American Journal of Botany

American Journal of Botany 99(2): 291–311. 2012.

TARGETED ENRICHMENT STRATEGIES FOR NEXT-GENERATION PLANT BIOLOGY[1]

RICHARD CRONN[2,6], BRIAN J. KNAUS[2], AARON LISTON[3], PETER J. MAUGHAN[4], MATTHEW PARKS[3], JOHN V. SYRING[5], AND JOSHUA UDALL[4]

[2]Pacific Northwest Research Station, USDA Forest Service, Corvallis, Oregon 97331 USA; [3]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331 USA; [4]Department of Plant and Wildlife Sciences, Brigham Young University, Provo, Utah 84602 USA; and [5]Department of Biology, Linfield College, McMinneville, Oregon 97128 USA

- *Premise of the study:* The dramatic advances offered by modern DNA sequencers continue to redefine the limits of what can be accomplished in comparative plant biology. Even with recent achievements, however, plant genomes present obstacles that can make it difficult to execute large-scale population and phylogenetic studies on next-generation sequencing platforms. Factors like large genome size, extensive variation in the proportion of organellar DNA in total DNA, polyploidy, and gene number/redundancy contribute to these challenges, and they demand flexible targeted enrichment strategies to achieve the desired goals.
- *Methods:* In this article, we summarize the many available targeted enrichment strategies that can be used to target partial-to-complete organellar genomes, as well as known and anonymous nuclear targets. These methods fall under four categories: PCR-based enrichment, hybridization-based enrichment, restriction enzyme-based enrichment, and enrichment of expressed gene sequences.
- *Key results:* Examples of plant-specific applications exist for nearly all methods described. While some methods are well established (e.g., transcriptome sequencing), other promising methods are in their infancy (hybridization enrichment). A direct comparison of methods shows that PCR-based enrichment may be a reasonable strategy for accessing small genomic targets (e.g., ≤50 kbp), but that hybridization and transcriptome sequencing scale more efficiently if larger targets are desired.
- *Conclusions:* While the benefits of targeted sequencing are greatest in plants with large genomes, nearly all comparative projects can benefit from the improved throughput offered by targeted multiplex DNA sequencing, particularly as the amount of data produced from a single instrument approaches a trillion bases per run.

Key words: target enrichment; genome reduction; hybridization; genotyping-by-sequencing; microfluidic PCR; multiplex PCR; transcriptome sequencing.

With only modest fanfare, next-generation sequencing is poised to transform plant biology. It is now possible to isolate genomic DNA or RNA from any plant or any plant community (e.g., roots from soil) and obtain a nearly complete sample of all nucleotides contained within the sample of interest, irrespective of genome, transcriptome, or metagenome complexity. This flood of information will transform nearly every aspect of plant biology, as researchers utilizing genetic, biochemical, physiological, developmental, or species occurrence information can gain access to the rich and variable genomic variation that drives the phenotypes and processes we observe. The engines of this transformation—specifically, next-generation sequencing (NGS) platforms—have increased their capacity by almost an order of magnitude each year for the past 5 years (Mardis, 2011), reaching a current capacity that exceeds hundreds of billions of bases per instrument run. In the near future, this wealth of information will translate into more accurate results, higher precision and power for statistical analyses, and greater insight into complex processes, such as physiological responses to stress, the process of adaptation, and biotic community complexity.

For model plants and plants possessing small genomes, the "future" has already arrived, as whole genome sequencing has been used to evaluate interindividual variation at a genome scale in *Arabidopsis* (Schneeberger et al., 2011), rice (Arai-Kichise et al., 2011), and soybean (Lam et al., 2011); to dissect the genetic basis of complex adaptive traits like serpentine soil tolerance (Turner et al., 2010) or the circadian clock (Ashelford et al., 2011); and to gain insights into processes that were previously difficult to evaluate, such as the frequency of alternative splicing in the complete transcriptome (Filichkin et al., 2009). In contrast, for nonmodel plants and plants possessing large genomes, we are at a crossroads where complete genomes can be sequenced but not readily assembled and where comparative genome-scale analysis of a large number of individuals is not cost effective for most studies. Next-generation sequencing has certainly evolved to a point where total DNA from nonmodel organisms can be evaluated for sequence variation at highly abundant targets like organelle genomes and rDNA (Meyers

[1]Manuscript received 27 July 2011; revision accepted 17 December 2011.

We thank Jonathan Wendel and Lex Flagel for input on hybridization experiments and Katherine Armes, Tara Jennings, Sara Sundholm, and Jennifer Swanson for laboratory assistance. Tara Jennings and Shannon Straub provided critical input and editorial suggestions for early drafts of this manuscript. The research described in this review was financially supported by the USDA-AFRI Plant Genome program (grant 2010-65300-20166 to R.C.), the National Science Foundation (grants ATOL-0629508 and DEB-0317103 to A.L. and R.C.), the M. J. Murdock Charitable Trust (to J.V.S.), and our respective institutions. The use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U. S. Government.

[6]Author for correspondence (e-mail: rcronn@fs.fed.us)

doi:10.3732/ajb.1100356

and Liston, 2010; Steele and Pires, 2011; Straub et al., 2011, 2012). Low depth or "genome skimming" (Straub et al., 2012) surveys of high-copy targets are increasingly feasible, even for a large number of samples (e.g., Parks et al., 2009; Straub et al., 2012), and they are certain to meet the needs for diverse areas of plant research, including species identification/DNA barcoding, phylogeography, and phylogeny.

Beyond high copy and repetitive targets, most of the unique fraction of the nuclear genome is generally inaccessible to "genome skimming", and studies that depend on accurate information on gene presence/absence, gene structure, and accurate detection of single nucleotide polymorphisms (SNPs) are not amenable to low depth strategies. For accurate resolution of these features, sequencing depth across the genome needs to be sufficiently high—typically in the range of 15× to 30× (Schatz et al., 2010)—that a reference sequence can be assembled, and structural and sequence variants can be accurately detected relative to background sequencing error. At this depth, the highest capacity next-generation sequencers currently available can sequence a single ~1 Gbp genome per lane; this genome size is exceeded by 70% of plants documented in the Plant *C*-value database (Zonneveld et al., 2005; http://www.kew.org/cvalues/; accessed October 2011).

An alternative to sequencing complete genomes is to reduce genomic complexity in a sample by targeting a portion of the genome for selective enrichment, while eliminating the remaining (majority) of the genome. Many targeted sequencing strategies have been developed to take advantage of the growing capacity of next-generation sequencers (also see Garber, 2008; Turner et al., 2009; Mamanova et al., 2010; Davey et al., 2011). Most methods were developed to enable selective isolation and sequencing of targets from the human genome, which is large and complex (3.3 Gbp; 22 000 genes; Pertea and Salzberg, 2010). These enrichment methods build on traditional molecular biological approaches that have been used for decades—PCR-based enrichment, hybridization-based enrichment, restriction enzyme-based enrichment, and physical isolation of mRNA—and can be used to target known regions of the genome, such as dispersed gene fragments, long contiguous segments, and anonymous regions that can be reliably isolated from the background of the larger genome. The major departure of these methods from their historical roots is in modifications to accommodate large targets (kilobases to megabases) to capitalize on the high capacity afforded by NGS platforms. For the near-future, targeted sequencing of genomic partitions—genes, organellar genomes, transcriptomes, and possibly exomes—represents a powerful, cost-effective approach for obtaining accurate DNA sequences for comparative genetic analysis from large-genomed plants.

In this paper, we summarize available genome reduction strategies and show how they can be used to enrich genomic DNA and total RNA preparations for specific targets: partial-to-complete organellar genomes; known nuclear targets for SNPs and microsatellites; and anonymous nuclear targets. These sequences in turn can be used to address population genetic, phylogenetic, and comparative genomic questions. All methods have been tested extensively in human genomics applications, and they are increasingly being adapted to study plants. Here, we highlight how these strategies have been successfully applied to study different aspects of plant comparative biology including sequencing partial-to-complete organelle genomes for population genomic, phylogeographic, and phylogenomic analysis; sequencing nuclear loci to identify and genotype polymorphic markers in population- and taxon-specific discrimination and studies of gene function and adaptation, as well as the construction of dense genetic linkage maps; sequencing transcribed mRNAs for gene discovery, polymorphism identification, and functional gene expression analysis.

In light of the diversity and peculiarities of plant genomes (highly variable size; highly variable organization; high redundancy through replicative transposition and polyploidy), it seems certain that no single genome reduction method will be "best" for all applications. However, by combining genome reduction methods with other approaches (such as low coverage genome skimming), plant biologists can exploit the power of massively parallel target enrichment for a diversity of targets, ranging from comparatively small (e.g., targeted resequencing of regions in the range of tens to hundred of kbp from populations), to very large (e.g., Mbp targets from smaller numbers of individuals). The benefits of massively parallel target enrichment are most evident in plants with large genomes, but comparative studies of individuals, populations, and taxa with *any* size genome can benefit from the improved throughput and reduced analytical complexity offered by genome reduction approaches.

THE BASICS OF GENOME REDUCTION

For studies using next-generation sequencing, the amount of sequencing required to adequately characterize a genomic target, and ultimately complete the study, depends on three important factors: the *specificity* of target enrichment, the *enrichment factor* across targeted regions, and the *uniformity* of target enrichment. Additional factors that relate to efficiency and cost are the ability of enrichment methods to scale to "next-generation capacity" with minimal effort and the compatibility of enrichment methods with multiplex sequencing approaches that enable the simultaneous sequencing of multiple samples.

Specificity and *enrichment factors* can be calculated from the experimental sequencing data. Given N total bases sequenced in an experiment, N_T is the number of bases mapping inside the target region, N_G is the number of bases mapping outside the target region, G is the genome size, and L is the target region size:

$$\text{Specificity} = \frac{N_T}{N} \times 100 \qquad \text{Enrichment} = \frac{N_T \times (G - L)}{N_G \times L}.$$

Specificity is simply the proportion of "on-target" reads relative to the total pool of sequence reads, and the enrichment factor is the ratio of the coverage of the targeted region vs. the coverage of the genome outside the target region (the unenriched fraction). *Uniformity* of enrichment is a measure of the variation in sequencing depth across sites within a contiguous target region, as well as the average depth among multiple targeted regions. Measures of uniformity follow common measures of dispersion, such as the coefficient of variation.

The *ability to scale* to next-generation capacity is important because sequencing on these platforms becomes increasingly expensive if capacity is not reached. Achieving the right balance between methodological simplicity and scalability requires trade-offs that can only be determined on a project-by-project basis. For example, the simplest and most specific method of target enrichment is traditional single-plex PCR because target enrichment can be optimized for every locus examined. However, the effort required to isolate many loci by PCR scales proportionately with the number of targets examined. If the goal is to isolate hundreds to thousands of loci, then the benefits of traditional PCR (high specificity) will be outweighed by the labor, time, input DNA, and inevitable gap filling that would be required across multiple samples. Methods with lower specificity and uniformity may be preferred in many situations if they scale efficiently to kb- or Mb-sized targets.

The compatibility of enrichment methods with multiplex sequencing is a final and important factor in targeted enrichment because the reduced complexity of enriched libraries requires multiplex sequencing to be cost effective. Nearly all multiplex sequencing approaches involve adding a unique "barcode" nucleotide sequence to one of the platform-specific adapters or PCR primers so that multiple libraries can be mixed and sequenced in a common sequencing reaction; after sequencing, barcode identifiers are used to associate a sequence with a specific sample. The location of a molecular barcode varies by method and platform (Fig. 1A–C). For example, barcodes can be added to adapters so that they are sequenced with the genomic DNA insert in a single sequencing reaction (Binladen et al., 2007; Cronn et al., 2008; Hamady et al., 2008; Smith et al., 2010). This type of "internal" tagging can place identifiers on the distal end of the 5′ adapter so that they are read immediately before the insert. Libraries containing inserts smaller than the read length of the instrument (e.g., PCR amplicons, small RNAs) can use barcodes that are located on the 5′ end of the 3′ adapter (Vivancos et al., 2010) or even include unique pairs of barcodes on the 5′ and 3′ adapters to increase the barcode complexity (Roche Diagnostics Corporation, 2009; Galan et al., 2010). A different strategy to internal barcode tagging is for barcodes to be added to adapters so that they are sequenced separately of the insert sequencing reaction. This method of "index" tagging is currently limited to the Illumina platform, and it has the advantage of being fully compatible with Illumina's base-calling pipeline (Kircher et al., 2011). A large number of barcoded adapters are available for each of these approaches, so barcode availability is typically less of a concern for high multiplexing than is the difficulty of balancing equimolar input of large numbers of templates (e.g., Craig et al., 2008; Galan et al., 2010).

Beyond these factors, additional considerations in genome reduction may profoundly impact the cost and feasibility of a study, for example:

(1) Are the targeted regions of interest (ROIs) known genes or genomic regions? If so, targeted sequencing requires a reference for primer or probe design. The reference used for oligonucleotide synthesis must be sufficiently similar that the enrichment method (primed amplification; hybridization) works at the desired efficiency. Reference availability (e.g., organellar genome sequences; EST resources; gene sequences) may be limiting for some plant groups, but these resources are expected to grow at a remarkable rate, particularly as efforts like the 1 KP Project (one thousand plant transcriptomes; http://www.onekp.com/project.html) near completion.

(2) Are the regions of interest anonymous? If so, the targeting method needs to be sufficiently selective such that targets have a high probability of being captured across multiple samples. This selectivity is usually accomplished through the choice of restriction enzymes that have different recognition sequences, or different sensitivities to methylation.

(3) How complex is the target pool? Is the goal to enrich highly similar paralogs from a gene family or homoeologous loci from polyploid genomes, or to instead target divergent genes with low genic redundancy? Analytical approaches used for isolation may be minimally influenced by these decisions, but the choice of sequencing platform may be crucial to success, as read length may have a marked impact on the power of identifying unique loci and alleles in a collection of highly similar sequences.

(4) What are the desired data? Contiguous chromosomal regions, such as complete organellar genomes? Dispersed genomic loci representing complete or partial genes? Segregating sites dispersed evenly across the genome? Expressed gene sequences with transcript abundance information? The enrichment methods outlined here produce different kinds of data, and this may limit their application, particularly if the resulting sequences are too short for analysis, or if the resulting depth is insufficient for variant detection or quantitation.

PCR-BASED ENRICHMENT

The universal familiarity with PCR makes it a common starting point for exploring the utility of next-generation sequencing. As an enrichment method, PCR provides a reference point from which targeted enrichment can be compared, because it has high target specificity, reproducibility, and sequencing uniformity (Mamanova et al., 2010). Successful PCR requires a high degree of sequence conservation in the priming sites, and this can limit amplification to genic regions with comparatively low mutation rates and low rates of rearrangement/insertion-deletions events. Most often, primers are designed to target specific loci (Cronn et al., 2008; Bundock et al., 2009; Njuguna et al., 2010), although degenerate primers can be used to target allelic variants (Yuan et al., 2009; Kawakami et al., 2010), exons, promoters, or poly(A) sites (Senapathy et al., 2010) with reduced target specificity. Amplicons of any size range can conceivably be sequenced on one of the available next-generation platforms (Table 1). Although direct sequencing of full-length amplicons is only feasible for DNA molecules that fall within the "optimal" read length of the instrument, this is usually much smaller than the modal sequence length for sequencers that deliver mixed read lengths (e.g., 454/Roche; Roche Diagnostics Corporation, 2009), or it is equal to the cycle number for sequencers that deliver fixed read lengths (Illumina, SOLiD). Amplicons larger than the optimal length can be partially sequenced by direct sequencing or completely sequenced following fragmentation and conversion into a library of randomly sheared products.

