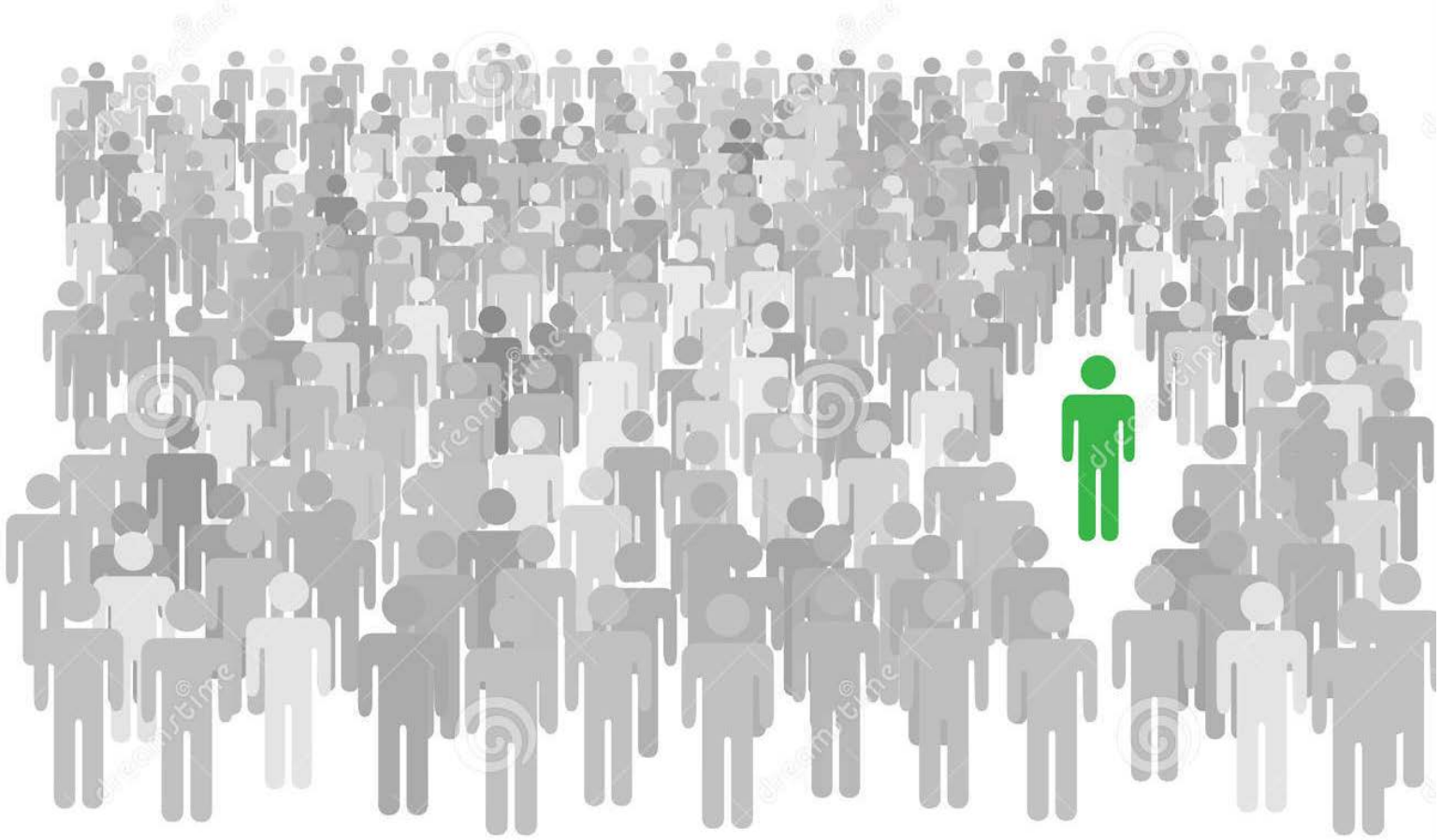
Population genomics



John Wang
April 29, 2020
Intro NGS

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:

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:

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:

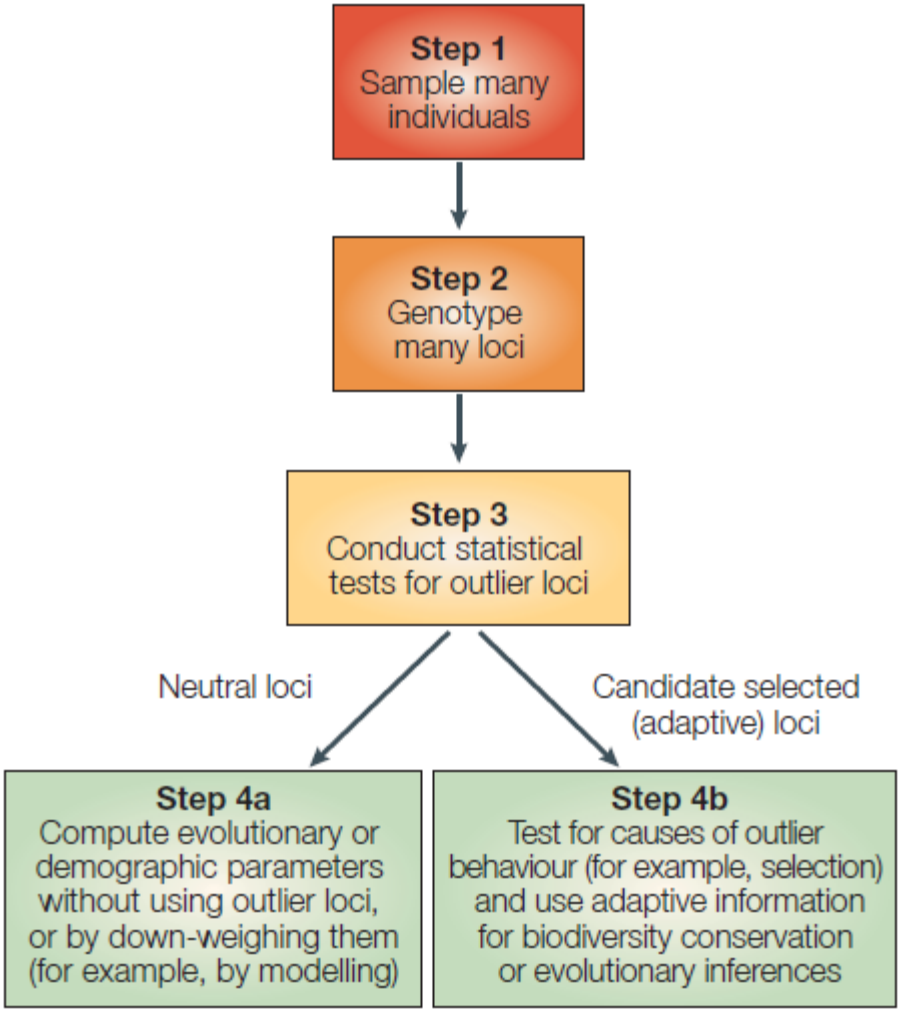


Figure 1 | **Flow chart of the four main steps in the population-genomic approach.** The approach summarized

Population genomics:

Outlier detection -- Fst

Neutral loci

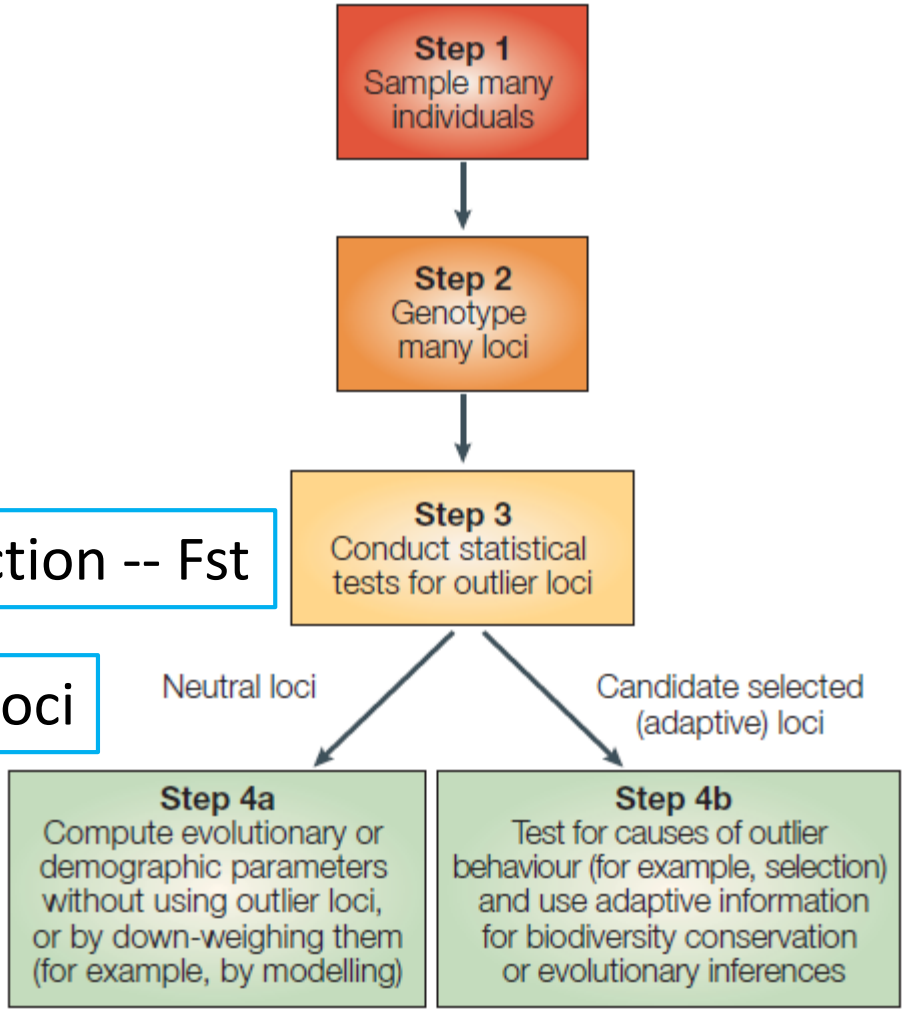


Figure 1 | **Flow chart of the four main steps in the population-genomic approach.** The approach summarized

Population genomics:

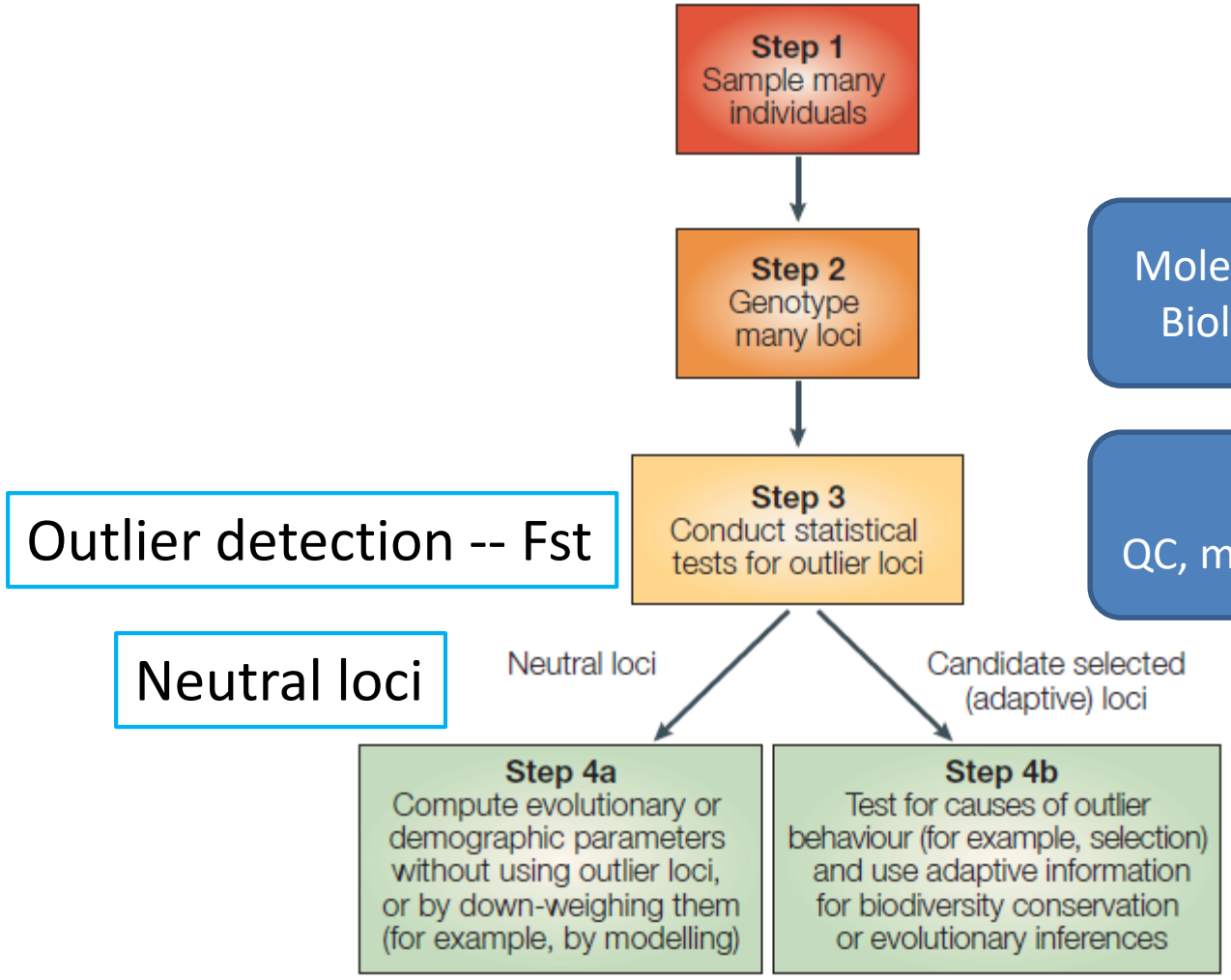


Figure 1 | **Flow chart of the four main steps in the population-genomic approach.** The approach summarized

Molecular Biology

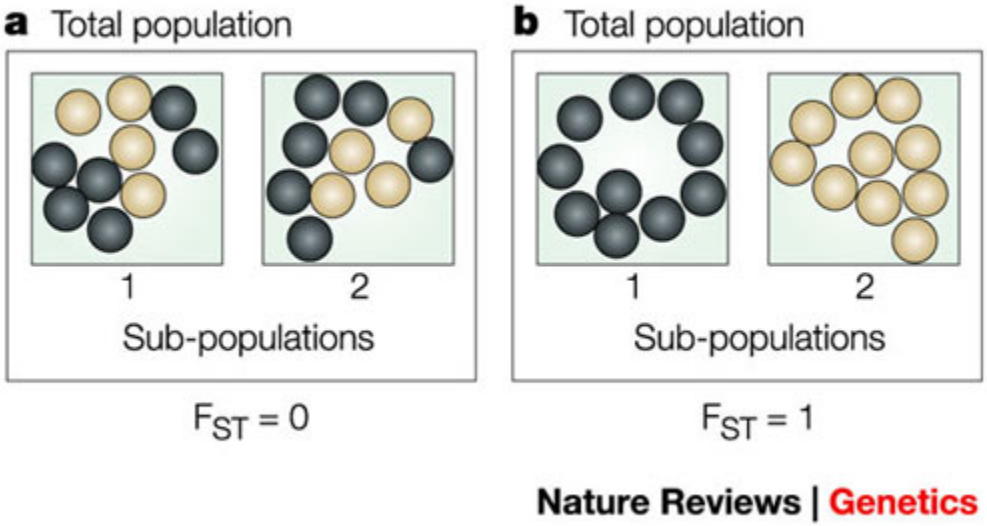
Sequencing strategy

Bioinformatics
QC, mapping, variant calling …

Outlier detection -- Fst

Neutral loci

Fixation index – F_{ST}



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

Some population genomics/genetics questions

Some population genomics/genetics questions

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

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]

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]

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]

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]

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]

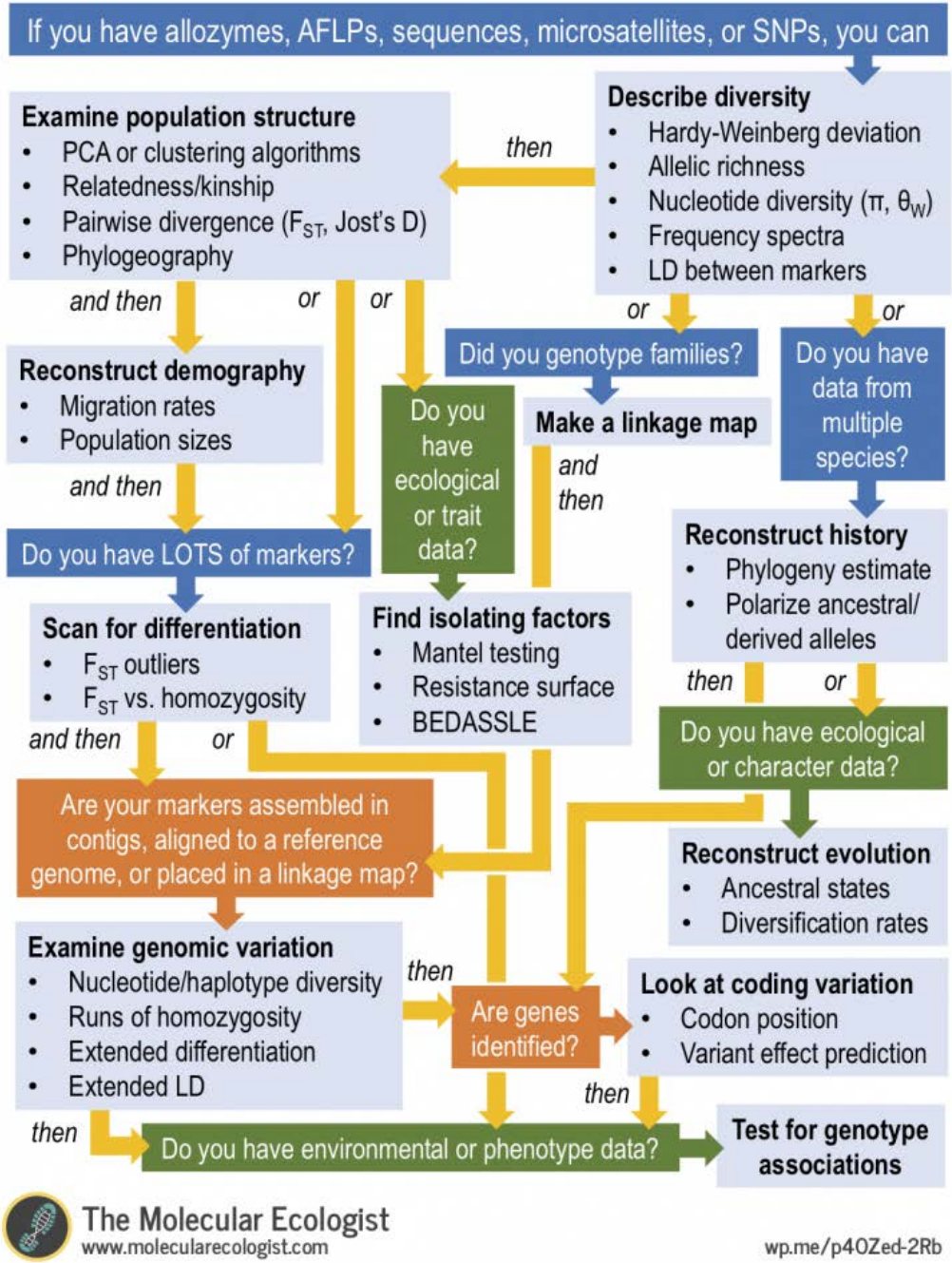
What was past demographic history? [out of Africa]

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

What can I do with this genetic data?



The Molecular Ecologist
www.molecularecologist.com

wp.me/p4OZed-2Rb

Approaches

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

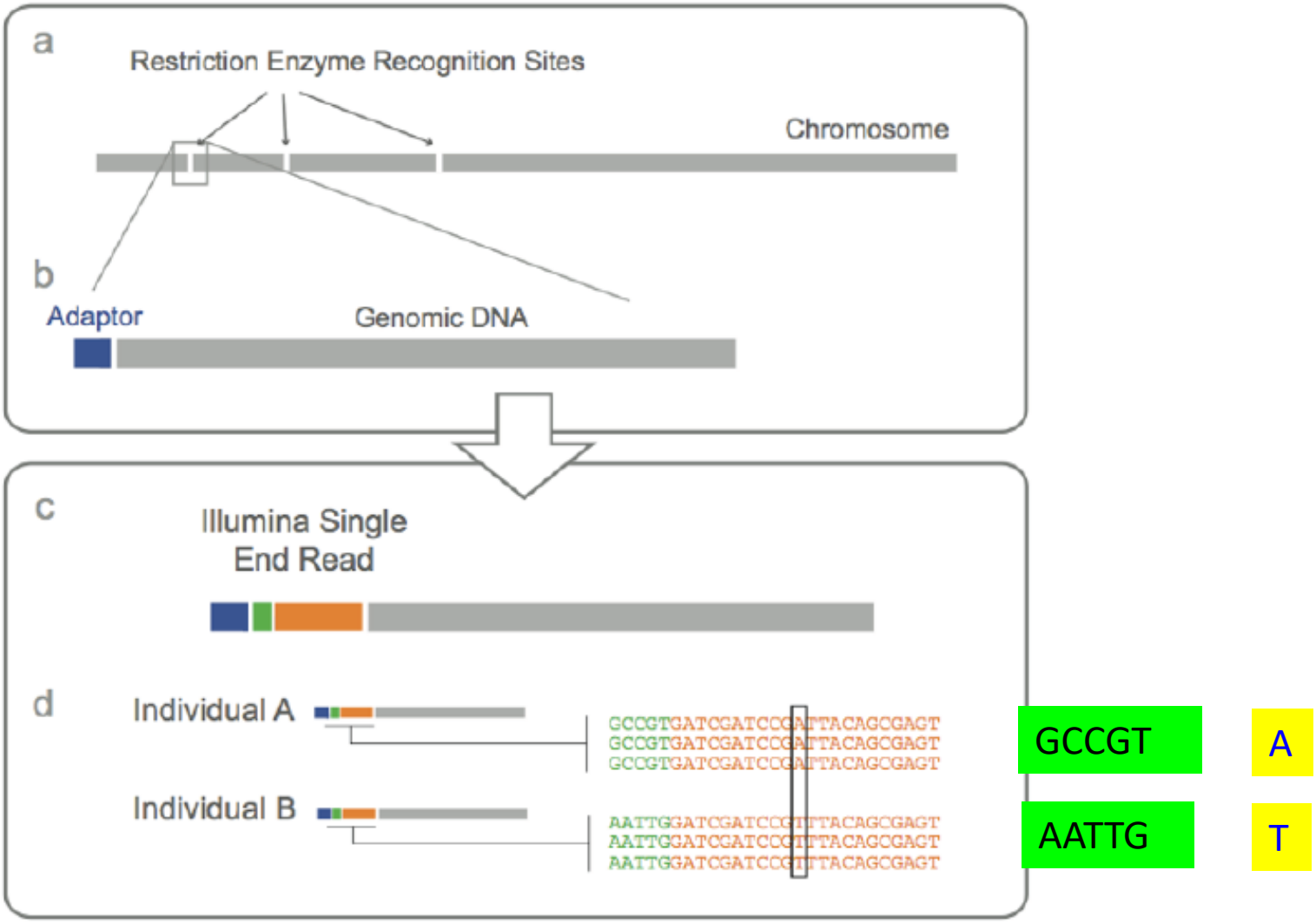
Reduced representation sequencing approaches

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

Genotyping-by-sequencing

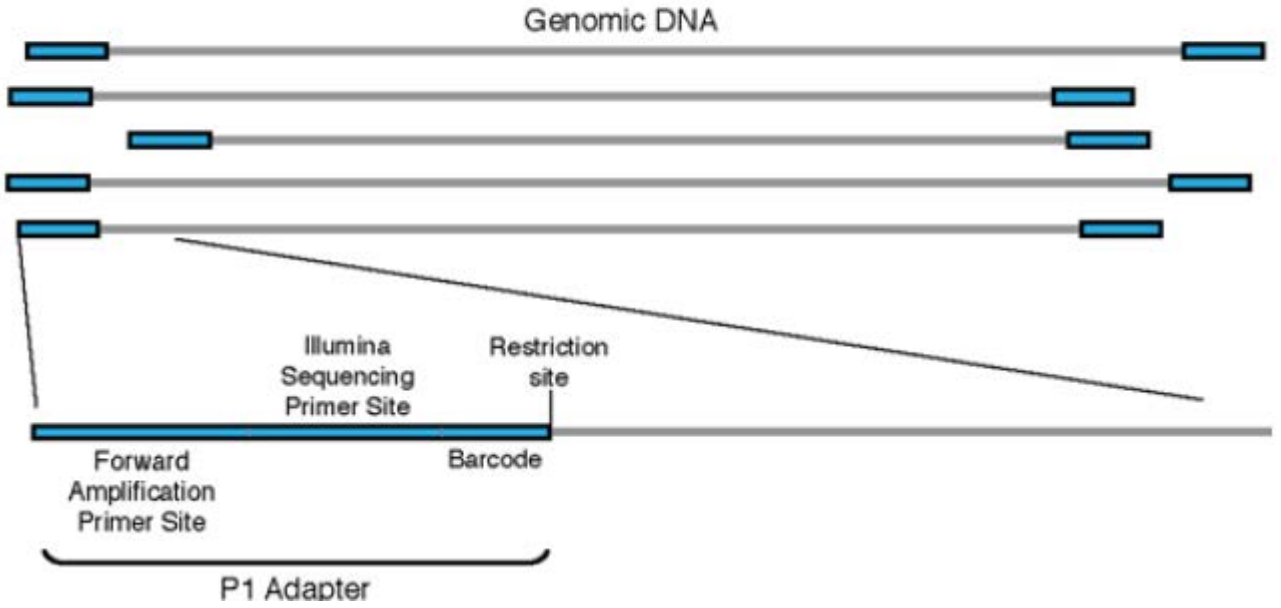
e.g., RADseq

Restriction site associated DNA (RAD) marker sequencing with barcoding

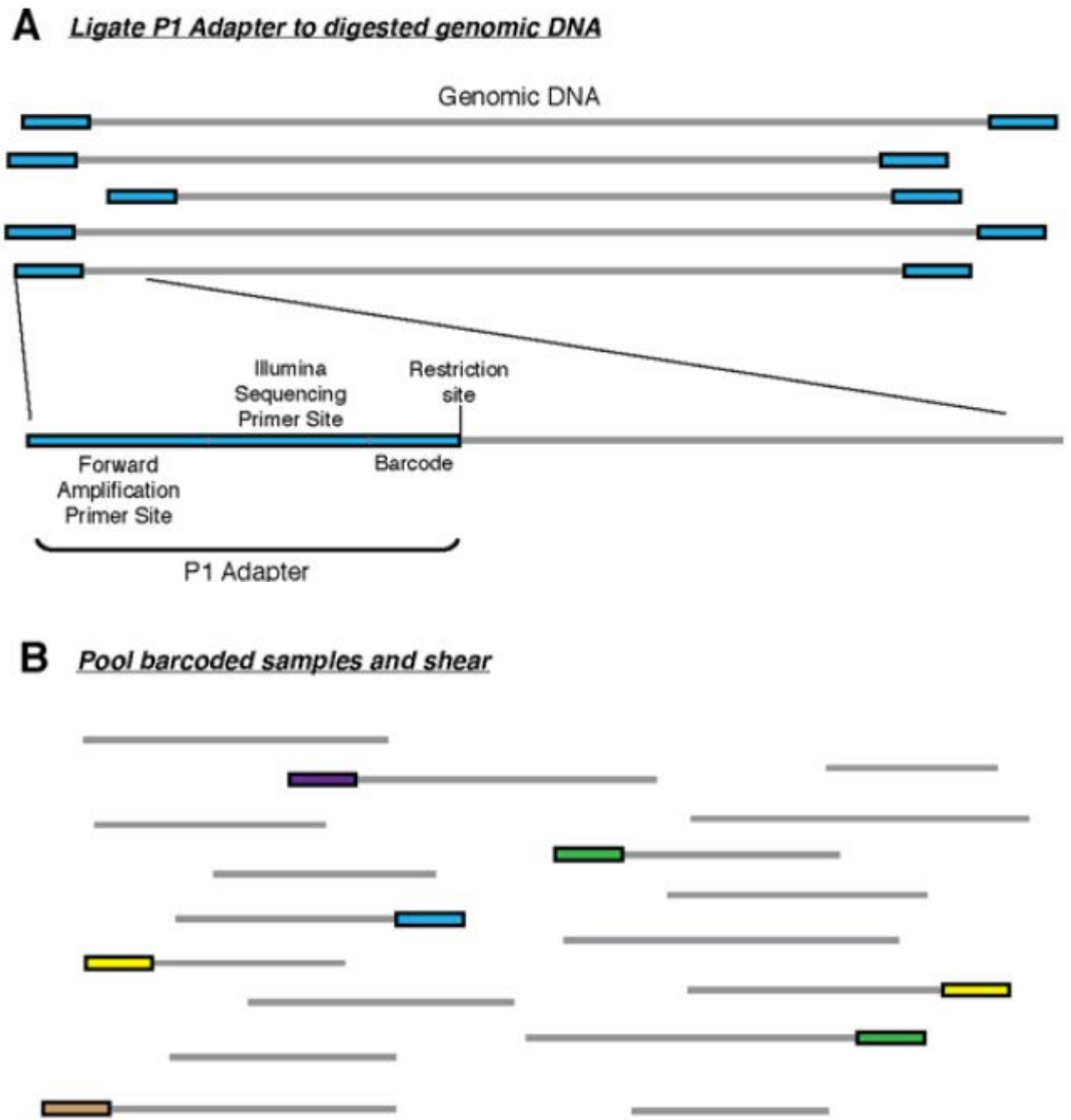


Baird et al 2008, PLoS One

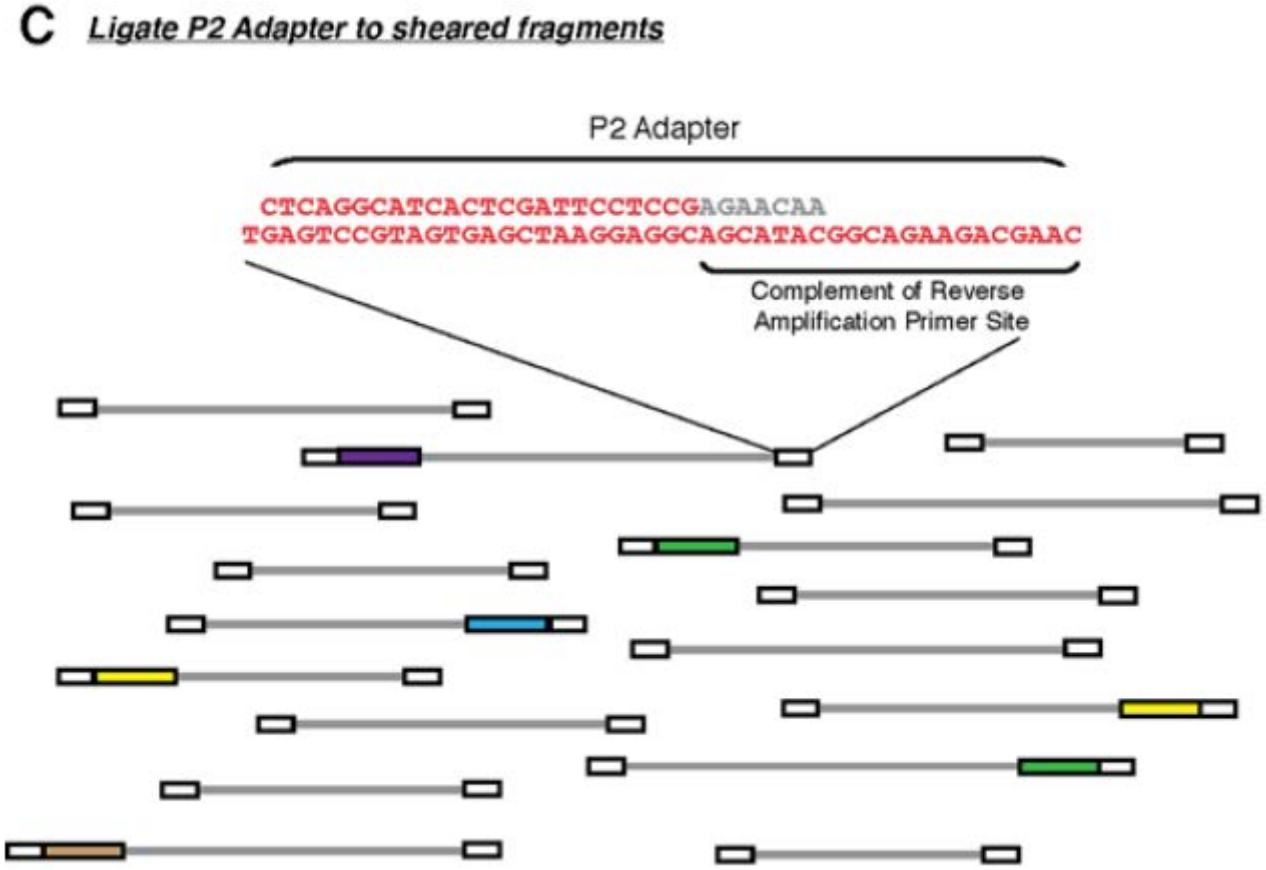
A ***Ligate P1 Adapter to digested genomic DNA***



Baird et al 2008, PLoS One

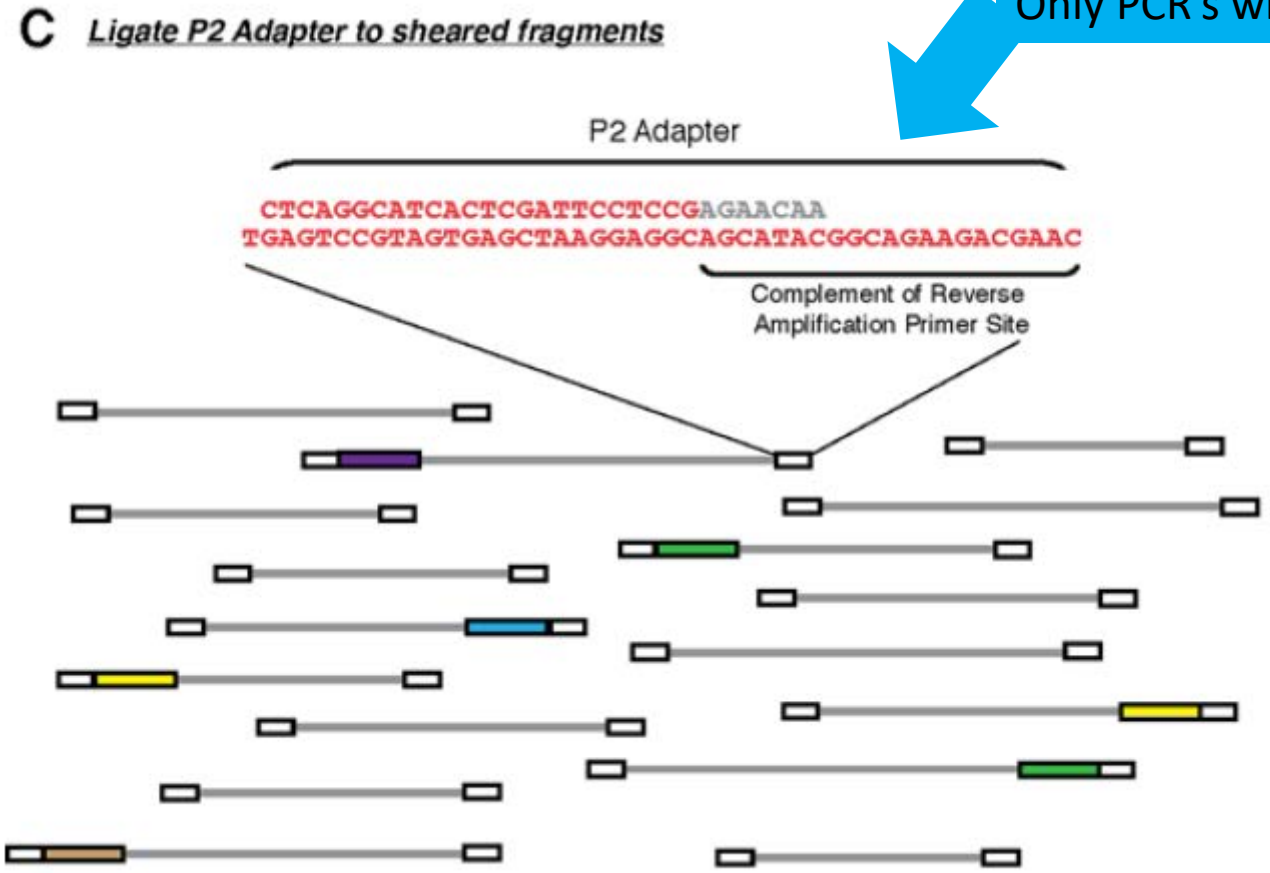


Baird et al 2008, PLoS One

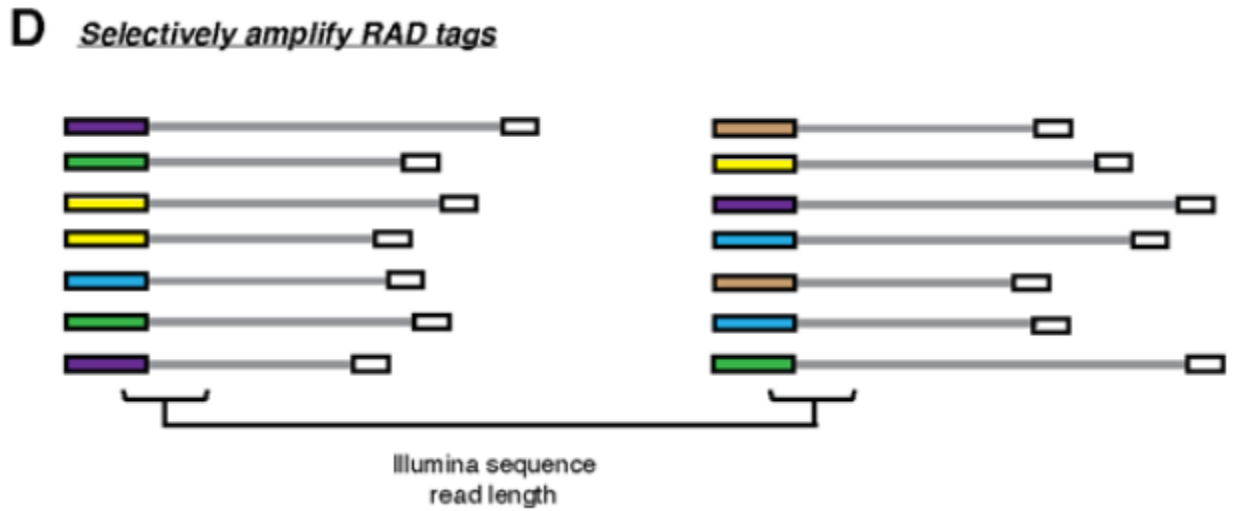


Baird et al 2008, PLoS One

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



Baird et al 2008, PLoS One



Baird et al 2008, PLoS One

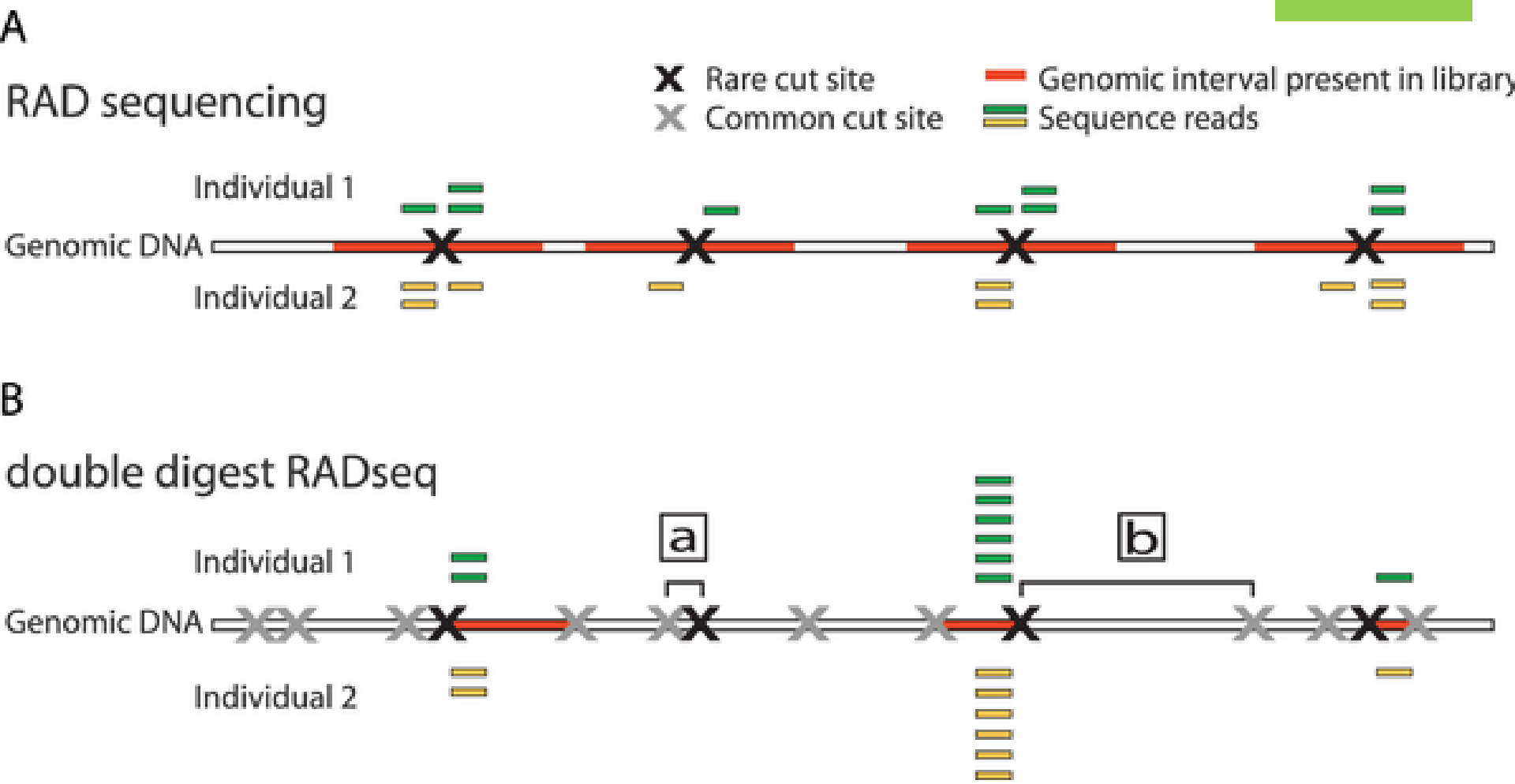
RAD markers sequence a subset of the genome

Slide courtesy of E. Johnson

Baird et al 2008, PLoS One

There are several variants of RAD-seq

ddRAD
ezRAD
2bRAD
...

