OPEN ACCESS Freely available online

PLOS ONE

TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline

Jeffrey C. Glaubitz[1]*, Terry M. Casstevens[1], Fei Lu[1], James Harriman[1], Robert J. Elshire[1¤], Qi Sun[2], Edward S. Buckler[1,3]

1 Institute for Genomic Diversity, Cornell University, Ithaca, New York, United States of America, **2** Biotechnology Resource Center Bioinformatics Facility, Cornell University, Ithaca, New York, United States of America, **3** USDA Agricultural Research Service, Ithaca, New York, United States of America

Abstract

Genotyping by sequencing (GBS) is a next generation sequencing based method that takes advantage of reduced representation to enable high throughput genotyping of large numbers of individuals at a large number of SNP markers. The relatively straightforward, robust, and cost-effective GBS protocol is currently being applied in numerous species by a large number of researchers. Herein we describe a bioinformatics pipeline, TASSEL-GBS, designed for the efficient processing of raw GBS sequence data into SNP genotypes. The TASSEL-GBS pipeline successfully fulfills the following key design criteria: (1) Ability to run on the modest computing resources that are typically available to small breeding or ecological research programs, including desktop or laptop machines with only 8–16 GB of RAM, (2) Scalability from small to extremely large studies, where hundreds of thousands or even millions of SNPs can be scored in up to 100,000 individuals (e.g., for large breeding programs or genetic surveys), and (3) Applicability in an accelerated breeding context, requiring rapid turnover from tissue collection to genotypes. Although a reference genome is required, the pipeline can also be run with an unfinished "pseudo-reference" consisting of numerous contigs. We describe the TASSEL-GBS pipeline in detail and benchmark it based upon a large scale, species wide analysis in maize (*Zea mays*), where the average error rate was reduced to 0.0042 through application of population genetic-based SNP filters. Overall, the GBS assay and the TASSEL-GBS pipeline provide robust tools for studying genomic diversity.

Citation: Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, et al. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9(2): e90346. doi:10.1371/journal.pone.0090346

Editor: Nicholas A. Tinker, Agriculture and Agri-Food Canada, Canada

Received November 13, 2013; **Accepted** January 28, 2014; **Published** February 28, 2014

Funding: This work was supported by the National Science Foundation (www.nsf.gov) under the Plant Genome Research Program (PGRP) (grant numbers DBI-0820619 and IOS-1238014) and the Basic Research to Enable Agricultural Development (BREAD) project (ID:IOS-0965342), as well as by the USDA-ARS (www.usda.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jcg233@cornell.edu

¤ Current address: AgResearch Limited, Grasslands Research Centre, Palmerston North, New Zealand

Introduction

The advent of next generation sequencing has elicited a revolution in biology buoyed by an advancing tidal wave of raw sequence data[1–4]. By combining the power of next generation sequencing with reduced representation[5], which focuses sequencing resources on the ends of restriction fragments, it is now possible to quickly genotype unprecedented numbers of samples even in large genome species[6–8]. Several low cost, high throughput methods that combine next generation sequencing with reduced-representation have been developed (e.g., [9–19]). Because of its relative simplicity and robustness, the genotyping by sequencing (GBS) method of Elshire et al. [12] or close derivatives thereof have already been applied in numerous species by many researchers (e.g., [7,17,20–31]). For the first time, generation of copious quantities of genotypic data for genetic experiments is no longer a bottleneck. Instead, the new bottleneck is the efficient bioinformatics analysis of the vast and ever-expanding sea of data. Opportunities to apply markers to breeding or conservation biology are now often limited only by the availability of appropriate bioinformatics tools.

To address this bioinformatics bottleneck, we implemented a GBS analysis pipeline in the Java program TASSEL[32] (version 4) that is specifically tailored to the GBS protocols of Elshire et al. [12] or Poland et al. [20]. However, the TASSEL-GBS pipeline is not limited to the specific restriction enzymes utilized in those protocols: it currently accepts 15 single restriction enzymes and 15 restriction enzyme pairs, and new enzymes are easily added. Furthermore, the TASSEL-GBS pipeline should work on nearly any restriction enzyme and barcoding approach (e.g., [33]), provided that sequence reads commence with the barcode immediately followed by the remnant of the restriction enzyme cut site (Figure 1A). Compared to other available pipelines for similar purposes [16,25,34–38] the TASSEL-GBS pipeline is specifically designed to efficiently handle large quantities of data from large numbers of samples: to date, we have analyzed more than 45,000 maize samples. The TASSEL-GBS pipeline was designed for species with a reference genome; however, it is possible to use incomplete genome assemblies consisting of numerous contigs as a pseudo-reference. For species without a reference genome, an alternative approach, appropriate for small to medium scale studies, has already been implemented in TASSEL [22].

