 doi:10.1093/bfgp/elt051

Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes

Cory D. Hirsch, Joseph Evans, C. Robin Buell and Candice N. Hirsch

Advance Access publication date 6 January 2014

Abstract

Technology and software improvements in the last decade now provide methodologies to access the genome sequence of not only a single accession, but also multiple accessions of plant species. This provides a means to interrogate species diversity at the genome level. Ample diversity among accessions in a collection of species can be found, including single-nucleotide polymorphisms, insertions and deletions, copy number variation and presence/absence variation. For species with small, non-repetitive rich genomes, re-sequencing of query accessions is robust, highly informative, and economically feasible. However, for species with moderate to large sized repetitive-rich genomes, technical and economic barriers prevent *en masse* genome re-sequencing of accessions. Multiple approaches to access a focused subset of loci in species with larger genomes have been developed, including reduced representation sequencing, exome capture and transcriptome sequencing. Collectively, these approaches have enabled interrogation of diversity on a genome scale for large plant genomes, including crop species important to worldwide food security.

Keywords: *exome capture; re-sequencing; genome diversity; RNA-Seq*

INTRODUCTION

Numerous plant genomes have been sequenced and more are currently being sequenced. The genome sequence of Arabidopsis (*Arabidopsis thaliana*) was completed in 2000 and represented the first high quality, finished plant genome sequence [1]. Iterative improvements to the initial genome assembly have been made. The current version of the Arabidopsis genome sequence, TAIR10, represents ~119 Mb of the genome with a limited number of gaps (http://arabidopsis.org/portals/genAnnotation/gene_structural_annotation/agicomplete.jsp). In 2002, draft genome sequences of indica and japonica rice (*Oryza sativa*) were published by two independent groups [2, 3], making rice the second plant species and first food crop with a genome sequence. In 2005, the International Rice Genome Sequencing Project released an accurate finished genome sequence of the japonica subspecies incorporating genetic map-based information, bolstering the usefulness

Corresponding author. Candice N. Hirsch, Department of Agronomy and Plant Genetics, University of Minnesota, 518 Borlaug Hall, 1991 Upper, Buford Circle, Saint Paul, MN 55108. E-mail: cnhirsch@umn.edu

Cory D. Hirsch, PhD is a National Science Foundation National Plant Genome Initiative Postdoctoral Research Fellow in the Department of Plant Biology at the University of Minnesota. His current research interests involve utilizing systems biology approaches to understand anthocyanin biosynthesis in potato.

Joseph Evans earned his PhD in Biochemistry from Texas A&M University in 2012, and is currently a postdoctoral research associate at Michigan State University. His research interests involve applying genomic techniques to explore breeding improvement opportunities in grass species.

C. Robin Buell earned her PhD in Biology/Molecular Biology from Utah State University in 1992, and did postdoctoral research at Michigan State University and the Carnegie Institution of Washington. She has been on the faculty at Louisiana State University and The Institute for Genomic Research. Currently, she is a Professor of Plant Biology at Michigan State University and has research interests in the genome biology of plants.

Candice N. Hirsch earned her PhD in Plant Breeding and Plant Genetics from the University of Wisconsin in 2010, and was a postdoctoral research associate at Michigan State University until 2013. She is currently an Assistant Professor at the University of Minnesota in the Department of Agronomy and Plant Genetics working on maize translational genomics.

of the rice genome in functional research [4]. As with Arabidopsis, development of new technologies and datasets have permitted improvements to the 'gold standard' rice genome sequence [5]. Another important plant reference genome sequence is that of maize (*Zea mays*) [6]. The maize genome (~2.3 Gb) is substantially larger than the genomes of either Arabidopsis or rice, which is attributed to retrotransposon insertions within the last three million years [7]. A draft genome sequence of maize in which the genic regions were targeted for 'finishing' and improved sequence quality was released in 2009 [6].

In the last decade, the number of plant species with genome sequences has grown rapidly, in large part due to the lower cost of sequencing and improvements in assembly techniques [8]. Plants with medicinal properties (*Cannabis sativa* [9]), trees (*Populus trichocarpa* [10] and *Carica papaya* [11]) and tuber bearing plants (*Solanum tuberosum* [12]) have genome sequences available, allowing new methods and techniques to be explored in these once non-model species. The assembly of the 20-Gb Norway spruce (*Picea abies*) genome sequence [13] demonstrates the major improvements made in high-throughput sequencing technologies and the associated software necessary to assemble large repetitive genomes.

The availability of a reference genome opens new areas for understanding genome composition, diversity, dynamics and evolution that are specifically enabled through comparative approaches. For example, reference genome sequences of sorghum (*Sorghum bicolor*) and rice, along with available resources from the wheat D genome (*Aegilops tauschii*) revealed accelerated genome evolution in Triticeae [14]. Comparative genomics have also been used to study evolution and domestication in maize [6].