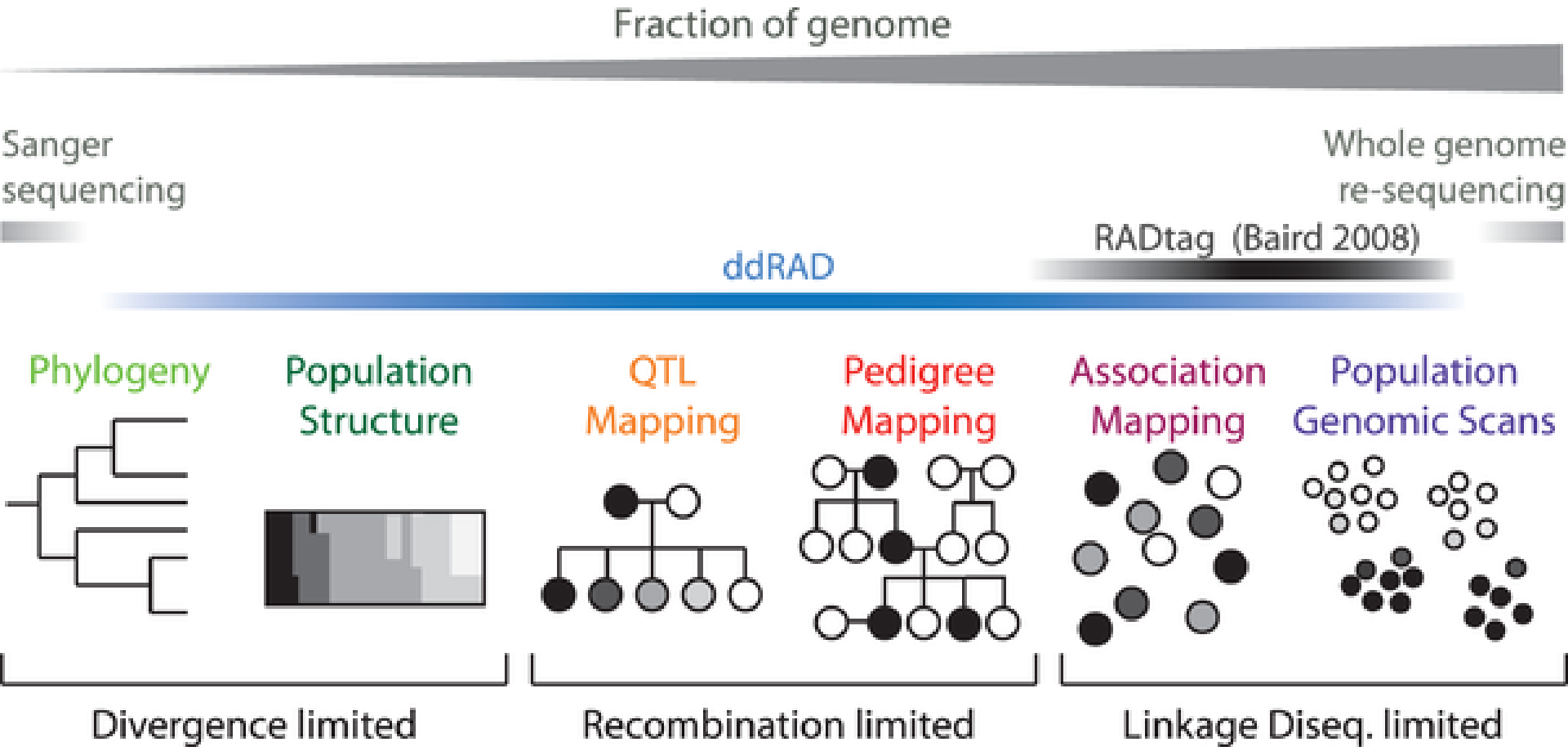


Peterson et al 2012, PLoS One

ddRAD-seq

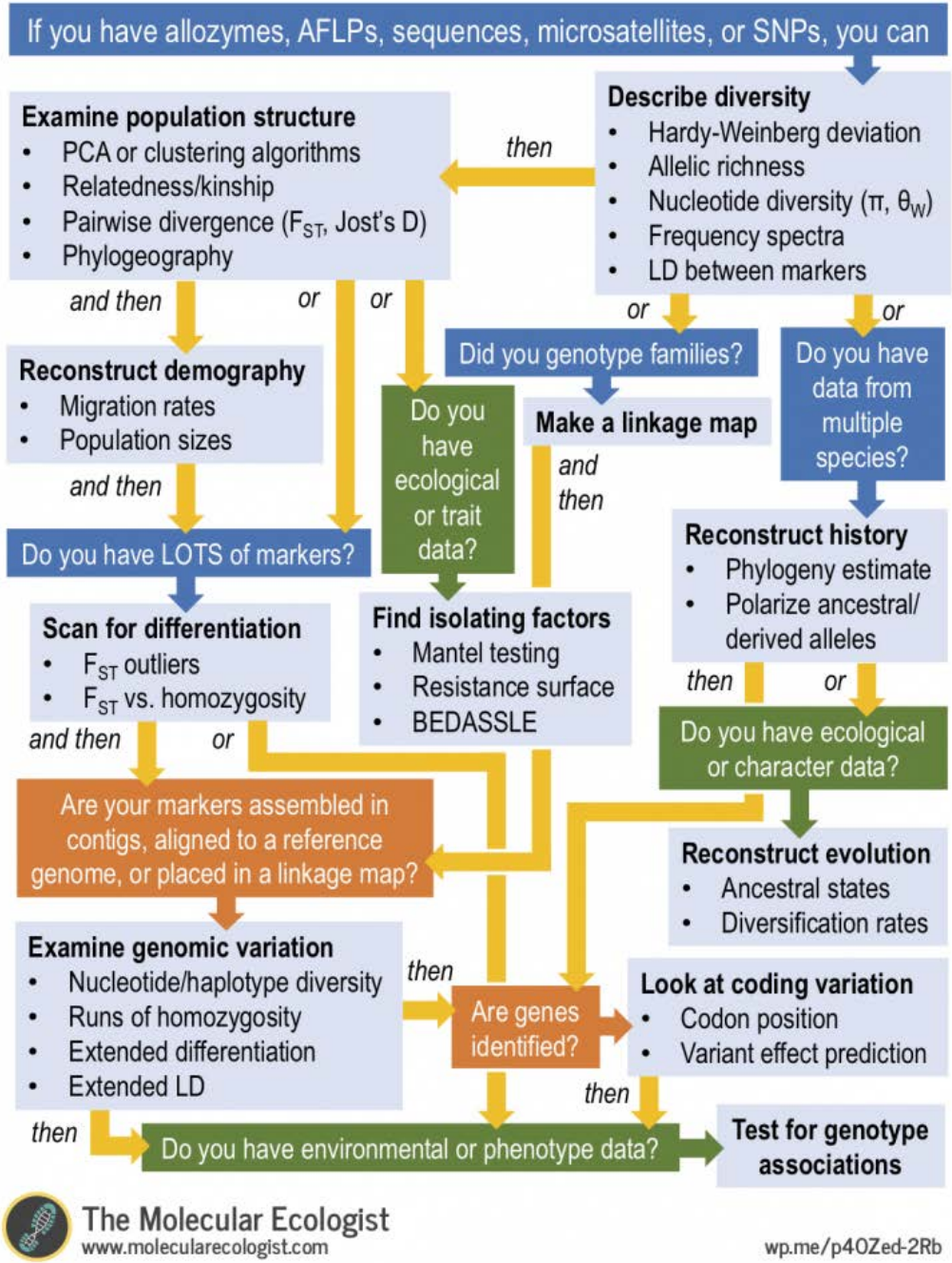
Easier tunability; better reproducibility

But PCR duplicates unknowable



Peterson et al 2012, PLoS One

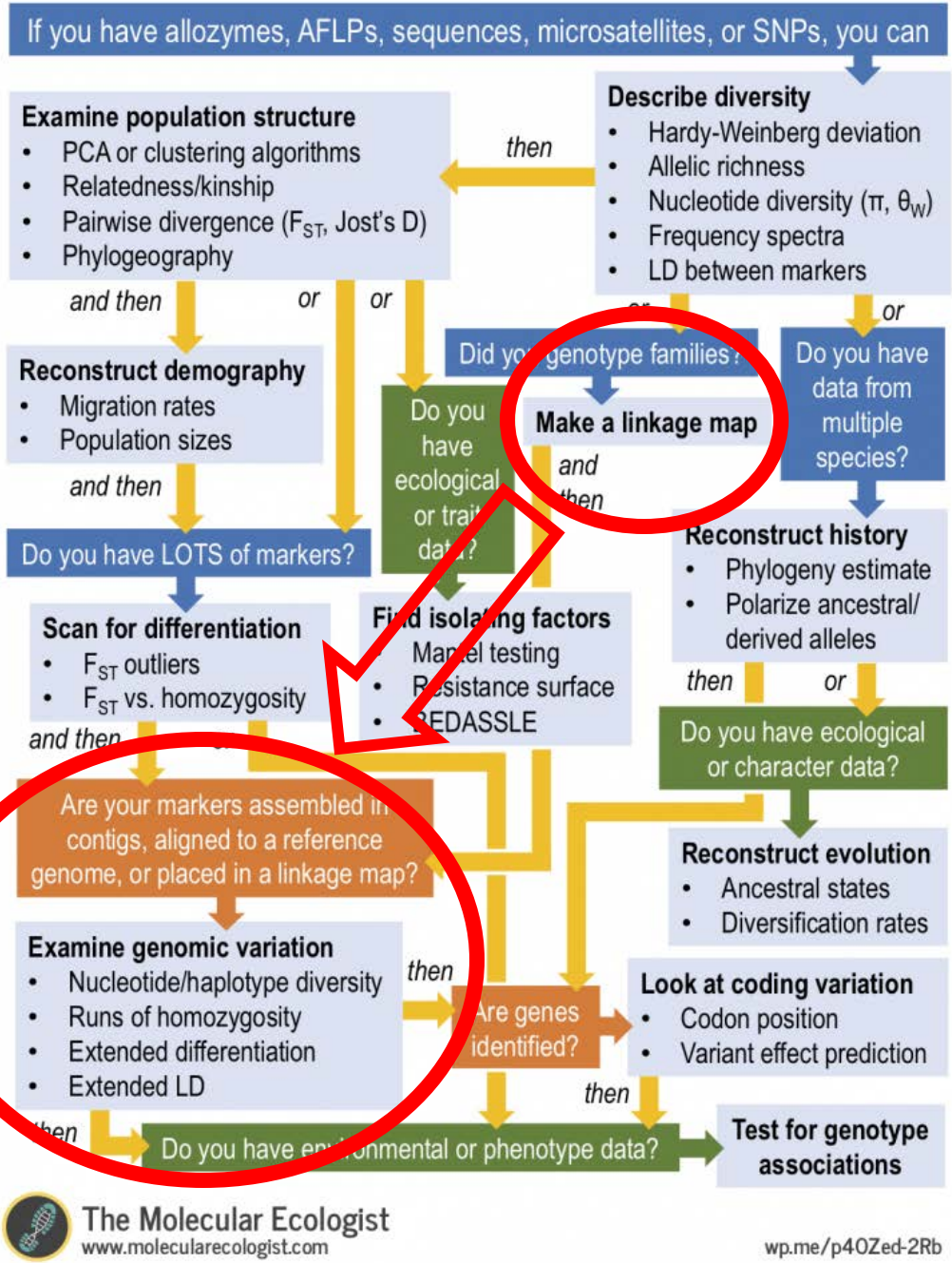
What can I do with this genetic data?



The Molecular Ecologist
www.molecularecologist.com

wp.me/p4OZed-2Rb

What can I do with this genetic data?



Our RADseq example

Monogyne

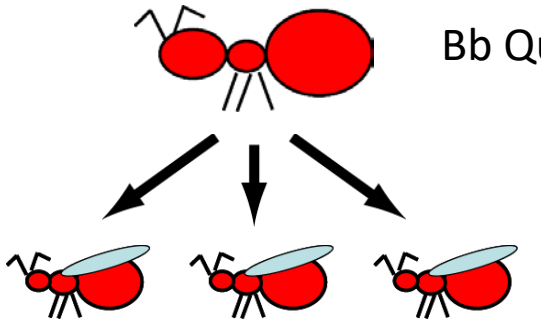
Polygyne

Fire ant *Solenopsis invicta*

Goals:

1) Build genetic map

2) test for the presence of a supergene

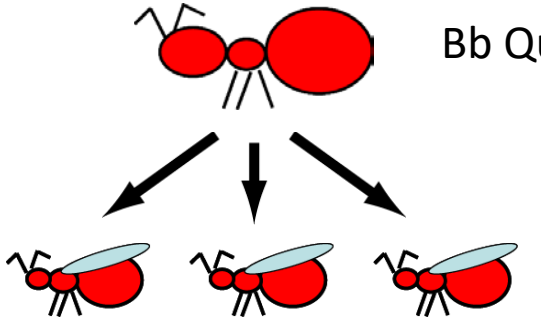


Bb Queen

88 males

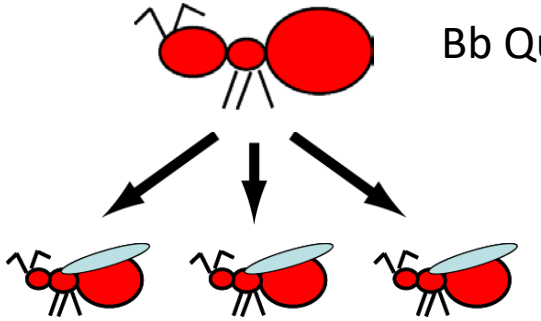
Isolate DNA, barcode each
Sequence RAD tags with Illumina

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



88 males

Isolate DNA, barcode each
Sequence RAD tags with Illumina



Bb Queen

88 males

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

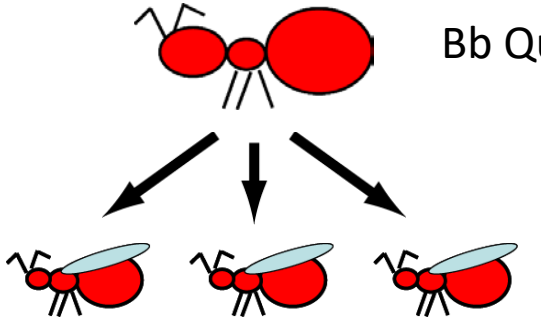
Isolate DNA, barcode each
Sequence RAD tags with Illumina

Filter

12,684 RAD tags with multiple alleles

4,232 w/ unambiguous genotype info

2,724 w/ <25% missing data



Bb Queen

88 males

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

Isolate DNA, barcode each
Sequence RAD tags with Illumina

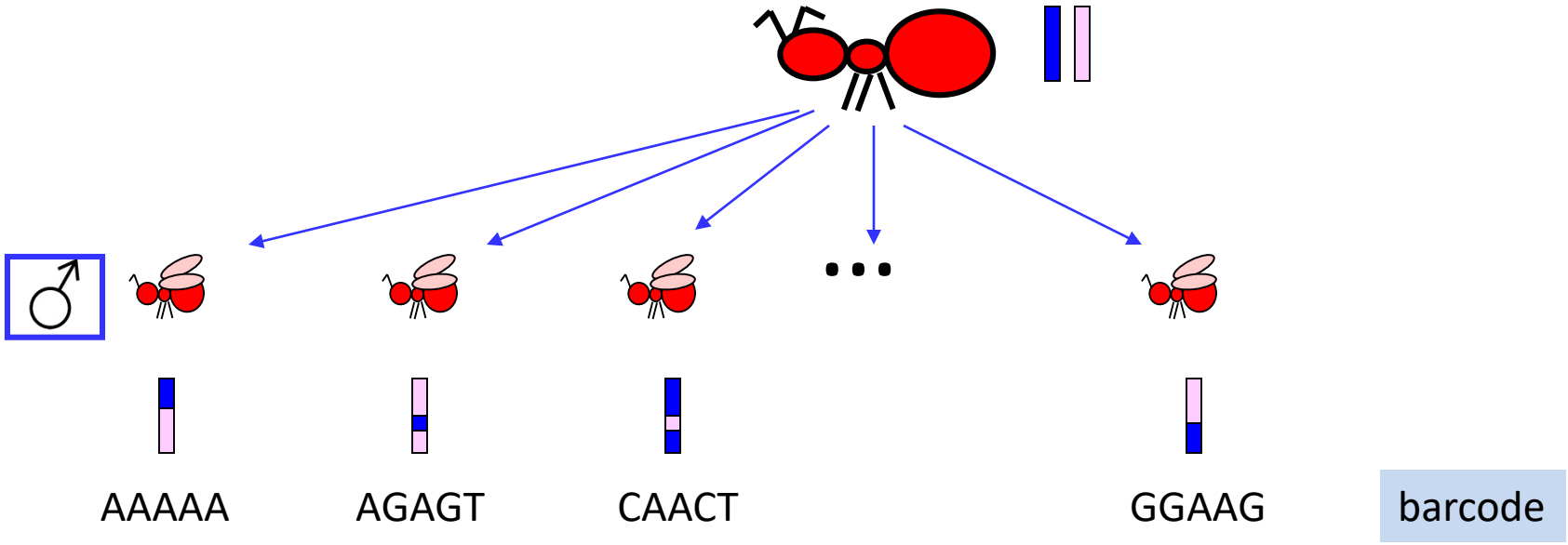
Filter

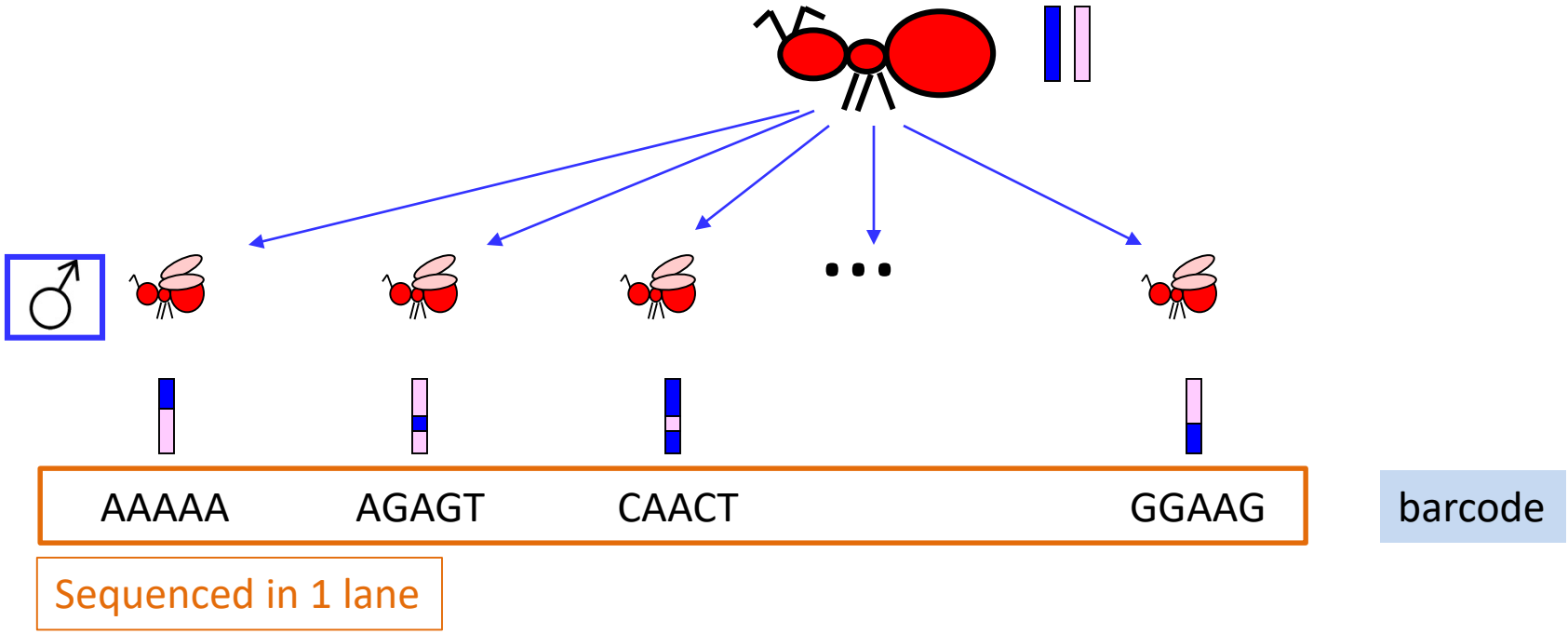
12,684 RAD tags with multiple alleles

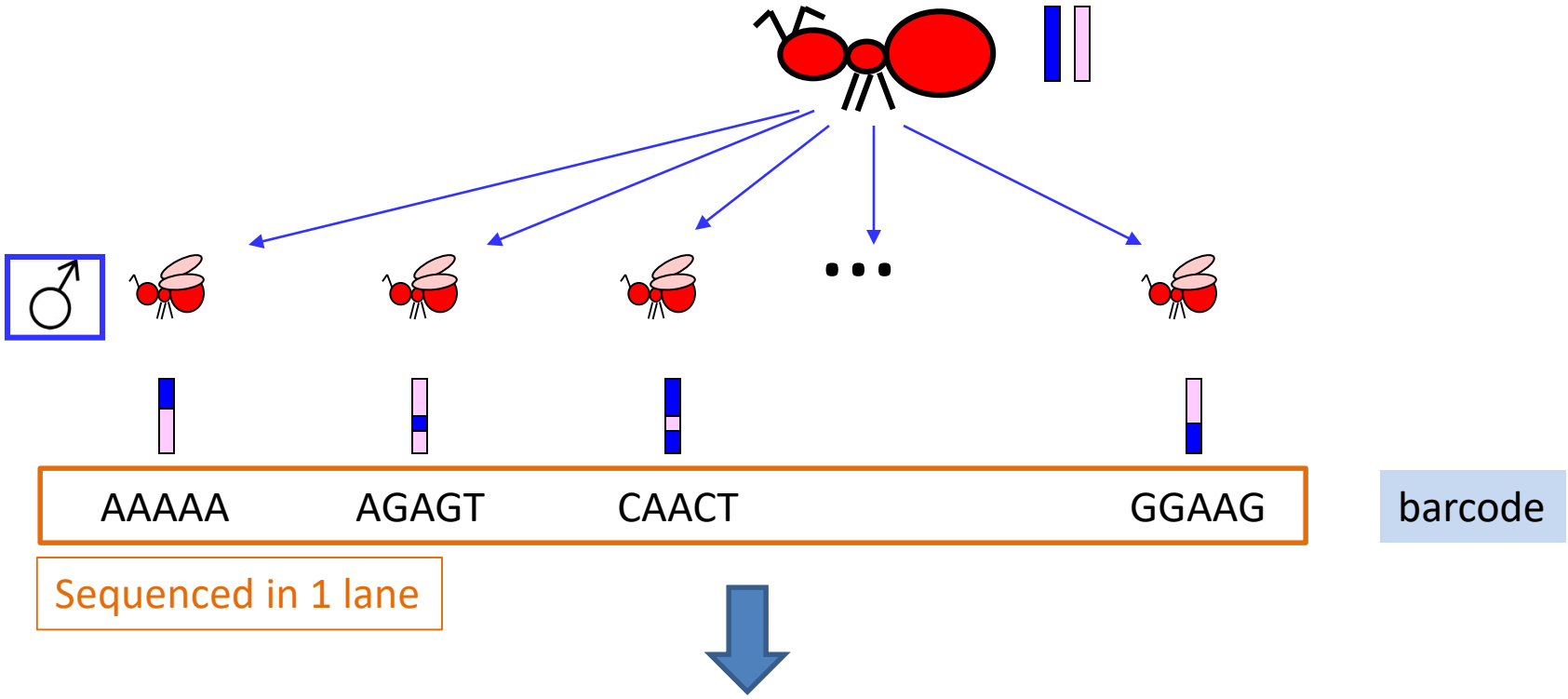
4,232 w/ unambiguous genotype info

2,724 w/ <25% missing data

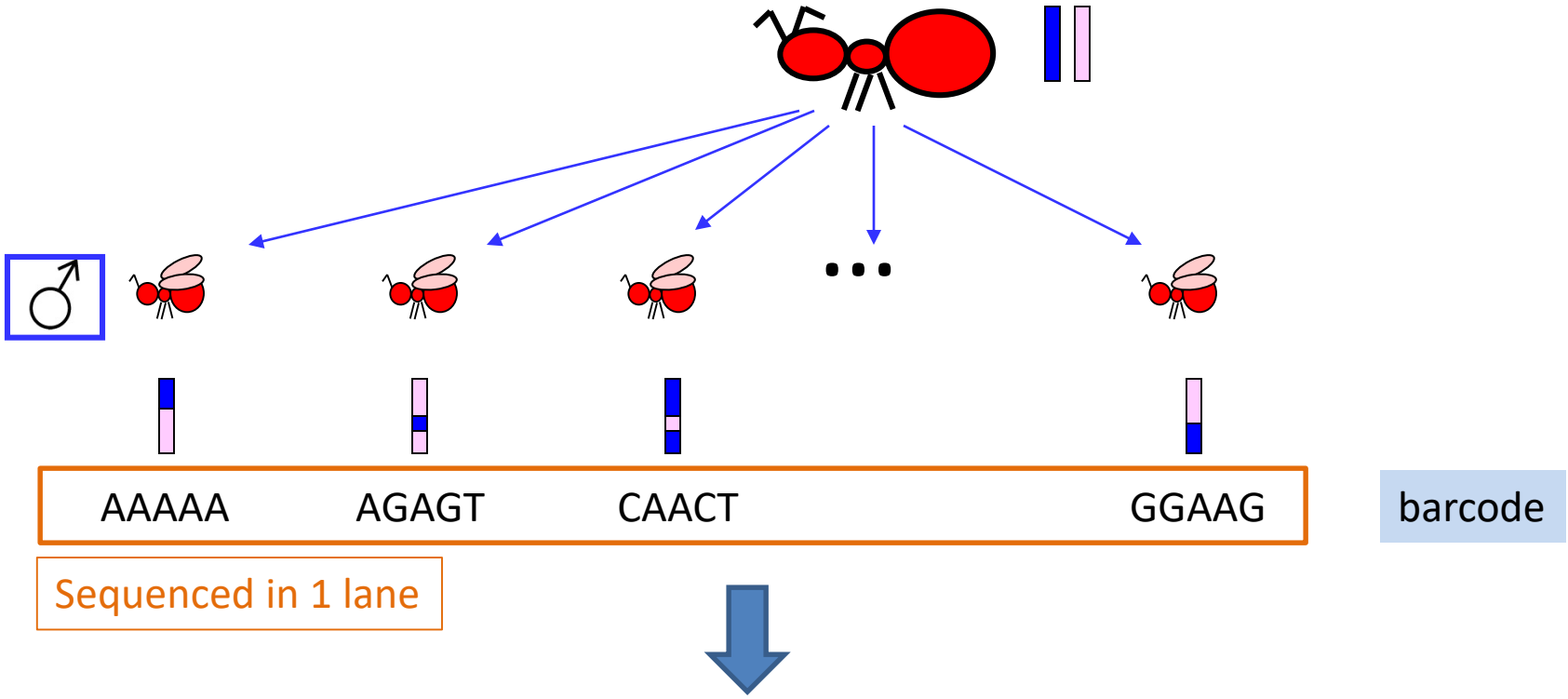
16 linkage groups
(+1 locus outlier)


AAAAA
AGAGT
CAACT
GGAAG
barcode


AAAAA
AGAGT
CAACT
GGAAG
barcode
Sequenced in 1 lane

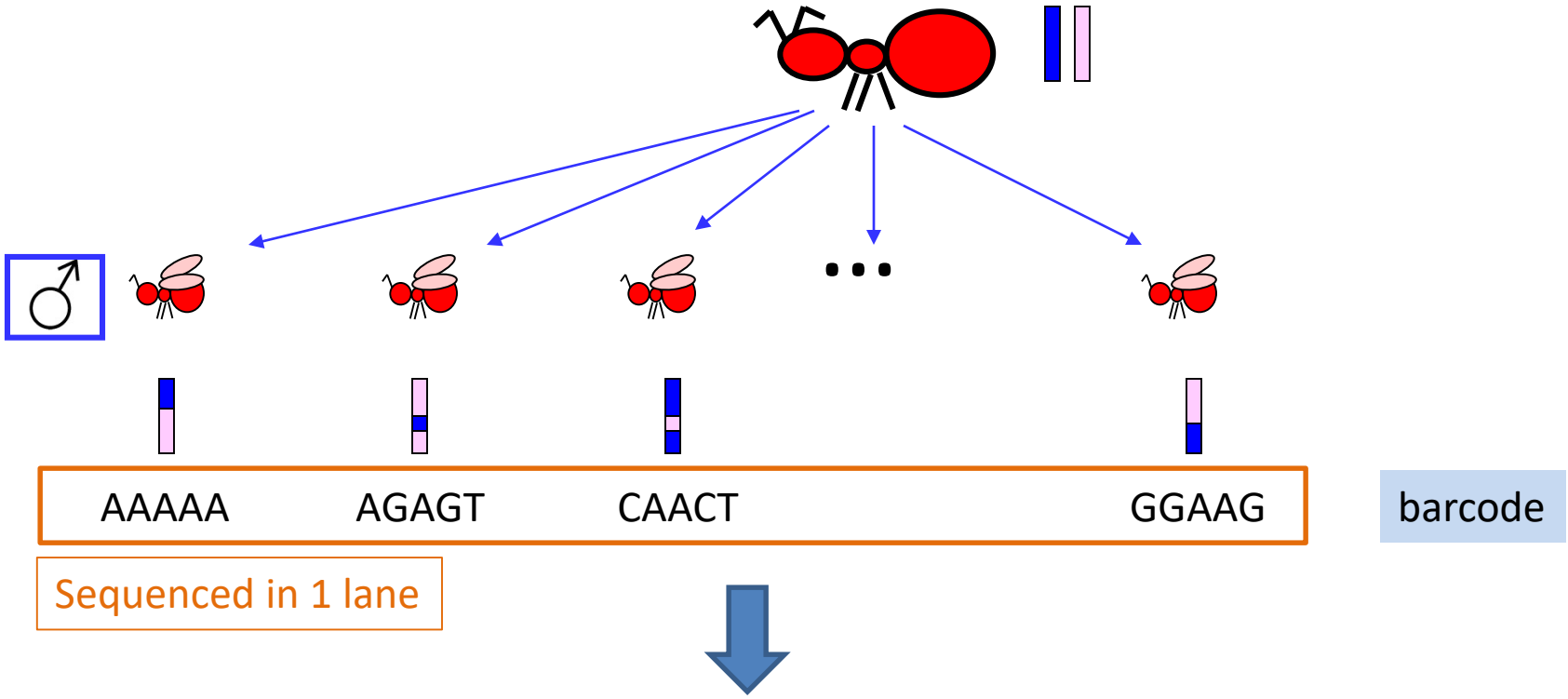


1) Split by barcode
2) Trim off barcode
3) Quality trimmer
4) Chop to 50 bp
5) Merge identical
6) Make fake FASTQ file
7) Run MAQ or bwa



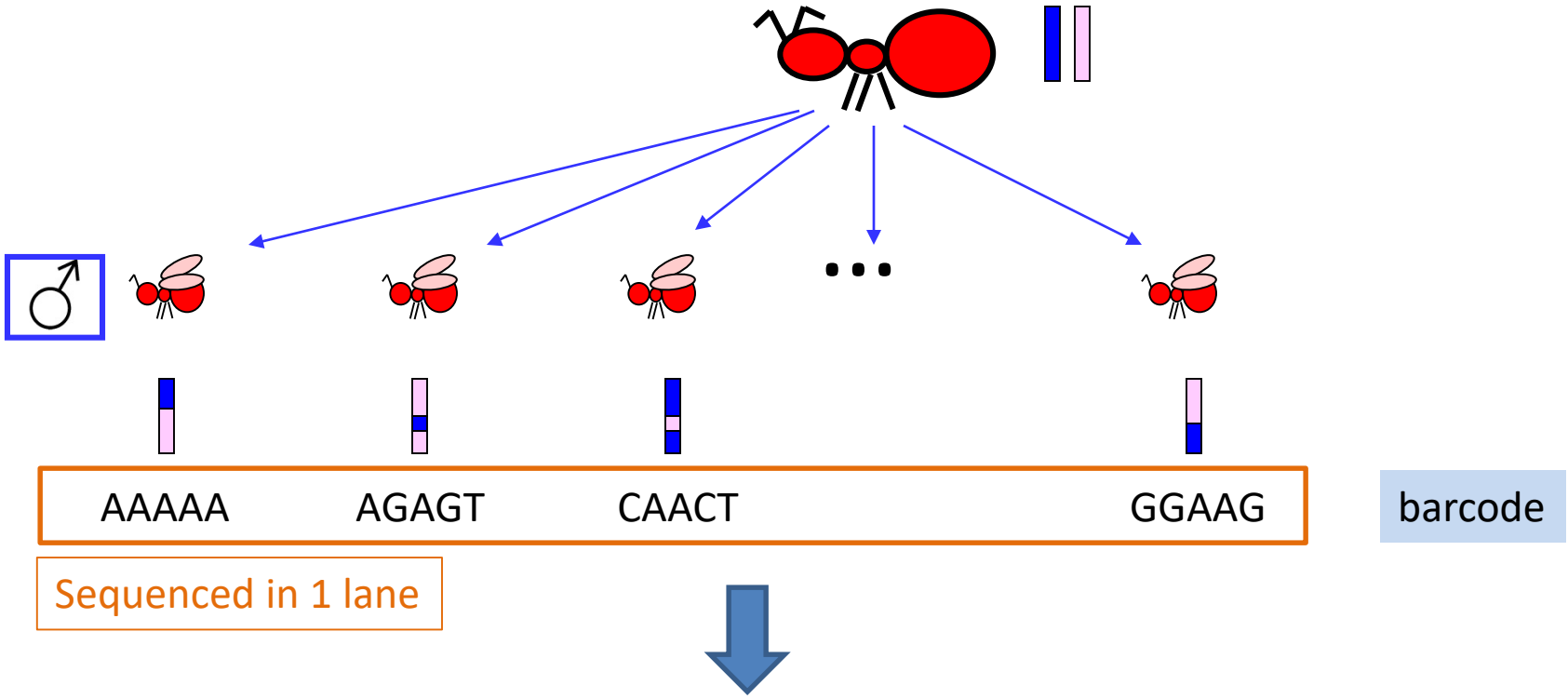
1) Split by barcode
2) Trim off barcode
3) Quality trimmer
4) Chop to 50 bp
5) Merge identical
6) Make fake FASTQ file
7) Run MAQ or bwa

GGAAGaattcacgtac … gtacgt … acgtacgt … acgt(76)



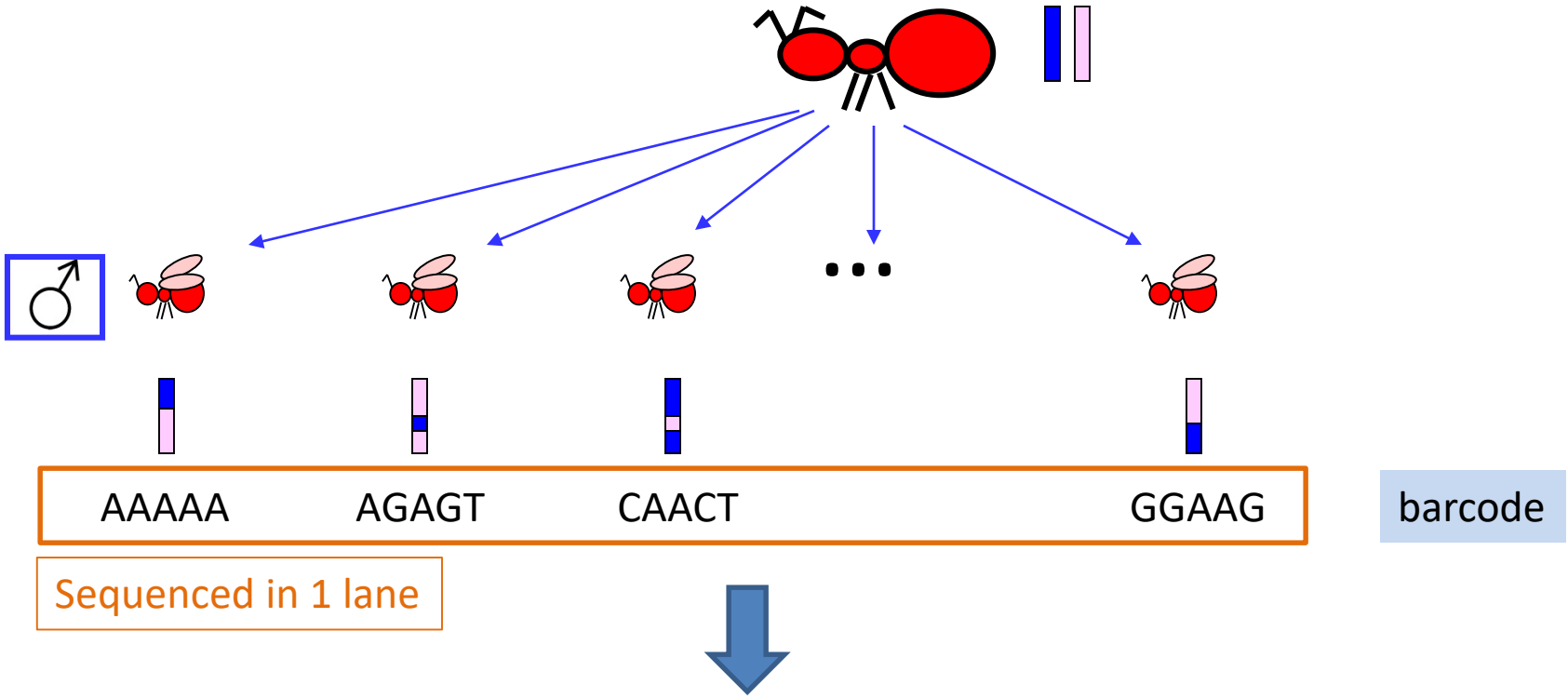
1) Split by barcode
2) Trim off barcode
3) Quality trimmer
4) Chop to 50 bp
5) Merge identical
6) Make fake FASTQ file
7) Run MAQ or bwa

GGAAGaattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt … acgt(76)



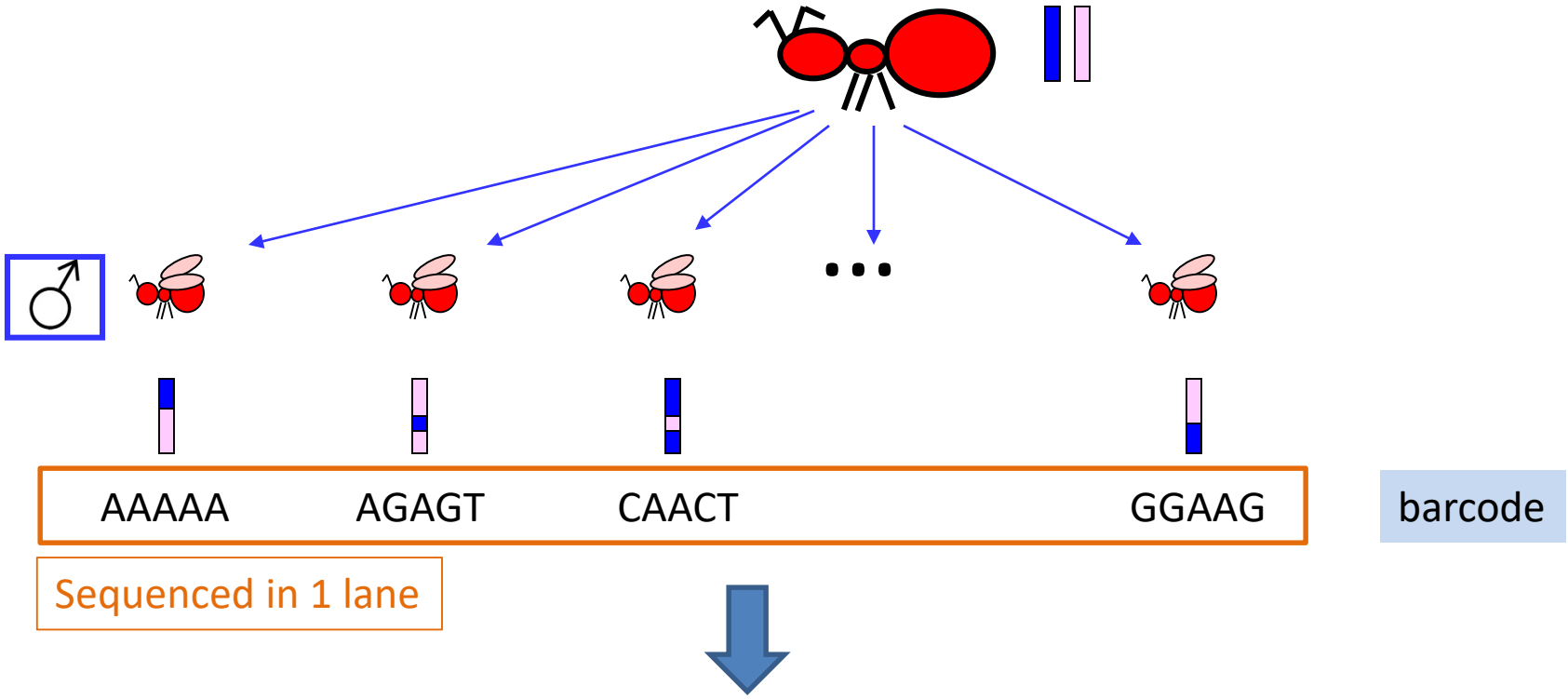
1) Split by barcode
2) Trim off barcode
3) Quality trimmer
4) Chop to 50 bp
5) Merge identical
6) Make fake FASTQ file
7) Run MAQ or bwa

```
GGAAGaattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt(63)
```



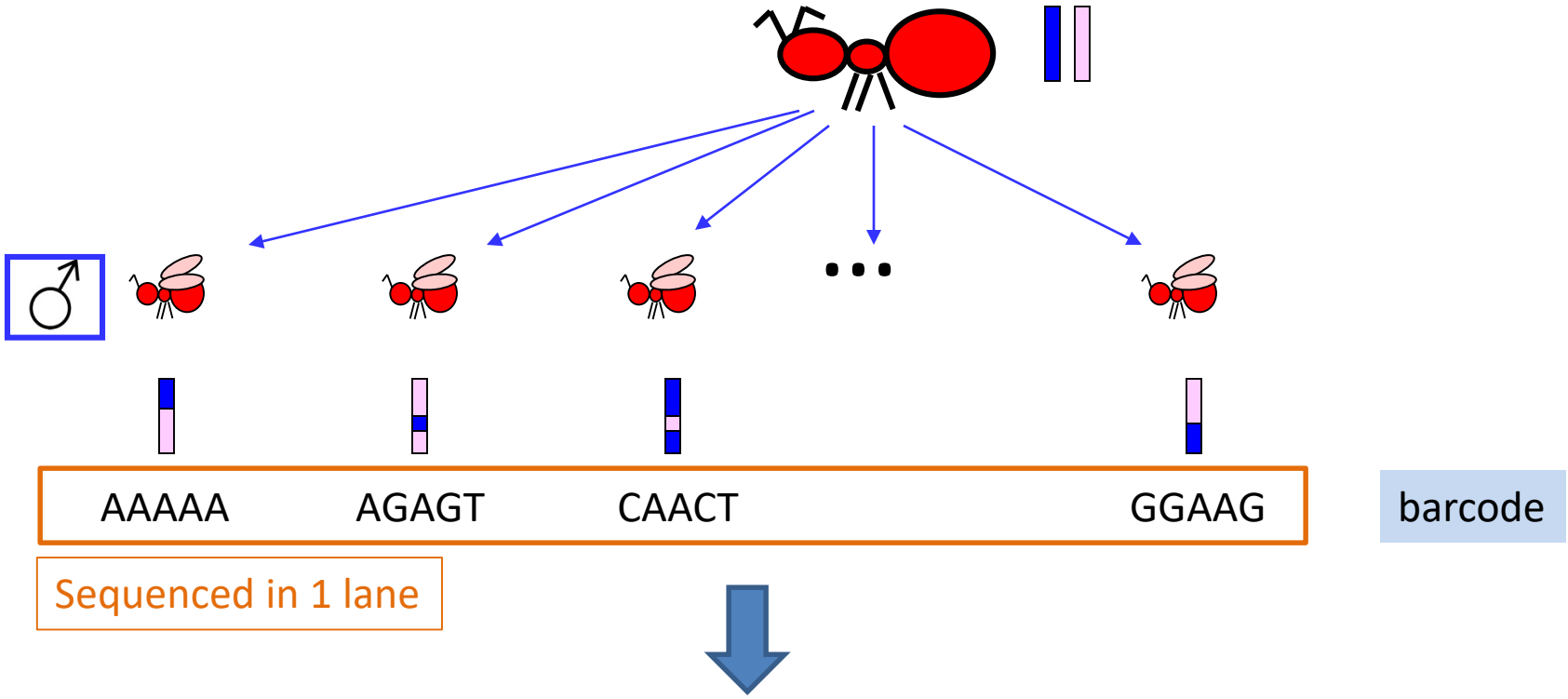
1) Split by barcode
2) Trim off barcode
3) Quality trimmer
4) Chop to 50 bp
5) Merge identical
6) Make fake FASTQ file
7) Run MAQ or bwa

```
GGAAGaattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt(63)
_____aattcacgtac … gtacgt(50)
```



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

```
GGAAGaattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt(63)
_____aattcacgtac … gtacgt(50)
_____aattcacgtac … gtacgt(50)
```



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

```
GGAAGaattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt … acgt(76)
_____aattcacgtac … gtacgt … acgtacgt(63)
_____aattcacgtac … gtacgt(50)
_____aattcacgtac … gtacgt(50)
```

8) Insert data into MySQL database (persistence)
9) New automatic scripts for filtering and QC of SNPs
10) MSTMap to determine genetic linkage

http://creskolab.uoregon.edu/stacks/

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

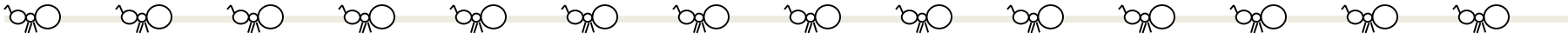
Stacks can be used to generate mappable markers from RAD-seq data. Thousands of markers can be generated from a single generation, F1 map as well as markers for traditional F2 and backcross designs. Stacks can export data to JoinMap, OneMap, or R/qtl. These data can be used for examining

