Proteomics & Bioinformatics

Fu and Dong, J Proteomics Bioinform 2015, 8:8
http://dx.doi.org/10.4172/jpb.1000368

Research Article Open Access

paSNPg: A GBS-Based Pipeline for Protein-Associated SNP Discovery and Genotyping in Non-Model Species

Yong-Bi Fu* and Yibo Dong

Plant Gene Resources of Canada, Saskatoon Research Centre, Agriculture and Agri-Food Canada, Saskatoon, Canada

Abstract

Genotyping-by-sequencing (GBS) has recently developed as a feasible genomic approach for exploring genome-wide genetic variation for population and evolutionary genomic analyses of non-model species. To facilitate the acquisition of function-associated genetic variation data in natural populations, we present a GBS-based pipeline called paSNPg for protein-associated SNP (paSNP) discovery and genotyping in non-model organisms. The pipeline was developed through the expansion of the published npGeno utility to assemble nuclear contigs from raw GBS sequence data, separate protein-associated contigs from all assembled contigs based on published PEP (or Predictions on Entire Proteomes) sequence data sets, and call paSNPs across assayed samples based on protein-associated contigs. Testing the pipeline with two GBS sequence data sets, *Arabidopsis thaliana* and *Oryza sativa*, revealed its potential use in exploring protein-associated genetic variation from genomic DNAs of non-model species.

Keywords: Genotyping-by-sequencing; Non-model species; SNP discovery; SNP genotyping; Function-associated genetic marker; Protein-associated SNP.

Abbreviations: GBS: Genotyping-By-Sequencing; NE: Nuclear Exon; Nu: Nuclear; NGS: Next Generation Sequencing; paSNP: Protein-associated SNP; PEP: Predictions on Entire Proteomes; SNP: Single Nucleotide Polymorphism

Introduction

There is a growing interest in acquisition of function-associated genetic marker data from natural populations for population and evolutionary genomic analyses of non-model species [1-4]. The function-associated markers sample the coding regions of a genome and are more sensitive and informative to characterize adaptive genetic diversity patterns and investigate their associations with ecological factors than the commonly applied amplified fragment length polymorphism (AFLP) markers that largely are selectively neutral [5,6]. Large efforts have been currently directed toward the development of function-associated single nucleotide polymorphism (SNP) through transcriptomic analyses such as RNA-seq technology [4,7]. However, these expressed polymorphisms are dependent on the assayed materials and their developmental stages [1,3,4].

Genotyping-by-sequencing (GBS) has recently developed as a feasible genomic approach for exploring genome-wide genetic variation for population genomic analysis of non-model species [8-11], thanks to the advances in next generation sequencing (NGS) [12]. The GBS approach is a combined one-step process of SNP marker discovery and genotyping through genome reduction with restriction enzymes [13] and SNP calls without a reference genome [8,9,14,15]. This approach has displayed a major advantage of being rapid, high throughput, and cost-effective for a genome-wide analysis of genetic variability in a range of non-model species [14-16]. To our knowledge, however, little attention has been made to separate function-associated SNPs from GBS SNP data and no specific computational pipelines are currently available for acquiring function-associated SNPs from GBS sequence data [17].

Recently we have developed a genetic diversity focused GBS (gd-GBS) protocol for plant genetic diversity analysis [18]. It uses two restriction enzymes to reduce genome complexity, applies Illumina multiplexing indexes for sample barcoding, and has a custom bioinformatics pipeline called npGeno for SNP genotyping. The npGeno pipeline takes sample fastq input, constructs contigs from sequence reads from all samples, calls SNPs using the constructed contigs as a reference, filters resulting SNPs, and formats data outputs. However, the pipeline does not separate nuclear SNPs from exon regions of a nuclear genome, and the resulting nuclear SNP markers could be either selective (i.e., function-associated) or neutral.

Here we present a new GBS-based pipeline called paSNPg for the discovery and genotyping of protein-associated SNPs (paSNPs) from GBS sequence data. The pipeline was developed through the modification of the existing npGeno utility and the addition of new computational steps to identify protein-associated contigs based on published PEP (i.e., Predictions for Entire Proteomes) [19] sequence data for paSNP discovery and genotyping. Specifically here, we will describe the paSNPg pipeline, provide an empirical test on the performance of the pipeline using the GBS sequence data sets of two model plants, and discuss the advantages and limitations of paSNPg in its applications. It is our hope that this GBS-based pipeline will serve as a useful tool for generating paSNP data from genomic DNAs, rather than RNAs expressed in specific tissues of certain developmental stage, in natural populations for population and evolutionary genomic analyses of non-model organisms.

