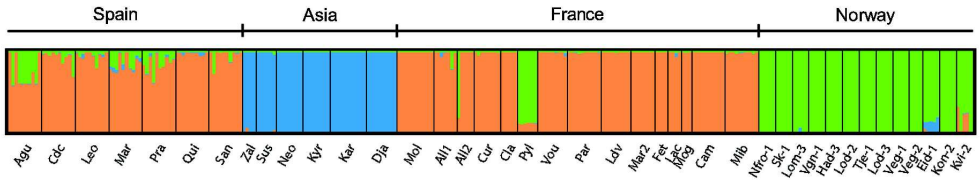
Genetic differentiation was calculated between pairs of populations using different markers and different estimates (see methods). Pearson correlation coefficients were calculated for all pairs. Their significance was tested using mantel tests, 1000 permutations.



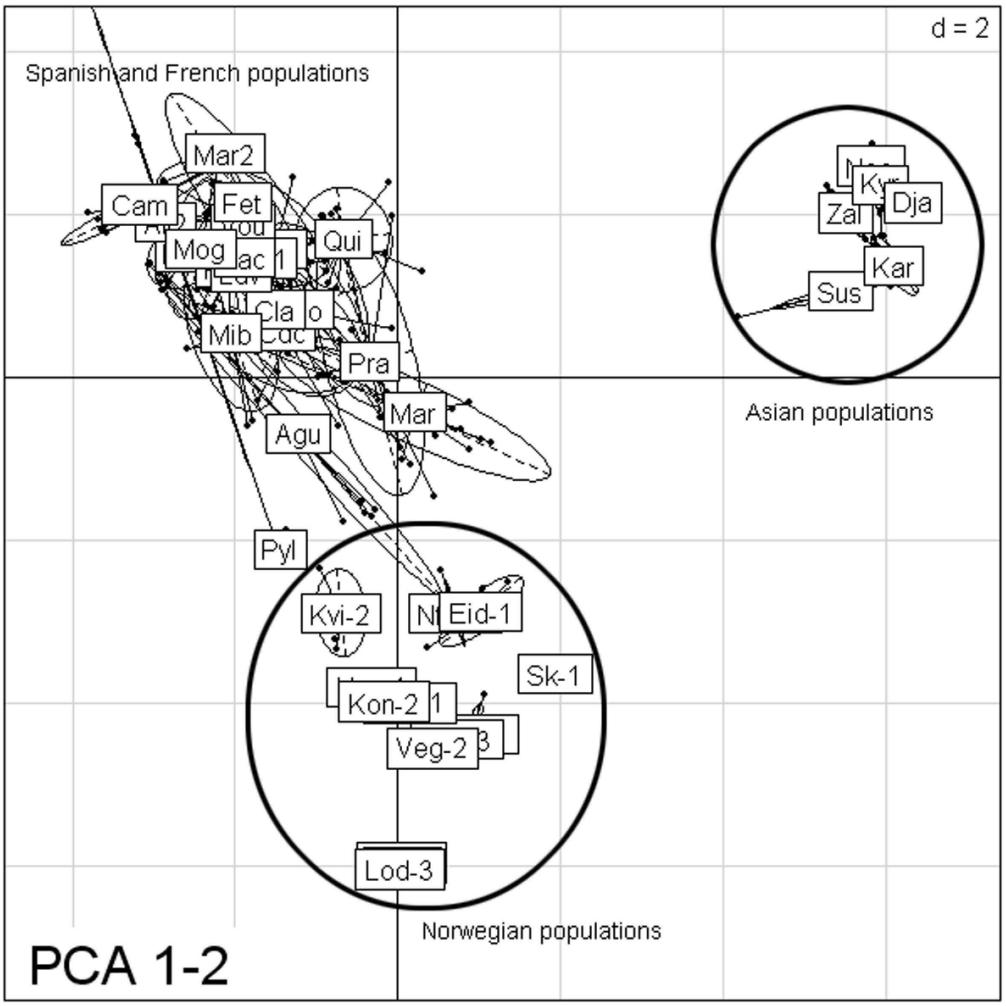

164x135mm (600 x 600 DPI)



