Choosing the right platform for your agriculture genomic applications

Microarrays vs. sequencing-based genotyping—pros and cons

Introduction

Rapid advances in next-generation sequencing (NGS) have allowed agricultural scientists to help build an extensive resource of genomes, which will become a "genomic library of the living world". Academic scientists, animal breeders, and commercial seed companies are beginning to tap into this vast genomic library resource to power agrigenomic strategies. By applying genotyping of genomic markers to the characterization and selection of important traits, they aim to improve productivity and commercial viability.

This technical note consolidates experience of sequence-based genotyping approaches as described in peer-reviewed journals and compares the performance of microarrays for agriculture genotyping applications to assist in making genotyping technology decisions.

Overview of sequence-based genotyping

Genomic selection and association mapping or linkage disequilibrium (LD) mapping techniques require large numbers of markers to accurately estimate the trait associated with the genotypes. This requires that the technology used in obtaining the genotypic information be cost-effective, efficient, and high-throughput. Whole-genome sequencing and targeted genotyping using sequence capture are expensive and low-throughput methods of generating genotypic data and remain impractical for use in routine applications. In the goal to identify cost-effective sequencing, sequencing-based genotyping approaches such as restriction site–associated DNA sequencing (RADseq) using restriction enzyme (RE) [1], and genotyping-by-sequencing (GBS) [2] have evolved and are referenced for their potential in research and routine-use applications.

The sequencing-based approaches rely on barcoding to multiplex samples and reduce genotyping costs. The techniques use restriction enzymatic digestion to target restriction sites and low-copy genomic regions to reduce genomic complexity. This offers the capability to avoid regions with repetitive sequences that are predisposed to producing ambiguous or false single-nucleotide polymorphisms (SNPs) and also increase sequencing costs. Genotypic data obtained using sequencing varies in quality and quantity, and is highly dependent on both the genome size and structure of the organism and the populations being evaluated. The complexity in the genomic structure such as ploidy level, amount of GC content and repetitive sequences, genetic diversity within populations being studied, and the mating systems within the populations has a direct impact on the cost, accuracy, and efficiency of the sequencing technique to collect genotypic data both accurately and easily. The sequencing-based genotyping techniques are useful in genotyping species where marker discovery is lagging or incomplete. Both GBS and RADseq can be used on as few as 96 samples without access to a reference genome or previously discovered markers. The techniques also are ideal for screening thousands of polymorphisms to understand the consequences of genetic variation, which traditionally has relied on a very small number of markers such as microsatellites and amplified fragment length polymorphisms (AFLPs). Sequencing-based genotyping techniques have been used to perform marker discovery. Results from the various experiments on species such as barley, maize, wheat, bovine, and trout among others have been published in journals.

The capability to use the sequencing-based genotyping technologies for routine genotyping remains elusive for several reasons, some of which are listed in this technical note. A special issue on GBS techniques in the journal *Molecular Ecology* summarized the state of the newer GBS techniques as incomplete and not sufficiently scalable across different plant and animal species [3].

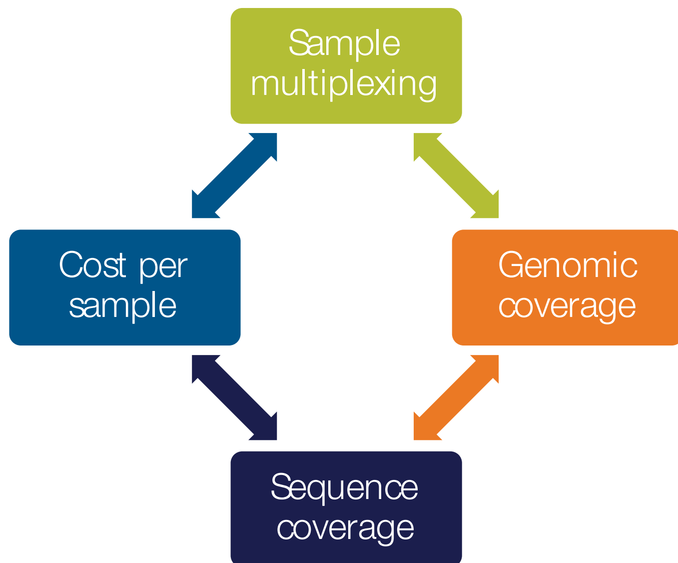

Figure 1. All applications on NGS platforms must balance four key parameters.

Key experimental factors in sequence-based genotyping

All NGS platforms have a limit to the number of bases of sequence that can be output from each sequencing run. This finite output capacity means that sequencing-based genotyping runs must balance four key parameters: level of sample multiplexing, genomic coverage, sequence coverage, and cost per sample (Figure 1).

Multiplexing of samples is critical because the finite output capacity of the sequencer must be shared across all the samples included in the run. More samples means less bases of sequence capacity available to each sample.

Genomic coverage is important because it dictates the percentage of the genome that is assayed and, therefore, the number of markers in the genome that are accessible. Greater genomic coverage is achieved at the expense of one of the other parameters because it uses more of the output capacity of the sequencer.

