Analysis of genotyping-by-sequencing (Gbs) data

Kagale, Sateesh; Koh, Chushin; Clarke, Wayne E.; Bollina, Venkatesh; Parkin, Isobel A. P.; Sharpe, Andrew G.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1007/978-1-4939-3167-5_15
Plant Bioinformatics, pp. 269-284, 2016

Analysis of Genotyping-by-Sequencing (GBS) Data

Sateesh Kagale[1], Chushin Koh[1], Wayne E. Clarke[2], Venkatesh Bollina[2], Isobel A.P. Parkin[2] and Andrew G. Sharpe[1]

1. National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK, S7N0W9, Canada

2. Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, S7N0X2, Canada

Summary

The development of genotyping-by-sequencing (GBS) to rapidly detect nucleotide variation at the whole genome level, in many individuals simultaneously, has provided a transformative genetic profiling technique. GBS can be carried out in species with or without reference genome sequences yields huge amounts of potentially informative data. One limitation with the approach is the paucity of tools to transform the raw data into a format that can be easily interrogated at the genetic level. In this chapter we describe bioinformatics tools developed to address this shortfall together with experimental design considerations to fully leverage the power of GBS for genetic analysis.

1. Introduction

It was a significant achievement when the first plant genome sequence of *Arabidopsis thaliana* was published in 2000 [1] and heralded the application of genomics tools to plant research. The choice of this first species, with one of the smallest plant genomes and limited dispersed repetitive DNA, was partly driven by the cost and efficiency of available sequencing technologies. Today the transformative advances in sequencing platforms and chemistries, which have led to dramatic reductions in cost per base, have played a major role in deciphering multiple complex genomes. To date as many as 55 plant genomes have been sequenced and made publicly available [2] (http://www.phytozome.net/). Combined with such reference genome sequences next generation sequencing (NGS) has allowed a multitude of new approaches to be applied to the identification, analyses and visualisation of fundamental genetic variation. Identifying and utilising natural and induced genetic variation remains a prime objective in plant research with important implications in population genetics, evolution and crop breeding. The most abundant and perhaps most informative variation that can be exploited are single nucleotide

polymorphisms (SNPs) that have proven ideal markers for the study of plant genomes [3].

A number of approaches have been described to capture genome wide natural and induced genetic variation by NGS. The majority of these approaches rely on the use of reduced representation, which delimits the portion of large and complex genomes to be assessed to a manageable size. Initially proposed by Altshuler et al., [4] reduced representation allowed a high density SNP map to be generated for a genome previously thought to be too large for such analyses. However, it has been the combination of reduced representation, NGS and multiple indexing of samples that has provided the ability to study extremely large genomes at reasonable cost. The relative simplicity and cost-effectiveness of the genotype-by-sequencing (GBS) approach has encouraged its application in multiple species, including both model and non-model plants [5-8]. Also the increased marker density that is offered has led to its growing use in the anchoring of genome sequence assemblies, effectively removing the necessity to generate expensive and error prone physical maps [9-11]. The only current limitation is the bioinformatic and computational burden that is generated, with regards to both data processing and storage.

GBS now takes many forms, the first GBS data was generated using restriction site associated DNA sequencing (RAD-seq) [12] which utilised a single restriction enzyme combined with shearing of the digested DNA to capture a suitable portion of the genome. By optimising enzyme choice and eliminating the necessity for DNA shearing the Cornell group simplified the approach and allowed more extensive multiplexing, which reduced costs further [13]. There have been several modifications to the basic protocols, predominantly incorporating the use of two enzyme digestion, including 2b-RAD [14], ddRAD-seq [15] and a variant to the Cornell GBS approach

by Poland et al. [6] that utilizes methylation sensitive enzymes to further reduce the representation of the target genome. There have been several reviews describing the different approaches to GBS in plants [16-19].

