Commentary

Whole genome sequencing (WGS) meets biogeography and shows that genomic selection in forest trees is feasible

Recently developed approaches to high throughout sequencing (HTS) and single nucleotide polymorphism (SNP) detection offer unprecedented power to uncover functionally important genetic variation at the level of whole genomes and entire metapopulations in wild species. In a landmark paper published in this issue of *New Phytologist*, Slavov *et al.* (pp. 713–725) put these novel approaches to work in *Populus trichocarpa* (black cottonwood), an ecologically and economically important forest tree with an assembled and annotated genome (Tuskan *et al.*, 2006). This study is important because it takes whole genome sequencing (WGS) in wild species to a level that allows a range of topics in the modern life sciences to be addressed, from biogeography to molecular plant domestication and breeding, with a level of precision normally only seen in human medical genetics.

'*Slavov* et al. *have demonstrated that whole genome resequencing in wild, outcrossing forest trees is now ready to tackle major issues of general scientific and applied interest.*'

Making use of Illumina-based WGS and Infineum SNP detection assays, Slavov *et al.* uncover significant geographic differentiation in *Populus trichocarpa* at multiple spatial scales across the species' range in western North America. They go on to demonstrate significant covariation of genetic polymorphisms and geographic latitude. Interestingly, this covariation is stronger for nongenic SNPs at the range-wide scale and for genic SNPs in the core of the species' range. This points to neutral, demographic processes ('isolation by distance') as drivers of differentiation over larger geographic areas and to locally varying selection as a potential driver of differentiation at a smaller spatial scale. Next, Slavov *et al.* examine patterns of recombination and linkage disequilibrium (LD) along the genome and across the species' range. Patterns along chromosomes indicate the presence of LD in this outcrossing species over much larger physical distances than previously thought

(3–6 kilo bases), and levels of recombination along chromosomes covaried with several DNA sequence and epigenetic features. LD across the species' range, however, varied across populations, thus pointing to important variation in effective population sizes (N_e) and contributing to our understanding of the biogeographic history of this important forest tree.

Current HTS approaches suitable for population genetics have been around for several years now (Metzker, 2010), but surprisingly few studies have managed to go beyond mere bioinformatic problem-solving exercises and address major biological or applied issues in wild, undomesticated species (Hohenlohe *et al.*, 2010; Gompert *et al.*, 2012). Most current HTS studies of wild species make use of genotyping-by-resequencing approaches such as restriction site associated DNA sequencing (RAD-seq; Hohenlohe *et al.*, 2010), which provide sufficient coverage depth at a manageable cost by sequencing low-complexity DNA libraries rather than whole genomes (but see Turner *et al.* (2010) and Rubin *et al.* (2010) for rare examples of WGS studies involving populations of wild species). The Slavov *et al.* study stands out, because they convincingly manage to interpret WGS data for a wild, largely undomesticated forest tree (*Populus trichocarpa*) in terms of biogeography, domestication and breeding.

With regard to biogeography, their genomic data allow the authors to revisit long-standing questions on the phylogeographic history of plants in western North America (Soltis *et al.*, 1997). In particular, the lack of a north–south gradient in diversity (and thus, of a gradient in N_e) speaks against migration from a single southern source and is consistent with colonization from multiple refugia following the last ice age (Soltis *et al.*, 1997). Based on high genomic diversity and N_e of populations in the center of the species' range, Slavov *et al.* argue that these may represent genomic 'melting pots', *sensu* Petit *et al.* (2003) or de Carvalho *et al.* (2010). Recent admixture in such 'melting pots' is expected to leave specific signatures in admixed genomes, with large tracts of DNA originating from each participating source gene pool (Stölting *et al.*, 2012). It will be interesting to follow up on the 'melting pot' hypothesis by tracking the likely origin and size of chromosomal blocks present in central populations.

With a view towards tree domestication, recent work in humans and forest trees indicates that functionally important, potentially adaptive traits often have a polygenic basis (Pritchard *et al.*, 2010; Neale & Kremer, 2011). Hence, HTS and SNP genotyping in trees offers great potential in genomic selection programs, that is, assessing the sum of gene effects (breeding values) for important traits directly from high-density molecular genetic data. The finding that LD extends over relatively large chromosomal distances (Slavov *et al.*) is encouraging, as it implies that fewer markers than expected are required to tag LD blocks for this purpose. This may hold true for *Populus trichocarpa* and for many other tree species with similar sets of life history traits.