Sequence coverage (or sequence "depth") indicates the mean number of reads per sequence across the data set. In reality, some sequences will be read more often and some less often, or not at all. The sequence coverage affects the percentage of gaps in the data and also genotype accuracy. Accurate genotype calls typically require 30x or more coverage across each SNP. Increasing sequence coverage also forces a trade-off somewhere else to balance the use of the sequencer's capacity.

Of course, sample multiplexing, genomic coverage, and sequence coverage can all be improved by investing in more runs on the sequencer, but this rapidly increases cost.

This technical note discusses the impact of experimental approaches in each of the newer sequencing techniques, the impact of genomic complexity on the number of markers, and the application range. The variability in experimental approaches can dramatically increase costs within any genotyping program 5-fold for fewer than 30,000 markers.

Applications

Genomic coverage is highly dependent on the choice of genotyping technology and method, and this choice ultimately depends on the application of interest. Each method offers different levels of coverage of the plant or animal genome. This impacts the number of markers that can be accessed and dictates which method is appropriate for the target application. These range from population genomic scans to determining the phylogeny. Figure 2 shows how the different sequencing- and array-based genotyping methods map to applications and the relative cost associated with covering the genome.

The number of markers covered by each sequencing-based method depends on experimental parameters such as type of restriction enzymes, DNA quality and quantity, and analysis techniques. The number of markers available for each application shown in Figure 2 is a function of the fraction of the genome that was sequenced. The number of markers that are required for accurate genotyping is a function of the number of genome-wide levels of LD, the recombination events captured in the pedigree, and the divergence among groups [4]. Genomic coverage can be increased by changing the restriction enzyme in the assay to increase the number of tags. However, as noted

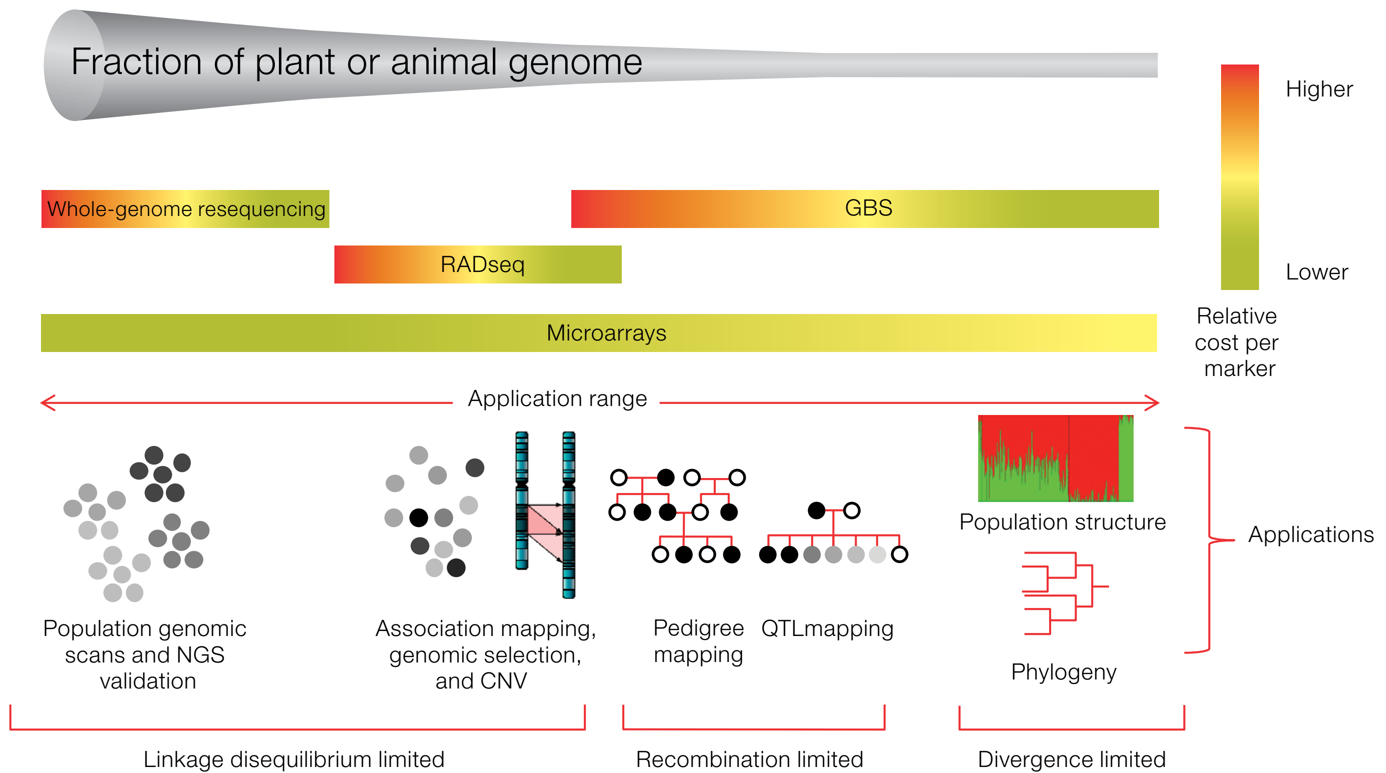

Figure 2. Representation of the optimal genomic coverage for various agrigenotyping applications and the scalability of cost versus genotyping methodology. The color gradient shows the cost per marker associated with obtaining the desired number of markers relative to each method. Sequencing costs increase with an increase in the fraction of the genome being sequenced. Consequently, cost per marker is low or comparable with cost of arrays when sequencing only a few hundred markers for phylogeny or population structure. Microarrays offer a cost-effective technology to validate markers discovered using resequencing or any of the sequencing-based genotyping techniques offering the capability to screen different populations.

