Genotyping-by-Sequencing in Potato

15

Chiheb Boudhrioua, Maxime Bastien, Gaétan Légaré, Sonia Pomerleau, Jérôme St-Cyr, Brian Boyle and François Belzile

Abstract

Genotyping-by-sequencing (GBS) is one of the most cost-effective approaches to sequencing in potato (SNP) discovery and genotyping. The reduction of genome complexity that is central to the GBS approach is useful in the analysis of many plant genomes in which large size and polyploidy can prove challenging. In previous work in our lab, GBS has been explored and optimized on a tetraploid potato using two different enzymatic approaches (*Ape*KI and *Pst*I/*Msp*I) and two modes of genotyping (diploid and tetraploid) (Bastien et al., Submitted). This chapter describes the GBS process, starting with library preparation to sequencing data analysis and single nucleotide polymorphism (SNP) calling and filtering of GBS-derived markers. It also presents examples of the obtained results, an assessment of marker quality and their potential uses.

15.1 Introduction

Large genomes and polyploidy, either recent or due to more ancient events, are two factors that contribute to the complexity of genome analysis in many crop species. Potato, with a genome of ~800 Mb is not particularly large, but its autotetraploid nature does represent a challenge for many analyses (Barrell et al. 2013). With four allelic copies of each gene, there is the potential for more than two alleles at a given locus, there are five possible genotypic classes (AAAA, AAAB, AABB, ABBB and BBBB) for each locus and, because it is reproduced clonally, heterozygosity is very common, contrary to the situation in species in which varieties are fixed lines.

For these reasons, the development of efficient large-scale genotyping approaches in potato is somewhat lagging behind many other important crops. In terms of highly parallel and high-throughput assays, the two most commonly used

C. Boudhrioua · M. Bastien · F. Belzile (✉)
Département de Phytologie and Institut de Biologie Intégrative et Des Systèmes (IBIS), Université Laval, Quebec City, QC G1V 0A6, Canada
e-mail: francois.belzile@fsaa.ulaval.ca

G. Légaré · S. Pomerleau · J. St-Cyr · B. Boyle
Plateforme d'Analyses Génomiques, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC G1V 0A6, Canada

S. K. Chakrabarti et al. (eds.), *The Potato Genome*,
Compendium of Plant Genomes, https://doi.org/10.1007/978-3-319-66135-3_15

approaches in crops are genotyping arrays ("SNP chips") and next-generation sequencing of a selected fraction of the genome ("complexity reduction" methods) (Bajgain et al. 2016). To date, there have been two arrays developed for genotyping potato. The Infinium 8303 Potato Array (also known as the "SolCap array") was the first developed using sequencing data obtained from the transcriptomes of six varieties (Felcher et al. 2012). In total, this number of SNPs was distributed in about 4500 genes as, on average, 1.8 SNP markers/genes were included in the array design. More recently, a 20 K Infinium array (aka the "SolSTW array") was developed (Vos et al. 2015). The latter comprises 4454 SNPs from the SolCAP array as well as an additional 15,138 SNPs derived from the targeted sequencing of 807 genes (Uitdewellingen et al. 2013). It must be remembered, however, that even in relatively large sets of accessions, not all of these SNPs will prove informative. For example, of the over 8 K SNPs on the SolCAP array, only 3763 yielded a complete characterization of all five possible genotypes ("tetraploid mode") among a set of 250 potato clones (Hirsch et al. 2013). Similarly, for the SolSTW array, Vos et al. (2015) successfully called genotypes at slightly over 15 K SNPs among a large and diverse set of 569 potato clones.

Complexity reduction approaches typically rely on capturing a reproducible subset of the genome, usually through the use of restriction enzymes (Davey et al. 2011). Although there are slight differences in methodology, both genotyping-by-sequencing (GBS) and RAD-Seq rely on the sequencing of a set of restriction fragments of a given size (usually between 150 and 400 bp). In potato, two such complexity reduction approaches have been published to date. Uitdewellingen et al. (2013) used a somewhat atypical GBS approach in which DNA was fragmented and captured via in-solution hybridization using probes derived from selected genes. Sequencing the captured genomic segments from 84 potato accessions allowed the detection of close to 130 K variants located within a limited set of 807 genes. This represents an atypical GBS approach as it relied on sequence capture to restrict the sequencing effort to a non-random portion of the genome (i.e. 807 genic regions). In a recent RAD-Seq effort reported by Jiang et al. (2016), the authors sought to identify optimal enzyme combination leading to the minimization of chloroplast and rDNA sequences in their RAD-Seq libraries. The most favourable enzyme combination (*Eco*RI and *Msp*I) made it possible to call ~5 K informative SNPs in a set of 12 potato genotypes.

In our own GBS work, we have explored different enzyme combinations and determined the number, read depth and amount of missing data that result from these in potato (Bastien et al. submitted). In addition, we have examined how the chosen genotyping mode (diploid or tetraploid) affects the number of informative SNP markers obtained, as well as the ability to impute missing data among the resulting data sets. As described in what follows, we recommend the use of the *Ape*KI protocol in diploid SNP calling mode (i.e. AA, AB, BB) when there is a need to maximize the number of SNPs and genome coverage (e.g. for GWAS), albeit at the expense of a full resolution of the genotypic state. When it is important to benefit from a full characterization of the genotypic state (tetraploid mode) (e.g. for QTL mapping), we recommend a two-enzyme protocol (*Pst*I/*Msp*I); although it results in fewer informative SNP loci, each of these benefits from deep read coverage sufficient to call the full array of possible genotypes.