Pipeline description

Motivation: The paSNPg pipeline was initiated when an effort was

***Corresponding author:** Yong-Bi Fu, Plant Gene Resources of Canada, Saskatoon Research Centre, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada, Tel: 1-306-385-9298; Fax: 1-306-385-9489; E-mail: yong-bi.fu@agr.gc.ca

Received July 02, 2015; **Accepted** July 27, 2015; **Published** Aiugust 01, 2015

Citation: Fu YB, Dong Y (2015) paSNPg: A GBS-Based Pipeline for Protein-Associated SNP Discovery and Genotyping in Non-Model Species. J Proteomics Bioinform 8: 190-194. doi:10.4172/jpb.1000368

made to identify and separate nuclear SNPs that are selective from those SNPs generated by the npGeno pipeline for a genetic diversity analysis [18]. The original thought was to identify function-associated contigs through a Blast2GO analysis [20] and to use those identified function-associated contigs as reference for function-associated SNP calls. This approach was found to be effective, but it requires a separate, lengthy analysis with Blast2GO, making the acquisition of function-associated SNPs less user-friendly. Thus, we adopted the Blast2GO approach using specific taxon's PEP sequence data sets available to identify protein-associated contigs and expanded the existing npGeno utility to create a new independent pipeline paSNPg that allows for automatic acquisition of paSNPs from GBS sequence data.

Overview of the pipeline: The complete pipeline package is illustrated in supplementary Figure S1A and has five defined subdirectories (Input_data, Output_results, Pep_database, Scripts, Threshold_set), three shell scripts (paSNPg.sh, minpaSNPg.sh, remove_missingSNPs.sh) and a guide documentation (Getting Started with paSNPg.pdf; Figure S1B). The functional view of the pipeline is shown in Figure 1 with input, computational, and output components. Four input files are necessary for running paSNPg: Fastq input, Pep_database, Pident_Plength.txt and Missing_threshold.txt. GBS sequence data and PEP sequence data located in Input_data are the primary input, while two .txt files are provided to define the running parameters for contig alignment and missing data. All custom shell and Perl scripts are resided in the Scripts subdirectory.

The pipeline is automatically run with a shell script paSNPg.sh that controls two major pairs of computational shell scripts: generate_nuclear/exon_contigs.sh and call_nuclear/exon_SNP.sh. The first shell script pair generate_nuclear/exon_contigs.sh was developed to perform de novo assembly of nuclear (Nu) contigs using Minia [21] based on collapsed reads with Fastx_collapser [22] from fastq files of all samples and generates Nu_contigs.fasta. The Perl script select_exon_contigs.pl was written to (1) identify nuclear exon (NE) contigs NE_contigs.fasta from Nu_contigs.fasta by using the BLAST algorithm [23] to compare the sequences against available PEP data set(s) and using matching criteria Pident_plength.txt and (2) output NE-contigs information associated with PEP into NE_contigs-information.txt. The second script pair call_nuclear/exon.SNP.sh was developed to identify and generate SNP genotype data for all samples using Bowtie2 [24] and SAMtools [25] and based on either Nu_contigs.fasta or NE_contigs.fasta. The Perl script screen_sampleSNP.pl analyzes outputs from Bowtie2 and SAMtools and obtains SNP genotypes of each sample. The identify_SNP.pl script identifies SNPs and generates all SNP genotype data from every sample for the whole nuclear genomes and for the exon regions only. This script generates four output files associated with Nu_SNP* or NE_SNP*. Note that all NE_SNP* files are the output for paSNPs. The fasta_format.pl script transforms N*_SNP_hap.txt data to N*_SNP_hap.fasta files. The formatted fasta files can be used by other software such as PGDSpider [26] for format conversions required for other specific population genomic analyses.

