Novel algorithms for population-scale analysis of plant genomes

Annual Plan Review

Kevin Murray

Borevitz Lab, CPEB, ANU

24 July 2015

Understanding Plant Function and Variation

- We want to understand function of plant systems
- Genetic variation in these systems important

irri.org

Existing genetic approaches

- Forward genetics
  - Select trait/phenotype
  - Randomly mutate genome
  - Screen, map loci affecting phenotype
  - (read backwards is reverse genetics)
- Association mapping in populations
  - Select parents by phenotype, cross
  - Examine variation in phenotype across population
  - Associate phenotype with genotype, map loci

Existing genetic approaches

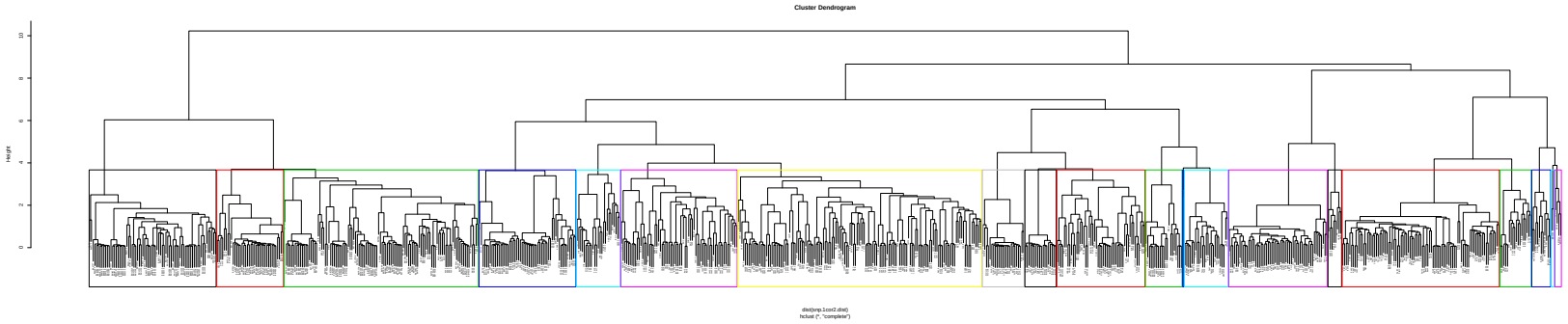
- Forward genetics
  - Select trait/phenotype
  - Randomly mutate genome
  - Screen, map loci affecting phenotype
  - (read backwards is reverse genetics)
- Association mapping in populations
  - Select parents by phenotype, cross
  - Examine variation in phenotype across population
  - Associate phenotype with genotype, map loci
- These approaches **diversity limited**: "missing heritability"

Missing heritability in the field?

- Find more diversity in the field!
- Sample natural populations
  - Ecological hypotheses of trait selection, adaptation
  - Sample widely as possible across non-uniform genetic diversity

Missing heritability in the field?

- Find more diversity in the field!
- Sample natural populations
  - Ecological hypotheses of trait selection, adaptation
  - Sample widely as possible across non-uniform genetic diversity
- Now **complexity limited**: complex kinship & population structure
- Mandates development of economic, accurate large scale population genomics

