Bioinformatics workshop

3rd SoyaGen Annual Meeting

7 December 2018

Basic data handling, diversity analysis and phylogenetic trees, parental line selection and confirmation of hybrid status with TASSEL

Martine Jean
3rd SoyaGen Annual Meeting, December 7th, 2018
Université Laval, Québec

Overview

- **Getting started**
  - Opening and saving SNP files in TASSEL
  - Getting an overview of your dataset
  - Clarifying some genetic concepts and limitations of GBS analysis
- **Basic data filtrations**
  - Filtering datasets with the "Filter Genotype Table" functions
  - Filtering low-quality SNP loci
    - Removing loci not assigned to a chromosome (scaffolds)
    - Removing loci with excess of heterozygous genotypes
    - Removing loci with low minor allele frequency
  - Filtering low-quality samples
- **Basic genetic analyses**
  - Selecting parental lines
  - Testing putative F1 hybrids
  - Using haplotypes to predict allelic states at important agronomic genes

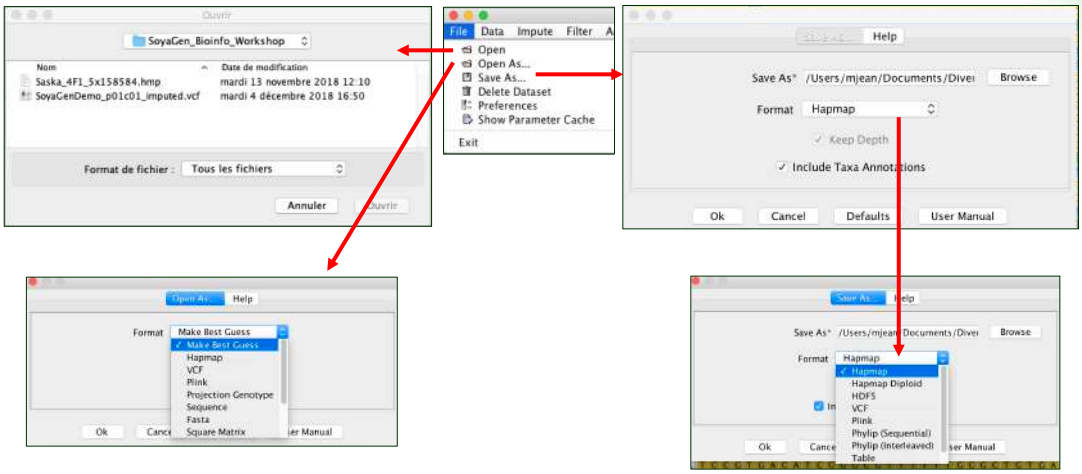
Overview

- **Getting started**
  - Opening and saving SNP files in TASSEL
  - Getting an overview of your dataset
  - Clarifying some genetic concepts and limitations of GBS analysis
- **Basic data filtrations**
  - Filtering datasets with the "Filter Genotype Table" functions
  - Filtering low-quality SNP loci
    - Removing loci not assigned to a chromosome (scaffolds)
    - Removing loci with excess of heterozygous genotypes
    - Removing loci with low minor allele frequency
  - Filtering low-quality samples
- **Basic genetic analyses**
  - Selecting parental lines
  - Testing putative F1 hybrids
  - Using haplotypes to predict allelic states at important agronomic genes

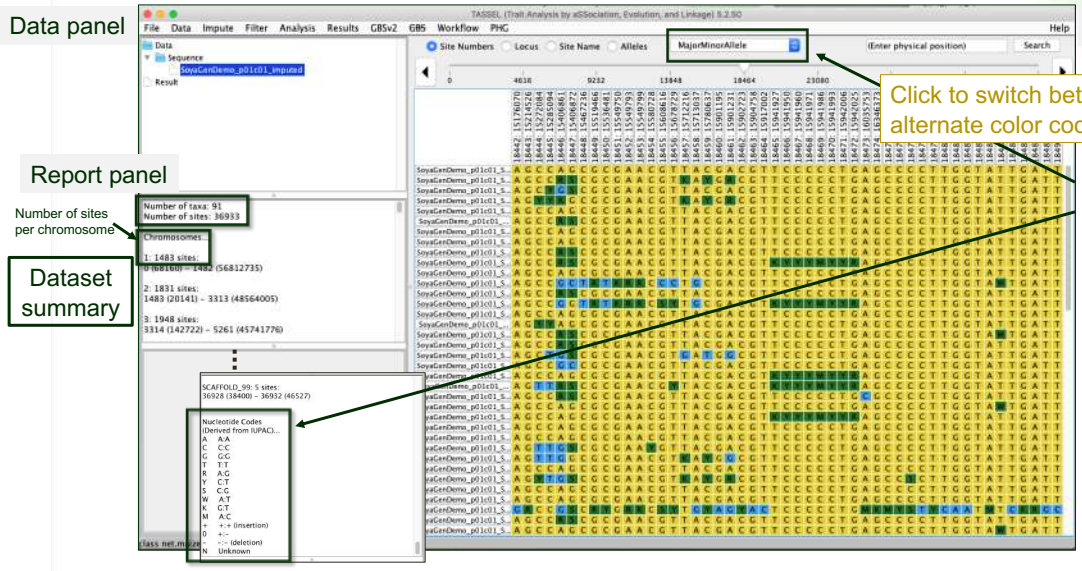
Getting started: Opening and saving SNP files in TASSEL

Opening SNP files

Saving SNP files



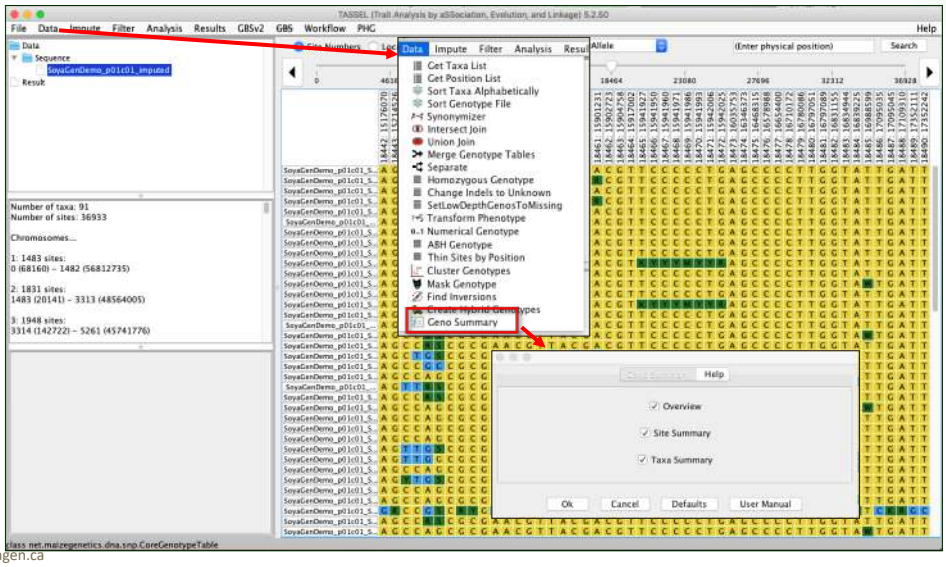
Getting an overview of your dataset



www.soyagen.ca

5

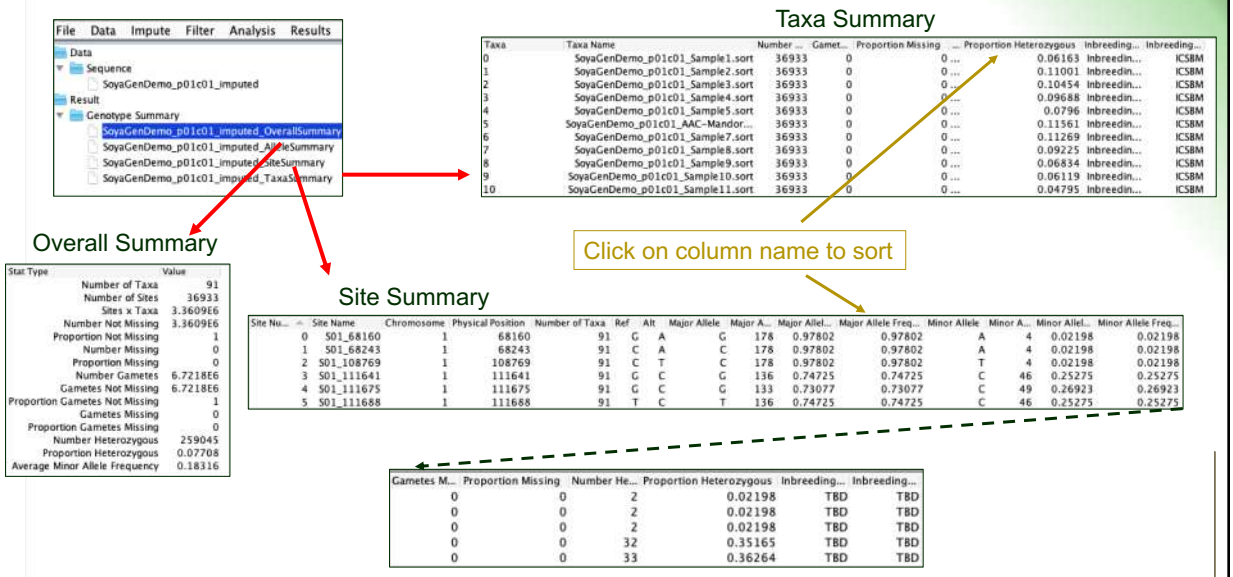
Using the "Geno Summary" function to get a more detailed overview of your dataset



www.soyagen.ca

6

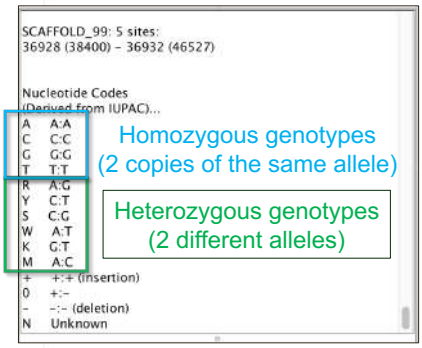
Outputs from the "Geno Summary" function



www.soyagen.ca

7

Clarifying some genetic concepts and limitations of GBS analysis



	Allele frequency					
Allele1	**0,01**	**0,25**	**0,49**	**0,51**	**0,75**	**0,99**
Allele2	**0,99**	**0,75**	**0,51**	**0,49**	**0,25**	**0,01**

