# Using local PCA to summarize how relatedness varies along the genome
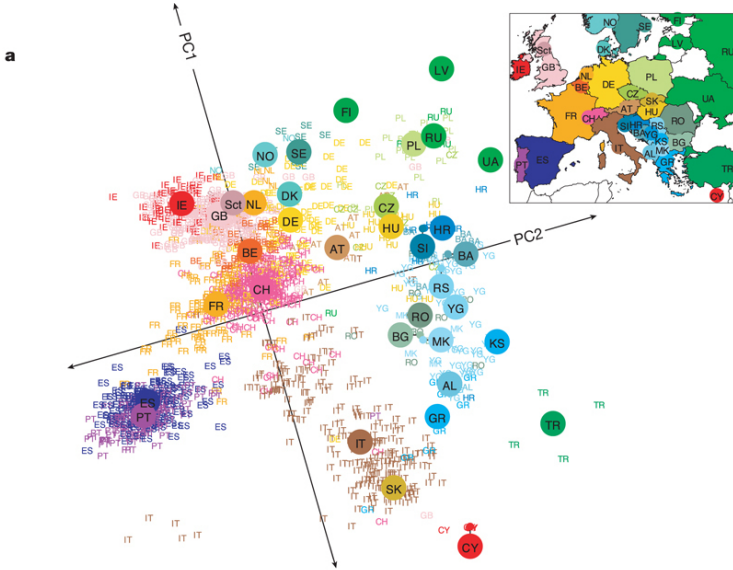
Peter Ralph and Han Li

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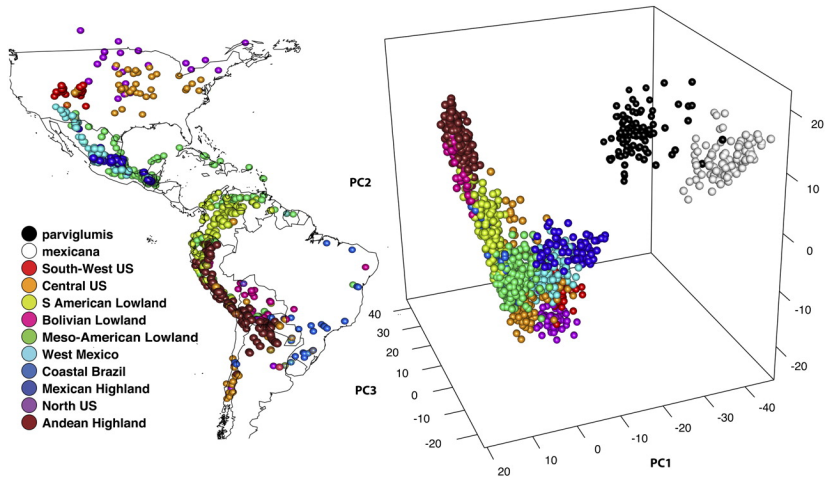*and* University of Southern California

June 26, 2017

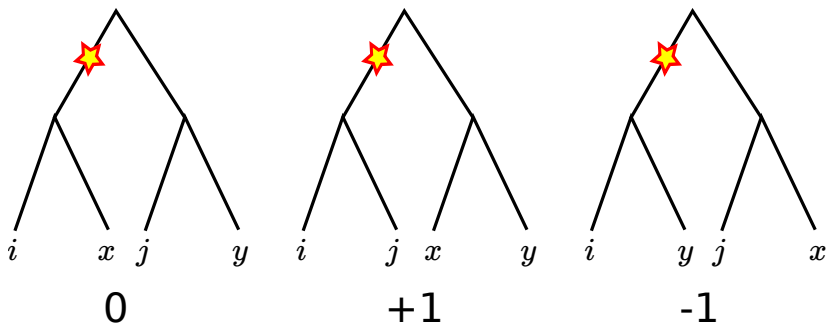# Principal Components Analysis (PCA)



(Novembre et al 2008)

# Principal Components Analysis (PCA)



(van Heerwaarden et al 2010)

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.
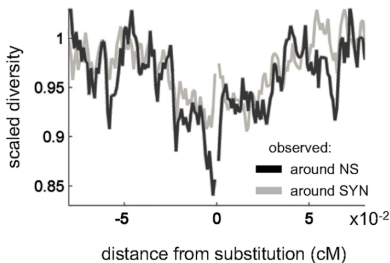ex: local adaptation, or background selection.



distance from substitution (cM)

(Elyashiv et al 2016)

# "Population structure"

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

### but: linked selection

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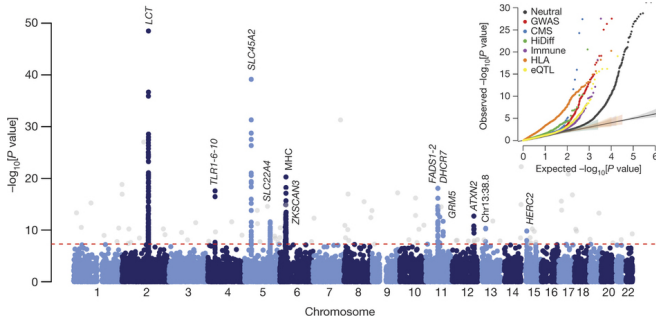ex: local adaptation, or background selection.