The description of the GBS approach provided here will be subdivided into five major procedures: (1) DNA extraction; (2) GBS library preparation; (3) sequencing; (4) GBS data analysis; and (5) further SNP filtering. We will conclude this chapter by an example in order to illustrate the results that can be obtained.

15.2 Materials

15.2.1 DNA Extraction

1. Reagents
   - DNeasy 96 plant kit, Qiagen or equivalent
   - Liquid nitrogen

2. Required lab equipment
   - Equipment for tissue grinding: TissueLyser
   - Water bath or heating block (65 °C)
   - Vortexer
   - Centrifuge with Plate Rotor 2 × 96 (max. 6000 rpm).

15.2.2 GBS Library Preparation and Sequencing

1. Oligonucleotides
   - The oligonucleotides used to prepare barcoded adapters are ordered as normal oligonucleotides at the 25-nM scale with standard desalting (to be shipped dried). Order oligonucleotides to prepare barcoded oligonucleotides in complementary plates, one for the top and one for the bottom strand. Having corresponding wells in two different plates makes the production of double-stranded adapters much easier (Tables 15.1 and 15.2).
   - The oligonucleotides used to prepare the common adapter are ordered as normal oligonucleotides at the 1-µmole scale with standard desalting. For each adapter, two oligonucleotides are ordered in complementary pairs and must be annealed to form the double-stranded adapter (Table 15.1).
2. Enzymes
   We purchase *Msp*I (R0106L), Hi-fidelity *Pst*I (R3140L), *Ape*KI (R0643L), T4 DNA ligase (M0202L) and Q5 High-fidelity polymerase (M0491L) from New England Biolabs.
3. Solutions
   - Elution buffer (EB): 10 mM Tris-Cl pH 8.0
   - 10X Annealing buffer (10X AB): 500 mM NaCl, 100 mM Tris-Cl pH 8.0
   - 80% ethanol freshly prepared
4. Other reagents
   - Qiaquick PCR Purification Kit or equivalent
   - Axygen PCR Clean Up kit or equivalent
   - Quant-iT Picogreen dsDNA assay kit or equivalent
5. Required lab equipment
   - Thermocycler
   - Magnet for magnetic bead purification
   - BluePippin or Pippin prep
   - Bioanalyzer or equivalent
   - Ion Proton sequencer
6. Recommended lab equipment
   - Mix mate
   - Repeater stream with advanced combitips
   - Ion CHEF

Table 15.1 Oligonucleotide sequences

Oligonucleotide	Sequence
Top barcoded oligo *Pst*I	5'-CCCTGCGTGTCTCCGACTCAG-**[Barcode]**-GATTGCA
Bottom barcoded oligo *Pst*I	5'-ATC-**[Barcode Reverse Complement]**-CTGAGTCGGAGACACGCAGGG
Top common adapter *Msp*I	5'-CGAGATCGGAAGAGCGGGGAGCTTAAGC
Bottom common adapter *Msp*I	5'-CCTCTCTATGGGCAGTCGGTGATCCCGCTCTTCCGATCT
Top barcoded oligo *ApeK*I	5'-CCCTGCGTGTCTCCGACTCAG-**[Barcode]**-GAT
Bottom barcoded oligo *ApeK*I	5'-CWGATC-**[Barcode Reverse Complement]**-CTGAGTCGGAGACACGCAGGG
Top common adapter *ApeK*I	5'-CWGAGATCGGAAGAGCGGGGAGCTTAAGC
Bottom common adapter *ApeK*I	5'-CCTCTCTATGGGCAGTCGGTGATCCCGCTCTTCCGATCT
Ion forward PCR primer	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG
Ion reverse PCR primer	5'-CCACTACGCCTCCGCTTTCCTCTCTCTATGGGCAGTCGGTGAT

