MadMaize Final Project

BCB 546X
Dec 5, 2018

Laura Tibbs, Ben Cortes, Qi Mu, Jialu Wei, Andy Herr

Blanca *et al. BMC Genomics* (2015) 16:257
DOI 10.1186/s12864-015-1444-1

BMC Genomics

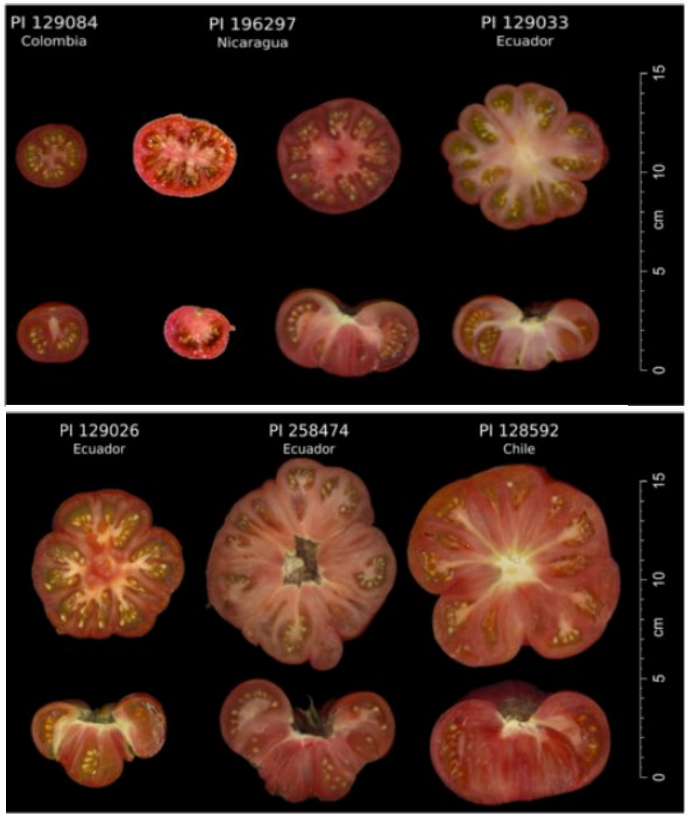
RESEARCH ARTICLE Open Access

Genomic variation in tomato, from wild ancestors to contemporary breeding accessions

José Blanca[1†], Javier Montero-Pau[1†], Christopher Sauvage[2], Guillaume Bauchet[2,3], Eudald Illa[4], María José Díez[1], David Francis[4], Mathilde Causse[2], Esther van der Knaap[4*†] and Joaquín Cañizares[1*†]

Overview

- How has the genome of tomatoes changed following domestication?
- Used a genomic approach to answer this question
  - Genotyping of 1008 lines of tomatoes (7720 SNPs)
    - Tomato Infinium Array
  - Principal component analysis
  - Diversity and genetic differentiation
  - Phylogenetic mapping
  - Genotyping fruit weight and shape loci
- Concluded that tomato domestication occurred in a two-step process
  - 1st domestication in S. America was followed by a 2nd in Mesoamerica



Tomatoes from SLC vintage subgroup - Blanca et al., 2015, http://www.ars.usda.gov

Data Pre-Processing in R

Initial dataset: 1008 samples genotyped at 7720 SNPs (Table S1-S2)

1. Remove markers with > 10% missing data
    - They removed 240 SNPs, we removed 251
2. Remove markers with major allele frequency > 0.95
    - They removed 1137 SNPs, we removed 998

Data Pre-Processing in R, cont.

3. For all downstream analyses except rarefaction, remove “SNPs that mapped closer than 0.1 cM...based on the genetic maps of Sim et al.”

- Problem #1: Sim et al had multiple genetic maps and no consensus map
  - We made consensus map from most complete genetic maps in Sim et al. using R package LPMerge
- Problem #2: 1175 SNPs from our dataset were NOT in the genetic maps in Sim et al.
  - We thinned the SNPs that were in the maps using 0.1 cM criteria, and kept the rest
- Result: 2959 SNPs in our data, 2313 in theirs

Data Pre-Processing in R, cont.

