Published December 12, 2012

Genotyping-by-Sequencing for Plant Breeding and Genetics

Jesse A. Poland* and Trevor W. Rife

Abstract

Rapid advances in "next-generation" DNA sequencing technology have brought the US$1000 human (*Homo sapiens*) genome within reach while providing the raw sequencing output for researchers to revolutionize the way populations are genotyped. To capitalize on these advancements, genotyping-by-sequencing (GBS) has been developed as a rapid and robust approach for reduced-representation sequencing of multiplexed samples that combines genome-wide molecular marker discovery and genotyping. The flexibility and low cost of GBS makes this an excellent tool for many applications and research questions in plant genetics and breeding. Here we address some of the new research opportunities that are becoming more feasible with GBS. Furthermore, we highlight areas in which GBS will become more powerful with the continued increase of sequencing output, development of reference genomes, and improvement of bioinformatics. The ultimate goal of plant biology scientists is to connect phenotype to genotype. In plant breeding, the genotype can then be used to predict phenotypes and select improved cultivars. Furthering our understanding of the connection between heritable genetic factors and the resulting phenotypes will enable genomics-assisted breeding to exist on the scale needed to increase global food supplies in the face of decreasing arable land and climate change.

Next-Generation Genotyping

DRIVEN BY THE QUEST for a $1000 human genome, rapid advances in next-generation sequencing (NGS) output have provided technology with the ability to greatly transform the way we think about plant genomics and breeding. With the introduction of massively parallel sequencing, raw sequencing output is doubling roughly every 6 mo (Fig. 1). The availability of inexpensive sequencing technology has transformed the way genomes are sequenced (Xu et al., 2011; Wang et al., 2011), polymorphisms are discovered (Mardis, 2008; Futschik and Schlötterer, 2010; You et al., 2011; Nielsen et al., 2011), gene expression is analyzed (Geraldes et al., 2011; Harper et al., 2012), and populations are genotyped (Baird et al., 2008; Elshire et al., 2011; Davey et al., 2011; Truong et al., 2012; Poland et al., 2012a; Wang et al., 2012). Sequencing is rapidly becoming so inexpensive that it will soon be reasonable to use it for every genetic study. Next-generation sequencing applications have the potential to revolutionize the field of plant genomics and the practice of applied plant breeding.

One of the primary objectives of functional genomics in agricultural species is to connect phenotype to genotype and use this knowledge to make phenotypic predictions and select improved plant types. To do this on a genome-wide scale requires large populations with dense molecular markers across the genome. To put the power of NGS to work for plant breeding and genomics,

Published in The Plant Genome 5:92–102.
doi: 10.3835/plantgenome2012.05.0005

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

J.A. Poland, USDA-ARS, Hard Winter Wheat Genetics Research Unit and Dep. of Agronomy, Kansas State Univ., 4008 Throckmorton Hall, Manhattan KS, 66506; T.W. Rife, Interdepartmental Genetics, Kansas State Univ., 4024 Throckmorton Hall, Manhattan KS, 66506. Received 29 May 2012. *Corresponding author (jesse.poland@ars.usda.gov).

Abbreviations: AM, association mapping; GBS, genotyping-by-sequencing; GS, genomic selection; HMM, hidden Markov model; MSG, multiplexed shotgun genotyping; NGS, next-generation sequencing; PAV, presence–absence variation; RAD, restriction association DNA; SNP, single nucleotide polymorphism.

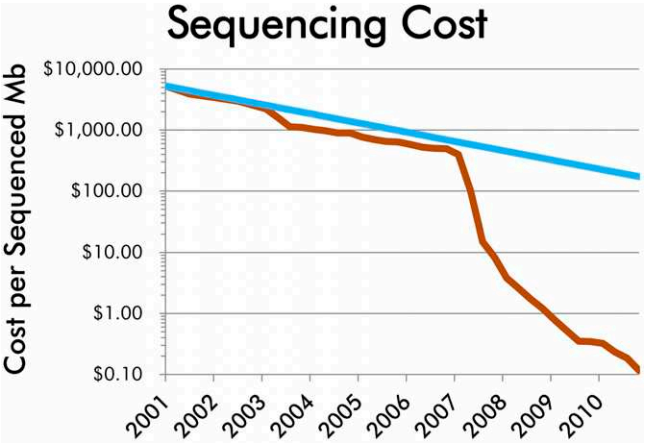

Figure 1. A comparison of actual sequencing capacity (orange) to what would be expected if sequencing technology was following Moore's Law (blue). The significant decrease in 2007 coincides roughly with the introduction of next-generation sequencing technology. Data is from the National Human Genome Research Institute (Wetterstrand, 2012).

new approaches for sequence-based genotyping have been developed. One promising approach is genotyping-by-sequencing (GBS), which uses enzyme-based complexity reduction (using restriction endonucleases to target only a small portion of the genome) coupled with DNA barcoded adapters to produce multiplex libraries of samples ready for NGS sequencing. This approach has been demonstrated to be robust across a range of species and capable of producing tens of thousands to hundreds of thousands of molecular markers (Elshire et al., 2011; Poland et al., 2012a). The flexibility of GBS in regards to species, populations, and research objectives makes this an ideal tool for plant genetics studies. As the phenomenal increase in NGS output continues, many research questions that were once out of reach will be resolved through the application of these approaches.

All-in-One

The two key components for genotyping germplasm are finding DNA sequence polymorphisms and assaying the markers across a full set of material. Classically, this has been a two-step process involving marker discovery followed by assay design and genotyping. An important strength of sequence-based genotyping approaches is that the marker discovery and genotyping are completed at the same time. This facilitates exploration of new germplasm sets or even new species without the upfront effort of discovering and characterizing polymorphisms. Another key component of GBS datasets is that the raw data is dynamic. The raw sequences obtained from GBS can be reanalyzed, uncovering further information (e.g., new polymorphisms, annotated genes, etc.) as bioinformatics techniques improve, reference genomes develop, and the collection of sequence data increases. Each of these factors adds additional value to the same raw dataset.

One of the first and broadly adapted applications for using NGS was for single nucleotide polymorphism (SNP) and presence–absence variation (PAV) discovery in diverse populations with and without reference genomes (Baird et al., 2008; Wiedmann et al., 2008; Gore et al., 2009a, 2009b; Huang et al., 2009; Deschamps et al., 2010; Hyten et al., 2010; You et al., 2011; Nelson et al., 2011; Hohenlohe et al., 2011; Byers et al., 2012). These studies have focused on assaying a few key genotypes with a reduced-representation approach (Baird et al., 2008) or with whole-genome resequencing (Huang et al., 2009). While highly effective for SNP discovery, this approach is limited in the number of lines assayed and does not simultaneously assay the markers across the full population of interest.

