Online Supporting Materials

The online supporting materials have four sections:

(1) The gd-GBS protocol;
(2) Online supporting references;
(3) Supporting tables and figures; and
(4) Supplemental information attachments for the pipeline, npGeno, and for Principal Coordinates Analysis of flax SNP data, respectively.

Section 1: The gd-GBS Protocol

Following the schematic representation of GBS (Figure 1), we divide the gd-GBS procedures into major parts (or components) and each part has many steps. The major steps are shown in Figure 2. Following are the detailed procedures of the original gd-GBS protocol. Related references are given in Section 2 and supporting tables and figures are given in Section 3.

Part I. Sample Preparation

1. DNA Extraction

(1) Extract genomic DNA from 15–20 mg ground freeze-dried leaf tissue using a Qiagen DNeasy Plant Mini Kit (Qiagen, Toronto, ON, Canada). To maximize the final concentration perform two elutions of 30 µL with AE Buffer into separate tubes. In flax the highest concentration of DNA tends to come out in the second elution.
(2) Quantify samples using Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen, Eugene, OR, USA) and adjust to 20 ng/µL with Sigma BPC Grade water (cat. no. W3513)

2. Prepare Adapters (Adapter sequences can be found in Table S1 and illustrated in Figure S1.)

(1) Make 10X Annealing Buffer stock: [S1 in supplemental references below]

1 mL	1 M Tris HCl, pH 8	100 mM final conc.
1 mL	5 M NaCl	500 mM final conc.
0.2 mL	0.5 M EDTA	10 mM final conc.
7.8 mL	H_2O (Sigma BPC Grade)	

Dilute stock to 1X.
(2) Resuspend adapters to 200 µM in IDTE (IDT, Coralville, IA, USA) or 1X TE pH 8.
(3) Make 100 µL of 40 µM double stranded adapter:
20 µL PstI_Adapter 2A (200 µM)
20 µL PstI_Adapter 2B (200 µM)
10 µL 10X Annealing Buffer
50 µL H_2O (Sigma BPC Grade)
Repeat with MspI_Adapter 1A and MspI_Adapter 1B.
(4) In a thermocycler, incubate at 97.5 °C for 2.5 min, and then cool at a rate of not greater than 3 °C per min until the solution reaches a temperature of 21 °C. Hold at 4 °C [S1 in supplemental references].

(5) Prepare working strength concentrations of annealed adapters from this annealed stock.
For 100 reactions combine:
PstI Adapter 2 (4 µM): 5 µL of 40 µM double stranded adapter stock, with 45 µL IDTE.
PstI Adapter 2 (0.1 µM): 2.5 µL of 4 µM double stranded PstI Adapter 2, with 97.5 µL IDTE.
MspI Adapter 1 (10 µM): 37.5 µL of 40 µM double stranded MspI Adapter 1 stock, 112.5 µL IDTE.

(6) Prepare Adapter Mix [S1]:
5 µL of Adapter Mix is required per sample. For 50 reactions combine:

50 µL	0.1 µM PstI Adapter 2	0.02 µM final conc.
75 µL	10 µM MspI Adapter 1	3 µM final conc.
125 µL	1X Annealing Buffer	

Store Adapters and Adapter Mix at −20 °C.

3. Double Digest – *Pst*I/*Msp*I

(1) Make a Restriction Master Mix [S1]:

2.0 µL	10X NEB Cut Smart Buffer	1X final conc.
0.4 µL	PstI-HF 20U/µL	8U
0.4 µL	MspI 20U/µL	8U
7.2 µL	H_2O (Sigma BPC Grade)	

(2) Add 10 µL DNA (200 ng) to 10 µL Restriction Master Mix (20 µL total volume).
(3) Digest for 3 h at 37 °C. Hold at 4 °C [S1].
Reactions can be stored at 4 °C overnight.

4. Adapter Ligation

(1) Prepare a Ligation Master Mix [S2]:

1.0 µL	T4 DNA Ligase (Invitrogen)	1 U/µL
8.0 µL	5X T4 DNA Ligase Buffer	1X final
5.0 µL	Adapter Mix	0.02 µM, 3 µM final
4.0 µL	10 mM ATP	1 mM final
2.0 µL	H2O (Sigma BPC Grade)	

(2) Add 20 µL Ligation Master Mix to 20 µL Digestion.
(3) Incubate at room temperature (23 °C) for 2 h, then heat-kill at 65 °C for 10 min [S1].
Completed ligation can be stored at −20 °C.

5. Clean Samples

Clean the ligation with Agencourt AMPure XP Beads (Beckman Coulter, Mississauga, ON, Canada) following the standard AMPure XP Bead (cat. no. A63880) Protocol (summarized below) using a 1.5X ratio of bead solution volume to ligation solution volume to remove short DNA fragments such as unligated adapters and adapter-adapter ligation products [S1].

(1) Bring AMPure XP Beads to room temperature.
(2) Prepare fresh 70% ethanol (400 µL per sample is required).

(3) Transfer digestion to 0.8ml 96-well storage plate.
(4) Vortex the Agencourt AMPure XP bottle for 30 s to resuspend magnetic particles.
(5) Add Agencourt AMPure XP beads equal to 1.5X the volume of the solution (60 µL).
(6) Mix reagent and digestion thoroughly by pipette mixing 10 times (or shake at 1800 rpm for 2 min).
(7) Incubate for 5 min at room temperature.
(8) Place the plate onto a magnetic stand for 2 min, or until the solution has cleared.
(9) With the plate on the magnetic stand, remove the solution from the reaction plate and discard. Do not disturb the pellet of separated magnetic beads. If beads are drawn out, leave a few microlitres of supernatant behind.
(10) With the plate on the magnetic stand, add 200 µL of 70% ethanol to each well of the reaction plate.
(11) Incubate for 30 s at room temperature. Remove and discard the supernatant.
(12) Repeat for a total of two washes.
(13) Be sure to remove all of the ethanol from the bottom of the well as it is a known PCR inhibitor.
(14) With the plate on the magnetic stand, dry ~ 5 min to ensure all traces of ethanol are removed.
(15) Remove from the magnetic rack and add 30 µL of elution buffer (reagent grade water) to each tube and pipette mix 10 times (or shake at 1800 rpm for 2 min).
(16) Incubate at room temperature for 2 min.
(17) Place the reaction plate onto the magnetic stand for 2 min, or until the supernatant has cleared to separate beads from the solution.
(18) Transfer the eluent to new tubes (0.2 mL strip tubes or 96-well plate).