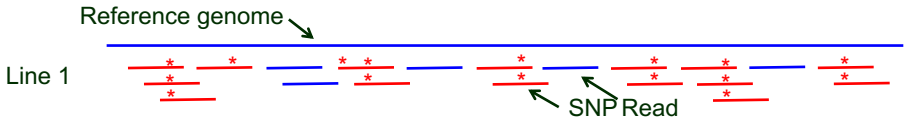
- Chromosomes vs scaffolds
  - Scaffolds are genomic sequences that have not yet been assigned to a chromosome
  - They are not useful for most genetic analyses
- Genotypes vs alleles
  - **Minor allele** vs **major allele**
  - Minor allele frequency (MAF)
- Limitations of GBS analysis
  - Large amount of missing data
  - False homozygotes due to low read depth
  - False heterozygotes due to genome duplication and structural variation

www.soyagen.ca

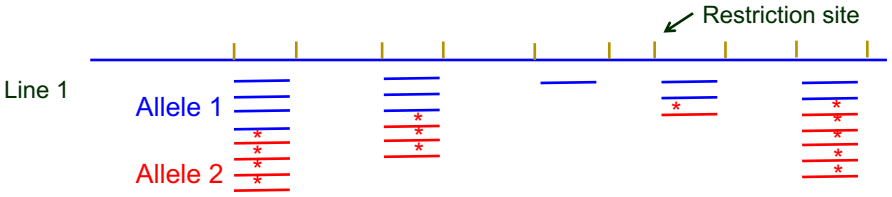
8

Clarifying some limitations of GBS analysis

Whole genome resequencing



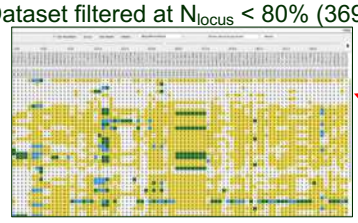
Genotyping-by-sequencing (GBS, Elshire *et al.* 2011)



- Despite the increase in read depth achieved with GBS, no or very few reads will be available at some loci in some samples
  - This will create missing data and false homozygous genotypes

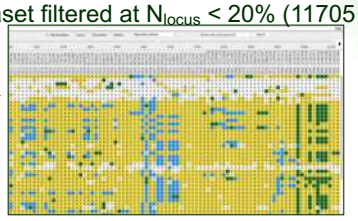
Dealing with the large amount of missing data of GBS datasets: filtration vs imputation

Dataset filtered at N_{locus} < 80% (36933 SNP)

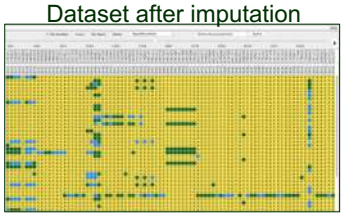


Lines with high N frequency

Dataset filtered at N_{locus} < 20% (11705 SNP)

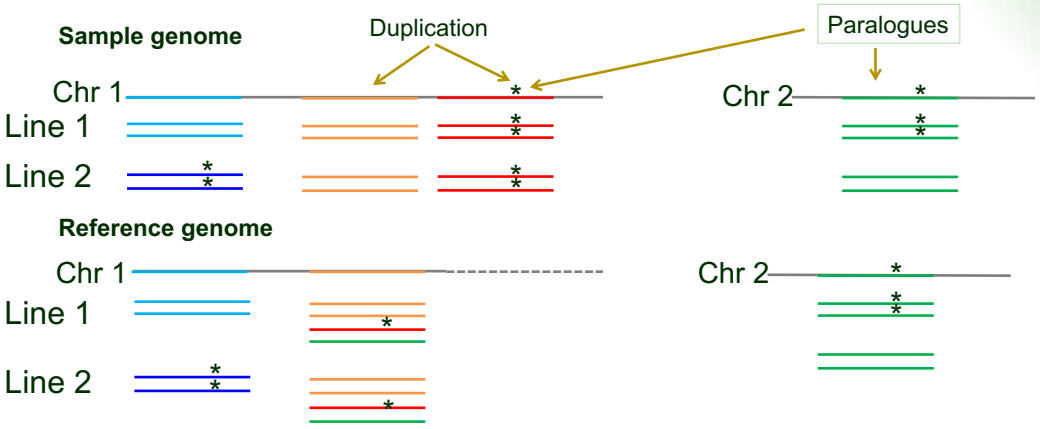


Dataset after imputation



- Basic filtration thresholds for missing data per locus
  - N_{locus} < 80% (dataset to be imputed)
  - N_{locus} < 10-20% (dataset to be used without imputation)
- Imputed data are accurate enough for most genetic analyses
  - Identifying lines displaying large numbers of missing data before imputation can be useful as some of their imputed genotypes may be less accurate
- Filtering a dataset to remove loci with high missing data frequency will create a smaller dataset
  - However, removing loci with high missing data frequency will also remove most of the false homozygous genotypes

Clarifying some limitations of GBS analysis

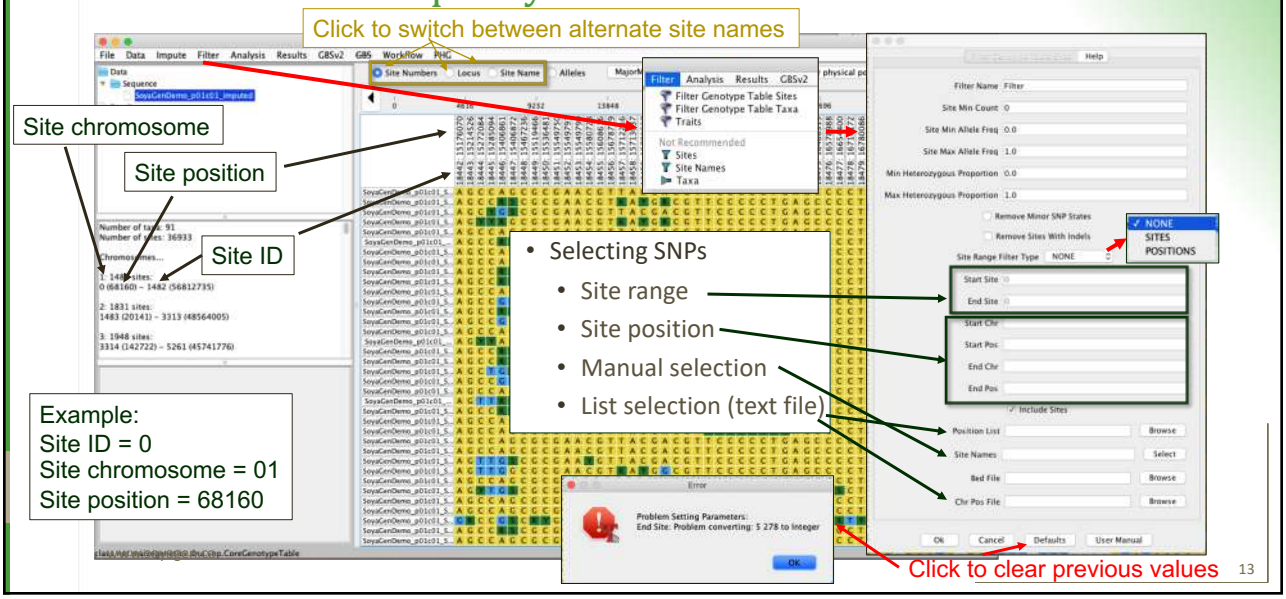


- Paralogues and line-specific structural variations (duplication, insertion, deletion) can be scored as heterozygous genotypes by SNP calling softwares
- Removing loci displaying large amount of heterozygous genotypes will eliminated most of those artefactual loci

