COST Action FP1202 – STSM Report

STSM Applicant: Dragos POSTOLACHE, Institute of Plant Genetics (CNR-IGV), Sesto Fiorentino, Florence (Italy).

Host: Professor Martin LASCOUX, Uppsala University, Evolutionary Biology Centre (Sweden).

STSM Topic: *Adaptation of Abies alba populations located in the southern peripheral species distribution range*.

Period STSM: from 10th of September 2013 to 1st of October 2013.
STSM type: Regular (from Italy to Sweden)

Purpose of the STSM

The main goal of Short Term Scientific Mission (STSM) at the Uppsala University, Evolutionary Biology Centre (EBC), was to perform statistical analysis on a dataset of 406 genotyping candidate genes (or SNPs) in Silver fir *(Abies alba)* populations located in the southern peripheral species distribution range.

Adult trees of Silver fir (*Abies alba* Mill.*)* were sampled in 8 populations, along an altitudinal gradient in each population (see Table 1). The sampled Silver fir populations are located in the southern species distribution range (see Fig.1).

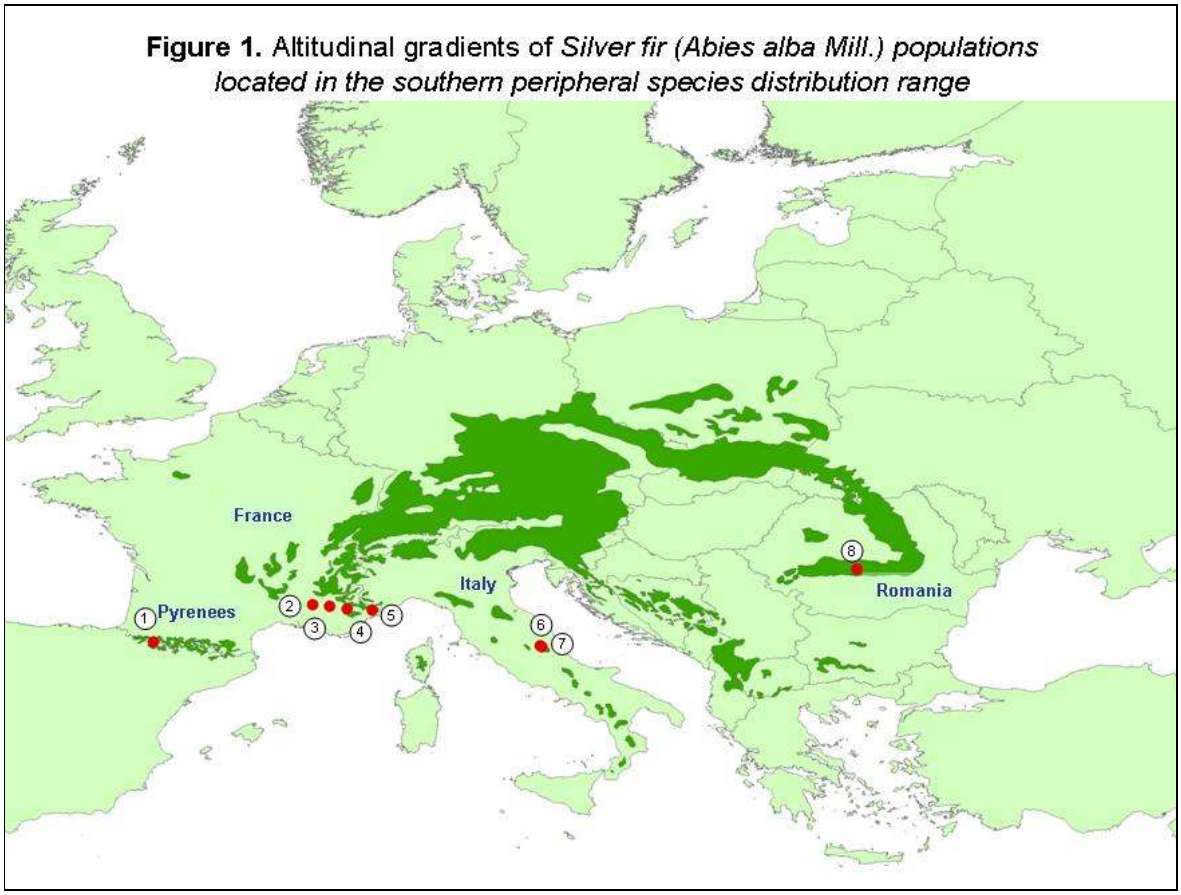


Figure 1. Altitudinal gradients of *Silver fir (Abies alba Mill.) populations located in the southern peripheral species distribution range*