The key objective of the GBS approach, therefore, is not merely to discover polymorphisms and then transfer these to a fixed assay, but to simultaneously discover polymorphisms and obtain genotypic information across the whole population of interest. It is this combined one-step approach that makes GBS a truly rapid and flexible platform for a range of species and germplasm sets and perfectly suited for genomic selection (GS) in plant breeding programs. As sequencing output continues to increase, GBS will evolve first to lower levels of complexity reduction (to capture more sequence variants) and then to whole-genome resequencing (to capture all variants). Whole-genome resequencing has been applied in *Arabidopsis thaliana* (L.) Heynh., rice (*Oryza sativa* L.), and maize (*Zea mays* L.) (Huang et al., 2009; Ashelford et al., 2011; Gan et al., 2011; Chia et al., 2012; Jiao et al., 2012; Xu et al., 2012), although it quickly becomes less manageable with larger, more complex genomes that lack a solid reference genome (Morrell et al., 2011). The level of multiplexing has also been limited in this approach, increasing per-sample cost.

As GBS can be readily used for de novo discovery and application of new molecular polymorphisms, it is particularly powerful for new sets of germplasm and uncharacterized species. In many ways the greatest advantage of sequence-based genotyping approaches is the reduction of ascertainment bias associated with marker discovery in panels differing from the target population. This is an obvious advantage for association studies in which differing allele frequencies greatly influence the power and precision of the study (Myles et al., 2009; Hamblin et al., 2010). For breeding applications, informative polymorphisms can be discovered as novel germplasm is introduced into the breeding pool. The use of an unrepresentative marker panel in surveying molecular diversity is highly problematic for getting a true representation of molecular diversity present in a target population. Most GBS approaches use methylation-sensitive enzymes. If these enzymes target differentially methylated regions of the genome, ascertainment bias could potentially be introduced in different sets of germplasm, but evidence for this has yet to be seen. While markers discovered with GBS should have little bias across sets of germplasm, it is also unknown how uniformly they are spaced across the genome. Evidence from Poland et al. (2012a), however, indicated that GBS markers were

Table 1. A technical comparison of current genotyping methods using next-generation sequencing of multiplex barcoded libraries. Adapted from Wang et al. (2012). Flavors of genotyping using next-generation sequencing of multiplex DNA-barcoded reduced-representation libraries.

Method	Random shearing	Size selection	Fragment size	Enzymes†	Multiplexing level‡	Analysis tool(s)	Reference
Multiplex shotgun genotyping	No	Yes	Size selected	*Mse*I	96 (up to 384)	Burrows-Wheeler alignment tool	Andolfatto et al., 2011
Restriction association DNA sequencing (RAD-seq)	Yes	Yes	Size selected	*Sbf*I *Eco*RI	96	Custom Perl scripts	Baird et al., 2008
Double digest RAD-seq	No	Yes	Size selected	*Eco*RI and *Msp*I	48§	MUSCLE¶	Peterson et al., 2012
2b-restriction association DNA	No	No	33–36 bp	*Bsa*XI#	NA††	Custom Perl scripts	Wang et al., 2012
Genotyping-by-sequencing	No	No	<350 bp	*Ape*KI‡‡	48 (up to 384)	TASSEL§§	Elshire et al., 2011
Genotyping-by-sequencing – two enzyme	No	No	<350 bp	*Pst*I and *Msp*I	48 (up to 384)	TASSEL	Poland et al., 2012a
Sequence-based genotyping	No	Yes	Size selected	*Eco*RI and *Mse*I *Pst*I and *Taq*I	32	Burrows-Wheeler alignment tool and unified genotyper	Truong et al., 2012
Restriction enzyme sequence comparative analysis	No	Yes	Size selected	*Mse*I *Nla*III	NA¶¶	Burrows-Wheeler alignment tool and Samtools	Monson-Miller et al., 2012

†All of these approaches can use different enzymes. Shown are the enzyme(s) used in the initial study.

‡All of these methods have the possibility to increase the number of multiplexed samples using additional unique barcodes. The multiplex level as reported in the reference paper. Given in parenthesis are subsequent increases.

§Combinatorial barcoding is possible, placing a barcode on each end of the DNA fragment. Using a set of 48 adapter P1 barcodes and × 12 polymerase chain reaction (PCR) 2 indices it is possible to uniquely label 576 individuals (48 [adapter P1 barcodes] × 12 [PCR2 indices]). This method would require paired-end sequencing.

¶MUSCLE, multiple sequence comparison by log-expectation.

#Uses type IIB restriction endonucleases.

††NA, not applicable.

‡‡Has been successfully applied to using *Pst*I and *Hin*dIII (E. Buckler and R. Elshire, personal communication, 2012).

§§TASSEL, trait analysis by association, evolution, and linkage.

¶¶96-plexing reported but unpublished.

uniformly spaced across the chromosomes of both wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.).

Many Flavors

The use of reduced-representation sequencing for targeting small portions of the genome was first demonstrated by Altshuler et al. (2000). This approach was then later combined with NGS and DNA barcoded adapters to sequence multiplex libraries in parallel. There are many variations of this approach and GBS is one specific method for genotyping using NGS of multiplex DNA-barcoded reduced-representation libraries (Table 1). Furthermore, the combination of enzymes that can be used for complexity reduction is almost endless. Davey et al. (2011) has thoroughly reviewed several approaches of complexity reduction including complexity reduction of polymorphic sequences (van Orsouw et al., 2007) and deep sequencing of reduced representation libraries (van Tassell et al., 2008).

The use of restriction enzymes for targeted reduction of genome complexity combined with NGS was first described by Baird et al. (2008) and termed restriction association DNA (RAD). Restriction association DNA methods use a restriction enzyme to generate genomic fragments, which are then ligated to an adaptor containing a forward primer for amplification, sequencing platform primer sites, and a unique DNA barcode that enables sample multiplexing (Baird et al., 2008; Craig et al., 2008; Cronn et al., 2008). The samples are pooled, randomly sheared, and size selected to create a uniform collection of similarly-sized DNA fragments (Baird et al., 2008). The fragments are then ligated to a Y adaptor that ensures only fragments containing the first adaptor will be amplified (Baird et al., 2008). Restriction association DNA markers provided a robust method to discover polymorphisms and map variation in a population (Miller et al., 2007).