Table 15.2 Barcode sequences

CTAAGGTAAC	TCTATTCGTC	TCGCATCGTTC	CGGACAATGGC	TTCCTACCAGTC	CTAGGACATTC
TAAGGAGAAC	AGGCAATTGC	TAAGCCATTGTC	TTGAGCCTATTC	TCAAGAAGTTC	CTTCCATAAC
AAGAGGATTC	TTAGTCGGAC	AAGGAATCGTC	CCGCATGGAAC	TTCAATTGGC	CCAGCCTCAAC
TACCAAGATC	CAGATCCATC	CTTGAGAATGTC	CTGGCAATCCTC	CCTACTGGTC	CTTGGTTATTC
CAGAAGGAAC	TCGCAATTAC	TGGAGGACGGAC	TCCACCTCCTC	TGAGGCTCCGAC	TTGGCTGGAC
CTGCAAGTTC	TTCGAGACGC	TAACAATCGGC	**CAGC**ATTAATTC	CGAAGGCCACAC	CCGAACACTTC
TTCGTGATTC	TGCCACGAAC	CTGACATAATC	TCTGGCAACGGC	TCTGCCTGTC	TCCTGAATCTC
TTCCGATAAC	AACCTCATTC	TTCCACTTCGC	TCCTAGAACAC	CGATCGGTTC	CTAACCACGGC
TGAGCGGAAC	CCTGAGATAC	AGCACGAATC	TCCTTGATGTTC	TCAGGAATAC	CGGAAGGATGC
CTGACCGAAC	TTACAACCTC	CTTGACACCGC	TCTAGCTCTTC	CGGAAGAACCTC	CTAGGAACCGC
TCCTCGAATC	AACCATCCGC	TTGGAGGCCAGC	TCACTCGGATC	CGAAGCGATTC	CTTGTCCAATC
TAGGTGGTTC	TCGACCACTC	TGGAGCTTCCTC	TTCCTGCTTCAC	**CAGC**CAATTCTC	TCCGACAAGC
TCTAACGGAC	CGAGGTTATC	TCAGTCCGAAC	CCTTAGAGTTC	CCTGGTTGTC	CGGACAGATC
TTGGAGTGTC	TCCAA**GCTGC**	TAAGGCAACCAC	CTGAGTTCCGAC	TCGAAGGCAGGC	TTAAGCGGTC
TCTAGAGGTC	TCTTACACAC	TTCTAAGAGAC	TCCTGGCACATC	CCTGCCATTCGC	TTCGCAATGAAC
TCTGGATGAC	TTCTCATTGAAC	TCCTAACATAAC	CCGCAATCATC	TTGGCATCTC	TTCCGCACGGC
CGAAGGCCACAC	TTGGAGGCCAGC	TTGGCCAATTGC	TCTAGTTCAAC		

15.3 Methods

15.3.1 DNA Extraction

1. High molecular weight genomic DNA is extracted from 50 mg (fresh weight) of young leaves. If used fresh, tissues need to be frozen (with liquid nitrogen) just prior to sample grinding. More conveniently, leaf samples (cuttings, punches) are dried directly in wells/Eppendorf tubes in the presence of silica gel. Grinding is performed either with small disposable plastic pestles in Eppendorf tubes or using a mixer mill for 96-well plates, in which case one tungsten bead is included in each well containing leaf tissue. DNA of the highest purity can be obtained using a commercial kit, but CTAB-based protocols can also be used successfully.
2. The DNA concentration (ng/μL) of each sample is measured with a spectrophotometer (Nanodrop 1000, Fisher Scientific) for samples devoid of RNA contamination (prepared with a kit). For samples obtained with CTAB-based protocol, we may have some residual RNA. In the latter case, using a fluorometric quantification method (e.g. PicoGreen) may prove more precise (see note 1). A total of 200 ng per sample is used for the preparation of the GBS libraries.

15.3.2 GBS Library Preparation

This part will consist of (1) common and barcoded adapter preparation (see note 2); (2) complexity reduction using enzymes; and (3) multiplexing using barcoded adapters. The described protocol is largely inspired from the original procedure developed in the Poland Lab (Poland et al. 2012). We have mainly optimized and improved the procedure over time. In what follows, we will describe a "standard" procedure based on 96-plex library preparation and sequencing (see note 3).

1. Double-stranded barcoded adapter preparation (Stock BC adapter plate—0.1 μM final)
   - Re-suspend dried single-stranded oligonucleotides to 100 uM in EB.
   - In a PCR plate, make 100 μl of 10 μM double-stranded barcoded adapters by mixing:
     - 10 μL of top single-stranded oligo at 100 μM
     - 10 μL of bottom single-stranded oligo at 100 μM
     - 10 μL of 10X AB
     - 70 μL of H_20
   - Seal the plate, mix using a mixmate, then spin down.
   - In a thermocycler, heat to 95 °C for 1 min, then cool down to 30 °C at the rate of 1 °C per minute, then hold at 4 °C. (see note 4).
   - Dilute 1/10 using 1X AB (see note 5).
   - Repeat step 4 once to bring barcoded adapters to 0.1 μM.
2. Common adapter preparation (10 μM final):
   - Re-suspend dried single-stranded oligonucleotides to 100 μM in EB.
   - In a PCR plate, make 100 μL of 10 μM double-stranded common adapter by mixing:
     - 10 μL of top single-stranded oligo at 100 μM
     - 10 μL of bottom single-stranded oligo at 100 μM
     - 10 μL of 10X AB
     - 70 μL of H20
   - Seal the plate, mix with mixmate, then spin down.
   - In a thermocycler, heat to 95 °C for 1 min and then cool at the rate of 1 °C per minute, then hold at 4 °C.
3. Make working adapter plates:
   - Each well in the working adapter plates will have 0.02 μM of a unique barcoded adapter and 1 μM of the common adapter.
   - In a 96-well plate, add:
     - 20 μL barcoded adapters at 0.1 μM (from 1)

- 10 μL common adapter at 10 μM (from 2)
- 10 μL 10X AB
- 60 μL water

– Mix well and spin down.