Overview

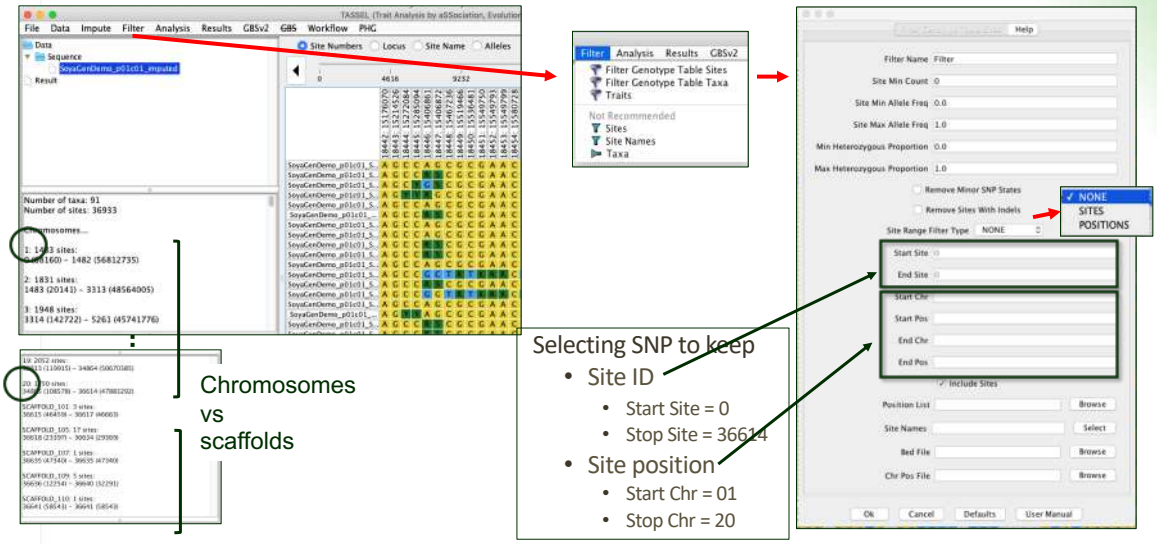
- **Getting started**
  - Opening and saving SNP files in TASSEL
  - Getting an overview of your dataset
  - Clarifying some genetic concepts and limitations of GBS analysis
- **Basic data filtrations**
  - Filtering datasets with the "Filter Genotype Table" functions
  - Filtering low-quality SNP loci
    - Removing loci not assigned to a chromosome (scaffolds)
    - Removing loci with excess of heterozygous genotypes
    - Removing loci with low minor allele frequency
  - Filtering low-quality samples
- **Basic genetic analyses**
  - Selecting parental lines
  - Testing putative F1 hybrids
  - Using haplotypes to predict allelic states at important agronomic genes

Using the "Filter Genotype Table Sites" function to filter low-quality SNP and create SNP subsets



13

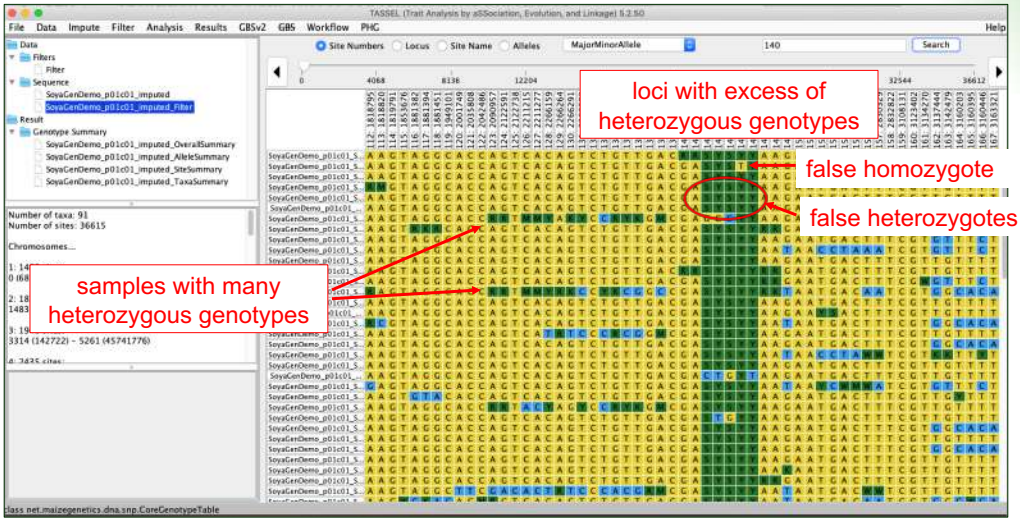
Identifying and removing SNP loci not assigned to a chromosome



www.soyagen.ca

14

Identifying loci with excess of heterozygous genotypes



Removing loci with excess of heterozygous genotypes

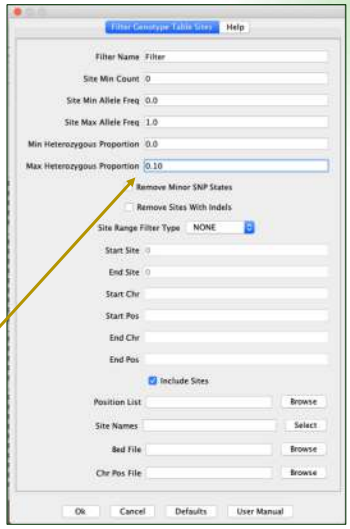
Site Summary

Site Number	Site Name	Chromoso...	Physical Po...
23847	S15_1959...	15	19591353
33521	S19_1839...	19	18391489
1913	S02_8311...	02	8311495
8571	S05_3624...	05	36243358
8572	S05_3624...	05	36243387
8573	S05_3624...	05	36243393
8591	S05_3625...	05	36253774
8592	S05_3625...	05	36253786
8594	S05_3625...	05	36253819
9121	S06_7139...	06	7139654
9122	S06_7139...	06	7139724
9123	S06_7139...	06	7139731
15982	S10_1260...	10	1260836
21356	S13_4080...	13	40807071
21357	S13_4080...	13	40807080
21358	S13_4080...	13	40807086
23308	S15_6759...	15	6759025
23511	S15_1216...	15	12163313

....

Only 1 homozygous call

Gametes M...	Proportion...	Number He...	Propor...	Inbreeding...	Inbreeding...
0	0	90	0.98901	TBD	TBD
0	0	90	0.98901	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD
0	0	89	0.97802	TBD	TBD

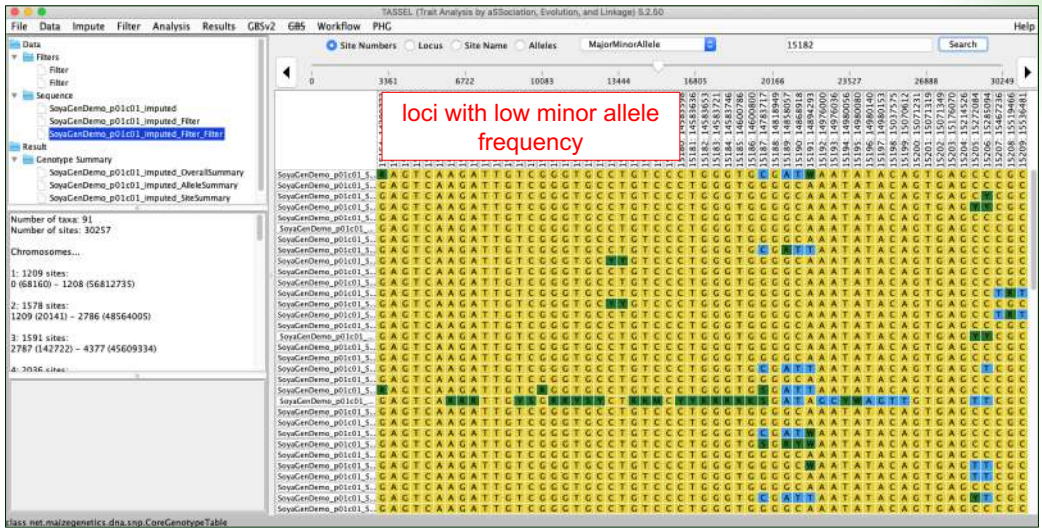


- In a collection of inbred lines, the frequency of heterozygous genotypes should be very low
- SNP loci with more than 10% heterozygous calls are likely to be SNP calling artefacts and should be removed

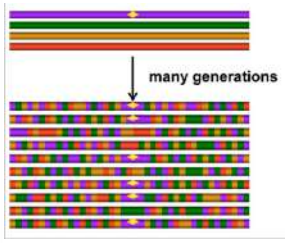
Basic filtration threshold

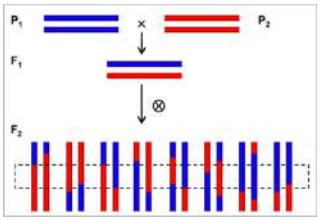
- H < 10% (collection of inbred lines)

Identifying SNP loci with low minor allele frequency



Selecting the appropriate filtration threshold for low minor allele frequency





(Zhou et al 2008 The Plant Genome 1:5–20)

- Genome-wide association analysis (GWAS)
  - Basic threshold: minMAF > 0,05
  - Minimum minor allele frequency (minMAF) needed to detect reliable marker-trait association
    - 5/100 alleles = combination of 5 homo to 10 hetero per 100 samples
  - Special cases: Very small or very large populations
    - minMAF=0,05 may be too relax if N_{taxa} < 20 (0,05x20=1)
    - minMAF=0,05 may be too strict if N_{taxa} > 1000 (0,05x1000=50)
- Biparental QTL mapping analysis
  - Basic threshold: minMAF > 0,30
  - Maximum deviation tolerate from the expected allele frequency = 0,5 to avoid creating false linkage groups during genetic map construction
- Most other analyses
  - Basic threshold: minMAF > 0,05

