

# Population genomics



John Wang  
May 5, 2016  
Introduction to Next  
-Generation Sequencing (NGS) Data and Analysis



# Population genetics:

The study of polymorphism (within species) and divergence (between species). [Hartl 2003]



Aims to understand evolution

Changes in frequency of alleles within populations over space and time

- natural selection
- genetic drift
- gene flow
- mutation
- recombination

# Population genomics:



The field of population genomics surveys patterns in the genome within and among populations to make inferences about evolution and the genome. (Nosil and Buerkle 2010)

Population genomics can be broadly defined as the **simultaneous study of numerous loci or genome regions** to better understand the roles of evolutionary processes (such as mutation, random genetic drift, gene flow and natural selection) that influence variation across genomes and populations. (Luikart et al 2003)

Population genomics is the use of genome-wide sampling to identify and to **separate locus-specific effects** (such as selection, mutation, assortative mating and recombination) from **genome-wide effects** (such as drift or bottlenecks, gene flow and inbreeding) to improve our understanding of microevolution. (Black et al 2001)

# Population genomics:

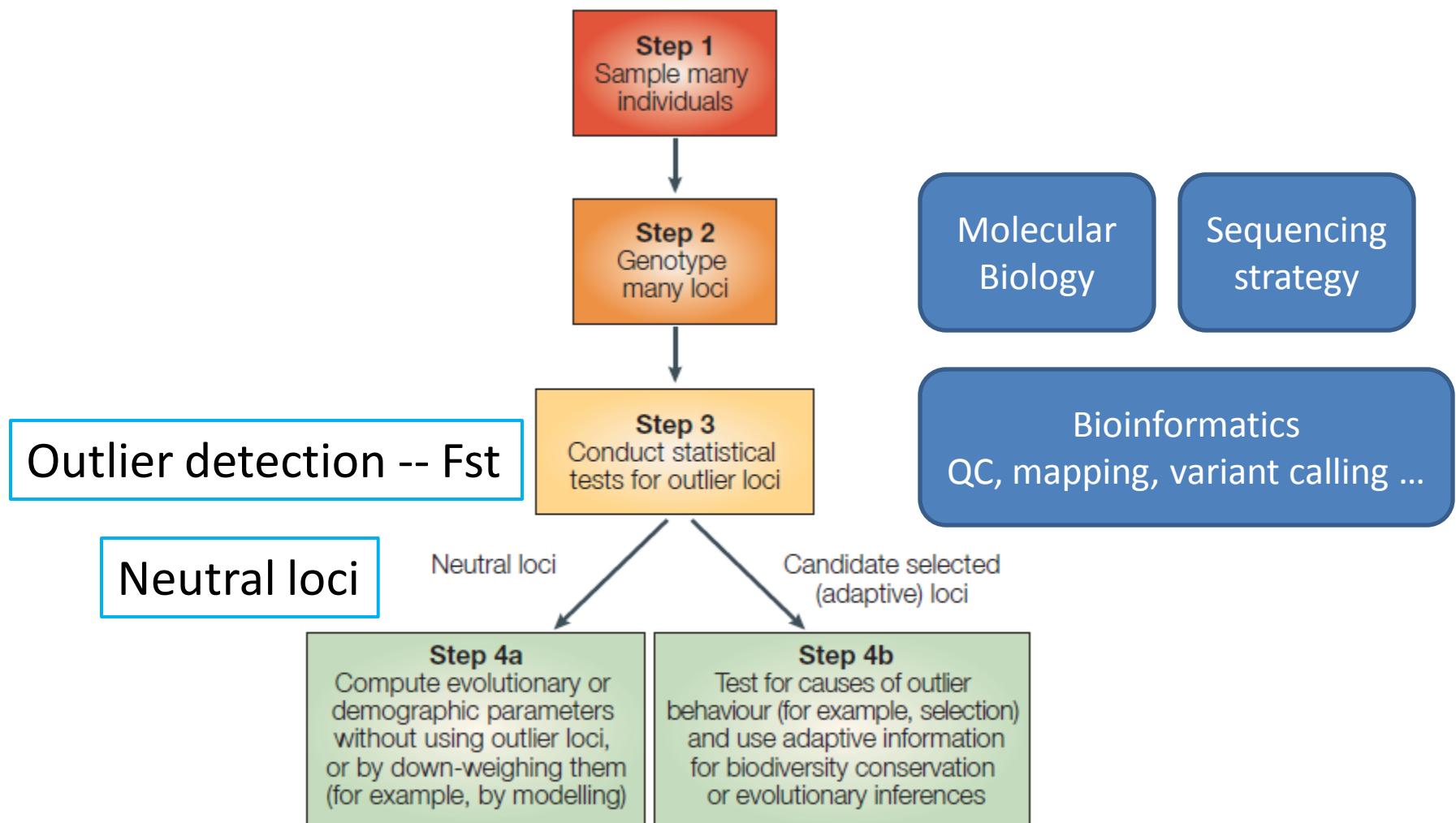
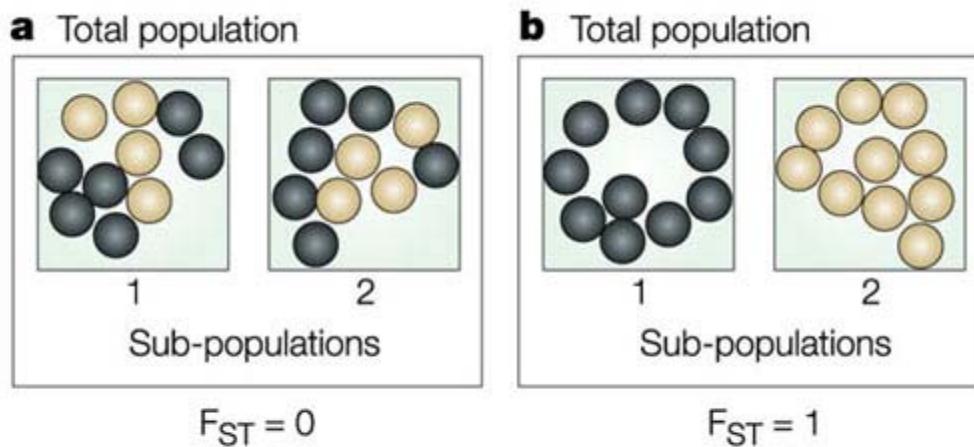


Figure 1 | Flow chart of the four main steps in the population-genomic approach. The approach summarized

# Fixation index – $F_{ST}$



Nature Reviews | Genetics

- Measure of population differentiation.
- Most widely used index of genetic divergence b/w populations.
- High  $F_{ST}$  → two populations are fixed for allelic differences.  
Drift or selection
- $F_{ST}$  scans across genome (locus-by-locus → e.g., manhattan plots)

# Some population genomics/genetics questions

What genes are under selection? [speciation genes, domestication]

Has selection been acting or can drift be sufficient explanation? [allele frequency clines]

What gene contributes to a trait? [agriculturally important genes]

What was past demographic history? [out of Africa]

Is this one big population or many small ones? [climate change]

“Just so” stories of selection are easy to invoke...

... but there might be alternative scenarios, e.g. genetic surfing

# Approaches

- Reduced representation sequencing (various)
- Low coverage sequencing → Pool-seq
- Whole genome re-sequencing (WGS)

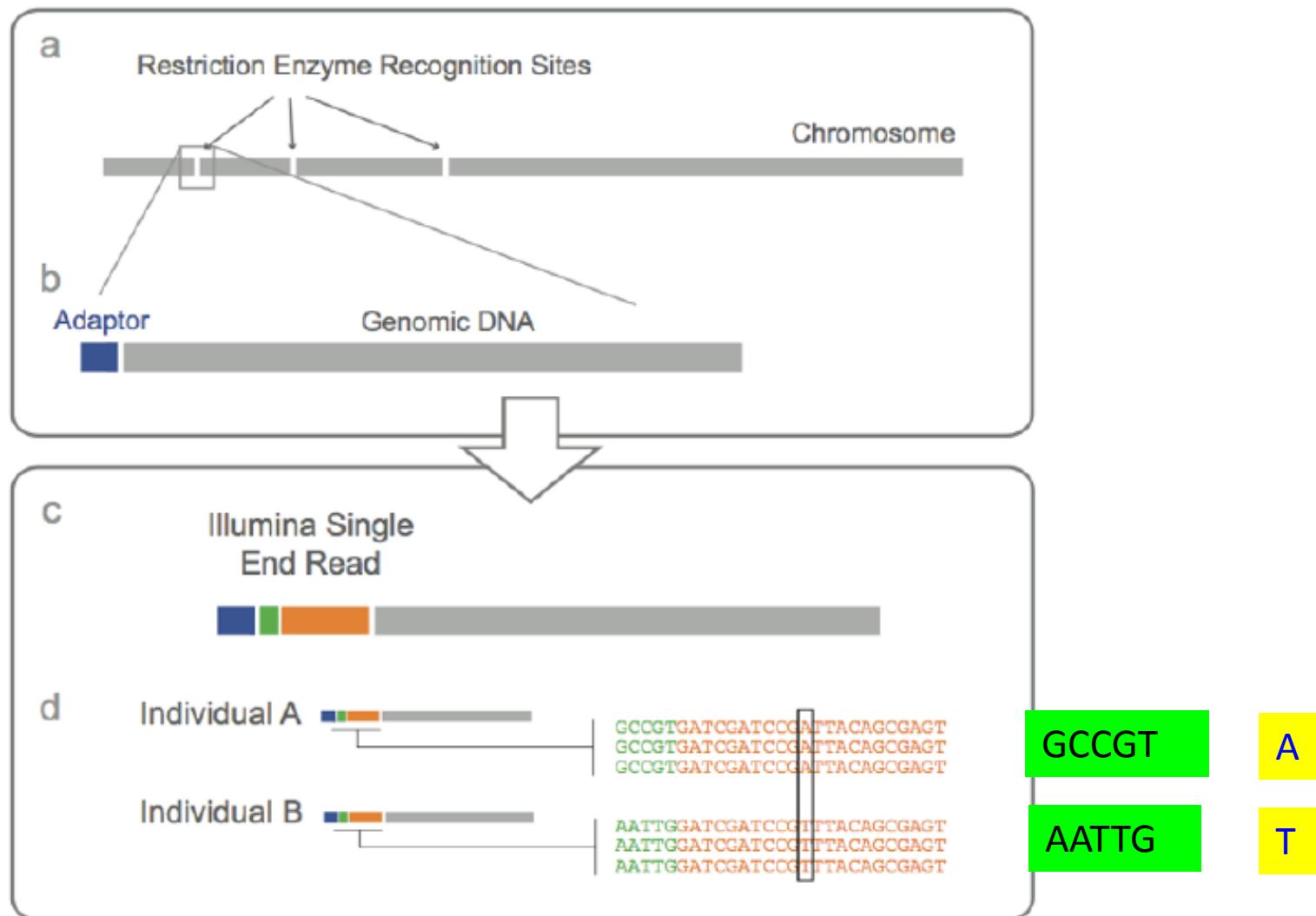
# Reduced representation sequencing approaches

- RAD-seq [genotyping-by-synthesis]
- Exon capture (targeted sequencing)
- Transcriptome (RNA-seq)
- Restriction enzyme size selection

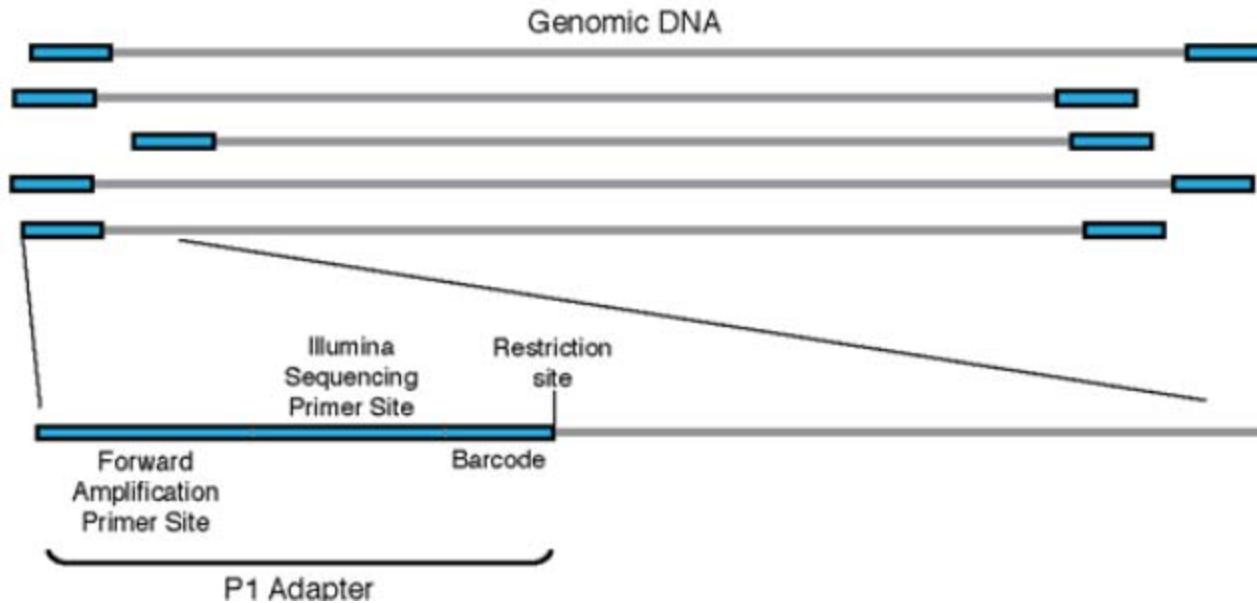
# Genotyping-by-sequencing

e.g., RADseq

# Restriction site associated DNA (RAD) marker sequencing with barcoding



**A** *Ligate P1 Adapter to digested genomic DNA*

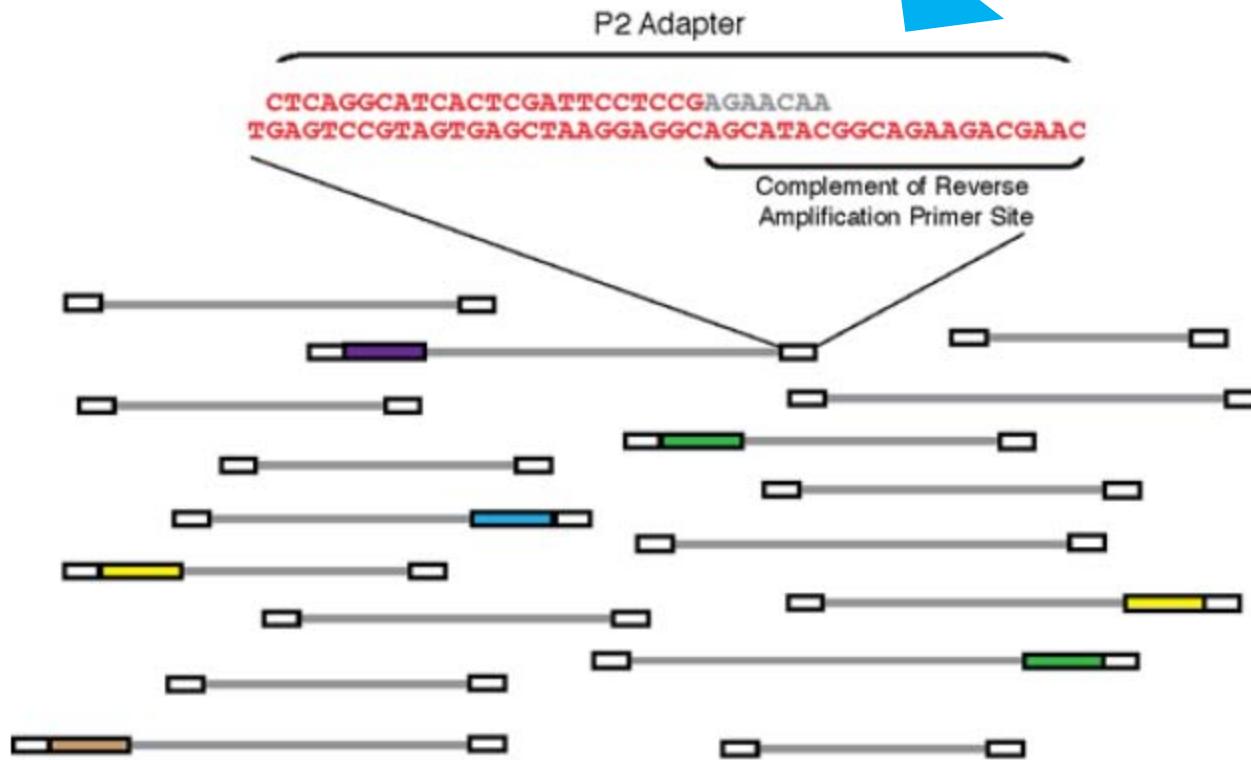


**B** *Pool barcoded samples and shear*

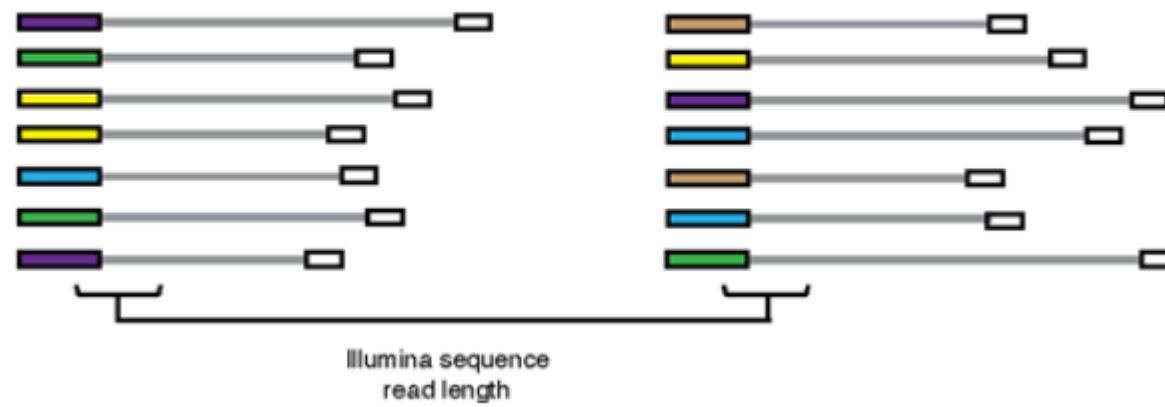


This is really clever:  
Only PCR's what you want

C *Ligate P2 Adapter to sheared fragments*



**D** *Selectively amplify RAD tags*



# RAD markers sequence a subset of the genome



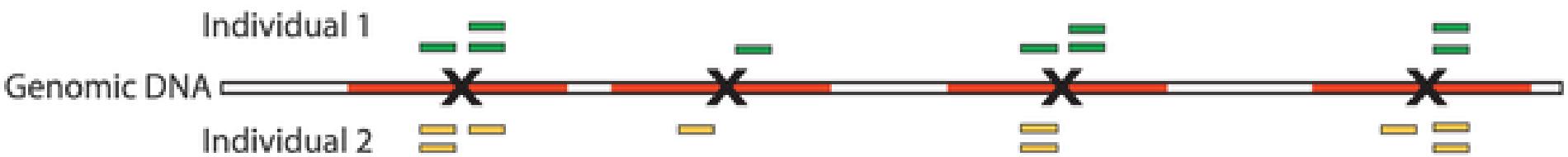
# There are several variants of RAD-seq

ddRAD  
ezRAD  
2bRAD  
...

A

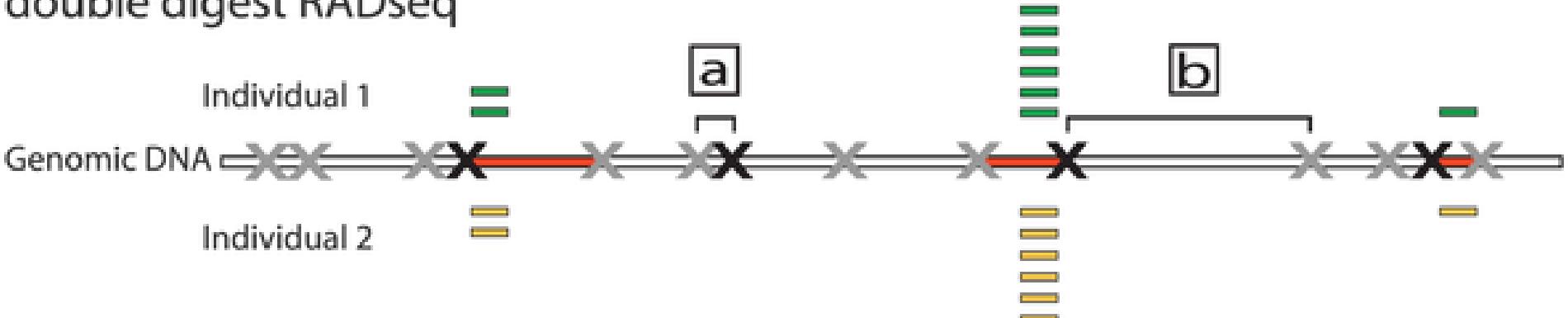
## RAD sequencing

X Rare cut site      — Genomic interval present in library  
X Common cut site      — Sequence reads



B

## double digest RADseq



# ddRAD-seq

Easier tunability; better reproducibility

But PCR duplicates unknowable

Fraction of genome

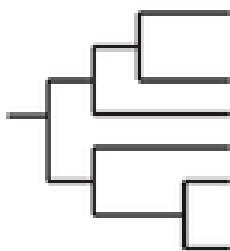
Sanger  
sequencing

Whole genome  
re-sequencing

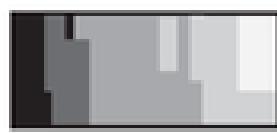
RADtag (Baird 2008)

ddRAD

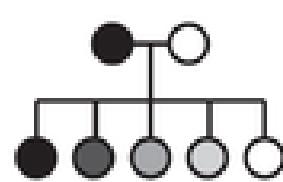
Phylogeny



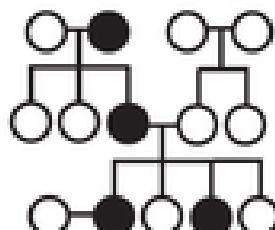
Population  
Structure



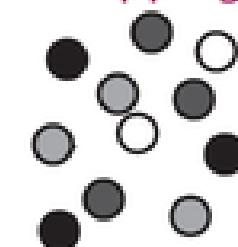
QTL  
Mapping



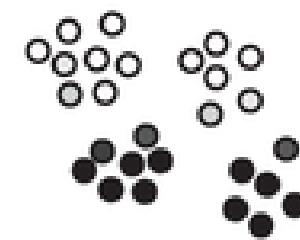
Pedigree  
Mapping



Association  
Mapping



Population  
Genomic Scans



Divergence limited

Recombination limited

Linkage Diseq. limited

# Our RADseq example



Monogyne

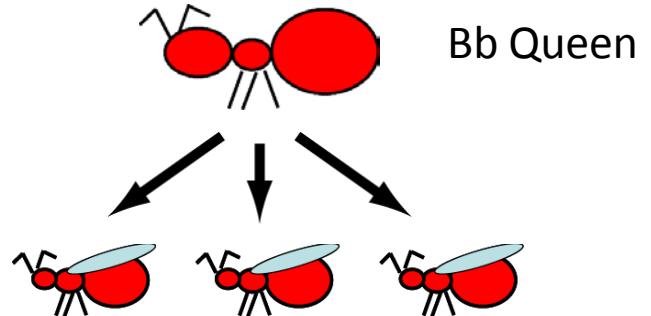


Polygyne

Fire ant *Solenopsis invicta*

Goals:

- 1) Build genetic map
- 2) test for the presence of a supergene

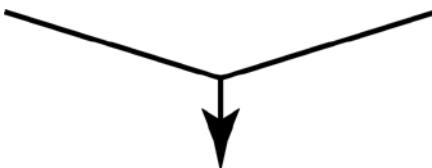


88 males

192,236 RAD tags in genome  
(96,118 EcoRI sites)

Isolate DNA, barcode each  
Sequence RAD tags with Illumina

Filter

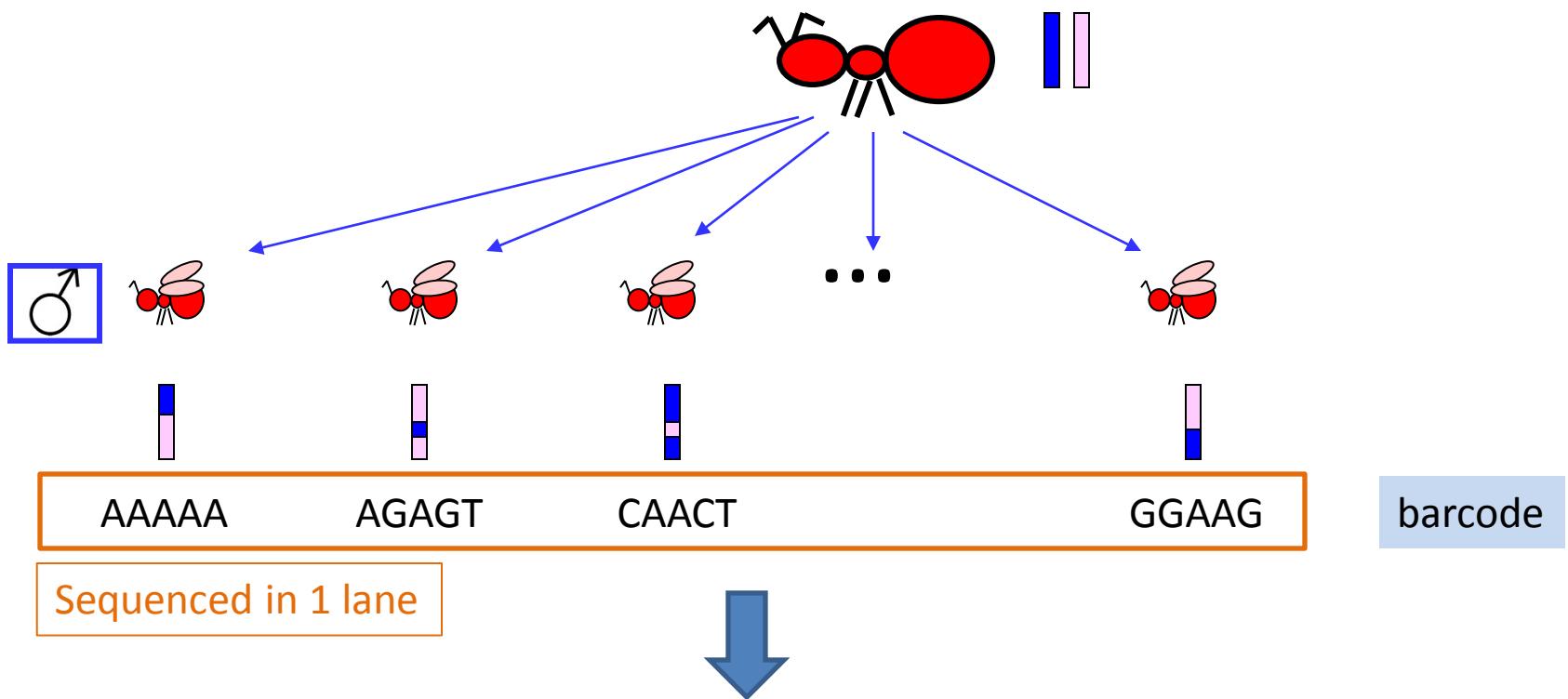


12,684 RAD tags with multiple alleles

4,232 w/ unambiguous genotype info

2,724 w/ <25% missing data

16 linkage groups  
(+1 locus outlier)



- 1) Split by barcode
- 2) Trim off barcode
- 3) Quality trimmer
- 4) Chop to 50 bp
- 5) Merge identical
- 6) Make fake FASTQ file
- 7) Run MAQ or bwa
  
- 8) Insert data into MySQL database (persistence)
- 9) New automatic scripts for filtering and QC of SNPs
- 10) MSTMap to determine genetic linkage

# Stacks

Stacks is a software pipeline for building loci from short-read sequences, such as those generated on the Illumina platform. Stacks was developed to work with restriction enzyme-based data, such as RAD-seq, for the purpose of building genetic maps and conducting population genomics and phylogeography.