6. PCR Amplification with Indexed Primers

Prior to PCR Amplification, prepare a Sample Sheet (see Part II, Step 5, below and Figure S2) to plan which indexed primers (Table S1) will be used for each sample. The Illumina MiSeq uses a green laser to read G/T bases and a red laser to read A/C bases. At each cycle at least one nucleotide of each colour must be read or the index read could fail [S3]. It is important to balance the primers so the library has an equal (or nearly equal) number of G/T and A/C nucleotides at each base position in the index sequence (Table S2). PCR steps are illustrated in Figure S1.

Phusion High-Fidelity DNA Polymerase PCR Protocol (New England Biolabs, Whitby, ON, Canada):

13.5 µL	H_2O (Sigma BPC Grade)	
5.0 µL	2 mM dNTP	200 µM final conc.
10.0 µL	5X HF Buffer (NEB)	1X final conc.
0.5 µL	Phusion Polymerase 2U/µl (NEB)	1U final conc.
0.5 µL	ddP1 25 µM (Indexed Primer)	0.25 µM final conc.
0.5 µL	ddP2 25 µM (Indexed Primer)	0.25 µM final conc.
20 µL	Ligation	

PCR Program:

98 °C 30s
(98°C 10s, 65 °C 30s, 72 °C 30s) × 14 cycles
72 °C 5 min
Quantify PCR reactions using a Quant-iT™ PicoGreen® dsDNA Assay Kit.

Part II. Library Assembly

1. Prepare Samples for Size Selection

Size selection is automated using a Pippin Prep (Sage Science, Beverly, MA, USA) which can only run four lanes of 30 µL at one time. If more than four samples are going to be sequenced they are combined and concentrated using the Zymo Research (Irvine, CA, USA) DNA Clean & Concentrator-5 Kit (cat. no. D4013) as summarized below:

(1) Pool up to 4 PCR reactions in each group. The total volume of the pool has to be less than 200 µL (4 × 50 µL = 200 µL). Combine samples of similar concentrations in each pool based on PicoGreen quantification after PCR.
(2) Add 1000 µL DNA Binding Buffer to each pool.
(3) Transfer mixture to a provided Zymo-Spin Column in a Collection Tube.
(4) Add 200 µL DNA Wash Buffer to the column. Centrifuge for 30 s. Repeat the wash step.
(5) Transfer the column to a 1.5 mL microtube and add 15 µL DNA Elution Buffer to the column matrix. Incubate at room temperature for 1 min.
(6) Centrifuge for 30 s to elute the DNA.

Optional: To determine the size range of the fragments, run samples on an Agilent Bioanalyzer using an Agilent High Sensitivity DNA Kit (Agilent, Santa Clara, CA, USA).

2. Size Selection with Pippin Prep

(1) Combine 10 µL from each of 3 cleaned samples for each lane (for a total 12 accessions per lane, 48 samples per Pippin Prep gel). If fewer than 48 samples are being cleaned, adjust volumes accordingly to evenly divide samples among all lanes. Add 10 µL Pippin Prep loading solution to bring loading volume to 40 µL.
(2) Set the Pippin Prep to "broad" size selection with a mean of 500 bp and a range of ±100 bp (400–600 bp). The size range must include both the insert size and 140 bp of adapter/primer sequence.
(3) Follow the "Pippin Prep DNA Size Selection System Operations Manual" for preparing the cassette and loading samples.
(4) Run on a 2% agarose cartridge (containing ethidium bromide) with External Marker B.
(5) Collect 40 µL of each sample from the Elution Modules after the run.

3. Pool Samples for Sequencing Library

(1) Quantify samples using a Quant-iT™ PicoGreen® dsDNA Assay Kit.
(2) Convert quantity from ng/µl to nM using the formula [S4]:

$$\left(\frac{concentration\ in\ ng/\mu l}{660\ \text{g}/mol \times \text{average library size in bp}}\right) \times 10^6 = concentration\ in\ nM$$

(3) Adjust concentration of all samples to either 2 nM or 4 nM with 10mM Tris pH 8.5, 0.1% Tween 20 (Teknova, Hollister, CA, USA).
(4) Pool samples by combining 5 µL of each in one 1.5mL tube. Shake at 1000 rpm for 2 min.

4. Prepare Libraries for Sequencing

The denaturing and diluting of the library should be prepared on the same day as the sequencing. Once denatured, the library can be stored for up to one week. Below is the summary of the method for library preparation as outlined by Illumina [S5] for a 4 nM library. To prepare a 2 nM library see the manual. Please refer to the latest version of "Preparing Libraries for Sequencing on the MiSeq" [S5] for any recent modifications or updates provided by Illumina.