Getting started with Stacks

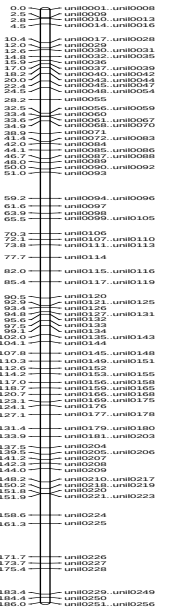
Stacks Manual

Frequently Asked Questions

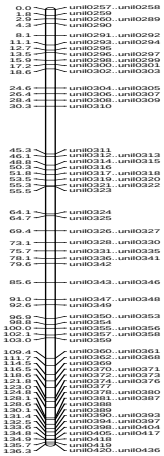
Fire ant genetic map v1



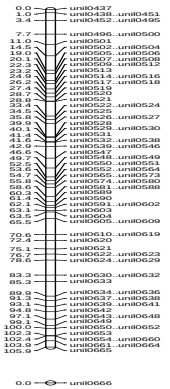
LG1



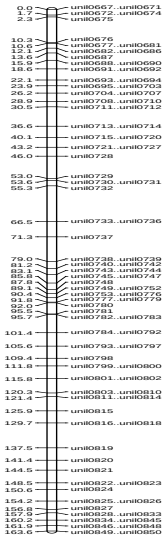
LG2



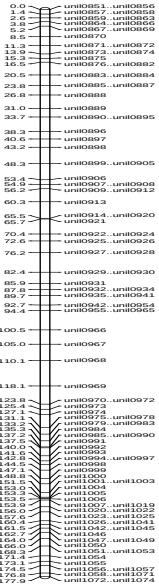
LG3



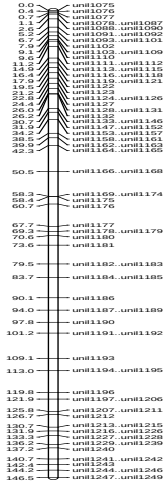
LG4



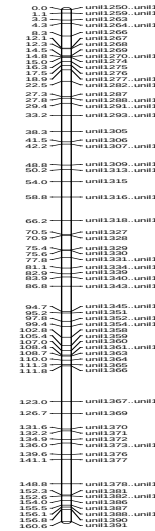
LG5



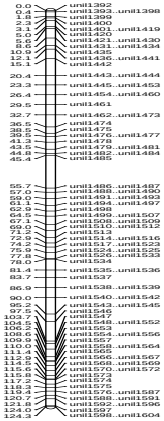
LG6



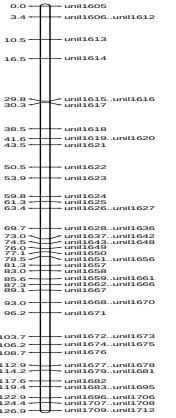
LG7



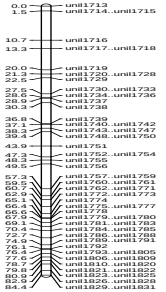
LG8



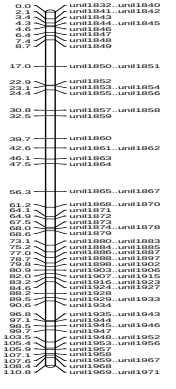
LG9



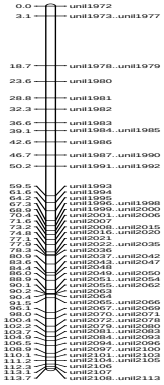
LG10



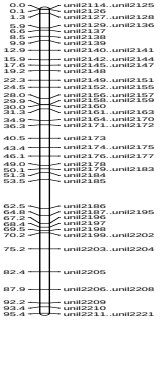
LG11



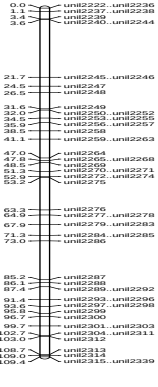
LG12



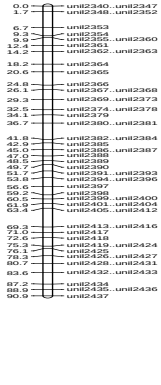
LG13



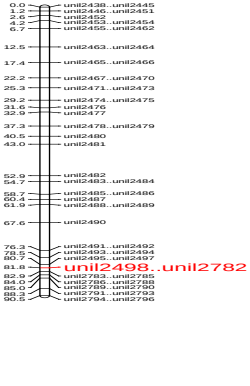
LG14



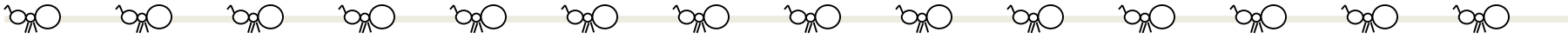
LG15



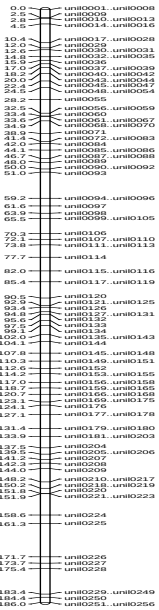
LG 16 or S



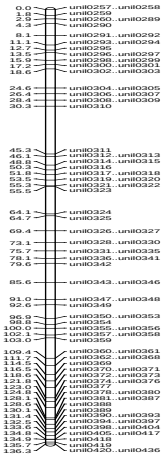
Fire ant genetic map v1



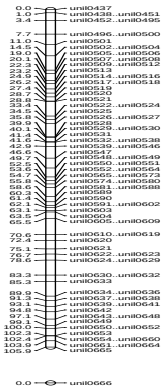
LG1



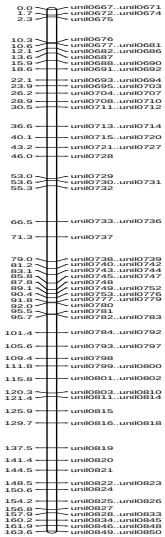
LG2



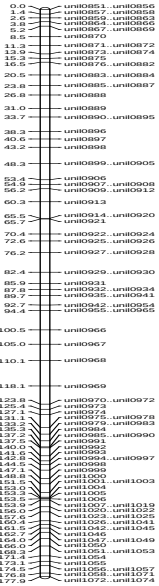
LG3



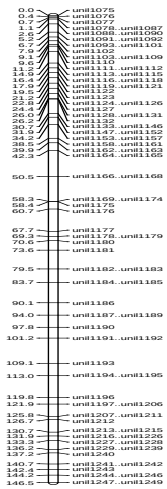
LG4



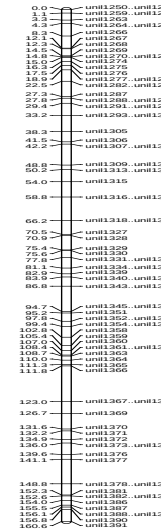
LG5



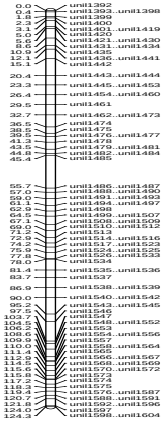
LG6



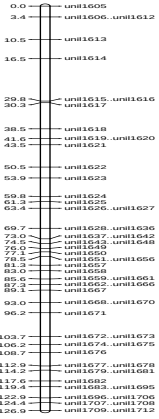
LG7



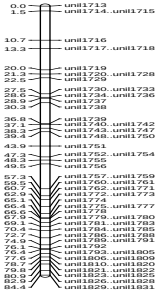
LG8



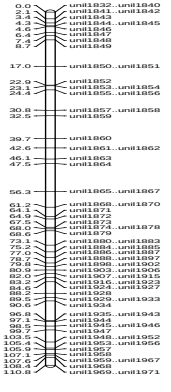
LG9



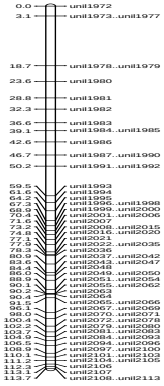
LG10



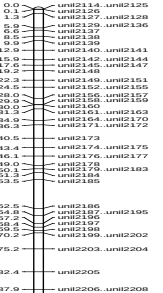
LG11



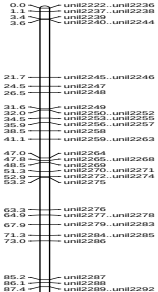
LG12



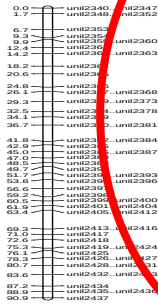
LG13



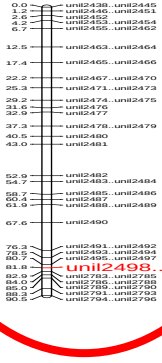
LG14



LG15

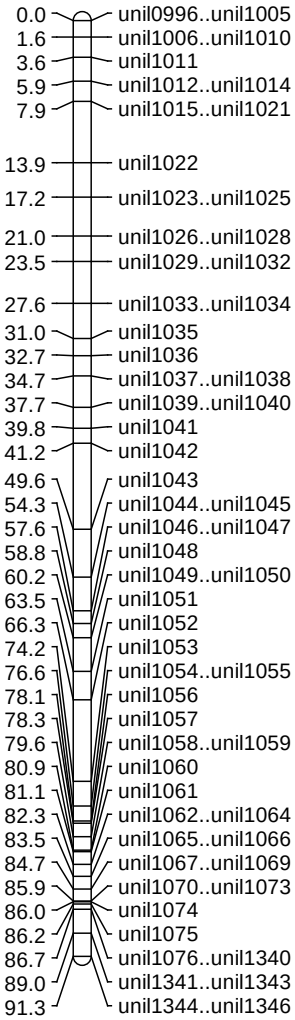


LG 16 or S



Large non-combining region around *Gp-9*

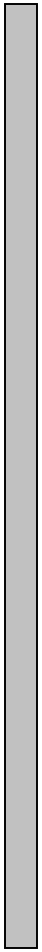
LG 16 = SB or Sb



Genetic map

Physical map

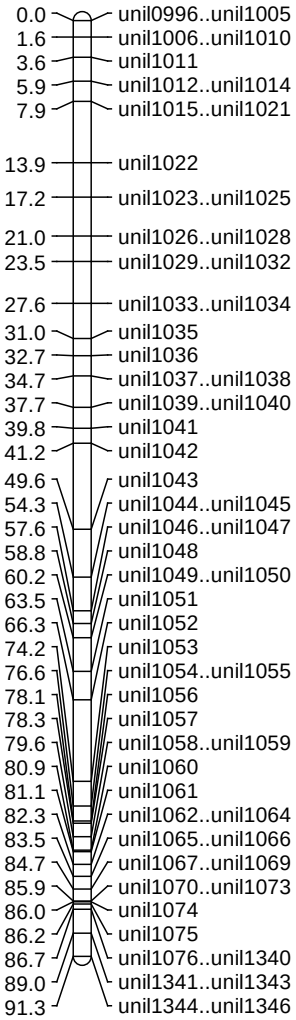
LG 16 (~23 Mbp)



Large non-combining region around *Gp-9*

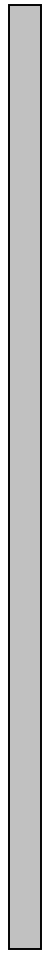
LG 16 = SB or Sb

LG 16 (~23 Mbp)



Genetic map

Physical map



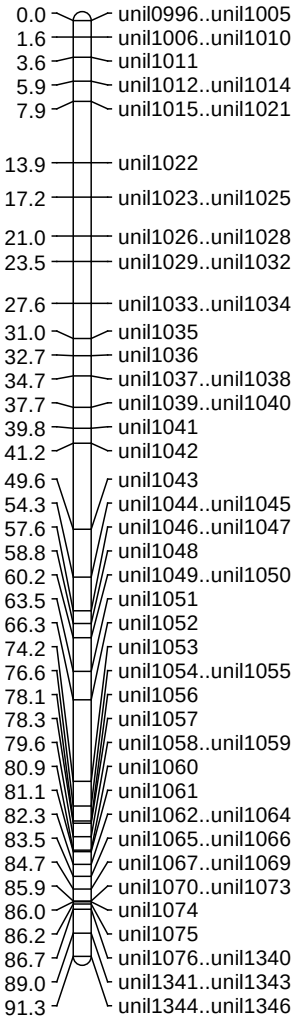
Gp-9

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

Large non-combining region around *Gp-9*

LG 16 = SB or Sb

LG 16 (~23 Mbp)



Genetic map

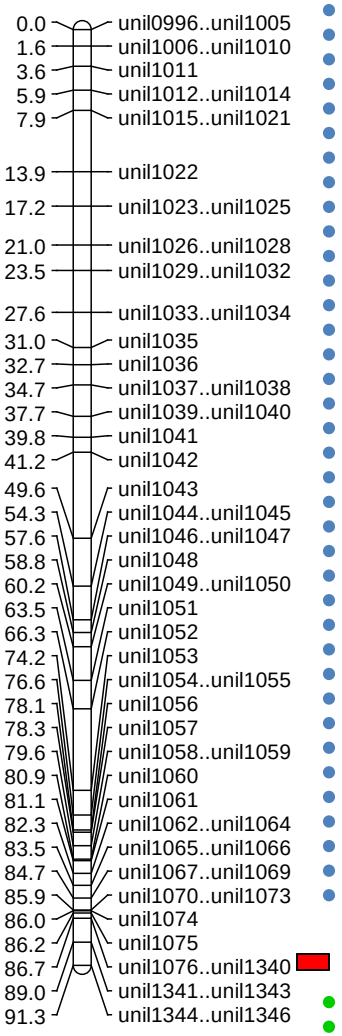
Physical map

Gp-9
- 265 markers non-recombining
- 44 scaffolds
- ~12.7 Mbp

12.7 Mbp
(55%)

Large non-combining region around *Gp-9*

LG 16 = SB or Sb



LG 16 (~23 Mbp)

Genetic map *Physical map*

8 Mbp

12.7 Mbp
(55%)

2.3 Mbp

Gp-9
- 265 markers non-recombining
- 44 scaffolds
- ~12.7 Mbp

2nd ant also has supergene

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

Purcell et al 2014

2nd ant also has supergene

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

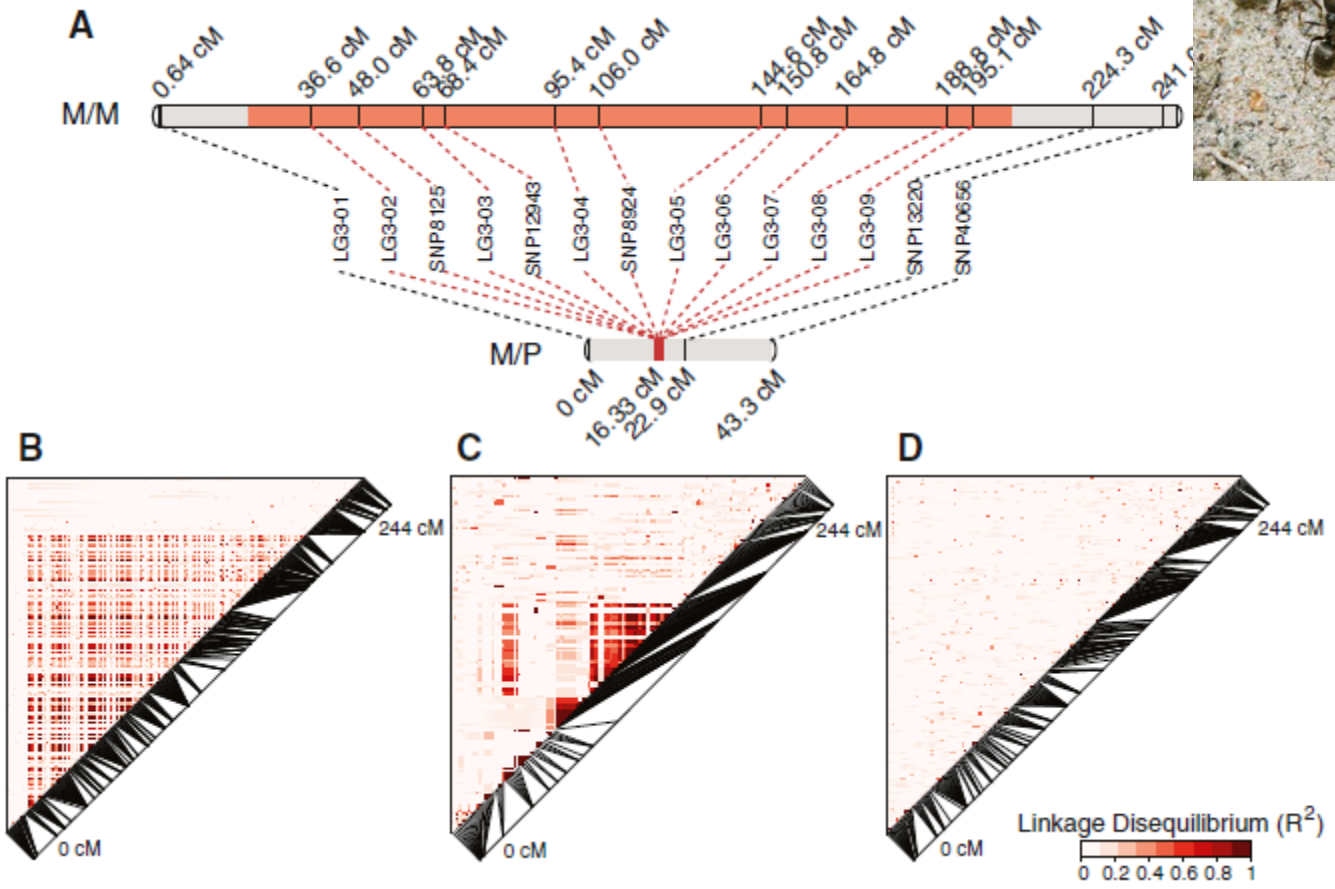
Formica selysi
Also monogyne and polygyne social forms

Purcell et al: GBS
Linkage map

Then population LD verified supergene

Purcell et al 2014

Linkage disequilibrium → supergene



Purcell et al 2014

RAD-seq in phylogeography

Wyeomyia smithii mosquito

larvae live in

Sarracenia purpurea carnivorous pitcher plant

photo: Leonora bug level

Emerson et al 2010

RAD-seq in phylogeography

Wyeomyia smithii mosquito

larvae live in

Sarracenia purpurea carnivorous pitcher plant

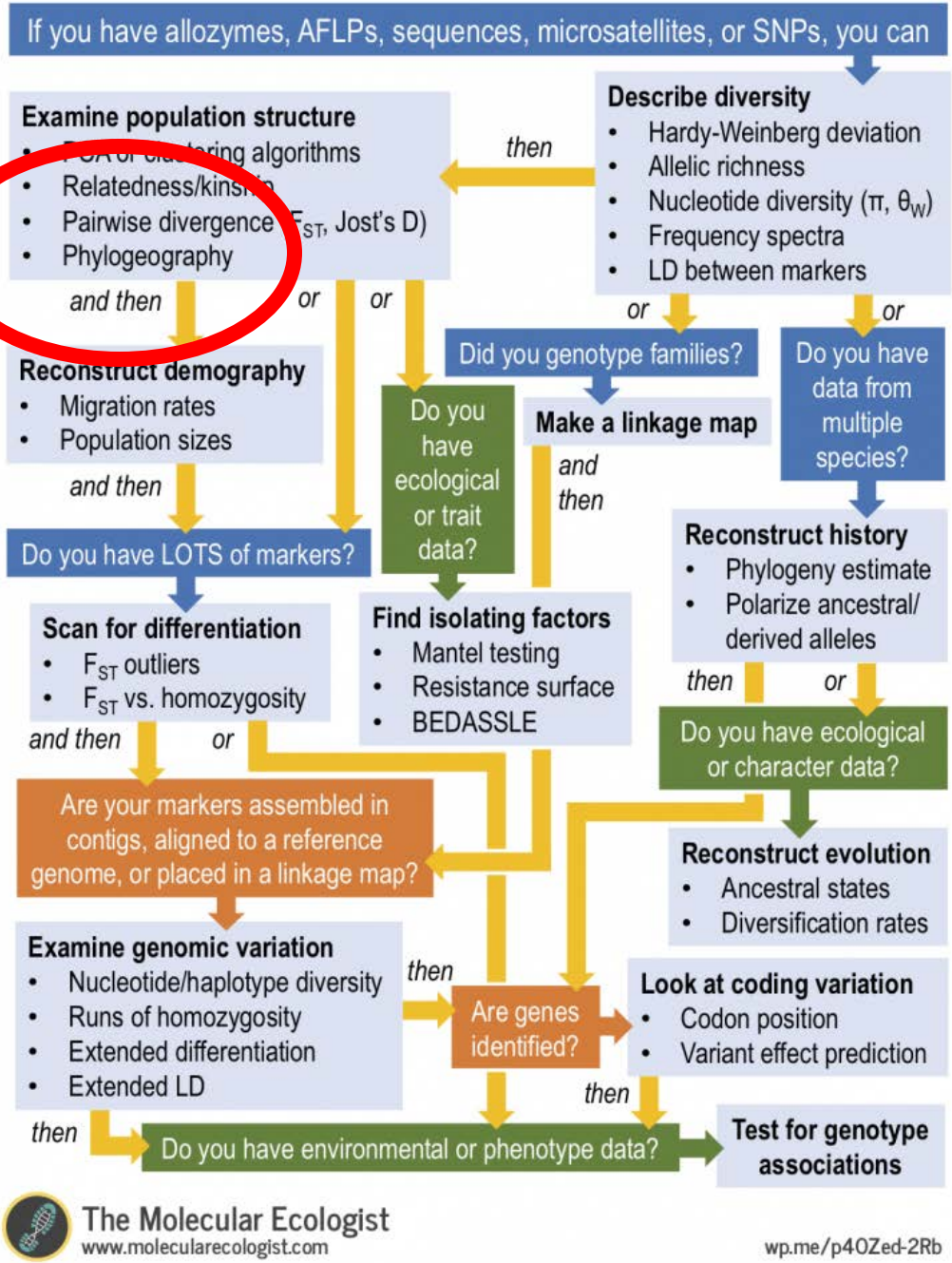
photo: Leonora bug level

What is their population structure in the Eastern USA?

Is there an expansion pattern associated with glacial retreat?

Emerson et al 2010

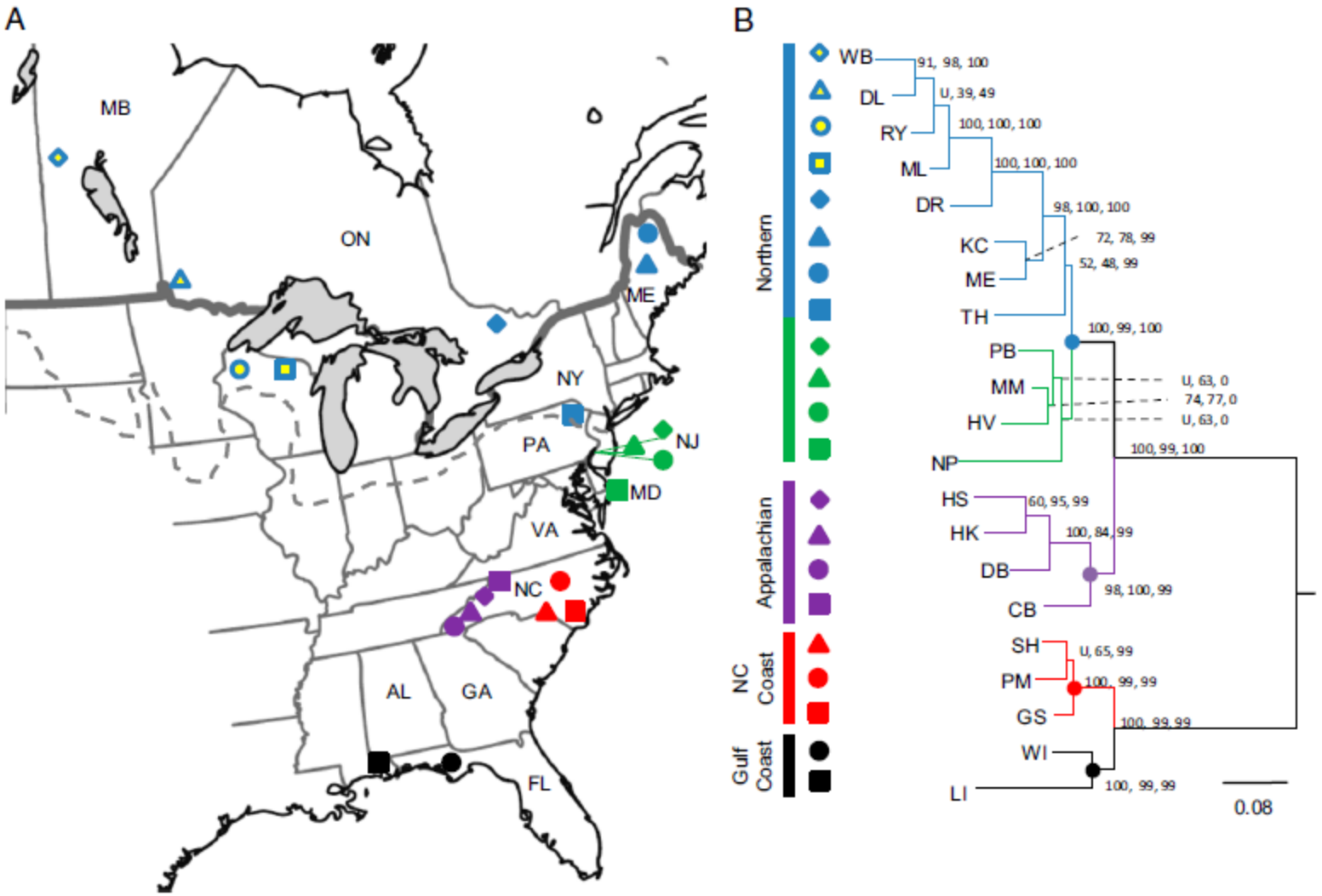
What can I do with this genetic data?



The Molecular Ecologist
www.molecularecologist.com

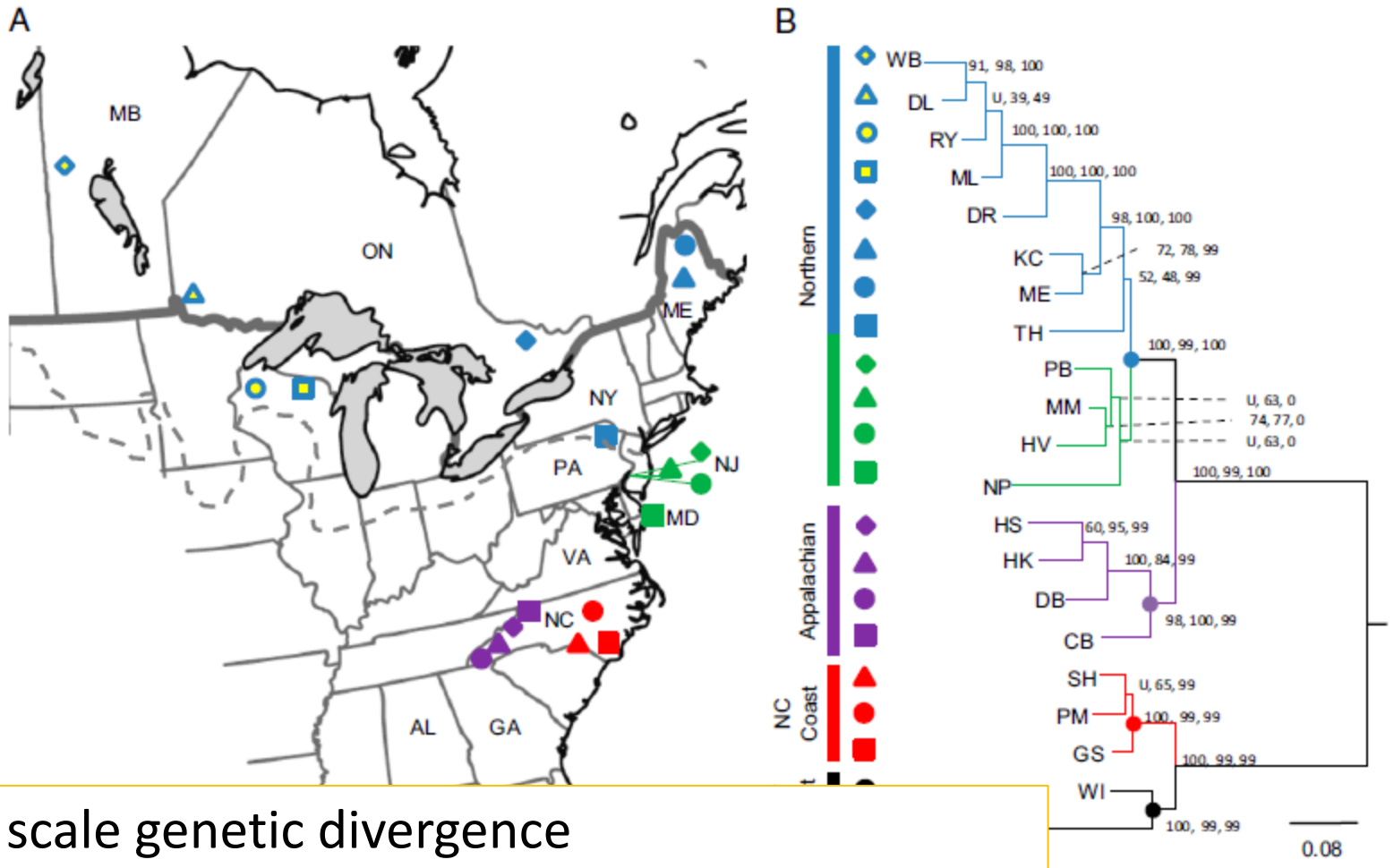
wp.me/p4OZed-2Rb

W. smithii samples and phylogenetic tree from RAD



Emerson et al 2010

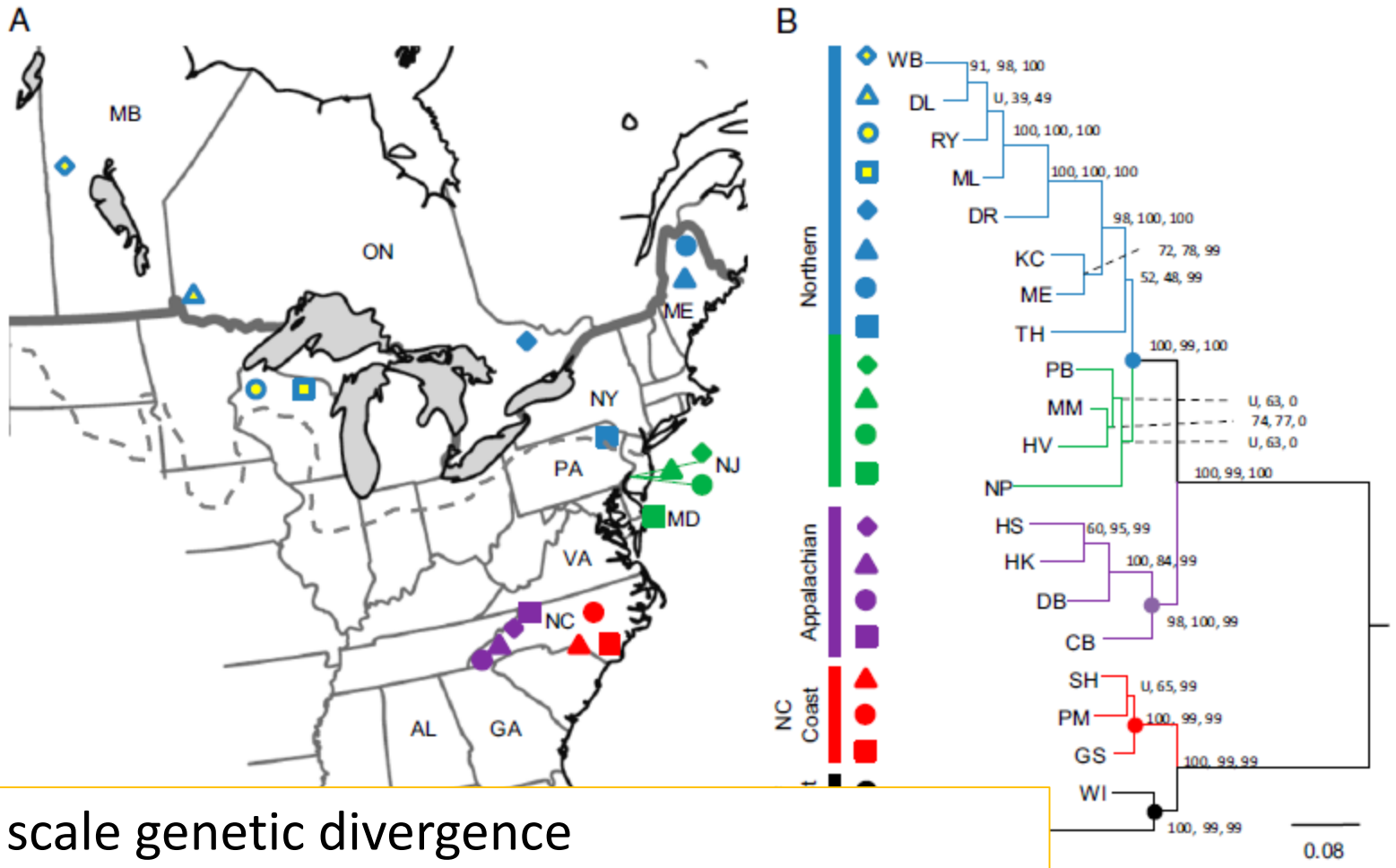
W. smithii samples and phylogenetic tree from RAD



Fine scale genetic divergence
Evidence for glacial retreat associated expansion

Emerson et al 2010

W. smithii samples and phylogenetic tree from RAD



Fine scale genetic divergence
Evidence for glacial retreat associated expansion

Future – local photoperiodic response adaptations

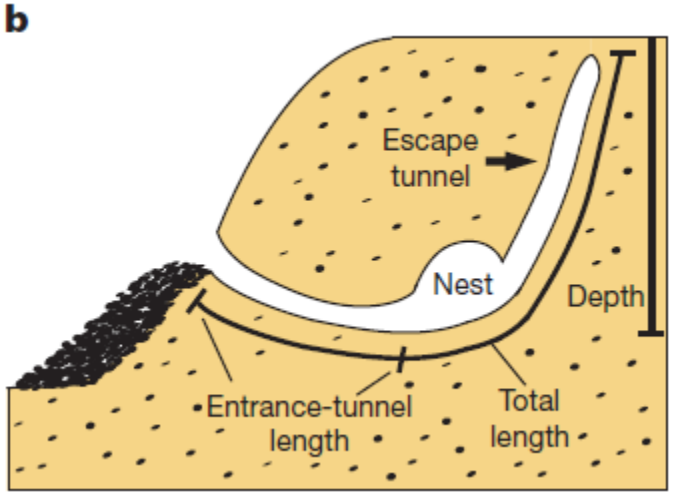
Emerson et al 2010

402 | NATURE | VOL 493 | 17 JANUARY 2013

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

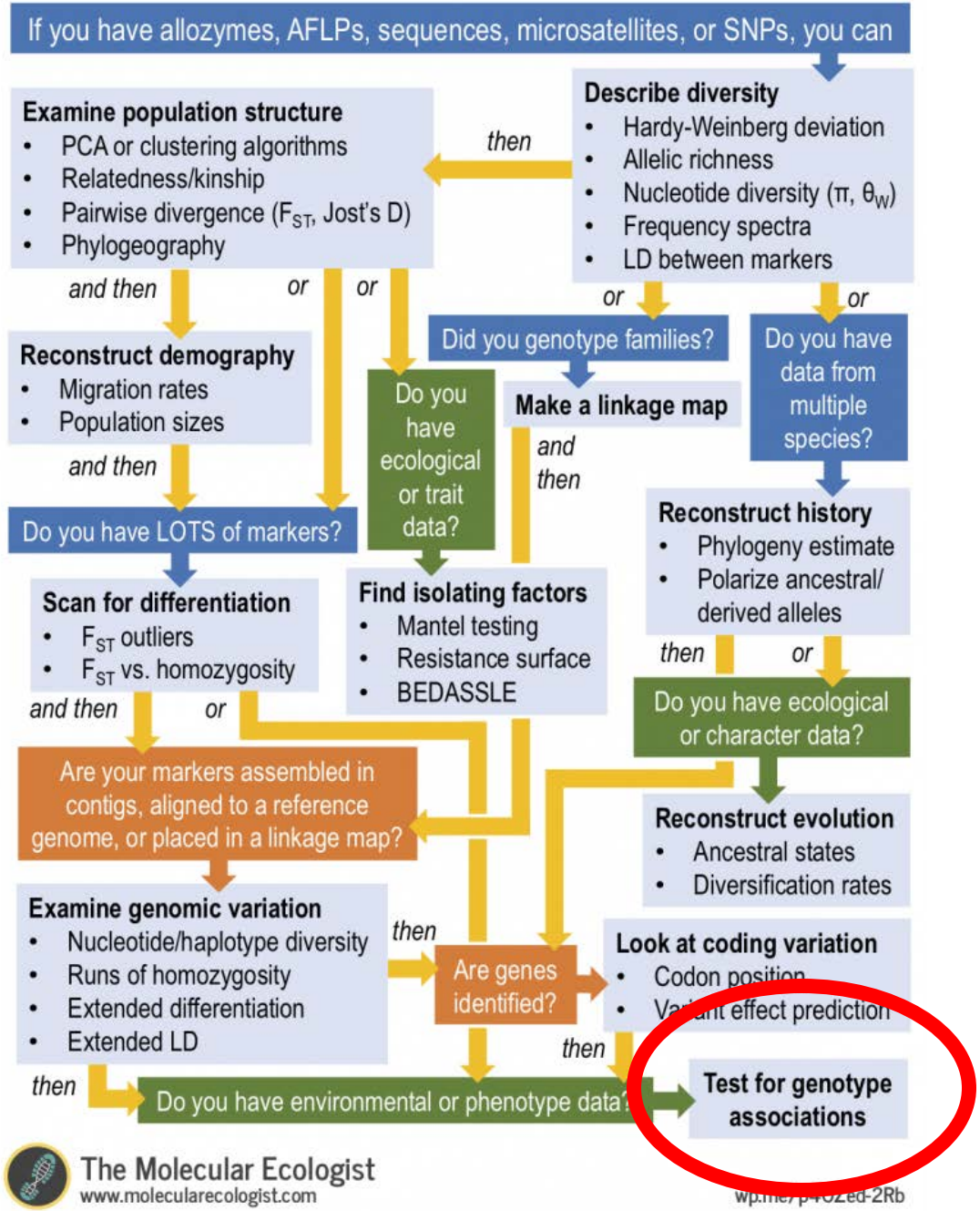
Jesse N. Weber[1]†, Brant K. Peterson[1,2] & Hopi E. Hoekstra[1,2]

Vera Domingues/Hopi Hoekstra



Weber et al 2013

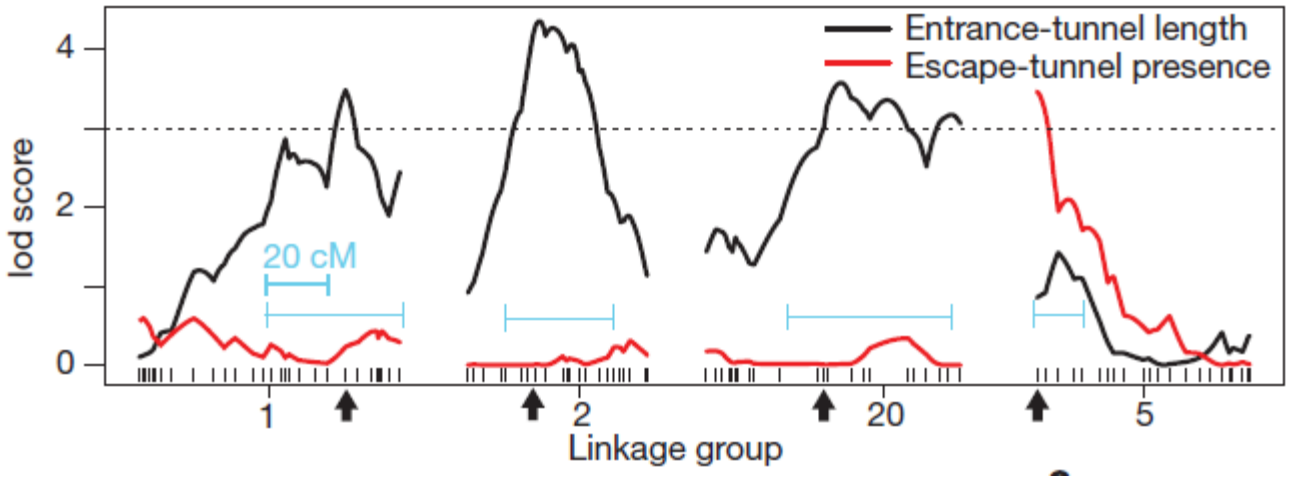
What can I do with this genetic data?



Tunnel building at least 4 loci

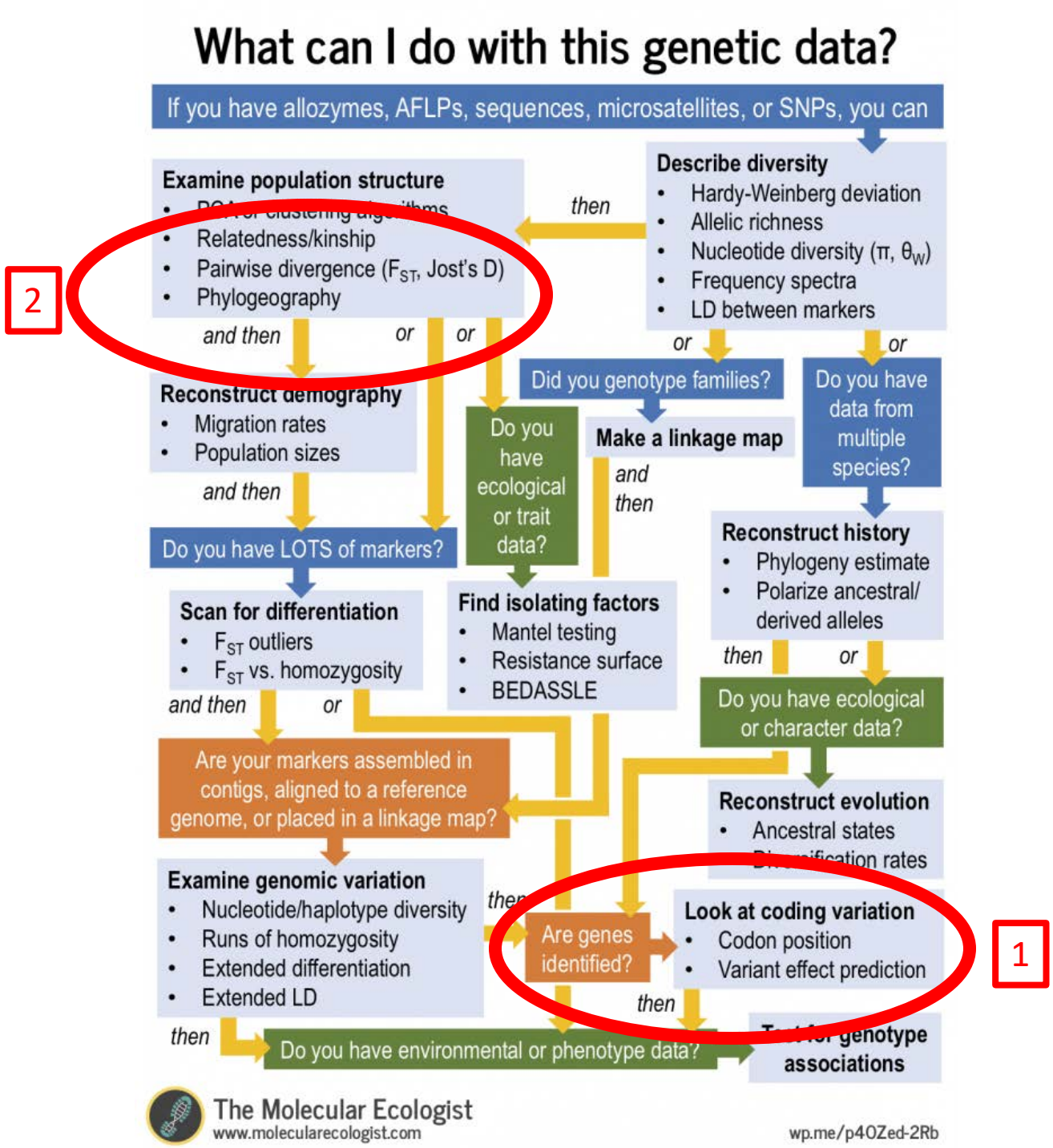
Crossed 2 species of mice:
oldfield mice (escape tunnels) x deer mice (simple nests)

QTL mapping with ddRAD-seq



Weber et al 2013

Exon sequencing


What can I do with this genetic data?
If you have allozymes, AFLPs, sequences, microsatellites, or SNPs, you can
Describe diversity
Hardy-Weinberg deviation
Allelic richness
Nucleotide diversity (π, θW)
Frequency spectra
LD between markers
then
Examine population structure
Relatedness/kinship
Pairwise divergence (FST, Jost's D)
Phylogeography
2
and then
or
or
or
or
Did you genotype families?
Do you have data from multiple species?
Reconstruct demography
Migration rates
Population sizes
and then
Do you have ecological or trait data?
Make a linkage map
and then
Do you have LOTS of markers?
Reconstruct history
Phylogeny estimate
Polarize ancestral/ derived alleles
Scan for differentiation
FST outliers
FST vs. homozygosity
Find isolating factors
Mantel testing
Resistance surface
BEDASSLE
then
or
and then
or
Do you have ecological or character data?
Are your markers assembled in contigs, aligned to a reference genome, or placed in a linkage map?
Reconstruct evolution
Ancestral states
Examine genomic variation
Nucleotide/haplotype diversity
Runs of homozygosity
Extended differentiation
Extended LD
then
Are genes identified?
Look at coding variation
Codon position
Variant effect prediction
1
then
then
Do you have environmental or phenotype data?
genotype associations
The Molecular Ecologist
www.molecularecologist.com
wp.me/p4OZed-2Rb

Targeted (re)sequencing or exon capture

Targeted (re)sequencing or exon capture

- Target for specific regions (e.g., exons)

Targeted (re)sequencing or exon capture

- Target for specific regions (e.g., exons)
- Subset of the genome can answer your question

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)

Targeted (re)sequencing or exon capture

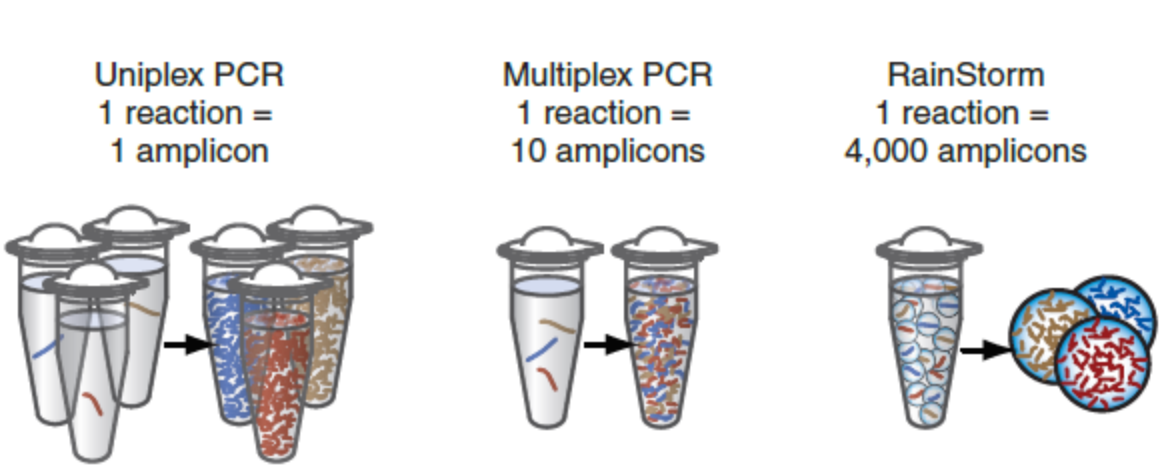
- Target for specific regions (e.g., exons)
- Subset of the genome can answer your question
- Reduce effort and cost
    - Reduced data storage (sometimes never analyzed)
- Increase sample size