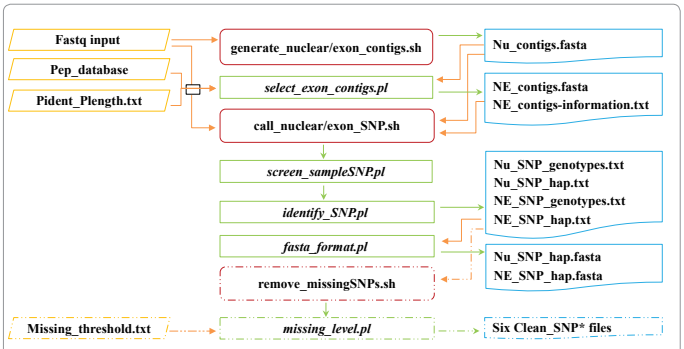

Figure 1: Flow chart of the paSNPg pipeline for protein-associated SNP discovery and genotyping. The pipeline has input components on the left, computational components in the middle, and output components on the right. The brown arrows reflect the input and green narrows the output. The dash lines show the separate run of some components if needed. Explanations for each component and file are given in the text.

The optional remove_missingSNPs.sh script serves as cleaning GBS SNPs with a given level of missing observations across the assayed samples through the missing_level.pl script. Specifically, it removes SNPs with a missing level given in Missing_threshold.txt or higher across the samples and outputs three Clean_NE_SNP and three Clean_Nu_SNP text files into the Output_results subdirectory. The optional script minpaSNPg.sh performs a modified version of the npGeno utility for automatic SNP genotyping from GBS sequence data without acquisition of protein-associated contigs and paSNPs.

Availability and requirements: The complete pipeline package is placed on online supplementary file (Figure S1A) and is available for download. It was developed for use on a Linux operating system, as it is dependent on a number of freely available Unix-like programs. They are (1) Minia (http://minia.genouest.org/); (2) Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml); (3) SAMtools v0.1.18 (http://samtools.sourceforge.net/); (4) Perl in Linux (http://www.perl.org/get.html); (5) Fastx_collapser (http://hannonlab.cshl.edu/fastx_toolkit/); and (6) Blast+ (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download). These programs need to be downloaded from the sources mentioned above and installed in a Linux server following their respective documented installation instructions, including the setting of their proper execution paths. Thus, the access to a computing facility such as Linux servers and basic operational skills in a Linux environment are also required for its use.

Another requirement is the search for and acquisition of published PEP sequence data for specific paSNPg application. The steps used to obtain the plant PEP data sets from public PEP databases are illustrated in Figure S2 and these steps can be applied to select and acquire PEP sequence data sets of various taxa close or related to the species of interest. The acquired individual pep.all.fa files should be compressed using the tar command as Pep_database.tar.bz2 and placed in the subdirectory Input_data before running the pipeline.

Application: The pipeline was mainly designed to analyze GBS sequence data with a de-multiplexed set of paired-end fastq files that were generated through Illumina MiSeq or HiSeq instruments from multiple samples of one or more populations. Such marker data acquisition is compatible with most marker data sets acquired in natural populations for population or evolutionary genomic analyses.

The steps required to copy the pipeline into a Linux server and install six dependent, free software and the procedures to operate the pipeline are described in the user guide Getting Started with paSNPg.pdf. Extra effort is needed to upload the GBS sequence data with a de-multiplexed set of paired-end fastq (or fastq.gz format) files into the subdirectory Input_data. To start the pipeline, run the shell script paSNPg.sh by typing: ./paSNPg.sh at the command prompt and nine output files will be generated and placed in the subdirectory Output_results. It is recommended to run the remove_missingSNPs.sh script with a defined level of missing data to obtain related clean SNP data for

further analysis. The running time is largely dependent on the extent of GBS sequence data, and the major computational effort is placed on nuclear contig assembly and identification of protein-associated contigs.

The pipeline is run with several conservative default settings. First, Minia uses default options with kmer_size of 100 bp and min-abundance of 80% sample size to identify reliable contigs. Second, the default settings for the percentages of identical matches and alignment length defined in Pident-Plength.txt are 75% and 99%, respectively. Third, Missing_threshold.txt is used to remove the SNP loci having a level of missing observations or higher. A level of 10-20% is recommended, but the default setting is 0%.