4.1. Denature DNA for a 4 nM Library

(1) Combine sample DNA and freshly diluted 0.2 N NaOH (diluted from 1 N NaOH stock, Teknova, Hollister, CA, USA) in a 1.5 mL tube:
5 µL 4 nM sample DNA Pool
5 µL 0.2 N NaOH
(2) Vortex briefly, then centrifuge the sample solution at 280 x g for 1 min.
(3) Incubate 5 min at room temperature to denature the DNA into single strands.
(4) Add pre-chilled HT1 (Illumina, San Diego, CA, USA) to the tube containing denatured DNA to make 20 pM denatured library:
10 µL Denatured DNA
990 µL Pre-chilled HT1 (Illumina)

4.2. Dilute Denatured DNA for a 4 nM Library

Use the following instructions to dilute the 20 pM DNA further to give 600 µL of the desired input concentration.

(1) Dilute the denatured DNA to the desired concentration using the following table: Libraries diluted to 8 pM were optimal for flax.

Final Concentration	6 pM	8 pM	10 pM	12 pM	15 pM	20 pM
20 pM denatured DNA	180 µL	240 µL	300 µL	360 µL	450 µL	600 µL
Pre-chilled HT1	420 µL	360 µL	300 µL	240 µL	50 µL	0 µL

(2) Invert several times to mix and then pulse centrifuge the DNA solution.
(3) Place the denatured and diluted DNA on ice until you are ready to load your samples onto the MiSeq reagent cartridge.

4.3. Prepare PhiX Control

Combine the following volumes to dilute the PhiX library to 4 nM:
2 µL 10 nM PhiX library (Illumina)
3 µL 10 mM Tris-Cl, pH 8.5 with 0.1% Tween 20 (Teknova)

4.4. Denature the PhiX Control

(1) Combine the following volumes of 4 nM PhiX library and freshly diluted 0.2 N NaOH in a 1.5 mL tube:

5 µL 4 nM PhiX library
5 µL 0.2 N NaOH

(2) Vortex briefly to mix the 2 nM PhiX library solution.
(3) Centrifuge to 280 x g for 1 min.
(4) Incubate for 5 min at room temperature to denature the PhiX library into single strands.
(5) For a 20 pM PhiX library combine:
10 µL Denatured PhiX library
990 µL Pre-chilled HT1 (Illumina)

If you are using MiSeq Reagent Kit v3, use your PhiX control at a 20 pM concentration. No further dilution is required. See "Preparing Libraries for Sequencing on the MiSeq" [S5] for PhiX dilution instructions for use with a v2 MiSeq Reagent Kit.

NOTE: You can store the denatured 20 pM PhiX library up to 3 weeks at −20°C. After 3 weeks, cluster numbers tend to decrease.

4.5. Combine Sample Library and PhiX Control

Illumina recommends a low-concentration PhiX control spike-in at 1% for most libraries.
For low diversity libraries, increase the PhiX control spike-in to at least 5%.

(1) Combine the following volumes of denatured PhiX control library and your denatured sample libraryfor a 5% spike-in:
30 µL Denatured PhiX control
570 µL Denatured and diluted sample library
(2) Set the combined sample library and PhiX control aside on ice until you are ready to load it onto the MiSeq reagent cartridge.

5. Sample Sheet Preparation for GBS (without using Illumina Experiment Manager)

Follow the "MiSeq Sample Sheet Quick Reference Guide" [S6] to set up a basic sample sheet (Figure S2) with the following categories: Header, Reads, Settings and Data.

Header: The Workflow, Application, Assay and Chemistry parameters will stay the same for all GBS sequencing runs. The Experiment Name and Date can be changed with each library.

[Header]	
IEMFileVersion	4
Experiment Name	
Date	
Workflow	GenerateFASTQ
Application	FASTQ Only
Assay	TruSeq HT
Description	
Chemistry	Amplicon

Reads: The number of reads will change depending on the size of fragment and the type of MiSeq Kit. The sample sheet below is set up for a 600-cycle v3 MiSeq Kit (300 cycles read in each direction).

If the fragment to be sequenced is only 200 bp the number of reads in the sample sheet would need to be changed to less than 200.

[Reads]
301
301

Settings: In the sample sheet you need to identify the adapter sequences to avoid reporting sequence beyond the sample DNA. The TruSeq adapter sequences are listed in Illumina's MiSeq Sample Sheet Quick Reference Guide [S6]. For GBS it is also important to add the sequence of the restriction site so that these bases are not included in the final sequence (PstI and MspI restriction sites are underlined below).

[Settings]	
ReverseComplement	0
Adapter	CCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2	CTGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Data: The data section will need to be edited for each new MiSeq run. The FASTQ files will be separated based on the information in this section. The sample name and index sequences need to be entered for each sample. The 6-base index sequences for i7 primers (column titled "index") need to be entered into the sample sheet as reverse complements because of the way the sequence is read during sequencing (see Table S2).

[Data]							
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2

Name your sample sheet with the barcode off the MiSeq reagent cartridge and save as a .csv file. The barcode number is located on the reagent cartridge label. See Figure S2 for an example of a Sample Sheet.

Part III. Sequencing on the MiSeq System

(1) Refer to the latest version of Illumina's MiSeq System User Guide [S7] for instructions on how to load reagents, consumables, and the Sample Sheet and how to initiate and monitor a run. In general, preparing the instrument for a run involves (1) loading the final library onto the reagent cartridge, (2) loading the flow cell and reagent cartridge onto instrument, and (3) uploading the Sample Sheet to instrument before initiating a run. Priming and FCA sites are illustrated in Figure S1.

(2) The MiSeq Reporter Software [S8] generates an output data file in FASTQ format with adapter sequences trimmed and files separated by index as part of the run files. They can be found in the folder {MiSeq Run Name (a series of numbers and letters)}\Data\Intensities\BaseCalls

(3) Download run data and FASTQ files.