Linking genotypes and phenotypes is another major area of interest in plant genetics and breeding that has benefitted greatly from the availability of reference genome sequences and identification of genome variation within a species. Currently, if the species of interest has a reference genome sequence and the genome is of limited size, high-throughput re-sequencing is employed to detect variation at the whole genome level (Figure 1A) [15]. Numerous types of variants can be detected through re-sequencing including single-nucleotide polymorphisms (SNPs) and insertions and deletions (InDels). Structural polymorphisms such as copy number variation (CNV) and presence/absence variation (PAV) can be detected with re-sequencing approaches using read depth [16] and *de novo* assembly of unmapped reads [17]. The application of re-sequencing has lead to the identification of millions of SNPs and InDels that have revealed the genetic diversity within many species (e.g. [18, 19]). A genome-wide association study (GWAS) approach tests thousands to millions of genetic variants (typically SNPs) for significant associations to phenotypes measured on a large number of individuals [20]. The first GWAS conducted on plants was completed in Arabidopsis for over 100 phenotypes [21]. GWAS has since been successfully implemented in crops including barley (*Hordeum vulgare*) [22], maize [23], rice [24] and tomato (*Solanum lycopersicum*) [25]. These variant datasets also serve as useful molecular markers for breeding efforts [18, 26] and have provided insights into crop domestication, gene function and selection sweeps [26–28]. However, there are technical aspects of re-sequencing that need to be taken into account such as sequence read errors, ploidy levels and heterozygosity that can confound accurate variant detection [29, 30].

With large plant genomes, there is both a cost and data analysis impediment to whole-genome re-sequencing. Reduced representation approaches have been employed for a number of large and complex plant genomes to identify variants that can be linked to phenotypes, thus furthering our understanding of genome evolution, genome biology and the genetic architecture underlying complex plant phenotypes. In this review, we discuss approaches currently used to interrogate genomic variation within large and complex plant genomes that reduce the complexity of genome assayed for variation and in parallel, the complexity of the resulting datasets (Figure 1).

REDUCED REPRESENTATION-BASED APPROACHES TO ASSESSING GENOME DIVERSITY

The large repetitive nature of many plant genomes can prevent whole genome re-sequencing approaches from being cost effective and can be technically challenging. However, numerous approaches have been developed that can dramatically reduce the sequence space, primarily by eliminating or greatly reducing the repetitive sequence content (Figure 2). Reduced representation approaches include exome capture,

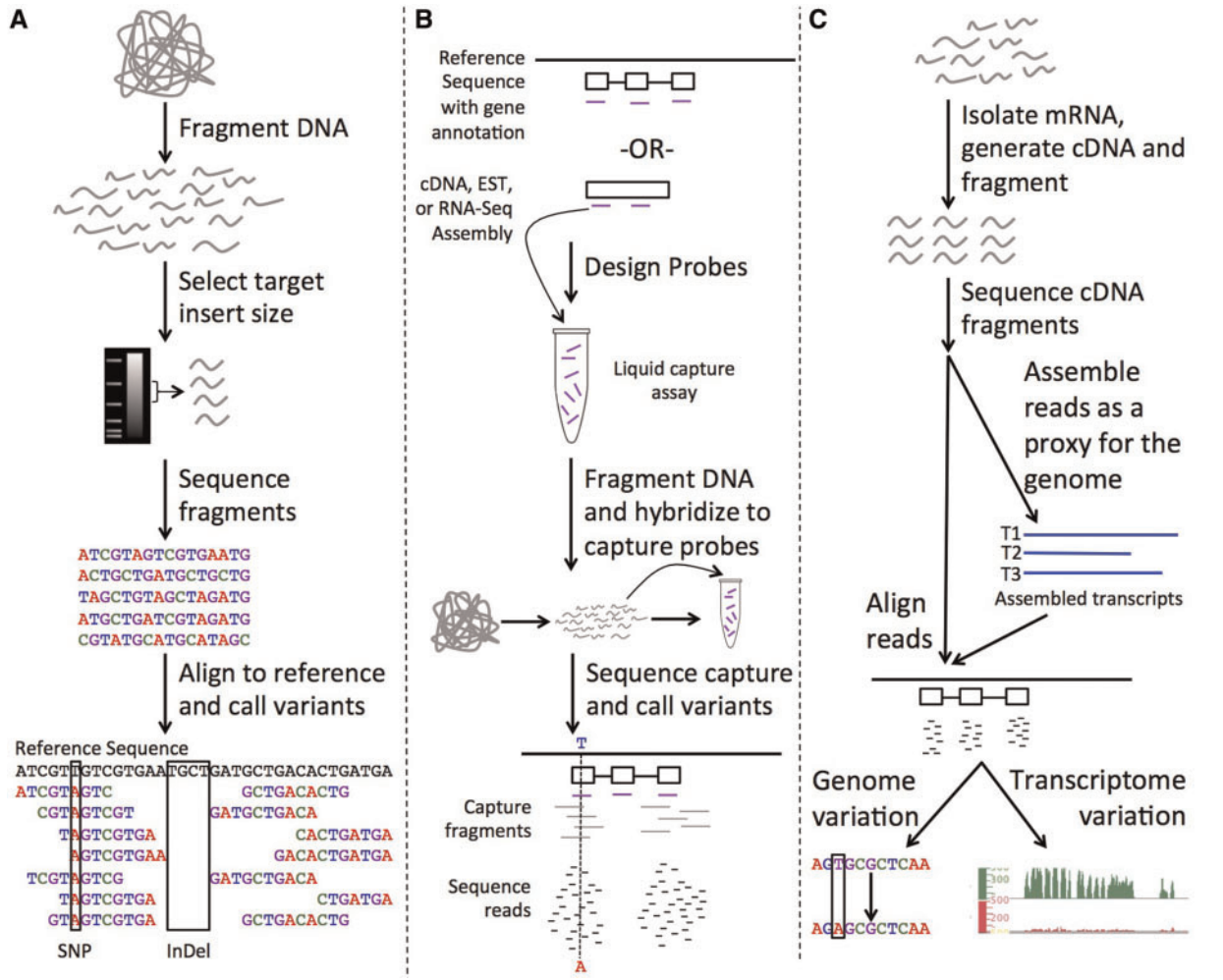

Figure 1: Representative technology approaches to interrogate genome diversity. (**A**) Whole-genome re-sequencing, (**B**) exome capture and (**C**) transcriptome sequencing. EST, Expressed Sequence Tag.

genotyping-by-sequencing and transcriptome sequencing as a proxy for genome sequencing.