Table 1 Populations used in this study

Nr.	Population	Number of samples per altitudinal gradients		Latitude (decimal degrees) WGS84	Longitude (decimal degrees) WGS84
		bottom	top		
1	Ossau Valley French Pyrenees	70	74	42.855	-0.457778
2	Ventoux (France)	246	290	44.17511	5.2437
3	Lurs (France)	55	56	44.11422	5.83912
4	Issole (France)	49	47	44.0242	6.46244
5	Vesubie (France)	38	40	43.97074	7.36577
6	Valle della Corte (Italy)	46	48	42.70347	13.37576
7	Colle dell'Abete (Italy)	47	46	42.66772	13.42677
8	Arges Fagaras Mountains (Romania)	94	93	45.4411	24.6947
TOTAL number of analyzed samples = 1339		645	694		

Description of the work carried out during the STSM

The genotyped data set of 406 SNPs used in this study come from the KASP genotyping assay of 763 SNPs in 316 candidate genes selected by Roschanski et al. (2013) which comprises SNPs from putative candidate genes for drought tolerance.

The work during the first days of my STSM was focused in the preparation of the input files for subsequent statistical analysis. Hence we tried to group genotyped data according to sampled populations and altitudinal gradients. At the end of processing KASP genotyped data we have prepared the first input file that included 406 SNPs genotyped in 1448 samples grouped in 16 populations. Each population included samples either from the bottom part of the altitudinal gradient and from the top.

The software GenAlEx v6.5 (Peakall & Smouse 2012) was used to analyze the first input file for the presence of monomorphic SNPs and missing data. Samples with genotyped data <50% were removed.

Based on the analysis of the first input file we prepared the second input file that included 270 polymorphic SNPs in 175 genes genotyped in 1339 samples (see Table 1).

We have also prepared a third input file consisting of one SNP per each gene (175 SNPs) genotyped in 1339 samples.

Before staring to test the effects of selection by looking for outliers it is important to estimate the effects of demography.

According to (Li *et al.* 2011) one of the remaining challenges in population genetics is to develop approaches to disentangle selection from demography, which is due to the fact that selection events are often associated with demographic changes.

In order to tackle this challenge the "two-step approach" still remains one of the best strategies. In the first step approaches to estimate the distribution of demographic and genetic parameters are used, followed by those looking for outliers (Li *et al.* 2011).

The first step in our case was to investigate population genetic structure using the software STRUCTURE version 2.3 (Pritchard *et al.* 2000; Falush *et al.* 2003) and the package *adegenet* for the R software (Jombart, T. 2008).

The second step will be to use different statistics to estimate association between allele frequencies in delineated cluster of populations and altitude in order to detect candidate genes potentially involved in local adaptation.

Description of the main results obtained

Genetic structure analysis results

The software STRUCTURE was used as a Bayesian clustering approach with two different models. The admixture model was used, in which the fraction of ancestry from each cluster is estimated for each individual and allowed for correlated allele frequencies, as well as the "locprior" model when population identity is used as *a priori* information for clustering.

Three independent runs for each K value ranging from 1 to 20 were performed after a burn-in period of 10^4 steps followed by 4×10^4 Markov Chain Monte Carlo replicates.

To identify the optimal number of clusters (K) that best explained the data, the rate of change of L(K) (ΔK) between successive K values was calculated following Evanno et al. (2005) using the web application "StructureHarvester" (Earl & von Holdt 2012).

The optimal number of clusters calculated with ΔK method following Evanno et al. (2005) is K=3 for data set with 270 SNPs by using admixture model and K=2 for data set with 175 SNPs genotyped in 1339 samples by using locprior model (see Fig. 2 and Fig.3).