The common feature of all the approaches is the type and volume of data that is produced, since all have exploited the Illumina sequencing platforms, generating millions of sequence reads usually of 100 bp or less for each indexed sample. Thus the bioinformatics pipeline described in the following chapter would be applicable to any of the published protocols in either single-end or paired-end read format. All the methods can be used in the absence of a reference genome; however, the use of a reference genome is generally far more effective in ensuring the robust identification of genome wide SNPs. The following chapter will focus on the analyses of GBS data where there is access to a complete or draft genome; although tools (section 2.1) that have been developed to analyse GBS in the absence of a reference genome are listed.

2. Materials

In this chapter, we discuss a Bioinformatics pipeline (**Figure 1**) that is designed to identify genetic variants such as SNPs and insertions/deletions (InDels) from NGS data generated by most major RAD and GBS approaches. This pipeline uses a suite of publicly available software and custom Perl scripts. There are alternative pipelines that have been developed and are listed in section 2.1.

2.1. Publicly available Software and tools for GBS:

1. **Trimmomatic** (http://www.usadellab.org/cms/?page=trimmomatic) is a multithreaded command line tool that can be used for trimming adapter sequences and low quality

regions from Illumina sequencing reads [20].

2. **Bowtie2** (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) is an ultrafast short read alignment tool that can be used for aligning sequencing reads against a reference genome [21]. It should be noted that other alignment tools are available for this application, most commonly BWA [22].
3. **SAMtools** (http://samtools.sourceforge.net/) is a package of utilities designed for manipulating alignments in the SAM (Sequence alignment/Map) or BAM (Binary alignment/Map) format, including sorting, merging, indexing and generating alignments in a per-position format [23].
4. **BCFtools** (http://samtools.github.io/bcftools/) is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart (BCF).
5. **GATK (Genome Analysis Toolkit) genotyper** (http://www.broadinstitute.org/gatk/) provides a wide variety of tools for variant discovery and genotyping [24-26].
6. **STACKS** (http://creskolab.uoregon.edu/stacks/) allows *de novo* assembly of short read GBS data and the identification of genetic variation in the absence of a reference genome [27].
7. **TASSEL-GBS** (http://www.maizegenetics.net/) is an implementation of a GBS analysis pipeline in the TASSEL software package [28].

2.2. In house tools:

A set of utility Perl scripts (listed in **Table 1**) were written to perform various tasks associated with data processing, read alignment and SNP discovery. These scripts are open source and freely available upon request.

3. Methods

The basic workflow for variant discovery using NGS data generated by RAD-seq and GBS approaches can be divided into three sequential steps: (1) raw data processing, (2) read alignment to a reference genome or *de novo* assembly of the sequence tags, and (3) variant discovery and annotation. In general, these three steps are shared by most of the currently availably genotyping pipelines. In the following subsections, each of these steps are reviewed to provide background information for the available bioinformatics tools that are customised to perform various tasks associated with these steps.

3.1 Raw data processing

RAD-seq and GBS employ a highly multiplexed sequencing strategy for constructing reduced representation libraries for the Illumina NGS platform (*see* **Note 1**). Demultiplexing is the first key step of processing raw sequencing data, which separates reads into their corresponding samples based on barcode matching. Demultiplexing of Illumina reads is generally carried out using Illumina CASAVA or MiSeq reporter software; however, CASAVA cannot demultiplex RAD-seq and GBS reads which contain customised inline barcodes in only one of the adapter sequences. We have developed a Perl script *util_barcode_splitter.pl* (**Table 1**) to demultiplex RAD-seq and GBS reads.