Exome capture

Exome capture sequencing, a technique designed to capture only the exonic regions of a genome, provides a mechanism to reduce the sequence space of a complex genome, and thereby increase the depth of genic sequence coverage in a re-sequencing project [34]. Exome capture sequencing was originally conducted in humans, which have ~30 Mb of coding sequence distributed across ~3000 Mb of total sequences [34]. Since its initial development, exome capture has become a routine feature of cancer and human genetic disease research [35–37].

The basis of exome capture sequencing is the use of targeted oligonucleotide sequences to bind complementary sequences from whole genomic DNA, followed by high-throughput sequencing (Figure 1B) [38]. By using capture probes to target exon sequences of an organism, genic material is enriched in sequencing [38]. Several different capture methods may be employed to enrich sequences. One method, called solid phase capture, involves fixing probes to a solid surface in a manner similar to a microarray [38, 39]. Query genomic DNA is hybridized to the surface bound capture probes and non-specific hybridization is removed through stringent washes. The captured DNA is then eluted and sequenced using a high-throughput sequencing platform. Capture assays are increasingly eliminating fixing of the probes to a solid surface in favor of a liquid capture. Liquid phase capture involves suspending biotinylated DNA or RNA probes in solution [40], capturing genomic DNA through annealing, binding the biotinylated probes to a

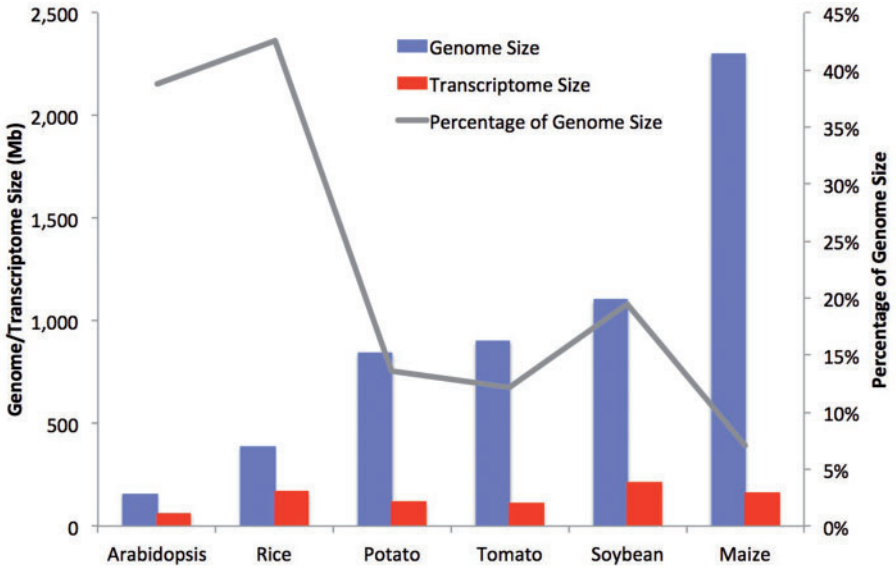

Figure 2: Estimated genome and transcriptome size variation of representative plant species with reference genome sequences. Genome sizes reflect estimated genome size, not genome assembly size and transcriptome sizes reflect current gene annotations for each species (Arabidopsis (*A. thaliana*) [31] ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff, rice (*O. sativa*) [4] ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/all.gff3, potato (*S. tuberosum*) [12] http://potato.plantbiology.msu.edu/pgsc_download.shtml PGSC_DM_v4.03_pseudomolecules_genes.gff3, tomato (*S. lycopersicum*) [32] ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.3_release/ITAG2.3_gene_models.gff3, soybean (*Glycine max*) [33] http://www.phytozome.net/dataUsagePolicy.php?org=Org_Gmax/Gmax_189_gene.gff3 and maize (*Z. mays*) [6] http://ftp.maizesequence.org/current/filtered-set/ZmB73_5b_FGS.gff). Note that transcriptome size is relatively constant between plant species with highly variable genome sizes, highlighting the increased utility of targeting the genic sequence space in species with large and complex genomes.

streptavidin substrate, eluting the captured sequences from the probes and then sequencing the captured DNA fragments [41].

Initial exome capture experiments in humans and mice showed the importance of reducing nonspecific binding, initially performed by the inclusion of species-specific sequence from the common repeat element C_0t-1 repeats [38, 42, 43]. As a result, in early plant experiments, the capture was performed in two steps. The initial capture used probes designed to target species-specific repeat sequences; after binding, the probes and sequences were discarded, removing the repeat sequences from the pool of genetic material [44]. A second capture step was then performed with probes targeting the sequences of interest. With repeat sequences removed, nonspecific binding was minimized, thereby allowing targeted material to be sequenced efficiently. [44]. Recent technology improvements bypass this two-step capture, thereby reducing costs by not requiring an additional hybridization step or the development of an initial set of capture probes that target repetitive elements, allowing researchers to perform the exome capture in a single step [45]. Combined with liquid-phase capture, this advancement allows researchers to rapidly perform exome capture experiments on a wide variety of large repetitive plant genomes [45–47].