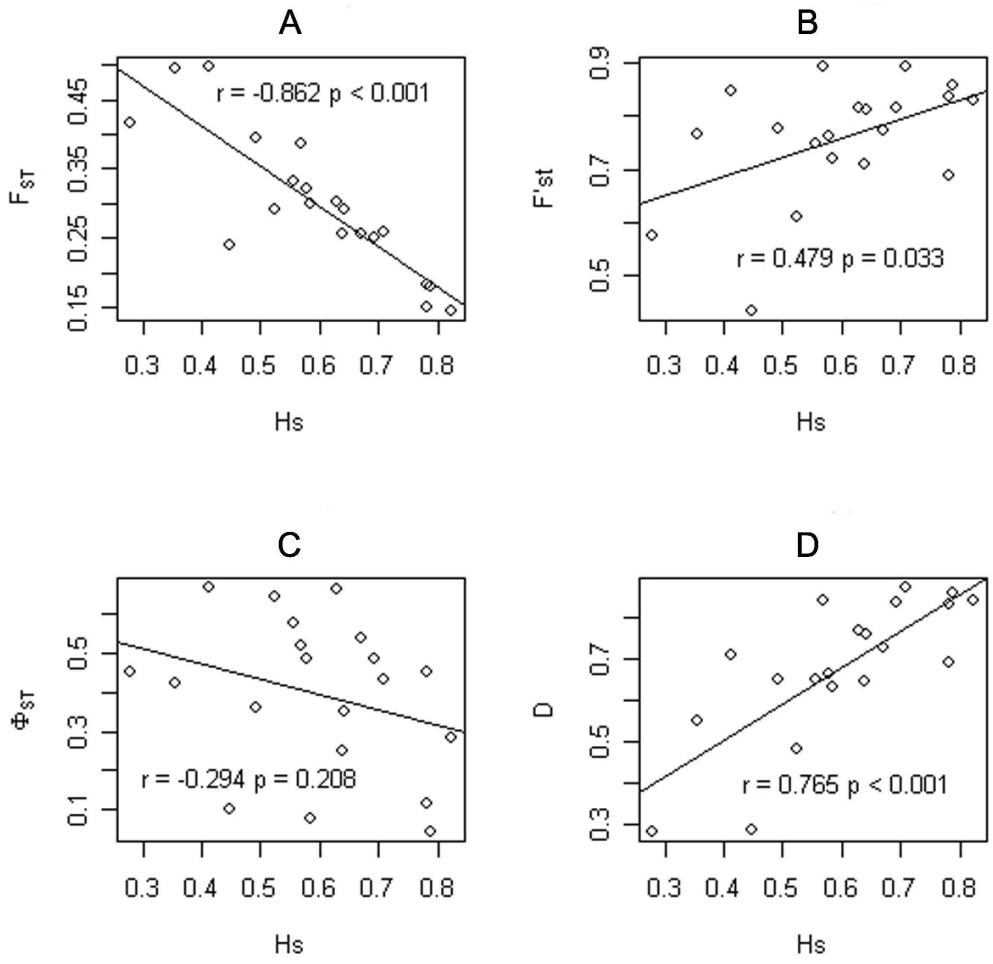

99x96mm (600 x 600 DPI)



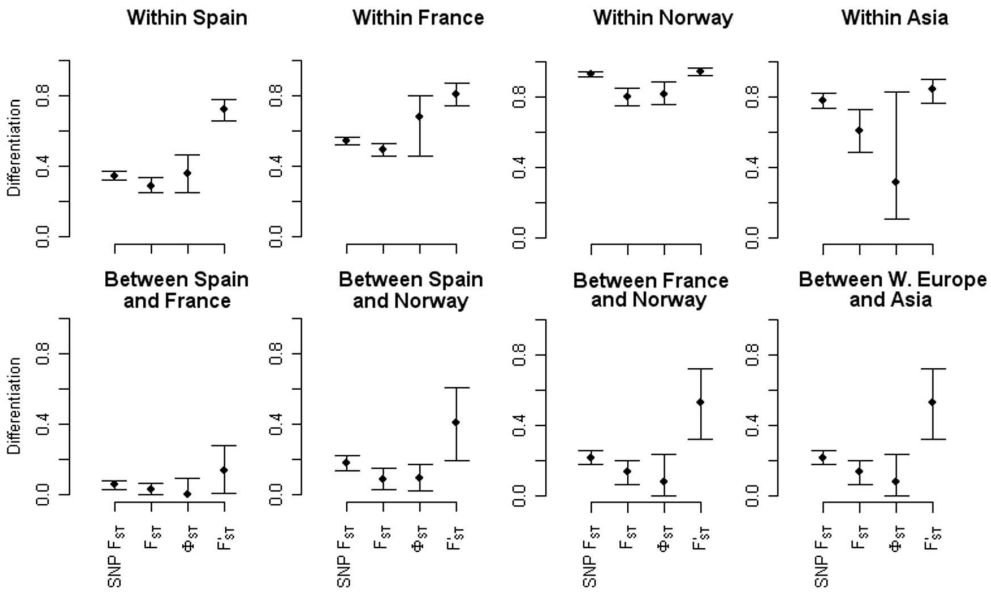

158x29mm (600 x 600 DPI)