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

1,000's

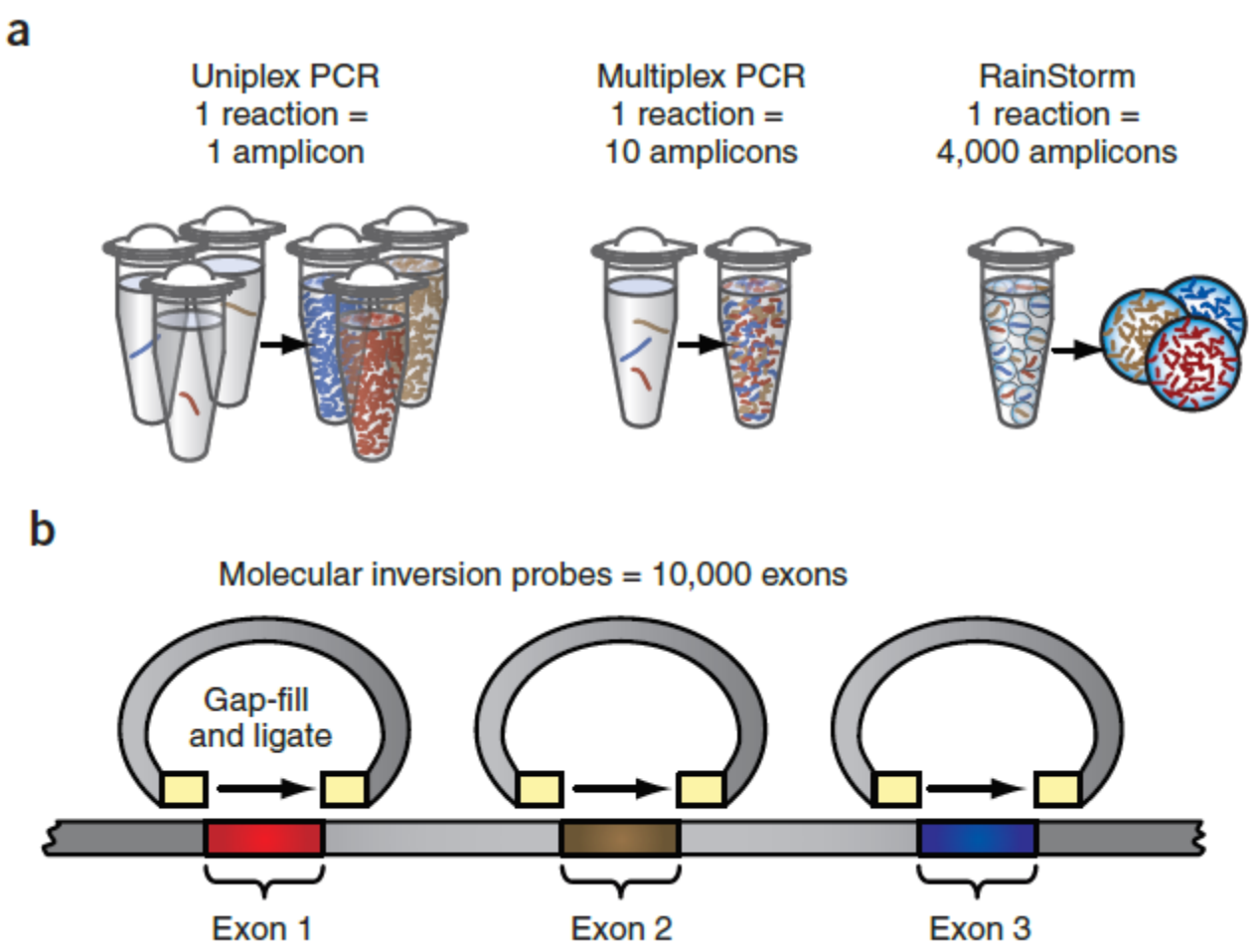


Mamanova et al 2010

Exon capture

1,000's

MIP
10,000's

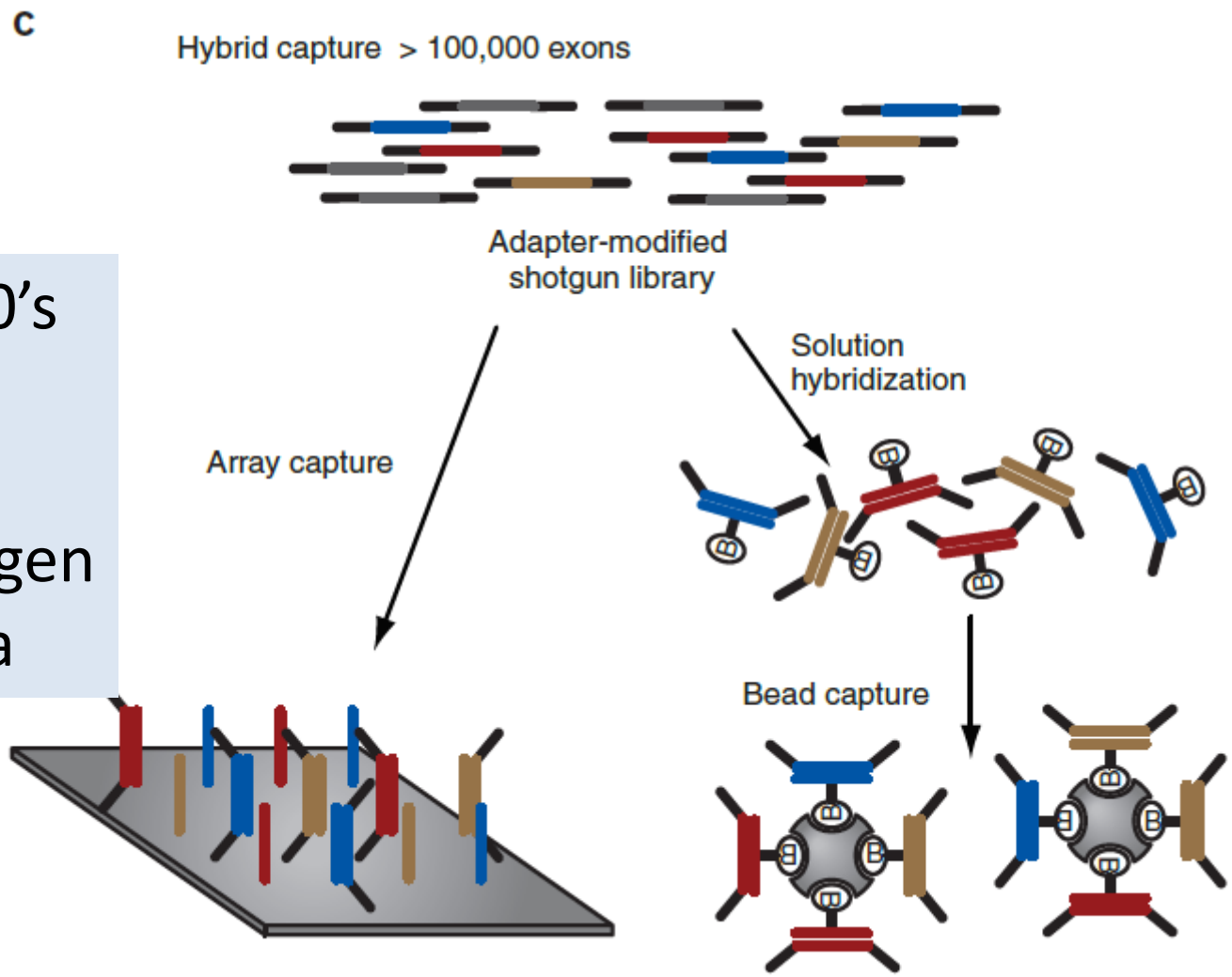


Mamanova et al 2010

Exon capture

100,000's

Agilent
Nimblegen
Illumina



Mamanova et al 2010

Exon capture – applications

Bamshad et al 2011

Exon capture – applications

- Mendelian disease discovery
  - Since many are protein coding mutations

  - Strategies:
    - Family based
    - Extreme phenotypes (e.g., cardiovascular health)

Bamshad et al 2011

Exon capture – applications

- Mendelian disease discovery
  - Since many are protein coding mutations

  - Strategies:
    - Family based
    - Extreme phenotypes (e.g., cardiovascular health)

- Non-human organsims
  - Anything you can imagine with many (protein coding) loci

  - Cloning homozygous lethals (via heterozygotes)
  - Climate change driven subdivision

Bamshad et al 2011

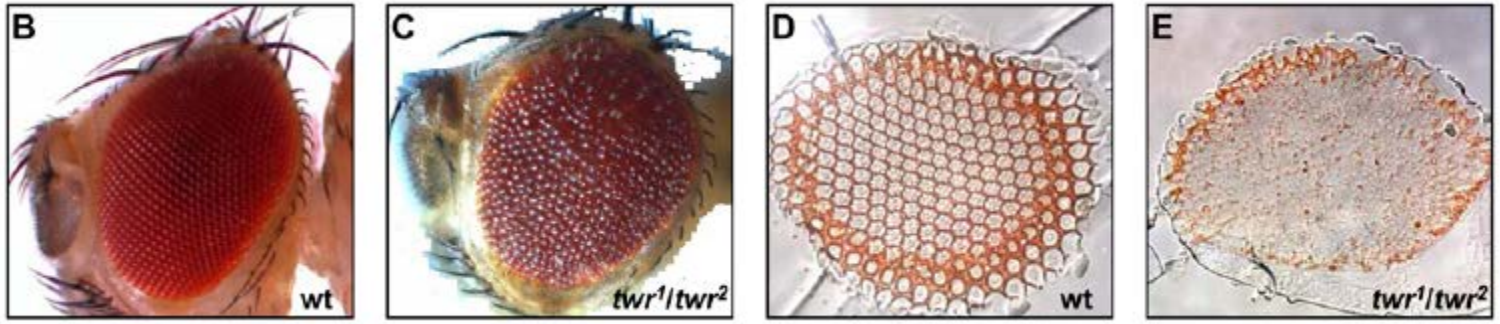
Exon capture – applications

- Mendelian disease discovery
  - Since many are protein coding mutations
  - Strategies:
    - Family based
    - Extreme phenotypes (e.g., cardiovascular health)
- Non-human organsims
  - 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

Bamshad et al 2011

Exon capture – homozygous lethal case study
find in heterozygote

twisted bristles roughened eye (*twr*) mutant



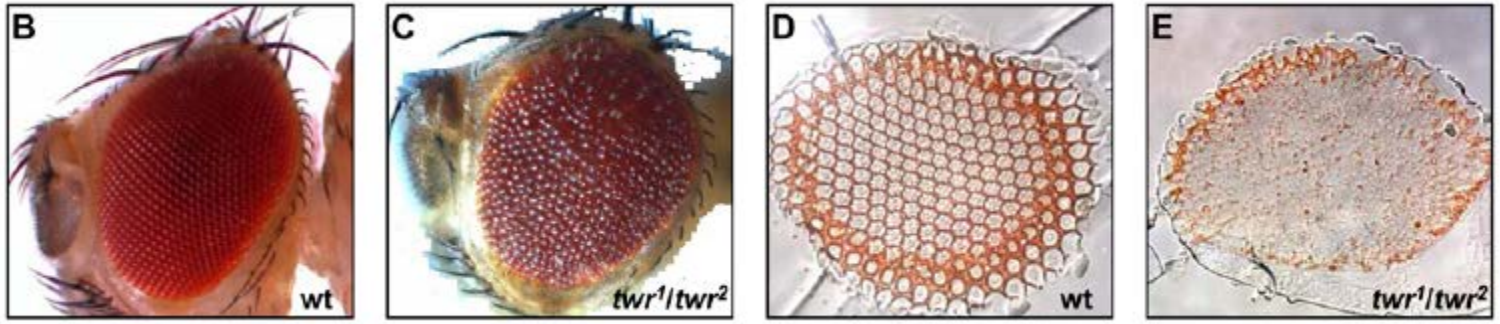
twr/twr → lethal

twr1/twr2 → viable

Wang and Chen 2010 Gen Res

Exon capture – homozygous lethal case study
find in heterozygote

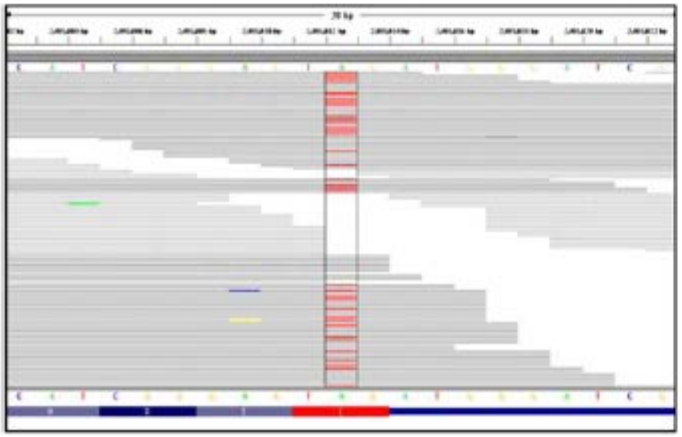
twisted bristles roughened eye (*twr*) mutant

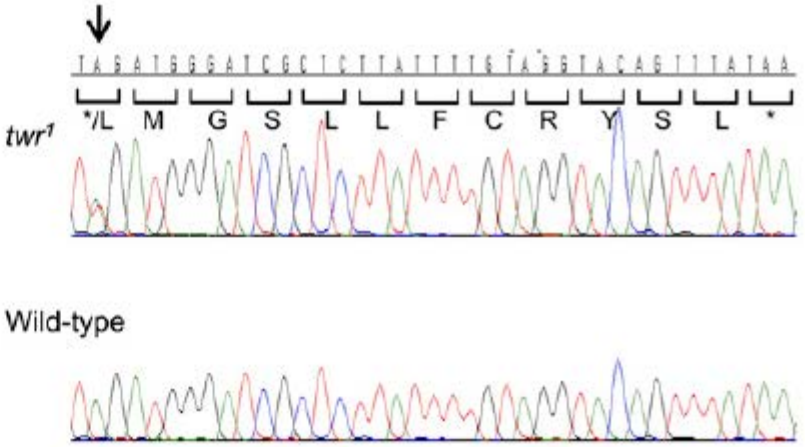


twr/twr → lethal

twr1/twr2 → viable

F

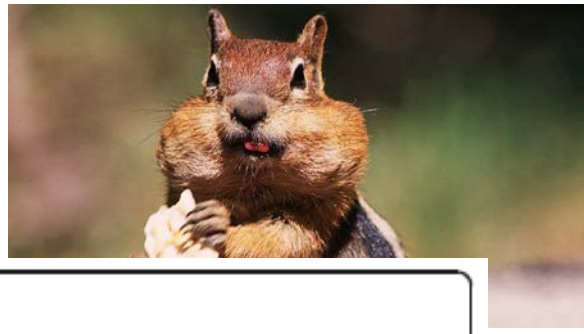


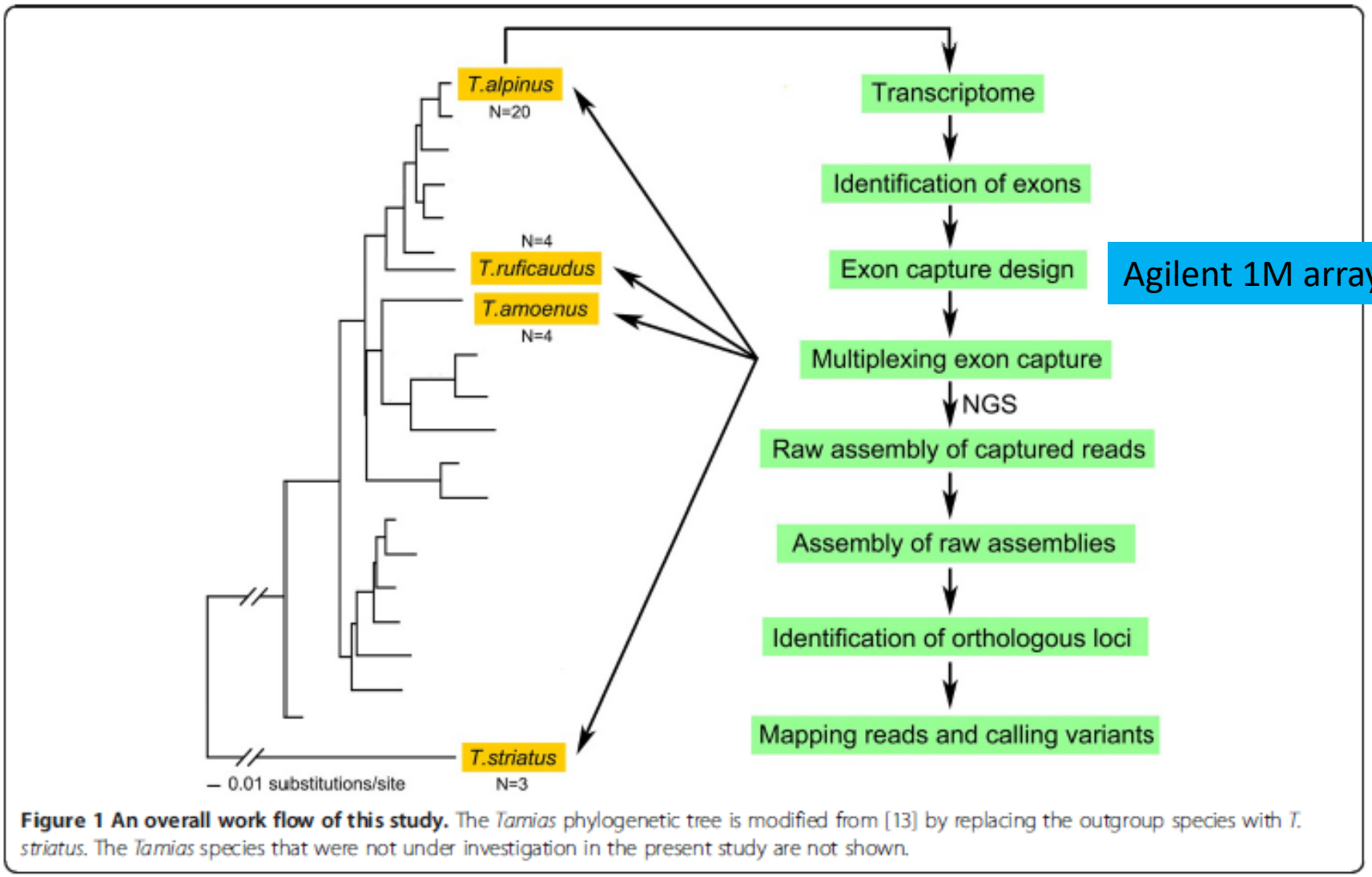


Wang and Chen 2010 Gen Res

Exon capture – chipmunk with no reference genome

Exon capture – chipmunk with no reference genome



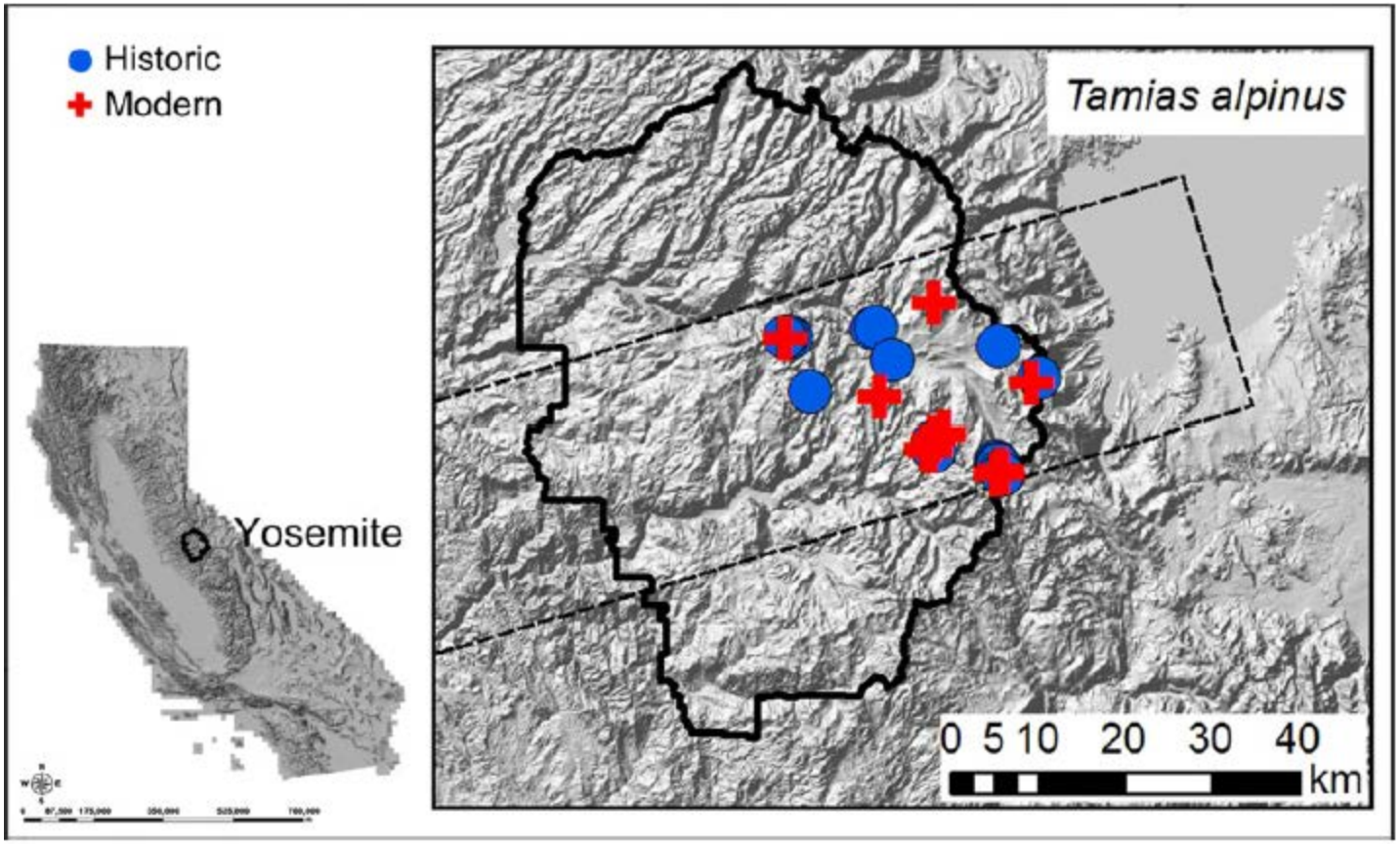


Agilent 1M arrays

Figure 1 An overall work flow of this study. The *Tamias* phylogenetic tree is modified from [13] by replacing the outgroup species with *T. striatus*. The *Tamias* species that were not under investigation in the present study are not shown.

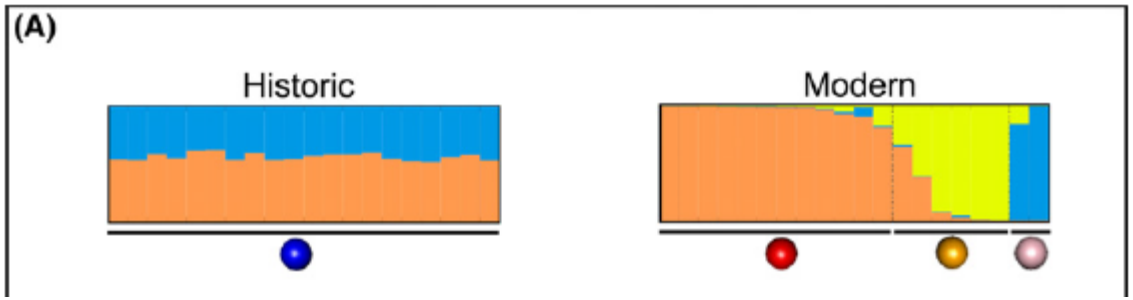
Bi et al 2012

Exon capture – chipmunk →
museum samples and range contraction



Bi et al 2013

Exon capture – chipmunk →
museum samples and range contraction



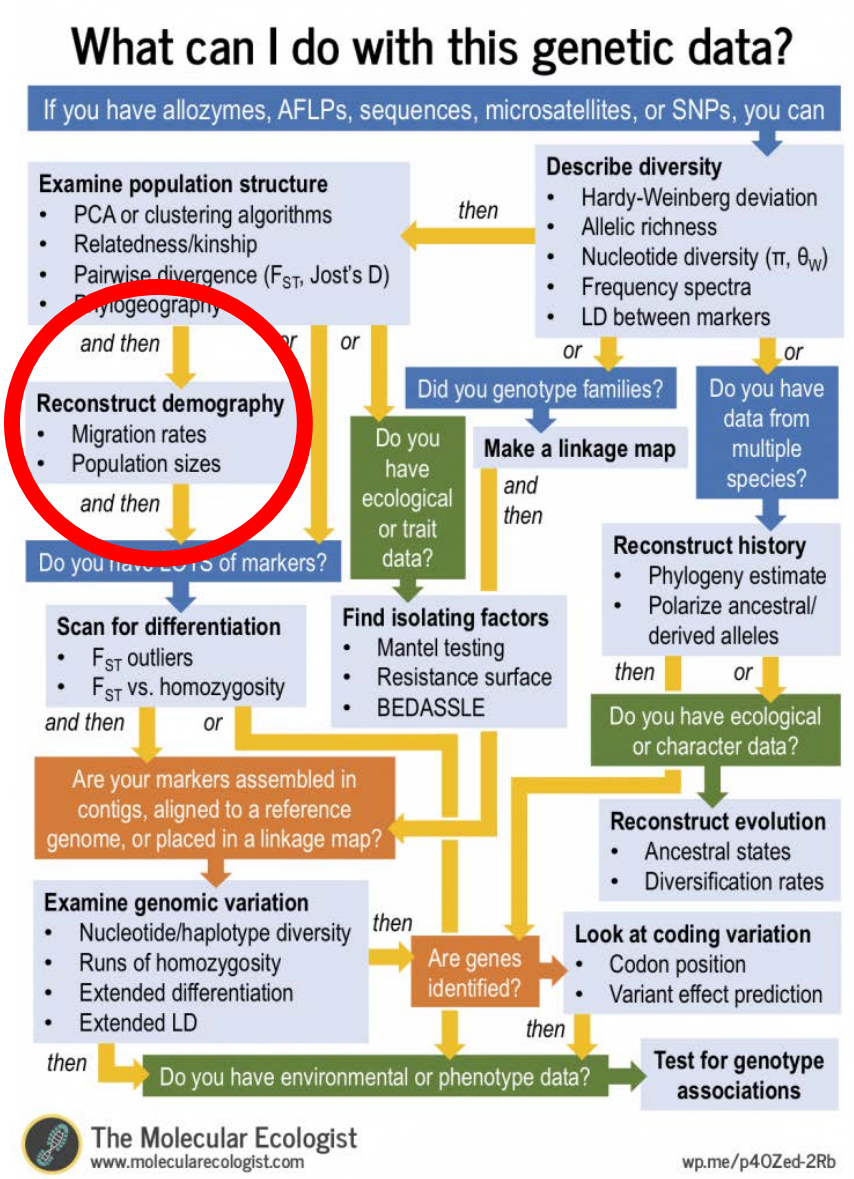
STRUCTURE analysis reveals more modern subdivision

Consistent with reduced gene flow associated with climate change

Bi et al 2013

RE delimited sequencing

RE delimited sequencing



GENOME RESEARCH
2011

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

Francesca Luca,[1] Richard R. Hudson,[1,2,3] David B. Witonsky,[1] and Anna Di Rienzo[1,3]

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

GENOME RESEARCH
2011

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

Francesca Luca,[1] Richard R. Hudson,[1,2,3] David B. Witonsky,[1] and Anna Di Rienzo[1,3]

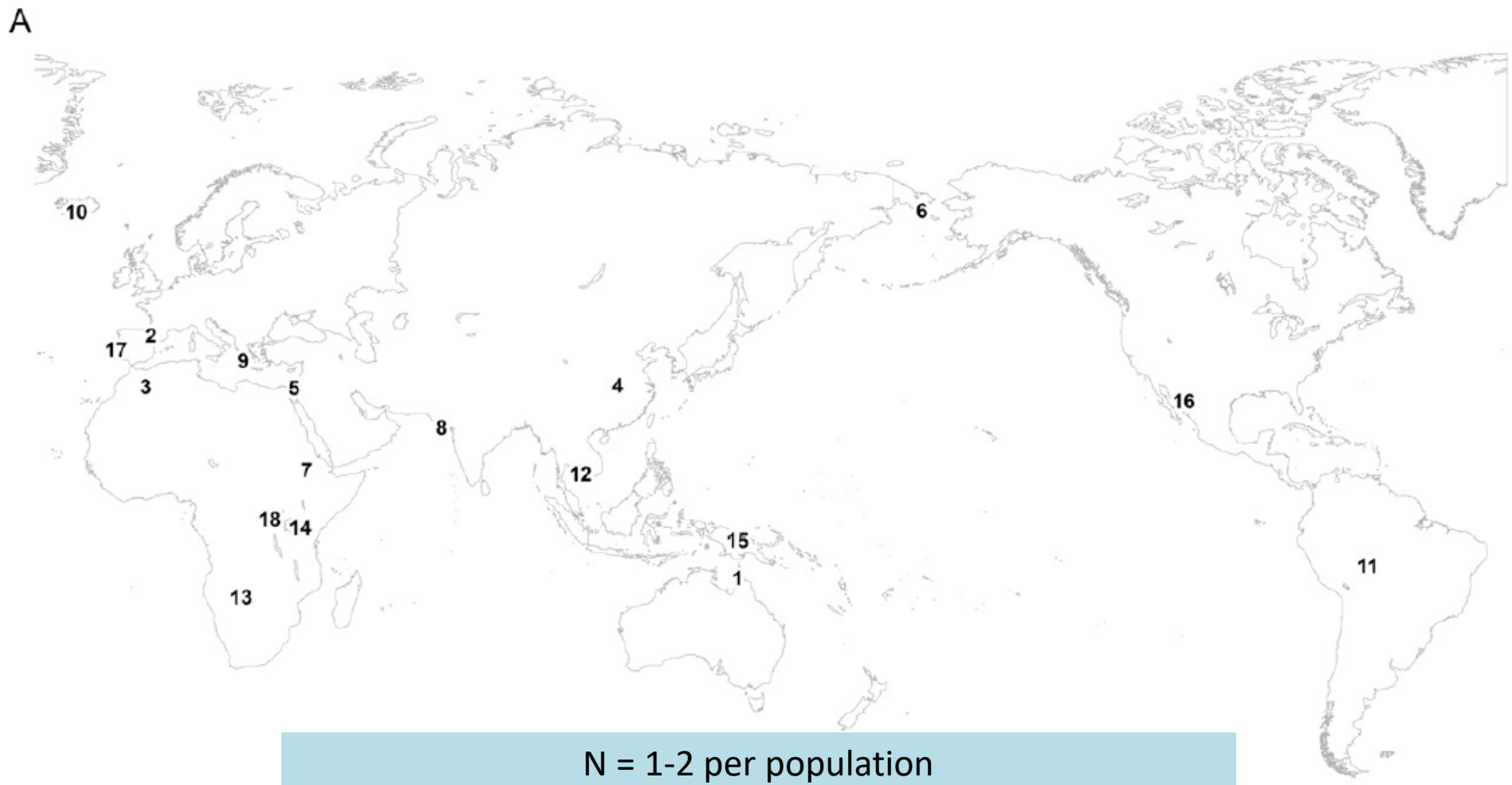
Reduced representation simply by digestion (Van Tassell et al 2008)

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



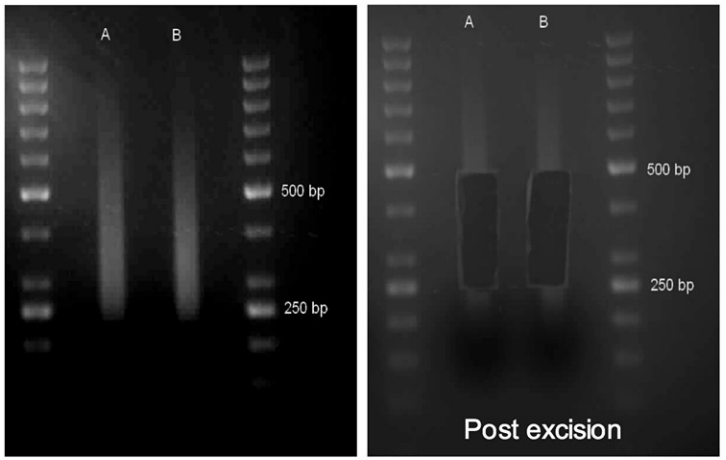
N = 1-2 per population
Use power of many loci to estimate population parameters
• nucleotide diversity

Luca et al 2011

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 ≤4 mismatch

Bioinformatics – mapping, filtering

Strategy:

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

Filtering: QC or clean up data

- Remove 2 nts from RE <GTAC>
- 3rd nt also biased
- keep phred score ≥20
- if >2 alleles (keep best 2)

Bioinformatics – mapping, filtering

Strategy:

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

Filtering: QC or clean up data

- Remove 2 nts from RE <GTAC>
- 3rd nt also biased
- keep phred score ≥20
- if >2 alleles (keep best 2)

Caveats:

- Cannot know pcr duplicates
- Loss due to RE polymorphism

Bioinformatics – mapping, filtering

Strategy:

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

RsaI RsaI <GTAC>

Filtering: QC or clean up data

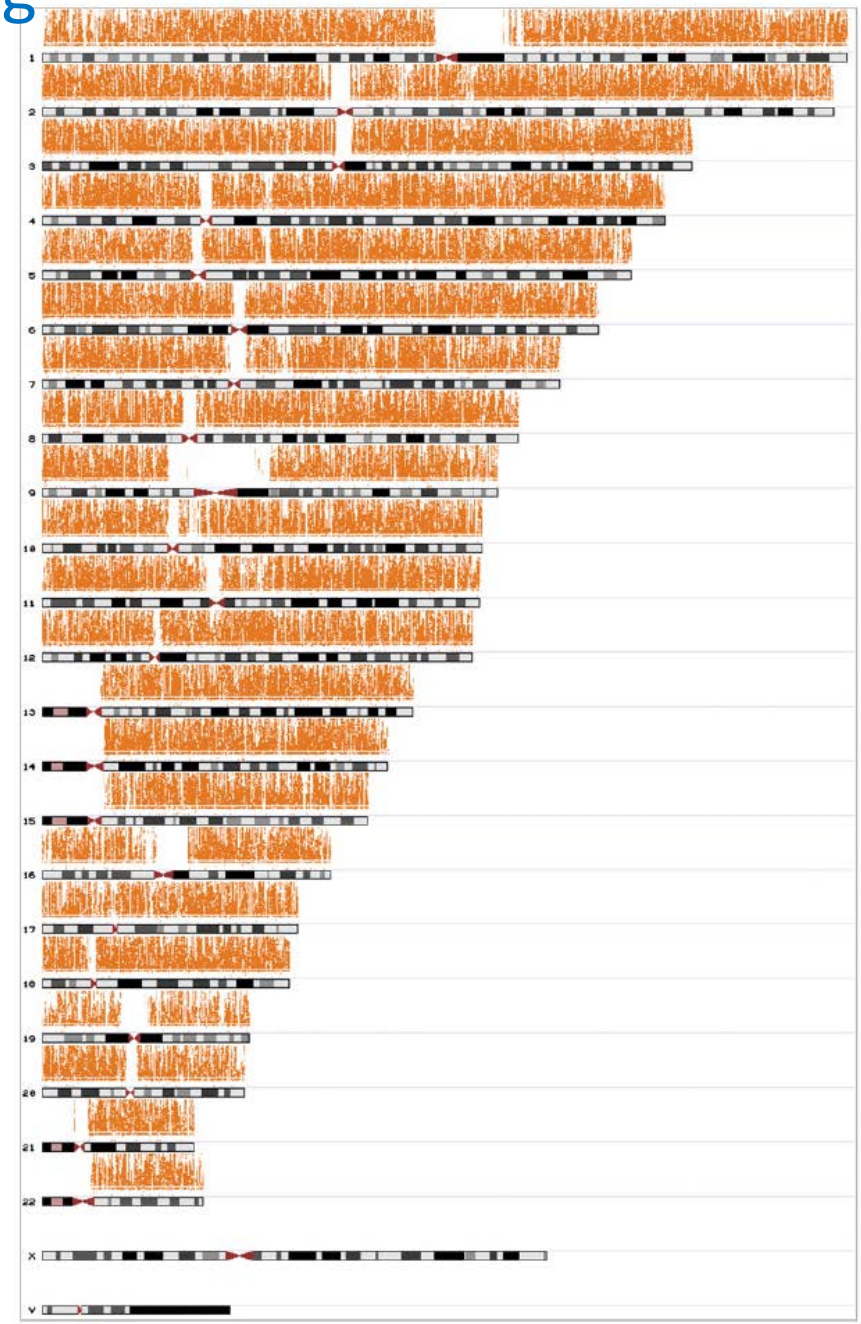
- Remove 2 nts from RE <GTAC>
- 3rd nt also biased
- keep phred score ≥20
- if >2 alleles (keep best 2)

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 (π):

Calculate: Proportion of heterozygous sites within each genome is estimate of its π population

0.61-1.08/1000 <lower than previous estimates>

Test serial founder effect:

Expect negative correlation

π ~ geographic distance

Data analysis #1

This method: 1000's of sites per genome

Nucleotide diversity (π):

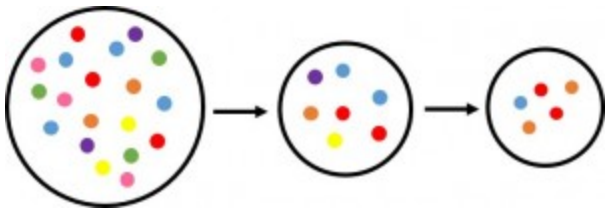
Calculate: Proportion of heterozygous sites within each genome is estimate of its π population

0.61-1.08/1000 <lower than previous estimates>

Test serial founder effect:

Expect negative correlation

π ~ geographic distance



Data analysis #1

This method: 1000's of sites per genome

Nucleotide diversity (π):

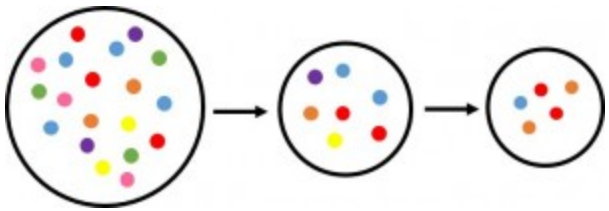
Calculate: Proportion of heterozygous sites within each genome is estimate of its π population

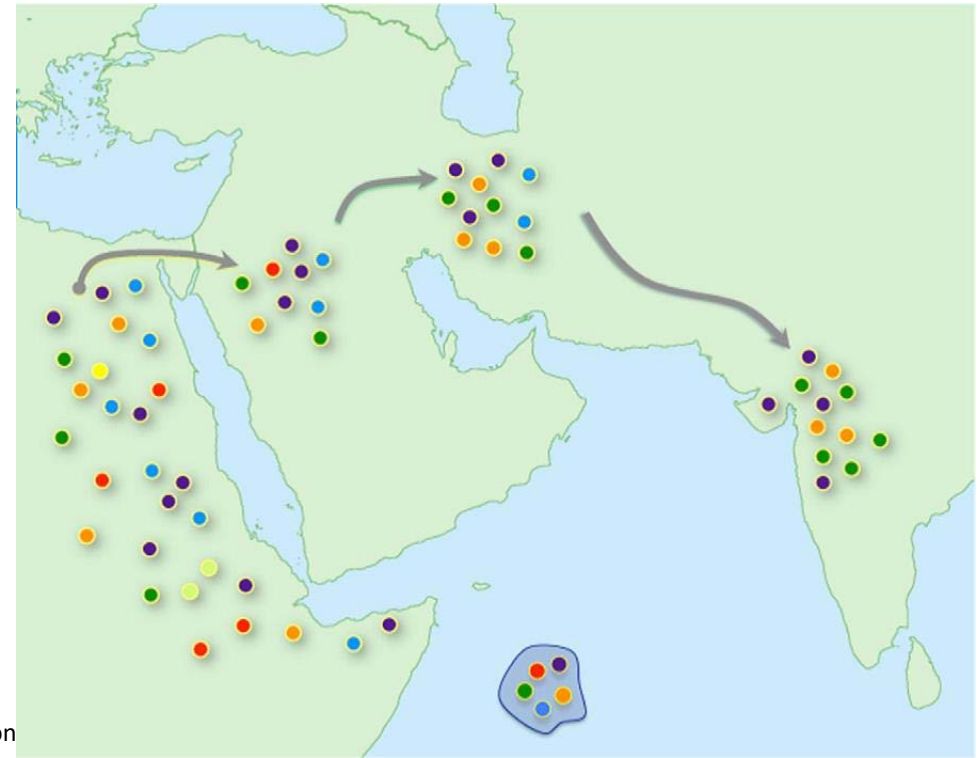
0.61-1.08/1000 <lower than previous estimates>

Test serial founder effect:

Expect negative correlation

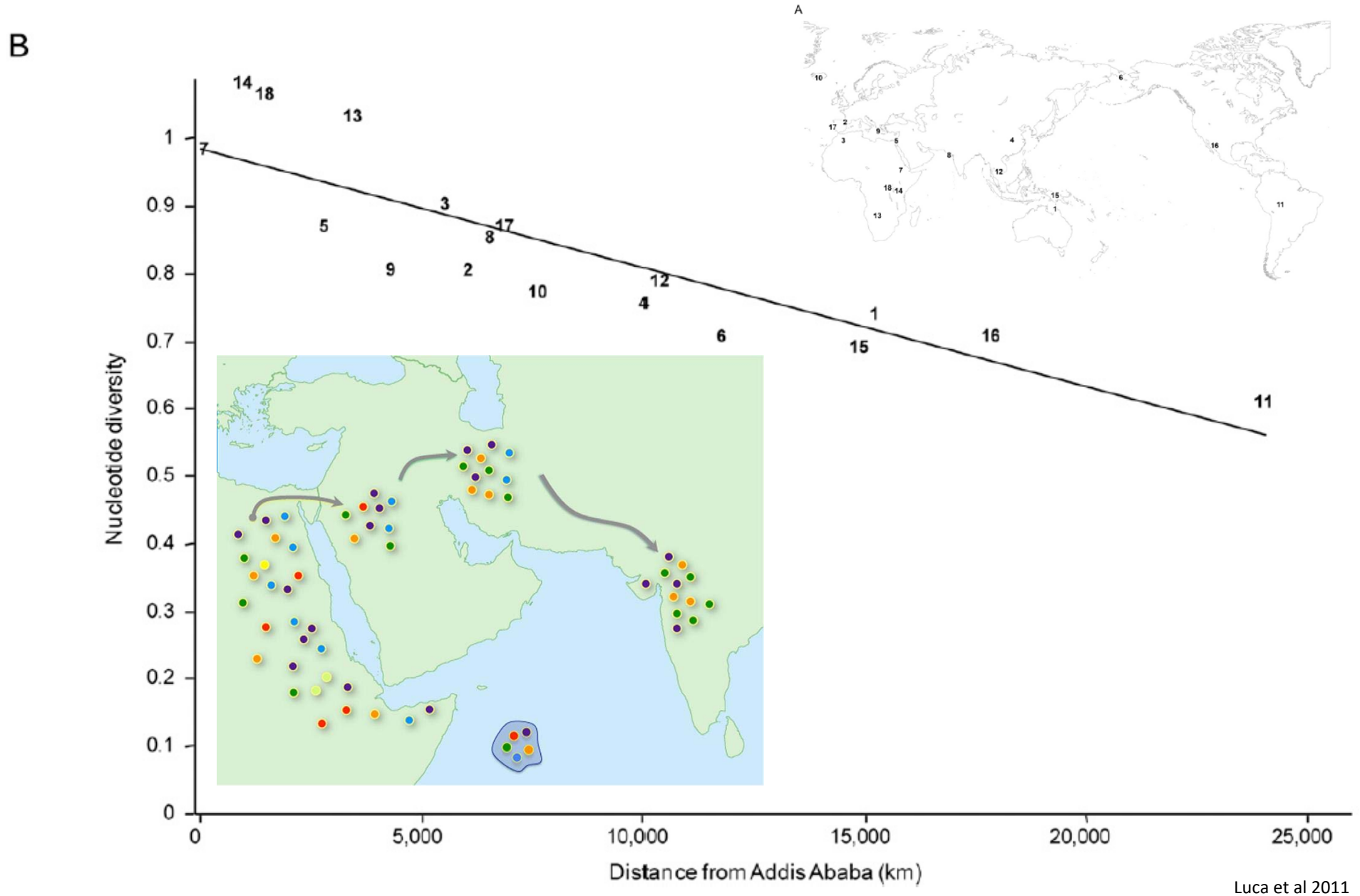
π ~ geographic distance





Emma Whittington

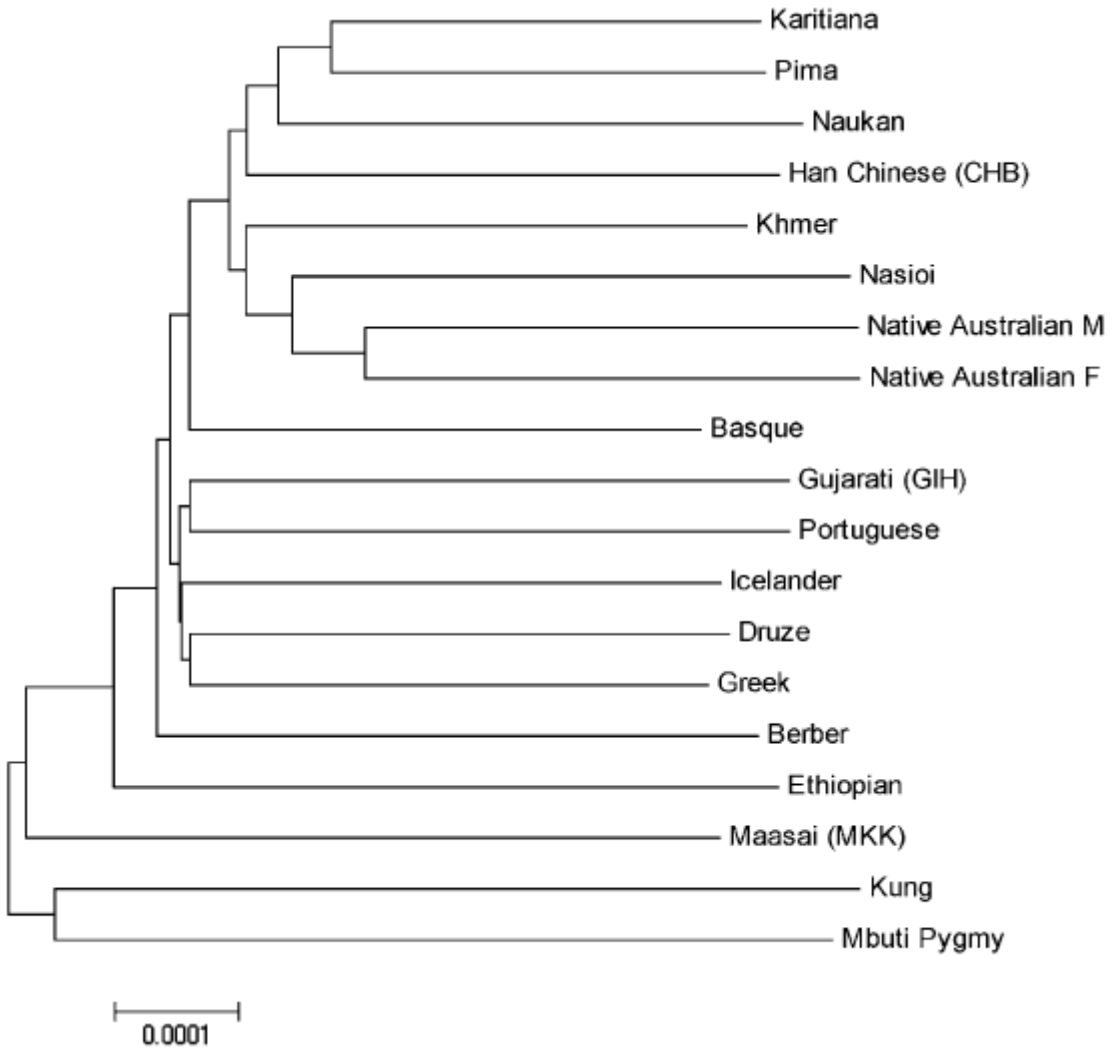
Data consistent with serial founder effect