First-generation RAD analysis had drawbacks similar to older restriction enzyme-based marker technologies: the requirement of species-specific arrays, a hybridization for every comparison, and limitations for assaying presence-absence variation (Baird et al., 2008). Combining the progressive features of RAD with NGS, however, resulted in the discovery of new markers at a significantly decreased cost (Baird et al., 2008). The simultaneous discovery of SNP markers during RAD sequencing facilitated robust mapping of many polymorphisms and precise assignment of chromosomal regions to mapping parents, allowing for detection of recombination locations. The RAD approach has recently been modified to use restriction enzymes that cut upstream and downstream of a target site (Wang et al., 2012). This new methodology produces uniform length tags, allows nearly all of the restriction sites to be surveyed, and permits marker intensity adjustment

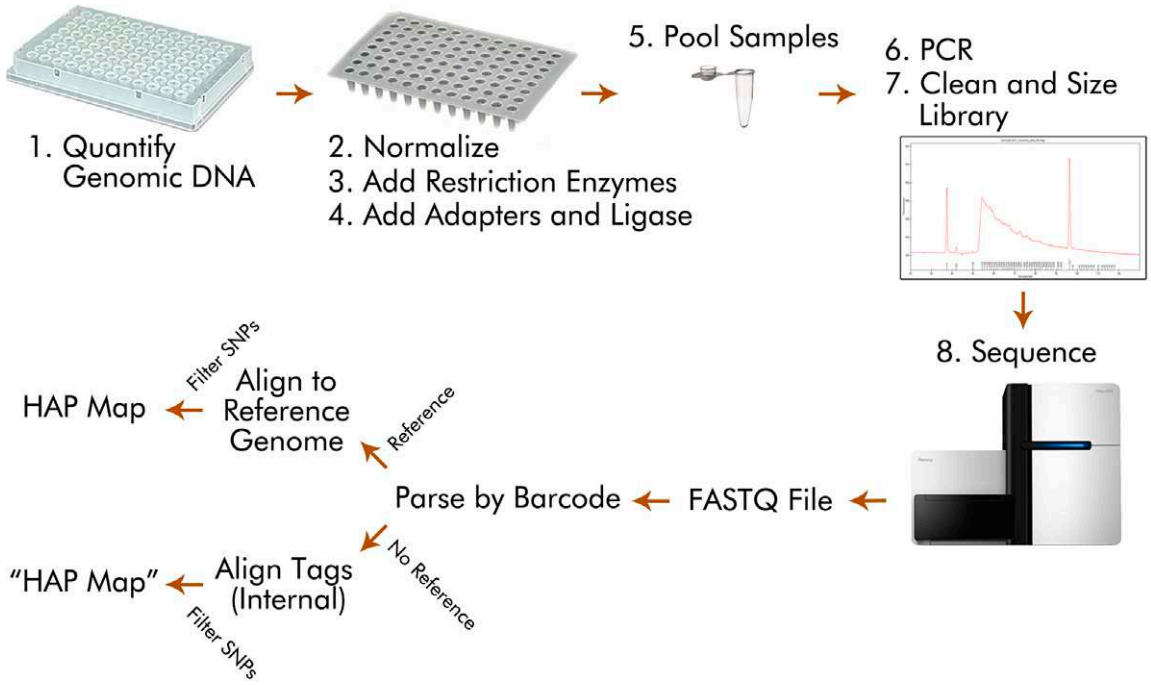

Figure 2. Schematic overview of steps in genotyping-by-sequencing (GBS) library construction, sequencing, and analysis. (1) Genomic DNA is quantified using fluorescence-based method. (2) Genomic DNA (gDNA) is normalized in a new plate. Normalization is needed to ensure equal representation of all samples and equal molarity of gDNA and adapters. (3) A master mix with restriction enzyme(s) and buffer is added to the plate and incubated. (4) The DNA barcoded adapters are added along with ligase and ligation buffers. (5) Samples are pooled and cleaned. (6) The GBS library is polymerase chain reaction (PCR) amplified. (7) The amplified library is cleaned and evaluated on a capillary sizing system. (8) Libraries are sequenced. Data analysis: Following a sequencing run, FASTQ files containing raw data from the run are used to parse sequencing reads to samples using the DNA barcode sequence. Once assigned to individual samples, the reads are aligned to a reference genome. In the case of species without a complete reference genomic sequence, reads are internally aligned (alignment of all sequence reads will all other reads from that library) and single nucleotide polymorphisms (SNPs) identified from 1 or 2 bp sequence mismatch. Various filtering algorithms can then be used to distinguish true biallelic SNPs from sequencing errors.

(Wang et al., 2012). The next flavor of sequence-based genotyping was multiplexed shotgun genotyping (MSG), which required only one gel purification, eliminated DNA shearing, required less starting DNA, and implemented a hidden Markov model (HMM) to determine points of chromosomal recombination (Andolfatto et al., 2011). Multiplexed shotgun genotyping used a single common cutting restriction enzyme and produced a limited complexity reduction suitable for the smaller genome (approximately 130 Mb) of *Drosophila simulans* (Andolfatto et al., 2011). In the context of a reference genome, the HMM imputation approach was highly effective for tracing parental origin and defining recombination break points (Andolfatto et al., 2011).

The original GBS protocol was developed to simplify and streamline the construction of RAD libraries (Elshire et al., 2011). The strength of the GBS protocol is its simplicity: using inexpensive adapters, allowing pooled library construction, and avoiding shearing and size selection (Fig. 2). The GBS approach removed the need for size selection by using a short polymerase chain reaction extension of the multiplexed library. Instead of the Y adapters used in the RAD protocol, the original GBS protocol used a single restriction enzyme, a barcoded adaptor, and a common adaptor (Elshire et al., 2011). Although all combinations of adapters can ligate to the DNA fragments, only those that contained one of each barcode are able to be amplified and sequenced (Davey et al., 2011).

The original GBS approach was recently extended to a two-enzyme version that combines a rare- and a common-cutting restriction enzyme to generate uniform libraries consisting of a forward (barcoded) adaptor and a reverse (Y) adaptor on alternate ends of each fragment (Poland et al., 2012a). The use of two enzymes in this GBS approach enables the capture of most fragments associated with the rare-cutting enzyme. The use of a Y adaptor on the common restriction site avoids amplification of more common fragments, a preferential situation for larger, more complex genomes. Following the original work on wheat and barley, this GBS approach has been successfully applied in several species including cotton (*Gossypium hirsutum* L.), oat (*Avena sativa* L.), sorghum [*Sorghum bicolor* (L.) Moench], and rice with little to no change in protocol (Poland, unpublished data, 2012).

