

(51) International Patent Classification:
G02C 3/02 (2006.01)

(21) International Application Number:

PCT/US20 13/0202 11

(22) International Filing Date:

4 January 2013 (04.01.2013)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/582,861 4 January 2012 (04.01.2012) US

(71) Applicant: DOW AGROSCIENCES LLC [US/US];
9330 Zionsville Road, Indianapolis, IN 46268 (US).

(72) Inventors: BUYARAPU, Ramesh; 6103 Eagles Nest Blvd, Zionsville, IN 46077 (US). TANG, Shunxue; 3442 Burlingame Blvd, Camel, IN 46074 (US). ARORA, Kanika; 2903 Spring Meadow Ct, Indianapolis, IN 46268 (US). ELANGO, Navin; 8385 Chadwood Lane West Drive; Apt. IB, Indianapolis, IN 46268 (US). KUMPATLA, Siva P.; 14524 Baldwin Lane, Carmel, IN 46032 (US). MARRI, Pradeep; 14304 Camden Lane, Carmel, IN 46074 (US). TANG, Jennifer Changhong; 4901 S. Cobblestone Dr., Zionsville, IN 46077 (US). MCEWAN, Robert; 8645 Malaga Drive, Apt ID, Indiana-

polis, IN 46250 (US). EVANS, Clive; 9860 East State Road 47, Sheridan, IN 46069 (US). PARLIAMENT, Kelly; 844 Bridle Circle, Carmel, IN 46032 (US).

(74) Agent: LEE, Yung-Hui; Dow AgroSciences LLC, 9330 Zionsville Rd, Indianapolis, Indiana 46268 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

[Continued on nextpage]

(54) Title: HAPLOTYPE BASED PIPELINE FOR SNP DISCOVERY AND/OR CLASSIFICATION

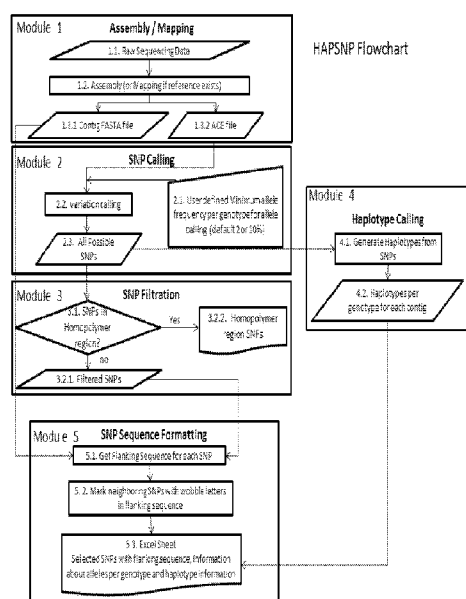


Figure 1

(57) Abstract: This invention is related to systems and methods for discovery and/or classification of single-nucleotide polymorphism (SNP) markers. The SNP sequences identified and/or classified using the systems and methods disclosed can be useful for phenotype or trait association studies. In particular, a haplotype based pipeline for SNP discovery and/or classification (HAPSNP) is provided and the disclosed systems and methods can be especially useful for polyploid and complex plant genomes.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG). **Published:**

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

Declarations under Rule 4.17:

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(Hi))*

HAPLOTYPE BASED PIPELINE FOR SNP DISCOVERY AND/OR CLASSIFICATION

FIELD OF THE INVENTION

[0001] This invention is generally related to the field of bioinformatics, and more specifically the field of discovery and/or classification of single nucleotide polymorphism (SNP) markers in an organism.

BACKGROUND OF THE INVENTION

[0002] Single nucleotide polymorphism (SNP) markers have become markers of choice for marker assisted selection (MAS) in crop improvement programs because of their higher abundance, amenability for automation and availability of high throughput genotyping platforms. However, current methodology for identifying SNPs in plants has many limitations, including a very high rate of false positives. This problem is especially challenging for plants with complex genomes. Thus, there remains a need for methodology which can identify and/or classify SNPs efficiently and accurately.

SUMMARY OF THE INVENTION

[0003] This invention is related to systems and methods for discovery and/or classification of single-nucleotide polymorphism (SNP) markers. The candidate SNP markers identified and/or classified using the systems and methods disclosed can be useful for phenotype or trait association studies. In particular, a haplotype based pipeline for SNP discovery and/or classification (HAPSNP) is provided and the disclosed systems and methods can be especially useful for polyploid and complex plant genomes.

[0004] In one aspect, provided is a computerized system for discovery and/or classification of single nucleotide polymorphism (SNP) markers in an organism. The system comprises:

- (a) an input device and an output device/interface;
- (b) an analysis system interface coupled to memory of a computer;
- (c) an operating system optionally comprising a database;
- (d) a controller module for SNP calling or Loci identification; and
- (e) a filtration engine for removing unreliable SNPs.

[0005] In one embodiment, the system further comprises at least one of alignment module, assembly/mapping module, haplotype calling module, and SNP sequence formatting module. In a further or alternative embodiment, the system comprises all of

assembly/mapping module(s), haplotype calling module, and SNP sequence formatting module. In another embodiment, the system further comprises a SNP marker classification module. In one embodiment, the system comprises an automatic sequencer or DNA sequencing machine. In another embodiment, the input device is selected from the group consisting of automated sequencer, sequencing data input device, and sequencing data storage device. In another embodiment, the output interface comprises a list of candidate SNP markers.

[0006] In one embodiment, the database described herein contains information selected from the group consisting of SNPs, contigs with at least one SNP, and haplotypes with at least two SNPs. In another embodiment, the filtration engine for removing unreliable SNPs comprises at least four SNP filters to remove unreliable SNPs and generate reliable SNPs. In another embodiment, the filtration engine for removing unreliable SNPs comprises at least five SNP filters to remove unreliable SNPs and generate reliable SNPs. In another embodiment, the filtration engine for removing unreliable SNPs comprises at least six SNP filters to remove unreliable SNPs and generate reliable SNPs. In another embodiment, the filtration engine for removing unreliable SNPs comprises at least five SNP filters to remove unreliable SNPs and generate reliable SNPs.

[0007] In one embodiment, the assembly/mapping module converts raw sequence data into contig FASTA files and/or ACE files. In another embodiment, the haplotype calling module generates haplotype data by examining patterns of SNP loci across contigs. In another embodiment, the SNP sequence formatting module generates candidate SNP markers with flanking sequences. In a further or alternative embodiment, the candidate SNP markers have a validation success rate of at least or greater than 60%. In a further embodiment, the candidate SNP markers have a validation success rate of from 60% to 80%. In a further embodiment, the candidate SNP markers have a validation success rate of about 75%.

[0008] In another embodiment, the computerized system provided comprises a controller module for SNP calling, and further comprises an assembly/mapping module, a haplotype calling module, and a SNP sequence formatting module. In another embodiment, the computerized system provided comprises a controller module for Loci identification, and further comprising an alignment module and a SNP sequence formatting module.

[0009] In another aspect, provided is a method for use in a computerized system for discovery and/or classification of single nucleotide polymorphism (SNP) markers in an

organism. The method comprises:

- (a) assembling/mapping sequence data using an assembly/mapping module;
- (b) identifying all possible SNPs using a SNP calling module; and
- (c) generating reliable SNPs using a SNP filtration module.

[0010] In one embodiment, the method further comprises determining haplotype using at least one haplotype calling module. In a further or alternative embodiment, the method further comprises formatting candidate SNP markers using at least one SNP sequence formatting module.

[0011] In one embodiment, the computerized system of the method comprises a system described herein. In another embodiment, the method provides candidate SNP markers having a validation success rate of at least or greater than 60%. In a further embodiment, the candidate SNP markers have a validation success rate of from 60% to 80%. In a further embodiment, the candidate SNP markers have a validation success rate of about 75%. In a further or alternative embodiment, the method provides candidate SNP markers having a validation success rate of at least two folds as compared to a publicly available program. In another embodiment, the method provides candidate SNP markers having a validation success rate of at least one and half folds (*i.e.*, at least 50% increase) as compared to a publicly available program. In a further embodiment, the publicly available program is QualitySNP. The QualitySNP program is disclosed in Tang *et al.*, BMC Bioinformatics 7:438 (2006), the content of which is incorporated in its entirety.

[0012] In one embodiment, the system or method disclosed provides that the candidate SNP markers are classified into at least two types using at least one SNP marker classification module. In a further embodiment, the candidate SNP markers are classified into at least three types. In another embodiment, the system or method disclosed provides that at least one type of the candidate SNP markers with at least or greater than 60% validation success rate. In a further embodiment, the candidate SNP markers have a validation success rate of from 60% to 80%. In a further embodiment, the candidate SNP markers have a validation success rate of about 75%. In another embodiment, the system or method disclosed provides that at least one type of the candidate SNP markers with at least or greater than two folds validation success rate as compared to a publicly available program. In a further embodiment, the publicly available program is QualitySNP.