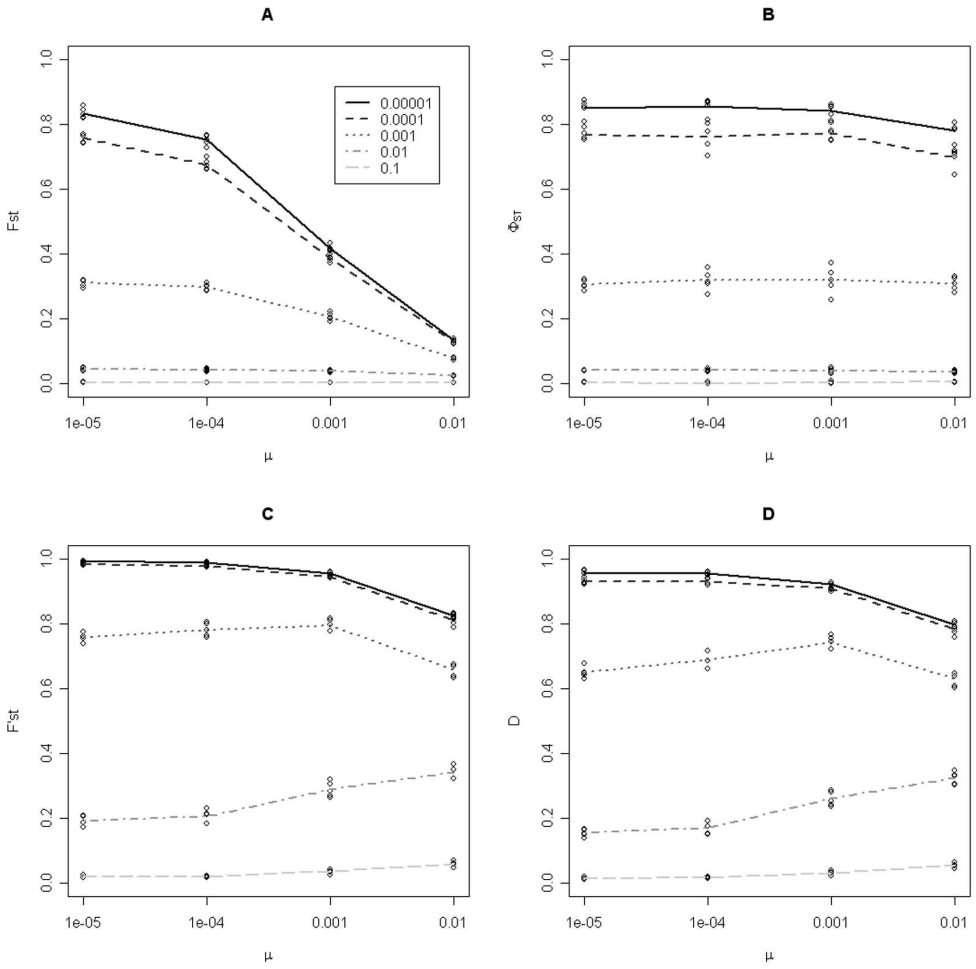

99x99mm (600 x 600 DPI)