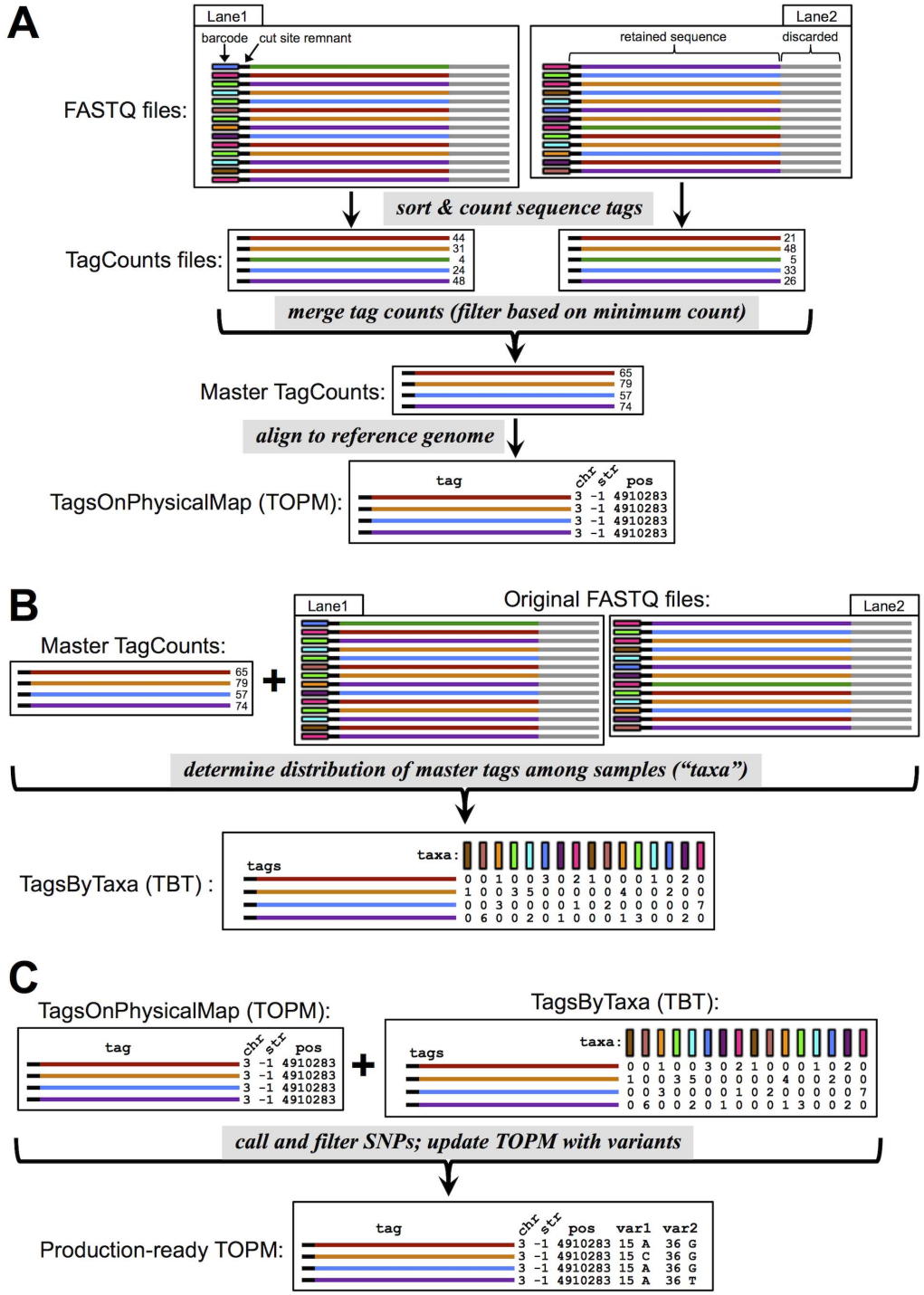

Figure 1. Schematic representation of the TASSEL-GBS Discovery Pipeline. (A) Barcoded sequence reads are processed and collapsed into a set of unique sequence tags, with one TagCounts file produced per input FASTQ file. The separate TagCounts files are then merged to form a ''master'' TagCounts file, which retains only those tags present at or above an experiment-wide minimum count. This master tag list is then aligned to the reference genome and a TagsOnPhysicalMap (TOPM) file is generated, containing the genomic position of each tag with a unique, best alignment. (B) The barcode information in the original FASTQ files is then used to tally the number of times each tag in the master tag list is observed in each sample (''taxon'') and these counts are stored in a TagsByTaxa (TBT) file. (C) The information recorded in the TOPM and TBT is then used to discover SNPs at each ''TagLocus'' (set of tags with the same genomic position) and filter the SNPs based upon the proportion of taxa covered by the TagLocus, minor allele frequency, and inbreeding coefficient (F_{IT}). For each retained SNP, the allele represented by each tag in the corresponding TagLocus is recorded in the TOPM file, along with its relative position in the locus. The end product of the Discovery Pipeline is a ''production-ready'' TOPM that can then be used by the Production Pipeline to call SNPs.
doi:10.1371/journal.pone.0090346.g001

Herein, we describe the TASSEL-GBS pipeline in detail, as implemented in TASSEL 4, and we benchmark the pipeline based upon a large-scale, species-wide analysis of maize (*Zea mays*). There were three primary motivations behind the development of this software: (1) Support the use of GBS in high-throughput, accelerated plant breeding, (2) Accommodate the high genomic diversity that is frequently encountered in species critical to agriculture and conservation, and (3) Provide an analysis platform that can be run in many contexts and with modest computational resources, such as those typically available in the developing world.

Terminology

A read is a single sequence in the FASTQ output file generated by the GBS assay

A ***good, barcoded read*** is a sequence read with a perfect match to one of the barcodes provided in a barcode key file and with no N's in the sequence following the barcode up to the trim length. Under the current implementation, reads are trimmed to 64 bp (not including the barcode).

A ***tag*** refers to a unique sequence (excluding the barcode) up to a specified length (currently 64 bp) from one or more "good, barcoded reads". A given tag is typically observed in numerous good, barcoded reads of identical sequence (up to the trim length).

For our purposes, a ***taxon*** refers to a nameable entity from which one or more DNA samples can be taken.

Design Considerations

Separation of SNP discovery and production SNP calling

Genomics-assisted, accelerated plant breeding usually consists of two separate phases of analysis: a survey of genetic diversity within a species or large breeding program to discover useful markers, followed by usage of these markers to rapidly advance generations. During the advancement cycle, time is of the essence, as thousands of samples need to be processed as quickly as possible so that decisions can be made for the next breeding cycle. The TASSEL-GBS software, by its division into Discovery (Figure 1) and Production pipelines (Figure 2), mirrors the two phases of accelerated plant breeding.

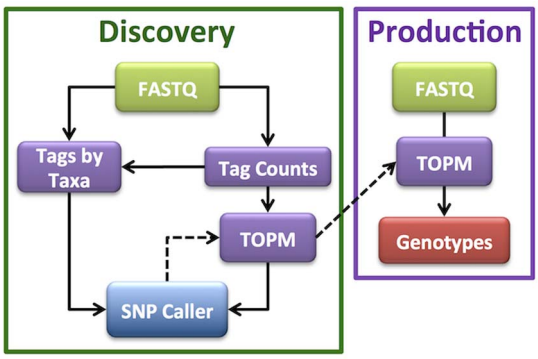

Figure 2. Relationship between the TASSEL-GBS Discovery and Production pipelines. The Discovery Pipeline is run periodically on all FASTQ files generated to date in a species, and the ascertained and filtered SNPs are stored in a "production-ready" TOPM. The Production pipeline utilizes this production-ready TOPM to quickly call SNPs either for the original samples in the Discovery Build, or for subsequent, post-Discovery samples.
doi:10.1371/journal.pone.0090346.g002

The aim of the Discovery Pipeline (Figure 1) is to use the cumulative sequence data from all available samples run to date in a species (or breeding population) to discover SNPs. These SNPs are stored in a "TagsOnPhysicalMap" (TOPM) data structure, containing all of the potentially useful, unique, sequence tags, the genomic positions of the subset of the tags with unique best alignment positions, and the alleles that each useful tag represents for each discovered SNP. The "production-ready" TOPM (populated with variants for each useful tag) can then be in used in the single-step Production Pipeline to quickly produce genotypes, by determining which useful tags are present in each sample. A production-ready TOPM will be applicable to a breeding population as long as the genetic diversity present in the founders of the breeding population is well represented in the individuals that comprised the corresponding Discovery Build.

Our general approach is to periodically perform a comprehensive "Discovery Build" including all samples run to date in our study species. Performing a Discovery Build on a large number of samples is a multistep process that, depending on the number of samples, can require considerable computing resources (or time). For example, the most recent Discovery Build that we performed in maize (AllZeaGBSv2.6), comprising 31,978 samples (plus 758 blank negative controls, where TE buffer was substituted in place of a DNA sample), took 495 CPU-hours on 64 core Linux machine with 512GB of RAM (where each core was a 1.4 GHz AMD Opteron Processor 6272), plus additional time for staging of all of the input FASTQ files, etc. The Production Pipeline, in contrast, provides an avenue by which genotypes for the set of SNPs discovered in the most recent Discovery Build can be quickly generated for new samples. Running the Production Pipeline on a single FASTQ file containing sequence reads from 95 or 383 samples (plus a blank negative control) requires approximately 1 CPU-hour on a MacBook Pro with a 2.6 GHz Intel Core i7 processor and 16GB of RAM running OS X.