4. Remove duplicate accessions:
    - “When an accession was genotyped more than once and both genotypes were inconsistent...all data for the accession was removed from the analysis… unless it was clear based on the passport information which genotype was correct.”
    - Otherwise, “one randomly chosen genotype representative of the uniquely named accession was used.”
    - Problem: when is it “clear” that one genotype is correct?
    - Result: 950 samples in our data, 952 in theirs
        - Likely different samples due to random selection

Overall:

- Fairly similar but not identical data
- When there are discrepancies in results, this is probably why

Fig 1-2: Principal Component Analysis (PCA)

- Graphically shows population structure and relatedness of samples
- Eigensoft v3.0 on HPC-class to calculate
- R to graph

Fig 1 (theirs):

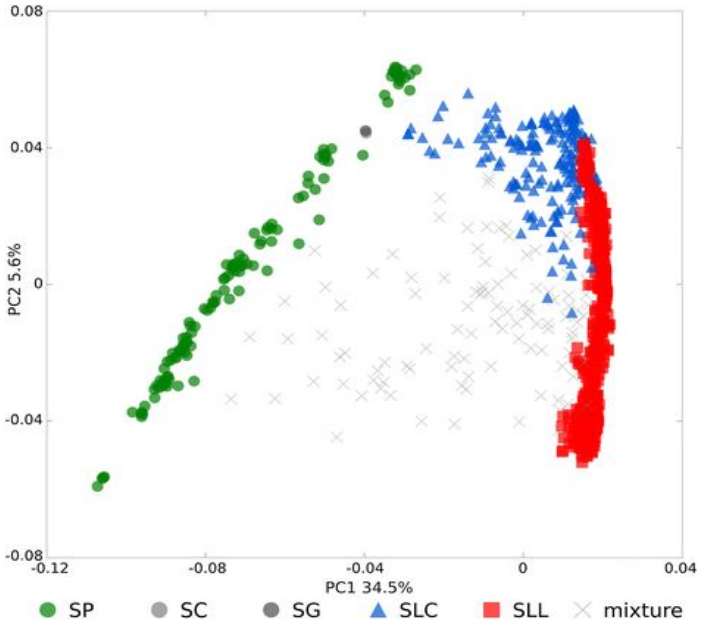


Fig 1 (ours):

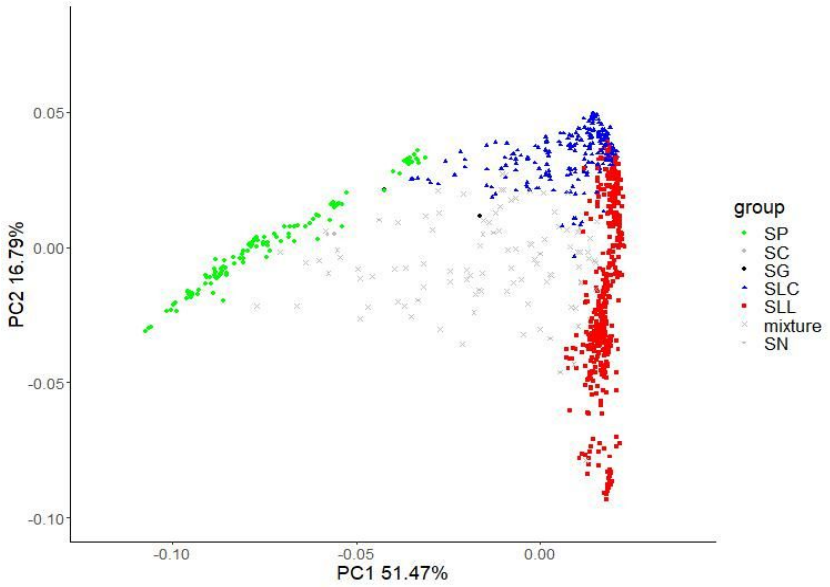


Fig 1-2: Principal Component Analysis (PCA), cont

Fig 2 (theirs):

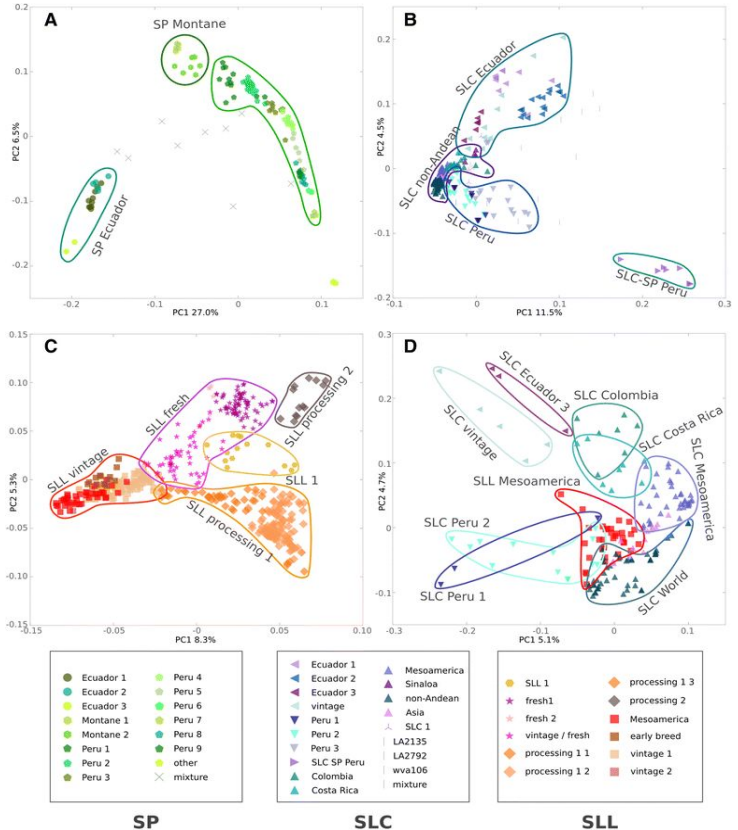


Fig 2 (ours):