Capture probe sequences can be designed from a number of templates. The most common templates are previously sequenced genomic DNA [34, 38, 44] or predicted gene annotations [45, 48] to enrich probes in genic sequences. The use of probes designed to a reference sequence was used to identify regions of genomic heterogeneity between closely related individuals of the reference cultivar in soybean, showing the resolving power of exome capture even in closely related self-pollinating populations [45]. In plants, however, it is becoming more common for exome capture probe sets to be designed using *de novo* transcript assemblies [49], cDNA libraries [47, 50] or whole-genome sequencing reads aligned to a closely related species [51]. Probes designed based on cDNA sequence or

RNA-Sequencing (RNA-Seq) transcript assemblies allow researchers to develop capture probes with very little information available about their genome of interest, relying instead on transcript sequences as proxies for the genome sequence. During probe design, exon boundaries must be taken into consideration. If a capture probe spans an exon boundary, the corresponding intronic sequence in the genomic DNA will not anneal to the probe and the probe will have poor capture efficiency [49]. Despite this concern, in all examined cases, capture probes successfully enriched targeted sequences [44, 45, 47–51] and in some cases enriched the genic sequences >10-fold [51]. This technique has been demonstrated successfully in Loblolly pine (*Pinus taeda* L.), where the large genome size (21.7 Gb) and repetitive sequence content precluded efficient use of whole-genome sequencing [49].

Enrichment of genic sequences and the reproducibility of capture targets from sample to sample makes exome capture sequencing an appealing option for high-throughput variant detection [41, 48, 49]. For variant detection, sequence reads from the captured template are aligned to a reference sequence, such as the transcript assembly used to design the probes [46] or the reference genome sequence for that species [45]. Care must be taken during variant detection for capture assays developed from RNA-Seq transcript assemblies or complementary DNA (cDNA) sequences, as genes or alleles that are expressed lowly or not at all can be excluded from the probe design process, which can result in a bias during capture of genomic sequence [50]. Additionally, with polyploid species, the presence of homeologs can complicate read alignments, resulting in a higher false-positive rate [47]. Despite these challenges, transcriptome-derived probes have been used successfully in exome capture studies in polyploid species. In allohexaploid bread wheat (*Triticum aestivum*), researchers using a cDNA-derived exome capture probe set were able to identify and differentiate both homeologous (between subgenomes) and varietal (between varieties) SNPs [50], demonstrating specificity with a transcriptome sequence-based approach for probe design.

Exome capture sequencing can also be used to evaluate structural variation [52]. Several analysis packages are available to identify CNV from read depth variation in exome capture datasets [53–55], but it should be noted that these packages are designed for use with sequence data from diploid organisms, and may not be applicable to polyploid plant species. One limitation of exome capture is that it is reliant on the probe design and the capture process can only capture sequences that have probes designed for them. As a result, genes that are absent in the reference accession or otherwise missing from the probe set will not be captured and while absence variants versus the reference accession can be identified, novel presence variants cannot.

Genotyping-by-sequencing

In general, genotyping-by-sequencing refers to any genotyping determined through sequencing. This can include 'skim' whole-genome re-sequencing, as was performed by Huang *et al*. [56], in which 0.02X coverage of recombinant inbred lines (RILs) was generated to genotype a RIL population and identify quantitative trail loci (QTL). Alternatively, reduced representation genotyping-by-sequencing approaches can be used in species with large repetitive genomes through restriction enzyme digestion, followed by high-throughput sequencing of genomic sequences adjacent to the enzyme cut-site. This technique allows for increased sequence coverage of nonrepetitive sequences. Restriction site-associated DNA (RAD) markers are commonly generated by digesting genomic DNA with a restriction enzyme, ligating adaptors, sample pooling, randomly shearing the DNA, size selection, a second round of adapter ligation, polymerase chain reaction (PCR) amplification of fragments containing both adapters, followed by library construction and sequencing. Sequencing RAD markers was first described in stickleback fish [57]. The multiplexing capability of RAD marker sequencing can greatly reduce the cost of sequencing library preparation, while maintaining identification of a large number of markers [58]. In addition, a reference genome is not required when applying RAD marker sequencing. Genotyping samples using RAD marker sequencing have been useful in linkage mapping and QTL identification in barley [59], eggplant (*Solanum melongena*) [60] and perennial ryegrass (*Lolium perenne*) [61, 62].

A modified version of RAD marker sequencing was used by Elshire *et al*. [63], which omitted the shearing step, included less sample handling, fewer purification steps, no size selection and improved barcoding to simplify the method, which is referred to simply as GBS [63]. In GBS assays using the Elshire method, samples are digested, barcoded, pooled and sequenced in a multiplexed fashion and

SNPs and InDels can be identified. Furthermore, haplotype blocks can be determined based on genotype scores and imputation can be performed when necessary to be used as markers for downstream analysis [64]. Variations of the Elshire GBS protocol have been used for marker discovery and to study genome diversity, population structure, genomic selection and genetic diversity in plant species such as barley, grapevine, maize, switchgrass (*Panicum virgatum*) and wheat [63, 65–70]. GBS is most easily utilized in homozygous diploid species with a reference sequence. However, in species without a reference genome the restriction site adjacent sequence tags can be considered dominant markers [63]. GBS has added value to mapping and breeding populations as shown by increasing marking density thereby allowing previously unknown QTL to be identified [71].

