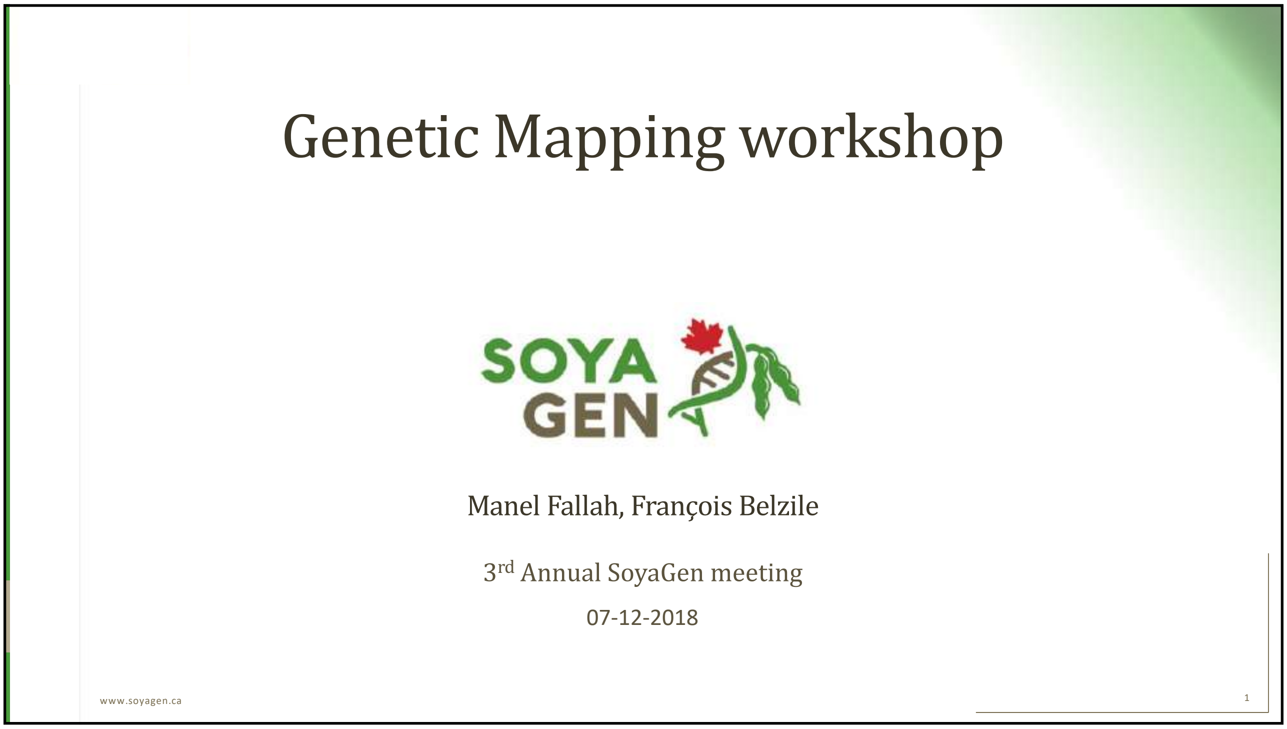
SOYA GEN

Bioinformatics workshop

3rd SoyaGen Annual Meeting

7 December 2018

UNIVERSITÉ LAVAL

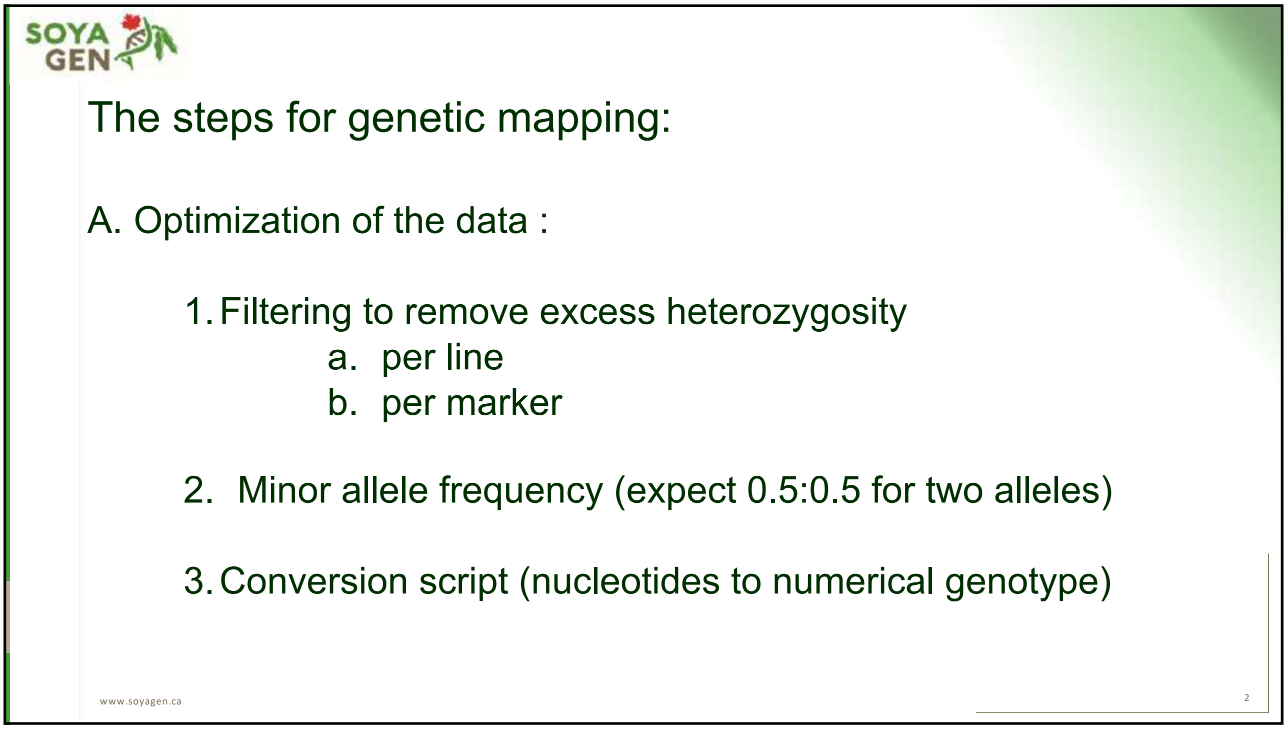


Genetic Mapping workshop

Manel Fallah, François Belzile

3rd Annual SoyaGen meeting

07-12-2018



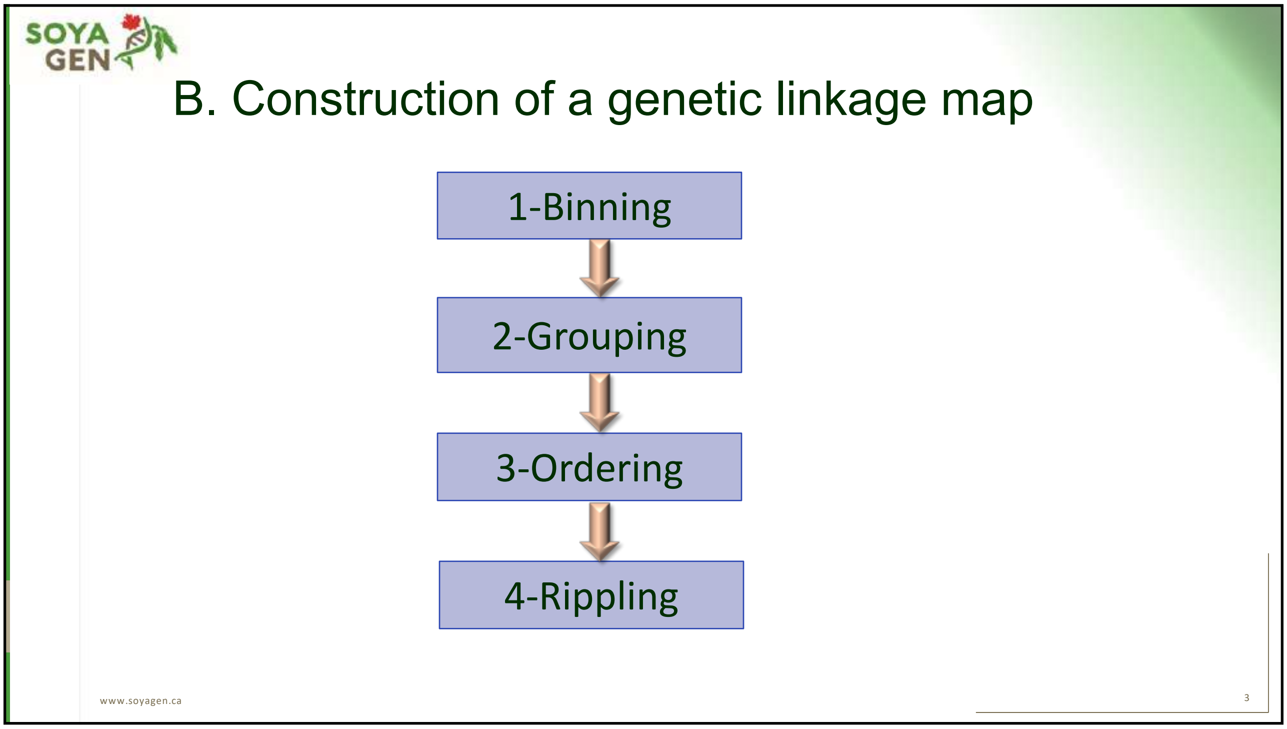
The steps for genetic mapping:

A. Optimization of the data :

1. Filtering to remove excess heterozygosity
   a. per line
   b. per marker

2. Minor allele frequency (expect 0.5:0.5 for two alleles)

3. Conversion script (nucleotides to numerical genotype)



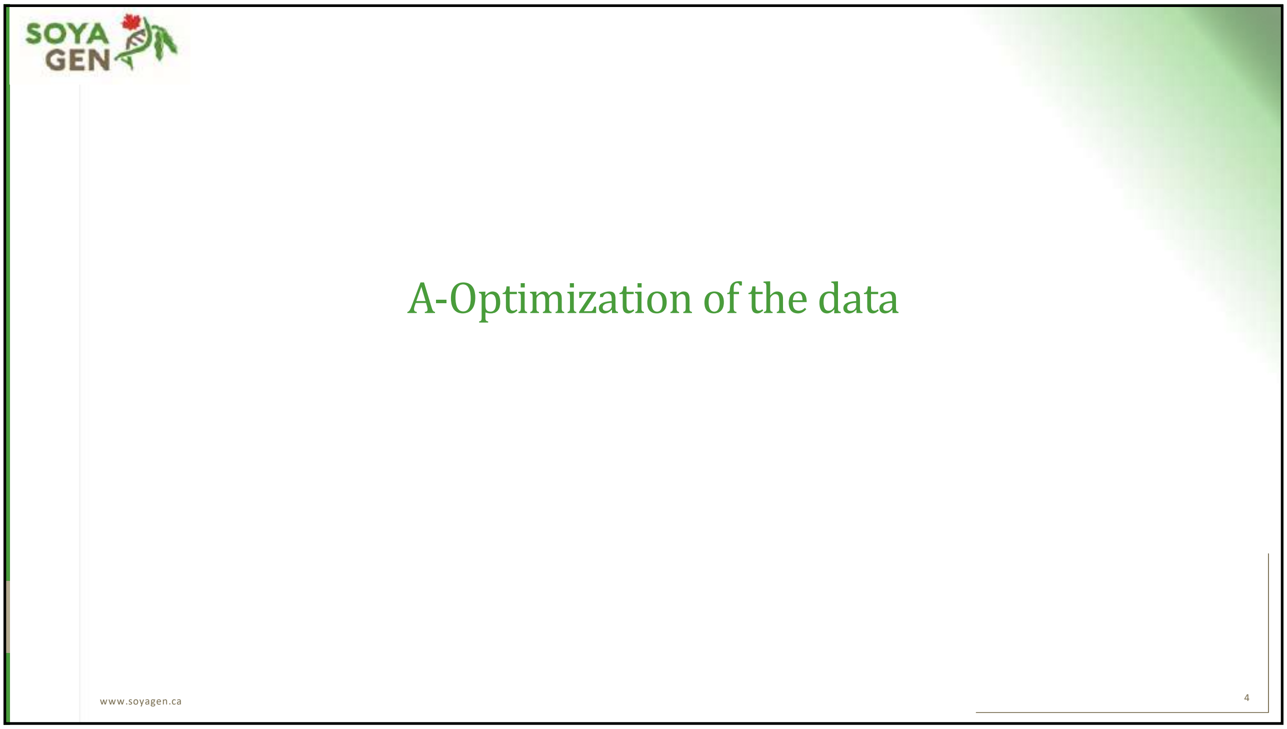
B. Construction of a genetic linkage map

1-Binning

2-Grouping

3-Ordering

4-Rippling

