



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



Solanaceae Coordinated
Agricultural Project



Downstream analysis with SNP markers

Part I: Introduction to computer software for data analysis

Sung-Chur Sim

The Ohio State University, OARDC

SolCAP workshop

Outline

Part I: Introduction to computer software

- MicroSatellite Analyzer (MSA)
- Graphical GenoType (GTT)
- STRUCTURE
 - ✓ What can you do using the software?
 - ✓ Where can you download the software?
 - ✓ How can you format input data ?

Part II: The use of STRUCTURE for association mapping

- Detail steps to generate a Q-matrix using STRUCTURE

MicroSatellite Analyzer: MSA

- 🍷 An independent analysis tool for large data sets (Dieringer and Schlötterer 2003)
 - Descriptive statistics per population and locus (e.g. allelic richness, heterozygosity, and Shannon index of diversity)
 - F_{ST} , F_{IS} , and F_{IT} based on the Weir and Cockerham method
 - F_{ST} per locus and population pair ; P-value for F_{ST} determined by permuting genotypes among groups
 - Genetic distance including Nei's standard genetic distance
 - Converts your data into the formats of GENEPOP, STRUCTURE, ARLEQUIN, etc.
- 🍷 Version 4.05 available for Windows, Linux, and Mac:
(http://i122server.vu-wien.ac.at/MSA/MSA_download.html)



microsatellite analyzer

Search

Instant is on

About 109,000 results (0.29 seconds)

Advanced search

- Everything
- Images
- Videos
- More

Wooster, OH

Change location

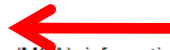
Show search tools

Microsatellite Genotyping

www.AppliedBiosystems.com Capture Success With Our Accurate Analysis Applications. Contact Us!

Sponsored links

MSA download



Microsatellite Analyzer (MSA). information about MSA can be found here. For downloading please click on the required OS icon. ...

i122server.vu-wien.ac.at/MSA/MSA_download.html - Cached - Similar

CS lab software

CS lab software. **Microsatellite Analyzer (MSA)**. Provides basic summary ...

i122server.vu-wien.ac.at/CSlab_software.htm - Cached

Show more results from vu-wien.ac.at

Analysis Software

Genealogical **analysis** of linked **microsatellite** and unique event polymorphism haplotypes. Population splitting model and growth models allowed. ...

www2.hawaii.edu/~khayes/Software_links.htm - Cached - Similar

microsatellites - Evolutionary Genetics Software Links by Sergios ...

microsatellite analysis software list, URL, compiled by David McDonald ... PCAGEN, Windows, Principal Components Analysis of microsatellite data. ...

softlinks.amnh.org/microsatellites.html - Cached - Similar

Automated Microsatellite Analysis

File Format: PDF/Adobe Acrobat - Quick View

Fast, Automated Analysis from lane finding through allele scoring with SagaGT **Microsatellite Analysis**. Software. • An exclusive Allele Fingerprinting ...

www.licor.com/bio/PDF/genomics/microsat.pdf - Similar

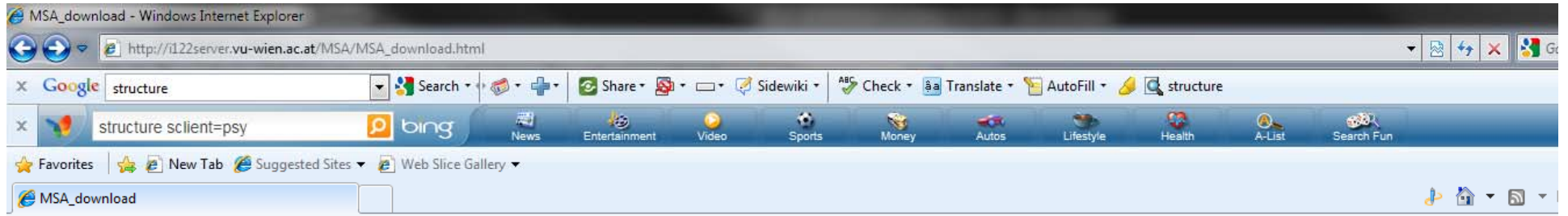
Microsatellite Analysis on the Applied Biosystems 3130 Series ...

File Format: PDF/Adobe Acrobat - Quick View

Quality Fragme
STR, AFLP, TRFI
48 hours, Flexible
www.laragen.com

Microsatellite
Superior Sensitivi
DNA **microsatell**
www.aati-us.com

See your ad here



Microsatellite Analyzer (MSA)

information about MSA can be found [here](#)

For downloading please click on the required OS icon.
You will receive a folder containing the following items:

executables
documentation
sample input file

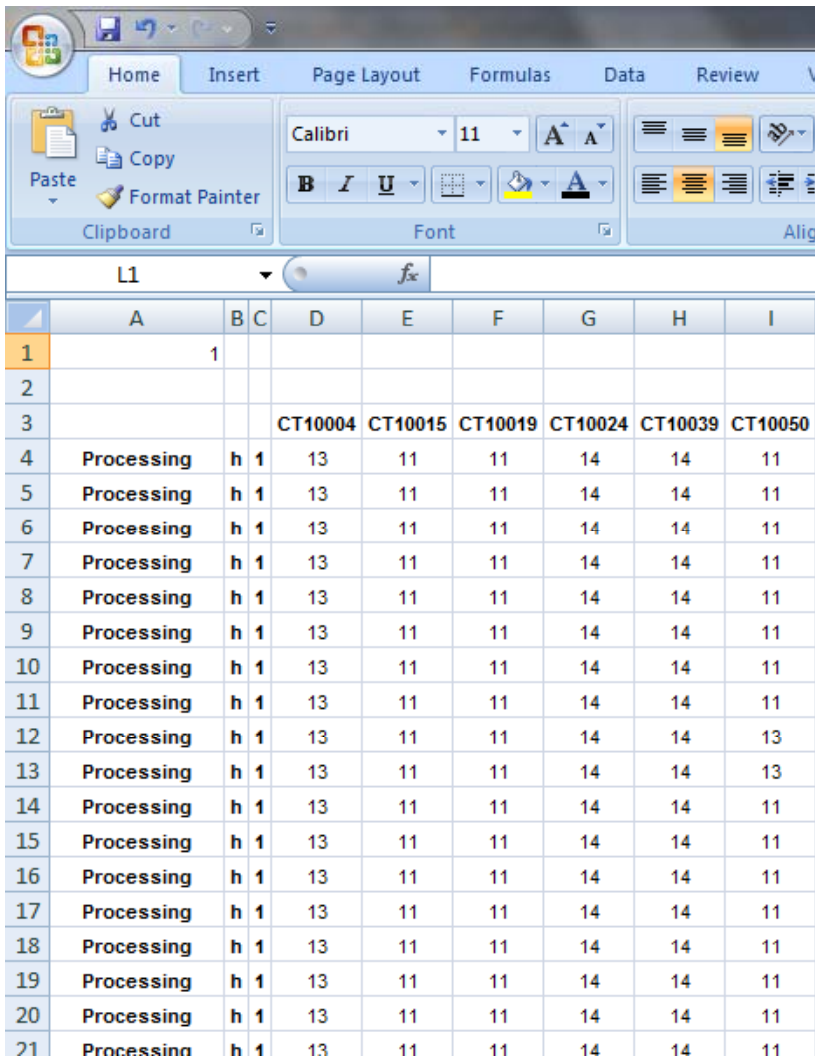


contact: [Daniel Dieringer](#)

