Tree Genetics & Genomes (2018) 14: 39
https://doi.org/10.1007/s11295-018-1251-3

REVIEW

RADseq approaches and applications for forest tree genetics

Thomas L. Parchman[1] · Joshua P. Jahner[1] · Kathryn A. Uckele[1] · Lanie M. Galland[1] · Andrew J. Eckert[2]

Received: 30 August 2017 / Revised: 30 April 2018 / Accepted: 2 May 2018 / Published online: 21 May 2018

Abstract

As tree species vary extensively in genome size, complexity, and resource development, reduced representation methods have been increasingly employed for the generation of population genomic data. By allowing rapid marker discovery and genotyping for thousands of genomic regions in many individuals without requiring genomic resources, restriction site-associated DNA sequencing (RADseq) methods have dramatically improved our ability to bring population genomic perspectives to non-model trees. The rapid recent increase in studies of trees utilizing RADseq suggests that it is likely to become among the most common approaches for generating genome-wide data for a variety of applications. Here we provide a practical review of RADseq and its application to research areas of tree genetics. We briefly review RADseq laboratory methods and consider analytical approaches for assembly, variant calling, and bioinformatic processing. To guide considerations for study design, we use in silico analyses of eight available tree genomes to illustrate how expected marker number and density vary across laboratory approaches and genome sizes, and to consider the ability of RADseq designs to query coding regions. We review the empirical use of RADseq for different research objectives, considering its strengths and limitations. Many studies have used RADseq data to perform genome scans for selection, although limited marker density and linkage disequilibrium will often compromise its utility for such analyses. Regardless of this limitation, RADseq offers a powerful and inexpensive technique for generating genome-wide SNP data that can greatly contribute to research spanning phylogenetic and population genetic inference, linkage mapping, and quantitative genetic parameter estimation for tree genetics.

Keywords Genome scan · Genotyping by sequencing · Linkage mapping · Phylogenetics · Population genomics · RADseq

Understanding how genetic variation is shaped across landscapes and genomes is critical for inferring the action of evolutionary processes on forest tree populations, and for predicting and managing their response to environmental change (González-Martínez et al. 2006; Aitken et al. 2008; Alberto et al. 2013; Sork et al. 2013). Until recently, population genetic analyses in tree species were often limited to traditional molecular markers (e.g., SSRs), Sanger sequencing of modest numbers of genes, or to relatively expensive genotyping approaches that were dependent on genomic or transcriptomic resources for development (e.g., Eckert et al. 2013; Geraldes et al. 2013; Pavy et al. 2013). The emergence of DNA sequencing technologies that allow inexpensive and massively parallel sequencing of short DNA reads has changed this rapidly (Mardis 2013) and is driving a steady increase in genomic resources. Transcriptomes have been sequenced, assembled, and annotated for many tree species (e.g., Pinosio et al. 2014; Yeaman et al. 2014). Whole genome reference sequences have been published for a growing number of angiosperms (e.g., *Eucalyptus grandis*, Myburg et al. 2014; *Malus domestica*, Velasco et al. 2010; *Populus trichocarpa*, Tuskan et al. 2006; *Prunus persica*, Verde et al. 2013) and even for a number of the massive-genomed conifers (e.g., *Picea abies*, Nystedt et al. 2013; *Picea glauca*, Birol et al. 2013; *Pinus taeda*, Neale et al. 2014; *Pinus lambertiana*, Stevens et al. 2016).

By increasing the ability to genotype far larger numbers of loci and individuals, high-throughput sequencing technolo-

Communicated by S. C. González-Martínez

✉ Thomas L. Parchman
tparchman@unr.edu

[1] Department of Biology and Program in Ecology, Evolution, and Conservation Biology, University of Nevada Reno, Reno, NV 89557, USA

[2] Department of Biology, Virginia Commonwealth University, Richmond, VA 23284, USA

gies are also improving our understanding of how evolutionary processes shape genetic variation across populations, species, and genomes of tree species (Neale and Kremer 2011; Sork et al. 2013; Holliday et al. 2016; Ingvarsson et al. 2016; Yeaman et al. 2016). Substantial genomic resource development has bolstered population genomic perspectives in a number of well-studied angiosperm systems. A quality reference genome (Tuskan et al. 2006), high-density SNP genotyping arrays (Geraldes et al. 2013), and whole genome resequencing studies (e.g., Slavov et al. 2012; Evans et al. 2014) have further developed *Populus* as a model tree system for molecular, functional, and evolutionary genomics. Similar resources have been developed for *Eucalyptus* (Myburg et al. 2014; Silva-Junior et al. 2015; Silva-Junior and Grattapaglia 2015) as well as several fruit tree species (e.g., Bianco et al. 2014), and additional tree species with manageable genome sizes will soon follow. Whole genome resequencing produces ideal data for population genomics and has been recently employed for tree species with smaller genome sizes and quality reference genomes (Slavov et al. 2012; Evans et al. 2014; Silva-Junior and Grattapaglia 2015; Wang et al. 2016). However, it is not currently cost-effective for most species, particularly those with large genomes, or for studies requiring large numbers of individuals. As a result, investigators have often turned to reduced representation methods that sample subsets of genomes. Targeted capture approaches allow high-throughput sequencing of predetermined genomic regions but require genomic resources and the design of capture arrays (Jones and Good 2016; but see Puritz and Lotterhos 2017). As protein coding regions can represent small fractions of genome space, targeted capture approaches for such regions (referred to as here as exome sequencing) have emerged as promising for population genomic studies in tree species (e.g., Zhou et al. 2014; Holliday et al. 2016; Lu et al. 2016; Yeaman et al. 2016). The ability to focus sequencing on exons can be advantageous for some investigations, although such non-uniform sampling of genome space may compromise others.

Restriction enzyme-guided sequencing approaches, such as restriction site-associated DNA sequencing (RADseq, Miller et al. 2007; Baird et al. 2008) and genotyping by sequencing (GBS, Elshire et al. 2011) have become the most popular reduced representation methods for non-model organisms. The terms RADseq and GBS have both been used as umbrella terms describing a family of methods that use restriction enzymes to guide complexity reduction and sequencing; we collectively refer to these methods here as RADseq. Because library preparation is simply based on restriction enzyme digest and subsequent adaptor ligation, RADseq methods can be implemented with or without prior genomic resources and can be used to rapidly and inexpensively generate data for large numbers of individuals. In addition, because it simultaneously allows SNP discovery and genotyping in the genomic regions sequenced, RADseq suffers less from ascertainment bias than some alternative genotyping approaches. A diversity of laboratory methods offer flexibility in the number of loci that can be genotyped, the number of individuals that can be multiplexed, and thus overall project design and cost (see Davey et al. 2011; Puritz et al. 2014b; Andrews et al. 2016). Because of these attributes, RADseq is rapidly bringing genome-wide perspectives and increased resolution to basic and applied areas of ecological and evolutionary genetic research in natural populations of a diverse range of non-model taxa, including a diversity of tree species (Narum et al. 2013; Andrews et al. 2016).

Here we provide a practical review of methodological and analytical aspects of RADseq and its application to research areas in tree genetics. We begin with a brief overview of variability in the methods through which RADseq libraries are produced, and how this variation can be considered for project design. We use in silico analyses with different RADseq approaches across forest tree species spanning a continuum of genome sizes ranging from small (*Prunus persica*, 0.26 Gb) to large (*Pinus taeda*, 22 Gb) to illustrate how expected marker numbers and densities vary across genomes and laboratory methods and then link these patterns to their suitability for different research objectives. We review characteristics of RADseq data, potential sources of genotyping error, and bioinformatic approaches to alignment and genotype inference, including methods that account for statistical uncertainty for lower coverage sequencing data. Lastly, we consider applications of RADseq to different research objectives and review research examples in forest tree genetics that have and may continue to be well served by such data. While RADseq may not be the ideal method for analyses requiring saturated marker densities, it represents a rapid and affordable means for generating genome-wide SNP data that can contribute substantially to our understanding of genetic variation in trees for a diversity of research objectives.

How RADseq approaches work

The RADseq family of methods share the attributes of using restriction enzyme digests and barcoded adaptor ligation to guide high-throughput sequencing of subsets of genomes for many samples. The original RADseq method (Miller et al. 2007; Baird et al. 2008) is most widely used, but a variety of alternative methods have more recently been introduced (for reviews see Davey et al. 2011; Puritz et al. 2014b; Andrews et al. 2016). Each begins with digestion of high molecular weight genomic DNA with one or a combination of restriction enzymes, so that DNA sequencing is guided by the genomic distribution of restriction enzyme cut sites. Variation in the length and GC content of recognition sites gives rise to variation in the frequency and genomic distribution of cut sites (Davey et al. 2011). Thus, the degree of complexity reduction

can be tuned by choosing enzymes that cut more or less frequently. After digestion, customized adaptors that facilitate PCR amplification and sequencing by synthesis are ligated to fragments. DNA barcodes, each corresponding to an individual sample, are embedded within adaptors to allow samples to be pooled and simultaneously sequenced. After barcoded adaptor ligation, samples can typically be pooled for the remaining laboratory steps, further reducing the time and monetary investment required. PCR is then used to amplify fragments to produce DNA libraries suitable for sequencing on short-read technologies such as the Illumina platform.

The original RADseq uses a single restriction enzyme and mechanical shearing to optimize fragment sizes for sequencing; size selection is not used and all fragments are sequenced (Miller et al. 2007; Baird et al. 2008). For most other approaches, a subset of the fragment size distribution is extracted to ensure that fragments are within the optimal size range for sequencing and to further reduce the sampled portion of the genome, thereby increasing coverage depth per locus or allowing larger numbers of individuals to be multiplexed. Size selection can be partially accomplished through PCR cycles that eliminate long fragments (e.g., GBS; Elshire et al. 2011), but is more commonly achieved using automated electrophoresis extractions with a device such as Pippin Prep (Sage Science, Inc.). Fragment libraries are then used to generate single-end or paired-end sequencing data, usually on an Illumina instrument. The resulting reads begin with bases associated with barcodes and restriction enzyme cut sites and extend into the tagged genomic regions.

Multiple RADseq methods exist, all having features that allow the number and density of markers to be adjusted for particular research scenarios. Laboratory protocols implemented for tree genetics have predominantly been the original RADseq (Miller et al. 2007; Baird et al. 2008), ddRADseq (double digest restriction site-associated DNA sequencing, Peterson et al. 2012), and GBS (Elshire et al. 2011) (Table 1). For review of other methodological variants, see Davey et al. (2011) and Andrews et al. (2016). The methods vary by their use of one or two restriction enzymes, the manner in which adaptors are ligated to fragments, the point in library preparation where multiplexing occurs, and other aspects related to complexity reduction. Here we provide only this brief overview, as descriptions and comparisons of alternative laboratory methods have been thoroughly reviewed elsewhere (see Davey et al. 2011; Poland and Rife 2012; Puritz et al. 2014b; Andrews et al. 2016). Although the purchase of customized barcoded oligos entails an initial cost for labs first establishing RADseq workflows, the expense of constructing libraries is otherwise limited to the modest cost of restriction enzymes, PCR reagents, and plastics. Thus, the production of genome-wide genotypic data with RADseq represents a far less substantial time and monetary commitment than past methods required. Instead, research progress is more often limited by the ability and availability of trained investigators to thoroughly execute the bioinformatic analysis of data.

Research design for RADseq with forest trees

The number of sampled genomic regions will depend on genome size, the RADseq method utilized, choice of restriction enzymes, degree of size selection, genetic diversity, and parameter settings applied during the filtering of data. The preferred density of loci should depend on the research objective and the pattern and extent of linkage disequilibrium (LD) in the populations under study. Researchers seeking higher density SNPs for locus-specific parameter estimates (i.e., genome scans, association mapping) can employ methods with the least complexity reduction in order to maximize marker density. In contrast, many research objectives in tree genetics are less dependent on maximizing marker density and can benefit more from genotyping a reduced set of loci in larger numbers of individuals. Prediction of the number of loci expected for different RADseq protocols can be valuable for optimizing sequencing effort based on the number of samples, desired marker density, and sequencing coverage depth required to best suit research goals. While probabilistic prediction can be accomplished based on genome size and GC content (Davey et al. 2011; `radcounter` tool available at www.wiki.ed.ac.uk/display/RADSequencing), considerable phylogenetic variation exists among eukaryotic lineages in the frequency of enzyme recognition sequences independent of GC content (Davey et al. 2011; Herrera et al. 2015). Several software packages exist for the in silico prediction of the number of genomic regions a RADseq protocol will query for taxa with available reference genomes (`simRAD`, Lepais and Weir 2014; `ddradseqtools`, Mora-Márquez et al. 2017). For taxa where reference genomes are not available, genomes of similar size for closely related taxa can often be used for this purpose (e.g., Chafin et al. 2017).

To characterize the expected numbers and densities of loci, we used `simRAD`, following example code of Lepais and Weir (2014), to perform in silico predictions for eight tree species representing a continuum of genome size (Table 2). We utilized reference genomes for peach (*Prunus persica*, Verde et al. 2013), black cottonwood (*Populus trichocarpa*, Tuskan et al. 2006), grand eucalyptus (*Eucalyptus grandis*, Myburg et al. 2014), domesticated apple (*Malus domestica*, Velasco et al. 2010), a basal angiosperm (*Amborella trichopoda*, Albert et al. 2013), valley oak (*Quercus lobata*; Sork et al. 2016), Douglas-fir (*Pseudotsuga menziesii*, Neale et al. 2017), and loblolly pine (*Pinus taeda*, Neale et al. 2014). We predicted the number of fragments that would be sequenced with the two RADseq methods most commonly used for tree species (RADseq and ddRADseq; Table 1) across a

Table 1 Examples of studies utilizing RADseq methods for forest tree genetics. Studies are organized by the type of analysis and by the year of publication. For each study, the RADseq method, restriction enzyme(s), and assembly method are listed

