Supplementary Information

Genetic structure of the American ginseng (*Panax quinquefolius* L.) in Eastern Canada using reduced-representation high-throughput sequencing

Simon Joly, Annie Archambault, Stéphanie Pellerin, Andrée Nault

Material and Methods

Sampling

The geographic distance between the sampled populations can be found in Table S1.

Molecular work

Dried leaf tissues were ground with a TissueLyser (Qiagen, Toronto, Ontario) and total DNA was extracted using EZ-10 Spin Column Genomic DNA kits for Plants (BioBasics, Markham, Ontario). We used a reduced-representation sequencing approach to obtain an estimate of population structure across the genome. We used the complexity reduction of polymorphic sequences (CRoPS) strategy (Davey et al. 2011). In short, total DNA was digested with restriction enzymes and adaptors were ligated to cut fragments. A fraction of these were then amplified using selective primers and the pool of amplified fragments was sequenced by high-throughput sequencing.

To minimise costs, we used a population level approach in which all individuals from a population were marked with the same common barcode and were pooled together in equimolar ratio prior to sequencing (Gompert et al. 2010). Although individual genotype information is lost with this approach, population allele frequencies and population structure can be measured. Our approach slightly differs from that of Gompert et al. (2010). First, the ten individuals per population were pooled only before fragment size selection to avoid population-level biases due to PCR amplification. Second, to maximize coverage and sequence length, the fragment size selected was slightly larger, between 300 and 400 bp, and was performed using a Pippin Prep (Sage Science, Beverly, MA, USA).

Protocol

Genomic DNA (180 µg) was digested for 3 hours at 37 °C with two restriction enzymes: a 4bp-cutter (Mse1, T/TAA; 5 Units) and a 6bp-cutter (EcoR1, G/AATTC; 3 Units) with 1X NEB4 buffer (New England Biolabs) and 4 µg of BSA. Two double-stranded adaptors were prepared with the oligonucleotides listed in Table S2. Resuspended EcoRI_adapter1 and EcoRI_adapter2 oligonucleotides were mixed together, heated at 95 °C for 5 minutes, and slowly cooled down to make the double stranded adaptor. The same procedure was applied to MseI_adapter1 and MseI_adapter2 oligonucleotides. *EcoRI* adaptors were diluted to a final concentration of 5 µM, while *MseI* adaptors were diluted to a final concentration of 50 µM. Ligation of the double-stranded adaptors to digested DNA (10 µL) was performed at 16 °C for 3 hours with 1X NEB4 buffer (New England Biolabs), 0.15 µM *EcoRI* adaptor, 1.5 µM *MseI* adaptor, 200 Units of T4 DNA ligase (New England Biolabs), and 1mM ATP in a total volume of 20 µL. Ligated fragments were then diluted 10 fold using 0.1X TE buffer.

Selective amplifications were performed using primers that contain the selective primer, a multiplex identifier (MID) and the library tags necessary for the amplicon sequencing. The forward EcoRI primers thus consisted of the LibL-A sequence, the key sequence and the selective EcoRI primer (+C), whereas the reverse MseI primers thus consisted of the LibL-B sequence, the key sequence and the selective EcoRI primer (+AC) (Table S3). Oligos were purified by HPLC to ensure integrity of the primers.

Amplification of individual plants was achieved using iProof high-fidelity DNA polymerase (BioRad Laboratories) to reduce PCR errors that could negatively affect 454 pyrosequencing results. The reaction mix included 1X HF buffer, 2.5 mM of $MgCl_2$, 200 µM od dNTPs, 300µM of each primer, 16.2 ng of the digested template, and 0.5 U of iProof polymerase; cycling conditions were identical as those of Gompert et al (2010). Amplifications were checked on agarose gel, and combined by population in equimolar ratio. Fragments of size between 300 and 400 bp were selected using a Pippin Prep (Sage Science, Beverly, MA, USA) for each pool and then sequenced in one quarter of a Roche 454 run (Genome Québec Innovation Centre, Montréal, Canada).