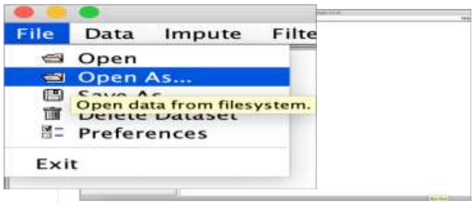


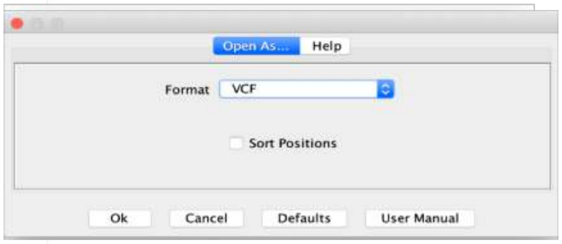
A-Optimization of the data

1. Heterozygosity filter (TASSEL):

➢ Enter the vcf file in TASSEL

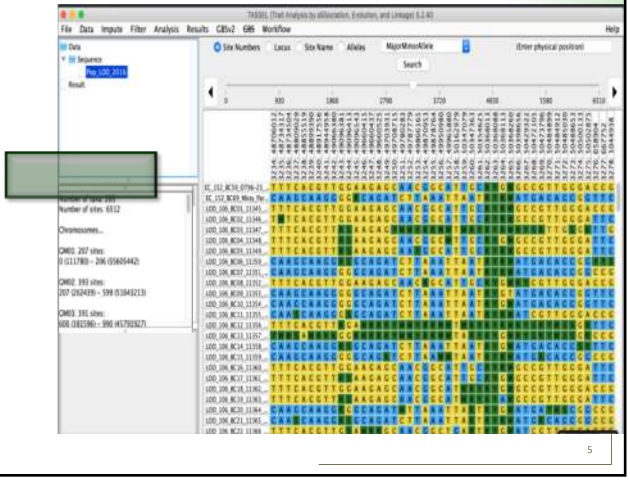
To load the data file, push on **file** (the left corner of the window) then **open as**





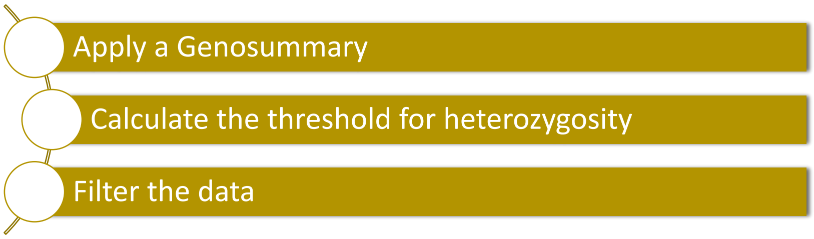
Format VCF

□ Sort Positions



The heterozygosity filter

- ➢ The point of heterozygosity filters is to eliminate lines and markers that have an excessive amount of heterozygosity.
- ➢ There will always be some heterozygosity (biological and technical causes)
- ➢ The amount of “normal” heterozygosity depends on the type of lines (e.g. F5 vs RILs)