[0013] In some embodiments, the organism for the systems or methods described herein

comprises a polyploid genome. In some embodiments, the organism is a plant. In some embodiments, the plant is selected from the group consisting of cotton, canola, corn, soybean, sunflower, and wheat. In another embodiment, the method provided further comprises classifying candidate SNP markers using a SNP marker classification module. In a further embodiment, at least one type of the candidate SNP markers has a validation success rate of at least 60%. In a further embodiment, at least two types of the candidate SNP markers have a validation success rate of at least 60%.

[0014] In another aspect, provided is a method for use in a computerized system for discovery and/or classification of single nucleotide polymorphism (SNP) markers in an organism in combination with a system for genotyping-by-sequencing (GBS). The method comprises:

- (a) aligning genotyping-by-sequencing (GBS) data using an alignment module;
- (b) generating reliable SNPs using a SNP filtration module; and
- (c) formatting candidate SNP markers using a SNP sequence formatting module.

[0015] In one embodiment, the computerized system of the method comprises a system described herein. In another embodiment, the method provided further comprises identifying SNP loci using a Loci identification module. In another embodiment, the method provides candidate SNP markers having a validation success rate of at least or greater than 60%. In a further embodiment, the candidate SNP markers have a validation success rate of from 60% to 80%. In a further embodiment, the candidate SNP markers have a validation success rate of about 75%. In a further or alternative embodiment, the method provides candidate SNP markers having a validation success rate of at least two folds as compared to a publicly available program. In another embodiment, the method provides candidate SNP markers having a validation success rate of at least one and half folds (*i.e.*, at least 50% increase) as compared to a publicly available program. In a further embodiment, the publicly available program is QualitySNP. The QualitySNP program can be obtained from the world wide website bioinformatics.nl/tools/snpweb as disclosed in Tang *et al.*, BMC Bioinformatics 7:438 (2006), the content of which is incorporated in its entirety.

[0016] In one embodiment, the system or method disclosed provides that the candidate SNP markers are classified into at least two types using at least one SNP marker classification module. In a further embodiment, the candidate SNP markers are classified into at least three types. In another embodiment, the system or method disclosed provides

that at least one type of the candidate SNP markers with at least or greater than 60% validation success rate. In a further embodiment, the candidate SNP markers have a validation success rate of from 60% to 80%. In a further embodiment, the candidate SNP markers have a validation success rate of about 75%. In another embodiment, the system or method disclosed provides that at least one type of the candidate SNP markers with at least or greater than two folds validation success rate as compared to a publicly available program. In a further embodiment, the publicly available program is STACKs.

[0017] In some embodiments, the organism for the systems or methods described herein comprises a polyploid genome. In some embodiments, the organism is a plant. In some embodiments, the plant is selected from the group consisting of cotton, canola, corn, soybean, sunflower, and wheat. In some embodiments, the plant is *G. hirsutum*, *G. barbadense*, or *G. mustelinum*.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] Figure 1 shows an exemplary embodiment of the HAPSNP pipeline provided herein. Five modules of the system are illustrated: (1) Assembly/mapping; (2) SNP calling; (3) SNP filtration; (4) Haplotype calling; and (5) SNP sequence formatting.

[0019] Figure 2 shows an exemplary system provided herein.

[0020] Figure 3 shows exemplary input sequences from raw sequencing data.

[0021] Figure 4 shows an exemplary output screen shot after the assembly/mapping module.

[0022] Figure 5 shows an exemplary output screen for possible SNPs after the SNP calling module.

[0023] Figure 6 shows an exemplary output screen for homopolymer region SNPs after the SNP filtration module.

[0024] Figure 7 shows an exemplary output screen for filtered SNP after the SNP filtration module.

[0025] Figure 8 shows an exemplary output screen after the haplotype calling module.

[0026] Figure 9 shows an exemplary output screen after the SNP sequence formatting module.

[0027] Figure 10 shows an exemplary output screen after both the haplotype calling and the SNP sequence formatting module.

[0028] Figure 11 shows an example of Type I, II, and III SNPs in cotton identified using

an exemplary system and method provided herein. Figure 11A shows a typical distribution of Type I SNPs; Figure 11B shows a typical distribution of Type II SNPs; Figure 11C shows a typical distribution of Type III SNPs.

[0029] Figure 12 shows an exemplary embodiment of the HAPSNP pipeline provided to be combined with genotyping-by-sequencing (GBS). Four modules of the system are illustrated: (1) Alignment; (2) Loci identification; (3) SNP filtration; and (4) SNP sequence formatting.

DETAILED DESCRIPTION OF THE INVENTION

[0030] Unlike diploids, SNP marker development in polyploid crop species is very challenging due to the existence of multiple sub-genomes in the nucleus. Due to the presence of duplicated loci in the sub-genomes, it is very difficult to distinguish true SNPs from allelic variations in homologs and false SNPs from non-allelic variations in paralogs.

[0031] Previously, transcriptome and genome complexity reduction techniques combined with high throughput sequencing technologies have been used to enable rapid development of informative SNP markers. SNP mining programs (for example, AutoSNP) have been developed to use allelic frequency as a measure of SNP confidence, but allelic frequency alone is not a good measure of SNP quality especially in polyploid crops. High genomic complexity, narrow genetic base, polyploid nature, and lack of reference genome are major factors to hinder development of candidate SNP markers in cultivated cotton, canola and other species.

[0032] As used herein, the phrase "candidate SNP markers" refers to SNP sequences identified to be validated using biological and/or other assays as associated with traits or phenotypes of an organism, for example plants. As used herein, the phrase "plant" includes dicotyledons plants and monocotyledons plants. Examples of dicotyledons plants include tobacco, Arabidopsis, soybean, tomato, papaya, canola, sunflower, cotton, alfalfa, potato, grapevine, pigeon pea, pea, Brassica, chickpea, sugar beet, rapeseed, watermelon, melon, pepper, peanut, pumpkin, radish, spinach, squash, broccoli, cabbage, carrot, cauliflower, celery, Chinese cabbage, cucumber, eggplant, and lettuce. Examples of monocotyledons plants include corn, rice, wheat, sugarcane, barley, rye, sorghum, orchids, bamboo, banana, cattails, lilies, oat, onion, millet, and triticale.

[0033] As used herein, the phrase "linkage analysis" refers to a method used to identify SNPs close or adjacent to one another in the same contig, chromosome, or a stretch of

sequence defined otherwise. Methods for construction of contigs are well known in the art, for example, see the CAP3 program disclosed in Huang, X. and A. Madan "CAP3: A DNA Sequence Assembly Program." *Genome Research* 9(9): 868-877 (1999), the content of which is incorporated by reference in its entirety.

[0034] As used herein, the phrase "polymorphism" refers to a difference of DNA bases in genomes/chromosomes of organisms. In some embodiments, the polymorphism may reside within coding sequence of an open reading frame. Alternatively, it may reside within non-coding sequences. As used herein, all bases that have variations from genomes/chromosomes of organisms can be considered as polymorphism, which will be distinguished from errors introduced by human manipulation such as sequencing error or mutation introduced during amplification.

[0035] As used herein, the phrase "haplotype" refers to a group of SNPs that are generally inherited together. Haplotypes can have stronger correlations with traits or phenotypic effects compared with individual SNPs, and therefore may provide increased diagnostic accuracy in some cases (see *e.g.*, Stephens *et al.* (2001) *Science* 293: 489-493).

[0036] In the field of bioinformatics, FASTA format was introduced by Bill Pearson and David Lipman in 1988 for representing either nucleotide or amino acid sequences (see Pearson and Lipman, "Improved tools for biological sequence comparison" (1988) *Proc. Natl. Acad. Sci. USA* 85:2444-2448). Basically, a sequence in FASTA format is a text-based format beginning with a single-line description containing a greater-than symbol (>) in first column, followed by lines of sequence data. In addition, an ACE file is a generally used data to represent sequence assembly.

[0037] To increase the efficiency of SNP detection from homologous sequences, and reduce the risk of high false positive rate due to the presence of homeologous sub-genomes in polyploid genomes like cotton and canola, the systems and methods disclosed herein provides a SNP detection pipeline which utilizes the haplotype information to distinguish homologous loci from paralogous loci.

[0038] In one embodiment, this Haplotype Based Pipeline for SNP Discovery and/or Classification (HAPSNP) provided herein uses high throughput sequence data assembly tools along with multiple custom scripts to decipher the contig assembly sequence by (i) identifying putative SNPs initially; (ii) generating haplotype information and allelic frequency of loci in respective genotypes; and (iii) enhancing the ability to identify high

quality SNPs using the haplotype information and allelic frequency. This exemplary pipeline functions well for SNP marker discovery using the sequence information from biparental resources, for example, in both cotton and canola. SNPs identified from this pipeline can be converted into genotyping assays and can be validated with a success rate of up to 60-80% polymorphism rate across various genotypes.

[0039] The efficiency of the HAPSNP provided herein is relatively high in (i) high assay validation rate (60-80%) compared to other SNP mining programs (<25%) for polyploid species; and (ii) more robust in handling huge datasets for allele mining (>10 Million sequences) compared to other SNP mining programs (<1 million sequences). The utility of the exemplary HAPSNP provided herein can be extended to other complex diploids, polyploid crop species and targeted de-novo or re-sequencing projects to identify true SNPs and also to analyze the multiple types of sequence data (for example, from 454 Life Science Corporation, Applied Biosystems (ABI), and/or Illumina Inc.) from more than multiple parental (>2) sources.