Direct sequencing of small PCR products—Direct sequencing of small amplicons precludes the need for a labor-intensive DNA fragmentation step, and it reduces the cost of library preparation. Direct sequencing on the Roche/454 platform (Bundock et al., 2009; Rigola et al., 2009; Roche Diagnostics Corporation, 2009) is optimal for amplicons that are generally less than the modal read length; this length is currently 500–700 bp, depending on the chemistry used. Amplicons greater than or equal to the modal length can be sequenced, but sequencing quality across amplicons may be diminished and sequence yield can be reduced. An important consideration for direct sequencing on the Roche/454 is that amplicons should be within ±10% length, because emulsion PCR (emPCR) can impart a selective

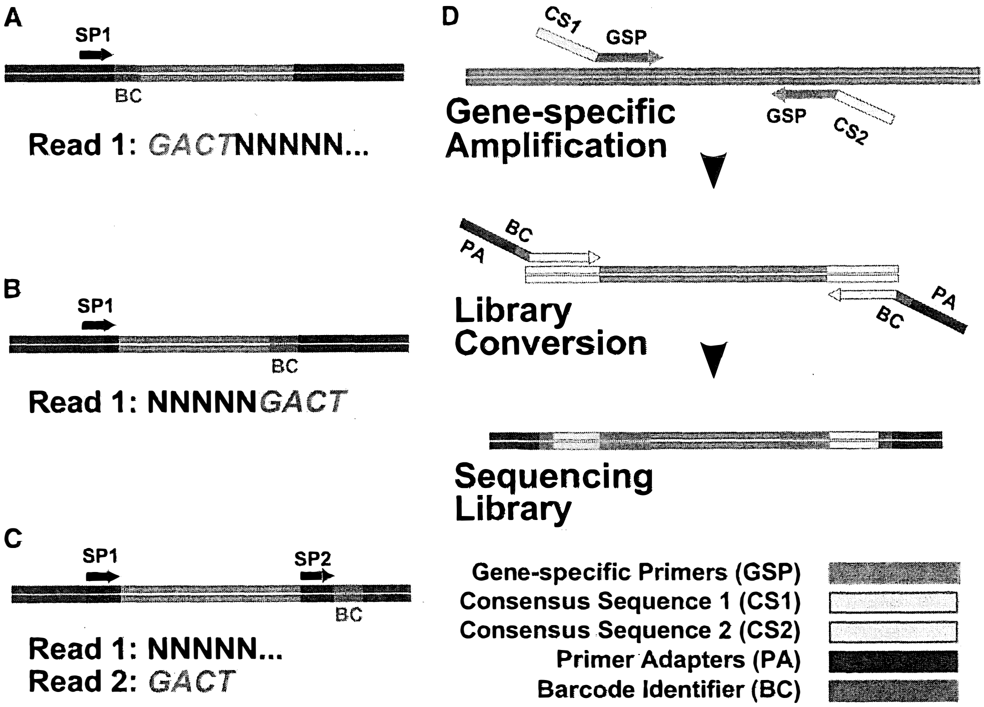

Fig. 1. Options for multiplex sequencing of enriched targets. (A) Internal barcodes (BC) can be added to the 5′ adapter so that they are read immediately off the sequencing primer (SP1). In this example, the barcode identifier "GACT" precedes the insert sequence (N) in Read 1. (B) Barcodes can be added to the 3′ adapter so that they are read immediately downstream of the insert. This method requires that the length of the insert be shorter than the sequencing read length. (C) Barcodes can be added within an adapter so that they are read in a separate "index" sequencing reaction with a second sequencing primer (SP2). This method is compatible with inserts of any length. (D) Fusion PCR primers can be used for direct sequencing of barcoded PCR amplicons. Fusion primers include a sequence that targets a specific gene (GSP), and either a linker that connects directly to attachment adapters for the sequencing platform (2 primers, direct amplification), or a conserved sequence (CS) that is targeted by a second pair of primer adapters (PA) that include the attachment sequence (4 primers, indirect coamplification). In each case, 5′ and/or 3′ internal barcode identifiers (BC) are added to facilitate high multiplexing.

bias for shorter amplicons. Direct sequencing of PCR products on microread platforms (like the Illumina Genome Analyzer or HiSeq2000) has not been reported for amplicons in plants, but it is gaining popularity in metagenomics applications due to the potential for deep sequencing and detection of rare 16S rDNA variants (Caporaso et al., 2010; Bartram et al., 2011). In these applications, paired-end sequences up to 100 bp per read (200 bp total) have been used to sequence multiplexed environmental samples. Application notes from the manufacturer show that it is possible to directly sequence pooled PCR products with single-end sequencing up to 150 bp, or paired-end sequencing up to 300 bp. For planning purposes, it is important to note that paired-end sequencing on the Illumina does require additional informatics steps to merge paired reads into a single

TABLE 1. Sequencing capacity of selected next-generation platforms. The last three columns provide estimates of multiplex levels and numbers of PCR reactions for a hypothetical example assuming a 150-kbp target, amplified in thirty 5-kbp fragments at the desired coverage depth.

Platform, read length	Instrument capacity (Gbp)	Minimum samples/run[a]	Minimum sample capacity (Gbp)	Desired coverage depth	Maximum multiplex	Amplicons (5 kb each) required to match capacity	Per-sample sequencing cost, US$[e] (kb/$)
Illumina GAIIx, 150 bp single-end SE	45[b]	7	6.4	50	857	25715	2200 (1864)
Illumina HiSeq2000, 100 bp SE	285[b]	7	40.7	50	5428	81429	3000 (6800)
Roche/454 GS FLX Titanium XL+, avg. 700 bp SE	0.7[c]	16	0.04	5	58	875	2100 (19)
ABI SOLiD 5500xl, 75 bp SE	56[d]	6	9.3	50	1244	18667	2000 (8350)

[a] "Run" signifies the maximum available physical division of sequencing surface currently available. For Illumina platforms, one lane per flow cell is used for calibration/quality control purposes.

[b] http://www.illumina.com/systems/sequencing.ilmn (accessed 15 July 2011)

[c] http://454.com/products/gs-flx-system/index.asp (accessed 15 July 2011)

[d] http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html (accessed 15 July 2011)

[e] Prices estimated for the GAIIx (http://htseq.uoregon.edu/), Roche/454 (http://www.genome.duke.edu/cores/sequencing/), and all other platforms (http://www.molecularecologist.com/next-gen-table-2b/) (all sites accessed 15 July 2011)

contig; these steps are not required on the Roche/454. It is also important to note that the sequencing read lengths stated include the sequence of the gene-specific PCR amplification primers, because most methods usually resequence the primers along with the insert. If primer regions need to be excluded, it is possible to redesign sequencing primers to minimize wasted sequencing effort (Caporaso et al., 2010).

Sequencing of long PCR products—An alternative to direct sequencing of small amplicons is to amplify large targets that are randomly sheared and constructed into libraries that are sequenced on NGS platforms. This approach has a number of advantages over amplification and sequencing of small target amplicons. For example, direct sequencing of shorter products may result in excessive coverage of amplicon ends and reduced coverage uniformity (Harismendy and Frazer, 2009). While there is no theoretical limit to the length that an amplicon can be sequenced, a long PCR cannot extend efficiently beyond 35 kbp (Cheng et al., 1994; Keeney, 2011), and reactions targeting amplicons in the range of 3–10 kbp may prove more robust. In general, most studies have found increased efficiency and reduced costs in amplifying longer regions of DNA, followed by fragmentation through nebulization or sonication during library construction (Mamanova et al., 2010).

In light of the effort associated with amplifying large numbers of long PCR products by traditional PCR, it is not surprising that long PCR-based targeted enrichment strategies have been coupled with next-generation sequencing in only a handful of cases. These studies have incorporated a "brute-force" approach that entails individual amplification of numerous long products, which are then pooled and tagged with barcoded adapters prior to sequencing (Fig. 2A). In plant phylogenomic and population genomic studies, the chloroplast genome has been the primary target to date. Full plastome sequences have been amplified in 3–4-kbp pieces for sequencing on the Illumina platform in *Pinus*, the broader Pinaceae, and *Fragaria* (Rosaceae), with application to phylogenetic resolution at low taxonomic levels (Cronn et al., 2008; Parks et al., 2009; Njuguna et al., 2010). Similarly, complete and partial plastome sequences have also been applied to the analysis of intraspecific estimates of diversity and differentiation in a number of pine species (Whittall et al., 2010). These studies highlighted a challenge associated with sequencing and assembly of fragmented amplicons on the Illumina platform, as coverage depths drop significantly in the 30-bp proximal to primer sites (Fig. 2; Cronn et al., 2008; see also Whittall et al., 2010; Njuguna et al., 2010). Similar decreases were reported by Knaus et al. (2011) when primers were spaced at greater distances (e.g., 75–100 bp) in an effort to minimize this phenomenon. Increasing amplicon overlap to a minimum of 100 bp should resolve this problem and is deserving of consideration in primer design (Harismendy and Frazer, 2009).

The nuclear genome is a more challenging target in plants due to its complexity and variability (Kellogg and Bennetzen, 2004) and the lack of a reference sequence for the vast majority of plant taxa. PCR-based enrichment strategies have been applied to the nuclear genome in vertebrate studies, suggesting that the technical challenges associated with complex plant genomes should not be insurmountable. For example, Craig et al. (2008) PCR-amplified 5-kbp nuclear regions totaling 120 kbp from 46 individuals to interrogate genetic variants within ENCyclopedia Of DNA Elements (ENCODE Project; http://www.genome.gov/10005107), with subsequent sequencing performed on the Illumina GA platform. Applicability to nonmodel organisms has also been explored to some degree in vertebrates, as well. Babik et al. (2009) used tagged, degenerate PCR primers in conjunction with sequencing on the 454 FLX platform to amplify and successfully genotype a major histocompatibility complex (MHC) exon in 79 bank vole accessions. Notably, while Babik et al. (2009) predicted that up to 1000 individuals could be genotyped using their methods in a single 454 FLX sequencing run, they found greater than two orders of magnitude differences in the number of sequence reads obtained per individual due to errors in pooling.

Similar to the nuclear genome, PCR-based enrichment paired with next-generation platforms has not been explored to date for plant mitochondrial targets. Again, this lack is largely due to the paucity of mitochondrial genome reference sequences and the inherent structural variability of plant mitochondrial genomes (for example, Alverson et al., 2010; but also see Duminil et al., 2002). The more compact and structurally consistent mitochondrial genomes of animals have been sequenced using long-PCR approaches on both Illumina and 454 platforms (Ermini et al., 2008; Jex et al., 2010; Zaragoza et al., 2010; Knaus et al., 2011).

Multiplex and microfluidic amplification of PCR products—Standard nonmultiplexed PCR represented a reasonable enrichment strategy for sample preparation when the earliest NGS sequencers were released, but the dramatic rise in sequencing capacity makes this strategy less cost effective. Multiplex PCR amplification of numerous targets in a single reaction has been developed as a possible alternative (Edwards and Gibbs, 1994; Markoulatos et al., 2002), and multiplexing kits are now available for many applications in human genetics. There are many challenges in developing high-multiplex amplification, such as off-target priming and primer–dimer formation, and the inability to accurately quantify individual amplicon concentrations. Efforts to reduce primer interactions have led to the development of methods that use common amplification primers of chemically tagged strands (Varley and Mitra, 2008), the use of common "selector" adapters (Dahl et al., 2007), and the immobilization of primers onto solid surfaces (Meuzelaar et al., 2007). These modifications permit the simultaneous amplification of targets ranging in the hundreds to low thousands per reaction. These approaches could prove cost effective for specific applications such as targeted exon sequencing.

An alternative to multiplexing amplicons in a common reaction is to independently amplify products in microfluidic reactors, and then pool the products after amplification for multiplexed sequencing. One of these commercially available platforms—the Access Array System (Fluidigm, San Francisco, CA)—uses a specialized 48 × 48 microchannel plate that combines 48 amplicon primer sets with 48 DNA samples in 36 nL wells. The small reaction chamber provides sufficient scaling of the PCR reaction such that only 15 U of *Taq* polymerase is required to generate 2304 amplicons. After PCR, reactions are combined across amplicons, and 48 pools of amplicons (one per individual) are collected, quantified, and multiplex sequenced on NGS platforms. PCR amplicons from three to four Access Array plates (6912–9216 amplicons) can be simultaneously sequenced in one-half of a Roche/454 picotiter plate. Sequencing depth per amplicon is typically between 50× to 150×, and reads representing each individual sample are identified via their multiplex identifier (MID) barcode sequence (Fig. 1; Roche Diagnostics Corporation, 2009). As with traditional

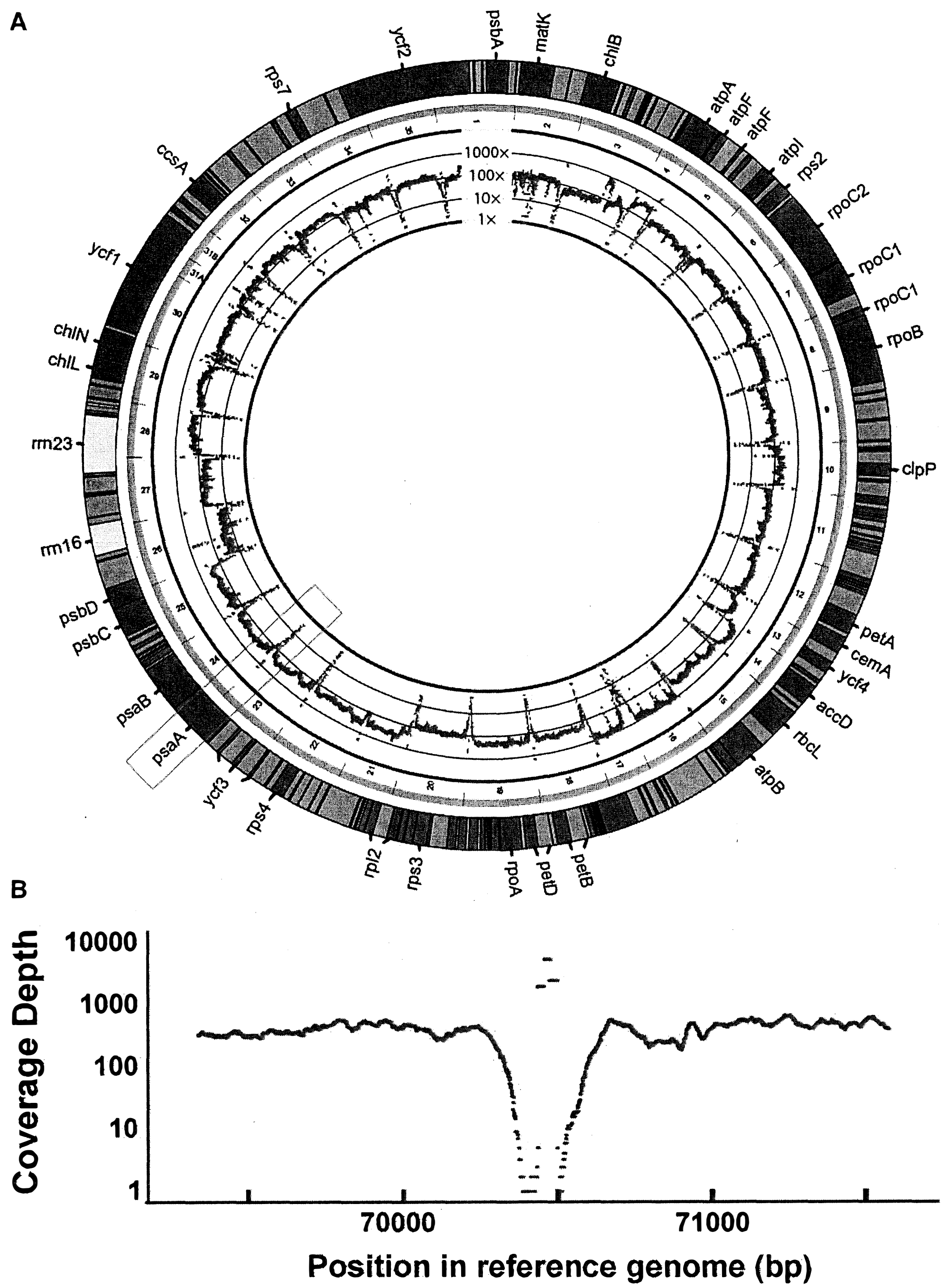

Fig. 2. Coverage depth of PCR-amplified chloroplast genomes sequenced on Illumina GAIIx (Cronn et al., 2008; Parks et al., 2009). Plastome assembly is from *Pinus flexilis* (FJ899576, Parks et al., 2009), and the total length is 117.3 kbp. (A) Coverage depth across entire plastome. Plastome map regions include protein coding (blue), noncoding (green), rRNA (yellow), and tRNA (orange) regions. Coding loci >500 bp are labeled. Amplicon coverage depth scale is shown in the inner track. Amplicon and primer junction locations (hash marks) are shown between plastome map and coverage depth plot tracks. (B) Coverage depth at junction between amplicons 23 and 24, indicated by red box in (A).