WEB 14756

http://i122server.vu-wien.ac.at/MSA/MSA_download.html

Input format



The screenshot shows an Excel spreadsheet with the following data:

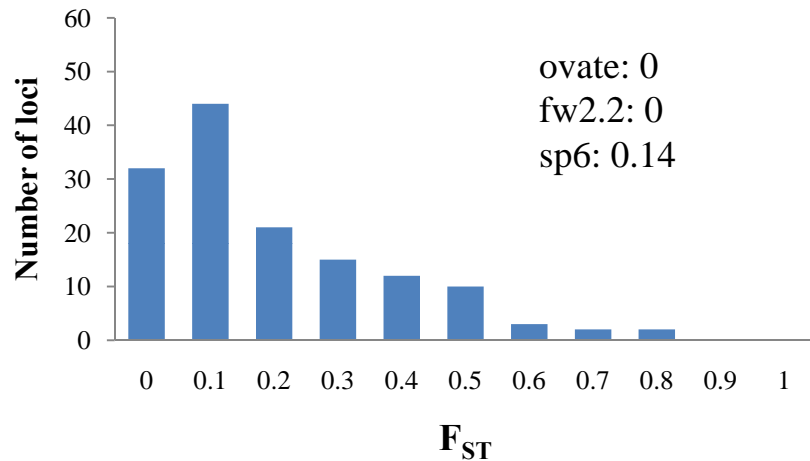
	A	B	C	D	E	F	G	H	I
1		1							
2									
3				CT10004	CT10015	CT10019	CT10024	CT10039	CT10050
4	Processing	h	1	13	11	11	14	14	11
5	Processing	h	1	13	11	11	14	14	11
6	Processing	h	1	13	11	11	14	14	11
7	Processing	h	1	13	11	11	14	14	11
8	Processing	h	1	13	11	11	14	14	11
9	Processing	h	1	13	11	11	14	14	11
10	Processing	h	1	13	11	11	14	14	11
11	Processing	h	1	13	11	11	14	14	11
12	Processing	h	1	13	11	11	14	14	13
13	Processing	h	1	13	11	11	14	14	13
14	Processing	h	1	13	11	11	14	14	11
15	Processing	h	1	13	11	11	14	14	11
16	Processing	h	1	13	11	11	14	14	11
17	Processing	h	1	13	11	11	14	14	11
18	Processing	h	1	13	11	11	14	14	11
19	Processing	h	1	13	11	11	14	14	11
20	Processing	h	1	13	11	11	14	14	11
21	Processing	h	1	13	11	11	14	14	11

One or two column format

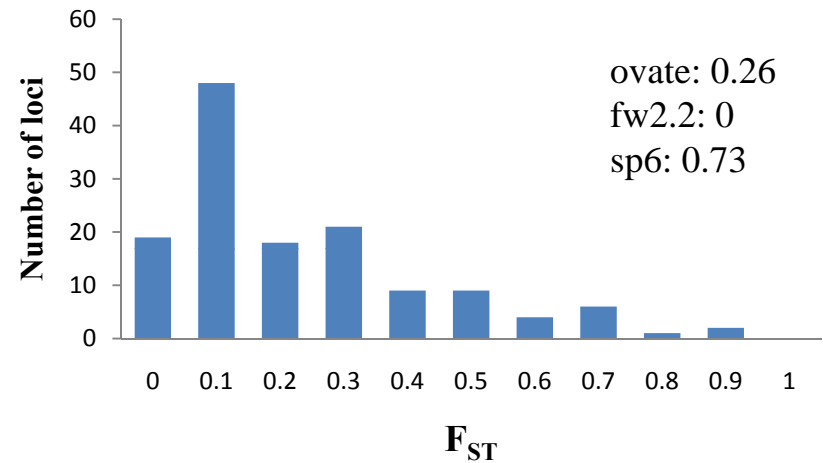
- Specify **one (1) or two (2)** column format in the cell A1
- Enter name of population in the first column (no empty cell)
- Specify **inbred (h) or outbred (d)** for your species in the second column (no empty cell)
- Enter group number of population (no empty cell)
- SNP data converted from letter codes to numerical coding
- Missing data can be indicated by -1, nd, dot(.), or empty cell
- Save your data in the format **“TAB DELIMITED”**

Identify loci that distinguish populations

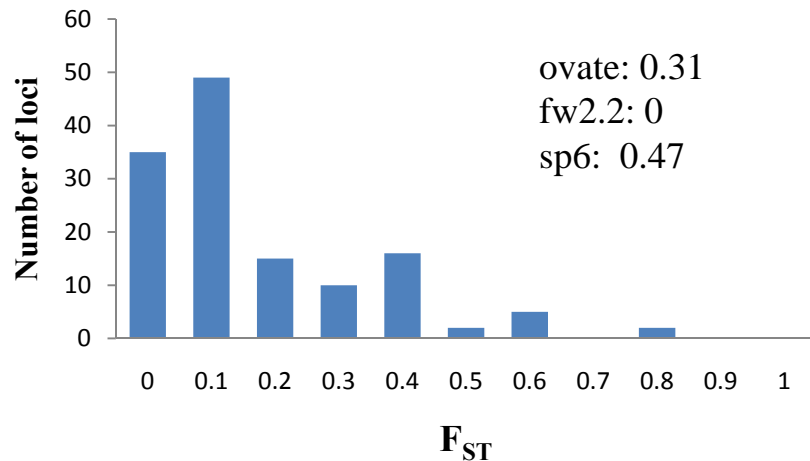
Processing vs. Freshmarket



Processing vs. Heirloom



Freshmarket vs. Heirloom



$F_{ST} = 0$: an allele of a gene is fixed or the gene is under balancing selection

$F_{ST} = 1$: a gene under diversifying selection

Graphical GenoType: GGT

- 🍅 A tool for representing molecular marker data by graphical representation and color coding of chromosomes
 - Useful for evaluation of plant material and selection of a desired genotype
- 🍅 Advanced genetic analyses
 - Marker-trait association
 - Genetic distance
 - Linkage disequilibrium
- 🍅 Version 2.0 available for Windows
(http://www.plantbreeding.wur.nl/UK/software_ggt.html)

Web Images Videos Maps News Shopping Gmail more



Graphical genotype

Search

Instant is on

About 979,000 results (0.43 seconds)

Advanced search

Everything

Images

Videos

More

Wooster, OH

Change location

Show search tools

Graphical genotypes

GGT stands for **Graphical GenoTypes**. A graphical representation of molecular marker data can be an important tool in the process of selection and evaluation ...
www.plantbreeding.wur.nl/UK/software_ggt.html - Cached - Similar

GCP McClintock Bioinformatics Resource - Concept of graphical ...