It is of note that when utilizing RAD sequencing or GBS, as the complexity of the genome increases, through increased ploidy and/or heterozygosity, increased read depths are needed and computational approaches are more difficult to implement.

Transcriptome sequencing

An alternative reduced representation method in large and complex plant genomes is to use transcriptome sequencing also known as RNA-Seq (Figure 1C). In contrast to other reduced representation approaches, RNA-Seq provides information about genome diversity as well as transcriptome diversity for the sampled tissue(s). The major limitation to this approach is that an allele/gene must be expressed in the sampled tissue(s) in order to interrogate its diversity. RNA-Seq has been used to identify genome level variation including SNPs, InDels, simple sequence repeats (SSRs) and gene level PAV in plant species with reference genome sequences such as maize [23, 72, 73], potato [74] and tomato [75]. For plant species with large and complex genomes lacking a high quality reference sequence such as alfalfa (*Medicago sativa*) [76, 77], *Hevea brasiliensis* (source of commercial natural rubber) [78], *Panicum hallii* (emerging model for switchgrass) [79], rye (*Secale cereale*) [80] and various medicinal plant species [81, 82], RNA-Seq has been used to generate a transcriptome assembly first, which served as a proxy for the genome sequence, followed by interrogation of genome, transcriptome and metabolome diversity. It should be noted that in polyploidy species, there are additional challenges to assembling a transcriptome, namely that the assembly of diverse alleles, splice variants and gene family members can be complicated by the presence of transcripts from additional homologous (autopolyploid) or homeologous (allopolyploid) chromosomes.

RNA-Seq is a powerful tool for identifying molecular markers in a cost effective manner. For example, in potato, minimal genomic resources were available despite it being a model species in the Solanaceae. Following the release of the potato genome sequence [12], RNA-Seq reads and Sanger-derived expressed sequence tags in six cultivars were used for SNP discovery [74]. The SNPs were used to generate an Infinium SNP array [83] that has been widely used by the potato community to provide insights into the effects of North American potato breeding on allele frequencies, understanding the genetic basis of diversification and trait improvement [84], identification of polymorphisms related to the steroidal glycoalkaloid metabolic pathway [85] and development of genetic maps for QTL mapping studies [83, 86]. Likewise in tomato, SNP molecular markers were identified using RNA-Seq reads from accessions that span tomato market classes [75], from which a large SNP genotyping array was developed for evaluation of an expanded panel of tomato accessions [87].

RNA-Seq data can also be used to identify molecular markers in species that lack a reference sequence, such as in alfalfa, a widely grown perennial forage with complex genetics and minimal genomic resources. RNA-Seq of two tissues from two alfalfa genotypes with variable cellulose and lignin concentrations was used for *de novo* assembly, generation of a gene index and identification of SNP and SSR polymorphisms [76]. Using the SNPs and SSRs as well as differential gene expression analysis, candidate genes responsible for the differences in cell wall composition were identified. Likewise, RNA-Seq of 27 alfalfa genotypes used for *de novo* transcriptome assembly and SNP detection identified nearly 900 000 SNPs and InDels. Analysis of diversity from wild species in cultivated alfalfa revealed that ~95% of the polymorphic sites present in the wild species were also present in the cultivated lines following domestication and breeding efforts [77].

Although transcriptome assembly is commonly implemented in species lacking a reference sequence, it can also be used in species with a reference sequence to identify nonreference genes. In a maize study that evaluated seedling RNA-Seq from 21 diverse lines, transcriptome assembly of reads that did

not align to the reference genome sequence identified 1321 high confidence novel transcripts. Over half of the sequences were present only in a subset of the lines thus showing they are dispensable in nature [73]. The study also provided insights into the mechanisms underlying heterosis, with the identification of over 2000 transcripts specific to two major heterotic groups involved in US grain hybrids and tight clustering of two distinct heterotic groups based on SNPs identified from the RNA-Seq reads.

Bulk segregant analysis (BSA) is an approach that has been used to identify the genomic position of genes underlying mutant phenotypes by measuring the allele frequency of genetic markers in pools of mutant and wild-type plants. A variant of BSA was developed in maize, which utilized RNA-Seq data (BSR-Seq; [72]). SNP markers were discovered *de novo* and mutant and wild-type pools were quantitatively genotyped using the RNA-Seq reads. In addition to the cost savings of performing RNA-Seq rather than whole genome re-sequencing, BSR-Seq allowed for quantification of gene expression levels within the significant region that was used to identify a candidate gene supported by independent transposon-induced mutant alleles.

Understanding how gene expression variation is controlled can provide valuable insights into the basis of phenotypic diversity. One method to understand the quantitative regulation of gene expression variation is with expression QTL (eQTL) studies, which test for linkages between variation in expression and genetic polymorphisms. RNA-Seq is a powerful method for performing eQTL studies, as quantitative expression variation, allele-specific expression, alternative splice form variation and genomic molecular markers can be evaluated [88, 89]. Using RNA-Seq for GWAS studies is also useful due to the fact that both genomic and expression level polymorphisms can be evaluated [90]. In a study examining the genetic architecture of maize oil biosynthesis, RNA-Seq from 368 diverse maize lines identified over one million SNPs that were used for a GWAS [23]. The study identified 74 significant loci that were subsequently examined using eQTL mapping, linkage mapping and co-expression analysis. Interestingly, the co-expression analysis indicated that one-third of the genes affected the phenotype via transcriptional regulation, highlighting the value of identifying expression variation in addition to sequence variation.