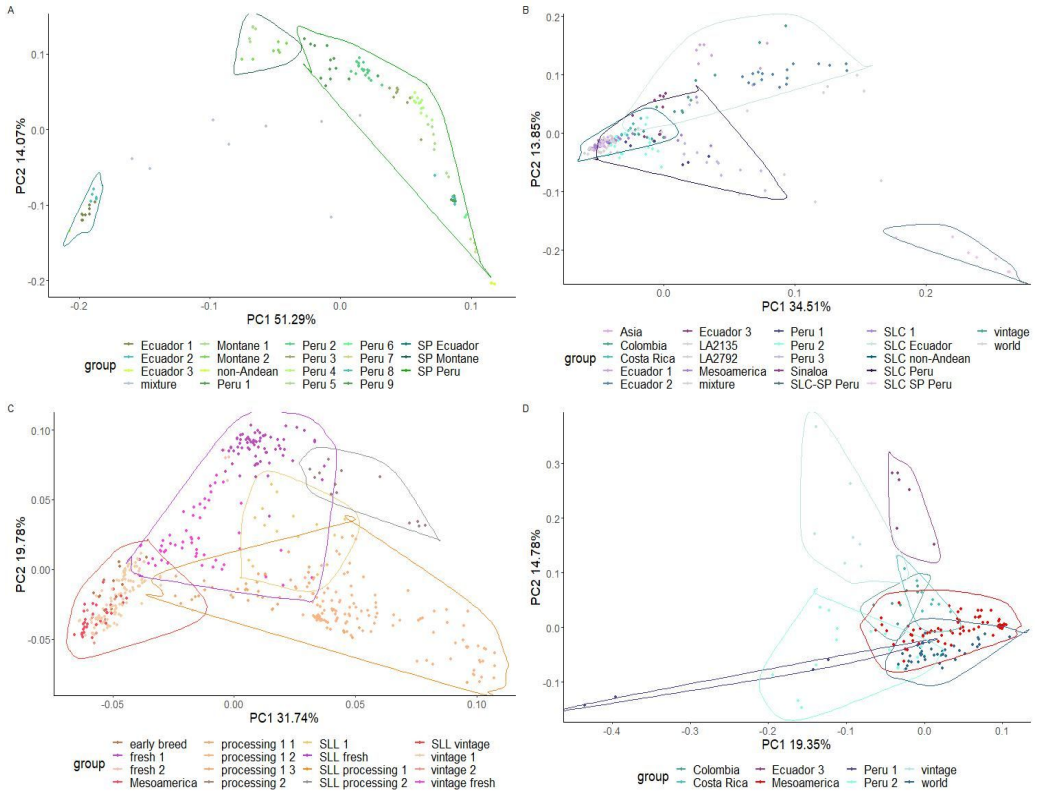
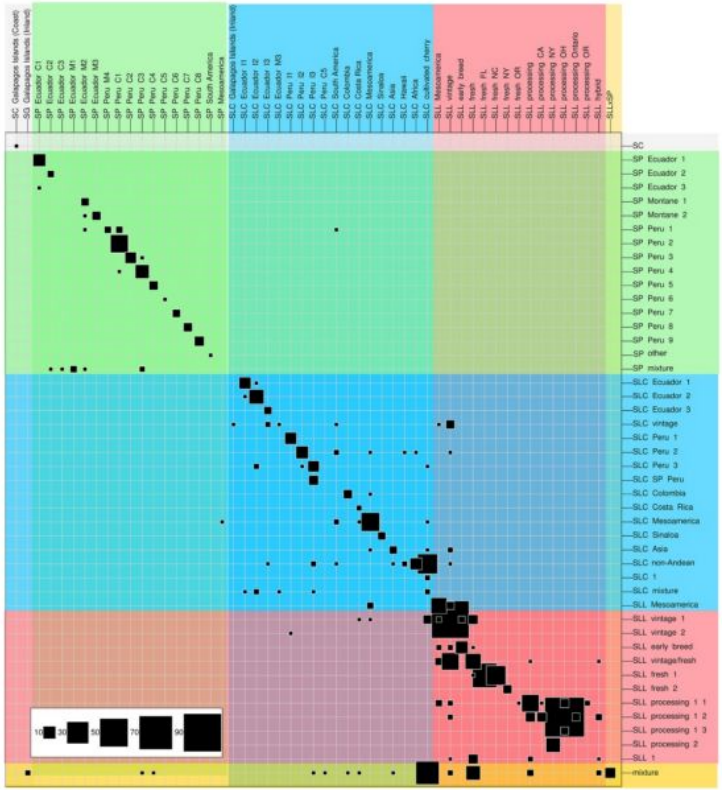