Jun 27, 2008 ... The concept of '**graphical genotypes**', which was first proposed by Young ... (c) Example of a **graphical genotype** of a single IRRI-derived BC1 ...
mcclintock.generationcp.org/index.php?... - Cached - Similar

[PDF] Analysis of mosquito genome structure using graphical genotyping

File Format: PDF/Adobe Acrobat - View as HTML
by DW Severson - Cited by 8 - Related articles
demonstrate the utility of using **graphical genotyping** to easily and quickly evaluate the entire **Graphical genotypes** for all individuals examined in both ...
www.nd.edu/~dseverso/Pubs/Severson_InsMolBiol_1995.pdf - Similar

Restriction fragment length polymorphism maps and the concept of ...

by ND Young - 1989 - Cited by 176 - Related articles
fulness of this concept, **graphical genotypes** for individu- als from backcross and F2 populations ... **graphical genotype**, the primary goal would be to trans- ...
www.springerlink.com/index/j371v2552428g844.pdf

Flapjack - Graphical Genotyping

New software tools for **graphical genotyping** and haplotype visualization are required that can routinely handle the large data volumes generated by high ...
bioinf.scri.ac.uk/flapjack/ - Cached - Similar

PDF - Graphical Genotype of Maize Inbred B86 Revealed by RFLPs

by S Fahr - 1993 - Cited by 1 - Related articles
Graphical genotype for maize inbred B86 based on 178 DNA probes in combination with restriction enzymes EcoRI, EcoRV, and HindIII. ...
onlinelibrary.wiley.com/doi/10.1111/i.1439-0523.1993.tb00565.x/pdf

http://www.plantbreeding.wur.nl/UK/software_ggt.html

Plant Breeding

Education

Research

Publications

News & Calendar

About Plant Breeding

Work at

Phone book

Links

Contact

Latest update: February 2010

» [DOWNLOAD GGT 2.0 \(Versie: 2010\)](#)



GGT, what is it? [Click to find out more..](#)

Reference:

- GGT 2.0: Versatile Software for Visualization and Analysis of Genetic Data *Journal of Heredity* 2008 99(2):232-236

Please cite either paper if you have used GGT 2.0 in research leading to a scientific publication

There is also a [POSTER](#) [PDF, updated Jan '06] that explains the features and possible use of GGT 2.0 and detailed instructions can be read in the [GGT 2.0 user manual](#)

GGT Updates :

update Feb 2010

- No new developments are planned for GGT 2.0 but occasional support will still be possible
- New Build with extended expiration date of 2015
- fixed estimation of phase in DH with many missing data points

Input format

- 🍅 Two data files derived from locus and map data
- 🍅 Locus file
 - Contains data on marker alleles using the MapMaker or JoinMap type of coding
 - A plain text file

Locus file

```
; This file was used as input for the JOINMAP mapping software
; use the BUIL GGT FILE option to merge '.loc' and '.map' files into a
'.ggt' file
;
; Fri, 10 Jan 1997, 11:54
; grouping file: mylvuniq.grp
; original file: mylvuniq.loc
; linkage group: 1
```

```
name = lvuniq-1|
popt = RI9
nloc = 72
nind = 103
```

```
E42M32-231#1
```

```
aaaaa -ca-a caccc aacca cacac caaca aaaca ccca accaa accca
caaac aaaaa accaa acccc accaa cccac acaaa caccc caaaa ccaac
ccc
```

```
E33M61-740
```

```
aaaaa aaaca caccc aacca aacac caaca caaca acaaa acacc accaa
cacaa acaaa accca ccccc accaa accac accaa caaca caaaa caaac
```

Input format

- 🍅 Two data files derived from locus and map data
- 🍅 Locus file
 - Contains data on marker alleles using the MapMaker or JoinMap type of coding
 - A plain text file
- 🍅 Map file
 - Specifies marker positions on a linkage map
 - A plain text file

Map file

```
; Genetic map file of a Barley RIL population  
; chromosome 1
```

```
chrom 1
```

E33M55-508	0.0
E39M61-574	1.8
E35M48-228	4.0
E33M61-740	14.6
E35M54-93	14.6
E41M40-112	20.7
E42M51-267	23.3
E42M32-231#1	26.5
E42M40-287	28.5
E33M61-120	29.2
E37M32-00	38.0

Input format

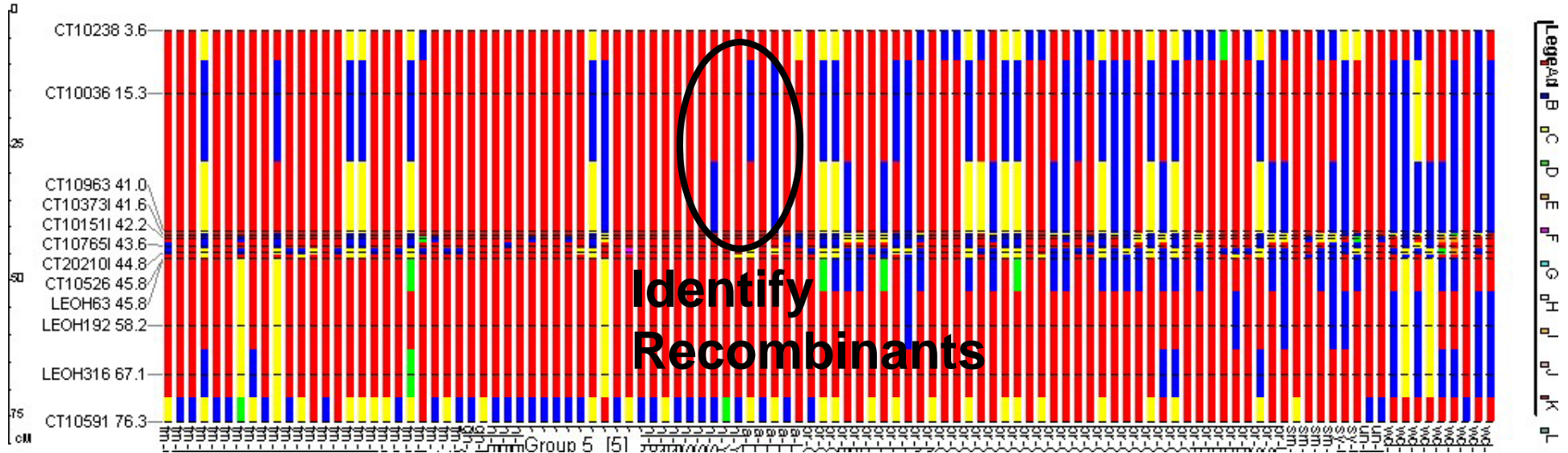
- 🍅 Two data files derived from locus and map data
- 🍅 Locus file
 - Contains data on marker alleles using the MapMaker or JoinMap type of coding
 - A plain text file
- 🍅 Map file
 - Specifies marker positions on a linkage map
 - A plain text file
- 🍅 Build a GGT file by merging the locus and map files using the ‘Build GGT-file’ option
- 🍅 The GGT file can also be prepared from an Excel spreadsheet