Species	Citation	Method	Type of analysis
Actinidia chinensis (kiwifruit)	Liu et al. 2017	RAD; *Eco*RI, R	Linkage mapping
Populus tremula (Salicaceae)	Zhigunov et al. 2017	ddRAD; *Hind*III-*Nla*III, RC	Linkage mapping, QTL
Gleditsia triacanthos (Fabaceae)	Gailing et al. 2017	RAD; *Sbf*I, D	Linkage mapping
Citrus (mandarin)	Imai et al. 2017	GBS; *Pst*I, RC	Linkage mapping, QTL
Quercus rubra (Fagaceae)	Konar et al. 2017	ddRAD; *Eco*RI-*Mse*I, D	Linkage mapping
Ficus carica (common fig)	Mori et al. 2017	ddRAD; *Pst*I-*Msp*I, R	Linkage mapping
Olea europaea ssp. europaea (olive)	Marchese et al. 2016	GBS; *Ape*KI, D	Linkage mapping
Populus deltoides × *P. simonii* (Salicaceae)	Mousavi et al. 2016	RAD; *Eco*RI, RC and D	Linkage mapping
Populus deltoides × *P. simonii* (Salicaceae)	Tong et al. 2016	RAD; *Eco*RI, RC	Linkage mapping
Ziziphus jujuba (jujube)	Zhang et al. 2016	GBS[a]; *Mse*I-*Hae*III-*Eco*RI, R	Linkage mapping
Prunus persica (peach)	Bielenberg et al. 2015	GBS; *Ape*KI, R	Linkage mapping, QTL
Pinus balfouriana (Pinaceae)	Friedline et al. 2015	ddRAD; *Eco*RI-*Mse*I, D	Linkage mapping
Prunus avium (sweet cherry)	Guajardo et al. 2015	GBS; *Ape*KI, RC	Linkage mapping
Citrus grandis (pummelo)	Guo et al. 2015	RAD; *Eco*RI, RC	Linkage mapping
Elaeis guineensis (oil palm)	Pootakham et al. 2015a	ddRAD; *Pst*I-*Msp*I, R	Linkage mapping, QTL
Hevea brasiliensis (rubber tree)	Pootakham et al. 2015b	ddRAD; *Pst*I-*Msp*I, R	Linkage mapping
Callitris glaucophylla (Cupressaceae)	Sakaguchi et al. 2015	ddRAD; *Eco*RI-*Bgl*II, RC	Linkage mapping
Actinidia chinensis (kiwifruit)	Scaglione et al. 2015	ddRAD; *Sph*I-*Mbo*I, R	Linkage mapping
Malus × *domestica* (apple)	Gardner et al. 2014	ddRAD; *Hind*III-*Msp*I, R	Linkage mapping, QTL
Ziziphus jujuba (jujube)	Zhao et al. 2014	RAD; *Eco*RI, D	Linkage mapping
Robinia pseudoacacia (Fabaceae)	Verdu et al. 2016	ddRAD; *Eco*RI-*Mse*I, D	Marker development
Cedrus atlantica (Pinaceae)	Karam et al. 2015	RAD; *Pst*I, D	Marker development
Pinus tabuliformis, *P. densata*, and *P. yunnanensis* (Pinaceae)	Pan et al. 2015	GBS/ddRAD; *Hpa*II, *Pst*I/*Eco*RI-*Mse*I, D	Marker development, method optimization
Frangula alnus (Rhamnaceae)	De Kort et al. 2014c	RAD-PE; *Sbf*I, D	Marker development
Pinus contorta and *Picea glauca* (Pinaceae)	Chen et al. 2013	GBS; *Ape*KI, D	Marker development, method optimization
Quercus section *Cyclobalanopsis* (Fagaceae)	Deng et al. 2018	RAD; *Pst*I, D	Phylogenetics
Aurantioideae (Rutaceae)	Nagano et al. 2018	ddRAD; *Bgl*II-*Eco*RI, D and RC	Phylogenetics
Quercus sect. *Quercus* (Fagaceae)	Fitz-Gibbon et al. 2017	RAD; *Pst*I, D and RC	Phylogenetics
Coffea (Rubiaceae)	Hamon et al. 2017	GBS; *Pst*I, RC	Phylogenetics
Quercus sect. *Quercus* (Fagaceae)	McVay et al. 2017	RAD; *Pst*I, D	Phylogenetics
Quercus sects. *Quercus* and *Protobalanus* (Fagaceae)	Pham et al. 2017	RAD; *Pst*I, D	Phylogenetics
Morella (Myricaceae)	Liu et al. 2015a	RAD; *Eco*RI, D	Phylogenetics
Diospyros (Ebenaceae)	Paun et al. 2016	RAD; *Sbf*I, D	Phylogenetics
Quercus sects. *Quercus*, *Lobatae*, and *Protobalanus* (Fagaceae)	Hipp et al. 2014	RAD; *Pst*I, D	Phylogenetics
Quercus sects. *Quercus*, *Lobatae*, and *Protobalanus* (Fagaceae)	Hipp et al. 2013	RAD; *Pst*I, D	Phylogenetics
Quercus chrysolepis and *Q. tomentella* (Fagaceae)	Ortego et al. 2017	ddRAD; *Eco*RI-*Mse*I, D	Phylogenetics, introgression
Quercus series *Virentes* (Fagaceae)	Eaton et al. 2015	RAD; *Pst*I, D	Phylogenetics, introgression
Rhizophora mangle (Rhizophoraceae)	Hodel et al. 2017	ddRAD; *Eco*RI-*Mse*I, D	Phylogeography
Quercus series *Virentes* (Fagaceae)	Cavender-Bares et al. 2015	RAD; *Pst*I, D	Phylogeography
Pinus strobiformis × *P. flexilis*	Menon et al. 2018	ddRAD; *Eco*RI-*Mse*I, D	Hybridization
Citrus maxima × *C. reticulata* (pummelo and mandarin)	Oueslati et al. 2017	GBS; *Ape*KI, RC	Hybridization
Populus alba × *P. tremula* (Salicaceae)	Christe et al. 2016	RAD; *Pst*I, RC	Hybridization
Betula nana, *B. pubescens*, and *B. pendula* (Betulaceae)	Zohren et al. 2016	RAD; *Pst*I, R and RC	Hybridization

Table 1 (continued)

Species	Citation	Method	Type of analysis
Populus alba × *P. tremula* (Salicaceae)	Lindtke et al. 2014	ddRAD; *Eco*RI-*Mse*I, R and RC	Hybridization
Populus alba × *P. tremula* (Salicaceae)	Stölting et al. 2013	RAD; *Pst*I, RC	Hybridization
Betula nana, *B. pubescens*, and *B. pendula* (Betulaceae)	Wang et al. 2013	RAD; *Pst*I, D, R, and RC	Hybridization, marker development
Populus alba × *P. tremula* (Salicaceae)	Caseys et al. 2012	RAD; *Pst*I, RC	Hybridization
Cornus florida (Cornaceae)	Pais et al. 2017	ddRAD; *Pst*I-*Msp*I, D	Genome scan, landscape genomics
Tsuga mertensiana (Pinaceae)	Johnson et al. 2017a	ddRAD; *Sph*I-*Mlu*CI, D	Landscape genomics
Alnus glutinosa (Betulaceae)	De Kort et al. 2014b	GBS; *Pst*I, D	Landscape genomics
Picea engelmannii × *P. glauca* (Pinaceae)	El-Dien et al. 2015	GBS; *Ape*KI, D	Genomic selection
Picea sitchensis (Pinaceae)	Fuentes-Utrilla et al. 2017	RAD/SD-RAD; *Pst*I/*Pst*I-*Alw*I, D	Genomic selection, QTL analysis
Picea engelmannii × *P. glauca* (Pinaceae)	Ratcliffe et al. 2015	GBS; D	Genomic selection
Pinus albicaulis (Pinaceae)	Lind et al. 2017	ddRAD; *Mse*I-*Eco*RI, D	Association mapping
Banksia attenuata (Proteaceae)	He et al. 2016	RAD; *Eco*RI, D	Association mapping
Pinus contorta (Pinaceae)	Parchman et al. 2012	ddRAD; *Eco*RI-*Mse*I, D	Association mapping
Tsuga mertensiana (Pinaceae)	Johnson et al. 2017b	ddRAD; *Sph*I-*Mlu*CI, D	Parentage analysis, dispersal distance estimation
Milicia (Moraceae)	Daïnou et al. 2016	RAD; *Sbf*I, D	Species delimitation
Alnus glutinosa (Betulaceae)	De Kort et al. 2016	GBS; *Pst*I, D	Heritability, evolvability
Castanopsis carlesii (Fagaceae)	Sun et al. 2016	RAD; *Sbf*I, D	Population genetics, genome scan
Alnus glutinosa (Betulaceae)	De Kort et al. 2014a	GBS; *Pst*I, D	Seed zone delineation

Naming of sequencing approaches was taken from Andrews et al. (2016) or from the cited literature: GBS, genotyping by sequencing (Elshire et al. 2011); RAD, restriction site-associated DNA sequencing (Miller et al. 2007; Baird et al. 2008); ddRAD, ddRADseq (Peterson et al. 2012); SD-RAD, second digestion RADseq (as described in Fuentes-Utrilla et al. 2017); RAD-PE, paired-end RADseq (Etter et al. 2011)

D de novo assembly, *R* reference-based assembly to the genome of the same species, *RC* reference-based assembly to a congener's genome

[a] Three enzymes were used in this protocol

continuum of complexity reduction involving different enzymes or size selection windows. Genome sizes ranged from 0.26 to 22 Gb, and enzyme recognition sites ranged from four to eight bases in length (Table 2). The original RADseq (Baird et al. 2008) uses single restriction enzymes without further complexity reduction. For this, we used the `in.silico.digest` function to digest genomes at specified cut sites without size selection. The use of two restriction enzymes and subsequent size selection with ddRADseq (Peterson et al. 2012) increases complexity reduction. For ddRADseq, we used `in.silico.digest` with pairs of enzymes, `adapt.select` to subset fragments properly flanked by the two cut sites, and `size.select` to extract fragments from different size windows (Table 2). Predictions based on the number of restriction sites in genome sequences represent the theoretical maximum number of sampled genomic regions. Empirically, this number may be reduced, and the number of retained polymorphisms will depend on biological (e.g., mutation rate, genetic diversity) and experimental (e.g., sampling scheme) attributes, as well as bioinformatic choices made during analysis (e.g., filtering).

Variation for a given method across species illustrates, as expected, that numbers of predicted loci are positively related to genome size (Table 2). This relationship, however, is not always strongly linear, due to GC content and other aspects of genome architecture influencing the distribution of enzyme recognition sites (Table 2, Fig. 1). Within species, laboratory method, choice of restriction enzymes, and size selection interval create pronounced variation in marker densities (spanning three orders of magnitude; Table 2). Enzymes with larger recognition sites (e.g., *Sbf*I, 8 bp recognition site) typically result in fewer loci than those with shorter sites (e.g., *Eco*RI, 6 bp recognition site; Table 2), although this pattern will not hold for all enzymes (see Herrera et al. 2015). The base composition of sites also matters; *Eco*RI and *Pst*I both have 6 bp recognition sites, yet *Eco*RI, with a higher GC content recognition site, generates substantially more loci than *Pst*I across all species surveyed (Table 2, Fig. 1). Such variation among taxa and enzymes highlights the value of in silico prediction for RADseq research design.

The original RADseq (Baird et al. 2008) does not employ size selection and thus generates the highest marker densities

Table 2 Number of loci predicted for 15 different RADseq laboratory protocols across eight tree species with available reference genomes: *Prunus persica* (Verde et al. 2013); *Populus trichocarpa* (Tuskan et al. 2006); *Eucalyptus grandis* (Myburg et al. 2014); *Malus domestica* (Velasco et al. 2010); *Amborella trichopoda* (Albert et al. 2013); *Quercus lobata* (Sork et al. 2016); *Pseudotsuga menziesii* (Neale et al. 2017); *Pinus taeda* (Neale et al. 2014). The number of predicted loci for each protocol is given in thousands, with marker density listed parenthetically (thousands of loci per gigabase). The sizes of enzyme recognition sites are listed parenthetically after enzyme names, followed by size selection windows for ddRADseq protocols. For the six angiosperm genomes, assembly length is reported instead of estimated genome size

Protocol	Genus							
	Prunus	*Populus*	*Eucalyptus*	*Malus*	*Amborella*	*Quercus*	*Pseudotsuga*	*Pinus*
	Genome size/ assembly length							
	0.23 Gb	0.43 Gb	0.69 Gb	0.70 Gb	0.71 Gb	0.76 Gb	19 Gb	22 Gb
RAD: *Nsi*I (4)	211 (917)	423 (984)	623 (903)	624 (891)	667 (939)	704 (926)	16,853 (887)	17,308 (787)
RAD: *Eco*RI (6)	142 (617)	254 (591)	431 (625)	390 (557)	354 (499)	400 (526)	9495 (500)	10,346 (470)
RAD: *Pst*I (6)	94.8 (412)	152 (353)	222 (322)	208 (297)	149 (210)	161 (212)	4276 (225)	4309 (196)
RAD: *Sbf*I (8)	3.22 (14.0)	5.17 (12.0)	8.65 (12.5)	8.75 (12.5)	12.5 (17.6)	4.95 (6.51)	179 (9.42)	181 (8.23)
ddRAD: *Eco*RI, *Mse*I (6,4), 250–450 bp	17.2 (74.8)	24.5 (57.0)	55.6 (80.6)	44.8 (64.0)	44.2 (62.3)	39.7 (52.2)	1241 (65.3)	1320 (60.0)
ddRAD: *Eco*RI, *Mse*I (6,4), 300–350 bp	4.70 (20.4)	6.57 (15.3)	14.6 (21.2)	12.8 (18.3)	11.8 (16.6)	10.4 (13.7)	322 (16.9)	355 (16.1)
ddRAD: *Eco*RI, *Mse*I (6,4), 350–400 bp	3.35 (14.6)	4.70 (10.9)	11.9 (17.2)	9.40 (13.4)	9.26 (13.0)	7.84 (10.3)	265 (13.9)	288 (13.1)
ddRAD: *Eco*RI, *Mse*I (6,4), 400–450 bp	2.79 (12.1)	3.47 (8.07)	8.58 (12.4)	6.32 (9.03)	6.74 (9.49)	5.86 (7.71)	191 (10.1)	208 (9.45)
ddRAD: *Eco*RI, *Mse*I (6,4), 450–500 bp	1.93 (8.39)	2.64 (6.14)	6.82 (9.88)	5.86 (8.37)	5.05 (7.11)	4.18 (5.50)	176 (9.26)	149 (6.77)
ddRAD: *Eco*RI, *Sph*I (6,6), 250–450 bp	6.16 (26.8)	12.1 (28.1)	19.2 (27.8)	18.0 (25.7)	19.6 (27.6)	14.8 (19.5)	425 (22.4)	432 (19.6)
ddRAD: *Eco*RI, *Sph*I (6,6), 300–350 bp	1.54 (6.70)	2.90 (6.74)	4.50 (6.52)	4.60 (6.57)	4.83 (6.80)	3.48 (4.58)	98.6 (5.19)	104 (4.73)
ddRAD: *Eco*RI, *Sph*I (6,6), 350–400 bp	1.47 (6.39)	3.05 (7.09)	4.91 (7.12)	3.90 (5.57)	4.46 (6.28)	3.68 (4.84)	117 (6.16)	110 (5.00)
ddRAD: *Eco*RI, *Sph*I (6,6), 400–450 bp	1.44 (6.26)	2.99 (6.95)	4.80 (6.96)	4.18 (5.97)	4.40 (6.20)	3.48 (4.58)	93.0 (4.89)	90.6 (4.12)
ddRAD: *Eco*RI, *Sph*I (6,6), 450–500 bp	1.36 (5.91)	2.77 (6.44)	4.20 (6.09)	4.31 (6.16)	3.92 (5.52)	3.19 (4.20)	83.1 (4.37)	91.1 (4.14)
ddRAD: *Eco*RI, *Sbf*I (6,8), 250–450 bp	0.22 (0.96)	0.31 (0.72)	0.45 (0.65)	0.45 (0.64)	0.87 (1.23)	0.24 (0.32)	9.06 (0.48)	9.21 (0.42)

when executed with frequently cutting enzymes (Table 2). RADseq designs utilizing stronger complexity reduction reduce cost per individual sample and can still effectively sample genome space for research objectives that do not require maximizing marker density, such as analyses of population structure, gene flow, phylogenetic inference, or QTL mapping (Peterson et al. 2012). Increased complexity reduction results in the ability to multiplex more samples and/or to achieve

