Genotyping-by-Sequencing on Pooled Samples and its Use in Measuring Segregation Bias during the Course of Androgenesis in Barley

Sébastien Bélanger, Patricio Esteves, Isabelle Clermont, Martine Jean, and François Belzile*

Abstract
Estimation of allelic frequencies is often required in breeding but genotyping many individuals at many loci can be expensive. We have developed a genotyping-by-sequencing (GBS) approach for estimating allelic frequencies on pooled samples (Pool-GBS) and used it to examine segregation distortion in doubled haploid (DH) populations of barley (Hordeum vulgare L.). In the first phase, we genotyped each line individually and exploited these data to explore a strategy to call single nucleotide polymorphisms (SNPs) on pooled reads. We measured both the number of SNPs called and the variance of the estimated allelic frequencies at various depths of coverage on a subset of reads containing 5 to 25 million reads. We show that allelic frequencies could be cost-effectively and accurately estimated at a depth of 50 reads per SNP using 15 million reads. This Pool-GBS approach yielded 1984 SNPs whose allelic frequency estimates were highly reproducible (CV = 10.4%) and correlated ($r = 0.9167$) with the “true” frequency derived from analysis of individual lines. In a second phase, we used Pool-GBS to investigate segregation bias throughout androgenesis from microspores to a population of regenerated plants. No strong bias was detected among the microspores resulting from the meiotic divisions, whereas significant biases could be shown to arise during embryo formation and plant regeneration. In summary, this methodology provides an approach to estimate allelic frequencies more efficiently and on materials that are unsuitable for individual analysis. In addition, it allowed us to shed light on the process of androgenesis in barley.

Doubled haploid populations are often used in plant breeding programs of many species to quickly achieve complete homozygosity in a single generation and begin selection. In barley, these attractive genetic materials can be generated via androgenesis (the process leading from immature microspores via embryos to entire plants) in less than 1 yr. One documented drawback of DH lines is that they often exhibit segregation distortion in a significant portion of the genome (Li et al., 2010) and this can significantly affect the breeding outcome by altering the frequency of alleles in the progeny (Zhang et al., 2012). Such biases have usually been observed to be in favor of the alleles of the parent most amenable to in vitro culture (Sayed et al., 2002). Although it has been possible to document segregation distortion in many studies in barley (Manninen 2000; Sayed et al., 2002; Li et al., 2010), it remains totally unknown when and how these biases arise.

To measure segregation distortion over the course of androgenesis, the allelic frequency would need to be measured on biological materials that do not provide enough DNA to be studied individually (e.g., microspores). To bypass this limitation, a possible solution consists of genotyping and directly estimating the allelic frequency within the population. In this case, the development of a genotyping approach that is capable of measuring allelic frequencies at a large

Département de Phytologie and Institut de Biologie Intégrative et des Systèmes, Université Laval, Quebec City, QC, Canada. Received 28 Oct. 2014. Accepted 10 Nov. 2015. *Corresponding author (francois.belzile@fsaa.ulaval.ca).

Abbreviations: DH, doubled haploid; GBS, genotyping-by-sequencing; MAF, minor allele frequency; Pool-GBS, genotyping-by-sequencing using pooled samples; SDR, segregation distortion region; SNP, single nucleotide polymorphism; UNEAK, Universal Network-Enabled Analysis Kit.
number of molecular markers and on a composite DNA sample obtained from a representative population of microspores, embryos, or plants would be essential.

Recently, analyses of pooled samples have been investigated in various next-generation sequencing methods such as genome resequencing (Turner et al., 2011; Zhu et al., 2012; Rollstab et al., 2013) or complexity reduction approaches such as reduced representation libraries (Van Tassell et al., 2008), restriction site-associated DNA sequencing (Gautier et al., 2013), and GBS (Byrne et al., 2013). For example, Rollstab et al. (2013) investigated a pooled whole-genome resequencing (Pool-Seq) approach on a nonmodel species, Arabidopsis halleri (L.) O’Kane & Al-Shehbaz, exploiting the reference genome of a related model species [Arabidopsis thaliana (L.) Heynh.]. This resequencing approach proved a powerful tool to measure allelic frequencies at over 2 million SNPs in three pools of 20 individuals. Validation of these estimates was conducted on a limited scale (nine SNP loci) and indicated that the estimates derived from Pool-Seq were highly accurate as long as the read coverage exceeded 20×.