## Download Stacks

Version 1.19

[Recent Changes \[updated April 23, 2014\]](#)

## Stacks Pipeline

---



### Genetic Maps

Stacks can be used to generate mappable markers from RAD-seq data. Thousands of markers can be generated from a single generation, F1 map as well as markers for traditional F2 and backcross designs. Stacks can export data to JoinMap, OneMap, or R/qtl. These data can be used for examining

### Getting started with Stacks

[Stacks Manual](#)

### Frequently Asked Questions

# Fire ant genetic map v1



**LG1**

**LG2**

**LG3**

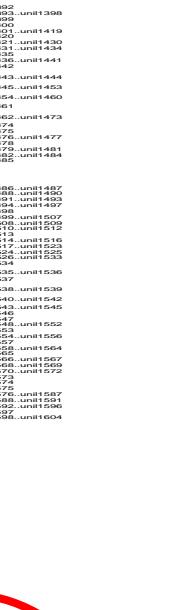
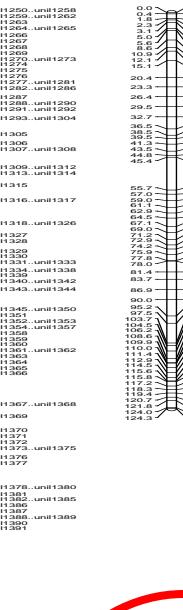
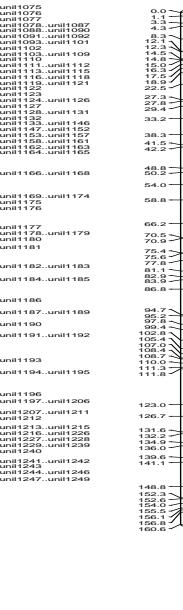
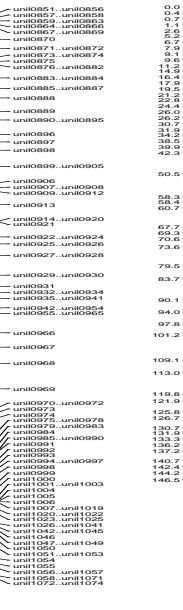
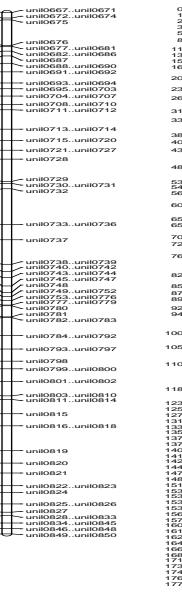
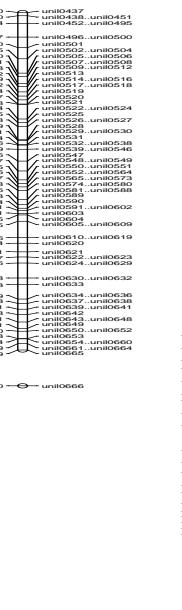
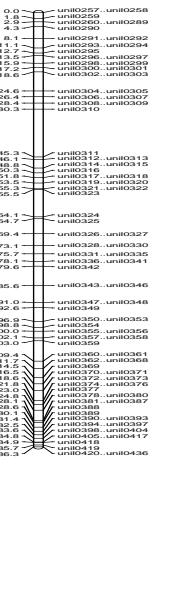
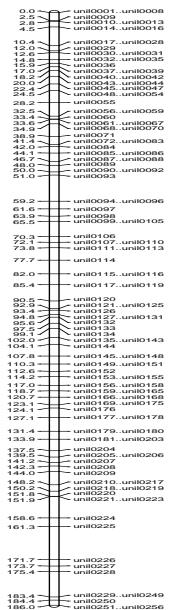
**LG4**

**LG5**

**LG6**

**LG7**

**LG8**



**LG9**

**LG10**

**LG11**

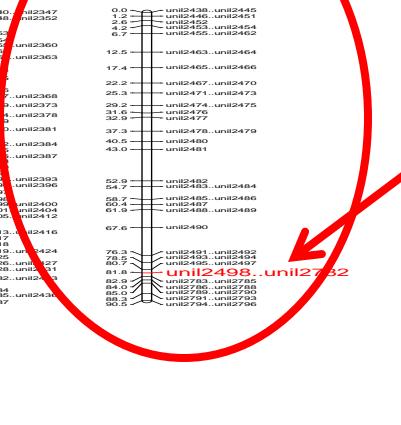
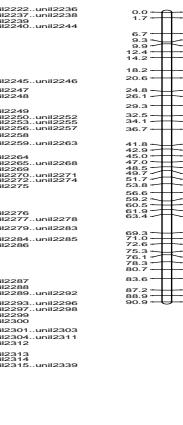
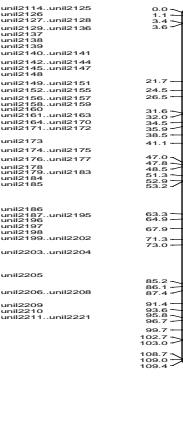
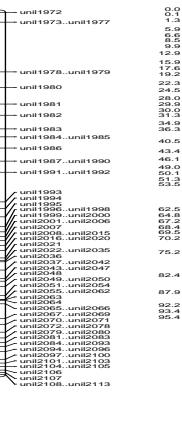
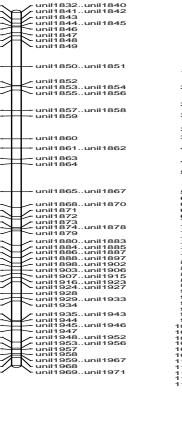
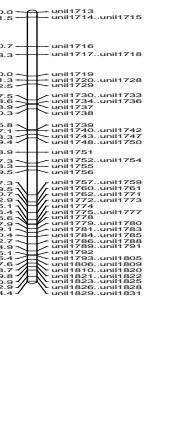
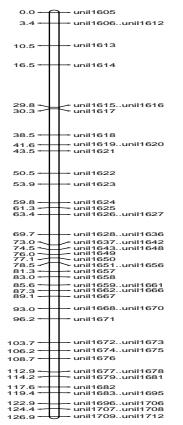
**LG12**

**LG13**

**LG14**

**LG15**

**LG 16 or S**



**Gp-9**

# Large non-combining region around *Gp-9*



LG 16 = SB or Sb

LG 16 (~23 Mbp)

0.0	unil0996..unil1005
1.6	unil1006..unil1010
3.6	unil1011
5.9	unil1012..unil1014
7.9	unil1015..unil1021
13.9	unil1022
17.2	unil1023..unil1025
21.0	unil1026..unil1028
23.5	unil1029..unil1032
27.6	unil1033..unil1034
31.0	unil1035
32.7	unil1036
34.7	unil1037..unil1038
37.7	unil1039..unil1040
39.8	unil1041
41.2	unil1042
49.6	unil1043
54.3	unil1044..unil1045
57.6	unil1046..unil1047
58.8	unil1048
60.2	unil1049..unil1050
63.5	unil1051
66.3	unil1052
74.2	unil1053
76.6	unil1054..unil1055
78.1	unil1056
78.3	unil1057
79.6	unil1058..unil1059
80.9	unil1060
81.1	unil1061
82.3	unil1062..unil1064
83.5	unil1065..unil1066
84.7	unil1067..unil1069
85.9	unil1070..unil1073
86.0	unil1074
86.2	unil1075
86.7	unil1076..unil1340
89.0	unil1341..unil1343
91.3	unil1344..unil1346

Genetic map

Physical map



Gp-9

- 265 markers non-recombinig
- 44 scaffolds
- ~12.7 Mbp

## 2<sup>nd</sup> ant also has supergene



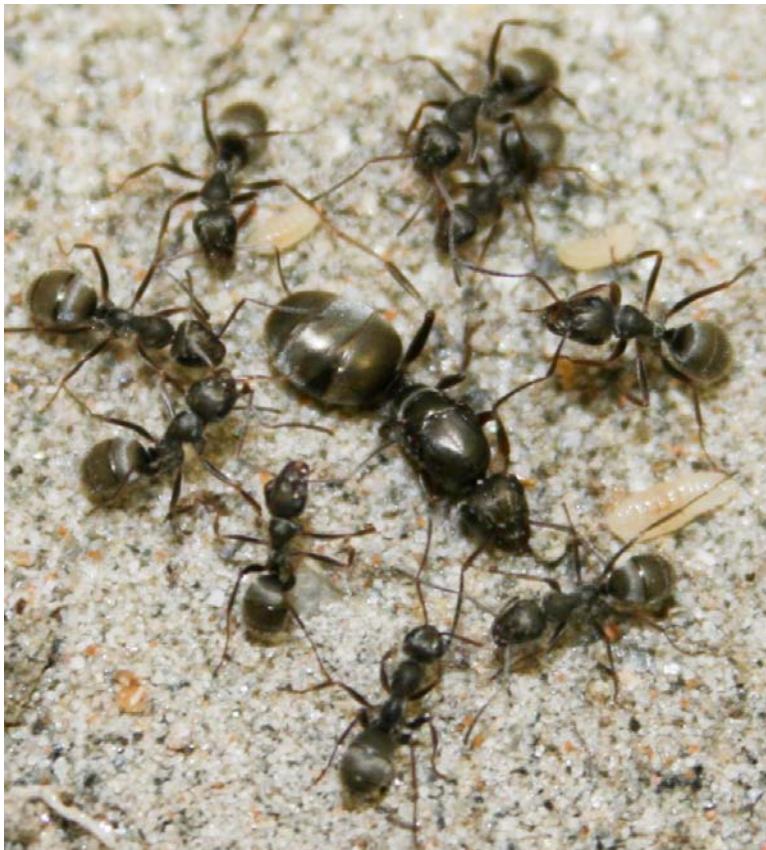
Fire ant case just RAD-seq  
but not really population genomics

*Formica selysi*

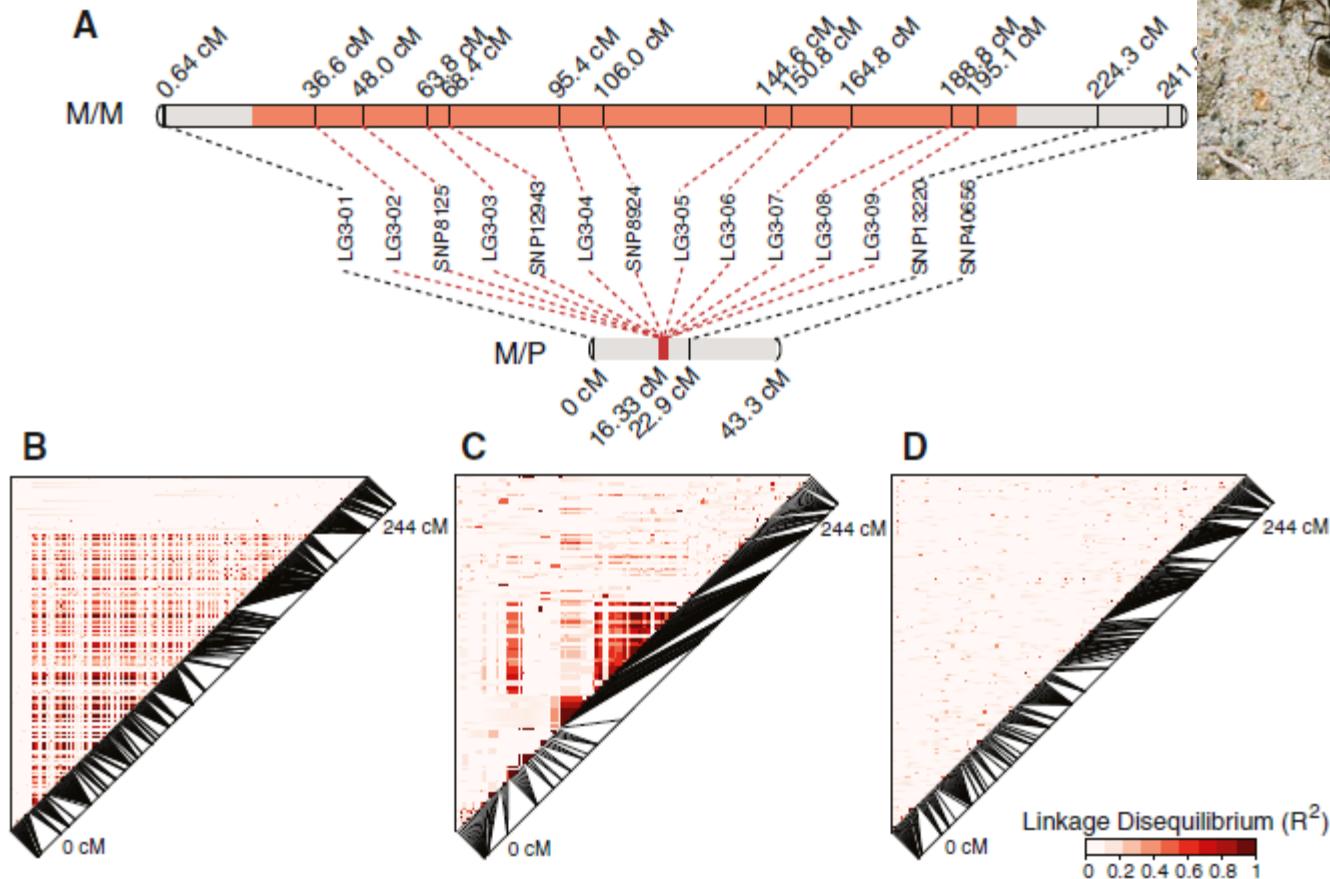
Also monogyne and polygyne social forms

Purcell et al: GBS  
Linkage map

Then population LD verified supergene



# Linkage disequilibrium → supergene



# RAD-seq in phylogeography

*Wyeomyia smithii* mosquito

larvae live in

*Sarracenia purpurea* carnivorous  
pitcher plant

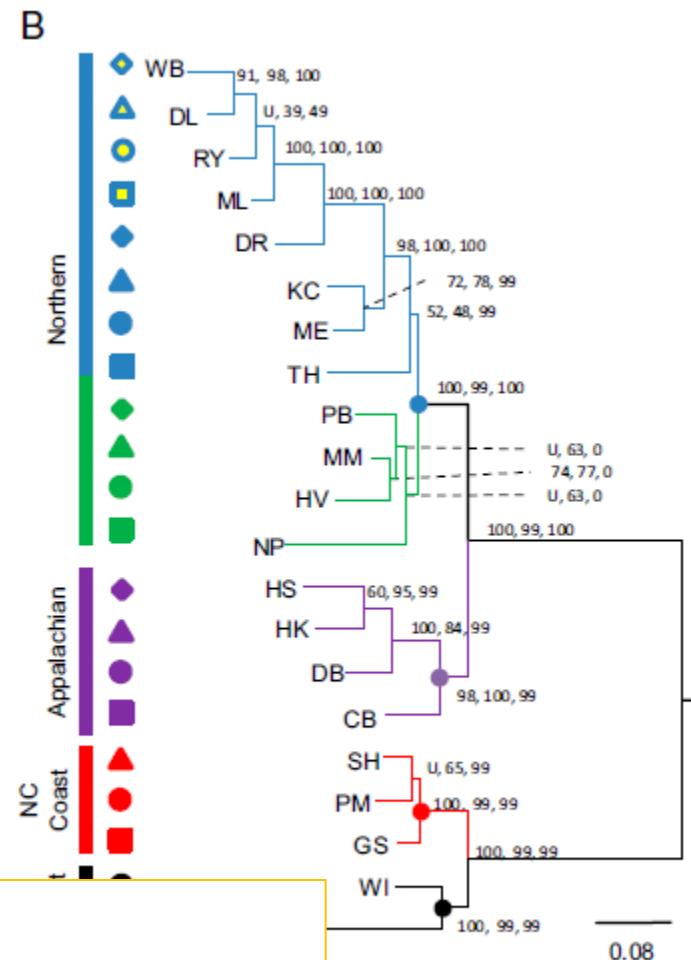
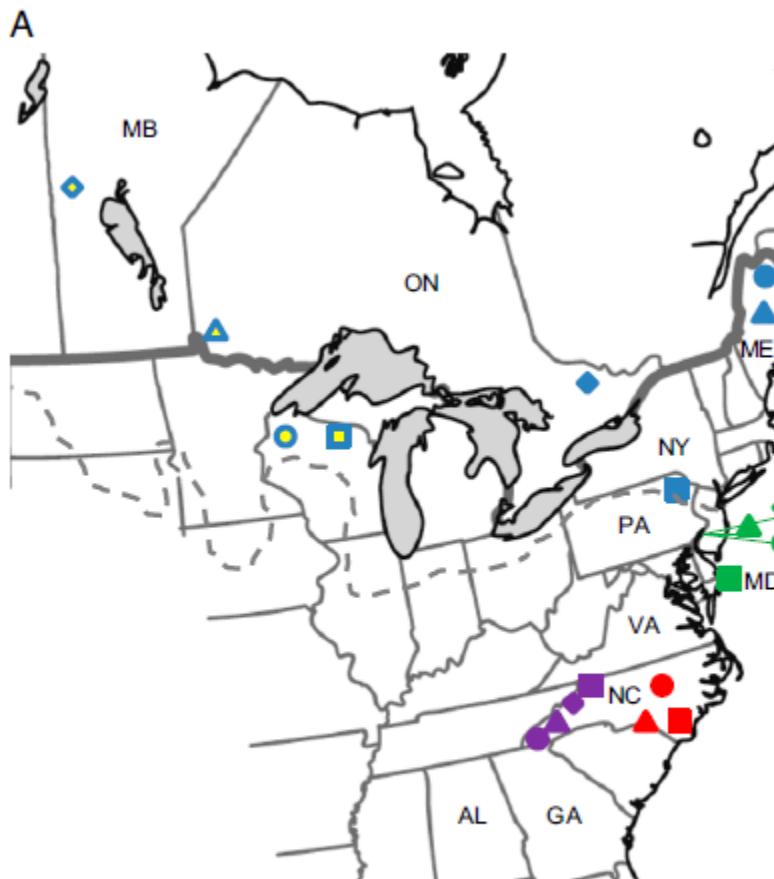


photo: Leonora bug level

What is their population structure in the Eastern USA?

Is there an expansion pattern associated with glacial retreat?