Luca et al 2011

Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)

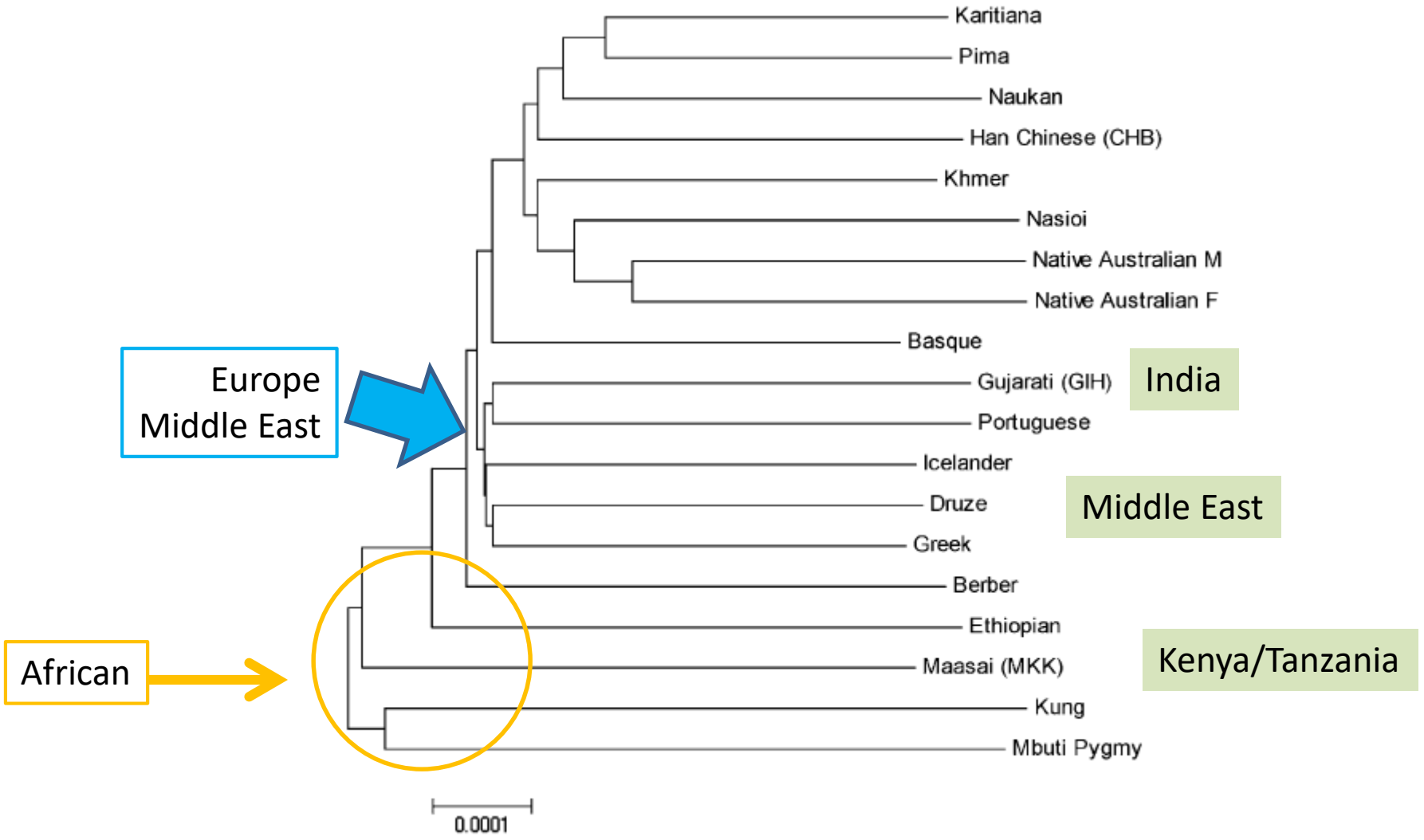


Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)

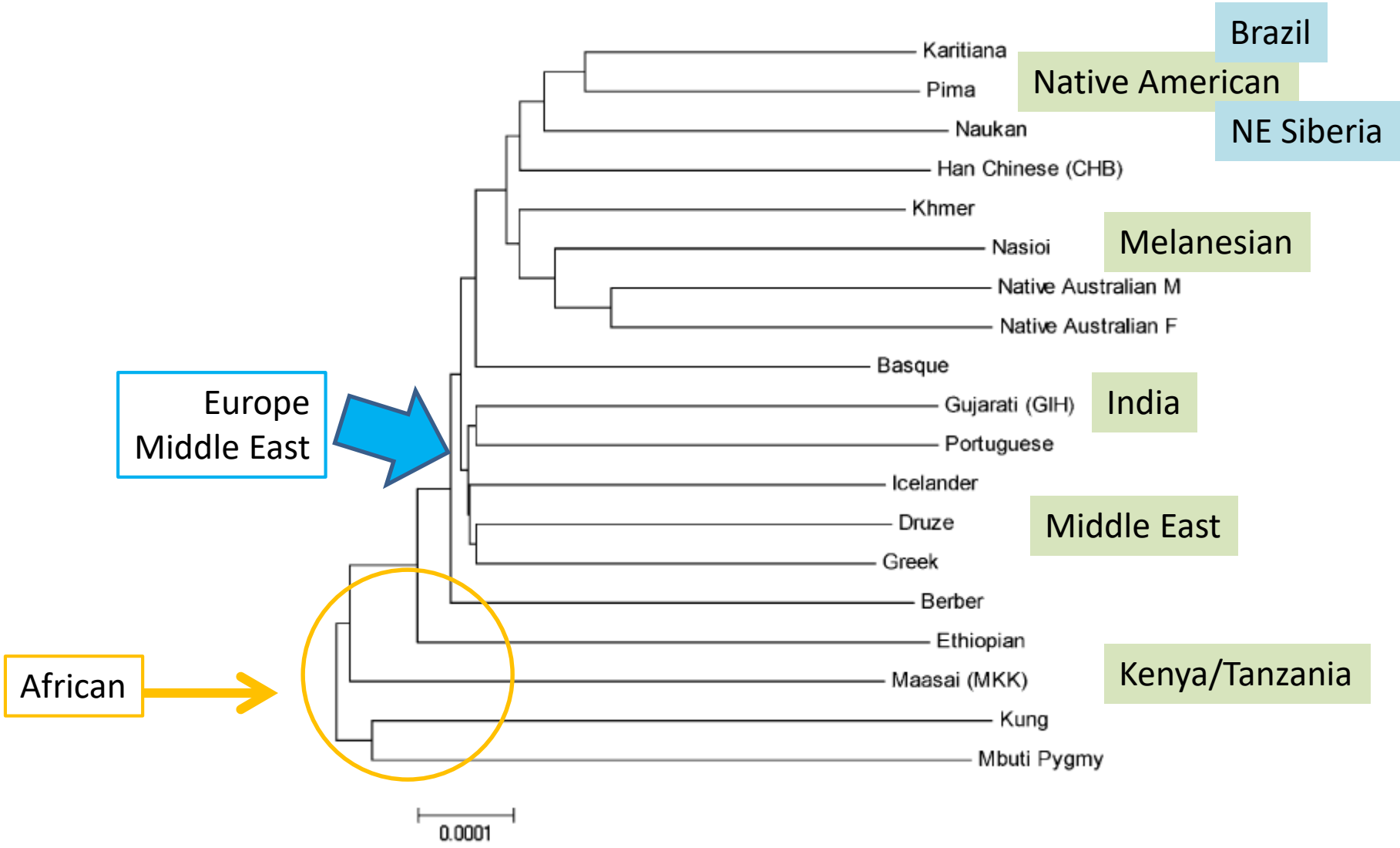
Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)



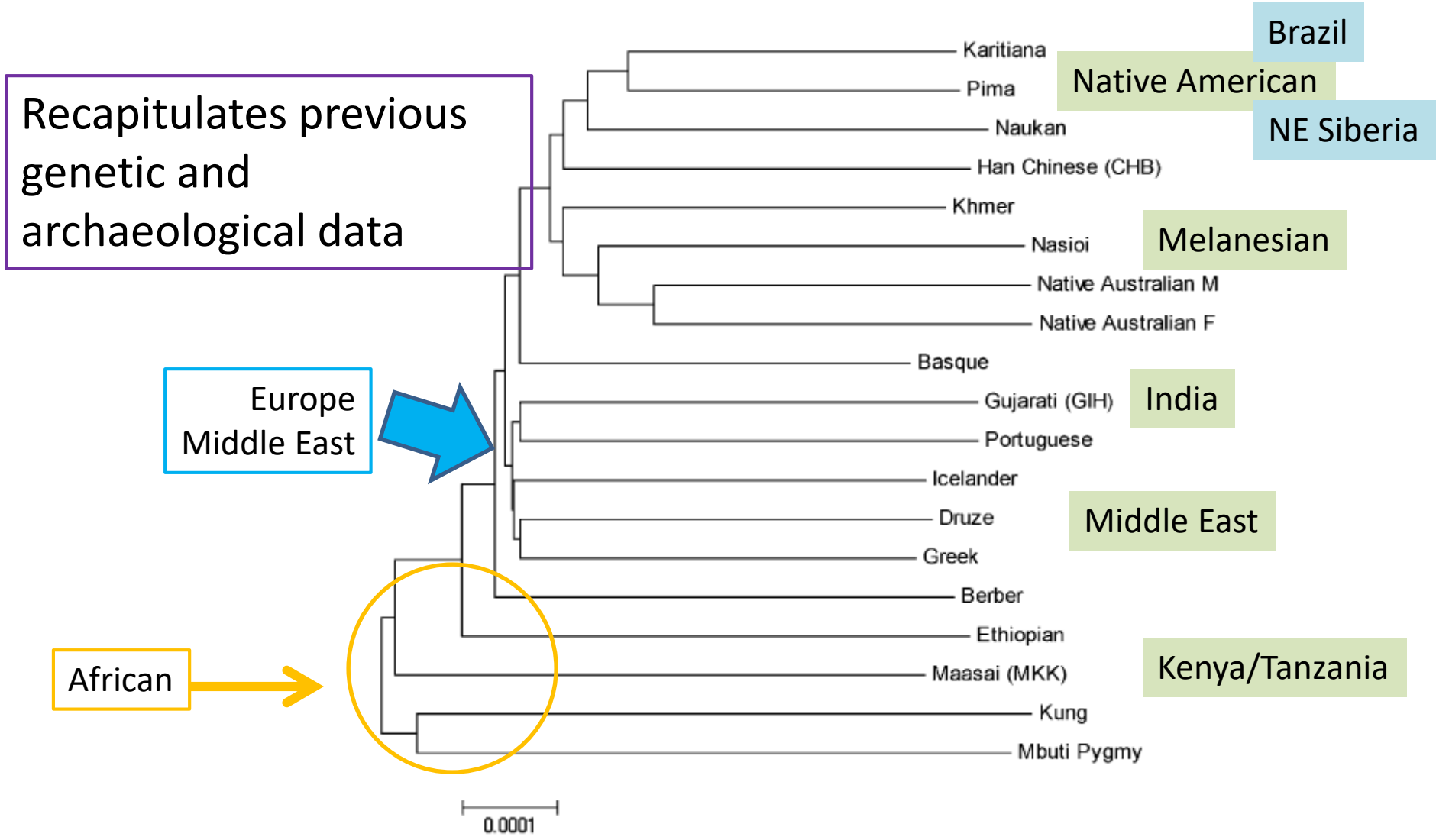
Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)



Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)

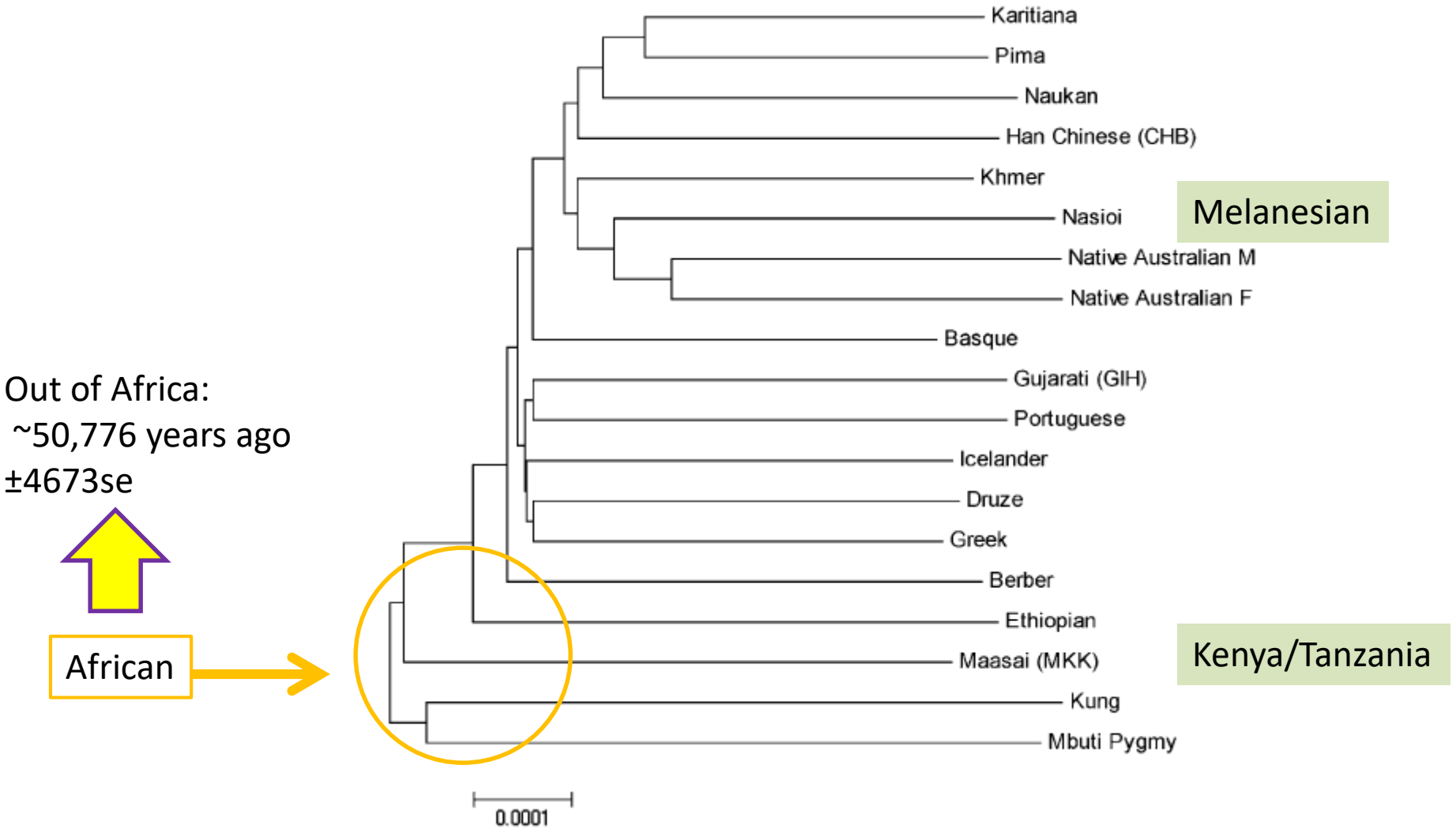


Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)

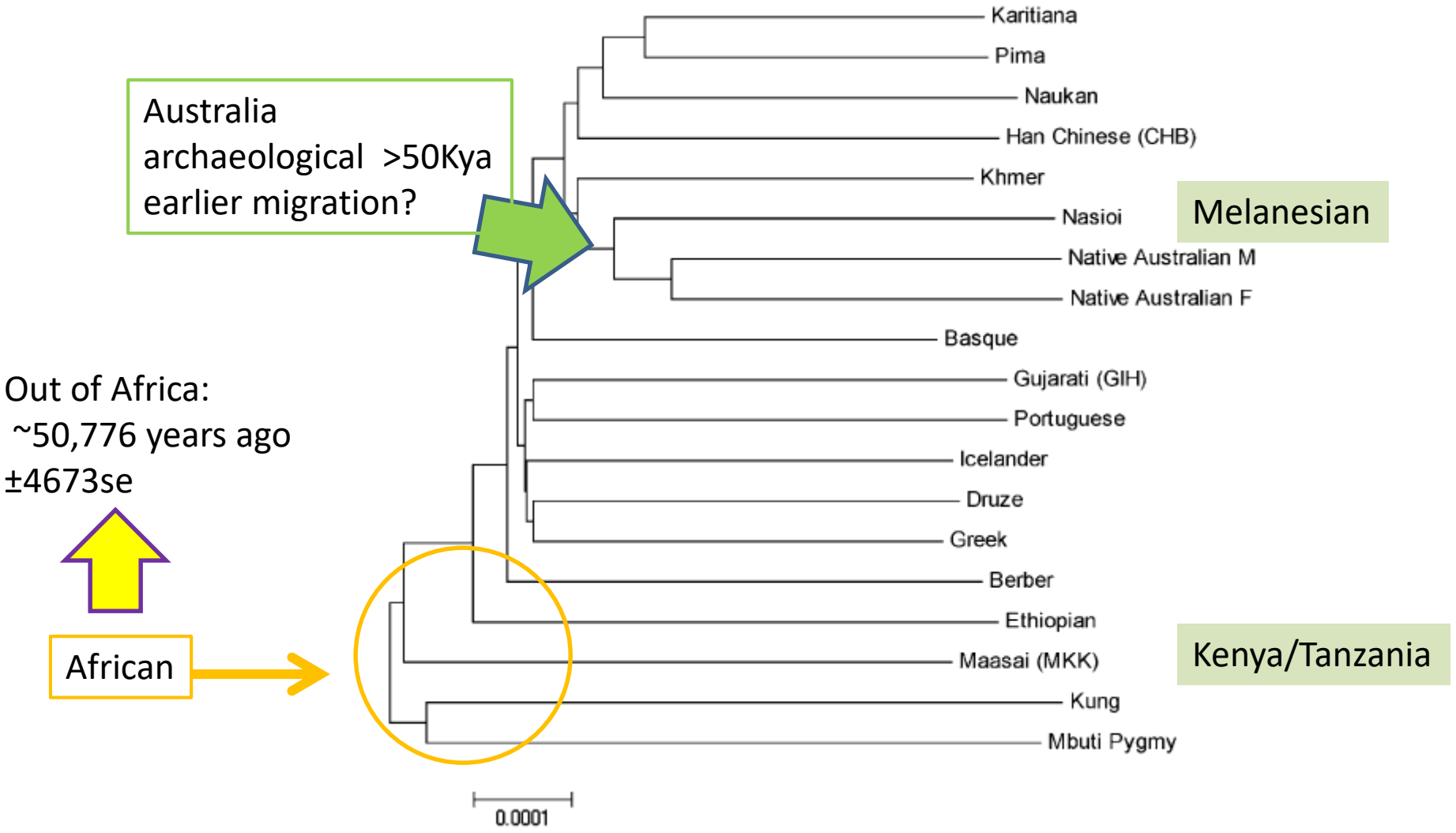
Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)



Data analysis #2: Population divergence and split times

Pairwise differences then phylogenetic tree (neighbor joining)



7 OCTOBER 2011 VOL 334 SCIENCE

An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia

Morten Rasmussen,[1,2]* Xiaosen Guo,[2,3]* Yong Wang,[4]* Kirk E. Lohmueller,[4]* Simon Rasmussen,[5] Anders Albrechtsen,[6] Line Skotte,[6] Stinus Lindgreen,[1,6] Mait Metspalu,[7] Thibaut Jombart,[8] Toomas Kivisild,[9] Weiwei Zhai,[10] Anders Eriksson,[11] Andrea Manica,[11] Ludovic Orlando,[1] Francisco M. De La Vega,[12] Silvana Tridico,[13] Ene Metspalu,[7] Kasper Nielsen,[5] María C. Ávila-Arcos,[1] J. Víctor Moreno-Mayar,[1,14] Craig Muller,[15] Joe Dortch,[16] M. Thomas P. Gilbert,[1,2] Ole Lund,[5] Agata Wesolowska,[5] Monika Karmin,[7] Lucy A. Weinert,[8] Bo Wang,[3] Jun Li,[3] Shuaishuai Tai,[3] Fei Xiao,[3] Tsunehiko Hanihara,[17] George van Driem,[18] Aashish R. Jha,[19] François-Xavier Ricaut,[20] Peter de Knijff,[21] Andrea B Migliano,[9,22] Irene Gallego Romero,[19] Karsten Kristiansen,[2,3,6] David M. Lambert,[23] Søren Brunak,[5,24] Peter Forster,[25,26] Bernd Brinkmann,[26] Olaf Nehlich,[27] Michael Bunce,[13] Michael Richards,[27,28] Ramneek Gupta,[5] Carlos D. Bustamante,[12] Anders Krogh,[1,6] Robert A. Foley,[9] Marta M. Lahr,[9] Francois Balloux,[8] Thomas Sicheritz-Pontén,[5,29] Richard Villems,[7,30] Rasmus Nielsen,[4,6]† Jun Wang,[2,3,6,31]† Eske Willerslev[1,2]†

7 OCTOBER 2011 VOL 334 SCIENCE

An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia

Morten Rasmussen,[1,2]* Xiaosen Guo,[2,3]* Yong Wang,[4]* Kirk E. Lohmueller,[4]* Simon Rasmussen,[5]
Anders Albrechtsen,[6] Line Skotte,[6] Stinus Lindgreen,[1,6] Mait Metspalu,[7] Thibaut Jombart,[8]
Toomas Kivisild,[9] Weiwei Zhai,[10] Anders Eriksson,[11] Andrea Manica,[11] Ludovic Orlando,[1]
Francisco M. De La Vega,[12]
J. Víctor Moreno-Mayar,[1,14]
Agata Wesolowska,[5] Monika
Fei Xiao,[3] Tsunehiko Hanih
Peter de Knijff,[21] Andrea B
David M. Lambert,[23] Søren
Michael Bunce,[13] Michael F
Anders Krogh,[1,6] Robert A.
Richard Villems,[7,30] Rasmu

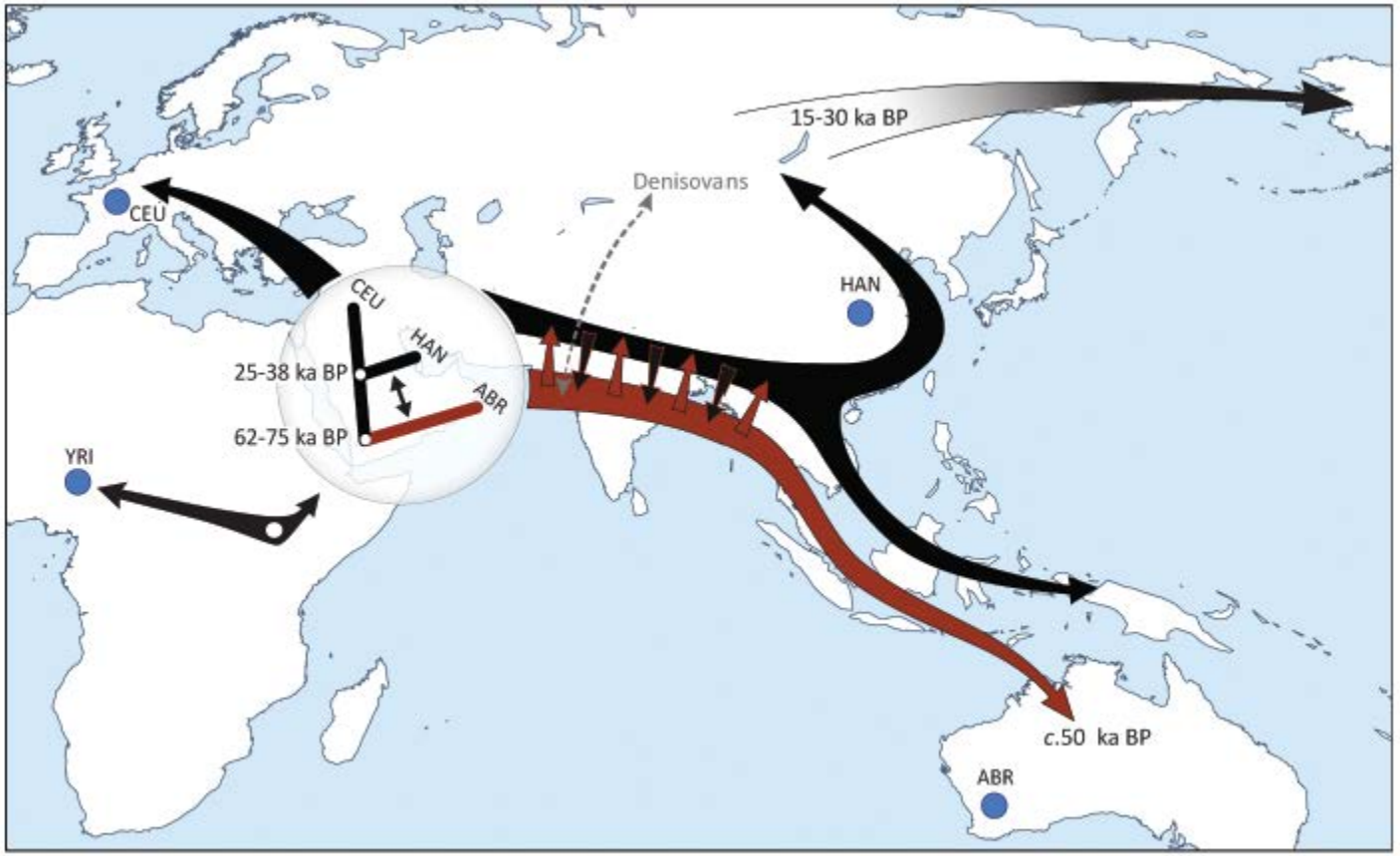


Fig. 2. Reconstruction of early spread of modern humans outside Africa. The tree shows the divergence of

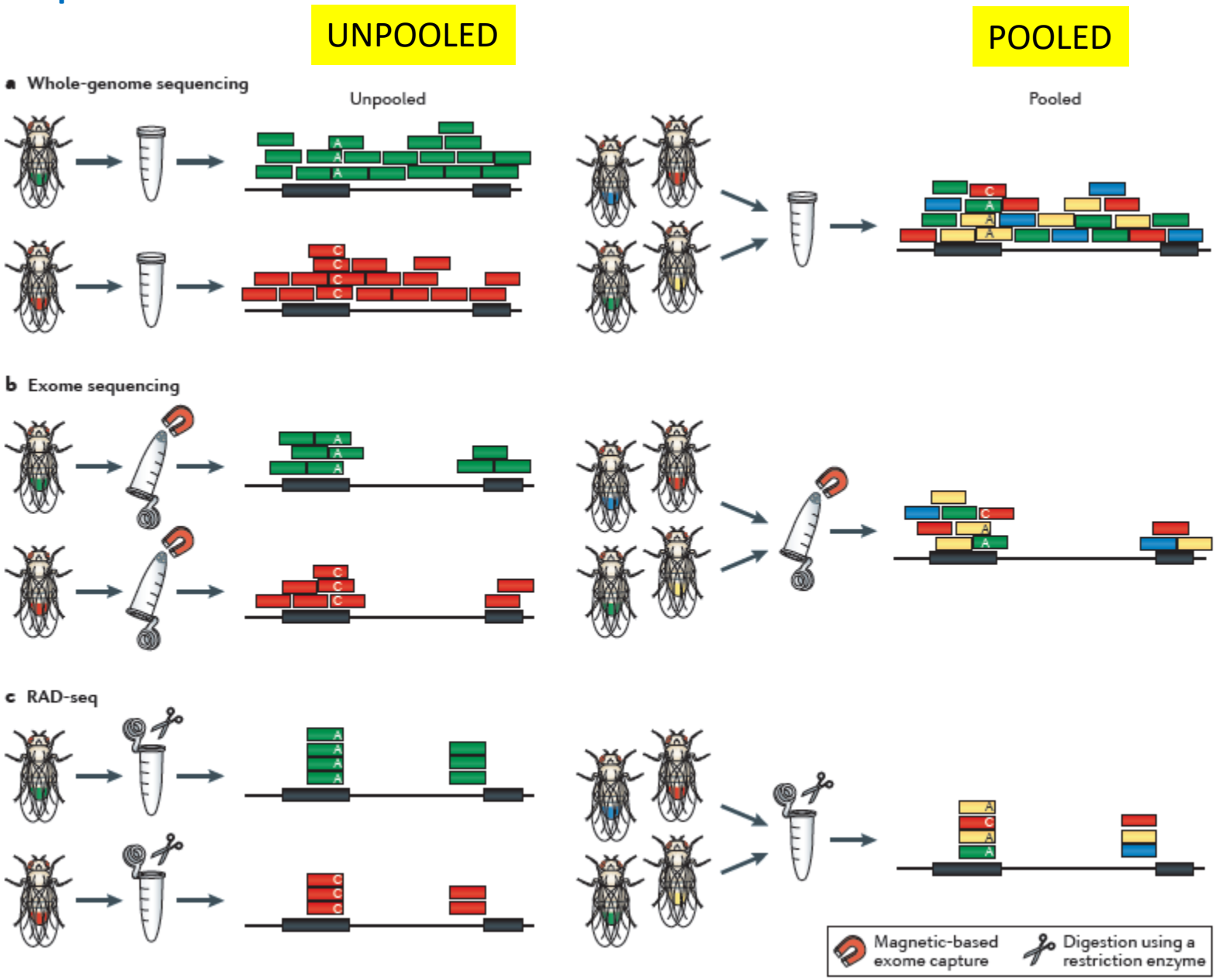
Low coverage approaches

PoolSeq

Eponymous:
pool samples then sequence

often ~1x depth for each sample

PoolSeq



Schlotterer et al 2014

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

Schlotterer et al 2014

PoolSeq – advantages

Schlotterer et al 2014

PoolSeq – advantages

Better estimate of allele frequencies (polymorphism data)

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

Pool → not every individual, BUT basically only 1x

Schlotterer et al 2014

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]

Schlotterer et al 2014

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]

Schlotterer et al 2014

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…

Schlotterer et al 2014

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

Schlotterer et al 2014

PoolSeq – concerns

Schlotterer et al 2014
Cutler & Jensen 2010

PoolSeq – concerns

SNP call versus Illumina error rate

Need enough depth, but assaying superficially

Schlotterer et al 2014
Cutler & Jensen 2010

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

Schlotterer et al 2014
Cutler & Jensen 2010

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,

Schlotterer et al 2014
Cutler & Jensen 2010

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

Schlotterer et al 2014
Cutler & Jensen 2010

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)

Schlotterer et al 2014
Cutler & Jensen 2010

PoolSeq – some software

Population genetics		
PoPoolation	Estimates variation within populations	39
PoPoolation2	Estimates differentiation between multiple populations	132
Pool-HMM	Detects selective sweeps from the allele frequency spectrum using a hidden Markov model	133
npstat	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
Stacks	Developed for population genomics with RAD-seq; may also be used with pooled RAD-seq data	135
Bayenv2	Estimates differentiation between populations	79
SelEstim	Detects and measures selection	136
KimTree	Infers population histories	137
Haplotype information		
harp	Estimates frequencies of known haplotypes using read counts; supports a sliding window approach	107
PoolHap	Estimates frequencies of known haplotypes using a regression on allele frequencies	106
eALPS	Estimates the abundance of individuals in pools given the genotypes of at least some individuals	109
LDx	Estimates linkage disequilibrium between pairs of SNPs spanned by single-end or paired-end reads	138
Forward genetic screens		
SHOREmap	Identifies causative recessive variants from a large pool of recombinants that have the recessive genotype	44
CloudMap	A cloud-based pipeline for localizing mutations	139
MULTIPOOL	Identifies candidate loci from bulk segregant analysis in which progeny are grouped by phenotype and sequenced as pools	140
Fishyskeleton	Detects mutations in zebrafish from a pool of mutant F_2 fish	141
NGM	A web-based tool for localizing mutations from a small pool of F_2 population	142
SNPtrack	A web-based tool for localizing mutations using a hidden Markov model	143

… there are others

Schlotterer et al 2014

PoolSeq – some applications

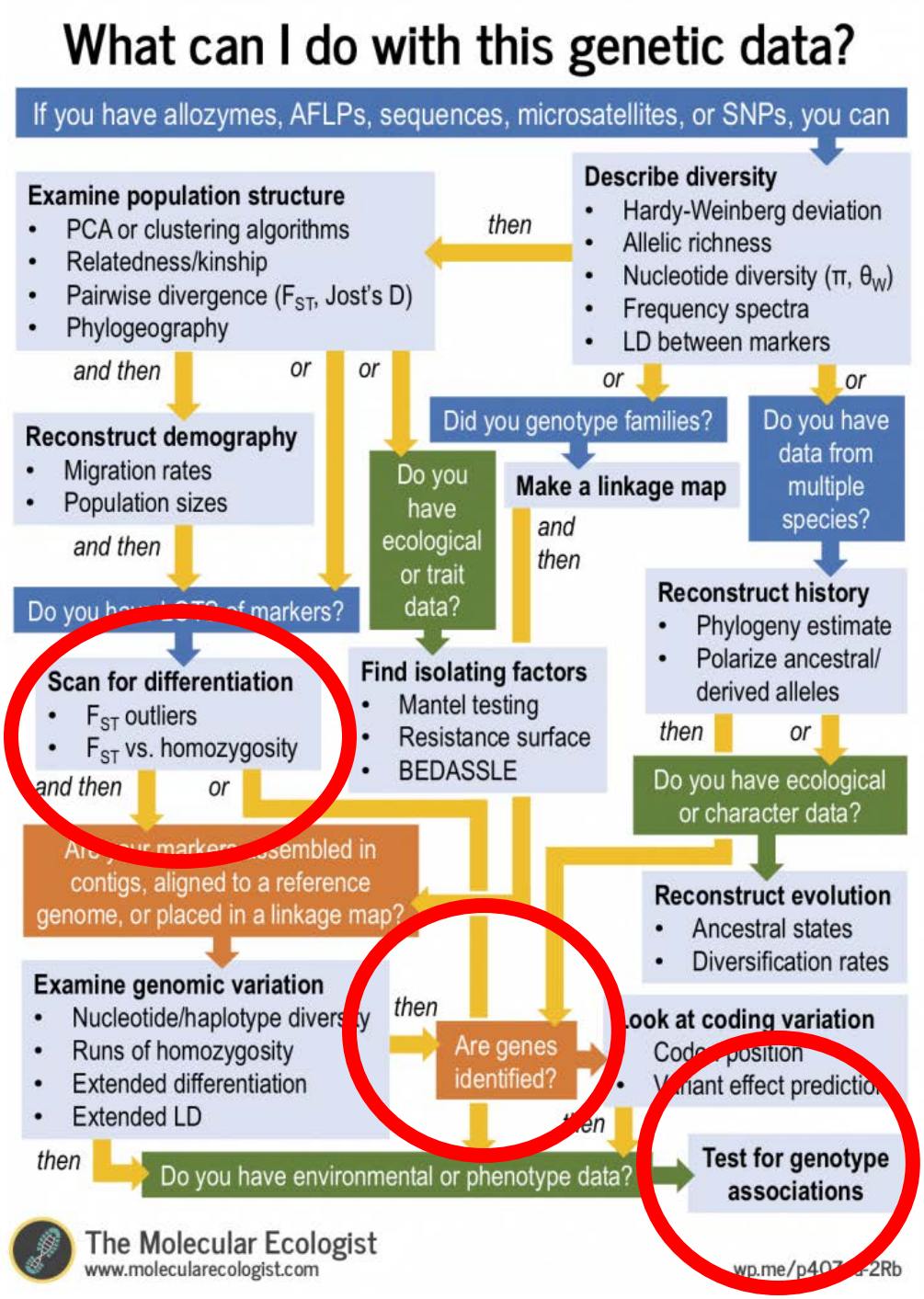
- Induced mutation
- QTL
- GWAS
- Evolve and re-sequence
- Domestication gene identification
- Recombination rate analysis

…

PoolSeq – some application

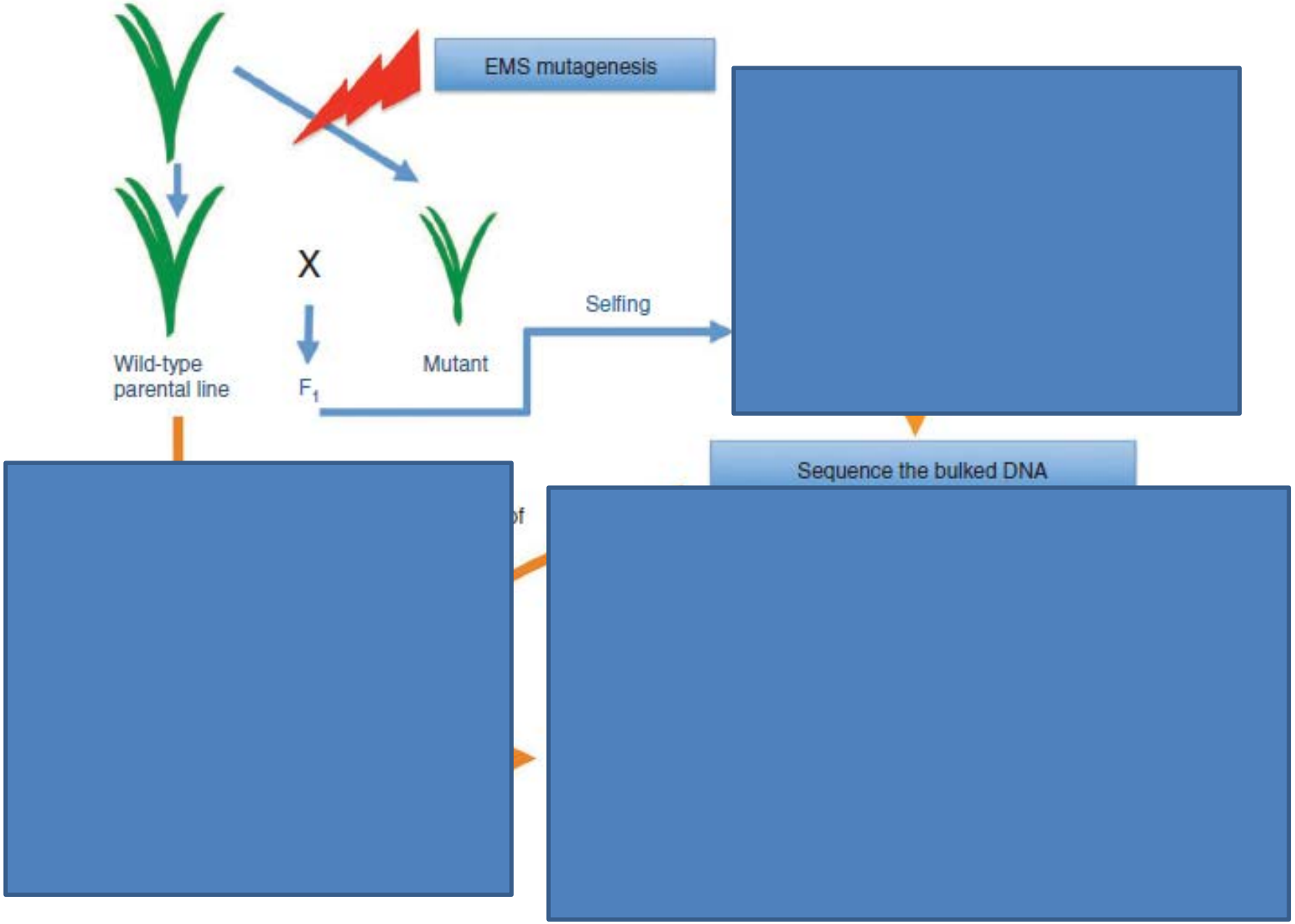
- Induced mutation
- QTL
- GWAS
- Evolve and re-sequence
- Domestication gene identification
- Recombination rate analysis