Complexity reduction approaches represent an attractive method to produce a large catalog of SNPs when genome size is large and not conducive to complete resequencing, as is the case for most cereals. A pooled GBS analysis was used for the first time by Byrne et al. (2013) in perennial ryegrass (Lolium perenne L.). These authors showed that it was possible to obtain consistent allelic frequency estimates at the population level as long as the pooled DNA sample contained equal contributions of DNA from the individuals composing the population. Allelic frequencies were successfully estimated based on the read coverage for each allele at individual SNPs and it was shown that these frequency estimates were highly correlated in replicated pools. It was not shown, however, if these estimates were in close agreement with the “true” allelic frequencies estimated via the analysis of individuals.

In barley, GBS has become a highly used genotyping tool to characterize alleles at thousands of SNPs using a two-enzyme protocol (Poland et al., 2012). Considering that a complete reference genome for barley is still in development (International Barley Genome Sequencing Consortium, 2012), most of the GBS work done to date has relied on the use of analytical pipelines capable of calling SNPs in the absence of a reference genome such as the Universal Network-Enabled Analysis Kit (UNeAK) (Lu et al., 2013). In essence, UNeAK identifies pairs of sequence tags (called “tag pairs”) that differ at a single base position. In principle, these sequence tags represent the two alleles at a SNP locus. Since UNeAK tallies the read count of each tag, this allows the estimation of the allelic frequency at the population level by the number of reads for each allele.

The major goal of this study was to identify when segregation distortion arises over the course of androgenesis. To perform these analyses, we first developed an efficient and simple GBS methodology to measure allelic frequencies on composite DNA samples (Pool-GBS). Using a set of 76 DH barley lines, we established an analytical strategy aiming to maximize the number of SNPs called and to minimize both the variability of the estimated allelic frequencies and the cost of the analysis. The resulting estimates were highly correlated with the true allelic frequencies but only when care was taken to ensure the uniform contribution of each member of the population to the pooled sample. We then used this method to accurately estimate allelic frequencies throughout androgenesis. We found that biases were absent at the microspore stage and arose strictly during the in vitro phase of the process.

Materials and Methods

Plant Material

Doubled haploid lines were produced by isolated microspore culture (Esteves and Belzile, 2014; Esteves et al., 2014) from two crosses involving barley cultivars HV779STP07 × Oceanik (#1126) and Myriam × UL163 (#1114). Among the progeny of Cross #1126, we randomly selected three independent and partially overlapping sets of 76 DH plants (Fig. 1a). For the first set of lines, a 3-cm segment from the tip of the first leaf was collected from each plant individually. For the second set of DH lines, we collected a 3-cm segment from the tip of the first leaf of each plant to form a single composite sample of leaf tissue (the “uniform” pool). For the third set of DH lines, we collected leaf segments of unequal size on plants at various developmental stages with limited effort put toward ensuring a uniform contribution of each plant to the composite sample (the “nonuniform” pool). The second cross (#1114) was used to investigate segregation bias during androgenesis. We collected four samples (Fig. 1b) corresponding to a population of microspores (i) before and (ii) after a stress treatment meant to induce embryogenesis, (iii) developing embryos, and (iv) regenerated plants. The first three were collected as a single composite sample containing 1 × 10⁶ microspores or 100 embryos, whereas leaf segments from regenerated plants (green or albino) were collected and analyzed for each plant individually.

Genomic DNA Isolation, GBS Library Construction, and Sequencing

All genomic DNA isolation was performed using the DNeasy Plant Kit (Qiagen Sciences, Germantown, MD) as per the manufacturer’s instructions. For the three sets of DH lines from Cross #1126, genomic DNA was isolated from ~50 mg of fresh leaves of individual DH lines or from ~50 mg of lyophilized ground leaves of the composite samples (one uniform and one nonuniform pool). For the four samples collected during androgenesis, DNA was isolated from pools of ~1 × 10⁶ microspores before and after stress treatment, from 100 freshly ground embryos, and from ~50 mg of fresh leaves of individual green or albino plants. For all samples, DNA was quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE).
Two different types of GBS libraries were prepared using the two-enzyme protocol (PstI–MspI) described by Poland et al. (2012): one to genotype each line individually and a second to genotype an entire population as a pool. In the first case, GBS analysis of individual lines was performed for each DH line, the F1 plant, and the parental lines of Cross #1126 (79-plex) and on 94 freshly regenerated DH plants and the two parents of Cross #1114 (96-plex). For Pool-GBS, libraries were prepared using 200 ng of genomic DNA from five composite DNA samples originating from the two crosses (identified in parentheses): (i) the uniform pool of leaf tissue (#1126), (ii) the nonuniform pool of leaf tissue (#1126), (iii) untreated microspores (#1114), (iv) stress-treated microspores (#1114), and (v) embryos (#1114). Single-end sequencing was performed on an Illumina HiSeq2000 (Illumina, San Diego, CA) at the McGill University-Genome Quebec Innovation Center (Montreal, Canada). Depending on the aim of the experiment, libraries were sequenced to a depth varying between 13.5M and 216M reads.