4. Normalize DNA and prepare sample plates:
   – Quantify sample genomic DNA (see note 1).
   – Prepare sample plates so each well contains 10 μL of DNA at a 20 ng/μL concentration (200 ng total). These plates will be used directly for further steps so ensure they are compatible with available thermocyclers.
5. Restriction digest:
   This protocol uses a double-digest with *Pst*I and a second enzyme *Msp*I. Barcoded adapters will be ligated to the *Pst*I overhang while the common adapter will be ligated to the *Msp*I overhang (see note 6).
   – To each well of the sample plates prepared in 4 add (see note 7):
     - 3 μL CutSmart buffer (supplied with NEB restriction enzymes)
     - 5 units *Pst*I HiFi
     - 5 units *Msp*I
     - Complete to 30 μL with water
   – Mix well and spin down.
   – Incubate in a thermocycler at 37 °C for 2 h, then hold at 8 °C (see note 8).
   – Proceed immediately with adapter ligation.
6. Ligate adapters to cut genomic DNA:
   The ligation is carried out directly in the same reaction plate without the need for reaction clean-up.
   – To each well of the restriction digest plates prepared in 5, add (see note 9):
     - 5 μL of 10X T4 DNA ligase reaction buffer (supplied with T4 DNA ligase)
     - 400 units of T4 DNA ligase
     - 5 μL from the corresponding well of the working adapter plate prepared in step 3 (see note 10)
     - Complete to 50 μL with water
   – Mix well, spin down, and incubate at 22 °C for 2 h, then 65 °C for 20 min and hold at 8 °C when completed (see note 11).
7. Pool and clean samples:
   – Pool 5 μL from 48 reaction wells into a 1.7 mL tube (columns 1–6).
   – Repeat step 1 for the other 48 reaction wells (columns 7–12).
   – Add 1.2 mL of Qiagen PB buffer to each 1.7 mL tube.
   – Mix well using a vortex and spin down.
   – Load 750 μL on a Qiaquick column.
   – Spin for 15 s.
   – Discard flow-through.
   – Repeat steps 5–7 until the complete volume from the two tubes has been loaded to the column.
   – Wash column with 750 μL of PE, spin 1 min, discard flow-through.
   – Rotate column and spin 1 min to remove all traces of PE.
   – Transfer column in a new 1.7 mL tube.
   – Add 30 μL of EB to the center of the column, let stand for 1 min., then spin 1 min to elute the pooled library.
8. Size the library using a BluePippin:
   – Add 10 μL of BluePippin buffer to the eluted library from 7.
   – Follow BluePippin instructions for loading on a 2% cassette (BEF2010).
   – We set elute times from 46–60 min.
   – You should retrieve about 50–60 μL per library that would be sufficient for multiple PCR reactions.
9. PCR amplification and enrichment:
   Appropriate primers complementary to the ligated adapters are added and PCR is performed to amplify the pool of restriction fragments (see note 12).
   – For each library prepare the amplification mix:
     - 22.9 μL of water
     - 10 μL of 5X Q5 buffer
     - 10 μL of Q5 enhancer solution
     - 1 μL of 10 mM Dntp
     - 0.3 μL of 10 μM FWD IonExpress Primer
     - 0.3 μL of 10 μM REV IonExpress Primer
     - 5 μL of DNA from step 3.2.8
     - 0.5 μL of Q5 polymerase

- Mix well and spin down.
- Run the following PCR Program:
  - 75 °C for 5 min.
  - 5 cycles of:
    98 °C 10 s
    55 °C 30 s
    72 °C 30 s
  - 7 cycles of:
    98 °C 10 s
    65 °C 30 s
    72 °C 30 s
  - 72 °C 5 min
  - Hold at 4 °C
- Add 50 µL of Axygen PCR clean-up kit and mix well, transfer to a 1.5 mL tube.
- Let stand for 5 min at room temperature.
- Put on magnet for 2 min.
- Remove the liquid without disturbing the magnetic beads.
- While keeping the tube on the magnet, wash the pellet twice with 1 mL of freshly prepared 80% ethanol.
- Remove all traces of ethanol and let dry for 10–15 min.
- Remove from magnet.
- Re-suspend dried beads in 30 µL of EB, let stand for 2 min.
- Put on magnet and wait 5 min for beads to pellet.
- Transfer your eluted library to a new tube. Be careful not to carry over beads.

10. Quality control:
    - We perform a Nanodrop quantification right after purification. Expect between 5 and 20 ng/µL.
    - The most important quality control is a Bioanalyzer trace (or equivalent). High quality libraries will look like Bart Simpson's hairdo, meaning relatively sharp edges with spikes on top. There should be no primer dimers located around 100–110 nt. Background after 400 bases should be flat. A large camelback hump from 500–2000 bases is indicative of PCR over-cycling (Fig. 15.1).
    - Quantify the library with PicoGreen or equivalent (see note 13). Convert concentration from ng/µL to nM (see note 14). Dilute library to 200 pM.