# "Population structure"

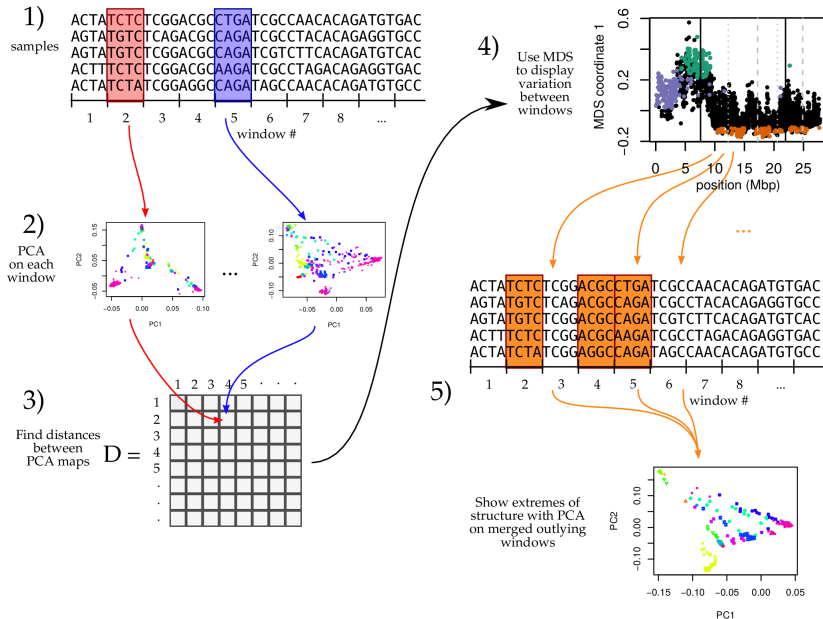is historical patterns of interbreeding, migration, and population sizes.

but: linked selection

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

Goal: describe shared, large-scale variation variation.

# Our method



1) samples

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

# "LOSTRUCT"

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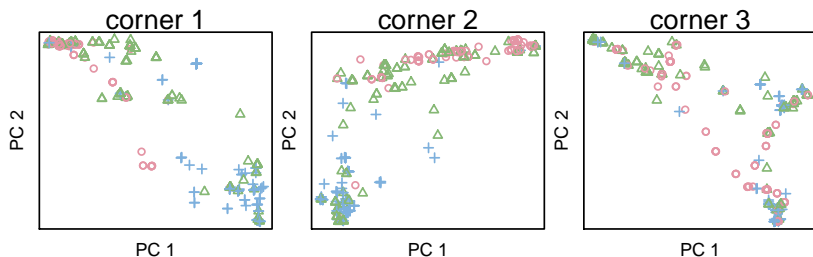- and a script interface
- https://github.com/petrelharp/local_pca
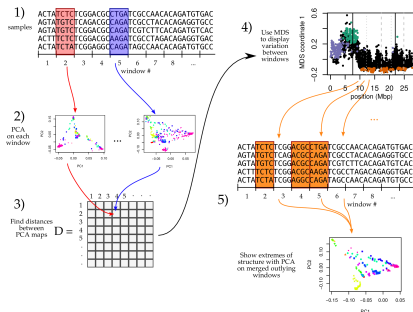
# SIMULATION: LOCAL ADAPTATION

- 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
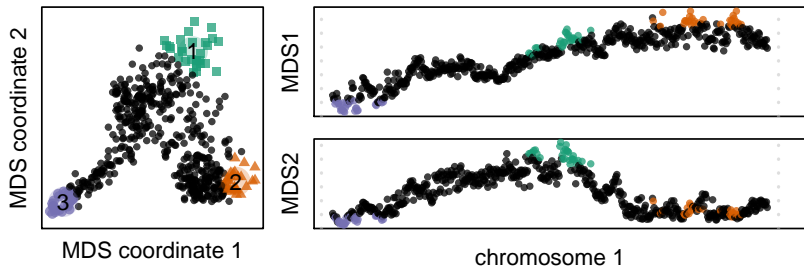
# SIMULATION: LOCAL ADAPTATION