Fresh Market

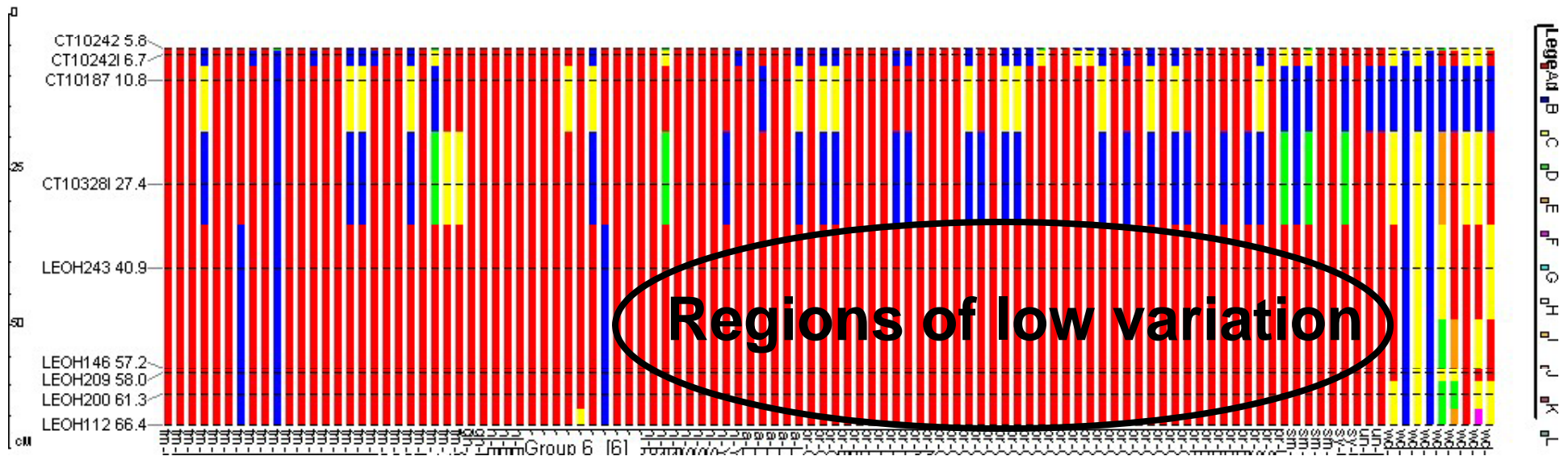
Heirloom

Processing

Wild



Chromosome 5

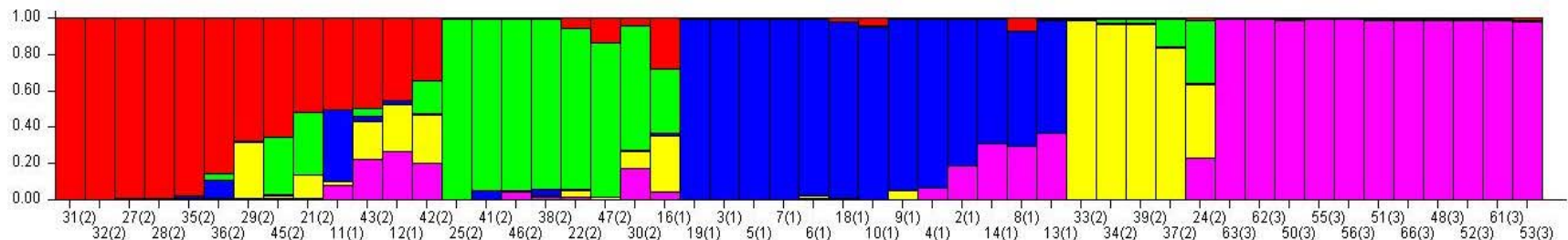


Chromosome 6

STRUCTURE

- 🍷 A model-based clustering method (Pritchard et al. 2000)
 - Inferring population structure using multi-locus genotype data
 - Generating a Q-matrix to correct for population subdivision during marker-trait association analysis in complex populations (e.g. breeding populations)
 - Identifying migrants and admixed individuals

🍷 Version 2.3.3 available for Windows, Linux, and Mac:
(<http://pritch.bsd.uchicago.edu/structure.html>)





structure

Instant is on

About 357,000,000 results (0.15 seconds)

Advanced search

- Everything
- Images
- Videos
- News
- Books
- More

Wooster, OH
Change location

Any time
Latest
Past 2 days

All results
Page previews

More search tools

Something different
composition
morphology
conformation
geometry
dynamics

Structure

Oct 13, 2010 ... 1270–1279) combined X-ray crystallography and electron microscopy to get the complete high-resolution open **structure** of group II chaperonin ...
[Current Issue](#) - [Login](#) - [Permissions](#) - [Conferences](#)
www.cell.com/structure/ - [Cached](#)

Structure - Wikipedia, the free encyclopedia

Structure is a fundamental, if intangible, notion referring to the recognition, observation, nature, and stability of patterns and relationships of entities ...
[Types of structure](#) - [Biological structure](#) - [See also](#) - [References](#)
en.wikipedia.org/wiki/Structure - [Cached](#) - [Similar](#)

Software For Inferring Population Structure

The program **structure** is a free software package for using multi-locus genotype data to investigate population **structure**. Its uses include inferring the ...
pritch.bsd.uchicago.edu/structure.html - [Cached](#) - [Similar](#)

Software for Genetic Analysis

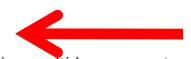
The program **structure** is a free software package for using multi-locus genotype data to investigate population **structure**. Its uses include inferring the ...
pritch.bsd.uchicago.edu/software.html - [Cached](#) - [Similar](#)

ScienceDirect - Structure, Volume 18, Issue 9, Pages 1067-1220 (8 ...

Sep 8, 2010 ... The online version of **Structure** on ScienceDirect, the world's leading platform for high quality peer-reviewed full-text publications in ...
www.sciencedirect.com/science/journal/09692126 - [Similar](#)

EXPRESS: Clothing for Women and Men : Shop the Hottest Clothes at ...