PCR, reactions can also be multiplexed on the Fluidigm system, with up to 10 primer pairs pooled per well of a 48 × 48 IFC chip, extending the number of amplicons that can be screened to 23 040 per reaction (48 individuals × 48 wells × 10 target amplicons/well). Coverage statistics provided by Fluidigm show that if a 48 × 480 amplicon pool is sequenced on one lane of the Illumina HiSeq 2000, amplicons would be sequenced at an average of 8680× depth. This is an unnecessarily high sequencing depth, so a reasonable strategy would be to pool the products from many Access Arrays experiments and then sequence them simultaneously.

The second commercially available platform—the RainDance system (RainDance Technologies, Lexington, Massachusetts, USA)—creates microscopic "reaction chambers" by merging picoliter droplets of primer pairs with the remaining reactants in a typical PCR reaction (reagents, template DNA;

Tewhey et al., 2009). Once merged, droplets are maintained in an emulsion and processed as an emPCR sample. This platform has many of the advantages of the Fluidigm (reduced primer pair interactions; reduced reagent needs), but it permits the amplification of significantly more unique amplicons per reaction. One published study using RainDance (Tewhey et al., 2009) reported excellent coverage depths across 3725 of 3976 amplicons when sequenced on the Roche/454 and Illumina platforms. As with traditional PCR and the Access Array, it may be possible to mix multiple primers per droplet (up to five) to further increase the number of targets to ca. 20 000 (Tewhey et al., 2009) per sample.

Considerations—Although PCR-based enrichment can be integrated with next-generation sequencing platforms, there are a number of factors that may make this a less than ideal coupling. First, as with all PCR-based applications, nontarget amplification (i.e., pseudogenes, paralogs) (Alvarez and Wendel, 2003; Arthofer et al., 2010), PCR recombination (Bradley and Hillis, 1997; Cronn et al., 2002), and PCR bias (Mutter and Boynton, 1995; Kanagawa, 2003) can act singly or in concert to negatively impact sequencing results. Second, difficulties in accurate quantification and equimolar pooling of PCR amplicons from multiple reactions and across highly multiplexed PCR samples are well documented (Craig et al., 2008; Galan et al., 2010; Elshire et al., 2011). Third, PCR enrichment is prone to failure at the level of individual reactions due to poor template quality or inhibitory contaminants. When combined with the high primer specificity required by PCR, the challenge of "hole filling" may be the greatest in PCR among the currently available methods.

Finally, and perhaps most important, is the difficulty of scaling PCR to the capacity of NGS platforms. Consider a project that targets a 150-kbp region by PCR-amplifying thirty 5-kbp regions, that the products are sequenced on the minimum sample unit possible (1/16 of a picotiter plate for the Roche/454; one lane for the Illumina and SOLiD platforms), and a reasonable sequencing depth for the resulting products is maintained (5× for the Roche/454, 50× for microread sequencers). In this example, it would take 875 and 81429 5-kbp amplicons to meet the sequencing capacity of the Roche/454 and Illumina HiSeq 2000, respectively (Table 1). If smaller amplicons of 500 bp are used, then 8750 (454) and 814 290 (HiSeq) amplicons must be generated to complete the same task. Obviously, NGS platforms do not need to be filled to capacity to perform well; however, the farther an experiment is from the limit of a given instrument, the higher the effective sequencing costs are for an experiment.

HYBRIDIZATION-BASED ENRICHMENT

With the rapid growth of NGS platforms (Mardis, 2011), the "front end" of targeted high-throughput sequencing was quickly recognized as a significant bottleneck and one that would worsen with the anticipated rate of growth in NGS platforms. Many groups looked to conventional hybridization-based methodologies (Lovett et al., 1991; Parimoo et al., 1991) as a potential solution for efficiently enriching multiple targets simultaneously from complex eukaryotic genomes. The development of microarray technologies, and particularly high-density custom oligonucleotide arrays synthesized using digital photolithography (Hughes et al., 2001), offered the ability to make dense pools of oligonucleotide probes to drive hybridization-based sequence capture for homologous, cohybridizing targets. The confluence of high-density oligonucleotide synthesis and NGS technologies has set the stage for transforming hybridization into a capture method with broad potential in the plant sciences, and one that is likely to displace PCR from a starring role in targeted enrichment.

Hybridization-based enrichment takes advantage of the high specificity of oligonucleotide probes (DNA or RNA) to hydrogen bond to complementary sequences, a feature exploited in Southern hybridization (Southern, 1975), sequence capture (Lovett et al., 1991; Parimoo et al., 1991; Bashiardes et al., 2005), and microarray technology (Hughes et al., 2001). By adjusting hybridization conditions (probe length, hybridization temperature, buffer ionic strength), targets can be enriched at varying levels of stringency. Targets are usually enriched from a background genome, but hybridization with complementary double-stranded baits can also be used to selectively remove unwanted portions of the genome (e.g., target depletion; Fu et al., 2010). Hybridization-based enrichment is either conducted with probes on a solid support ("on-array"; Albert et al., 2007; Okou et al., 2007; Fu et al., 2010) or in-solution (Porreca et al., 2007; Gnirke et al., 2009). On-array approaches use probes fixed to a solid support such as a nylon filter, glass slide, or microarray. When complex DNA (usually a library prepared for a specific NGS platform) is applied to the array, the desired fragments hybridize to homologous probes, nontargeted fragments are washed away, and the target is enriched with PCR prior to sequencing. Solution hybridization is similar to solid phase except that probes are free and typically biotinylated to facilitate capture. Following hybridization of probes and the DNA pool, biotin probe–target hybrids are captured with streptavidin beads, nontargeted DNA is washed away, and targets are eluted and enriched by PCR.

As an enrichment method, hybridization offers an attractive alternative to large-scale PCR, and it has been widely adopted in human genomics as a method for rapid screening of a large number of predefined genomic targets (Gnirke et al., 2009; Bainbridge et al., 2010; Mamanova et al., 2010; Bansal et al., 2011). Many commercial suppliers offer probe synthesis services that can be used to enrich targets as small as 1–2 Mbp (ca. 10 000–20 000 probes) or as large as complete exomes (>30 Mbp). Solution hybridization is accomplished in a single tube, so the process scales to large numbers of samples, multiwell formats, and robotic automation (Fisher et al., 2011). In most commercial applications, hybridization probes are biotinylated, single-stranded RNAs that are 120 bp in length. RNA probes have significant advantages because RNA–DNA hybrids have a higher affinity and melting temperature than DNA–DNA hybrids and single-stranded RNA probes lack a probe complement that can reanneal and reduce the effective concentration of probes (Sambrook and Russell, 2001). Although they are less efficient, double-stranded DNA probes can also be used in some applications (see below).

In their original application, hybridization reactions were typically run unmultiplexed (Gnirke et al., 2009; Bainbridge et al., 2010; Mamanova et al., 2010). It was quickly discovered, however, that enrichment factors were sufficiently high that complete human exomes (30 Mbp) could be captured and sequenced in multiplex, with 8-plex providing excellent results (Nijman et al., 2010). Similar approaches have recently been attempted in loblolly pine (*Pinus taeda*; Pinaceae), and it appears that successful enrichment of thousands of loci can be accomplished from 8-plex hybridizations (L. Neves and M. Kirst, University of Florida, personal communication). Considering that these preliminary experiments targeted 6 Mbp from a genomic background of 22 Gbp, the limit of multiplexing seems to be primarily determined by the cumulative target size, not the size of the background genomic complexity of the multiplex pool. Experiments with high multiplexing of nonbarcoded human DNAs seem to support this notion; Bansal and colleagues (Bansal et al., 2011) successfully enriched 600-kbp targets from pools of 100 individuals (a total genome pool of over 300 Gbp), providing an efficient method to identify single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms in defined coding regions, and the authors predict that multiplexing of larger pools (e.g., 400 human samples) is possible.

Applications of large-scale hybridization-based enrichment in plants are still in their infancy, but published studies to date highlight the power of this approach. For example, Fu et al. (2010) used sequential on-array hybridization to deplete repetitive elements from *Zea mays* genomic libraries, then enrich the libraries for unique target loci (Fu et al., 2010). In this example, genomic DNA libraries from inbreds B73 and Mo17 were hybridized to a repeat subtraction array containing 720000 probes representing the highly repetitive fraction of the *Z. mays* genome. Unbound sequences were recovered and rehybridized on one of two capture arrays, with targets representing a 2.2-Mbp interval on chromosome 3 of B73, or 43 widely dispersed genes. Following sequencing on the Roche/454, these authors observed that 22–36% of the resulting reads were on target, that 98% of the targeted bases were sequenced at least once, and that the targets were enriched 1800- to 3000-fold. More recently, Saintenac et al. (2011) used solution hybridization to target nearly 3500 dispersed loci (3.5 Mbp) from the larger genomes of allotetraploid wheats (*Triticum dicoccoides* and *T. durum* cv. Langdon, each nearly 10 Gb/1C) that were barcoded, pooled, and hybridized in a single reaction. In this example, nearly 60% of the total Illumina-based reads aligned to reference, and the overall enrichment factor for the experiment was 2900-fold. Of 3497 full-length reference (cDNA) sequences, 2273 were represented with a median depth of 10×. A key outcome of both of these studies is that a large number of polymorphic positions were identified, with over 2500 representing the two accessions of *Z. mays* (Fu et al., 2010) and nearly 19000 representing the two allotetraploid *Triticum* species (Saintenac et al., 2011). Similarly, both studies found that probes targeting genes with known paralogues tended to enriched all copies with high efficiency. This feature was advantageous in the case of allotetraploid *Triticum*, as the authors partitioned polymorphism into allelic variation (~4400 SNPs), and into differences between the homoeologous A- and B-genomes (~15000 polymorphic sites; Saintenac et al., 2011).

We have explored hybridization-based enrichment as an alternative to PCR-based amplification of complete chloroplast genomes from conifers (Parks, 2011; M. Parks et al., unpublished manuscript; see Appendix S1 with the online version of this article). For comparison, the 34 novel chloroplast genomes sequenced, assembled and described in Parks et al. (2009) required over 1200 long-PCR reactions; many species could not be included in this analysis because they failed the large number of PCRs required or because there was insufficient template DNA. Using hybridization enrichment, we have since enriched and sequenced over 100 conifer chloroplast genomes (M. Parks et al., unpublished manuscript), with the majority of genomes enriched in a single experiment where barcoded samples were hybridized in 4-plex reactions (88 samples, 22 reactions). We have also used hybridization to enrich nuclear genomic targets for population genetic applications, such as repeat enrichment for microsatellite development in conifers (Jennings et al., 2011) and exon enrichment for SNP validation in sagebrush (Bajgain et al., 2011). Lessons learned from these early development efforts illustrate the enormous potential of hybridization methods for routine large-scale target enrichment for population genomic and phylogenomic analysis. Here, we highlight four key findings.

Finding 1: Short, untiled probes can enrich large targets—Our earliest experiments used 39 pooled, 3′-biotinylated PCR primers as hybridization probes for targets in the *Pinus* chloroplast genome. At 18 to 36 bp, these probes are small relative to probes used in commercial solution hybridization kits (typically 120 bp). Despite their short length, however, these oligonucleotides proved to be excellent hybridization probes, enriching targets to 400× above background levels and enriching flanking regions 600 bp upstream and downstream of the probe (Fig. 3A). These 39 probes totaled 950 bp in length, but they enriched targets 54 kbp in size (Fig. 3B), a value that accounts for 45% of the *Pinus* chloroplast genome. These results and others (Gnirke et al., 2009; Mamanova et al., 2010) show that the size of flanking, off-target sequence enrichment is determined by the insert size of the input genomic library (~600 bp in our case). If off-target DNA sequences are desirable (as they often are in population and phylogenetic studies), this flanking DNA can be enriched simply by increasing the size of the insert size of the input library.

Finding 2: Pooled PCR products can serve as enrichment probes—To enrich complete chloroplast genomes, we reasoned that PCR amplicons spanning a complete chloroplast genome could also serve as hybridization probes. To test this hypothesis, we PCR amplified the chloroplast genome from *Pinus thunbergii* in 35 separate reactions (as indicated in Fig. 2), concatenated the PCR products via ligation, and used ϕ29 polymerase to simultaneously amplify and biotinylate the concatemerized probes. These concatemers were used as solution hybridization probes to enrich complete chloroplast genomes individually or in 4-plex reactions. Our results from 111 hybridization experiments show that randomly concatenated PCR products made excellent hybridization probes, as chloroplast DNA increased in abundance from 1–4% in unenriched samples, to as high as 80% in hybridized samples (Fig. 3C). Across experiments, chloroplast genomes were enriched to an average of 52%, and the resulting assemblies averaged over 90% complete (M. Parks et al., unpublished manuscript).

Finding 3: Hybridization enriches degraded targets that may not be amplifiable by long-PCR—Our efforts to complete a comprehensive chloroplast genome phylogeny for *Pinus* (M. Parks et al., unpublished manuscript) motivated us to use older specimen tissue samples that contained degraded DNA that was suboptimal for long-PCR amplification. Even though these DNAs showed low PCR success, we were able to make small insert Illumina libraries for many of these samples and enrich chloroplast DNA using hybridization (Table 2). In these tests, the *Pinus* DNAs with the poorest PCR success (0–6.3%) at eight diagnostic loci could be enriched by hybridization to the point that chloroplast genome assemblies exceeded 40.5 kb. The trend we observed was one of a threshold effect, where

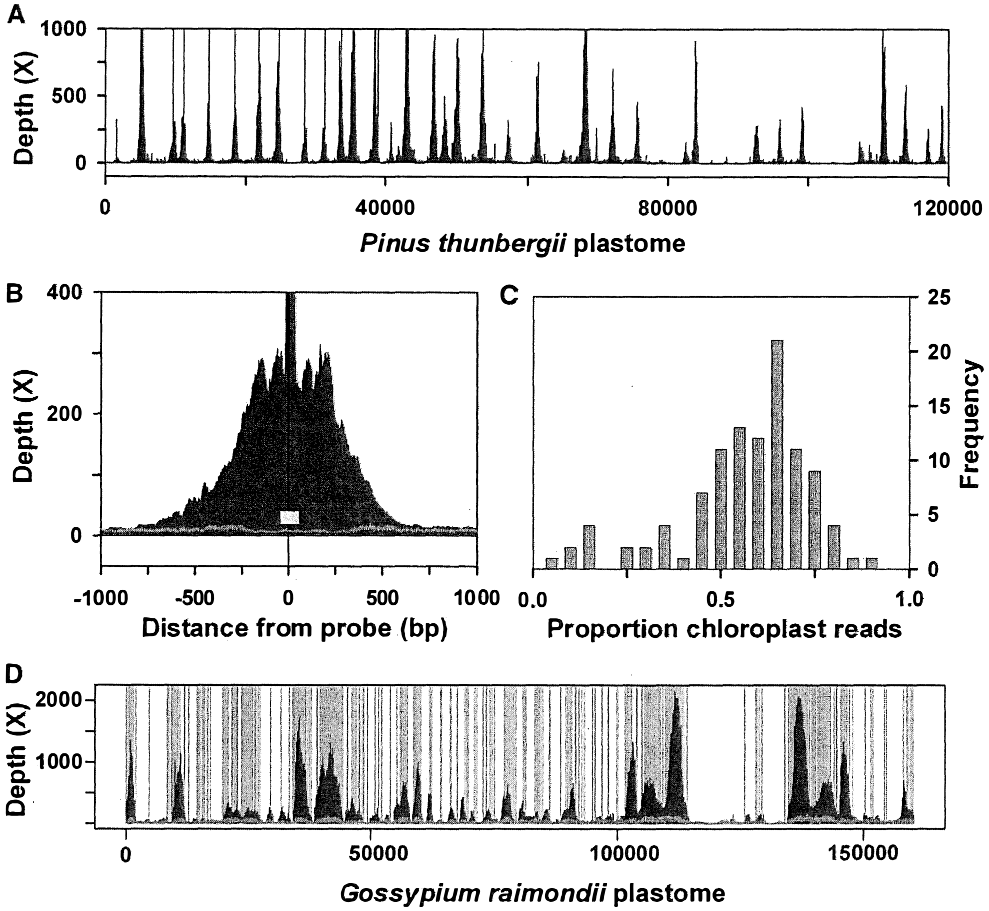

Fig. 3. Hybridization-based enrichment of chloroplast genomes from *Pinus thunbergii* and seed plants. (A) Plot showing sequencing depth by chloroplast genome position in a hybridization experiment with 39 short oligonucleotide probes. Probe locations show significant enrichment (blue peaks), with an enrichment factor of 400×. (B) Plot showing average sequencing depth by flanking nucleotide sites 1000 bp upstream (–) and downstream (+) of the probe. The location of the probe is shown in yellow, the average background for the experiment is in green, and enriched sequences are in blue. (C) Results from 111 hybridization experiments in conifers. Conifers show a native chloroplast representation of less than 0.05; the average representation for 111 hybridization experiments was 0.52. (D) Plot showing sequencing depth for the *Gossypium raimondii* chloroplast genome, following enrichment using chloroplast genome probes derived from *Pinus thunbergii*. The average background for the experiment is shown in green, and enriched sequences are in blue. Enriched targets correspond to regions that have >80% pairwise sequence identity between these divergent genomes (red shading).

samples showing 40–50% PCR success were "good enough" to yield hybridization-based assemblies >90% complete (Table 2). These results highlight a unique role for hybridization enrichment, namely, the capture of genomic targets from rare, degraded, or forensic specimens (Knapp and Hofreiter, 2010). It may be no coincidence that hybridization-based approaches played a central role in the remarkable enrichment of targets from Neanderthal specimens ranging in age from ~38 kyr to 70 kyr (Briggs et al., 2009; Burbano et al., 2010).