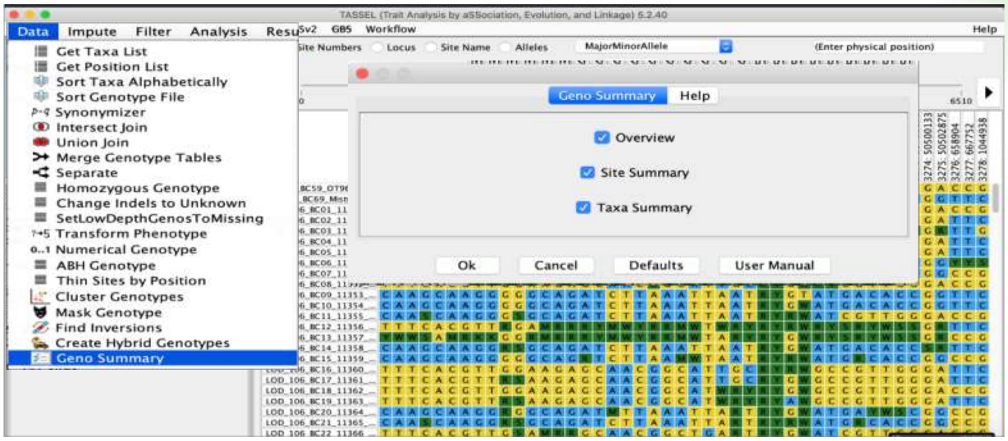


Apply a Genosummary

Calculate the threshold for heterozygosity

Filter the data

✓The steps of the Genosummary: **Data** then **Geno Summary**



Geno Summary | Help

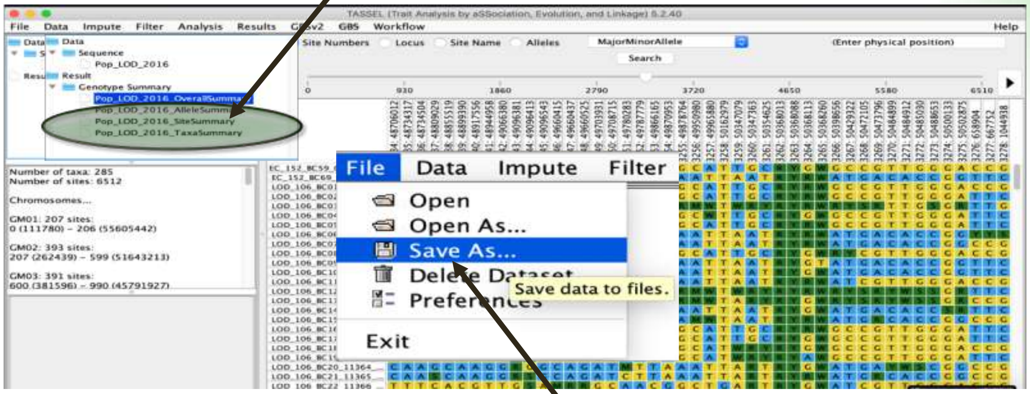
☑ Overview

☑ Site Summary

☑ Taxa Summary

The Genosummary results

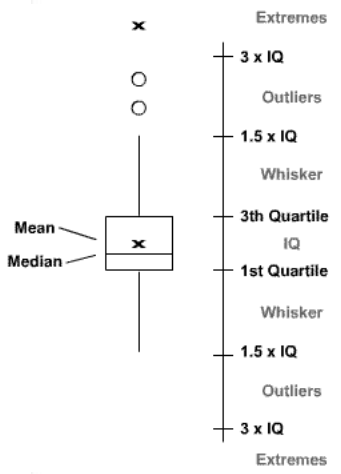
The two Result files to export to Excel to determine an appropriate threshold (i.e. how much heterozygosity is "too much")



```
Number of taxa: 285
Number of sites: 6512

Chromosomes...

GM01: 207 sites:
0 (111780) – 206 (55605442)

GM02: 393 sites:
207 (262439) – 599 (51643213)

GM03: 391 sites:
600 (381596) – 990 (45791927)
```

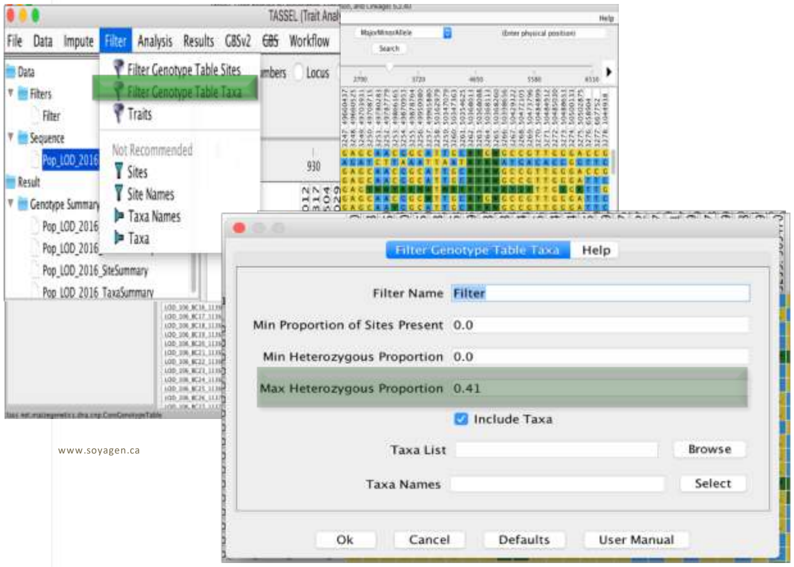
Save as "Table"

Extreme heterozygosity values are determined using an inter-quartile-range (IQR) approach



- The distribution of heterozygosity (found both in lines and in markers) is plotted in the form of a box plot
- Extreme outliers are lines/markers whose proportion of heterozygosity exceeds a value corresponding to >1.5 x IQR
- The critical value used to filter the dataset (both for markers and lines) is that which corresponds to 1.5 x IQR
- This analysis is carried out in Excel

a-Filtering the lines that have "too much" heterozygosity



Filter Genotype Table Taxa | Help

Filter Name: Filter

Min Proportion of Sites Present: 0.0

Min Heterozygous Proportion: 0.0

Max Heterozygous Proportion: 0.41

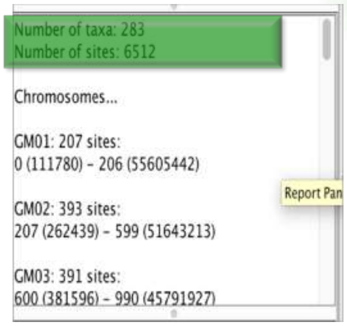
☑ Include Taxa

Taxa List: [Browse]

Taxa Names: [Select]

Ok | Cancel | Defaults | User Manual

✓ Verify if the filter was correctly done



Number of taxa: 283
Number of sites: 6512

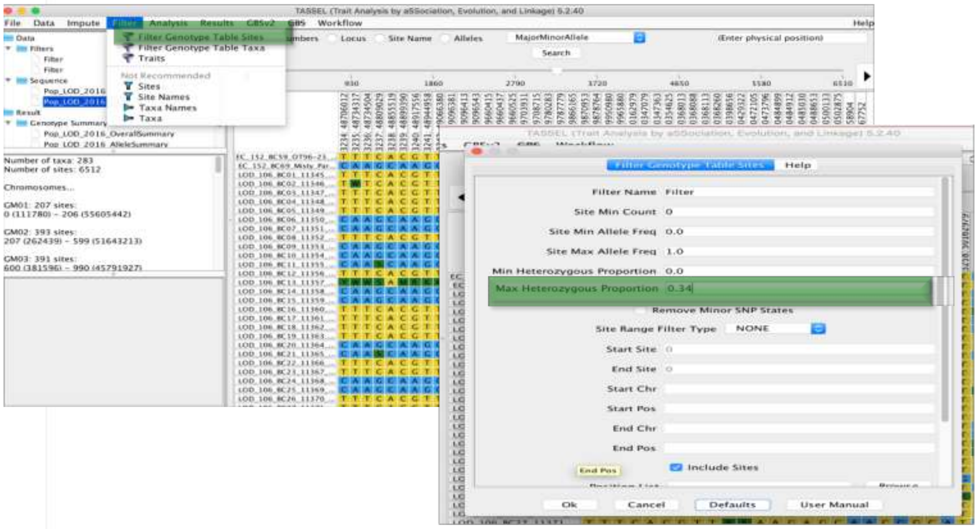
Chromosomes...