**Single Nucleotide Polymorphism Discovery and Genotype Calling**

Single nucleotide polymorphism discovery and genotype calling were performed using the UNEAK pipeline (Lu et al., 2013) either on individual DH lines or on artificial and true pools for Cross #1126. In an exploratory phase, an artificial pool was obtained by simply combining the reads from all 76 DH lines from Cross #1126 to form a single dataset (Fig. 2a). The resulting set of 125M reads was then split into five equal sets of 25 million reads to be used as replicates (Fig. 2b). To explore the impact of sequencing depth in Pool-GBS, we extracted a subset of increasing size (5–25 million; Fig. 2c) from each set of 25 million reads and used each subset to call SNPs. The ensuing analyses were conducted in five replicates, each provided by one of the original sets of 25 million reads. For the analysis of true pools (uniform and nonuniform), we used a total of 15 million reads. For the study of segregation bias during the course of androgenesis, SNP discovery and genotype calling was performed both on pooled libraries from microspores and embryos (15 million reads each) and 94 individual regenerated plants. We ran the UNEAK pipeline simultaneously on all of these different datasets to ensure that the naming of SNP loci (“tag pairs” in UNEAK terminology) was the same throughout all the analyses. Next, UNEAK was run using the default parameters except for the minimal minor allele frequency (MAF) and the minimal tag count, which were both set to zero rather than their default values (0.05 and 5 respectively). Furthermore, the source code of UNEAK was modified to suppress the read count limit originally set at 127 reads per allele at each SNP locus.

**Single Nucleotide Polymorphism Filtering**

When analyzing individual lines, we kept SNPs that (i) were homozygous for complementary alleles in the parental lines, (ii) had no more than 10% missing data, and (iii) had a MAF of \( \geq 0.05 \). For the artificial pooling analysis, we kept only those SNPs that (i) had data in all pools, (ii) were homozygous for complementary alleles in the parental lines, (iii) had at least 10 reads, (iv) had a MAF of \( \geq 0.05 \). For the analysis of true pools, only SNP loci with a minimum read depth of 50 reads per SNP and a MAF of \( \geq 0.5 \) were retained.
Genetic Mapping of SNPs, Estimation of Allelic Frequencies, and Segregation Bias Analysis

The genetic map position of SNPs was determined by performing a basic local alignment search tool search using the first sequence of each tag pair as query against the barley (cultivar Morex) genome assembly using BarleyMap (Cantalapiedra et al., 2015). Based on the parental genotypes, alleles were coded as originating from a specific parent. Allelic frequencies were estimated on individual markers or averaged across all markers located within successive 5-cM intervals on the genetic map. The frequency of each allele was estimated either based on the genotypes called for the individual lines or on the read counts for each allele in the various pools. For each set of SNPs, χ² tests were performed to assess segregation distortion and, given the large number of tests, the resulting p-values were transformed into q-values using an R script (qvalue); segregation bias was declared to be significant at \( q \leq 0.05 \).

Impact of Pool Size on SNP Depth of Coverage and on the Variability of Estimated Allelic Frequencies

For the artificial pool resulting from combining data for individual DH lines (Cross #1126), the number of SNPs with a mean read count (across the five replicates) of at least 10, 50, 100, or 200 was determined. Only SNPs that were present in all five replicates were counted. For each set of SNPs meeting the required minimum read depth, we calculated the mean and SD for the frequency of alleles inherited from one parent (HV779STP07). The CV for the
estimated allelic frequencies across the five replicates was then computed to assess the variability of these estimates.

**Comparison of Allelic Frequencies Estimated Individually versus in Pools and between Pool Replicates**

The allelic frequencies estimated either through the analysis of 76 individual DH lines or five artificial pools of 5 to 25 million reads were compared at four levels of read depth (10, 50, 100, and 200) by computing Pearson’s correlation.

**Results**

**Optimized Strategy for Estimating Allelic Frequencies Based on the Number of Reads**

To enable comparisons and validation, allelic frequencies were first measured by performing GBS on the two parental lines, the F1 plant (Cross #1126), and 76 individual DH lines. The sequencing of the resulting 79-plex *PstI/MspI* GBS library yielded a total of 130 million reads (an average of ~1.7 million reads per line). A set of 3815 SNPs were segregating in this population and 2802 of these had previously been anchored on the map of the barley genome. In this case, the allelic frequency at a SNP locus was equivalent to the proportion of DH lines that were homozygous for one or the other allele at this locus. This frequency, derived from the genotypes of individual lines, was taken to reflect the true allelic frequency and was the gold standard against which we compared the allelic frequencies estimated using Pool-GBS analysis.