Favoring calling a large number of SNPs versus depth per SNP

In a GBS assay, the tradeoff, for a given genome size, between number of SNPs genotyped and the depth of coverage at each SNP is controlled by the level of multiplexing and the choice of restriction enzyme(s) [12,20]. One of our primary motivations for performing GBS (in maize and other organisms) is to enable GWAS, which requires a high density of markers, so that each causative polymorphism stands a reasonable chance of being in LD with one or more markers [39]. Hence, we favor increasing the number of markers at the expense of depth and thus designed the TASSEL-GBS pipeline with low coverage data in mind. The resultant missing data and under-calling of heterozygotes can be compensated for by redundant coverage of haplotypes at high marker density, facilitating imputation.

The likelihood of success of this imputation-based strategy depends on the number and length of homozygous, identical by descent (IBD) segments present in the study population. This, in turn, depends on the demographic history of the population. Imputation is least challenging in biparental populations consisting of RILs [40]. Imputation can also be relatively straightforward in a set of homozygous inbred individuals descending from a limited number of founders (e.g., modern maize lines; [24]). Even in outcrossing species, extensive homozygous IBD stretches can be present if a population bottleneck of sufficient severity occurred at some point in the demographic history of the study population

[41]. Such bottlenecks are common in outcrossing crop species, associated with domestication (e.g., [42]), modern improvement (e.g., [43]), or the development of a breeding population [44]. For unrelated (or distantly related) individuals (either inbred or outcrossed), large sample sizes improve the prospect of successful imputation: the more individuals genotyped, the more likely a given haplotype will be present in multiple individuals in homozygous form. Marker density and depth of coverage per marker need only be high enough to permit recognition of these homozygous IBD segments.

The TASSEL-GBS pipeline is thus optimized for low sequencing depth (0.5 to 3×) over a large number of markers in a large sample of individuals. However, it is flexible enough for the analysis of higher coverage data either from a small genome species, or from a large genome species where a lower level of multiplexing, replicate runs of the same library preps, and/or a less frequently cutting restriction enzyme (or enzyme combination) are used.

Capacity for large numbers of markers and samples

Our analysis strategy favoring large numbers of markers scored at low depth in a large sample of individuals requires the TASSEL-GBS pipeline to be able to handle very large data structures (Table 1). The largest data structure is the "TagsByTaxa" (TBT) object, which records the observed depth in each individual for each potentially useful sequence tag (where "taxa" in "TagsByTaxa" refers to an individual sample from a particular GBS library prep). Our most recent Discovery Build in maize (AllZeaGBSv2.6) comprised 97.5 million potentially useful tags from 31,978 samples (plus 758 blank negative controls), with the depth for each tag in each sample recorded as a single byte (prior to compression). In order to efficiently store and retrieve data from a matrix of this size (3.2 TB of uncompressed, raw data), we used the HDF5 storage format (http://www.hdfgroup.org). We also implemented a rapid and efficient run length compression algorithm to further decrease the storage size. At low depth and with high genetic diversity (numerous sequence tags per locus), the TBT is a sparse data matrix consisting mostly of zeros; our run length compression algorithm takes advantage of this. The sequence tags themselves are stored in a binary format requiring only 16 bytes to hold 64 bases (2 bits per base), prior to compression. Thus, 64 base tags can be conveniently held in two "longs" in Java. The 64 base upper limit on tag lengths in the current implementation of TASSEL-GBS will be lifted in the near future, which will be helpful for study organisms with limited diversity.

Capability to run on modest computing infrastructure

GBS provides an unprecedented opportunity for genomic markers to be used by researchers working in numerous species, including researchers in the developing world. Many of these potential users do not have access to big memory computers or large clusters. The TASSEL-GBS Discovery and Production pipelines can be run on a Linux, Mac or Windows computer with 8–16 GB of RAM. The main demand with respect to the amount of RAM required is the number good, barcoded reads in a typical FASTQ file produced from GBS, which currently ranges from 200–300 million. The small memory footprint required by the TASSEL-GBS pipeline, in relation to its high capacity in terms of number of markers and individual samples, renders it useful to a broad array of users without access to sophisticated computing infrastructure, a target user group which may include breeders carrying out genomic selection experiments.

Avoiding redundant alignment of identical reads

Genotyping approaches based on whole genome, rather than reduced-representation, sequencing typically align all or most of the reads produced to the reference genome prior to calling SNPs (e.g., [36]). In contrast, the TASSEL-GBS pipeline first collapses all of the reads into a master tag list containing all of the sequence tags present at or above a user-specified minimum count, tallied across all of the samples in the Discovery Build (Figure 1A). Each tag in this master tag list is then aligned to the reference genome. This strategy dramatically reduces the computation time devoted to alignment, and permits the use of more computationally expensive alignment algorithms. As the master tag list for our most recent maize Discovery Build (AllZeaGBSv2.6) consisted of 97.5 million tags, distilled from more than 46.8 billion sequence reads, a 392-fold reduction in computational time devoted to alignment was achieved in our case.

Favoring allelic redundancy over quality scores

Quality scores produced by the Illumina base caller are strongly negatively correlated with position in the read. In most whole genome sequencing approaches the position of a particular SNP in different reads is essentially random. In contrast, a GBS SNP has a consistent position in each read, as all of the GBS tags from a particular genomic location (a "TagLocus") that are used to discover and call SNPs originate from the same restriction enzyme cut site and have the same strand orientation. Consequently, the more distal SNPs in a GBS tag tend to have lower quality scores. If quality scores were used to filter GBS reads, these more distal SNPs might end up with lower depth of coverage. Furthermore, the quality scores frequently are not indicative of true quality [45,46]. Hence, rather than using quality scores to filter out bad reads, the TASSEL-GBS pipeline instead relies on the number of times a given tag has been observed as an indicator of sequence quality. GBS sequence tags that occur in a minimum, user-specified, number of reads across all of the samples in a Discovery Build are deemed as potentially useful and are kept for further processing (alignment to the genome and SNP calling). Illumina

Table 1. Size of the key data structures used by the TASSEL-GBS pipeline for a recent maize "Discovery Build" (AllZeaGBSv2.6).

Data Structure	Data Points	Compressed Size	Uncompressed size
Sequencing Files	4,679 Gnt[1]	3.9 TB	11.6 TB
Tags by Taxa (TBT)	3.2 trillion[2]	82.0 GB	3,198 GB
Tags on Physical Map (TOPM)	10.2 billion[3]	6.44 GB	14.0 GB

[1]Giganucleotides
[2]Read depths for 97,502,532 tags across 32,736 taxa (including 758 blank negative controls.
[3]105 data points per tag (with each base counted as one data point) times 97,502,532 tags.
doi:10.1371/journal.pone.0090346.t001

quality scores are ignored, and therefore do not need to be tracked throughout the pipeline.

Population genetic-based filtering of putative SNPs

Putative SNPs from GBS may be of low quality for multiple reasons. The sequencing error rate for a SNP may be high because of its distance from the read start and/or its immediate sequence context [47,48]. Alternatively, paralogous sequence tags from different loci may be mistakenly aligned to a single TagLocus, resulting in spurious SNPs. To detect and filter out error-prone SNPs, the TASSEL-GBS pipeline relies on population-genetic parameters such as the minor allele frequency (MAF) and, in particular, the inbreeding coefficient (or "index of panmixia"), F_{IT}. Filtering based upon minimum MAF can remove spurious SNPs arising solely from sequencing error. Artifactual SNPs originating from paralogous tags will tend to be excessively heterozygous and can thus be distinguished on the basis of low F_{IT}.