# *W. smithii* samples and phylogenetic tree from RAD



Fine scale genetic divergence

Evidence for glacial retreat associated expansion

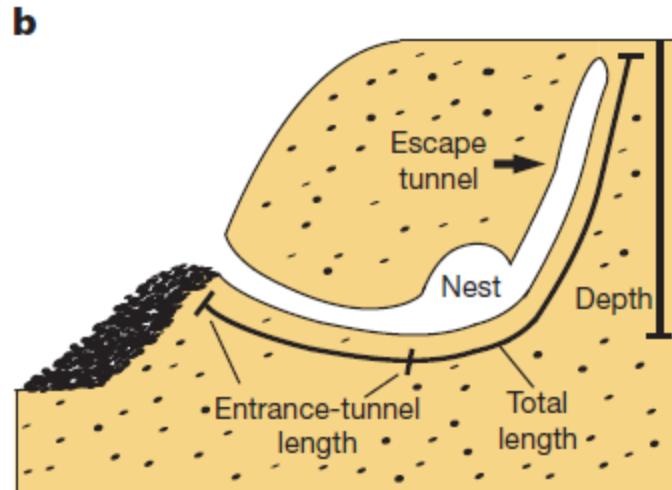
Future – local photoperiodic response adaptations

# Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice

Jesse N. Weber<sup>1†</sup>, Brant K. Peterson<sup>1,2</sup> & Hopi E. Hoekstra<sup>1,2</sup>



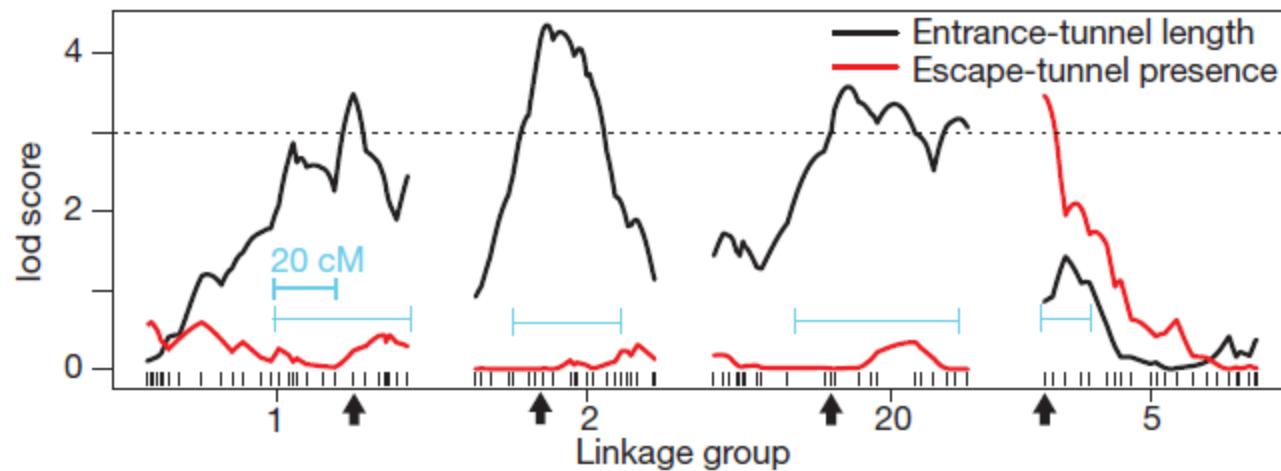
Vera Domingues/Hopi Hoekstra



# Tunnel building at least 4 loci

Crossed 2 species of mice:  
oldfield mice (escape tunnels) x deer mice (simple nests)

QTL mapping with ddRAD-seq



# Exon sequencing

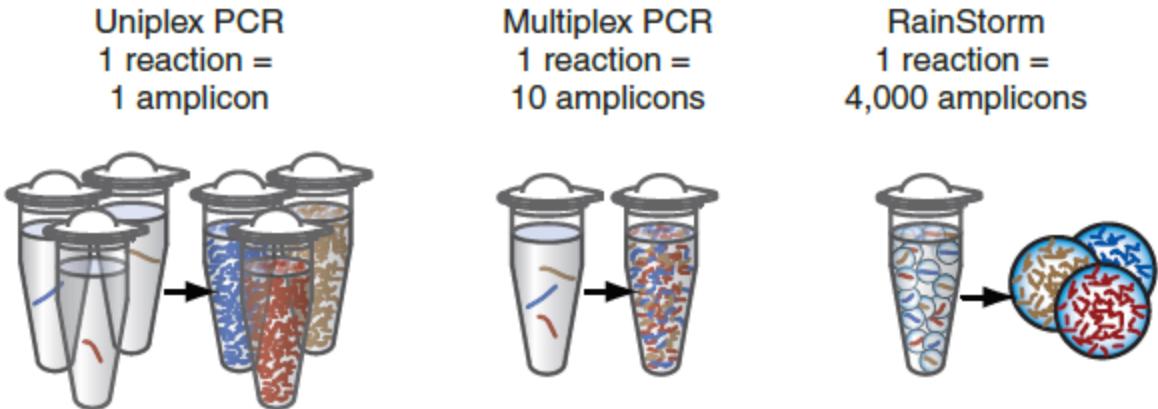
# Targeted (re)sequencing or **exon capture**

- Target for specific regions (e.g., exons)
- Subset of the genome can answer your question
- Reduce effort and cost
  - Reduced data storage (sometimes never analyzed)
- Increase sample size
- Something of a stop gap until WGS is too cheap
  - Genome too big

# Exon capture

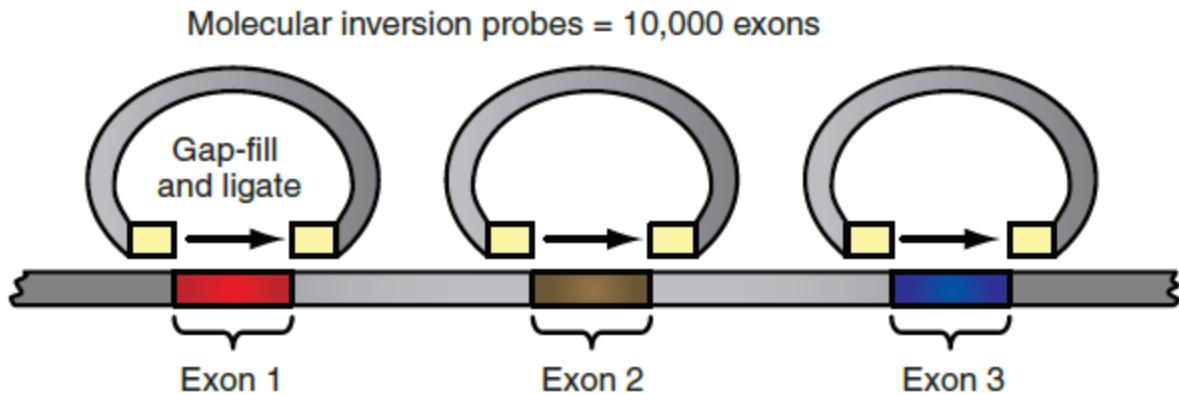
a

1,000's



b

MIP  
10,000's



# Exon capture

C

Hybrid capture > 100,000 exons

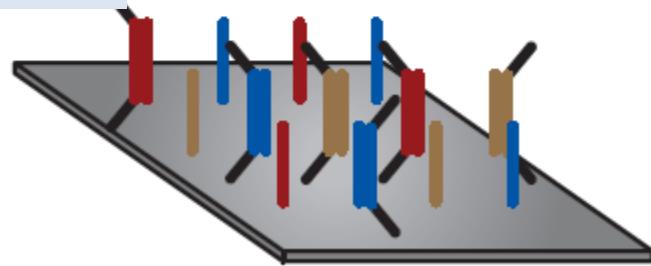


Adapter-modified shotgun library

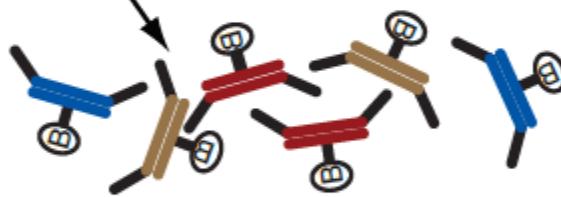
100,000's

Agilent  
Nimblegen  
Illumina

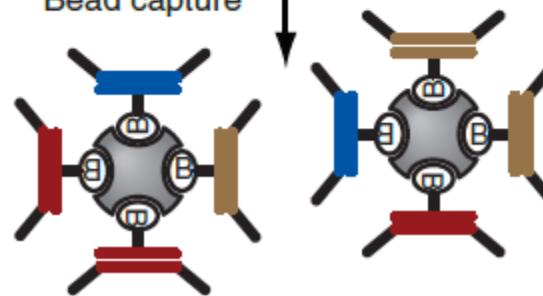
Array capture



Solution hybridization



Bead capture



# Exon capture – applications

- Mendelian disease discovery

Since many are protein coding mutations

Strategies:

Family based

Extreme phenotypes (e.g., cardiovascular health)

- Non-human organisms

Anything you can imagine with many (protein coding) loci

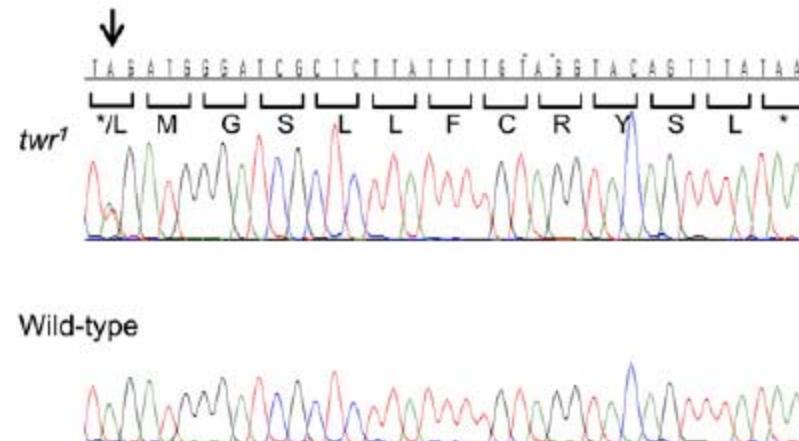
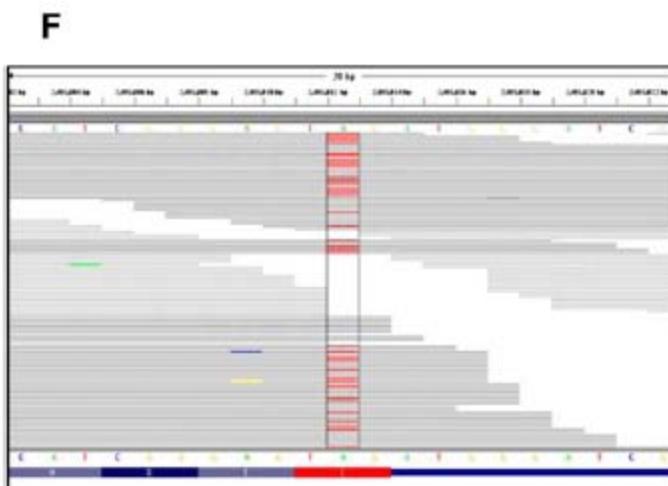
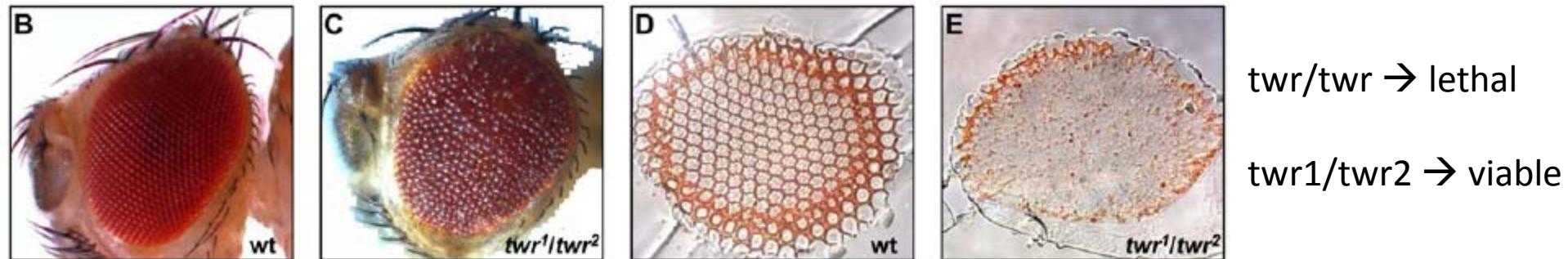
Cloning homozygous lethals (via heterozygotes)

Climate change driven subdivision

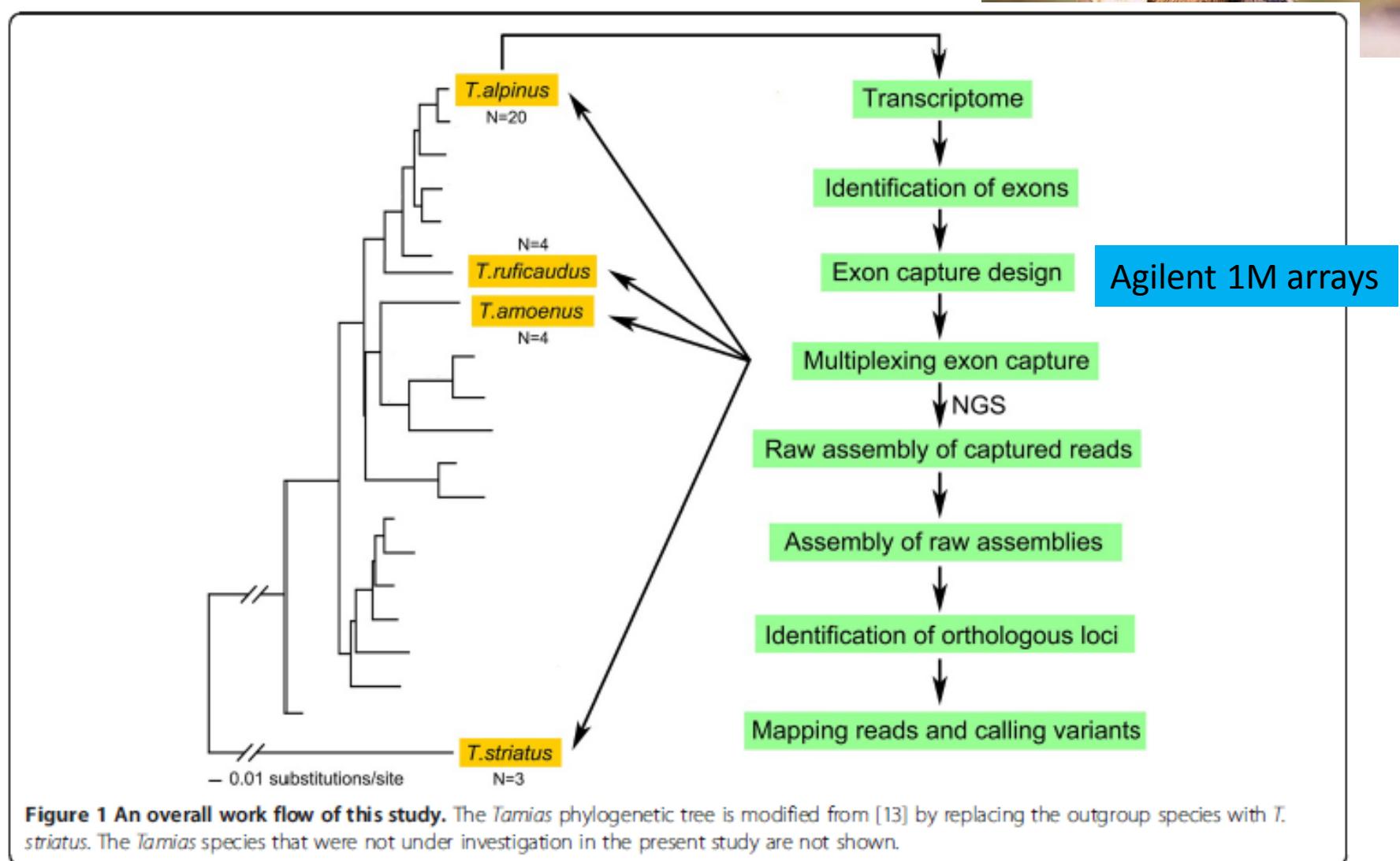
- Caveat: nontrivial (although not huge) cost of setup

# Exon capture – homozygous lethal case study find in heterozygote

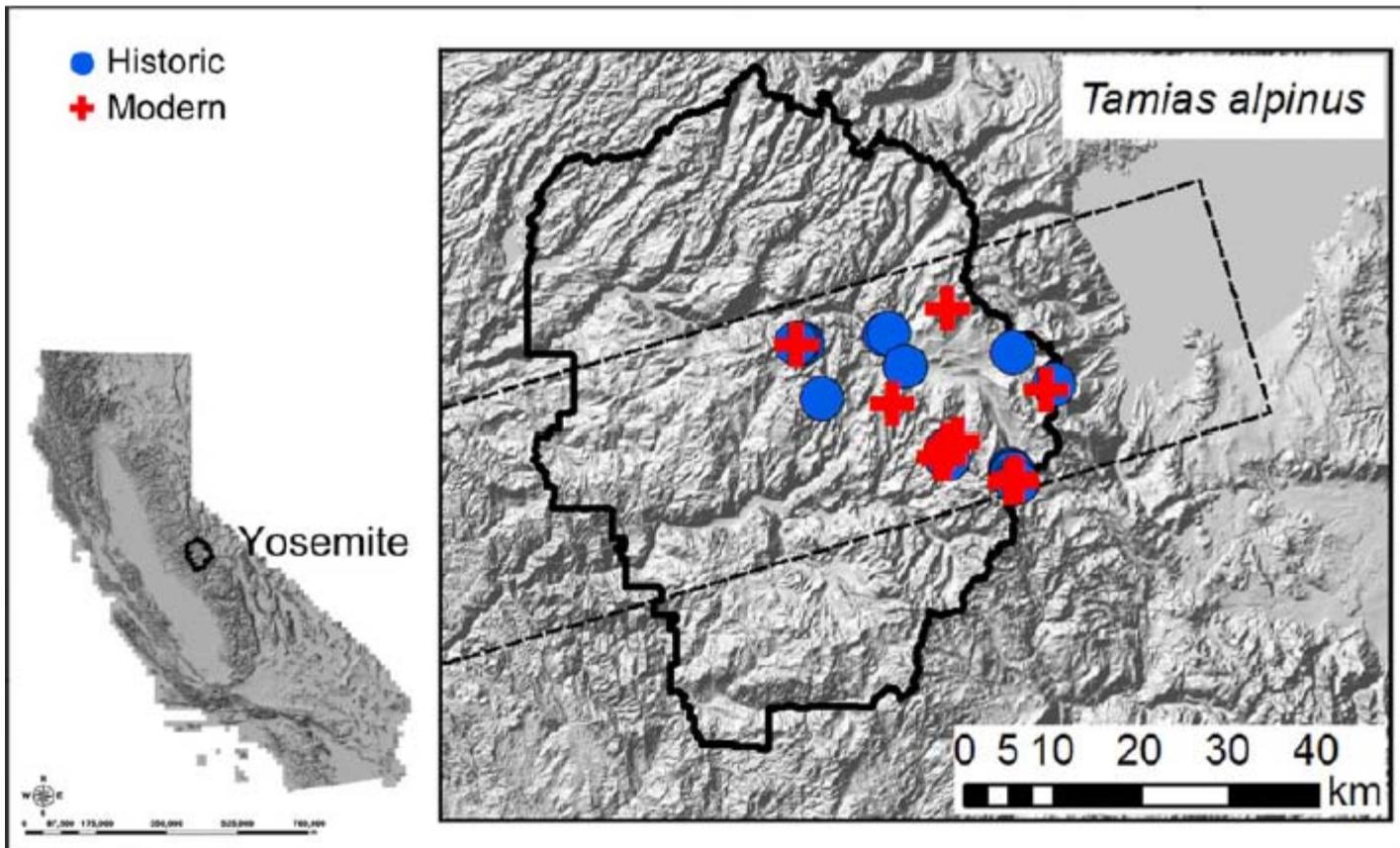
*twisted bristles roughened eye (twr) mutant*



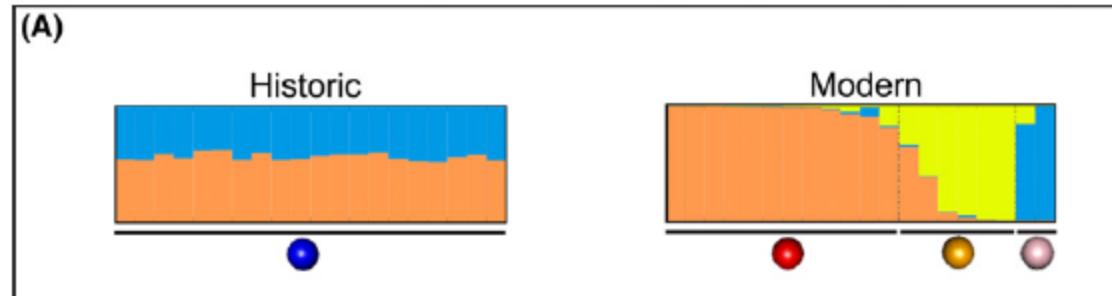
# Exon capture – chipmunk with no reference genome



# Exon capture – chipmunk → museum samples and range contraction



# Exon capture – chipmunk → museum samples and range contraction



STRUCTURE analysis reveals more modern subdivision  
Consistent with reduced gene flow associated with  
climate change

# RE delimited sequencing

# A reduced representation approach to population genetic analyses and applications to human evolution

Francesca Luca,<sup>1</sup> Richard R. Hudson,<sup>1,2,3</sup> David B. Witonsky,<sup>1</sup> and Anna Di Rienzo<sup>1,3</sup>

Reduced representation simply by digestion (Van Tassell et al 2008)

Population genomics goals:

Test serial founder model of human global dispersal

Estimate timing of out of Africa

First estimate of the colonization of Australia

Nucleotide diversity

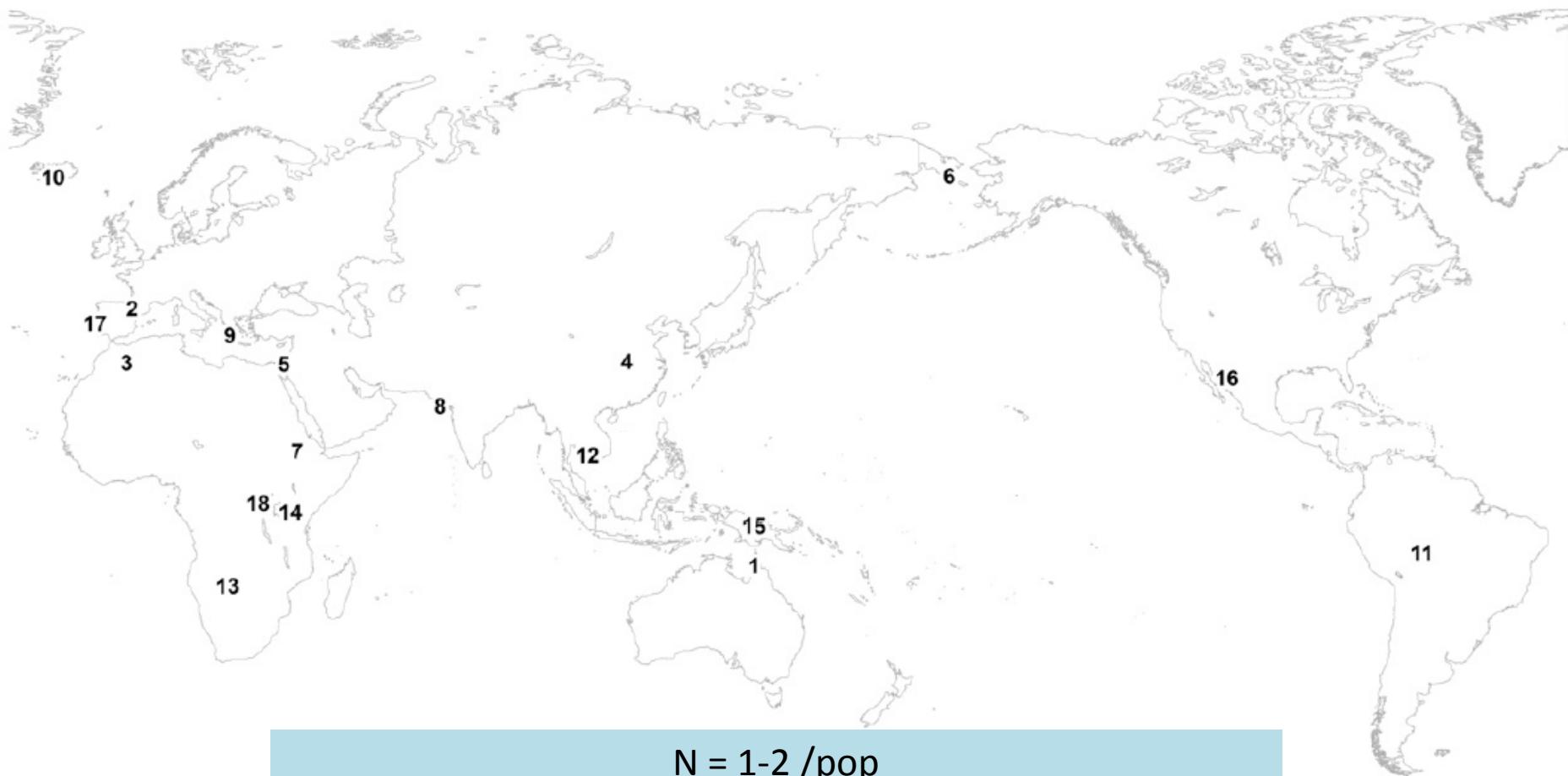
Heterozygosity

Allele frequency spectrum

...