above, any move to increase genomic coverage will be at the expense of lower sample multiplexing, lower sequence coverage, or higher cost per sample.

Fraction of plant or animal genome

- **Population genomic scans and sequencing validation:** Arrays and sequencing-based genotyping techniques have been used in conducting population scans and verifying markers that have been discovered using NGS. Sequencing-based genotyping strategies are susceptible to calling false SNPs because of inherent errors within the sequencing technology, copy number variants that are not mapped to the reference genome, and from paralogs or homologs. False SNPs can be eliminated through the use of deeper sequence coverage, although this increases cost per sample, or through use of double haploids or a high-quality reference sequence, but these can lead to more complicated informatics pipelines with stringent filtering criteria that discard a majority of the sequenced data. Markers are verified by running a large diverse set of samples within the population to identify informative and reproducible polymorphic markers. High-density Applied Biosystems™ Axiom™ microarrays have been applied successfully to verify sequencing discoveries and eliminate false SNPs that are the result of sequencing errors in a number of species, including chicken [5] and salmon [6]. Microarrays offer an easy way to evaluate millions of markers across a diverse population and also verify markers that may have been discovered through the different sequencing techniques such as RADseq, RNAseq, and resequencing [7].

- **Association mapping, genomic selection, and copy number applications:** Association mapping (AM) techniques use large numbers of polymorphic markers to overcome the challenges and limitations in QTL mapping. AM rely on the LD and the recombination present within the existing gene pool to perform phenotype-genotype correlation among random mating populations, general collection of lines, or germplasm [8]. In AM studies, additional markers increase the likelihood of finding or tagging causal variants [9], so more markers is better. Although AM can be done by sequencing-based genotyping methods, arrays can usually genotype high densities of markers more cost-effectively and with much greater data quality and completeness.

Dense sets of markers are also used in genomic selection where the effects of markers are simultaneously estimated in a genotyped and phenotyped test or training population and then used to predict values of selection candidates. The accuracy of genomic selection increases with an increase in marker density. 50,000 markers are expected to be sufficient for accurate estimation of the relationships [10]. Copy number variation detection offers the capability to study and identify heritable variation in complex traits.

- **Pedigree and quantitative trait loci (QTL) mapping:** In contrast to association mapping, QTL mapping looks at the effects of multiple genes on quantitative traits such as the QTL in salmon controlling resistance to sea lice or the size of seed. QTL identification is based on bi-parental crosses and requires identification of the individual genes in the chromosomal regions through fine mapping, requiring large numbers of crosses to generate sufficient numbers of meiotic events. Pedigree genotyping makes use of the breeding material in the QTL detection, covering multiple generations and linking many crosses through their common ancestors in the pedigree. This offers identification and use of most alleles present in an ongoing breeding program.
- **Phylogeny and population mapping:** Population structures can vary across accessions maintained by the research institution and across germplasm maintained by breeders. Different population structure requires different genome-wide analysis study (GWAS) methods. Surveying population structure and performing phylogeny through construction of genetic or linkage maps provide information on the recombination rate across the genome. Understanding population structure also helps in selecting the appropriate markers and density. Population analysis can be completed by studying markers in a very small fraction of the genome.

Workflow: A comparison of the workflows for array-based technologies and sequencing-based genotyping techniques is shown in Figure 3. The sequencing-based genotyping technologies rely on the use of barcodes to multiplex samples (e.g., 96 samples sequenced in a single lane will require 96 sample barcodes). Library preparation requires choice of an RE that is appropriate for the species and the number of markers required. This process requires optimization to avoid issues such as primer dimer that can increase sequencing costs. After library preparation, the actual sequencing can take anywhere from 11 hours to 11 days, depending on the instrument capability and the percentage of the genome that is sequenced. Higher sample multiplexing is possible but, as noted earlier, must be balanced against the size of the genome, the percentage of the genome that is sequenced, and the sequence coverage in a single lane. After sequencing, the data is filtered and barcodes are demultiplexed to retrieve the markers for each sample.

An important consideration in using any genotyping technology is the computing infrastructure and the analysis pipelines. Analysis pipelines need to be customized to the species of interest, experimental approach, populations being studied, and the underlying technology itself.