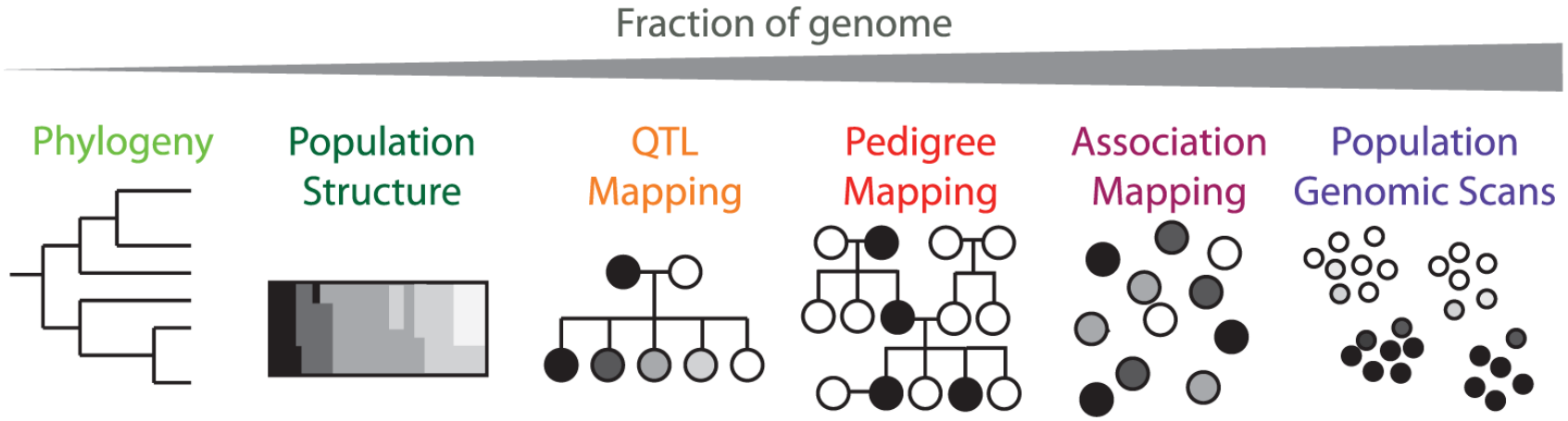


Large-scale genome analysis

- Moving from 100s to 1000s and 10000s of samples *per PhD!*

Large-scale genome analysis

- Moving from 100s to 1000s and 10000s of samples *per PhD!*
- Efficient algorithms to analyse large-scale genomic data
  - Reference & alignment free: *less bias, de novo*
  - Platform/protocol agnostic: *future proof*
  - Computationally efficient: *not the bottleneck*
  - Cross scale: *one tool to rule them all*



after Peterson *et al.* [1]

k-mer analysis

- Analyse k-length words of sequences

$k = 3$

```
ACGTGT
ACG
 CGT
  GTG
   TGT
```

k-mer analysis

- Analyse k-length words of sequences
- Computationally and biologically appropriate
  - Fast
  - Constant-memory (using `khmer`)
  - Scalable and parallelisable
  - Cross-scale

$k = 3$

```
ACGTGT
ACG
 CGT
  GTG
   TGT
```

Thesis overview

- *in silico*, experiment-driven software development
  - New tools, and new combinations of tools (pipelines)
  - Multi-layered analysis of large population sequencing projects

Thesis overview

- *in silico*, experiment-driven software development
  - New tools, and new combinations of tools (pipelines)
  - Multi-layered analysis of large population sequencing projects
- Chapter 1: k-mer based clustering of next gen sequencing
  - First-pass basic clustering
  - *kinship/relatedness*

Thesis overview

- *in silico*, experiment-driven software development
  - New tools, and new combinations of tools (pipelines)
  - Multi-layered analysis of large population sequencing projects
- Chapter 1: k-mer based clustering of next gen sequencing
  - First-pass basic clustering
  - *kinship/relatedness*
- Chapter 2: Machine learning for population genomics
  - Detailed analysis of k-mer genetic distance
  - *population structure, visualisation, sample classification*

Thesis overview

- *in silico*, experiment-driven software development
  - New tools, and new combinations of tools (pipelines)
  - Multi-layered analysis of large population sequencing projects
- Chapter 1: *k*-mer based clustering of next gen sequencing
  - First-pass basic clustering
  - *kinship/relatedness*
- Chapter 2: Machine learning for population genomics
  - Detailed analysis of *k*-mer genetic distance
  - *population structure, visualisation, sample classification*
- Chapter 3: Population "genome-typing by sequencing"
  - Pan-genome variant calling
  - *genome-wide genotyping through whole genome sequencing*

k-mer based clustering

- Extend alignment-free sequence comparison to raw NGS data
- Have released new software package: `kWIP`

Welcome to kWIP's documentation!