Fig.2 Evanno method for detecting ΔK for 270 SNPs genotyped 1339samples (Admixture model)

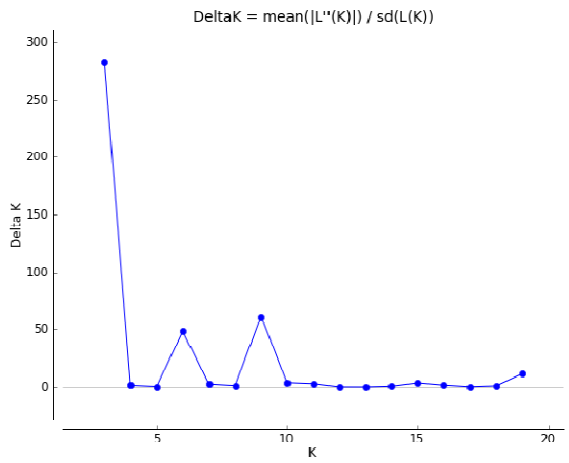
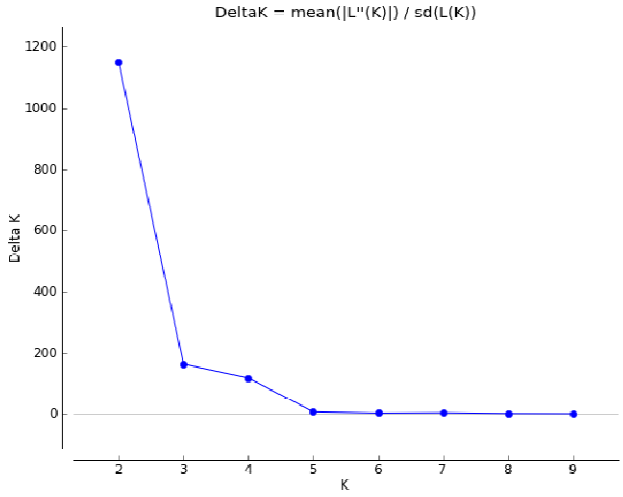
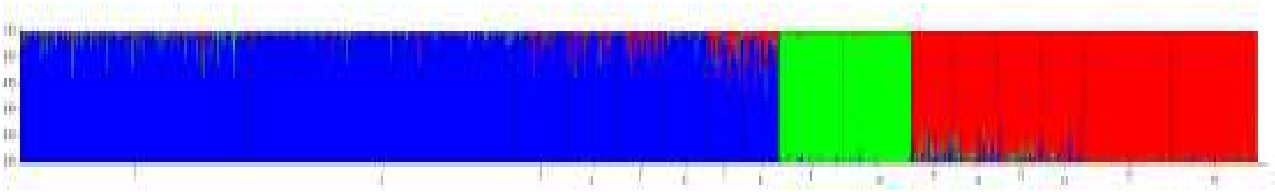


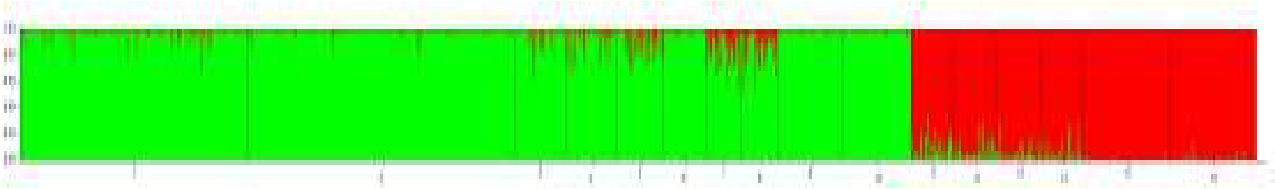
Fig.3 Evanno method for detecting ΔK for 175 SNPs genotyped in 1339 samples (LOCPRIOR model)



K=3



K=2



The number of clusters was different and higher when we apply the mode of posterior probability distribution of the data (Pritchard et al.2000) (see Fig. 4 and Fig.5)