Fig 4: Passport/Genetic Classification Comparison

Theirs



Ours

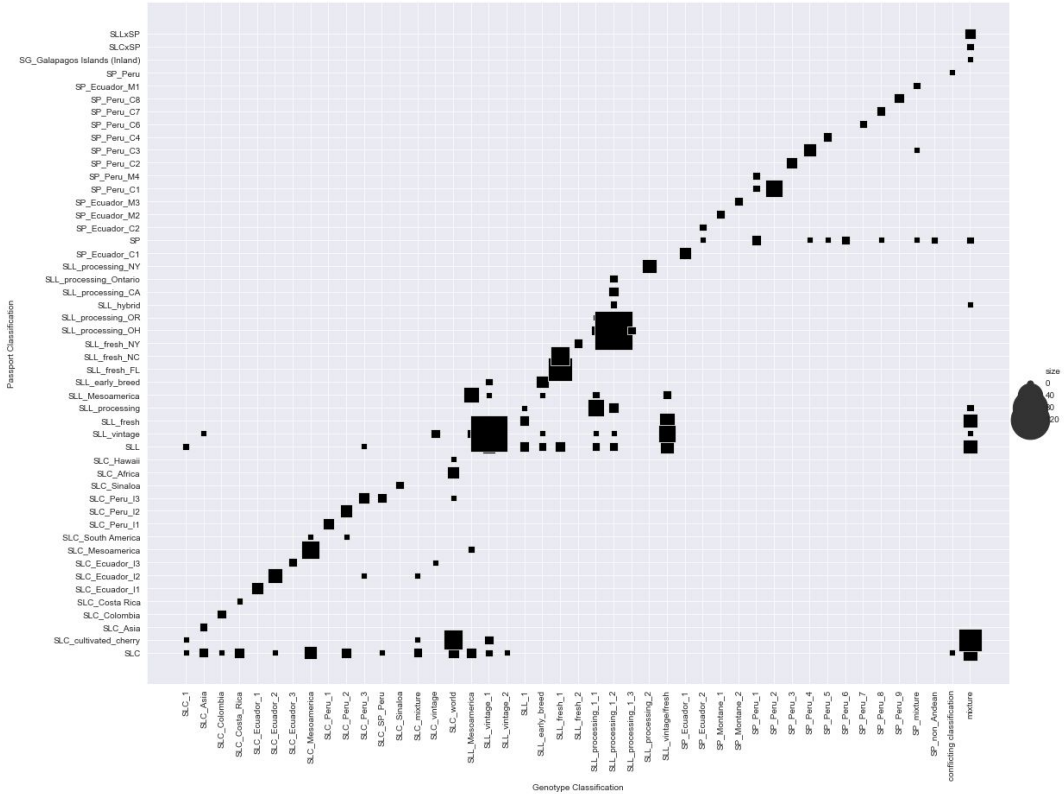


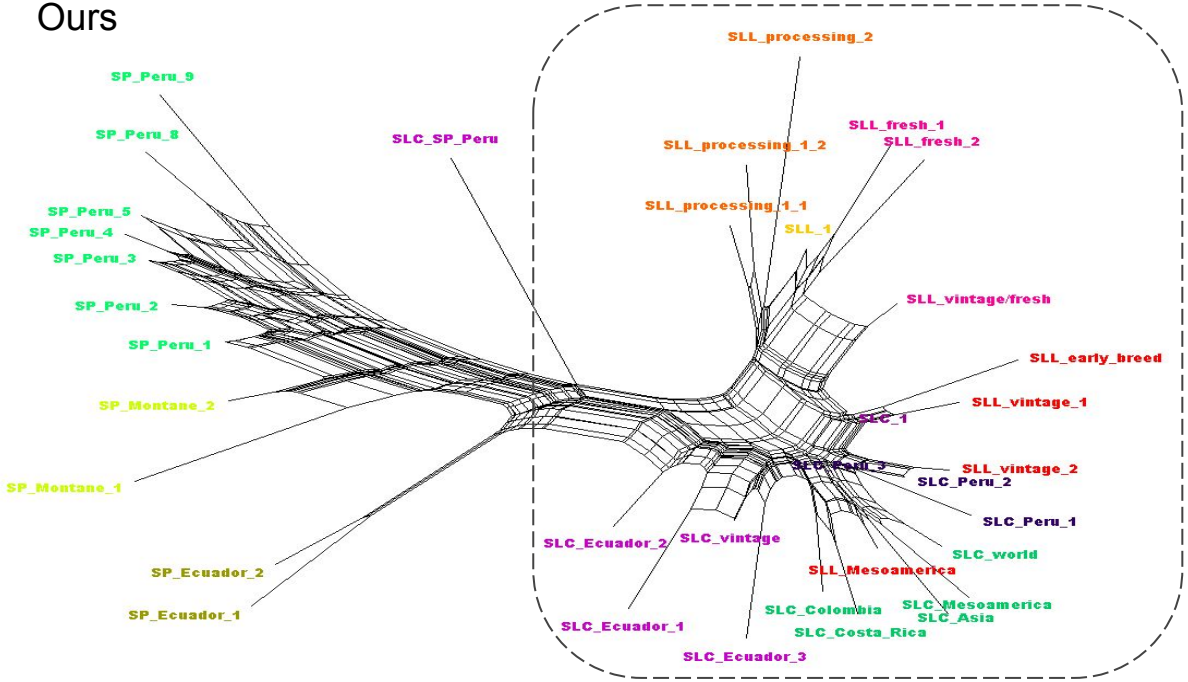
Fig 5: Neighbor network for the genetic subgroups