Removing loci with low minor allele frequency

Site Summary

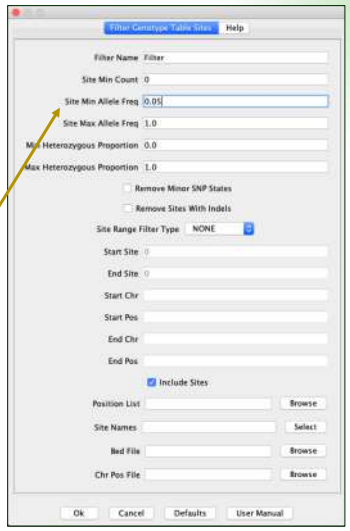
1 minor allele = 1 heterozygote

Site Number	Site Name	Chromoso...	Physical Po...
7	S01_1251...	01	125159
46	S01_6692...	01	669257
48	S01_6694...	01	669442
69	S01_1199...	01	1199174
93	S01_1783...	01	1783133
118	S01_2288...	01	2288627
123	S01_2290...	01	2290523
125	S01_2293...	01	2293052
129	S01_2730...	01	2730684
137	S01_2832...	01	2832822
139	S01_3123...	01	3123402
154	S01_3278...	01	3278993
165	S01_3550...	01	3550442
167	S01_3650...	01	3650455
182	S01_3947...	01	3947327
183	S01_3947...	01	3947363

....

Minor Allele	Minor Allel...	Minor Allel...	Minor ...	A
T	1	0.00549	0.00549	
C	1	0.00549	0.00549	
T	1	0.00549	0.00549	
C	1	0.00549	0.00549	
C	1	0.00549	0.00549	
T	1	0.00549	0.00549	
A	1	0.00549	0.00549	
T	1	0.00549	0.00549	
C	1	0.00549	0.00549	
A	1	0.00549	0.00549	
A	1	0.00549	0.00549	
C	1	0.00549	0.00549	
C	1	0.00549	0.00549	
G	1	0.00549	0.00549	
A	1	0.00549	0.00549	
A	1	0.00549	0.00549	

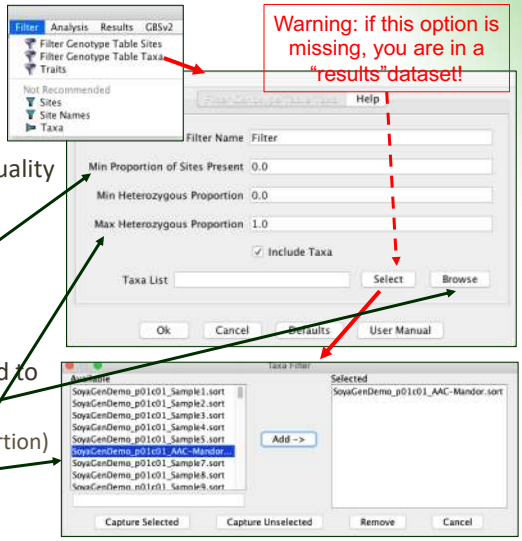
- **Basic filtration thresholds**
  - minMAF > 0,05 (GWAS)
  - minMAF > 0,30 (QTL mapping)
  - minMAF > 0,05 (most other analyses)



Filtering low-quality samples

Taxa Summary

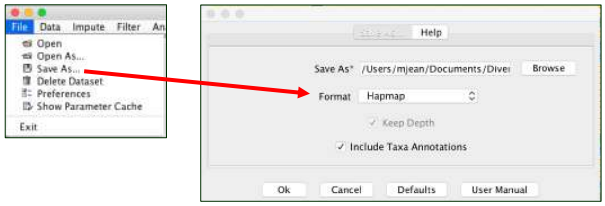
Taxa Name	Number of ...	Gam...	Prop...	Number He...	Proportion Heterozygous
SoyaGenDemo_p01c01_Sample2.sort	18803	0	0	2645	0.14067
SoyaGenDemo_p01c01_Sample26.sort	18803	0	0	2610	0.13881
SoyaGenDemo_p01c01_AAC-Mandor.sort	18803	0	0	2455	0.13056
SoyaGenDemo_p01c01_Sample3.sort	18803	0	0	2408	0.12806
SoyaGenDemo_p01c01_Sample4.sort	18803	0	0	1995	0.1061
SoyaGenDemo_p01c01_Sample7.sort	18803	0	0	1993	0.10599
SoyaGenDemo_p01c01_S03-W4.sort	18803	0	0	1831	0.09738
SoyaGenDemo_p01c01_Sample41.sort	18803	0	0	1686	0.08967



- The "Taxa Summary" output can be used to identify low-quality samples
  - Unexpected excess of heterozygous genotypes in a cultivar is usually indicative of a technical problem
    - DNA sample contamination
    - Accidental outcross
- The "Filter Genotype Table Taxa" function can than be used to remove those samples from the dataset
  - Filtration thresholds (missing data or heterozygous proportion)
  - Sample list in a text file
  - Manual sample selection

Renaming file in Excel: Step 1

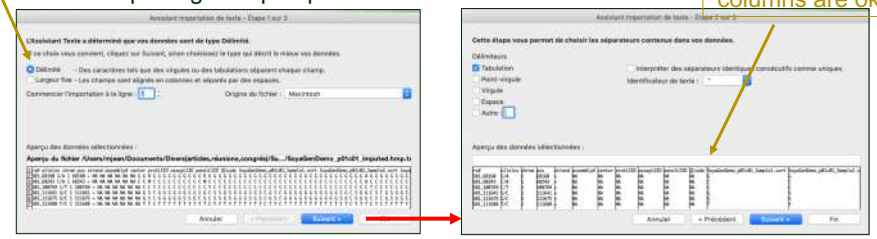
Exporting a SNP file from TASSEL in Hapmap (hmp) format



Select “Delimited”

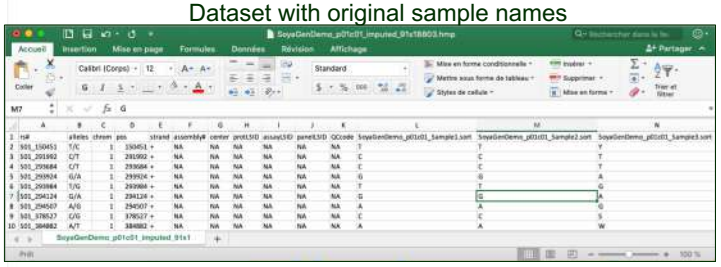
Importing a Hapmap file in Excel as a text file

Make sure that columns are ok

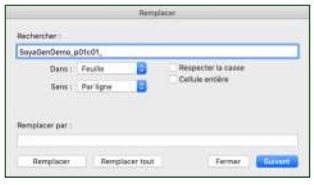


Renaming file in Excel: Step 2

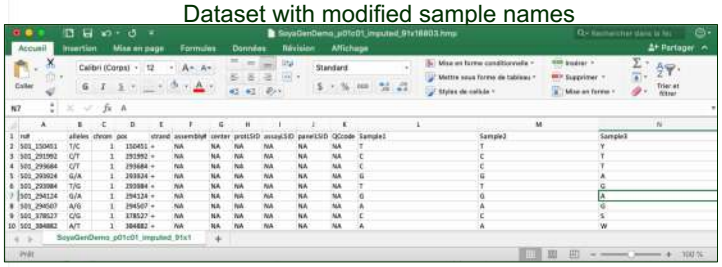
Dataset with original sample names



Using the “replace” function from Excel to modify sample names



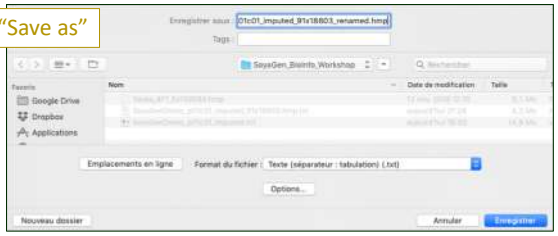
Dataset with modified sample names

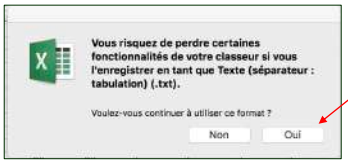


Renaming file in Excel: Step 3

Exporting a Hapmap file as a text file from Excel

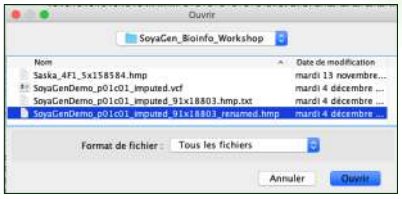
Select “Save as”



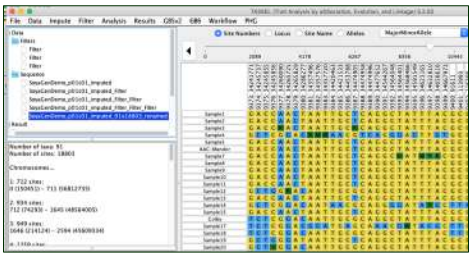


Click ”yes”