These population-genetic filters are most powerful if a substantial proportion of the samples consist of inbred lines. Among inbred samples, both error-prone SNPs and spurious SNPs originating from paralogous tags will appear to be excessively heterozygous. The Discovery SNP caller in the TASSEL-GBS pipeline allows the user to specify which samples are highly inbred, and uses this subset of inbreds to calculate F_{IT} and apply the minimum F_{IT} filter. Additional, related, filters can also be applied enforcing a minimum "inbred coverage" (proportion of the inbred samples to be non-missing at the SNP) and maximum "inbred heterozygosity score" ($=$ *nInbredHets*/[*nInbredsGT1ReadHomoMin* + *nInbredHets* + 0.5], where *nInbredHets* is the number of inbred taxa that are scored as heterozygous for the SNP, and *nInbredsGT1ReadHomoMin* is the number of inbred taxa with a read depth >1 for the SNP that are scored as minor allele homozygotes).

Implementation

Discovery Pipeline

The Discovery Pipeline consists of multiple steps, with each step (with the exception of alignment to the reference genome) being carried out by a TASSEL "plugin" that can be run from the TASSEL 4 Standalone command line interface. Detailed documentation on the function and usage of each individual plugin is available at www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf. Rather than describe each individual plugin, here we describe the main functions of the pipeline (Figure 1) and their key features.

In order to have maximal power to discover and filter SNPs, we advocate running the Discovery Pipeline (i.e., performing a "Discovery Build") at the species-wide level, with all samples sequenced to date, across multiple FASTQ files. Each FASTQ file contains GBS data from multiple samples distinguished by DNA barcodes at the beginning of each read [12]. We currently perform GBS at 384 plex (3072 samples per flowcell), and run each GBS library prep on one or more flowcell lanes depending on the desired sequencing depth.

Collapsing reads into a master tag list. The TASSEL-GBS Discovery Pipeline first reads through each of the available FASTQ files and generates one output "TagCount" file per input FASTQ file (Figure 1A). Each output, binary TagCount file contains a sorted list of all the unique sequence tags observed in the corresponding FASTQ file, the length of each tag in bases, and the number of times each tag was observed. This list can be used as a key-value map where the tag sequence and length together serve as the key and the corresponding tag count serves as the value. Each nucleotide is encoded in two bits allowing tag sequences to be stored in Java longs (64 bits, or 32 bases per long). Degenerate bases or N's are not permitted and quality scores are not retained. Potentially chimeric sequences are eliminated by trimming the sequence at the corresponding restriction enzyme site, if present.

After this initial pass through all of the available FASTQ files, the individual TagCounts files are merged into a single, master TagCount file containing a list of all tags of interest for a species. Only tags occurring at or above a (user-specified) minimum number of reads across all of the FASTQ files in the experiment are retained in the output master tag list. The more times a particular tag has been observed, the less likely it contains a sequencing error. The minimum tag count controls the tradeoff between the amount of sequencing errors admitted into the analysis versus the minimum allele frequency of interest. We do not expect to eliminate all sequencing errors at this step. We are usually able to filter out or correct most of them in subsequent steps of the pipeline, or in further, downstream processing customized to the biology of the study population(s). Furthermore, tags containing one or more sequencing errors can still be useful to score SNPs at other, non-error positions.

Alignment of tags to the reference genome. Alignment of each tag of interest in the master tag list to the reference genome is carried out with third party software. To facilitate this, the master tag list file is converted from TagCount format into FASTQ format (with fake, uniformly high quality scores). Currently, SAM format [36] output alignment files produced by the free software programs Bowtie2 [49] or BWA [36] can be read by the TASSEL-GBS pipeline and converted into a "TagsOnPhysicalMap" (TOPM) file that can be used for SNP calling.

The TOPM contains all of the tags present in the master TagCount file and genomic positions for the subset of tags that align to a unique best position in the genome. Retention of all of the master tags in the TOPM, even those without unique best genomic positions, affords future incorporation of additional information regarding the positions of unplaced tags, either from alternative aligners or from genetic mapping evidence. The TOPM is sorted by tag sequence and functions as a key-value map where a tag sequence can be used as a key to retrieve its corresponding physical position. It is also possible to programmatically traverse the tags within a TOPM in their physical position order through the use of an in-memory, primitive treemap. At the SNP calling step of the Discovery Pipeline (see below), the alleles represented by each useful tag at each useful SNP ("variants") are added to the TOPM. Three file formats of the TOPM are supported: text, binary, and HDF5.

Determining the distribution of tags across individual samples. With a TOPM file available, the next ingredient needed to discover and call SNPs is a matrix that records the number of times each tag in the master tag list was observed in each DNA sample, which we refer to as a TagsByTaxa (TBT) file (Figure 1B). In order to maximize the capacity of our pipeline, we currently store the TBT in HDF5 format (http://www.hdfgroup.org). The HDF5 format facilitates extremely fast read and write access to large data sets. To construct a TBT, the DNA sample of origin of each good, barcoded read in the set of input FASTQ files is determined based upon its barcode. If a good, barcoded read matches one of the tags in the master tag list, its depth in the appropriate taxon is incremented in the output TBT, up to a maximum depth of 127. Since the matrix is often extremely sparse, a custom run length encoding compression algorithm was implemented that provides a high level of compression, with minimal reduction in access speed.

SNP discovery and initial filtering. SNP discovery (Figure 1C) is performed for each set of tags that align to the exact same starting genomic position and strand, where the starting genomic position of a tag is defined by the cut site remnant at the beginning of the tag. Such tags, originating from the same restriction enzyme cut site and with the same orientation (but not necessarily of the same length), collectively comprise a "TagLocus". To call SNPs and ensure that indels are handled consistently, a *de novo* multiple sequence alignment of all the tags in each TagLocus is performed using the BioJava 3.0 API [50], which implements the CLUSTAL W algorithm [51]. For each SNP in the resulting "TagLocusAlignment", the allele represented by each tag is determined and the TBT file is consulted to tally the observed depths of each allele in each taxon. The genotype of the SNP in each taxon is then determined either by a binomial likelihood ratio method of quantitative SNP calling (the default; for details see Supplementary Text S1) or, optionally, following the method of Hohenlohe et al. [52].

After genotypes are obtained for a potential SNP, initial filtering is then performed based upon user settings for minimum minor allele frequency, and for a minimum coefficient of panmixia, or inbreeding relative to the entire population, F_{IT}, (where $F_{IT} = 1 - Ho/He$, Ho = observed heterozygosity, He = expected heterozygosity = $2q(1\text{-}q)$, and q = minor allele frequency). Error-prone SNPs and spurious SNPs from paralogy often appear excessively heterozygous, with lower F_{IT} than expected. If the user supplies a "pedigree file" that indicates the expected inbreeding coefficient (F) of each taxon, then only inbred taxa, with an expected inbreeding coefficient greater than or equal to the user-specified minimum coefficient of panmixia (*minF* parameter), are used in the calculation of F_{IT}. Inbred samples, available in many crop species and model organisms, can greatly add to the power of this filter. If enough inbred samples are available, then, additional filtering of SNPs can then be optionally performed enforcing a minimum "inbred coverage" (proportion of the inbred samples to be non-missing at the SNP) and a maximum "inbred heterozygosity score" (defined above).