Part IV. SNP Calling

The SNP calling is conducted using the pipeline, npGeno, we developed for this effort. Figure S3 shows the process of running npGeno and explanations on its components and outputs. To run the pipeline, read "Getting Started with npGeno.pdf" in the attached files (or section 4 below). There are several steps required for generating the SNP genotype and unphased haplotype data as below:

(1) Read the text file "Getting Started with npGeno.pdf" attached in the pipeline folder (Section 4 below and attached files) to get familiar with the program.
(2) Install the dependent freely available software on a Linux operating system according to their respective installation manuals: Minia, Bowtie2, SAMtools, and Perl (if not available) in Linux, and set up paths to access these programs. Download the executable fastx_collapser to the folder where Minia resides. Test if installed software is working by typing: minia, bowtie2, samtools or perl separately.
(3) Create a directory for the npGeno program and copy the whole pipeline to this directory.
(4) Upload all FASTQ data into the directory npGeno resides.
(5) If needed, adjust the related parameters for the output file *Clean_SNP_Genotypes.txt* as instructed for *Missing_threshold.txt* and *SNP_position_threshold.txt* in the subfolder "Threshold_set". Also, generate and provide *Sample_sheet.csv* in the subfolder "Threshold_set". Make sure *Sample_sheet* information matches with the names of those FASTQ data.
(6) Start the pipeline by running the shell file *npGeno.sh* by typing: ./npGeno.sh in command line.
(7) Seven output files are generated in the subfolder "Output_results" of the directory npGeno resides. They are *allcontigs.fa*, *All_SNP_Genotypes.txt*, *Clean_SNP_Genotypes.txt*, *All_SNP_hap.txt*, *Clean_SNP_hap.txt*, *Clean_genotype_STRUCTURE.txt* and *Clean_haplotype_MEGA.txt*.

Part V. Conventional Genetic Diversity Analysis

With these six output data files, particularly with *Clean_SNP_Genotypes.txt* and *Clean_SNP_hap.txt*, one could proceed to perform a conventional genetic diversity of assayed samples to estimate heterozygosity, infer genetic relationship and structure, or quantify genetic distance and differentiation, depending on the study objectives. This can be done using commonly applied population genetic analysis tools such as GenAlEx, AMOVA, STRUCTURE, or R packages. We assume the researcher is familiar with these commonly used computer programs.

Depending on the amount of SNP data generated from one experiment, we highly recommend the use of *Clean_SNP_Genotypes.txt* or *Clean_SNP_hap.txt* data for the genetic diversity analysis, as these datasets should have more accurate SNP information than the data *All_SNP_Genotypes.txt* or *All_SNP_hap.txt*. However, the latter ones can also be directly used or modified to use, depending on the level of error tolerance. If needed, two formatted datasets *Clean_genotype_STRUCTURE.txt* and *Clean_haplotype_MEGA.txt* can be used by other software such as PGDSpider [S9] for further format conversions required for specific diversity analyses.

Section 2: Online Supporting References

Following references are cited in the gd-GBS procedures and may also be cited in the main text:

S1. Peterson, B.K.; Weber, J.N.; Kay, E.H.; Fisher, H.S.; Hoekstra, H.E. Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* **2012**, *7*, e37135.

S2. Poland, J.A.; Brown, P.J.; Sorrells, M.E.; Jannink, J.-L. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE* **2012**, *7*, e32253.

S3. Illumina. Nextera® XT DNA Sample Preparation Guide. Part # 15031942 Rev. C, 2012a. Available online: http://supportres.illumina.com/documents/myillumina/900851dc-01cf-4b70-9e95-d590531c5bd4/nextera_xt_sample_preparation_guide_15031942_c.pdf (accessed on 21 July 2014).

S4. Illumina. Converting ng/µl to nM When Calculating dsDNA Library Concentration. Support Bulletin, January 25, 2014. Available online: https://my.illumina.com/MyIllumina/Bulletin/HhytK0bcEUqiBTj2rZbZAQ/converting-ngl-to-nm-when-calculating-dsdna-librar]. (accessed on 21 July 2014).

S5. Illumina.Preparing Libraries for Sequencing on the MiSeq®. Part No. Part # 15039740 Rev. D. 2013a. Available online: http://supportres.illumina.com/documents/documentation/system_documentation/miseq/preparing-libraries-for-sequencing-on-miseq-15039740-d.pdf (accessed on 21 July 2014).

S6. Illumina. MiSeq® Sample Sheet Quick Reference Guide. Part # 15028392 Rev. E. 2012b. Available online: http://supportres.illumina.com/documents/myillumina/090a4fe9-6e17-4ae6-8965-e6e125065142/miseq_samplesheet_quickrefguide_15028392_e.pdf (accessed on 21 July 2014).

S7. Illumina. MiSeq® System User Guide. Part # 15027617 Rev. L. 2013b. Available online: http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq_system_user_guide_15027617_l.pdf (accessed on 21 July 2014).

S8. Illumina. MiSeq Reporter User Guide. Part # 15028784 Rev. J. 2013d. Available online: http://supportres.illumina.com/documents/documentation/software_documentation/miseqreporter/miseqreporter_userguide_15028784_j.pdf (accessed on 21 July 2014).

S9. Lischer, H.E.L.; Excoffier, L. PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **2012**, *28,* 298–299.

Section 3: Supporting Tables and Figures

Table S1. Enzyme-specific adapter indexing PCR primer sequences*.