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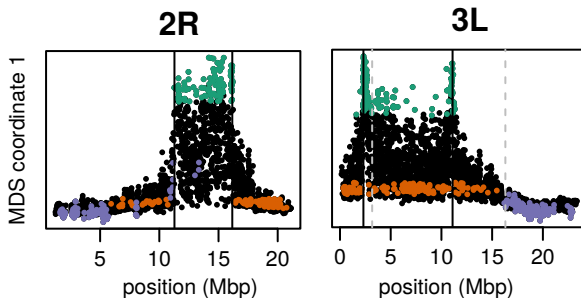- 25MB (0.625M) with 1000 evenly spaced loci with $s = \pm 0.001$,
- with simuPOP $+$ msprime

# 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)
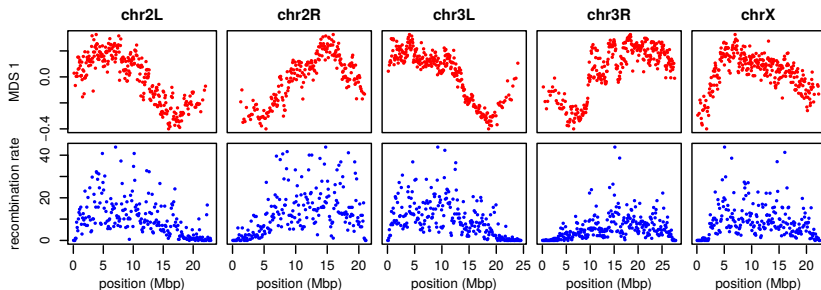- without less common inversion haplotypes: linked selection?

# 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?

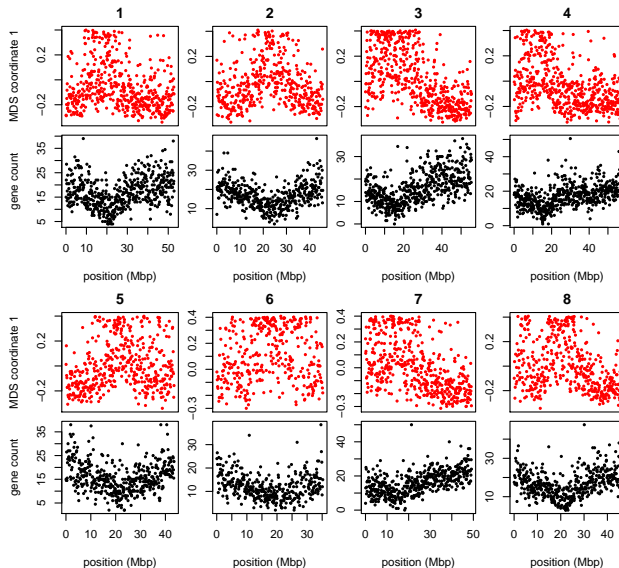# DATA: AFRICAN *D. melanogaster*

- ▶ 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)
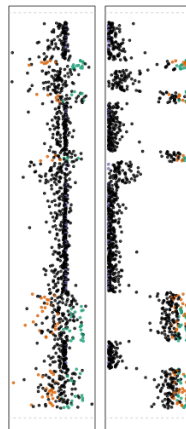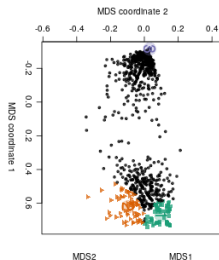- ▶ without less common inversion haplotypes: linked selection?
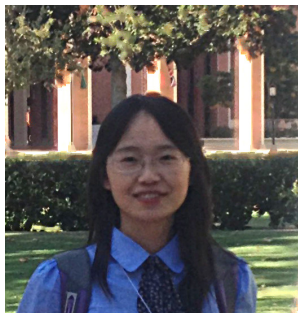
▶ 263 pan-Mediterranean samples – WGS – 100 Kb windows

# CONCLUSIONS



- there isn't always a single "population structure"
- (more) evidence for widespread effects of linked selection
- `lostruct` is a visualization tool
- applicable to other summary strategies
- try it out: `https://github.com/petrelharp/local_pca`

# Thanks

Han Li – USC – bioRxiv:070615



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

# Thanks

Han Li – USC – bioRxiv:070615



**University of Oregon**
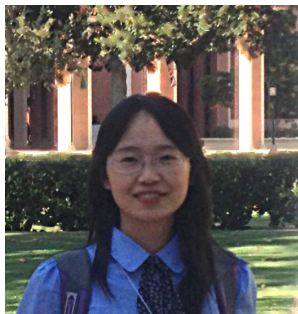


John Pool, Russ Corbett-Detig
Peter Chang, Matilde Cordeiro
Graham Coop, Jeremy Berg, Yaniv
Brandvain, Chuck Langley
Jaime Ashander, Jerome Kelleher

# 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 $\approx$ low recomb regions
- ▶ missing data?
  filter; but user beware