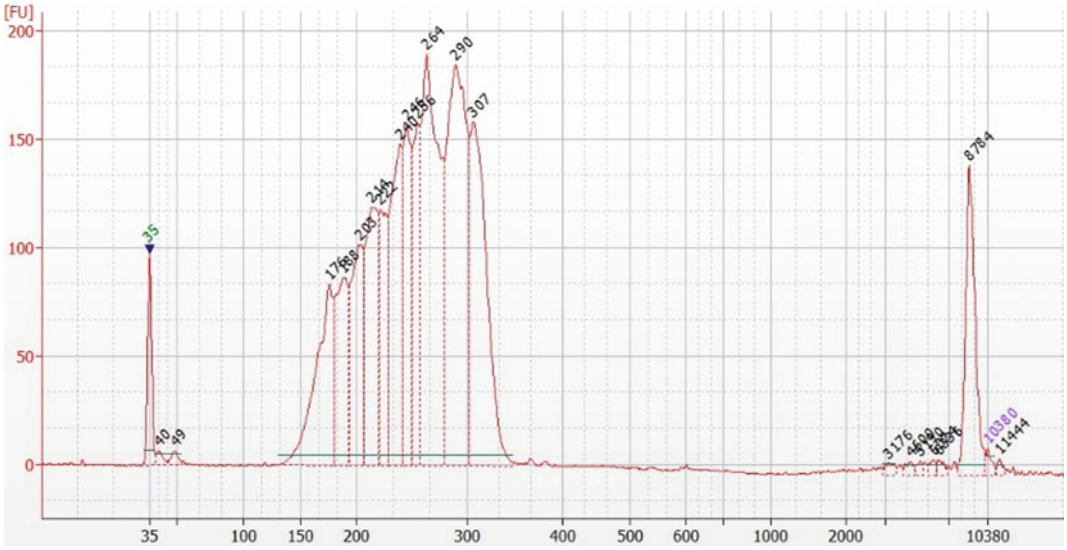

Fig. 15.1 An example of a high quality library

15.3.3 Sequencing

Typically, each 96-plex library is sequenced on a single Ion PI chip yielding >70 M reads with a median length of 140–160 bp. If deeper coverage is required, the same library can be loaded onto additional chips to provide a larger number of reads per sample.

1. Load Ion CHEF and perform Ion Proton Sequencing:
   The sequencing reaction will proceed from the barcoded adapter. Follow the manufacturer's instructions to load the Ion CHEF and Ion Proton Sequencer.
   – Load 25 μL of a 200 pM GBS library. Our experience has shown that it generates good sequencing runs (Fig. 15.2).
   – Run the FastqCreator plugin when the sequencing run is completed to generate the fastq file.
   – Compress the fastq file using gzip to move the data from the Ion Server to the data analysis server.

15.3.4 GBS Data Analysis

For this analysis, two different modes can be used to call variants: a diploid mode or a tetraploid mode (see note 15). This step is carried out using various bioinformatics tools (SABRE, BWA, PLATYPUS, VCFtools, etc.) included in the Fast-GBS pipeline (https://bitbucket.org/jerlar73/fastgbs). For calling SNPs in tetraploid mode, an additional software, Freebayes (https://github.com/ekg/freebayes), is needed.

Prior to feeding the fastq files into the pipeline, check the quality (see note 16) of the raw sequences using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) or Galaxy (https://usegalaxy.org/).

1. Diploid mode:
   – As explained in the Fast-GBS page, we first need to create four directories: **refgenome**, **data**, **barcodes** and **results** and put the appropriate files in the first three: (i) the reference genome with the companion index file; (ii) the raw Fastq sequences in compressed format (.gz); and (iii) the barcode sequences with the corresponding sample name.
   – Using the appropriate parameter file, the Fast-GBS pipeline is run. Default options can be used, however, one can change them depending on the nature of the data. Also, some basic filtrations are included in the pipeline by default:
   • Minimum read length to keep: 50 nucleotides
   • Minimum size of bam file (per sample): 3000 kilobytes

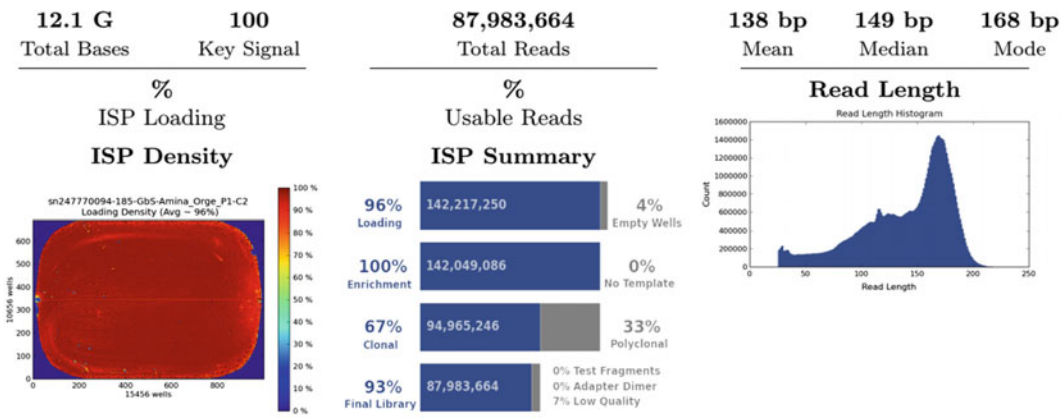

Fig. 15.2 A run summary for one chip Ion Torrent (Proton)

- Sequencing depth (minimum number of reads supporting a variant): two reads
- Maximum amount of missing data tolerated per locus: 80%
- The Fast-GBS variant file (.vcf) is stored in the **results** directory already created. By default, with Fast-GBS, genotypes are called in diploid mode.

2. Tetraploid mode:
   – As in diploid mode, the same steps are used to run Fast-GBS but only to generate the alignment files (.s4.bam) needed for the rest of the analysis.
   – Using the alignment files (.s4.bam) as an input, genotype calls in tetraploid mode are conducted using the default parameters of Freebayes. The ploidy level must be set to four and a minimum of three reads supporting an alternate allele is required to call a polymorphism.