To simulate a Pool-GBS analysis, we began by combining the reads from all 76 DH lines into one artificial pool and 125 million of these reads were divided into five sets of precisely 25 million reads meant to serve as replicates (Fig. 2). To determine the optimal depth of sequencing for a pool, we first examined the impact of the number of reads on the number of SNPs called at four depths of coverage (10, 50, 100, and 200 total reads per SNP). We observed that increasing the number of reads from 5 to 25 million led to a steady increase in the number of SNPs called at all depths, except at 10 reads per SNP, where it stayed high and constant at around 2300 SNPs (Fig. 3). At 50 reads per SNP, a similarly high number of SNPs was reached only when sequencing 25 million reads. At the deeper depths of coverage (100 or 200 reads/SNP), a large decrease in the number of SNPs called was observed at all depths of sequencing, with the number of SNPs called at a coverage of 200 reads being only half that called at 10 or 50 reads/SNP at a depth of sequencing of 25M reads. Interestingly, the efficiency of SNP calling (number of SNPs called per million reads), seen as the slope, decreased as the depth of sequencing increased. This indicates that there was a decreasing return, measured as the number of additional SNPs that can be called, for every additional million reads that were generated and analyzed.

Second, we measured the variability of the estimated allelic frequencies obtained. As expected, increasing the number of reads led to a decrease in the variability of the estimated allelic frequencies (Fig. 4). Except for the lowest coverage (10 reads per SNP), the estimated allelic frequencies were relatively stable from one replicate to the next, as the coefficients of variation were close to or below 10% under most conditions. Such variability in the estimates of allelic frequencies could also be expressed in the form of a correlation between the five different replicates. At coverages above 50 reads per SNP, correlations were quite good, ranging between 0.91 and 0.98. As expected, there was a trade-off between the number of SNPs called and the precision of the estimated allelic frequencies. We feel that a depth of sequencing of 15 million reads provides a good compromise between these two, as close to 2000 SNPs could be called with a low CV and a high correlation between replicates (CV = 10.44%; r = 0.935).

**Accuracy of Estimated Allelic Frequencies**

To determine the accuracy of the estimated allelic frequencies obtained via Pool-GBS analysis, we compared these estimates to the values obtained in the individual analysis at four levels of read depth by computing Pearson’s correlation for SNPs that were shared in the two analytical approaches. Given the different parameters that were used to call and filter SNPs on individual lines and pools, not all SNPs scored in one analysis were scored in the other (typically ~95% of SNPs were shared). We observed two general patterns. First, for a given depth of sequencing (horizontally in Fig. 5), correlations generally increased as the depth of coverage required to call a SNP increased. However, this improvement in accuracy came at the cost of drastic reductions in the number of SNPs called. For example, at 15 million reads, the number of shared SNPs went from 2176 (at ≥10 reads per SNP) down to only 711 SNPs (at ≥200 reads per SNP), whereas the correlation remained quite high in all cases, ranging between 0.915 and 0.926. Second, for a given depth of coverage required to call a SNP (vertically in Fig. 5), no clear trend was observed in the strength of the correlations, as these were relatively constant across all depths of sequencing (5–25 million reads). At all depths of coverage other than ≥10 reads per SNP, increasing the depth of sequencing led to important increases in the number of SNPs. For example, at ≥50 reads per SNP, only 957 SNPs were called with 5 million reads; this number increased to around 1900 SNPs at 15 million reads. These results indicate that the compromise identified earlier (15 million reads, ≥50 reads per SNP) allows the calling of a large number of shared SNPs (1894) for which the estimated allelic frequencies were very highly correlated (r = 0.9167) with the true frequency estimated by analyzing individual DH lines.
Validation of the Pool-GBS Approach

We validated the capacity of the Pool-GBS approach to correctly identify chromosomal regions exhibiting a significant segregation bias in the population of DH lines. In all cases, the frequency derived from the genotypes of individual lines was taken to reflect the true allelic frequency and was compared to the frequency estimated using the abundance of reads corresponding to each allele. Overall, a total of five segregation distortion regions (SDRs), one each on chromosomes 2H, 3H, 4H, 5H, and 7H, were detected in this population of DH lines (Supplementary Fig. S1). The location and size of the affected region, the magnitude of distortion, and the parent of origin of the favored allele are shown in Table 1. As shown in Fig. 6a, using individual analysis, a very strong segregation bias was observed on chromosome 2H, reaching a peak (at ~50 cM) where 95% of the alleles were inherited from one parent (Oceanik). When an artificial pool composed of a portion of the same reads (15 million of 125 million reads) was analyzed using the Pool-GBS approach (15 M reads, ≥50 reads per SNP), a very similar picture emerged (Fig. 6b), albeit with a few clear outliers. To reduce such "noise", we computed a mean frequency across all SNP markers located in successive 5-cM bins. As seen in Fig. 6c, the allelic frequencies estimated with the artificial pool closely mirrored the pattern of allelic frequencies inferred from the individual lines all along the chromosome.