GM01: 207 sites:
0 (111780) - 206 (55605442)

GM02: 393 sites:
207 (262439) - 599 (51643213)

GM03: 391 sites:
600 (381596) - 990 (45791927)

b-Filtering the markers that have “too much” heterozygosity



Number of taxa: 283
Number of sites: 6512

Chromosomes...

GM01: 207 sites:
0 (111780) - 206 (55605442)

GM02: 393 sites:
207 (262439) - 599 (51643213)

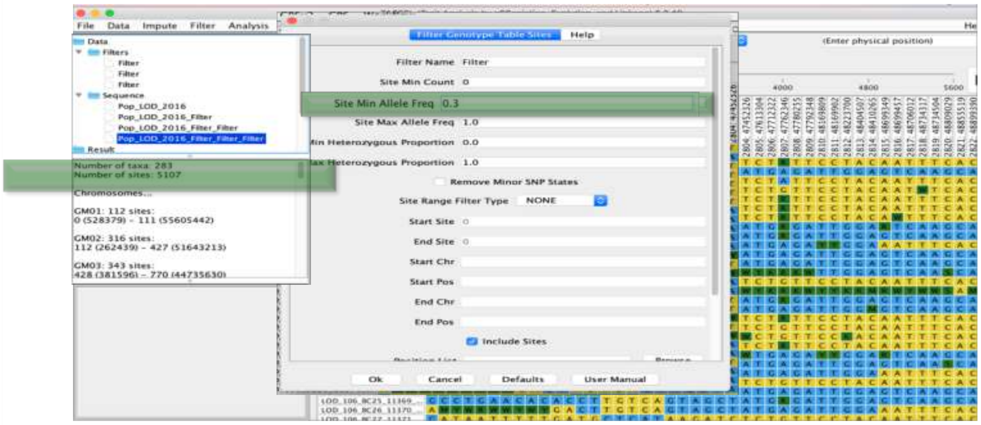
GM03: 391 sites:
600 (381596) - 990 (45791927)

Filter Genotype Table Sites

Filter Name: Filter
Site Min Count: 0
Site Min Allele Freq: 0.0
Site Max Allele Freq: 1.0
Min Heterozygous Proportion: 0.0
Max Heterozygous Proportion: 0.34
□ Remove Minor SNP States
Site Range Filter Type: NONE
Start Site: 0
End Site: 0
Start Chr:
Start Pos:
End Chr:
End Pos:
☑ Include Sites

2-Filtering markers for Minor Allele Frequency

(expectation = 0.5; tolerate down to 0.3)



Number of taxa: 283
Number of sites: 5107

Chromosomes...

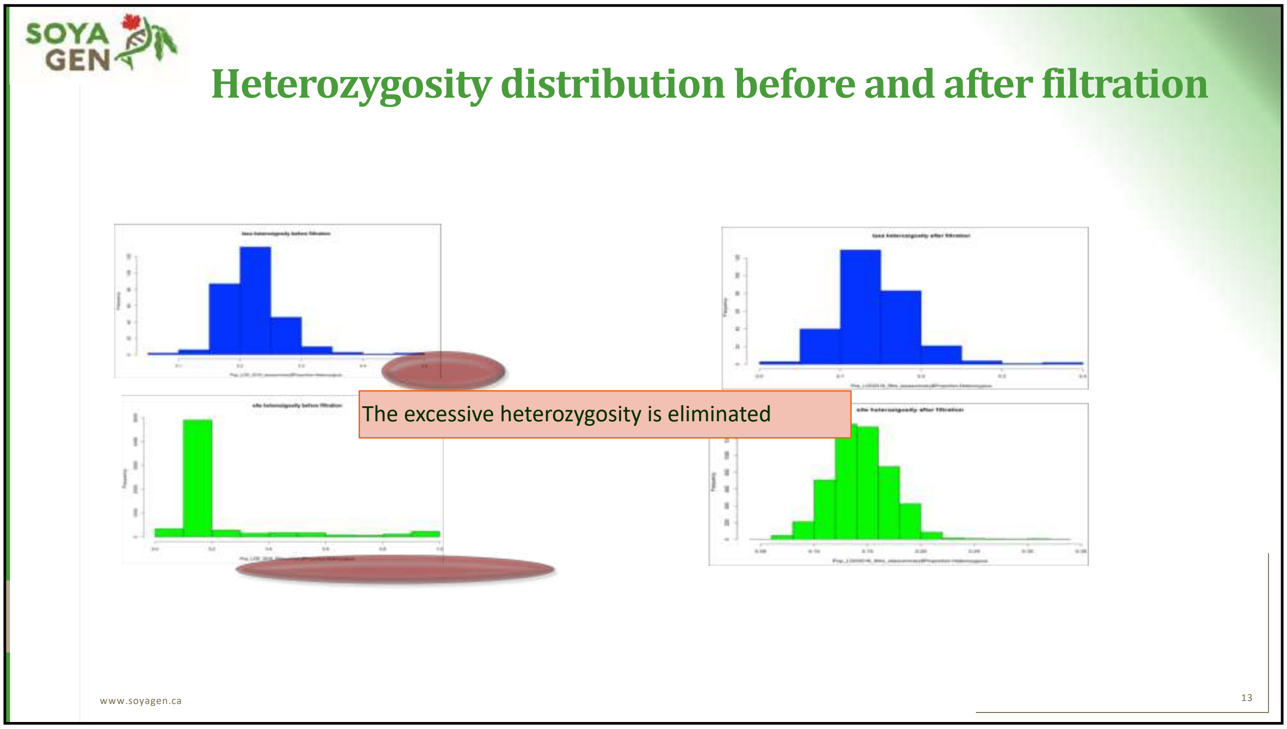
GM01: 112 sites:
0 (528379) - 111 (55605442)

GM02: 316 sites:
112 (262439) - 427 (51643213)

GM03: 343 sites:
428 (381596) - 770 (44735630)

Filter Genotype Table Sites

Filter Name: Filter
Site Min Count: 0
Site Min Allele Freq: 0.3
Site Max Allele Freq: 1.0
Min Heterozygous Proportion: 0.0
Max Heterozygous Proportion: 1.0
□ Remove Minor SNP States
Site Range Filter Type: NONE
Start Site: 0
End Site: 0
Start Chr:
Start Pos:
End Chr:
End Pos:
☑ Include Sites