# Samples: 19 individuals (18 pops) around the world

A



$$N = 1-2 / \text{pop}$$

Use power of many loci to estimate population parameters

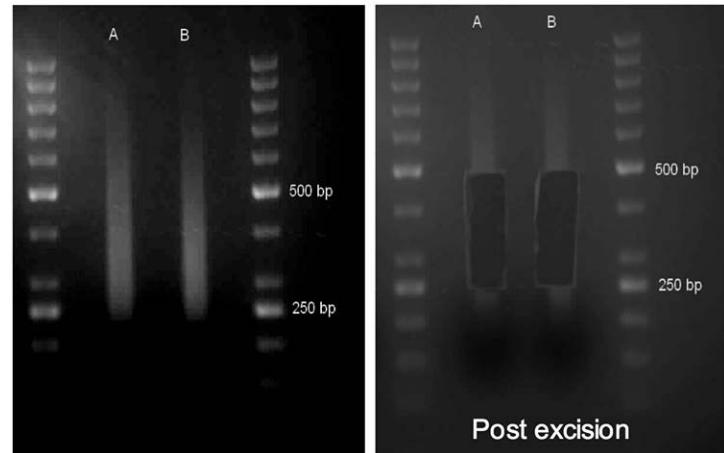
- nucleotide diversity

# Molecular protocol – relatively simple

- 1) Cut with RsaI restriction enzyme  
~9 million sites total  
less near CpG islands  
not near repeats



- 2) Select 70-75 bp fragments  
(really 50-100 bp)  
~51,000 sites in human genome



Amador, D. M., Wang, C., Holland, K. H. and Mou, Z. (2013)

- 3) Illumina single end sequencing  
36 bp (?)  
1 individual/1-2 lanes  
wanted ~50x coverage  
(got <20x)



# Bioinformatics – mapping, filtering

## Strategy:

- Map to reduced genome
- Limited to 36 bp flanking RsaI sites
- Removed redundant loci with  $\leq 4$  mismatch



Filtering: QC or clean up data

- Remove 2 nts from RE <GTAC>
- 3<sup>rd</sup> nt also biased
- keep phred score  $\geq 20$
- if >2 alleles (keep best 2)

## Caveats:

- Cannot know pcr duplicates
- Loss due to RE polymorphism

## RESULT:

- 61.7% uniquely mapping
- Even/random coverage



# Data analysis #1

This method: 1000's of sites per genome

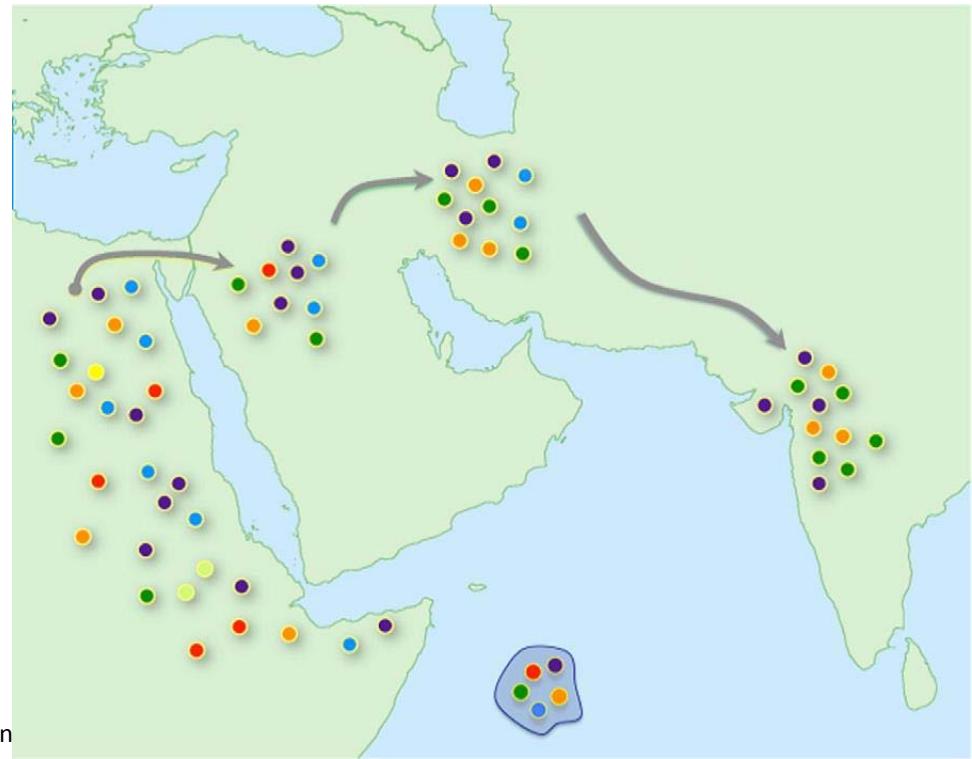
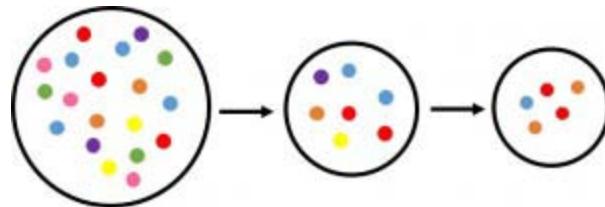
Nucleotide diversity ( $\pi$ ):

Calculate: Proportion of heterozygous sites within each genome  
is estimate of its  $\pi$  population

0.61-1.08/1000 <lower than previous estimates>

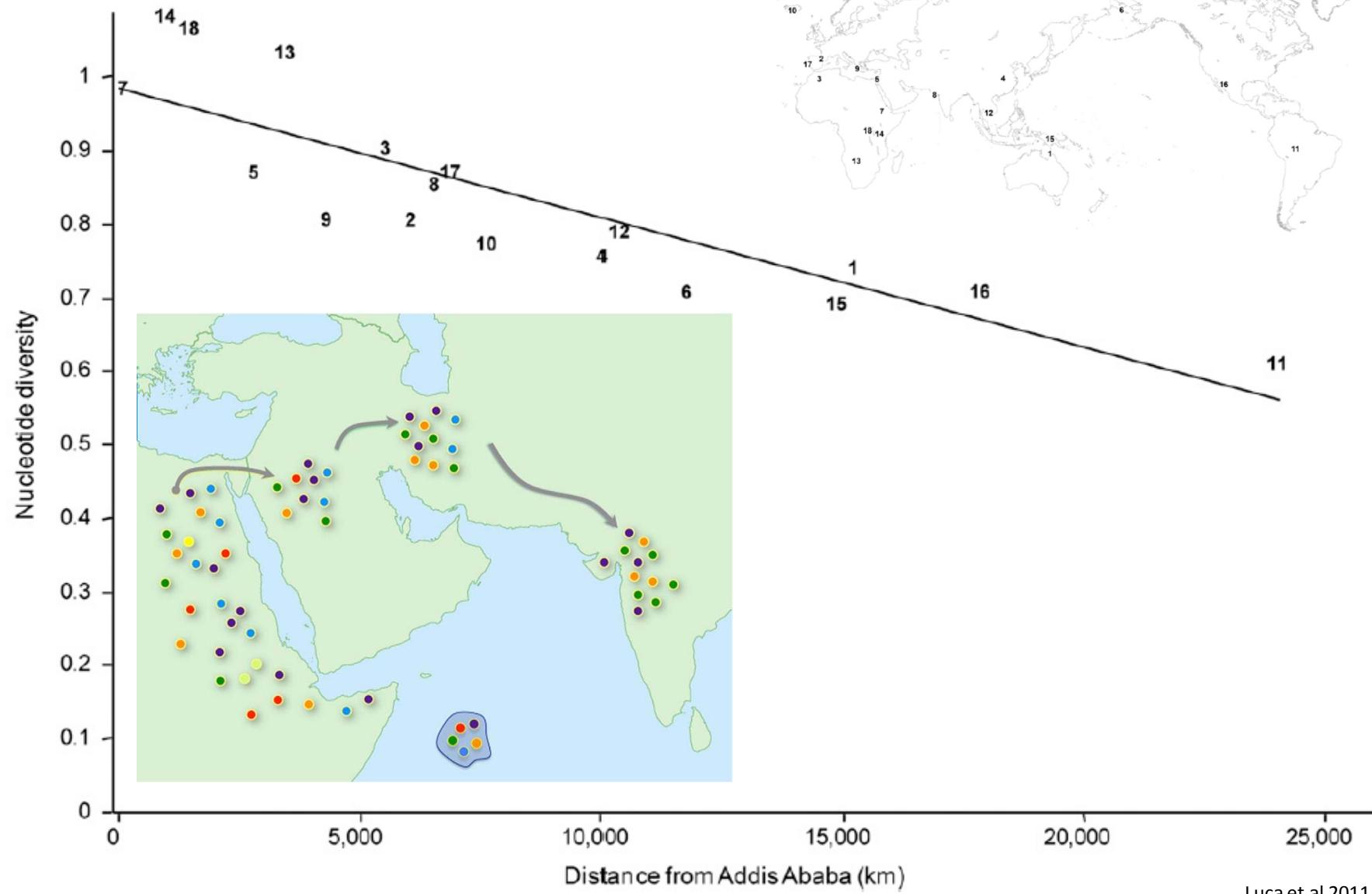
Test serial founder effect:

Expect negative correlation  
 $\pi \sim$  geographic distance



# Data consistent with serial founder effect

B



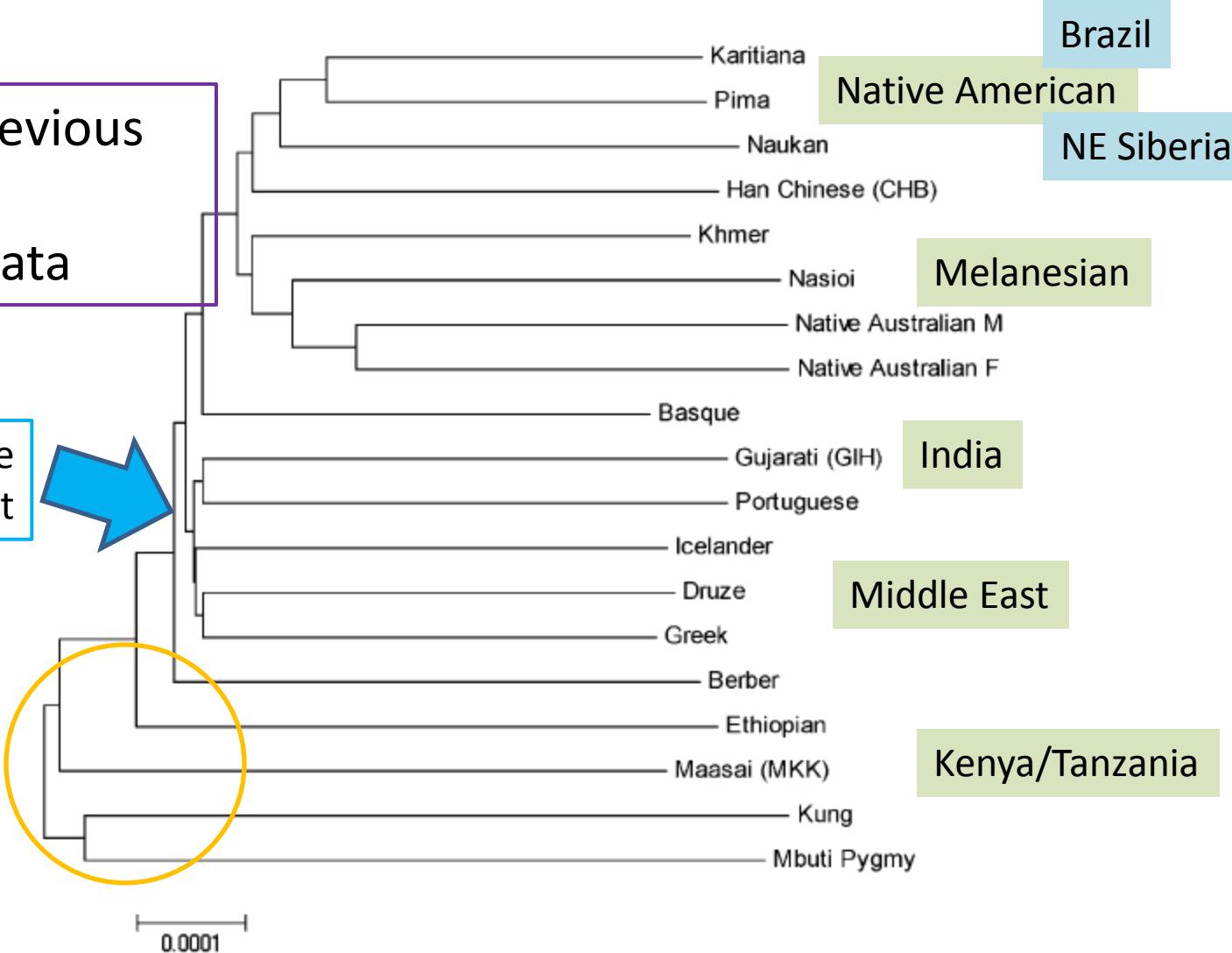
# Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)

Recapitulates previous genetic and archaeological data

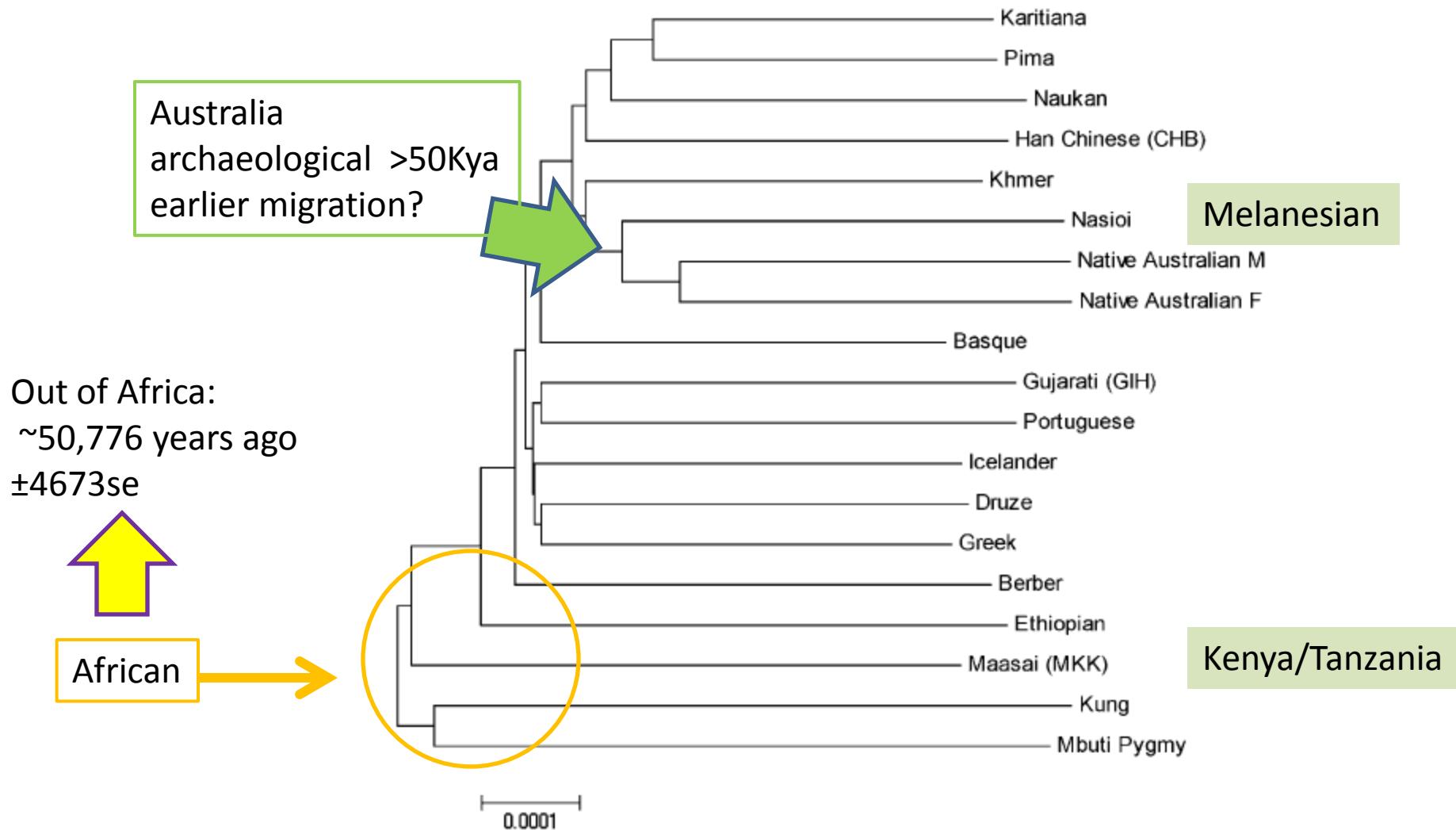
Europe  
Middle East

African



# Data analysis #2: Population divergence and split times

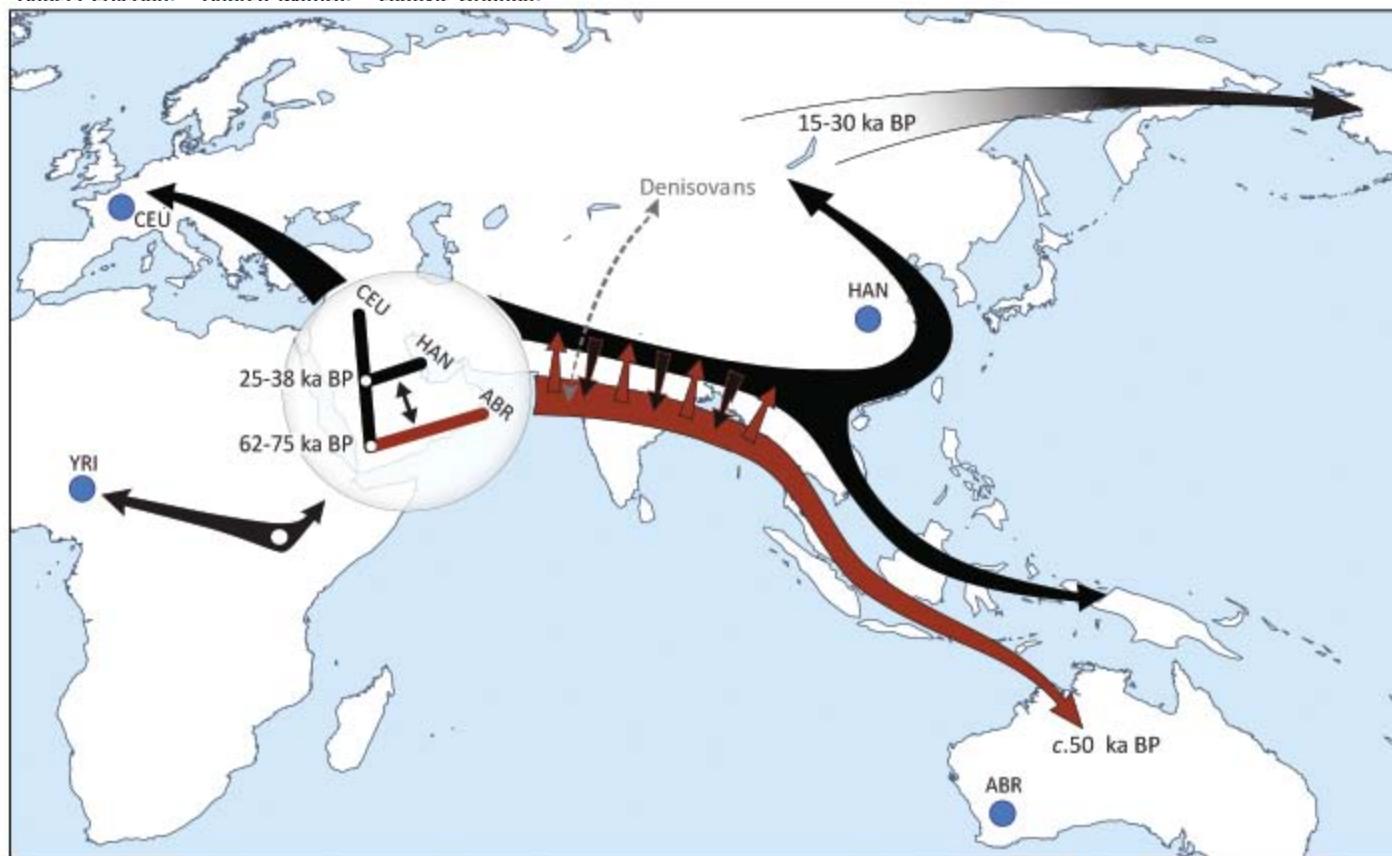
Pairwise differences then phylogenetic tree (neighbor joining)



# An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia

Morten Rasmussen,<sup>1,2\*</sup> Xiaosen Guo,<sup>2,3\*</sup> Yong Wang,<sup>4\*</sup> Kirk E. Lohmueller,<sup>4\*</sup> Simon Rasmussen,<sup>5</sup> Anders Albrechtsen,<sup>6</sup> Line Skotte,<sup>6</sup> Stinus Lindgreen,<sup>1,6</sup> Mait Metspalu,<sup>7</sup> Thibaut Jombart,<sup>8</sup> Toomas Kivisild,<sup>9</sup> Weiwei Zhai,<sup>10</sup> Anders Eriksson,<sup>11</sup> Andrea Manica,<sup>11</sup> Ludovic Orlando,<sup>1</sup>

Francisco M. De La Vega,<sup>12</sup>  
J. Víctor Moreno-Mayar,<sup>1,14</sup>  
Agata Wesolowska,<sup>5</sup> Monik Fei Xiao,<sup>3</sup> Tsunehiko Hanih Peter de Knijff,<sup>21</sup> Andrea B David M. Lambert,<sup>23</sup> Søren Michael Bunce,<sup>13</sup> Michael F Anders Krogh,<sup>1,6</sup> Robert A. Richard Villemans,<sup>7,30</sup> Rasmus



**Fig. 2.** Reconstruction of early spread of modern humans outside Africa. The tree shows the divergence of

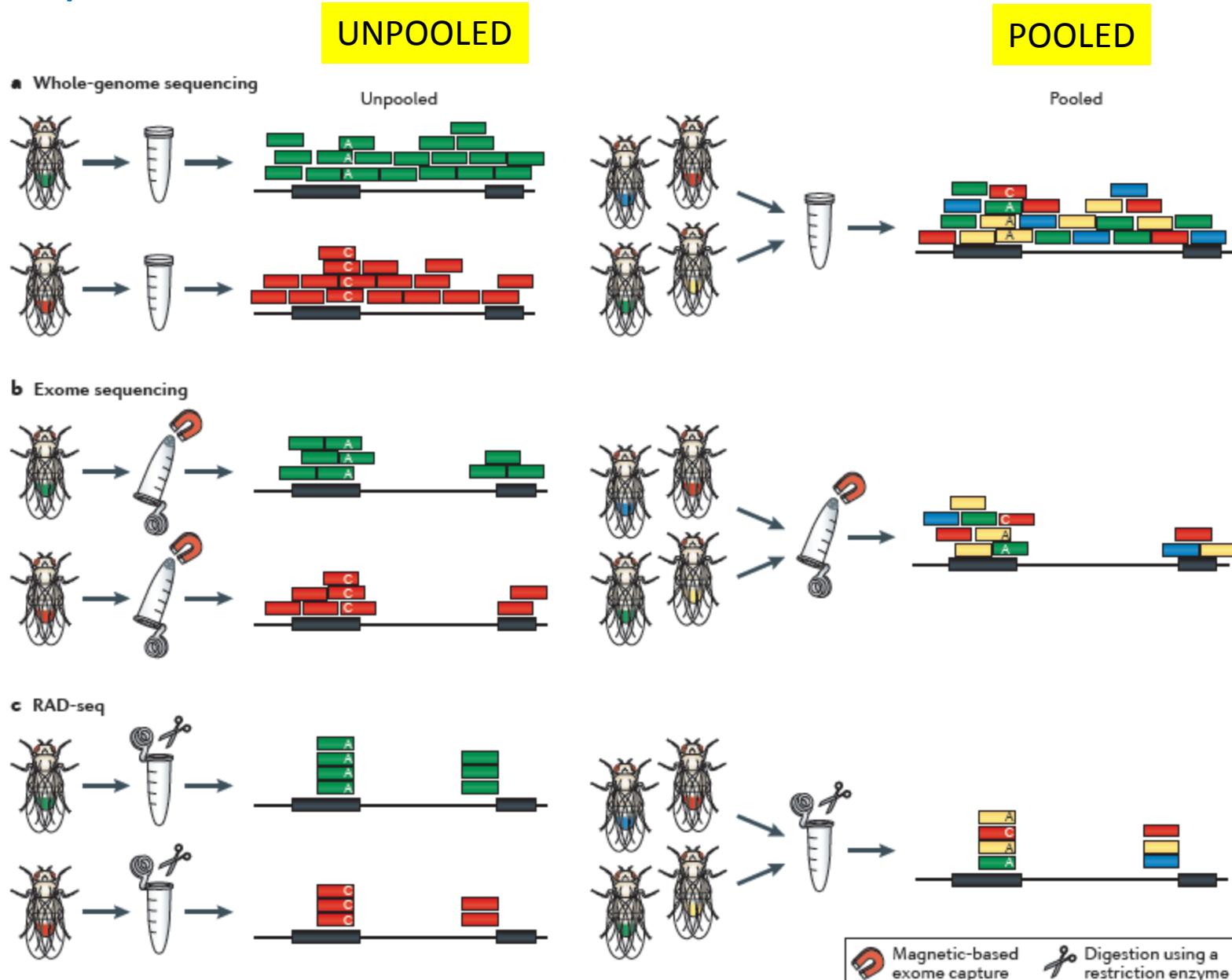
# Low coverage approaches

# PoolSeq

Eponymous:  
pool samples then sequence

often ~1x depth for each sample

# PoolSeq



# PoolSeq – applications

- Identification of functionally diverged alleles in mapping experiments
- Characterization of alleles involved in the adaptation of populations to their local environment
- Analysis of temporal allele frequency trajectories

# PoolSeq – advantages

Better estimate of allele frequencies (polymorphism data)

Individual → often redundant coverage (e.g., 10x)

Pool → not every individual, BUT basically only 1x

Protocol easier

[ No individualization ]

Cheaper

[ Barcoding/library construction entails a big cost ]

Ideally many individuals

>100

>50 okay...

