INSTITUTE FOR GENOMIC DIVERSITY

Cranioleuca

Genotyping by Sequencing (GBS)
UNEAK Pipeline

Prepared by Katie Hyma (keh233@cornell.edu)

6/17/2013

http://www.igd.cornell.edu/index.cfm/page/projects/GBS.htm

Contents

Sample Details

189 uniquely named samples (189 total) and 0 blank(s) digested with enzyme(s) PstI

Library Plate	Library Plate ID	Flowcell	Lane:
Brumfield_1	450017033	D25GWACXX	5
Brumfield_2	450017034	D25GWACXX	6
Brumfield_6	450017038	D260TACXX	2

Analysis Notes

The GBS UNEAK analysis pipeline (Tassel Version: 3.0.139 Date: November 8, 2012) was run on these samples
An overview of GBS and the GBS pipeline can be found here: http://www.maizegenetics.net/Table/Genotyping-By-Sequencing/. See the following sections for a list of options used for this analysis.

We are providing SNP calls output from the GBS Bioinformatics pipeline (see description of data files below), along with the keyfile(s) used to associate barcodes with sample IDs while running the GBS pipeline.

***Please note,** the GBS pipeline is continually being developed, and as such the results from this run can only be reproduced with the GBS pipeline version noted above, and with the options listed in the following section.

Description of Data Files

We are providing SNP calls in both HapMap and VCF format. Currently SNP calling as implemented for hapmap and VCF files are independent, and as such variations between the two files can be expected.

HapMap files

HapMap is the standard file format generated by the TASSEL GBS pipeline. In the folder labeled “hapMap” you will find four files

1. HapMap.fas.txt – this file contains the fasta format sequences for all tag pairs. The name of the sequences indicates the Tag Pair membership and tag length. Tags shorter than 64 bp have been padded with tailing “A”.
2. HapMap.hmc.txt – this file contains the genotype information for each individual, for each tax pair, along with the total allele counts and the allele frequency
3. HapMap.hmp.txt – this file is a hapmap file with SNP calls that can be imported into TASSEL, or any other program that accepts hapmap file format input.
4. HapMap.filtered.hmp.txt – this file is the result of filtering the HapMap.hmp.txt file on missingness. See section below for parameters.

Here we have filtered on missingness only (see the options listed below for the plugin GBSHapMapFiltersPlugin). We highly recommend filtering your data with parameters that are appropriate for your species/project. HapMap files can be easily opened and filtered in TASSEL 4.0. http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119

VCF files

VCF is an alternative format for holding SNP information that retains information on depth of coverage for each allele, and can be output from the GBS pipeline, but is not a format currently supported by TASSEL. We are providing one file in the folder labeled “vcf” (all.mergedSNPs.vcf.gz) that contains all SNP calls, and has been compressed with bgzip.

You can work with VCF files output from the GBS pipeline using VCFtools (http://vcftools.sourceforge.net/).

VCFtools is currently only available for Linux/Unix platforms. If you do not have access to a Linux computer and wish to use VCFtools, you can purchase machine hours on CBSU’s BioHPC Computing Lab (http://cbsu.tc.cornell.edu/lab/Pricing.aspx).

GBS reference pipeline options used for analysis

plugin	option	value	description
UMergeTaxaTagCountPlugin	-m	200000000	Maximum tag number in the merged TagCount file. Default: 60000000
UMergeTaxaTagCountPlugin	-m	200000000	Maximum tag number in the merged TagCount file. Default: 60000000
UmergeTaxaTagCountPlugin	-c	5	Minimum count of a tag must be present to be output. Default: 5
UmergeTaxaTagCountPlugin	-t		Merge identically named taxa or not. -t n = do not merge. Default: merge
UTagCountToTagPairPlugin	-e	0.03	Error tolerance rate in the network filter. Default: 0.03
UMapInfoToHapMapPlugin	-mnMAF	0.05	Minimum minor allele frequency. Default: 0.05
UMapInfoToHapMapPlugin	-mxMAF	0.5	Maximum minor allele frequency. Default: 0.5
UMapInfoToHapMapPlugin	-mnC	0	Minimum call rate (proportion of taxa covered by at least one tag)
UMapInfoToHapMapPlugin	-mxC	1	Maximum call rate. Default: 1 (proportion of taxa covered by at least one tag)
GBSHapMapFiltersPlugin	-mnSCov	0.8	Minimum site coverage i.e. the minimum call rate for a SNP to be included in the output where call rate is the proprotion of the taxon genotypes for that SNP that are not N" (where N=missing). Default: 0.1"
GBSHapMapFiltersPlugin	-mnTCov	0.1	Minimum taxon coverage i.e. the minimum SNP call rate for a taxon to be included in the output; where call rate is the proportion of the SNP genotypes for a taxon that are not "N" (where N = missing). Default: 0.1
GBSHapMapFiltersPlugin	-mnF		Minimum value of F (inbreeding coefficient). Not tested by default. DO not invoke this option unless you are working with inbred lines or an inbreeding species.
GBSHapMapFiltersPlugin	-mnMAF	0.01	Minimum minor allele frequency Default: 0.0 (no filtering);
GBSHapMapFiltersPlugin	-mxMAF	1	#Maximum minor allele frequency. Default: 1.0 (no filtering);
GBSHapMapFiltersPlugin	-hLD		#Specifies that samples should be filtered for high LD. Default: false (off).tbt2vcfPlugin
tbt2vcfPlugin	-mnMAF	0.05	Minimum minor allele frequency (default: 0.0)

