Using local PCA to summarize how relatedness varies along the genome

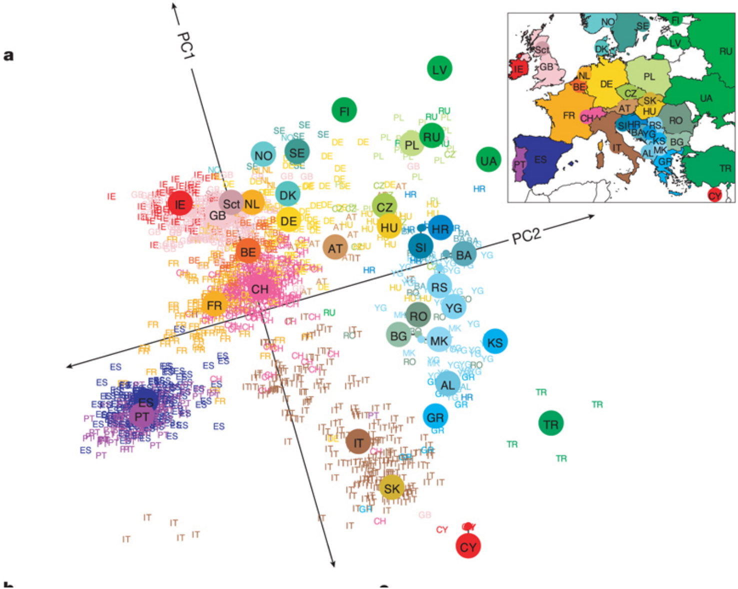
Peter Ralph and Han Li

University of Oregon – Institute of Ecology and Evolution
and University of Southern California

June 26, 2017

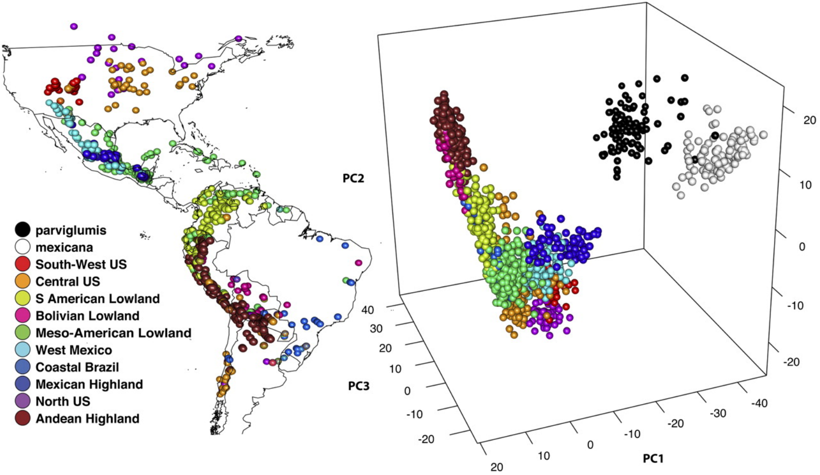
bioRxiv:070615

Principal Components Analysis (PCA)



(Novembre et al 2008)

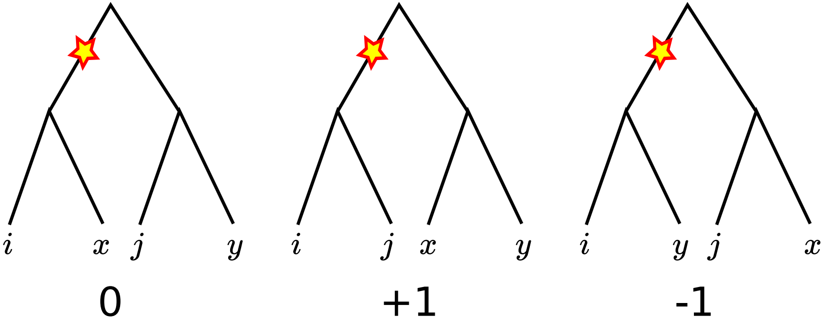
Principal Components Analysis (PCA)



(van Heerwaarden et al 2010)

. . . DESCRIBES THE COVARIANCE MATIX

Genetic covariance between samples *i* and *j* is
the average over loci
and reference samples *x*, *y* of:



. . . so summarizes average patterns of relationships
caused by population structure.