The options for tailoring GBS to any species or desired application are almost endless. A range of enzymes have been evaluated in maize with success in varying the level of complexity reduction (E. Buckler, personal communication, 2012). With a varied level of complexity reduction, it is possible to increase coverage

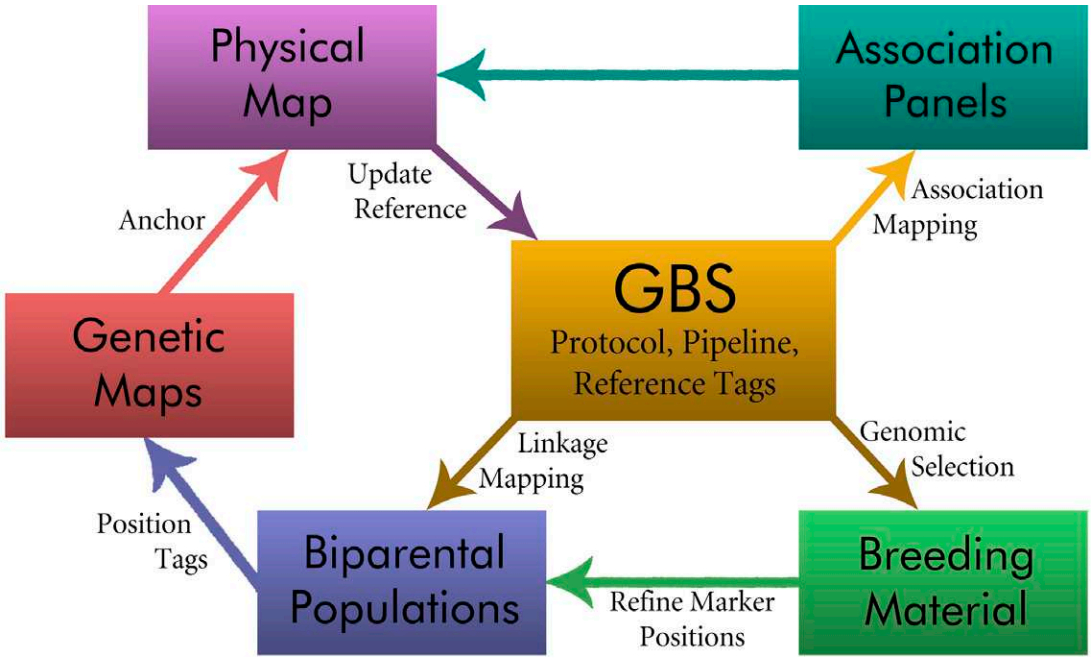

Figure 3. Integration of genotyping-by-sequencing (GBS) in the context of plant breeding and genomics for a species without a completed reference genome.

of a target genome or increase the multiplexing level of a target population. The interplay of these two factors will determine the optimal approach for the species under investigation. For species with large genomes or no reference genome, the use of rare-cutting restriction enzymes (i.e., 6 bp or greater target site) with methylation sensitivity can assist in creating a higher level of complexity reduction by targeting fewer sites. This will lead to higher sampling depth of the same genomic sites and reduce the amount of missing data (Fig. 3).

Hand in Hand with the Reference Genome

Sequence-based genotyping greatly benefits from a well-characterized (sequenced) reference genome. A reference genome makes ordering and imputing low coverage marker data generated through GBS and other sequence-based genotyping approaches straightforward. This has been seen in many of the reported uses of sequence-based genotyping. The MSG approach used by Andolfatto et al. (2011) made use of the *D. simulans* reference genome to first align tags to the reference and then call SNPs. Using a physical map framework, the parent-of-origin was then imputed across all SNPs segregating in the population. This approach is very robust for assigning parent-of-origin in biparental populations. Likewise, Huang et al. (2009) used the reference genome of rice to first align NGS tags and subsequently call SNPs. The physical ordering of these markers greatly enabled and simplified the imputation and assignment of parent-of-origin for segregating populations.

Although GBS approaches greatly benefit from a reference genome, the rapid discovery and ordering (through genetic mapping) of sequence-based molecular markers can assist with the development and refinement of a reference genome. High-density genetic maps developed through GBS can be used to anchor and order physical maps and refine or correct unordered sequence contigs. In *D. simulans*, Andolfatto et al. (2011) were able to assign 8 Mb to linkage groups, which comprised 30% of the unassembled *D. simulans* genome or about 6% of the total genome. This is a substantial improvement of an already well-characterized genome. Likewise, in current efforts in much larger, more complex genomes including barley (5.5 Gb) and wheat (16 Gb) (Arumuganathan and Earle, 1991), high-density GBS maps are being used to assist with anchoring and ordering large numbers of assembled but unanchored and unordered contigs (International Barley Sequencing Consortium, 2012). This approach appears very promising, creating a positive feedback loop in which the development of the reference genome assisted by GBS markers leads to better SNP calling and order-based imputation for GBS datasets.

Maps Made Easy

The combination of GBS with a well-defined reference genome makes the development of genetic maps for characterizing segregating populations exceptionally straightforward. In the absence of a solid reference genome, a high-density reference genetic map can serve the same purpose. For characterizing a new population, there will no longer be any need to place markers on linkage groups, calculate recombination frequencies, or order markers. With a reference genome, markers can be ordered along the physical chromosome. This ordering can then be used to precisely place recombination break points. The power of such approaches has been

highlighted in recent papers with model species including *D. simulans* (Andolfatto et al., 2011), rice (Huang et al., 2010), and maize (Elshire et al., 2011). Even at low coverage, the placement of sparse markers on the physical map can be used to narrow points of recombination to 100 to 200 kb intervals (Huang et al., 2009; Xie et al., 2010). This approach can be extended to populations with heterozygous chromosomal segments such as F_2 or BC_1 populations. Andolfatto et al. (2011) demonstrated a HMM that accurately inferred heterozygous states from low-pass sequence-based genotyping. These same approaches have successfully been applied in maize (P. Bradbury, personal communication, 2012).

In the absence of a solid reference genome, the same ease of genetic mapping can be accomplished through development of a reference genetic map for the species of interest. Genotyping-by-sequencing markers and other framework markers can be integrated to develop a high-density genetic map (Poland et al., 2012a). For new populations, GBS tags can be used to make genotype calls based on the reference map without the need to construct a de novo map. The extremely large number of markers produced with GBS allows sufficient coverage for most populations even if only a fraction of the total markers are used.

These same approaches for developing genetic maps and graphical genotypes can be broadly applied to the characterization of populations of interest for breeding and germplasm improvement including elite breeding lines, segregating populations for selection, near-isogenic lines, and alien-introgression lines. The use of a variety of algorithms to correctly infer the heterozygous or homozygous state of chromosome regions will add value to inferences and conclusions for molecular breeding and selection (Andolfatto et al., 2011). Other algorithms can be used for phasing markers in segregating and outcrossing populations. This will generally, however, require known marker order of the GBS SNPs.