15.3.5 Further SNP Filtering

Using VCFtools (http://vcftools.sourceforge.net/) and an in-house script (see note 17), quality filters are applied to the raw variants file in order to select SNPs of superior quality. In potato, it depends on the enzymatic approach, the genotype mode and the eventual use of these variants. The main filters used are:

– Preserve SNP markers only, i.e. eliminate indels.
– A filter based on the number of reads supporting each genotype call. When using the *Ape*KI GBS protocol, 11 reads are required to call a genotype in either diploid or tetraploid mode. Using *Pst*I/*Msp*I, a more stringent filter can be applied in tetraploid mode with 11 reads supporting a homozygote and 53 reads per heterozygous genotype.
– A filter based on the proportion of missing data. The thresholds for missing data can be fixed between 10–20%. The choice of threshold depends on the population and the number of markers needed. For example, if we are studying a panel of cultivated potato, linkage disequilibrium (LD) can be much shorter compared to a population derived from a biparental cross, thus more markers are needed.
– The minor allele frequency (MAF) is the frequency of the less common allele in a population. The choice of the MAF threshold will depend on the nature of your population and eventual use of the data. To describe population structure and kinship, we may be interested in keeping even rare alleles as these may refine the relationships between lines. In such a case, we can use a MAF as low as 1%. For an association panel, we more typically use minimal MAF values between 5 and 10%.

15.3.6 Example

To illustrate the type of results obtained with such a protocol, we applied GBS on two sets of potato germplasm. To first compare the efficacy of two different GBS library preparation protocols (*Ape*KI and *Pst*I/*Msp*I), we used a small set of 8 clones (Set A). In a second stage, we genotyped a much larger collection (Set B) of 375 clones representing the extent of diversity present in a public potato breeding program in the province of Quebec in Canada.

1. DNA was extracted from 50 mg of fresh young leaves using the DNeasy 96 Plant kit (Qiagen). DNA concentrations were normalized to 20 ng/μl and subsequently used for library preparation.
2. Eight potato samples (Set A) were sequenced as part of a 48-plex GBS library. The GBS libraries were prepared with both the *Ape*KI and *Pst*I/*Msp*I enzymes.
3. For set B (375 clones), three 96-plex and one 87-plex *Ape*KI libraries were prepared.
4. For all libraries, single-end sequencing was performed on an Illumina HiSeq 2000. Since this initial work, we have adopted the Ion

Torrent sequencing technology, and the protocols described herein are for this type of sequencing.

5. For Set A, sequencing yielded approximately 19.1 million reads and 19.6 million reads in total with *Ape*KI and *Pst*I/*Msp*I respectively. About 72.1% and 75.4% of the reads were successfully mapped to the potato reference genome v.4.03.
6. These reads obtained after preparing and sequencing GBS libraries with two different restriction-enzyme combinations (*Ape*KI and *Pst*I/*Msp*I) were used to call genotypes in either diploid or tetraploid mode. We kept only SNPs with fewer than 12.5% missing data. SNPs with a minor allele frequency below 10% (diploid) or a minor allele count (tetraploid) below three were also removed (Fig. 15.3).
7. In diploid mode, 11 reads were required to keep a genotype with the two restriction-enzyme combinations. Markers were then filtered on the percentage of missing data from 0 to 50% (Table 15.3). In general, *Ape*KI yielded 2.5 to 3 times more markers than *Pst*I/*Msp*I but these genotype calls were based on 2 to 3 times fewer reads.
8. In tetraploid mode, a stringent filter was applied with 11 reads required to support a homozygous call and 53 reads needed to distinguish the three heterozygous genotype classes. Markers were then filtered according to the percentage of missing data, from 0–50% (Table 15.4). This filter considerably reduced the number of markers obtained using *Ape*KI compared to *Pst*I/*Msp*I. These results are due to the fact that this enzyme cuts frequently in the genome, thereby increasing the number of loci examined and concomitantly reducing the number of reads supporting a genotype call at each locus. Thus, under conditions used in this study, *Pst*I/*Msp*I would be recommended over *Ape*KI to call genotypes in tetraploid mode, while the *Ape*KI protocol maximizes the the number of markers that can be called in diploid mode.
9. A total of 670.3 million reads obtained for GBS libraries prepared for Set B (375 lines) were used to call markers in diploid mode. After removing genotype calls supported by fewer than 11 reads, indels and markers with more than 20% missing data, 42,786 markers were left. Markers having a minor allele frequency below either 1% or 5% were also eliminated, yielding respectively 22,545 and 15,424 SNPs.
10. To assess the accuracy of GBS-derived genotype calls, we selected 52 lines from Set B. Among these, 126 markers with diploid genotype calls for these lines were in

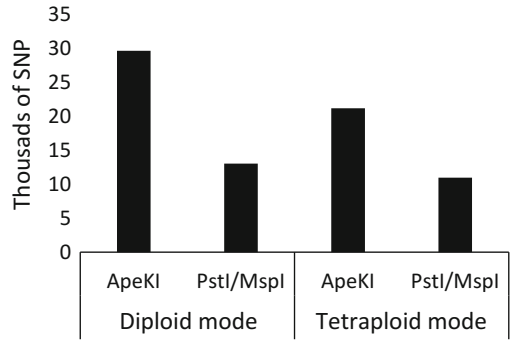

Fig. 15.3 Number of informative SNP markers obtained among eight potato genotypes using two restriction-enzyme combinations and two SNP-calling modes. In all cases, SNP markers with fewer than 12.5% missing data and with a minor allele frequency above 10% (diploid) or a minor allele count (tetraploid) above three were kept