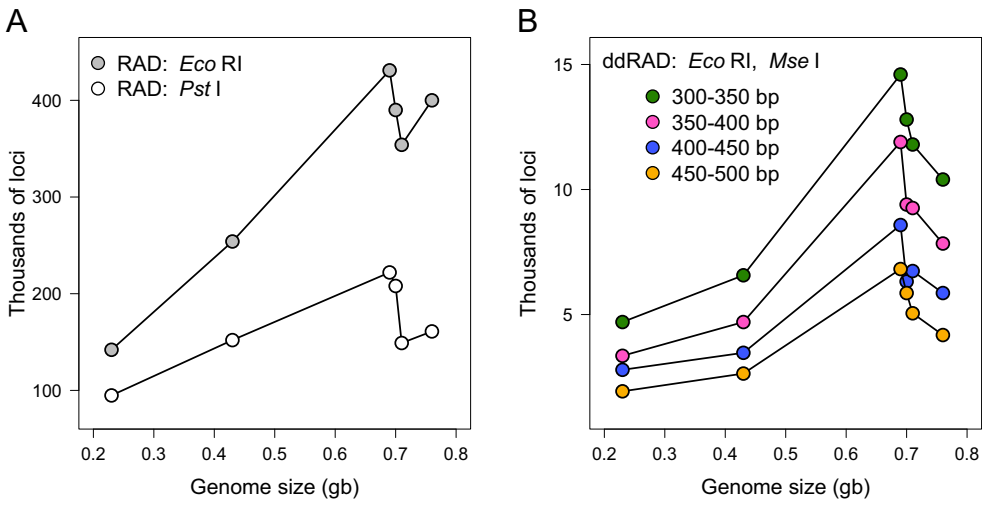

Fig. 1 The relationship between genome size and the number of loci predicted using `simRAD` for six angiosperm species (from smallest to largest genome size: *Prunus persica*; *Populus trichocarpa*; *Eucalyptus grandis*; *Malus domestica*; *Amborella trichopoda*; *Quercus lobata*). The relationship is shown for **a** two original RADseq laboratory protocols using enzymes with a 6 bp restriction site and **b** four ddRADseq protocols that use the same two restriction enzymes (*Eco*RI, *Mse*I) but vary by size selection window (more information in Table 2)

higher coverage depth. Complexity reduction can be controlled to varying degrees by modifying restriction enzymes and/or size selection strategies. The original RADseq employed with restriction enzymes with infrequent recognition sites can reduce complexity substantially (e.g., *Sbf*I, Table 2). Transposons and other repetitive elements can be abundant in plant genomes and are often heavily methylated (Lisch 2009, 2013; Nystedt et al. 2013; Wegrzyn et al. 2014). The use of methylation-sensitive enzymes (e.g., *Eco*RI, *Ape*KI, *Hap*II, *Msp*I) can further reduce complexity by enriching the representation of non-repetitive genomic regions (Elshire et al. 2011; Poland and Rife 2012; Pan et al. 2015). For example, with the original RADseq using *Pst*I (methylation-sensitive 6-cutter) applied to *Cedrus atlantica* (16 Gb genome), 17% of SNPs occurred in protein coding regions and only 3.6% annotated to transposable element families, suggesting substantial enrichment for non-repetitive genomic regions (Karam et al. 2015). As repetitive regions are common in tree genomes and can lead to downstream artefacts with assembly and variant calling, the use of methylation-sensitive enzymes could be effective for complexity reduction and improving analysis quality.

With ddRADseq, both enzyme choice and the range of fragment sizes selected can produce substantial variation in the number of sampled genomic regions (Peterson et al. 2012; Table 2). This can be valuable for sequencing fewer loci at higher depth or across larger numbers of individuals. For any ddRADseq design, narrower size selection windows will also reduce the number of predicted loci. Depending on the enzymes used, different size regions of the fragment size distribution can also contain variable densities of loci. With size selection of 50 base intervals, ranging from 300–350 to 450–500, marker density declines sharply as fragment sizes increase for ddRADseq with *Eco*RI and *Mse*I (Table 2, Fig. 1). In contrast, with the enzymes *Eco*RI and *Sph*I, the number of predicted loci remains similar across the same windows (Table 2). Given its complexity reduction flexibility, ddRADseq is well suited to cost-effective genotyping designs for analyses requiring fewer markers in a broad range of tree species. As the oligos for a given enzyme pair represent an upfront cost to establishing a RADseq workflow, the ability to alter marker densities through size selection changes without enzyme modification makes ddRADseq an attractive option for labs working on multiple species.

To illustrate how commonly employed ddRADseq designs sample genic and non-genic portions of genomes, as both are often implicated as important to the genetic architecture of phenotypes (e.g., Mei et al. 2018), we mapped predicted fragments generated above with in silico ddRADseq analyses (*Eco*RI, *Mse*I; size selection window: 250–450 bp) to the reference genomes for *P. trichocarpa*, *E. grandis*, and *P. taeda*. These species differ greatly in genome size and complexity, as well as the number of predicted RADseq fragments (Table 2). We quantified the ability to uniquely map RADseq fragments to the reference genome from which they originated and the position of these mapped fragments relative to annotated genes. We used BLASTN (default parameter settings) to map fragments to each genome sequence, quantified patterns in the returned list of hits, and summarized the position of these mappings relative to genic regions. In general, the majority of fragments generated in silico could be uniquely mapped to the genome from which they originated. For example, only 2 to 3% of the fragments had a perfect, full-length match to multiple genomic positions. Mapping abilities were efficient even when allowing for conditions comparable to single-end sequencing of fragments containing variants (50% query length with 98.5% identity which translates to 2–6 variants depending upon size of the query), with only 5 to 10% of fragments having second best blast hits of sufficient similarity to create high levels of uncertainty in positional homology.

Mapped fragments overlapped roughly 25% of annotated gene regions for *Populus* and *Eucalyptus* (Fig. 2). We did not examine this pattern for *Pinus*, as the annotations are less certain. The median distance to the closest annotated gene region across all fragments, moreover, varied from 1978 bp (*Populus*) to 5697 bp (*Eucalyptus*), with a 2.9-fold increase in the median distance to the closest gene between species differing 1.5-fold in genome size. Importantly, the ddRADseq prediction for these analyses involves substantial complexity reduction; laboratory methods generating higher marker densities (Table 2) will result in concomitant increases in the coverage of genic regions. These results are consistent with the hypothesis that RADseq-generated fragments can confidently cover portions of genic and non-genic regions of angiosperms. With genome sizes up to 50 times larger, the sampling of genic regions will be lower for conifers, although the use of methylation-sensitive enzymes could improve the recovered fraction (e.g., Karam et al. 2015). These results should be interpreted cautiously, as they rely on qualities of the assembled and annotated genome sequences, do not consider variation in rates of mutation or recombination across genomes, and do not contextualize the degree of coverage within gene regions (e.g., one RADseq fragment will not tag all relevant variation segregating within a locus). Nonetheless, they highlight how genomic resources emerging across multiple clades of trees can be leveraged to gauge the strengths and weaknesses of RADseq designs for particular evolutionary and ecological questions.

Are certain methods better choices for specific research objectives and tree species? The original RADseq used with frequently cutting enzymes will generate the highest marker densities and should be preferred for studies aimed at detecting selection on genome variation, association mapping, or

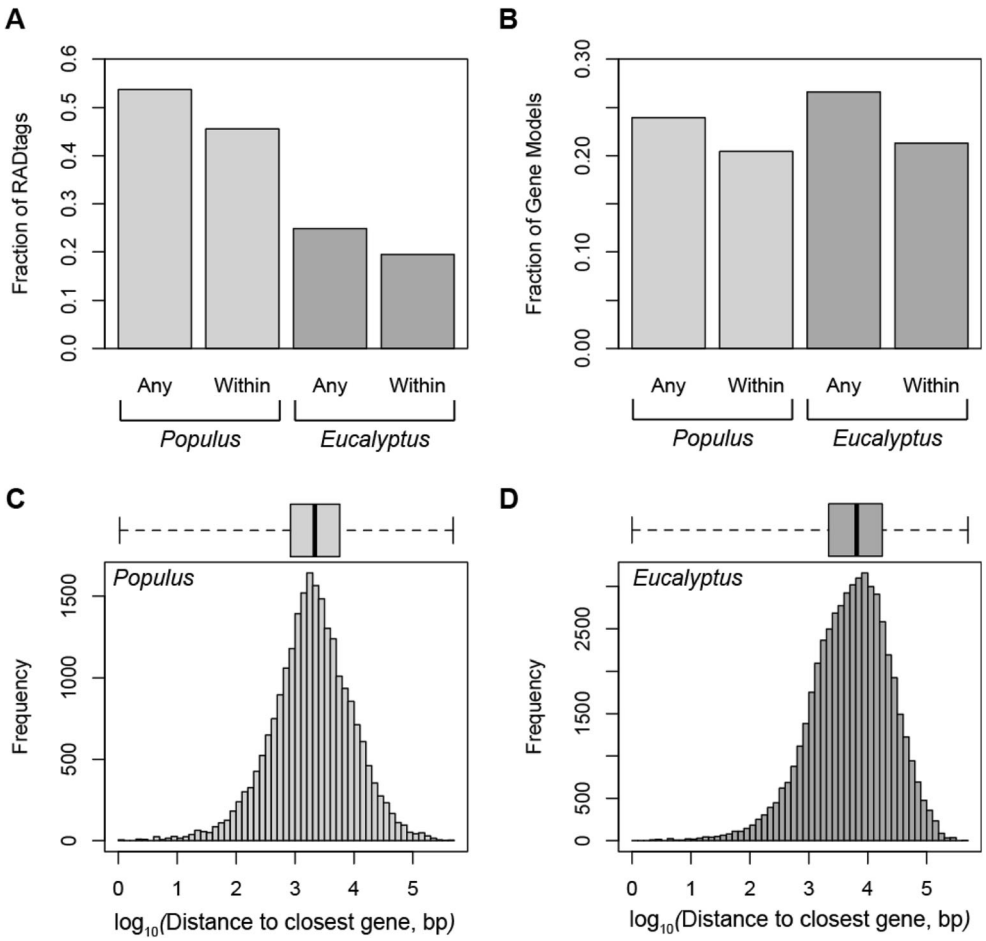

Fig. 2 Summary statistics of the overlap between in silico-predicted ddRADseq fragements and genic regions for *Populus trichocarpa* and *Eucalyptus grandis* reveal that the percentage of RADseq fragments overlapping (Any) or within (Within) genes decreases as a function of genome size (**a**), but that the fraction of genes associated with a RADseq fragment is fairly constant (**b**). Median distances (bp) to the closest gene (i.e., midpoint of gene model to midpoint of the RADseq fragment) also increased with genome size (**c**, **d**), with the median being 1978 bp (or 3.296 in $\log_{10}$ space) and 5697 bp (or 3.756 in $\log_{10}$ space) for *P. trichocarpa* and *E. grandis*, respectively

other analyses benefitting from maximum marker density. Such methods could be relatively cost-effective in smaller-genomed angiosperm species. For the large-genomed conifers, they will sample enormous numbers of genomic regions (Table 2), limiting the number of individuals that can be effectively multiplexed on individual sequencing lanes. All of the methods represented in Table 2 have been used for complexity reduction in tree species. The large genome sizes of conifers can require more substantial complexity reduction to achieve cost-effective multiplexing, and this has been achieved by using infrequently cutting and/or methylation-sensitive restriction enzymes with original RADseq (e.g., Karam et al. 2015; Pan et al. 2015) as well as with ddRADseq approaches employing size selection (e.g., Johnson et al. 2017a; Lind et al. 2017; Menon et al. 2018). Many RADseq studies of tree species have used methylation-sensitive enzymes (*Eco*RI, *Ape*KI, *Hap*II, *Msp*I; Table 1), which may be effective in reducing representation of repetitive genomic regions. In silico prediction for designs utilizing methylation-sensitive enzymes will overestimate the number of predicted loci, and the extent of this overestimation will likely increase with the genomic fraction consisting of repetitive elements (e.g., Pan et al. 2015). Overall, the ability to alter complexity reduction gives RADseq great flexibility for genotyping designs for a wide array of research objectives across tree species with variable genome sizes and patterns of LD. Given the diversity of RADseq outcomes that are dependent upon experimental choices and structural attributes of the genomes under scrutiny, we suggest that researchers carefully match their questions to expected data outcomes. The growing availability of genomic resources across major clades of forest trees should increasingly allow researchers to explore outcomes in silico prior to protocol implementation.

Analytical approaches and considerations for RADseq data

Nearly all RADseq projects use Illumina instruments for sequencing, and data sets of hundreds of millions or billions of reads are the starting point for analyses. While past population genetic analysis tools were often implemented with graphical user interface (GUI) software, the larger size and complexity of RADseq and other genomic data sets require many analyses to be executed on high-performance computing systems or at least on server nodes

with large amounts of RAM. This necessitates that researchers are effective working from the command line in Unix or Linux environments, and have a working knowledge of a scripting language such as Perl or Python. The bioinformatic skills needed for responsible analysis of such data are significant and represent a transition in expertise for tree genetics investigators accustomed to molecular marker systems of the past. Building such expertise within research groups may require substantial time and personnel investment and can be an under-appreciated cost in constructing research budgets.

Utilization of reference genomes

One reason that RADseq datasets have proliferated in non-model trees is that the approach allows marker discovery and genotyping in the absence of genomic resources. Nonetheless, there are numerous advantages to utilizing reference genomes for the analysis and interpretation of RADseq data. First, alignment of reads, or consensus sequences from clustered genomic regions, to a reference genome allows filtering of contaminants from library preparation or from endosymbionts residing in plant tissue (e.g., bacterial and fungal endophytes). Second, reference genome alignment can reduce genotyping error caused by the misalignment of paralogous regions. Third, reference genomes allow the ordering of loci across scaffolds and the empirical measurement of LD decay, facilitating an understanding of how parameter estimates vary across genomes (e.g., Hohenlohe et al. 2012; Stölting et al. 2013; Ruegg et al. 2014) and providing context for interpreting genome scan or mapping analyses (e.g., Roesti et al. 2012; Epstein et al. 2016). Finally, as illustrated above, reference genomes can be used for in silico prediction for tailoring RADseq methods to research objectives.

Despite the advantages of processing RADseq data using reference genomes, the extent to which aligning to a reference genome can minimize genotyping bias is not always clear. Mapping raw reads to a reference genome can bias allele frequencies towards states found in the reference, reduce the number of called SNPs, or bias nucleotide diversity estimates downward (Pool et al. 2010), issues that will be exaggerated with more divergent reference genomes. Paris et al. (2017) suggest that building loci de novo before mapping consensus sequences to reference genomes may have the advantages of merging reads into biologically informed contigs and avoiding bias that may occur from aligning reads individually to reference genomes. In such cases, alignment of contig consensus sequences to reference genomes can still allow evaluation of potential alignment or genotyping error. Useful de novo assembly approaches have been developed, and several procedures have been proposed for optimizing de novo assembly of RADseq data (Mastretta-Yanes et al. 2015; Paris et al. 2017). A number of studies have compared de novo and reference-based approaches in terms of genotyping bias and effects on downstream analyses. Shafer et al. (2017) used RADseq data for Galapagos sea lions (*Zalophus wollebaeki*) to compare de novo and reference-based assembly with draft genomes of *Z. wollebaeki* and three related marine mammals. Reference-based approaches led to analytical results more consistent with expectations from simulations, but marker density decreased as the genetic distance to the reference increased. In similar analyses for *Betula nana* and two other *Betula* species, reference-based assemblies limited the over-assembly of paralogous regions, but de novo methods led to higher marker density (Wang et al. 2013). The benefits of reference genomes are less clear for phylogenetic data, where reads from multiple species are often aligned to a single reference. Fitz-Gibbon et al. (2017) compared maximum likelihood phylogenetic trees for California white oaks generated after de novo assembly and assembly to the *Quercus lobata* genome. Although these two approaches produced highly concordant phylogenies, further studies assessing different methods of alignment should be useful.