“Population structure”

is historical patterns of interbreeding, migration, and population sizes.

but: linked selection

locally distorts resulting genealogical patterns.
ex: local adaptation, or background selection.

“Population structure”

is historical patterns of interbreeding, migration, and population sizes.

but: linked selection

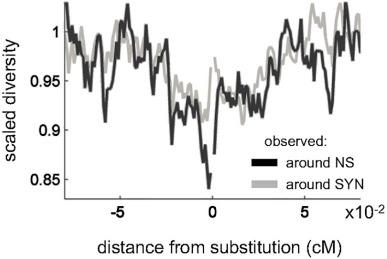
locally distorts resulting genealogical patterns.
ex: local adaptation, or background selection.

“Population structure”

is historical patterns of interbreeding, migration, and population sizes.

but: linked selection

locally distorts resulting genealogical patterns.
ex: local adaptation, or background selection.



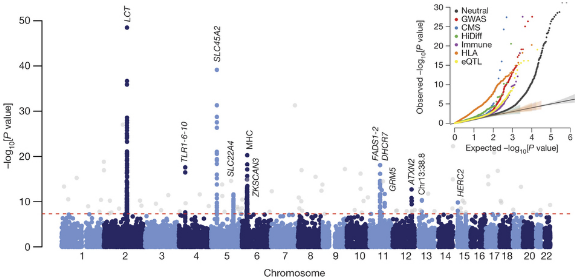
(Elyashiv et al 2016)

“Population structure”

is historical patterns of interbreeding, migration, and population sizes.

but: linked selection

locally distorts resulting genealogical patterns.
ex: local adaptation, or background selection.



(Mathieson et al 2015)

“Population structure”

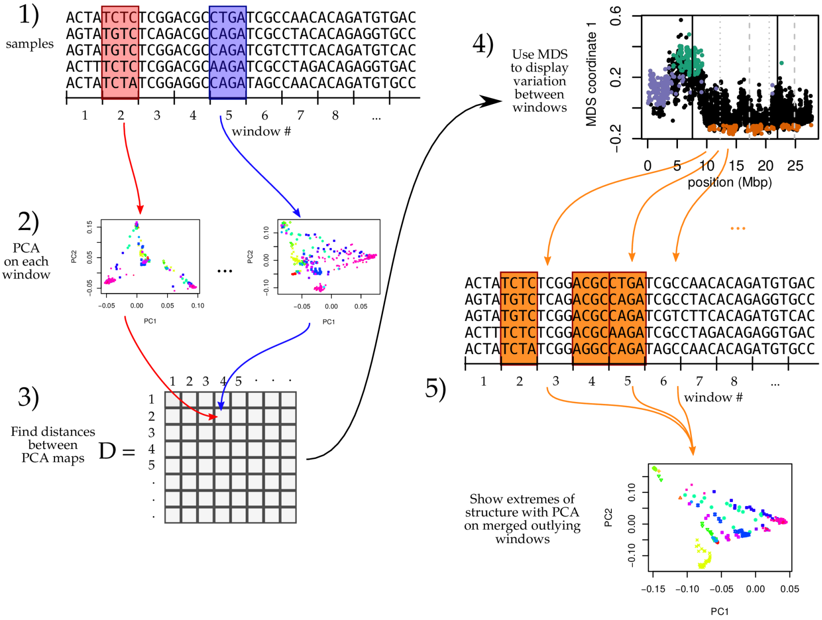
is historical patterns of interbreeding, migration, and population sizes.

but: linked selection

locally distorts resulting genealogical patterns.
ex: local adaptation, or background selection.

Question: is there significant shared variation
in patterns of relatedness?