Fig. 4 Pritchard method of mean likelihood L(K) for 270 SNPs genotyped in 1339 samples. (Admixture model)

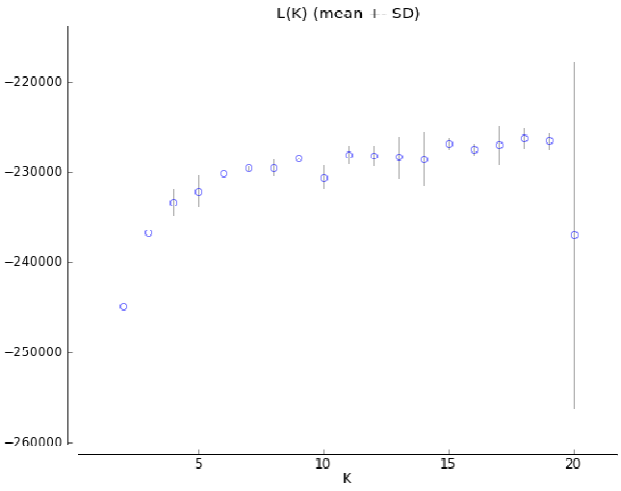
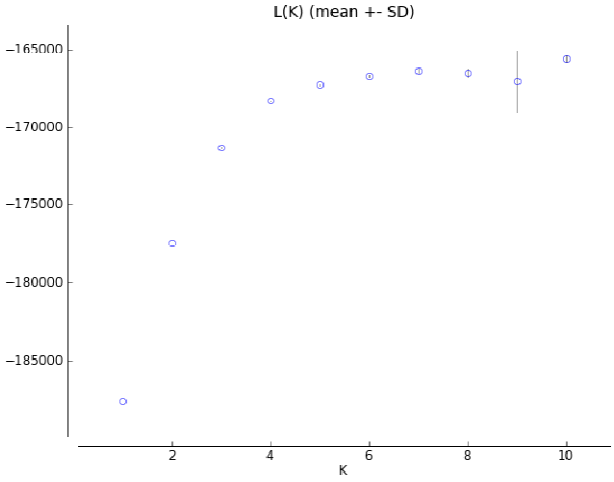


Fig. 5 Pritchard method of mean likelihood L(K) for 175 SNPs genotyped in 1339 samples. (LOCPRIOR model)

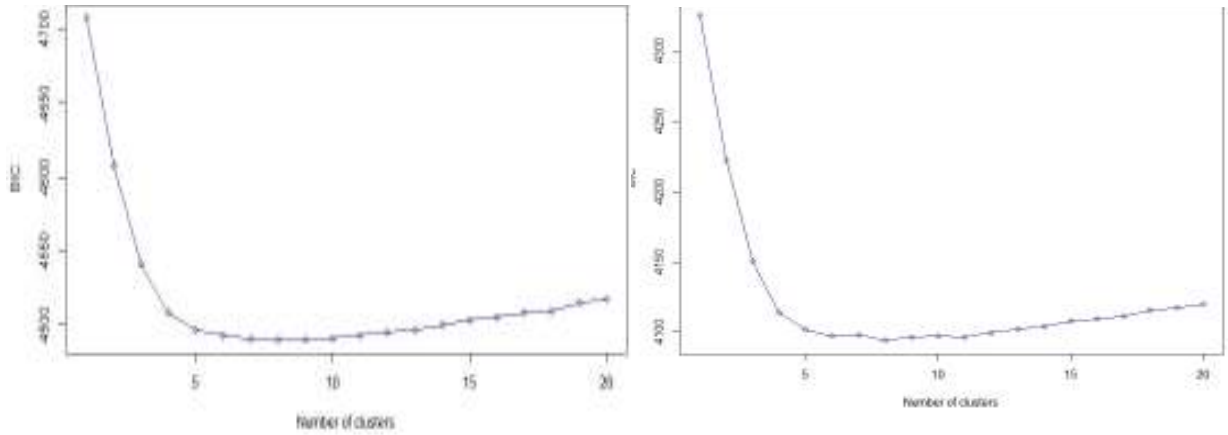


Detecting the number of clusters with the multivariate approach (Jombart, T. 2008) which is based on the lowest BIC (Bayesian Information Criterion) gave similar results for both input files of 270 SNPs and 175 SNPs genotyped in 1339 samples (see Fig. 6 anf Fig.7).