Unbiased (genome wide)

Compared to RADseq or exome seq

## PoolSeq – concerns

SNP call versus Illumina error rate

Need enough depth, but assaying superficially

Loss of low frequency alleles

But such alleles are informative for many pop genetics analyses:

hitchhiking

quantification of positive and purifying selection

demographic parameters

Bad for:

Heterozygosity, CNVs, inversions and transposable elements,

No haplotype info, no linkage disequilibrium (LD) information

e.g., evidence for African lactose tolerance

Variance in pooling (Less of a problem with greater pooling)

# PoolSeq – some software

<b>Population genetics</b>		
<a href="#">PoPopulation</a>	Estimates variation within populations	39
<a href="#">PoPopulation2</a>	Estimates differentiation between multiple populations	132
<a href="#">Pool-HMM</a>	Detects selective sweeps from the allele frequency spectrum using a hidden Markov model	133
<a href="#">npstat</a>	Computes a wide range of population genetic estimators; may be used in conjunction with an external SNP caller; every contig needs to be analysed separately	134
<a href="#">Stacks</a>	Developed for population genomics with RAD-seq; may also be used with pooled RAD-seq data	135
<a href="#">Bayenv2</a>	Estimates differentiation between populations	79
<a href="#">SelEstim</a>	Detects and measures selection	136
<a href="#">KimTree</a>	Infers population histories	137
<b>Haplotype information</b>		
<a href="#">harp</a>	Estimates frequencies of known haplotypes using read counts; supports a sliding window approach	107
<a href="#">PoolHap</a>	Estimates frequencies of known haplotypes using a regression on allele frequencies	106
<a href="#">eALPS</a>	Estimates the abundance of individuals in pools given the genotypes of at least some individuals	109
<a href="#">LDx</a>	Estimates linkage disequilibrium between pairs of SNPs spanned by single-end or paired-end reads	138
<b>Forward genetic screens</b>		
<a href="#">SHOREmap</a>	Identifies causative recessive variants from a large pool of recombinants that have the recessive genotype	44
<a href="#">CloudMap</a>	A cloud-based pipeline for localizing mutations	139
<a href="#">MULTIPOOL</a>	Identifies candidate loci from bulk segregant analysis in which progeny are grouped by phenotype and sequenced as pools	140
<a href="#">Fishyskeleton</a>	Detects mutations in zebrafish from a pool of mutant F <sub>2</sub> fish	141
<a href="#">NGM</a>	A web-based tool for localizing mutations from a small pool of F <sub>2</sub> population	142
<a href="#">SNPtrack</a>	A web-based tool for localizing mutations using a hidden Markov model	143

... there are others

# PoolSeq – some applications

Induced mutation

QTL

GWAS

Evolve and re-sequence

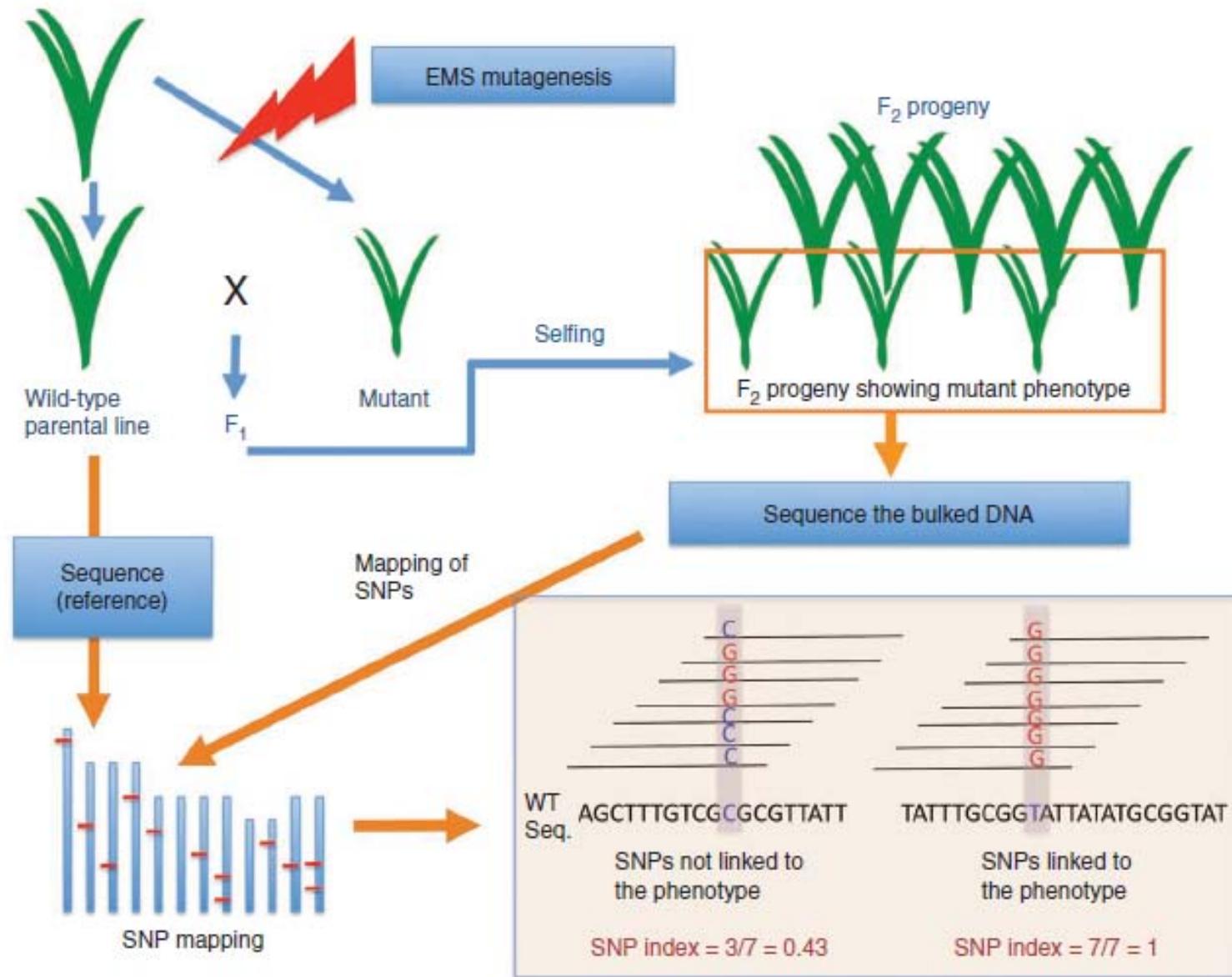
Domestication gene identification

Recombination rate analysis

...

# PoolSeq – MutMap for induced mutation

Induced mutation



## A recessive mutant trait due to a single gene mutation

Mapping mutation within same strain, so less noise

**SNP index:** The ratio of the number of reads containing a mutant SNP to the total number of reads containing the site of the SNP.

SNP index:

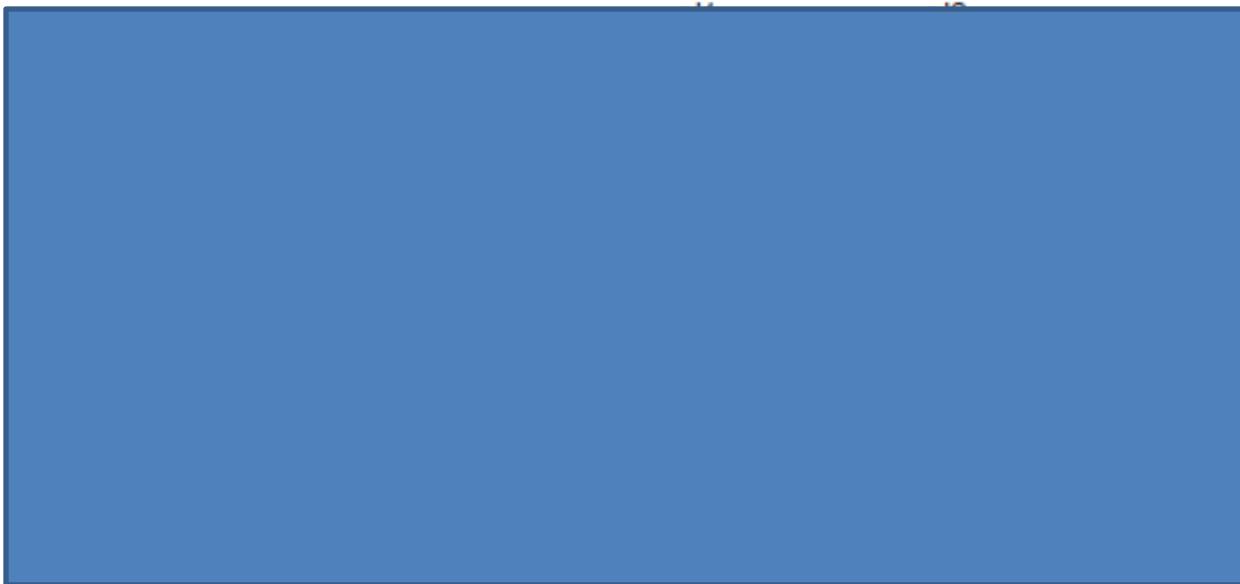
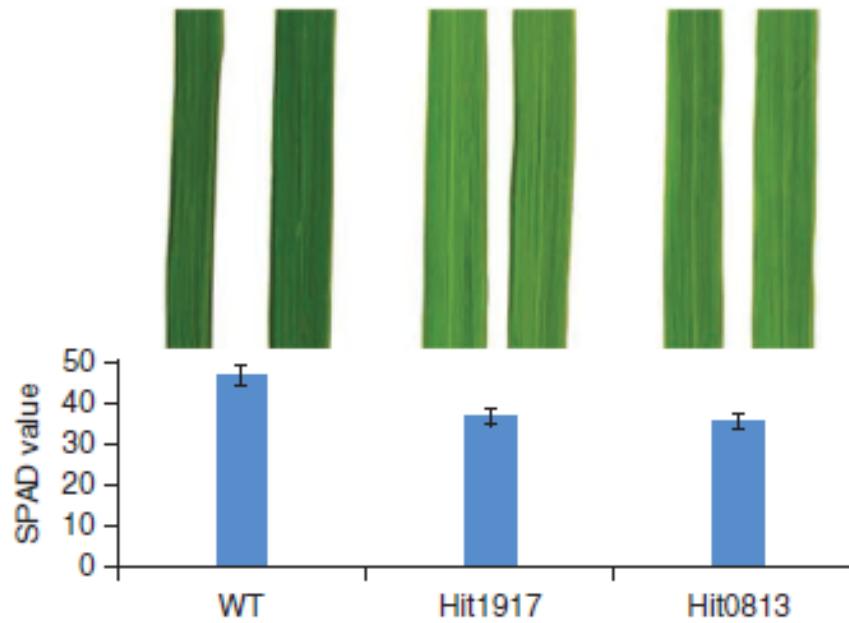
100% for a SNP tightly linked to the causal gene  
50% for a SNP unlinked to the mutation

>50% for a mutant SNP loosely linked to the causal mutation  
<50% for the wild type allele

# PoolSeq – MutMap for induced mutation

Induced mutation

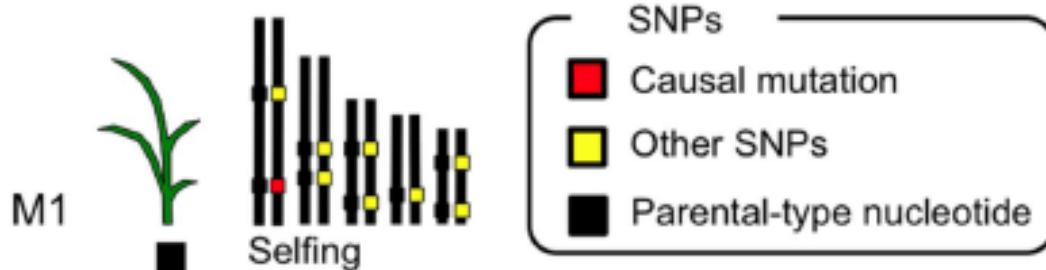
a



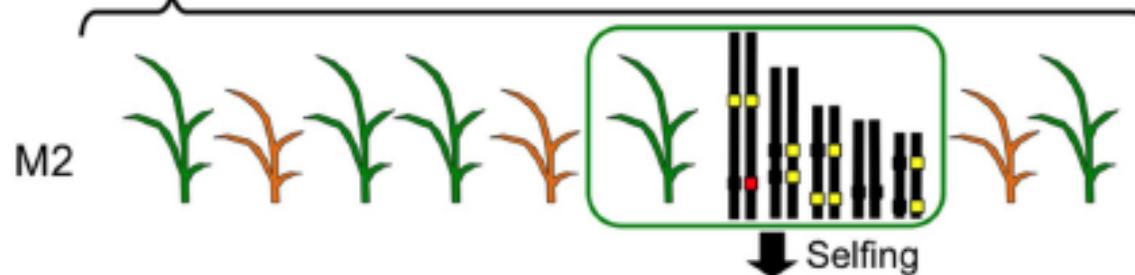
# Variant: MutMap+ → No need to cross

Induced mutation

A



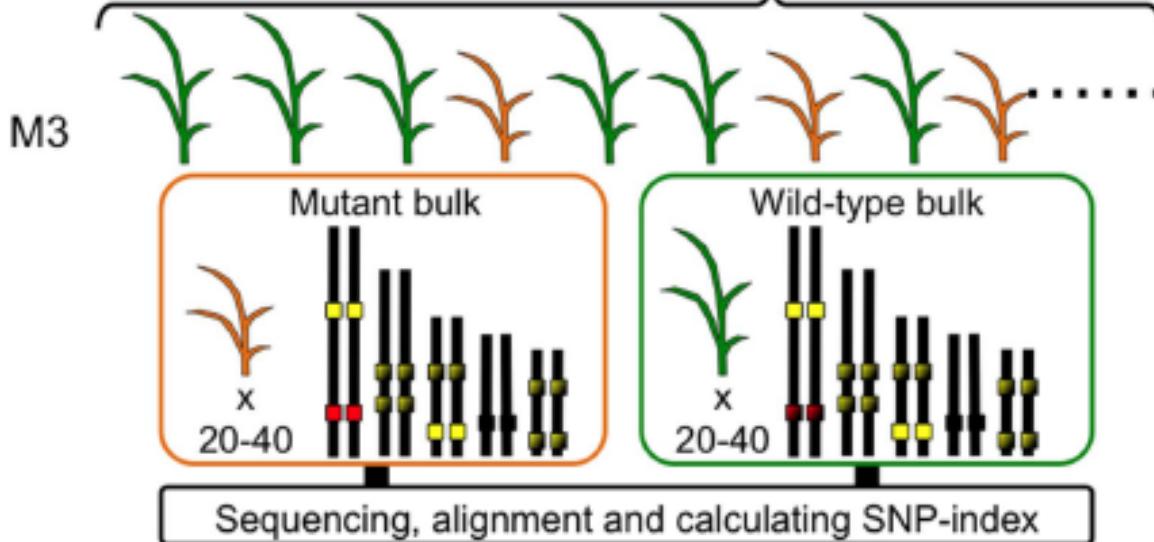
B

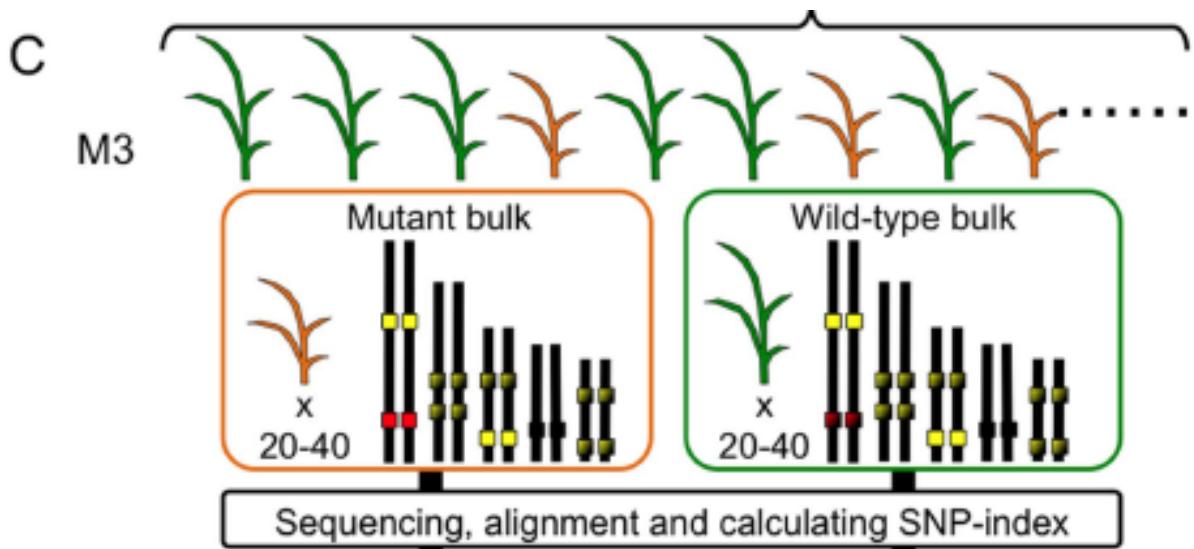


Some wt → all wt  
ignore

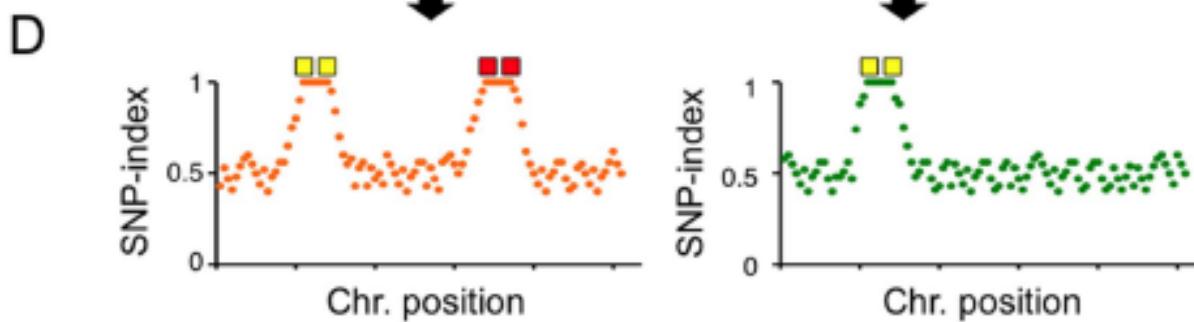
wt → wt + mut  
keep

C

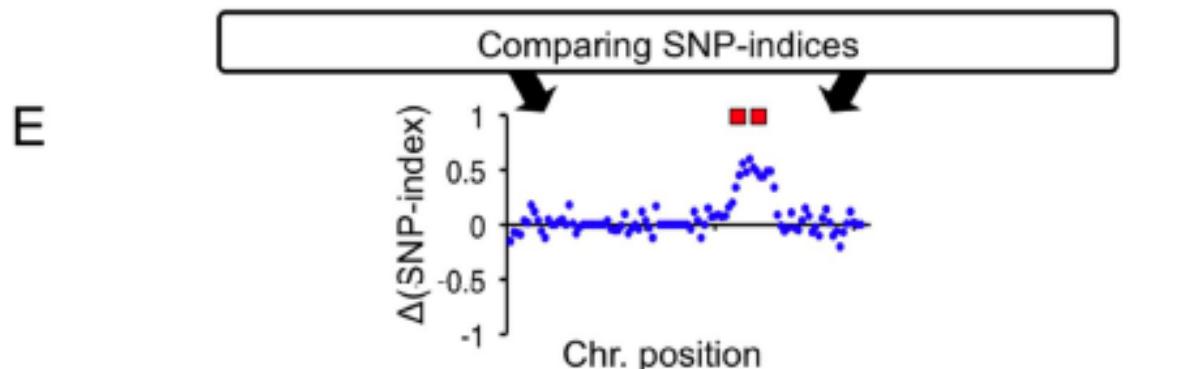




Induced mutation



Some SNPs fixed  
due to chance or  
drift



$\Delta\text{SNP index}$

# NGS and Poolseq

OLD:  
many graduate students and many years

NEW:  
single graduate student and fewer years

PI's and peons ostensibly happier

# Poolseq

xQTL mapping  
QTL-seq

# X-QTL mapping

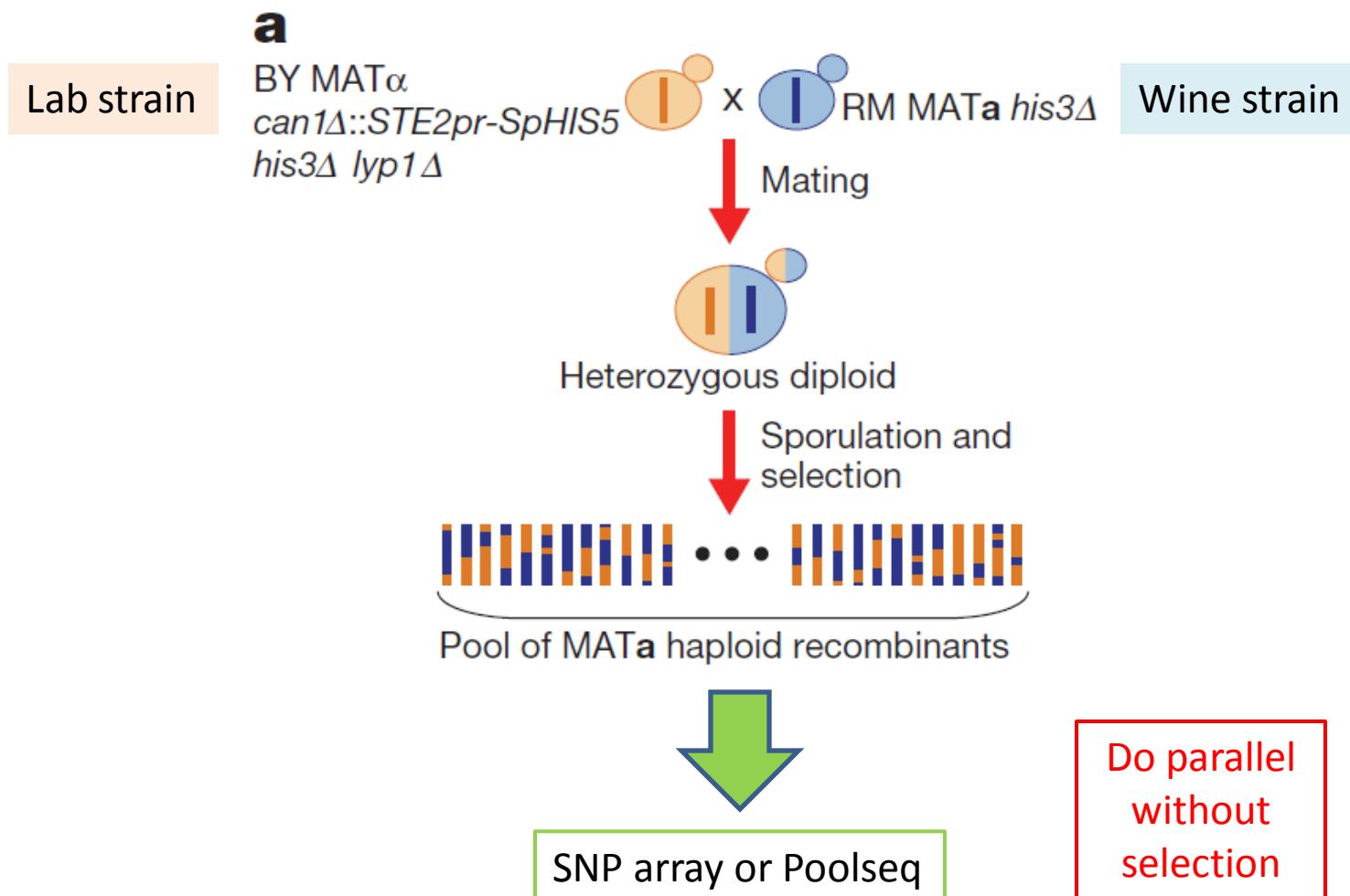
Highly complex traits → Multiple loci  
Often each with small effect

Need large populations for QTL mapping

## Dissection of genetically complex traits with extremely large pools of yeast segregants

Ian M. Ehrenreich<sup>1,2,3</sup>, Noorossadat Torabi<sup>1,4</sup>, Yue Jia<sup>1,3</sup>, Jonathan Kent<sup>1</sup>, Stephen Martis<sup>1</sup>, Joshua A. Shapiro<sup>1,2,3</sup>, David Gresham<sup>1</sup>†, Amy A. Caudy<sup>1</sup> & Leonid Kruglyak<sup>1,2,3</sup>