Finally, to test the robustness of the Pool-GBS analysis, an independent and partially overlapping set of DH lines was used to obtain a uniform pool of leaf material from which DNA was extracted and a single GBS library prepared. A total of 19.3 million reads was obtained from this true pool and a set of exactly 15 million reads was analyzed using a minimal coverage of 50 reads per SNP. This yielded a total of 2102 SNPs. At this depth of sequencing and the filtering described above, the SNPs retained were supported by an average of 243 reads (median = 237 reads). As shown in Fig.
6d, the mean allelic frequency (in 5-cM bins, as in Fig. 6c) faithfully reproduced the significant segregation bias in the “left half” of chromosome 2H. Highly similar segregation biases were also observed for the other SDRs using Pool-GBS analysis on this true pool (Supplementary Fig. S1). Furthermore, within SDRs, the frequency of the favored allele estimated using this true pool ranged between 0.72 and 0.93, and it differed by no more than 0.04 relative to the frequency inferred from individual analysis (Table 1). In summary, we conclude that the Pool-GBS approach allowed us to estimate the allelic frequencies within a DH population correctly using a much smaller number of reads (15 million) than was required to achieve the same result when analyzing individual lines (125 million).

Importance of Uniform Pools
To investigate the importance of the uniformity in composition of the composite samples, we compared the true pool described above that had been created by methodically collecting a single 3-cm segment from the first leaf of each line to produce a “uniform” pool to a second composite derived from “nonuniform” sampling achieved by collecting leaf segments from adult plants in the greenhouse without consideration for the developmental stage and size of these segments. Overall, the Pool-GBS analysis performed with both the uniform and nonuniform pools gave similar estimates of allelic frequencies. However, the nonuniform pool displayed a lower sensitivity for detecting segregation bias. Indeed, although the strongest biases could still be observed using the nonuniform pool (on 2H and 5H), lower accuracy in allelic frequency estimates led to a failure to detect significant biases on chromosomes 3H, 4H, and 7H (Fig. 7). This suggests that to maximize sensitivity and accuracy, care must be taken to ensure that the different individuals making up a pool contribute equally to the composite DNA sample.
Segregation Biases Arising during Androgenesis in Barley

To determine for the very first time when segregation distortion arises during barley androgenesis, we estimated allelic frequencies over the course of this process. Allelic frequencies were estimated using the Pool-GBS approach on microspores (both before and after stress treatment) as well as on embryos of Cross #1114. In addition, segregation distortion was also examined in regenerated plants of the same cross via a conventional GBS analysis conducted on 94 individual regenerated DH plants and the two parental lines. In the Pool-GBS analyses, 13.5 million to 22.8 million reads were obtained per library and SNPs were called at a minimal depth of 50 reads per SNP (Table 2). In contrast, the 94 regenerated plants and both parents were analyzed individually using a total of 217.2 million reads (~2.3M reads per line). The number of SNPs found to be segregating was remarkably stable from pool to pool and between the Pool-GBS and individual plant analysis, with this number ranging between 1117 and 1132 (Table 2). It is interesting to note that despite the much smaller number of reads used to call segregating SNPs in the three pooled analyses (13.5–22.8 million reads) compared to individual analysis (217 million reads), the same amount of useful information could be derived.

These SNP data were used to identify chromosomal regions exhibiting a significant segregation bias in the populations of microspores, embryos, and regenerated plants. As was done earlier, to minimize the impact of outliers, we calculated the mean allelic frequency across markers in successive 5-cM bins along the barley chromosomes. We first observed that no significant biases were observed in untreated microspores and that only 4 of the 163 tested bins exhibited a significant distortion among treated microspores. These SDRs were located on four chromosomes (1H, 5H, 6H, and 7H) but encompassed less than 2% of the barley genetic map (20 cM of 990.1 cM). The magnitude of the distortion was typically limited, with the frequency of the favored
Table 1. Segregation distortion observed in a barley doubled haploid population derived from a cross between HV779STP07 and Oceanik.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Individual lines</th>
<th>Real uniform pool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak (cM)</td>
<td>Range (Start–end)</td>
</tr>
<tr>
<td>2H</td>
<td>50.5</td>
<td>7–94</td>
</tr>
<tr>
<td>3H</td>
<td>131.5</td>
<td>108–155</td>
</tr>
<tr>
<td>4H</td>
<td>47.0</td>
<td>0–94</td>
</tr>
<tr>
<td>5H</td>
<td>40.5</td>
<td>31–50</td>
</tr>
<tr>
<td>7H</td>
<td>89.5</td>
<td>42–137</td>
</tr>
</tbody>
</table>

† Frequency of the favored allele at the single nucleotide polymorphism loci showing the greatest degree of segregation distortion.
‡ Favored parental allele at the single nucleotide polymorphism loci showing the greatest degree of segregation distortion.
§ O, Oceanik; H, HV779STP07.