CONCLUSIONS

The range of diversity in plant species is highly variable, attributable to multiple sources, such as the mechanism of reproduction, with outcrossing species having higher levels of heterozygosity relative to self-pollinating species. Additionally, there is massive variation in genome size due to differences in repetitive sequence content. Characterizing this diversity is important to understand the genetic architecture underlying phenotypic diversity, mechanisms of genome selection and evolution and the genetic basis of biological phenomena such as heterosis. As with the range of diversity in plant species, the availability of genomic tools is also highly variable and different technological approaches are necessary to interrogate genome diversity in different plant species. In this article, we reviewed multiple reduced representation approaches to characterize genome diversity, each with positive and negative aspects depending on the plant species being studied (Table 1).

The ideal approach to studying genome diversity is deep coverage whole-genome re-sequencing. Using this approach, genomic diversity (SNPs, InDels, PAV and CNV) can be explored over the entire genome. Whole-genome re-sequencing requires access to a reference genome sequence, which is not available for all plant species. Additionally, for species with large repetitive genomes, this approach is not cost effective with current technologies. For researchers working in species with large genomes that may or may not have extensive genomic tools, reduced representation methods can prove very useful. For example, despite the initial effort to design the capture array and the potential capture bias, exome capture is a cost effective way to characterize numerous levels of genome diversity in only low copy exon portions of the genome. Genotyping-by-sequencing offers an alternative to SNP genotyping arrays, allowing for an increased number of polymorphic loci to be genotyped and requires no initial set up cost, or access to a reference sequence. However, there is the possibility for a large amount of missing data, InDels and structural variants are less readily detected and bioinformatics expertise is necessary. Finally, using the transcriptome as a proxy to the genome is powerful for species that lack a reference genome sequence. However, with RNA-Seq only genes that are expressed can be interrogated and only alleles that are expressed will be identified. Additionally, CNV cannot be evaluated

Table 1: Advantages and disadvantages of approaches to interrogate genome diversity

	Advantages	Disadvantages
Whole Genome		
Whole-genome re-sequencing	-SNPs, InDels, CNV and PAV genome variation can be assayed -The entire genome is assayed	-Reference sequence needed -Can be expensive and not cost effective
Reduced representation		
Exome capture	-No reference sequence needed -SNPs, InDels, CNV and PAV[a] genome variations can be assayed	-*A priori* knowledge required -Capture bias -Initial capture design effort
Genotyping-by-sequencing (RAD/GBS)	-No reference sequence needed -Large number of polymorphic loci assayed -No *a priori* knowledge required	-Large amount of missing data -InDels and structural variants are less readily detected
Transcriptome sequencing	-No reference sequence needed -Can be assembled and used as a proxy for genic regions of the genome -Assay genome and transcriptome variation -Large number of polymorphic loci assayed	-Only assay-expressed genes -Allele-specific expression can result in mis-genotyped individuals -Cannot assay structural variation

Note: [a]Relative to the sequence space from which the probe set was designed

and there is not an absolute relationship between transcriptome and genome level PAV.

The choice of which approach to use with complex genomes depends on the required information content for downstream analyses, the cost per sample, the total number of samples to be assayed, and the desired level of investment in technology and/or bioinformatics. For some applications such as genetic map construction or QTL identification, GBS may provide ample resolution for a relatively small cost per sample thereby enabling large numbers of samples to be screened. In contrast, RNA-Seq will provide a substantially higher level of resolution of variant detection per sample in addition to expression abundance estimations yet costs substantially more per sample than GBS, potentially restricting the sample size. Thus, with new technologies and analysis tools, the prospect for exploring genome diversity in any plant species is now a reality. However, it should be noted that with each technology and analysis tool, there are limitations. Understanding these technological limitations is necessary to understand the limitations of scientific conclusions.

Key Points

- Current technologies permit access to genome sequences of hundreds of individuals within a species.
- For plant species with large genomes, alternative approaches to access genic regions of the genome are well developed, cost effective and being used in a wide number of species.
- New insights into diversity have, and will continue to enable major advances in all aspects of plant science including breeding, biology and evolution.

FUNDING

Genome diversity work in the Buell lab is funded by the Department of Energy Great Lakes Bioenergy Research Center (Department of Energy Office of Biological and Environmental Research grant number DE-FC02-07ER64494); the United States Department of Agriculture (2009-85606-05673); the National Science Foundation Plant Genome Research Program (IOS-1237969); Funding for CDH is provided by a National Science Foundation National Plant Genome Initiative Fellowship (120274).

References

1. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;**408**:796–815.
2. Goff SA, Ricke D, Lan T-H, *et al*. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* 2002;**296**: 92–100.
3. Yu J, Hu S, Wang J, *et al*. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 2002;**296**: 79–92.