- Neighbor-Net
  - A distance based method for constructing phylogenetic network that is based on the Neighbor-Joining (NJ) algorithm
  - It uses genetic distance data instead of sequence data
- Genetic distance:
  - A measure of the population differentiation due to genetic structure
  - Gst: most common, but not very accurate in certain conditions
  - Dest: differentiation estimator - authors used, custom python scripts & library (hard to implement)
  - Fst: fixation index, common - I used, R packages
- Data processing
  - Removed groups with <5 individuals, as well as mixture genetic subgroups
  - Final: 760 individuals, 36 subgroups
- Figure visualization
  - SplitsTree software - used by authors and I.

Fig 5A: Neighbor network for the genetic subgroups

Theirs

Ours



Theirs:

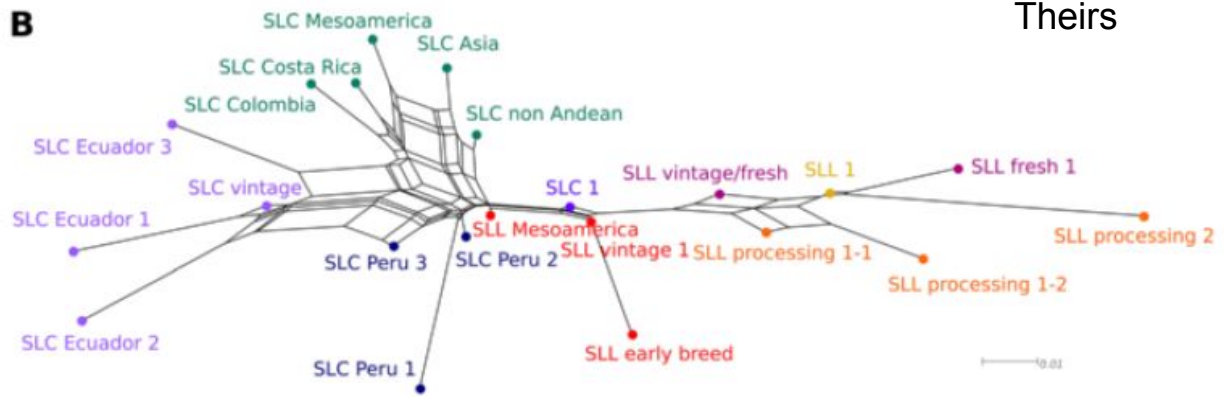
- Custom python scripts
- Custom python library for genetic distance: Dest
- SplitsTree v4.13.1

Ours:

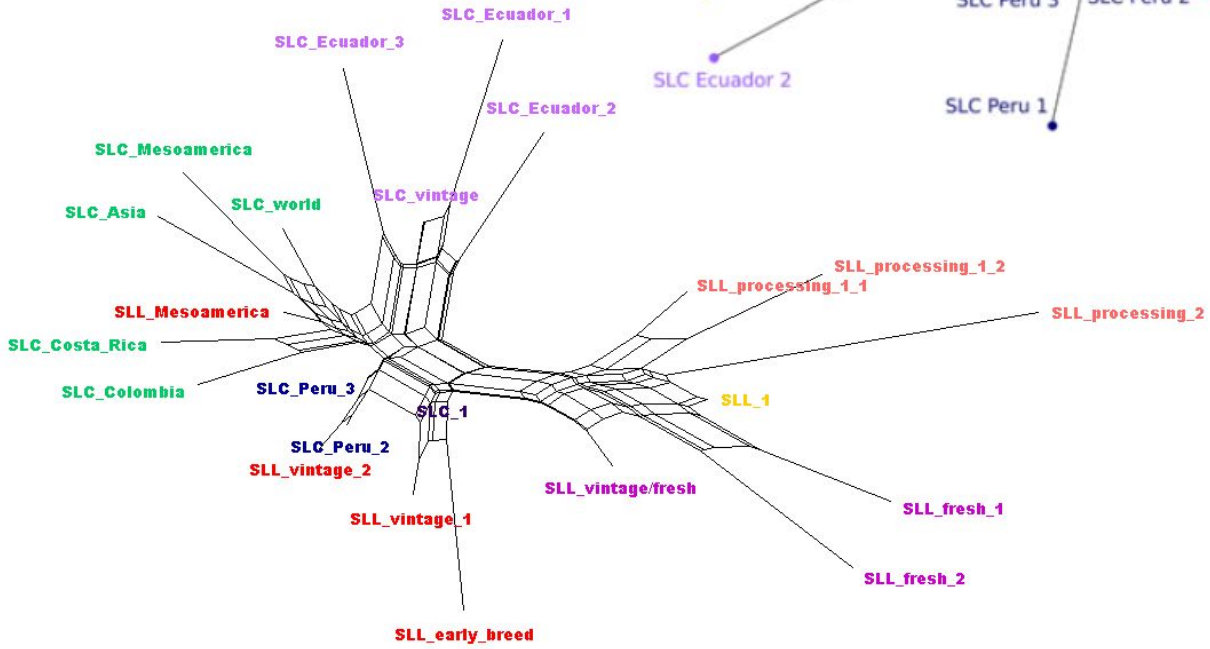
- R script with adegenet, hierfstat, and phangorn packages
- Genetic distance: Fst
- SplitsTree v4.14.6

Fig 5B: Neighbor network for the genetic subgroups

Theirs



Ours

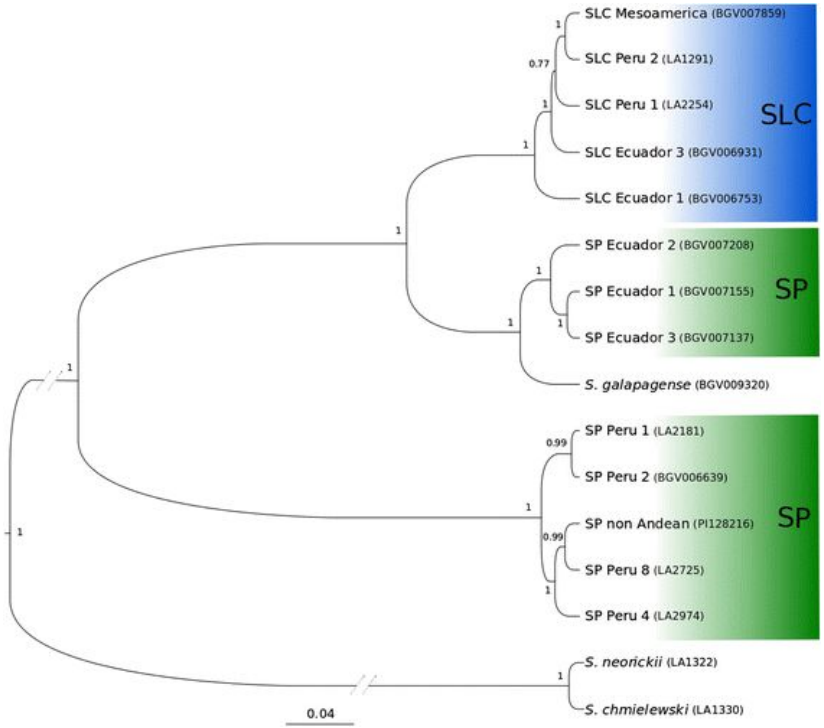


- Reproducibility
  - Technical aspect
  - Overall trend

Fig 6: Phylogenetic tree based on SNP data

theirs:

- *Beast2* - SNAPP : finite-sites model likelihood algorithm within MCMC



ours:

- *TASSEL*: Neighbor Joining Clustering
- *R*: visualizing with ggtree()