Due to a current lack of reference genomes for study species or their close relatives, most RADseq studies of forest trees to date have used de novo approaches (Table 1). Nonetheless, studies of *Populus*, *Eucaplyptus*, and a number of fruit tree species have utilized available high-quality reference genomes (e.g., Lindtke et al. 2014; Mori et al. 2017; Oueslati et al. 2017; Table 1). In many instances where reference genomes do not exist for the species under study, genomes of closely related taxa have been used (Table 1). Studies of population structure and hybridization in different *Populus* species have often utilized the *P. trichocarpa* genome (e.g., Stölting et al. 2013; Christe et al. 2016; Table 1), and fruit tree genomes have been used for related natural and domesticated varieties (e.g., Oueslati et al. 2017; Nagano et al. 2018). Although few in number and having assemblies characterized by large numbers of small scaffolds, draft conifer genomes could be useful for RADseq data from closely related species. For non-model trees with smaller genomes, even low-coverage shotgun sequencing could be used to generate reasonable contig assemblies to achieve some benefits of a reference genome. As costs and methods for genome assembly continue to improve, reference genomes should increasingly contribute to the analysis and interpretation of RADseq data.

Software pipelines for alignment and variant calling

Among the software pipelines designed specifically for assembly and variant calling with RADseq data, `Stacks` (Catchen et al. 2011, 2013) has been most widely used. Alternatives have more recently been developed, all of which share similar workflows but have unique capabilities (Table 3). Initial analysis steps involve removing barcodes,

Table 3 Examples of assembly and variant calling pipelines designed specifically for RADseq data

Pipeline	Notable characteristics
AftrRAD[a]	De novo assembly pipeline that can incorporate indel variation
	Evaluates loci for excess heterozygosity to remove paralogous loci
	Computationally efficient option if multicore computing resources are not available
dDocent[b]	Bash wrapper for assembly, variant calling, and filtering
	Can be implemented for any RADseq library protocol or sequence type
	Reference optional assembly that incorporates indel variation
	Bayesian model to account for genotype uncertainty (FreeBayes[k])
	Haplotype-based option for paralog identification
fastGBS[c]	Reference-based assembly pipeline that can incorporate indel variation
	Able to utilize sequencing reads of different lengths from different sequencing platforms
GBS-SNP-CROP[d]	Reference optional pipeline tailored for paired-end GBS datasets
	Designed for studies of crop species with few genomic resources, including polyploids
GibPSs[e]	De novo pipeline that can be implemented for any RADseq protocol or sequence type
	Capable of analyzing fully, partially, or non-overlapping reads in the same analysis
	Uses a distance-based network clustering approach to identify loci
	Incorporates three filters for the removal of paralogous loci
ipyrad[f]	De novo or reference-based assembly pipeline that can incorporate indel variation
	Designed for studies with greater phylogenetic breadth
	Filters paralogous loci based on the number of heterozygous states
RADIS[g]	Series of Perl scripts that utilize Stacks to perform most pipeline steps
	Designed for phylogenetics (analyses conducted with RAxML[l])
	Allows for rapid comparison of how alternative parameter settings affect tree topology
Stacks[h]	Most commonly used pipeline for de novo and reference-based assembly
	Detects PCR duplicates with paired-end sequencing of original RADseq libraries
	Calculates common population genetic summary statistics
TASSEL-GBS[i]	Designed for studies with exceptionally high numbers of individuals and loci
	Requires a reference genome (or a pseudo-reference)
TASSEL-UNEAK[j]	A network-based pipeline for de novo assemblies, designed to also handle polyploidy
	Currently supported in TASSEL v3.0[m], but not in later versions

[a] Sovic et al. (2015)
[b] Puritz et al. (2014a)
[c] Torkamaneh et al. (2017)
[d] Melo et al. (2016)
[e] Hapke and Thiele (2016)
[f] Eaton (2014)
[g] Cruaud et al. (2016)
[h] Catchen et al. (2011, 2013)
[i] Glaubitz et al. (2014)
[j] Lu et al. (2013)
[k] Garrison and Marth (2012)
[l] Stamatakis (2014)
[m] Bradbury et al. (2007)

associating reads with the correct sample, and filtering contaminants and low-quality reads. Assembly either begins by aligning reads to a reference genome or by clustering or assembling reads de novo. Most pipelines can utilize de novo and reference-based approaches, although some require reference genomes (e.g., fastGBS, TASSEL-GBS). Some can handle indels and SNPs (e.g., dDocent, ipyrad), while others disregard indels to speed computation (e.g., Stacks, GibPSs). Following mapping or clustering, variable sites are identified and genotypes are called or assigned likelihoods. Information on variant positions is commonly stored in Variant Call Format (VCF), which can be further filtered using

packages such as vcftools (Danecek et al. 2011). Stacks can additionally produce population genetic summary statistics and locus-specific statistical tests (Catchen et al. 2013), further contributing to its popularity. Other software commonly used in population genetic analyses can also be used for flexible assembly (e.g., bwa; Li and Durbin 2009) and variant calling (samtools, Li et al. 2009; GATK, McKenna et al. 2010). Although the above pipelines typically have easy to follow tutorials, investigators should have the bioinformatic expertise to thoroughly explore all aspects of data quality and to perform careful filtering of variants. It is worth noting that most of these methods require users to set coverage thresholds for calling genotypes, treat individual genotypes failing to meet such thresholds as missing data, and do not incorporate information on uncertainty (further discussed below).

Sources of genotyping error

PCR duplicates, when one allele is randomly over amplified with respect to another, artificially increase homozygosity (Davey et al. 2011; Puritz et al. 2014b) and can occur in any sequencing by synthesis approach that utilizes PCR for library preparation. The original RADseq allows the detection of PCR duplicates with paired-end sequencing because random shearing ensures reads from the fragmented ends will start and stop at different positions (Hohenlohe et al. 2013). Although other methods cannot directly detect PCR duplicates, laboratory modification can reduce PCR artefacts (Puritz et al. 2014b), and replicate libraries can be used to evaluate error introduced during laboratory preparation (e.g., Mastretta-Yanes et al. 2015). PCR duplicates, however, should not systematically affect allele frequency estimates, as they are not related to allelic variation. The mis-assembly of reads from paralogous or repetitive regions can also cause error by falsely identifying heterozygous genotypes, and this could be exacerbated in tree genomes with many instances of gene duplication and large repetitive fractions, especially those of conifers (Nystedt et al. 2013; Wegrzyn et al. 2014). This problem can be ameliorated with reference genome use, but bioinformatic methods also exist for filtering of loci with exceptionally high coverage, unbalanced read counts in heterozygotes, or abnormally high polymorphism (Table 3; Gayral et al. 2013; Hapke and Thiele 2016; Verdu et al. 2016).

Because RADseq samples genomes are based on restriction cut sites, the data are subject to complications arising from segregating mutations within cut sites. Allelic dropout (ADO) occurs when such mutation causes an allele to not be sequenced, leading to patterns of missing data which can bias allele frequency estimates and cause heterozygous genotypes to be erroneously identified as homozygous. ADO can bias estimates of genetic diversity downward, increase F_{ST} estimates, and lead to false positives in F_{ST} genome scans (Arnold et al. 2013; Gautier et al. 2013; Cariou et al. 2016). Investigators should be cautious about this source of bias for locus-specific parameter estimates for inferring selection and aware that RADseq is not ideal for comparing genetic diversity estimates across systems. As ADO is driven by polymorphism itself (groups with higher polymorphism are more likely to harbor restriction site mutations), taxa and genomic regions with higher polymorphism are more prone to the bias caused by ADO (Cariou et al. 2016; Cooke et al. 2016). Cariou et al. (2016) used simulations to illustrate that the bias has limited effect on diversity estimates for taxa with polymorphism rates below 2%, although with higher levels of polymorphism, ADO could substantially bias estimates of genetic diversity. ADO can affect some types of analyses more than others, but its influence can be minimized by removing loci for which reads do not occur in all (or most) samples (Davey et al. 2013; Puritz et al. 2014b). GBStools (Cooke et al. 2016) is capable of detecting and correcting for allelic dropout (based on differences in coverage among loci), although it relies on relatively high coverage depth. As some forest tree populations harbor high levels of genetic diversity, investigators working with such systems should be particularly aware of the influence ADO could have on parameter estimates, and make attempts to evaluate and minimize this influence. Additional discussion of ADO and other sources of RADseq genotyping bias can be found in Davey et al. (2013), Puritz et al. (2014b), Mastretta-Yanes et al. (2015), Fountain et al. (2016), and Andrews et al. (2016).

As with any high-throughput sequencing approach, bioinformatic processing of RADseq data can substantially influence the characteristics of retained loci and downstream analyses (Mastretta-Yanes et al. 2015; Rodríguez-Ezpeleta et al. 2016; Shafer et al. 2017). Careful filtering is essential for minimizing the influence of sequencing, alignment, and genotyping error, and may involve consideration of coverage depth, mapping and genotype quality, departures from HWE, counts of reads in heterozygotes, and other parameters. Loci with low minor allele frequencies (e.g., MAF < 0.03) are often filtered to guard against calling variants introduced from sequencing error. However, low-frequency variants can illustrate unique aspects of population history (e.g., Gompert et al. 2014), and their removal can adversely influence some population genetic analyses (Linck and Battey 2017), particularly those based on the site frequency spectrum (SFS). Shafer et al. (2017) compared parameter estimates generated from the same data using > 300 different combinations of genotyping pipelines and filtering parameters. They reported substantial variation in resulting summary statistics (e.g., π, Het_{obs}, F_{IS}), and especially inferred demographic parameters, highlighting the potential effect of assembly, variant calling, and filtering methods and parameters used with RADseq data. Yet, aside from the unique consequences of ADO, most of the challenges associated with variant calling from high-throughput sequencing approaches are not unique to

RADseq. As for any high-throughput sequencing data, the characteristics of raw data and the range of parameter values used for bioinformatic analyses should be carefully considered (Catchen et al. 2013; Mastretta-Yanes et al. 2015). Bioinformatic approaches and parameter choices should be carefully reported, as well as metrics associated with coverage depth and accepted levels of missing data in variants retained for analyses. Mastretta-Yanes et al. (2015) demonstrate the ability to detect genotyping errors and optimize `Stacks` using sequencing replicates, and Paris et al. (2017) illustrate a method using `Stacks` that does not require additional sequencing. Researchers should also consider approaches that incorporate statistical uncertainty into genotype inference and downstream analyses.

Incorporating genotype uncertainty

High-throughput sequencing data can be used to infer genotypes probabilistically based on read quality and coverage depth, with explicit statistical models (in contrast to previous genotyping methods for which implicit models and methods might have masked genotype uncertainty). As with any high-throughput sequencing method, RADseq data are characterized by stochastic variation in sequence coverage depth across individuals and loci. Statistical uncertainty in genotypes arises from sequencing and alignment errors, variation in coverage depth, and finite sampling of alleles from individuals. For many of the software pipelines in Table 3, genotypes are categorically called at loci with user-specified coverage thresholds. This fails to propagate information on genotype uncertainty, and results in large amounts of data being discarded (or treated as missing) that do not meet coverage thresholds but that could improve estimation of population-level parameters. Consequently, many authors have argued that genotype uncertainty should be modeled probabilistically and incorporated in statistical genetic methods, particularly for the analysis of low to medium coverage data (Nielsen et al. 2011; Buerkle and Gompert 2013; Fumagalli et al. 2013). Sequencing quality, alignment quality, and coverage depth can be incorporated into genotype likelihoods that include information on uncertainty, can be associated with multi-sample information, and can lead to more accurate genotype calls (Nielsen et al. 2011). Genotype likelihoods can be calculated for RADseq data with several commonly used programs (e.g., `samtools`, Li et al. 2009; `GATK`, McKenna et al. 2010; `ANGSD`, Korneliussen et al. 2014), and analyses conducted directly on the likelihoods can allow the downstream incorporation of uncertainty (e.g., Skotte et al. 2013; Vieira et al. 2013). `ANGSD` contains a suite of programs for estimating genotype likelihoods and conducting population genetic analyses with them.

Bayesian models have also been developed that account for coverage depth and quality during the estimation of genotype probabilities, and that improve parameter estimation for low-coverage sequencing data (e.g., Gompert et al. 2012, 2014; Nielsen et al. 2012; Fumagalli et al. 2013). These approaches are appealing because they can incorporate population-level priors (e.g., allele frequencies) into posterior estimates of genotype probabilities. Bayesian genotype probabilities can be used to call genotypes based on probability thresholds or can be directly used in downstream analyses, thereby incorporating genotype uncertainty. Recent studies have described and used such models to estimate allele frequencies and genotype probabilities (Gompert et al. 2012; Nielsen et al. 2012), to quantify population structure (Fumagalli et al. 2013), and to estimate admixture coefficients (Skotte et al. 2013; Gompert et al. 2014; Lindtke et al. 2014). These approaches allow investigators to avoid discarding large amounts of data that do not pass coverage depth thresholds set by many genotype calling approaches, and can result in a lower frequency of genotypes being treated as missing data.

Given a fixed budget, researchers face a trade-off among coverage depth, the number of individuals, and the number of sampled genomic regions. Although higher genotype certainty may be preferred for inferences relying on specific individual genotypes, sequencing larger numbers of individuals at lower depth can improve the estimation of population-level parameters. This is because all individuals contain information about a population, and allele frequencies, not genotypes per se, are used for many analyses. Similarly, because all loci contribute information about an individual, larger numbers of loci will improve the estimation of parameters for individuals (e.g., admixture coefficients, Skotte et al. 2013; Gompert et al. 2014; Lindtke et al. 2014). Indeed, studies using such models to investigate the trade-off between coverage depth and sample size have found that, for a fixed amount of sequencing, estimates of population and individual-level parameters have less bias and more precision with larger numbers of individuals and relatively low sequencing coverage (1-2X; Buerkle and Gompert 2013; Fumagalli et al. 2013). In practice, the optimal investment of sequencing among the number of individuals, the number of loci, and coverage depth will depend on the purpose of the work and the analytical methods to be employed. As research in tree genetics often benefits from sampling large numbers of individuals, low to medium coverage projects could often be preferable and would be best leveraged by the use of methods incorporating uncertainty.

Applications of RADseq to forest tree genetics

RADseq data have been generated for research on diverse tree species and for a variety of applied and basic objectives (Table 1). The number of publications using RADseq in tree genetics has increased sharply after early examples, with the majority occurring over the last 2 years (Table 1). This trend suggests that the use of RADseq will continue to increase, and

that it will be among the most common approaches for generating SNP data for tree genetics for the near future. Below we highlight several research areas where RADseq data are contributing to genetic analyses of trees and other non-model organisms.