4. The International Rice Genome Sequencing Consortium. The map-based sequence of the rice genome. *Nature* 2005; **436**:793–800.
5. Kawahara Y, de la Bastide M, Hamilton J, *et al*. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 2013;**6**:4.
6. Schnable PS, Ware D, Fulton RS, *et al*. The B73 maize genome: complexity, diversity, and dynamics. *Science* 2009;**326**:1112–5.
7. SanMiguel P, Gaut BS, Tikhonov A, *et al*. The paleontology of intergene retrotransposons of maize. *Nat Genet* 1998;**20**: 43–5.
8. Schatz M, Witkowski J, McCombie WR. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol* 2012;**13**:243.
9. Bakel HV, Stout JM, Cote AG, *et al*. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol* 2011;**12**:R102. http://genomebiology.com/2011/12/10/r102 (02 January 2014, date last accessed).
10. Tuskan GA, DiFazio S, Jansson S, *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006;**313**:1596–604.
11. Ming R, Hou S, Feng Y, *et al*. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 2008;**452**:991–6.
12. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* 2011; **475**:189–95.
13. Nystedt B, Street NR, Wetterbom A, *et al*. The Norway spruce genome sequence and conifer genome evolution. *Nature* 2013;**497**:579–84.
14. Luo MC, Deal KR, Akhunov ED, *et al*. Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc Natl Acad Sci USA* 2009;**106**:15780–5.
15. Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. *Nat Rev Genet* 2012;**13**: 85–96.
16. Duan J, Zhang J-G, Deng H-W, *et al*. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* 2013;**8**: e59128.
17. Lai J, Li R, Xu X, *et al*. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 2010;**42**: 1027–30.
18. McHale LK, Haun WJ, Xu WW, *et al*. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol* 2012;**159**:1295–308.
19. Mace ES, Tai SS, Gilding EK, *et al*. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat Commun* 2013;**4**:2320.
20. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 2013;**9**:29.
21. Atwell S, Huang YS, Vilhjálmsson BJ, *et al*. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 2010;**465**:627–31.
22. Wang M, Jiang N, Jia T, *et al*. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *Theor Appl Genet* 2012;**124**:233–46.
23. Li H, Peng Z, Yang X, *et al*. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet* 2013;**45**:43–50.
24. Huang X, Zhao Y, Wei X, *et al*. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* 2012;**44**:32–9.
25. Ranc N, Munos S, Xu J, *et al*. Genome-wide association mapping in tomato (*Solanum lycopersicum*) Is possible using genome admixture of *Solanum lycopersicum* var. *cerasiforme*. *G3* 2012;**2**:853–64.
26. Hufford MB, Xu X, van Heerwaarden J, *et al*. Comparative population genomics of maize domestication and improvement. *Nat Genet* 2012;**44**:808–11.
27. Xu X, Liu X, Ge S, *et al*. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 2012;**30**:105–11.
28. Lam H-M, Xu X, Liu X, *et al*. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* 2010;**42**:1053–9.
29. Kumar S, Banks TW, Cloutier S. SNP discovery through next-generation sequencing and its applications. *Int J Plant Genomics* 2012;**2012**:831460. http://www.hindawi.com/journals/ijpg/2012/831460/ (02 January 2014, date last accessed).
30. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2012;**13**:36–46.
31. Bennett MD, Leitch IJ, Price HJ, *et al*. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann Bot* 2003;**91**:547–57.
32. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 2012;**485**:635–41.
33. Schmutz J, Cannon SB, Schlueter J, *et al*. Genome sequence of the palaeopolyploid soybean. *Nature* 2010;**463**:178–83.
34. Ng SB, Turner EH, Robertson PD, *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;**461**:272–6.
35. Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**:745–55.
36. Marian AJ. Molecular genetic studies of complex phenotypes. *Transl Res* 2012;**159**:64–79.
37. Veltman JA, Brunner HG. Applications of next-generation sequencing *de novo* mutations in human genetic disease. *Nat Rev Genet* 2012;**13**:565–75.
38. Hodges E, Xuan Z, Balija V, *et al*. Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* 2007; **39**:1522–7.
39. Okou DT, Steinberg KM, Middle C, *et al*. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 2007;**4**:907–9.
40. Gnirke A, Melnikov A, Maguire J, *et al*. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009;**27**:182–9.
41. Bainbridge MN, Wang M, Burgess DL, *et al*. Whole exome capture in solution with 3 Gbp of data. *Genome*

Biol 2010;**11**:R62. http://genomebiology.com/2010/11/6/R62 (02 January 2014, date last accessed).