Pipeline testing

We performed an empirical evaluation on the pipeline using the GBS sequence data sets of two model plants *A. thaliana* and *O. sativa*, as paSNPs can be obtained from such data sets using PEP sequence data of its own species or those of other related taxon for an informative comparison. The test GBS data for these plant species were generated following exactly the gd-GBS protocol [18]. Briefly, 12 *Arabidopsis* races and 12 cultivated rice accessions were obtained from various sources (Table 1), their seeds were grown in a greenhouse and young leaf tissue was separately collected from a single seedling of each *Arabidopsis* race and rice accession, and total genomic DNA was extracted and quantified. Further steps were performed on DNA samples: preparation with digestion of two restriction enzymes *Pst*I and *Msp*I, library assembly for the barcoding and pooling of 24 DNA samples, and sequencing on a MiSeq instrument. The sequencing output and barcoding information for each sample are summarized in Table 1.

The paSNPg pipeline was run with default settings in a Linux server on each GBS sequence data set, considering the availability of different PEP sequence data sets. Searching public PEP databases revealed 38 plant species with released pep.all.fa files (Figure S2B). To assess the potential use of the pipeline for non-model species, three PEP sequence data sets were defined for each model species: Pep1 stands for the pep.all.fa data obtained from its own species, Pep2 for all pep.all.fa data available from its closely related species within a genus, and Pep3 for all pep.all.fa data available from distantly related genera, families, and taxa. Specifically for the test with *Arabidopsis* GBS data, Pep2 had only the Arabidopsis_lyrata.v.1.0.23.pep.all.fa file and Pep3 had 36 other plant pep.all.fa files (Figure S2B). For the test with rice, Pep1 had both *Oryza_sativa* and *Oryza_indica* pep.all.fa files, Pep2 consisted of eight other Oryza*.pep.all.fa files and Pep3 had 28 other plant pep.all.fa files (Figure S2B). A total of six separate runs of paSNPg were made; each run for *Arabidopsis* or rice GBS data consumed about 10 or 13 hours, respectively; and 85-90 percent of the computational time was on Nu contig assembly and NE contig identification. The test outcomes of paSNPs with missing data are shown in Table 1 and Figure 2, and

Sample	I7_index	I5_index	Raw reads	Filter reads	paSNPs [a]			Heterozygous paSNPs		
					Pep1	Pep2	Pep3	Pep1	Pep2	Pep3
Arabidopsis										
Col0	CAGATC	AGTCAA	583757	514127	202	190	130	8	8	6
Col1	ACTTGA	AGTCAA	686519	602112	207	191	134	4	3	2
Col2	GATCAG	AGTCAA	588852	518482	202	187	129	3	0	0
Col3	CTTGTA	AGTCAA	727682	641888	208	194	139	3	3	1
Col4	GGCTAC	GTAGAG	420552	372465	196	180	127	8	8	5
Col5	TGACCA	GTAGAG	759351	675564	211	196	131	3	3	0
Col6	ACAGTG	GTAGAG	606352	539896	195	189	135	4	3	2
Col7	GCCAAT	GTAGAG	647679	576263	211	196	140	2	2	1
Bur0	CAGATC	GTAGAG	589549	525992	159	159	121	1	5	5
Tsu1	ACTTGA	GTAGAG	669306	593685	204	177	135	2	5	7
LER	GATCAG	GTAGAG	637718	566801	204	178	133	0	4	4
WS4	CTTGTA	GTAGAG	522965	464454	186	171	121	7	10	8
Mean			620024	549311	199	184	131	4	5	3
Rice										
R1120	GGCTAC	CTTGTA	468561	406905	719	706	328	65	66	33
R971	TGACCA	CTTGTA	581679	507890	692	682	331	78	79	42
R286	ACAGTG	CTTGTA	471639	411051	661	652	324	44	40	20
R242	GCCAAT	CTTGTA	552139	479092	689	663	320	59	51	25
R237	CAGATC	CTTGTA	425615	370887	714	708	345	64	51	32
R614	ACTTGA	CTTGTA	628054	543484	738	725	337	52	42	26
R423	GATCAG	CTTGTA	420596	365303	700	703	339	62	55	28
R1662	CTTGTA	CTTGTA	482911	419658	672	662	334	56	57	30
R1409	GGCTAC	AGTCAA	419646	366602	700	690	330	45	37	25
R1570	TGACCA	AGTCAA	544887	474002	683	674	353	36	32	26
R735	ACAGTG	AGTCAA	560799	486134	622	612	282	38	31	19
R163	GCCAAT	AGTCAA	469506	409684	700	690	336	64	62	28
Mean			502169	436724	691	681	330	55	50	28

[a] Pep1 stands for the PEP sequence data set from its own species, Pep2 for the PEP sequence data set(s) from its closely related species within a genus, and Pep3 for the PEP sequence data sets from distantly related genera, families, and taxa. Total paSNPs obtained using Pep1, Pep2 and Pep3 for 12 *Arabidopsis* samples are 229, 203 and 146; for 12 rice samples 802,789 and 370, respectively.