Importing the modified SNP file in TASSEL



Dataset with renamed lines



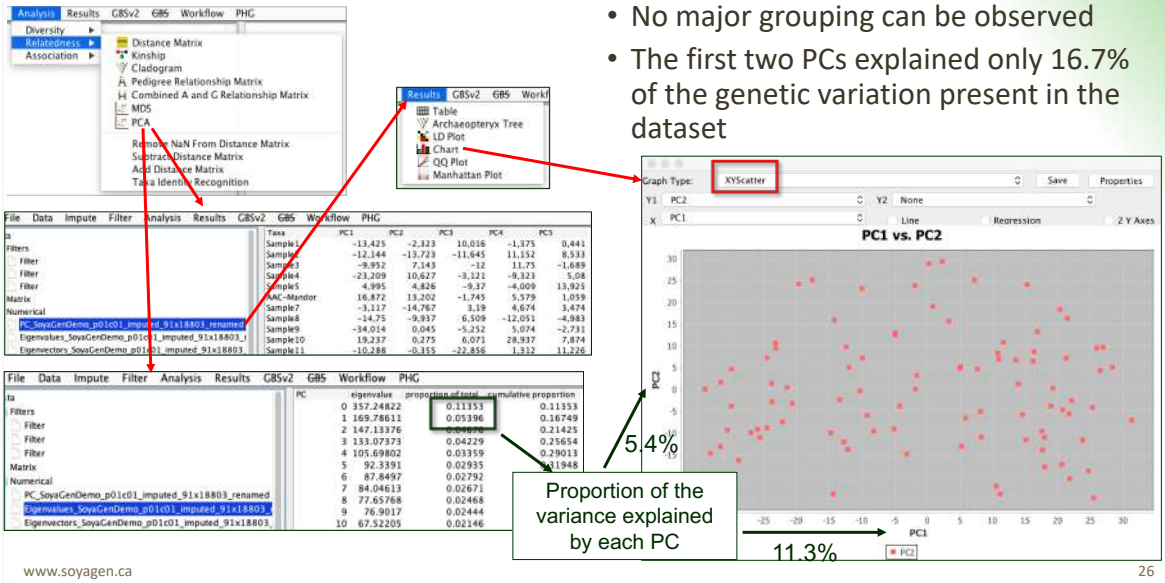
Overview

- **Getting started**
  - Opening and saving SNP files in TASSEL
  - Getting an overview of your dataset
  - Clarifying some genetic concepts and limitations of GBS analysis
- **Basic data filtrations**
  - Filtering datasets with the “Filter Genotype Table” functions
  - Filtering low-quality SNP loci
    - Removing loci not assigned to a chromosome (scaffolds)
    - Removing loci with excess of heterozygous genotypes
    - Removing loci with low minor allele frequency
  - Filtering low-quality samples
- **Basic genetic analyses**
  - Selecting parental lines
  - Testing putative F1 hybrids
  - Using haplotypes to predict allelic states at important agronomic genes

Selecting parental lines to maximize allele diversity

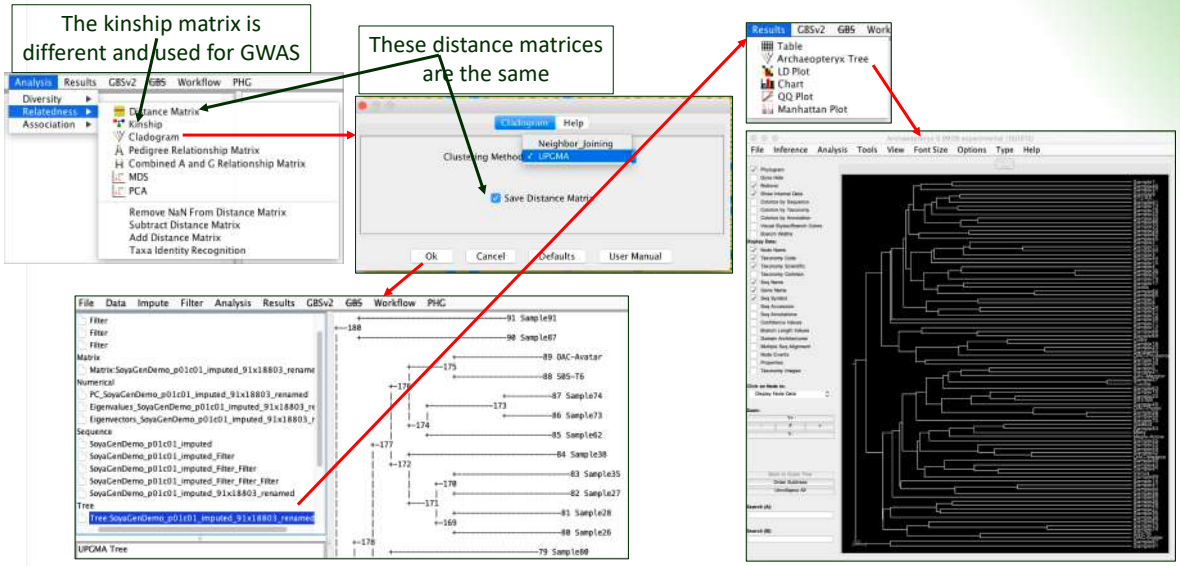
- Objective
  - Identify parental combinations that will maximize allele diversity
- Protocol
  - Step 1. Use the “PCA” and “Chart” functions to create a PCA plot
  - Step 2. Use the “Cladogram” function create a distance matrix and calculate the branching pattern needed to make a tree
  - Step 3. Create a tree with the “Make tree” function to visualize the overall relationship between the lines from the dataset
  - Step 4. Examine the distance matrix to compare lines and select parental combinations that will maximize allele diversity

Using the “PCA” and “Chart” functions to create a PCA plot



- No major grouping can be observed
- The first two PCs explained only 16.7% of the genetic variation present in the dataset

Using the "Cladogram" and "Archaeopteryx Tree" functions to create a tree



How the distance matrix is calculated by TASSEL

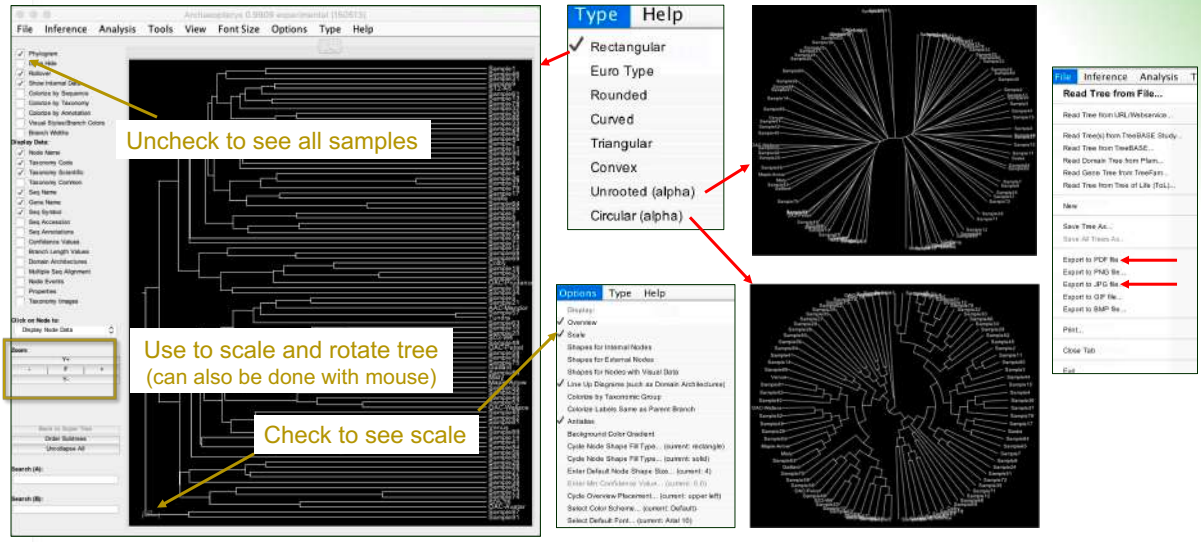
Genotypes:

Line_1	Line_2	Line_3	Line_4	Line_5	
A	G	R	A	A	0
A	A	A	A	G	1
A	A	A	A	R	0.5
A	A	A	R	R	0.5
A	A	A	R	A	0.5
	1/5=0.2	0.5/5=0.1			2.5/5=0.5

Distance matrix:

Taxa	Line_1	Line_2	Line_3	Line_4	Line_5
Line_1	0	0.2	0.1	0.2	0.4
Line_2	0.2	0	0.1	0.4	0.6
Line_3	0.1	0.1	0	0.3	0.5
Line_4	0.2	0.4	0.3	0	0.5
Line_5	0.4	0.6	0.5	0.5	0

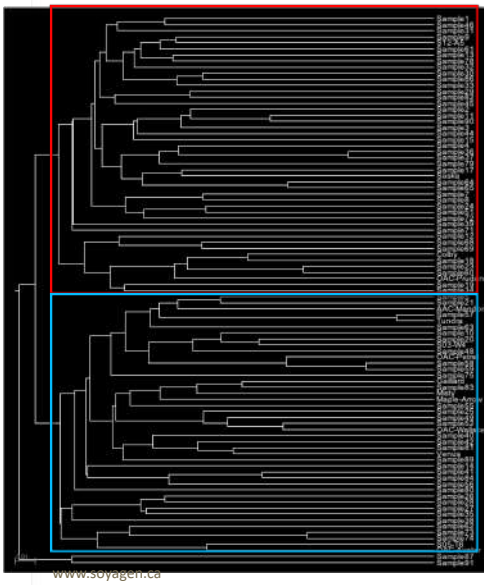
Formatting and saving the tree



www.soyagen.ca

29

Using the tree to visualize the overall relationship between the lines from the dataset



