# Genetic structure for ten species between St. Hippolyte and the Harvard Forest based on Genotyping-by-sequencing

# Methods

## DNA extraction and library preparation and sequencing

Leaves were immediately placed in presence of silica gel on the field. The DNA was extracted using the plant DNA kit from QIAGEN (Mississauga, Ontario). DNA concentrations were determined on gel and 200 ng of each sample in 10 µL was sent to IBIS (University Laval) for Genotyping-by-sequencing (GBS) library preparation. The library was prepared following Elshire et al. (2011) using the enzyme combination SbfI (CCTGCA/GG) – MspI (C/CGG). This combination was chosen to reduce the number of fragments and thus maximize the sequencing depth, a conservative strategy for obtaining good SNP calling since the samples involved several species with various genome sizes. The libraries (single ends; 100 bp) were sequenced at the Genome Quebec Innovation Centre (Montreal, Canada) on one lane of a HiSeq 2000 Illumina sequencer.

## Bioinformatics

The reads were filtered for quality using Trimmomatic vers. 0.35 (Bolger et al. 2014) with the following sequential steps: adapter trimming (seed mismatch = 3; clip threshold = 6), removing leading and trailing nucleotides with phred scores < 15, removing the remaining nucleotides of a read after the mean nucleotide phred score within a sliding window of 5 nucleotides drops below 15, and finally remove all reads < 50 bp. Sequence quality was inspected before and after filtering using FastQC vers. 0.11.4 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The software Stacks vers. 1.35 (Catchen et al. 2013) was used to generate loci *de novo* and perform SNP calling. Amongst most critical parameters, we set –m to 3, –M to 3, and - max\_locus\_stacks to 4 in ustacks, and –n to 5 in cstacks. These parameters were chosen following Mastretta-Yanes et al. (2015) and considering that the populations compared are geographically isolated. Most other parameters were set to default. The loci assembly and SNP calling was performed independently for the different species.

## Population structure

Only the loci that were present in at least 60% of the individuals in each population were used in the genetic analyses. We estimated locus-based *F*ST statistics (*Phi*ST) with Stacks to quantify the overall population structure. However, to allow a finer description of the genetic structure, we estimated the genetic distances between the individuals. We used the genpofad distance that has the advantage of incorporating polymorphic SNPs in distances and that provides an accurate estimate of true genetic distances (Joly et al. 2015). The genpofad distance was estimated from the full loci sequences obtained from Stackes. We then estimated a neighbour-joining tree (Saitou and Nei 1987) separately for each species to test whether the individuals within each populations were more similar with each other than to individuals from the other population.

# Results

The sequencing resulted in a total of 175,511,015 reads of 100 nucleotides. After cleaning, a total of 172,821,478 reads remained (98.48%). The reads that passed the filter were of good quality (Fig. 1) and no adapters were detectable in the sequences. In general, the reads were relatively well partitioned between the individuals (Fig. 2).

We obtained between 264 (in *Prunus*) and 2188 (in *Vaccinium*) loci per species (Fig. 3). Interestingly, these numbers do not seem to correlate with the genome size of the different species (Table 1). Note that few loci were recovered for all individuals in each species (Fig. 3). Nevertheless, the sequencing depth per loci was relatively good (Fig. 4), with a mean depth above 27 for all species, which shows that no individual as a particularly poor sequencing depth.

## Population structure

Locus based *FST* was similar across species with a mean *FST* that ranged from 0.10 to 0.19 (Fig. 5). The genealogies of individuals support the population structure (Fig. 6). For five species out of ten, the two populations were reciprocally monophyletic (Fig. 6); individuals from a population were closer to individuals from the same population than to individuals from the other population. For the remaining species, only one individual typically grouped with the other population. There are two exceptions. *Populus grandidentata* doesn’t show clear genetic structure among populations and the individuals are much less differentiated than for the other species (Fig. 6). The other exception is *Prunus pensylvanica*, for which two individuals from Massachusetts represent outliers and are more clearly more distant than the remaining individuals (Fig. 6).

# References

Bainard J.D., Husband B.C., Baldwin S.J., Fazekas A.J., Gregory T.R., Newmaster S.G., Kron P. 2011. The effects of rapid desiccation on estimates of plant genome size. Chromosome Res. 19:825–842.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics.:btu170.

Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22:3124–3140.

Costich D.E., Ortiz R., Meagher T.R., Bruederle L.P., Vorsa N. 1993. Determination of ploidy level and nuclear DNA content in blueberry by flow cytometry. Theor. Appl. Genet. 86:1001–1006.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 6:e19379.

Horjales M., Redondo N., Blanco A., Rodríguez M.A. 2003. Cantidades de DNA nuclear en árboles y arbustos. NACC Nova Acta Científica Compostel. Bioloxía.:23–33.

Joly S., Bryant D., Lockhart P.J. 2015. Flexible methods for estimating genetic distances from single nucleotide polymorphisms. Methods Ecol. Evol. 6:938–948.

Mastretta-Yanes A., Arrigo N., Alvarez N., Jorgensen T.H., Piñero D., Emerson B.C. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol. Ecol. Resour. 15:28–41.