tbt2vcfPlugin	-mnLCov	0	Minimum locus coverage (proportion of Taxa with a genotype) (default: 0.0)
MergeDuplicateSNP_vcf_Plugin	-ak	3	Maximum number of alleles that are kept for each marker across the population; default: 3

Analysis Details

Summary of Reads and Tags found in each Sequencing Lane

**Please note that blanks must be named "blank," case insensitive to be automatically excluded from summary calculations, and to be excluded from the failed sample list **

FastQ file	Barcodes found in lane	Total # of reads per lane	Total number of good barcoded reads
D25GWACXX_5_fastq.gz	95	234489676	214777572
D25GWACXX_6_fastq.gz	94	297132667	151180269
D260TACXX_2_fastq.gz	4	282004668	10498717

Total number of Tags after Merging

- 45286319

Failed Samples

Failed samples are defined as those with less than 10% of the mean reads coming from the lane on which they were sequenced.

Flowcell	Lane	SampleID	Library Plate	Row	Column
C24L5ACXX	7	63442	Brumfield_5	D	6
D260TACXX	2	4139	Brumfield_6	D	9

Summary of Resulting SNPs

- HapMap SNPs (unfiltered) : 105148
- HapMap SNPs (filtered) : 35865
- VCF SNPs: 101825

VCFtools version [v0.1.10] was used to calculate Depth and Missingness from the file all.mergedSNPs.vcf.gz