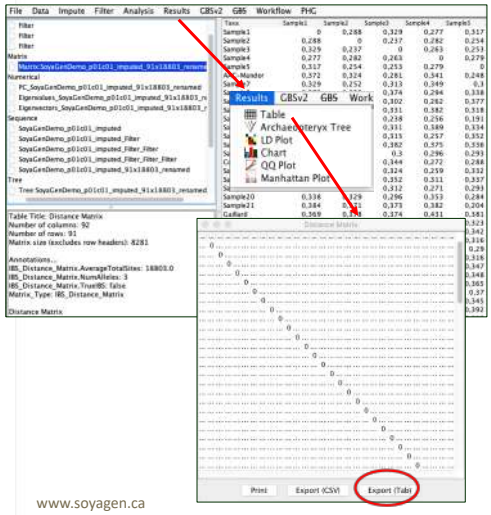
- The dataset is composed of two main groups of lines and several subgroups can be also observed in each main group
  - Two lines (Sample87and Sample91) are very different from the others and form a small third group
- Crossing lines from different groups should maximize diversity
- However, a cladogram is not the best tool to identify the most divergent lines
- Examining the distance matrix used to generate the cladogram is a more precise way to compare lines and select the most divergent parental combination

www.soyagen.ca

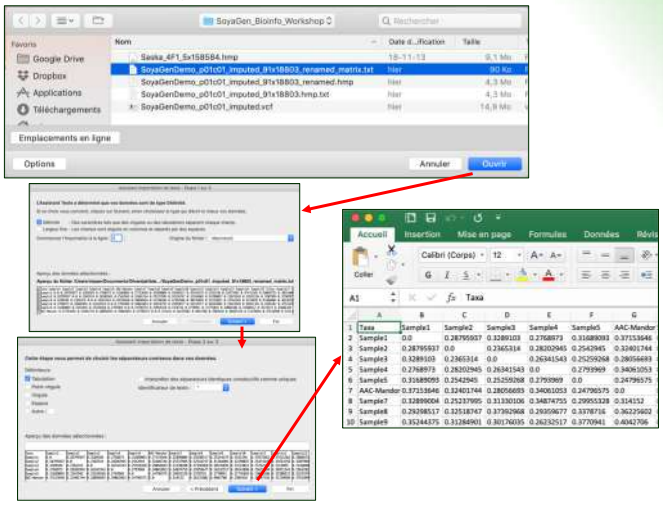
30

Color-coding the distance matrix in Excel: Step 1

Exporting the distance matrix in text format

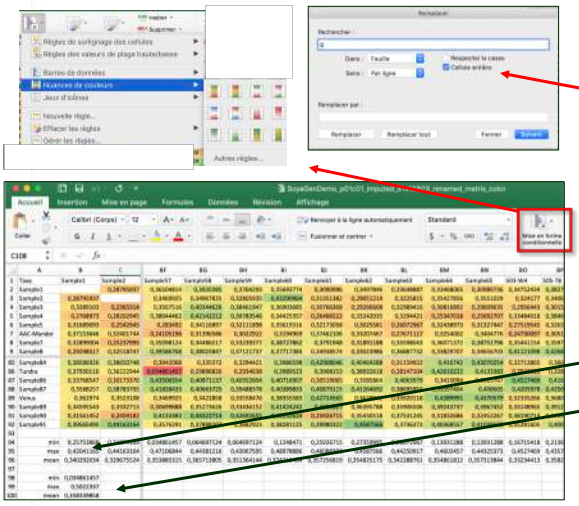


Importing the distance matrix in Excel



www.soyagen.ca

Color-coding the distance matrix in Excel: Step 2

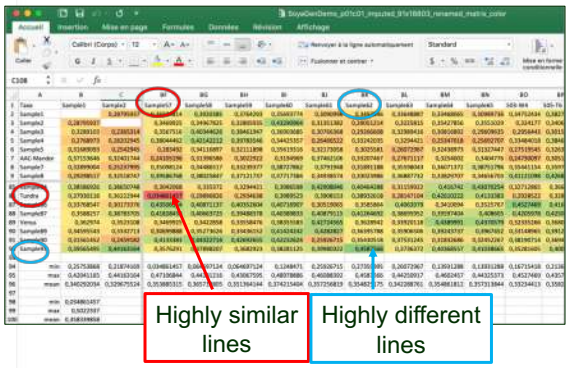


- How to color-coding the distance matrix in Excel
  - Step 1. Change "." to "," with the "Replace" menu
  - Step 2. Select the "Entire cell" option from the "Replace" menu, then replace "0" by nothing
  - Step 3. Color-code the matrix using the "Conditional Formatting" menu and the "Color Scales" option
- Calculating the min, max and mean distance value per line and for the entire dataset will help you find the most similar and different lines

www.soyagen.ca

32

Using a distance matrix to compare lines and select parental combination that maximize allele diversity

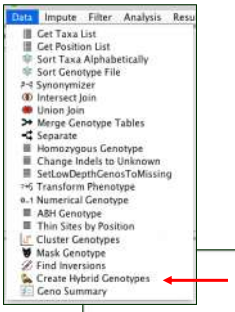


- Use the color-code the distance matrix to visualize the level of similarity between the lines
- Use the min and max values per line to identify the lines that are the most similar or different to your line of interest
  - Unexpected very high similarity between lines may indicate sample mislabelling or breeding problems
- Use the max value for the entire dataset to identify the lines that would make the most divergent parental combination

Hybridity testing

- Objectives
  - Discriminating F1 and self
- Protocol
  - Step 1. Merge the F1 dataset to a reference dataset with the “Merge Genotype Table” function
  - Step 2. Generate a tree from the merged dataset
  - Step 3. Remove SNP loci unique to a single dataset
  - Step 4. Create a tree to visualize the lines
  - Step 3. Compare the tree position of the putative F1 hybrids to the ones of its parental lines to discriminating true F1 hybrids from accidental selfing of the maternal parent

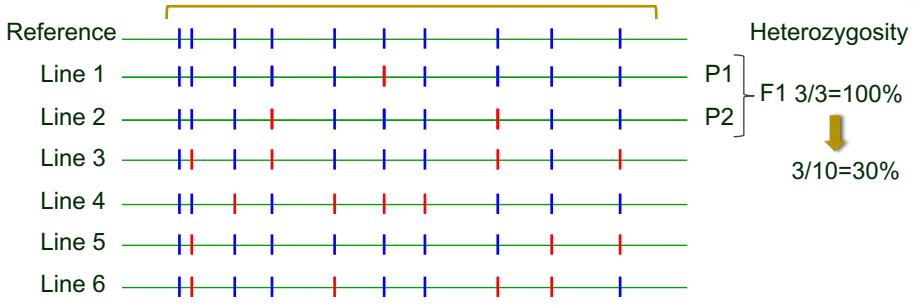
An example: discriminating simulated F1 hybrids and simulated accidental selfing



- A simulated F1 dataset was derived from the SoyaGen dataset with the "Create Hybrid Genotypes" function
  - 1 simulated selfing
  - 4 simulated F1 hybrids