To illustrate the effectiveness of these population genetic-based SNP filters, as applied in our most recent maize Discovery Build (AllZeaGBSv2.6, with 31,978 samples), we focused on the subset of 5,254 samples from the maize Nested Association Mapping (NAM) population [53]. The NAM population is a series of 25 biparental, F_2-derived RIL families all with a common female parent, the inbred line B73. Error-prone SNPs can be identified in a biparental family through their tendency, when they are in fact not segregating, to *appear* to be weakly polymorphic, but with segregation ratios significantly deviating from the 1:1 expectation. In contrast, non-segregating SNPs in a given family that display no spurious polymorphism are free of genotypic error in that family. The availability of the 25 biparental NAM families provides tremendous power to detect error-prone SNPs, and thus to examine the effectiveness of our filters.

We used the 5,254 NAM RIL samples to estimate error rates for three alternative sets of chromosome 10 SNPs discovered in the full set of 31,978 maize samples comprising our AllZeaGBSv2.6 Discovery Build. The three sets of chromosome 10 SNPs were obtained after application of three different filtering regimes: (1) no filters other than MAF >= 0.001, (2) a minimal filter only for MAF >= 0.01, and (3) our "standard" maize Discovery Build filters of MAF >= 0.001, minimum F_{IT} in inbred samples of 0.8, inbred coverage >0.15, and inbred heterozygosity score <0.21. The SNPs were discovered and filtered based upon all 31,978 maize samples (including the NAM RILs) and their allele frequencies were then separately calculated in each of the 25 NAM families. To minimize sampling error, allele frequency was only estimated for a particular SNP-NAM family combination if at least 19 RILs in that family had non-missing genotypes for the SNP (n >= 19). Thus, by "SNP-NAM family combination" we mean a particular SNP (e.g., "S10_2918") in a particular NAM family (e.g., "B73 × B97") that has at least 19 non-missing genotypes, regardless of whether it is polymorphic or not within that family. Hence, for each SNP, allele frequencies were calculated in 25 or fewer families, depending on the amount of missing data in each family. NAM family-specific minor allele calls for a SNP were classified as errors if the family-specific MAF was greater than zero but less than 0.25, and the SNP significantly deviated from 1:1 segregation in that family at $p<0.001$ (binomial test). The overall error rate for a SNP was then estimated as the total number of these error calls divided by the total number of calls for that SNP in NAM families with n >= 19 where the SNP significantly deviated from 1:1 segregation at $p<0.001$ (including the monomorphic SNPs). A pictorial explanation of this method of estimating error rates, using a single NAM family for illustrative purposes, is provided in Supplementary Text S2.

Compared to non-filtered SNPs (Figures 3A and 3B), application of our "standard" maize Discovery Build filters (MAF >= 0.001, minimum F_{IT} in inbred samples of 0.8, inbred coverage >0.15, inbred heterozygosity score <0.21) (Figures 3E and 3F) greatly increased the proportion of SNP-NAM family combinations displaying either appropriate 1:1 segregation (Figure 3F), or no polymorphism at all (Figure 3E). In contrast, a minimal filter based only on MAF (MAF >= 0.01) (Figures 3C and 3D) was far less effective at removing error-prone SNPs than our standard filters. Furthermore, our standard filters clearly shifted the distribution of error rates (estimated from the NAM samples) toward zero (Figure 4) and reduced the mean error rate (Table 2) relative to either no filtering (other than MAF >= 0.001) or minimal filtering (only for MAF >0.01). Extremely low estimates of mean and median error rates after application of our standard filters (0.0042 and zero respectively; Table 2) indicate that, for the most part, highly reliable SNP genotypes are produced by the GBS assay and the TASSEL-GBS pipeline, at least for inbred samples (where under-calling of heterozygotes due to low coverage is not an issue).

Error prone SNPs that are not removed by our standard filters (e.g., polymorphic SNPs with family specific MAFs <0.25 in Figure 3F) can be easily removed from biparental RIL families by filtering for appropriate allele frequencies and/or based on their relatively low levels of linkage disequilibrium (LD) with neighboring SNPs. In addition, error-prone SNPs identified in biparental families can be excluded from analyses of the remaining, non-biparental samples in a Discovery Build. If biparental populations are not available in your study species, it should be possible to use half-sib (e.g., open pollinated) families, or any population having a prior expected allele frequency range for polymorphic markers, to filter out error-prone SNPs. Alternatively, recently bottlenecked populations with high levels of extended LD can be used to filter out error-prone SNPs, which should display relatively low levels of LD with their neighboring SNPs.

After SNP calling and filtering, the input TOPM is updated with variants, and this "production-ready" TOPM is then saved to a new file (Figure 1C). For each SNP that passed the filtering step, the allele that is represented by each tag in the corresponding TagLocus is recorded in the "production-ready" TOPM, as well as the relative position of the SNP with respect to the genomic position of the TagLocus. The production TOPM produced by our AllZeaGBSv2.6 Discovery Build contained 955,690 useful SNPs.