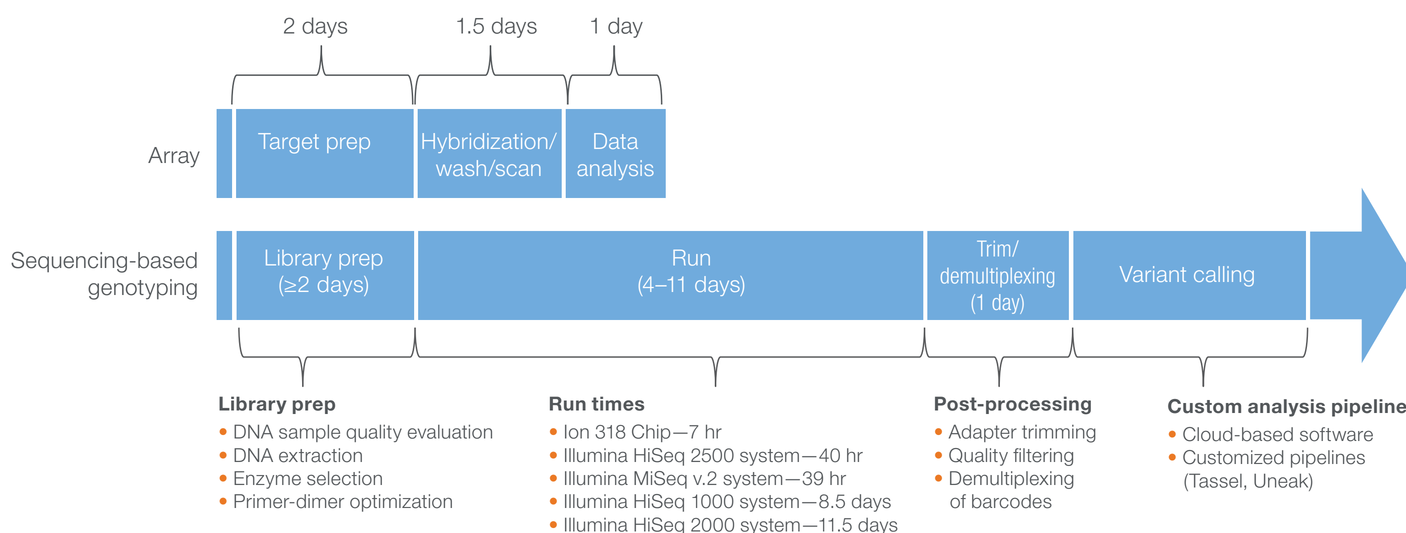

Figure 3. Workflow comparison of array- and sequencing-based genotyping techniques.

A recent publication from the James Hutton Institute concluded that an important practical outcome of the work on using GBS for studying barley was that the GBS data was more challenging to handle and subsequently to analyze than the current multiplex SNP assay technology that was run in the lab [11]. Computing infrastructure, bioinformatics specialists to maintain custom analysis pipelines and software to perform alignment and analysis, and time required to extract useful genotyping data are among the many challenges in adopting sequencing-based genotyping techniques. Analysis of the data from the sequencing techniques is often performed in the "cloud" to minimize local data storage and computing requirements. Variant calling is often performed through customized software pipelines that call the genotypes. Each storage technology has costs associated with transfer, storage, and retrieval of the data, which can affect the cost of a genotyping program.

In contrast, array-based genotyping techniques can easily genotype millions of data points per sample using a desktop workstation, keeping infrastructural costs low and operating efficiency high. The simplicity and ease of use of array-based genotyping allows microarrays to be used in all scenarios and environments, and are especially attractive in routine breeding applications when processing large numbers of samples and turnaround time are both important.

DNA quality and quantity in library preparation: The demands placed by the sequencing-based genotyping techniques on the amount of DNA concentration and the quality of the DNA remains one of the more serious challenges in using the method for practical applications. DNA sequencing requires several micrograms (µg) of purified, high molecular weight genomic DNA that is free of contaminants and symbionts. Bacterial contamination can impact sequencing because random amplification of the DNA material means both the bacterial DNA and the biological sample that is being genotyped are sequenced simultaneously.

Table 1. Amount of DNA required for genotyping biological samples in microarrays compared with sequencing-based genotyping techniques. Sequencing-based genotyping techniques require between 2x and 30x the amount of concentration that is required for microarrays.

Amount of DNA recommended	Microarrays	Sequencing-based genotyping methods
Diploid plants and animals	100 ng	200–3,000 ng
Polyploid plants and animals	150 ng	

Number and type of markers: The single biggest difference between arrays and sequencing-based genotyping is the capability to target specific chromosomal regions or specific SNPs with arrays, as shown in Figure 4. Microarray-based technologies can target any number of markers within specific chromosomal regions with a design strategy that employs markers that are equally spaced across the genome or at higher spacing in specific regions of the genome if necessary. This flexibility offers capability to use arrays in GWAS [12], QTL mapping, association mapping, and genomic selection with a certain amount of ascertainment bias. Ascertainment bias can be reduced by performing SNP discovery in a diverse set of cultivars [13].

Figure 4. Marker selection technique in arrays and sequencing-based genotyping. Arrays are able to target haplotype blocks within the genome using markers that target those specific regions. In contrast, sequence-based genotyping techniques can target areas across the genome through markers that are located near restriction sites.

