TECHNICAL NOTE

# Isolation, characterization and genotyping of single nucleotide polymorphisms in the non-model tree species *Frangula alnus* (Rhamnaceae)

**Hanne De Kort · Katrien Vandepitte · Joachim Mergeay · Olivier Honnay**

**Abstract** The genomic content of the non-model plant *Frangula alnus* remains largely unexplored. However, because of its occurrence over a wide latitudinal range and its invasive nature in the US and Canada, *F. alnus* may serve as an excellent study species in tree conservation genomic research. We used pooled paired-end sequencing of Restriction-site Associated DNA to discover SNPs in the insect pollinated shrub *F. alnus*. We identified 7,383 SNPs from which the corresponding contig sequences were annotated to the available plant genomes. An enrichment analysis showed an excess of putatively adaptive and plastic gene ontology terms.

The alder buckthorn (*Frangula alnus*), an insect pollinated small tree, has a wide geographical distribution, covering large parts of Eurasia (Hampe et al. 2003). The exploration of the *F. alnus* genome may therefore yield interesting insights into its adaptive potential in the face of global change (De Kort et al. 2012). Furthermore, the species is invasive in Northern America, where it has probably been introduced before the nineteenth century (Wyman 1971). After numerous introductions into private and public gardens, the species became widespread and naturalized from the early twentieth century and has started to invade natural ecosystems from the mid-twentieth century (Lee and Thompson 2012).

Except for the 16 polymorphic nuclear microsatellite markers that have been developed for the species (Rigueiro et al. 2009), the *Frangula alnus* genome remains unexplored. We applied a next-generation sequencing approach to increase genomic resources available for population genomics, functional genomics and conservation genomics. More specifically, Paired-End Restriction-site Associated DNA sequencing (RAD-PE) was performed on a HiSeq2000 (Illumina) at Floragenex Inc. (USA) as described by Etter et al. (2011). To reduce the cost of RAD library preparation and sequencing, we screened a pooled set of eight individuals, which were sampled across the species' latitudinal distribution range to reduce ascertainment bias. The samples originated from Tuscany (Italy), Picardy (France), Wallonia and Flanders (Belgium), and Medelpad (Sweden). Velvet (v1.0.18) was used to assemble consensus *F. alnus* contigs from paired-end data. Alignments were formatted into sorted sequence pileup files, using the SAMtools package (v0.1.14). A custom Floragenex sequence utility parsed the pileups and cataloged putative SNP variants into a variant call format (VCF 4.1) file. To minimize false positives, biallelic SNPs flanked by minimal 50 bp free of polymorphism at either side were called at a minimum sequencing depth of $20\times$ with a minimal minor allele frequency threshold of 0.05. The RAD-PE procedure generated 33.5 million reads equivalent to 19.1 Mbp of sequence, covering approximately 4 % of the species' genome size (Horjales

H. De Kort (✉) · K. Vandepitte · O. Honnay
Plant Conservation and Population Biology, Biology Department, University of Leuven, Kasteelpark Arenberg 31, 3001 Heverlee, Belgium
e-mail: hanne.dekort@bio.kuleuven.be

J. Mergeay
Research Institute for Nature and Forest, Gaverstraat 4, 9500 Geraardsbergen, Belgium

**Table 1** Enrichment results of the *Arabidopsis* orthologs of the 306 annotated SNP sequences against the TAIR database

| GO Term | a/p | P value | Sample frequency | Background frequency |
|---|---|---|---|---|
| GO:0050896 response to stimulus | a | 2.08E−05 | 98/265 (37.0 %) | 6,604/30,328 (21.8 %) |
| GO:0051716 cellular response to stimulus | a | 1.10E−04 | 53/265 (20.0 %) | 2,800/30,328 (9.2 %) |
| GO:0006950 response to stress | a | 1.70E−04 | 67/265 (25.3 %) | 4,013/30,328 (13.2 %) |
| GO:0065007 biological regulation | p | 1.01E−03 | 80/265 (30.2 %) | 5,397/30,328 (17.8 %) |
| GO:0050789 regulation of biological process | p | 1.62E−03 | 74/265 (27.9 %) | 4,901/30,328 (16.2 %) |
| GO:0042221 response to chemical stimulus | a | 5.15E−03 | 60/265 (22.6 %) | 3,778/30,328 (12.5 %) |
| GO:0006952 defense response | a | 7.40E−03 | 34/265 (12.8 %) | 1,665/30,328 (5.5 %) |
| GO:1901700 response to oxygen-containing compound | a | 7.61E−03 | 44/265 (16.6 %) | 2,455/30,328 (8.1 %) |

Only significant putative adaptive (a) and plastic (p) GO terms are shown ($p < 0.01$)