- step 1. generate segregating population of very large size
- step 2. selection based phenotyping (drug resistance or cell sorter)
- step 3. measure pooled allele frequencies (NGS)



# 4-NQO resistance case study

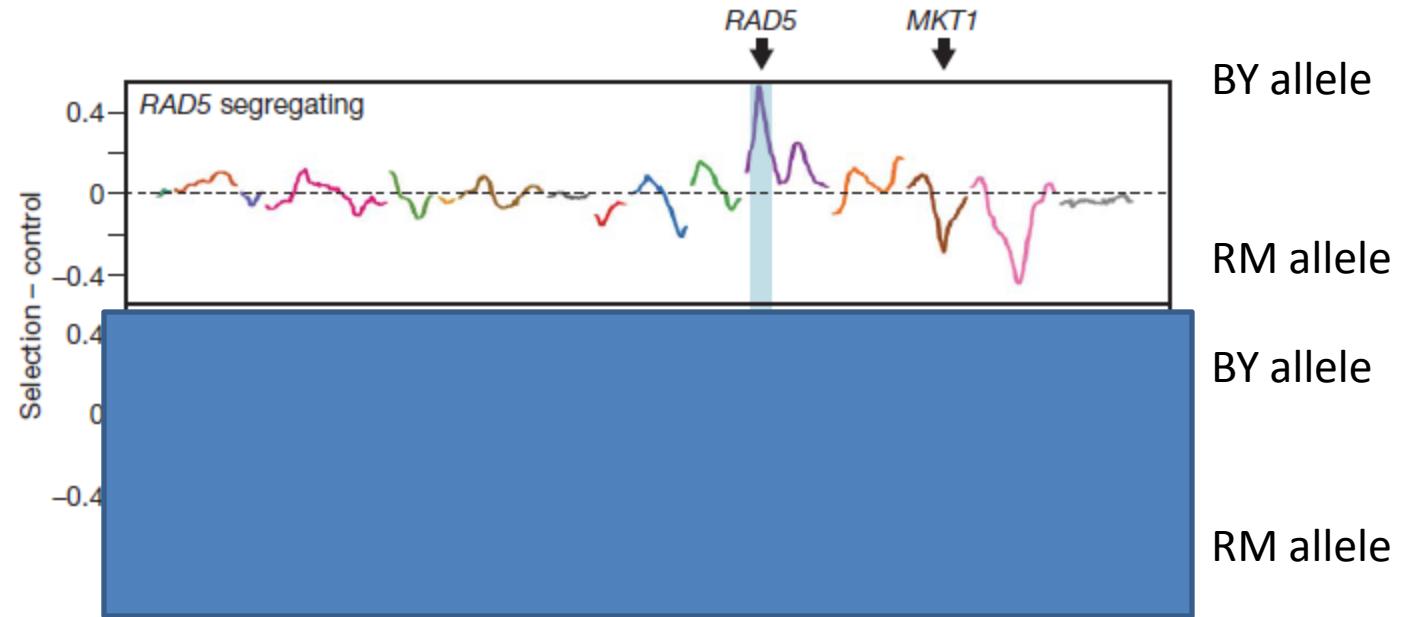
4-nitroquinoline  
DNA damaging agent

Previously:  
Conventional QTL  
RAD5 (DNA repair gene)

Follow up backcrossing revealed a 2<sup>nd</sup> gene:  
MKT1

# 4-NQO resistance case study

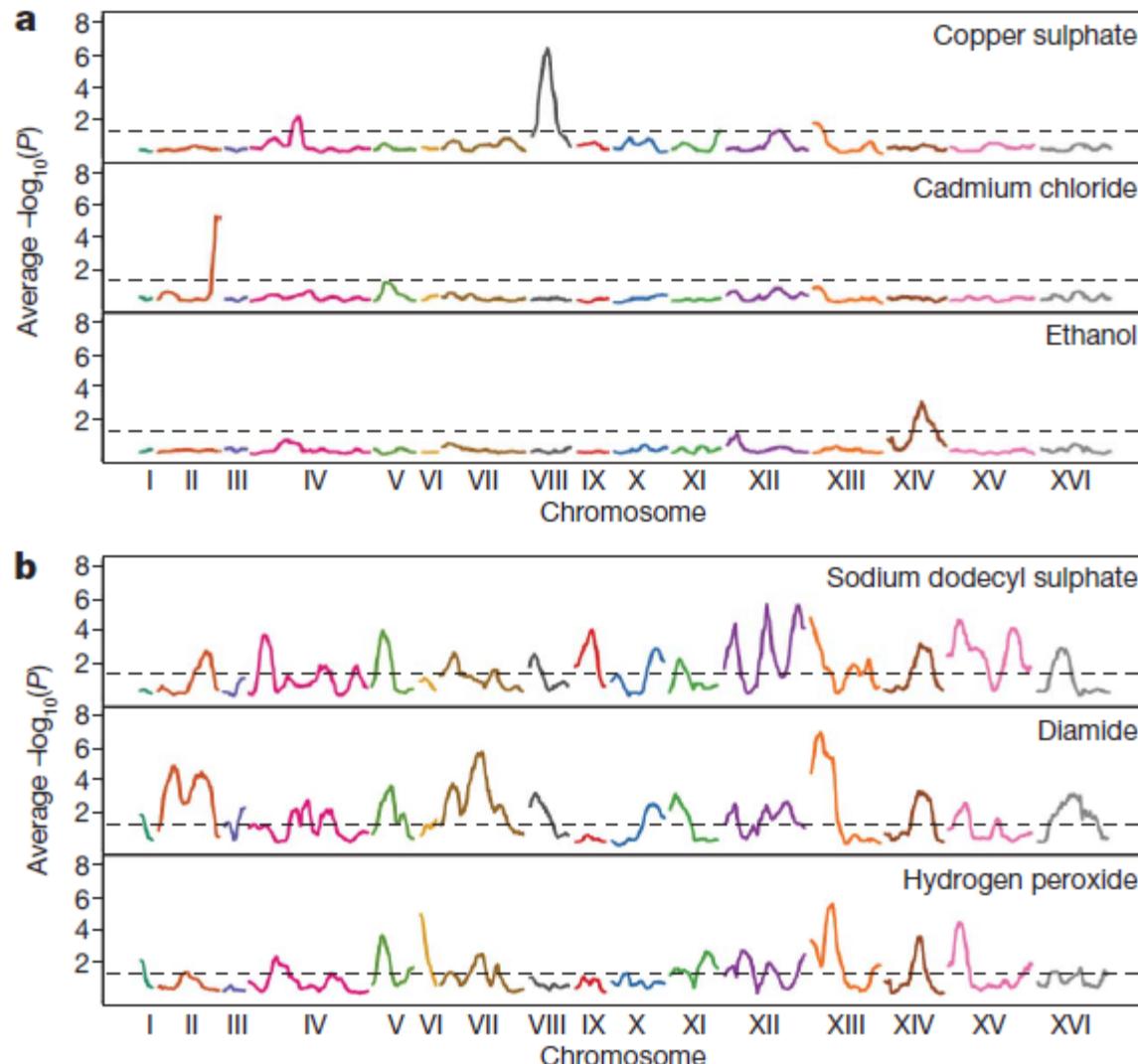
14 loci 70%



Very powerful

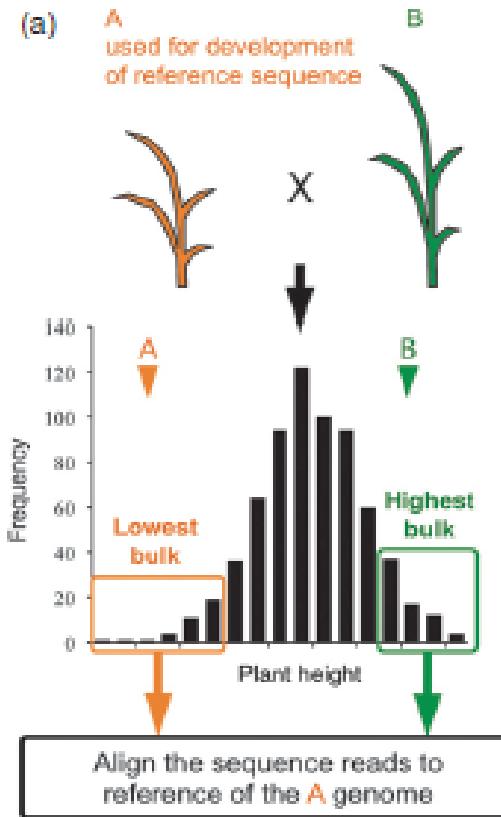
Basically never uncover this with single gene approach

# Other examples of simple and complex traits



xQTL mapping  
is advertised to require big N  
(more recombinants better resolution)

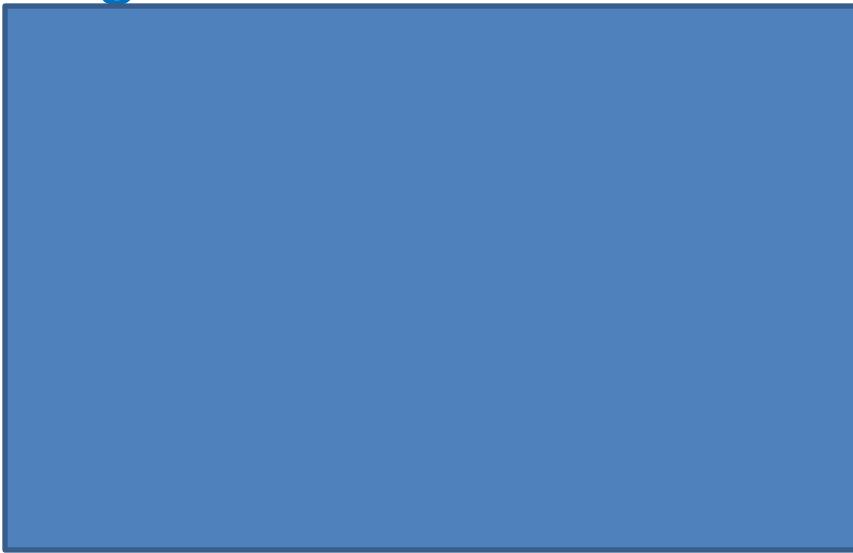
QTL-seq  
results suggest only 20-50 can be informative  
(although finding the exact gene is unclear)



- N = 20-50
- Larger genome
- Works on crops
- ~MutMap with SNP index

# Case study: seed vigor

(a)



# Poolseq

## GWAS

taking advantage of historical recombination

Basically, the more recombinants, the better map resolution.

F2 crosses or Recombinant Inbred Lines have limited recombination events.

Wild samples have history to help

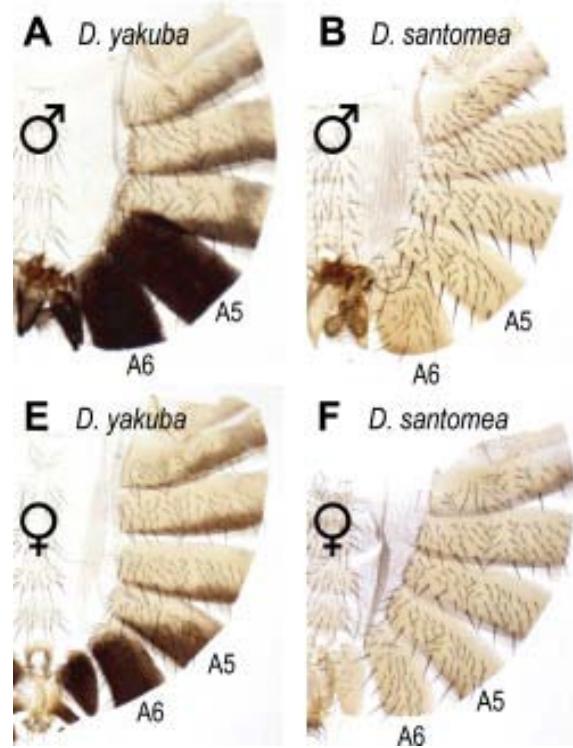
*Drosophila* body pigmentation example

# *Drosophila* pigmentation background

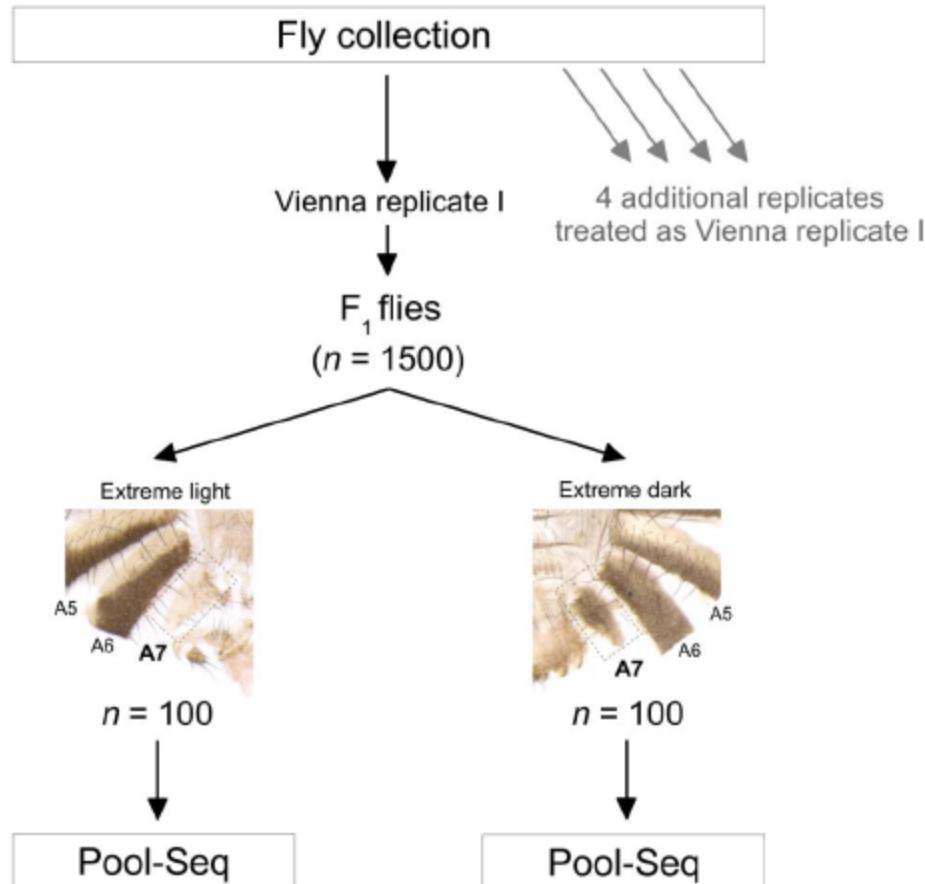
## *D. yakuba* x *D. santomea*

- There is variation in abdominal pigmentation
- Probably sexual selection

- Hybridization experiments identified: *tan* and *yellow*
- regulatory differences



# *D. melanogaster* → natural variation in pigmentation



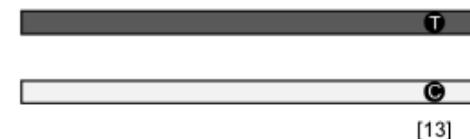
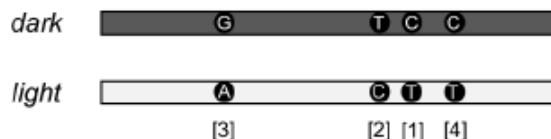
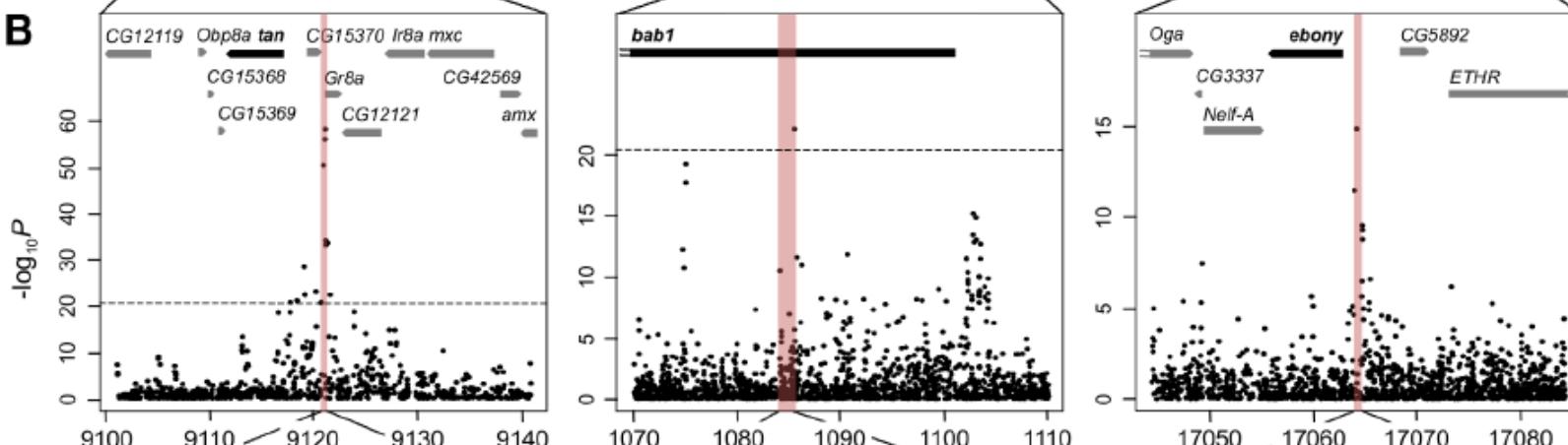
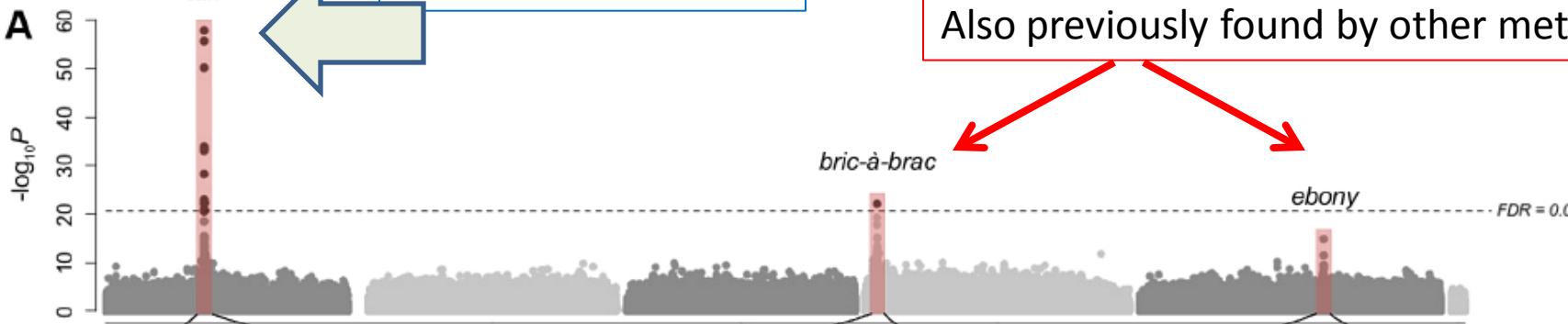
2 pops from Vienna, Austria  
3 pops from Bolzano, Italy

Grow in lab controlled conditions 1 gen  
Sort by color  
Take extremes  
Pool-seq with **replicates**

Matches *D. yakuba*/  
*santomea* result

All regulatory

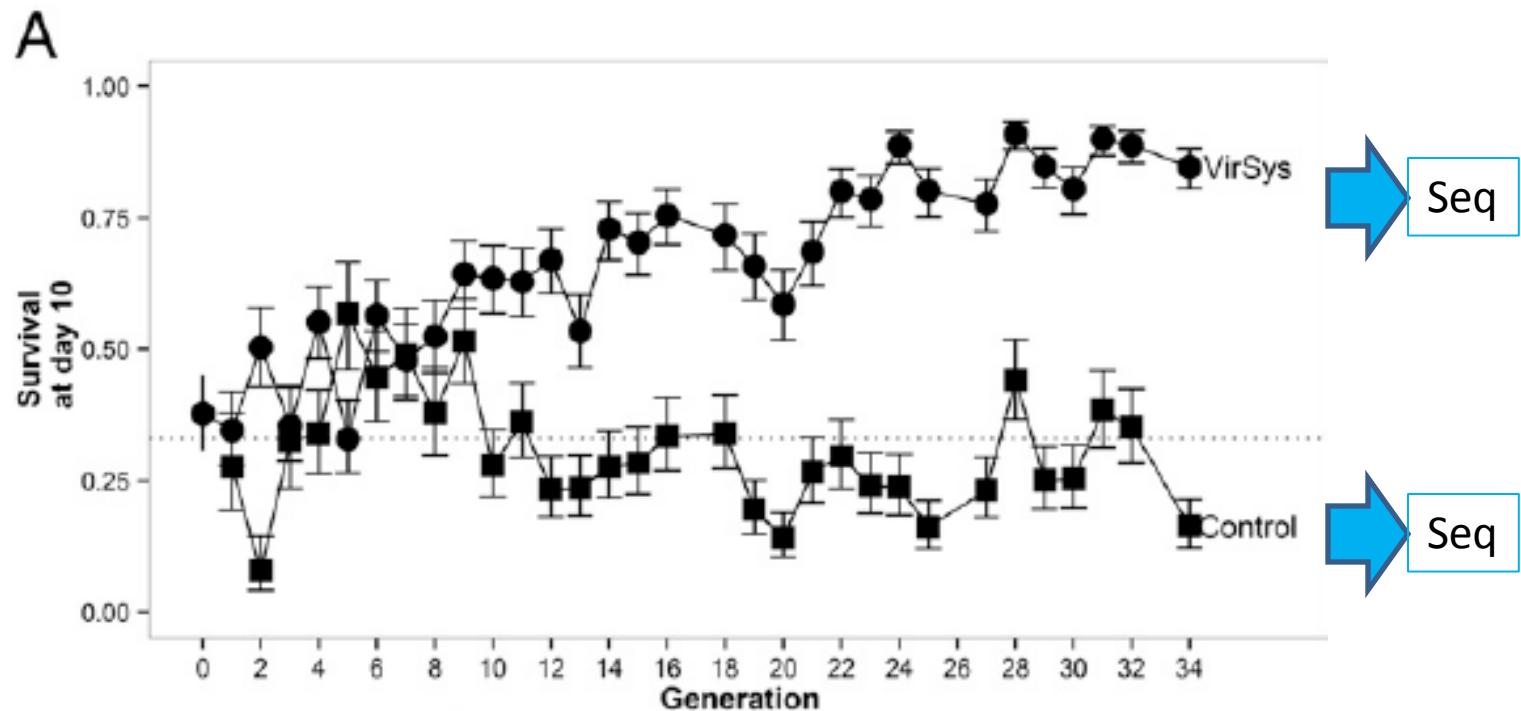
Also previously found by other methods



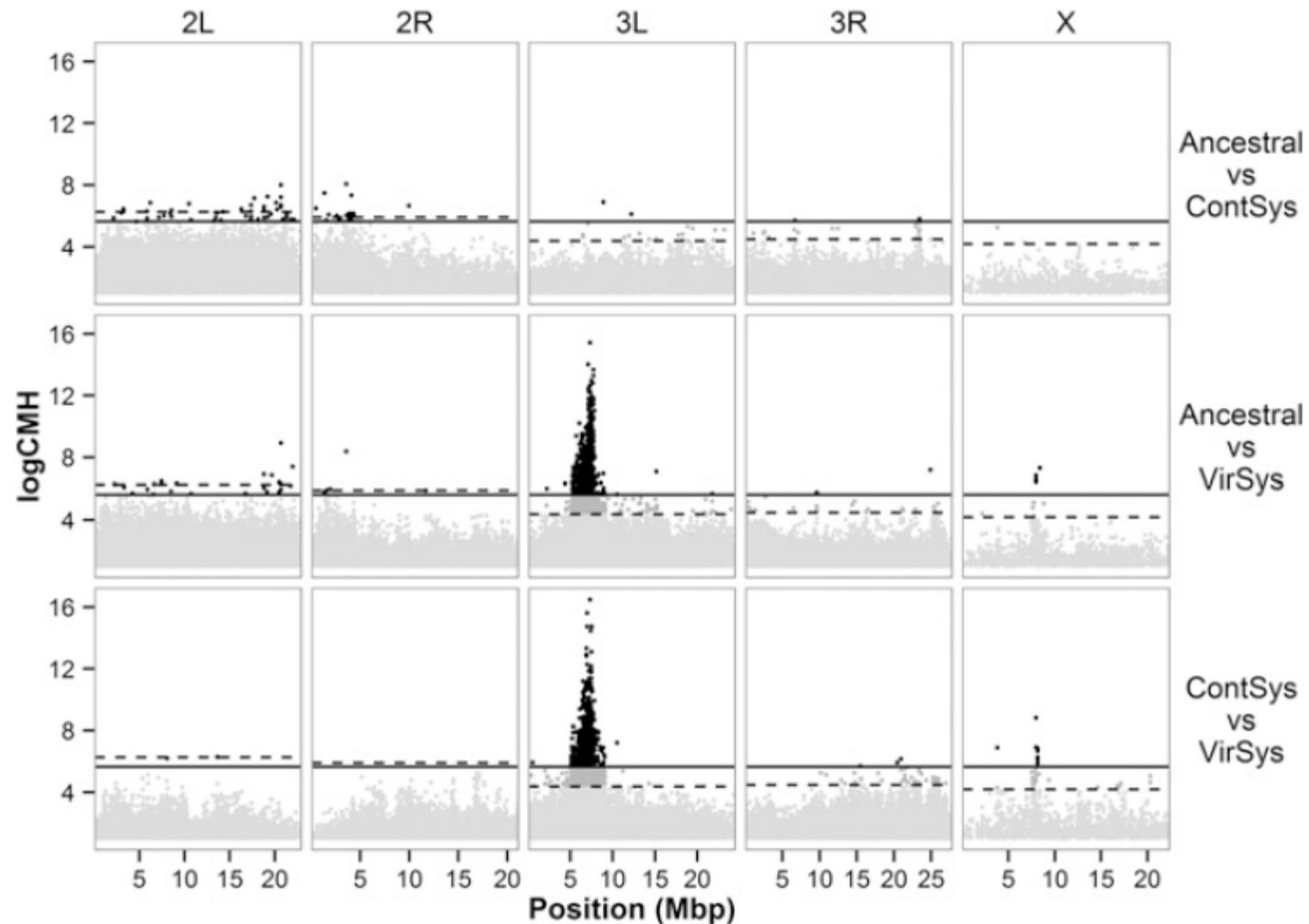
# Evolve and re-sequence

# Evolve and re-sequence

Evolve *Drosophila* lines to be more resistant to C virus  
4x replicates [virus, control1, control2]