Be a trendsetter with the latest in women's and men's clothing from Express. Shop Jeans, Shirts, Tops, Dresses and Accessories for women and men.
www.express.com/ - [Cached](#) - [Similar](#)



Software For Inferring Population Structure - Windows Internet Explorer

http://pritch.bsd.uchicago.edu/structure.html

Google structure software

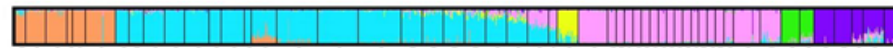
structure scient=psy

Software For Inferring Population Structure

[Home](#) [Software](#) [Lab Members](#) [Publications](#) [Data](#) [Contact Information](#)

Structure

The program *structure* is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPS, microsatellites, RFLPs and AFLPs.



Download [Structure 2.3.3](#).



What to cite: The basic algorithm was described by Pritchard, Stephens & Donnelly (2000). Extensions to the method were published by Falush, Stephens and Pritchard (2003) and (2007) and by Hubisz, Falush, Stephens and Pritchard (2009).


Contributors: [Daniel Falush](#), [Melissa Hubisz](#), [Matthew Stephens](#), [Jonathan Pritchard](#), [Peter Donnelly](#), [William Wen](#), [Mike Trienis](#), [Pall Melsted](#).

Questions and Discussion: We have now started a [Google Groups](#) forum devoted to Structure. This replaces the [Genetic Software Forum](#) which is no longer active.

<http://pritch.bsd.uchicago.edu/structure.html>

Input format

	A	B	C	D	E	F
1			CT10153	CT10162	CT10184	CT10187
2	Campbell28	1	14	12	-1	13
3	Campbell28	1	14	12	-1	13
4	Fla7060	1	12	13	12	13
5	Fla7060	1	12	13	12	13
6	Fla7547	1	12	12	12	13
7	Fla7547	1	12	12	12	13
8	Fla7771	1	14	12	12	13
9	Fla7771	1	14	12	12	13
10	Fla7775	1	14	13	12	13
11	Fla7775	1	14	13	12	13
12	Fla7600	1	14	12	13	13
13	Fla7600	1	14	12	13	13
14	Floradade	1	14	12	12	13
15	Floradade	1	14	12	12	13
16	HC23E-2(93)	1	14	12	13	13
17	HC23E-2(93)	1	14	12	13	13
18	HC353-1	1	12	13	13	13
19	HC353-1	1	12	13	13	13
20	HC84173	1	12	13	12	13
21	HC84173	1	12	13	12	13
22	HC98248	1	14	12	13	13
23	HC98248	1	12	12	13	13
24	HC99471-3	1	12	12	13	13
25	HC99471-3	1	12	12	13	13
26	HCEBR2	1	14	12	13	13
27	HCEBR2	1	14	12	13	13
28	Ohio-MR13	1	14	12	12	13
29	Ohio-MR13	1	14	12	12	13

 A matrix where the data for individuals are in rows, the loci are in column

- ***n* consecutive rows** have the data for each individual of *n*-ploid species
- **Integer** should be used for coding genotype
- Missing data should be indicated by a number which doesn't occur elsewhere in the data (e.g. -1)
- The data file should be a **text file (.txt)** not an excel file (.xls) for running STRUCTURE

Summary

- 🍅 Three computer programs, MSA, GGT, and STRUCTURE were introduced for SNP data analysis by providing the following information:
 - What can the programs do?
 - Where can you download them?
 - How can you format input data for each program?



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



Solanaceae Coordinated
Agricultural Project



Downstream analysis with SNP markers

Part II: The use of STRUCTURE software for association mapping of bacterial spot resistance in tomato

Sung-Chur Sim

The Ohio State University, OARDC

SolCAP workshop

Bacterial spot in tomato

- 🍅 A disease complex caused by species of *Xanthomonas* bacteria.
- 🍅 Five physiological races: T1-T5
- 🍅 Sources of resistance from close relatives of cultivated tomato (*Solanum lycopersicum* L.) or *S. pimpinellifolium*
 - Hawaii 7998 (T1)
 - Hawaii 7981 (T3)
 - PI128216 (T3)
 - PI114490 (T1, T2, T3, and T4)



Association analysis models incorporate a correction for population structure

Unified mixed model (Yu et al. 2006)

$$\begin{matrix}
 Y = \mu & \text{REPy} & + & \mathbf{Qw} & + & \text{Marker}\alpha & + & Zv & + & \text{Error} \\
 \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1_n \end{bmatrix} & + & \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} & + & \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} & + & \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} & + & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} & + & \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
 \end{matrix}$$

Adding a matrix, Qw , of population structure can correct for pseudo-linkage and can add insight to which crosses, pedigrees, subpopulations have the highest breeding value

STRUCTURE analysis

Format marker data

The screenshot shows a Microsoft Excel spreadsheet titled "InputData_SpotPopulation [Compatibility Mode] - Microsoft Excel". The spreadsheet contains a table of marker data. The columns are labeled A through X, and the rows are numbered 1 through 27. The data consists of numerical values for each marker ID. The cell at row 12, column R (R12) is highlighted with a black border.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23
2	6111R1	11	13	11	13	12	13	13	11	11	12	12	12	13	11	-1	14	13	11	14	13	12	13	
3	6111R1	11	13	14	13	13	11	14	14	14	12	12	11	12	11	-1	14	13	11	14	13	12	13	
4	6111R2	11	13	11	13	12	13	13	11	11	12	12	12	13	11	11	11	13	11	14	13	12	13	
5	6111R2	11	13	14	13	12	13	13	14	14	12	12	11	13	11	11	11	13	11	11	13	12	13	
6	6111R3	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	11	13	11	14	13	12	13	
7	6111R3	11	13	14	13	12	13	13	11	14	12	12	12	13	11	11	11	13	11	14	13	12	13	
8	6111S1	11	13	14	13	12	13	13	11	14	12	12	12	13	11	-1	11	13	11	11	13	12	13	
9	6111S1	11	13	14	13	12	13	13	11	14	11	12	12	13	11	-1	11	11	11	11	13	12	13	
10	6111S2	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	11	13	11	14	13	12	13	
11	6111S2	11	13	14	13	13	11	14	11	14	12	12	12	13	11	11	14	11	11	11	13	12	13	
12	6115S3	11	13	14	13	12	13	13	11	14	12	-1	12	13	11	11	11	13	11	14	13	12	12	
13	6115S3	11	13	14	13	13	11	14	11	14	12	-1	12	13	11	11	11	13	14	11	13	12	12	
14	6115S4	11	13	11	13	13	11	14	11	14	12	12	12	13	11	11	11	13	14	14	13	12	12	
15	6115S4	11	13	14	13	13	11	14	11	14	12	13	12	13	11	11	14	13	14	14	13	12	12	
16	6117R1	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	14	13	11	11	13	12	13	
17	6117R1	11	13	14	13	12	13	13	11	14	12	12	12	13	11	11	14	11	11	11	13	12	13	
18	6117R2	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	11	13	11	14	13	12	12	
19	6117R2	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	14	13	14	14	13	12	12	
20	6117S1	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	14	13	11	14	13	12	12	
21	6117S1	11	13	14	13	13	11	14	11	14	12	12	12	13	11	11	14	11	14	11	13	12	12	
22	6117S2	11	13	14	13	12	13	13	11	11	12	12	12	13	11	11	11	13	11	14	13	12	13	
23	6117S2	11	13	14	13	13	11	14	14	14	12	13	12	13	11	11	11	11	14	11	13	12	13	
24	6117S3	11	13	11	13	13	11	14	11	14	12	12	12	13	11	11	11	13	11	14	13	12	12	
25	6117S3	11	13	11	13	13	11	14	11	14	12	13	12	13	11	11	11	13	14	14	13	12	12	
26	6117S4	11	13	11	13	12	13	13	11	14	12	12	12	13	11	11	14	13	11	11	13	12	12	
27	6117S4	11	13	14	13	13	11	14	11	14	12	12	12	13	11	11	14	11	14	11	13	12	12	