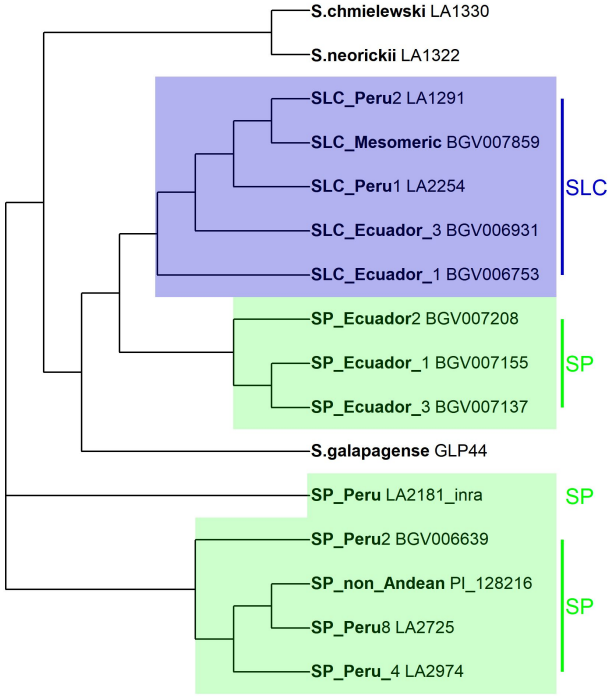
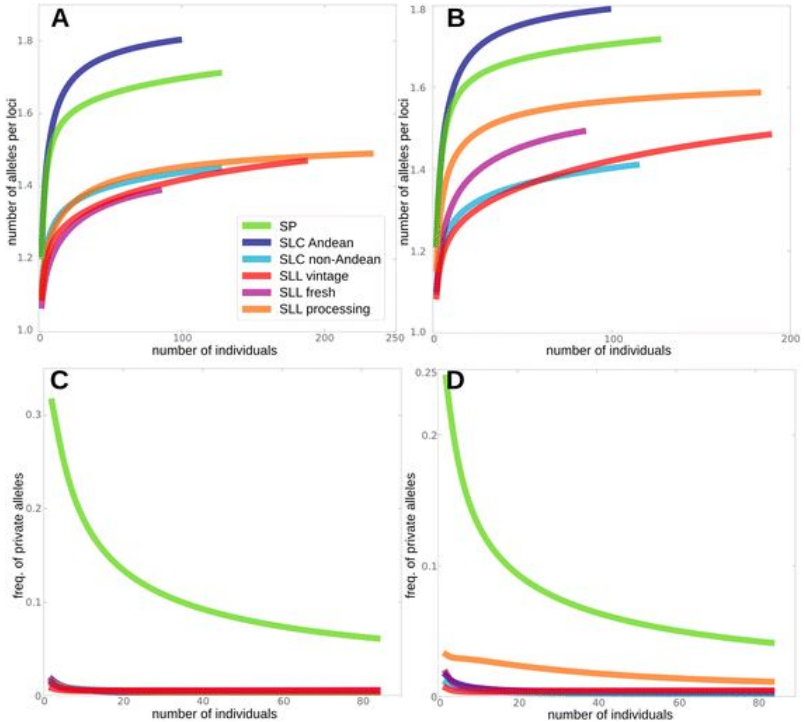


Fig 7: Rarefaction analysis

2959 SNPs 6471 SNPs



theirs

To estimate genetic diversity:

Problem: sample size differ across populations

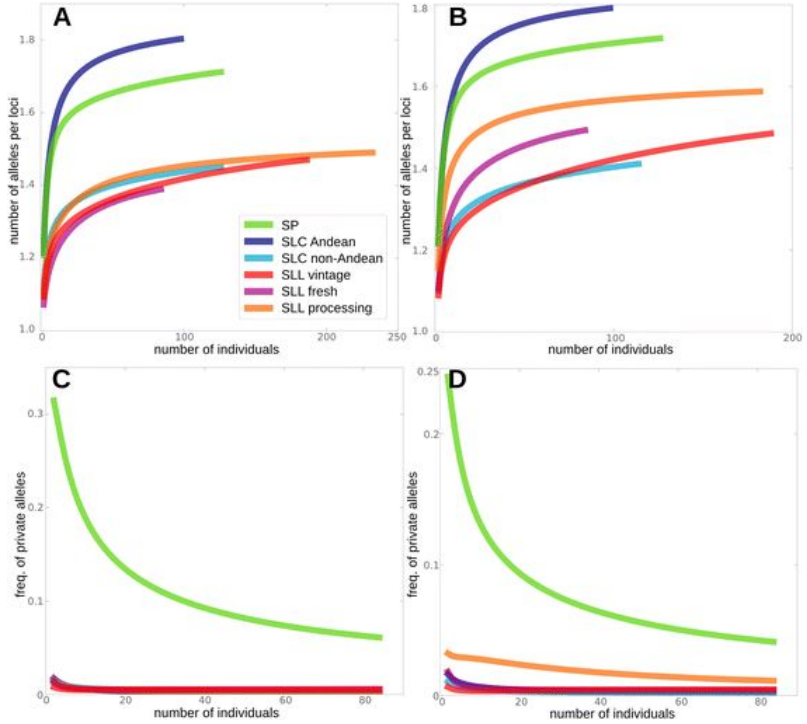
Rarefaction approach:
averaging across all subsamples of size x