# Evolve and re-sequence

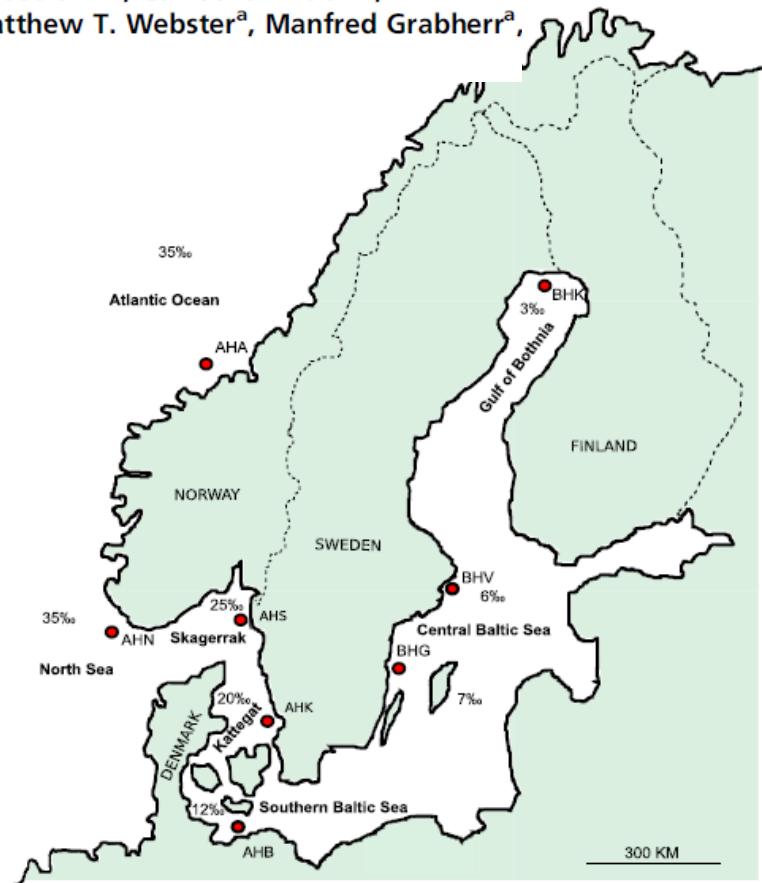


*pastrel* gene (and lesser extent *Ubc-E2H*)

# Ecology (reverse ecology) and local adaptation

## Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring

Sangeet Lamichhaney<sup>a,1</sup>, Alvaro Martinez Barrio<sup>a,1</sup>, Nima Rafati<sup>a,1</sup>, Görel Sundström<sup>a,1</sup>, Carl-Johan Rubin<sup>a</sup>, Elizabeth R. Gilbert<sup>a,2</sup>, Jonas Berglund<sup>a</sup>, Anna Wetterbom<sup>b</sup>, Linda Laikre<sup>c</sup>, Matthew T. Webster<sup>a</sup>, Manfred Grabherr<sup>a</sup>, Nils Ryman<sup>c</sup>, and Leif Andersson<sup>a,d,3</sup>

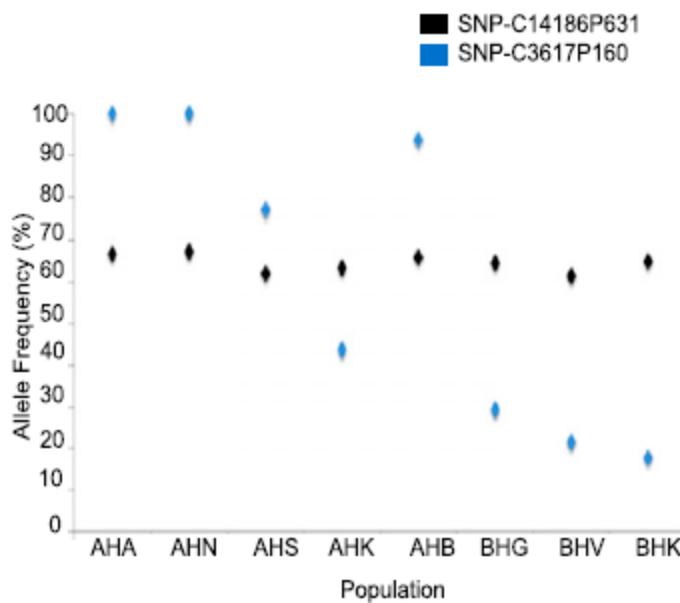


# Ecology (reverse ecology) and local adaptation



8 Pool-seq populations  
Each with 50 individuals  
30x coverage  
Map against transcriptome

A



2 SNPs with different  
allele frequency patterns

Most undifferentiated  
~440,000  
Few highly differentiated <1%  
~3800

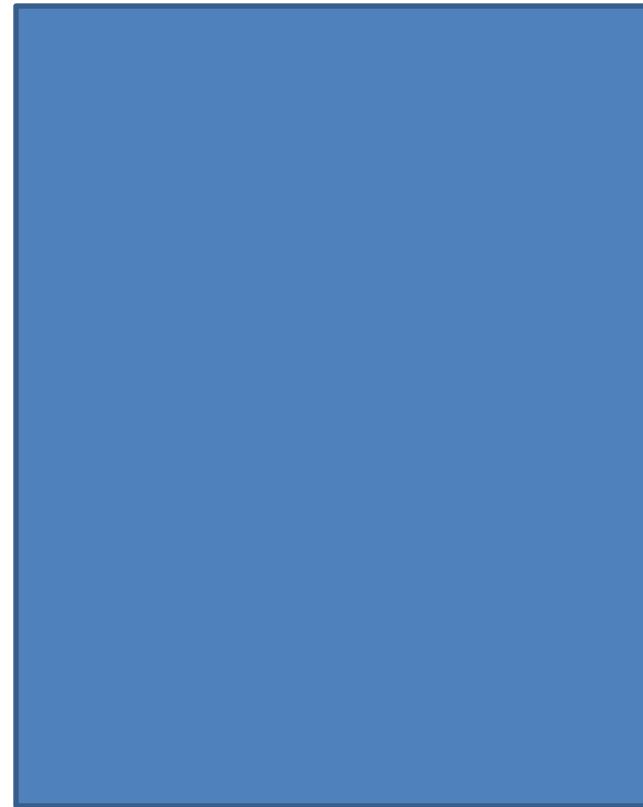
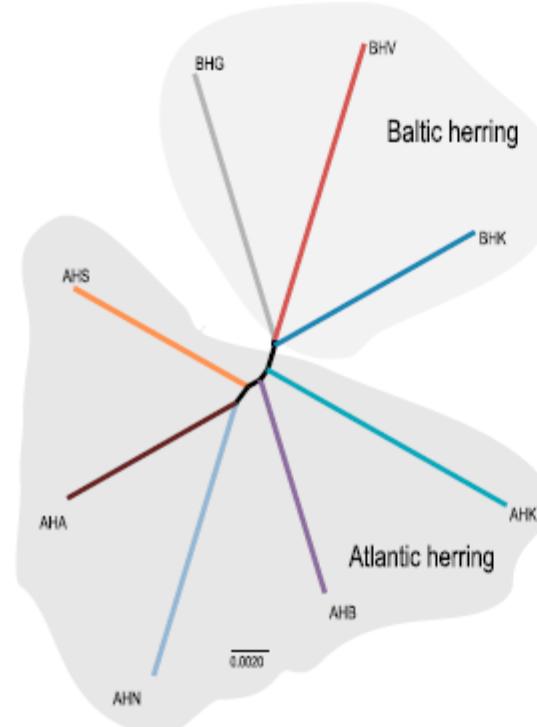
# Little difference between Baltic and Atlantic if using all SNPs...



All SNPs →  
(star phylogeny)

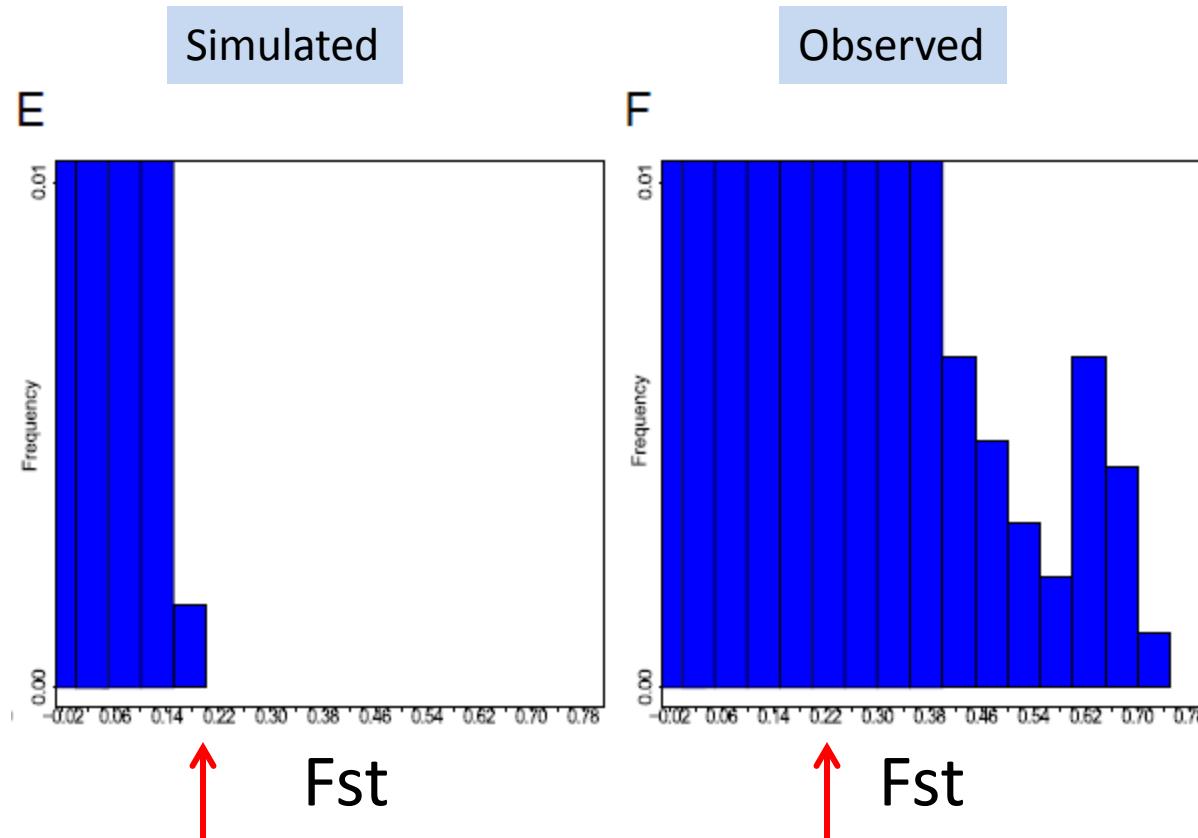
High Fst SNPs →  
separate populations

B



# Are high Fst SNPs from drift or selection?

Simulation analysis to test



Unlikely drift → therefore selection!

local adaptation

Now need to find causative genes and mechanism

Lamichhaney et al 2012

# Domestication – artificial selection → Wolf vs Dog



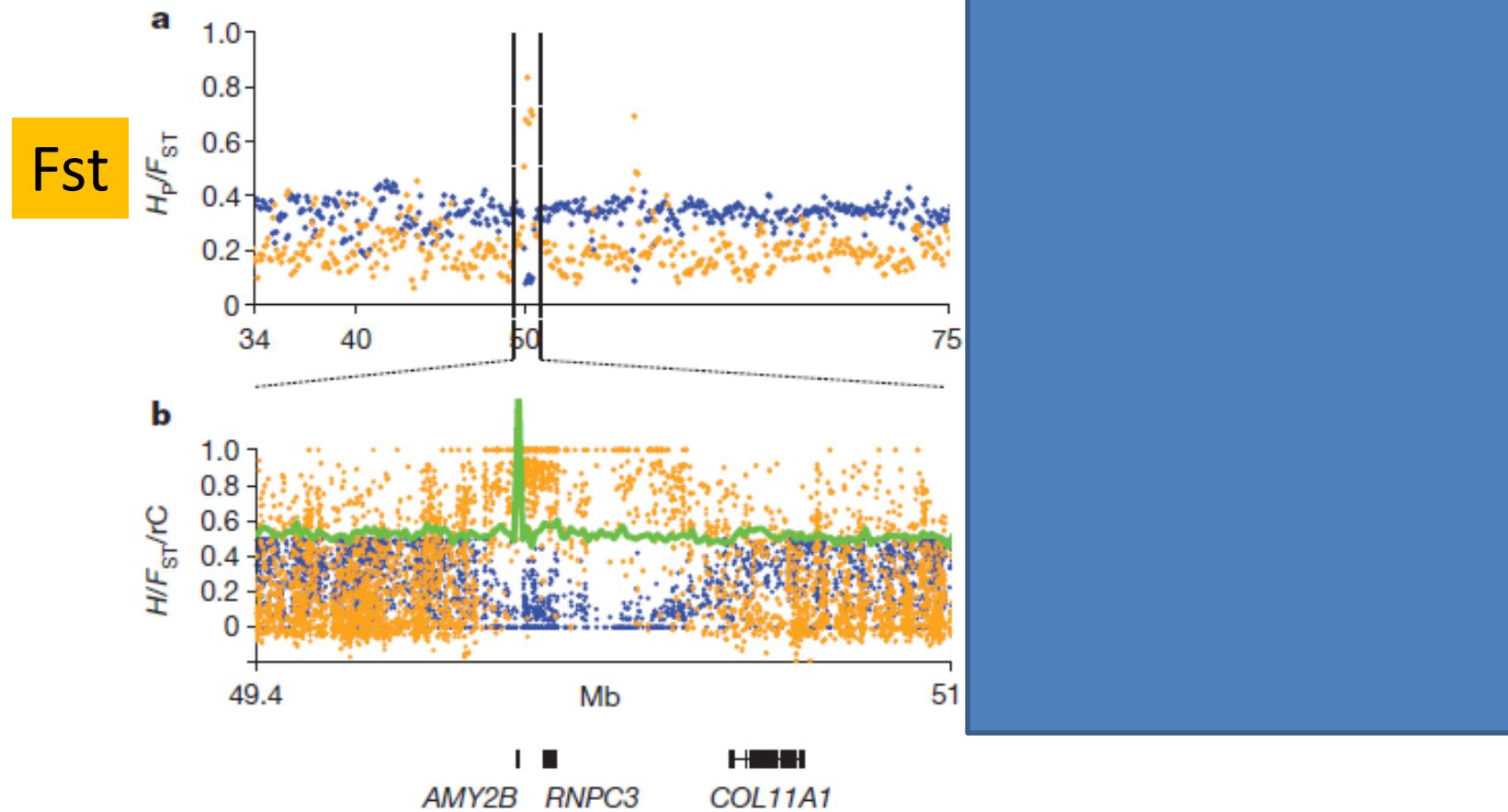
By Retron - self-made now, Public Domain,  
<https://commons.wikimedia.org/w/index.php?curid=3865957>



Charles M. Schulz

# Dog adaptation to starch diet

36 candidate domestication loci  
here → Amylase



also maltase activity

# A basic science question a very population genetics question

Recombination analysis using Pool-seq



Photo James Gaither

*Mimulus guttatus*  
(Common yellow monkeyflower)

# Recombination analysis using Pool-seq



Earlier, mentioned that Pool-seq not so good for recombination/LD studies because **too short**

However, monkeyflower has high nt diversity (2.9%)  
i.e., within 1 Illumina read, there can be  $\geq 2$  nt differences



# Samples and analysis sketch



98 samples

4 localities

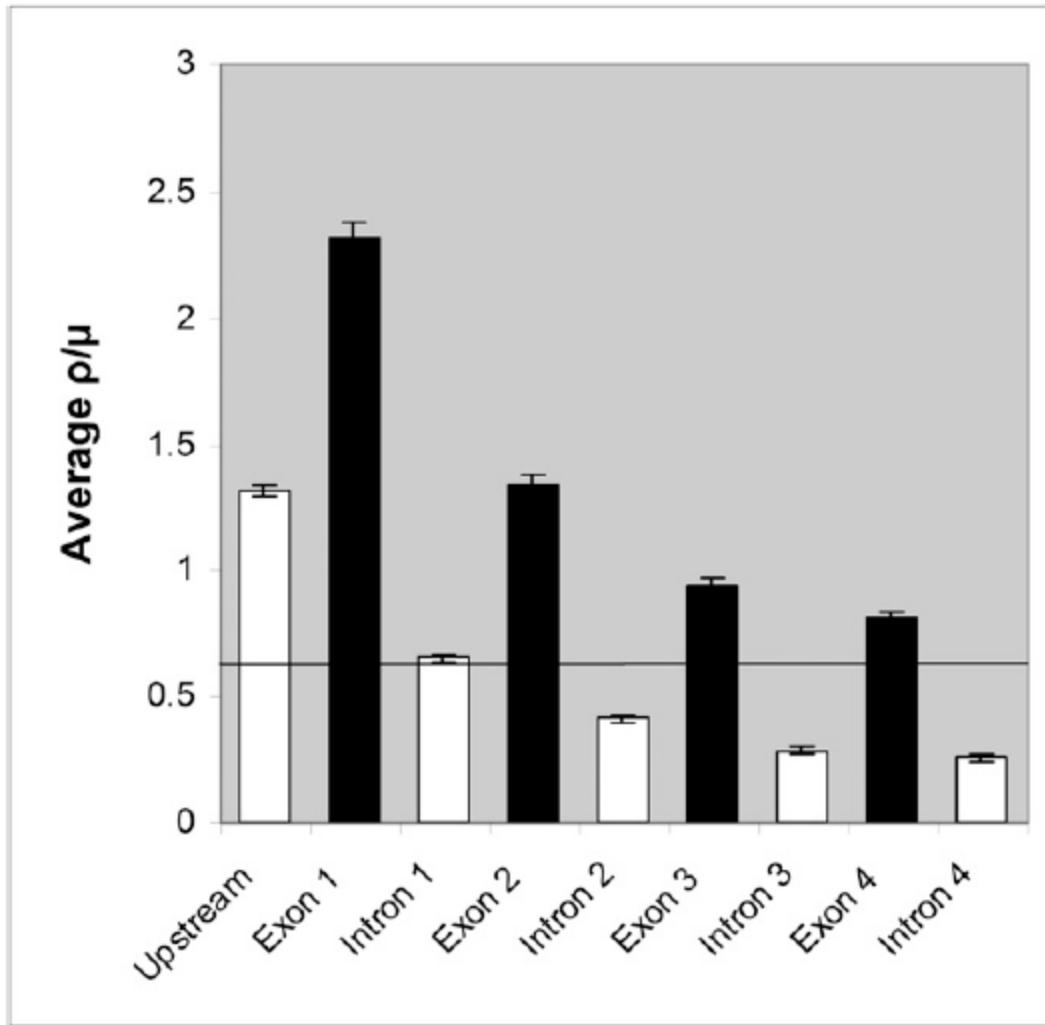
4 Pool-seq runs

PE 75 bp runs (of ~200 bp fragments)

After QC → 255x coverage

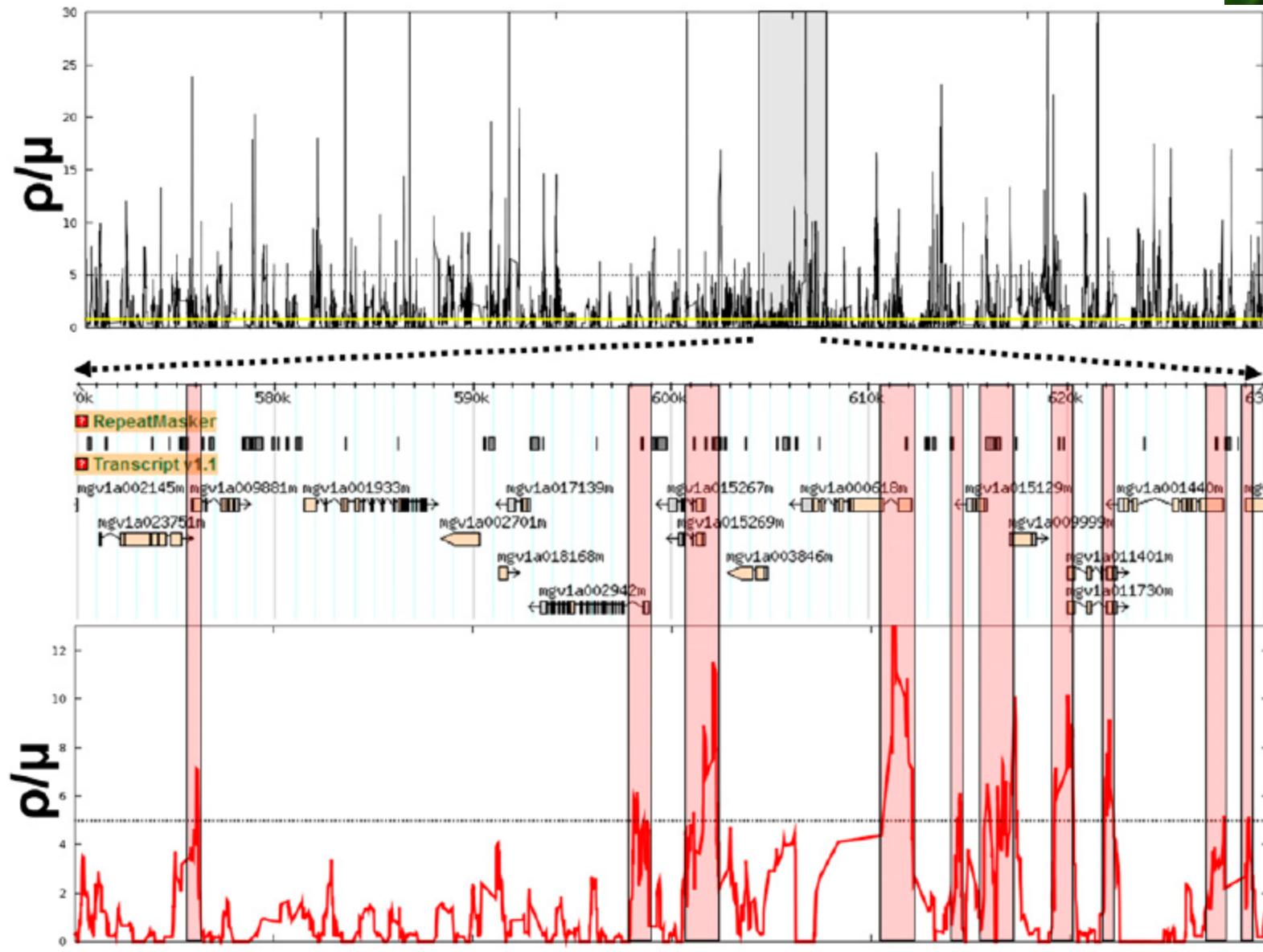
Examined pairs of SNPs within 50 bp

# Interesting recombination patterns



- Exons > introns
- Polarity  
 $5' > 3'$  recomb. rate

# Recombination landscape example



# Whole genome (re)sequencing

Can do lots of stuff  
Here, demographic history,  
Scan for regions of diversity,  
Identify gene for phenotype