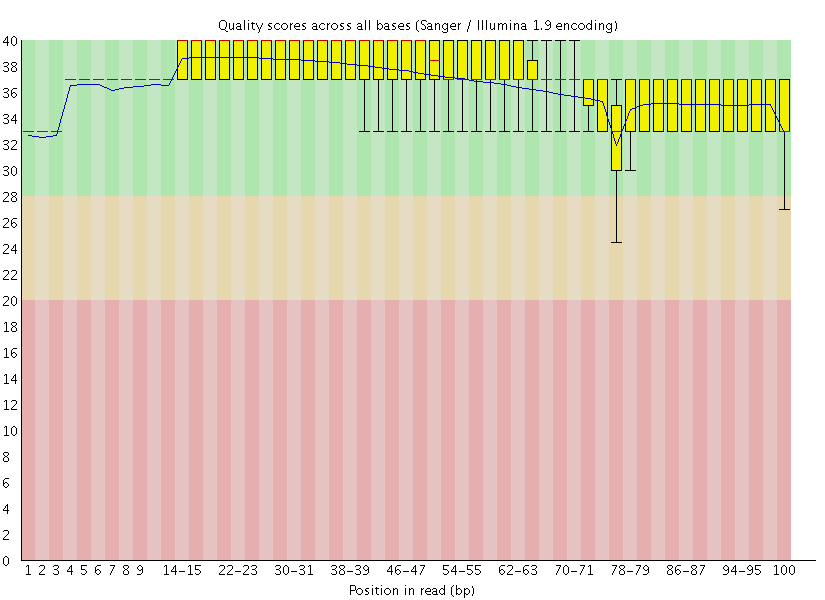
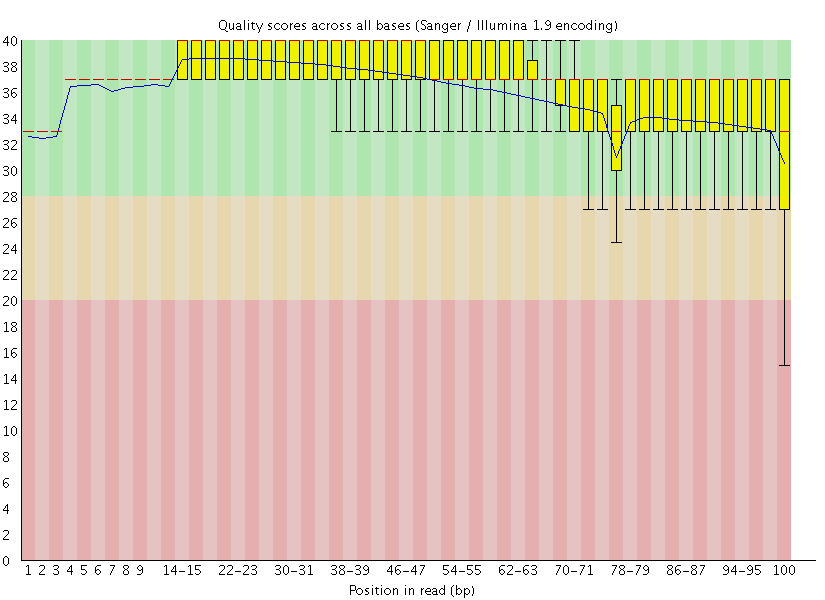
Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Siljak-Yakovlev S., Pustahija F., Šolić E.M., Bogunić F., Muratović E., Bašić N., Catrice O., Brown S.C. 2010. Towards a genome size and chromosome number database of Balkan flora: C-values in 343 taxa with novel values for 242. Adv. Sci. Lett. 3:190–213.

Zonneveld B.J.M., Leitch I.J., Bennett M.D. 2005. First nuclear DNA amounts in more than 300 angiosperms. Ann. Bot.

**Table 1. Approximate haploid genome size in mega base pairs (Mbp) for the species studied**

|  |  |  |  |
| --- | --- | --- | --- |
| Species | Haploid genome size (1c) in Mbp | Method | Reference |
| *Acer pensylvanicum* | ? |  |  |
| *Alnus incana* | 553 Mbp | Flow cytometry | (Siljak-Yakovlev et al. 2010) |
| *Fagus grandifolia* | 528 Mbp | Flow cytometry | (Bainard et al. 2011) |
| *Lonicera nitida (!)* | 929 Mbp | Flow cytometry | (Zonneveld et al. 2005) |
| *Populus (genus mean)* | 484 Mbp | Flow cytometry | Angiosperm DNA C-values database |
| *Prunus sp. (mean of diploids)* | 262 Mbp | Flow cytometry | Angiosperm DNA C-values database |
| *Quercus rubra* | 831 Mbp | Flow cytometry | (Horjales et al. 2003) |
| *Spiraea alba* | ? |  |  |
| *Vaccinium myrtilloides* | 616 Mbp | Flow cytometry | (Costich et al. 1993) |
| *Viburnum lantanoides (genus mean)* | 3907 Mbp | Flow cytometry | Angiosperm DNA C-values database |



**Figure 1**. Per base sequence quality (Phred score) before (left) and after (right) data filtering.



**Figure 2**. Number of reads for each individuals included in the analysis.



**Figure 3.** Number of loci obtained for each species and number of individuals for which the loci was obtained.



**Figure 4.** Number of sequences per locus obtained for each species.



**Figure 5.** Locus-based *FST* distribution of all loci between the two populations for all species. The blue lines indicate the genome mean *FST*.



**Figure 6.** Neighbour-joining phylogenies of individuals of genetic distances for all individuals.