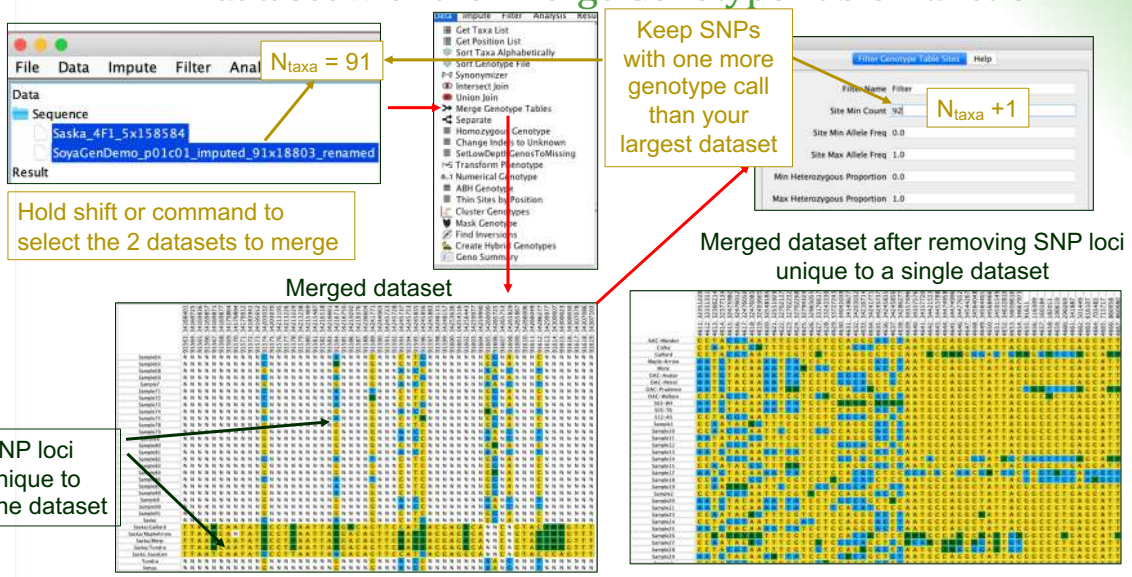
	42495: 13148477	42496: 13148514	42497: 13166274	42498: 13166275	42499: 13166280	42500: 13166720	42501: 13175447	42502: 13175448	42503: 13175465	42504: 13197027	42505: 13202727	42506: 13212339	42507: 13212402	42508: 13212535	42509: 13212557	42510: 13212761	42511: 13224464	42512: 13242558	42513: 13242594	42514: 13248959	42515: 13248970	42516: 13249073	42517: 13249093	42518: 13249210	42519: 13264910	42520: 13264921	42521: 13264936	42522: 13282125	42523: 13282188	42524: 13282204	42525: 13294324	42526: 13306228	42527: 13308026	42528: 13308209	42529: 13310531	42530: 13310698	42531: 13310702	42532: 13316711	42533: 13316717	42534: 13316729	42535: 13316735	42536: 13316763
Saska_SoyaGen	C	T	C	T	A	T	C	A	G	T	T	A	A	C	C	C	A	G	G	A	A	T	C	G	G	G	T	A	T	A	T	A	G	G	A	G	A	A	T	A	T	T
Saska/Tundra	C	T	C	T	A	T	C	A	G	T	T	A	A	C	C	C	A	G	K	A	A	T	C	G	G	G	T	A	T	A	Y	A	G	G	A	G	A	A	Y	A	K	Y
Saska/Misty	Y	T	C	T	A	T	C	A	G	T	T	A	A	C	C	C	A	G	G	A	A	T	C	G	G	G	T	A	T	A	Y	A	G	R	A	G	A	A	Y	A	K	T
Saska/Gaillard	C	T	C	T	A	T	C	A	G	T	T	A	A	C	C	C	A	G	G	A	A	T	C	G	G	G	T	A	T	A	T	A	G	G	A	G	A	A	T	A	T	T
Saska/MapleArrow	Y	T	C	T	A	T	C	A	G	T	T	A	A	C	C	C	A	G	G	A	A	T	C	G	G	G	T	A	T	A	Y	A	G	R	A	G	A	A	Y	A	K	T

In GBS analysis, the proportion of heterozygosity observed in F1 hybrids will always seem low



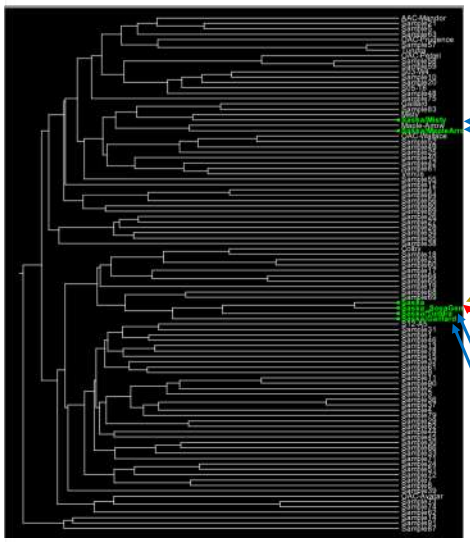
- In GBS analysis, the number of heterozygous genotypes detected in the F1 hybrids is limited by the number of SNPs polymorphic between the two parents among the total number of SNPs available in the dataset
  - In the example shown, one might expect 100% of SNP loci to be heterozygous in an F1 hybrid
  - However, it would only be 30% because most SNPs are not polymorphic between P1 and P2.

Merging a simulated F1 dataset to a reference dataset with the "Merge Genotype Table" function



Discriminating true F1 from accidental selfing of the maternal parent

Tree generated from the merged dataset



Simulated F1 hybrids

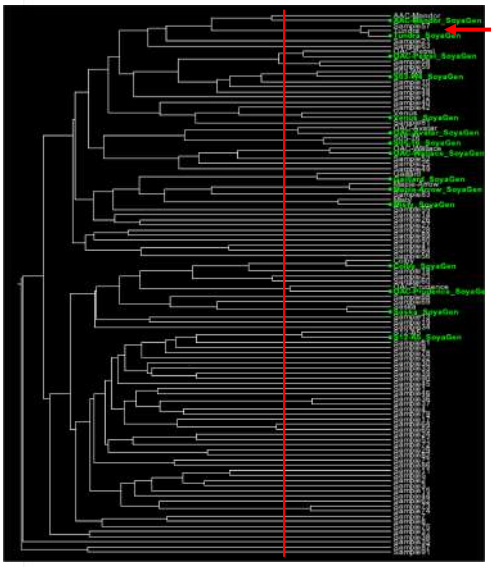
Simulated maternal parent

Simulated selfing (replicate of the same line)

Simulated F1 hybrids

- Selfing are very similar to maternal parent
- F1 hybrids are near one of their parents
  - Sometime the male, sometime the female
- The other parent can usually be confirmed by checking the distance matrix

Using a tree to identify unexpected highly related lines



- Some lines in the dataset seem to be as similar or more similar than replicates of the same line
  - Possible explanations:
    - Mislabelled samples
    - Accidental selfing of a maternal parent
    - Nearly isogenic lines (NILs) or very highly similar sister-lines

www.soyagen.ca (arbitrary line to facilitate comparison)

39

Using haplotypes to predict allelic states at important agronomic genes

- Objectives
  - Select lines with a desired trait using haplotype information around the gene of interest
- Protocol
  - Step 1. Get the genomic position of your gene of interest from a database
  - Step 2. Select SNPs surrounding your gene of interest in your dataset
  - Step 3. Use the “Linkage Disequilibrium” and ”LD plot’ functions to create a LD plot
  - Step 4. Identify SNPs that always co-segregate (LD=1) with your gene of interest
  - Step 5. Create haplotypes with the “Cluster genotype” function
  - Step 6. Create a tree to visualize lines belonging to each cluster
  - Step 7. Use reference lines to identify the allele present in uncharacterized lines

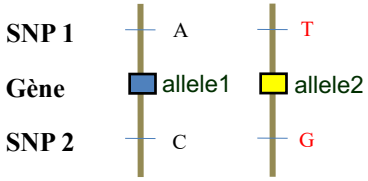
www.soyagen.ca

40

Late maturity cultivar | Early maturity cultivar

Gene	Gene model	Allele	Count	%
E1	Glyma.06g207800	*E1*	1	1.5
		e1-nl	39	58.2
		e1-as	25	37.3
		e1-fs	2	3.0

(Tardivel et al. 2018, in press)

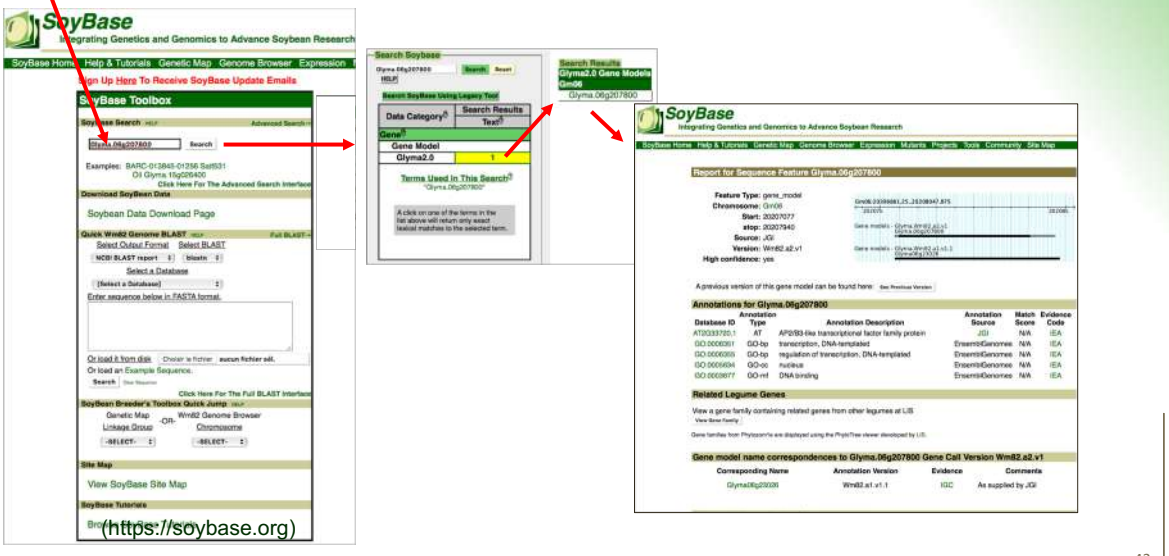


An example: finding haplotypes for the E1 gene

- The E1 gene is one of the most important genes controlling maturity in soybean
- Among the 136 Canadian lines characterized by Tardivel et al. (2108), most carry either the *e1-nl* or *e1-as* alleles.
- KASP assays could be used to determine the allele present at the E1 gene in multiple lines
  - Costly and specific to a single allele of a single gene
- SNPs from GBS dataset can often be grouped to create haplotypes specific for each allele of a gene of interest
  - Haplotypes for multiple genes of interest can be assessed in the same GBS dataset