[0040] The HAPSNP pipeline provided herein can be implemented for single nucleotide variation detection in any organism including plants and it can also be used for formatting of the SNP sequence information to suit assay designing for multiple genotyping chemistries, for example, Illumina Inc.'s GoldenGate assay, Infinium® iSelect® beadchip, or KBioscience's KASPar® assay.

[0041] An exemplary embodiment of the HAPSNP pipeline is shown in Figure 1, where five modules of the system are illustrated: (1) Assembly/mapping; (2) SNP calling; (3) SNP filtration; (4) Haplotype calling; and (5) SNP sequence formatting.

[0042] For module (1) Assembly/mapping, the input can be raw sequencing data. The raw sequencing data can be generated either for *de novo* or re-sequencing purposes through the next generation sequencing (NGS) instruments which can be initially quality filtered according to the standard criteria set by NGS instrument manufacturers. In some embodiments, when a reference genome or sequence exists, sequences from two or more sources (for example, genotypes and/or parental lines) can be assembled into contigs using *de novo* assembly programs (for example, Celera Assembler), or mapped to a reference genome or sequence using other programs (for example, Mosaik program). In one embodiment, the assembled or mapped data is converted in .ace file format for further processing. Thus, the output of module (1) can be either contig FASTA files or .ace files.

[0043] For module (2) SNP calling, the input typically includes .ace files as illustrated in Figure 1. In one embodiment, all possible loci with single nucleotide variations are identified by a custom designed script in the contig regions. In a further or separate embodiment, the systems and methods provided herein allow user to set the sequencing depth at SNP position for each allele and allelic frequency per genotype required for SNP allele calling. The major function of module (2) is to remove most of the false SNPs from sequencing errors, and this function is critical for distinguishing the allelic variants (homologous, true SNPs) from the non-allelic (homeologous or paralogous, false SNPs) variants and for haplotype calling (see description for module (4) below). The output of module (2) may include all possible SNPs and contigs generated from SNPs.

[0044] For module (3) SNP filtration, the input typically includes all identified SNPs from module (2). The major function of module (3) is to remove SNPs found in homopolymer stretches where sequencing technology is prone to errors (for example, especially from 454 Life Science Corporation). Other technology/project specific filters are contemplated to be applied in module (3) to further reduce false positives. The SNP filtration module provided is different than existing programs because filtration used here does not depend on (*i.e.*, independent from) numbers of SNPs, frequency of duplication of SNPs, or size of the population as in existing programs. Further, the HAPSNP pipeline provided herein allows users to choose and/or create customized SNP filtration unit within the SNP filtration module for a specific purpose, for example, for particular crops including cotton, canola, corn, wheat, sunflower, or soybean.

[0045] For module (4) Haplotype calling, the input typically includes all possible SNPs or contigs generated from SNPs from module (3). The information from module (3) is used to generate the haplotype information for each contig. As provided herein, each haplotype is defined as a unique combination of alleles in contiguous series of SNP locations found in a contig. Haplotypes can be generated for each contig by examining the patterns of SNP loci across contigs. SNPs with more than two haplotypes in any of the genotypes (most common in polyploids) or with the same two haplotypes in all the genotypes are considered false SNPs as they are potentially non-allelic variations between paralogs and eliminated for further validation. Thus, the major function of module (4) is to greatly enhance the percentage of true SNPs for validation after haplotype generation. The output of module (4) may include haplotypes generated from contigs/SNPs. Module (4) can optionally include a haplotype

filtration unit to filter out false or undesired haplotype.

[0046] For module (5) SNP sequence formatting, the input typically includes filtered SNPs, Contigs (FASTA files), and haplotypes generated from contigs/SNPs. In some embodiment, the contig sequences are used to get flanking sequence for each filtered SNP. In some embodiments, each of the filtered SNP loci is converted to [Allele 1/Allele 2] format and the flanking sequences are formatted to fit for assay design and validation with, for example, KASPar®, Infinium®, or GoldenGate genotyping technology. In some embodiments, if there are multiple SNP loci in the same contig, the SNP other than the selected position (10 bases upstream and downstream) can be converted into ambiguous bases or wobbles (R, Y, M, K, M, S, W, H, B, V, D, N) to avoid assay design in the flanking SNP region. The SNPs that are away from the selected SNP position can be converted to major allele. This process reduces the risk of failure during assay validation. The output of module (5) typically includes selected SNPs with flanking sequences, for example, listed in an Excel spreadsheet.

[0047] In some embodiments, the systems or methods disclosed herein further comprises at least one SNP marker classification module. In some embodiments, candidate SNP markers are classified into at least two types using the SNP marker classification module. In other embodiments, candidate SNP markers are classified into at least three types using the SNP marker classification module. The classification can be based on association with genotype or other criteria as demonstrated in examples herein.

[0048] Major advantage of the systems and methods provided herein include at least one of the following: (1) the HAPSNP pipeline disclosed can handle large sequencing data generated from NGS instruments; (2) the HAPSNP pipeline disclosed can use sequencing depth at SNP position and allele frequency to assure the quality of allele calling and distinguish the allelic variations from non-allelic variations between paralogs, (3) the HAPSNP pipeline disclosed can implement haplotype information to further enhance the percentage of true SNPs, and (4) the HAPSNP pipeline disclosed can format the SNP sequence information to suit assay design with multiple genotyping platforms.

[0049] The HAPSNP pipeline provided includes a data storage/database and retrieval system for SNPs/haplotypes integrated with operation system and analysis system as shown in Figure 2. The input device may include raw sequencing data from genomic DNA, expression sequence tag (ESTs), genome sequence tags (GSTs), and/or nucleic acid

information from other sources such as FASTA files. In some embodiments, the HAPSNP pipeline provided herein allows users to input specific sequence data as desired. The output device may include unit for generating Excel spreadsheet to be displayed in a computer screen, database for SNP/haplotype of contig/alignments (before and after filtration), and/or user-friendly interface, for example, a web-based interface or e-mail notification system.

[0050] The HAPSNP pipeline disclosed can be particularly useful for SNP discovery in polyploid species including cotton, canola, and soybean. Further, the HAPSNP pipeline disclosed is powerful enough to identify SNPs from a set of two parents and also to generate haplotype information for "Genotyping by Sequencing" projects used in either quantitative trait locus (QTL) mapping or trait introgression programs, or even for hybrid crops. The utility of the systems and methods disclosed can be extended to analyze the data from multiple sequencing technologies and also multiple parental sources to identify candidate SNP loci for assay validation in, for example, cotton and canola. In addition, the systems and methods disclosed can be used to analyze the NGS data from targeted re-sequencing projects in, for example, soybean, corn, sunflower, and cotton.

EXAMPLES

Example 1

Comparison of the Improved HAPSNP pipeline to Existing Programs

[0051] Sequencing data from cotton can be imported directly into the assembly/mapping module of the HAPSNP pipeline provided as shown in Figure 3. After assembly/mapping (for example, see Module 1 of Figure 1), the output .ace (or ACE) file can be input into the SNP calling module (see Module 2 of Figure 1 and Figure 4). The SNP calling module determines all possible SNPs based on sequence comparison among all input sequences and optionally a reference sequence is considered for sequence comparison. Contig sequences and identifiers can be included in all SNPs as output after SNP calling as shown in Figure 5. These SNPs/contigs are then subject to SNP filtration (for example, see Module 3 of Figure 1). The SNP filtration module can also determine whether SNPs are in a homopolymer region. If yes, the homopolymer region SNPs can be displayed as shown in Figure 6. After SNP filtration, false positive SNPs are removed and input into a SNP sequence formatting module as shown in Figure 1.

[0052] Separately, all possible SNPs are subject to the haplotype calling module (for

example, see Module 4 of Figure 1). The haplotype calling module can optionally include a haplotype filtration unit which is independent from the SNP filtration module. The haplotype information can be input into the SNP sequence formatting module to be considered for association with genotypes after combination with filtered SNPs (see Figure 9 for an example of a haplotype output).

[0053] Finally, the SNP sequence formatting module (for example, see Module 5 of Figure 1) compiles filtered SNP with flanking sequences together with haplotype information (optionally filtered) to determine contigs containing "candidate SNP markers" (see Figure 10 as an example of the output of candidate SNP markers with contig identifiers).

[0054] As shown in Figure 10, the output of the HAPSNP pipeline provided herein can include (1) contig identifier information, (2) contig sequence information, (3) SNP sequence information, and (4) haplotype designation.

Table 1. Exemplary Validation Success Results		
Crop	Pipeline	Validation Success
Cotton	QualitySNP*	27.0 %
Canola	Canada SNP Discovery Consortium**	32.8 %
Cotton	HAPSNP pipeline	68.7%
Canola	HAPSNP pipeline	60.5%***
* The QualitySNP program is disclosed in Tang <i>et al.</i>, BMC Bioinformatics 7:438 (2006). ** The Canada SNP Discovery Consortium is part of the International HapMap Project available at the world wide website //snp.cshl.org/. *** A combination of two studies.		