The marker data file used in this example is available on the workshop URL: <http://pbgworks.org/tomato-workshop> (file name: STRUCTURE_InputData.txt)

STRUCTURE analysis

Format marker data



Decide how long to run STRUCTURE
(burnin and MCIC)

Burnin length: how long to run the simulation before collecting data to minimize the effect of the starting configuration
(Recommendation: 10,000 ~100,000)

MCIC length: how long to run the simulation after the burnin to get accurate parameter estimates
(Recommendation: 500,000~1,000,000)

STRUCTURE analysis

Format marker data



Decide how long to run STRUCTURE
(burnin and MCIC)



Run simulations 20 times for each of several different Ks



Identify the best K based on the log likelihood values from the 20 simulations for each K using the non-parametric and/or ΔK methods

Inference of best K (number of populations)

🍷 The log likelihood for each K, $\ln P(D) = L(K)$

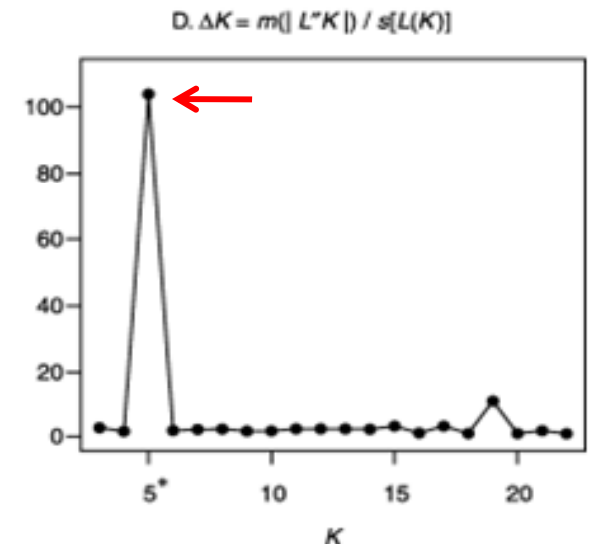
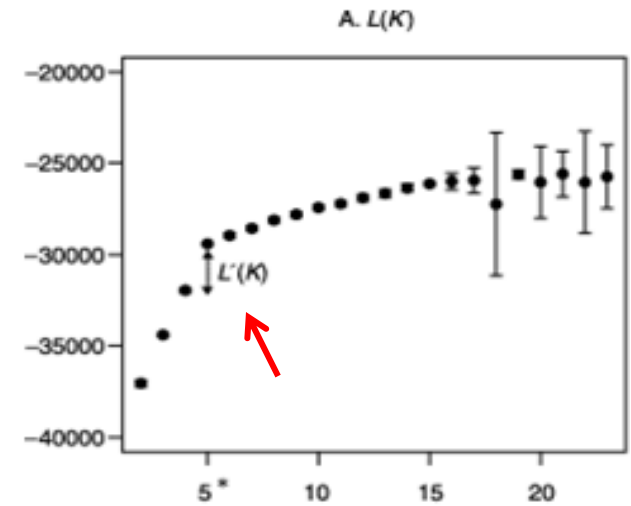
🍷 Two approaches to determine the best K

1. Use of L(K): When K is approaching a true value, L(K) plateaus (or continues increasing slightly) and has high variance between runs (Rosenberg et al. 2001, Evanno et al. 2005).

➡ *nonparametric test (Wilcoxin test)*

2. Use of an ad hoc quantity (ΔK): Calculated based on the second order rate of change of the likelihood (ΔK) (Evanno et al. 2005). The ΔK shows a clear peak at the true value of K.

➡ $\Delta K = m([L''K])/s[L(K)]$



Structure

File Project Parameter Set Plotting View Help

Project - T1

- Project Data
- Project Information
- Simulation Summary
- Parameter Sets
 - 500000
 - Settings
 - Results
 - 500000_run_100 (K=8)
 - 500000_run_101 (K=9)
 - 500000_run_102 (K=9)
 - 500000_run_103 (K=9)
 - 500000_run_104 (K=9)
 - 500000_run_105 (K=9)
 - 500000_run_106 (K=9)
 - 500000_run_107 (K=9)
 - 500000_run_108 (K=9)
 - 500000_run_109 (K=9)
 - 500000_run_10 (K=4)
 - 500000_run_110 (K=9)
 - 500000_run_111 (K=9)
 - 500000_run_112 (K=9)
 - 500000_run_113 (K=9)
 - 500000_run_114 (K=9)
 - 500000_run_115 (K=9)
 - 500000_run_116 (K=9)
 - 500000_run_117 (K=9)