The numbers of clusters detected with the multivariate approach was most plausible and close to reality if we take into consideration the postglacial history of Silver fir revealed by paleobotanic and genetic studies (Liepelt *et al.* 2009).

Fig. 6 Value of BIC versus number of clusters for 270 SNPs genotyped in 1339 samples.

Fig. 7 Value of BIC versus number of clusters for 175 SNPs genotyped in 1339 samples.



Different statistical analyses realized with package *adegenet*, analysing 270 SNPs genotyped in 1339 individuals and taking into consideration the most plausible scenario of grouping our individuals in 4 clusters are reported in Table 2, Fig. 8, Fig.9, Fig.10, Fig.11.

Table 2 Inferred BIC grouping of individuals in clusters for K=4

Nr.	Population	Cluster I	Cluster II	Cluster III	Cluster IV
1	Ossau Valley French Pyrenees_Bottom			70	
2	Ossau Valley French Pyrenees_Top			74	
3	Ventoux (France)_Bottom	246			
4	Ventoux (France)_Top	290			
5	Lurs (France)_Bottom	55			
6	Lurs (France)_Top	56			
7	Issole (France)_Bottom	49			
8	Issole (France)_Top	47			
9	Vesubie (France)_Bottom	38			
10	Vesubie (France)_Top	36	4		
11	Valle della Corte (Italy)_Bottom		46		
12	Valle della Corte (Italy)_Top		48		
13	Colle dell'Abete (Italy)_Bottom		47		
14	Colle dell'Abete (Italy)_Top		46		
15	Arges Fagaras Mountains (Romania)_Bottom				94
16	Arges Fagaras Mountains (Romania)_Top				93