Finding 4: Hybridization has a broad phylogenetic reach—To push the limits of hybridization-based enrichment, we used concatenated PCR amplicon-based probes from *Pinus thunbergii* to enrich chloroplast genome DNA from a diploid cotton (*Gossypium raimondii*; Fig. 3D). Remarkably, we found levels of chloroplast enrichment in *G. raimondii* similar to those we obtained in *Pinus*; the native cpDNA representation (5.9%) was increased to 46.3% in the enriched sample, and 59.1 kbp of sequence was enriched >5× the mean depth of the unenriched sample. By comparing the nucleotide sequence of the *Pinus* and *Gossypium* chloroplast genomes, we found that hybridization enrichment was greatest in 168 regions that had >80% pairwise sequence identity (Fig. 3D). These results provide evidence that heterologous probes can enrich conserved targets from highly

TABLE 2. Hybridization-based enrichment of chloroplast genomes from samples of degraded tissue of *Pinus* species.

Taxon	% PCR success (8 loci)	% Genome assembled after enrichment	Assembly size (bp)
P. fenzeliana Hand.-Mazz.	0.0	45.5	54637
P. bhutanica Grierson, Long & Page	6.3	33.8	40524
P. durangensis Martinez	37.5	40.4	48476
P. massoniana Lamb.	37.5	99.8	119762
P. balfouriana Balf.	50.0	94.3	113210
P. johannis M. F. Robert	56.3	97.2	116607
P. hwangshanensis W. Y. Hsai	75.0	99.9	119874
P. chiapensis (Martinez) Andresen	75.0	97.6	117060
P. edulis Engelm.	93.8	96.8	116153
P. pumila (Pall.) Regel	93.8	95.8	114943

divergent lineages. More experimentation is needed to define the limits of this kind of "heterologous enrichment", but it is clear that carefully selected probe pools should be effective well beyond the original source species, and in the case of conserved gene targets (chloroplast gene sequences, conserved orthologous gene pools), they may show success across large evolutionary distances.

Considerations—Despite the advantages of hybridization-based enrichment, it is important to note that hybridization may show limited success in cases where unique insertions are present in the sample pool but absent in the probe pool. For this reason, other methods like low coverage genome skimming (Straub et al., 2012) can be recommended as a first pass strategy to gain sequences for abundant targets (e.g., chloroplast or mitochondrial genome DNA) and to evaluate the frequency of insertion/deletion in the targets and taxa of interest. Likewise, medium-depth genomic sequencing (e.g., 5–10×) can be used to identify putative SNPs and indels in conserved low copy nuclear sequences (S. C. K. Straub and A. Liston, Oregon State University, unpublished data), and these can also be used in the design of probes for targeted enrichment.

Hybridization probes can be made from reagents as simple as short PCR primers and double-stranded PCR products, but probe construction and biotinylation proceed by different paths depending on whether the probe pool is single- vs. double-stranded, or RNA vs. DNA. Appendix S1 (see online version of this article) provides example protocols for synthesizing single- and double-stranded DNA probe. Melting temperature equivalence and equimolar representation should be maintained across the pooled probes, as deviation from equivalence in these factors can result in the over-representation of favored targets. Finally, the use of blocking agents is critical to achieving a high enrichment factor in hybridization. The ideal blocking agent would be the highly repetitive fraction from the genome of the organisms being hybridized; since this is rarely available, the use of less-specific blocking agents is recommended (Sambrook and Russell, 2001). In our experience, the most important blocking agents to include are complementary, nonextendable oligonucleotides that mask the platform-specific adapter sequences used in library construction. Inclusion of these oligonucleotides will prevent "daisy-chaining" of targets that occurs through cross-hybridization between common adapter sequences of different inserts (Gnirke et al., 2009).

RESTRICTION-ENZYME-BASED ENRICHMENT

Single nucleotide polymorphisms (SNPs), defined as single-base changes, are the most abundant type of sequence variation in eukaryotic genomes (Garg et al., 1999; Batley et al., 2003). The high frequency of SNPs in most species offers the possibility of constructing extremely dense genetic maps (for map-based gene cloning and haplotype-based association studies), conducting F_{ST}-based outlier tests, and conducting phylogeographic and phylogenetic analysis with a large number of unlinked loci. Historically, the discovery of SNPs and SNP genotyping in large populations has been expensive and time-consuming, limiting their utilization in nonmodel species. Methods for targeted SNP discovery and genotyping based on restriction site conservation have been reported in the literature and have been validated in plant species in the last few years, including restriction-site-associated DNA (RAD) tags (Miller et al., 2007; Baird et al., 2008; Davey et al., 2011), genomic reduction based on restriction site conservation (GR-RSC) (Maughan et al., 2009), and genotyping by sequencing (GBS) (Davey et al., 2011; Elshire et al., 2011) (Fig. 4; Table 3). These methods rely on the discriminatory power of the restriction endonucleases to produce homologous restriction fragments among the individual samples being assayed. When paired with NGS platforms, these methods provide a cost-effective means to identify large numbers of high-confidence SNPs with broad application across diverse genomes.

Restriction-site-associated DNA (RAD) tags—The first description of this methodology by Miller et al. (2007) predated the era of readily available NGS platforms and thus relied on microarray hybridization to interrogate thousands of genetic (RAD-tag) loci in paired-sample comparisons. The recent use of NGS technology replaces the more technically challenging step of microarray hybridization (Baird et al., 2008). In the RAD technique (Fig. 4A), DNA samples are individually subjected to restriction digest using a single endonuclease (e.g., *Sbf*I), and the resulting restriction fragments are ligated with a forward adapter containing DNA sequences for forward amplification and Illumina sequencing, as well as a barcode for sample identification. Following ligation, samples are pooled and sheared to produce random fragments averaging ~500 bp. Fragments of a specific size (300–700 bp) are size-selected using agarose gel electrophoresis, and a 3′ adenine is added to facilitate the ligation of a Y-shaped reverse adapter to the fragments. This Y-adapter blocks amplification of DNA fragments that lack the forward adapter, thus a final PCR amplification step with forward and reverse primers ensure that only RAD tags are amplified. Amplified DNA fragments are sequenced using standard NGS protocols (Hohenlohe et al., 2011; Davey et al., 2011). After sequencing, sequence reads are bioinformatically deconvoluted into sample sets based on sample-specific barcodes. SNPs between samples and a sequence reference are identified by pairwise grouping of sequence data from each sample. Various stringency parameters, including read coverage and alignment matches among reads are used to identify high-confidence SNPs (Pfender et al., 2011). Similarly, genotypic calls for individual samples segregating within populations are based on comparison of the individual sample RAD tag sequences with the reference RAD tag sequence.

The RAD process produces two types of genetic markers. Sequence polymorphisms within the restriction recognition site leads to dominant markers (RAD tag cluster is present in one individual, but not the other), whereas SNPs outside of the restriction recognition site, but within the sequence of the RAD tags itself, are biallelic SNPs that segregate in a codominant fashion. Dominant RAD polymorphisms are often discarded, as null alleles require deep sequencing for accurate detection (Chutimanitsakun et al., 2011; Davey et al., 2011). RAD analysis has successfully been applied to linkage map development and QTL analysis for reproductive traits in barley (Chutimanitsakun et al., 2011) and stem rust resistance in *Lolium perenne* (Pfender et al., 2011).

RAD sequencing has recently been applied to more challenging questions of population genetic and phylogeographic analysis of wild populations of animals. An important facet of these studies is that the populations were wild, so ancestral haplotypes are unknown, and the imputation of missing genotypes is difficult to address. The reliance on one restriction site, combined with tiled reads across adjacent sheared regions, makes it

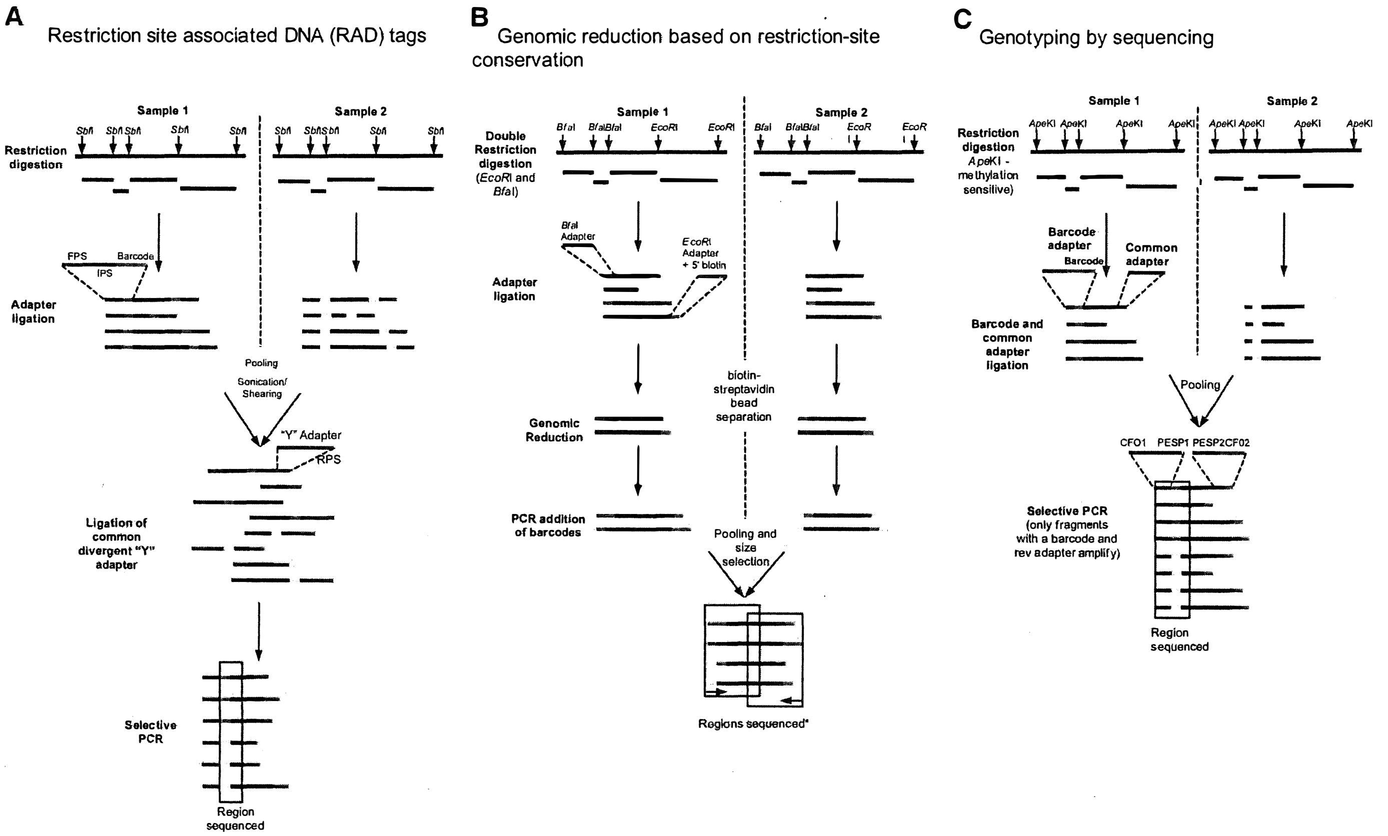
A
Restriction site associated DNA (RAD) tags
Sample 1
Sample 2
SbfI
Restriction digestion
FPS
Barcode
IPS
Adapter ligation
Pooling
Sonication/ Shearing
"Y" Adapter
RPS
Ligation of common divergent "Y" adapter
Selective PCR
Region sequenced
B
Genomic reduction based on restriction-site conservation
Sample 1
Sample 2
BfaI
EcoRI
EcoR
Double Restriction digestion (EcoRI and BfaI)
BfaI Adapter
EcoRI Adapter + 5' biotin
Adapter ligation
biotin-streptavidin bead separation
Genomic Reduction
PCR addition of barcodes
Pooling and size selection
Regions sequenced*
C
Genotyping by sequencing
Sample 1
Sample 2
ApeKI
Restriction digestion ApeKI - methylation sensitive)
Barcode adapter
Barcode
Common adapter
Barcode and common adapter ligation
Pooling
CFO1
PESP1
PESP2CF02
Selective PCR (only fragments with a barcode and rev adapter amplify)
Region sequenced

Fig. 4. Targeted sequencing via restriction enzyme-based enrichment. (A) Restriction-site-associated DNA (RAD) tags, FPS = forward priming site, IPS= Illumina sequencing priming site, RPS = reverse priming site. (B) Genomic reduction based on restriction-site conservation. 454-pyrosequencing adapters are added during the em-PCR library phase (not shown). Fragments are bidirectionally sequenced. (C) Genotyping by sequencing. CFO1and CF02 = Complement (Ilumina) flowcell binding site oligo 1 and 2, respectively. PESP1 and PESP2 = paired-end sequencing primer site 1 and 2, respectively.

possible to assemble the larger sequences required for these kinds of unpedigreed populations (Etter et al., 2011). Hohenlohe et al. (2010) used RAD to identify over 45 000 SNPs from pelagic and freshwater stickleback fish, and this information was used to test biogeographic hypotheses regarding the origin of freshwater populations and to identify genomic regions that colocalize with QTLs known to influence stickleback phenotypes. Similarly, Emerson et al. (2010) extended RAD to examining the phylogeography of the pitcher plant mosquito (*Wyeomyia smithii*) across its range in eastern North America. The unprecedented volume of data available for this analysis provided a high degree of population discrimination, allowing the authors to document the northward migration of this mosquito from refugia in the southern Appalachian Mountains to their current range. Finally, RAD has been used for SNP marker development to document and measure the frequency of hybridization in introduced rainbow trout (*Oncorhynchus mykiss*) and native westslope cutthroat trout (*Oncorhynchus clarkii lewisi*) in the western United States (Hohenlohe et al., 2011). These authors used relatively simple measures—excessively high observed heterozygosity and deviations from Hardy–Weinberg proportions—to discriminate true SNPs from differences in homoeologues in this tetraploid genome, strategies that could be widely applied in the analysis of angiosperm genomes. A common finding in all of these studies is that RAD (and presumably related technologies) can provide genomics-scale insights for nonmodel species when no prior genomic information is available.