[0055] The HAPSNP pipeline of this example is compared to publicly unmodified program including QualitySNP and Consortium in either cotton or canola. As shown in Table 1, the HAPSNP pipeline provided can increase validation success of candidate SNP markers more than two folds from about 27-33% to about 60-69%.

[0056] For the canola project, 1499 SNP markers are validated out of 4,568 candidates from the Canada SNP Discovery Consortium, thus a 32.8% validation success. Using the HAPSNP pipeline of this example, the validation success of 60.5% is calculated based on a combination of two studies $[(1374+5285)/(2171+8830) = 60.5\%]$: first set (for example, Type

I) - 1374 validated SNP markers out of 2171 candidates SNP markers (resulting a 63% validation success by itself), and second set (for example, Type II) - validated SNP markers 5285 out of 8830 candidates SNP markers (resulting a 60% validation success by itself). Thus, the HAPSNP pipeline provided is superior to existing programs for SNP discovery and/or validation and is able to achieve more than 60% validation success rate.

[0057] Potential SNPs with high confidence score can be classified into Type I, II and III SNPs using the allelic information for that locus. Type I SNPs are variations where alleles are in homologous condition in each genotype. Type II SNPs have heterologous alleles in one genotype and homologous allele in other genotype. Type III SNPs are typically derived from paralogous or homeologous sequences in the genome, and have heterologous alleles within each genotype. These SNPs can be further filtered and formatted with flanking sequence information to fit for multiple SNP genotyping assay formats including GoldenGate®, KASPar®, Infinium® etc. Figures 11A-C show typical distributions of Type I, II, and III SNPs in cotton using the systems and methods provided herein.

Example 2

HAPSNP pipeline in combination with GBS

[0058] Single nucleotide polymorphism (SNP) markers have become markers of choice for marker assisted selection (MAS) in crop improvement programs because of their higher abundance, amenability for automation and availability of high throughput genotyping platforms. Complexity reduction approaches combined with high throughput sequencing technologies have enabled rapid development of informative SNP markers. Genotyping-by-Sequencing (GBS) methods offer high throughput approaches for SNP discovery and genotyping. However, high genomic complexity, narrow genetic base, polyploid nature and lack of reference genome hinder development of candidate SNP markers in cultivated cotton and other non-model plant species. To increase the efficiency of SNP detection from homologous sequences, and reduce the risk of high false positive rate due to the presence of homeologous sub-genomes in polyploid genomes like cotton and canola, the HAPSNP pipeline disclosed herein is used to combine GBS data/system to distinguish homologous loci from paralogous loci. This particular embodiment of HAPSNP pipeline can extract exact homology matches from high throughput sequencing data using STACKs program and is designed with multiple custom scripts to decipher the homologous sequence tags to provide at least one of the following advantages: (i) identifying putative SNPs; (ii) generating

haplotype information and allelic frequency of loci across multiple genotypes; (iii) enhancing the ability to identify high quality SNPs using the haplotype information and allelic frequency; (iv) facilitating redundancy check within the SNP dataset; and (v) providing SNP sequence in assay convertible format. SNPs identified from this pipeline are converted into genotyping assays and are validated with a success rate of up to 75% polymorphism rate across various genotypes. The efficiency of this pipeline is relatively high due to (i) high assay validation rate (~75%) compared to other SNP mining programs (<25% for polyploid species); and (ii) its robustness in handling huge datasets for allele mining (>10 Million sequences) compared to other SNP mining programs (<1 million sequences). The utility of this pipeline can be extended to other complex diploids, polyploid crop species and targeted de-novo or re-sequencing projects to identify true SNPs.

[0059] Unlike diploids, SNP marker development in polyploid crop species is very challenging due to the existence of multiple sub-genomes in the nucleus. Due to the presence of duplicated loci in the sub-genomes, it is very difficult to distinguish true SNPs from allelic variations as in homologs and false SNPs from non-allelic variations as in paralogs. This HAPSNP pipeline provided in this example can be implemented for single nucleotide variation detection in any crop and it can also be used for formatting of the SNP sequence information to suit assay designing for multiple genotyping chemistries such as Illumina GoldenGate, Infinium, iSelect, TaqMan or KASPar assays. Figure 12 represents the flowchart of this GBS-HAPSNP pipeline. The utility of this pipeline can also be extended for routine genotyping from GBS experiments in complex polyploids, including *G. hirsutum*, *G. barbadense*, or *G. mustelinum*.

We claim:

1. A computerized system for discovery and/or classification of single nucleotide polymorphism (SNP) markers in an organism, comprising,
 - (a) an input device and an output device/interface;
 - (b) an analysis system interface coupled to memory of a computer;
 - (c) an operating system comprising a database;
 - (d) a controller module for SNP calling or Loci identification; and
 - (e) a filtration engine for removing unreliable SNPs.
2. The computerized system of claim 1, further comprising at least one of alignment module, assembly/mapping module, haplotype calling module, and SNP sequence formatting module.
3. The computerized system of claim 1, wherein the input device is selected from the group consisting of automated sequencer, sequencing data input device, and sequencing data storage device.
4. The computerized system of claim 1, wherein the output interface comprises a list of candidate SNP markers.
5. The computerized system of claim 1, wherein the database contains information selected from the group consisting of SNPs, contigs with at least one SNP, and haplotypes with at least two SNPs.
6. The computerized system of claim 1, wherein the filtration engine for removing unreliable SNPs comprises at least four SNP filters to remove unreliable SNPs and generate reliable SNPs.
7. The computerized system of claim 2, wherein the SNP sequence formatting module generates candidate SNP markers with flanking sequences.

8. The computerized system of claim 4, further comprises a SNP marker classification module.
9. The computerized system of claim 8, wherein the candidate SNP markers are classified into at least three types using the SNP marker classification module.
10. The computerized system of claim 9, wherein at least one type of the candidate SNP markers has a validation success rate of at least 60%.
11. The computerized system of claim 1, comprising a controller module for SNP calling, and further comprising an assembly/mapping module, a haplotype calling module, and a SNP sequence formatting module.
12. The computerized system of claim 1, comprising a controller module for Loci identification, and further comprising an alignment module and a SNP sequence formatting module.
13. A method for use in a computerized system for discovery and/or classification of single nucleotide polymorphism (SNP) markers in an organism, comprising,
 - (a) assembling/mapping sequence data using an assembly/mapping module;
 - (b) identifying all possible SNPs using a SNP calling module; and
 - (c) generating reliable SNPs using a SNP filtration module.
14. The method of claim 13, wherein the computerized system comprises a system of claim 1.
15. The method of claim 13, further comprising determining haplotype using a haplotype calling module.
16. The method of claim 13, further comprising formatting candidate SNP markers using a SNP sequence formatting module.

17. The method of claim 13, wherein the computerized system comprises a system of claim 11.
18. The method of claim 13, wherein the method provides candidate SNP markers having a validation success rate of at least 60%.
19. The method of claim 13, wherein the method provides candidate SNP markers having a validation success rate of at least two folds as compared to a publicly available program.
20. The method of claim 19, wherein the publicly available program is QualitySNP.
21. The method of claim 13, wherein the organism comprises a polyploidy genome.
22. The method of claim 21, wherein the organism is a plant.
23. The method of claim 22, wherein the plant is selected from the group consisting of cotton, canola, corn, soybean, sunflower, and wheat.
24. The method of claim 13, further comprising classifying candidate SNP markers using a SNP marker classification module.
25. The method of claim 25, wherein at least one type of the candidate SNP markers has a validation success rate of at least 60%.
26. The method of claim 25, wherein at least two types of the candidate SNP markers have a validation success rate of at least 60%.
27. A method for use in a computerized system for discovery and/or classification of single nucleotide polymorphism (SNP) markers in an organism in combination with a system for genotyping-by-sequencing (GBS), comprising,
 - (a) aligning genotyping-by-sequencing (GBS) data using an alignment module;

- (b) generating reliable SNPs using a SNP filtration module; and
- (c) formatting candidate SNP markers using a SNP sequence formatting module.

- 28. The method of claim 27, wherein the computerized system comprises a system of claim 1.
- 29. The method of claim 27, further comprising identifying SNP loci using a Loci identification module.
- 30. The method of claim 27, wherein the computerized system comprises a system of claim 12.
- 31. The method of claim 27, wherein the method provides candidate SNP markers having a validation success rate of at least 60%.
- 32. The method of claim 27, wherein the method provides candidate SNP markers having a validation success rate of at least two folds as compared to a publicly available program.
- 33. The method of claim 32, wherein the publicly available program is STACKs.
- 34. The method of claim 27, wherein the organism comprises a polyploidy genome.
- 35. The method of claim 34, wherein the organism is a plant.
- 36. The method of claim 35, wherein the plant is selected from the group consisting of cotton, canola, corn, soybean, sunflower, and wheat.
- 37. The method of claim 36, wherein the plant is *G. hirsutum*, *G. barbadense*, or *G. mustelinum*.