Adapters	Oligo Sequence (5'-3')
*Pst*I_Adapter 2A	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCA
*Pst*I_Adapter 2B	/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
*Msp*I_Adapter 1A	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
*Msp*I_Adapter 1B	/5Phos/CGAGATCGGAAGAGCGAGAACAA
Indexed Primers	
ddP2-1	CAA GCA GAA GAC GGC ATA CGA GAT CGT GAT GTG ACT GGA GTT CAG ACG TGT GC
ddP2-2	CAA GCA GAA GAC GGC ATA CGA GAT GTA GCC GTG ACT GGA GTT CAG ACG TGT GC
ddP2-3	CAA GCA GAA GAC GGC ATA CGA GAT GCC TAA GTG ACT GGA GTT CAG ACG TGT GC
ddP2-4	CAA GCA GAA GAC GGC ATA CGA GAT TGG TCA GTG ACT GGA GTT CAG ACG TGT GC
ddP2-5	CAA GCA GAA GAC GGC ATA CGA GAT CAC TGT GTG ACT GGA GTT CAG ACG TGT GC
ddP2-6	CAA GCA GAA GAC GGC ATA CGA GAT ATT GGC GTG ACT GGA GTT CAG ACG TGT GC
ddP2-7	CAA GCA GAA GAC GGC ATA CGA GAT GAT CTG GTG ACT GGA GTT CAG ACG TGT GC
ddP2-8	CAA GCA GAA GAC GGC ATA CGA GAT TCA AGT GTG ACT GGA GTT CAG ACG TGT GC
ddP2-9	CAA GCA GAA GAC GGC ATA CGA GAT CTG ATC GTG ACT GGA GTT CAG ACG TGT GC
ddP2-10	CAA GCA GAA GAC GGC ATA CGA GAT TAC AAG GTG ACT GGA GTT CAG ACG TGT GC
ddP1-1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG GCT ACA CAC TCT TTC CCT ACA CGA CG
ddP1-2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TTG TAA CAC TCT TTC CCT ACA CGA CG
ddP1-3	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GTC AAA CAC TCT TTC CCT ACA CGA CG
ddP1-4	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TAG AGA CAC TCT TTC CCT ACA CGA CG
ddP1-5	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC AAC TAA CAC TCT TTC CCT ACA CGA CG
ddP1-6	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG AGT GGA CAC TCT TTC CCT ACA CGA CG
ddP1-7	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC ACC GGA CAC TCT TTC CCT ACA CGA CG
ddP1-8	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ATA ATA CAC TCT TTC CCT ACA CGA CG

 Indexes are in red.

Table S2. Index sequences*.

	Index ID	Index in Primer Sequence	Index Sequence for Sample Sheet
i7	ddP2-1	CGTGAT	ATCACG
	ddP2-2	GTAGCC	GGCTAC
	ddP2-3	GCCTAA	TTAGGC
	ddP2-4	TGGTCA	TGACCA
	ddP2-5	CACTGT	ACAGTG
	ddP2-6	ATTGGC	GCCAAT
	ddP2-7	GATCTG	CAGATC
	ddP2-8	TCAAGT	ACTTGA
	ddP2-9	CTGATC	GATCAG
	ddP2-10	TACAAG	CTTGTA
i5	ddP1-1	GGCTAC	GGCTAC
	ddP1-2	CTTGTA	CTTGTA
	ddP1-3	AGTCAA	AGTCAA
	ddP1-4	GTAGAG	GTAGAG
	ddP1-5	CAACTA	CAACTA
	ddP1-6	GAGTGG	GAGTGG
	ddP1-7	CACCGG	CACCGG
	ddP1-8	TATAAT	TATAAT

*Oligonucleotide sequences © 2007-2013 Illumina, Inc.

Figure S1. The double-digest gd-GBS procedure: sequences* of adapters, PCR primers and sequencing primers showing complementary alignment and priming. (**a**) Adapters ligate to complementary sequence on the digested genomic DNA (gDNA). Note that only the *Pst*I/*Msp*I fragment is shown. (**b**) First round of amplification completes the complementary sequence for PCR Primer 1. Reverse strand synthesis cannot occur since the priming site for PCR Primer 1 is not present in the top strand due to the Y-adapter. (**c**) PCR Primer 1 binds to the newly synthesized top strand. Synthesis can now occur on both strands, as normal, in subsequent rounds of PCR amplifying fragments with *Pst*I Adapter 2 and *Msp*I Adapter 1. (**d**) The complete doubly stranded amplicon prior to denaturation includes Flow Cell Annealing (FCA) sequences, Indexes, Index and Multiplex Read priming sites. (**e**) The amplicons which now form the library are denatured prior to paired-end sequencing. The bottom strand is bound to the flow cell at FCA 1 and the gDNA and Index 1 regions are sequenced. During bridge amplification, which binds FCA 2 to the flow cell, Index 2 is sequenced. Once bridge amplification is complete, the top second strand is read, resulting in a pair of reads from both ends of the original gDNA fragment.