Our method



1) samples

ACTATCTCTCGGACGCCTGATCGCCAACACAGATGTGAC
AGTATGTCTCAGACGCCAGATCGCCTACACAGAGGTGCC
AGTATGTCTCGGACGCCAGATCGTCTTCACAGATGTCAC
ACTTTCTCTCGGACGCAAGATCGCCTAGACAGAGGTGAC
ACTATCTATCGGAGGCCAGATAGCCAACACAGATGTGCC

2) PCA on each window

3) Find distances between PCA maps D =

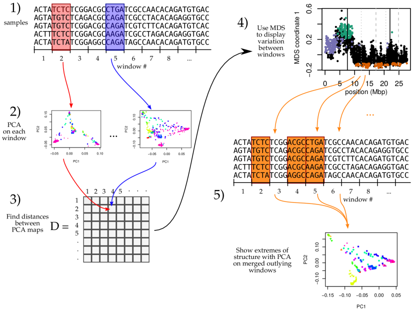
4) Use MDS to display variation between windows

5) Show extremes of structure with PCA on merged outlying windows

ACTATCTCTCGGACGCCTGATCGCCAACACAGATGTGAC
AGTATGTCTCAGACGCCAGATCGCCTACACAGAGGTGCC
AGTATGTCTCGGACGCCAGATCGTCTTCACAGATGTCAC
ACTTTCTCTCGGACGCAAGATCGCCTAGACAGAGGTGAC
ACTATCTATCGGAGGCCAGATAGCCAACACAGATGTGCC

“LOSTRUCT”

- an R package
- with templated Rmarkdown reports
- and a script interface
- `https://github.com/petrelharp/local_pca`

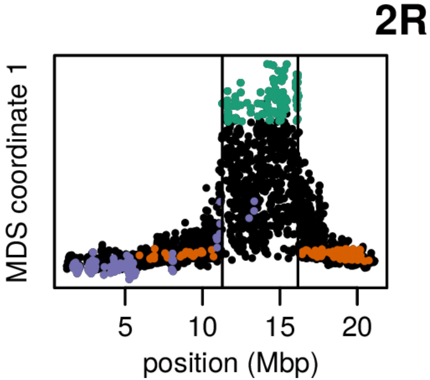

1)
samples
ACTATCTCTCGGACGCCTGATCGCCAACACAGATGTGAC
AGTATGTCTCAGACGCCAGATCGCCTACACAGAGGTGCC
AGTATGTCTCGGACGCCAGATCGTCTTCACAGATGTCAC
ACTTTCTCTCGGACGCAAGATCGCCTAGACAGAGGTGAC
ACTATCTATCGGAGGCCAGATAGCCAACACAGATGTGCC
1 2 3 4 5 6 7 8 ...
window #
2)
PCA on each window
PC1
PC2
...
3)
Find distances between PCA maps
D =
1 2 3 4 5 . . .
4)
Use MDS to display variation between windows
MDS coordinate 1
−0.2 0.2 0.6
0 5 10 15 20 25
position (Mbp)
...
ACTATCTCTCGGACGCCTGATCGCCAACACAGATGTGAC
AGTATGTCTCAGACGCCAGATCGCCTACACAGAGGTGCC
AGTATGTCTCGGACGCCAGATCGTCTTCACAGATGTCAC
ACTTTCTCTCGGACGCAAGATCGCCTAGACAGAGGTGAC
ACTATCTATCGGAGGCCAGATAGCCAACACAGATGTGCC
1 2 3 4 5 6 7 8 ...
window #
5)
Show extremes of structure with PCA on merged outlying windows
PC1
PC2
−0.15 −0.10 −0.05 0.00 0.05
−0.10 0.00 0.10

Data: African *D. melanogaster*

- DPGP (Langley et al 2012; Pool et al 2012; Lack et al 2015)
- 380 mostly African samples – WGS – 9 Kb windows
- large, segregating inversions (Corbett-Detig & Hartl 2012; Langley et al 2012)
- without less common inversion haplotypes: linked selection?