Genomic reduction based on restriction-site conservation (GR-RSC)—Maughan et al. (2009) developed GR-RSC in an attempt to identify SNPs in *Amaranthus* (an Andean crop of regional importance) and later validated the methodology to simultaneously discover and genotype SNPs in an *Arabidopsis* recombinant inbred line (RIL) population. GR-RSC is based on restriction-site conservation across related individuals, removal of >90% of the genome via biotin–streptavidin paramagnetic bead separation and size selection via gel electrophoresis, followed by >10-fold sequencing coverage of the remaining genome via high throughput sequencing (Fig. 4B). In short, DNA is double digested with restriction endonucleases that recognize 4-base and 6-base recognition sites. Subsequently, adapters are ligated to the ends of the digested DNA fragments. The adapter ligated to the end of the 6-base recognition site is end-labeled with a 5′-biotin molecule, while the adapter on the 4-base recognition site is unlabeled. Genomic reduction is accomplished by removing the nonlabeled DNA fragments from the reaction using a biotin–strepavidin paramagnetic bead separation. DNA barcode sequences are then added to the DNA fragments using PCR primers complementary to the adapter sequences. Equimolar amounts of each individual PCR sample are pooled together and size-selected (500–650 bp) via electrophoresis in preparation for standard high throughput sequencing. Similar to RAD methods, the use of incorporated DNA barcodes allows for the assignment of individual reads to specific DNA sample pools, which can in turn be used for SNP discovery and genotyping

TABLE 3. Summary of experiments utilizing targeted sequencing via restriction enzyme-based enrichment.

Study	Method	Organism	Population type (No. individuals)	Sequencing platform (chemistry)	No. reads sequenced (No. plates or lanes)[a]	No. markers discovered	Mapping strategy	Biallelic markers mapped	Dominant markers mapped	QTL analysis
Pfender et al., 2011	RAD	*Hordeum vulgare*	DH (93)	Illumina (1X36 bp)	49 500 000	530	de novo (JoinMap 4)	445	NA	Stem rust
Chutimanitsakun et al., 2011	RAD	*Lolium perenne*	F$_1$ (193) ♂	Illumina (1X36 bp)	29 715 175	1156[b]	de novo (JoinMap 4)	305	NA	Reproductive fitness[c]
			F$_1$ (193) ♀			1216[b]	de novo (JoinMap 4)	329		
Maughan et al., 2009	GR-RSC	*Amaranthus* sp.	Pop1-3 (2)[d]	Roche/454	1 272 089 (1)	140	NA	NA[e]	NA	NA
			Pop1-4 (2)[d]			5433				
			Pop2-4 (2)[d]			11 038				
			Pop2-3 (2)[d]			11 047				
Maughan et al., 2010	GR-RSC	*Arabidopsis thaliana*	RIL (60)	Roche/454	3 098 246 (2)	6159	de novo (JoinMap 4)	1555	NA	NA
			RIL (60)	Illumina (1X76 bp)	16 476 819 (1)	701	de novo (JoinMap 4)	311		
Elshire et al., 2011	GBS	*Zea mays*	RIL (276)	Illumina (1X86 bp)	145 836 644 (6)	NA	Reference genome/ Framework map	25 185	167 497	NA
		Hordeum vulgare	DH (43)		27 500 000 (1)	NA	Framework map	NA	24 186	

[a] Reads with identifiable multiplex barcode DNA sequence.

[b] Two linkage maps with different numbers of map markers were produced for the male and female parent.

[c] Traits included: final leaf number, plant height, spike number, floret number, grain number, hundred grain weight, and grain yield.

[d] SNP discovery only.

[e] A total of 411 were subsequently converted and mapped by Maughan et al. (2011) using KASPar genotyping chemistry.

via direct comparison of reads from an individual to a reference sequence. The reliability of the method was supported by the development of five highly supported linkage groups that were collinear with the *Arabidopsis* reference genome (r = 0.981) (Maughan et al., 2009).

One notable difference between GR-RSC and RAD methods is the contig lengths produced by the differing sequencing technology used (Maughan et al., 2010). As expected, contigs produced by 454-pyrosequencing were significantly larger than those produced via Illumina sequencing, although the size of contigs is a continually moving target on all platforms. The value of the increased sequence information is an important consideration if PCR-based SNP assays are desired for individual SNP loci. For example, using the flanking sequence derived from the 454-pyrosequencing data, Maughan et al. (2011) developed 411 individual SNP assays for *Amaranthus*, based on KBioscience KASPar genotyping chemistry, which they genotyped on a Fluidigim nanofluidic chip (Fluidigm Corp., South San Francisco, CA). The use of nanofluidic genotyping reduced the cost per data point to US$0.05, which is comparable to the cost achieved through genotyping by sequencing using GR-RSC (Maughan et al., 2011). A single 96 × 96 Fluidigm integrated fluidic chip is capable of producing 9216 genotypic data points in a single run (~3 h), and minimal operator technical expertise is required.

Genotyping-by-sequencing—Elshire et al. (2011) recently described a genotyping approach for high diversity species, termed genotyping-by-sequencing (GBS). GBS relies on the use of methylation-sensitive restriction endonucleases (e.g., *Ape*KI) to avoid repetitive regions of the genome, while targeting lower copy regions of the genome, thus simplifying the computational challenge associated with restriction fragment alignment from species with high levels of genetic diversity (Fig. 4C). DNA is cut with *Ape*KI and ligated with a "common" adapter and a "barcode" adapter, which consists of a 4–8-bp barcode on the 3′ end immediately upstream of the compatible sticky ends. Modulation of the barcode size results in fewer sequence-phasing errors in the subsequent fragment sequencing. Individual samples are pooled, and a common set of PCR primers (complementary to adapter sequence, minus the barcode sequence) are used to amplify the pooled library. The PCR primers also contain sequences that allow the PCR products to bind to the Illumina sequencing flow cell and prime DNA sequencing reactions. Amplified fragments are purified and checked for appropriate fragment size (170–350 bp) and the contamination of adapter dimers. Single-end sequencing is performed using standard Illumina sequencing; while the authors used 1 × 86 bp reads, these could be extended using newer chemistry. After sequencing, reads are pooled based on barcode sequence information, and a set of reference sequence tags is identified. Reads from segregating lines within the population are then sorted into a presence/absence genotyping table based on the reference sequence tags and the parental source of the tag determined. A binomial test is used to test for segregation of the presence/absence scores against an independent framework map established from previously mapped SNPs. When pairs of tags aligned to the same unique position of a reference genome and cosegregated with the same framework SNP, the tags are merged into a single bialleic GBS marker, tested for cosegregation with the framework SNPs (Fisher's exact test), and then incorporated into a high density framework map and ordered according to their positions in the reference genome.

While similar in concept to RAD and GR-RSC, the GBS protocol is simpler to perform, requiring no sonication, paramagnetic bead separation, size selection, gel purification, or specialized equipment. A pre-established and moderately dense framework map was instrumental in determining the relative map position of the tags, especially in light of the short read lengths (64 bp) and the diverse nature of the maize genome. The availability of a reference genome allowed for the identification of biallelic tags and the physical mapping of the tags to the reference genome. Accurate genotyping of presence/absence tags requires deep and uniform sequencing across all samples. Notable was a bias toward the sequencing of smaller restrictions fragments (<64 bp), possibly the result of preferential amplification of small fragments during library construction, and/or the requirements for optimal cluster formation of the Illumina flow cell. We note that tags were mapped based on tests of linkage to reference SNPs, and not through de novo linkage mapping. The possibility of constructing a de novo map with GBS data are suggested, but such an approach is not presented.

Considerations—As with all new sequencing methodologies, targeted sequencing via restriction enzyme-based enrichment has some known limitations and may have additional unknown limitations. First, only a limited number of restriction enzyme-based enrichment experiments have been reported in the botanical literature, and all have been with diploid plant species. Undoubtedly the application of these methods to allopolyploid or autopolyploid species will be significantly more complex. Improvements to assembly and mapping algorithms will be needed to avoid mis-assembling paralogous/orthologous regions, especially in light of the fact that increased sequence data will be required to cover the increased size of the genomes of polyploid plant species. Second, researchers should recognize that the depth of coverage required to accurately call genotypes of heterozygous lines or populations is higher than what is required to accurately genotype lines or populations consisting of homozygous lines at the same level of confidence. Indeed, a minimum of four reads spanning the SNP in question would be required to achieve a 95% confidence level (P = 0.046) of accurately distinguishing a homozygous genotype from a heterozygous genotype. Illumina sequencing, with its significantly increased read numbers, may be the preferred sequencing platform for populations with high heterozygosity. Third, researchers often think in terms of cost per data point when evaluating genotyping-by-sequencing strategies. We note that data point cost is directly related to the level of genetic diversity in the population being genotyped. Populations derived from a narrow genetic base will exhibit few polymorphisms; consequently, the cost per data point will increase. Fourth, targeted sequencing via restriction enzyme-based enrichment cannot target specific chromosomal regions or specific SNPs, thus these methods are cost prohibitive for studies targeting a small number of discrete genetic loci. Indeed, if restriction enzyme-based enrichment strategies are used in linkage mapping or association mapping studies of agronomic traits (i.e., QTL), postdiscovery efforts would be needed to convert the linked markers into another SNP assay format. Last, before a large-scale implementation of restriction enzyme-based enrichment for genotyping, user-friendly bioinformatic tools, capable of handling data files from the various high throughput sequencers, are urgently needed to facilitate de novo SNP discovery and automated genotyping—especially in light of the voluminous data expected in future NGS platforms.

TRANSCRIPTOME-BASED ENRICHMENT

One of the most widely used genome reduction strategies is to focus on the transcribed portion of the genome, or the transcriptome. The transcriptome comprises a relatively small fraction of the total size of plant genomes, ranging from ~25% for angiosperms with compact, gene-dense genomes (e.g., *Arabidopsis thaliana, Medicago truncatulata*) to ~1% or less for large, highly repetitive genomes from conifers (e.g., *Pinus taeda*) (Rabinowicz et al., 2005) (Table 4). Transcriptome sequencing (often called RNA-seq) provides an efficient route to discover and describe gene and transcript structure, catalog polymorphism in exons and noncoding regions flanking exons for mapping and phenotypic associations, and quantify expression patterns that may be developmentally or environmentally regulated (Lister et al., 2009; Wang et al., 2009; Wilhelm and Landry, 2009). Transcriptome sequencing offers the promise of sequencing tens of thousands of genes without prior sequence knowledge, and it uniquely offers a means to discover novel differentially spliced transcripts ("isoforms"). For gene expression studies, the sensitivity of transcript detection permits quantitation over a range that spans many orders of magnitude. In light of the comparative ease of producing this kind of data—only standard molecular biology kits are required—transcriptome sequencing provides "one-stop shopping" for the entry of nonmodel organisms into high-throughput sequencing projects.

A strength of transcriptome sequencing is that it focuses sequencing resources on the expressed portion of the genome without a need for prior sequence knowledge, in contrast to PCR- and hybridization-based approaches, which require advance knowledge of target sequences and the design/synthesis of oligonucleotide primers or probes. Restriction methods are similar to transcriptome sequencing in that they require no prior sequence knowledge, but restriction methods are sensitive to enzymatic biases (recognition sites, methylation sensitivity, incomplete digestion). Restriction methods also produce anonymous and often dominant data, where transcriptome sequencing yields codominant variation from identifiable gene sequences (excepting cases of allelic or homoeologue expression dominance; e.g., Adams et al., 2003).

Transcriptome sequencing begins with total RNA extraction from tissue(s) of interest. Since ribosomal RNA makes up the vast majority of total RNA in most preparations (often >90%; Raz et al., 2011), most library construction methods reduce rRNA abundance and enrich the protein-coding mRNA fraction by oligo(dT) selection of poly(A)+ mRNA. This approach has the advantage of enriching the polyadenylated portion of the transcriptome (which includes the majority of the expressed gene space) so that it makes up the majority of the RNA pool; conversely, it has the disadvantage of undersampling nonpolyadenylated RNAs. If nonpolyadenylated RNAs are of interest, rRNA can be selectively depleted using hybridization-based probes (e.g., RiboMinus™ from Invitrogen, Ribo-Zero™ from Epicenter). Since a large proportion of the transcriptome is made up of a relatively small number of highly expressed transcripts, many transcriptome sequencing studies use a duplex-specific nuclease to perform "double-stranded normalization" (Shagin et al., 2002), a process that evens the representation of transcripts in an RNA pool. While double-stranded normalization makes it easier to sample rare transcripts on lower-throughput sequencing platforms (e.g., Roche/454), it does distort the relative abundance of transcripts in a pool, and this type of quantitative information is key for developing detailed transcriptome atlases (Severin et al., 2010; Li et al., 2010).

Library preparation is specific to each sequencing technology but typically involves fragmentation, first strand synthesis with reverse transcriptase and random hexamer or oligo(dT) priming, second strand synthesis, and ligation to platform-specific adapters to create a double-stranded DNA library. Strand-specific library construction methods are also available (summarized in Levin et al., 2010), and these offer benefits in characterizing novel transcriptomes (unambiguously identifies transcribed strands), as well as de novo transcriptome assembly

TABLE 4. Summary of genome size, gene and transcript content of plants at Phytozome, and predictions for loblolly pine. Estimated transcriptome sizes for all taxa assume the average transcript length of *Arabidopsis thaliana* (2343 bp; Gan et al., 2011).

Taxon	Genome size (Mbp)	Loci	Transcripts[a]	Transcriptome size (Mbp, estimated)	Transcriptome/genome ratio
Chlamydomonas reinhardtii	112	17114	17114	40	0.36
Arabidopsis thaliana	135	27416	35386	83	0.61
Carica papaya	135	27332	27796	65	0.48
Volvox carteri	138	14491	14542	34	0.25
Cucumis sativus	203	21491	32528	76	0.38
Arabidopsis lyrata	207	NA	32670	77	0.37
Selaginella moellendorffii	212	22273	22285	52	0.25
Prunus persica	227	27864	28702	67	0.30
Medicago truncatula	241	50962	53423	125	0.52
Brachypodium distachyon	272	25532	32255	76	0.28
Citrus clementina	296	25385	35976	84	0.28
Aquilegia coerulea	302	25784	27583	65	0.21
Citrus sinensis	319	25376	46147	108	0.34
Mimulus guttatus	321	26718	28282	66	0.21
Oryza sativa	372	40838	51258	120	0.32
Ricinus communis	400	NA	31221	73	0.18
Populus trichocarpa	403	40668	45033	106	0.26
Setaria italica	405	35471	40599	95	0.23
Physcomitrella patens	480	32272	38354	90	0.19
Vitis vinifera	487	26346	26346	62	0.13
Eucalyptus grandis	691	44974	54935	129	0.19
Sorghum bicolor	697	34496	36338	85	0.12
Manihot esculenta	760	30666	34151	80	0.11
Glycine max	975	66153	NA	155	0.16
Zea mays	2,400	80000	106000	248	0.10
Pinus taeda[b]	21600	224300	NA	526	0.02

[a] Includes alternative isoforms.
[b] Estimated in Rabinowicz et al., 2005.

because it reduces the memory footprint (Grabherr et al., 2011) and helps identify antisense transcripts. After sequencing, reads can be assembled de novo to explore novel transcript discovery using a number of available software packages (Zerbino and Birney, 2008; Birol et al., 2009; Martin et al., 2010; Robertson et al., 2010; Grabherr et al., 2011). Once a reference is available, reads can be aligned to it using existing software (Langmead et al., 2009; Li et al., 2009; Li and Durbin, 2010; Lunter and Goodson, 2010; Trapnell et al., 2010).