Raw sequencing data often contain various types of errors and artefacts, such as base calling errors, low quality bases, adaptor contamination and duplicate reads [29]. Thus it is necessary to perform quality assessment and correction of reads by filtering or trimming of low quality reads or regions. There are numerous publicly available software that can be used for pre-processing of sequencing reads, such as Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), PRINSEQ (http://prinseq.sourceforge.net/), FastqMcf (http://code.google.com/p/ea-utils/wiki/FastqMcf), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and cutadapt

(http://code.google.com/p/cutadapt/). In our pipeline (**Figure 1**), we have adopted Trimmomatic, which is a fast, multithreaded command line tool that can be used to (i) remove adapter sequences, (ii) trim leading and trailing low quality regions (below a user defined quality threshold), (iii) scan the read with a user defined base-pair size sliding window and cut when the average quality per base has dropped below a threshold, and (iv) keeping only those read-pairs where both reads were longer than the specified minimal length. Trimmomatic is also designed to handle 'read-through' for paired-end data. A 'read-through' is when a fragment size smaller than the read length is sequenced and hence results in overlapping read-pairs that include both the target fragment and adapter sequence. It is essential to remove one of the reads in this case in order to avoid over-stating read-depth for variant calling.

Amplification by polymerase chain reaction (PCR) is often used for target enrichment during the preparation of libraries for next-generation sequencing. PCR duplicates resulting from the original DNA templates being sequenced many times can have a detrimental effect on the quality of variant calls especially when the coverage is low (*see* **Note 2**). Computational methods for the detection and removal of PCR duplicates have become available that generally rely on the observation of identical alignment positions of reads to the reference genome. Read mapping being a computationally intensive process (*see* **Note 3**), the development of an alternate method for detection of PCR duplicates based on direct comparison of read sequences is essential, especially when the proportion of PCR duplicates is very high. To this end, we have developed a Perl script *util_find_uniq_reads.pl* (**Table 1**) that compares read sequences and removes duplicate reads.

3.2 Read alignment to a reference genome

After read clean-up, alignment of short reads to a reference genome is the first step in a high-

throughput genotyping workflow. In the absence of a reference genome, paired-end sequencing data generated by RAD-seq or GBS approaches can be assembled *de novo* using software packages such as STACKS [27], UNEAK [30] or RApiD [31] to produce mini-contigs that can be used as a reference for read mapping and genotyping (*see* **Note 4**). In the last few years, a myriad of efficient short-read alignment programs, such as MAQ [32], mrsFast [33], STAMPY [34] Bowtie2 [35], BWA [22] and SOAP2 [36], have been developed. Most of these widely used aligners utilize hashing algorithms (MAQ, mrsFast, STAMPY) or Burrows–Wheeler transform (BWT) [37] based indexing (Bowtie2, BWA and SOAP2) for short read mapping. The hash-based aligners use hash tables to store the information of either the reference genome or short reads. A major drawback of the hash-based aligners is that they require prohibitive amount of memory (*see* **Note 3**). The second generation BWT-based aligners are preferred as they consume only a limited amount of memory [38, 39].

In our genotyping workflow (**Figure 1**), we have adopted Bowtie2 which is faster, more sensitive, and more accurate than BWA and SOAP2 across a wide range of parameter settings [35]. Bowtie2 supports both local and global (end-to-end) modes of alignment of short reads [35]. A local alignment considers only a short segment of the read and clips unaligned characters from one or both ends of the read to maximise the alignment score. Conversely, global alignment involves alignment of all characters in the read. In our experience, local mode of alignment of the reads is faster and useful for mapping reads generated by GBS, although less accurate (due to increased multi-mapping) than global alignment. GBS does not involve size fractionation of the sequencing library and hence sometimes results in the generation of fragments that are either too short to be useful or result in paired-end sequencing reads that overlap completely. On the other hand, the RAD-seq protocol includes a size fractionation step and most reads generated by this

non-overlapping approach can be aligned in an end-to-end manner. An example of the variation in the distribution of predicted enzyme sites for both RAD-seq (*Eco*RI) and GBS (*Pst*I and *Msp*I), together with a representation of relative genome coverage of each method, has been demonstrated for the *Brassica oleracea* genome [11]. RAD captured a greater portion of the genome with a high percentage of the potential sites being tagged and sequenced, while GBS coverage was impacted by the degree of cytosine methylation.