Fig. 6. Segregation distortion on chromosome 2H measured using four different approaches. Allelic frequencies were estimated (a) individually using a total of 125 million reads from 76 doubled haploid (DH) barley lines, (b) on an artificial pool of 15 million reads, (c) on an artificial pool of 15 million reads, and (d) on a uniformly sampled pool of 15 million reads. The x-axis shows the position (in cM) of markers or marker bins along the chromosome and the y-axis shows the estimated allelic frequency. Each symbol indicates the allelic frequency of a parental allele (black circle, HV779STP07; red circle, Oceanik) at each single nucleotide polymorphism (SNP) (a and b) or averaged over all SNPs located within a 5-cM bin (c and d). The dashed blue horizontal lines indicate the critical allelic frequencies above or below which segregation was declared to be significant.
allele ranging from 0.64 to 0.72. We conclude that no strong bias was observed in microspores either before or after the stress treatment used to induce embryogenesis (Table 3; Supplementary Fig. 2). Thus neither the meiotic divisions nor the stress treatment induced a differential transmission of alleles on a genome-wide level.

In stark contrast, however, very strong biases were observed in the embryos and regenerated plants. As detailed in Table 3, five and seven regions of significant distortion were observed in embryos and plants, respectively. The size of these SDRs ranged from as little as 14 cM to as large as 77 cM. Similarly, the frequency of the predominant allele varied between 0.65 and 0.95, with the allele from Myriam being favored over the allele from UL163 in most cases. Interestingly, there was a high degree of overlap between the SDRs identified at the embryo stage and those identified at the plant stage. Indeed, four of the five SDRs observed at the embryo stage overlapped with SDRs identified in the regenerated plants. In addition, these SDRs were similar in both their size and the magnitude of the allelic bias (Table 3). There were, nonetheless, a few instances where SDRs differed markedly between the embryo stage and the regenerated plant stage. As shown in Fig. 8 (top panel), only a very modest distortion was observed on chromosome 6H in embryos, whereas a much stronger and extensive area of distortion was seen in plants. Inversely, on chromosome 1H, the magnitude and extent of segregation bias was much more pronounced in embryos than in plants (Fig. 8, bottom panel).

Fig. 7. Impact of sample uniformity for pooled genotype-by-sequencing (Pool-GBS) analysis. These graphs show the estimated allelic frequencies on chromosome 7H in true pools of 15 million reads resulting from the sampling of doubled haploid barley lines in a nonuniform (left) or uniform (right) fashion. The x-axis shows the position (in cM) of marker bins along the chromosome and the y-axis shows the estimated allelic frequency. Each symbol indicates the allelic frequency of a parental allele (black circle, HV779STP07; red circle, Oceanik) averaged over all single nucleotide polymorphisms located within a 5-cM bin. The dashed blue horizontal lines indicate the critical allelic frequencies above or below which segregation was declared to be significant.

Table 2. Number of reads and single nucleotide polymorphisms (SNPs) with their associated mean coverage for barley samples investigated during the androgenesis process of the cross between Myriam and UL163.