Contents:

kWIP

- The k-mer Weighted Inner Product

kWIP

- The k-mer Weighted Inner Product
- Algorithm:
  - For each sample: count all k-mers into a Count-Min Sketch

kWIP

- The k-mer Weighted Inner Product
- Algorithm:
  - For each sample: count all k-mers into a Count-Min Sketch
  - For each analysis set, i.e "population":
    - Calculate the informational entropy of CMS bins (H)
    - For each pair of samples A and B, calculate $\sum_{i=0}^{n} A_i \cdot B_i \cdot H_i$

kWIP

- The k-mer Weighted Inner Product
- Algorithm:
  - For each sample: count all k-mers into a Count-Min Sketch
  - For each analysis set, i.e "population":
    - Calculate the informational entropy of CMS bins (H)
    - For each pair of samples A and B, calculate $\sum_{i=0}^{n} A_i \cdot B_i \cdot H_i$
- The software:
  - `C++`, >2000 lines of code
  - Uses `khmer` for k-mer counting & hashing
  - Parallelised, $\approx$ 10 hrs for 96 rice samples.
  - GNU GPL licensed, source code released on GitHub
- Paper in Prep

kWIP Experiments

- 3000 rice genomes:
  - 3000 rice lines from known families
  - Analysing in sets of ≈ 200, from all major groups
  - Recover known grouping w/ kWIP, not w/ unweighted IP
  - Sensitive to read depth
- Simulation
  - Fake population genome sequencing studies
  - Experiments in progress, early results positive
  - Test limitations of kwip
- Protocol optimisation
  - Effect of varying k
  - Effect of CMS size
  - More appropriate normalisation

Machine learning for Population Genomics

- How to maximise amount of information *k*-mer abundance provides?
  - Pairwise genetic distance: `kWIP`
  - Admixture
  - Online clustering (not all pairs)
  - Visualisation of distance and confidence

Machine learning for Population Genomics

- How to maximise amount of information *k*-mer abundance provides?
  - Pairwise genetic distance: `kWIP`
  - Admixture
  - Online clustering (not all pairs)
  - Visualisation of distance and confidence
- Experiments
  - Detect known introgression and admixture in 3000 rice lines dataset
  - Investigate on-line classifier for novel rice samples: "who am I"
  - Develop HTML5 visualisation of `kWIP` output

Progress

- First-pass basic clustering
  - `kWIP` implemented & optimised
  - Experiments in progress, some complete
  - Paper in prep

Progress

- First-pass basic clustering
  - `kWIP` implemented & optimised
  - Experiments in progress, some complete
  - Paper in prep
- Machine Learning
  - Collaborations started
  - Initial experiments planned

Progress

- First-pass basic clustering
  - `kWIP` implemented & optimised
  - Experiments in progress, some complete
  - Paper in prep
- Machine Learning
  - Collaborations started
  - Initial experiments planned
- Population pan-genome variant calling: *"genome-typing by sequencing"*
  - Initial collaborations started (DIB-lab)
  - Evaluating published tools

Thanks

- Justin, Norman, Sylvain, Gavin and Barry
- Cheng Soon Ong, Christfried Webers
- C. Titus Brown, Michael Crusoe, Camille Scott (DIB-lab) @ UC Davis
- Kenneth McNally/IRRI
- Yourselves