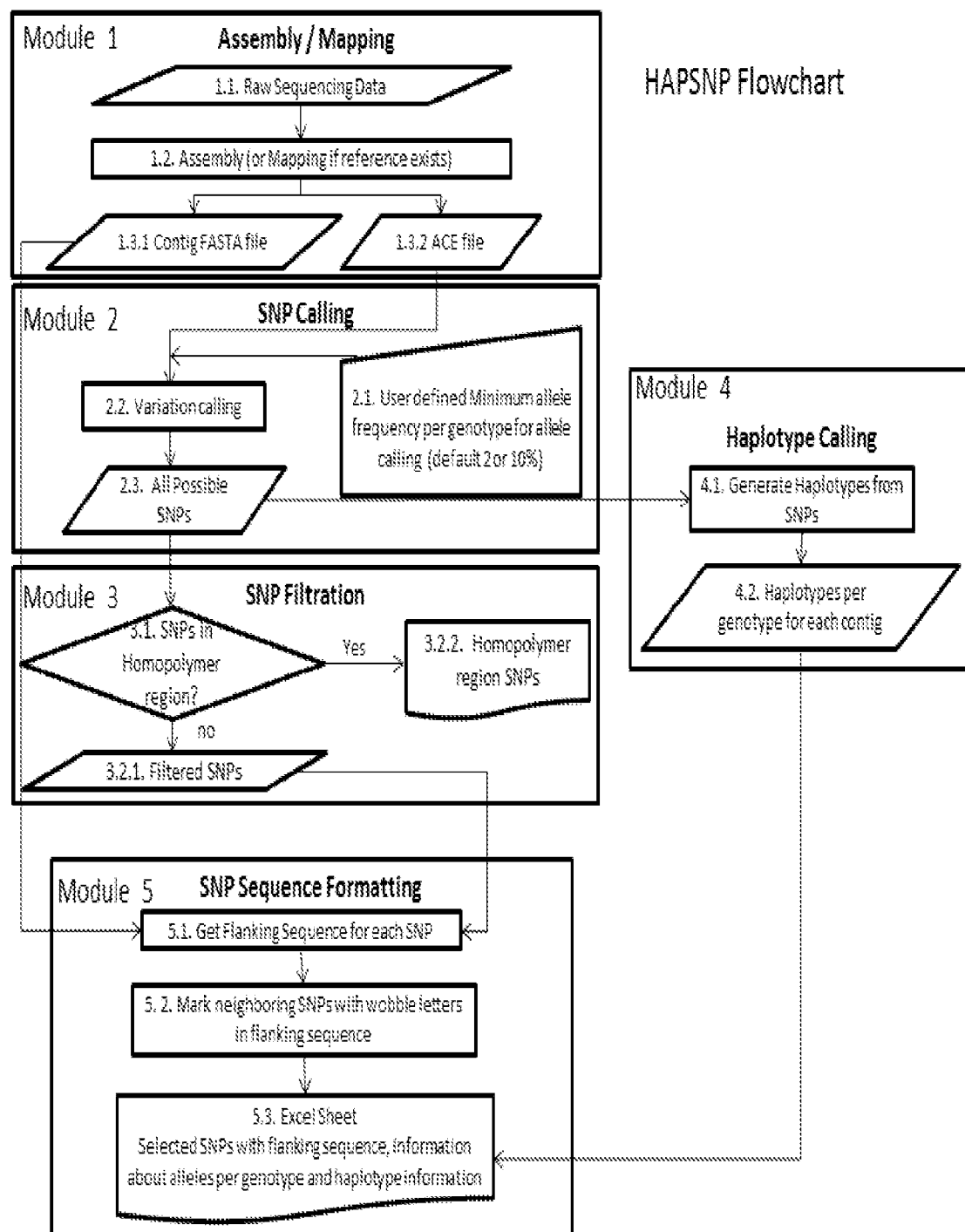


Figure 1

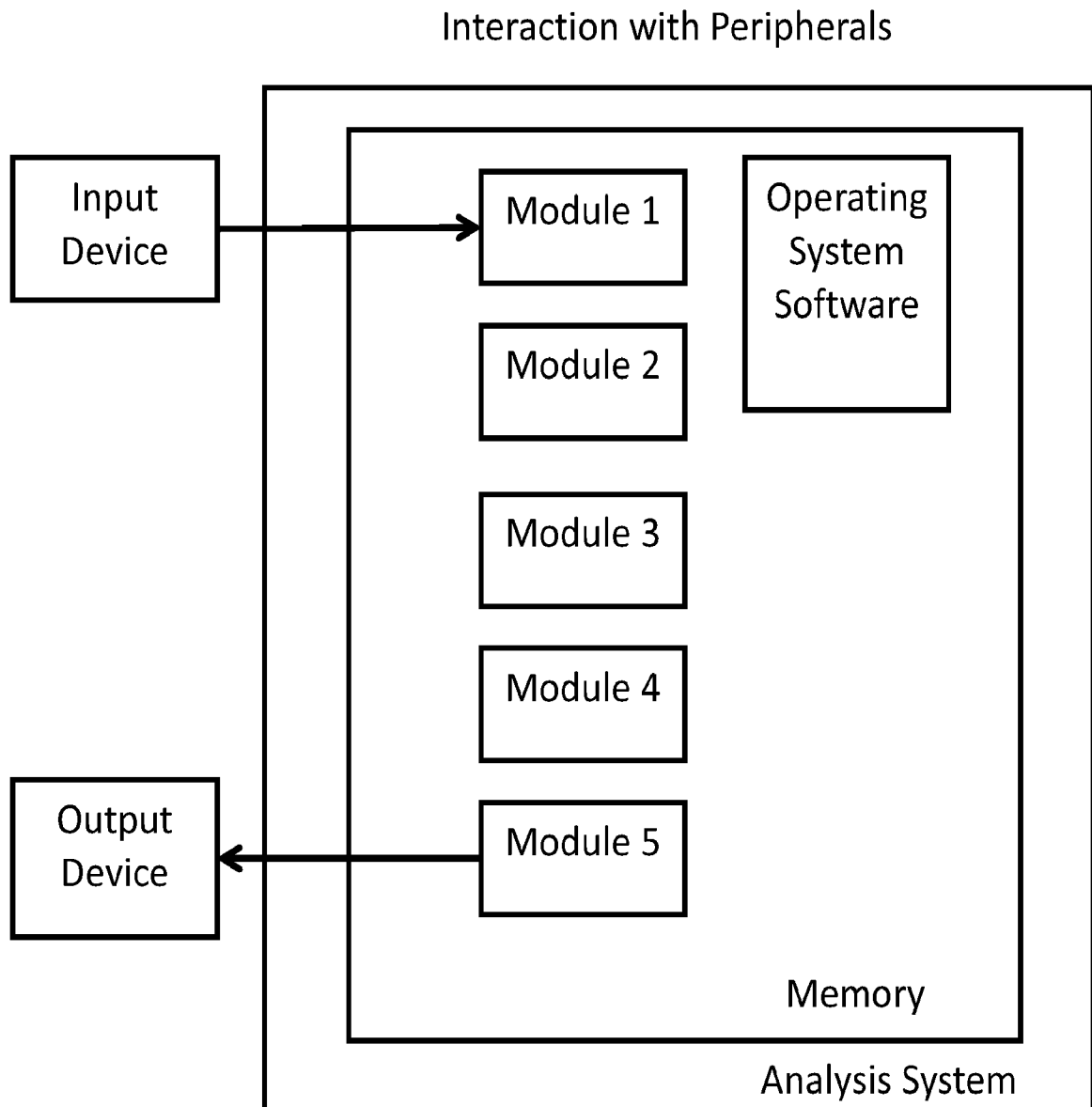


Figure 2

Module 1.1 – Input Screenshot

>DAS_P72_7
ACAAGATTGATCTGCTCCATCATCAAGTCATCAGCTCCAGAACATCTCCGACGAACGTAGAT
>DAS_P72_8
CTCGTGTGCTATGTCGTGTGATACTTAGTCTAGCTACAGACTTAAGTACC
>DAS_P72_9
AGCTTGCGGCTAGTCGTGTGGCCTAAGTCAGATTCGATCACGGCCAAGGCACACAGGG
>DAS_P72_10
GTGTGTGACACACAGCCGTGTCACCTGGCCGTGTGTACACAC
>DAS_P72_11
GTACATTGATATATACTATAGTACAGTGATCGGTGTCATGTAG
>DAS_P72_12
GTGTGTCTCATCGTGTGTACACACGGCCATGTTGCACGACCNCCTAT
>DAS_P72_13
GTGGTACTGAGGAGATGGCATATCTGCTATAATTGAGCTTGTAG
>DAS_P72_14
ATATCTTCATATGCTACGTAGAGGATCTAACTACATAGATTCTACTCC
>DAS_P72_15
CTGACAGTGTGTGTACACGGACTCACACACGGTGGTGTGACT
>DAS_P72_16
AGACGAGACTATACATACTATAGCATGGCGATACATTATAC
>DAS_P72_17
GTGTCAGAGACATGCATAGGCGATCGAGAGATGAGGCGAGT
>DAS_P72_18
AGGTGTGTGACTAGTCGTGTGTGACACACAGCCATGCGACATGGCTGTGTGTCTCT
>DAS_P72_19
GTCTCAGTCGTGTGCGACACATGGCTATGTGGCACGGTCGTGTGTCCNCGTC
>DAS_P72_20
ACTGAGAGATGTAGATATATCAGAGATTAGATACCTCACACAACACTACTACGCAAGGCACACA
>DAS_P72_21
CTATACATACTATAGCATGGCGATACATTATACACTACTACGCGACAC
>DAS_P72_22
GCAGCATGCTGATGAAGGACCAGAGGCACAAGGCTATGTCTTGCATCCAACTACAGTGTCTGATGTGA