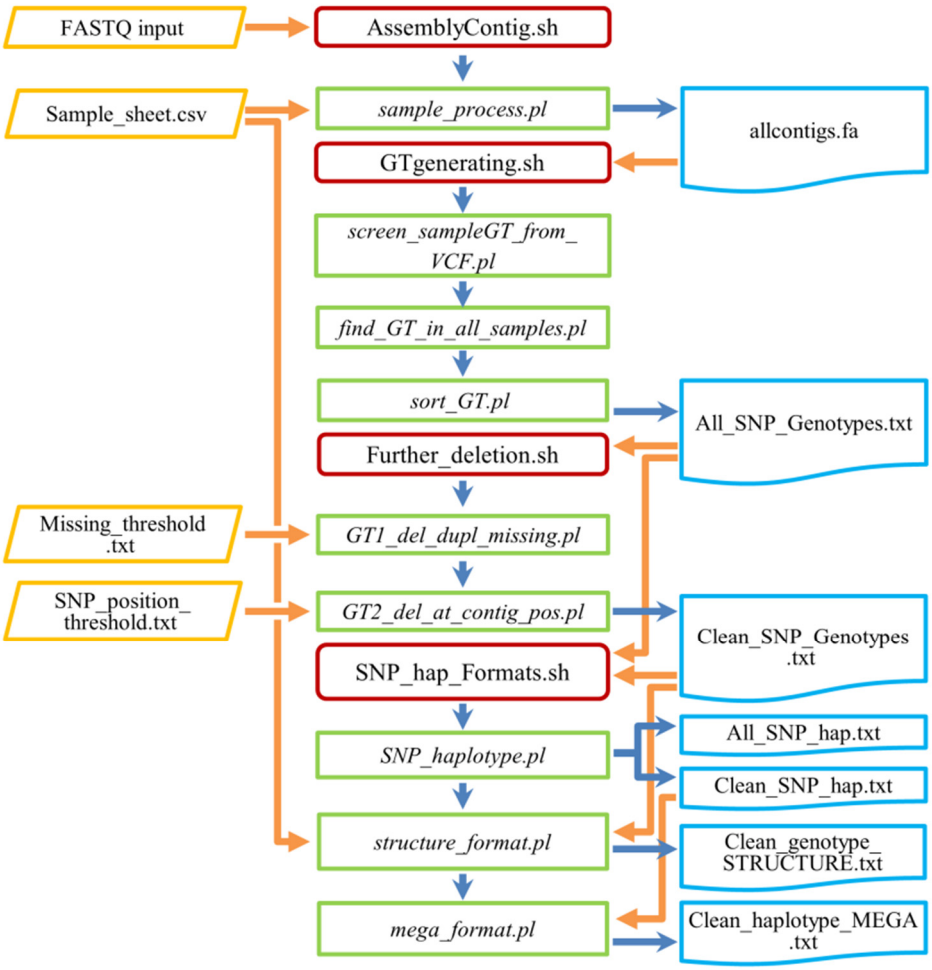
Note: Lavender = digested genomic DNA, Orange = Adapter 1 (ligating sequence is underlined) including the Illumina Multiplex 2 read and Index 1 read priming sequences (Illumina primers are bold and underlined), Green = Adapter 2 (ligating sequence is underlined) including the Illumina Multiplex 1 priming sequence (the Illumina primer is underlined), Purple (italics) = index sequences, Red = Illumina Flow Cell Annealing 1 sequence, Blue = Illumina Flow Cell Annealing 2 sequence. *Oligonucleotide sequences © 2007–2013 Illumina, Inc.

Figure S2. Example Sample Sheet for flax gd-GBS application.

[Header]									
IEMFileVersion	4								
Experiment Name									
Date									
Workflow	GenerateFASTQ								
Application	FASTQ Only								
Assay	TruSeq HT								
Description									
Chemistry	Amplicon								
[Reads]									
251									
251									
[Settings]									
ReverseComplement	0								
Adapter	CCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCA								
AdapterRead2	CTGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT								
[Data]									
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
1	CN 101174	flax	A01	ddP2_1	ATCACG	ddP1_3	AGTCAA		
2	CN 101265	flax	A02	ddP2_2	GGCTAC	ddP1_3	AGTCAA		
3	CN 18989	flax	A03	ddP2_3	TTAGGC	ddP1_3	AGTCAA		
4	CN 100917	flax	A04	ddP2_4	TGACCA	ddP1_3	AGTCAA		
5	CN 101245	flax	A05	ddP2_5	ACAGTG	ddP1_3	AGTCAA		
6	CN 113643	flax	A06	ddP2_6	GCCAAT	ddP1_4	GTAGAG		
7	CN 18974	flax	A07	ddP2_7	CAGATC	ddP1_4	GTAGAG		
8	CN 19003	flax	A08	ddP2_8	ACTTGA	ddP1_4	GTAGAG		
9	CN 33399	flax	B01	ddP2_9	GATCAG	ddP1_4	GTAGAG		
10	CN 52732	flax	B02	ddP2_10	CTTGTA	ddP1_4	GTAGAG		
11	CN 101388	flax	B03	ddP2_1	ATCACG	ddP1_5	CAACTA		
12	CN 101392	flax	B04	ddP2_2	GGCTAC	ddP1_5	CAACTA		
13	CN 101397	flax	B05	ddP2_3	TTAGGC	ddP1_5	CAACTA		
14	CN 101405	flax	B06	ddP2_4	TGACCA	ddP1_5	CAACTA		
15	CN 18991	flax	B07	ddP2_5	ACAGTG	ddP1_5	CAACTA		
16	CN 98935	flax	B08	ddP2_6	GCCAAT	ddP1_6	GAGTGG		
17	CN 97325	flax	C01	ddP2_7	CAGATC	ddP1_6	GAGTGG		
18	CN 97871	flax	C02	ddP2_8	ACTTGA	ddP1_6	GAGTGG		
19	CN 98475	flax	C03	ddP2_9	GATCAG	ddP1_6	GAGTGG		
20	CN 98986	flax	C04	ddP2_10	CTTGTA	ddP1_6	GAGTGG		

Figure S3. Flow chart of the pipeline npGeno and explanation.

A: Run npGeno.sh



B: Explanation:

1. *npGeno.sh*: The shell script to run npGeno program.
2. *FASTQ input*, *Sample_sheet.csv, Missing_threshold.txt* and *SNP_position_threshold.txt* are input files for running npGeno.
3. *AssemblyContig.sh*: The shell script to perform *de novo* assembly contigs as a reference from FASTQ files of all samples.

3.1 *sample_process.pl*: The perl script to process samples in sets of smaller size to run Minia from *Sample_sheet.csv*.

4. *GTgenerating.sh*: The shell script to generate SNPs genotype data for all samples using Bowtie2 and SAMtools.

4.1 *screen_sampleGT_from_VCF.pl:* The Perl script to analyze outputs from Bowtie2 and SAMtools and get SNP genotypes of each sample.

4.2 *find_GT_in_all_samples.pl:* The Perl script to merge all SNPs genotype data from every sample.

4.3 *sort_GT.pl:* The Perl script to sort and output the SNPs genotype for all samples.

5. *Further_deletion.sh*: The shell script to process the genotype result from *GTgenerating.sh* using various filters.