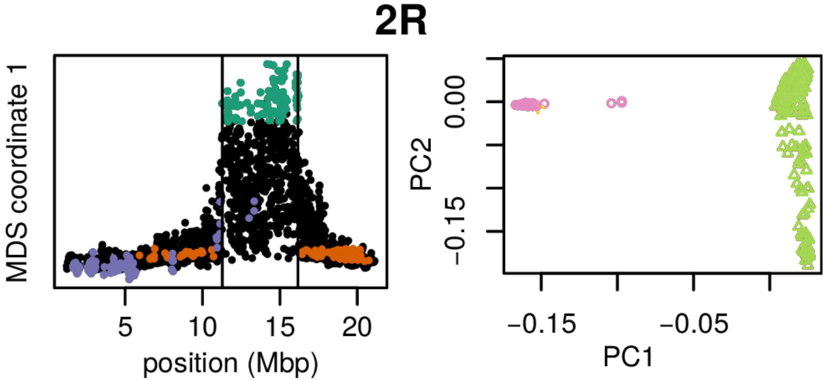
DATA: AFRICAN *D. melanogaster*

- DPGP (Langley et al 2012; Pool et al 2012; Lack et al 2015)
- 380 mostly African samples – WGS – 9 Kb windows
- large, segregating inversions (Corbett-Detig & Hartl 2012; Langley et al 2012)
- without less common inversion haplotypes: linked selection?



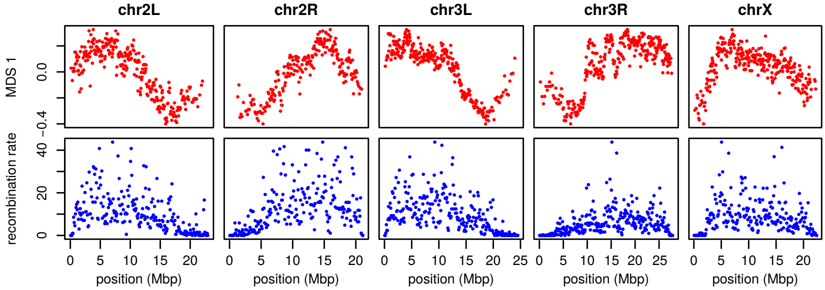
DATA: AFRICAN *D. melanogaster*

- DPGP (Langley et al 2012; Pool et al 2012; Lack et al 2015)
- 380 mostly African samples – WGS – 9 Kb windows
- large, segregating inversions (Corbett-Detig & Hartl 2012; Langley et al 2012)
- without less common inversion haplotypes: linked selection?



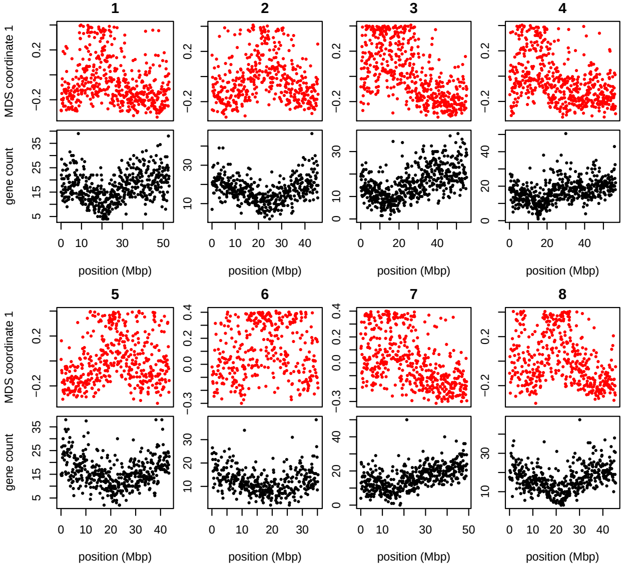
Data: African *D. melanogaster*

- DPGP (Langley et al 2012; Pool et al 2012; Lack et al 2015)
- 380 mostly African samples – WGS – 9 Kb windows
- large, segregating inversions (Corbett-Detig & Hartl 2012; Langley et al 2012)
- without less common inversion haplotypes: linked selection?



Data: *Medicago truncatula* Hapmap (Tang et al 2014)

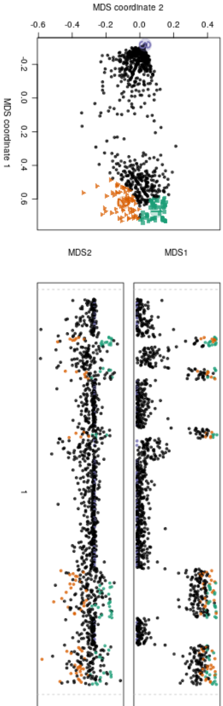
- 263 pan-Mediterranean samples – WGS – 100 Kb windows



Patterns are not driven by:

- recombination rate variation
- polymorphism
- PC switching
- missingness

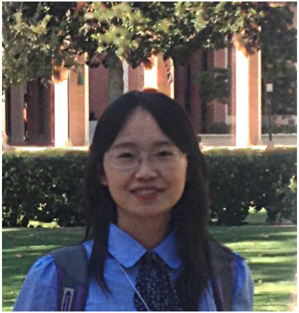
Conclusions



- There may be more than one "population structure"
- We have (more) evidence for strong, widespread effects of linked selection
- The method is applicable to other summary strategies
- `lostruct` is a visualization tool
  try it out: `https://github.com/petrelharp/local_pca`

Thanks

Han Li – USC – bioRxiv:070615

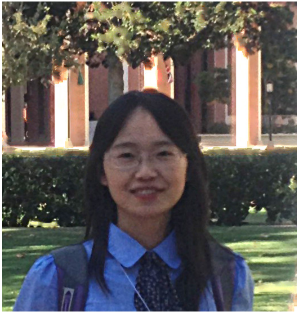


John Pool, Russ Corbett-Detig
Peter Chang, Matilde Cordeiro
Peter Tiffin, Tim Paape
Graham Coop, Jeremy Berg, Yaniv Brandvain, Chuck Langley
Jaime Ashander, Jerome Kelleher

Funding: NSF ABI

Thanks

Han Li – USC – bioRxiv:070615



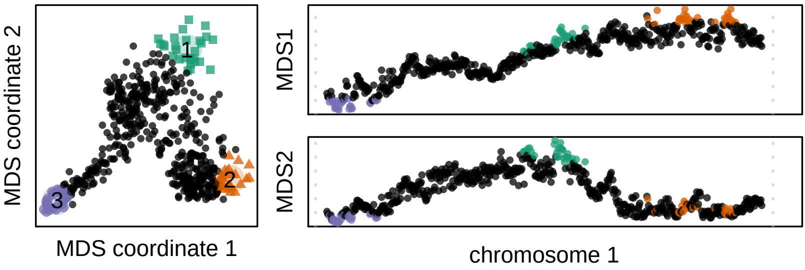
John Pool, Russ Corbett-Detig
Peter Chang, Matilde Cordeiro
Peter Tiffin, Tim Paape
Graham Coop, Jeremy Berg, Yaniv Brandvain, Chuck Langley
Jaime Ashander, Jerome Kelleher

Funding: NSF ABI

University of Oregon

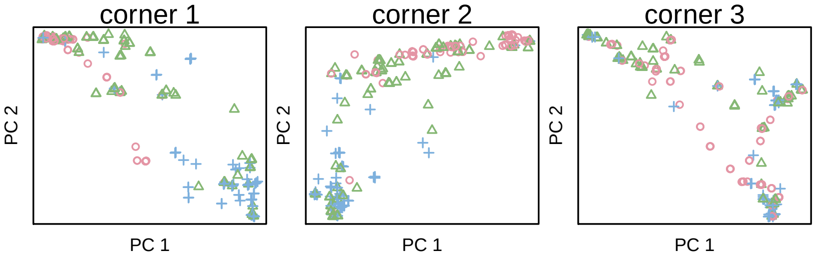
Simulation: local adaptation

- three populations: (hot, dry) – (hot, wet) – (cold, wet)
- clustered genes: (hot/cold) – (wet/dry)
- 1000 diploids in each population and 1% migration
- 25MB (0.625M) with 1000 evenly spaced loci with $s = \pm 0.001$,
- with simuPOP + msprime



Simulation: local adaptation

- three populations: (hot, dry) – (hot, wet) – (cold, wet)
- clustered genes: (hot/cold) – (wet/dry)
- 1000 diploids in each population and 1% migration
- 25MB (0.625M) with 1000 evenly spaced loci with $s = \pm 0.001$,
- with simuPOP + msprime



Robust?

- PC switching?
  uses distance metric insensitive to PC order
- window choice?
  weakly: method selects size maximizing information while minimizing noise
- mutation rate variation?
  normalizes by matrix norm to capture just structure
- recombination rate variation?
  do windows in cM if possible
  else inversions ≈ low recomb regions
- missing data?
  filter; but user beware