Summary of Project T1

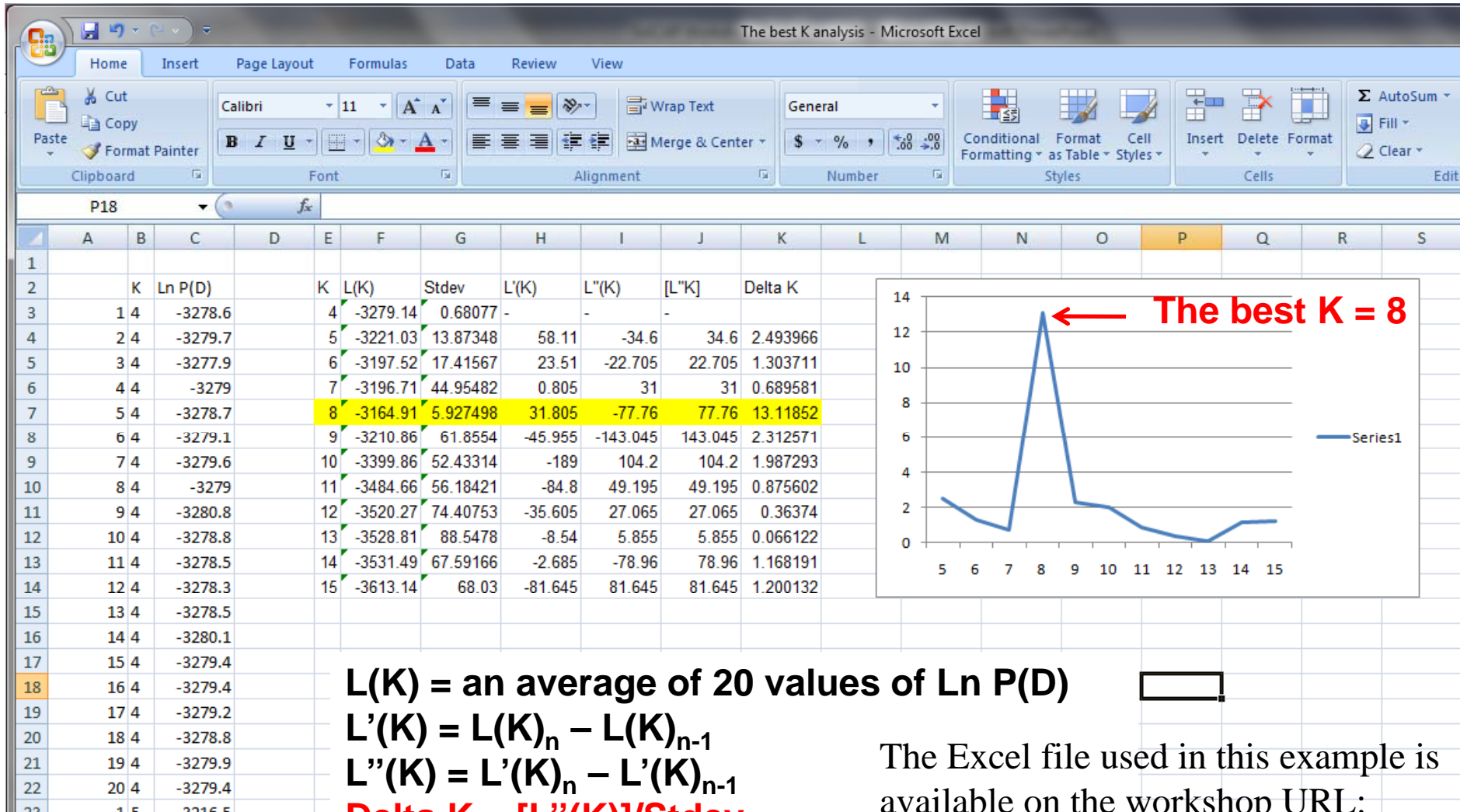
File

Summary of Simulations

Parameter ...	Run Name	K	Ln P(D)	Var[LnP(D)]	α_1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6	Fst_7
500000	500000_run_9	4	-3278.6	206.2	0.2052	0.3857	0.4197	0.6008	0.2779	-	-	-
500000	500000_run_8	4	-3279.7	207.8	0.2056	0.2752	0.4202	0.3860	0.6039	-	-	-
500000	500000_run_7	4	-3277.9	204.8	0.2047	0.5993	0.3861	0.2764	0.4203	-	-	-
500000	500000_run_6	4	-3279.0	206.8	0.2052	0.4188	0.6014	0.3874	0.2763	-	-	-
500000	500000_run_5	4	-3278.7	205.8	0.2046	0.2757	0.3860	0.4199	0.6002	-	-	-
500000	500000_run_4	4	-3279.1	207.0	0.2043	0.3842	0.5991	0.2756	0.4204	-	-	-
500000	500000_run_3	4	-3279.6	208.0	0.2060	0.6010	0.2762	0.3868	0.4202	-	-	-
500000	500000_run_2	4	-3279.0	206.9	0.2045	0.3843	0.6003	0.4195	0.2769	-	-	-
500000	500000_run_20	4	-3280.8	210.4	0.2047	0.5984	0.2766	0.3868	0.4198	-	-	-
500000	500000_run_1	4	-3278.8	206.3	0.2051	0.4198	0.2768	0.6013	0.3864	-	-	-
500000	500000_run_19	4	-3278.5	206.0	0.2041	0.5977	0.3862	0.4193	0.2777	-	-	-
500000	500000_run_18	4	-3278.3	204.9	0.2047	0.4190	0.6000	0.2746	0.3856	-	-	-
500000	500000_run_17	4	-3278.5	205.7	0.2039	0.4193	0.5983	0.3855	0.2763	-	-	-
500000	500000_run_16	4	-3280.1	209.1	0.2051	0.5992	0.3861	0.4203	0.2767	-	-	-
500000	500000_run_15	4	-3279.4	207.5	0.2059	0.4196	0.2770	0.6027	0.3875	-	-	-
500000	500000_run_14	4	-3279.4	207.4	0.2051	0.4187	0.3867	0.2751	0.6023	-	-	-
500000	500000_run_13	4	-3279.2	206.9	0.2055	0.4193	0.3863	0.2768	0.6024	-	-	-
500000	500000_run_12	4	-3278.8	206.1	0.2048	0.5975	0.2769	0.4198	0.3849	-	-	-
500000	500000_run_11	4	-3279.9	207.2	0.2044	0.4187	0.2748	0.3849	0.6022	-	-	-
500000	500000 run 10	4	-3279.4	207.6	0.2060	0.2757	0.3890	0.4195	0.6033	-	-	-

Log likelihood values

Inference of best K using the delta K method



$L(K)$ = an average of 20 values of $\ln P(D)$

$L'(K) = L(K)_n - L(K)_{n-1}$

$L''(K) = L'(K)_n - L'(K)_{n-1}$

Delta K = $[L''(K)]/Stdev$

The Excel file used in this example is available on the workshop URL:
<http://pbgworks.org/tomato-workshop>
 (file name: The best K analysis.xls)

STRUCTURE analysis

Format marker data



Decide how long to run STRUCTURE
(burnin and MCIC)



Run simulations 20 times for each of several different Ks



Identify the best K based on the log likelihood values from the
20 simulations for each K using the non-parametric and/or ΔK
methods