Sequencing-based genotyping techniques rely on random sampling of DNA pools, and the number of markers is proportional to the number and size of the regions being sequenced. The genomic regions are expected to be bias-free to the extent that the restriction sites are conserved in the population being studied. As a result, markers are not conserved across samples, and no two samples provide the same set of markers.

This leads to missing data and requires complex informatics to recover the missing data through imputation. Markers that are not conserved across samples have to be imputed from the reference genome or haplotypes from related lines using very high–coverage sequencing (18x or higher).

Factors to consider in choice of sequencing methodology

It is important to consider the various factors that can affect the genotyping before deciding which approach to employ for any of the above mentioned applications.

- **Heterozygote call errors:** The sequencing-based genotyping techniques, especially GBS, rely on low coverage to keep costs low and achieve a large number of markers that can be used for association mapping. The downside of this experimental approach is significant undercalling of heterozygotes, which compromises genotype accuracy. GBS undercalls heterozygotes by as much as 50%. A study on required coverage in DNA sequencing predicts that for each heterozygous diploid, a depth of 13.5x is required to detect both alleles at least once for 99.75% of positions [14]. To detect each allele at least twice, a depth of 18x would be required. Increasing the sequencing coverage results in a higher cost per sample and makes it more expensive than microarrays. Studies in *Vitis* (grape) demonstrated 30–50% heterozygote undercalling when genotyped at a mean depth of 5.7x [15]. The heterozygote call accuracy on arrays is determined by the probe design, and this is highly predictable, leading to genotype call accuracies approaching 100%. Further design methodologies used with arrays are capable of genotyping markers in genomics regions that have greater than 60% GC content.
- **Genomic coverage:** The number of markers delivered by any technology is expected to provide uniform coverage of the genome. Sequencing-based genotyping techniques exhibit missing data, which results in nonuniform genome coverage. The missing data is a function of the experimental conditions and genome structure, and results from a combination of library complexity (i.e., number of unique sequence tags) and sequence coverage of the library. The amount of missing data is directly correlated to the level of sample multiplexing during library preparation and the type of enzymes used for RE digestion. The choice of the RE in the sequencing technique impacts allele dropout, which impacts population genetics statistics. Rare markers require enzymes that cut less frequently but then result in fewer markers. The use of enzymes that cut more frequently results in more markers but at much lower coverage, resulting in large amounts of missing data.

To minimize costs and increase throughput in sequencing techniques, genomic complexity needs to be reduced. The downside of using complexity reduction is that the genotypic data that is acquired has a significant amount of missing data [16]. Genotypic data can be missing because of inherent variation within the genomic structure such as presence-absence variation, variability in polymorphic restriction sites, and differential methylation that can impact the methylation-sensitive enzymes used in reduced-representation sequencing techniques. Missing data becomes critical in QTL mapping where genotype data quality of the parental lines is extremely important for genotype calling of the mapping population. The parental lines then need to be sequenced at a very high depth of coverage.

Figure 5 demonstrates the relationship between sequence coverage and missing data, as shown in a recent publication comparing the sequencing-based genotyping on different platforms [17]. The study demonstrated that at 10x coverage it is possible to get 1,000 markers with 50% data missing and increase the number of markers to 30,000 markers with 90% data missing through low coverage.

The amount of genotypic data that is expected versus what is achieved can vary widely. A recent study in maize using GBS demonstrated skewed coverage of genomic positions' distribution, with a disproportionate number of regions having no representation for which information was initially expected [18]. This limited the scope and application of the sequencing technique and was found to be unusable for fine mapping for association studies. Genotypic data for a majority of the desired sites was achievable only through a substantial increase in read depth, impacting the cost of sequencing.

The missing data are recovered using imputation techniques by aligning the data to reference genomes that require significant investment, advanced analysis, and complicated pipelines that can filter, sort, and align sequence data. The lack of an easy-to-use, uniform informatics pipeline remains the second biggest hurdle in adopting the sequencing-based genotyping techniques

Figure 5. Data from a recent publication on application of GBS in barley showing relationship between number of sequence reads, sequencing depth, and read calls. (A) The average percentage of missing data per SNP in each sequenced sample is plotted as a function of the number of sequence reads in that sample. **(B)** Histogram of missing data per SNP. **(C)** The number of SNP calls plotted against the minimum depth at a variant position in a given sample to make a successful genotype call. All SNP calls were made with the SAMtools pipeline. The minor allele frequency was set to 30%, and the maximum rate of missing data was set to 50%. The sequencing platforms used for this study include Illumina™ HiSeq™ 2000 system (black), Ion Proton™ System (red), and Ion PGM™ System (green). doi:10.1371/journal.pone.0076925.g002.

for routine use. Imputation works well with closely related individuals, but in highly diverse samples, the missing data is replaced by alleles from the nearest neighbor [19]. The sequencing-based genotyping techniques also miss low-frequency alleles when there is a high proportion of missing data. The alternative is to pursue a higher depth of sequencing, which results in a higher cost per sample.