References

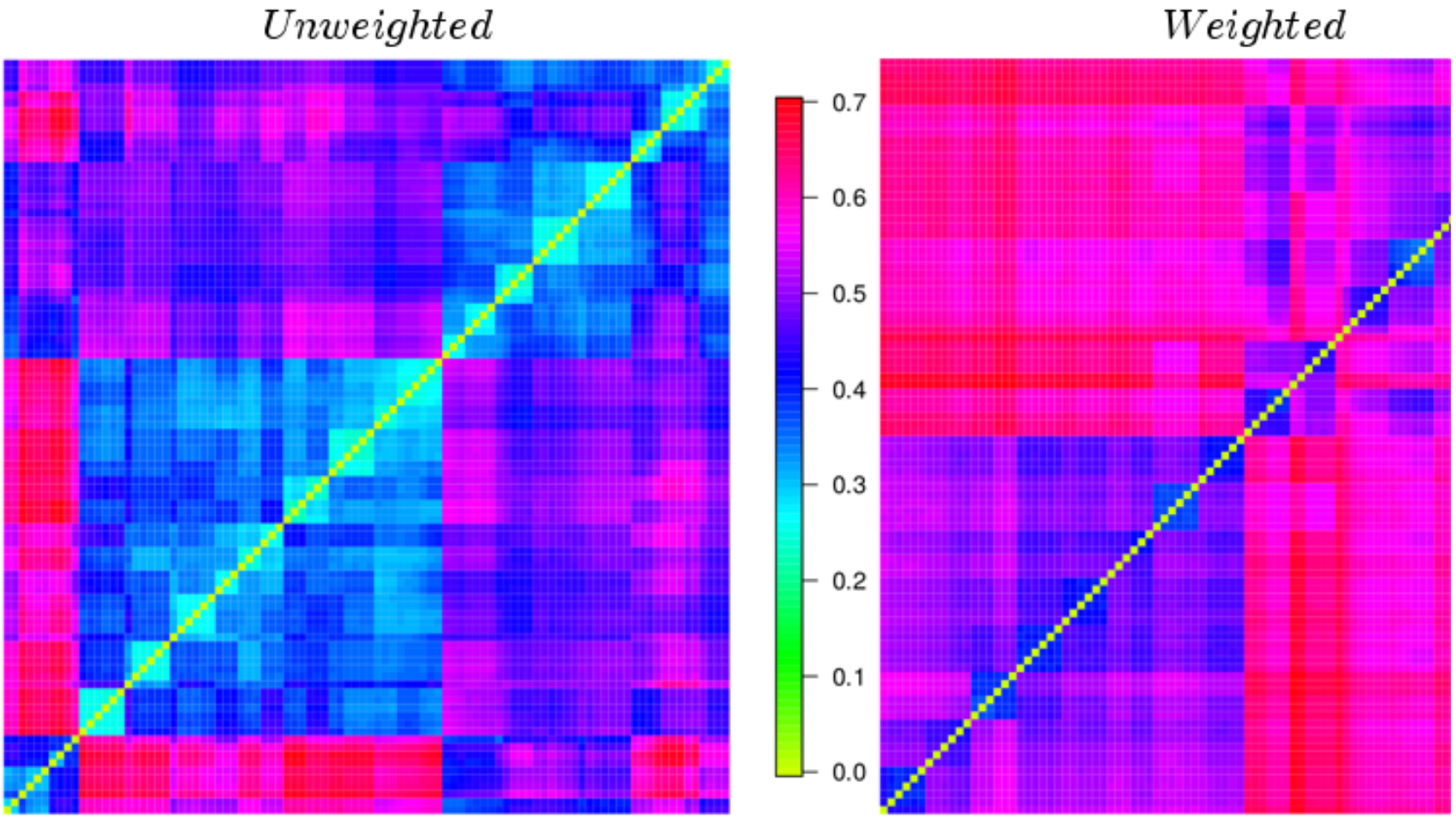
- Peterson, B. K. *et al.* Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* **7,** e37135 (2012).
- Brachi, B., Morris, G. P. & Borevitz, J. O. Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology* **12,** 232 (2011).

.