Fig.8 Plot with corresponding populations to inferred BIC clusters for K=4

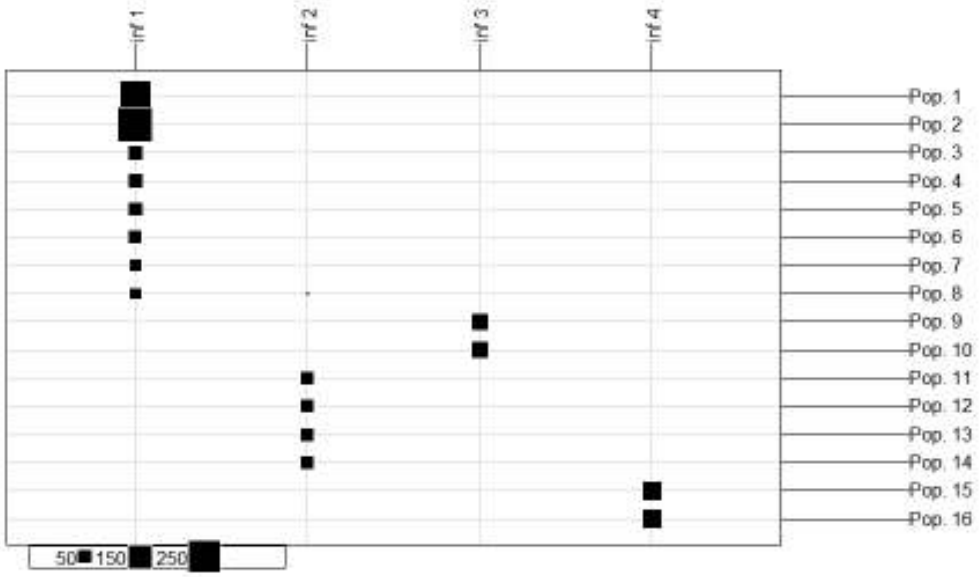


Fig.9 Membership probabilities for each individual to clusters when K=4

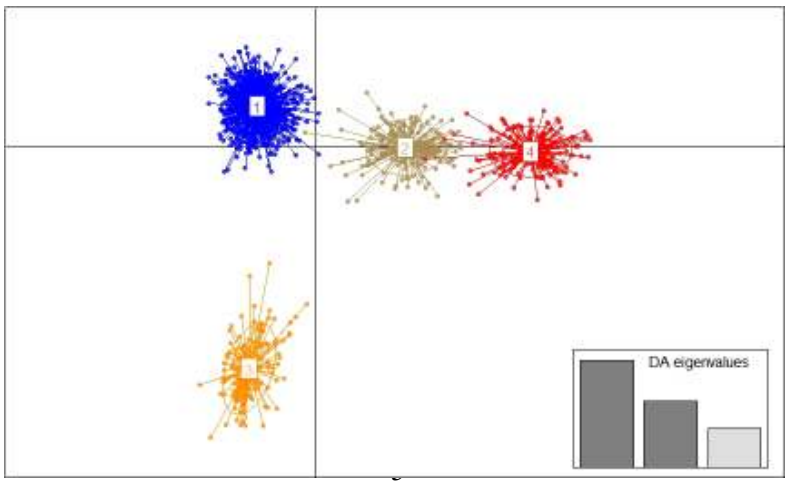


Fig.10 Plot with grouping individuals based on retain discriminant functions when K=4

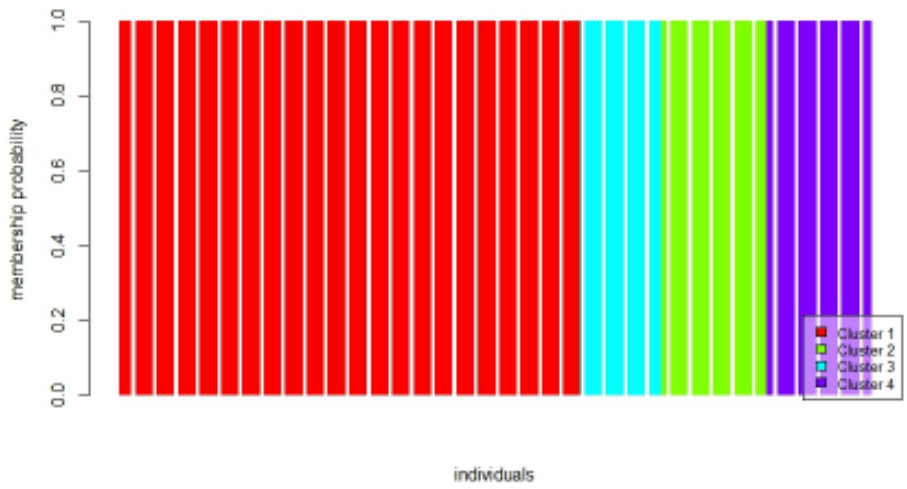
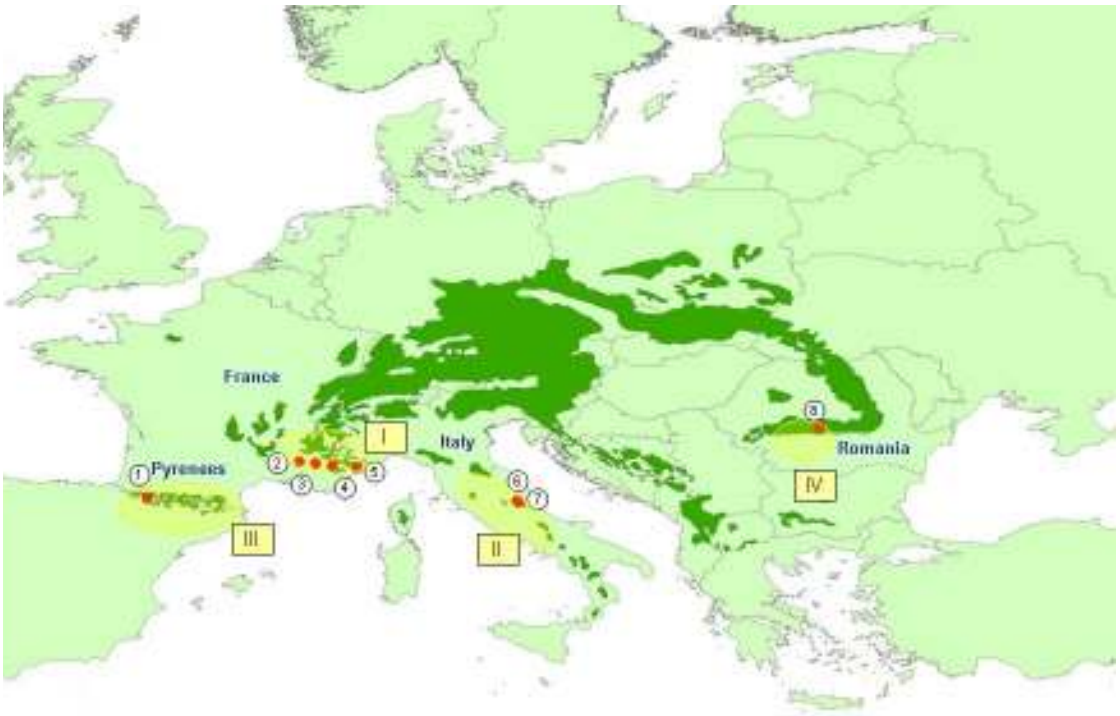