Mapping Single Genes

Genotyping-by-sequencing and other sequence-based genotyping approaches can be very powerful for mapping single genes. The de novo discovery of high-density markers in a population of interest has the potential to circumvent the cumbersome process of marker discovery and testing for fine mapping of target genes and mutations. In the absence of a reference map, RAD markers have been used in bulked segregant analysis to quickly identify linked markers (Baird et al., 2008). For single genes of interest, this can be a valuable approach to rapidly identify segregating polymorphisms. In lupin (*Lupinus angustifolius* L.), Yang et al. (2012) were able to identify 30 markers linked to an anthracnose resistance gene. One advantage of GBS for mapping single genes in F_2 or similar populations is that the per-sample cost will be low enough that individual samples can be used rather than bulks. This will allow correction or removal of any individuals that were incorrectly phenotyped while confirming segregation of linked markers. Depending on the application, there will be a balance between finding markers linked to the gene of interest using GBS and developing single marker assays from the resulting data. Considering breeding approaches, it can still be optimal to prescreen populations with markers for known single genes (with large effects) for smaller investment in time and sample costs before conducting whole genome profiling. Selected plants carrying desired genes can then be genotyped using GBS for GS.

An Excess of Markers

While preselection of breeding populations for single markers for important genes is a viable breeding strategy, sequencing capacity is becoming so inexpensive and readily available that it will soon be reasonable to generate whole-genome profiles on any germplasm of interest. Previously, scientists spent a majority of their time developing and working with a small number of markers. Many projects today still require only a small number of markers to complete. Genotyping-by-sequencing, however, can readily generate tens of thousands of usable markers, which can be selectively filtered into the few required for a target experiment. While statistical geneticists will always prefer to have as many markers as possible, GS models have diminishing returns on additional markers once the population has reached the point of "marker saturation" (Jannink et al., 2010; Heffner et al., 2011). On the other hand, for association mapping (AM) studies, additional markers increase the likelihood of finding and tagging causal polymorphisms (Cockram et al., 2010). The current limitation for the generated data is computational. There are new algorithms and developments in cluster computing to provide the computational resources needed to make these quantitative genetics questions more manageable (Stanzione, 2011). Quantitative geneticists and bioinformatics personnel will be needed to manage breeding data and develop models. At the same time, bioinformatics training will become a more central component to any plant breeding and genetics curriculum.

Filling in the Blanks

The "catch" to GBS and sequence-based genotyping in general is that datasets often have a significant amount of missing data due to low coverage sequencing (Davey et al., 2011). Biologically, missing genotyping calls in GBS datasets can be the result of presence–absence variation, polymorphic restriction sites, and/or differential methylation. On the other hand, the technical issue of missing data with GBS is a combination of (i) library complexity (i.e., number of unique sequence tags) and (ii) sequence coverage of the library.

Library complexity is directly related to the species' genome under investigation and the choice of enzyme(s) used for complexity reduction. Enzymes with a shorter recognition site will naturally produce more fragments than those with a longer recognition site. Methylation-sensitive enzymes will greatly reduce the number of fragments in species with large portions of repetitive

DNA. In barley, libraries constructed using *Pst*I and *Msp*I generate around 500,000 to 600,000 unique tags, while in wheat around 1.5 million tags are generated (Poland, unpublished data, 2012). The actual number of sequence tags present in a raw dataset is substantially higher partly due to allelic variants but largely due to sequencing errors, many of which can be nonrandom. This can and will generate many versions of "unique" tags.

The level of missing data is based on the sequencing coverage, which is a function of the library complexity, the multiplexing level, and the output of the sequencing platform (Andolfatto et al., 2011). The multiplexing level and the number of independent sequences generated from the sequencing platform will determine the average number of reads per sample. Higher multiplexing levels will reduce the data per sample while increased sequencing output (when using the same multiplexing level) will understandably increase the data per sample. One key component of GBS on different sequencing platforms is the number of *independent* reads. Post-Sanger sequencing platforms generally rely on a large number of short sequence reads to produce gigabases of sequence data (Metzker, 2009). The new platforms are continually increasing the sequencing output, a function of more and longer reads. For GBS, however, generating longer reads is less advantageous than generating more reads. More sequence reads provides more data per sample. Alternatively, increasing read numbers allows higher multiplexing levels with static amounts of data per sample. For GBS, 10 Gb of sequence data generated from 100 million reads of 100 bp would be preferable to 10 million reads of 1000 bp. While increasing the number of reads is clearly advantageous for GBS, longer reads are also beneficial, leading to the discovery of more polymorphisms (particularly in species with limited diversity) and assisting GBS applications in polyploids where secondary, genome-specific polymorphisms are needed to differentiate a segregating SNP from homeologous sequences on other genomes.

Missing data can be dealt with by (i) sequencing to higher depth or (ii) imputing. The logical approach to removing missing data is to sequence to a higher depth by reducing the multiplexing level or sequencing the library multiple times. This can be very effective (Fig. 4), but has the drawback of increasing per-sample cost. For important AM panels or parents of a breeding program, however, the additional investment to generate higher coverage of the tags is likely worthwhile. For breeding applications using GBS with targeted selection, other approaches to minimize the impact of missing data are preferable. Since a majority of the breeding population will be discarded, minimizing genotyping cost will take preference over minimizing missing data.

The second approach is imputation of missing data. Depending on the genome, the type of GBS libraries, and the overall size of the datasets, imputation can give very accurate results. There are many imputation algorithms (Marchini et al., 2007; Purcell et al., 2007; Browning and Browning, 2007), most of which are targeted toward haplotype reconstruction on a reference genome. Other approaches such as a random forest model (Breiman, 2001) can be used to impute unordered markers (as is the situation in wheat). Sequencing diverse, key individuals in the population (parents or representatives of kinship clusters) can greatly improve imputation accuracy by defining known haplotypes for the population.

Finally, a matrix of realized relationships among individuals in a breeding population can be constructed without imputation. For very high-density genotyped data generated by GBS, the marker coverage is sufficient to saturate the genomic linkage disequilibrium present in most breeding programs. From this perspective, it is only necessary to determine a pairwise identity between individuals for the markers that are present in both individuals. With high marker density, there will still be tens of thousands of pairwise comparisons between two individuals, well beyond the saturation point for most elite breeding material. Imputation with the simple marker mean can still produce accurate GS prediction models. From a GS perspective, kinship-based marker imputation can be used to optimize the realized relationship matrix in the presence of a high level of missing data (Poland et al., 2012b). This approach has been shown to improve the relationship estimates and give more accurate GS model predictions.

Association Mapping

Genotyping-by-sequencing has the potential to be an excellent tool for genotyping of diverse panels for AM. One key to applying GBS for AM is addressing the missing data problem. As previously noted, higher coverage sequencing will reduce the amount of missing data at the expense of increased per-sample costs. For a high-value AM panel that will be well characterized and extensively phenotyped and serve as a community resource population, the additional cost of sequencing several times to achieve high coverage is likely worth the investment. This will produce a very well-characterized genetic population. At a high coverage, imputation of missing data will become a very precise exercise, particularly on populations with extensive linkage disequilibrium. Depending on the species under interrogation, the GBS markers will need to be ordered via a physical reference map or through genetic mapping.