Linkage mapping

Genetic linkage maps have long been used to describe the genomes of trees and to detect quantitative trait loci (QTL) associated with economically or ecologically important phenotypes (reviewed by Ritland et al. 2011). Conifers in particular are highly amenable to linkage mapping due to the availability of breeding populations, and because the haploid megagametophyte tissue of seeds directly captures the products of maternal meioses (Cairney and Pullman 2007). The increased number of loci that can be readily generated by RADseq is enabling the construction of denser linkage maps than previous approaches allowed, improving genomic resources that remain useful for many analyses in tree genetics. Linkage mapping with RADseq can be a useful alternative for characterizing genomes of species where whole genome sequencing is still cost prohibitive and challenging. Linkage maps can also improve genome assemblies, because they can be used for ordering of scaffolds assembled through de novo sequencing and assembly (Bartholomé et al. 2015; Fierst 2015). Higher density maps could be especially useful for improving assemblies for the ongoing conifer genome projects, which, due to massive genome sizes and high repeat content, are currently producing assemblies containing enormous numbers of unordered scaffolds (Nystedt et al. 2013; Neale et al. 2014; Stevens et al. 2016).

Linkage mapping has been the most common application of RADseq in trees to date (Table 1) and has resulted in substantially higher mapping densities than previous most marker systems allowed (e.g., Friedline et al. 2015; Mousavi et al. 2016; Fuentes-Utrilla et al. 2017). For example, Friedline et al. (2015) used RADseq data generated from megagametophytes of four maternal trees to map 20,655 contigs to unique positions along the 12 linkage groups of *Pinus balfouriana*, creating one of the densest genetic maps published for any tree species to date. Linkage maps generated with RADseq data have also been used to map QTL segregating with economically or ecologically important phenotypes in forest and fruit tree species (Bielenberg et al. 2015; Pootakham et al. 2015b; Fuentes-Utrilla et al. 2017). Higher density linkage maps should also be useful for characterizing genome-wide patterns of diversity, divergence, and LD in natural populations. For example, linkage maps with binned markers (e.g., Friedline et al. 2015) could be used to study patterns of LD within and among populations and to associate these patterns with other population genetic summaries (e.g., heterozygosity, SFS). This could improve analyses for a variety of questions outlined below and also provide resources to assess the efficacy of RADseq to cover and describe genomic variation.

Population genetic structure and history

The evolutionary histories of recently diverged populations and species are often hard to disentangle due to recent divergence, periods of ongoing gene flow, and the variable effect of evolutionary processes across genomes. Higher density SNP data generated with modern genotyping platforms have substantially improved resolution for understanding genetic differentiation across the landscape and across species boundaries. For non-model organisms and limited research budgets, RADseq data have proven useful for documenting fine-scale or cryptic patterns of population genetic structure that were undetectable with fewer loci (e.g., Larson et al. 2014; Benestan et al. 2015; Alter et al. 2017). Such data can also produce more precise estimates of ancestry, which can improve our understanding of how hybridization and introgression vary across genomes and populations (e.g., Caseys et al. 2012; Mandeville et al. 2015).

Features of tree biology such as high rates of outcrossing and long-distance pollen and seed dispersal often limit population genetic differentiation, and forest tree populations are often thought to be large and unstructured (Petit and Hampe 2006). More thorough resolution offered by higher density genotypic data may challenge this view (e.g., Slavov et al. 2012) and should increase our understanding of the extent to which historical, geographical, and environmental variation shape population structure. To date, surprisingly few studies have utilized RADseq to investigate landscape genetic structure in trees. Sun et al. (2016) were able to delineate populations of two varieties of *Castanopsis carlesii* (Fagaceae) in southeastern China, finding support for climatically driven divergence in the face of recurrent gene flow. Johnson et al. (2017a) employed ddRADseq to recover the postglacial colonization history of mountain hemlock (*Tsuga mertensiana*) on Alaska's Kenai Peninsula, finding support for subtle population structure in spite of high connectivity and long-distance dispersal. Their results suggest that mountain hemlock could respond to climate change via dispersal across elevation and latitude, highlighting how such perspective could inform conservation and/or management strategies.

Increased precision in ancestry estimates has also been demonstrated in studies applying RADseq to analyze admixture and reproductive isolation in hybrid zones. Zohren et al. (2016) documented unidirectional introgression from two diploid birch species into a third tetraploid species using RADseq, with improved resolution compared to past analyses of microsatellites. Additionally, a pair of studies estimated intra- and interspecific ancestry for hundreds of individuals across replicate *Populus* hybrid zones in Europe,

finding evidence that strong postzygotic selection against certain hybrid classes maintains reproductive isolation (Lindtke et al. 2014; Christe et al. 2016). In contrast, Menon et al. (2018) documented a history of divergence with gene flow and an abundance of advanced generation hybrids between *Pinus strobiformis* and *P. flexilis*, with admixture coefficients correlated with environmental variables. These patterns are consistent with a lack of strong isolating barriers and highlight the important role of extrinsic factors for evolutionary patterns within hybrid zones. Given the improved resolution RADseq data can have for describing genetic structure and admixture, such analyses of forest trees are likely to increase, especially since an understanding of population structure and admixture is often important for other analyses (e.g., GWAS, local adaptation).

In addition to characterizing patterns of contemporary population structure, RADseq datasets are facilitating analyses of historical demographic processes (e.g., Nice et al. 2013; O'Loughlin et al. 2014). Compared to traditional molecular marker systems, the greatly increased number of markers accessible with RADseq can allow more accurate evaluation of alternative demographic scenarios and facilitate parameter estimation (e.g., population expansion, migration, time; Robinson et al. 2014; Jeffries et al. 2016). RADseq data have been used with approaches that compare the observed SFS to alternatives generated with coalescent simulations (e.g., `fastsimcoal2`; Excoffier et al. 2013) or diffusion approximations (`δaδi`; Gutenkunst et al. 2009). Approximate Bayesian computation (ABC; Shafer et al. 2015; Elleouet and Aitken 2018) has also been used to analyze more complex models. Recent studies have used such approaches to evaluate models of demographic history in tree species (Eaton et al. 2015; Izuno et al. 2017; Menon et al. 2018), and such analyses could contribute perspective on historical processes underlying divergence and speciation as RADseq datasets accumulate. However, the estimation of such demographic parameters could in some cases be sensitive to individual sampling effort and bioinformatic approaches used for RADseq analyses (e.g., Elleouet and Aitken 2018), probably due to the difficulty of accurately quantifying the distribution of low-frequency variants in the SFS (Shafer et al. 2017). Although RADseq holds promise for demographic reconstruction, investigators should carefully consider a variety of genotyping and filtering approaches to assess robustness of results.

Molecular quantitative genetics

Research in forest genetics often benefits from quantitative genetic parameter estimates to understand geographic variation in local adaptation, to predict the phenotypic response to natural or artificial selection, and to understand how populations might respond to environmental change (Cornelius 1994; Savolainen et al. 2007; Lind et al. 2018). Traditionally, approaches for estimating quantitative genetic parameters, including heritability, have relied on controlled crosses, progeny designs within common gardens, or relatedness estimates from pedigrees in select natural populations (Falconer and Mackay 1996; Hill 2010). Common garden approaches for estimating quantitative genetic parameters have a long history in forest trees (reviewed in Morgenstern 1996; Lind et al. 2018), but require substantial time investment, and are often limited to phenotypes in the seedling phase. Moreover, as common gardens can have less environmental and more additive genetic variance, heritabilities estimated using progeny arrays can be elevated relative to natural populations (Conner et al. 2003; Castellanos et al. 2015). Such estimates can also be affected because the assumption that offspring are half-sibs can be violated in open pollination designs, and also do not account for variance in realized relatedness caused by segregation and recombination (Hill and Weir 2011). Alternatively, relatedness estimates from genetic markers can be used with phenotypic measurements to quantify heritability in natural populations (Andrew et al. 2005; Visscher et al. 2008; Gienapp et al. 2017).

Marker-based approaches for estimating relatedness in natural populations were suggested some time ago (Ritland 1996; Ritland 2000), but early applications suffered from low precision due to small marker numbers (Thomas et al. 2002; Coltman 2005; Csilléry et al. 2006). With higher density data, such as that generated with RADseq, realized genomic relatedness matrices can provide precise estimates of relatedness without the need for pedigree information. Heritability estimates from genomic data often agree with (Robinson et al. 2013; Bérénos et al. 2014; Gienapp et al. 2017), or are more precise than those estimated from pedigrees (El-Dien et al. 2016). For example, relatedness estimates based on 7338 SNPs allowed for separation of additive and non-additive factors and improved heritability estimates compared to a pedigree model for open-pollinated white spruce families (El-Dien et al. 2016). The number of loci needed for precise estimation of relatedness will vary by species, with larger numbers of loci needed for taxa with low LD and larger genomes (Wang 2016). Given sufficient genomic sampling, the ability to estimate the realized genomic relationship matrix directly means that quantitative genetic approaches can be applied in natural populations of trees, assuming phenotyping can also be effectively accomplished in populations under study. RADseq offers a cost-effective approach for supporting field-based estimation of quantitative genetic parameters in mature populations of forest trees, where estimates could be obtained for adult phenotypes in the natural populations in which they evolve. This could improve understanding of the evolutionary potential of populations and eventually even inform assisted migration (e.g., Aitken and Whitlock 2013).

Phenotypic selection has long been used in tree breeding programs, but has been difficult and expensive as a result of

long-life spans and late age at maturity. Thus, foresters have been greatly interested in the possible use of genotypic information to improve tree breeding. Although genome-wide association analyses have detected variants explaining variation in key phenotypes, markers have typically explained very small proportions of phenotypic variation (e.g., Eckert et al. 2009; Quesada et al. 2010), limiting their ability to guide tree breeding. In contrast to genome-wide association approaches, genomic selection simultaneously estimates effects over all markers to predict phenotype and estimate breeding values based on genome-wide estimates of relatedness among individuals (Grattapaglia and Resende 2011; Grattapaglia 2014). Genomic selection relies on precise and repeatable measures of phenotype, but further benefits from the highest marker density possible. Genetic and phenotypic data are analyzed with mixed models to estimate predictors of breeding value in a training population, before genotypic data alone are used with these models to estimate values in the breeding population. Many phenotypes of interest to tree breeders can be measured precisely, and emerging methods for high-throughput phenotyping could further improve capabilities (Desta and Ortiz 2014).

Over the last decade, animal breeders have found great success with the application of genomic selection for breeding value estimation (Daetwyler et al. 2010; Hayes and Goddard 2010; Meuwissen et al. 2016), and interest has grown in its application to crop and tree breeding programs (Grattapaglia and Resende 2011; Grattapaglia 2014). Genomic selection can be employed in any population for which precise and reproducible phenotypic measures and high-density markers can be obtained. It has been applied in a number of forest and fruit tree species where SNP chip data are available, and resulted in relatively high phenotypic prediction accuracies (Kumar et al. 2012; Resende et al. 2012a, 2012b, 2012c; Zapata-Valenzuela et al. 2012). More recently, RADseq approaches have been employed for genomic selection in trees, with promising results (El-Dien et al. 2015; Ratcliffe et al. 2015; Fuentes-Utrilla et al. 2017). Because RADseq data can be inexpensively and rapidly obtained for large numbers of trees, it could be valuable for genomic selection in non-model trees for which other genomic resources, such as high-density SNP chips, are not available or not cost-effective.

Phylogenetic inference

RADseq has been increasingly used for phylogenetic inference in groups where traditional sequencing approaches failed to resolve relationships (Cruaud et al. 2014; Darwell et al. 2016; Massatti et al. 2016). RADseq data are well suited for phylogenetic inference because it samples SNPs with a genome-wide distribution, often recovers extraordinarily large numbers of phylogenetically informative markers, and does not require prior genomic information on the species under study (Ree and Hipp 2015; Leaché and Oaks 2017). RADseq has excelled in producing well-resolved and highly supported phylogenies for young clades and adaptive radiations characterized by recent divergence (Wagner et al. 2013; Cruaud et al. 2014; Ebel et al. 2015; Darwell et al. 2016). In addition, analyses of both empirical and simulated data have illustrated that RADseq can be useful for deeper divergence (e.g., up to 60 Ma, Rubin et al. 2012; Cariou et al. 2013; Eaton et al. 2017). Although the retention of orthologous restriction sites across sampled groups will depend on divergence time, effective population size, and other factors, RADseq could be a powerful approach for phylogenetic analyses of tree species from many groups.

Despite this promise, there are aspects of RADseq data that cause issues for phylogenetic inference, including ADO and the recognition of paralogous loci. ADO will cause the systematic loss of shared loci between clades and will result in larger amounts of missing data for RADseq studies sampling more strongly divergent lineages, although these patterns of missing data can themselves be phylogenetically informative. While RADseq is typically viewed as more appropriate for younger groups that share orthologous restriction sites at sufficient rates (Rubin et al. 2012; Cariou et al. 2013), it has also been successful at resolving deeper relationships despite larger amounts of missing data arising from ADO (e.g., Eaton et al. 2017). The extent of information loss is affected by tree shape and taxonomic sampling breadth, both of which can be optimized with a well-designed sampling scheme (Eaton et al. 2017). Due to the form of the data, most studies concatenate RADseq loci rather than use multilocus methods to analyze them independently (Ree and Hipp 2015). As this ignores among locus variation in genealogy, evolutionary rate, and substitution patterns, problems with inference could arise from concatenation (Liu et al. 2015b). Alternatively, coalescent-based methods account for unique evolutionary history across loci but must rely on other simplifying assumptions in the process (Leaché and Oaks 2017). While future development of analytical approaches should improve understanding of best practices for utilizing RADseq data for phylogenetic inference, the positives of the approach are leading to a rapid increase in its use. `Stacks` and `ipyrad` have been most commonly used to process RADseq data for phylogenetic inference, and the latter was developed specifically for this purpose. `RADIS` (Cruaud et al. 2016) allows for efficient exploration of how assembly and variant calling parameters in `Stacks` affect tree topology. For detailed reviews on the analytical aspects of applying of RADseq data to phylogenetic inference see Ree and Hipp (2015) and Leaché and Oaks (2017).

RADseq data have increased phylogenetic resolution and resolved patterns of diversification in a wide variety of plant groups exhibiting recent divergence (Hou et al. 2015; Mort et al. 2015; Massatti et al. 2016), including trees (Liu et al.

2015a; Paun et al. 2016; Hamon et al. 2017). For example, RADseq data strongly resolved relationships of subgroups and species pairs in *Coffea*, a young genus that had been challenging due to shallow sequence divergence (Hamon et al. 2017). RADseq data has also improved the ability to resolve complicated phylogenetic histories within and among sections of oaks (*Quercus*), a group known for extensive hybridization (Grant 1981). For example, Hipp et al. (2014) resolved phylogenetic relationships within and among the white, red, and golden oaks (sections *Quercus*, *Lobatae*, and *Protobalanus*), producing a more highly resolved topology than past analyses. Such studies demonstrate that RADseq has potential to refine our understanding of the evolutionary histories of tree lineages, especially for young groups in which phylogenetic resolution has been elusive.