5.1 *GT1_del_dupl_missing.pl:* The Perl script to delete duplicated sites and missing data in the output genotype data from GTgenerating.sh according to the setting in *Missing_threshold.txt.*

5.2 *GT2_del_at_contig_pos.pl:* The Perl script to delete the SNPs located within a specific number of bases from both ends of each contig according to the setting in *SNP_position_threshold.txt.*

6. *SNP_hap_formats.sh*: The shell script to process the genotype results from *Further_deletion.sh* using various filters.

6.1 *SNP_haplotype.pl***:** The Perl script to generate SNP haplotype data according to the SNP genotype data from *All_SNP_Genotypes.txt*, *Clean_SNP_Genotypes.txt* and VCF files of all samples.

6.2 *structure_format.pl*: The Perl script to convert *Clean_SNP_Genotypes.txt* into a STRUCTURE format file according to the sample setting in *Sample_sheet.csv*.

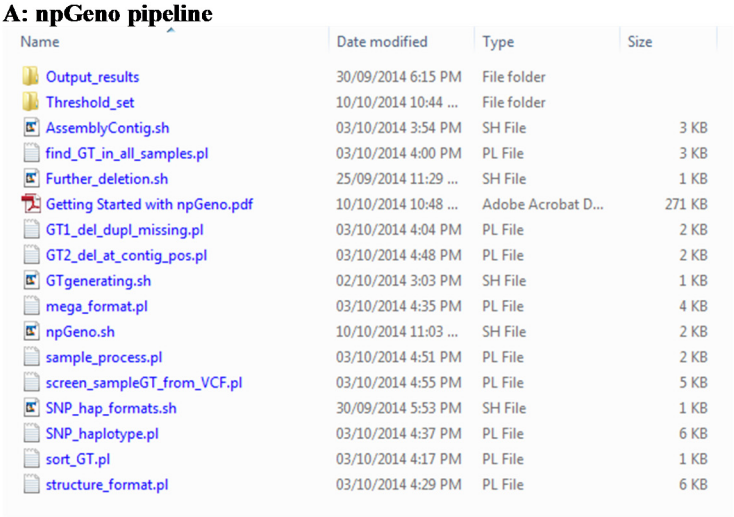
6.3 *mega_format.pl*: The Perl script to transform *Clean_SNP_hap.txt* into a MEGA format file.

7. *allcontigs.fa*, *All_SNP_Genotypes.txt*, *Clean_SNP_Genotypes.txt*, *Clean_genotype_STRUCTURE.txt*, *All_SNP_hap.txt*, *Clean_SNP_hap.txt* and *Clean_haplotype_MEGA.txt* are output files.

Section 4: Supplemental File Attachments for npGeno and a PCoA Analysis

1. File Attachment for the npGeno Pipeline and "Getting Started with npGeno.pdf"

A: npGeno pipeline



Name	Date modified	Type	Size
Output_results	30/09/2014 6:15 PM	File folder	
Threshold_set	10/10/2014 10:44 ...	File folder	
AssemblyContig.sh	03/10/2014 3:54 PM	SH File	3 KB
find_GT_in_all_samples.pl	03/10/2014 4:00 PM	PL File	3 KB
Further_deletion.sh	25/09/2014 11:29 ...	SH File	1 KB
Getting Started with npGeno.pdf	10/10/2014 10:48 ...	Adobe Acrobat D...	271 KB
GT1_del_dupl_missing.pl	03/10/2014 4:04 PM	PL File	2 KB
GT2_del_at_contig_pos.pl	03/10/2014 4:48 PM	PL File	2 KB
GTgenerating.sh	02/10/2014 3:03 PM	SH File	1 KB
mega_format.pl	03/10/2014 4:35 PM	PL File	4 KB
npGeno.sh	10/10/2014 11:03 ...	SH File	2 KB
sample_process.pl	03/10/2014 4:51 PM	PL File	2 KB
screen_sampleGT_from_VCF.pl	03/10/2014 4:55 PM	PL File	5 KB
SNP_hap_formats.sh	30/09/2014 5:53 PM	SH File	1 KB
SNP_haplotype.pl	03/10/2014 4:37 PM	PL File	6 KB
sort_GT.pl	03/10/2014 4:17 PM	PL File	1 KB
structure_format.pl	03/10/2014 4:29 PM	PL File	6 KB

B: Getting Started with npGeno.pdf

Steps to Use npGeno:

1. Familiarize yourself with npGeno by reading *Getting Started withnpGeno.pdf* (this file) attached in the pipeline folder.
2. Install all required free software, set up paths to access those computer programs, and test if installed software is working by typing: minia, bowtie2, SAMtools, or perl separately.
3. Create a directory for the npGeno pipeline and copy the whole pipeline to this directory.
4. Upload all FASTQ data into the same directory npGeno resides.
5. If needed, adjust the related parameters for the output file *Clean_SNP_Genotypes.txt* by editing *Missing_threshold.txt* and *SNP_position_threshold.txt* in the subfolder "Threshold_set". Also, generate and place *Sample_sheet.csv* in the subfolder "Threshold_set" in the same directory of npGeno.
6. Start the pipeline by running the shell file *npGeno.sh* by typing: ./npGeno.sh at the command prompt.
7. Seven output files are generated in the subfolder "Output_results" in the same directory of npGeno.

Prerequisite:

1) Minia (http://minia.genouest.org/). Extend k-mer length to 100 by typing: make clean && make k=100
2) Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
3) SAMtools (http://samtools.sourceforge.net/)
4) Perl in Linux (http://www.perl.org/get.html)
5) Fastx_collapser (http://hannonlab.cshl.edu/fastx_toolkit/). Download it to the same directory of Minia.