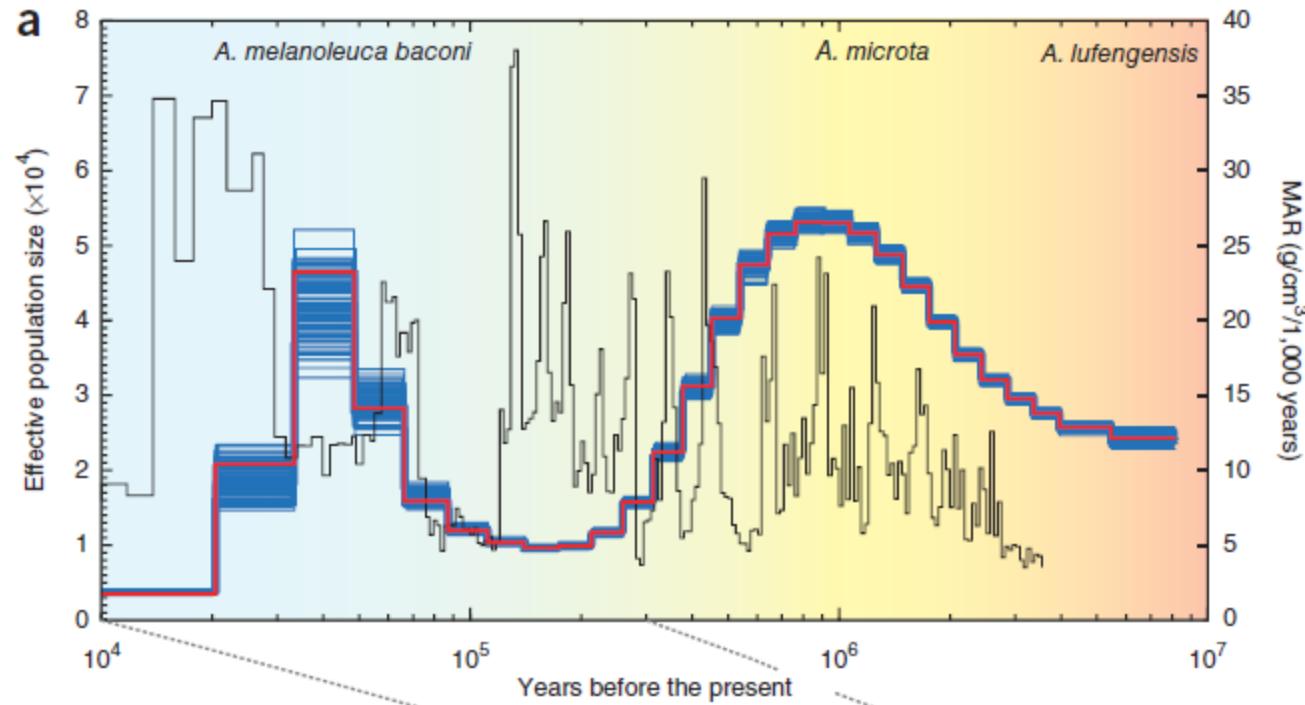
# Panda demographic history

34 wild panda + ref seq

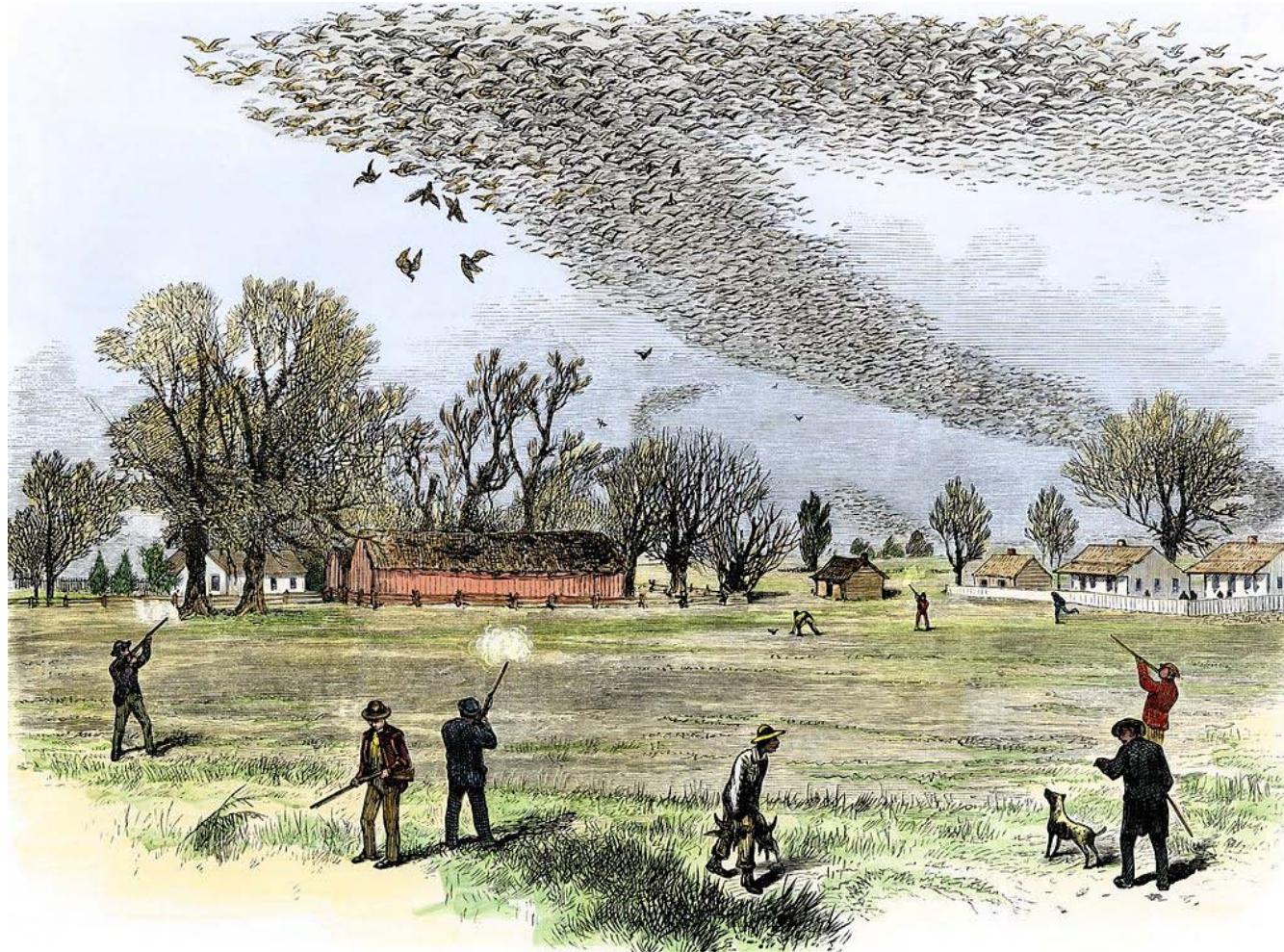


Pairwise sequentially Markovian coalescent (PSMC)

Can infer past history from 1 or few individuals

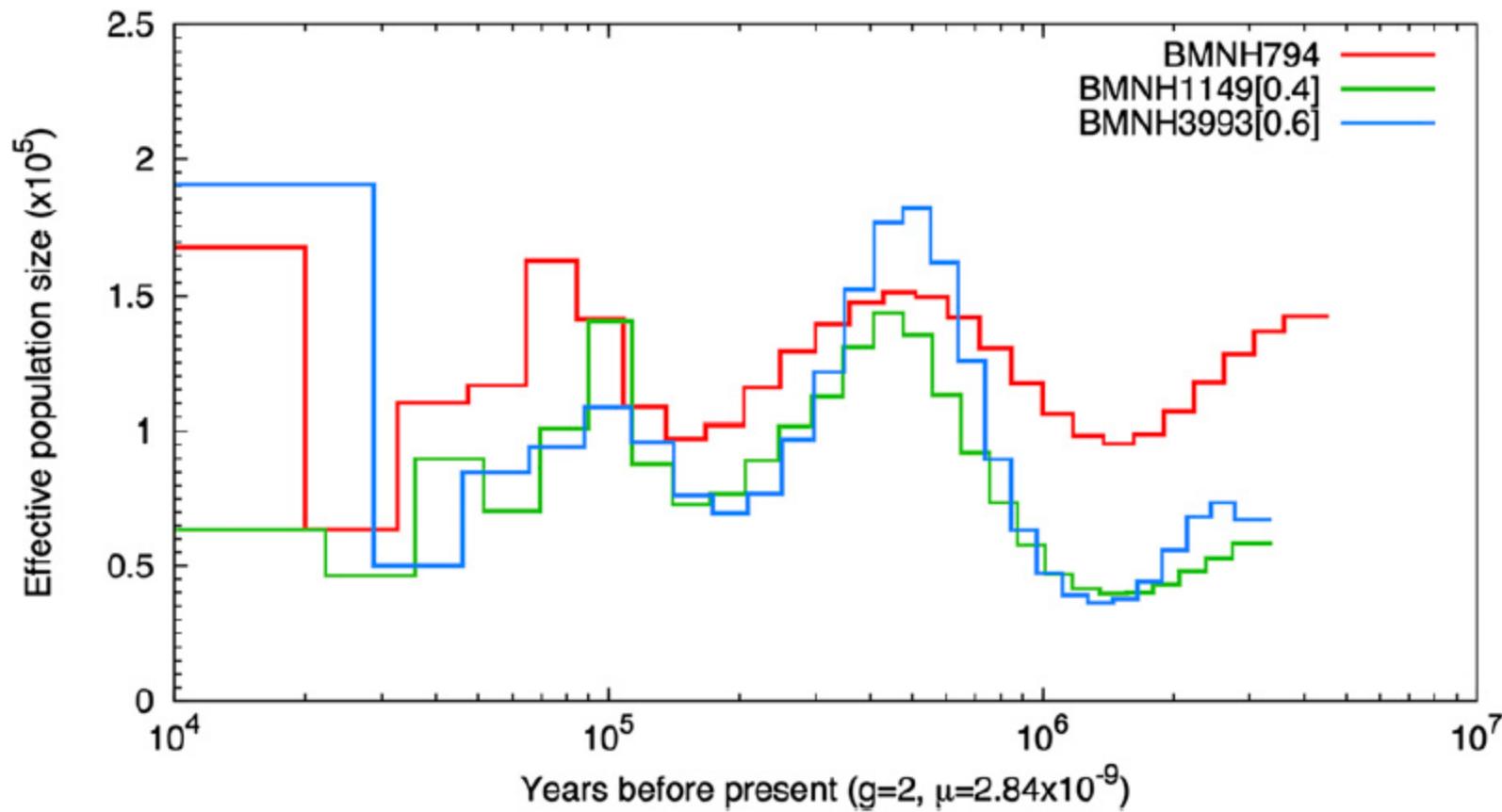


# Passenger pigeon demographic history



# Passenger pigeon demographic history (+ museum DNA)

N=3 from museum, PSMC

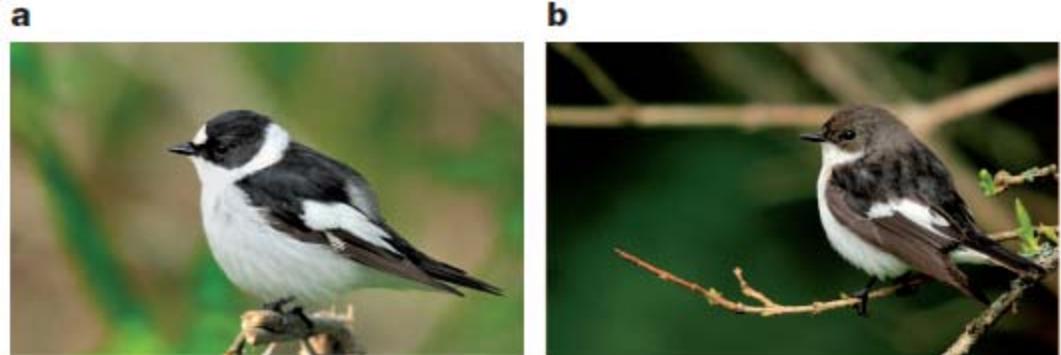


$N_e$  (effective population size; 100,000's) << census N (billions)

Past fluctuations → climatic, food-resource [acorn], and other ecological variations

Increased extinction risk

# Speciation islands? (divergence islands)



collared

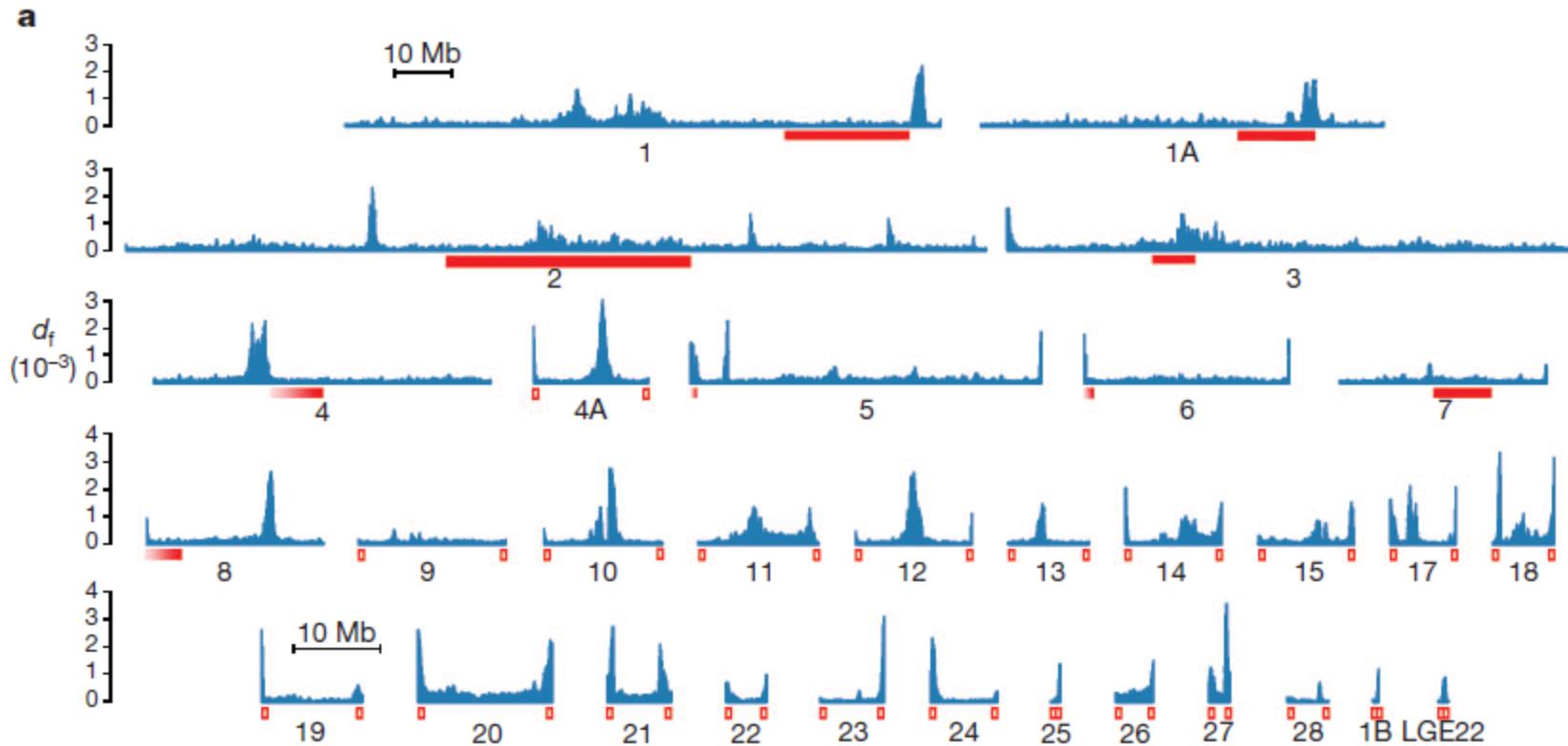
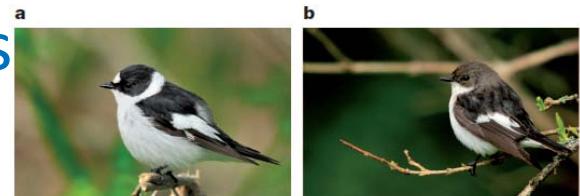
pied

flycatchers  
diverged < 2Mya  
separated, 2<sup>o</sup> contact  
hybridize

10 re-sequenced individuals + 1 ref seqs (collared)

Compared genomewide divergence to each other

# Divergence islands often near centromeres and telomeres



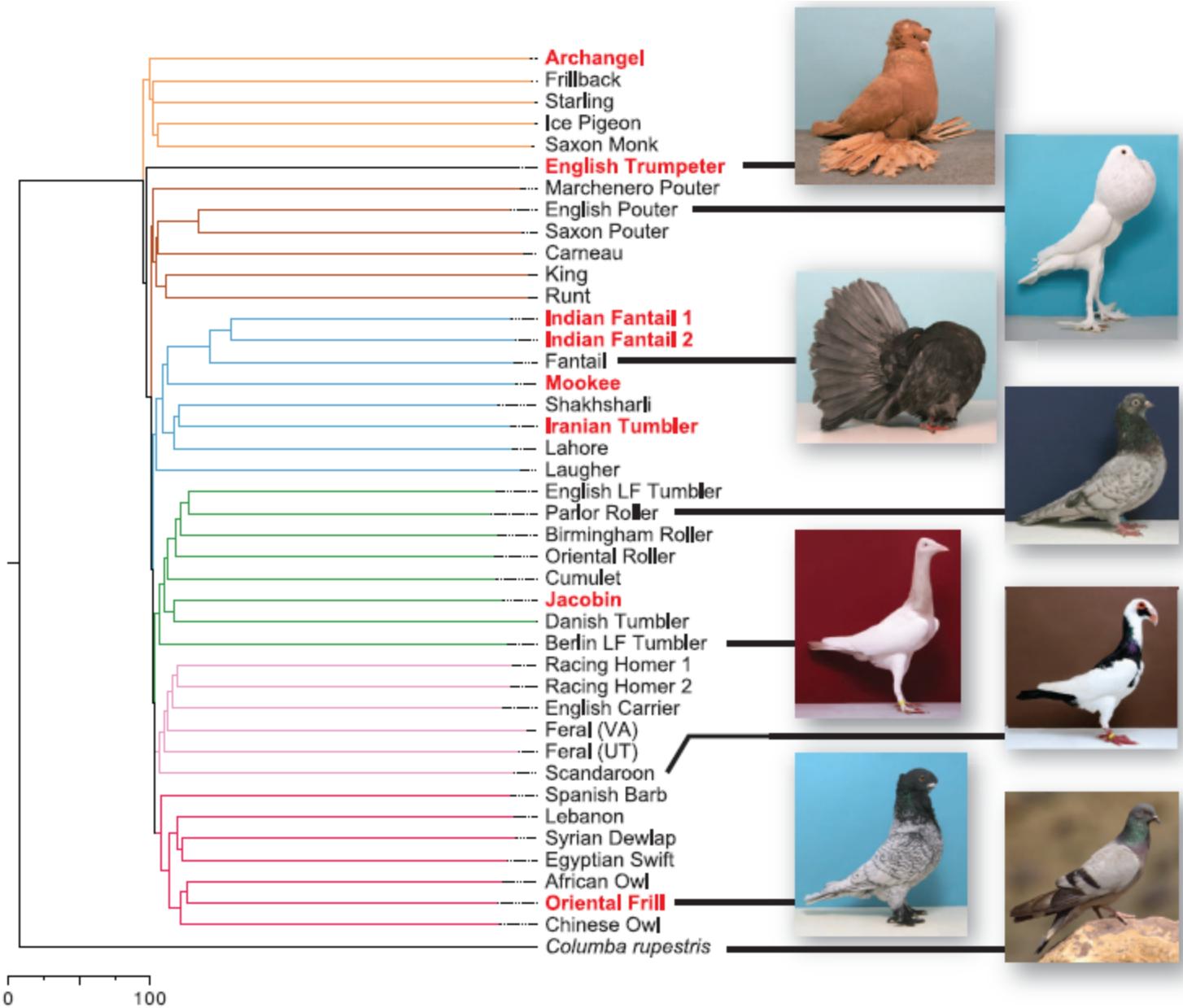
Authors postulate meiotic drive

# Genomic Diversity and Evolution of the Head Crest in the Rock Pigeon

Michael D. Shapiro,<sup>1\*</sup> Zev Kronenberg,<sup>2</sup> Cai Li,<sup>3,4</sup> Eric T. Domyan,<sup>1</sup> Hailin Pan,<sup>3</sup>  
Michael Campbell,<sup>2</sup> Hao Tan,<sup>3</sup> Chad D. Huff,<sup>2,5</sup> Haofu Hu,<sup>3</sup> Anna I. Vickrey,<sup>1</sup>  
Sandra C. A. Nielsen,<sup>4</sup> Sydney A. Stringham,<sup>1</sup> Hao Hu,<sup>5</sup> Eske Willerslev,<sup>4</sup>  
M. Thomas P. Gilbert,<sup>4,6</sup> Mark Yandell,<sup>2</sup> Guojie Zhang,<sup>3</sup> Jun Wang<sup>3,7,8\*</sup>

1 ref + 40 additional re-sequenced pigeon genomes

# 1 ref + 40 additional re-sequenced pigeon genomes



# What is the genetic basis for head crests?

Recessive mendelian trait



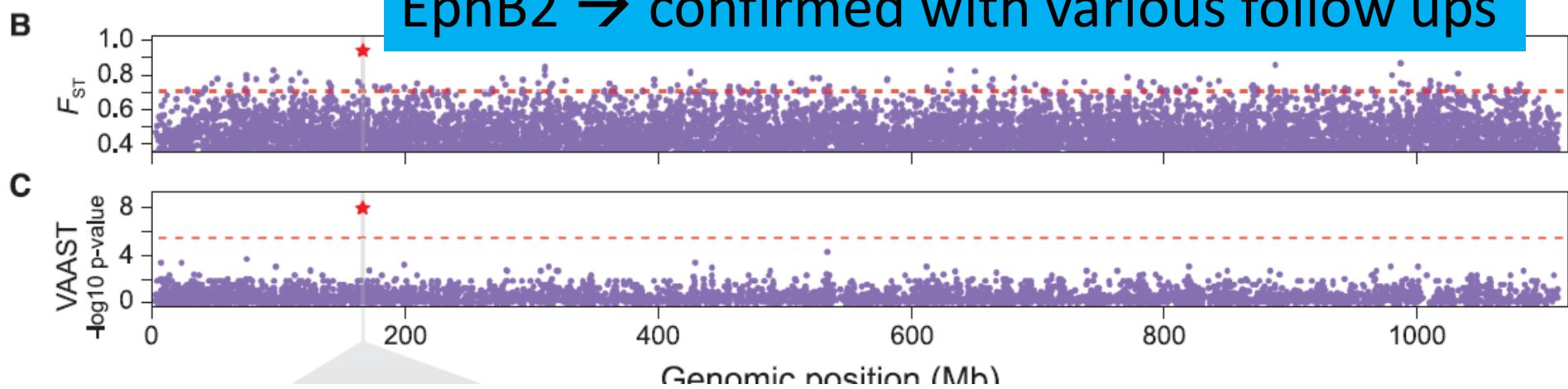
Peak crest

Shell crest

Mane

Hood

EphB2 → confirmed with various follow ups



## Concluding remarks 1

- Population genomics can be very powerful
- Find the right question (like all of science)
  - Better to know some population genetics
- Studies often stronger when combined with mechanism

## Concluding remarks 2

- There are economical and expensive ways
- Getting cheaper
  - Trend to whole genomes
  - More samples
- No plug and play; no established pipelines
- Finally, hope you get some ideas for your projects