Genetic basis of phenotype and adaptation

Genome-wide data are needed for the discovery and characterization of adaptive genetic diversity within a forward genetics-based framework. This is because it is impossible to fully describe the genetic architecture of phenotype based on patterns of trait variation in controlled settings (e.g., a common garden). For example, heritability estimates do not quantify the number of loci underlying trait variation (Visscher et al. 2008), and studies linking genotypic to phenotypic variation or trying to establish the genetic architecture of adaptation do not have precise a priori expectations, even in well-studied systems, for the number of loci involved. Additionally, even if an a priori number of true positives could be estimated, there remains the issue of detectability, which is linked to the unknown values for the frequencies of causative alleles within natural populations (e.g., rare alleles with consequential effects are difficult to detect in realistic sample sizes). Thus, empirical genome-wide scans aimed at characterizing adaptive genetic diversity are optimized as interrogation of standing levels of genomic variation increases.

Genome scans for population differentiation outliers, as well as gene-environment associations, have been successfully utilized with RADseq data for multiple species. Inferences have ranged from the detection of genomic regions potentially involved in adaptation (e.g., Hohenlohe et al. 2010; Guo et al. 2016; Laporte et al. 2016) to association analyses linking phenotypic and genotypic variation (e.g., Comeault et al. 2014; Slavov et al. 2014; Brelsford et al. 2017). Despite the popularity of RADseq for such approaches, recent debate has identified concerns with naïve applications of the data to identify genetic architectures underlying fitness-related variation (Catchen et al. 2017; Lowry et al. 2017a, 2017b; McKinney et al. 2017). Much of this debate centers upon the fact that reduced representation approaches, such as RADseq, by definition sample limited fractions of genomes and are underpowered for detecting genetic regions involved in adaptation (Tiffin and Ross-Ibarra 2014; Hoban et al. 2016). The take home from this debate is that the ability of RADseq data to detect genetic architectures of adaptation depends upon a set of nuanced issues related to achievable marker density, unknown patterns of LD, and the true genetic architecture of adaptation.

As highlighted by Lowry et al. (2017a, 2017b), thorough characterization of the genetic architecture of adaptation will rarely be achievable with RADseq, as markers are often separated by physical distances that far exceed the average extent of LD. In addition, if it is unlikely to identify many adaptive variants a priori yet a large number are discovered, one potential explanation is that an unrecognized number of false positives are involved. Nonetheless, Catchen et al. (2017) and McKinney et al. (2017) highlight numerous cases where RADseq data were useful inputs for the detection of adaptive genetic variation. This includes examples where RADseq loci were implicated in replicate cases of parallel evolution (e.g., Hohenlohe et al. 2010; Gagnaire et al. 2013; McGee et al. 2016), tagged genomic regions containing genes previously implicated in adaptation (Hohenlohe et al. 2010; Nadeau et al. 2014), or identified loci explaining substantial percentages of phenotypic variation (Comeault et al. 2014; Brelsford et al. 2017). Thus, when RADseq can generate marker densities that effectively span the extent of LD, it could be quite useful for the initial detection of potentially adaptive variation. Also embedded within this debate are issues related to the unknown genetic architecture of adaptation relative to the size and complexity of genomes (reviewed by Lind et al. 2018 for trees), the evolutionary processes affecting genetic architectures (e.g., hard versus soft sweeps; Pritchard et al. 2010), and research goals (e.g., partial discovery versus complete description). The entanglement of unknown biological patterns with theoretical expectations that can only be tested with genome-wide data, of which RADseq is one cost-effective source, has thus created a research environment where investigators must carefully consider trade-offs and interactions among study design, interpretation of empirical results, biology of the system, and expectations from theory.

For forest trees, a long history of common garden and reciprocal transplant studies has demonstrated adaptation to local environments (reviewed by Savolainen et al. 2007, 2013; Lind et al. 2018). Before the advent of high-throughput sequencing, attempts to quantify the molecular genetic basis of adaptation often employed candidate gene approaches, where modest numbers of genes of known or hypothesized function were analyzed for association with phenotype or environment (e.g., Eckert et al. 2009, 2010; Quesada et al. 2010; Holliday et al. 2010). Given that the genetic architecture of adaptation in trees is often expected to be polygenic, RADseq offers a cost-effective avenue to generate genome-wide data for large numbers of samples without a priori declarations about the numbers and types of genetic regions underlying adaptive

variation. As detailed above, the success of scans based on these data depends on a number of nuanced quantities and considerations. LD has often been reported as very low in trees (e.g., Neale and Ingvarsson 2008; Sork et al. 2016), although recent resequencing work illustrates that LD may be higher than previously assumed and can vary substantially across tree genomes (e.g., Pyhäjärvi et al. 2011; Slavov et al. 2012; Silva-Junior and Grattapaglia 2015). RADseq methods with minimal complexity reduction can generate marker densities that span the extent of LD across small to moderately sized angiosperm genomes (Table 2), and have been able to detect genetic regions potentially involved in adaptation. For example, Stölting et al. (2013) applied ~ 38,000 SNPs (1 SNP per ~ 13 kb) to characterize introgression across hybridizing *Populus alba* and *P. tremula* populations and quantified heterogeneous divergence consistent with differential introgression of genomic regions involved in isolation or adaptive divergence. Pais et al. (2017), moreover, reported evidence for environmentally driven divergence at loci potentially involved in adaptation to abiotically divergent conditions in *Cornus florida* populations.

In contrast, genome scans using RADseq data have less often been applied to conifers, perhaps due to concerns that marker density and rapidly decaying LD (Pyhäjärvi et al. 2007; Neale and Ingvarsson 2008) could prevent identification of genetic architectures within their enormous and complex genomes. Maybe surprisingly so, RADseq data used for such analyses in multiple conifer species have yielded positive results consistent with identification of some portion of the genetic architecture underlying trait variation and adaptation. For example, Parchman et al. (2012) detected 11 SNPs that explained more than 50% of the phenotypic variation in serotiny across populations of *Pinus contorta*. Similarly, Lind et al. (2017) detected loci in *P. albicaulis* associated with a water availability gradient across the Lake Tahoe Basin and characterized signals of selection for these loci as being subtle and coordinated allele frequency shifts across populations.

For many non-model tree species, RADseq remains an economical and powerful avenue for producing high-density SNP data and will likely continue to be utilized for initial attempts at detecting genomic regions involved in adaptation. This is because whole genome resequencing for large numbers of samples has yet to become cost-effective for most tree species, and other reduced representation methods also have limitations and biases (Catchen et al. 2017; McKinney et al. 2017; discussed further below). Ultimately, as we gain a better understanding of LD in populations of forest trees (cf. Pyhäjärvi et al. 2011; Slavov et al. 2012; Silva-Junior and Grattapaglia 2015), our understanding for the utility of RADseq derived data for querying patterns of adaptation should increase. In the meantime, forest geneticists need to understand the limits of RADseq, carefully match expected patterns in RADseq data with expectations from theory, strive to provide estimates of LD when possible, and integrate across multiple lines of evidence (i.e., outliers from genome scans alone are limited evidence regardless of how markers were generated) when making inferences about adaptation (see Lowry et al. 2017b for useful guidelines for applying RADseq to analyze the genetic basis of adaptation).

Prospectus and conclusions

RADseq has rapidly facilitated the generation of high-density population genomic data for inference in many research areas of tree genetics. Nonetheless, choosing among alternative high-throughput methods requires consideration of research objectives, genomic resource availability, genome characteristics, and the number of samples to be analyzed. SNP genotyping arrays developed for trees with ample genomic resources will continue to be valuable because they produce clean and reproducible data, often in annotated coding regions, across populations and studies (e.g., Geraldes et al. 2013; Porth et al. 2013; McKown et al. 2014; Plomion et al. 2016). Although they do not maximize marker density, SNP arrays have been usefully applied for genome-wide association (e.g., Porth et al. 2013) and genomic selection (e.g., Kumar et al. 2012; El-Dien et al. 2016). Genome resequencing could provide ideal data for population genomics and is promising for analyses of adaptation in angiosperms with ample genomic resources. Resequencing studies in such trees have shed light on population genetic structure (e.g., Slavov et al. 2012), the genetic basis of adaptation (e.g., Evans et al. 2014), and genome-wide variation in LD (e.g., Wang et al. 2016). However, genome resequencing is currently out of reach for tree species with larger genomes and for research requiring large numbers of individuals. Targeted sequence capture is less limited by genome size, and targeted exon sequencing has been successfully applied for analyses of adaption in angiosperms and conifers (Holliday et al. 2016; Yeaman et al. 2016). However, targeted capture approaches, whether they focus on exons or genome variation more broadly, often require genomic resources and investment for designing capture probes that is not trivial (Jones and Good 2016). Both targeted exon capture and RNA sequencing can provide dense genotypic data in coding regions but are not ideally suited to characterizing genome-wide patterns of variation important to many population genetic inferences. In addition, investigations of genetic architecture of adaptation with such data often make the explicit assumption that causal variation resides largely within coding regions (Stern and Orgogozo 2008). This may not be true for species with large and complex genomes (Mei et al. 2018) and could

lead to biased representations of effect size distributions even when causal variation is partly genic.

RADseq is best distinguished from the above alternatives by its low cost, laboratory flexibility and simplicity, representative sampling of coding and non-coding variation, and applicability to any tree species. For non-model trees, the leap from traditional molecular marker systems to high-density RADseq data has allowed genome-wide perspectives that were far out of reach prior to its emergence. The diversity and flexibility of laboratory methods allow datasets to span a continuum of marker densities and sampling effort, ensuring that RADseq can be applied to a wide range of ecological, evolutionary, and applied issues. For some trees, RADseq studies may be designed to effectively cover the range of LD for analyses aimed at detecting selection, although limited marker densities will compromise its utility for such applications. Regardless, many research goals in tree genetics are less limited by incomplete genomic sampling. RADseq is well suited for phylogenetic inference and range-wide analyses of population genetic variation that can improve our understanding evolutionary history, guide conservation and management, and even inform analyses of the genetic basis of adaptation. It could also enable quantitative genetic approaches in natural populations and contribute to linkage mapping to enhance genomic resource development and the characterization of LD in non-model trees.

Investigators should understand the limits of RADseq and carefully weigh the costs and benefits of alternative approaches when matching genotyping methods with research goals and budgets. The increasing availability of reference genomes will enable alternative methods, but will also improve the design of RADseq studies, as well as bioinformatic processing and interpretation of analyses. Groups applying RADseq and other high-throughput approaches should strive to build necessary expertise for the responsible analysis of such data. When possible, investigators should explore in silico patterns prior to empirical assessment of populations, choose methods carefully to suit research objectives, report and summarize patterns of LD, and attempt to assess and report homology of RADseq loci to available genome resources. As few tree species have substantially developed genomic resources, and many are characterized by moderate to large genomes with considerable complexity, RADseq should remain among the most cost-effective methods for generating SNP data into the near future.

Acknowledgements We thank Santiago González-Martínez for inviting this review, and Chris Nice and C. Alex Buerkle for discussion and constructive comments on portions of the manuscript. During the writing of the manuscript, Thomas Parchman was supported by the National Science Foundation (DEB-1344250), Andrew Eckert was supported by the National Science Foundation (EF-1442486) and the United States Department of Agriculture (USDA 2016-67013-24469), and Kathryn Uckele was supported with a National Science Foundation Graduate Research Fellowship.

References

Aitken SN, Whitlock MC (2013) Assisted gene flow to facilitate local adaptation to climate change. Annu Rev Ecol Evol Syst 44:367–388

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. Evol Appl 1:95–111

Albert VA, Barbazuk WB, Der JP, Leebens-Mack J, Ma H, Palmer JD, Rounsley S, Sankoff D, Schuster SC, Soltis DE, Soltis PS et al (2013) The Amborella genome and the evolution of flowering plants. Science 342:1241089

Alberto FJ, Aitken SN, Alía R, González-Martínez SC, Hänninen H, Kremer A, Lefèvre F, Lenormand T, Yeaman S, Whetten R, Savolainen O (2013) Potential for evolutionary responses to climate change—evidence from tree populations. Glob Chang Biol 19: 1645–1661

Alter SE, Munshi-South J, Stiassny ML (2017) Genomewide SNP data reveal cryptic phylogeographic structure and microallopatric divergence in a rapids-adapted clade of cichlids from the Congo River. Mol Ecol 26:1401–1419

Andrew RL, Peakall R, Wallis IR, Wood JT, Knight EJ, Foley WJ (2005) Marker-based quantitative genetics in the wild? The heritability and genetic correlation of chemical defenses in *Eucalyptus*. Genetics 171:1989–1998

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22:3179–3190

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376

Bartholomé J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion JM (2015) High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. New Phytol 206:1283–1296

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). Mol Ecol 24:3299–3315

Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. Mol Ecol 23:3434–3451

Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, Salvi S, Jansen J, Viola R, Gut I et al (2014) Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus*× *domestica* Borkh). PLoS One 9:e110377

Bielenberg DG, Rauh B, Fan S, Gasic K, Abbott AG, Reighard GL, Okie WR, Wells CE (2015) Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) Batsch]. PLoS One 10:e0139406

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MMS, Keeling CI, Brand D, Vandervalk BP, Kirk H, Pandoh P, Moore RA, Zhao Y, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, MacKay J, Bohlmann J, Jones

SJM (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 29:2633–2635

Brelsford A, Toews DP, Irwin DE (2017) Admixture mapping in a hybrid zone reveals loci associated with avian feather coloration. Proc R Soc Lond B Biol Sci 284:20171106

Buerkle AC, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? Mol Ecol 22:3028–3035

Cairney J, Pullman GS (2007) The cellular and molecular biology of conifer embryogenesis. New Phytol 176:511–536

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol Evol 3:846–852

Cariou M, Duret L, Charlat S (2016) How and how much does RAD-seq bias genetic diversity estimates? BMC Evol Biol 16:240

Caseys C, Glauser G, Stölting KN, Christe C, Albrectsen BR, Lexer C (2012) Effects of interspecific recombination on functional traits in trees revealed by metabolomics and genotyping-by-resequencing. Plant Ecol Divers 5:457–471

Castellanos M, González-Rodríguez S, Pausas J (2015) Field heritability of a plant adaptation to fire in heterogeneous landscapes. Mol Ecol 24:5633–5642

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

Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140

Catchen JM, Hohenlohe PA, Bernatchez L, Funk WC, Andrews KR, Allendorf FW (2017) Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol Ecol Resour 17:362–365

Cavender-Bares J, González-Rodríguez A, Eaton DA, Hipp AA, Beulke A, Manos PS (2015) Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): a genomic and population genetics approach. Mol Ecol 24:3668–3687

Chafin TK, Martin BT, Mussmann SM, Douglas MR, Douglas ME (2017) FRAGMATIC: in silico locus prediction and its utility in optimizing ddRADseq projects. Conserv Genet Resour :1–4 https://doi.org/10.1007/s12686-017-0814-1

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA (2013) Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genet Genomes 9:1537–1544

Christe C, Stölting KN, Bresadola L, Fussi B, Heinze B, Wegmann D, Lexer C (2016) Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow. Mol Ecol 25:2482–2498

Coltman DW (2005) Testing marker-based estimates of heritability in the wild. Mol Ecol 14:2593–2599

Comeault AA, Soria-Carrasco V, Gompert Z, Farkas TE, Buerkle CA, Parchman TL, Nosil P (2014) Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in *Timema cristinae*. Am Nat 183:711–727