Input files:

1) Paired-end Illumina sequencing data files with FASTQ format are used.
2) Two input files in the "Threshold_set" subfolder with adjustable parameters for the output file *Cleaned_SNP_Genotypes.txt*:
   i) *Missing_threshold.txt* is used to remove the loci having a level of missing observations or higher; normally 10-20%. The default setting is 0%.
   ii) *SNP_position_threshold.txt* is used to delete the SNPs located within a specific number of bases from both ends of each contig; normally 10-20. The default setting is 20.
3) *Sample_sheet.csv* in the folder of "Threshold_set" is used to provide information on sample, population, and region (if any) for formatting output data into the output file *Clean_genotype_STRUCTURE.txt*. Also, the .csv file is used by *sample_process.pl* to subdivide large data into sets of smaller size. The .csv file can be generated using Microsoft Excel with five columns: "Sample_Name" (alpha-numeric), "Sample_ID" (numeric), "Population" (alpha-numeric), "Region" (numeric), and "Set" (numeric). The first two columns can be obtained from the MiSeq SampleSheet. Ensure the sample name and MiSeq sample number match those in the FASTQ files. For example, sample name "CN33133-1" with MiSeq sample number "3" on the MiSeq SampleSheet matches the FASTQ file CN33133-1_S3_L001_R1_001.fastq. The next two columns describe the population and regional information for each sample. "Region" should be set as "1" if there is no regional information. The column, "Set", is an ordinal number starting from "1" and used to separate the samples into multiple sets for parallel processing. The same number should be used to indicate all individual samples in the same set. The size of each set should be 15 GB or less (or ideally around 10 GB). See illustration in *Sample_sheet_flax_example.csv*

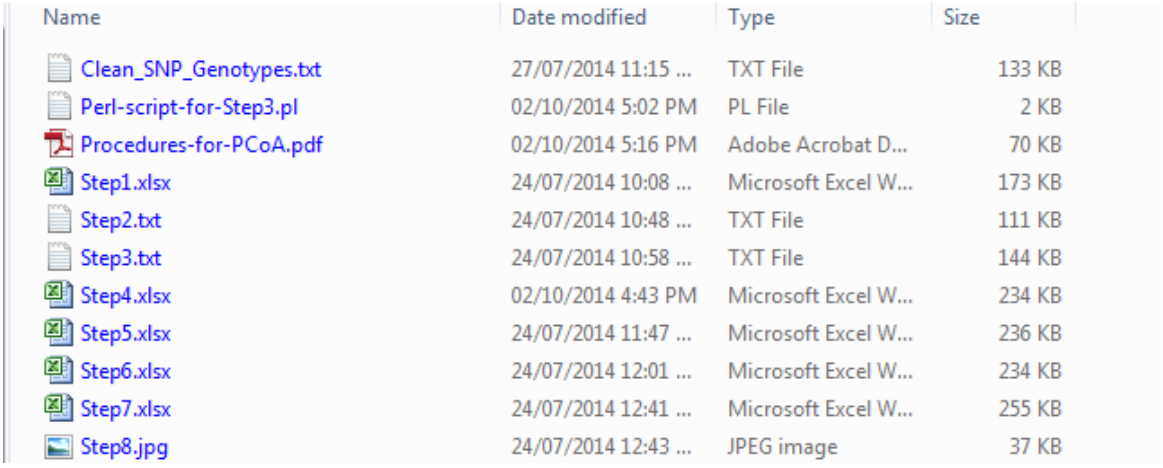
Output files:

allcontigs.fa consists of *de novo* assembly contigs from all samples as a reference for SNP genotyping.
All_SNP_Genotypes.txt includes all SNP genotype data for all samples.
All_SNP_hap.txt is unphased haplotype data corresponding to *All_SNP_Genotypes.txt*.
Clean_SNP_Genotypes.txt is the genotype data after removing SNPs showing the same genotypes for all samples, having a given level of missing observations and residing within a specific number of bases from both ends of each contig.
Clean_SNP_hap.txt includes unphased haplotype data corresponding to *Clean_SNP_Genotypes.txt*.
Clean_genotype_STRUCTURE.txt is a data file with a STRUCTURE format corresponding to *Clean_SNP_Genotypes.txt*.
Clean_haplotype_MEGA.txt is a data file with a MEGA format corresponding to *Clean_SNP_hap.txt*.

Note:

To get accurate genotype data, Minia should be run with the options of a higher kmer_size and larger min_abundance, as these two values are related to contig quality and data unbalance. Generally, kmer_size should be 100-150 for MiSeq data and min_abundance is half of the sample size or larger. In the flax example, kmer_size=100 and min_abundance=15. The option used for genome size should be 300000000 or larger to make Minia run faster, even it is over-estimated.

2. File Attachment for principal Coordinates Analysis of 20 flax Samples Based on Clean_SNP_Genotypes.txt



Name	Date modified	Type	Size
Clean_SNP_Genotypes.txt	27/07/2014 11:15 ...	TXT File	133 KB
Perl-script-for-Step3.pl	02/10/2014 5:02 PM	PL File	2 KB
Procedures-for-PCoA.pdf	02/10/2014 5:16 PM	Adobe Acrobat D...	70 KB
Step1.xlsx	24/07/2014 10:08 ...	Microsoft Excel W...	173 KB
Step2.txt	24/07/2014 10:48 ...	TXT File	111 KB
Step3.txt	24/07/2014 10:58 ...	TXT File	144 KB
Step4.xlsx	02/10/2014 4:43 PM	Microsoft Excel W...	234 KB
Step5.xlsx	24/07/2014 11:47 ...	Microsoft Excel W...	236 KB
Step6.xlsx	24/07/2014 12:01 ...	Microsoft Excel W...	234 KB
Step7.xlsx	24/07/2014 12:41 ...	Microsoft Excel W...	255 KB
Step8.jpg	24/07/2014 12:43 ...	JPEG image	37 KB