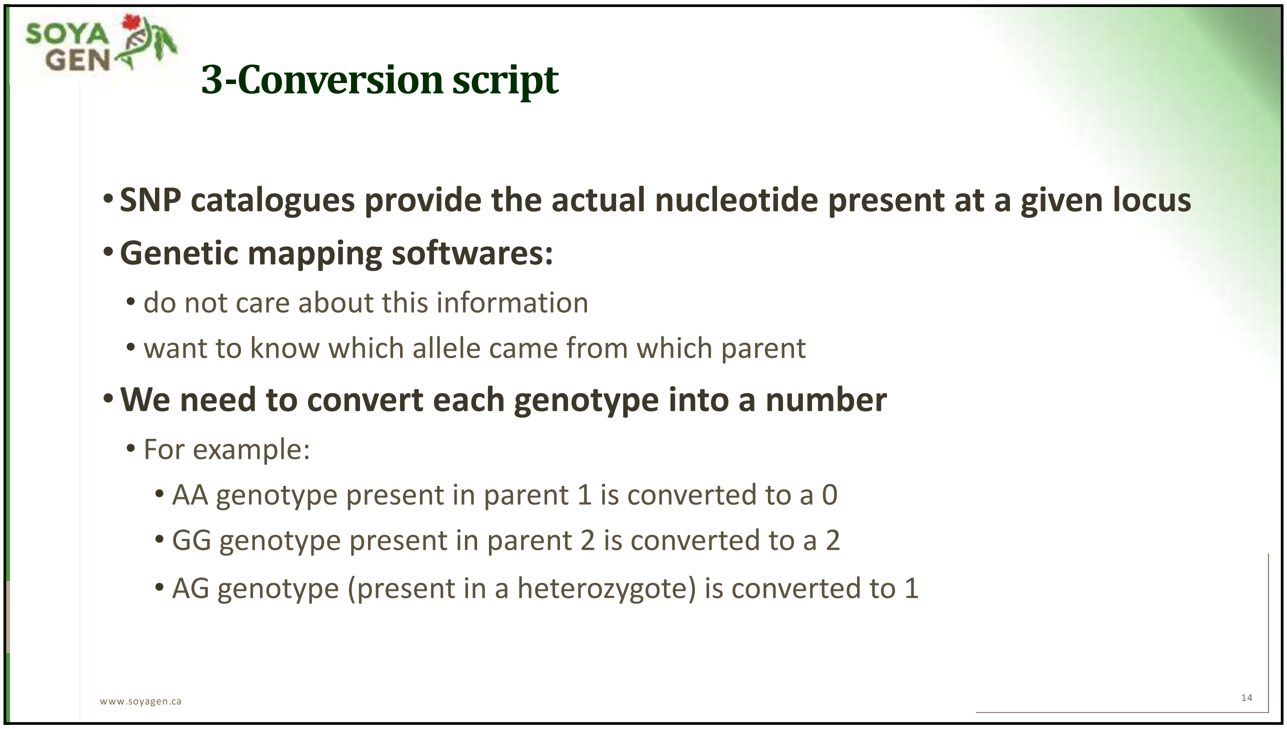
SOYA GEN

Heterozygosity distribution before and after filtration

The excessive heterozygosity is eliminated

www.soyagen.ca

13



SOYA GEN

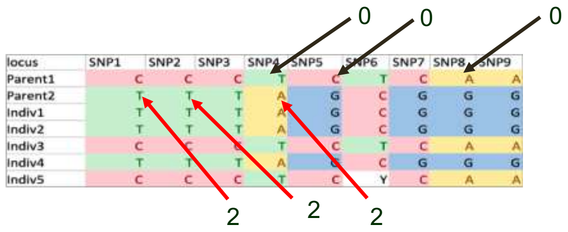
3-Conversion script

- **SNP catalogues provide the actual nucleotide present at a given locus**
- **Genetic mapping softwares:**
  - do not care about this information
  - want to know which allele came from which parent
- **We need to convert each genotype into a number**
  - For example:
    - AA genotype present in parent 1 is converted to a 0
    - GG genotype present in parent 2 is converted to a 2
    - AG genotype (present in a heterozygote) is converted to 1

www.soyagen.ca

14

How it works?

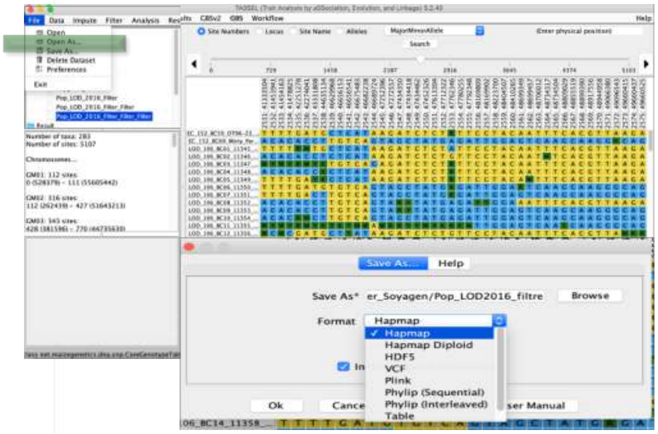


locus	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Parent1	C	C	C	T	C	T	C	A	A
Parent2	T	T	T	A	G	C	G	G	G
Indiv1	T	T	T	A	G	C	G	G	G
Indiv2	T	T	T	A	G	C	G	G	G
Indiv3	C	C	C	T	C	T	C	A	A
Indiv4	T	T	T	A	G	C	G	G	G
Indiv5	C	C	C	T	C	Y	C	A	A

locus	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Parent1	0	0	0	0	0	0	0	0	0
Parent2	2	2	2	2	2	2	2	2	2
Indiv1	2	2	2	2	2	2	2	2	2
Indiv2	2	2	2	2	2	2	2	2	2
Indiv3	0	0	0	0	0	0	0	0	0
Indiv4	2	2	2	2	2	2	2	2	2
Indiv5	0	0	0	0	0	1	0	0	0

The most important thing is the origin of the allele, not the nature of the nucleotide

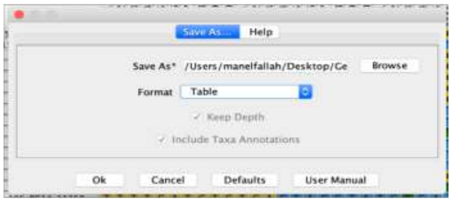
Prepare the file for the conversion



Save As* er_Soyagen/Pop_LOD2016_filtre

Format ✓ Hapmap

Save the filtered file in Hapmap format (.hmp.txt) and in Table format (.txt)



Save As* /Users/manelfallah/Desktop/Ge

Format Table

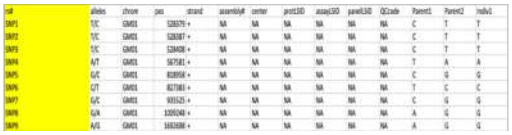
✓ Keep Depth

✓ Include Taxa Annotations

SOYA GEN

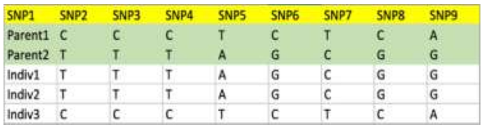
Prepare the file for the conversion

1-Open the hapmap file (.hmp.txt) in Excel



rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panelLSID	QCcode	Parent1	Parent2	Indiv1
SNP1	T/C	GM01	528379	+	NA	NA	NA	NA	NA	NA	C	T	T
SNP2	T/C	GM01	528387	+	NA	NA	NA	NA	NA	NA	C	T	T
SNP3	T/C	GM01	528408	+	NA	NA	NA	NA	NA	NA	C	T	T
SNP4	A/T	GM01	567581	+	NA	NA	NA	NA	NA	NA	T	A	A
SNP5	G/C	GM01	818958	+	NA	NA	NA	NA	NA	NA	C	G	G
SNP6	C/T	GM01	827383	+	NA	NA	NA	NA	NA	NA	T	C	C
SNP7	G/C	GM01	935525	+	NA	NA	NA	NA	NA	NA	C	G	G
SNP8	G/A	GM01	1009248	+	NA	NA	NA	NA	NA	NA	A	G	G
SNP9	A/G	GM01	1683698	+	NA	NA	NA	NA	NA	NA	A	G	G

2-Copy and transpose the rs# column in a new Excel sheet, delete the rs# cell and save as a text (.txt) file