We anticipate that the impact of the Slavov *et al.* study – and potential future studies arising from their data – will extend far beyond *Populus trichocarpa* to other species of *Populus* (poplars, aspens, cottonwoods) and to other groups of taxa not yet perceived as model species. This potential is most easily illustrated by a comparison to our own HTS data in two hybridizing Eurasian species of *Populus*, *P. alba* (white poplar) and *P. tremula* (European aspen) (Stölting *et al.*, 2012). The genomic 'coldspots' of recombination identified in *P. trichocarpa* by Slavov *et al.* have a remarkable tendency to coincide with genomic 'hotspots' of differentiation between *P. alba* and *P. tremula* in our study (Stölting *et al.*, 2012). The low-recombination region at 8 Mio base pairs on *Populus* chromosome VII (fig. 7 in Slavov *et al.*), for example, coincides exactly with a region of increased interspecific differentiation (e.g. F_{ST}) and reduced diversity (SNP density) in *P. alba* and *P. tremula* (Stölting *et al.*, 2012). Similar parallels can be found on several other chromosomes (e.g. VIII, X, XIV, XV, see the Supporting materials of both studies). For convenience, we have magnified and redrawn our interspecific genome scan results (Stölting *et al.*, 2012) for chromosome VII in Fig. 1 of this Commentary. Viewed in combination, the HTS data of Slavov *et al.* and Stölting *et al.* (2012) are suggestive of stronger selection and/or drift in genomic regions with low recombination, effectively reducing diversity and contributing to the maintenance of genomic differentiation between species in the face of gene flow (Smadja & Butlin, 2011). Of course, a caveat is in line here: Slavov *et al.* used a WGS approach whereas our study is based on *c.* 38 000 SNPs from RAD-seq (Stölting *et al.*, 2012), and recombination rates were not addressed directly in our study. Still, our comparison indicates a great potential for comparative genomics to 'tie together' and synthesize ecological and evolutionary genomic data between highly syntenic species of *Populus*; WGS-based genome scans of populations from both Eurasian species (*P. alba* and *P. tremula*) are currently underway in our laboratory.

Short-read HTS studies such as those by Slavov *et al.* or ourselves (Stölting *et al.*, 2012) are by no means limited to taxa with completely sequenced reference genomes (recall that standard HTS approaches in population genetics normally rely on annotated and assembled reference genomes for mapping sequence reads and detecting SNPs). Fig. 2 shows a re-analysis of our *Populus alba* and *P. tremula* HTS data (fully documented in Stölting *et al.*, 2012) with and without the use of the *P. trichocarpa* genome sequence for reference mapping. Polymorphism detection without the reference genome was achieved by *de novo* clustering of sequence reads, using sequence alignment criteria as in Stölting *et al.* (2012) and widely used software (see the legend of Fig. 2). After *de novo* sequence clustering and SNP calling, the *P. trichocarpa* genome was used to visualize and plot the results from both approaches (Fig. 2), again using chromosome VII as a convenient example. The results should encourage students with an interest in applying HTS to populations of nonmodel species. Although fewer SNPs are detected by *de novo* clustering compared with reference mapping (Fig. 2, red and blue curves), SNP densities exhibit highly similar patterns along chromosomes with both approaches (Pearson's correlation $r = 0.827$ for chromosome VII). In effect, our visualization and direct comparison of both strategies using the *P. trichocarpa* genome sequence confirms the validity and usefulness of *de novo* clustering of HTS reads, corroborating recent findings by Catchen *et al.* (2011).

In conclusion, Slavov *et al.* have demonstrated that whole genome resequencing in wild, outcrossing forest trees is now ready to tackle major issues of general scientific and applied interest. We anticipate that future work in *Populus* spp. will reveal: how strongly the genomic landscape of recombination (and thus the size of DNA blocks of relevance to selection programs) varies across individuals, populations, and species; how exactly drift and selection interact to shape the response of genomes to environmental shifts (e.g. climate change); how drift and selection interact to drive the build-up and maintenance of genomic 'islands' or 'continents' of differentiation between species, thus preventing or facilitating the breakdown of
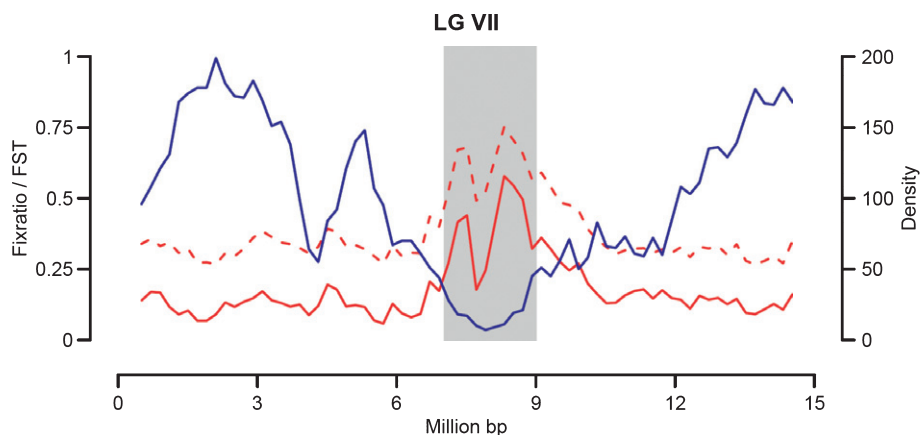
LG VII