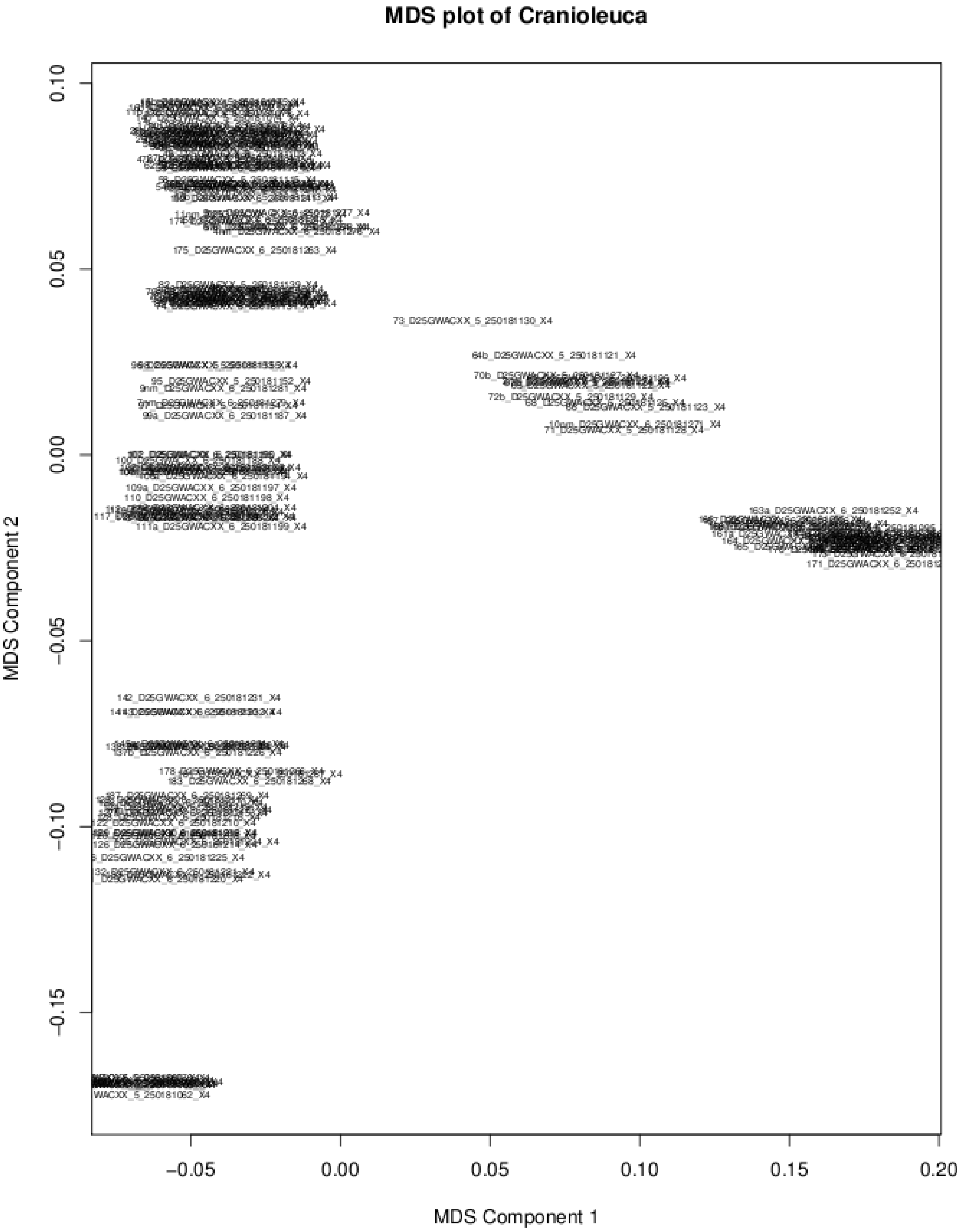
	Mean	Median	Standard Deviation
Individual (Taxa) Depth	4.7912	4.68942	1.795828
Site Depth	3.837864	2.54545	4.342437
Individual (Taxa) Missingness	0.502939	0.480108	0.09551876
Site Missingness	0.502939	0.534392	0.3598338

Multi Dimension Scaling (MDS) of genome-wide SNPs from the reference genome pipeline:

VCFtools version [v0.1.10] was used to filter the SNPs and generate an input file of remaining biallelic SNPs for use with Plink version [v1.07]/

- Filtered genotypes to those with sequencing depth between [3 and 127] **note that the GBS pipeline will count up to a maximum read depth of 127 **
- Filtered individuals with greater than [90]% missing data:
    - 125_D25GWACXX_6_250181213_X4
    - 134_D25GWACXX_6_250181223_X4
    - 144_D25GWACXX_6_250181233_X4
    - 147_D25GWACXX_6_250181236_X4
    - 148_D25GWACXX_6_250181237_X4
    - 150_D25GWACXX_6_250181239_X4
    - 154_D25GWACXX_6_250181243_X4
    - 155_D25GWACXX_6_250181244_X4
    - 160a_D25GWACXX_6_250181249_X4
    - 3nm_D25GWACXX_6_250181275_X4
    - 75b_D25GWACXX_5_250181132_X4
    - 76a_D25GWACXX_5_250181133_X4
    - 81a_D25GWACXX_5_250181138_X4
    - 83a_D25GWACXX_5_250181140_X4
    - 8nm_D25GWACXX_6_250181280_X4
- Filtered sites with more than [20]% missing data, resulting in:
    - [16261] filtered, biallelic SNPs

***Please note:** We are providing unfiltered VCF SNP calls, and you may like to filter those SNPs to something that makes sense for your species/dataset. For this MDS analysis we filter based on missingness and sequencing depth, but not on allele frequency. Please see the GBS options listed above for default allele frequency filters when calling SNPs**.**