Reads assembly

Adapter sequences were removed and poor quality nucleotides were filtered out from the reads by the sequencing facilities. Sequences were assembled in Geneious (Drummond et al. 2014) with the following parameters: min. overlap = 30 bp and min. overlap identity = 97%, word length and index word length of 10, max. ambiguity = 16, reanalyse threshold = 2, and gaps allowed (max. size = 3). To test whether the assembly algorithm affects the results, sequences were also assembled in SeqMan NGen (DNAStar Inc., Madison, WI) with the following parameters: Match Size = 21, Match Spacing = 75, Min. Match Percentage = 85, Match Score = 10, Mismatch Penalty = 15, Gap Penalty = 30, Max. Gap = 15, Min. contig length = 30.

Data filtering

To eliminate biases in subsequent analyses, we discarded contigs of chloroplast, mitochondrial, or ribosomal origin. To identify contigs of chloroplast origin, a similarity search was performed (megablast; threshold E-value = 1e-20) between the contigs consensus sequences and the chloroplast genome of *Panax ginseng* C. A. Mey (NCBI NC_006290). Sequences of mitochondrial origin were identified similarly (blastn; threshold E-value = 1e-20) using the *Arabidopsis thaliana* mitochondrion genome (NCBI NC_001284) as reference. Sequences of ribosomal origin were identified by similarity search (blastn; threshold E-value = 1e-10) on a BLAST database that consisted of American ginseng ribosomal sequences obtained from the NCBI nucleotide database with the search "txid4053[Organism] AND ribosomal". We also identified sequences of fungal or bacterial origin by similarity search (blastn; threshold E-value = 1e-20) on the NCBI nr nucleotide database obtained with the criteria "fungi[organism] AND bacteria[organism]". Finally, a similarity search (blastn; E-value 1e-20) on the TIGR Solanaceae repeat database version 3.2 was performed to identify contigs representing potential transposon or repeat sequences (e.g., centromeres, telomeres). Filtering results are given in Table S4.

An effort was made to discard contigs potentially representing gene families. Because there are numerous chloroplasts but only one nucleus per cell, the number of reads per chloroplast genes should be greater than that of nuclear genomic regions. Contigs with a higher coverage (i.e., more reads) than the mean for chloroplast contigs (43 reads) were thus discarded.

Contigs alignments that satisfied the initial filtering steps were then cleaned to remove sites and reads in the alignments that did not contain sufficient data. An initial look at the contig alignments revealed that some contig assemblies were problematic. Indeed, some contigs consisted of two main "blocks" of sequences joined together by one or two sequences that bridged the gap over the restriction site either because some fragments were not completely digested or because of point mutations in the restriction sites. Moreover, since most sequences contained information for only one of the "blocks", many pairwise distances are undefined. We chose to keep the largest block in number of nucleotides. Blocks were delimited by alignment "gaps" where data was missing for more than 60% of the sequences for more than 3 consecutive nucleotides. For the block retained, sequences that covered less than 75% of the block length and alignment positions with more than 10% missing data (gaps or Ns) were removed. Resulting alignments that contained six or more sequences were used in the subsequent analyses.

Homeologs detection

The next filtering step involved paralog detection. Although the American ginseng is tetraploid, the polyploid event probably predates the origin of *P. quinquefolius* given that it

forms a well supported clade with other polyploid species (Lee and Wen 2004). Therefore, gene homeologs (paralogs that originated from the polyploid event) are expected to predate the species origin. If a given contig includes paralogs sequences, the distribution of pairwise distances of its sequences should be bimodal with smaller distances expected among orthologs than among paralogs (including homeologs). We used population genetic theory to distinguish bimodal distance distributions resulting from paralogs from those due to allelic variation. According to the coalescent, the expectation for the age of the most recent common ancestor is $2N$e in a population, where N_e is the effective population size, and the upper 95% confidence interval for the coalescent time in a population is $4N_e$. Using these expectation, we used the 4x rule (Birky et al. 2010) to determine if the divergence between the paralogs is too large to be explained by intra-specific population divergence. In practice, if the nucleotide diversity (π) of the larger distance mode (an estimate of K = $8N$eμ) is more than 4 times greater than the nucleotide diversity of the within species diversity (an estimation of θ=$2N$eμ), the contig probably contains paralogs. More formally, contigs are considered to contain paralogs when K/theta = $8N$eμ/$2N$eμ > 4. We implemented this filter in R (R core team 2015) with the package 'mclust' (Fraley and Raftery 2006), which allows to fit one or two mode on the distance distributions using model-based clustering (Fraley and Raftery 2002). If the data fits one mode better than two, then the contig was automatically retained. If two modes fit the data better, we then tested if the mean nucleotide diversity of the larger mode is four times greater than that of the smaller mode, in which case the contig was excluded.