Fig. 1 Genomic differentiation between two Eurasian species of *Populus*, *P. alba* and *P. tremula*, along linkage group VII, based on single nucleotide polymorphisms (SNPs) detected by restriction site associated DNA (RAD) sequencing. Results are redrawn from Stölting *et al.* (2012), where the data are fully documented. Windowed analyses (window size 1 Mio base pairs, step size 200 kilo bases) of two measures of differentiation are indicated along the *x*-axis (window midpoints as physical distances in Mio base pairs): the proportion of SNPs fixed between species (fix ratio; red solid line) and F_{ST} (red hatched line). The blue curve indicates SNP density per window (number of SNPs per Mio base pairs). A region of particularly low recombination identified by Slavov *et al.* (this issue of *New Phytologist*, pp. 713–725; fig. 7) is highlighted in gray.
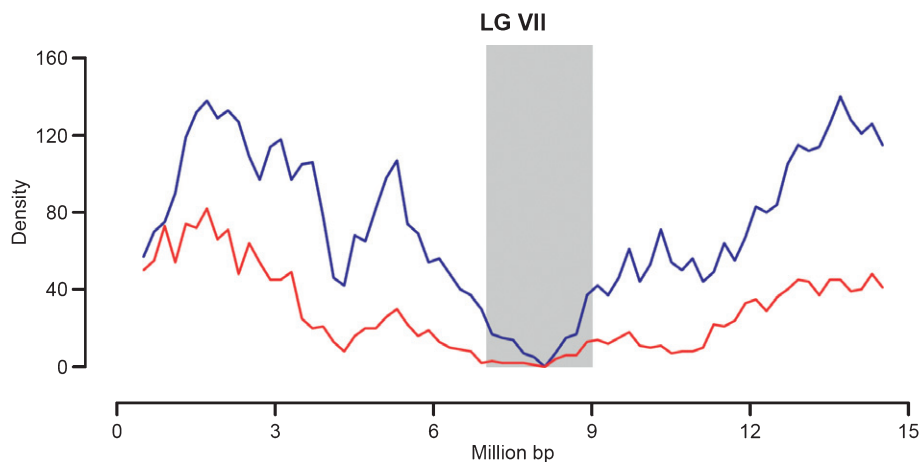
Fig. 2 Direct comparison of *genomic reference mapping* vs reference-free *de novo* clustering of high throughout sequencing (HTS) sequence reads, exemplified by restriction site associated DNA sequencing (RAD-seq) data for *Populus alba* and *P. tremula* (Stölting *et al.*, 2012). Single nucleotide polymorphism (SNP) densities (*y*-axis, number of SNPs per Mio base pairs) from both approaches are shown along *Populus* chromosome VII (*x*-axis, physical distance in Mio base pairs). SNP calling by genomic reference mapping (blue curve) was carried out with open source software following Stölting *et al.* (2012), using *P. trichocarpa* genome assembly v2 and a minimum threshold of eight reads per SNP, not permitting missing data in the final SNP table. Reference-free SNP calling (red curve) was achieved by *de novo* clustering of HTS reads with the Uᴄʟᴜsᴛ v3.0 software (http://drive5.com/usearch/usearch3.0.html) using the same statistical criteria. Window sizes are as in Fig. 1, total sample size was *N* = 14 haploid genomes per species, and all analyses were restricted to biallelic SNPs. A region of particularly low recombination identified by Slavov *et al.* (this issue of *New Phytologist*, pp. 713–725; fig. 7) is highlighted in gray.

reproductive barriers. *Populus* offers the great advantage of multiple ecologically and economically important species with highly syntenic genomes. Rapidly evolving bioinformatic approaches to HTS-based genotyping are starting to enable studies of genomic diversity in many other species that have not been 'on the radar' of molecular biology and genomics so far. Clearly, HTS (formerly known as 'next generation sequencing') is starting to change how we approach key issues in the fundamental and applied life sciences.

Christian Lexer* and Kai N. Stölting

Unit of Ecology & Evolution, Department of Biology, University of Fribourg, Chemin du Musée 10, CH-1700, Fribourg, Switzerland
(*Author for correspondence: tel +41 26 300 8868; email christian.lexer@unifr.ch)

References

de Carvalho D, Ingvarsson PK, Joseph J, Suter L, Sedivy C, Macaya-Sanz D, Cottrell J, Heinze B, Schanzer I, Lexer C. 2010. Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Molecular Ecology* 19: 1638–1650.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *Genes, Genomes, Genetics* 1: 171–182.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA. 2012. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution* 66: 2167–2181.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* 6: e1000862.

Metzker ML. 2010. Sequencing technologies – the next generation. *Nature Reviews Genetics* 11: 31–46.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* 12: 111–122.

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Chaddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M et al. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* 300: 1563–1565.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: sweeps, soft sweeps, and polygenic adaptation. *Current Biology* 20: R208–R215.

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464: 587–591.

Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA et al. 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist* 196: 713–725.

Smadja CM, Butlin RK. 2011. A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology* 20: 5123–5140.

Soltis DE, Gitzendanner MA, Strenge DD, Soltis PS. 1997. Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution* 206: 353–373.

Stölting KN, Nipper R, Lindtke D, Barbará T, Caseys C, Waeber S, Castiglione S, Lexer C. 2012. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*. doi:10.1111/mec.12011

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* 42: 260–263.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Rombauts S, Salamov A, Schein J et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604.