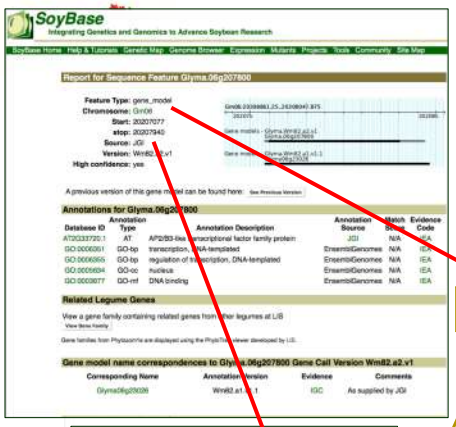
Gene	Gene model	Allele	Count	%
E1	Glyma.06g207800	*E1*	1	1.5
		e1-nl	39	58.2
		e1-as	25	37.3
		e1-fs	2	3.0

Finding the genomic position of a gene of interest in the SoyBase database



(https://soybase.org)

Selecting SNPs surrounding a gene of interest in your dataset

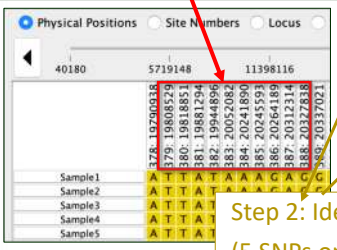


Step 1: Select Chromosome

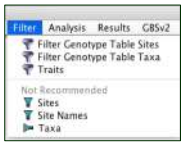
Position
Start Chr = 06
Stop Chr = 06

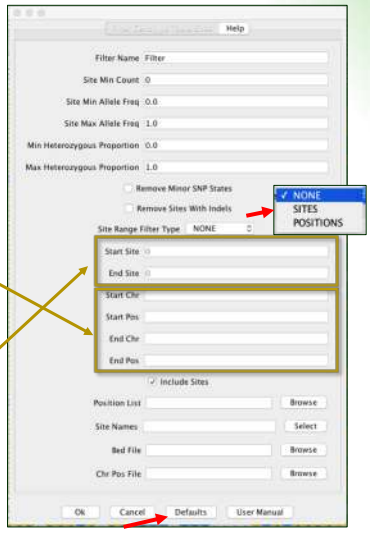
Step 3: Select SNP range

Site
Start Site = 379
Stop Site = 388



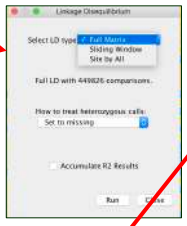
Step 2: Identify Site Range to keep
(5 SNPs on each side of the gene)

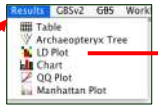


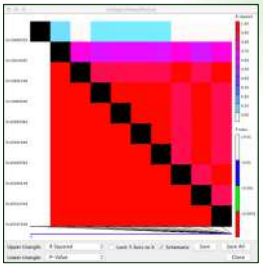


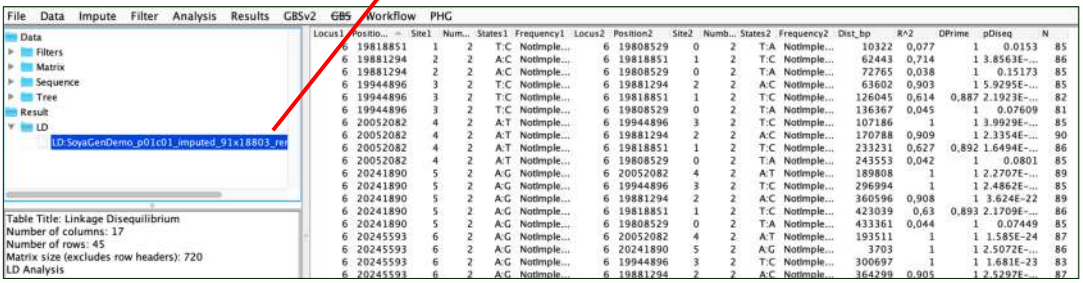
Click to clear previous values

Use the "Linkage Disequilibrium" and "LD plot' functions to create a LD plot



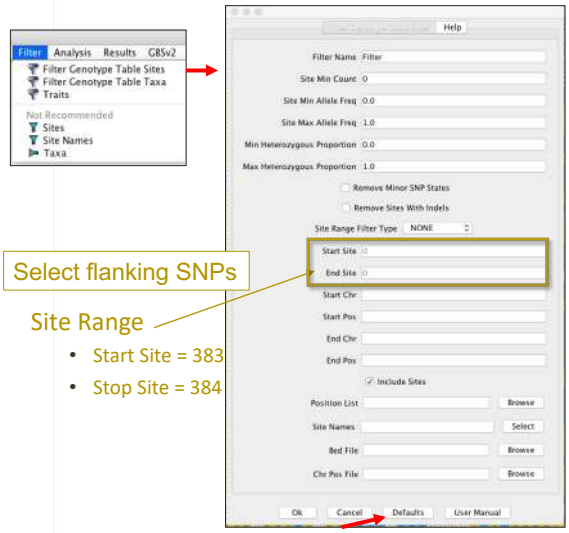




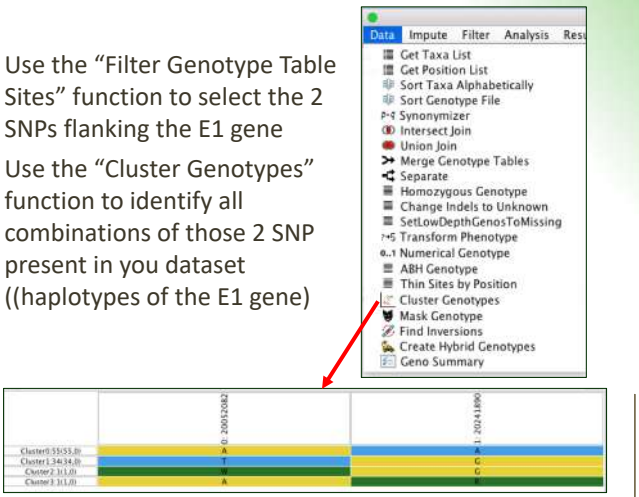


Use the LD plot to identify SNPs that always co-segregate with your gene of interest

Create haplotypes with the "Cluster genotype" function

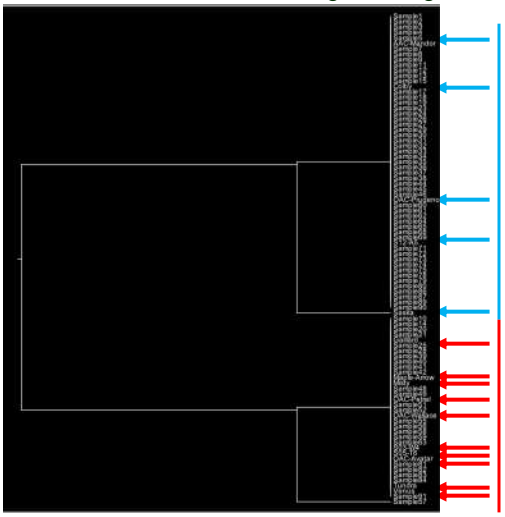


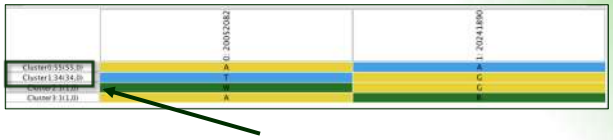
- Use the "Filter Genotype Table Sites" function to select the 2 SNPs flanking the E1 gene
- Use the "Cluster Genotypes" function to identify all combinations of those 2 SNP present in you dataset ((haplotypes of the E1 gene)



Using a tree made from flanking SNPs and reference lines to identify gene allele present in uncharacterized lines

Tree based on the 2 SNPs flanking the E1 gene

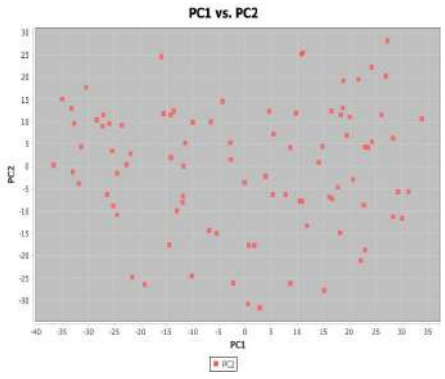


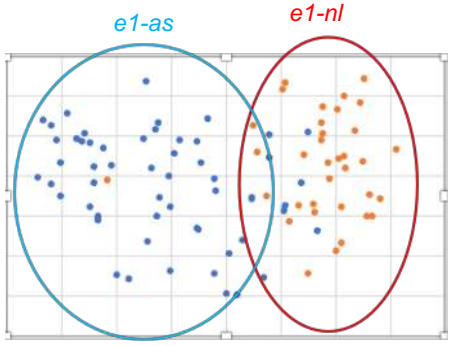


- There are only two main haplotypes in this dataset
- There are only two alleles found in the 15 lines characterized by Tardivel et al. (2018) present in this dataset
  - 5 with the *e1-as* allele
  - 10 with the *e1-nl* allele
- Uncharacterized lines are expected to carry the same allele as the reference lines found on the same branch of the tree

Color-coding sometimes highlights additional information about a dataset

- Selection for early maturity is probably a main factor explaining the genetic variation present in this dataset





PCA graph color-coded in Excel