Population genomics analyses

To assess the extent of population structure among the American ginseng populations, we used the Bayesian approach of Gompert et al. (2010) implemented in BAMOVA (Gompert and Buerkle 2011). The population model of BAMOVA uses a Bayesian hierarchical model to estimate loci specific and genome wide ϕ statistics (Excoffier et al. 1992). This allows estimating the amount of molecular variance partitioned among and within populations, and provides a measure of population genetic structure. A useful characteristic of the model for the present application is that sequences only need to be assigned to a population and not to a specific individual in the population.

For the analyses, only contigs represented by at least two sequences in each population were retained. Three independent chains of 500 000 generations were run with the parameters "-l 1 -v 0.2 –D 1 –a 0 –w 2000 –c 0.8", which were found to give the best chain mixing. The chain was sampled every 100 generations and the first 200 000 generations were discarded as burnin. Convergence and mixing of the BAMOVA analyses was assessed visually and statistically using the coda package (Plummmer et al. 2006) in R.

We also estimated the non-parametric F_{ST} measure based on sequence similarity. We followed Nordborg et al. (2005) such as

$$F_{ST} = 1 - \left(\frac{1}{k} \sum_{i=1}^{k} \widehat{\pi_{w,i}} \right) \Big/ \widehat{\pi_t},$$

where $\widehat{\pi_t}$ is the average number of pairwise differences per site for all pairs of accessions, and $\widehat{\pi_{w,i}}$ is the average number of pairwise differences per site for all pairs within population *i*. By doing this, we assumed that each sequence represents an independent draw from the population. F_{ST} values were also estimated between all pairs of populations and used to reconstruct a population tree by neighbour-joining (Saitou and Nei 1987) in R using the 'ape' package (Paradis 2012). We used the F_{ST} values for the tree construction because BAMOVA analyses sometimes had convergence issues when ran on pairs of populations, perhaps because of small sample sizes.

Results

The Figure S1 shows the untrimmed sequence length obtained from the 454 sequencing. The results from the filtering of the reads are presented in table S4.

References

Birky, C.W., Adams, J., Gemmel, M., and Perry, J. 2010. Using population genetic theory and DNA sequences for species detection and identification in asexual organisms. PLoS ONE **5**(5): e10609. doi:10.1371/journal.pone.0010609.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. **12**(7): 499–510. doi:10.1038/nrg3012.

Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., and Wilson, A. 2014. Geneious. Biomatters, Auckland, New Zealand. Available from http://www.geneious.com.

Excoffier, L., Smouse, P., and Quattro, J. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**: 479–491.

Fraley, C., and Raftery, A.E. 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**(458): 611–631. doi:10.1198/016214502760047131.

Fraley, C., and Raftery, A.E. 2006. MCLUST version 3: an R package for normal mixture modeling and model-based clustering.

Gompert, Z., and Buerkle, C.A. 2011. A hierarchical Bayesian model for next-generation population genomics. Genetics **187**(3): 903 –917. doi:10.1534/genetics.110.124693.

Gompert, Z., Forister, M.L., Fordyce, J.A., Nice, C.C., Williamson, R.J., and Buerkle, C.A. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. Mol. Ecol. **19**(12): 2455–2473. doi:10.1111/j.1365-294X.2010.04666.x.

Lee, C., and Wen, J. 2004. Phylogeny of *Panax* using chloroplast trnC–trnD intergenic region and the utility of trnC–trnD in interspecific studies of plants. Mol. Phylogenet. Evol. **31**(3): 894–903. doi:10.1016/j.ympev.2003.10.009.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobson, 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., and Bergelson, J. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. **3**(7): e196. doi:10.1371/journal.pbio.0030196.

Paradis, E. 2012. Analysis of phylogenetics and evolution with R. *In* 2nd edition. Springer, New York, USA.

Plummmer, M., Best, N., Cowles, K., and Vines, K. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News **6**(1): 7–11.

R core team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

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

Table S1. Pairwise geographic distances between populations in kilometres.