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Parent1	C	C	C	T	C	T	C	A	
Parent2	T	T	T	A	G	C	G	G	
Indiv1	T	T	T	A	G	C	G	G	
Indiv2	T	T	T	A	G	C	G	G	
Indiv3	C	C	C	T	C	T	C	A	

(Parental lines **must** be the first two individuals in this file)

SOYA GEN

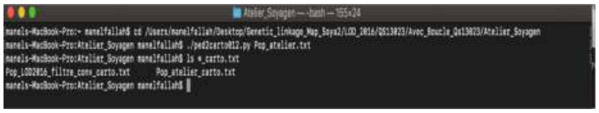
Run the script conversion

Run the bash command line : ./ped2carto.py Pop_atelier.txt

Script name

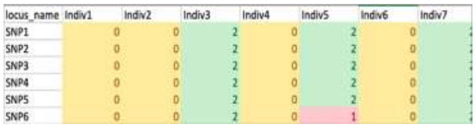
File Name

⚠ There must be a space between these



```
Atelier_Soyagen — -bash — 155x24
manels-MacBook-Pro:~ manelfallah$ cd /Users/manelfallah/Desktop/Genetic_linkage_Map_Soya2/LOD_2016/QS13023/Avec_Boucle_Qs13023/Atelier_Soyagen
manels-MacBook-Pro:Atelier_Soyagen manelfallah$ ./ped2carto012.py Pop_atelier.txt
manels-MacBook-Pro:Atelier_Soyagen manelfallah$ ls *_carto.txt
Pop_LOD2016_filtre_conv_carto.txt      Pop_atelier_carto.txt
manels-MacBook-Pro:Atelier_Soyagen manelfallah$ 
```

An output file with a new extension (_carto.txt) is going to be produced (in the case illustrated, it will be called “Pop_atelier_carto.txt”)



locus_name	Indiv1	Indiv2	Indiv3	Indiv4	Indiv5	Indiv6	Indiv7
SNP1	0	0	2	0	2	0	[illegible]
SNP2	0	0	2	0	2	0	[illegible]
SNP3	0	0	2	0	2	0	[illegible]
SNP4	0	0	2	0	2	0	[illegible]
SNP5	0	0	2	0	2	0	[illegible]
SNP6	0	0	2	0	1	0	[illegible]

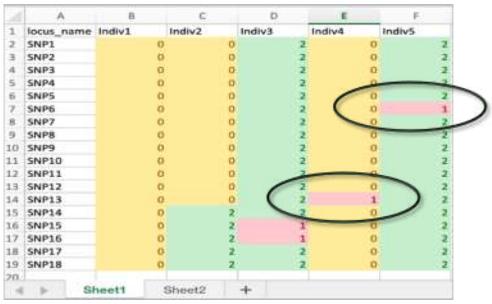
After running the script (rows = SNP loci, parents are eliminated and all genotypes are now 0, 1 or 2)

Elimination of "double recombinants"

When long stretches of one parental allele are interrupted by one or a few alternate genotype calls, these are most often genotyping errors (called "double recombinants"

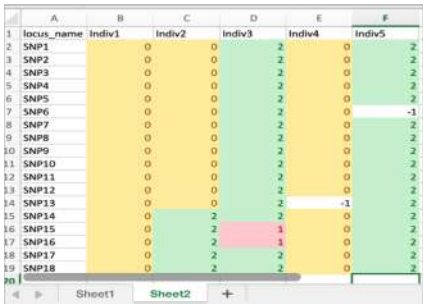
These must be eliminated and there is a tool for that!

Suspected errors are replaced with missing data (coded as "-1")



locus_name	Indiv1	Indiv2	Indiv3	Indiv4	Indiv5
SNP1	0	0	2	0	2
SNP2	0	0	2	0	2
SNP3	0	0	2	0	2
SNP4	0	0	2	0	2
SNP5	0	0	2	0	2
SNP6	0	0	2	0	1
SNP7	0	0	2	0	2
SNP8	0	0	2	0	2
SNP9	0	0	2	0	2
SNP10	0	0	2	0	2
SNP11	0	0	2	0	2
SNP12	0	0	2	0	2
SNP13	0	0	2	1	2
SNP14	0	2	2	0	2
SNP15	0	2	1	0	2
SNP16	0	2	1	0	2
SNP17	0	2	2	0	2
SNP18	0	2	2	0	2

→



locus_name	Indiv1	Indiv2	Indiv3	Indiv4	Indiv5
SNP1	0	0	2	0	2
SNP2	0	0	2	0	2
SNP3	0	0	2	0	2
SNP4	0	0	2	0	2
SNP5	0	0	2	0	2
SNP6	0	0	2	0	-1
SNP7	0	0	2	0	2
SNP8	0	0	2	0	2
SNP9	0	0	2	0	2
SNP10	0	0	2	0	2
SNP11	0	0	2	0	2
SNP12	0	0	2	0	2
SNP13	0	0	2	-1	2
SNP14	0	2	2	0	2
SNP15	0	2	1	0	2
SNP16	0	2	1	0	2
SNP17	0	2	2	0	2
SNP18	0	2	2	0	2

B. Construction of a genetic linkage map

1- Binning: Redundant markers are markers that are identical and do not provide additional information. The point is to remove redundant markers prior to generating the map file.

2-Grouping: Linkage groups are assembled based on anchor info (physical position).

3-Ordering: After the groups are correctly formed, the ordering is done "by input".

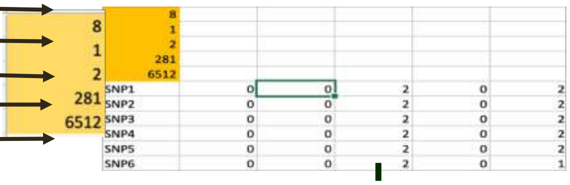
4-Rippling: After ordering, the position of each marker needs to be "fine tuned"

5-Outputting : Output the Results file.

1-Binning

a. Preparation of the bin file

Population type
Unit (cM)
Mapping function
Numbers of lines
Numbers of markers

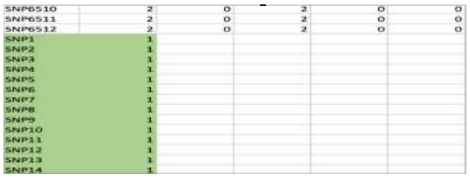


8					
1					
2					
281					
6512					
SNP1	0	0	2	0	2
SNP2	0	0	2	0	2
SNP3	0	0	2	0	2
SNP4	0	0	2	0	2
SNP5	0	0	2	0	2
SNP6	0	0	2	0	1

⚠ No Lines list

This is done by:

1) Copying the list of loci (column A)
2) Pasting the additional copy of loci at the bottom of the genotype table
3) Indicating in column B the appropriate chromosome number

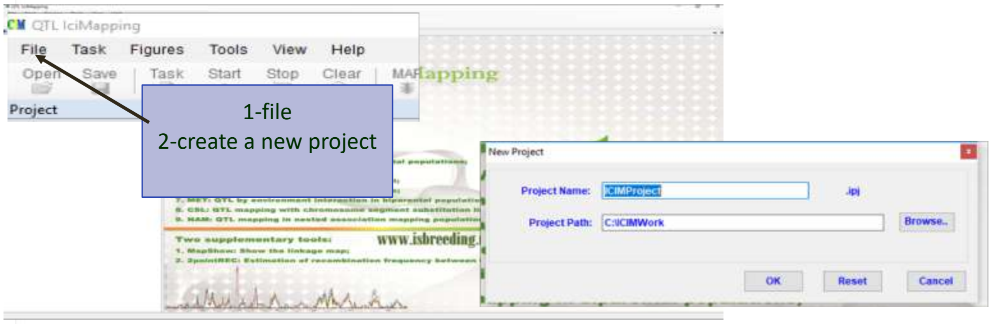


SNP6510	2	0	2	0	0
SNP6511	2	0	2	0	0
SNP6512	2	0	2	0	0
SNP1	1				
SNP2	1				
SNP3	1				
SNP4	1				
SNP5	1				
SNP6	1				
SNP7	1				
SNP8	1				
SNP9	1				
SNP10	1				
SNP11	1				
SNP12	1				
SNP13	1				
SNP14	1				