Fig.11 Most likely clusters in Silver fir populations located in the southern peripheral species distribution range.

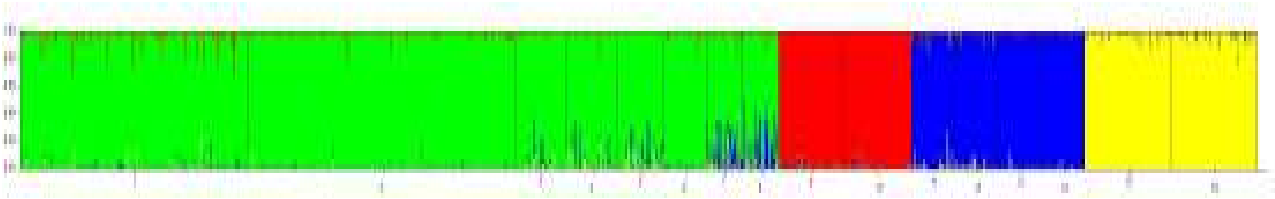


Our findings based on statistical multivariate approach reveal the existence of 4 clusters. Recently published research papers on Silver fir postglacial history recognize three major glacial refugia (Pyrenees, Apennine and Balkan Peninsula). Our clustering of all French populations in the same group seems to suggest that they represent a gene pool originated from a separate glacial refugia most probably located in the south eastern part of France and north western part of Italy.

If we take into consideration the postglacial history of Silver fir then our investigation of population genetic structure using the software STRUCTURE with two models (Admixture model and LOCIPRIOR model) and different number of loci (270SNPs or 175SNPs) show two most plausible clustering of individuals, with K=4 and K=5, as reported in figure 12.

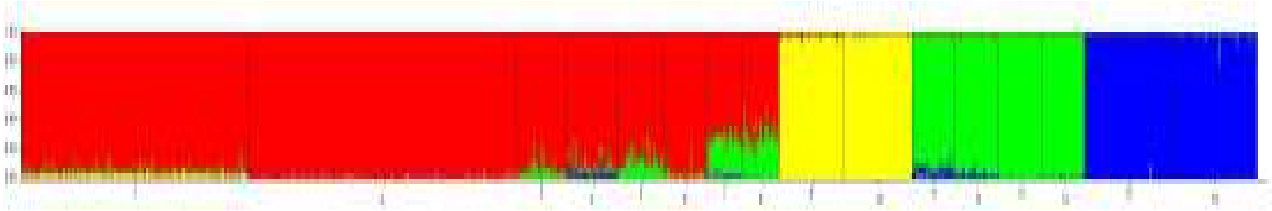
Fig. 12 Clustering analysis conducted in STRUCTURE K=4; K=5

K=4 (Admixture model) (270 SNPs, 1339samples)