Transcriptome sequencing for SNP detection—SNP discovery is the development of a polymorphic panel of SNPs that are used to assay populations or closely related species for nucleotide polymorphisms that are associated with a particular phenotype, to define population or geographic structure, or to track evolutionary history. SNP detection from transcriptome sequencing data are similar in some regards to other methods in that sequencing depth is an important index of the "quality" of a SNP; because of the extreme range of transcript representation in a transcriptome, however, read depths across SNPs are expected to be nonuniform.

During development of a transcriptome-sequencing-based SNP panel, researchers often consider pooling multiple unbarcoded samples (genotypes) to maximize SNP discovery per unit of sequencing cost. Although a theoretical foundation has been set to address this issue for DNA templates (Futschik and Schlötterer, 2010), this approach relies on simplifying assumptions that are violated in transcriptome sequencing (Cutler and Jensen, 2010). An important assumption is the equal sequencing representation of all individuals, necessitating equimolar pooling of target loci, something that cannot be accomplished in transcriptomes due to differences in transcript abundance among sampled individuals and possibly even among alleles within an individual (i.e., allelic dominance). Since the interpretations from pooled RNA templates are less clear than those from DNA templates, we recommend caution before adopting transcriptome pooling for polymorphism discovery.

RNA editing, primarily C-to-U substitutions, provide another departure of the transcriptome relative to the genome (Picardi et al., 2010), but can lead to false positive SNP identification when mapping RNA-derived sequence reads to a genomic reference. This can theoretically be circumvented by restricting SNP discovery projects to DNA vs. DNA or RNA vs. RNA comparisons, although the added sequence costs of RNA-based sequencing make this a less attractive alternative than rigorous SNP validation by other methods.

Transcriptome sequencing for expression profiling—Another goal of RNA-seq studies is to examine differential expression, or the relative transcript abundance and correlated changes in transcripts from a particularly pathway or network, and relate these to a phenotype (Anders and Huber, 2010; Robinson and Oshlack, 2010; Trapnell et al., 2010; Auer and Doerge, 2011; Di et al., 2011; Li et al., 2010). While polymorphism may be present in RNA-seq data, it is tolerated by mismatch parameterization during the alignment step and is typically ignored in downstream analyses (Langmead et al., 2009; Li and Durbin, 2010). The goal of differential expression analysis is to compare the number of sequencing reads mapped among different individuals for the same transcript or among different exons in the same locus when looking for alternative splicing.

A fundamental question to all next-generation sequencing projects is how many reads are needed to adequately answer project goals. For RNA-seq studies, the answer varies by application (transcript detection vs. transcript sequence assembly) and the complexity of the transcriptome, but generalizations can be made. For transcript detection using the Illumina platform, published studies report values between 3 to 100 million microreads (Wilhelm and Landry, 2009). For the simple transcriptome of *Saccharomyces cerevisiae*, it takes as few as four million reads to detect 80% of the known open reading frames (Wang et al., 2009). A comparative study of gene expression in more complex plant transcriptomes (*Arabidopsis thaliana*, *Brachypodium distachyon*, and *Zea mays*) showed that 32 million 1 × 32 bp reads were required to detect expression for 88% of the known cDNAs from these species (Priest et al., 2010). Our own work on the conifer Douglas-fir (*Pseudotsuga menziesii*; Pinaceae) indicates that 10 million mapped 1 × 80 bp Illumina microreads are required to detect 88% of the 38000 predicted transcripts assembled for this species (Fig. 5A; G. Howe, Oregon State University, and B. Knaus, unpublished data; http://www.fs.fed.us/pnw/olympia/silv/ccto/index.html). However, transcript detection alone does not address the issue of accurate quantification of expression, as a very large number of transcripts (~4300; Fig. 5B) are represented by five or fewer mapped reads in the total sample of ~10 million mapped microreads. These values are too low for accurate quantitative analysis, and the treatment of low abundance transcripts is currently a topic of active debate. If analysis is restricted to transcripts showing a reasonable minimum number of mapped reads (e.g., 25), a large number of loci (>20200) can still be retained for analysis. At a sample size of 10 million microreads, the current capacity of the Illumina HiSeq2000 sequencer (>200 million reads per lane) should allow at least 10 multiplexed transcriptomes to be surveyed in a single pool.

If the goal of the RNA-seq study is to instead assemble transcript sequences for polymorphism detection and sequence characterization, then the required sequencing depth is substantially greater. In their comparative study, Priest et al. (2010) found that ~94 million 1 × 32 bp reads were required to provide 1× coverage of 80% of the known transcriptome; this is nearly 3-fold more data than is required to detect transcript abundance. Due to the uneven representation of RNAs in the transcript pool, however, many genes can be assembled and screened for polymorphism with a much smaller sample of sequencing reads. In our example from Douglas-fir, 10 million microreads would provide >25× depth of coverage for ~3860 transcripts (Fig. 5B). This generalization is sensitive to many assumptions (e.g., low level of contaminating rRNA or adapter sequences; comparable transcript mapping densities across different samples), but it makes the point that if the goal is to opportunistically scan a large number of transcripts for sequence variation, transcriptome sequencing samples can also be multiplexed at moderately high levels (e.g., 10×). This example also highlights the converse situation; if the goal is to analyze sequence variation from *specific* low-abundance transcripts, very large amounts of sequencing will need to be applied to obtain adequate depth for assembly and accurate counts.

Considerations—Analysis of RNA-seq data poses a number of analytical challenges, many of which arise from to the biological complexity of RNA. First, the presence of multiple isoforms makes de novo assembly more computationally challenging, requiring as much as 1 GB of memory (RAM) per million input reads (Grabherr et al., 2011); at this scale, de novo assembly from one lane of Illumina HiSeq data could require ~200–300 GB of RAM. Second, since different isoforms share identical exons (Fig. 6), reference-based alignment of microreads arising from different isoforms will only map reads that are shared with the reference; reads that are specific to one or more novel isoforms (such as reads mapping to retained introns or unique splice junctions) will not map to the reference. It therefore appears that studies emphasizing novel isoforms require the use of all reads in an experiment, and these need to be mapped to databases of possible splice junctions (e.g., Li et al., 2010) or possibly de novo assemblies of multiple unique references (Gan et al., 2011). Third, the presence of nearly identical

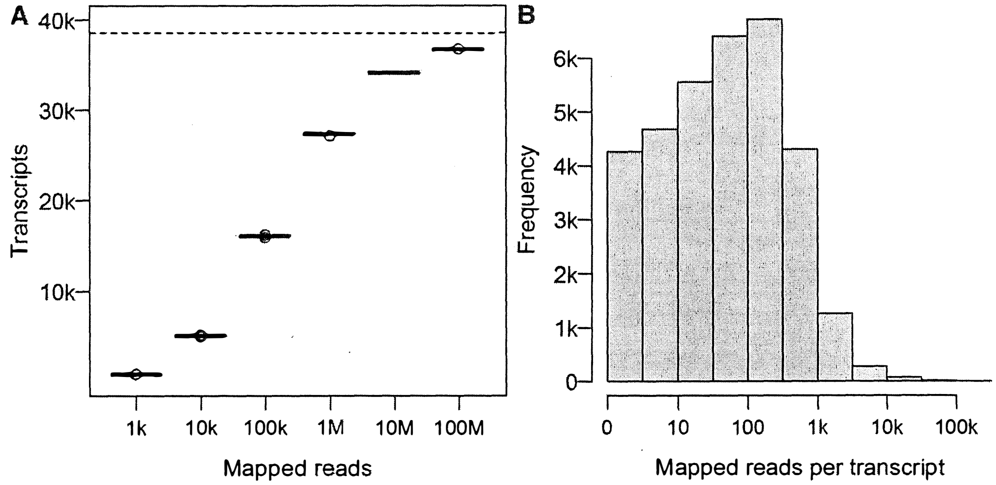

Fig. 5. Sequencing depth from transcriptome sequencing experiments. (A) Rarefaction plot showing the relationship of detected transcripts as a function of the size of a read pool, up to 100 million reads. The dashed line represents the maximum number of transcripts (38589) in the *Pseudotsuga menziesii* reference. Boxplots represent the variability in the number of transcripts detected in an RNA-seq experiment from needle tissue; outliers are indicated with circles. (B) Histogram of the number of reads mapping to each transcript from Douglas-fir needles in an RNA-seq experiment with 10 million mapped reads. In a data set of this size, 5100 transcripts are not sequenced, and 4750 transcripts are represented by five or fewer mapped reads. In contrast, 23060 transcripts are represented by 25 or greater mapped reads.

paralogous sequence strings across multiple sites in the genome (Oshlack et al., 2010) complicates transcriptome sequencing and counting studies. Because many reference-based aligners were developed with genomic data in mind, they handle these "multimap" situations by omitting a read if it maps to a parameterized maximum number of locations (e.g., N = 10). If the number of matching alignments is below this number, the read is randomly assigned to one of N possible positions. The net result of a multimap situation is that the true signal contained in differentially expressed transcripts can be diluted, with an underestimation in the abundance of overexpressed transcripts, and an overestimation of underexpressed transcripts. Paralogous genes that retain a high level of sequence similarity among copies can also be prone to multimap situations.

A final but important consideration of transcriptome sequencing is that gene expression and transcript abundance is dynamic. Not only can differentially spliced transcripts ("isoforms") arise from the same source gene (Fig. 6), the abundance of transcripts and isoforms fluctuate hourly, seasonally, by tissue type and developmental state, and in response to environmental conditions. A circadian rhythm to expression has been reported for 25% of the protein-coding genes in *Arabidopsis thaliana* (Hazen et al., 2009), while as many as 60% of genes in *Oryza sativa* subsp. *japonica* and *Populus trichocarpa* have been reported to be similarly regulated (Filichkin et al., 2011). A similarly large portion of transcribed genes fluctuate with developmental state (Li et al., 2010; Zenoni et al., 2010) and tissue type (Severin et al., 2010; Portnoy et al., 2011). Due to this wide range of expression variation, the choice of tissue collection is more complicated than for DNA-based studies, and care needs to be exercised so that comparative studies include samples from similar collection times, developmental states, and environments. These factors are of critical importance when designing transcript-counting experiments, as variation in collection time, developmental stage, and possibly taxonomic divergence may confound interpretations. Exploration of these factors, as well as approaches for the design and analysis of RNA-seq experiments, is an actively evolving area of research (Auer and Doerge, 2010).

Examples—Unlike the other methods detailed in this paper, transcriptome sequencing is well established in the literature, with over 50 published examples of transcriptome-sequencing studies focusing on plants (see online Appendix S2). Initial RNA-seq efforts focused on model organisms and single-taxon studies, but these have recently expanded to include diverse taxa and often compare transcriptome sequence and expression differences between varieties and subspecies within species, between congeneric species, and between closely related genera. RNA-seq has added substantially to our knowledge of already well-characterized models, such as rice (*Oryza sativa*; Poaceae). For example, Lu et al. (2010) used the Illumina platform to explore differences among two subspecies of rice, *Oryza sativa* subsp. *indica* and *O. s.* subsp. *japonica*. From libraries ranging from 23.6–30.9 million reads, they identified over 60000 SNPs and observed 3464 genes as differentially transcribed among the subspecies. Remarkably, this single study validated gene models for 46000 genes and identified over 15000 novel transcripts, 50% of which have no homolog in public protein databases. The impact on less-well characterized species is just as impressive. For example, RNA-seq was used to assemble a reference transcriptome of 20250 transcripts for big sagebrush (*Artemisia tridentata* subsp. *tridentata*; Asteraceae), and to identify SNPs between the reference and related sagebrush subspecies (Bajgain et al., 2011). The 20952 inferred SNPs identified in this study are currently being used to examine the evolutionary history of sagebrush and the molecular basis of adaptation between subspecies. RNA-seq is being used to gain an understanding of the evolutionary origin and developmental complexity of unique anatomical modifications, such as the characteristic "traps" of the bladderwort *Utricularia gibba* (Lentibulariaceae), which have been contrasted with other organs using the Roche/454 platform (Ibarra-Laclette et al., 2011). In this same study, de novo assembled contigs from chloroplast and mitochondrion genes and a supermatrix of 100 nuclear genes were also used to infer phylogeny. Genes involved in other key evolutionary transitions, such as the shift from obligate outcrossing to self-compatibility in *Eichhornia paniculata* (Pontederiaceae) (Ness et al., 2011), are also being evaluated for differential expression. As with all previously described studies, tens of thousands of SNPs were identified in *Eichhornia*, and these will enable detailed investigation into the evolution of breeding system in this group.

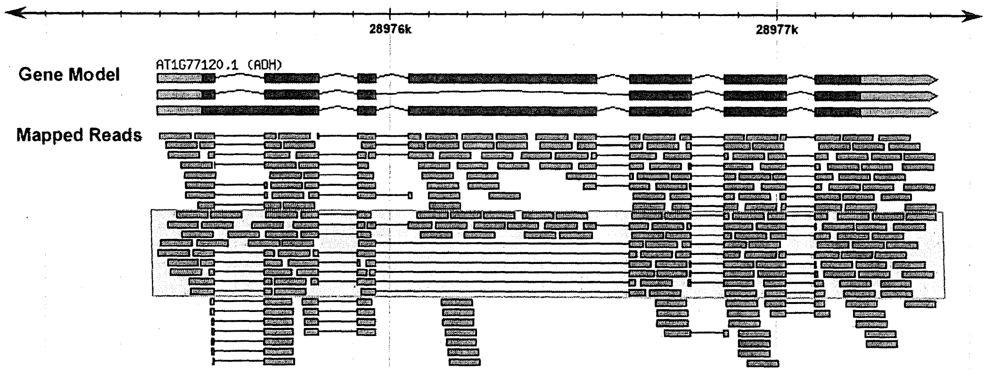

Fig. 6. Example of transcriptome read mapping on alternatively spliced genes. Three gene models are presented in blue (UTRs in grey), based on *Arabidopsis thaliana* alcohol dehydrogenase (ADH; top panel). Two hypothetical isoform models were created to demonstrate exon skipping (middle model) and intron retention (bottom model). Simulated reads (green) were mapped to each model to demonstrate the pattern of read alignment. A total of 250 reads were mapped, with 100 mapped to the original gene model, and 150 mapped to the alternative splicing models (highlighted in red box). Reads spanning splice junctions are indicated by filled rectangles joined by a horizontal line.

The flexibility and adaptability of RNA-seq makes it amenable to addressing a host of questions related to plant science, and the breadth of these applications is certain to expand. As the throughput of sequencers steadily increases and the cost of sequencing on a per-nucleotide basis drops, many groups are initiating large-scale, comparative transcriptome projects involving large numbers of samples. One of the most ambitious is the 1 KP Project (http://www.onekp.com/project.html), an effort to sequence more than one-thousand plant transcriptomes from a variety of tissues and across an evolutionary swath that includes all land plants. As transcriptome drafts from this project are finalized, they are sure to provide important reference transcriptomes to give plant researchers a large set of tools to develop PCR-based and hybridization-based markers for addressing taxon-specific questions.

TARGETED-ENRICHMENT: SCALING QUESTIONS TO METHODS

There are many options available for targeted enrichment when conducting NGS-based studies, but are some better suited for specific questions? The answer to this depends on many factors, including the nature of the study (physiological, population genetic, phylogenetic), the kinds of data produced, the number of samples included in the study, the number of loci targeted, the scaling efficiency of a method to high throughout sequencing, and the availability of computational resources or specialized instrumentation. To help address this question, we have summarized the level of biological complexity each method is suited to address and directly compared methods in a hypothetical scenario that involves analyzing a fixed number of individuals (96) for a range of target sizes (50–500 kbp) (Table 5; see Appendix S3 with the online version of this article).

Generally speaking, restriction digestion methods are the easiest and least expensive to perform among all enrichment methods; they are also efficient with regard to time and resources, requiring one reaction per sample. Restriction methods may have the narrowest range of application, being limited primarily to the detection of genetic variation in closely related individuals (e.g., mapping populations, natural populations). With increasing taxonomic and genetic divergence, the assumption of restriction site conservation is likely to be violated, and mutations within restriction sites and rearrangements around restriction sites will produce unsequenceable fragments (null alleles). These will increase the frequency of missing data and complicate downstream analyses (Davey et al., 2011). In limited cases, restriction enzyme enrichment methods may be useful at examining interspecific variation (such as in hybrid contact zones). Of the available methods, GBS and GR-RSC are the least expensive, while RAD is more expensive due to the cost of specialized adapters with modified nucleotides. All methods yield a large amount of data that is equally divided among the individuals included in a multiplex, so there is no difference in the cost of examining small (50 kbp) vs. large (0.5 Mbp) targets for 96 individuals.