⚠ 3-Convert from **(.txt)** to **(.bin)**

QTLIciMapping

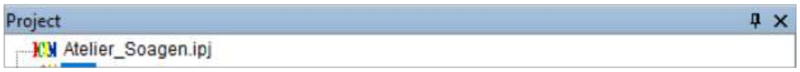
1-Create a new project



1-file
2-create a new project

Project Name: ICIMProject .ipj

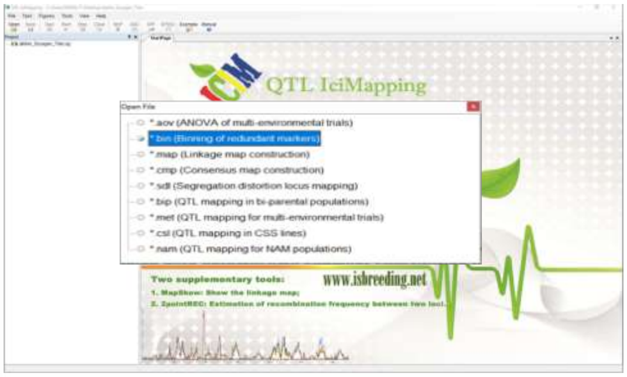
Project Path: C:\ICIMWork



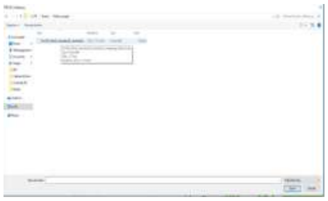
✓ The project created

2-Load the bin file

✓Open then choose the file type (here we choose bin)
✓Go to the working directory to choose the file
✓Select “Open” or “ouvrir”

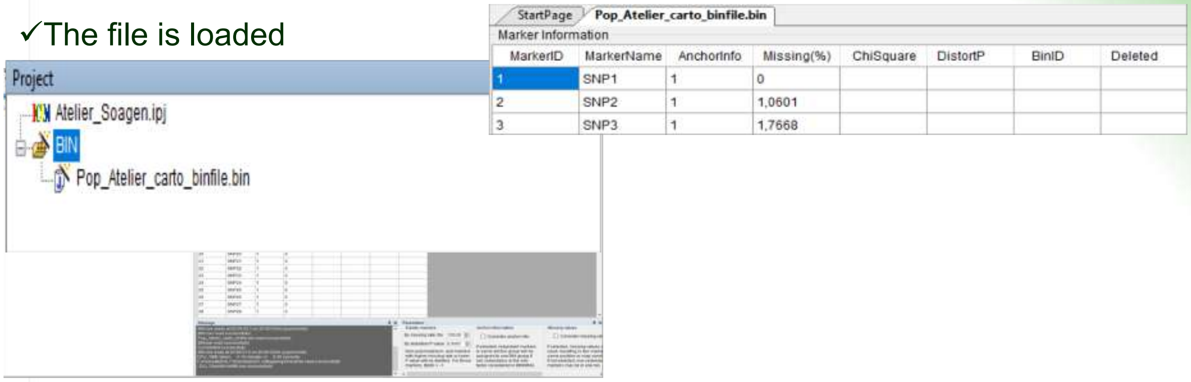

QTL IciMapping
Open File
Two supplementary tools:
www.isbreeding.net
1. MapShow: Show the linkage map;
2. 2pointREC: Estimation of recombination frequency between two loci.

- ○ *.aov (ANOVA of multi-environmental trials)
- ● *.bin (Binning of redundant markers)
- ○ *.map (Linkage map construction)
- ○ *.cmp (Consensus map construction)
- ○ *.sdl (Segregation distortion locus mapping)
- ○ *.bip (QTL mapping in bi-parental populations)
- ○ *.met (QTL mapping for multi-environmental trials)
- ○ *.csl (QTL mapping in CSS lines)
- ○ *.nam (QTL mapping for NAM populations)



3- Binning

✓The file is loaded



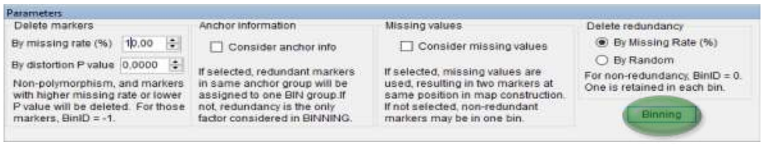
Project

- Atelier_Soagen.ipj
- BIN
  - Pop_Atelier_carto_binfile.bin

StartPage | Pop_Atelier_carto_binfile.bin

Marker Information

MarkerID	MarkerName	AnchorInfo	Missing(%)	ChiSquare	DistortP	BinID	Deleted
1	SNP1	1	0				
2	SNP2	1	1,0601				
3	SNP3	1	1,7668				



Parameters

Delete markers

By missing rate (%) 10,00

By distortion P value 0,0000

Non-polymorphism, and markers with higher missing rate or lower P value will be deleted. For those markers, BinID = -1.

Anchor information

☐ Consider anchor info

If selected, redundant markers in same anchor group will be assigned to one BIN group.If not, redundancy is the only factor considered in BINNING.

Missing values

☐ Consider missing values

If selected, missing values are used, resulting in two markers at same position in map construction. If not selected, non-redundant markers may be in one bin.

Delete redundancy

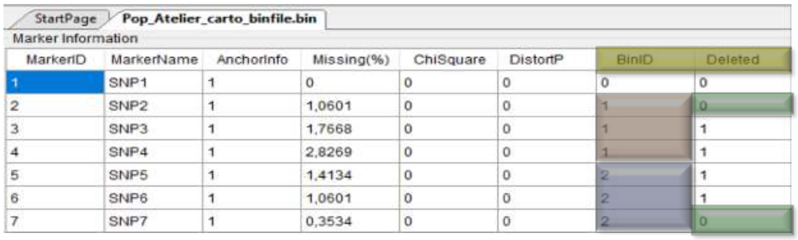
● By Missing Rate (%)

○ By Random

For non-redundancy, BinID = 0. One is retained in each bin.

The result file for the binning

- ✓ Two columns that are important: “**BinID**” and “**Deleted**”
- ✓ **BinID**:
  - * 0 : the marker is unique and it doesn’t belong to any linkage group
  - * Every number >0 indicates a group identification or BinID (markers with the same BinID belong to the same linkage group
- ✓ **Deleted**: 0 the marker was not deleted (saved for the map file)
  1 the marker has been deleted



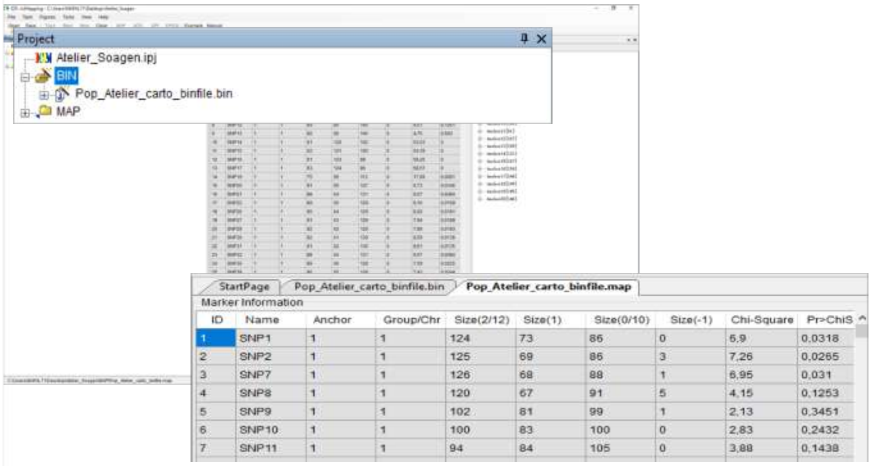
StartPage | Pop_Atelier_carto_binfile.bin