Table 1: List of *Arabidopsis thaliana* and *Oryza sativa* test samples, MiSeq sequencing barcodes and reads, and paSNP counts obtained using the paSNPg pipeline based on three different sets of published PEP data.

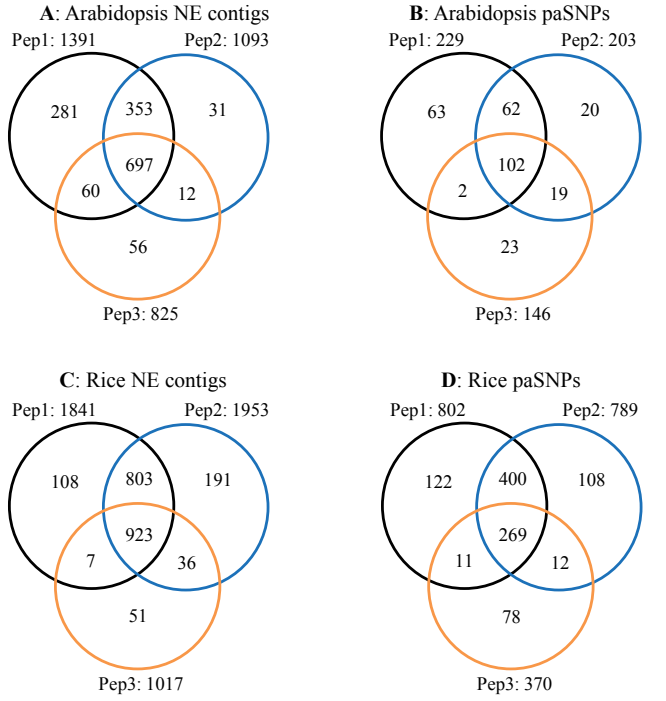

Figure 2: Venn diagrams showing the counts of nuclear exon (NE) contigs and paSNPs obtained through the application of the paSNPg pipeline to the *Arabidopsis* and rice MiSeq data using three different published PEP data sets. Pep1 stands for the PEP sequence data set from its own species, Pep2 for the PEP sequence data set(s) available from its closely related species within a genus, and Pep3 for the PEP sequence data sets available from distantly related genera, families, and taxa. Total nuclear contigs and SNPs with missing data obtained from *Arabidopsis* or rice GBS data were 8,307 and 1,126 or 20,738 and 5,438, respectively.

more results with 0, 10, and 20 percent of missing data are given in supplementary Table S1.

The pipeline testing with 12 *Arabidopsis* samples revealed a total of 8,307 nuclear contigs. Among those contigs, 1,391 (16.7%) were identified as NE contigs using Pep1, 1,093 (13.2%) using Pep2, and 825 (9.9%) using Pep3 (Figure 2A). Using Pep2 and/or Pep3 as for non-model species, one would find at least 697 (8.4%) NE contigs that were identified by using Pep1. For SNP discovery, a total of 1,126 nuclear SNPs with missing data were found, and 229 (20.3%) of them were identified as paSNPs using Pep1, 203 (18%) using Pep2, and 146 (13%) using Pep3 (Figure 2B). Using Pep2 and/or Pep3 as for non-model species would yield more than 102 (9%) paSNPs that were discovered by using Pep1. The counts of sample-wise paSNPs and heterozygous SNPs for *Arabidopsis* GBS data are given in Table 1.

The analysis of 12 rice samples revealed a total of 20,738 nuclear contigs. Among those contigs, 1,841 (8.9%) were identified as NE contigs using Pep1, 1,953 (9.4%) using Pep2, and 1,017 (4.9%) using Pep3 (Figure 2C). Using Pep2 and/or Pep3 as for non-model species would yield at least 923 (4.5%) NE contigs that were identified by using Pep1. For SNP discovery, a total of 5,438 nuclear SNPs with missing data were detected, and 802 (14.7%) of them were identified as paSNPs using Pep1, 789 (14.5%) using Pep2, and 370 (6.8%) using Pep3 (Figure 2D). Using Pep2 and/or Pep3 as for non-model species would yield more than 269 (4.9%) paSNPs that were discovered by using Pep1. The counts of sample-wise paSNPs and heterozygous SNPs for rice GBS data are given in Table 1.