…



PoolSeq – MutMap for induced mutation

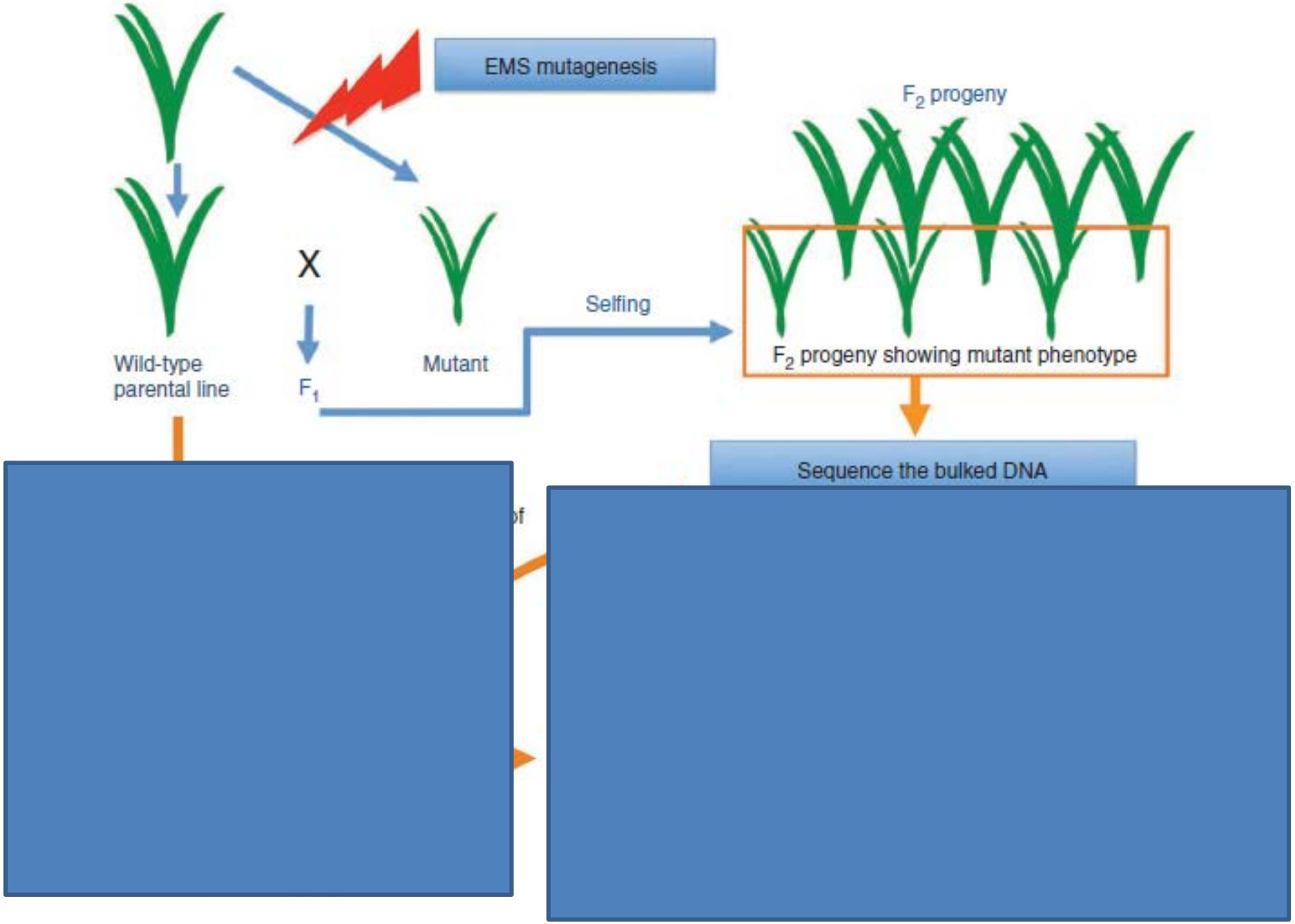
Induced mutation



Abe et al 2012

PoolSeq – MutMap for induced mutation

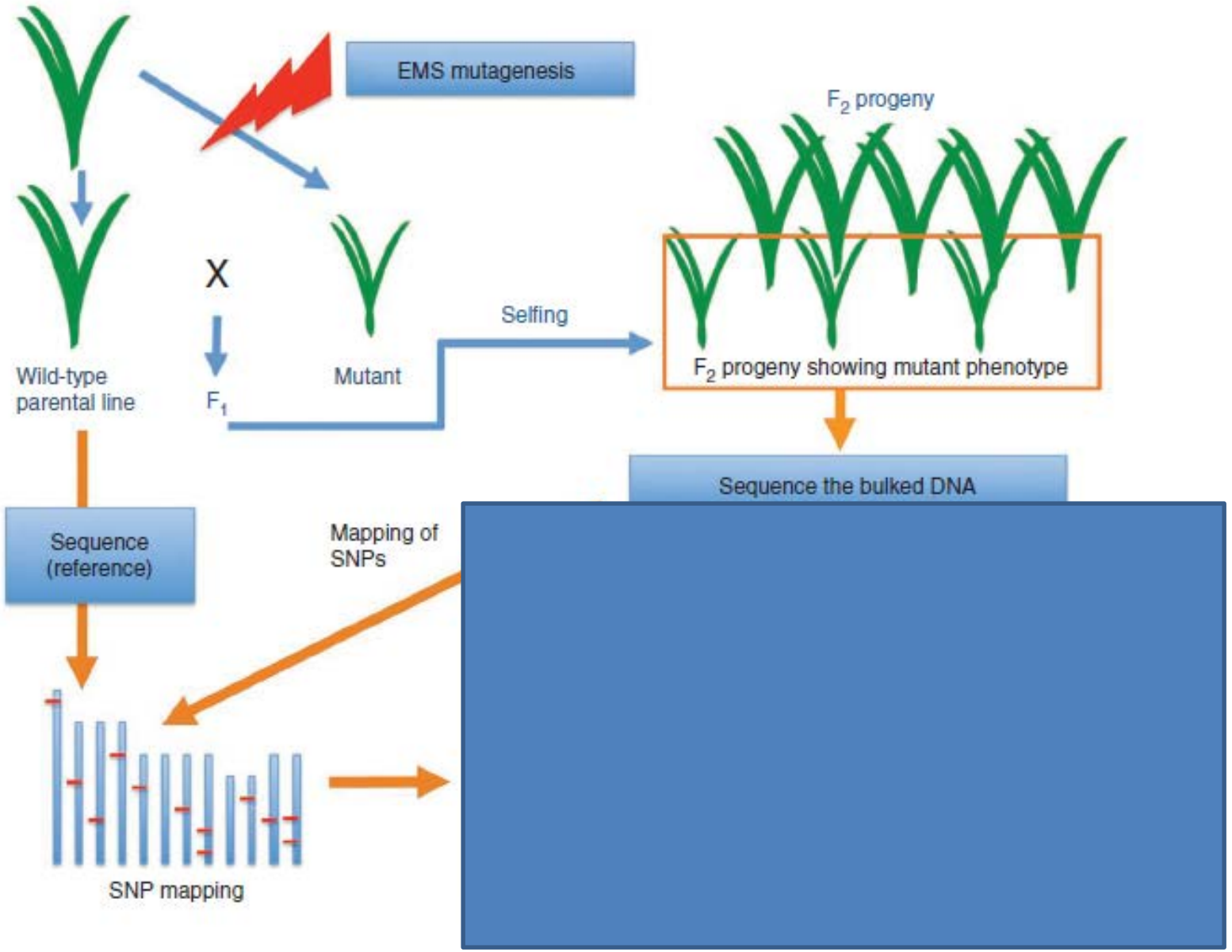
Induced mutation



Abe et al 2012

PoolSeq – MutMap for induced mutation

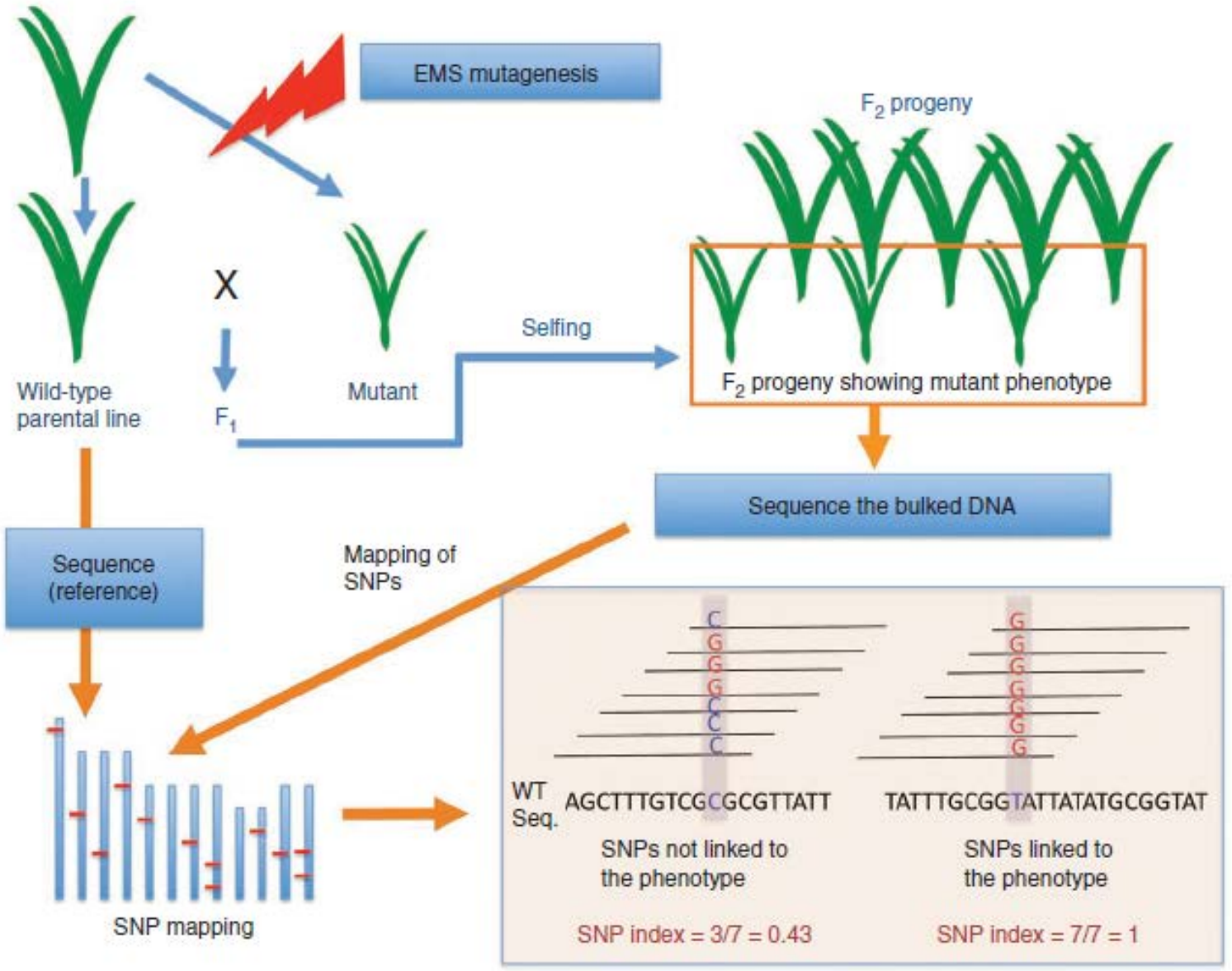
Induced mutation



Abe et al 2012

PoolSeq – MutMap for induced mutation

Induced mutation



Abe et al 2012

PoolSeq – MutMap for induced mutation

Induced mutation

A recessive mutant trait due to a single gene mutation

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

Some expectations

Induced mutation

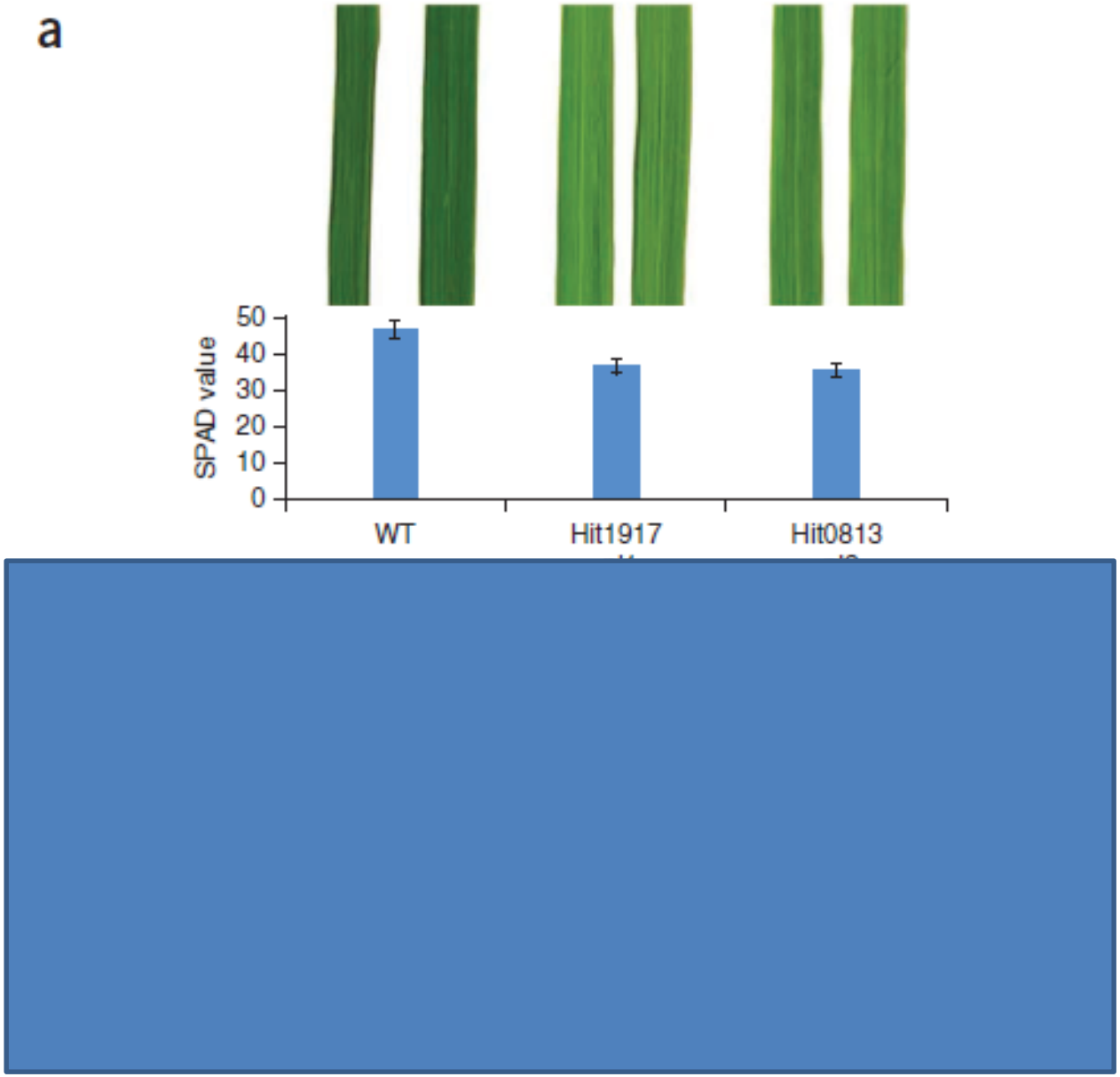
SNP index: The ratio of the number of reads containing a mutant SNP to the total number of reads containing the site of the SNP.

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

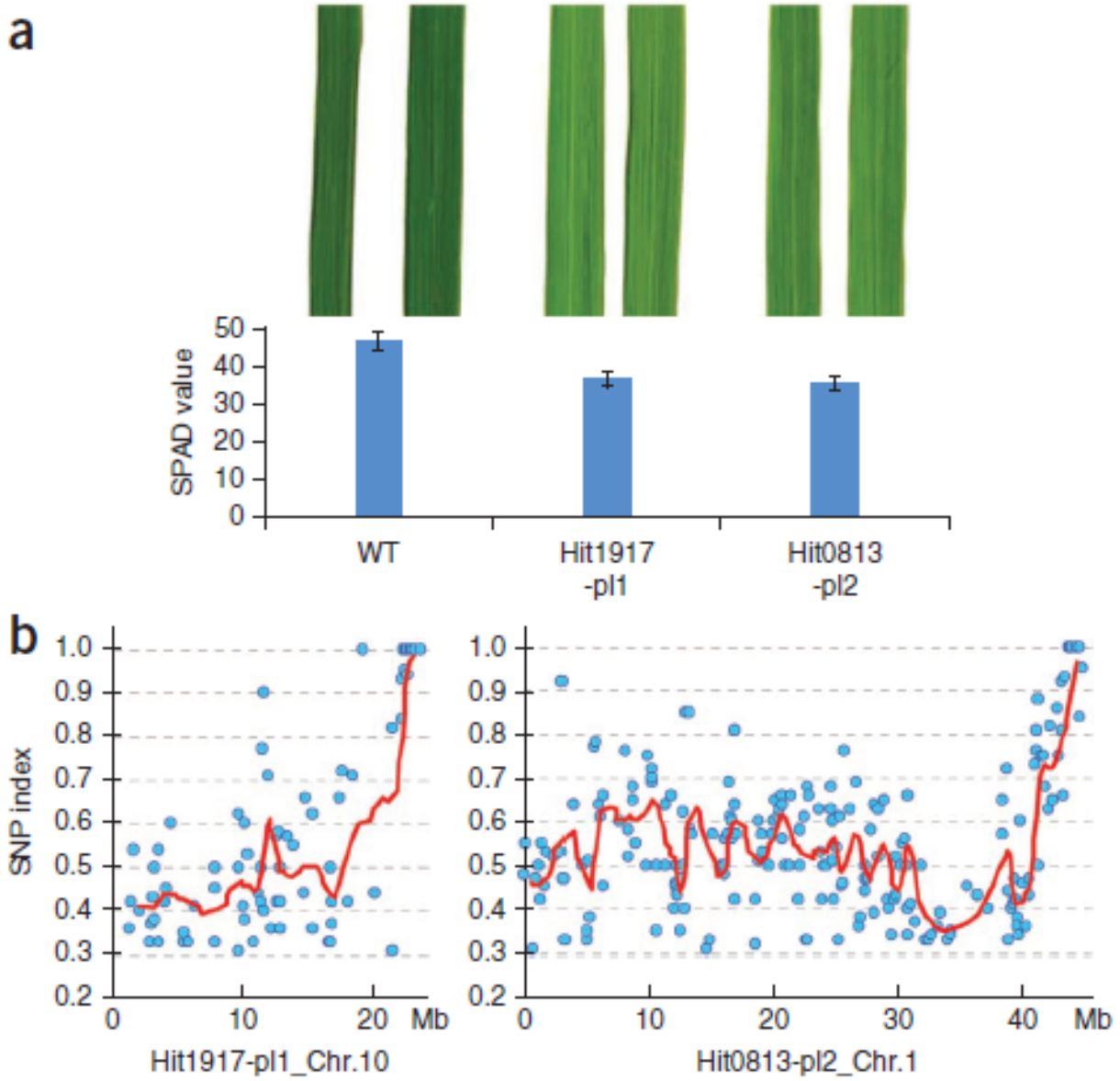
Induced mutation



Abe et al 2012

PoolSeq – MutMap for induced mutation

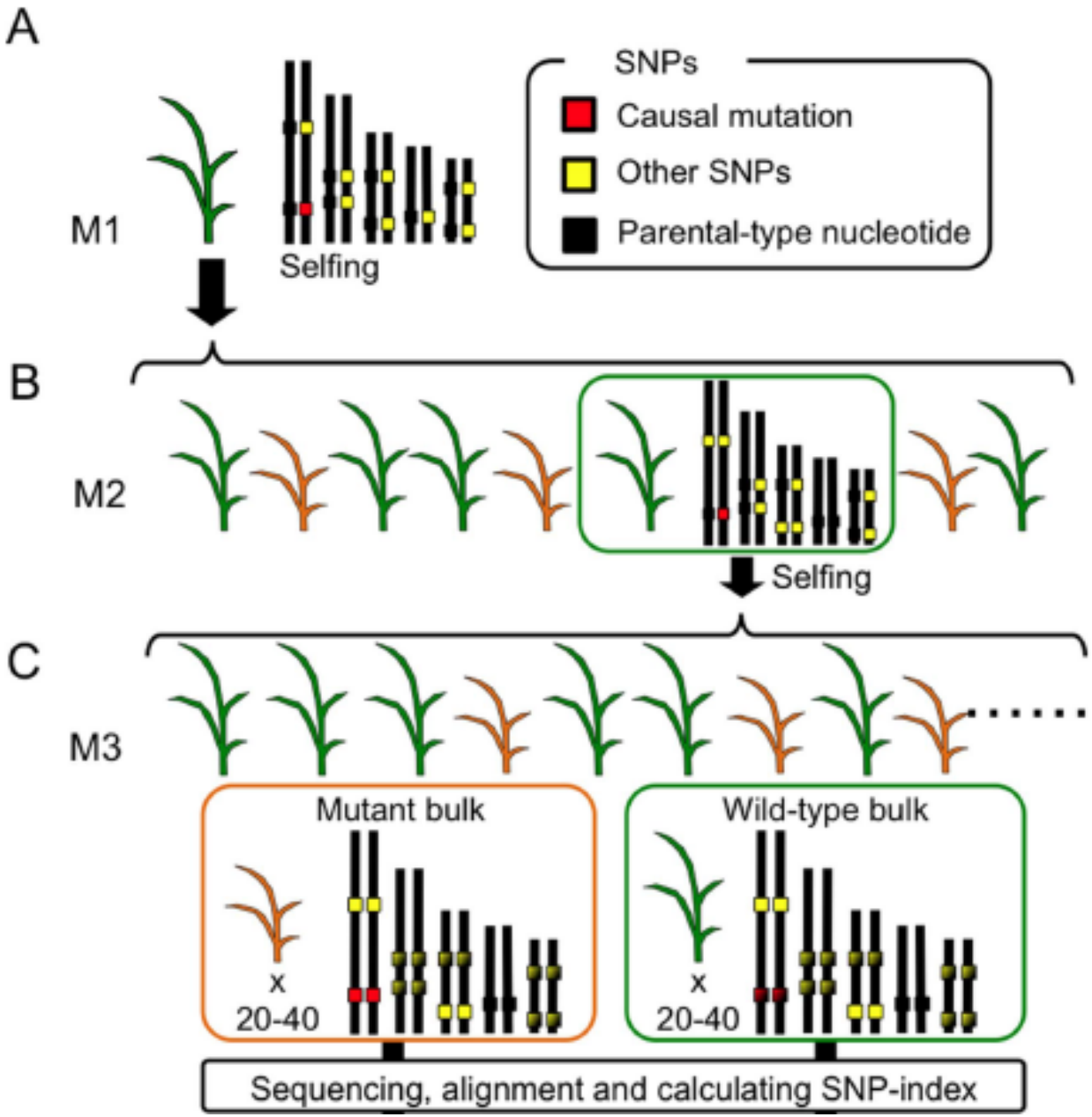
Induced mutation



Abe et al 2012

Variant: MutMap+ →No need to cross

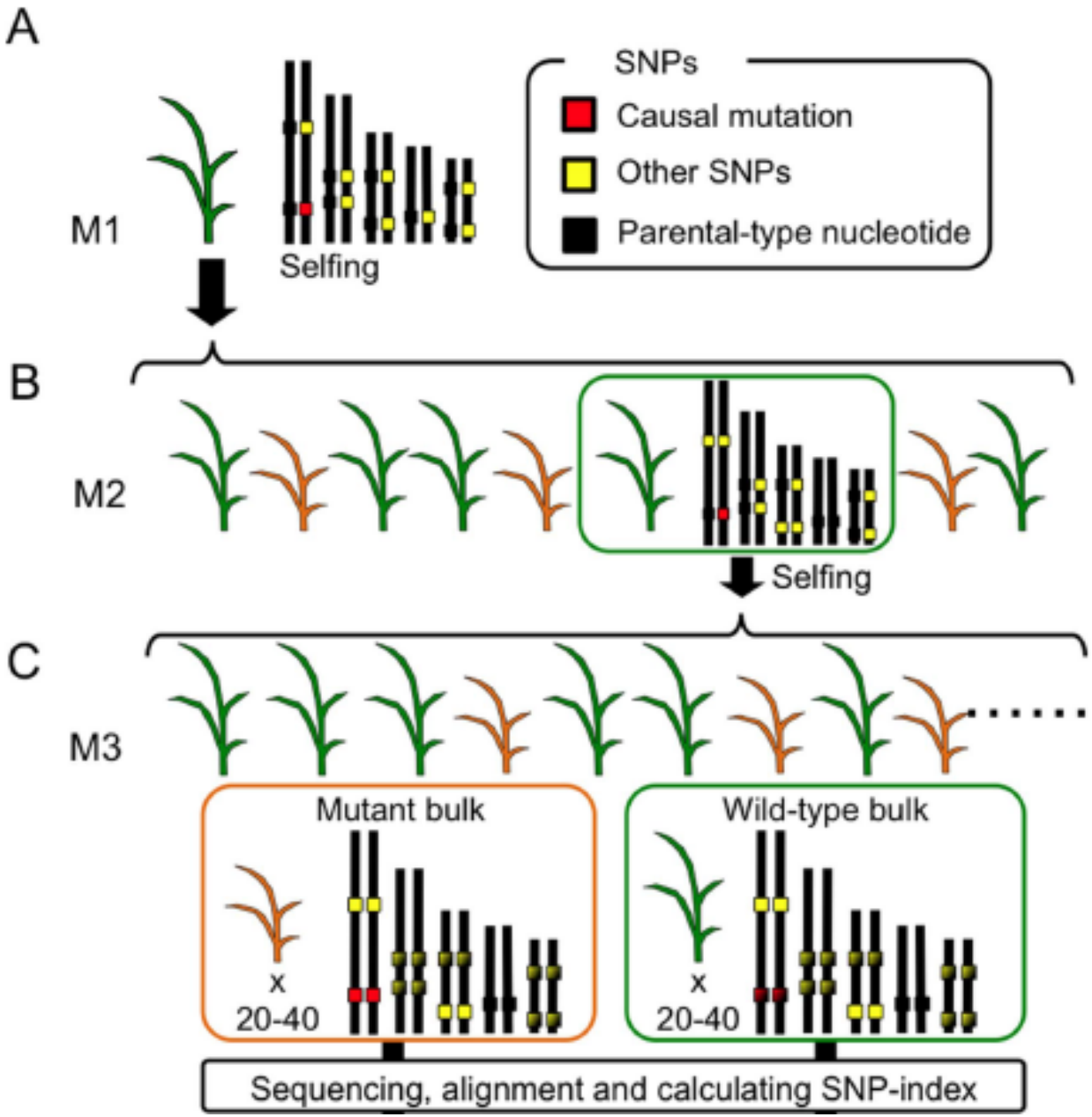
Induced mutation



Fekih et al 2013

Variant: MutMap+ →No need to cross

Induced mutation

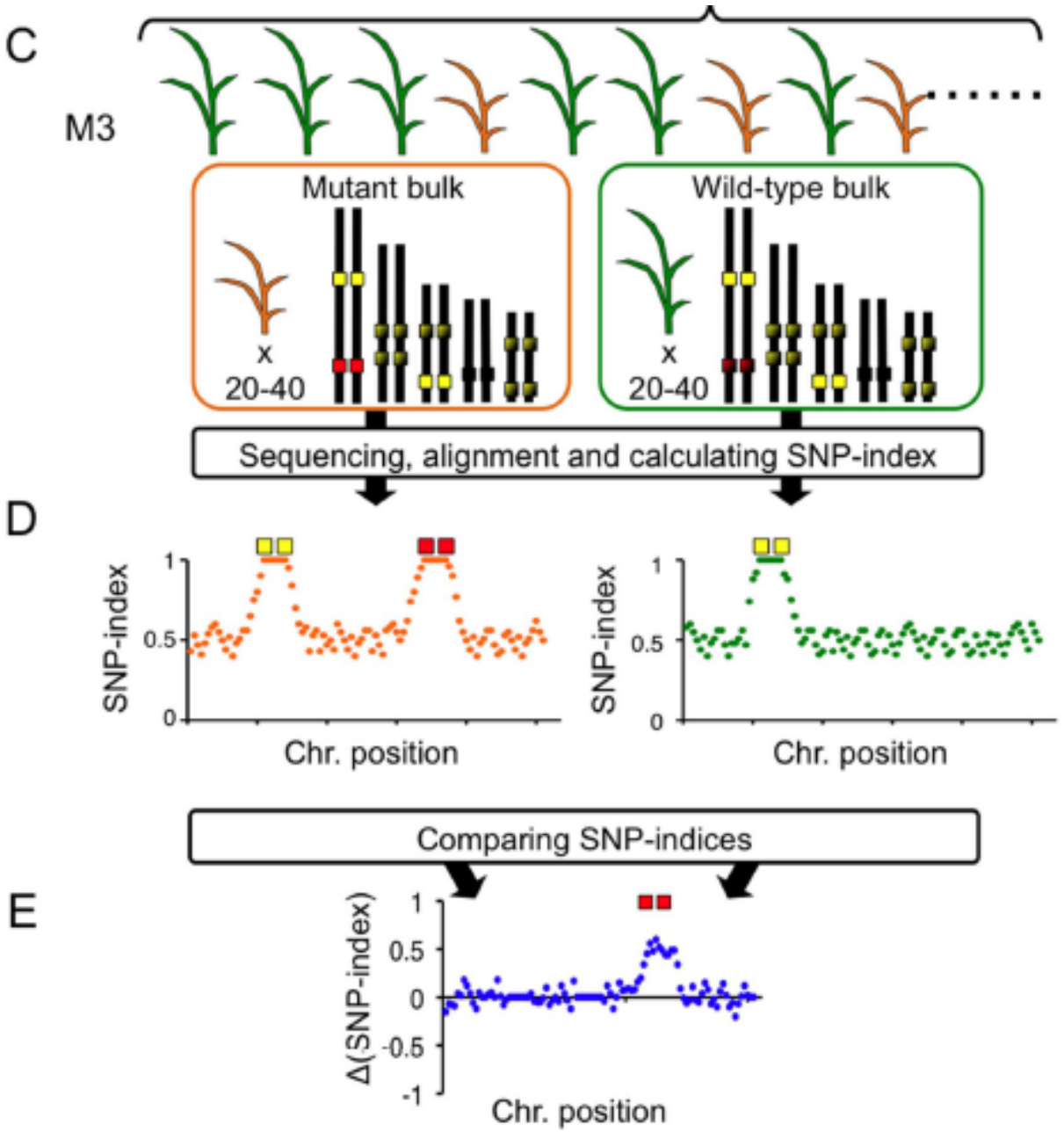


Some wt → all wt ignore

wt → wt + mut keep

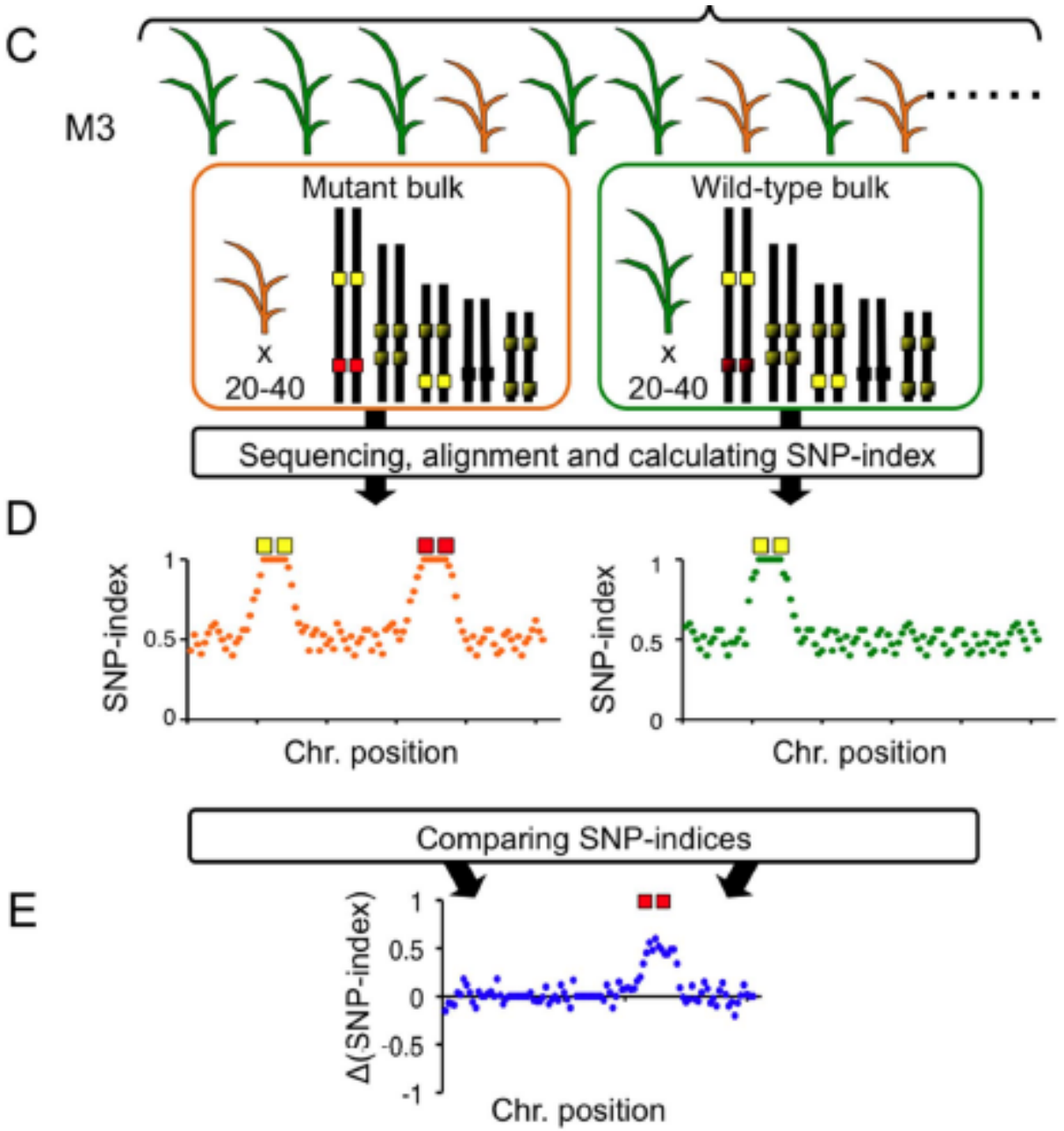
Fekih et al 2013

Induced mutation



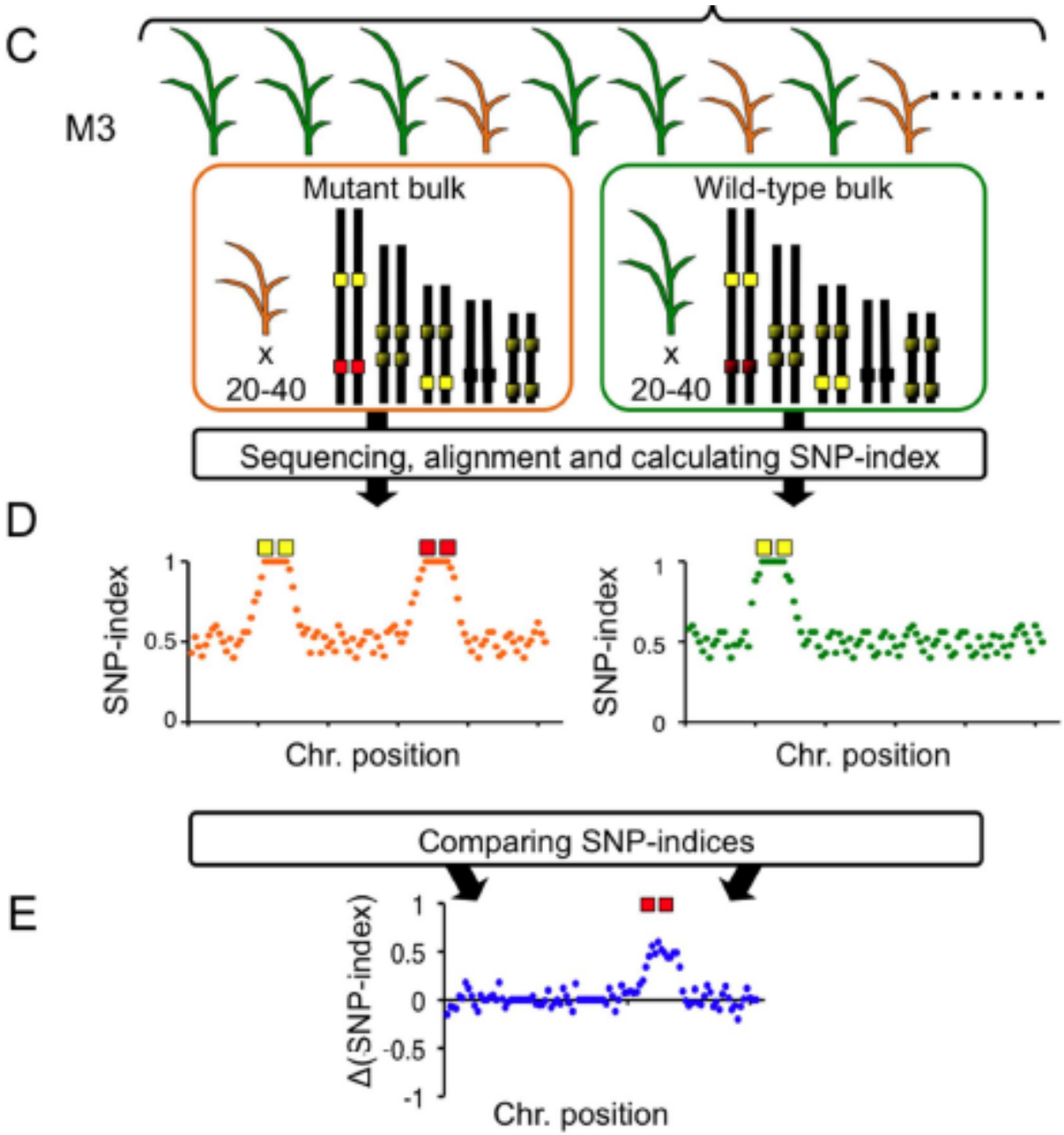
Fekih et al 2013

Induced mutation



Some SNPs fixed due to chance or drift

Fekih et al 2013



Induced mutation

Some SNPs fixed due to chance or drift

ΔSNP index

Fekih et al 2013

NGS and Poolseq

NGS and Poolseq

OLD:
many graduate students and many years

NGS and Poolseq

OLD:
many graduate students and many years

NEW:
single graduate student and fewer years

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

x-QTL

Highly complex traits → Multiple loci
Often each with small effect

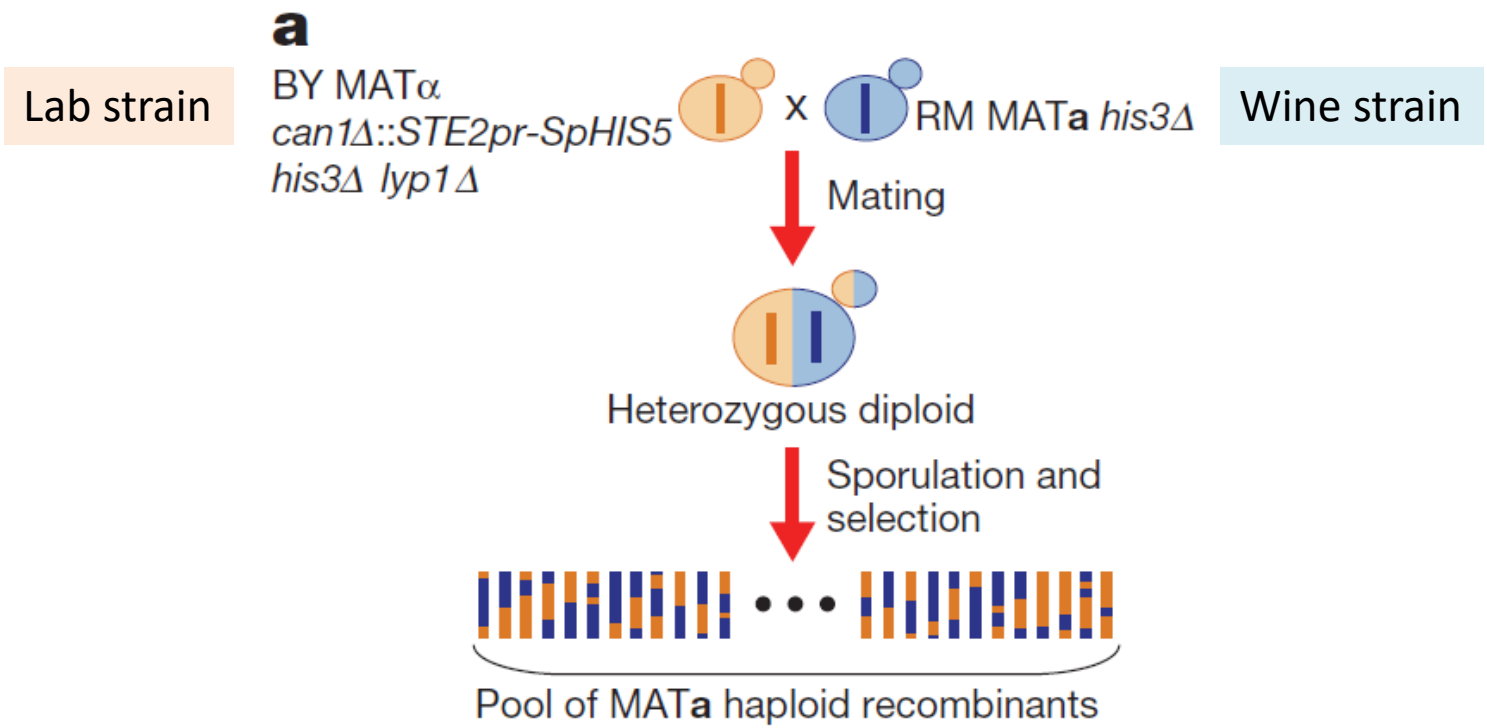
Need large populations for QTL mapping

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

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

x-QTL

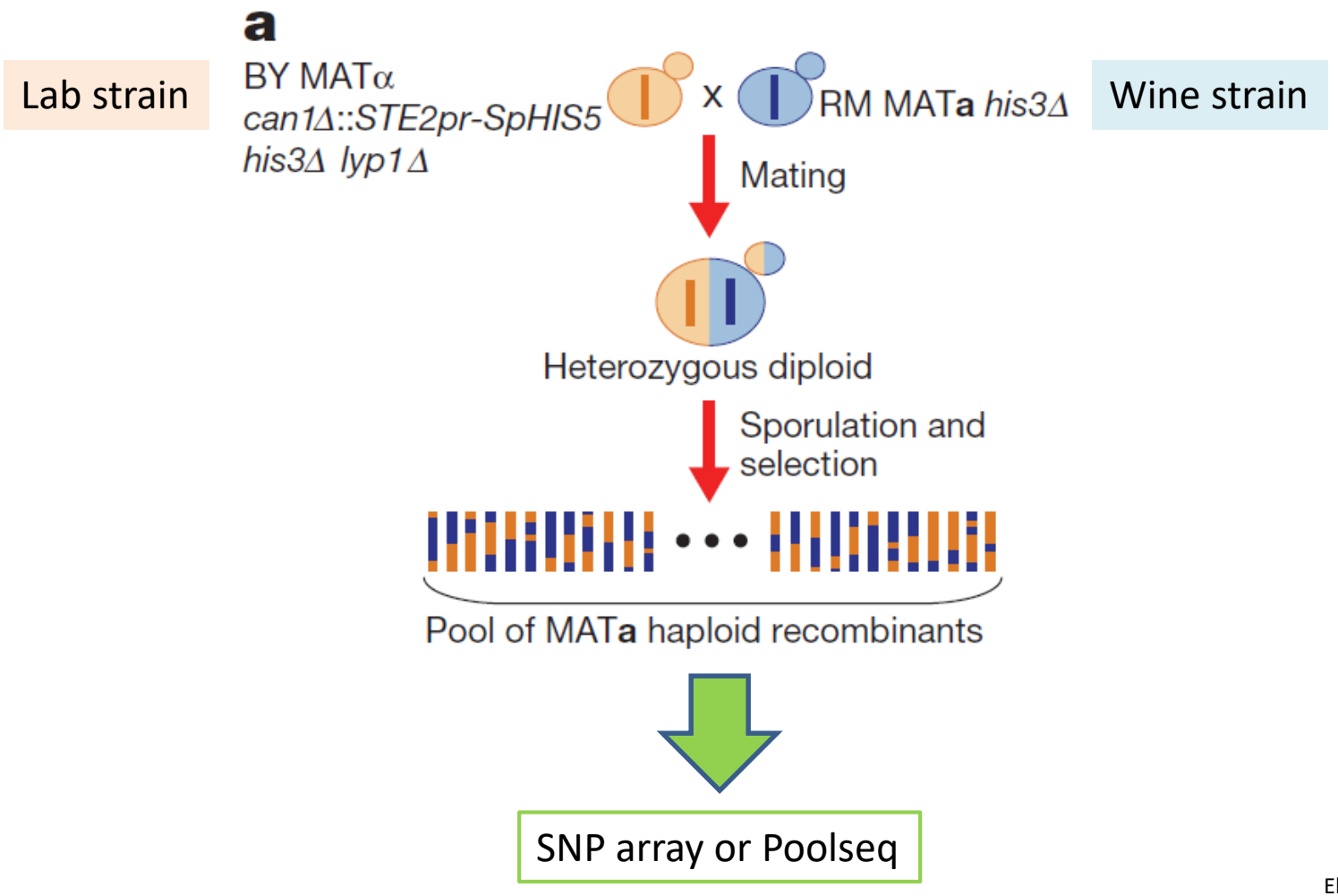
step 1. generate segregating population of very large size
step 2. selection based phenotyping (drug resistance or cell sorter)
step 3. measure pooled allele frequencies (NGS)



Ehrenreich et al 2010

x-QTL

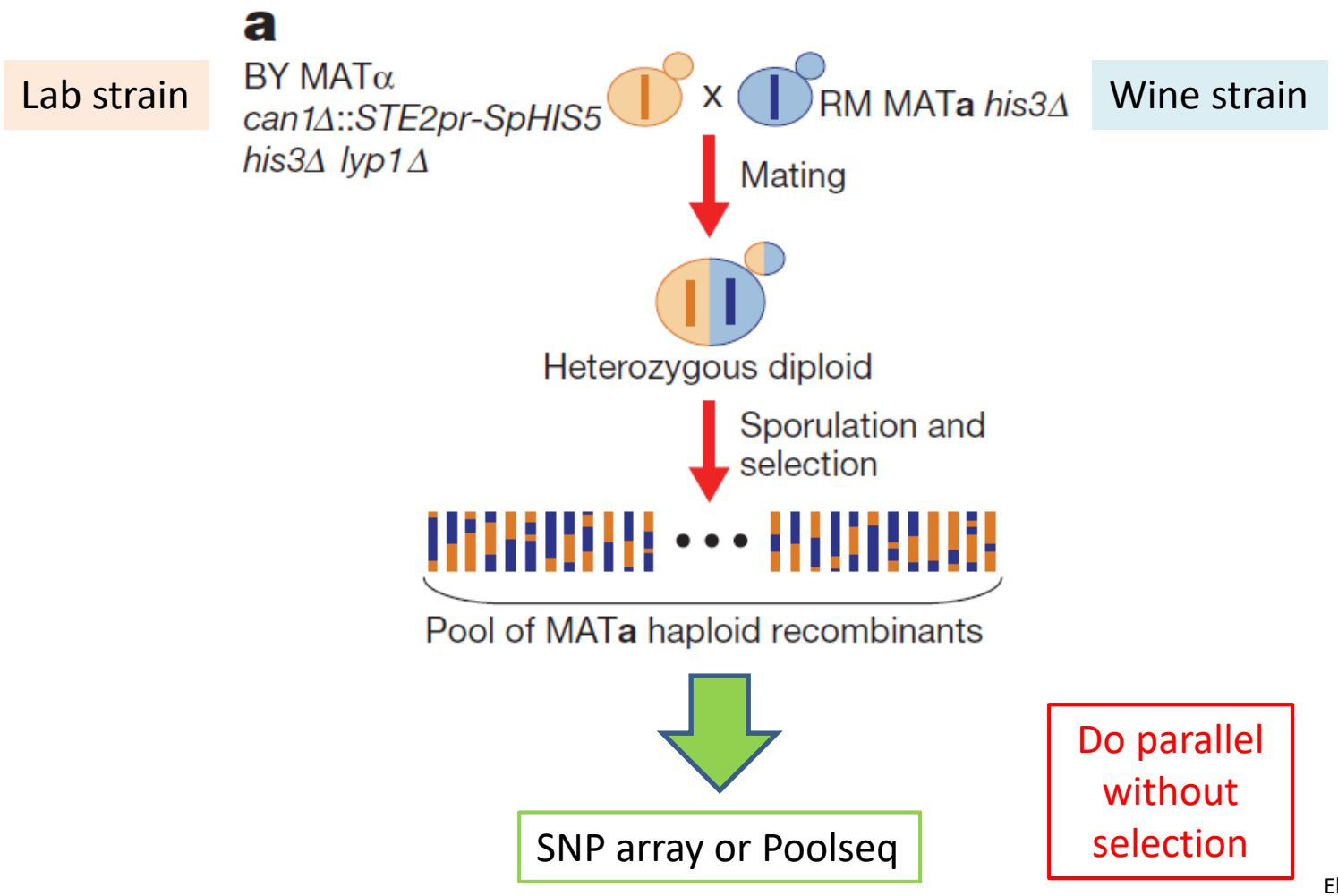
step 1. generate segregating population of very large size
step 2. selection based phenotyping (drug resistance or cell sorter)
step 3. measure pooled allele frequencies (NGS)