Multi-mapped reads are those that align to multiple locations within the reference genome sequence [40]. Most eukaryotic organisms, especially plants with polyploid genomes, carry orthologous and paralogous gene families that contain multiple isoforms of the same gene with nearly identical or similar sequences. Shorter reads being less specific tend to have more multi-mapping events. In polyploid plant species, the proportion of multi-mapped reads ranges from 20-60%. Discarding such a high proportion of multi-mapping reads will result in a significant loss of valuable information. Bowtie2 searches and reports all valid alignments that score better than a given cut-off. We use Perl utility scripts *bowtie2_extract_best_global_hit.pl* or *bowtie2_extract_best_local_hit.pl* to go through the SAM files and identify the best hit from multi-mapped reads as having the top most hit with at least X=6 (end-to-end) or X=12 (local) penalty score better than the runner up. The larger the X score the more confident a read is uniquely mapped but more alignments get discarded as a consequence.

Bowtie2 outputs alignments in SAM format which contains alignment data in human readable tab-delimited text. SAM files generally tend to be very large. BAM, a compressed binary version of SAM format, is a preferred format for the downstream variant detection analyses due to its relatively smaller size. We use the '*view*' command of SAMtools to convert mapped reads from

SAM to BAM format. For downstream analysis the alignments in BAM files must be sorted and indexed according to the chromosomal positions. To achieve this, we use the sort and index utilities of SAMtools.

3.3 Variant discovery

The next step after mapping reads to a reference genome is to call sequence variants (SNPs and InDels) from the processed BAM file. Multiple software tools for variant-calling are available, including SAMtools:mpileup/BCFtools [23], GATK [24-26], SOAP [41], SNVer [42] and GNUMAP [43]. A recent study performed systematic evaluation of these commonly used variant-calling bioinformatics pipelines and found a very poor concordance between variants called by each of these methods [44]. Each of the SNP calling methods are designed based on different sets of assumptions about the reference genome and reads, and their suitability in different situations depends upon various factors, including the nature of genotypes, presence or absence of multi-allelic SNPs, and sensitivity and specificity of detecting SNPs. In our variant-calling workflow, we have implemented two of the most commonly used SNP callers; SAMtools:mpileup/BCFtools [23] and GATK [24, 25]. Both of these pipelines also call InDels.

SAMtools:mpileup computes the likelihood of each possible genotype by generating a consensus sequence using the MAQ (Mapping and Assembly with Quality) model framework, which uses a general Bayesian framework for picking the base that maximizes the posterior probability with the highest Phred quality score, and outputs the information in the BCF format (binary variant call format). However, it does not call the variants. BCFtools does the actual calling and estimating allele frequency by applying the genotype likelihood information in BCF files. It generates output in the VCF (variant call format) format, which is the emerging standard for

storing variant data. Identification of InDels from paired-end reads is relatively more challenging than that of SNPs as incorrect placement of insertions or deletions during read alignment to a reference genome may lead to false positive SNPs. SAMtools:mpileup deploys a concept called Base Alignment Quality (BAQ; [45]) to provide an efficient and effective way to rule out false positive SNPs caused by alignment artefacts. With the BAQ strategy which is invoked by default in mpileup, the probability of a base being misaligned can be accurately measured. Although the combination of SAMtools:mpileup and BCFtools offers a straightforward way of calling SNPs and InDels, this approach is limited to only diploid calling as SAMtools:mpileup is designed to compute and handle only biallelic variants [45]. We have successfully used SAMtools:mpileup for variant-calling and genetic linkage mapping of populations produced from bi-parental crosses (Bollina et al., In preparation; [10, 11]).