MDS plot of Cranioleuca
MDS Component 1
MDS Component 2
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.05
0.00
0.05
0.10
0.15
0.20
175_D25GWACXX_6_250181263_X4
73_D25GWACXX_5_250181130_X4
64b_D25GWACXX_5_250181121_X4
72b_D25GWACXX_5_250181129_X4
71_D25GWACXX_5_250181128_X4
95_D25GWACXX_5_250181152_X4
9nm_D25GWACXX_6_250181281_X4
99a_D25GWACXX_6_250181187_X4
109a_D25GWACXX_6_250181197_X4
110_D25GWACXX_6_250181198_X4
111a_D25GWACXX_6_250181199_X4
163a_D25GWACXX_6_250181252_X4
171_D25GWACXX_6_2501812
142_D25GWACXX_6_250181231_X4
183_D25GWACXX_6_250181268_X4
122_D25GWACXX_6_250181210_X4
5_D25GWACXX_6_250181225_X4
WACXX_5_250181062_X4

Genotyping by Sequencing Resources

GBS Overview

http://www.igd.cornell.edu/index.cfm/page/projects/GBS.htm

http://www.maizegenetics.net/gbs-overview

GBS Frequently Asked Questions

http://www.igd.cornell.edu/index.cfm/page/projects/GBS/gbsfaq.htm

GBS Bioinformatics

http://www.maizegenetics.net/gbs-bioinformatics

Training

http://www.igd.cornell.edu/index.cfm/page/Education/workshops/GBSworkshop.htm

http://www.maizegenetics.net/training-sessions

TASSEL

General Information

http://sourceforge.net/projects/tassel/

http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119

Tassel 3.0 GBS pipeline documentation

http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf

Tassel 3.0/4.0 pipeline documentation

http://www.maizegenetics.net/tassel/docs/TasselPipelineCLI.pdf

Tassel Users Group

https://groups.google.com/forum/?fromgroups#!forum/tassel

How to Cite:

GBS:

GBS libraries were prepared and analyzed at the Institute for Genomic Diversity (IGD), according to Elshire et al. (2011), using the enzyme [enzyme] for digestion and creating a library with [number] unique barcodes.

GBS Bioinformatics:

The GBS UNEAK analysis pipeline [version], an extension to the Java program TASSEL (Bradbury et al. 2007), wasused to call SNPs from the sequenced GBS library with the following options [options].

SNP analysis:

VCFtools (v0.1.10) (Danecek et al 2011) was used to summarize data, to filter data and to generate input files for PLINK (Purcell et al2007), which were used for MDS (multidimensional scaling). Analyses were visualized using basic plotting functions in R version 2.15.2 (R Development Core Team 2008).

Bibliography:

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6(5): e19379. doi:10.1371/ journal.pone.0019379

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. “TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples.” Bioinformatics 23, no. 19 (June 22, 2007): 2633–2635.

Lu, F, Lipka, AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES and Costich DE. “Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol”. PLoS Genet 9(1):e1003215. Doi:10.1371/journal.pgen.1003215

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group. The Variant Call Format and VCFtools. Bioinformatics, 2011

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81, URL:http://pngy.mgh.harvard.edu/purcell/plink

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Tassel Pipeline Arguments:

Tassel Pipeline Arguments: -fork1 -UCreatWorkingDirPlugin -w . -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UQseqToTagCountPlugin -s 300000000 -w . -e PstI -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UFastqToTagCountPlugin -s 300000000 -w . -e PstI -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UMergeTaxaTagCountPlugin -w . -c 5 -m 200000000 -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UTagCountToTagPairPlugin -w . -e 0.03 -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UTagPairToTBTPlugin -w . -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UTBTToMapInfoPlugin -w . -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UMapInfoToHapMapPlugin -w . -mnMAF 0.05 -mxMAF 0.5 -mnC 0 -mxC 1 -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -UFasToTOPMPlugin -w . -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -GBSHapMapFiltersPlugin -hmp ./hapMap/HapMap.hmp.txt -o ./hapMap/HapMap.filtered.hmp.txt -mnTCov 0.1 -mnSCov 0.8 -mnMAF 0.01 -mxMAF 1 -sC 0 -eC 0 -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -tbt2vcfPlugin -i ./tagsByTaxa/tbt.bin -m ./mapInfo/fas.topm.txt -o ./vcf -ak 3 -mnMAF 0.05 -mnLCov 0 -s 1 -e 1 -endPlugin -runfork1

Tassel Pipeline Arguments: -fork1 -MergeDuplicateSNP_vcf_Plugin -i ./vcf/tagsByTaxa.c1 -o ./vcf/c1.mergedSNPs.vcf -ak 3 -endPlugin -runfork1