Ehrenreich et al 2010

x-QTL

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)



Ehrenreich et al 2010

4-NQO resistance case study

x-QTL

4-nitroquinoline
DNA damaging agent

Previously:

- Conventional QTL
  - RAD5 (DNA repair gene)

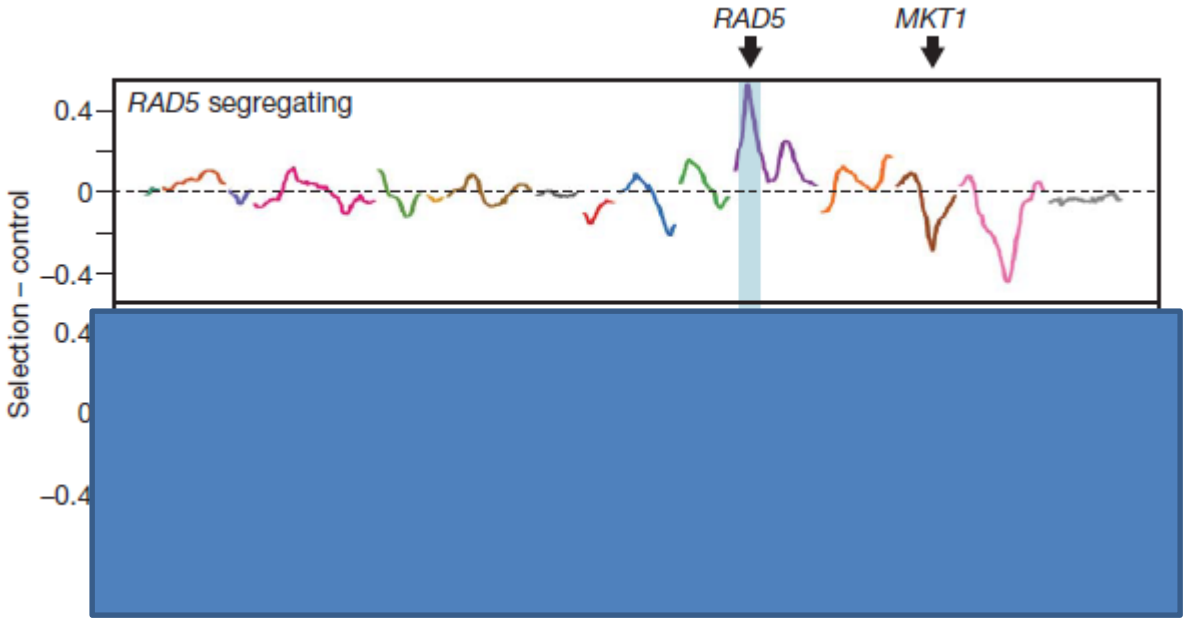
- Follow up backcrossing revealed a 2nd gene:
  - MKT1

Ehrenreich et al 2010

4-NQO resistance case study

x-QTL

14 loci 70%



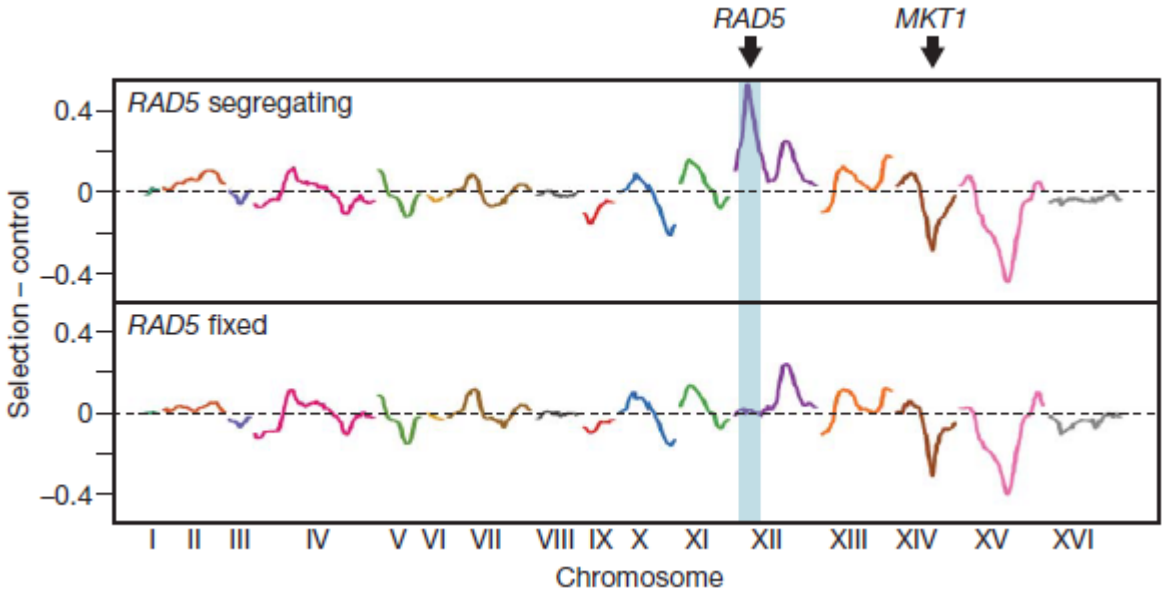
Ehrenreich et al 2010

4-NQO resistance case study

x-QTL

14 loci 70%

FIX RAD5



BY allele

RM allele

BY allele

RM allele

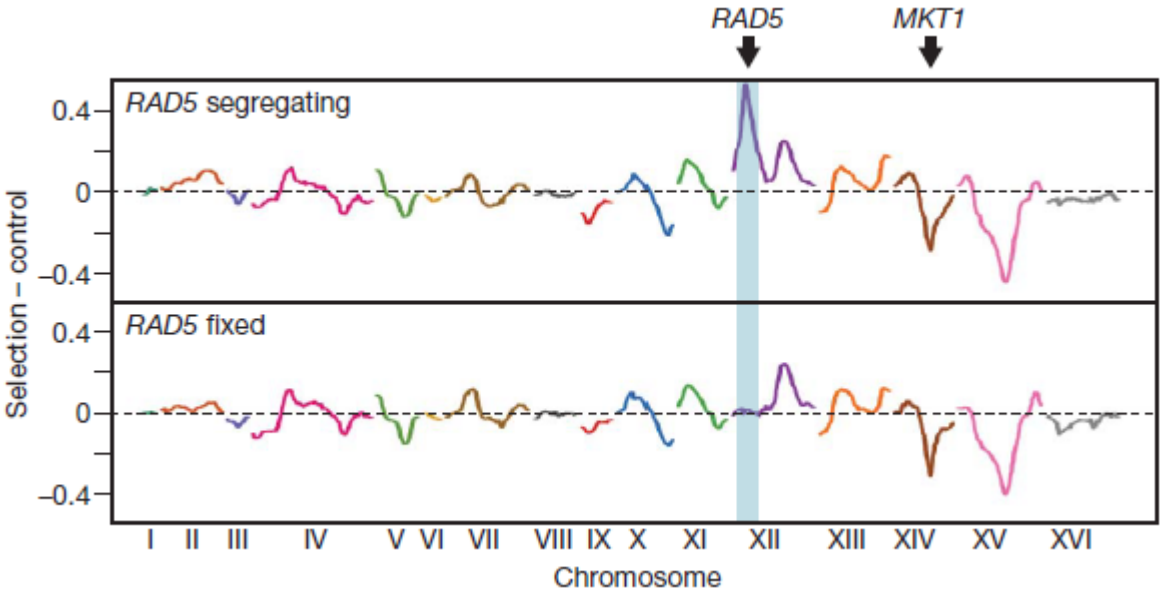
Ehrenreich et al 2010

4-NQO resistance case study

x-QTL

14 loci 70%

FIX RAD5



BY allele

RM allele

BY allele

RM allele

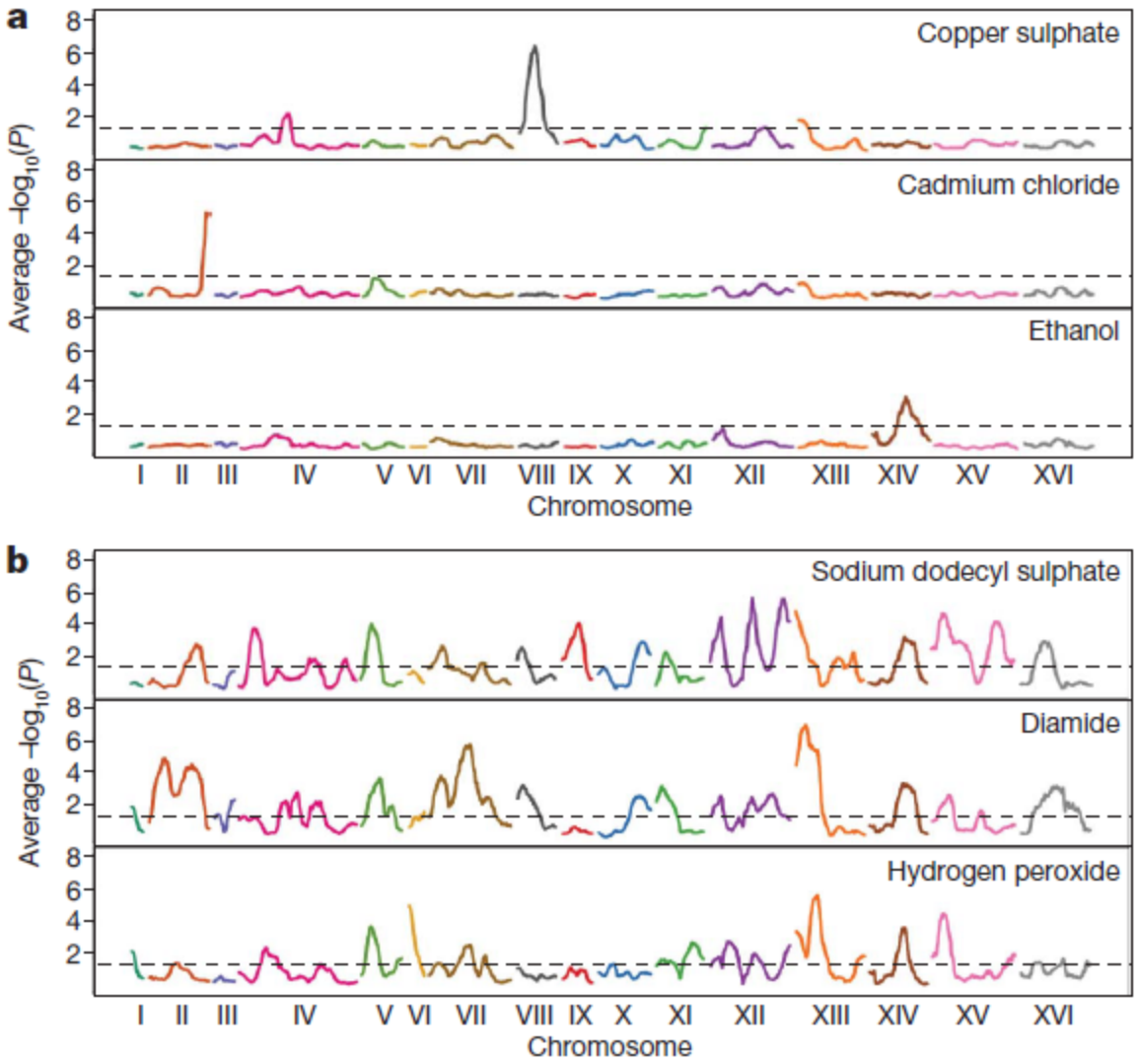
Very powerful

Basically, never uncover this with single gene approach

Ehrenreich et al 2010

Other examples of simple and complex traits

x-QTL



Ehrenreich et al 2010

xQTL mapping

is advertised to require big N

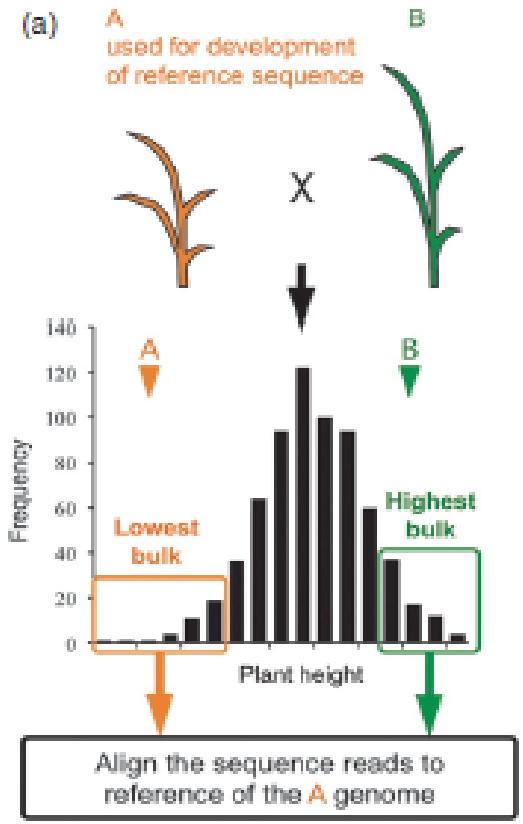
(more recombinants better resolution)

QTL-seq

results suggest only 20-50 can be informative

(although finding the exact gene is unclear)

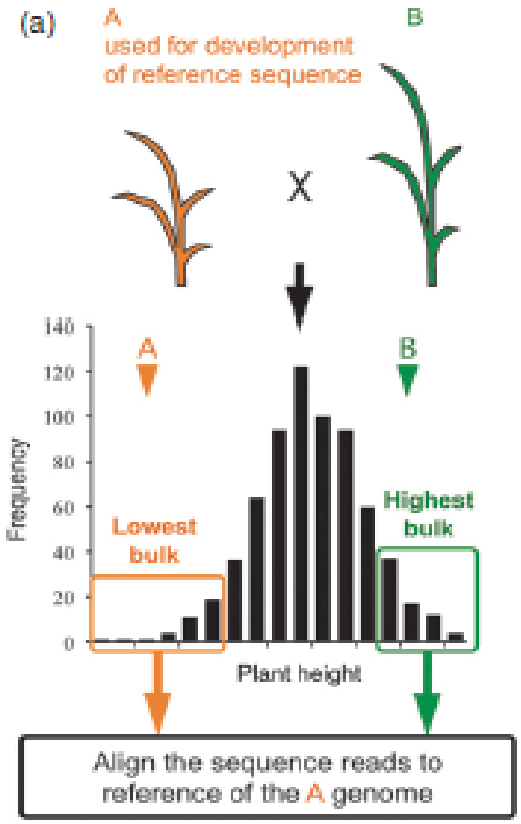
QTL-seq



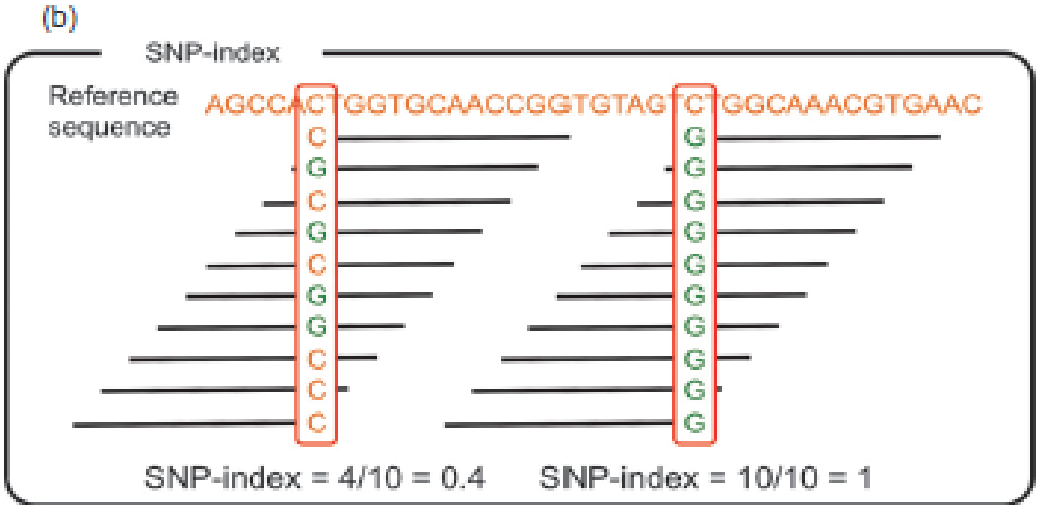
- N = 20-50
- Larger genome
- Works on crops
- ~MutMap with SNP index

Takagi et al 2013

QTL-seq

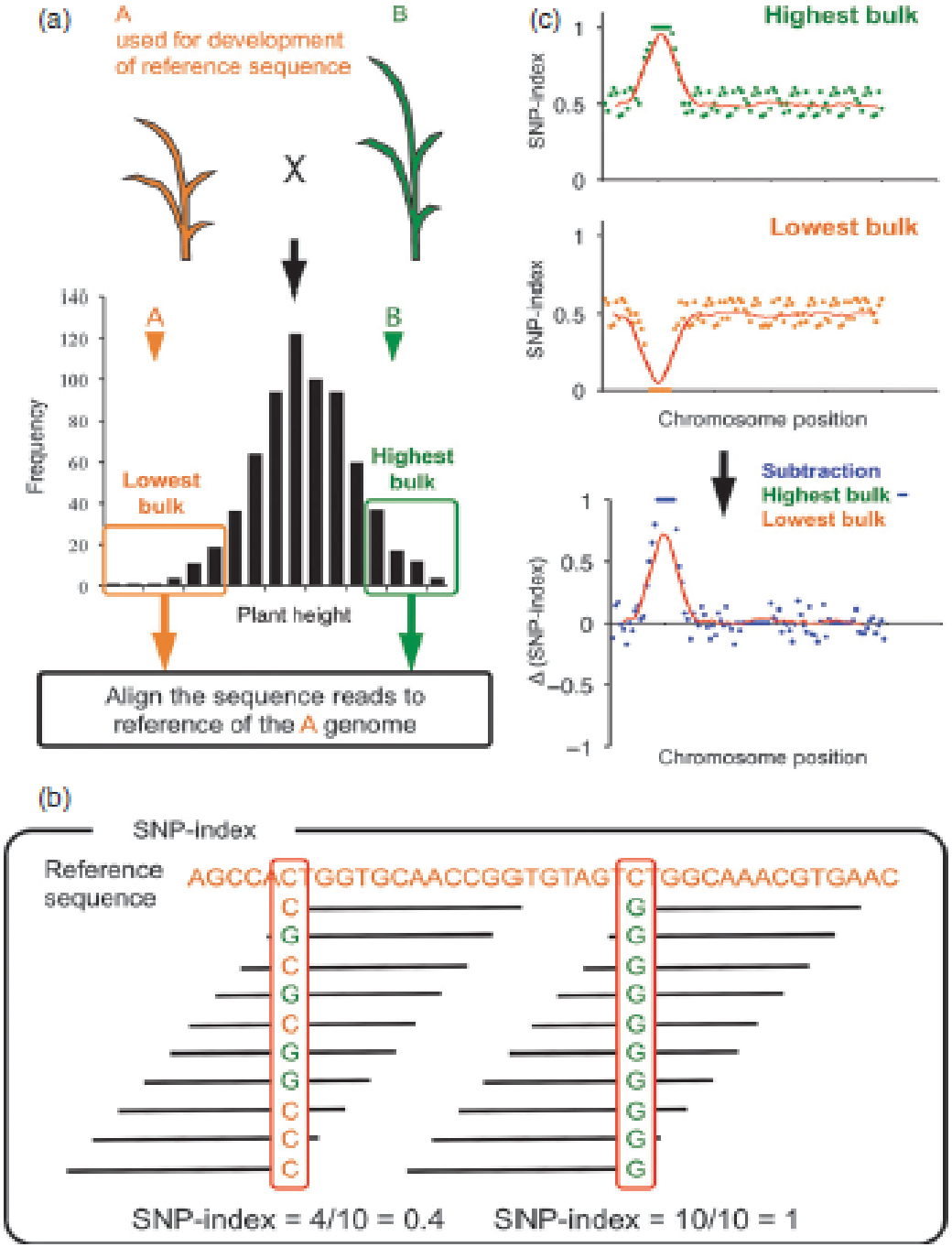


- N = 20-50
- Larger genome
- Works on crops
- ~MutMap with SNP index



Takagi et al 2013

QTL-seq



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

Takagi et al 2013

Case study: seed vigor

QTL-seq

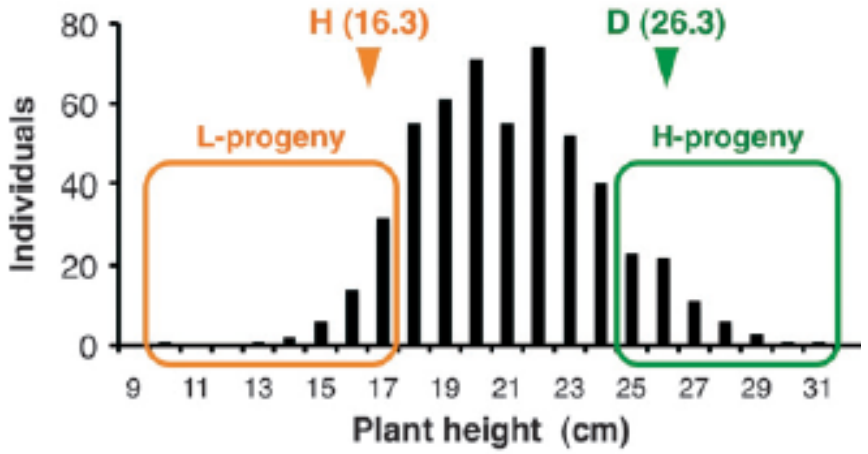
(a)

Case study: seed vigor

QTL-seq

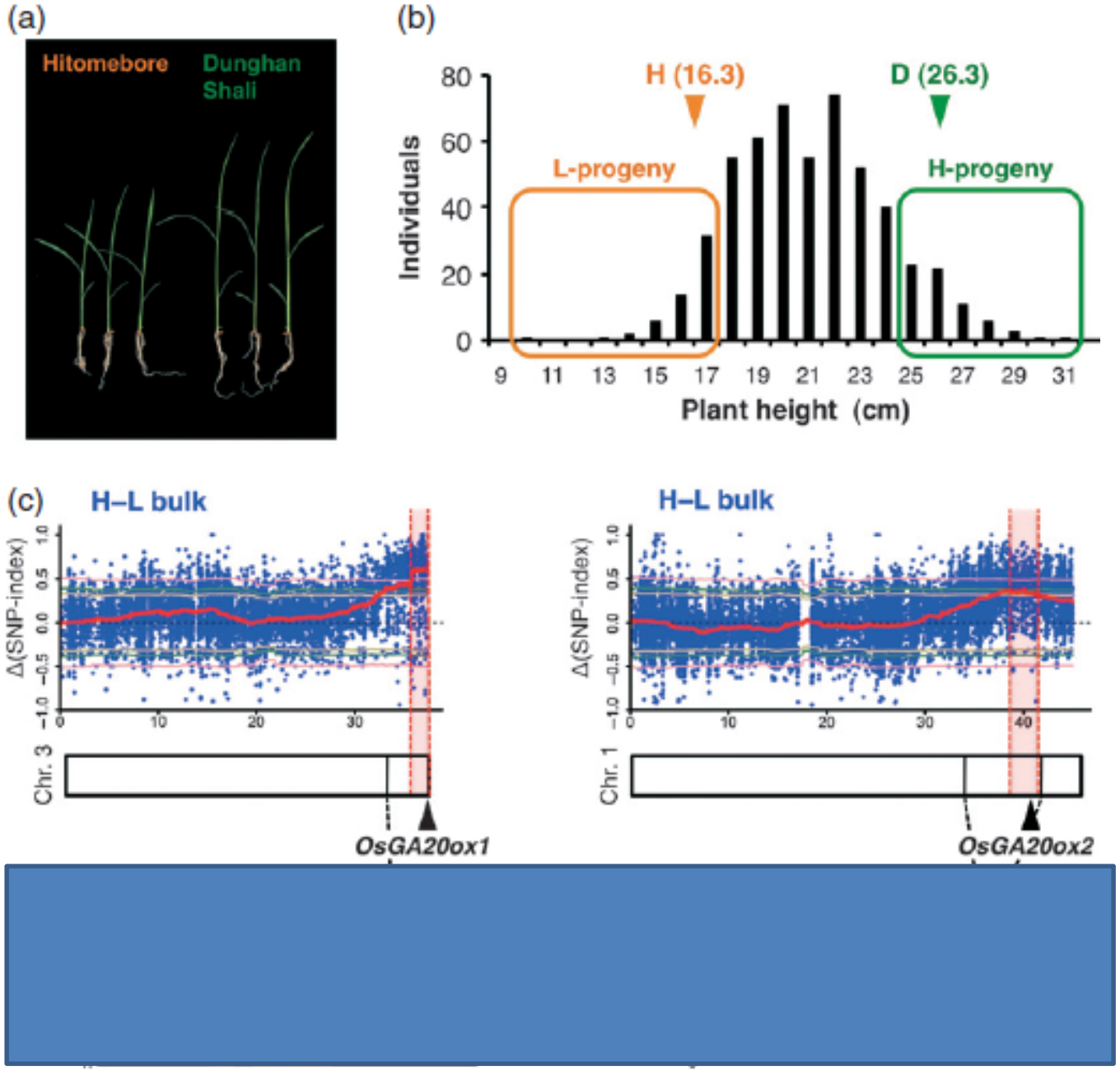
(a)

(b)



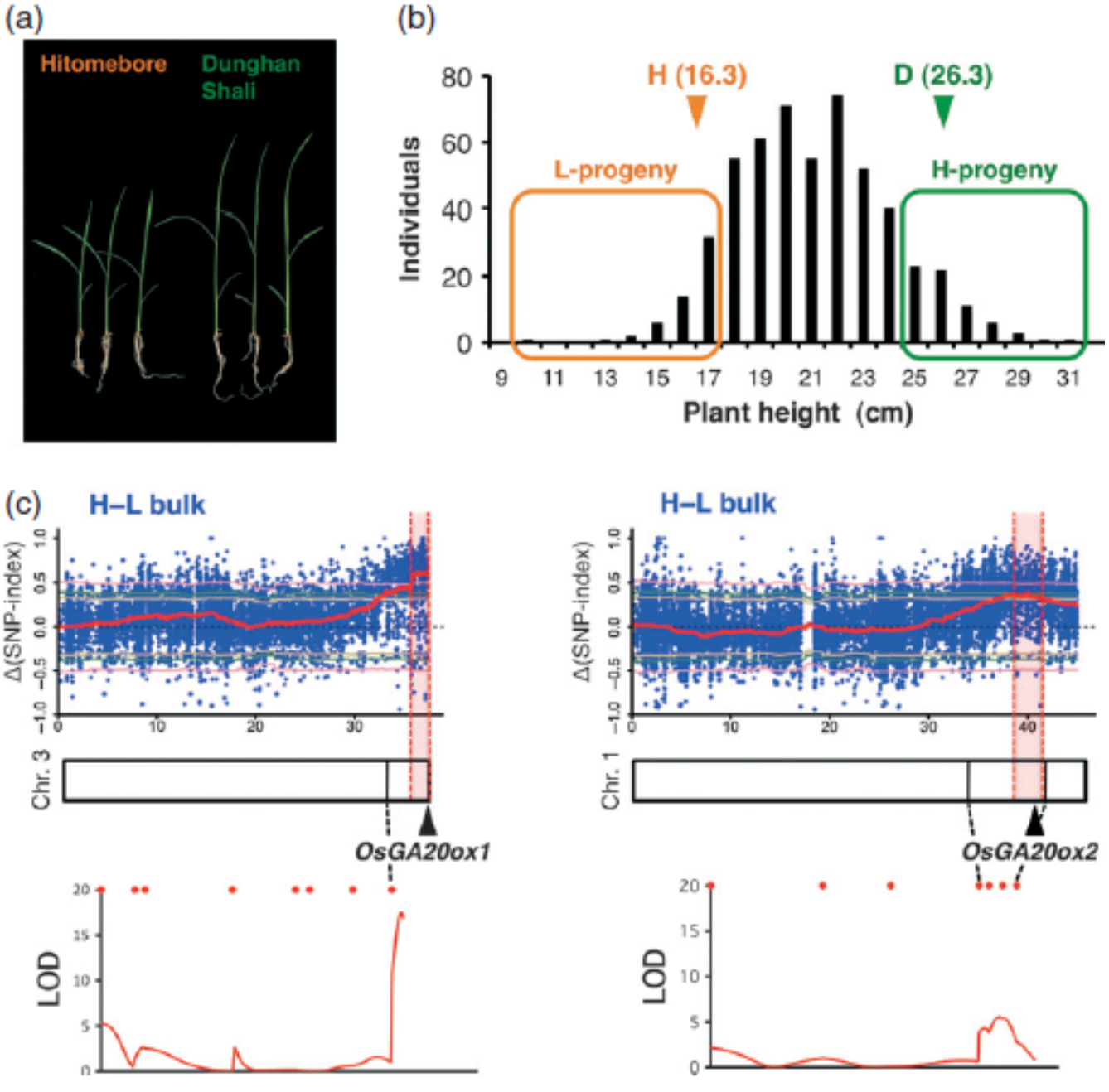
Case study: seed vigor

QTL-seq



Case study: seed vigor

QTL-seq



Poolseq

GWAS

taking advantage of historical recombination

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

Poolseq

GWAS

taking advantage of historical recombination

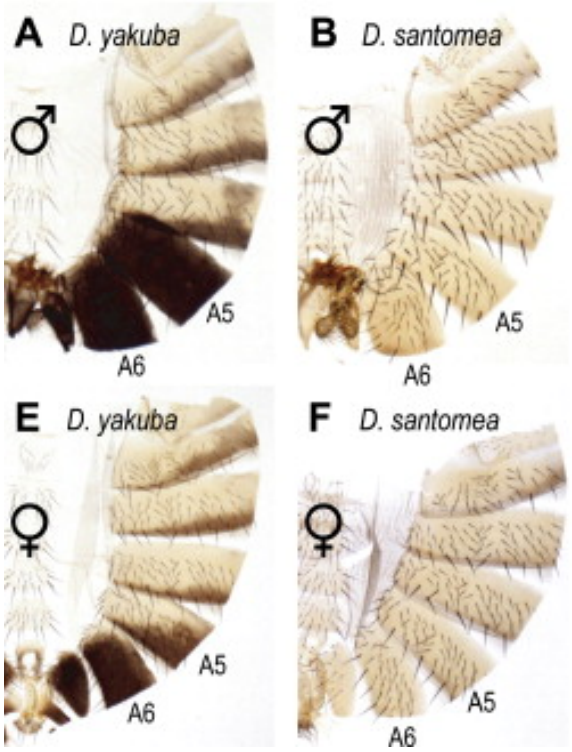
Basically, the more recombinants, the better map resolution.
F2 crosses or Recombinant Inbred Lines have limited recombination events.
Wild samples have history to help

Drosophila body pigmentation example

Drosophila pigmentation background

D. yakuba x *D. santomea*

- There is variation in abdominal pigmentation
- Probably sexual selection



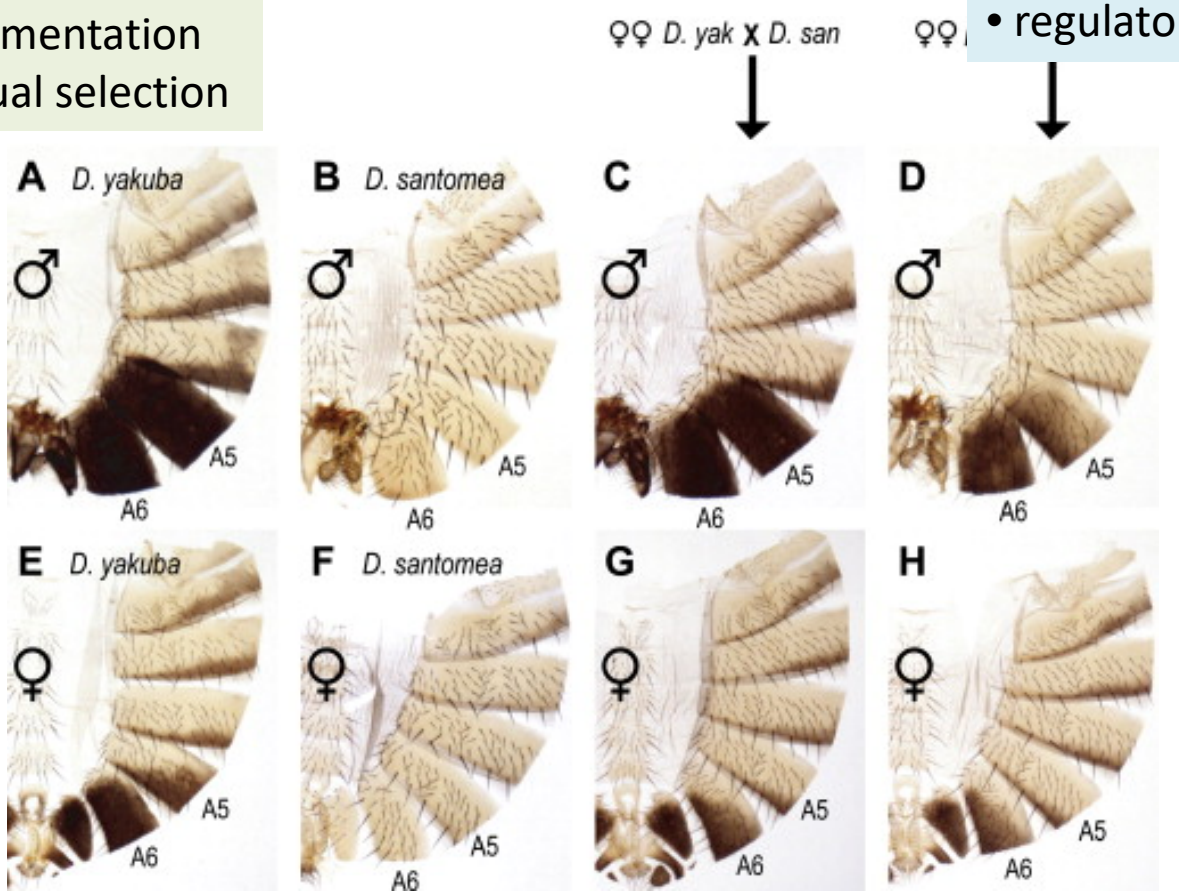
Jeong et al 2008

Drosophila pigmentation background

D. yakuba x *D. santomea*

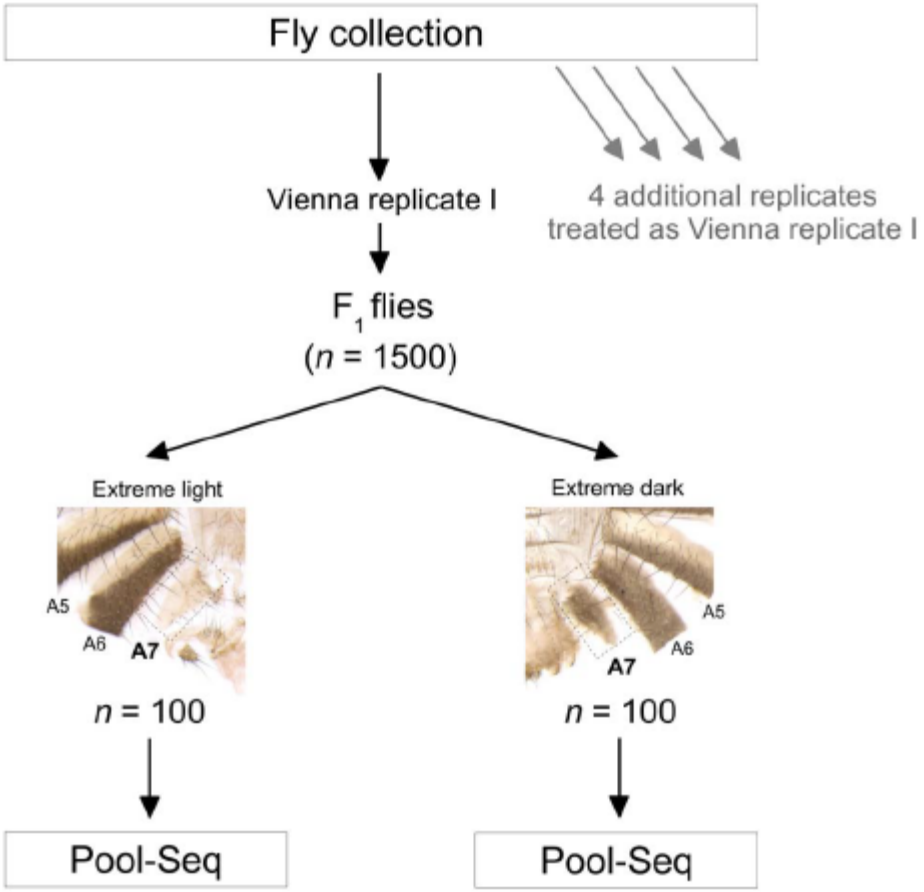
•There is variation in abdominal pigmentation
•Probably sexual selection

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



Jeong et al 2008

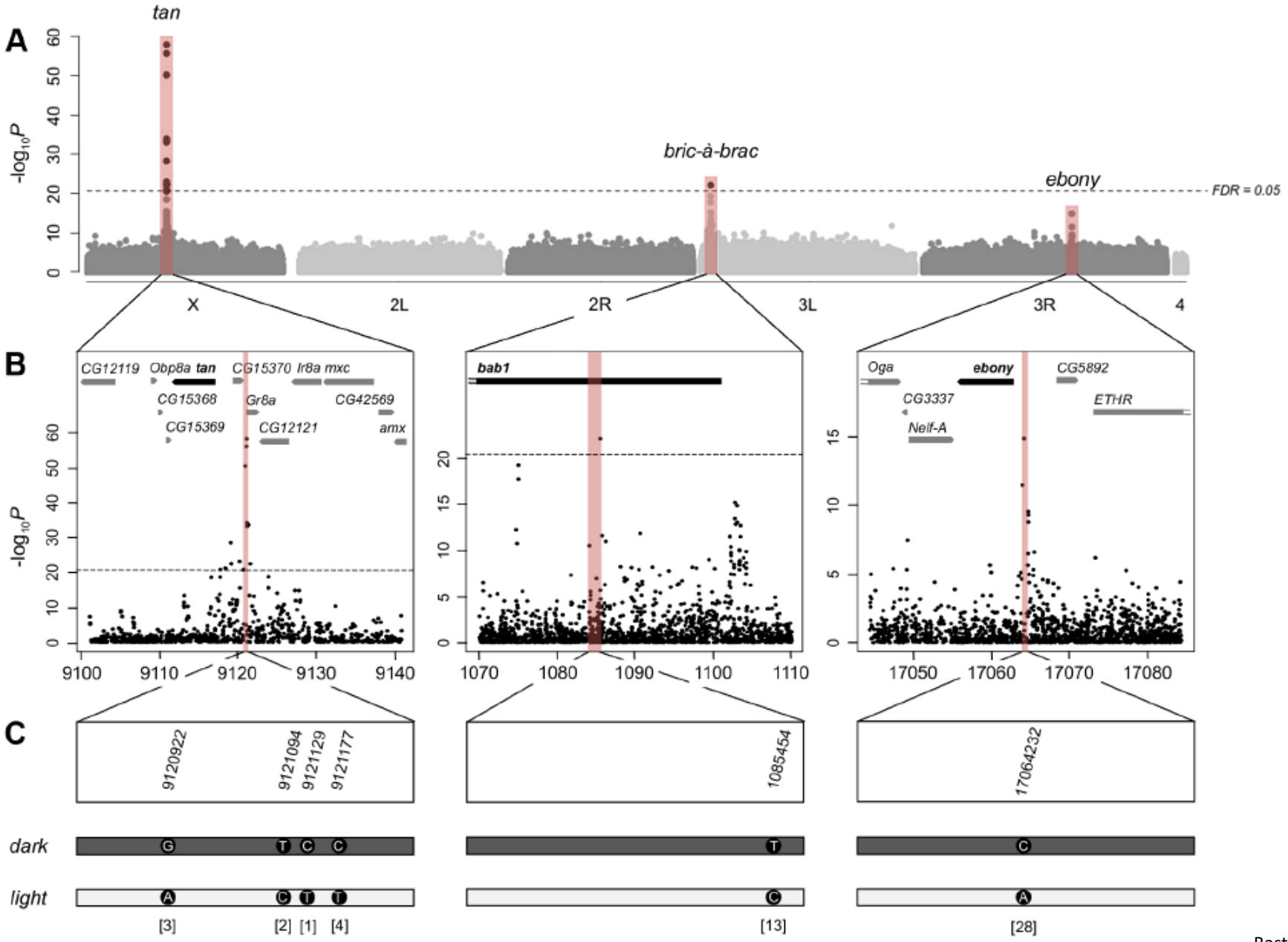
D. melanogaster → natural variation in pigmentation



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

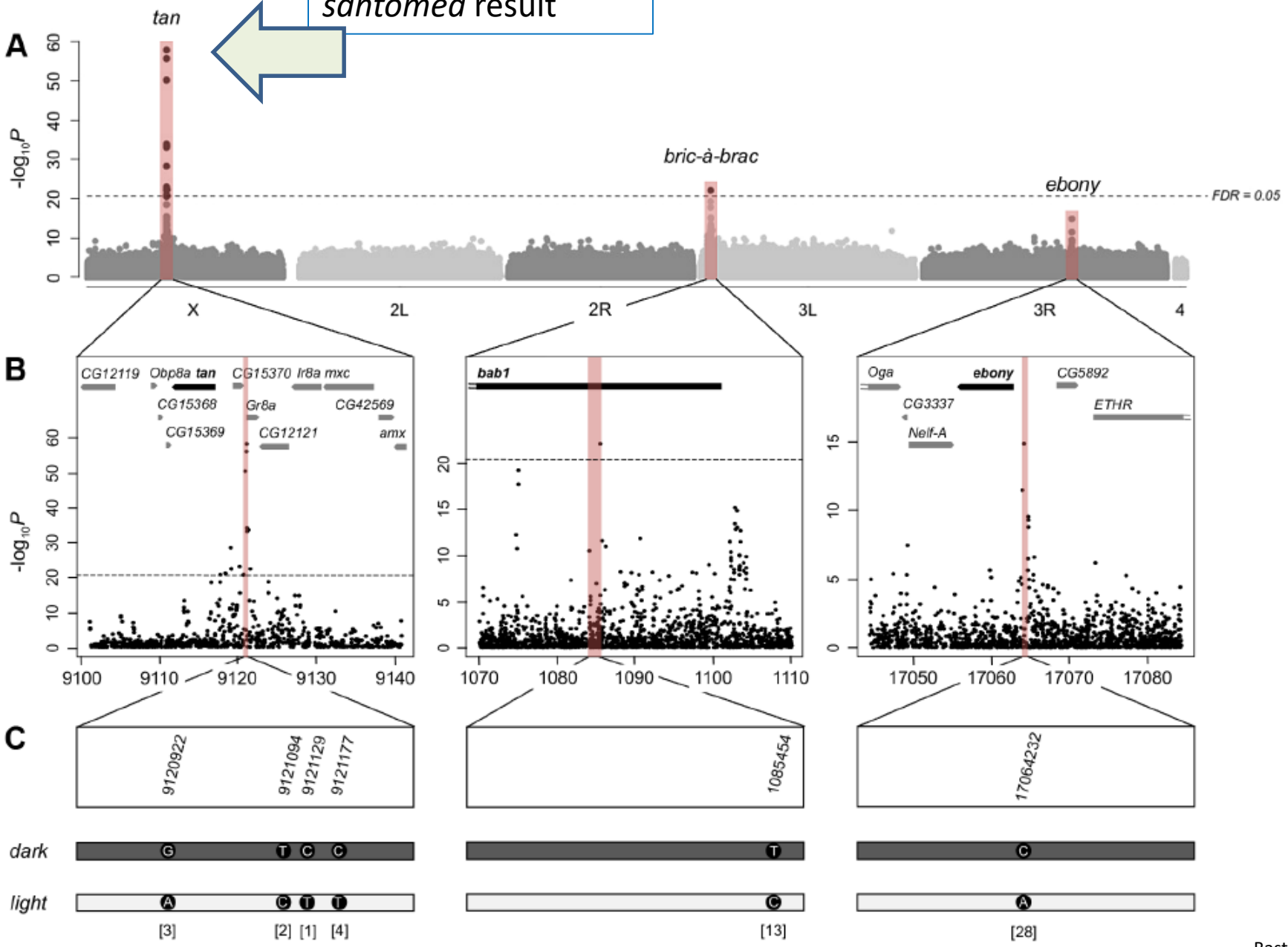
Grow in lab controlled conditions 1 gen
Sort by color
Take extremes
Pool-seq with replicates