GATK is similar to SAMtools but utilizes additional processing steps, such as local re-alignment around InDel loci in order to clean up alignment artefacts, marking non-informative duplicate reads, and quality recalibration of both base quality and variant quality to improve overall accuracy of variant-calling [24-26, 44]. GATK includes two variant calling tools, UnifiedGenotyper and HaplotypeCaller. The UnifiedGenotyper uses a Bayesian genotype likelihood model to estimate posterior probability of allele frequency at each locus. Additionally it utilizes information from multiple samples and supports SNP calling from non-diploid samples. The HaplotypeCaller, which combines a local *de novo* assembler with a more advanced hidden Markov model (HMM) likelihood function, outperforms the UnifiedGenotyper in discovering sequence variants. However, it currently supports only diploid calling and lacks multithreading support.

Filtering raw SNP candidates is an essential step in the genotyping workflow as its helps in reducing false positive calls made from biases in the sequencing data and removing those calls that do not fulfil specific thresholds for SNP and genotype properties. Filtering of false positive calls based on read depth and quality threshold is embedded within some of the currently available variant calling pipelines such as SAMtools and GATK. We perform additional filtering based on missing genotyping calls and minor allele frequency (MAF). The level of missing data depends upon sequencing coverage which is influenced by the multiplexing level and the output from sequencing platform [18, 46]. Missing data can be reduced by sequencing at higher depth and reducing the multiplexing level. An alternative method for replacing missing data is to impute missing values with plausible substitutes (*see* **Note 5**). In recent years, algorithms [47-49] have been developed for imputation of missing genotype data with great accuracy. MAF refers to the frequency at which the least common allele occurs in a given population [50]. We use the Perl utility script *filter_vcf.pl* (**Table 1**) to perform filtering based on missing genotype and MAF generally ignoring SNPs with a MAF less than 5%. The final output from the majority of the variant calling pipelines is generally in the VCF format which can be viewed using genomic viewers such as Tablet [51] or IGV [52] (**Figure 2**). We have also developed Perl scripts to generate genotype scores in tab delimited file formats for ease of downstream processing and analysis. The last step of our genotyping workflow involves merging SNPs based on identical segregation patterns. The cartoon in **Figure 3** depicts the logic as well as our approach for creating haplotypes blocks by merging closely linked SNP markers with identical segregation patterns to provide a recombination bin framework that can be easily incorporated into genetic mapping analysis.

3.4 Conclusion

The advent of very high throughput NGS platforms together with new technical methodologies to take advantage of these gains provided an opportunity for establishing high resolution genetic analysis in any species. The ability to profile large numbers of targeted loci for sequence variation in highly multiplexed sets of discrete individuals provided a platform for a range of applications. An initial limitation for the full deployment of these approaches have been the dearth of readily available bioinformatics tools to process the raw data to yield output that can be readily incorporated into classical genetic analyses. This chapter has outlined some of the recently available bioinformatics resources to enable researchers to establish GBS applications for genetic analysis in their laboratories, provided an example pipeline that could be utilized for this purpose, and also a description of key factors that need to be considered in experimental design.

4. Notes

1. Assessing sequencing data requirements

In many instances both RAD and GBS have been attempted with a number of restriction enzymes. However, the choice of a particular enzyme and the volume of sequencing data required depends on several factors such as, the genome size, sample multiplexing needs, GC content, frequency of the cut site (frequent to rare) and desired frequency of the sites throughout the genome. *In silico* analysis of a genome with a choice of an enzyme cut site would provide a glimpse prior to a selection. The RAD Counter tool provided on the RAD wiki website (https://www.wiki.ed.ac.uk/display/RADSequencing/Home;jsessionid=14E3C4ECD753766FC8E4EA41274A9BF1) provides the user with a simple Excel format to input relevant information

with respect to the above parameters to establish the optimal experimental design to ensure appropriate read depth is reached.