<table>
<thead>
<tr>
<th>Samples</th>
<th>Reads</th>
<th>SNPs</th>
<th>Mean coverage (reads per SNP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microspores before treatment</td>
<td>13,919,949</td>
<td>1132</td>
<td>104</td>
</tr>
<tr>
<td>Microspores after treatment</td>
<td>13,553,524</td>
<td>1117</td>
<td>57</td>
</tr>
<tr>
<td>Embryos</td>
<td>22,849,206</td>
<td>1127</td>
<td>162</td>
</tr>
<tr>
<td>Regenerated plants</td>
<td>217,200,947</td>
<td>1130</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 3. Segregation distortions observed in barley during androgenesis via isolated microspore culture derived from a cross between Myriam and UL163.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Chromosome</th>
<th>Peak (cM)</th>
<th>Range (Start–end)</th>
<th>Frequency</th>
<th>Parental line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryos</td>
<td>1H</td>
<td>39.5</td>
<td>8–71</td>
<td>0.84†</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>3H</td>
<td>47.5</td>
<td>87–112</td>
<td>0.71</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>4H</td>
<td>60.0</td>
<td>53–67</td>
<td>0.68</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>7H</td>
<td>52.5</td>
<td>43–62</td>
<td>0.70</td>
<td>Myriam</td>
</tr>
<tr>
<td>Regenerated plants</td>
<td>1H</td>
<td>44.5</td>
<td>36–53</td>
<td>0.65</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>2H</td>
<td>56.5</td>
<td>46–67</td>
<td>0.67</td>
<td>UL163</td>
</tr>
<tr>
<td></td>
<td>3H</td>
<td>37.5</td>
<td>2–73</td>
<td>0.95</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>4H</td>
<td>57.0</td>
<td>47–67</td>
<td>0.71</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>5H</td>
<td>148.5</td>
<td>128–169</td>
<td>0.73</td>
<td>UL163</td>
</tr>
<tr>
<td></td>
<td>6H</td>
<td>103.5</td>
<td>88–119</td>
<td>0.71</td>
<td>Myriam</td>
</tr>
<tr>
<td></td>
<td>7H</td>
<td>59.0</td>
<td>21–97</td>
<td>0.77</td>
<td>Myriam</td>
</tr>
</tbody>
</table>

† Frequency of the favored parental allele at the single nucleotide polymorphism loci showing the greatest degree of segregation distortion.
Discussion

A Simple and Efficient GBS Methodology to Estimate Allelic Frequencies in Pooled Samples

Allelic frequencies were accurately estimated using Pool-GBS analysis with as few as 15 million reads obtained from a segregating population of 76 barley DH lines. At a minimum depth of coverage of 50 reads per SNP, this analytical approach allowed us to estimate allelic frequencies at a large number of SNPs (1904) in a highly reproducible fashion ($CV = 10.4\%$ and $r = 0.935$ in five replicates). Furthermore, these frequencies estimated on the basis of read abundance were highly correlated ($r = 0.9167$) with the true frequency estimated by analyzing the DH lines individually and calculating the allelic frequency based on actual genotypes. This methodology was further validated by performing the analysis of the same population of DH lines using a composite DNA sample. Chromosomal regions exhibiting significant segregation distortion were also correctly identified. Compared to the analysis of individual lines, almost 10-fold fewer reads (15 million versus 125 million) were required to achieve the same or very similar result.

To date, in eukaryotes, most efforts to estimate allelic frequencies in pooled samples have focused on pooled whole-genome resequencing (Pool-Seq). For example, Rellstab et al. (2013) performed Pool-Seq on three populations of 20 individuals of *A. halleri* to estimate...
allelic frequencies and performed individual validation at nine loci. At a mean coverage of 104x, the estimated major allele frequencies in pools of 20 individuals differed by less than 0.04 from the frequencies measured by sequencing individuals. Similarly, in the present Pool-GBS analysis, a mean coverage of 243x was obtained (at a coverage threshold of ≥50 reads per SNP) and the differences in allelic frequencies measured individually or in pools ranged between 0.0 and 0.04 at the peak SNPs in the five genomic regions showing segregation distortion in Cross #1126. Rellstab et al. (2013) also explored the impact of the coverage depth on the accuracy of the estimated allelic frequencies. They reported that above 20x coverage, the coefficient of determination between the allelic frequencies estimated in pools and the true allelic frequency (assessed by individual sequencing) was >0.95 and that at 50x coverage, the R² values were in excess of 0.99. At the same coverage (50x), we obtained a lower R² value (0.84) using a pool of 15 million reads. However, this value was derived from the comparison of estimated and true allelic frequencies at a much larger number of loci than the nine loci used in Rellstab et al. (2013) and may thus better reflect the true degree of correlation between these data.

Although Pool-Seq represents a powerful approach, it would prove quite costly to perform in barley, whose genome is 20-fold larger than the A. halleri genome (5.1 Gb versus 255 Mb). It is for this reason that genome complexity reduction methods such as GBS have been preferred to perform SNP genotyping in species with large genomes. Byrne et al. (2013) were the first to explore pooled GBS analysis in perennial ryegrass, a highly heterozygous outcrossing species where composite samples were produced by collecting leaf tissue from many individual plants from each of eight varieties. At a depth of coverage of >20x, they found that the estimated allelic frequencies in four replicates were quite reproducible (r = 0.91). Although this demonstrated the reproducibility of the analytical approach used by Byrne et al. (2013), it did not directly test for accuracy relative to the true frequencies measured on individual plants. In our work, in addition to showing that the estimated allelic frequencies were very reproducible between replicates (r = 0.935 at ≥50x coverage), we also demonstrate that the estimated frequencies were highly correlated with the true frequencies (r = 0.92 at ≥50x coverage).