**Table 2** Properties of the 25 *Frangula alnus* populations sampled throughout the species native distribution range

| Populations | Regions | Latitude | Longitude | N | PLP | $H_E$ | $F_{IS}$ |
|---|---|---|---|---|---|---|---|
| De Pinte | Zandstreek (Belgium) | 50.98565 | 3.649552 | 30 | 0.84 | 0.32 | −0.038 |
| Evergem | Zandstreek (Belgium) | 51.16598 | 3.681092 | 36 | 0.82 | 0.30 | −0.015 |
| Moerbeke | Zandstreek (Belgium) | 51.14555 | 3.918045 | 27 | 0.84 | 0.31 | −0.024 |
| Sint-Katelijne-Waver | BDW (Belgium) | 51.07125 | 4.543055 | 24 | 0.85 | 0.31 | 0.010 |
| Herent | BDW (Belgium) | 50.91215 | 4.619936 | 38 | 0.85 | 0.30 | −0.016 |
| Zemst | BDW (Belgium) | 50.99224 | 4.398996 | 32 | 0.84 | 0.30 | −0.010 |
| Heist op den Berg | Kempen (Belgium) | 51.08424 | 4.793124 | 33 | 0.86 | 0.32 | −0.024 |
| Beringen | Kempen (Belgium) | 51.08847 | 5.305648 | 33 | 0.81 | 0.30 | −0.026 |
| Zoersel | Kempen (Belgium) | 51.25348 | 4.683878 | 37 | 0.85 | 0.31 | −0.015 |
| Holsbeek | BDO (Belgium) | 50.93981 | 4.830925 | 23 | 0.85 | 0.31 | −0.046 |
| Tongeren | BDO (Belgium) | 50.80119 | 5.480932 | 34 | 0.83 | 0.29 | −0.000 |
| Kortenaken | BDO (Belgium) | 50.86372 | 5.01641 | 29 | 0.82 | 0.31 | −0.032 |
| Thiérache | Ardennen (Belgium) | 50.00368 | 4.653925 | 12 | 0.80 | 0.29 | −0.007 |
| Regnièssart | Ardennen (Belgium) | 50.01361 | 4.603547 | 28 | 0.83 | 0.31 | −0.011 |
| Brûly-de-Pesche | Ardennen (Belgium) | 50.01711 | 4.469786 | 36 | 0.86 | 0.31 | −0.029 |
| Storåsmyran | Medelpad (Sweden) | 62.33861 | 16.87907 | 22 | 0.75 | 0.28 | −0.014 |
| Aborrtjärn | Medelpad (Sweden) | 62.42415 | 17.01771 | 28 | 0.79 | 0.27 | −0.038 |
| Ottsjön | Medelpad (Sweden) | 62.4421 | 17.23451 | 24 | 0.79 | 0.29 | −0.075 |
| Casina Rossa | Tuscany (Italy) | 43.16505 | 11.22042 | 16 | 0.69 | 0.23 | 0.035 |
| Famelunga | Tuscany (Italy) | 43.12181 | 11.17654 | 13 | 0.73 | 0.24 | 0.005 |
| Luriano | Tuscany (Italy) | 43.114894 | 11.157618 | 18 | 0.60 | 0.20 | −0.107 |
| Fosso bolza | Tuscany (Italy) | 43.123665 | 11.221991 | 15 | 0.65 | 0.23 | −0.004 |
| Boves | Picardy (France) | 49.85694 | 2.378333 | 7 | 0.75 | 0.28 | 0.088 |
| Saint-Michel | Picardy (France) | 49.9475 | 4.212778 | 7 | 0.75 | 0.30 | 0.036 |
| Hannapes | Picardy (France) | 49.98417 | 3.596389 | 17 | 0.83 | 0.31 | −0.019 |

et al. 2003). Reads were assembled into 56,404 contigs with a mean length of 323 ($N_{50} = 374$ bp) and with a high average sequencing depth of 39.7×. A total of 7,383 high-quality Illumina compatible SNPs, or ∼387 SNPs per Mbp of sequence, were called at a local sequencing depth of 18–1,579 (Online Resource 1, 2).

Blast2go was used to uncover the putative functions of the discovered SNP contigs. We applied a statistical significance threshold E value of $10^{-15}$ to minimize the number of false positive BLASTx hits. A total of 431 hits were retrieved with a mean E value of $10^{-32}$, from which the majority (306 SNP contigs) was mapped in *A. thaliana*. These hits were subsequently mapped to their corresponding gene annotations using the gene ontology database and several additional database files (Götz et al. 2008). The annotations were then mapped to higher level

categories (plant GO Slim), based on *A. thaliana* orthologs (AGI codes) developed by The Arabidopsis Information Resourse (TAIR). Ecologically relevant GO terms, such as "reproduction" and "response to abiotic stimulus" could be attributed to 130 SNP sequences (Online Resource 3,4). Moreover, 150 SNPs were found in sequences involved in gene expression such as "regulation of translation" and "transcription" (Online Resource 3,4), which may reflect relevant genetic variation underlying adaptive plasticity (Van Kleunen and Fischer 2005; Bell and Robinson 2012). To test whether the GO terms potentially associated with local adaptation or phenotypic plasticity were significantly enriched in our SNP dataset, we performed an enrichment analysis using AmiGO, which compares the proportion of GOs of the *Arabidopsis* orthologs of our annotation hits (405) to the complete TAIR database (Carbon et al. 2009). The analysis revealed significant enrichment of our annotated SNP dataset for several putatively adaptive and plastic GO terms (Table 1).