Conner JK, Franks R, Stewart C (2003) Expression of additive genetic variances and covariances for wild radish floral traits: comparison between field and greenhouse environments. Evolution 57:487–495

Cooke TF, Yee M, Muzzio M, Sockell A, Bell R, Cornejo OE, Kelley JL, Bailliet G, Bravi CM, Bustamante CD et al (2016) GBStools: a statistical method for estimating allelic dropout in reduced representation sequencing data. PLoS Genet 12:e1005631

Cornelius J (1994) Heritabilities and additive genetic coefficients of variation in forest trees. Can J For Res 24:372–379

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus JY (2014) Empirical assessment of RAD sequencing for interspecific phylogeny. Mol Biol Evol 31:1272–1274

Cruaud A, Gautier M, Rossi JP, Rasplus J-Y, Gouzy J (2016) RADIS: analysis of RAD-seq for interspecific phylogeny. Bioinformatics 32:3027–3028

Csilléry K, Johnson T, Beraldi D, Clutton-Brock T, Coltman D, Hansson B, Spong G, Pemberton JM (2006) Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. Genetics 173:2091–2101

Daetwyler HD, Hickey JM, Henshall JM, Dominik S, Gredler B, Van Der Werf JH, Hayes BJ (2010) Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim Prod Sci 50:1004–1010

Daïnou K, Blanc-Jolivet C, Degen B, Kimani P, Ndiade-Bourobou D, Donkpegan AS, Tosso F, Kaymak E, Bourland N, Doucet JL, Hardy OJ (2016) Revealing hidden species diversity in closely related species using nuclear SNPs, SSRs and DNA sequences—a case study in the tree genus *Milicia*. BMC Evol Biol 16:259

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Darwell CT, Rivers DM, Althoff DM (2016) RAD-seq phylogenomics recovers a well-resolved phylogeny of a rapid radiation of mutualistic and antagonistic yucca moths. Syst Entomol 41:672–682

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML (2013) Special features of RAD sequencing data: implications for genotyping. Mol Ecol 22:3151–3164

De Kort H, Mergeay J, Vander Mijnsbrugge K, Decocq G, Maccherini S, Bruun HHK, Honnay O, Vandepitte K (2014a) An evaluation of seed zone delineation using phenotypic and population genomic data on black alder *Alnus glutinosa*. J Appl Ecol 51:1218–1227

De Kort H, Vandepitte K, Bruun HH, Closset-Kopp D, Honnay O, Mergeay J (2014b) Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. Mol Ecol 23:4709–4721

De Kort H, Vandepitte K, Mergeay J, Honnay O (2014c) Isolation, characterization and genotyping of single nucleotide polymorphisms in the non-model tree species *Frangula alnus* (Rhamnaceae). Conserv Genet Resour 6:267–269

De Kort H, Vander Mijnsbrugge K, Vandepitte K, Mergeay J, Ovaskainen O, Honnay O (2016) Evolution, plasticity and evolving plasticity of phenology in the tree species *Alnus glutinosa*. J Evol Biol 29:253–264

Deng M, Jiang XL, Hipp AL, Manos PS, Hahn M (2018) Phylogeny and biogeography of East Asian evergreen oaks (*Quercus* section *Cyclobalanopsis*; Fagaceae): insights into the Cenozoic history of evergreen broad-leaved forests in subtropical Asia. Mol Phylogenet Evol 119:170–181

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601

Eaton DA (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849

Eaton DA, Hipp AL, González-Rodríguez A, Cavender-Bares J (2015) Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution 69:2587–2601

Eaton DA, Spriggs EL, Park B, Donoghue MJ (2017) Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst Biol 66:399–412

Ebel ER, DaCosta JM, Sorenson MD, Hill RI, Briscoe AD, Willmott KR, Mullen SP (2015) Rapid diversification associated with ecological

specialization in Neotropical *Adelpha* butterflies. Mol Ecol 24: 2392–2405

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, Clair JBS, Neale DB (2009) Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics 182:1289–1302

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185:969–982

Eckert AJ, Wegrzyn JL, Liechty JD, Lee JM, Cumbie WP, Davis JM, Goldfarb B, Loopstra CA, Palle SR, Quesada T, Langley CH (2013) The evolutionary genetics of the genes underlying phenotypic associations for loblolly pine (*Pinus taeda*, Pinaceae). Genetics 195: 1353–1372

El-Dien OG, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby YA (2015) Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 16:370

El-Dien OG, Ratcliffe B, Klápště J, Porth I, Chen C, El-Kassaby YA (2016) Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. Genes Genom Genet 6:743–753

Elleouet JS, Aitken SN (2018) Exploring approximate Bayesian computation for inferring recent demographic history with genomic markers in nonmodel species. Mol Ecol Resour 18:525–540

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Epstein B, Jones M, Hamede R, Hendricks S, McCallum H, Murchison EP, Schönfeld B, Wiench C, Hohenlohe P, Storfer A (2016) Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Nat Commun 7:12684

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: Walker JM (ed) Molecular methods for evolutionary genetics. Humana Press, Clifton, pp 157–178

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J et al (2014) Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nat Genet 46:1089–1096

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genet 9:e1003905

Falconer DS, Mackay TF (1996) Introduction to quantitative genetics, 4th edn. Longman, New York

Fierst JL (2015) Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools. Front Genet 6:1–8

Fitz-Gibbon S, Hipp AL, Pham KK, Manos PS, Sork V (2017) Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* subgenus *Quercus*). Genome 60: 1–13

Fountain ED, Pauli JN, Reid BN, Palsbøll PJ, Peery MZ (2016) Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol Ecol Resour 16:966–978

Friedline CJ, Lind BM, Hobson EM, Harwood DE, Mix AD, Maloney PE, Eckert AJ (2015) The genetic architecture of local adaptation I: the genomic landscape of foxtail pine (*Pinus balfouriana* Grev. & Balf.) as revealed from a high-density linkage map. Tree Genet Genomes 11:1–15

Fuentes-Utrilla P, Goswami C, Cottrell J, Pong-Wong R, Law A, A'Hara S, Lee S, Woolliams J (2017) QTL analysis and genomic selection using RADseq derived markers in Sitka spruce: the potential utility of within family data. Tree Genet Genomes 13:1–12

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R (2013) Quantifying population genetic differentiation from next-generation sequencing data. Genetics 195: 979–992

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L (2013) The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution 67:2483–2497

Gailing O, Staton ME, Lane T, Schlarbaum SE, Nipper R, Owusu SA, Carlson JE (2017) Construction of a framework genetic linkage map in *Gleditsia triacanthos* L. Plant Mol Biol Report 35:177–187

Gardner KM, Brown P, Cooke TF, Cann S, Costa F, Bustamante C, Velasco R, Troggio M, Myles S (2014) Fast and cost-effective genetic mapping in apple using next-generation sequencing. Genes Genom Genet 4:1681–1687

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet JM, Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol 22:3165–3178

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, Belkhir K, Cahais V, Loire E, Bernard A, Galtier N (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. PLoS Genet 9:e1003457

Geraldes A, DiFazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N et al (2013) A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. Mol Ecol Resour 13:306–323

Gienapp P, Fior S, Guillaume F, Lasky JR, Sork VL, Csilléry K (2017) Genomic quantitative genetics to study evolution in the wild. Trends Ecol Evol 32:897–908

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e90346

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

Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Mol Ecol 23:4555–4573

González-Martínez SC, Krutovsky KV, Neale DB (2006) Forest-tree population genomics and adaptive evolution. New Phytol 170:227–238

Grant V (1981) Plant speciation. Columbia University Press, New York

Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the way forward. In: Tuberosa R, Graner A, Frison E (eds) Genomics of plant genetic resources. Springer, New York, pp 651–682

Grattapaglia D, Resende MD (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255

Guajardo V, Solís S, Sagredo B, Gainza F, Muñoz C, Gasic K, Hinrichsen P (2015) Construction of high density sweet cherry (*Prunus avium* L.) linkage maps using microsatellite markers and SNPs detected by genotyping-by-sequencing (GBS). PLoS One 10:e0127750

Guo F, Yu H, Tang Z, Jiang X, Wang L, Wang X, Xu Q, Deng X (2015) Construction of a SNP-based high-density genetic map for pummelo using RAD sequencing. Tree Genet Genomes 11:1–11

Guo B, Li Z, Merilä J (2016) Population genomic evidence for adaptive differentiation in the Baltic Sea herring. Mol Ecol 25:2833–2852

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695

Hamon P, Grover CE, Davis AP, Rakotomalala JJ, Raharimalala NE, Albert VA, Sreenath HL, Stoffelen P, Mitchell SE, Couturon E, Hamon S, de Kochko A, Crouzillat D, Rigoreau M, Sumirat U, Akaffou S, Guyot R (2017) Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (*Coffea*) and insights into the evolution of caffeine content in its species. Mol Phylogenet Evol 109:351–361

Hapke A, Thiele D (2016) GIbPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. Mol Ecol Resour 16:979–990

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53:876–883

He T, D'Agui H, Lim SL, Enright NJ, Luo Y (2016) Evolutionary potential and adaptation of *Banksia attenuata* (Proteaceae) to climate and fire regime in southwestern Australia, a global biodiversity hotspot. Sci Rep 6:26315

Herrera S, Reyes-Herrera PH, Shank TM (2015) Predicting RAD-seq marker numbers across the eukaryotic tree of life. Genome Biol Evol 7:3207–3225

Hill WG (2010) Understanding and using quantitative genetic variation. Philos Trans R Soc Lond Ser B Biol Sci 365:73–85

Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res 93:47–64

Hipp AL, Manos PS, Cavender-Bares J, Eaton D, Nipper R (2013) Using phylogenomics to infer the evolutionary history of oaks. Int Oak J 24:61–71

Hipp AL, Eaton DA, Cavender-Bares J, Fitzek E, Nipper R, Manos PS (2014) A framework phylogeny of the American oak clade based on sequenced RAD data. PLoS One 9:e93975

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat 188:379–397

Hodel RGJ, Chen S, Payton AC, McDaniel SF, Soltis P, Soltis DE (2017) Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. Sci Rep 7:17598

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 Genet 6:e1000862

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philos Trans R Soc B 367:395–408

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Mol Ecol 22:3002–3013

Holliday JA, Ritland K, Aitken SN (2010) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytol 188:501–514

Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW (2016) Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in *Populus trichocarpa*. New Phytol 209:1240–1251

Hou Y, Nowak MD, Mirré V, Bjorå CS, Brochmann C, Popp M (2015) Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus *Diapensia* (Diapensiaceae). PLoS One 10:e0140175

Imai A, Yoshioka T, Hayashi T (2017) Quantitative trait locus (QTL) analysis of fruit-quality traits for mandarin breeding in Japan. Tree Genet Genomes 13:79

Ingvarsson PK, Hvidsten TR, Street NR (2016) Towards integration of population and comparative genomics in forest trees. New Phytol 212:338–344

Izuno A, Kitayama K, Onoda Y, Tsujii Y, Hatakeyama M, Nagano AJ, Honjo MN, Shimizu-Inatsugi R, Kudoh H, Shimizu KK, Isagi Y (2017) The population genomic signature of environmental association and gene flow in an ecologically divergent tree species *Metrosideros polymorpha* (Myrtaceae). Mol Ecol 26:1515–1532

Jeffries DL, Copp GH, Handley LL, Olsén KH, Sayer CD, Hänfling B (2016) Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. Mol Ecol 25:2997–3018

Johnson JS, Gaddis KD, Cairns DM, Konganti K, Krutovsky KV (2017a) Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. Am J Bot 104: 439–450

Johnson JS, Gaddis KD, Cairns DM, Krutovsky KV (2017b) Seed dispersal at alpine treeline: an assessment of seed movement within the alpine treeline ecotone. Ecosphere 8:e01649

Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. Mol Ecol 25:185–202

Karam MJ, Lefévre F, Dagher-Kharrat MB, Pinosio S, Vendramin G (2015) Genomic exploration and molecular marker development in a large and complex conifer genome using RADseq and mRNAseq. Mol Ecol Resour 15:601–612

Konar A, Choudhury O, Bullis R, Fiedler L, Kruser JM, Stephens MT, Gailing O, Schlarbaum S, Coggeshall MV, Staton ME, Carlson JE, Emrich S, Romero-Severson J (2017) High-quality genetic mapping with ddRADseq in the non-model tree *Quercus rubra*. BMC Genomics 18:417

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15:1

Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C, Carlisle C (2012) Genomic selection for fruit quality traits in apple (*Malus x domestica* Borkh.). PLoS One 7:e36674

Laporte M, Pavey SA, Rougeux C, Pierron F, Lauzent M, Budzinski H, Labadie P, Geneste E, Couture P, Baudrimont M, Bernatchez L (2016) RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic eels. Mol Ecol 25:219–237

Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evol Appl 7:355–369

Leaché AD, Oaks JR (2017) The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu Rev Ecol Evol Syst 48: 69–84

Lepais O, Weir JT (2014) SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol Ecol Resour 14:1314–1321

Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25:1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Linck EB, Battey CJ (2017) Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. bioRxiv p188623 doi: https://doi.org/10.1101/188623

Lind BM, Friedline CJ, Wegrzyn JL, Maloney PE, Vogler DR, Neale DB, Eckert AJ (2017) Water availability drives signatures of local adaptation in whitebark pine (*Pinus albicaulis* Engelm.) across fine spatial scales of the Lake Tahoe Basin, USA. Mol Ecol 26:3168–3185

Lind BM, Menon M, Bolte CE, Faske TM, Eckert AJ (2018) The genomics of local adaptation in trees: are we out of the woods yet? Tree Genet Genomes 14:29

Lindtke D, Gompert Z, Lexer C, Buerkle CA (2014) Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. Mol Ecol 23: 4316–4330

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 132:43–66

Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61

Liu L, Jin X, Chen N, Li X, Li P, Fu C (2015a) Phylogeny of *Morella rubra* and its relatives (Myricaceae) and genetic resources of Chinese bayberry using RAD sequencing. PLoS One 10:e0139840

Liu L, Xi Z, Wu S, Davis CC, Edwards SV (2015b) Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci 1360:36–53

Liu CY, Li DW, Zhou JH, Zhang Q, Tian H, Yao XH (2017) Construction of a SNP-based genetic linkage map for kiwifruit using next-generation restriction-site-associated DNA sequencing (RADseq). Mol Breed 37:139

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A (2017a) Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour 17:142–152

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A (2017b) Responsible RAD: striving for best practices in population genomic studies of adaptation. Mol Ecol Resour 17:366–369

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9:e1003215

Lu M, Krutovsky KV, Nelson CD, Koralewski TE, Byram TD, Loopstra CA (2016) Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics 17: 730

Mandeville EG, Parchman TL, McDonald DB, Buerkle CA (2015) Highly variable reproductive isolation among pairs of *Catostomus* species. Mol Ecol 24:1856–1872

Marchese A et al (2016) The first high-density sequence characterized SNP-based linkage map of olive (*Olea europaea* L. subsp. *europaea*) developed using genotyping by sequencing. Aust J Crop Sci 10:857–863

Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem 6:287–303

Massatti R, Reznicek AA, Knowles LL (2016) Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: a case study in *Carex* sect. *Racemosae*. Am J Bot 103:337–347