Table 15.3 Number of markers and depth of coverage per scored genotype in diploid mode as a function of the percentage of missing data

% missing data	*Ape*KI		*Pst*I/*Msp*I	
	Number of markers	Mean read depth per genotype	Number of markers	Mean read depth per genotype
0	27,263	33	15,615	112
≤ 12.5	40,631	30	18,682	103
≤ 25	51,943	28	20,961	98
≤ 50	74,308	26	24,817	92

Table 15.4 Number of markers and depth of coverage per scored genotype in tetraploid mode as a function of the percentage of missing data

% missing data	*Ape*KI		*Pst*I/*Msp*I	
	Number of markers	Mean read depth per genotype	Number of markers	Mean read depth per genotype
0	199	98	6024	170
≤ 12.5	461	78	7335	157
≤ 25	753	70	8133	150
≤ 50	1621	62	9148	144

common with the Infinium 8 K array. Comparison between the two approaches was conducted on this data set and showed a match rate of 90.4% between the two genotyping approaches.

11. A phylogenetic tree for these 52 lines was created based on 15,202 high-quality markers with diploid genotype call and a minor allele frequency above 5% (Fig. 15.4). The tree showed that clones belonging to the same market classes tended to group together.
12. SNP catalogues obtained via a GBS approach can be used for several analyses and applications in potato such as association analysis, analysis of genetic diversity and structure (Bastien et al., submitted).

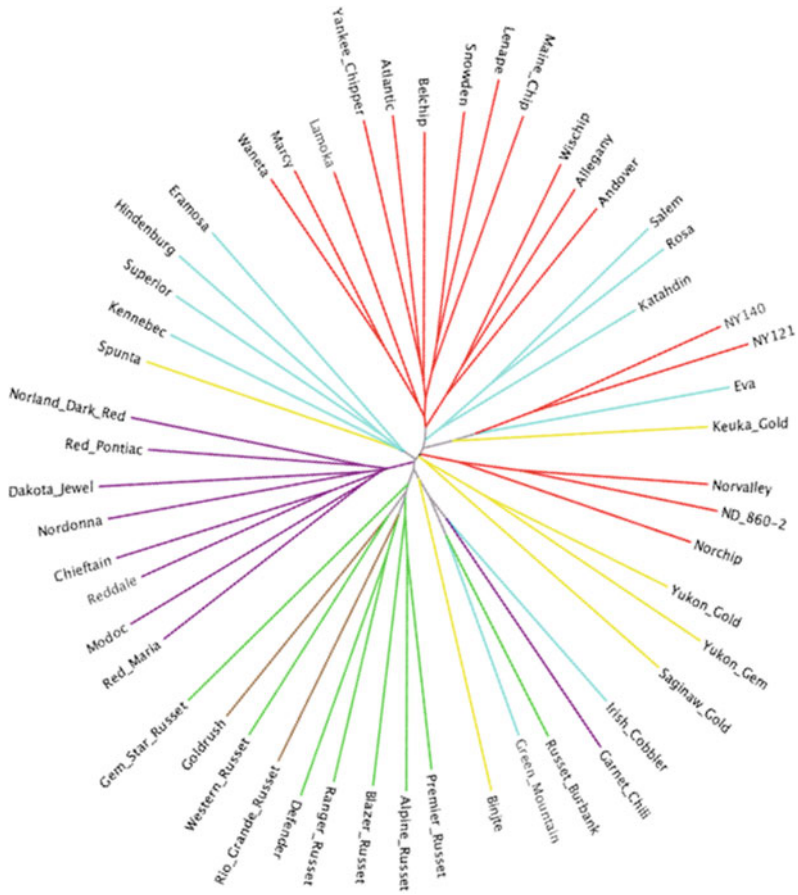

Fig. 15.4 Neighbour-Joining tree of 52 potato lines based on 15,202 SNP markers with diploid genotype calls. Color-coding reflects major market classes (red: chip processing; purple: pigmented; green: French fry processing; light blue: round white table; brown: table Russet; yellow: yellow flesh)

Notes

1. DNA concentration and quality are critical in order to produce a stable number of sequence tags from each sample. It is recommended that DNA be quantified using a florescence-based quantification method such as PicoGreen and Qubit. DNA quality can be assessed using a spectrophotometer and the observation of the 260/230 and 260/280 ratios which should be above 1.7. Do not necessarily throw away DNAs that do not meet the highest standards, ensure that they are well quantified and they might just work.
2. Table 15.1 lists primer sequences for both the *Ape*KI and *Pst*I/*Msp*I procedures. Use appropriate primers. Table 15.2 lists adapter sequences, adapter sequences containing bold characters are not suitable for the *Ape*KI procedure, therefore additional barcode sequences were added.
3. Different levels of multiplexing can be used: 48-plex, 96-plex, 192-plex and 384-plex. The choice will depend on the depth of coverage you want to achieve (decreases with increased multiplexing) and the budget you have (cost per sample decreases with increasing multiplexing). For Illumina library preparation, follow the procedures described either in Elshire et al. (2011) (*Ape*KI) or Poland et al. (2012) (*Pst*I/*Msp*I). To improve data quality, add a size-selection step using a BluePippin

apparatus (step 8 from our procedure) with time settings from 50–65 min because Illumina adapters are longer than Ion proton adapters.