Figure 3

Module 1.3.2 – Output Screenshot

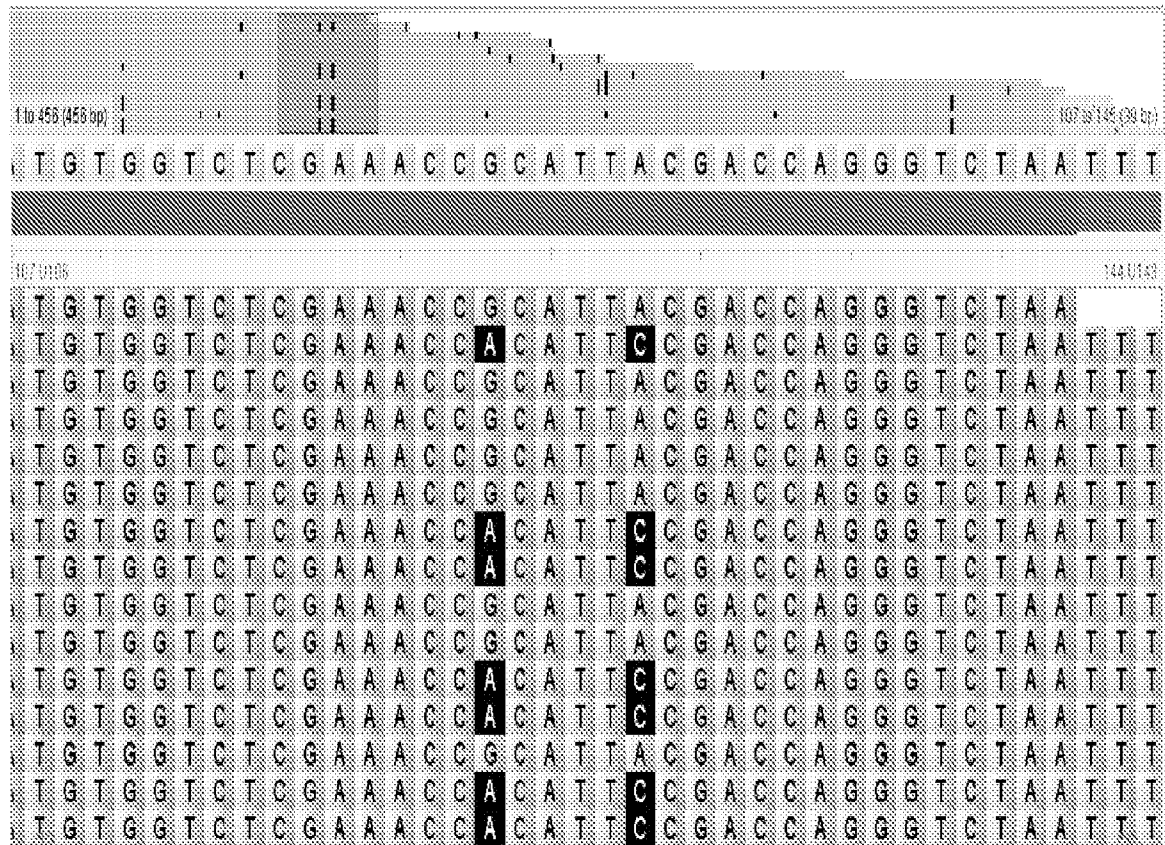


Figure 4

Module 2.3 – Output Screenshot

Contig	Pos	SNP	seq	P72_allele	P800_allele
7180000198765	300	[A/G]	ACATTCA1G		AG
7180000198765	324	[A/G]	ACATTCA1G		AG
7180000198765	404	[A/G]	ACATTCA1G		AG
7180000198766	96	[A/G]	AATTCTCAAG		A
7180000198766	126	[A/T]	AATTCTCAAA		T
7180000198766	150	[C/T]	AATTCTCAAC		T
7180000198766	195	[A/G]	AATTCTCAAG		G
7180000198766	375	[C/T]	AATTCTCAAC		T
7180000198766	453	[C/T]	AATTCTCAAC		T
7180000198768	239	[A/G]	GAATATCAAG		A
7180000198768	251	[G/T]	GAATATCAAG		GT
7180000198768	283	[C/T]	GAATATCACT		CT
7180000198768	294	[C/T]	GAATATCACT		T
7180000198768	346	[C/T]	GAATATCAAC		CT
7180000198770	13	[A/G]	AATTCTCAAG		AG
7180000198770	20	[A/G]	AATTCTCAAG		A
7180000198770	32	[G/T]	AATTCTCAAG		GT
7180000198770	45	[A/G]	AATTCTCAAG		G
7180000198770	49	[A/G/T]	AATTCTCAAGT		G
7180000198770	53	[A/G]	AATTCTCAAG		AG
7180000198770	75	[G/T]	AATTCTCAAGT		G
7180000198770	96	[C/T]	AATTCTCACT		CT
7180000198770	99	[A/G]	AATTCTCAAG		AG

Figure 5

Module 3.2.2 – Output Screenshot

Contig	Position	SNP	Sequence	P72_Allele	P800_Allele
7180000198765	397	[A/G]	ACATTCATAATTACATACTCTAGACAATAA	A	G
7180000198766	81	[C/T]	AATTCTCACTATTTCTTAGATTCATGGAC	C	T
7180000198770	16	[A/G]	AATTCTCAAATGGGA	[A/G]	AAGAGAGAG
7180000198770	159	[C/T]	AATTCTCAAATGGGAGAAGAGAGAAAA	A	CT
7180000198770	301	[C/T]	AATTCTCAAATGGGAGAAGAGAGAAAA	A	CT
7180000198770	471	[C/T]	AATTCTCAAATGGGAGAAGAGAGAAAA	A	CT
7180000198770	492	[A/T]	AATTCTCAAATGGGAGAAGAGAGAAAA	A	T
7180000198773	184	[A/G]	CAGAAGCTGTAAAGGAGGCTATTTGGC	A	AG
7180000198773	206	[A/G]	CAGAAGCTGTAAAGGAGGCTATTTGGC	G	AG
7180000198774	246	[A/G]	AAACCATCAAATTTCTCTTATAGGGATA	A	AG
7180000198774	366	[A/T]	AAACCATCAAATTTCTCTTATAGGGATA	A	T
7180000198776	170	[A/G]	AAATTCCAGTACCGAAAAATATCCAGGA	A	AG
7180000198776	314	[A/G]	AAATTCCAGTACCGAAAAATATCCAGGA	A	AG
7180000198776	348	[A/G]	AAATTCCAGTACCGAAAAATATCCAGGA	A	AG
7180000198785	199	[C/T]	ATGCTAAAAGAAGATTCAAGAATCATC	A	CT
7180000198786	315	[A/G]	CCTAAGTCATATTTATACATTAGACATTA	A	AG
7180000198787	187	[A/G]	CATGGAGATAAGTAGAGATAGACCGA	A	AG
7180000198787	374	[A/C]	CATGGAGATAAGTAGAGATAGACCGA	A	C
7180000198791	168	[A/G]	ATTCTCGGCATGGAGATAAGTAGAGAT	A	AG
7180000198791	195	[A/G]	ATTCTCGGCATGGAGATAAGTAGAGAT	A	AG
7180000198798	231	[G/T]	ATATCCGGGCTATGTCCGAAGGCATTCT	T	GT

Figure 6

Module 3.2.1 – Output Screenshot

Contig	Pos	SNP	seq	P72_P800	Haps_	Haps_	Haps_	Hap1 seq	#reads Ha	Hap2 seq	#reads Ha	Hap3 seq	#reads Ha	Hap4 seq	#reads Ha	Hap5 seq	#reads Ha
7180000198765	300	[A/G]	ACATTCA	G AG	1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
7180000198765	324	[A/G]	ACATTCA	G AG	1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
7180000198765	404	[A/G]	ACATTCA	G AG	1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
7180000198766	96	[A/G]	AATTCTC	A A	2	1	3	TATTGT	3	CGACGCC	5	A..	2				
7180000198766	126	[A/T]	AATTCTC	A T	2	1	3	TATTGT	3	CGACGCC	5	A..	2				
7180000198766	150	[C/T]	AATTCTC	C T	2	1	3	TATTGT	3	CGACGCC	5	A..	2				
7180000198766	195	[A/G]	AATTCTC	AG G	2	1	3	TATTGT	3	CGACGCC	5	A..	2				
7180000198766	375	[C/T]	AATTCTC	C T	2	1	3	TATTGT	3	CGACGCC	5	A..	2				
7180000198766	453	[C/T]	AATTCTC	C T	2	1	3	TATTGT	3	CGACGCC	5	A..	2				
7180000198768	239	[A/G]	GAATATC	AG A	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATT	2
7180000198768	251	[G/T]	GAATATC	G GT	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATT	2
7180000198768	283	[C/T]	GAATATC	T CT	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATT	2
7180000198768	294	[C/T]	GAATATC	CT T	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATT	2
7180000198768	346	[C/T]	GAATATC	C CT	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATT	2
7180000198770	13	[A/G]	AATTCTC	AG AG	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3