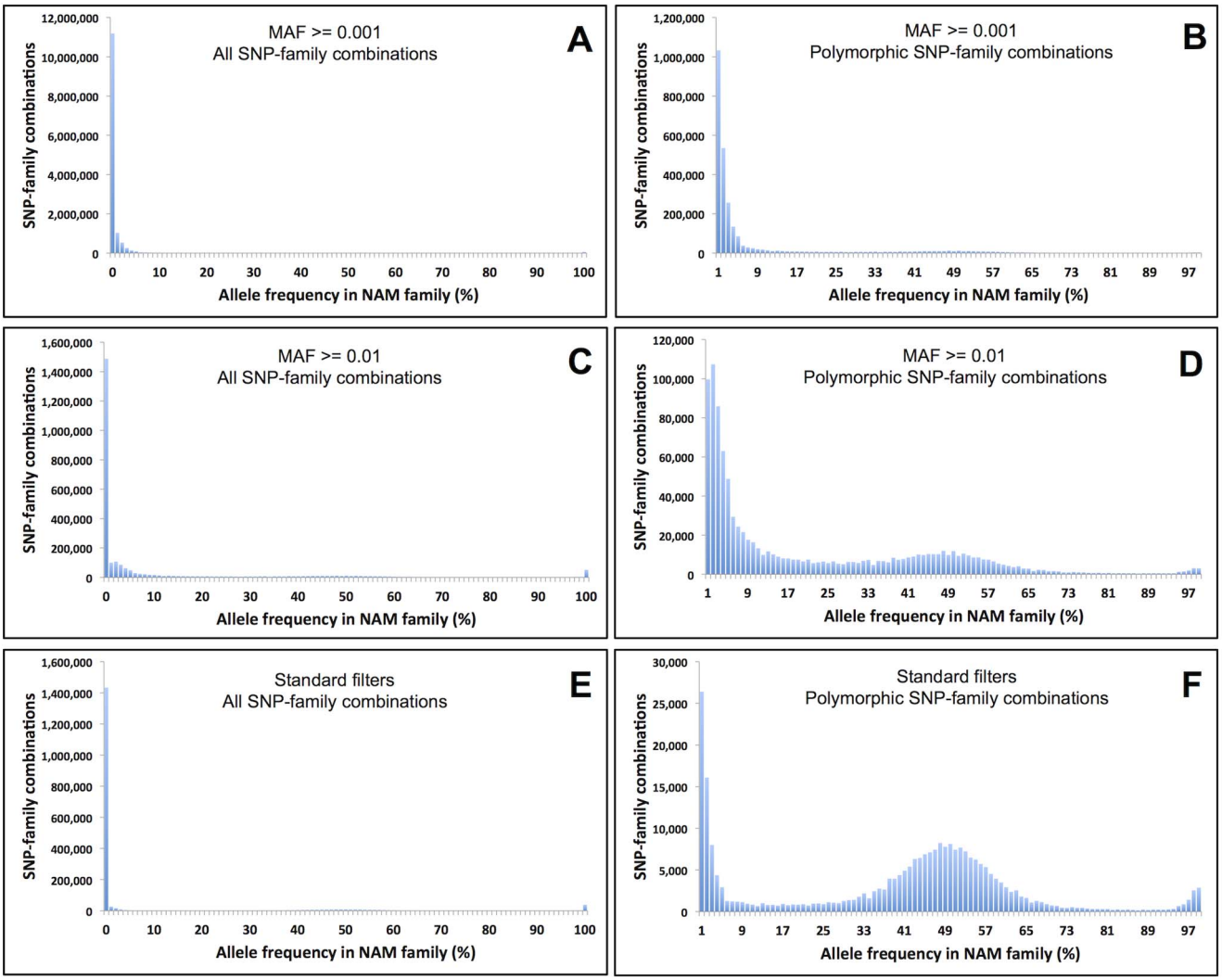

Figure 3. Within NAM family allele frequency distributions of chromosome 10 SNPs after different levels of filtering. Allele frequencies were calculated in each of the 25 Nested Association Mapping (NAM) families (collectively comprising 5,254 RILs) after application of the filters to the entire set of 31,978 maize samples in the AllZeaGBSv2.6 build. Allele frequencies were only estimated in a NAM family if at least 19 RILs had non-missing genotypes. Each histogram shows the allele frequency distribution for all the SNP-NAM family combinations with n >= 19. (A, B) No filter other than minimum MAF of 0.001. (C, D) A minimal filter only for MAF >= 0.01. (E, F) "Standard" maize build filters of MAF >= 0.001, minimum F_{IT} in inbred samples of 0.8, inbred coverage >0.15, and inbred heterozygosity score <0.21. (A, C, E) All SNP-family combinations: the error-free, monomorphic SNP-family combinations dwarf the segregating SNPs in all three cases. (B, D, F) Polymorphic SNP-family combinations only: omitting the monomorphic SNP-family combinations permits visualization of the remaining allele frequency distribution.
doi:10.1371/journal.pone.0090346.g003

Production Pipeline

In contrast to the multiple-step Discovery Pipeline, the Production Pipeline consists of a single step, utilizing the production-ready TOPM generated by the Discovery Pipeline to produce genotypes directly from input FASTQ files (Figure 2). The Production Pipeline determines the taxon of origin of each good, barcoded sequence read in each input FASTQ file and then checks if the read matches one of the useful tags in the production-ready TOPM. In this manner, allelic depths for each useful SNP in the TOPM are recorded for each taxon, allowing quantitative SNP calling to be performed, again either by our own binomial likelihood ratio method (for details see Supplementary Text S1) or, optionally, according to the method of Hohenlohe et al. [52]. If the GBS library preps for some samples have been run in replicate on multiple flow cell lanes (to increase depth), the corresponding allelic depth information is tallied across the replicates prior to SNP calling. Genotype files are produced in HapMap format as well as in or our own custom HDF5 format (which also records allelic depth). The ability to convert from this custom HDF5 format into VCF format [54], which also retains allelic depth, will be added to the TASSEL GUI in the near future.

Downstream Processing

Further processing of the genotype files, such as sub-setting out specific taxa or genomic regions of interest, filtering SNPs or taxa based upon coverage, or filtering of SNPs based on minor allele frequency, can be performed either with the TASSEL 4 GUI or the TASSEL 4 Standalone command line interface (the ability to filter for SNPs that are in LD with their neighbors will be added soon). Depending on the genome size, the exact molecular protocol and restriction enzyme used, and the sequence depth obtained, missing data can be common and actual heterozygotes

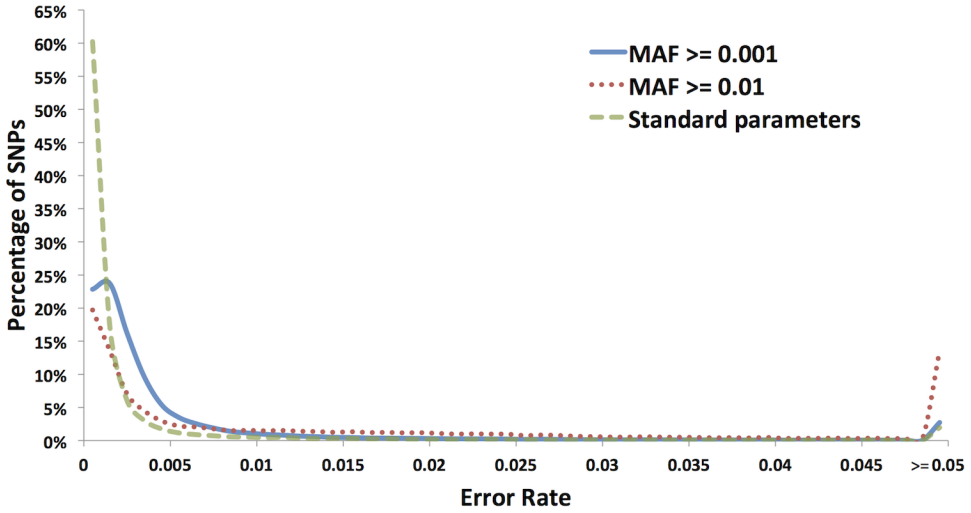

Figure 4. Error rate distribution of chromosome 10 SNPs for different levels of filtering. Error rates in the AllZeaGBSv2.6 Discovery build SNP calls were estimated using the NAM biparental families. NAM family-specific minor allele calls were defined as errors if the family-specific MAF was greater than zero but less than 0.25, and the SNP significantly deviated from 1:1 segregation in that family at $p<0.001$.
doi:10.1371/journal.pone.0090346.g004

can be substantially under-called as homozygotes. Hence, imputation of missing data (and, possibly, phasing of the final genotypes) is usually desirable. Since the optimal imputation approach depends greatly upon the biology of the species and the experimental design, this topic is beyond the scope of this paper. Numerous, general purpose imputation tools are already available [55]. Custom imputation approaches that we are developing for our maize experimental populations will be the subject of future publications.