Marker Information

MarkerID	MarkerName	AnchorInfo	Missing(%)	ChiSquare	DistortP	BinID	Deleted
1	SNP1	1	0	0	0	0	0
2	SNP2	1	1,0601	0	0	1	0
3	SNP3	1	1,7668	0	0	1	1
4	SNP4	1	2,8269	0	0	1	1
5	SNP5	1	1,4134	0	0	2	1
6	SNP6	1	1,0601	0	0	2	1
7	SNP7	1	0,3534	0	0	2	0

4-Load the Map file

✓The same steps to load the bin file but we choose map



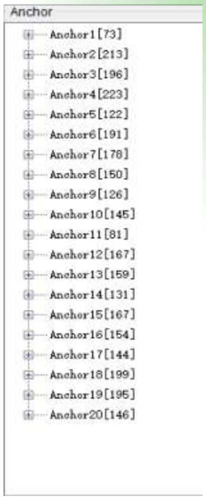
Project

- Atelier_Soagen.ipj
  - BIN
    - Pop_Atelier_carto_binfile.bin
  - MAP

StartPage | Pop_Atelier_carto_binfile.bin | Pop_Atelier_carto_binfile.map

Marker Information

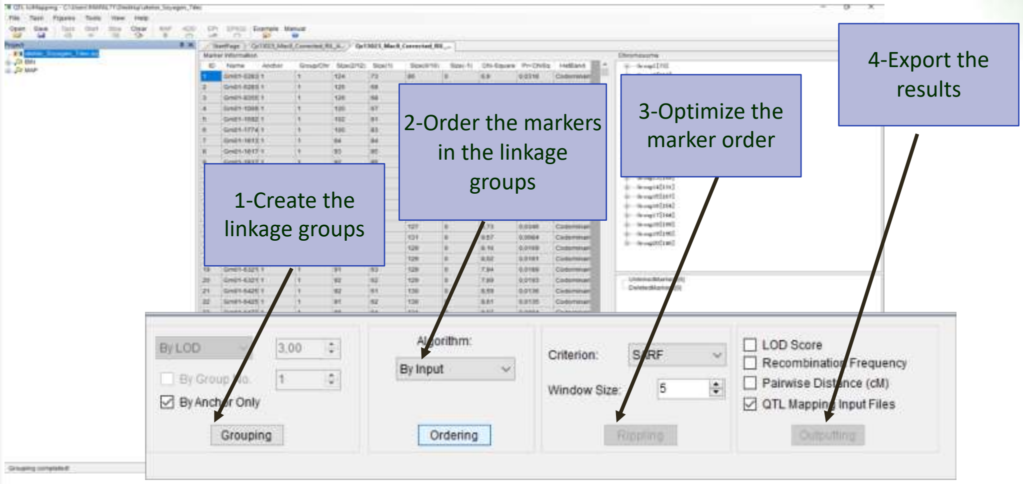
ID	Name	Anchor	Group/Chr	Size(2/12)	Size(1)	Size(0/10)	Size(-1)	Chi-Square	Pr>ChiS
1	SNP1	1	1	124	73	86	0	6,9	0,0318
2	SNP2	1	1	125	69	86	3	7,26	0,0265
3	SNP7	1	1	126	68	88	1	6,95	0,031
4	SNP8	1	1	120	67	91	5	4,15	0,1253
5	SNP9	1	1	102	81	99	1	2,13	0,3451
6	SNP10	1	1	100	83	100	0	2,83	0,2432
7	SNP11	1	1	94	84	105	0	3,88	0,1438



Anchor

- Anchor1[73]
- Anchor2[213]
- Anchor3[196]
- Anchor4[223]
- Anchor5[122]
- Anchor6[191]
- Anchor7[178]
- Anchor8[150]
- Anchor9[126]
- Anchor10[145]
- Anchor11[81]
- Anchor12[167]
- Anchor13[159]
- Anchor14[131]
- Anchor15[167]
- Anchor16[154]
- Anchor17[144]
- Anchor18[199]
- Anchor19[195]
- Anchor20[146]

5-The steps for mapping



1-Create the linkage groups

2-Order the markers in the linkage groups

3-Optimize the marker order

4-Export the results

By LOD 3.00

☐ By Group No. 1

☑ By Anchor Only

Algorithm: By Input

Criterion: SARF

Window Size: 5

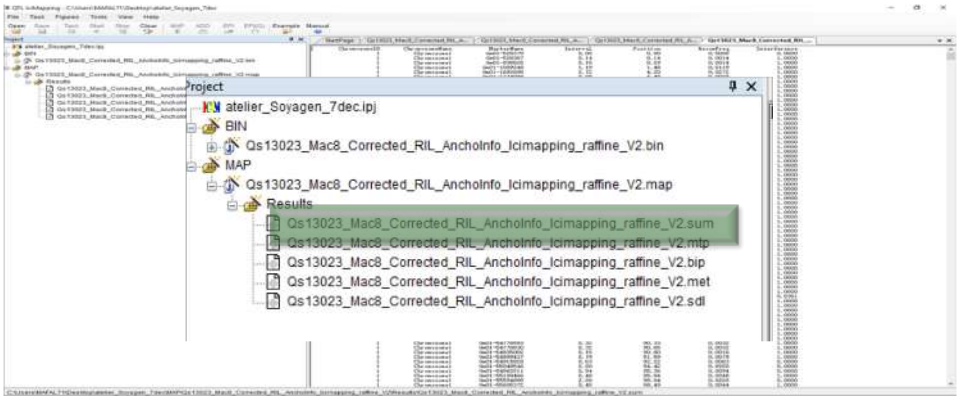
☐ LOD Score

☐ Recombination Frequency

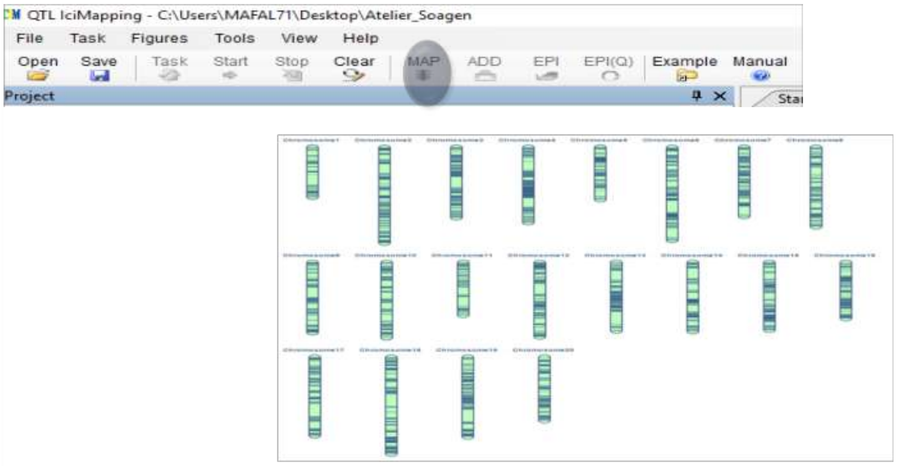
☐ Pairwise Distance (cM)

☑ QTL Mapping Input Files

6-The Output files


Project
atelier_Soyagen_7dec.ipj
BIN
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.bin
MAP
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.map
Results
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.sum
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.mtp
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.bip
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.met
Qs13023_Mac8_Corrected_RIL_AnchoInfo_Icimapping_raffine_V2.sdl

7-Draw the linkage map


QTL IciMapping - C:\Users\MAFAL71\Desktop\Atelier_Soagen