At the other end of the spectrum, transcriptome sequencing may be the most challenging and expensive enrichment method, although these disadvantages may be outweighed by its breadth of application, since the method can be used to address questions as narrowly focused as gene expression differences within an individual and as broad as sequence analysis for comparative phylogenetic studies. The technical challenge of transcriptome sequencing arises primarily from its requirement for fresh material, the precautions of working with RNA, and the informatic uncertainty imposed by alternative splicing of transcripts. From a cost perspective, the added costs for transcriptome sequencing are due to the materials required for poly(A) isolation and cDNA construction, although low-cost strand-specific RNA-Seq approaches have been reported (Zhong et al., 2011). On the positive side, RNA-seq requires one reaction per sample, and it scales efficiently with targets in vast excess of 500 kbp. In our estimates, the major expense in transcriptome sequencing is the cost of the actual sequencing run; for this reason, RNA-seq will benefit disproportionately from the growing capacity of NGS, and growth in capacity will translate into increased multiplexing capacity and reduced costs.

Hybridization-based enrichment has a similarly broad utility as transcriptome sequencing, since it can be used to enrich targets genomic DNA and cDNA libraries in a manner that preserves the relative ratios of the original target (or transcript) concentrations (Levin et al., 2009). Like transcriptome sequencing, hybridization enrichment is more challenging than PCR or restriction enzyme methods because the enrichment probes are usually RNA, and the process is lengthy. Hybridization enrichment is expensive for the isolation of small target pools, but it

Table 5. Comparative efficiencies of different methods for sequencing targets of different sizes from 96 samples. In general, restriction-enzyme-based methods have the narrowest range of application due to the requirement of restriction site conservation; restriction methods are also the least expensive of all methods described to date. Transcriptome sequencing has the widest range of applications, but it is the most expensive of the compared methods. PCR methods are intermediate in price, but they require a large number of reactions to execute.

	Focal area for different enrichment methods [a]			Approximate cost/sample to enrich and sequence targets (No. reactions required)	
Enrichment method	Differences between tissues/individuals	Differences between populations/species	Differences between species/genera	50 kbp	500 kbp
Short PCR (500 bp/amplicon)	–	±	+	$118 (9600)	$1836 (96000)
Long PCR (5000 bp/amplicon)	–	±	+	$163 (960)	$373 (9600)
Microfluidic Short PCR (500 bp/amplicon)	–	+	+	$53 (9600)	$528 (96000)
Hybridization (2 Mbp probe synthesis)	±	+	+	$186 (96)	$186 (96)
Restriction—GBS or GR-RSC	–	+	—	$25 (96)	$40 (96)
Restriction—RAD	–	+	–	$108 (96)	$124 (96)
Transcriptome	+	+	+	$334 (96)	$334 (96)

[a] Focal areas are noted as either "–" (method is not well suited for application), "±" (method can be applied but more efficient methods exist), or "+" (method is well suited for application).

is the *least* expensive method for enriching target pools greater than 500 kbp in complexity. Perhaps more than any other method, hybridization enrichment scales to large sample sizes, since methods exist for high-throughput library construction (Fisher et al., 2011) and multiple barcoded samples can be hybridized simultaneously (Nijman et al., 2010). Importantly, hybridization enrichment has become a standard application in human genomics, and this large community will drive improvements in product and software development. In contrast to transcriptome sequencing, the major expense in hybridization enrichment is the cost of biotinylated hybridization probes, so development of multiplex hybridization or cost-effective probe synthesis methods are the most important factors for controlling project costs.

The efficiency of PCR as an enrichment method depends primarily upon the number of loci sampled and the amplification platform, due to the proportional increase in primer and reagent costs with increasing targets. If the goal is to sequence a small pool of targets (e.g., 50 kbp), all PCR strategies—direct sequencing of short PCR products, sequencing of long PCR libraries, and microfluidic PCR—appear equally cost effective. Absent from our calculations is the time and expense associated with amplifying 9600 short (≤500 bp) or 960 long (≥5 kbp) PCR products for these studies, and these may be prohibitively high. Based on price alone, microfluidic PCR appears to be a cost-effective method for sequencing larger pools of targets up to 150 kbp. As noted, the number of amplicons required to enrich these targets by microfluidic methods (28 800 for short PCR products; 2880 for long PCR products) still demands more effort than other enrichment approaches. Due to the comparatively high cost per "PCR data point", we predict that the coupling of PCR with next-generation sequencing will be most productive when PCR amplicons are maximally informative (e.g., targeting previously sampled spacers or introns in phylogenetic studies) or if they are being compared to previous data that are difficult to produce using other approaches. PCR methods will be less attractive in studies where individual amplicons show limited variation, as is often the case in studies examining intraspecific genetic variation (e.g., linkage mapping, population-genetic, phylogeographic comparisons).

Irrespective of the method adopted for targeted sequencing, all of these approaches make efficient use of sequencing and data storage resources by maximizing the production of sequence reads for genes and targets of interest, and these efficiencies may yield underappreciated savings. Low-depth genome sequencing approaches can be used to sequence high-copy targets from a large numbers of samples, even if the targets of interest are rare in the total genome pool (e.g., 1% or lower for cpDNA; Straub et al., 2012). However, these approaches require a large investment into sequencing reagents and data storage capacity for microread sequences that are sampled at such low depth that they cannot be easily assembled or analyzed. This represents a hidden cost to the user, because lost sequencing capacity could be redirected toward gathering additional information (samples, targets) for a comparatively small increase in sample preparation costs.

At the moment, it is unclear which of these target-enrichment methodologies will become widely adopted in "next-generation" plant research, or if these approaches will be supplanted by even more efficient methods. What is certain is that enrichment methods, which were only recently dominated by low-to-medium throughput technologies, *will* be expanded and improved to take advantage of the stunning growth in NGS instruments. One only needs to look at the human genomics community to be inspired by the scale that next-generation target enrichment can be conducted in a single laboratory (e.g., hundreds of exome samples per week; Fisher et al., 2011) and to see that enrichment strategies are as relevant as they were when next-generation sequencing was first introduced. These methods are certain to redefine what is possible in future plant research, and they will help hasten an era that will be both exciting and unsettling, where sequencing run capacity will be measured in trillions of bases, transcriptome and draft genome sequences will be abundant for nearly all plant groups, and the cost to generate targeted data will be essentially free when calculated at the per-gene level.

LITERATURE CITED

Adams, K. L., R. C. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences, USA* 100: 4649–4654.

Albert, T. J., M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. A. Richmond, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nature Methods* 4: 903–905.

Alvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417–434.

Alverson, A. J., X. Wei, D. W. Rice, D. B. Stern, K. Barry, and J. D. Palmer. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* 27: 1436–1448.

Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11: R106.

Arai-Kichise, Y., Y. Shiwa, H. Nagasaki, K. Ebana, H. Yoshikawa, M. Yano, and K. Wakasa. 2011. Discovery of genome-wide DNA polymorphisms in a landrace cultivar of japonica rice by whole-genome sequencing. *Plant & Cell Physiology* 52: 274–282.

Arthofer, W., S. Schuler, F. M. Steiner, and B. C. Schlick-Steiner. 2010. Chloroplast DNA-based studies in molecular ecology may be compromised by nuclear-encoded plastid sequence. *Molecular Ecology* 19: 3853–3856.

Ashelford, K., M. E. E. Eriksson, C. M. M. Allen, R. D'Amore, M. Johansson, P. Gould, S. Kay, A. J. Millar, N. Hall, and A. Hall. 2011. Full genome re-sequencing reveals a novel circadian clock mutation in *Arabidopsis*. *Genome Biology* 12: R28.

Auer, P. L., and R. W. Doerge. 2010. Statistical design and analysis of RNA sequencing data. *Genetics* 185: 405–416.

Auer, P. L., and R. W. Doerge. 2011. A two-stage Poisson model for testing RNA-seq data. *Statistical Applications in Genetics and Molecular Biology* 10: 1–28. doi:10.2202/1544-6115.1627

Babik, W., P. Taberlet, M. J. Ejsmond, and J. Radwan. 2009. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources* 9: 713–719.

Bainbridge, M., M. Wang, D. Burgess, C. Kovar, M. Rodesch, M. D'Ascenzo, J. Kitzman, et al. 2010. Whole exome capture in solution with 3 Gbp of data. *Genome Biology* 11: R62.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3: e3376.

Bajgain, P., B. A. Richardson, J. Price, R. C. Cronn, and J. A. Udall. 2011. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). *BMC Genomics* 12: 370.

Bansal, V., R. Tewhey, E. M. LeProust, and N. J. Schork. 2011. Efficient and cost effective population resequencing by pooling and in-solution hybridization. *PLoS ONE* 6: e18353.

Bartram, A. K., M. D. Lynch, J. C. Stearns, G. Moreno-Hagelsieb, and J. D. Neufeld. 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology* 77: 3846–3852.

BASHIARDES, S., R. VEILE, C. HELMS, E. R. MARDIS, A. M. BOWCOCK, AND M. LOVETT. 2005. Direct genomic selection. *Nature Methods* 2: 63–69.

BATLEY, J., G. BARKER, H. O'SULLIVAN, K. J. EDWARDS, AND D. EDWARDS. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology* 132: 84–91.

BINLADEN, J., M. T. P. GILBERT, J. P. BOLLBACK, F. PANITZ, C. BENDIXEN, R. NIELSEN, AND E. WILLERSLEV. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2: e197.

BIROL, I., S. D. JACKMAN, C. B. NIELSEN, J. Q. QIAN, R. VARHOL, G. STAZYK, R. D. MORIN, ET AL. 2009. *De novo* transcriptome assembly with ABySS. *Bioinformatics* 25: 2872–2877.

BRADLEY, R. D., AND D. M. HILLIS. 1997. Recombinant DNA sequences generated by PCR amplification. *Molecular Biology and Evolution* 14: 592–593.

BRIGGS, A. W., J. M. GOOD, R. E. GREEN, J. KRAUSE, T. MARICIC, U. STENZEL, C. LALUEZA-FOX, ET AL. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325: 318–321.

BUNDOCK, P. C., F. G. ELIOTT, G. ABLETT, A. D. BENSON, R. E. CASU, K. S. AITKEN, AND R. J. HENRY. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal* 7: 347–354.

BURBANO, H. A., E. HODGES, R. E. GREEN, A. W. BRIGGS, J. KRAUSE, M. MEYER, J. M. GOOD, ET AL. 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328: 723–725.

CAPORASO, J. G., C. L. LAUBER, W. A. WALTERS, D. BERG-LYONS, C. A. LOZUPONE, P. J. TURNBAUGH, N. FIERER, AND R. KNIGHT. 2010. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences, USA* 108 (Suppl 1): 4516–4522.

CHENG, S., C. FOCKLER, W. M. BARNES, AND R. HIGUCHI. 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences, USA* 91: 5695–5699.

CHUTIMANITSAKUN, Y., R. W. NIPPER, A. CUESTA-MARCOS, L. CISTUE, A. COREY, T. FILICHKINA, E. A. JOHNSON, AND P. M. HAYES. 2011. Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. *BMC Genomics* 12: 4.

CRAIG, D. W., J. V. PEARSON, S. SZELINGER, A. SEKAR, M. REDMAN, J. J. CORNEVEAUX, T. L. PAWLOWSKI, T. LAUB, G. NUNN, AND D. A. STEPHAN. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* 5: 887–893.

CRONN, R., M. CEDRONI, T. HASELKORN, C. GROVER, AND J. F. WENDEL. 2002. PCR-mediated recombination in amplification products derived from polyploid cotton. *Theoretical and Applied Genetics* 104: 482–489.

CRONN, R., A. LISTON, M. PARKS, D. S. GERNANDT, R. SHEN, AND T. MOCKLER. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36: e122.

CUTLER, D. J., AND J. D. JENSEN. 2010. To pool, or not to pool? *Genetics* 186: 41–43.

DAHL, F., J. STENBERG, S. FREDRIKSSON, K. WELCH, M. ZHANG, M. NILSSON, D. BICKNELL, W. F. BODMER, R. W. DAVIS, AND H. JI. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proceedings of the National Academy of Sciences, USA* 104: 9387–9392.

DAVEY, J. W., P. A. HOHENLOHE, P. D. ETTER, J. Q. BOONE, J. M. CATCHEN, AND M. L. BLAXTER. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics* 12: 499–510.

DI, Y., D. W. SCHAFER, J. S. CUMBIE, AND J. H. CHANG. 2011. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology* 10(1): 24.

DUMINIL, J., M. H. PEMONGE, AND R. J. PETIT. 2002. A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial DNA. *Molecular Ecology Notes* 2: 428–430.

EDWARDS, M. C., AND R. A. GIBBS. 1994. Multiplex PCR: Advantages, development and applications. *Genome Research* 3: S65–S75.

ELSHIRE, R. J., J. C. GLAUBITZ, Q. SUN, J. A. POLAND, K. KAWAMOTO, E. S. BUCKLER, AND S. E. MITCHELL. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6: e19379.

ERMINI, L., C. OLIVIERI, E. RIZZI, G. CORTI, R. BONNAL, P. SOARES, S. LUCIANI, ET AL. 2008. Complete mitochondrial genome sequence of the Tyrolean Iceman. *Current Biology* 18: 1687–1693.

EMERSON, K. J., C. R. MERZ, J. M. CATCHEN, P. A. HOHENLOHE, W. A. CRESKO, W. E. BRADSHAW, AND C. M. HOLZAPFEL. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences, USA* 107: 16196–16200.

ETTER, P. D., J. L. PRESTON, S. BASSHAM, W. A. CRESKO, AND E. A. JOHNSON. 2011. Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE* 6: e18561.

FILICHKIN, S. A., G. BRETON, H. D. PRIEST, P. DHARMAWARDHANA, P. JAISWAL, S. E. FOX, T. P. MICHAEL, ET AL. 2011. Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. *PLoS ONE* 6: e16907.

FILICHKIN, S. A., H. D. PRIEST, S. A. GIVAN, R. SHEN, D. W. BRYANT, S. E. FOX, W. K. WONG, AND T. C. MOCKLER. 2009. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana. Genome Research* 20: 45–58.

FISHER, S., A. BARRY, J. ABREU, B. MINIE, J. NOLAN, T. DELOREY, G. YOUNG, ET AL. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology* 12: R1.

FU, Y., N. M. SPRINGER, D. J. GERHARDT, K. YING, C.-T. YEH, W. WU, R. SWANSON-WAGNER, ET AL. 2010. Repeat subtraction-mediated sequence capture from a complex genome. *Plant Journal* 62: 898–909.

FUTSCHIK, A., AND C. SCHLÖTTERER. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186: 207–218.

GALAN, M., E. GUIVIER, G. CARAUX, N. CHARBONNEL, AND J. F. COSSON. 2010. A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics* 11: 296.

GAN, X., O. STEGLE, J. BEHR, J. G. STEFFEN, P. DREWE, K. L. HILDEBRAND, R. LYNGSOE, ET AL. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana. Nature* 477: 419–423.

GARBER, K. 2008. Fixing the front end. *Nature Biotechnology* 26: 1101–1104.

GARG, K., P. GREEN, AND D. A. NICKERSON. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Research* 9: 1087–1092.

GNIRKE, A., A. MELNIKOV, J. MAGUIRE, P. ROGOV, E. M. LEPROUST, W. BROCKMAN, T. FENNELL, ET AL. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 27: 182–189.

GRABHERR, M. G., B. J. HAAS, M. YASSOUR, J. Z. LEVIN, D. A. THOMPSON, I. AMIT, X. ADICONIS, ET AL. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

HAMADY, M., J. WALKER, J. HARRIS, N. GOLD, AND R. KNIGHT. 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* 5: 235–237.