2. Removal of duplicate reads: advantages and limitations

Duplicate reads arising from PCR amplification during library preparation can result in perfect copies of the DNA template being sequenced multiple times. The proportion of duplicate reads can vary enormously and duplicate reads can artificially inflate read coverage which may have detrimental effect on the quality of variant calls. Hence the dataset used for variant calling should include only one copy per duplicate set of reads. Duplicate reads can be detected and removed by comparison of either the read sequences or their alignment coordinates. However, the risk of removal of identical or almost identical reads arising from duplicated genomic regions, especially in organisms carrying polyploid genomes, poses a serious challenge. Additionally, it is impossible to differentiate duplicate reads arising due to amplification bias and identical GBS tags originating from the same restriction site(s) at a particular genomic location. This is not an issue in the case of paired-end RAD tags as the additional DNA fragmentation combined with size fractionation step in RAD-sequencing protocol leads to the production of paired-end tags with at least one variable end. Thus we advise against removal of duplicate GBS tags, whereas the decision on removal of duplicate RAD tags should depend upon the ploidy status or the level of segmental duplication in the organism under consideration.

3. Computational Resources

The analysis of GBS and RAD data requires non-trivial computational resources. In order to reduce analysis time, the use of multiple CPU cores is recommended. Many desktop computers will be limited in the number of samples they can process by the available RAM. Additionally, the output of the analysis steps requires significantly more hard disk space than that of the raw

sequencing data. As an example of computational requirements, 96 GBS samples were processed using 16 CPU cores for Trimmomatic, Bowtie2, and GATK. The total time required to process the samples was approximately 13 hours and required at most 21GB of RAM. The samples were demultiplexed from 9.7GB of compressed fastq data and resulted in approximately 68 GB of uncompressed output using a pipeline optimized to reduce production of intermediary output files.

4. Single-end or paired-end mapping

Variant calling can be done using either single or paired-end data with resulting benefits in increased coverage with paired-end data. It is also difficult to accurately map single reads originating from regions with significantly higher sequence homology, such as repeat rich or duplicated genomic regions. Sequencing reads from both ends can partly overcome this difficulty. Filtering of paired-end sequencing data based on adapter contamination and quality as well as length thresholds results in the generation of a small proportion of single end reads. In such case both single-end and paired-end mapping followed by merging of separately generated SAM files before the variant discovery step is possible.

5. Data imputation

One issue with both RAD and GBS is the amount of missing data that can result from the sequencing, especially when this is carried out at a low level of coverage / depth. Hopefully such an outcome can be avoided in the first place by ensuring optimal levels of depth are reached by adopting an appropriate experimental design (see 5.1). However, when high levels of missing data result it is possible to adopt imputation approaches that are currently available for different experimental approaches with various population structures [49, 53]. As well, it is possible to limit the amount of missing data in some types of populations; for example bi-parental genetic

mapping populations as described in the main text. In this case the merging of SNP loci based on identical segregation patterns can be carried out to create haplotypes blocks with minimal missing data and a resultant recombination bin framework for genetic mapping analysis.

Acknowledgements

Table 1. List of utility Perl scripts designed to perform various tasks associated with genetic variant discovery using RAD-Seq and GBS data sets

Perl script	Utility
util_barcode_splitter.pl	Demultiplexes paired-end RADseq or GBS reads based on perfect match to barcodes
util_find_uniq_reads.pl	Compares read sequences and removes duplicate reads
bowtie2_extract_best_global_hit.pl	Goes through the SAM files and identifies the best hit from multi-mapped reads as having the top most hit with at least X=6 (or a user defined cut-off) penalty score better than the runner up.
bowtie2_extract_best_local_hit.pl	Goes through the SAM files and identifies the best hit from multi-mapped reads as having the top most hit with at least X=12 (or a user defined cut-off) penalty score better than the runner up.
filter_vcf.pl	Perform filtering based on missing genotype and minor allele frequency

Figure Legends

Figure 1. Bioinformatics workflow for genetic variant discovery using next generation sequencing based genotyping approaches such as RADseq and GBS.