Finally, 619 *F. alnus* samples were genotyped for 250 putative SNPs at KBioscience (UK) using KASPar technology. These 250 SNPs consisted of 70 putatively neutral SNPs (randomly chosen) and 180 putatively adaptive and plastic SNPs (given ecologically relevant GO terms with lowest annotation E-values). The samples were collected from 25 populations distributed throughout the species distribution range (Table 2). Data were automatically processed using inhouse KBioscience Kraken software. Genalex 6.5 (Peakall and Smouse 2012) was used to estimate the number of polymorphic loci and the expected heterozygosity in each population. Intra-population inbreeding coefficients ($F_{IS}$) were calculated with ARLEQUIN 3.5, using 10,000 permutations (Excoffier and Lischer 2010). A total of 205 SNPs (82 %) produced scorable patterns of which 186 (90.7 %) were polymorphic (Online Resource 5). Only a small fraction of the data was missing (5.8 %), and 85 duplicated samples produced almost identical calls (98.9 % identity), indicating robust genotyping. Expected heterozygosity per locus varied from 0.20 to 0.32 among locations. $F_{IS}$ values ranged from −0.107 to 0.088 with no populations showing a significant excess of homozygotes (Table 2).

In conclusion, we were able to find 7,383 candidate SNPs out of which over 200 were putatively associated with adaptive differentiation and plasticity in the non-model species *F. alnus*. The cost-effective pooled RAD procedure therefore serves as an ideal exploration tool to evaluate adaptive capacity of vulnerable and/or invasive populations of tree species without closely related reference.

## Data Archiving Statement

We have provided GenBank accession numbers for our contig sequences. The records will be released to the public database once they are processed by GenBank (http://www.ncbi.nlm.nih.gov/genbank/).

As soon as the sequences are released, they can be found in Genbank using one of the following project details:

| | |
|---|---|
| Submission ID | : SUB200632 |
| BioProject ID | : PRJNA203129 |
| Organism name/label | : Frangula alnus |

## References

Bell AM, Robinson GE (2012) Behavior and the dynamic genome. Science 332(6034):1161–1162

Carbon S, Ireland A, Mungall CJ et al (2009) AmiGo: online access to ontology and annotation data. Bioinformatics 25(2):288–289

De Kort H, Vandepitte K, Honnay O (2012) A meta-analysis of the effects of plant traits and geographical scale on the magnitude of adaptive differentiation as measured by the difference between $Q_{ST}$ and $F_{ST}$. J Evolut Ecol. doi:10.1007/s10682-012-9624-9

Etter PD, Preston JL, Bassham S et al (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS One 6(4):e18561

Excoffier L, Lischer HEL (2010) Arlequin suite ver. 3.5: a new series of programs to perform population genetics analysis under Linux and Window. Mol Ecol Res 10:564–567

Götz S, García-Gómez JM, Terol J et al (2008) High-throughput functional annotation and data mining with the Blast2go suite. Nucl Acids Res 36(10):3420–3435

Hampe A, Arroyo J, Jordano P, Petit RJ (2003) Rangewide phylogeography of a bird-dispersed Eurasian shrub: contrasting Mediterranean and temperate glacial refugia. Mol Ecol 12(12):3415–3426

Horjales M, Redondo N, Blanco A et al (2003) Cantidades de DNA nuclear en árboles y arbustos. Nova Acta Científica Compostelana (Bioloxía) 13:23–33

Lee TD, Thompson JA (2012) Effects of logging history on invasion of eastern white pine forests by exotic glossy buckthorn (*Frangula alnus* P. Mill.). For Ecol Manag 265:201–210

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28:2537–2539

Rigueiro C, Arroyo JM, Rodriguez R et al (2009) Isolation and characterization of 16 polymorphic microsatellite loci for *Frangula alnus* (Rhamnaceae). Mol Ecol Res 9(3):986–989

Van Kleunen M, Fischer M (2005) Constraints on the evolution of adaptive phenotypic plasticity in plants. New Phytol 166(1):49–60

Wyman D (1971) Shrubs and vines for American gardens. MacMillan Co, New York