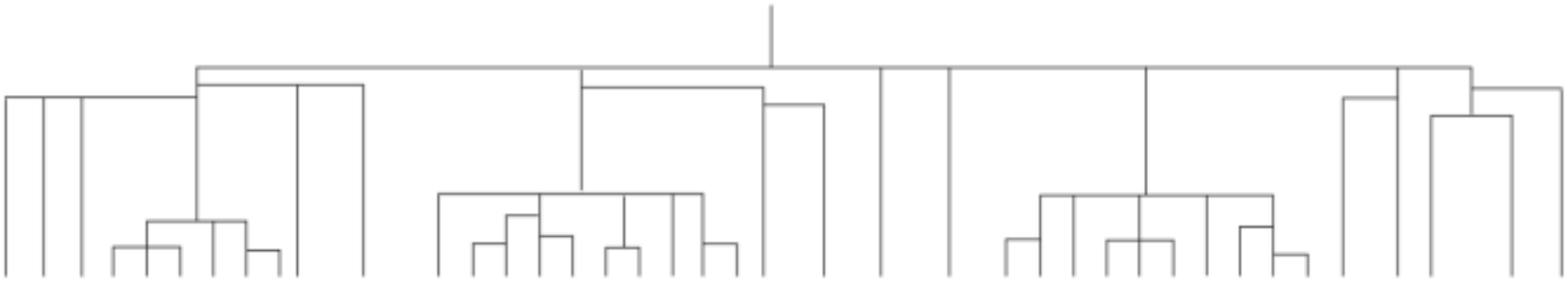
Side Projects

- Reduced Representation Sequence Filtering
  - de Bruijn graph based filter/normaliser
  - With GDU/ABC
- k-mer approaches to detect horizontal gene transfer
  - Collaboration with Adam Taranto