42. Albert TJ, Molla MN, Muzny DM, *et al*. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007;**4**:903–5.
43. Fairfield H, Gilbert GJ, Barter M, *et al*. Mutation discovery in mice by whole exome sequencing. *Genome Biol* 2011; **12**:R86. http://genomebiology.com/2011/12/9/R86 (02 January 2014, date last accessed).
44. Fu Y, Springer NM, Gerhardt DJ, *et al*. Repeat subtraction-mediated sequence capture from a complex genome. *Plant J* 2010;**62**:898–909.
45. Haun WJ, Hyten DL, Xu WW, *et al*. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol* 2011; **155**:645–55.
46. Mascher M, Richmond TA, Gerhardt DJ, *et al*. Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. *Plant J* 2013;**76**:494–505.
47. Saintenac C, Jiang DY, Akhunov ED. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol* 2011;**12**: http://genomebiology.com/2011/12/9/R88 (02 January 2014, date last accessed).
48. Zhou LC, Holliday JA. Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics* 2012;**13**:703.
49. Neves LG, Davis JM, Barbazuk WB, *et al*. Whole-exome targeted sequencing of the uncharacterized pine genome. *Plant J* 2013;**75**:146–56.
50. Winfield MO, Wilkinson PA, Allen AM, *et al*. Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol J* 2012;**10**:733–42.
51. Bundock PC, Casu RE, Henry RJ. Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. *Plant Biotechnol J* 2012; **10**:657–67.
52. Park G, Gim J, Kim A, *et al*. Multiphasic analysis of whole exome sequencing data identifies a novel mutation of ACTG1 in a nonsyndromic hearing loss family. *BMC Genomics* 2013;**14**:191. http://www.biomedcentral.com/1471-2164/14/191 (02 January 2014, date last accessed).
53. Love MI, Mysickova A, Sun RP, *et al*. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol* 2011;**10**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3517018/ (02 January 2014, date last accessed).
54. Sathirapongsasuti JF, Lee H, Horst BAJ, *et al*. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;**27**: 2648–54.
55. Deng XT. SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data. *BMC Bioinformatics* 2011;**12**:267. http://www.biomedcentral.com/1471-2105/12/267 (02 January 2014, date last accessed).
56. Huang X, Feng Q, Qian Q, *et al*. High-throughput genotyping by whole-genome resequencing. *Genome Res* 2009; **19**:1068–76.
57. Baird NA, Etter PD, Atwood TS, *et al*. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 2008;**3**:e3376.
58. Davey JW, Hohenlohe PA, Etter PD, *et al*. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 2011;**12**:499–510.
59. Chutimanitsakun Y, Nipper R, Cuesta-Marcos A, *et al*. Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. *BMC Genomics* 2011;**12**:4.
60. Barchi L, Lanteri S, Portis E, *et al*. A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *PLoS One* 2012;**7**:e43740.
61. Hegarty M, Yadav R, Lee M, *et al*. Genotyping by RAD sequencing enables mapping of fatty acid composition traits in perennial ryegrass (*Lolium perenne* (L.)). *Plant Biotechnol J* 2013;**11**:572–81.
62. Pfender WF, Saha MC, Johnson EA, *et al*. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theor Appl Genet* 2011;**122**:1467–80.
63. Elshire RJ, Glaubitz JC, Sun Q, *et al*. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 2011;**6**:e19379. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0019379 (02 January 2014, date last accessed).
64. Deschamps S, Llaca V, May GD. Genotyping-by-sequencing in plants. *Biology* 2012;**1**:460–83.
65. Poland JA, Brown PJ, Sorrells ME, *et al*. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 2012;**7**:e32253.
66. Lu F, Lipka AE, Glaubitz J, *et al*. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet* 2013; **9**:e1003215.
67. Crossa J, Beyene Y, Kassa S, *et al*. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 2013;**3**:1903–26, http://www.g3journal.org/content/early/2013/09/05/g3.113.008227.abstract (02 January 2014, date last accessed).
68. Tian F, Bradbury PJ, Brown PJ, *et al*. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 2011;**43**:159–62.
69. Gore MA, Chia JM, Elshire RJ, *et al*. A first-generation haplotype map of maize. *Science* 2009;**326**:1115–7.
70. Myles S, Chia JM, Hurwitz B, *et al*. Rapid genomic characterization of the genus *Vitis*. *PLoS One* 2010;**5**:e8219.
71. Spindel J, Wright M, Chen C, *et al*. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor Appl Genet* 2013;**126**:2699–716.
72. Liu S, Yeh CT, Tang HM, *et al*. Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS One* 2012;**7**:e36406.
73. Hansey CN, Vaillancourt B, Sekhon RS, *et al*. Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS One* 2012;**7**:e33071.
74. Hamilton JP, Hansey CN, Whitty BR, *et al*. Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics* 2011;**12**:302.
75. Hamilton JP, Sim S-C, Stoffel K, *et al*. Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome* 2012;**5**:17–29.

76. Yang SS, Tu ZJ, Cheung F, *et al*. Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC Genomics* 2011;**12**:199.
77. Li X, Acharya A, Farmer A, *et al*. Prevalence of single nucleotide polymorphism among 27 diverse alfalfa genotypes as assessed by transcriptome sequencing. *BMC Genomics* 2012;**13**:568.
78. Xia Z, Xu H, Zhai J, *et al*. RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. *Plant Mol Biol* 2011;**77**:299–308.
79. Meyer E, Logan TL, Juenger TE. Transcriptome analysis and gene expression atlas for *Panicum hallii* var. filipes, a diploid model for biofuel research. *Plant J* 2012; **70**:879–90.
80. Haseneyer G, Schmutzer T, Seidel M, *et al*. From RNA-Seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol* 2011;**11**:131.
81. Gongora-Castillo E, Fedewa G, Yeo Y, *et al*. Genomic approaches for interrogating the biochemistry of medicinal plant species. *Methods Enzymol* 2012;**517**:139–59.
82. Góngora-Castillo E, Childs KL, Fedewa G, *et al*. Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLoS One* 2012;**7**:e52506.
83. Felcher KJ, Coombs JJ, Massa AN, *et al*. Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One* 2012;**7**:e36347.
84. Hirsch CN, Hirsch CD, Felcher K, *et al*. Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. *G3* 2013;**3**:1003–13.
85. Manrique-Carpintero NC, Tokuhisa JG, Ginzberg I, *et al*. Sequence diversity in coding regions of candidate genes in the glycoalkaloid biosynthetic pathway of wild potato species. *G3* 2013;**3**:1467–79.
86. Hackett CA, McLean K, Bryan GJ. Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS One* 2013;**8**:e63939.
87. Sim SC, Durstewitz G, Plieske J, *et al*. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One* 2012;**7**:e40563.
88. Majewski J, Pastinen T. The study of eQTL variations by RNA-Seq: from SNPs to phenotypes. *Trends Genet* 2011;**27**: 72–9.
89. Gan X, Stegle O, Behr J, *et al*. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 2011;**477**: 419–23.
90. Delker C, Quint M. Expression level polymorphisms: heritable traits shaping natural variation. *Trends Plant Sci* 2011; **16**:481–8.