- **LD and polymorphic frequency:** The genomic diversity and mating systems within the populations from which genotypic data is collected can have a big impact on sequencing costs. Populations derived from a narrow genetic base exhibit fewer polymorphisms, requiring more sequencing and increasing the total cost. This is true in species such as tetraploid cotton that exhibits one polymorphism every 1,000–1,500 bases. The LD decay within the species also dictates the number of markers that are required for association mapping in diverse populations. Figure 6 demonstrates the impact of LD decay on marker resolution. Low resolution on marker density in species that have a high LD decay will result in inadequate coverage of the genome. Aquaculture species such as trout and plants, including maize, grape, and sugar beet, exhibit low LD and require a large number of fragments for association analysis. A recent study on whole-genome association studies in trout using sequencing-based genotyping techniques came to the conclusion that the rapid level of LD decay required higher levels of marker density to effectively conduct whole-genome association studies than what was available through the sequencing-based genotyping technologies [20].

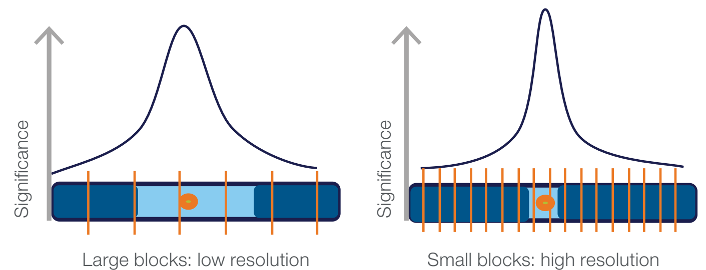

Figure 6. More SNPs are required to cover the genome and to yield a higher-resolution genetic map in species that have rapid LD decay.

- **Copy number applications:** Microarrays and sequencing-based genotyping technologies are being used to perform copy number studies to identify heritable variation in complex traits. Both technologies are capable of detecting copy number gains. Sequencing-based genotyping techniques have difficulty identifying copy number losses under low coverage because the segment loss shows up as low-coverage markers [21]. Higher coverage would allow detection of CNV losses, but costs were expected to increase by 40–50%.

- **Genomic complexity:** Polyploidy is one of the more complex attributes in plants and certain animals. 60–70% of angiosperms are polyploid, with the ploidy ranging from tetraploid in grapeseed to octoploid in strawberry and even more complex in sugarcane, where the ploidy can range from 12x to 16x. Polyploid species exhibit genome duplication. The challenges from polyploidy are as follows: (i) Polyploid species require increased sequence coverage to cover the increased size of genomes effectively increasing sequencing cost.

(ii) The genome assembly and mapping algorithms are complex and prone to errors especially when assembling paralogous and orthologous regions. In species that are both polyploid and heterozygous, imputation of data at each given locus requires a complicated analysis pipeline that cannot be used in routine breeding applications [22]. Furthermore, the higher depth of sequencing increases the total cost. Measuring the allele dosage information at each locus in the genome is important in genomic selection models. When using microarrays to genotype polyploid species, the signal contribution from the alternate subgenomes results in cluster compression. The polyploid species also exhibit varying ploidy because of complexity reduction created by interfering mutations. It is important to have an analysis pipeline that can automatically cluster and assign genotypes when used in routine applied applications to meet the tight breeding timelines. The Applied Biosystems™ Axiom™ GT1 algorithm uses Bayesian statistics to accurately assign genotypes to and cluster data from polyploid genomes. An example is shown in Figure 7. The automated pipeline offers ease of use and the capability to genotype thousands of markers across thousands of samples easily and accurately.

Guidance for choosing the right solution for your genotyping program

In light of advances in both microarray technology and sequencing techniques, scientists need to be aware of the challenges in using sequencing techniques and likewise the bias in both sequencing and array technologies. The following questions can help in making decisions on the appropriate technology to be used for the species under consideration in a research or genomic breeding program:

- Is the assay used for acquiring genotypic data compatible with the genomic structure of the species under consideration, and can it deliver a sufficient number of reliable markers?
- Are there specific chromosomal regions that need to be targeted, and what is the marker strategy that needs to be employed for covering the entire genome?
- What is the underlying LD structure of the species under consideration?
- Is the species a polyploid species, and what is the ploidy?
- What kind of informatics pipeline and expertise are required to incorporate the technology into the research or breeding program?
- How many hours will it take to call the genotypes and cluster the data?
- Are the populations under consideration inbred or diverse, unrelated individuals, and what is the expected level of heterozygosity?
- How many samples need to be genotyped, and are there any constraints on turnaround time or time to results?
- What depth of sequence coverage is required to accurately call the genotypes?
- What is the impact of data gaps, and how will you recover the missing genotypes?
- What is the cost of the assay taking into account the amount of missing data and the resources required for the bioinformatics pipeline and analysis?
- What is the throughput, turnaround time of the technology, and reliability of the assay and instrumentation used in the technology?
- What is the amount of bias that is acceptable in the breeding program, and are there ways to get around the bias?
- How many different technologies or assays need to be integrated into the research or breeding program for effective validation, marker trait, or routine use applications?