kWIP Distance Matrices

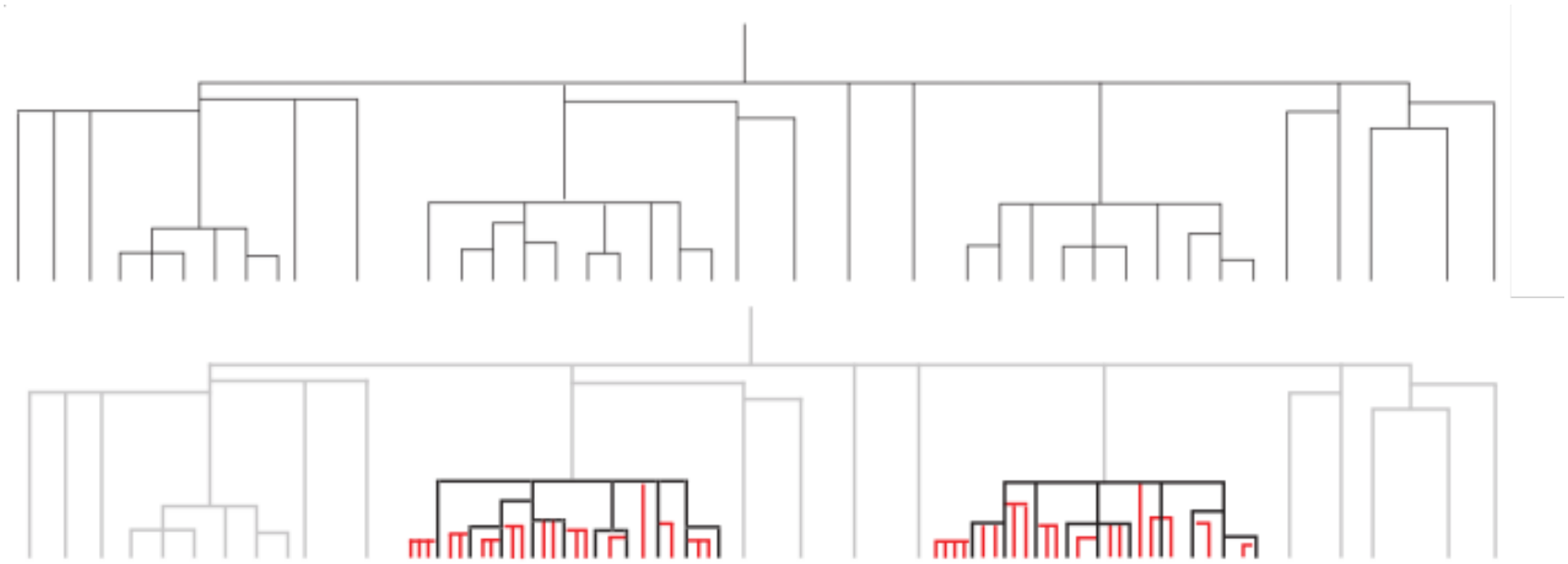


Population Re-structuring



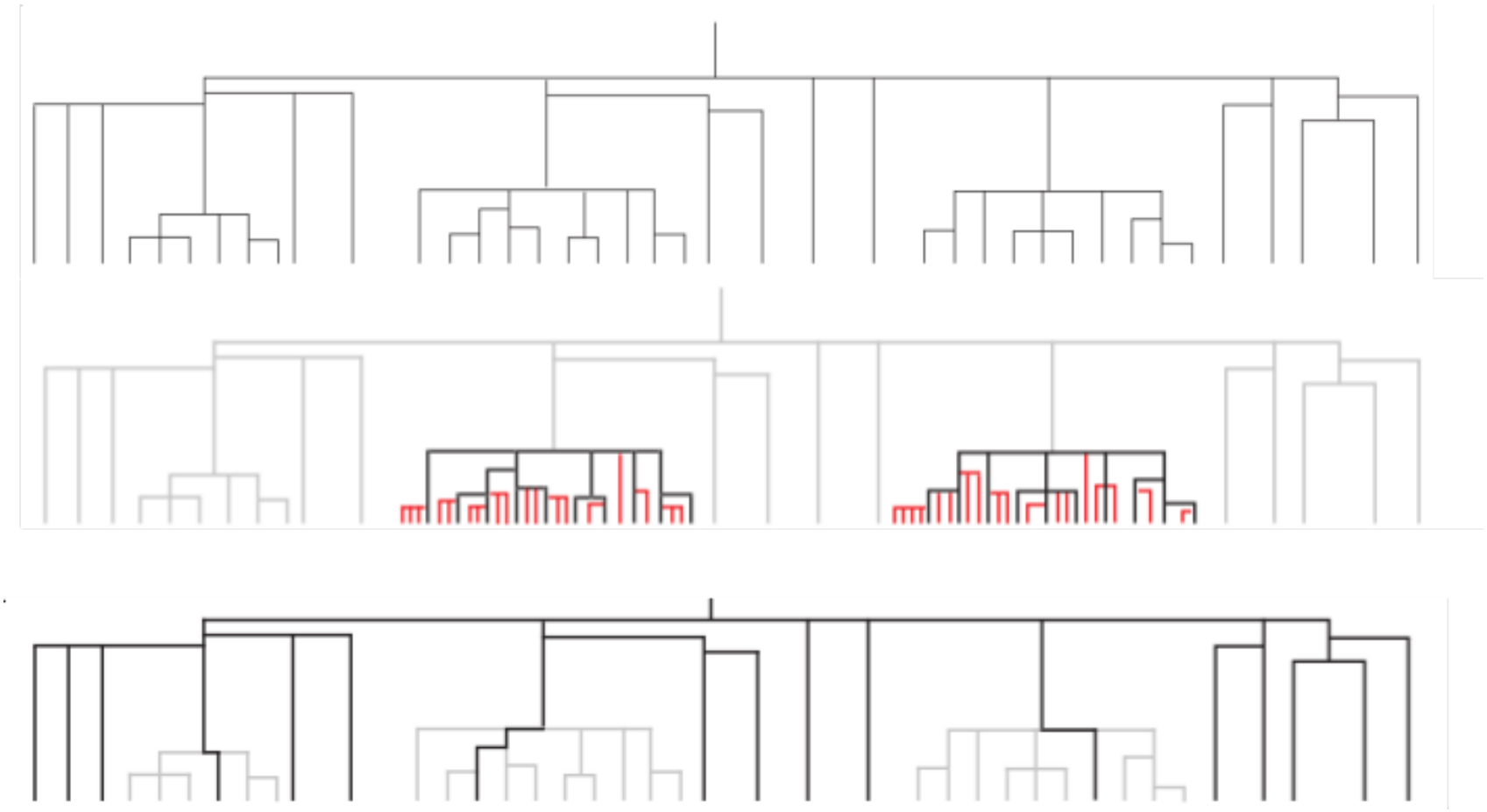
after Brachi *et al.* [2]

Population Re-structuring



after Brachi *et al.* [2]

Population Re-structuring



after Brachi *et al.* [2]