Importance of Uniform Pools

Previous studies have outlined the importance of ensuring a uniform contribution from each individual to produce unbiased estimations of allelic frequencies (Futschik and Schlötterer, 2010; Gautier et al., 2013). In this work, we observed that to properly estimate allelic frequencies and identify chromosomal regions exhibiting segregation distortion, it was necessary to ensure that each plant contributed equally to a composite sample. In the nonuniform pool, although we were able to document segregation distortion on some of the chromosomes (2H and 5H), inaccurate allelic frequency estimates prevented us from detecting the significant bias on chromosome 3H, 4H, and 7H. This emphasizes the need for care in preparing representative composite DNA samples.

Pooled Genotyping is Extremely Cost-effective

In a number of GBS studies, especially in species with large genomes, it is fairly common to aim for ~2 ×10⁶ reads per individual when analyzing 96-plex libraries (Sonah et al., 2013). For populations of the size examined in this work (i.e., 76 barley lines) this would require DNA extraction and library preparation to be performed on all 76 samples and would necessitate ~150 million reads. In contrast, the Pool-GBS analysis as performed here required the preparation of a single DNA sample and GBS library, a 75-fold reduction in all steps leading to library preparation. Also, only 15 million reads were required to provide coverage that was sufficient to estimate allelic frequencies accurately and consistently, a 10-fold reduction in terms of DNA sequencing. Thus, analyzing a single pool of this size represents a 10- to 15-fold lower cost compared to the analysis of individual lines. The advantages of such a pooled approach are obviously the greatest when there is a need to assess numerous populations, such as in the context of a breeding program, where it may be useful to cull DH populations that are severely distorted, or of ecological research where one is mainly interested in comparing allelic frequencies among numerous natural populations.

Segregation Bias Occurs Mainly during the Final Phases of Androgenesis

Although segregation bias is sometimes mentioned briefly when describing various segregating populations (e.g., in the context of gene or quantitative trait locus mapping), it has been the subject of relatively few detailed investigations. It is often reported as being more severe in progeny derived from androgenesis than from selfing in an F2 population (Sayed et al., 2002). Although this is true, an F2 population is not genetically fixed and not fully comparable with a DH population. Xu et al. (1997) have shown that segregation distortion in recombinant inbred populations can be quite similar to that seen in DH populations. In barley, segregation bias has most commonly been studied among DH progeny (Manninen, 2000; Sayed et al., 2002; Li et al., 2010). A limitation of previous studies can be found in the small numbers of DNA markers, which were often insufficient to provide extensive and uniform coverage of the genome and thus limited the capacity to detect significant segregation biases. For instance, Sayed et al. (2002) used only 43 simple sequence repeat markers grouped on four linkage groups with very sizeable gaps between markers. In contrast, Li et al. (2010) characterized SDRs in four DH barley populations with more than 500 markers per population distributed quite evenly on the genetic map. In this work, we demonstrate the feasibility of analyzing such populations as pools and estimating allelic frequencies for a large number of SNP
markers. Extensive genome coverage is thus achieved much more quickly and at a lower cost than is possible through the analysis of individual lines, irrespective of the marker technology used. In this work, we found several chromosomal regions exhibiting a significant distortion in the segregation of SNP markers (on all chromosomes) in DH populations of Cross #1114 (eight SDRs) and Cross #1126 (five SDRs). This is quite comparable in scale to previous studies in which the number of chromosomal regions exhibiting distortion ranged between 4 and 10 per population (Graner et al., 1991; Li et al., 2010). All biased regions identified in our DH populations overlapped or coincided with regions reported previously by Li et al. (2010) except for regions located on 1H (56 cM), 3H (125 cM), and 5H (150 cM).

Although the occurrence of segregation biases following androgenesis has been widely reported, it remained unknown at what stage such distortions arise during androgenesis. The most extensive study to date concerned segregation bias in maize (Zea mays L.), where this was examined in callus, embryos, haploid regenerated plants, and spontaneous diploid plants (Dufour et al., 2001). Our research in barley extends this work to the microspore stage, an innovation made possible thanks to the development of the Pool-GBS approach. Our results demonstrate that no important segregation biases were observed in the populations of microspores. Although four markers displaying a significant distortion were observed on four chromosomes in the population of stress-treated microspores, no association between these markers and the SDRs in embryos was observed. These observations suggest that the stress treatment applied to the microspores did not induce a selection producing segregation biases. We therefore provide strong evidence to support the view that segregation biases arising during androgenesis are largely a product of selective forces acting during the in vitro culture phase of this process.

References