Hardware Needs and Installation

The TASSEL-GBS pipeline is part of the TASSEL package, which is written in Java, so it can be run on Linux, Mac, or Windows operating systems. A minimum of 8 GB of RAM is required (at least 16 GB is recommended). Detailed installation instructions are provided at www.maizegenetics.net/tassel as well as in the TASSEL-GBS pipeline documentation (www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf). The source code is available from sourceforge.net/p/tassel/code/ci/master/tree, JUnit tests at sourceforge.net/p/tassel/maizegenetics4-test/ci/master/tree and a relatively small test data set at www.maizegenetics.net/tassel/GBSTestData.tar. The current version of TASSEL-GBS as described herein is implemented in TASSEL V4.3.5. Although we generally recommend using the latest version, users can revert to V4.3.5 or other versions by following the instructions posted in the following document www.maizegenetics.net/tassel/docs/UpdatingTasselStandaloneUsingGit.pdf, under the heading "To Update Packages to Older Releases…". The TASSEL-GBS pipeline will soon be available in TASSEL 5; major improvements to the pipeline (e.g., full VCF format support, allowing tag lengths greater than 64 bp, storage of tag depths per individual up to 10,000 rather than the current maximum of 127) will be implemented there.

Table 2. Comparison of error rates for chromosome 10 SNPs from the AllZeaGBSv2.6 build for different levels of filtering by the Discovery SNP caller.

Filter[1]	nSNPs	nSNPsTested[2]	avgErrorRate[3]	avg nSegregating[4]
MAF >= 0.001	694,517	680,623	0.00681	12,899
MAF >= 0.01	149,480	136,296	0.02218	12,818
Standard[5]	78,627	78,506	0.00420	7,192

[1]Filters applied to the entire build (31,978 non-blank samples)
[2]Minimum sample size of 19 in at least one maize Nested Association Mapping (NAM) family
[3]Average error rates estimated from 5,254 NAM RILs. Median error rates were zero for all three filters.
[4]Average number of chromosome 10 SNPs with n >= 19 and MAF between 0.25 and 0.75 across the 25 NAM families.
[5]MAF >= 0.001, minimum F_{IT} in inbred samples of 0.8, inbred coverage >0.15, inbred heterozygosity score <0.21.
doi:10.1371/journal.pone.0090346.t002

Strengths and Weaknesses

Strengths

The strengths of GBS and the TASSEL-GBS pipeline are the large number of markers potentially produced (depending on the biology of the species and the choice of restriction enzymes), low cost and minimal startup cost, and integration of SNP discovery with SNP calling.

The potentially large number of markers available from GBS makes GWAS feasible in study populations where linkage disequilibrium (LD) extends far enough so that causative polymorphisms stand a reasonable chance of being in LD with one or more markers. Alternatively, the large number of markers facilitates accurate projection of haplotypes from a set of more densely genotyped reference haplotypes [56] derived from whole genome sequencing (WGS) (e.g., [57,58]). This projection strategy is especially effective if the WGS reference haplotypes are representative of the founders of the study population [59].

Compared to alternative high-density marker technologies such as SNP arrays, GBS is relatively inexpensive, particularly if low coverage data suffices for the purpose of your study. Startup costs for GBS are minimal, as startup involves only (1) testing that your one of your candidate restriction enzymes (or enzyme pairs) produces a suitable GBS library, and (2) optimization of the ratio of sample DNA to the PCR adapters [12]. In contrast, startup for

a SNP array involves ascertainment of SNPs in a small discovery panel and assay design for each individual SNP.

The common practice of using of a small panel of individuals to discover SNPs for inclusion in a SNP array introduces an ascertainment bias that can severely distort estimates of key population genetic parameters gauging genetic diversity, sub-population differentiation and relatedness among individuals [60–62]. In contrast, the TASSEL-GBS Discovery pipeline integrates SNP discovery with SNP calling, using all available samples to date, and thus avoids the ascertainment bias that would arise from a small discovery panel. This type of ascertainment bias will also be minimal for new samples run through the Production Pipeline, provided that their genetic diversity is well-represented among the samples included in the Discovery Build. However, there might still be some subtle biases in either pipeline, caused by factors such as null alleles [63,64], alignment to the reference, and the use of inbreds only (rather than the full set of samples) to filter SNPs for F_{IT}.

Weaknesses

The main weakness of the GBS assay, when conducted at low coverage, is the amount of missing data. However, numerous imputation approaches are currently available [55] and yet more are currently in development, for a wide range of biological scenarios. As discussed above, the most appropriate imputation method and the probability of imputation success depends upon the biology of the study population. For some purposes, such as estimation of population allele frequencies [65], kinship, relatedness, and genetic distance, phylogenetic reconstruction [22], or germplasm quality control [24], imputation of missing data is usually not necessary.

Conclusions

The TASSEL-GBS pipeline for identifying and calling SNPs from next-generation, genotyping by sequencing data fulfills our design criteria better than any existing pipeline. It has a capacity for very large analyses involving tens of thousands of samples, yet can also be run at smaller scales. The pipeline permits rapid processing of the data, yet has a relatively modest memory footprint, allowing it to be run on desktop or laptop computers. This increases its usability by researchers in developing countries who may lack access to sophisticated computing resources. The separation of SNP discovery and genotyping into two phases reduces potential ascertainment biases and, more importantly, makes the TASSEL-GBS pipeline highly suitable for use in a genomics-assisted, accelerated breeding context, where rapid turnaround times from tissue collection to genotypes are essential. Furthermore, the high density of markers potentially available from the GBS assay should enable accurate genomic prediction over multiple generations.

Supporting Information

Text S1 Description of our binomial likelihood ratio method of quantitative SNP calling.
(PDF)

Text S2 Estimation of GBS SNP error rates using biparental RIL families.
(PDF)

Acknowledgments

We thank Yang (Jon) Zhang for contributions to the code for VCF output of genotypes and Sara J. Miller for assistance with manuscript formatting. We also thank the numerous users of the TASSEL-GBS pipeline (Katie Hyma in particular) who have made useful suggestions or have reported bugs.

Author Contributions

Conceived and designed the experiments: ESB JCG QS. Performed the experiments: ESB JCG TMC FL JH QS RJE. Analyzed the data: JCG ESB TMC QS RJE. Wrote the paper: JCG ESB.

References

1. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46. doi:10.1038/nrg2626.
2. Shendure J, Lieberman Aiden E (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30: 1084–1094. Available: http://www.ncbi.nlm.nih.gov/pubmed/23138308.
3. Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126: 1–11. Available: http://www.ncbi.nlm.nih.gov/pubmed/22948437. Accessed 11 November 2013.
4. Kilpinen H, Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. Trends Genet 29: 23–30. Available: http://www.ncbi.nlm.nih.gov/pubmed/23103023. Accessed 11 November 2013.
5. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513–516. Available: http://www.ncbi.nlm.nih.gov/pubmed/11029002.
6. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12: 499–510. Available: http://www.ncbi.nlm.nih.gov/pubmed/21681211. Accessed 6 November 2013.
7. Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J 5: 92. Available: https://www.crops.org/publications/tpg/abstracts/5/3/92. Accessed 11 November 2013.
8. Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol 22: 2841–2847. Available: http://www.ncbi.nlm.nih.gov/pubmed/23711105. Accessed 7 November 2013.
9. Van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, et al. (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 2: e1172. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2048665&tool=pmcentrez&rendertype=abstract. Accessed 11 November 2013.
10. Van Tassell C, Smith T, Matukumalli L (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5: 247–252. Available: http://www.nature.com/nmeth/journal/v5/n3/abs/nmeth.1185.html. Accessed 2013 Nov 11.
11. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2557064&tool=pmcentrez&rendertype=abstract. Accessed 2013 Nov 7.
12. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One 6: e19379. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract. Accessed 2011 Jul 18.
13. Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, et al. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res 21: 610–617. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3065708&tool=pmcentrez&rendertype=abstract. Accessed 2013 Nov 8.
14. Wang S, Meyer E, McKay JK, Matz M V (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Methods 9: 808–810. Available: http://www.ncbi.nlm.nih.gov/pubmed/22609625. Accessed 2013 Nov 8.
15. Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA, et al. (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One 7: e37565, . Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3360789&tool=pmcentrez&rendertype=abstract. Accessed 2013 Nov 8.
16. Monson-Miller J, Sanchez-Mendez DC, Fass J, Henry IM, Tai TH, et al. (2012) Reference genome-independent assessment of mutation density using restriction enzyme-phased sequencing. BMC Genomics 13: 72, . Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3305632&tool=pmcentrez&rendertype=abstract. Accessed 2013 Nov 11.
17. Chen Q, Ma Y, Yang Y, Chen Z, Liao R, et al. (2013) Genotyping by genome reducing and sequencing for outbred animals. PLoS One 8: e67500. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3715491&tool=pmcentrez&rendertype=abstract. Accessed 2013 Nov 11.