In such populations, GBS markers also have the advantage of being able to survey multiple haplotypes on a fine scale. When two or more SNPs are within the same tag, these SNP alleles are both evaluated concurrently. For PAVs, GBS also has the power to uncover these alleles. Array-based methods, particularly those applied to polyploid species, are limited in the ability to accurately survey PAVs as hybridization to a duplicated sequence will indicate an allele call (for the ancestral allele) even if the target locus is absent. Due to the context sequence accompanying a SNP, GBS enables discrimination between duplicated sequences. At higher sequencing coverage of the GBS library, PAV can then be

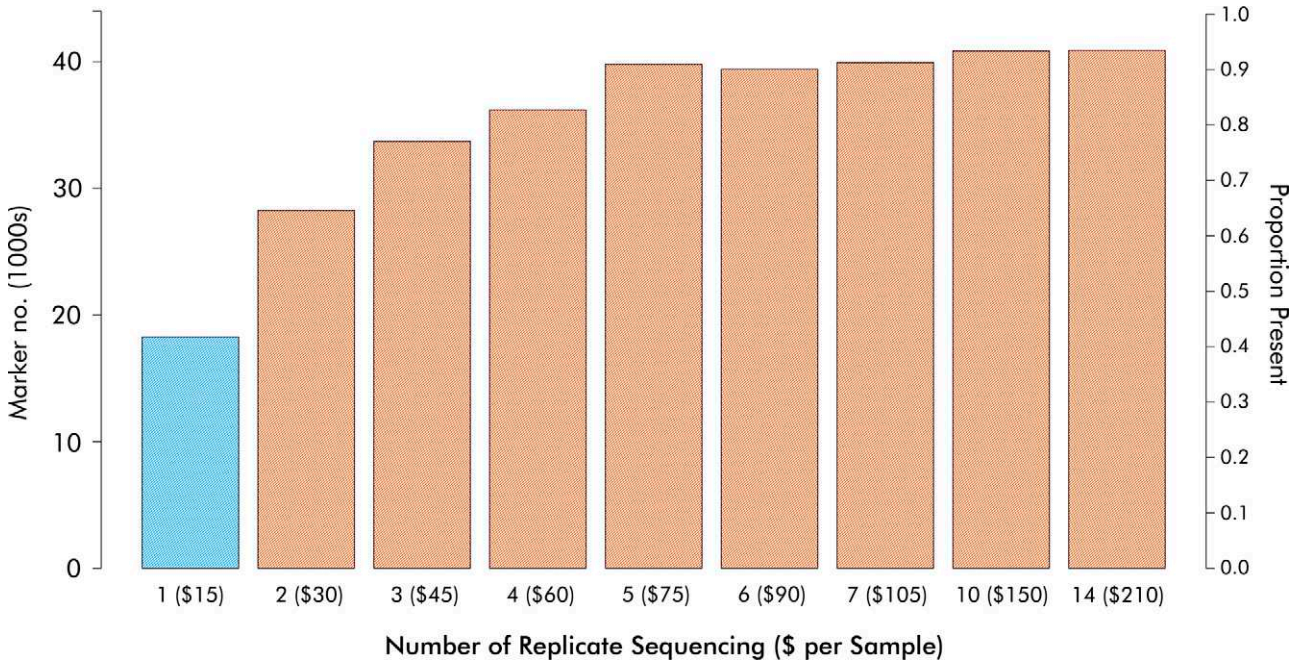

Figure 4. Removal of missing data in genotyping-by-sequencing by increasing coverage of the library via resequencing. In a set of international wheat breeding germplasm, several lines (samples) were replicated across two or more libraries. Replicating a sample two times increased the coverage of single nucleotide polymorphisms (SNPs) to 60% while five replications increase the coverage to over 90%. While very effective as a means to remove missing data, replicated sequencing increases the per-sample cost. The average per-sample cost is $15. In this situation for wheat, the number of replications is roughly equivalent to the sequencing coverage of the library (i.e., 5 replications give approximately 5x coverage). Data from J. Poland (unpublished data, 2012).

inferred by the absence of a given tag for a given sample in the pool of sequenced tags.

Genomic Selection

In the field of plant breeding, an important objective in the development of GBS is to create a low-cost genotyping platform capable of generating high-density genotypes. For GS in crop species, breeders need a fast, inexpensive, flexible method that will enable genotyping of large populations of selection candidates. A majority of the selection candidates are then discarded, creating a situation that is greatly benefited from low-cost genotyping. Genotyping-by-sequencing is quickly expanding to fill those requirements.

Genomic selection was proposed in 2001 by Meuwissen et al. as an approach to capture the full complement of small effect loci in genomic prediction models. Genomic selection takes advantage of dense genome-wide molecular markers by simultaneously fitting effects to all markers and avoiding statistical testing. By using these GS models, breeders are able to predict the performance of new experimental lines at early generations and generate suggested crosses and selections based on the model predictions (Jannink et al., 2010). Combined with a fast turnaround on generations, selection based on predicted breeding values determined by marker data provided by GBS could greatly increase gains in plant breeding programs (Meuwissen et al., 2001; Jannink et al., 2010).

The advantage of GBS for GS in breeding programs is the low per-sample cost needed for generating tens of thousands to hundreds of thousands of molecular markers. Poland et al. (2012b) have demonstrated the suitability for GBS markers in developing GS models in the complex wheat genome. They were able to demonstrate prediction accuracies for yield and other agronomic traits that are high enough to be suitable for breeding applications. The GBS markers also showed a significant improvement in the attained prediction accuracy over a previously used array of hybridization-based markers. The important finding of this work is the practical implications in breeding. The training population was genotyped without a priori knowledge of the population or SNPs and per-sample cost was below $20 (Poland et al., 2012b).

Putting Genotyping-by-Sequencing to Work

Looking forward, high-density markers from NGS will soon be applied to almost every genomic question. These marker datasets are low cost and dynamic, with data and genotyping results getting more robust and economical each year. Genotyping-by-sequencing has been shown to be a valid tool for genetic mapping (Baird et al., 2008; Elshire et al., 2011; Poland et al., 2012a), breeding applications (Poland et al., 2012b), and diversity studies (Fu, 2012; Lu et al., 2012). The ability to quickly generate robust datasets without considerable prior effort for marker discovery is quickly dispelling issues that have plagued researchers working with obscure or foreign species: a lack of defined and specific genetic tools for genome analysis (Allendorf et al., 2010).

Genotyping-by-sequencing is an ideal platform for studies ranging from quickly identifying single gene markers to whole genome profiling of association panels.