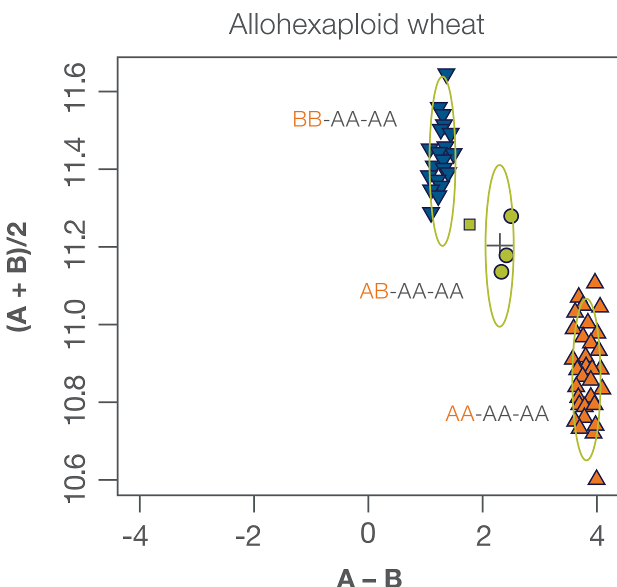

Figure 7. Data from the Applied Biosystems™ Axiom™ Genotyping Wheat Array demonstrating the cluster compression in hexaploid wheat because of signal from an alternate subgenome and the ability of the Axiom genotyping algorithm to resolve clusters.

Microarray technology continues to evolve with the Applied Biosystems™ Axiom™ 384HT format. This recent innovation in processing 384 samples simultaneously at extremely economical price points has allowed the technology to go from research to mainstream commercial agrigenomics. Microarrays remain the technology of choice in applications where turnaround time, ease of use, and data quality have the highest priority.

Array-based technologies offer the flexibility and affordability to consolidate multiple genotyping applications under a single platform. Innovations in assay and informatics analysis pipelines allow for unconstrained genotyping of all relevant markers of interest with results delivered in a few hours through an easy-to-use and simple workflow. Applied Biosystems™ Axiom™ Genotyping Solution offers complete solutions for applications ranging from genome-wide analysis to routine testing with the highest accuracy and reproducibility, a straightforward workflow, and the lowest cost.

A summary of the challenges with sequencing-based genotyping techniques is presented in Table 2.

Agrigenomics genotyping solutions provide breeders and researchers with a powerful cost-effective genotyping tool to identify, verify, and test or analyze complex genetic traits in plants or animals for faster and more precise breeding. Axiom genotyping begins with the selection of marker content from SNP library resources, continues with the design of a SNP array, and concludes with an array to identify sample genotype. This provides a functional genotyping tool for breeders and researchers to apply in marker-trait association, genome-wide association studies (GWAS), quantitative trait loci (QTL) analysis, and genomic selection programs.

Table 2. Comparison of features of sequencing-based genotyping techniques such as RADseq and GBS with Axiom Genotyping Arrays. Experimental conditions and genomic complexities that can negate the lower cost of the newer technologies are often advertised as a reason to switch from microarray technology.

<table>
<tr><th>Attribute</th><th>RADseq</th><th>GBS</th><th>Axiom Genotyping Arrays</th></tr>
<tr><td>DNA quality</td><td colspan="2">• Strict requirements for quality and quantity (200 ng–3,000 ng)
• RADseq library production requires several micrograms of purified, high molecular weight genomic DNA</td><td>• 100–200 ng of DNA</td></tr>
<tr><td>Marker selection</td><td colspan="2">• Markers located at or near restriction sites
• No ascertainment bias</td><td>• Capability to target specific chromosomes
• Ascertainment bias that can be reduced by resequencing of diverse lines</td></tr>
<tr><td>Application range</td><td colspan="2">• Inability to target specific chromosomes
• No control in variability associated with digestion, ligation, and amplification
• No control over digestion sites in samples (mutations, variability in genomes)
• No control over read depth for different parts of the region (RE sensitive to methylation)
• Missing data is impossible to avoid
• Capability to genotype 96 samples in species where a reference genome may not be available</td><td>• Broad applications based on SNPs on array
• NGS validation
• GWAS/marker trait association
• Genomic selection
• Routine-use applications
• Cannot be used for SNP discovery</td></tr>
<tr><td>Genomic complexities</td><td colspan="2">• Increasing genetic divergence may impact restriction sites and increase the frequency of missing data
• Large genomes require more sequencing
• Repeats confound assemblies</td><td>• No impact from restriction sites
• Capability to interrogate markers from large genomes or genomes with high heterozygosity</td></tr>
<tr><td>Polyploid capabilities</td><td colspan="2">• Polyploidy, high heterozygosity, and CNV—need high depth of seq.
• Homozygous 6x, heterozygous 13.5x ($P \geq 0.9975$)
• Assemblies may be chimeric
• Inability to detect CNV losses</td><td>• Capability to analyze allopolyploid data (from tetraploid to octoploid) in less than 1 hour</td></tr>
<tr><td>Number of SNPs</td><td>1,000–150,000; limited to positions around restriction sites</td><td>1,000–50,000; limited to positions around restriction sites</td><td>Multiple SNP tiers; 1,500–2.5 million SNPs</td></tr>
<tr><td>Turnaround time</td><td colspan="2">2–20 weeks (20–24 weeks advertised by service providers)</td><td>4–5 days including data analysis</td></tr>
<tr><td>Analysis</td><td colspan="2">• Missing data confounds imputation
• Complicated informatics</td><td>• Five-step analysis with Microsoft™ Windows™ platform–based Applied Biosystems™ Axiom™ Analysis Suite Software or command line–based Applied Biosystems™ Array Power Tools</td></tr>
</table>