To assess the extent of NE contigs and their paSNP loci that may be genuine, we mapped all the NE contigs to their reference genome sequences [27,28] using Bowtie2 with default settings and found that more than 95% NE contigs were mapped to the reference genomes. For *Arabidopsis*, there were 1,376 (98.9%) mapped NE contigs obtained using Pep1, 1080 (98.8%) using Pep2, and 811 (98.3%) using Pep3. For rice, there were 1,770 (96.1%) mapped NE contigs obtained using Pep1, 1866 (95.6%) using Pep2, and 994 (97.7%) using Pep3. The un-mapped NE contigs may be artificial from Minia assembly errors or genuine if the reported reference genome has gaps or errors. Further analysis of the un-mapped NE contigs revealed that 10 *Arabidopsis* and 13 rice un-mapped NE contigs were all identified by using Pep1, Pep2, or Pep3, implying these shared un-mapped NE contigs were also likely genuine. Together, we can reason that more than 99.5% *Arabidopsis* and 96.2% rice NE contigs identified by using Pep2 and/or Pep3 as for non-model species may be genuine. To determine if these NE contigs are real, however, requires further empirical assessments through Sanger sequencing.

These test outcomes demonstrate not only the effectiveness of the pipeline in acquiring paSNPs from model plant GBS sequence data, but also the potential utility in exploring paSNPs from a non-model organism by using published PEP sequence data of distant genus and taxon, as illustrated using Pep2 and Pep3. Clearly, having PEP sequence data from more closely related species or less distant taxa will increase the extent of paSNP acquisition. Also, the amount of paSNPs obtained is dependent on the species analyzed, as nearly three-fold more paSNPs were identified from rice, than *Arabidopsis*, GBS sequence data. The outcome of more paSNPs in rice may largely reflect its three-fold larger genome size and/or more efficient *Pst*I and *Msp*I digestions in monocot species.

Discussion

We made an effort to search the literature with the hope to acquire published GBS sequence data sets for further evaluation of paSNPg, but failed to obtain suitable GBS sequence data. Undoubtedly, the current version of the pipeline is preliminary without extensive real data tests and may carry some issues and/or limitations. First, the pipeline is not fully optimized for computational efficiency, particularly for larger GBS sequence data sets, as longer computation time is expected for contig assembly and separation. For a large GBS data set (more than 150 samples), one may consider to apply the npGeno utility [18] to assemble nuclear contigs by dividing the GBS sequence data set into subsets of smaller size (normally 10-15 GB/set) without a compromise of contig reliability. Second, the acquisition of paSNPs is highly dependent on the availability of published PEP sequence data and may not be optimal, as all of the paSNPs present in assayed samples may not be fully extracted from GBS SNP data. Currently, little is known about the efficiency of the pipeline in paSNP genotyping based on the variable nature of the PEP sequence data. Third, no efforts have been made yet to assess paSNPg's applicability with GBS sequence data generated from other NGS platforms and modifications may be needed for inputting data from other NGS platforms. Lastly, although the pipeline is technically effective for non-model diploid species, its application to other polyploidy species may generate bias in paSNP acquisition and needs to be evaluated further.

We are, however, confident that this pipeline will facilitate research efforts in acquiring protein-associated SNP data from genomic DNAs, rather than RNAs expressed in specific tissues of certain developmental stage, in natural populations for various population genomic analyses

of non-model species for the following reasons. First, the pipeline is automatic and user-friendly, and it can generate paSNP data from GBS sequence data without resorting to a separate Blast2Go call for function-associated contigs. Second, its computational time for function-associated contigs using only PEP data from related taxa is much faster than the separate Blast2Go analysis. Third, although our pipeline was tested on plant species, it is also applicable to any other non-model diploid organisms such as animal species for paSNP acquisition by including published PEP data of its related species or taxa. Forth, this pipeline allows for the detection and quantification of function-associated genetic variability from genomic DNAs, not RNAs expressed in specific tissues of certain developmental stage, in natural populations through the GBS approach without performing more complicated transcriptomic or RNA-seq analyses. The major advantage of the former lies mainly in the independence of developmental stage and the avoidance of possible sampling bias in natural populations. Fifth, our pipeline generates both nuclear SNP and paSNP genotype data from GBS sequence data that will facilitate the investigations on the role of neutral versus adaptive genetic variations in the genomic inferences of evolutionary processes.