Figure 7

Module 4.2 – Output Screenshot

Haps_P72	Haps_P800	Haps_total	Hap1 seq	#reads Hap1	Hap2 seq	#reads Hap2	Hap3 seq	#reads Hap3	Hap4 seq	#reads Hap4	Hap5 seq	#reads Hap5
1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
2	1	3	TATTGTT	3	CGACGCC	5	...A..	2				
2	1	3	TATTGTT	3	CGACGCC	5	...A..	2				
2	1	3	TATTGTT	3	CGACGCC	5	...A..	2				
2	1	3	TATTGTT	3	CGACGCC	5	...A..	2				
2	1	3	TATTGTT	3	CGACGCC	5	...A..	2				
2	1	3	TATTGTT	3	CGACGCC	5	...A..	2				
3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
17	12	24	GGAGGGG	58	AGAGGGG	13	AGAGGGG	6	GGAGGGG	4	AGAGGGG	3
17	12	24	GGAGGGG	58	AGAGGGG	13	AGAGGGG	6	GGAGGGG	4	AGAGGGG	3
17	12	24	GGAGGGG	58	AGAGGGG	13	AGAGGGG	6	GGAGGGG	4	AGAGGGG	3
17	12	24	GGAGGGG	58	AGAGGGG	13	AGAGGGG	6	GGAGGGG	4	AGAGGGG	3
17	12	24	GGAGGGG	58	AGAGGGG	13	AGAGGGG	6	GGAGGGG	4	AGAGGGG	3
17	12	24	GGAGGGG	58	AGAGGGG	13	AGAGGGG	6	GGAGGGG	4	AGAGGGG	3

Figure 8

Module 5.2 – Output Screenshot

Contig	Pos	SNP	seq	P72 : P800	Haps_P72	Haps_P800	Haps_total	Hap1 seq	#reads Ha	Hap2 seq	#reads Ha	Hap3 seq	#reads Ha	Hap4 seq	#reads Ha	Hap5 seq	#reads Ha
7180000198765	300	[A/G]	ACATTCA	G AG	1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
7180000198765	324	[A/G]	ACATTCA	G AG	1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
7180000198765	404	[A/G]	ACATTCA	G AG	1	4	5	TGGAG	3	CAGGA	2	CGAGG	1	CGGGA	1	CAAGG	1
7180000198766	96	[A/G]	AATTCTC	A A	2	1	3	TATTGTT	3	CGACGCC	5	A..	2				
7180000198766	126	[A/T]	AATTCTC	A T	2	1	3	TATTGTT	3	CGACGCC	5	A..	2				
7180000198766	150	[C/T]	AATTCTC	C T	2	1	3	TATTGTT	3	CGACGCC	5	A..	2				
7180000198766	195	[A/G]	AATTCTC	AG G	2	1	3	TATTGTT	3	CGACGCC	5	A..	2				
7180000198766	375	[C/T]	AATTCTC	C T	2	1	3	TATTGTT	3	CGACGCC	5	A..	2				
7180000198766	453	[C/T]	AATTCTC	C T	2	1	3	TATTGTT	3	CGACGCC	5	A..	2				
7180000198768	239	[A/G]	GAATATC	AG A	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
7180000198768	251	[G/T]	GAATATC	G GT	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
7180000198768	283	[C/T]	GAATATC	T CT	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
7180000198768	294	[C/T]	GAATATC	CT T	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
7180000198768	346	[C/T]	GAATATC	C CT	3	6	6	GACAGAT	7	AGCAGAC	3	..CGGATT	3	..CAGATC	3	..CATATTC	2
7180000198770	13	[A/G]	AATTCTC	AG AG	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	20	[A/G]	AATTCTC	AG A	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	32	[G/T]	AATTCTC	G GT	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	45	[A/G]	AATTCTC	AG G	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	49	[A/G/T]	AATTCTC	AGT G	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	53	[A/G]	AATTCTC	AG AG	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	75	[G/T]	AATTCTC	GT G	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3
7180000198770	96	[C/T]	AATTCTC	CT CT	17	12	24	GGAGGG	58	AGAGGG	13	AGAGGG	6	GGAGGG	4	AGAGGG	3

Figure 9

Module 5.3 – Output Screenshot

Contig	Pos	SNP	seq	P1	P2	Haps_1	Haps_P	Haps_to	Hap1_s	#reads	Hap2_s	#reads
7180000198795	288	[G/T]	GCTATCT G	T		1	1	2	GCTC	3	TTCT	2
7180000198795	319	[C/T]	GCTATCT C	T		1	1	2	GCTC	3	TTCT	2
7180000198795	377	[C/T]	GCTATCT T	C		1	1	2	GCTC	3	TTCT	2
7180000198795	380	[C/T]	GCTATCT C	T		1	1	2	GCTC	3	TTCT	2
7180000198809	228	[A/G]	GATTGAG	A		1	1	2	A.C	2	GGC	4
7180000198821	67	[A/T]	GAATGATA	T		1	1	2	T	2	A	3
7180000198845	94	[C/T]	CAAGTAGC	T		1	1	2	TTGAGT	3	CCAG.C	2
7180000198845	99	[C/T]	CAAGTAGC	T		1	1	2	TTGAGT	3	CCAG.C	2
7180000198845	210	[A/G]	CAAGTAGA	G		1	1	2	TTGAGT	3	CCAG.C	2
7180000198845	263	[A/G]	CAAGTAGG	A		1	1	2	TTGAGT	3	CCAG.C	2
7180000198845	431	[C/T]	CAAGTAGC	T		1	1	2	TTGAGT	3	CCAG.C	2
7180000198914	72	[A/G]	AAGAAATA	G		1	1	2	GGCGC	2	AATAT	4
7180000198914	137	[A/G]	AAGAAATA	G		1	1	2	GGCGC	2	AATAT	4
7180000198914	273	[C/T]	AAGAAATT	C		1	1	2	GGCGC	2	AATAT	4
7180000198926	189	[A/G]	AATTGAA G	A		1	1	2	A	3	G	3
7180000198929	196	[A/G]	GGTAAAT G	A		1	1	2	ACCG	3	GATA	3
7180000198929	205	[A/C]	GGTAAATA	C		1	1	2	ACCG	3	GATA	3
7180000198929	237	[C/T]	GGTAAATT	C		1	1	2	ACCG	3	GATA	3
7180000198954	424	[A/G]	AATTCAA A	G		1	1	2	AA	2	GG	2
7180000198954	443	[A/G]	AATTCAA A	G		1	1	2	AA	2	GG	2