applied
biosystems

In view of the challenges in genotyping-by-sequencing techniques from data management, computational requirements, and a custom informatics pipeline that needs to be tailored to every species and sample population, arrays by far remain the easiest technology for data quality, completeness, analysis, and application in routine breeding.

In summary, the Axiom Genotyping Solution for animal and plant genotyping offers the ability to customize genotyping content onto arrays for commercially important species. The Axiom Genotyping Solution includes species-specific and customized arrays with verified genomic content from the Applied Biosystems™ Axiom™ Genomic Database as well as complete reagent kits, data analysis tools, and a fully automated workflow utilizing the Applied Biosystems™ GeneTitan™ Multi-Channel (MC) Instrument.

References

1. Baird N. A., et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376 (2008). doi:10.1371/journal.pone.0003376
2. Elshire R. J., et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379 (2011). doi:10.1371/journal.pone.0019379
3. Narum S. R., et al. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology 22: 2841–2847 (2013). doi:10.1111/mec.12350
4. Peterson B. K., et al. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7(5):e37135 (2012). doi:10.1371/journal.pone.0037135
5. Kranis A., et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics 14:59 (2013). doi:10.1186/1471-2164-14-59
6. Houston R. D., et al. Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). BMC Genomics 15:90 (2014). doi:10.1186/1471- 2164-15-90
7. Affymetrix application note Mitigating sequencing errors, monomorphs, and poor performing markers during de novo SNP selection for genotyping applications (2013) P/N DNA02261 Rev. 1
8. Ersoz E. S., Yu J., Buckler E. S. Applications of linkage disequilibrium and association mapping in crop plants, in Genomics-Assisted Crop Improvement: vol 1: Genomics Approaches and Platforms, eds. Varshney R. K., Tuberosa R. Springer, pp. 97-119 (2007). doi:10.1007/978-1-4020-6295-7_5
9. Poland J. A., et al. Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5(3):92–102 (2012). doi:10.3835/plantgenome2012.05.0005
10. Meuwissen T., et al. Accelerating improvement of livestock with genomic selection. Annual Review of Animal Biosciences 1:221-237 (2013). doi:10.1146/annurev- animal-031412–103705
11. Hui L., et al. An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. BMC Genomics 15:104 (2014). doi:10.1186/1471-2164-15-104
12. Liu S., et al. Development of the catfish 250K SNP array for genome-wide association studies. BMC Research Notes 7:135 (2014). doi:10.1186/1756-0500-7-135
13. Cavanagh C. R., et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. PNAS 110(20):8057– 8062. doi:10.1073/pnas.1217133110
14. Wendl M. C., et al. Aspects of coverage in medical DNA sequencing. BMC Bioinformatics 9:239 (2008). doi: 10.1186/1471-2105-9-239
15. Hyma K. E. GBS Usage Cases: Non-model Organisms. (2013). http://cbsu.tc.cornell.edu/lab/doc/GBS_nonmodel_Sept_2013.pdf
16. Davey J. W., et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12(7):499–510 (2011). doi:10.1038/nrg3012
17. Mascher M., et al. Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. PLoS ONE 8(10):e76925 (2013). doi:10.1371/journal.pone.0076925
18. Beissinger T. M., et al. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193(4):1073–1081 (2013). doi:10.1534/genetics.112.147710
19. Huang X., et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics 42(11):961–967 (2010). doi:10.1038/ng.695
20. Rexroad C. E., et al. Estimates of linkage disequilibrium and effective population size in rainbow trout. BMC Genetics 10:83 (2009). doi:10.1186/1471-2156-10-83
21. Donato M.D., et. al. Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next- generation sequencing. PLoS ONE 8(5): e62137. doi:10.1371/journal.pone.0062137.
22. Li, X., et al. Applied genetics and genomics in alfalfa breeding. Agronomy 2:40-61 (2012). doi:10.3390/agronomy2010040

To find out more, visit **thermofisher.com/agrigenomics**

ThermoFisher
SCIENTIFIC

For Research Use Only. Not for use in diagnostic procedures. © 2020, 2021 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified. Illumina, HiSeq, and MiSeq are trademarks of Illumina, Inc. Microsoft and Windows are trademarks of Microsoft Corporation. **COL34241 0821**