Getting Started with npGeno.pdf

Steps to Use npGeno:

1. Familiarize yourself with npGeno by reading *Getting Started with npGeno.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).

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.