The genetic variant calling pipeline comprises three major steps, including raw data processing, read mapping to a reference genome, and variant discovery. Each of these steps is further divided into multiple sub-steps. The bioinformatics tools (shown in purple), input and output file formats (green), and the purpose, methodology or general outcome of each sub-step (bullet points) in the workflow are presented.

Figure 2. Genomeviewer (IGV; Thorvaldsdóttir et al. 2013) images illustrating alignment to the reference genome of short paired-end reads generated by RAD-seq (A) and GBS (B) approaches. The top two/three tracks represent the reference contig and positions of restrictions site(s): *Eco*RI (RAD-seq) or *Pst*I and *Msp*I (GBS). The following tracks show reads from each individual library aligned back to the reference using Bowtie2. Read bases that match the reference are displayed in gray and those that do not match (sequence variants) are shown in yellow.

Figure 3. Overview of the approach used for generating haplotypes by merging SNPs with identical segregation patterns.

As per the example shown in this cartoon, 5 RAD SNPs (at positions 100, 200, 300, 400 and 500 bp) were identified on scaffold1234. SNP#1 and SNP#2 have identical segregation pattern, except for the missing data points, so as SNP#3 to SNP#5. Instead of using all 5 SNPs for genetic mapping, we combine SNPs with identical scores. The locus name of each merged RAD

SNP (haplotype) provides additional information: the first part of the name includes the scaffold name, the next number indicates chronological order of SNP pattern identified in the scaffold, the next two numbers indicate the base pair positions between which this haplotype pattern was found, and the final number indicates the count of independent SNPs that had this pattern.

References

1. The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408 (6814): 796-815.

2. Michael, T.P., and Jackson, S. (2013). The First 50 Plant Genomes. *Plant Gen*. 6, 1-7

3. Ganal, M., Altmann, T., and Roder, M. (2009). SNP identification in crop plants. *Curr Opin Plant Biol* 12, 211-217.

4. Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., and Lander, E.S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-516.

5. Barchi, L., Lanteri, S., Portis, E., Valè, G., Volante, A., Pulcini, L., Ciriaci, T., Acciarri, N., Barbierato, V., Toppino, L., and Rotino, G.L. (2012). A RAD tag derived marker based eggplant linkage map and the location of qtls determining anthocyanin pigmentation. *PLoS ONE* 7, e43740.

6. Poland, J.A., and Rife, T.W. (2012). Genotyping-by-Sequencing for plant breeding and genetics. *Plant Gen*. 5, 92-102.

7. Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A.W., Pellicer, J., and Buggs, R.J.A. (2013). Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol Ecol* 22, 3098-3111.

8. Liu, H., Bayer, M., Druka, A., Russell, J., Hackett, C., Poland, J., Ramsay, L., Hedley, P., and Waugh, R. (2014). An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. *BMC Genomics* 15, 104.

9. Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., Cannon, S., Baek,

J., Rosen, B.D., Tar'an, B., Millan, T., Zhang, X., Ramsay, L.D., Iwata, A., Wang, Y., Nelson, W., Farmer, A.D., Gaur, P.M., Soderlund, C., Penmetsa, R.V., Xu, C., Bharti, A.K., He, W., Winter, P., Zhao, S., Hane, J.K., Carrasquilla-Garcia, N., Condie, J.A., Upadhyaya, H.D., Luo, M.-C., Thudi, M., Gowda, C.L.L., Singh, N.P., Lichtenzveig, J., Gali, K.K., Rubio, J., Nadarajan, N., Dolezel, J., Bansal, K.C., Xu, X., Edwards, D., Zhang, G., Kahl, G., Gil, J., Singh, K.B., Datta, S.K., Jackson, S.A., Wang, J., and Cook, D.R. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat Biotech* 31, 240-246.