	ON	CDPNQ 3684	CDPNQ 3715	CDPNQ 3690	CDPNQ 3677
CDPNQ 3684	200				
CDPNQ 3715	208	19			
CDPNQ 3690	131	70	81		
CDPNQ 3677	198	26	14	74	
CDPNQ 18468	248	77	58	132	59

Table S2 Oligonucleotides for preparation of double stranded adaptors.

Oligo name	**Modification**	**Sequence, 5' to 3'**
EcoRI_adapter1		CTCGTAGACTGCGTACC
EcoRI_adapter2	5' phosphorylated	AATTGGTACGCAGTCTAC
MseI adapter1	5' phosphorylated	TACTCAGGACTCAT
MseI adapter2		GACGATGAGTCCTGAG

Table S3 Sequences of the selective primers.

Oligo name	**Sequence, 5' to 3'**
LibL_A_MID1_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGACGAGTGCGTGACTGCGTACCAATTC
LibL_A_MID3_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGAGACGCACTCGACTGCGTACCAATTC
LibL_A_MID4_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGAGCACTGTAGGACTGCGTACCAATTC
LibL_A_MID5_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGATCAGACACGGACTGCGTACCAATTC
LibL_A_MID6_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGATATCGCGAGGACTGCGTACCAATTC
LibL_A_MID7_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGCGTGTCTCTAGACTGCGTACCAATTC
LibL_A_MID2_EcoRI_plus1	CCATCTCATCCCTGCGTGTCTCCGACTCAGACGCTCGACAGACTGCGTACCAATTC
LibL_B_MseI_plus2	CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGATGAGTCCTGAGTAAC

Table S4 Number of contigs removed by each filtering step.

Filter	Discarded contigs	Retained contigs
All contigs		13 235
Filter for chloroplast sequences	128	13 107
Filter for mitochondrion sequences	25	13 082
Filter for ribosomal sequences	120	12 962
Filter for transposons	91	12 871
Filter for Bacteria and Fungi	0	12 871
Filter for gene families	258	12 613
Filter for contigs with 6 sequences or more (after contig cleaning)	11 506	1 107
Filter for paralogs	177	751

Figures

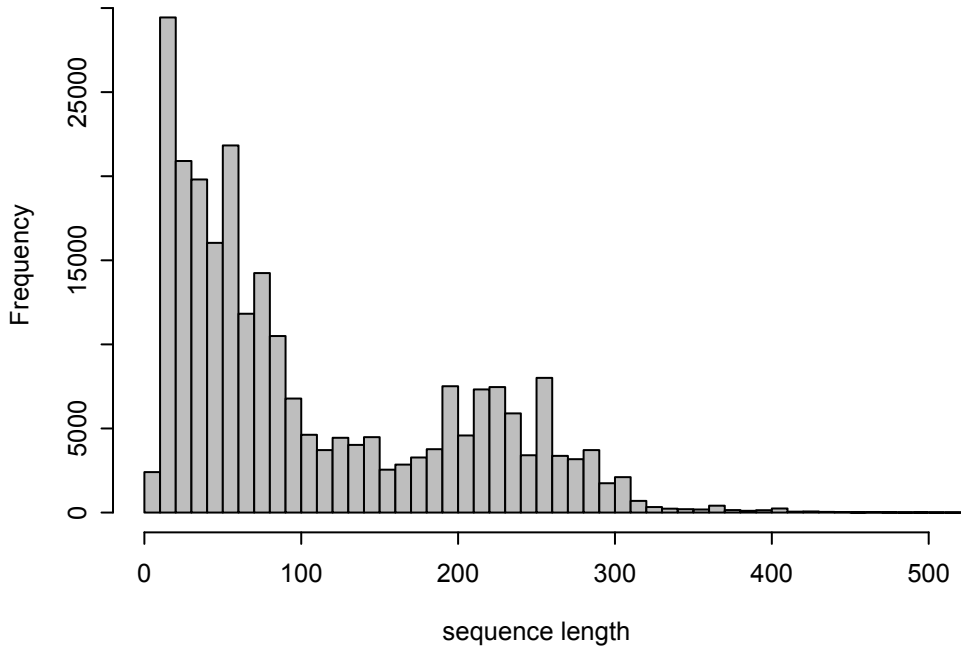

Fig S1 Distribution of untrimmed sequence lengths.