Perhaps one of the most exciting applications of GBS will be in the field of plant breeding. Theoretical and preliminary studies on genomic selection show great promise for accelerating the rate of developing new improved varieties. Genotyping-by-sequencing is providing a rapid and low-cost tool for genotyping these populations, allowing breeders to implement genomic selection on a large scale in their breeding programs. Current developments in sequencing output will drive per-sample cost below $10. Furthermore, there is no requirement for a priori knowledge of the species as the GBS methods have been shown to be robust across a range of species and SNP discovery and genotyping are completed together. This is a very important feature for moving genomics-assisted breeding into orphan crops with understudied genomes and commercial crops with large and complex genomes. Challenges remaining include data management as well as computational constraints on huge datasets, though the future looks promising. Genomic selection via GBS stands to be a major supplement to traditional crop development. The potential for GBS data to improve breeding systems through GS is enormous.

The application of sequence-based genotyping for a whole range of diversity and genomic studies will have an important place well into the future. Driven by applications across the whole spectrum of human, microbial, plant, and animal genomics, developments in NGS and genomics platforms must be put to use for plant breeding and genetics studies.

Acknowledgments

USDA-ARS and the USDA-NIFA funded Triticeae Coordinated Agriculture Project (T-CAP) (2011-68002-30029) provided support for T. Rife. This manuscript was greatly improved by the helpful comments of two anonymous reviewers. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References

Allendorf, F.W., P.A. Hohenlohe, and G. Luikart. 2010. Genomics and the future of conservation genetics. Nat. Rev. Genet. 11:697–709. doi:10.1038/nrg2844

Altshuler, D., V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, and E.S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516. doi:10.1038/35035083

Andolfatto, P., D. Davison, D. Erezyilmaz, T.T. Hu, J. Mast, T. Sunayama-Morita, and D.L. Stern. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21:610–617. doi:10.1101/gr.115402.110

Arumuganathan, K., and E.D. Earle. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9:415–415.

Ashelford, K., M.E. Eriksson, C.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 Biol. 12:R28. doi:10.1186/gb-2011-12-3-r28

Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko, and E.A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376. doi:10.1371/journal.pone.0003376

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. doi:10.1086/521987

Byers, R.L., D.B. Harker, S.M. Yourstone, P.J. Maughan, and J.A. Udall. 2012. Development and mapping of SNP assays in allotetraploid cotton. Theor. Appl. Genet. 124:1201–1214. doi:10.1007/s00122-011-1780-8

Chia, J.-M., C. Song, P.J. Bradbury, D. Costich, N. de Leon, J. Doebley, R.J. Elshire, B. Gaut, L. Geller, J.C. Glaubitz, M. Gore, K.E. Guill, J. Holland, M.B. Hufford, J. Lai, M. Li, X. Liu, Y. Lu, R. McCombie, R. Nelson, J. Poland, B.M. Prasanna, T. Pyhäjärvi, T. Rong, R.S. Sekhon, Q. Sun, M.I. Tenaillon, F. Tian, J. Wang, X. Xu, Z. Zhang, S.M. Kaeppler, J. Ross-Ibarra, M.D. McMullen, E.S. Buckler, G. Zhang, Y. Xu, and D. Ware. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44:803–807. doi:10.1038/ng.2313

Cockram, J., J. White, D.L. Zuluaga, D. Smith, J. Comadran, M. Macaulay, Z. Luo, M.J. Kearsey, P. Werner, D. Harrap, C. Tapsell, H. Liu, P.E. Hedley, N. Stein, D. Schulte, B. Steuernagel, D.F. Marshall, W.T.B. Thomas, L. Ramsay, I. Mackay, D.J. Balding, R. Waugh, and D.M. O'Sullivan. 2010. Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc. Natl. Acad. Sci. USA 107:21611–21616. doi:10.1073/pnas.1010179107

Craig, D.W., J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T.L. Pawlowski, T. Laub, G. Nunn, D.A. Stephan, N. Homer, and M.J. Huentelman. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods 5:887–893. doi:10.1038/nmeth.1251

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 Res. 36:e122. doi:10.1093/nar/gkn502

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. Nat. Rev. Genet. 12:499–510. doi:10.1038/nrg3012

Deschamps, S., M. la Rota, J.P. Ratashak, P. Biddle, D. Thureen, A. Farmer, S. Luck, M. Beatty, N. Nagasawa, L. Michael, V. Llaca, H. Sakai, G. May, J. Lightner, and M.A. Campbell. 2010. Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. Plant Gen. 3:53–68. doi:10.3835/plantgenome2009.09.0026

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. doi:10.1371/journal.pone.0019379

Fu, Y.-B. 2012. Genotyping-by-sequencing: A case study in barley. Workshop presented at: Genomics of Genebanks. Plant and Animal Genome Conference XX, San Diego, CA. 14–18 Jan. 2012. Workshop W362.

Futschik, A., and C. Schlötterer. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–218. doi:10.1534/genetics.110.114397

Gan, X., O. Stegle, J. Behr, J.G. Steffen, P. Drewe, K.L. Hildebrand, R. Lyngsoe, S.J. Schultheiss, E.J. Osborne, V.T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E.J. Belfield, N.P. Harberd, E. Kemen, C. Toomajian, P.X. Kover, R.M. Clark, G. Rätsch, and R. Mott. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477:419–423. doi:10.1038/nature10414

Geraldes, A., J. Pang, N. Thiessen, T. Cezard, R. Moore, Y. Zhao, A. Tam, S. Wang, M. Friedmann, I. Birol, S.J.M. Jones, Q.C.B. Cronk, and C.J. Douglas. 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. Mol. Ecol. Resour. 11:81–92. doi:10.1111/j.1755-0998.2010.02960.x

Gore, M.A., J.M. Chia, R.J. Elshire, Q. Sun, E.S. Ersoz, B.L. Hurwitz, J.A. Peiffer, M.D. McMullen, G.S. Grills, and J. Ross-Ibarra. 2009a. A first-generation haplotype map of maize. Science 326:1115–1117. doi:10.1126/science.1177837

Gore, M.A., M.H. Wright, E.S. Ersoz, P. Bouffard, E.S. Szekeres, T.P. Jarvie, B.L. Hurwitz, A. Narechania, T.T. Harkins, G.S. Grills, D.H. Ware, and E.S. Buckler. 2009b. Large-scale discovery of gene-enriched SNPs. Plant Gen. 2:121–133. doi:10.3835/plantgenome2009.01.0002

Hamblin, M.T., T.J. Close, P.R. Bhat, S. Chao, J.G. Kling, K.J. Abraham, T. Blake, W.S. Brooks, B. Cooper, C.A. Griffey, P.M. Hayes, D.J. Hole, R.D. Horsley, D.E. Obert, K.P. Smith, S.E. Ullrich, G.J. Muehlbauer, and J.-L. Jannink. 2010. Population structure and linkage disequilibrium in U.S. barley germplasm: Implications for association mapping. Crop Sci. 50:556–566. doi:10.2135/cropsci2009.04.0198