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

McGee MD, Neches RY, Seehausen O (2016) Evaluating genomic divergence and parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. Mol Ecol 25:260–268

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

McKinney GJ, Larson WA, Seeb LW, Seeb JE (2017) RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on breaking RAD by Lowry et al. (2016). Mol Ecol Resour 17:356–361

McKown AD, Guy RD, Klápště J, Geraldes A, Friedmann M, Cronk QC, El-Kassaby YA, Mansfield SD, Douglas CJ (2014) Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. New Phytol 201:1263–1276

McVay JD, Hipp AL, Manos PS (2017) A genetic legacy of introgression confounds phylogeny and biogeography in oaks. Proc R Soc Lond B 284:20170300

Mei W, Stetter MG, Gates DJ, Stitzer MC, Ross-Ibarra J (2018) Adaptation in plant genomes: bigger is different. Am J Bot 105: 16–19

Melo AT, Bartaula R, Hale I (2016) GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinformatics 17:1

Menon M, Bagley JC, Friedline CJ, Whipple AV, Schoettle AW, Leal-Saenz A, Wehenkel C, Molina-Freaner F, Flores-Renteria L, Gonzalez-Elizondo MS, et al. (2018) The role of hybridization during ecological divergence of southwestern white pine (*Pinus strobiformis*) and limber pine (*P. flexilis*). Mol Ecol 27:1245–1260

Meuwissen T, Hayes B, Goddard M (2016) Genomic selection: a paradigm shift in animal breeding. Anim Front 6:6–14

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–248

Mora-Márquez F, García-Olivares V, Emerson B, López de Heredia U (2017) ddradseqtools: a software package for in silico simulation and testing of double-digest RADseq experiments. Mol Ecol Resour 17:230–246

Morgenstern EK (1996) Geographic variation in forest trees: genetic basis and application of knowledge in silviculture. UBC Press, Vancouver

Mori K, Shirasawa K, Nogata H, Hirata C, Tashiro K, Habu T, Kim S, Himeno S, Kuhara S, Ikegami H (2017) Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (*Ficus carica* L.). Sci Rep 7:41124

Mort ME, Crawford DJ, Kelly JK, Santos-Guerra A, de Sequeira MM, Moura M, Caujapé-Castells J (2015) Multiplexed-shotgun-genotyping data resolve phylogeny within a very recently derived insular lineage. Am J Bot 102:634–641

Mousavi M, Tong C, Liu F, Tao S, Wu J, Li H, Shi J (2016) De novo SNP discovery and genetic linkage mapping in poplar using restriction site associated DNA and whole-genome sequencing technologies. BMC Genomics 17:656

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, Hussey SG, Pinard D, van der Merwe K, Singh P, van Jaarsveld I, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye CY, Li T, Sterck L, Vanneste K, Murat F, Soler M, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Pappas GJ, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, van de Peer Y, Rokhsar DS, Schmutz J (2014) The genome of *Eucalyptus grandis*. Nature 510:356–362

Nadeau NJ, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, Morrison A, McMillan WO, Jiggins CD, Papa R (2014) Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome Res 24:1316–1333

Nagano Y, Mimura T, Kotoda N, Matsumoto R, Nagano AJ, Honjo MN, Kudoh H, Yamamoto M (2018) Phylogenetic relationships of Aurantioideae (Rutaceae) based on RAD-Seq. Tree Genet Genomes 14:6

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol 22:2841–2847

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11: 149–155

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111–122

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JFD, Lorenz W, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, deJong PJ, Yorke JA, Salzberg SL, Langley CH (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59

Neale DB, McGuire PE, Wheeler NC, Stevens KA, Crepeau MW, Cardeno C, Zimin AV, Puiu D, Pertea GM, Sezen UU et al (2017) The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. Genes Genom Genet 7:3157–3167

Nice CC, Gompert Z, Fordyce JA, Forister ML, Lucas LK, Buerkle CA (2013) Hybrid speciation and independent evolution in lineages of alpine butterflies. Evolution 67:1055–1068

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443–451

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One 7:e37558

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584

O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, Lomas S, Burt A (2014) Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. Mol Biol Evol 31:889–902

Ortego J, Gugger PF, Sork VL (2017) Genomic data reveal cryptic lineage diversification and introgression in Californian golden cup oaks (section *Protobalanus*). New Phytol 218:804–818

Oueslati A, Slhi-Hannachi A, Luro F, Vignes H, Mournet P, Ollitrault P (2017) Genotyping by sequencing reveals the interspecific *C. maxima*/*C. reticulata* admixture along the genomes of modern citrus varieties of mandarins, tangors, tangelos, orangelos and grapefruits. PLoS One 12:e0185618

Pais AL, Whetten RW, Xiang QYJ (2017) Ecological genomics of local adaptation in *Cornus florida* L. by genotyping by sequencing. Ecol and Evol 7:441–465

Pan J, Wang B, Pei ZY, Zhao W, Gao J, Mao JF, Wang XR (2015) Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Mol Ecol Resour 15:711–722

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle C (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol 21:2991–3005

Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. Methods Ecol Evol 8:1360–1373

Paun O, Turner B, Trucchi E, Munzinger J, Chase MW, Samuel R (2016) Processes driving the adaptive radiation of a tropical tree (*Diospyros*, Ebenaceae) in New Caledonia, a biodiversity hotspot. Syst Biol 65:212–227

Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B, Pelgas B, Deslauriers M, Clément S, Lavigne P, Lamothe M, Cooke JEK, Jaramillo-Correa JP, Beaulieu J, Isabel N, Mackay J, Bousquet J (2013) Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. Mol Ecol Resour 13:324–336

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7: e37135

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. Annu Rev Ecol Syst 37:187–214

Pham KK, Hipp AL, Manos PS, Cronn RC (2017) A time and a place for everything: phylogenetic history and geography as joint predictors of oak plastome phylogeny. Genome 60:720–732

Pinosio S, González-Martínez S, Bagnoli F, Cattonaro F, Grivet D, Marroni F, Lorenzo Z, Pausas J, Verdú M, Vendramin G (2014) First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. Mol Ecol Resour 14:846–856

Plomion C, Bartholomé J, Lesur I, Boury C, Rodríguez-Quilón I, Lagraulet H, Ehrenmann F, Bouffier L, Gion JM, Grivet D, de Miguel M, de María N, Cervera MT, Bagnoli F, Isik F, Vendramin GG, González-Martínez SC (2016) High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). Mol Ecol Resour 16:574–587

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genome Res 20:291–300

Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrung S, Tangphatsornruang S (2015a) Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics 105:288–295

Pootakham W, Ruang-Areerate P, Jomchai N, Sonthirod C, Sangsrakru D, Yoocha T, Theerawattanasuk K, Nirapathpongporn K, Romruensukharom P, Tragoonrung S et al (2015b) Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). Front Plant Sci 6:1–12

Porth I, Klapšte J, Skyba O, Hannemann J, McKown AD, Guy RD, DiFazio SP, Muchero W, Ranjan P, Tuskan GA et al (2013) Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. New Phytol 200:710–726

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:208–215

Puritz JB, Lotterhos KE (2017) Expressed Exome Capture Sequencing (EecSeq): a method for cost-effective exome sequencing for all organisms with or without genomic resources. bioRxiv:223735. https://doi.org/10.1101/223735

Puritz JB, Hollenbeck CM, Gold JR (2014a) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2:e431

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014b) Demystifying the RAD fad. Mol Ecol 23:5937–5942

Pyhäjärvi T, Garcia-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177:1713–1724

Pyhäjärvi T, Kujala ST, Savolainen O (2011) Revisiting protein heterozygosity in plants—nucleotide diversity in allozyme coding genes of conifer *Pinus sylvestris*. Tree Genet Genomes 7:385–397

Quesada T, Gopal V, Cumbie WP, Eckert AJ, Wegrzyn JL, Neale DB, Goldfarb B, Huber DA, Casella G, Davis JM (2010) Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). Genetics 186:677–686

Ratcliffe B, El-Dien OG, Klápště J, Porth I, Chen C, Jaquish B, El-Kassaby Y (2015) A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. Heredity 115:547–555

Ree RH, Hipp AL (2015) Inferring phylogenetic history from restriction site associated DNA (RADseq). In: Hörandl E, Appelhans MS (eds) Next-Generation Sequencing in Plant Systematics. Koeltz Scientific Books, pp 181–204

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA et al (2012a) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128

Resende MFR, Muñoz P, Acosta JJ, Peter G, Davis JM, Grattapaglia D, Resende MDV, Kirst M (2012b) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol 193:617–624

Resende MFR, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012c) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503–1510

Ritland K (1996) A marker-based method for inferences about quantitative inheritance in natural populations. Evolution 50:1062–1073

Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. Mol Ecol 9:1195–1204

Ritland K, Krutovsky KV, Tsumura Y, Pelgas B, Isabel N, Bousquet J (2011) Genetic mapping in conifers. In: Plomion C, Bousquet J, Kole C (eds) Genetics, genomics and breeding of conifers. Edenbridge Science Publishers and CRC Press, pp 196–238

Robinson MR, Santure AW, DeCauwer I, Sheldon BC, Slate J (2013) Partitioning of genetic variation across the genome using multimarker methods in a wild bird population. Mol Ecol 22: 3963–3980

Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN (2014) Sampling strategies for frequency spectrum-based population genomic inference. BMC Evol Biol 14:254

Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, Álvarez P, Cotano U, Irigoien X (2016) Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. Mol Ecol Resour 16: 991–1001

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. Mol Ecol 21:2852–2862

Rubin BE, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. PLoS One 7:e33394

Ruegg K, Anderson EC, Boone J, Pouls J, Smith TB (2014) A role for migration-linked genes and genomic islands in divergence of a songbird. Mol Ecol 23:4757–4769

Sakaguchi S, Sugino T, Yoshihiko T, Ito M, Crisp MD, Bowman DMJS, Nagano AJ, Honjo MN, Yasugi M, Kudoh H et al (2015) High-throughput linkage mapping of Australian white cypress pine (*Callitris glaucophylla*) and map transferability to related species. Tree Genet Genomes 11:1–12

Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. Annu Rev Ecol Evol Syst 38:595–619

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. Nat Rev Genet 14:807–820

Scaglione D, Fornasiero A, Pinto C, Cattonaro F, Spadotto A, Infante R, Meneses C, Messina R, Lain O, Cipriani G et al (2015) A RAD-based linkage map of kiwifruit (*Actinidia chinensis* Pl.) as a tool to improve the genome assembly and to scan the genomic region of the gender determinant for the marker-assisted breeding. Tree Genet Genomes 11:1–10

Shafer ABA, Gattepaille LM, Stewart REA, Wolf JBW (2015) Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. Mol Ecol 24:328–345

Shafer A, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JB (2017) Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol 8:907–917

Silva-Junior OB, Grattapaglia D (2015) Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New Phytol 208:830–845

Silva-Junior OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. New Phytol 206:1527–1540

Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. Genetics 195:693–702

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 Phytol 196:713–725

Slavov GT, Nipper R, Robson P, Farrar K, Allison GG, Bosch M, Clifton-Brown JC, Donnison IS, Jensen E (2014) Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*. New Phytol 201:1227–1239

Sork V, Aitken S, Dyer R, Eckert A, Legendre P, Neale D (2013) Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genet Genomes 9:901–911

Sork VL, Fitz-Gibbon ST, Puiu D, Crepeau M, Gugger PF, Sherman R, Stevens K, Langley CH, Pellegrini M, Salzberg SL (2016) First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née (Fagaceae). Genes Genom Genet 6:3485–3495

Sovic MG, Fries AC, Gibbs HL (2015) AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. Mol Ecol Resour 15:1163–1171

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62:2155–2177

Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, Cardeno C, Paul R, Gonzalez-Ibeas D, Koriabine M, Holtz-Morris AE, Martínez-García PJ, Sezen UU, Marçais G, Jermstad K, McGuire PE, Loopstra CA, Davis JM, Eckert A, de Jong P, Yorke JA, Salzberg SL, Neale DB, Langley CH (2016) Sequence of the sugar pine megagenome. Genetics 204:1613–1626

Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. Mol Ecol 22:842–855

Sun Y, Surget-Groba Y, Gao S (2016) Divergence maintained by climatic selection despite recurrent gene flow: a case study of *Castanopsis carlesii* (Fagaceae). Mol Ecol 25:4580–4592

Thomas SC, Coltman DW, Pemberton JM (2002) The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. J Evol Biol 15:92–99

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. Trends Ecol Evol 29:673–680

Tong C, Li H, Wang Y, Li X, Ou J, Wang D, Xu H, Ma C, Lang X, Liu G et al (2016) Construction of high-density linkage maps of *Populus deltoides* × *P. simonii* using restriction-site associated DNA sequencing. PLoS One 11:e0150692

Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F (2017) Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18:5

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

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M et al (2010) The genome of the domesticated apple (*Malus domestica* Borkh.). Nat Genet 42:833–839

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F et al (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45:487–494

Verdu CF, Guichoux E, Quevauvillers S, De Thier O, Laizet Y, Delcamp A, Gévaudant F, Monty A, Porté AJ, Lejeune P et al (2016) Dealing with paralogy in RADseq data: in silico detection and single nucleotide polymorphism validation in *Robinia pseudoacacia* L. Ecol Evol 6:7323–7333

Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. Genome Res 23:1852–1861

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet 9:255–266

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol Ecol 22:787–798

Wang J (2016) Pedigrees or markers: which are better in estimating relatedness and inbreeding coefficient? Theor Popul Biol 107:4–13

Wang N, Thomson M, Bodles WJ, Crawford RM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJ (2013) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Mol Ecol 22:3098–3111

Wang J, Street NR, Scofield DG, Ingvarsson PK (2016) Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. Genetics 202:1185–1200

Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ et al (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196:891–909

Yeaman S, Hodgins KA, Suren H, Nurkowski KA, Rieseberg LH, Holliday JA, Aitken SN (2014) Conservation and divergence of gene expression plasticity following c. 140 million years of evolution in lodgepole pine (*Pinus contorta*) and interior spruce (*Picea glauca* × *Picea engelmannii*). New Phytol 203:578–591

Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, Nurkowski KA, Smets P, Wang T, Gray LK (2016) Convergent local adaptation to climate in distantly related conifers. Science 353:1431–1433

Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R (2012) SNP markers trace familial linkages in a cloned population of *Pinus taeda*—prospects for genomic selection. Tree Genet Genomes 8:1307–1318

Zhang Z, Wei T, Zhong Y, Li X, Huang J (2016) Construction of a high-density genetic map of *Ziziphus jujuba*. Tree Genet Genomes 12:1–10

Zhao J, Jian J, Liu G, Wang J, Lin M, Ming Y, Liu Z, Chen Y, Liu X, Liu M (2014) Rapid SNP discovery and a RAD-based high-density linkage map in jujube (*Ziziphus* Mill.). PLoS One 9:e109850

Zhigunov AV, Ulianich PS, Lebedeva MV, Chang PL, Nuzhdin SV, Potokina EK (2017) Development of F1 hybrid population and the high-density linkage map for European aspen (*Populus tremula* L.) using RADseq technology. BMC Plant Biol 17:180

Zhou L, Bawa R, Holliday JA (2014) Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood (*Populus trichocarpa*). Mol Ecol 23:2486–2499

Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, Nichols RA, Buggs RJ (2016) Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. Mol Ecol 25:2413–2426