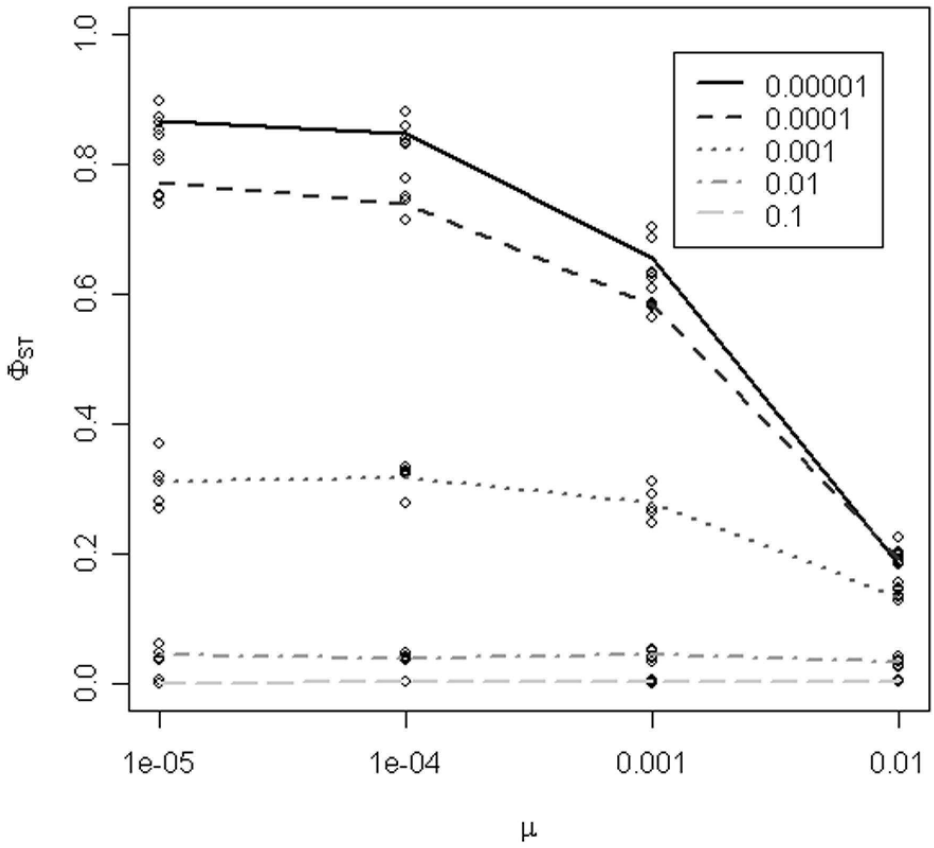

99x98mm (600 x 600 DPI)



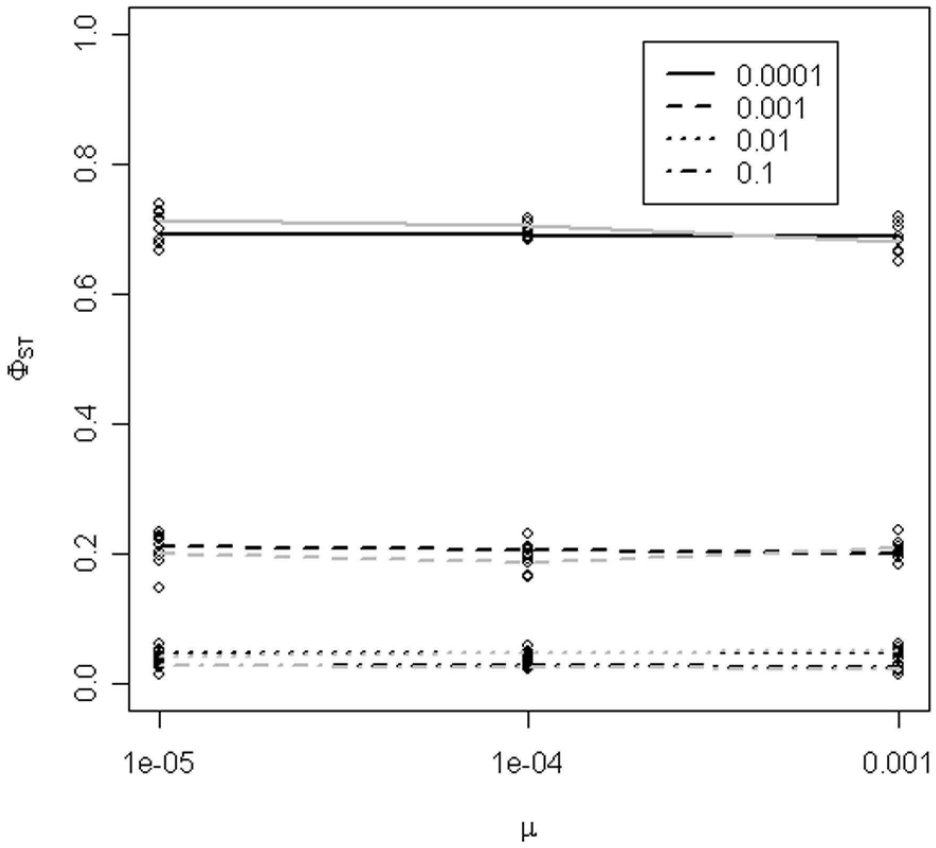

150x89mm (600 x 600 DPI)




150x149mm (600 x 600 DPI)




80x79mm (600 x 600 DPI)




80x79mm (600 x 600 DPI)