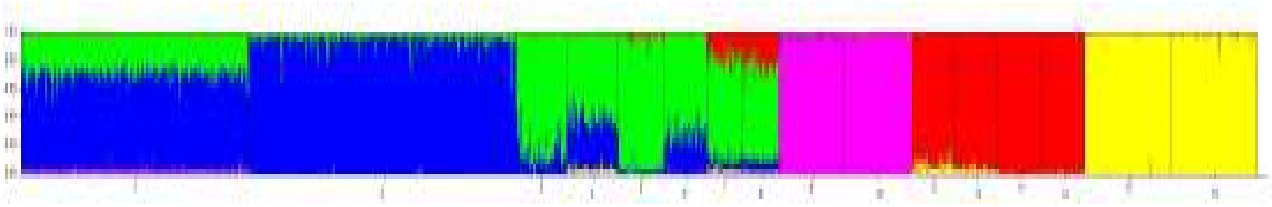


K=5 (Admixture model) (270 SNPs, 1339samples)

K=4 (LOCIPRIOR model) (175 SNPs, 1339samples)



K=5 (LOCIPRIOR model) (175 SNPs, 1339samples)



The clear genetic structuring will be of great help in our next steps of the analysis as we can treat some of the altitudinal gradients as “evolutionary replicates”. For example, SNPs showing high Fst values between high and low altitude samples in different inferred clusters can be seen as putative candidate genes involved in local adaptation.

Our next steps in the data analysis will be to take into consideration two scenarios of genetic clustering (K=4 and K=5) and try to estimate the association between allele frequencies in delineated cluster of populations and environmental gradients in order to detect candidate genes potentially involved in local adaptation.

Future collaboration with host institution (if applicable)

We agreed to continue our collaboration on the data analysis of Silver fir populations in order also to detect candidate genes potentially involved with local adaptation.

Foreseen publications/articles resulting or to result from the STSM (if applicable)

We expect to publish the results obtainedin the STSM in joint articles.

Confirmation by the host institution of the successful execution of the STSM

During his stay at EBC, Dragos has started the statistical analysis of the dataset of 406 genotyping candidate genes (or SNPs) in Silver fir *(Abies alba)* populations located in the southern part of the silver fir range. The data have first been curated thoroughly and standard population genetics analysis was carried out as shown by the results described in the previous sections. The analysis of the association between polymorphism and environmental variable remains to be done but should not be hard to implement now that the data are in very good shape and that the population structure analysis has been carried out.

Altogether this has been a very successful stay and Dragos has quickly been integrated to our team.

Acknowledgements:

The applicant would like to express his gratitude to Prof. Martin Lascoux for accepting to carry out STSM at the Uppsala University, Evolutionary Biology Centre (EBC) and for supervising data analysis.
The applicant would also like to extend his appreciation to Dr.Thomas Källman for providing support and guidance throughout the research.

Bibliography

Chen, J., Källman, T., Ma, X., Gyllenstrand, N., Zaina, G., Morgante, M., ... & Lascoux, M. (2012). Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (Picea abies). *Genetics*, *191*(3), 865-881.

Earl, D. A. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.*Conservation Genetics Resources*, *4*(2), 359-361.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, *14*(8), 2611-2620.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567-1587.

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

Heuertz, M., De Paoli, E., Källman, T., Larsson, H., Jurman, I., Morgante, M., ... & Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. *Genetics*, *174*(4), 2095-2105.

Li, J., Li, H., Jakobsson, M., Li, S. E. N., SjÖDin, P. E. R., & Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go?. *Molecular Ecology*, *21*(1), 28-44.

Liepelt, S., Cheddadi, R., de Beaulieu, J. L., Fady, B., Gömöry, D., Hussendörfer, E., ... & Ziegenhagen, B. (2009). Postglacial range expansion and its genetic imprints in Abies alba (Mill.)—A synthesis from palaeobotanic and genetic data. *Review of Palaeobotany and palynology*,*153*(1), 139-149.

Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update.*Bioinformatics*, *28*(19), 2537-2539.

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

Roschanski, A. M., Fady, B., Ziegenhagen, B., & Liepelt, S. (2013). Annotation and re-sequencing of genes from de novo transcriptome assembly of Abies alba (Pinaceae). *Applications in Plant Sciences*, *1*(1).