HARISMENDY, O., AND K. FRAZER. 2009. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *BioTechniques* 46: 229–231.

HAZEN, S. P., F. NAEF, T. QUISEL, J. M. GENDRON, H. CHEN, J. R. ECKER, J. O. BOREVITZ, AND S. A. KAY. 2009. Exploring the transcriptional landscape of plant circadian rhythms using genome tiling arrays. *Genome Biology* 10: R17.

HOHENLOHE, P. A., S. J. AMISH, J. M. CATCHEN, F. W. ALLENDORF, AND G. LUIKART. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* 11 (supplement 1): 117–122.

HOHENLOHE, P. A., S. BASSHAM, P. D. ETTER, N. STIFFLER, E. A. JOHNSON, AND W. A. CRESKO. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* 6: e1000862.

HUGHES, T. R., M. MAO, A. R. JONES, J. BURCHARD, M. J. MARTON, K. W. SHANNON, S. M. LEFKOWITZ, ET AL. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* 19: 342–347.

IBARRA-LACLETTE, E., V. A. ALBERT, C. A. PÉREZ-TORRES, F. ZAMUDIO-HERNÁNDEZ, M. D. ORTEGA-ESTRADA, A. HERRERA-ESTRELLA, AND L. HERRERA-ESTRELLA. 2011. Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biology* 11: 101.

ILLUMINA. 2011. Amplicon sequencing from FFPE tissues on the MiSeq system. Illumina, Inc., San Diego, California, USA.

JENNINGS, T. N., B. J. KNAUS, T. D. MULLINS, S. M. HAIG, AND R. C. CRONN. 2011. Multiplexed microsatellite recovery using massively parallel sequencing. *Molecular Ecology Resources* 11: 1060–1067. doi: 10.1111/j.1755-0998.2011.03033.x

JEX, A. R., R. S. HALL, D. T. J. LITTLEWOOD, AND R. B. GASSER. 2010. An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucleic Acids Research* 38: 522–533.

KANAGAWA, T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* 96: 317–323.

KAWAKAMI, T., S. C. STRAKOSH, Y. ZHEN, AND M. C. UNGERER. 2010. Different scales of Tyl/copia-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species. *Heredity* 104: 341–350.

KEENEY, S. 2011. Long-PCR amplification of human genomic DNA. *Methods in Molecular Biology* 688: 67–74.

KELLOGG, E. A., AND J. L. BENNETZEN. 2004. The evolution of nuclear genome structure in seed plants. *American Journal of Botany* 91: 1709–1725.

KIRCHER, M., P. HEYN, AND J. KELSO. 2011. Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics* 12: 382.

KNAPP, M., AND M. HOFREITER. 2010. Next generation sequencing of ancient DNA: Requirements, strategies and perspectives. *Genes* 1: 227–243.

KNAUS, B., R. CRONN, A. LISTON, K. PILGRIM, AND M. SCHWARTZ. 2011. Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. *BMC Ecology* 11: 10.

LAM, H.-M., X. XU, X. LIU, W. CHEN, G. YANG, F.-L. WONG, M.-W. LI, ET AL. 2011. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* 42: 1053–1059.

LANGMEAD, B., C. TRAPNELL, M. POP, AND S. L. SALZBERG. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

LEVIN, J. Z., M. F. BERGER, X. ADICONIS, P. ROGOV, A. MELNIKOV, T. FENNELL, C. NUSBAUM, L. A. GARRAWAY, AND A. GNIRKE. 2009. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biology* 10: R115.

LEVIN, J. Z., M. YASSOUR, X. ADICONIS, C. NUSBAUM, D. A. THOMPSON, N. FRIEDMAN, A. GNIRKE, AND A. REGEV. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* 7: 709–715.

LI, H., AND R. DURBIN. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595.

LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, G. ABECASIS, AND R. DURBIN. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

LI, P., L. PONNALA, N. GANDOTRA, L. WANG, Y. SI, S. L. TAUSTA, T. H. KEBROM, ET AL. 2010. The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* 42: 1060–1067.

LISTER, R., B. D. GREGORY, AND J. R. ECKER. 2009. Next is now: New technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology* 12: 107–118.

LOVETT, M., J. KERE, AND L. M. HINTON. 1991. Direct selection: A method for the isolation of cDNAs encoded by large genomic regions. *Proceedings of the National Academy of Sciences, USA* 88: 9628–9632.

LU, T., G. LU, D. FAN, C. ZHU, W. LI, Q. ZHAO, Q. FENG, Y. ZHAO, Y. GUO, W. LI, X. HUANG, AND B. HAN. 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research* 20: 1238–1249.

LUNTER, G., AND M. GOODSON. 2010. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21: 936–939.

MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER, A. KUMAR, E. HOWARD, J. SHENDURE, AND D. J. TURNER. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118.

MARDIS, E. R. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470: 198–203.

MARKOULATOS, P., N. SIAFAKAS, AND M. MONCANY. 2002. Multiplex polymerase chain reaction: A practical approach. *Journal of Clinical Laboratory Analysis* 16: 47–51.

MARTIN, J., V. M. BRUNO, Z. FANG, X. MENG, M. BLOW, T. ZHANG, G. SHERLOCK, M. SNYDER, AND Z. WANG. 2010. Rnnotator: An automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* 11: 663.

MAUGHAN, P. J., S. SMITH, D. FAIRBANKS, AND E. JELLEN. 2011. Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (*Amaranthus* sp.). *Plant Genome* 4: 92–101.

MAUGHAN, P. J., S. M. YOURSTONE, R. L. BYERS, S. M. SMITH, AND J. A. UDALL. 2010. SNP genotyping in mapping populations via genomic reduction and next-generation sequencing: Proof of concept. *Plant Genome* 3: 166–178.

MAUGHAN, P. J., S. M. YOURSTONE, E. N. JELLEN, AND J. A. UDALL. 2009. SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. *Plant Genome* 2: 260–270.

MEUZELAAR, L. S., O. LANCASTER, J. P. PASCHE, G. KOPAL, AND A. J. BROOKES. 2007. MegaPlex PCR: A strategy for multiplex amplification. *Nature Methods* 4: 835–837.

MEYERS, S. C., AND A. LISTON. 2010. Characterizing the genome of wild relatives of *Limnanthes alba* (meadowfoam) using massively parallel sequencing. *Acta Horticulturae* 859: 309–314 (ISHS).

MILLER, M. R., J. P. DUNHAM, A. AMORES, W. A. CRESKO, AND E. A. JOHNSON. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17: 240–248.

MUTTER, G. L., AND K. A. BOYNTON. 1995. PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic Acids Research* 23: 1411–1418.

NESS, R. W., M. SIOL, AND S. C. BARRETT. 2011. *De novo* sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. *BMC Genomics* 12: 298.

NIJMAN, I. J., M. MOKRY, R. VAN BOXTEL, P. TOONEN, E. DE BRUIJN, AND E. CUPPEN. 2010. Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nature Methods* 7: 913–915.

NJUGUNA, W., A. LISTON, R. CRONN, AND N. BASSIL. 2010. Multiplexed *Fragaria* chloroplast genome sequencing. *Acta Horticulturae* 859: 315–320 (ISHS).

OKOU, D. T., K. M. STEINBERG, C. MIDDLE, D. J. CUTLER, T. J. ALBERT, AND M. E. ZWICK. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* 4: 907–909.

OSHLACK, A., M. D. ROBINSON, AND M. D. YOUNG. 2010. From RNA-seq reads to differential expression results. *Genome Biology* 11: 220.

PARIMOO, S., S. R. PATANJALI, H. SHUKLA, D. D. CHAPLIN, AND S. M. WEISSMAN. 1991. cDNA selection: Efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proceedings of the National Academy of Sciences, USA* 88: 9623–9627.

PARKS, M. 2011. Plastome phylogenetics in the genus *Pinus* using massively parallel sequencing technology. Ph.D. dissertation, Oregon

State University, Corvallis, Oregon, USA. Website http://hdl.handle.net/1957/21691.

PARKS, M., R. CRONN, AND A. LISTON. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7: 84.

PERTEA, M., AND S. L. SALZBERG. 2010. Between a chicken and a grape: Estimating the number of human genes. *Genome Biology* 11: 206.

PFENDER, W. F., M. C. SAHA, E. A. JOHNSON, AND M. B. SLABAUGH. 2011. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne. Theoretical and Applied Genetics* 122: 1467–1480.

PICARDI, E., D. S. HORNER, M. CHIARA, R. SCHIAVON, G. VALLE, AND G. PESOLE. 2010. Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. *Nucleic Acids Research* 38: 4755–4767.

PORRECA, G. J., K. ZHANG, J. B. LI, B. XIE, D. AUSTIN, S. L. VASSALLO, E. M. LEPROUST, ET AL. 2007. Multiplex amplification of large sets of human exons. *Nature Methods* 4: 931–936.

PORTNOY, V., A. DIBER, S. POLLOCK, H. KARCHI, S. LEV, G. TZURI, R. HAREL-BEJA, R. FORER, V. H. PORTNOY, E. LEWINSOHN, Y. TADMOR, J. BURGER, A. SCHAFFER, AND N. KATZIR. 2011. Use of non-normalized, non-amplified cDNA for 454-based RNA sequencing of fleshy melon fruit. *The Plant Genome* 4: 36–46.

PRIEST, H. D., S. E. FOX, S. A. FILICHKIN, AND T. C. MOCKLER. 2010. Utility of Next-Generation sequencing for analysis of horticultural crop transcriptomes. *Acta Horticulturae* 859: 283–288.

RABINOWICZ, P. D., R. CITEK, M. A. BUDIMAN, A. NUNBERG, J. A. BEDELL, N. LAKEY, A. L. O'SHAUGHNESSY, ET AL. 2005. Differential methylation of genes and repeats in land plants. *Genome Research* 15: 1431–1440.

RAZ, T., P. KAPRANOV, D. LIPSON, S. LETOVSKY, P. M. MILOS, AND J. F. THOMPSON. 2011. Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* 6: e19287.

RIGOLA, D., J. VAN OEVEREN, A. JANSSEN, A. BONNE, H. SCHNEIDERS, H. J. VAN DER POEL, N. J. VAN ORSOUW, ET AL. 2009. High-throughput detection of induced mutations and natural variation using KeyPoint technology. *PLoS ONE* 4: e4761.

ROBERTSON, G., J. SCHEIN, R. CHIU, R. CORBETT, M. FIELD, S. D. JACKMAN, K. MUNGALL, ET AL. 2010. *De novo* assembly and analysis of RNA-seq data. *Nature Methods* 7: 909–912.

ROBINSON, M. D., AND A. OSHLACK. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11: R25.

ROCHE DIAGNOSTICS CORPORATION. 2009. Technical bulletin. Using multiplex identifier (MID) adaptors for the GS FLX Titanium chemistry—Extended MID set. *In* Roche Applied Science 7. Roche, Indianapolis, Indiana, USA.

SAINTENAC, C., D. JIANG, AND E. D. AKHUNOV. 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biology* 12: R88.

SAMBROOK, J., AND D. W. RUSSELL. 2001. Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA.

SCHATZ, M. C., A. L. DELCHER, AND S. L. SALZBERG. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research* 20: 1165–1173.

SCHNEEBERGER, K., S. OSSOWSKI, F. OTT, J. D. KLEIN, X. WANG, C. LANZ, L. M. SMITH, ET AL. 2011. Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proceedings of the National Academy of Sciences, USA* 108: 10249–10254.

SENAPATHY, P., A. BHASI, J. MATTOX, P. S. DHANDAPANY, AND S. SADAYAPPAN. 2010. Targeted genome-wide enrichment of functional regions. *PLoS ONE* 5: e11138.

SEVERIN, A. J., J. L. WOODY, Y. T. BOLON, B. JOSEPH, B. W. DIERS, A. D. FARMER, ET AL. 2010. RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biology* 10: 160.

SHAGIN, D. A., D. V. REBRIKOV, V. B. KOZHEMYAKO, I. M. ALTSHULER, A. S. SHCHEGLOV, P. A. ZHULIDOV, E. A. BOGDANOVA, ET AL. 2002. A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Research* 12: 1935–1942.

SMITH, A. M., L. E. HEISLER, R. P. ST.ONGE, E. FARIAS-HESSON, I. M. WALLACE, J. BODEAU, A. N. HARRIS, ET AL. 2010. Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Research* 38: e142.

SOUTHERN, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* 98: 503–517.

STEELE, P. R., AND J. C. PIRES. 2011. Biodiversity assessment: State-of-the-art techniques in phylogenomics and species identification. *American Journal of Botany* 98: 415–425.

STRAUB, S. C. K., M. FISHBEIN, T. LIVSHULTZ, Z. FOSTER, M. PARKS, K. WEITEMIER, R. C. CRONN, AND A. LISTON. 2011. Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

STRAUB, S. C. K., M. PARKS, K. WEITEMIER, M. FISHBEIN, R. C. CRONN, AND A. LISTON. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

TEWHEY, R., J. B. WARNER, M. NAKANO, B. LIBBY, M. MEDKOVA, P. H. DAVID, S. K. KOTSOPOULOS ET AL. 2009. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* 27: 1025–1031.

TRAPNELL, C., B. A. WILLIAMS, G. PERTEA, A. MORTAZAVI, G. KWAN, M. J. VAN BAREN, S. L. SALZBERG, B. J. WOLD, AND L. PACHTER. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511–515.

TURNER, T. L., E. C. BOURNE, E. J. VON WETTBERG, T. T. HU, AND S. V. NUZHDIN. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* 42: 260–263.

TURNER, E. H., S. B. NG, D. A. NICKERSON, AND J. SHENDURE. 2009. Methods for genomic partitioning. *Annual Review of Genomics and Human Genetics* 10: 263–284.

VARLEY, K. E., AND R. D. MITRA. 2008. Nested patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Research* 18: 1844–1850.

VIVANCOS, A. P., M. GÜELL, J. C. DOHM, L. SERRANO, AND H. HIMMELBAUER. 2010. Strand-specific deep sequencing of the transcriptome. *Genome Research* 20: 989–999.

WANG, Z., M. GERSTEIN, AND M. SNYDER. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57–63.

WHITTALL, J. B., J. SYRING, M. PARKS, J. BUENROSTRO, C. DICK, A. LISTON, AND R. CRONN. 2010. Finding a (pine) needle in a haystack: Chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19 (supplement 1): 100–114.

WILHELM, B. T., AND J. R. LANDRY. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48: 249–257.

YUAN, Y. W., C. LIU, H. E. MARX, AND R. G. OLMSTEAD. 2009. An empirical demonstration of using pentatricopeptide repeat (PPR) genes as plant phylogenetic tools: Phylogeny of Verbenaceae and the *Verbena* complex. *Molecular Phylogenetics and Evolution* 54: 23–35.

ZARAGOZA, M. V., J. FASS, M. DIEGOLI, D. LIN, AND E. ARBUSTINI. 2010. Mitochondrial DNA variant discovery and evaluation in human cardiomyopathies through next-generation sequencing. *PLoS ONE* 5: e12295.

ZENONI, S., A. FERRARINI, E. GIACOMELLI, L. XUMERLE, M. FASOLI, G. MALERBA, D. BELLIN, M. PEZZOTTI, AND M. DELLEDONNE. 2010. Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. *Plant Physiology* 152: 1787–1795.

ZERBINO, D., AND E. BIRNEY. 2008. Velvet: Algorithms for *de novo* short read assembly using De Bruijn graphs. *Genome Research* 18: 821–829.

ZHONG, S., J.-G. JOUNG, Y. ZHENG, Y.-R. CHEN, B. LIU, Y. SHAO, J. Z. XIANG, Z. FEI, AND J. J. GIOVANNONI. 2011. High-throughput Illumina strand-specific RNA sequencing library preparation. *Cold Spring Harbor Protocols*, doi: 10.1101/pdb.prot5652

ZONNEVELD, B. J. M., I. J. LEITCH, AND M. D. BENNETT. 2005. First nuclear DNA amounts in more than 300 angiosperms. *Annals of Botany* 96: 229–244.