4. Annealed oligonucleotides at a 10 μM concentration are very stable and can be stored at −20 °C indefinitely.
5. The original procedure recommended that adapters should be quantified after annealing to ensure that the double-stranded DNA formation was complete and they are at the correct concentration. Uniform concentration of adapters was believed critical to producing uniform numbers of reads between samples when sequencing the multiplexed library. We have not observed significant differences between wells and no longer perform quantification at this stage. Uniformity could be linked to the choice of the oligonucleotide provider.
6. For the *Ape*KI single digest, barcoded adapters will be ligated to one end of fragments while the common adapter will be ligated to the other end. Only those fragments will amplify at the PCR stage. Fragments with other combinations (barcoded-barcoded or common-common) will be lost. To each well of the sample plates prepared in step 4 add (see note 5):
   - 3 μL NEB 3.1 buffer (supplied with NEB *Ape*KI)
   - 5 Units *Ape*KI
   - Complete to 30 μL with water.
   - Mix well and spin down.
   - Incubate in a thermocycler at 75 °C for 2 h, then hold at 8 °C (see note 5).
   - Proceed immediately with adapter ligation.
7. It is easier to prepare a master mix (buffer, enzymes and water), then add 20 μL of it to each well of the samples plates. Prepare at least an extra 10% of the master mix. We use an Eppendorf stream repeater with 1 mL combitips advanced to distribute the mastermix.
8. The original procedure had a 20 min at 80 °C step to heat-inactivate the restriction enzymes. *Pst*I HiFi, *Msp*I and *Ape*KI cannot be heat-inactivated so this step is not required. Also note that adapters, by design, do not contain sites for restriction enzymes used and once ligated to a matching end, they are designed not to be recleaved.
9. It is easier to prepare a master mix (buffer, enzymes and water), then add 15 μL of it to each well of the samples plates. Prepare at least an extra 10% of the master mix. We use an Eppendorf stream repeater with 1 mL combitips advanced to distribute the mastermix. Remember that adapters must be added separately.
10. The original procedure called for adjusting adapter concentration depending on the species. We have used the specified concentrations of adapters with over 100 species covering a large portion of the life kingdom that includes fungi, insects, plants and animals without a single adjustment. However, restriction enzyme combinations might not be optimal for all species and this becomes particularly true when the restriction enzymes hit highly repeated elements, in this case, changing the restriction enzyme combinations is a better choice than trying to adjust the concentration of adapters. Also note that it is essential that the common adapter is added to at least 20-fold excess compared to the barcoded adapter to cover the difference in cut frequency between *Pst*I and *Msp*I.
11. Completed ligation can be safely stored at −20 °C.
12. Only fragments that have ligated adapters to both a *Pst*I cut-site and an *Msp*I cut-site will amplify. Keep the number of PCR cycles low to avoid undetectable PCR duplication events. It is better to perform multiple PCR reactions to increase yield rather than increasing the number of PCR cycles. We routinely perform three or four PCR reactions per GBS library. It is highly recommended to physically isolate pre-PCR and post-PCR operations to prevent contamination.
13. It is important to quantify libraries using a standardized methodology as this measurement will be used to load the precise amount of molecules on the sequencing instrument. Therefore, ensure that the methodology is

sensitive and falls well within the linear quantification range.

14. To convert ng/μL DNA concentration to nM: [nM DNA] = DNA concentration (ng/μL) x 10^6 (μL/L)/(Sample fragment size in bp x 656.4 (g/mole)).
15. A diploid model defines three marker classes for each SNP (AA, AB and BB); a tetraploid model has five marker classes (AAAA, AAAB, AABB, ABBB and BBBB).
16. The scoring system of each base is known as the Phred score. This score ranges from 0–64.
17. In-house scripts must be used to filter variants in tetraploid mode according to number of reads supporting the genotype and the minor allele frequency (MAF). Filtering according to missing data is possible using VCFtools.

References

Bajgain P, Rouse MN, Anderson JA (2016) Comparing Genotyping-by-Sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. Crop Sci 56(1):232–248

Barrel PJ, Meiyalaghan S, Jacobs JME, Conner AJ (2013) Applications of biotechnology and genomics in potato improvement. Plant Biotechnol J 11(8):907–920

Bastien M, Boudhrioua C, Fortin G, Belzile F Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. Submitted to Genome.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genetics 12:499–510

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(5):e19379

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CR, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS ONE 7(4):e36347

Hirsch CN, Hirsch CD, Felcher K, Coombs J, Zarka D, Van Deynze A, De Jong W, Veilleux RE, Jansky S, Bethke P, Douches DS, Buell CR (2013) Retrospective view of north American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. G3 (Bethesda):1003–1013

Jiang N, Zhang F, Wu J, Chen Y, Hu X, Fang O, Leach LJ, Wang D, Luo Z (2016) A highly robust and optimized sequence-based approach for genetic polymorphism discovery and genotyping in large plant populations. Theor Appl Genet 129:1739–1757

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink JL (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome 5(3):103–113

Uitdewilligen JGAML, Wolters A-MA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. PLoS ONE 8(5):e62355

Vos PG, Uitdewilligen JGAML, Voorrips RE, Visser RGF, van Eck HJ (2015) Development and analysis of a 20 K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. Theor Appl Genet 128:2387–2401