18. Morishige DT, Klein PE, Hilley JL, Sahraeian SM, Sharma A, et al. (2013) Digital genotyping of sorghum – a diverse plant species with a large repeat-rich genome. BMC Genomics 14: 448. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3716661&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 10.
19. Stolle E, Moritz RFA (2013) RESTseq–efficient benchtop population genomics with RESTriction Fragment SEQuencing. PLoS One 8: e63960. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3656931&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
20. Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3289635&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
21. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, et al. (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A 110: 453–458. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3545811&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 8.
22. Lu F, Lipka AE, Elshire RJ, Glaubitz JC, Cherney J, et al. (2013) Switchgrass genomic diversity, ploidy and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9: e1003215.
23. Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, et al. (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A 110: 5241–5246. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3612656&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
24. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, et al. (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14: R55. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3707059&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 7.
25. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, et al. (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8: e54603. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3553054&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 8.
26. De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One 8: e62137. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3656875&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
27. Rutkoski JE, Poland J, Jannink J-L, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda) 3: 427–439. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3583451&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
28. Kim S-I, Tai TH (2013) Identification of SNPs in closely related Temperate Japonica rice cultivars using restriction enzyme-phased sequencing. PLoS One 8: e60176, . Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3608622&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
29. Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, et al. (2013) Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava. Crop Sci 53: 1312. Available: https://www.crops.org/publications/cs/abstracts/53/4/1312. Accessed 2013 Nov 11.
30. Saintenac C, Jiang D, Wang S, Akhunov E (2013) Sequence-based mapping of the polyploid wheat genome. G3 (Bethesda) 3: 1105–1114. Available: http://www.ncbi.nlm.nih.gov/pubmed/23665877. Accessed 2013 Nov 7.
31. White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes glareolus) in Ireland. Mol Ecol 22: 2971–2985. Available: http://www.ncbi.nlm.nih.gov/pubmed/23701376. Accessed 2013 Nov 7.
32. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635. Available: http://www.ncbi.nlm.nih.gov/pubmed/17586829. Accessed 2011 Aug 9.
33. Mascher M, Wu S, Amand PS, Stein N, Poland J (2013) Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. PLoS One 8: e76925. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3789676&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 7.
34. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping Loci de novo from short-read sequences. G3 (Bethesda) 1: 171–182. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3276136&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 7.
35. Catchen J, Bassham S, Wilson T, Currey M, O'Brien C, et al. (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Mol Ecol 22: 2864–2883. Available: http://www.ncbi.nlm.nih.gov/pubmed/23718143. Accessed 2013 Nov 11.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2723002&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
37. Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19: 1124–1132. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2694485&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
38. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3083463&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
39. Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13: 174–180. Available: http://www.ncbi.nlm.nih.gov/pubmed/20089441. Accessed 2013 Nov 8.
40. Xie W, Feng Q, Yu H, Huang X, Zhao Q, et al. (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci U S A 107: 10578–10583. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2890813&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
41. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3154645&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
42. Harter AV, Gardner KA, Falush D, Lentz DL, Bye RA, et al. (2004) Origin of extant domesticated sunflowers in eastern North America. Nature 430: 201–205. Available: http://www.ncbi.nlm.nih.gov/pubmed/15241413.
43. Ofori A, Becker HC, Kopisch-Obuch FJ (2008) Effect of crop improvement on genetic diversity in oilseed Brassica rapa (turnip-rape) cultivars, detected by SSR markers. J Appl Genet 49: 207–212. Available: http://www.ncbi.nlm.nih.gov/pubmed/18670055.
44. Cowling WA (2013) Sustainable plant breeding. Plant Breed 132: 1–9. Available: http://doi.wiley.com/10.1111/pbr.12026. Accessed 2013 Nov 11.
45. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2532726&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
46. Eren AM, Vineis JH, Morrison HG, Sogin ML (2013) A filtering method to generate high quality short reads using illumina paired-end technology. PLoS One 8: e66643. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3684618&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
47. McElroy KE, Luciani F, Thomas T (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics 13: 74. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3305602&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
48. Allhoff M, Schönhuth A, Martin M, Costa IG, Rahmann S, et al. (2013) Discovering motifs that induce sequencing errors. BMC Bioinformatics 14 Suppl 5: S1. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3622629&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
49. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3322381&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 7.
50. Prlić A, Yates A, Bliven SE, Rose PW, Jacobsen J, et al. (2012) BioJava: an open-source framework for bioinformatics in 2012. Bioinformatics 28: 2693–2695. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3467744&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 10.
51. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 308517&tool = pmcentrez&rendertype = abstract.
52. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, et al. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6: e1000862. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2829049&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 7.
53. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. (2009) Genetic properties of the maize nested association mapping population. Science 325: 737–740. Available: http://www.ncbi.nlm.nih.gov/pubmed/19661427. Accessed 2011 Jun 14.
54. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156–2158. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3137218&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
55. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499–511. Available: http://www.ncbi.nlm.nih.gov/pubmed/20517342. Accessed 2013 Nov 6.

56. Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 (Bethesda) 1: 457–470. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3276165&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
57. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3498066&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
58. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44: 803–807. Available: http://dx.doi.org/10.1038/ng.2313. Accessed 2012 Nov 6.
59. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43: 159–162. Available: http://www.ncbi.nlm.nih.gov/pubmed/21217756. Accessed 2011 Jun 13.
60. Eller E (2001) Effects of ascertainment bias on recovering human demographic history. Hum Biol 73: 411–427. Available: http://muse.jhu.edu/journals/hub/summary/v081/81.5-6.eller.html. Accessed 2013 Nov 11.
61. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496–1502. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 1310637&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 6.
62. Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27: 2534–2547. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3107607&tool = pmcentrez&rendertype = abstract. Accessed 2013 Nov 11.
63. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, et al. (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol 22: 3165–3178. Available: http://www.ncbi.nlm.nih.gov/pubmed/23110526. Accessed 2013 Nov 11.
64. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22: 3179–3190. Available: http://www.ncbi.nlm.nih.gov/pubmed/23551379. Accessed 2013 Nov 7.
65. Alex Buerkle C, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? Mol Ecol 22: 3028–3035. Available: http://www.ncbi.nlm.nih.gov/pubmed/23174005. Accessed 2013 Nov 6.