Figure 10

Type I SNPs

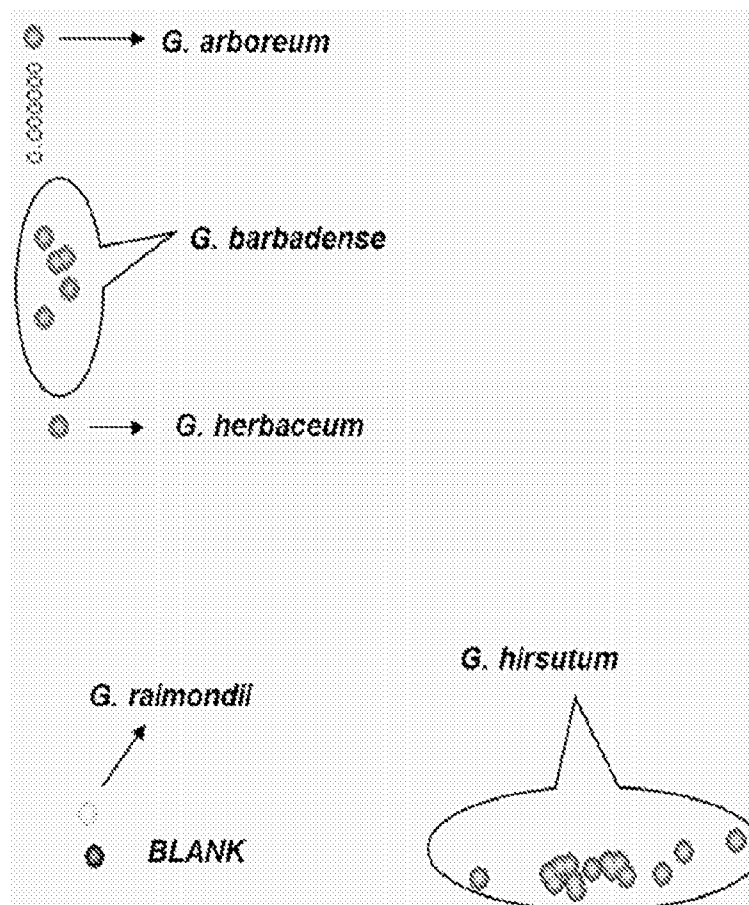


Figure 11A

Type II SNPs

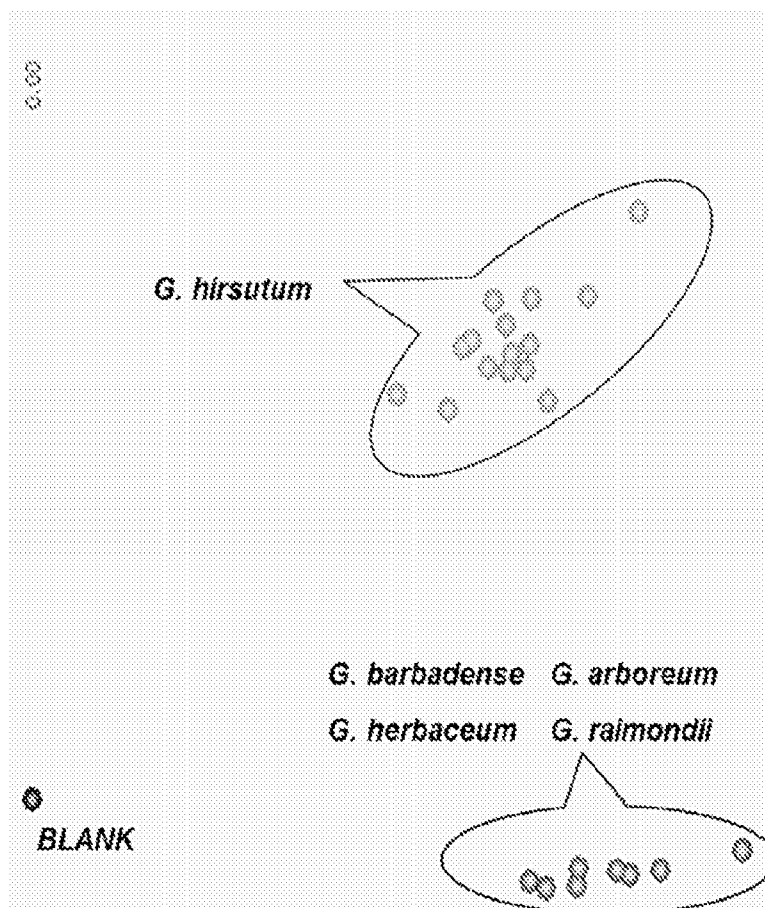


Figure 11B

Type III SNPs

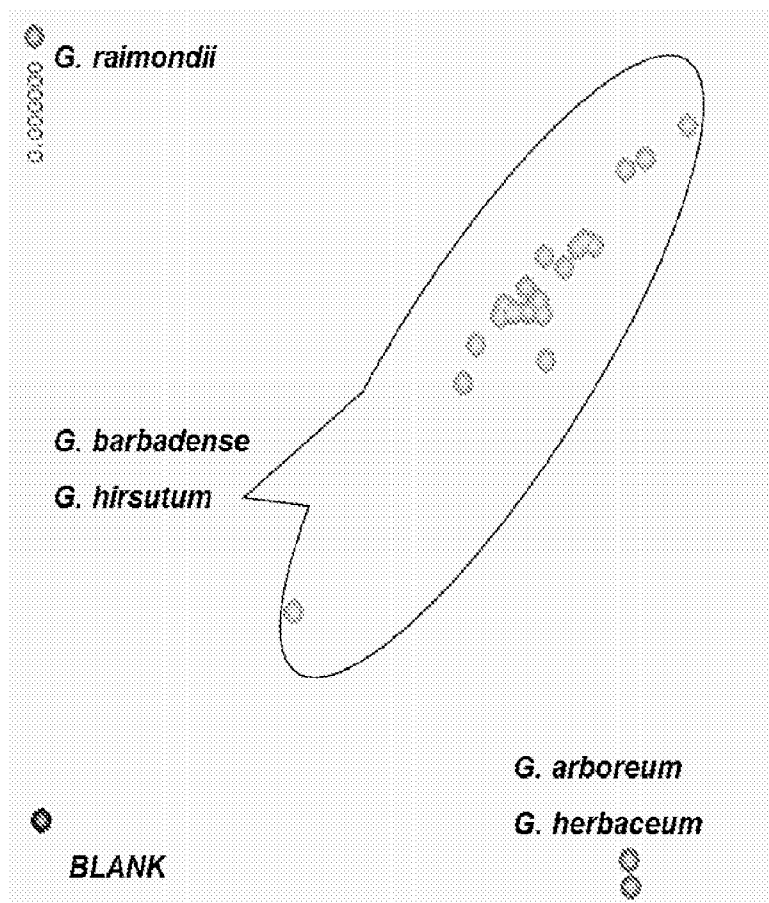


Figure 11C

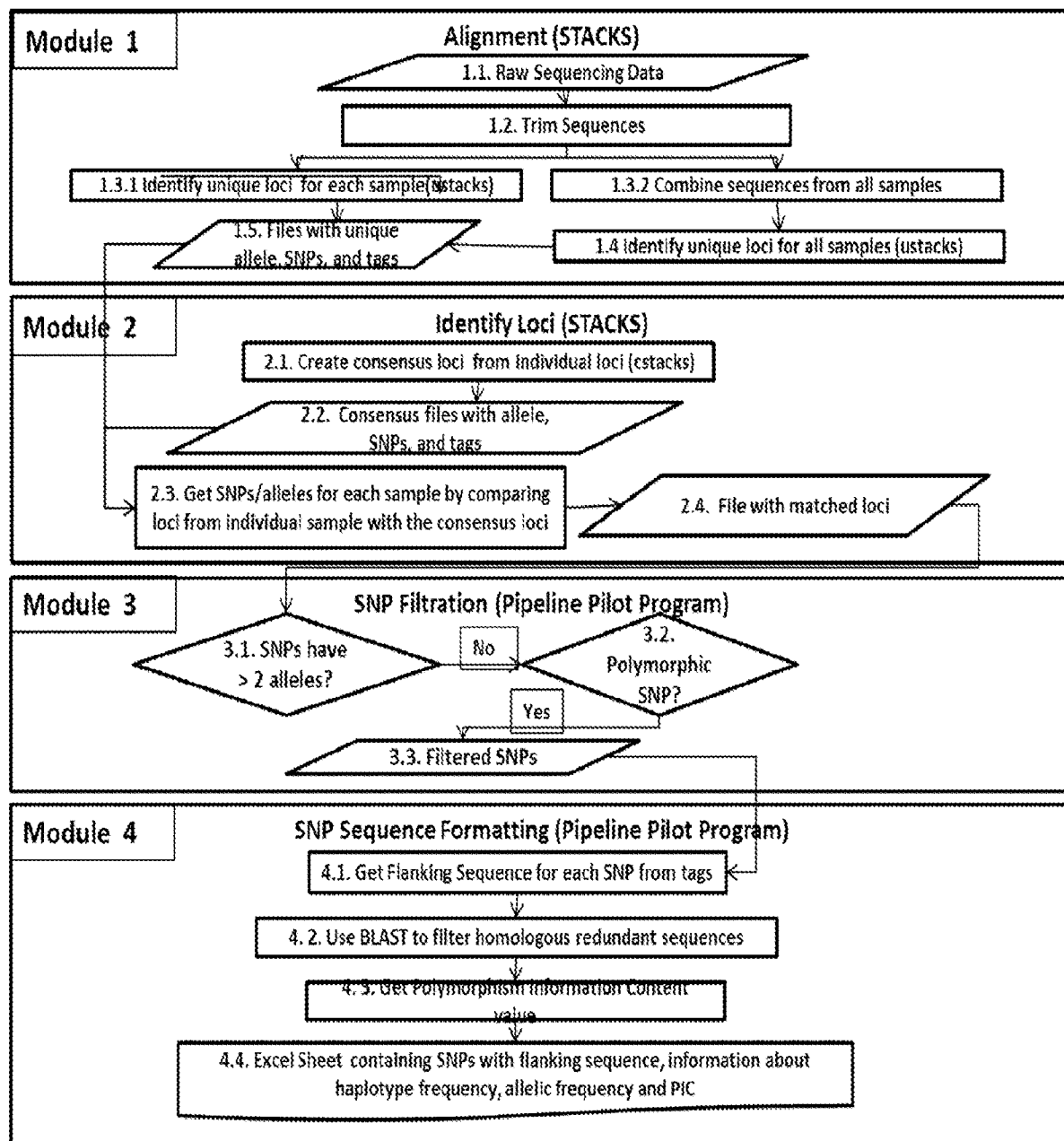


Figure 12