→ - Allelic richness
number of distinct alleles in the population

→ - Private allelic richness
number of alleles private to the population

Fig 7: Rarefaction analysis

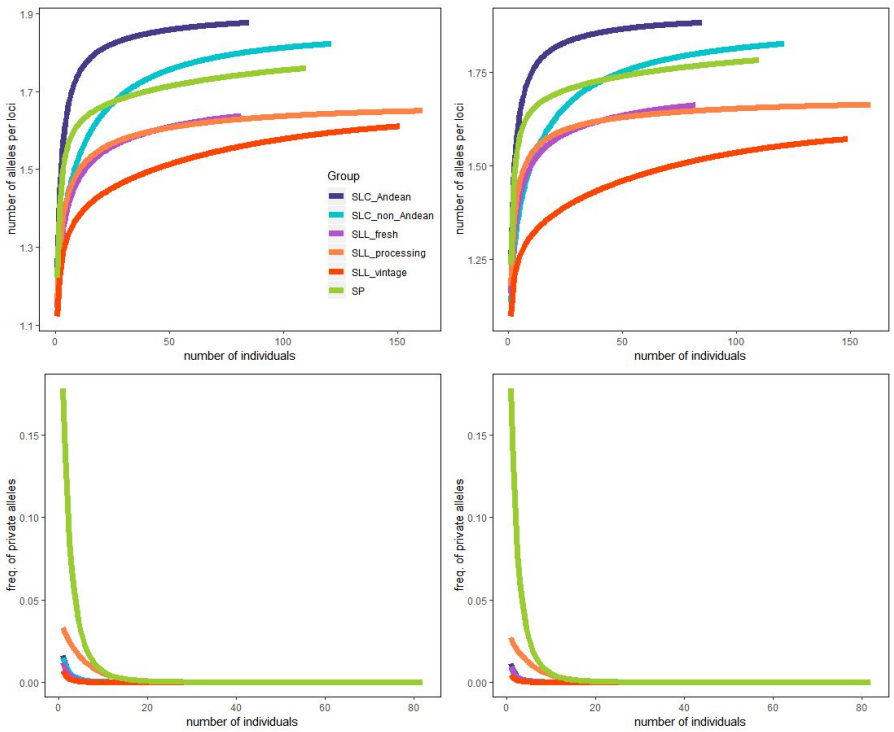
- *ADZE1.0*: analyzing
- *R*: data formatting; visualizing
- Two sets of markers



theirs

2959 SNPs

6471 SNPs

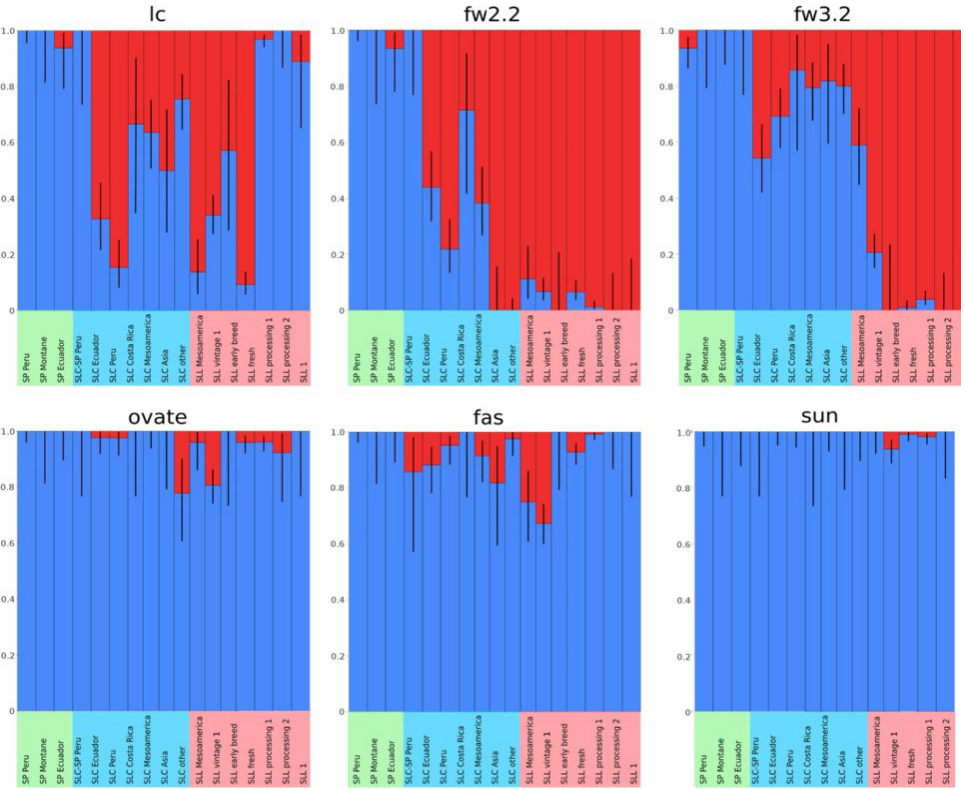


ours

Figure 8. Fruit weight and shape gene frequencies

- Created in R with ggplot2
- Hmisc package for binomial confidence intervals

Theirs



Ours