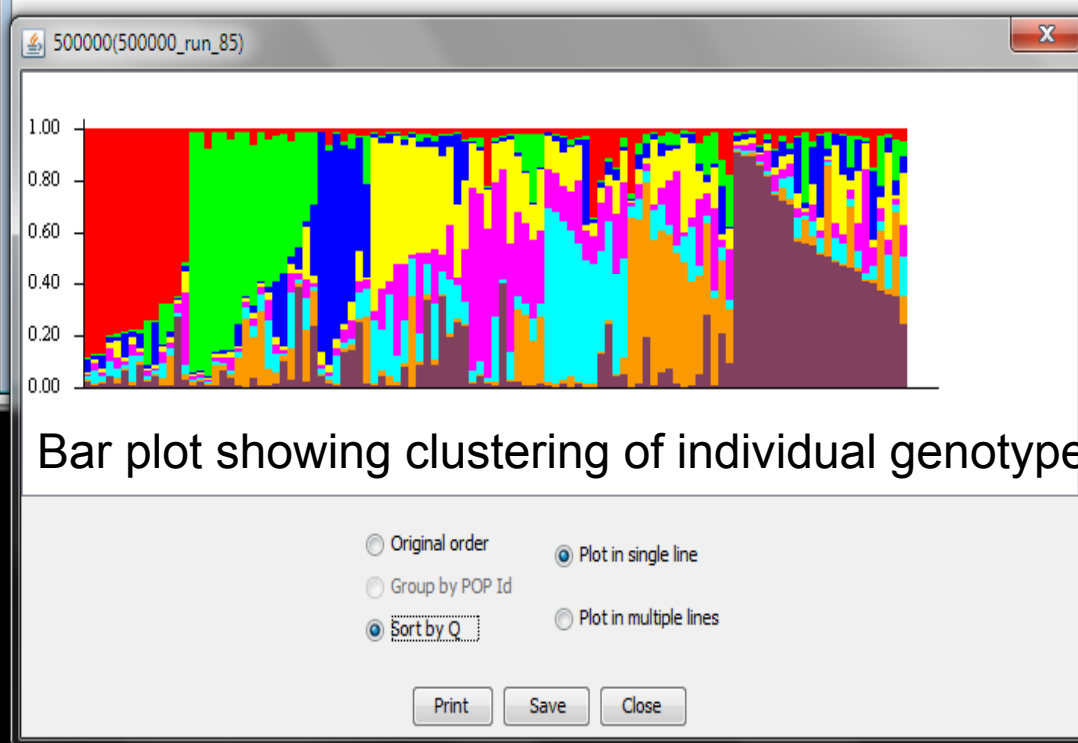
Retrieve a Q-matrix of the best K

- 500000_run_76 (K=7)
- 500000_run_77 (K=7)
- 500000_run_78 (K=7)
- 500000_run_79 (K=7)
- 500000_run_7 (K=4)
- 500000_run_80 (K=7)
- 500000_run_81 (K=8)
- 500000_run_82 (K=8)
- 500000_run_83 (K=8)
- 500000_run_84 (K=8)
- 500000_run_85 (K=8)
- 500000_run_86 (K=8)
- 500000_run_87 (K=8)
- 500000_run_88 (K=8)
- 500000_run_89 (K=8)
- 500000_run_8 (K=4)
- 500000_run_90 (K=8)
- 500000_run_91 (K=8)
- 500000_run_92 (K=8)
- 500000_run_93 (K=8)
- 500000_run_94 (K=8)
- 500000_run_95 (K=8)
- 500000_run_96 (K=8)
- 500000_run_97 (K=8)
- 500000_run_98 (K=8)
- 500000_run_99 (K=8)
- 500000_run_9 (K=4)

Simulation Result: 500000(500000_run_85)

Bar plot Data plot Histogram Triangle plot Tree plot

STRUCTURE by Pritchard, Stephens and Donnelly (2000)
and Falush, Stephens and Pritchard (2003)
Code by Pritchard, Falush and Hubisz
Version 2.3.1 (February 2009)



Structure

File Project Parameter Set Plotting View Help

500000_run_76 (K=7)
500000_run_77 (K=7)
500000_run_78 (K=7)
500000_run_79 (K=7)
500000_run_7 (K=4)
500000_run_80 (K=7)
500000_run_81 (K=8)
500000_run_82 (K=8)
500000_run_83 (K=8)
500000_run_84 (K=8)
500000_run_85 (K=8)
500000_run_86 (K=8)
500000_run_87 (K=8)
500000_run_88 (K=8)
500000_run_89 (K=8)
500000_run_8 (K=4)
500000_run_90 (K=8)
500000_run_91 (K=8)
500000_run_92 (K=8)
500000_run_93 (K=8)
500000_run_94 (K=8)
500000_run_95 (K=8)
500000_run_96 (K=8)
500000_run_97 (K=8)
500000_run_98 (K=8)
500000_run_99 (K=8)
500000_run_9 (K=4)

Simulation Result: 500000(500000_run_85)

Bar plot Data plot Histogram Triangle plot Tree plot

Inferred ancestry of individuals:



Label (%Miss) : Inferred clusters


1	6111R1	(2)	: 0.035	0.010	0.012	0.143	0.524	0.252	0.013	0.011
2	6111R2	(0)	: 0.033	0.009	0.018	0.133	0.245	0.536	0.010	0.015
3	6111R3	(0)	: 0.008	0.005	0.015	0.124	0.150	0.674	0.012	0.011
4	6111S1	(2)	: 0.018	0.006	0.026	0.116	0.150	0.669	0.008	0.008
5	6111S2	(0)	: 0.145	0.012	0.042	0.126	0.234	0.384	0.014	0.043
6	6111S3	(2)	: 0.017	0.006	0.317	0.051	0.026	0.055	0.007	0.520
7	6111S4	(2)	: 0.013	0.006	0.023	0.019	0.018	0.014	0.007	0.899
8	6111R1	(0)	: 0.014	0.004	0.012	0.670	0.041	0.238	0.009	0.011
9	6111R2	(0)	: 0.012	0.007	0.272	0.285	0.030	0.124	0.019	0.251
10	6111S1	(0)	: 0.015	0.006	0.039	0.302	0.053	0.087	0.009	0.489
11	6111S2	(0)	: 0.036	0.009	0.150	0.265	0.207	0.088	0.008	0.238
12	6111S3	(0)	: 0.007	0.013	0.010	0.027	0.023	0.011	0.009	0.899
13	6111S4	(0)	: 0.015	0.007	0.085	0.376	0.061	0.095	0.008	0.353
14	6124R1	(0)	: 0.074	0.007	0.067	0.028	0.061	0.013	0.006	0.744
15	6124R2	(0)	: 0.022	0.010	0.014	0.110	0.426	0.012	0.005	0.401
16	6124R3	(0)	: 0.014	0.008	0.023	0.013	0.017	0.009	0.004	0.911
17	6124R4	(0)	: 0.021	0.016	0.014	0.030	0.079	0.014	0.008	0.818

← Q-matrix

EN

SAS codes

```
%macro Mol(mark);  
proc mixed data = three;  
class &mark gen rep;  
model T1 = pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8 &mark /  
solution;   
random gen rep;   
%mend;  
    %Mol(M1);  
    %Mol(M2);  
    %Mol(M3);  
    %Mol(M4);  
run;
```

 **Marker α**

The SAS code used in this example is available on the workshop URL: <http://pbgworks.org/tomato-workshop> (file name: SAScode.txt)

Summary

- 🍅 STRUCTURE is a useful tool to detect population subdivision
- 🍅 The use of the Q-matrix can correct for subpopulations during association analysis in breeding populations; avoids detection of false-positives
- 🍅 The SNP resources from SolCAP are a powerful survey tool; we should be thinking beyond bi-parental populations toward analysis of complex breeding populations