10. Kagale, S., Chushin, K., Nixon, J., Bollina, V., Clarke, W.E., Tuteja, R., Spillane, C., Robinson, S.J., Links, M.G., Clarke, C., Higgins, E.E., Huebert, T., Sharpe, A.G., and Parkin, I.A.P. (2014). The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nat Commun* 5:3706.

11. Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon 1, Krishnakumar V, Bidwell SL, Denoeud F, Belcram H, Links MG, Just J, Clarke C, Bender T, Huebert T, Mason AS, Pires JC, Barker G, Moore J, Walley PG, Manoli S, Batley J, Edwards D, Nelson MN, Wang X, Paterson AH, King G, Bancroft I, Chalhoub B and Sharpe AG. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol* 15, R77.

12. Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376.

13. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for

high diversity species. *PLoS ONE* 6, e19379.

14. Wang, S., Meyer, E., McKay, J.K., and Matz, M.V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat Meth* 9, 808-810.

15. Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* 7, e37135.

16. Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., and Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Gen* 12, 499-510.

17. Deschamps, S., Llaca, V., and May, G.D. (2012). Genotyping-by-Sequencing in plants. *Biology* 1, 460-483.

18. Poland, J.A., Brown, P.J., Sorrells, M.E., and 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.

19. Edwards, D., Batley, J., and Snowdon, R. (2013). Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126, 1-11.

20. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. doi: 10.1093/bioinformatics/btu170.

21. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.

22. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760.

23. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.

24. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303.

25. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-498.

26. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., and DePristo, M.A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In Current Protocols in Bioinformatics (John Wiley & Sons, Inc.)

27. Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). Stacks: building and genotyping loci *de novo* from short-read sequences. *G3* 1, 171-182.

28. Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., and Buckler, E.S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9, e90346.

29. Dai, M., Thompson, R.C., Maher, C., Contreras-Galindo, R., Kaplan, M.H., Markovitz,

D.M., Omenn, G., and Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* 11 Suppl 4, S7.

30. Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., and Costich, D.E. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet* 9, e1003215.

31. Willing, E.M., Hoffmann, M., Klein, J.D., Weigel, D., and Dreyer, C. (2011). Paired-end RAD-seq for *de novo* assembly and marker design without available reference. *Bioinformatics* 27, 2187-2193.

32. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18, 1851-1858.

33. Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E., and Sahinalp, S.C. (2010). mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* 7, 576-577.

34. Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21, 936-939.

35. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-359.

36. Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009c). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966-1967.

37. Burrows, M., and Wheeler, D.J. (1994). A block-sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, 1994.

38. Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-

generation sequencing. *Brief Bioinform* 11, 473-483.

39. Huang, L., Popic, V., and Batzoglou, S. (2013). Short read alignment with populations of genomes. *Bioinformatics* 29, i361-370.

40. Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169-3177.

41. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res* 19, 1124-1132.

42. Wei, Z., Wang, W., Hu, P., Lyon, G.J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* 39, e132.

43. Clement, N.L., Snell, Q., Clement, M.J., Hollenhorst, P.C., Purwar, J., Graves, B.J., Cairns, B.R., and Johnson, W.E. (2010). The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* 26, 38-45.

44. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K., and Lyon, G.J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 5, 28.

45. Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics* 27, 1157-1158.

46. Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T., and Stern, D.L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic

mapping. *Genome Res* 21, 610-617.

47. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39, 906-913.

48. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84, 210-223.

49. Huang, B.E., Raghavan, C., Mauleon, R., Broman, K.W., and Leung, H. (2014). Efficient Imputation of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multi-Parental Crosses. *Genetics* 197, 401-404.

50. Robinson, M.R., Wray, N.R., and Visscher, P.M. (2014). Explaining additional genetic variation in complex traits. *Trends Genet* 30, 124-132.

51. Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet--next generation sequence assembly visualization. *Bioinformatics* 26, 401-402.

52. Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14, 178-192.

53. Marchini J, and Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11: 499–511.