Harper, A.L., M. Trick, J. Higgins, F. Fraser, L. Clissold, R. Wells, C. Hattori, P. Werner, and I. Bancroft. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat. Biotechnol. 30:798–802. doi:10.1038/nbt.2302

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4:65–75. doi:10.3835/plantgenome.2010.12.0029

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. Mol. Ecol. Resour. 11:117–122. doi:10.1111/j.1755-0998.2010.02967.x

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. Wang, J. Guan, D. Fan, Q. Weng, T. Huang, G. Dong, T. Sang, and B. Han. 2009. High-throughput genotyping by whole-genome resequencing. Genome Res. 19:1068–1076. doi:10.1101/gr.089516.108

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, M. Li, D. Fan, Y. Guo, A. Wang, L. Wang, L. Deng, W. Li, Y. Lu, Q. Weng, K. Liu, T. Huang, T. Zhou, Y. Jing, W. Li, Z. Lin, E.S. Buckler, Q. Qian, Q.-F. Zhang, J. Li, and B. Han. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961–967. doi:10.1038/ng.695

Hyten, D.L., Q. Song, E.W. Fickus, C.V. Quigley, J.-S. Lim, I.-Y. Choi, E.-Y. Hwang, M. Pastor-Corrales, and P.B. Cregan. 2010. High-throughput SNP discovery and assay development in common bean. BMC Genomics 11:475. doi:10.1186/1471-2164-11-475

International Barley Sequencing Consortium. 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature (in press).

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Briefings Funct. Genomics 9:166–177. doi:10.1093/bfgp/elq001

Jiao, Y., H. Zhao, L. Ren, W. Song, B. Zeng, J. Guo, B. Wang, Z. Liu, J. Chen, W. Li, M. Zhang, S. Xie, and J. Lai. 2012. Genome-wide genetic changes during modern breeding of maize. Nat. Genet. 44:812–815. doi:10.1038/ng.2312

Lu, F., A.E. Lipka, R.J. Elshire, J. Glaubitz, J. Cherney, M. Casler, E.S. Buckler, and D. Costich. 2012. Characterization of the genetic diversity of switchgrass using genotyping by sequencing. Poster presented at: Poster Session – Even Numbers. Plant and Animal Genome Conference XX, San Diego, CA. 14–18 Jan. 2012. Poster P0195.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39:906–913. doi:10.1038/ng2088

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–141. doi:10.1016/j.tig.2007.12.007

Metzker, M. 2009. Sequencing technologies – The next generation. Nat. Rev. Genet. 11:31–46. doi:10.1038/nrg2626

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

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 Res. 17:240–248. doi:10.1101/gr.5681207

Monson-Miller, J., D.C. Sanchez-Mendez, J. Fass, I.M. Henry, T.H. Tai, and L. Comai. 2012. Reference genome-independent assessment of mutation density using restriction enzyme-phased sequencing. BMC Genomics 13:72.

Morrell, P.L., E.S. Buckler, and J. Ross-Ibarra. 2011. Crop genomics: Advances and applications. Nat. Rev. Genet. 13:85–96.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, and E.S. Buckler. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202. doi:10.1105/tpc.109.068437

Nelson, J.C., S. Wang, Y. Wu, X. Li, G. Antony, F.F. White, and J. Yu. 2011. Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC Genomics 12:352. doi:10.1186/1471-2164-12-352

Nielsen, R., J.S. Paul, A. Albrechtsen, and Y.S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451. doi:10.1038/nrg2986

Peterson, B.K., J.N. Weber, E.H. Kay, H.S. Fisher, and H.E. Hoekstra. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135.

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253. doi:10.1371/journal.pone.0032253

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sanchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. (in press). doi:10.3835/plantgenome2012.06.0006

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795

Stanzione, D. 2011. The iPlant collaborative: Cyberinfrastructure to feed the world. Computer 44:44–52. doi:10.1109/MC.2011.297

Truong, H.T., A.M. Ramos, F. Yalcin, M. de Ruiter, H.J.A. van der Poel, K.H.J. Huvenaars, R.C.J. Hogers, L.J.G. van Enckevort, A. Janssen, N.J. van Orsouw, and M.J.T. van Eijk. 2012. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS ONE 7:e37565. doi:10.1371/journal.pone.0037565

van Orsouw, N.J., R.C.J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers, E. Verstege, H. Schneiders, H. van der Poel, J. van Oeveren, H. Verstegen, and M.J.T. van Eijk. 2007. Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 2:e1172. doi:10.1371/journal.pone.0001172

van Tassell, C.P., T.P.L. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C.T. Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren, and T.S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252. doi:10.1038/nmeth.1185

Wang, S., E. Meyer, J.K. McKay, and M.V. Matz. 2012. 2b-RAD: A simple and flexible method for genome-wide genotyping. Nat. Methods 9:808–810. doi:10.1038/nmeth.2023

Wang, X., H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. 43:1035–1039. doi:10.1038/ng.919

Wetterstrand, K.A. 2012. DNA sequencing costs: Data from the NHGRI large-scale genome sequencing program. National Human Genome Research Institute, Bethesda, MD. http://www.genome.gov/sequencingcosts (accessed 5 Mar. 2012).

Wiedmann, R.T., T.P.L. Smith, and D.J. Nonneman. 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. 9:81. doi:10.1186/1471-2156-9-81

Xie, W., Q. Feng, H. Yu, X. Huang, Q. Zhao, Y. Xing, S. Yu, B. Han, and Q. Zhang. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. USA 107:10578–10583. doi:10.1073/pnas.1005931107

Xu, X., X. Liu, S. Ge, J.D. Jensen, F. Hu, X. Li, Y. Dong, R.N. Gutenkunst, L. Fang, L. Huang, J. Li, W. He, G. Zhang, X. Zheng, F. Zhang, Y. Li, C. Yu, K. Kristiansen, X. Zhang, J. Wang, M. Wright, S. McCouch, R. Nielsen, J. Wang, and W. Wang. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30:105–111. doi:10.1038/nbt.2050

Xu, X., S. Pan, S. Cheng, B. Zhang, D. Mu, P. Ni, et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475:189–195. doi:10.1038/nature10158

Yang, H., Y. Tao, Z. Zheng, C. Li, M. Sweetingham, and J. Howieson. 2012. Application of next-generation sequencing for rapid marker development in molecular plant breeding: A case study on anthracnose disease resistance in *Lupinus angustifolius* L. BMC Genomics 13:318. doi:10.1186/1471-2164-13-318

You, F.M., N. Huo, K.R. Deal, Y.Q. Gu, M.-C. Luo, P.E. McGuire, J. Dvorak, and O.D. Anderson. 2011. Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genomics 12:59. doi:10.1186/1471-2164-12-59