Supplementary Information

1. A compressed file paSNPg.zip for the complete paSNPg pipeline package.
2. A pdf file consisting of the supporting Table S1, Figure S1, and Figure S2.

Acknowledgements

The authors would like to thank Gregory Peterson and Carolee Horbach for their technical assistance in generation of MiSeq data for the pipeline testing; Raju Datla, Daoquan Xiang, Gopalan Selvaraj, Lily Tang for providing the *Arabidopsis* testing materials; Hesham Agrama and Wenhui Yan for providing the cultivated rice testing materials; Matthew Links and Frank You for their assistance with access to Linux servers for bioinformatics analysis; and Gregory Peterson, Morgan Kirzinger and three anonymous journal reviewers for their helpful comments on an early version of the manuscript. This research was financially supported by an A-Base research project of Agriculture and Agri-Food Canada to YBF.

Authors' Contributions

YBF conceived the research, designed the pipeline, generated and analyzed the testing data, and wrote the manuscript. YD co-designed the pipeline, developed and tested the pipeline, analyzed the testing data, and contributed to the manuscript writing.

References

1. Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, et al. (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol Ecol Resour 11: 1-8.
2. DeFaveri J, Viitaniemi HM, Leder EH, Merilä J (2013) Characterizing genic and non-genic molecular markers: comparison of microsatellites and SNPs. Mol Ecol Res 13: 377-392.
3. Manel S, Holderegger R (2013) Ten years of landscape genetics. Trends Ecol Evol 28: 614-621.
4. De Wit P, Pespeni MH, Palumbi SR (2015) SNP genotyping and population genomics from expressed sequences – current advances and future possibilities. Mol Ecol 24: 2310–2323.
5. Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, et al. (2012) Adaptive genetic variation on the landscape: methods and cases. Annu Rev Ecol Evol Syst 43: 23-43.
6. Grover A, Sharma PC (2015) Development and use of molecular markers: Past and present. Critical Rev in Biotechnol. 28: 1-13.
7. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.
8. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2015) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379.
9. Fu YB, Peterson GW (2011) Genetic diversity analysis with 454pyrosequencing and genomic reduction confirmed the eastern and western division in the cultivated barley gene pool. Plant Genome 4: 226-237.
10. Grabowski PP, Morris GP, Casler MD, Borevitz JO (2014) Population genomic variation reveals roles of history, adaptation and ploidy in switchgrass. Mol Ecol 23: 4059–4073.
11. Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, et al. (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evol Appl 7: 355-369.
12. Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11: 31-46.
13. 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.
14. Peterson B, Weber JN, Kay EH, Fisher HS, Hoekstra H (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7: e37135.
15. 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.
16. Fu YB, Cheng B, Peterson GW (2014) Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. Genet Resour Crop Evol 61: 579-594.
17. Parisod C, Holderegger R (2012) Adaptive landscape genetics: Pitfalls and benefits. Mol Ecol 21: 3644-3646.
18. Peterson GW, Dong Y, Horbach C, Fu YB (2014) Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. Diversity 6: 665-680.
19. Carter P, Liu J, Rost B (2003) PEP: Predictions for Entire Proteomes. Nucleic Acids Res 31: 410-413.
20. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.
21. Chikhi R, Rizk G (2012) Space-efficient and exact de Bruijn graph representation based on a Bloom filter, WABI. In: Algorithms in Bioinformatics, XIII (eds Raphael B and Tang J), pp 263-248. Springer Verlag.
22. http://hannonlab.cshl.edu/fastx_toolkit/index.html
23. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10: 421.
24. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.
25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079.
26. Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28: 298–299.
27. Meinke DW, Cherry JM, Dean C, Roundsley SD, Koorneef M (1998) *Arabidopsis thaliana*: A model plant for genome analysis. Science 282: 679-682.
28. Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, et al. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res 36: D1028–D1033.

Citation: Fu YB, Dong Y (2015) paSNPg: A GBS-Based Pipeline for Protein-Associated SNP Discovery and Genotyping in Non-Model Species. J Proteomics Bioinform 8: 190-194. doi:10.4172/jpb.1000368