Bastide et al 2013

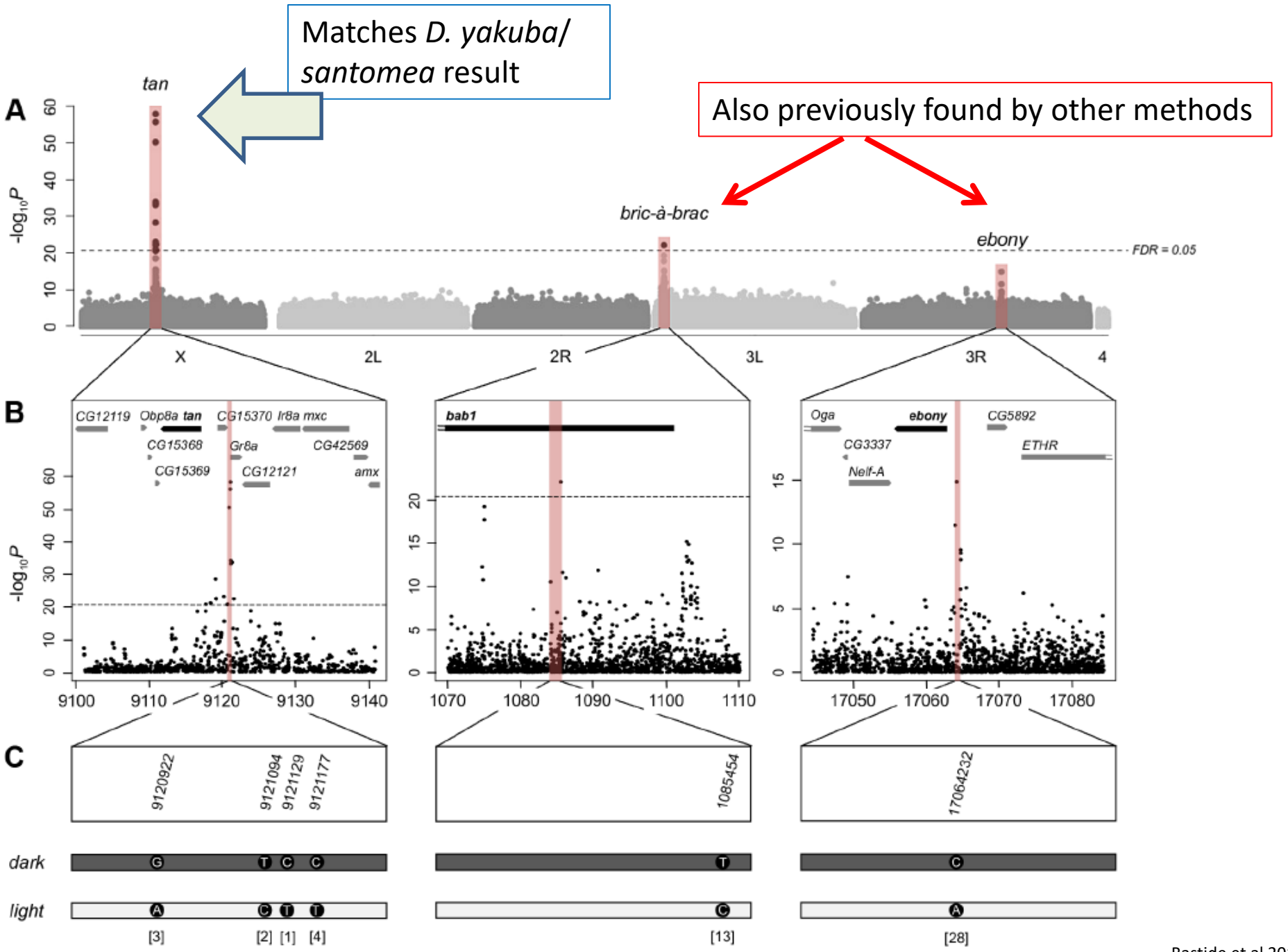


Bastide et al 2013

Matches *D. yakuba/ santomea* result



Bastide et al 2013

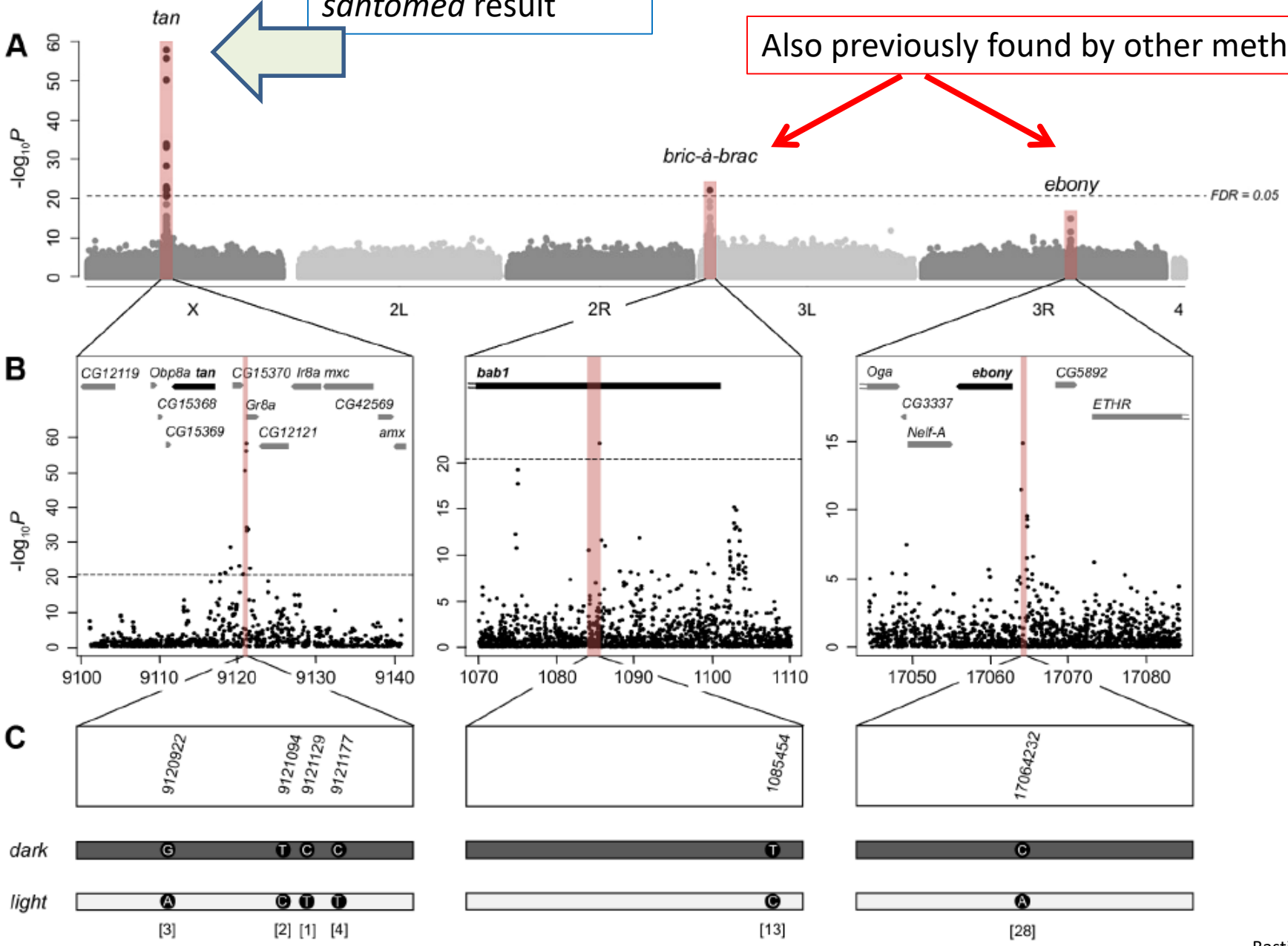


Bastide et al 2013

Matches *D. yakuba/ santomea* result

All regulatory

Also previously found by other methods



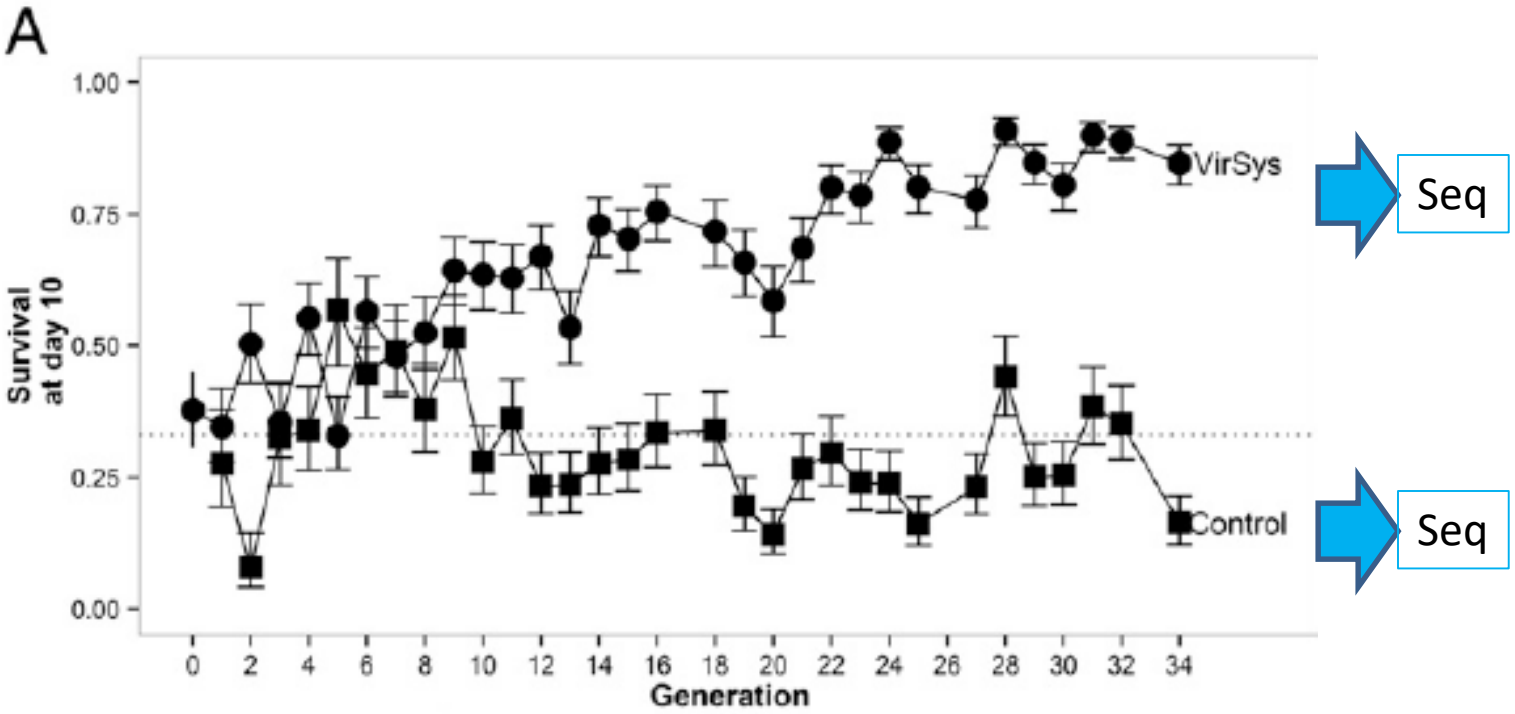
Bastide et al 2013

Evolve and re-sequence

Evolve and re-sequence

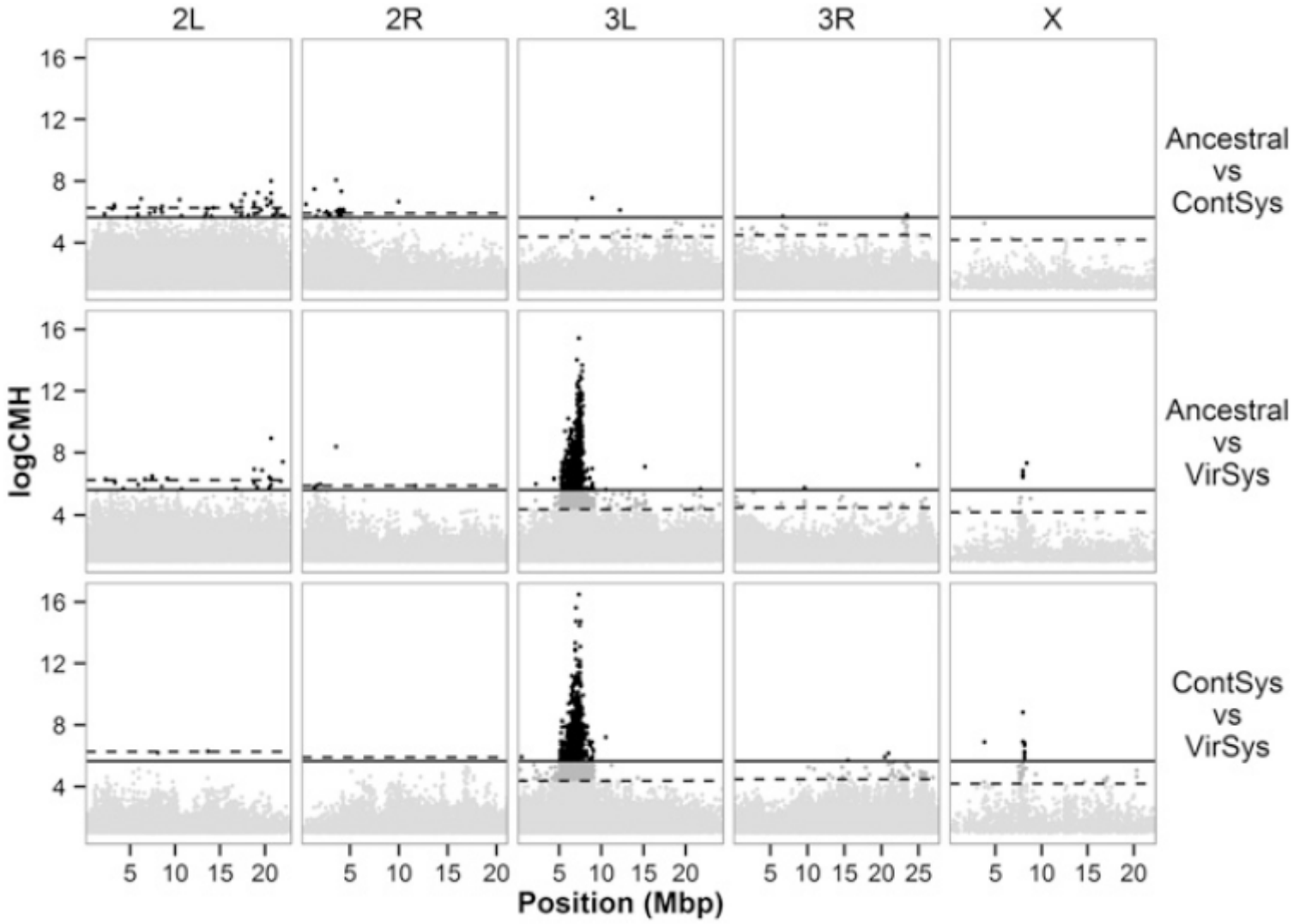
Evolve *Drosophila* lines to be more resistant to C virus

4x replicates [virus, control1, control2]



Martins et al 2014

Evolve and re-sequence



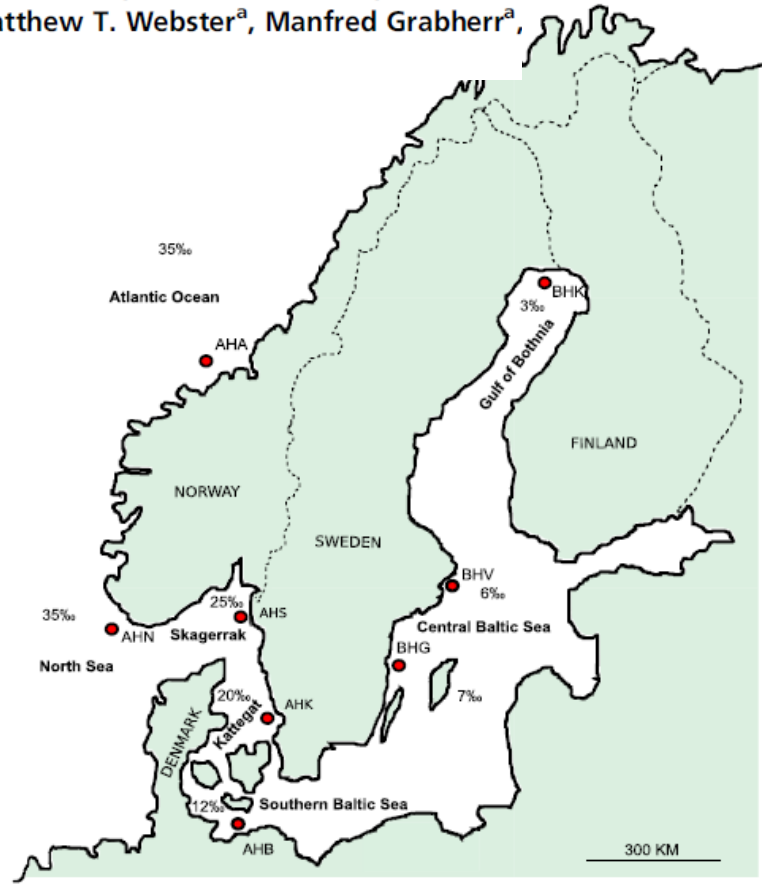
pastrel gene (and lesser extent *Ubc-E2H*)

Martins et al 2014

Ecology (reverse ecology) and local adaptation

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

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



Ecology (reverse ecology) and local adaptation

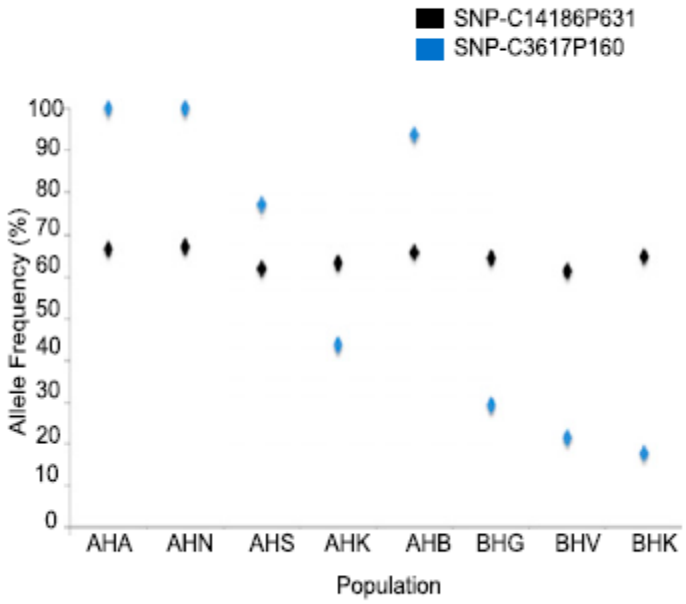
8 Pool-seq populations
Each with 50 individuals
30x coverage
Map against transcriptome

Lamichhaney et al 2012

Ecology (reverse ecology) and local adaptation

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

A

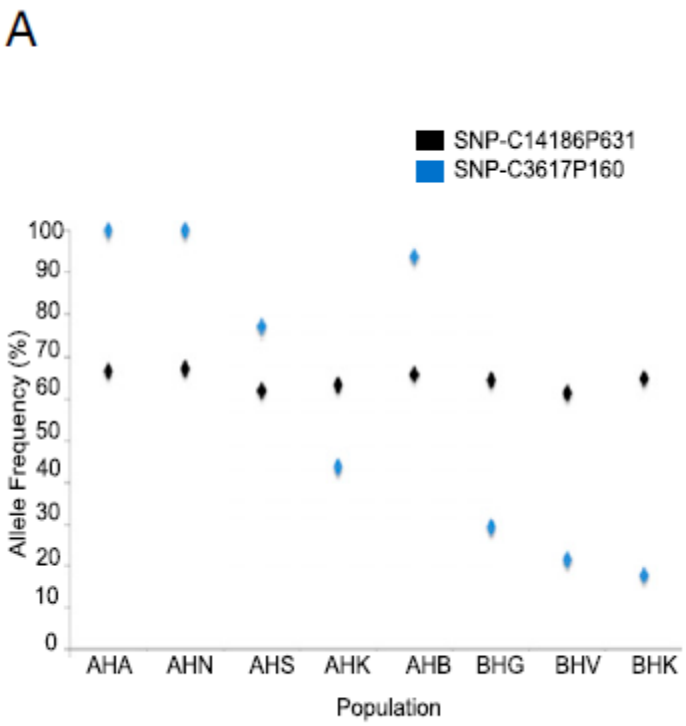


2 SNPs with different
allele frequency patterns

Lamichhaney et al 2012

Ecology (reverse ecology) and local adaptation

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



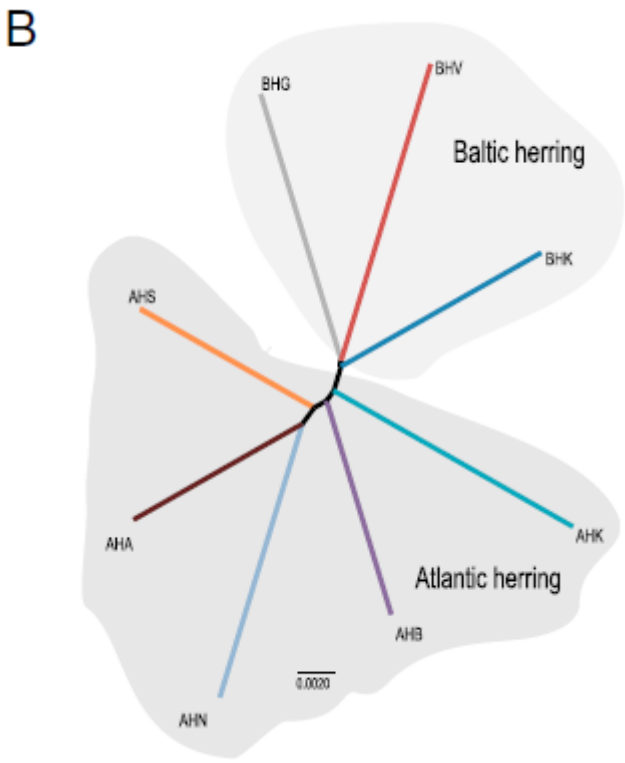
2 SNPs with different
allele frequency patterns

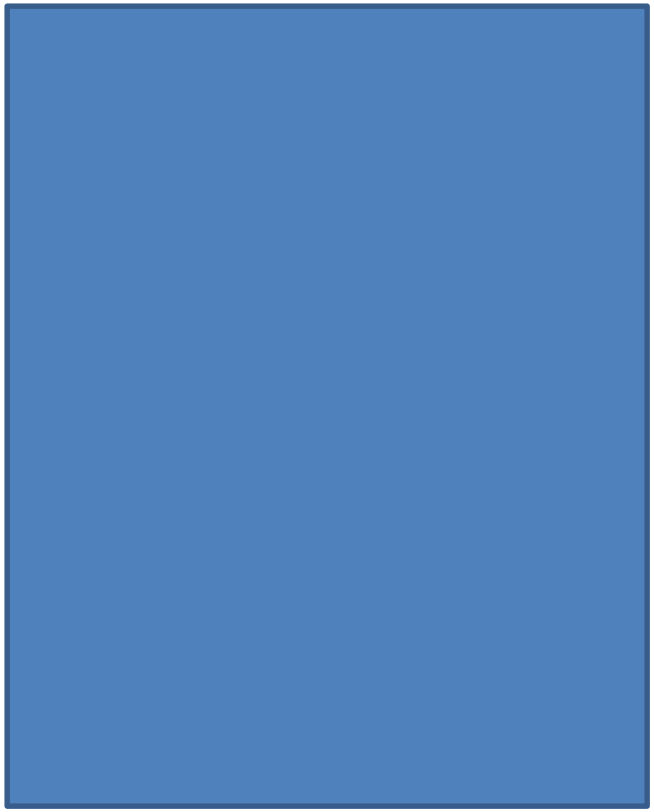
Most undifferentiated
~440,000
Few highly differentiated <1%
~3800

Lamichhaney et al 2012

Little difference between Baltic and Atlantic if using all SNPS…

All SNPs →
(star phylogeny)



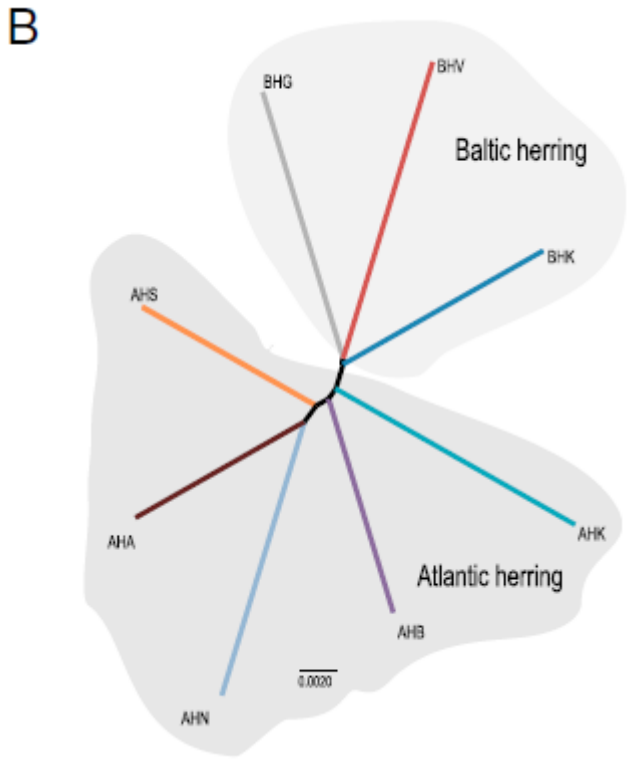


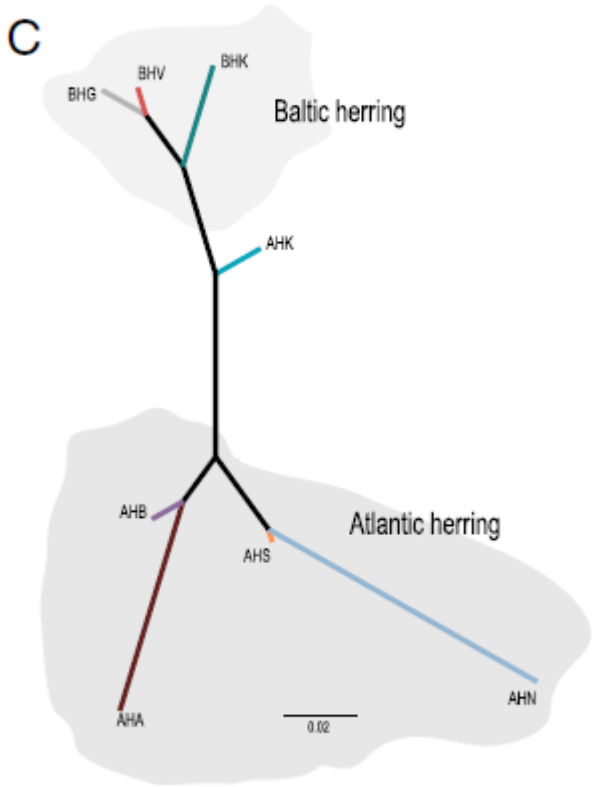
Lamichhaney et al 2012

Little difference between Baltic and Atlantic if using all SNPS…

All SNPs → (star phylogeny)

High Fst SNPs → separate populations





Lamichhaney et al 2012

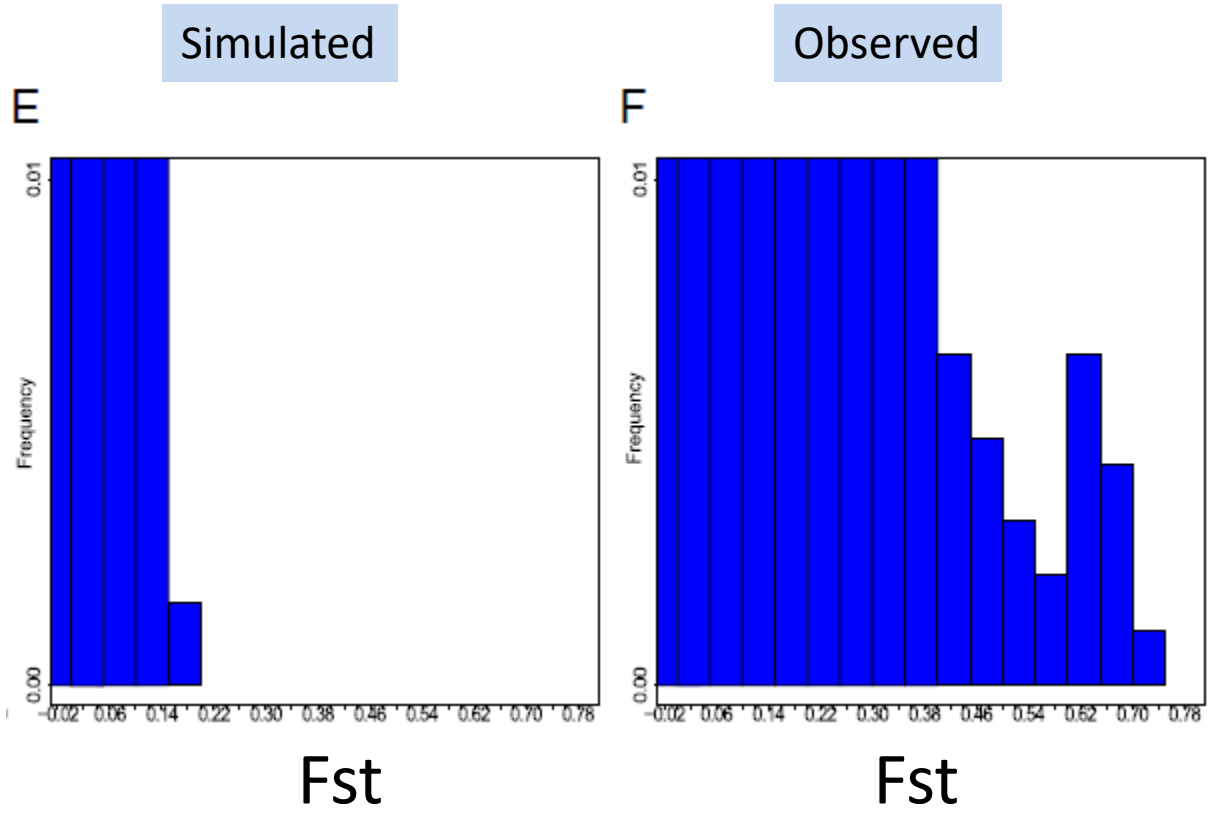
Are high Fst SNPs from drift or selection?

Simulation analysis to test

Lamichhaney et al 2012

Are high Fst SNPs from drift or selection?

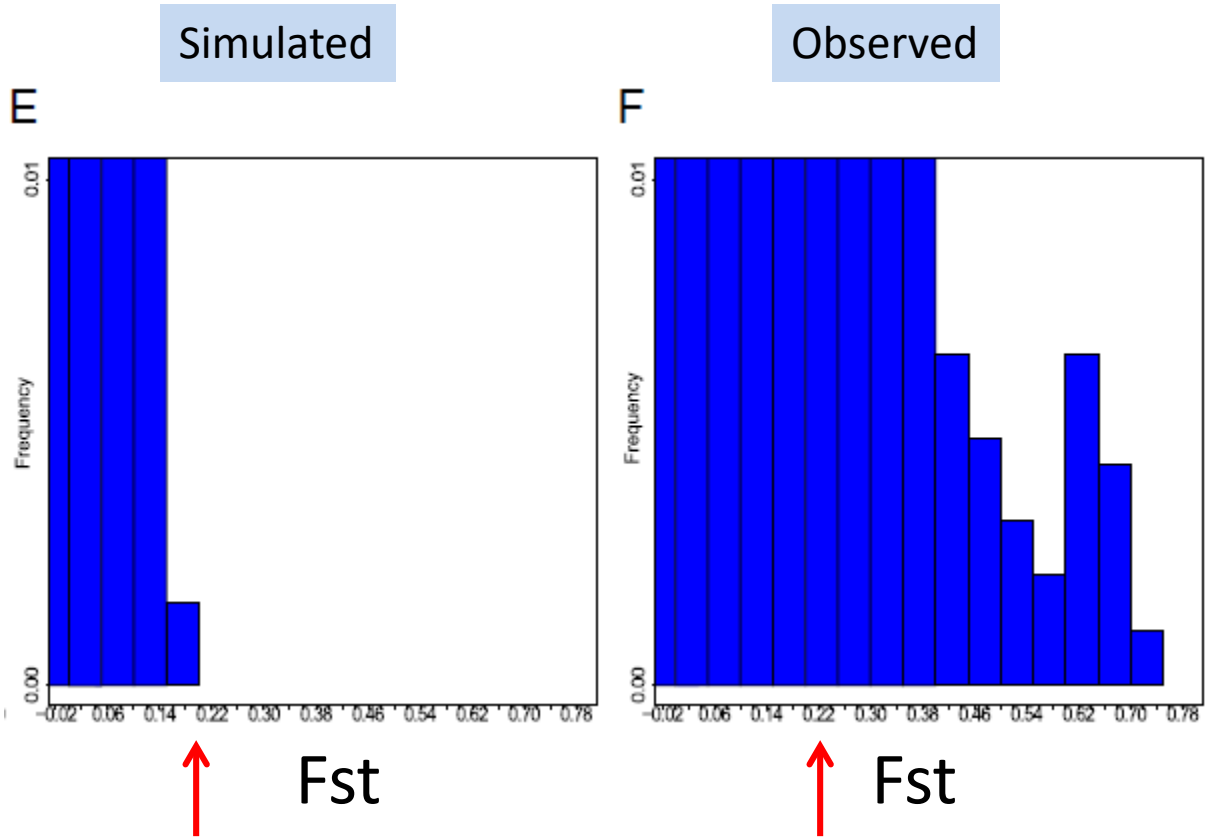
Simulation analysis to test



Lamichhaney et al 2012

Are high Fst SNPs from drift or selection?

Simulation analysis to test



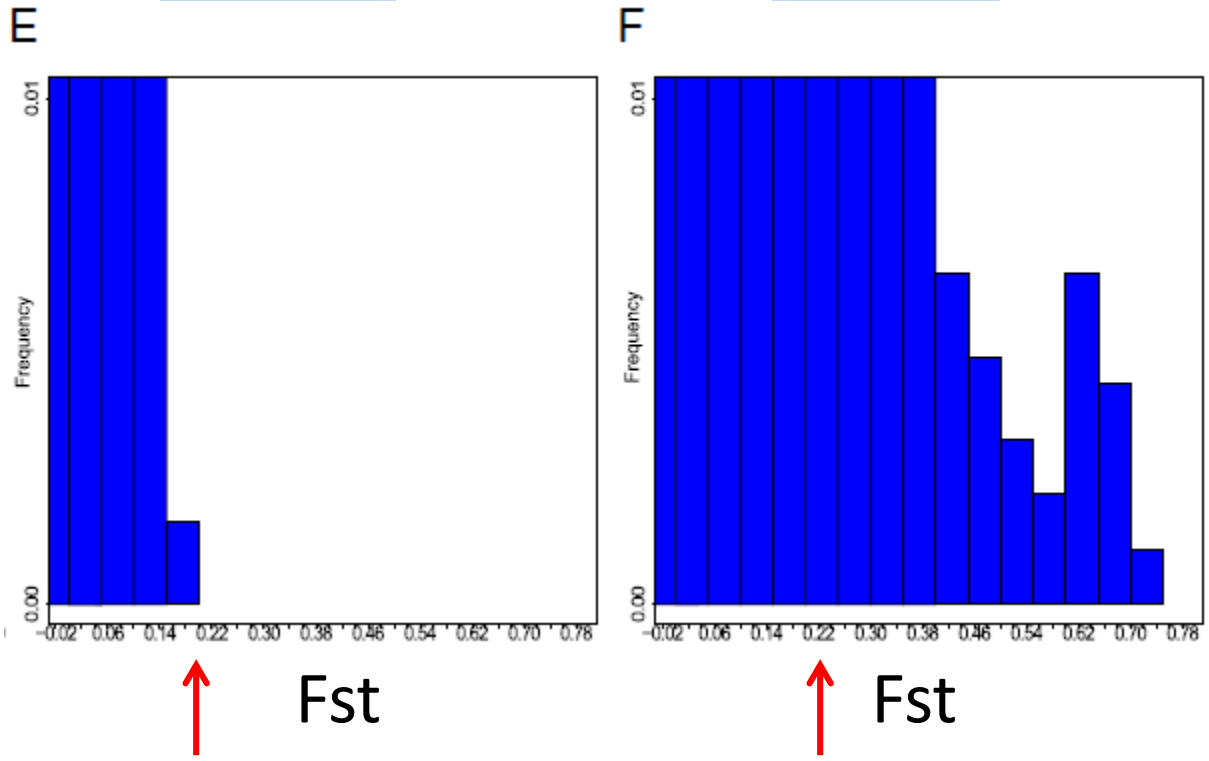
Lamichhaney et al 2012

Are high Fst SNPs from drift or selection?

Simulation analysis to test

Simulated

Observed

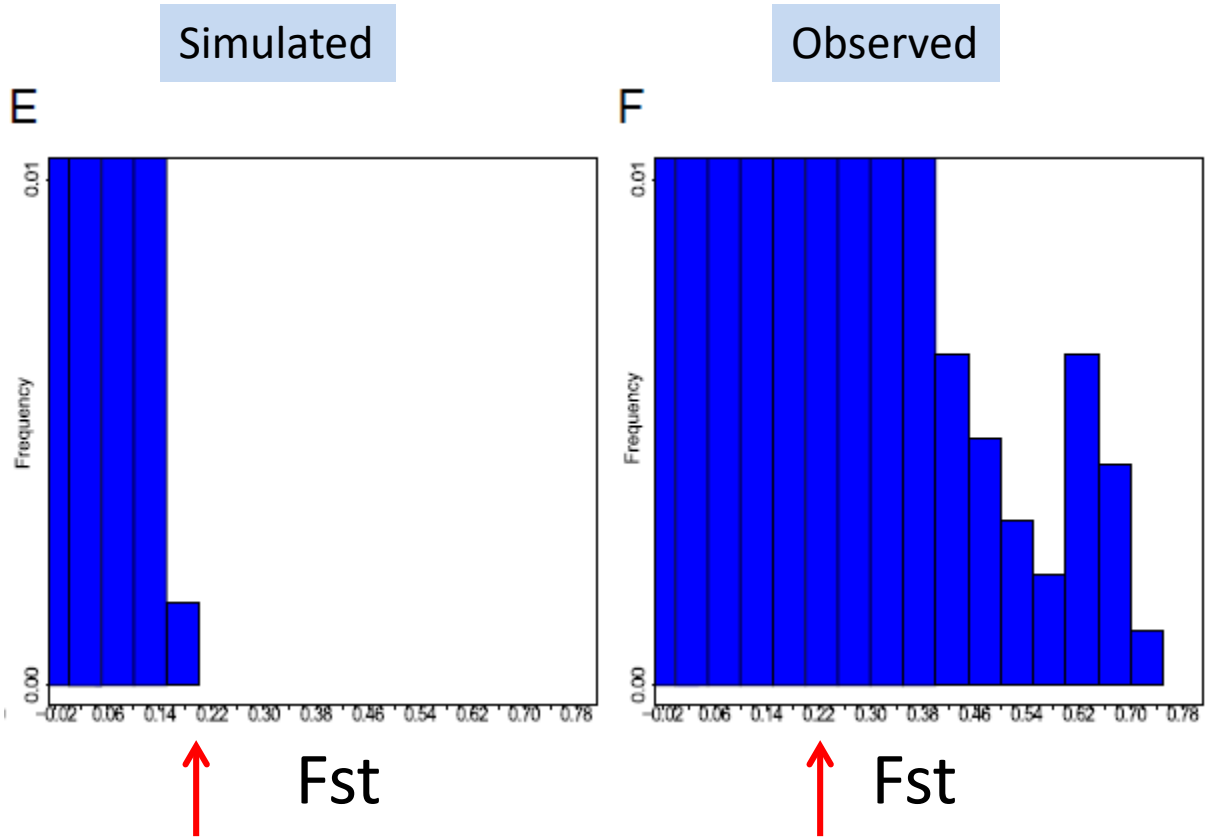


Unlikely drift → therefore selection!
local adaptation

Lamichhaney et al 2012

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

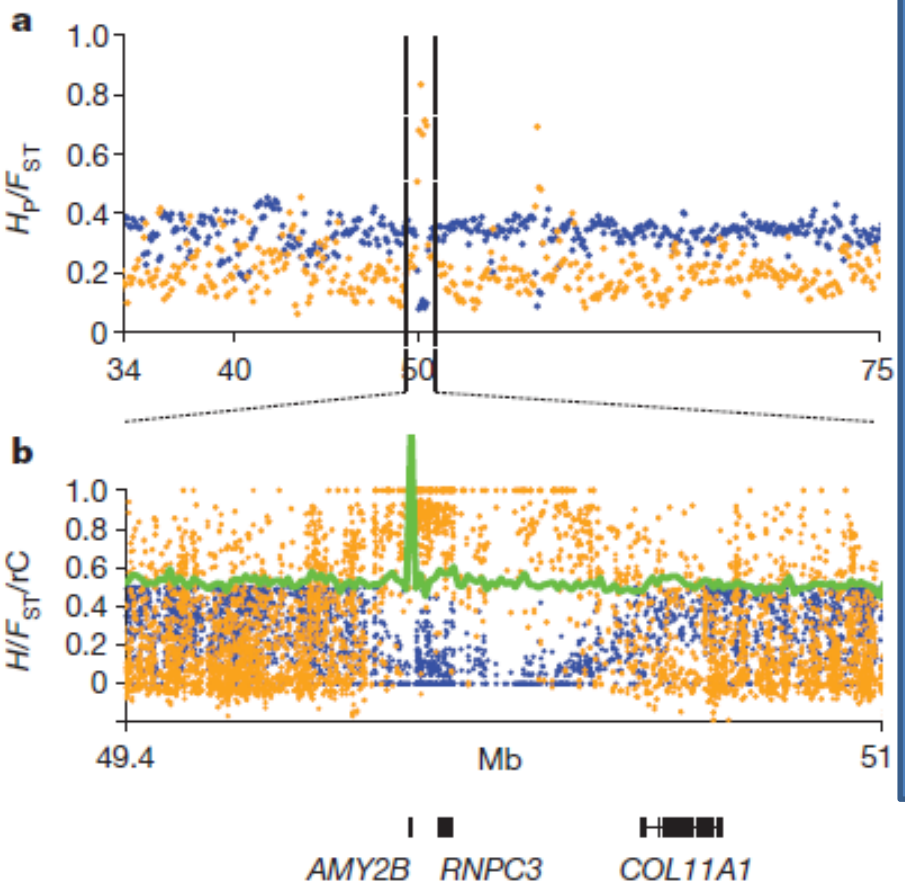
By Retron - self-made now, Public Domain, https://commons.wikimedia.org/w/index.php?curid=3865957

Charles M. Schulz

Dog adaptation to starch diet

36 candidate domestication loci here → Amylase

Fst

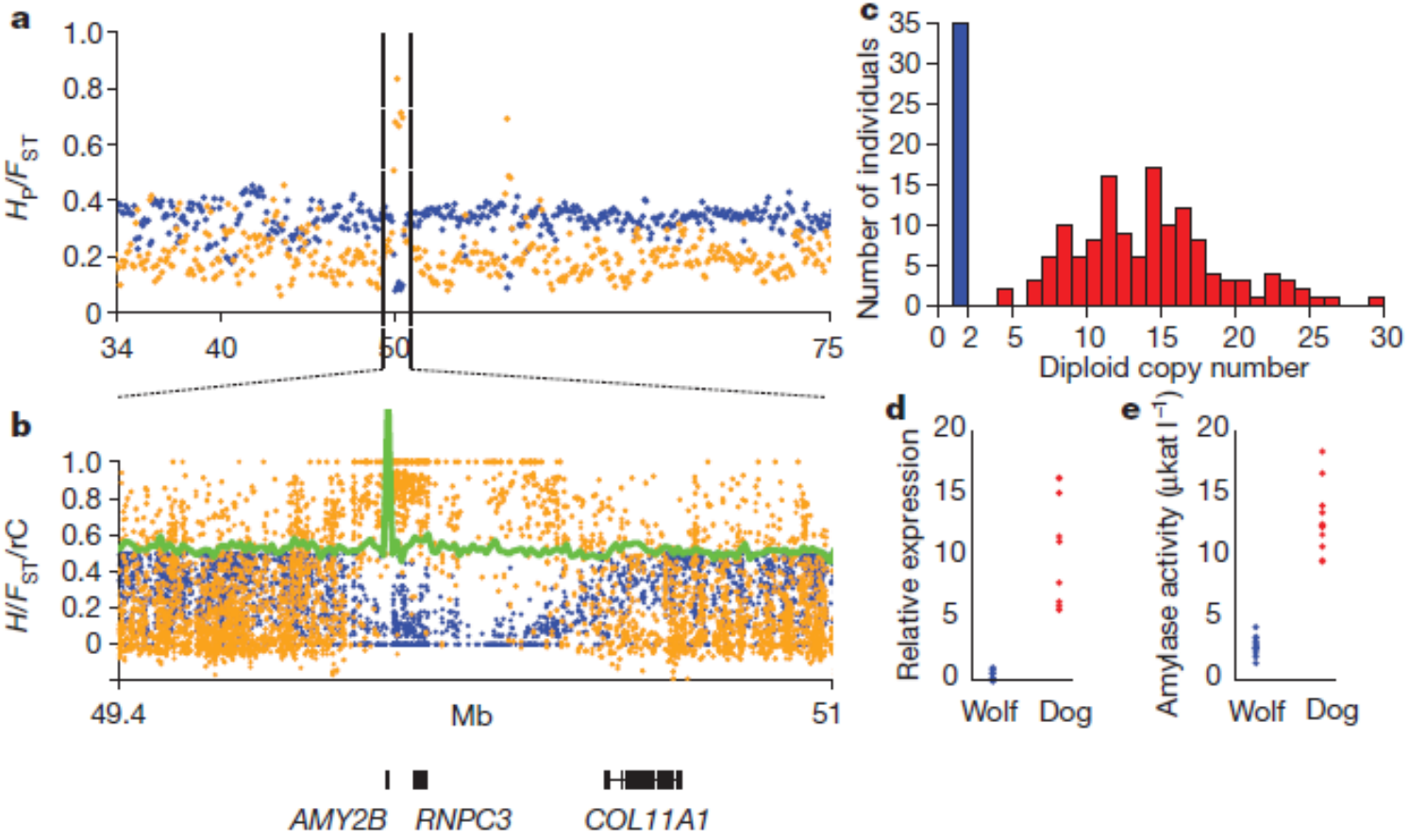


Axelsson et al 2013

Dog adaptation to starch diet

36 candidate domestication loci here → Amylase

Fst

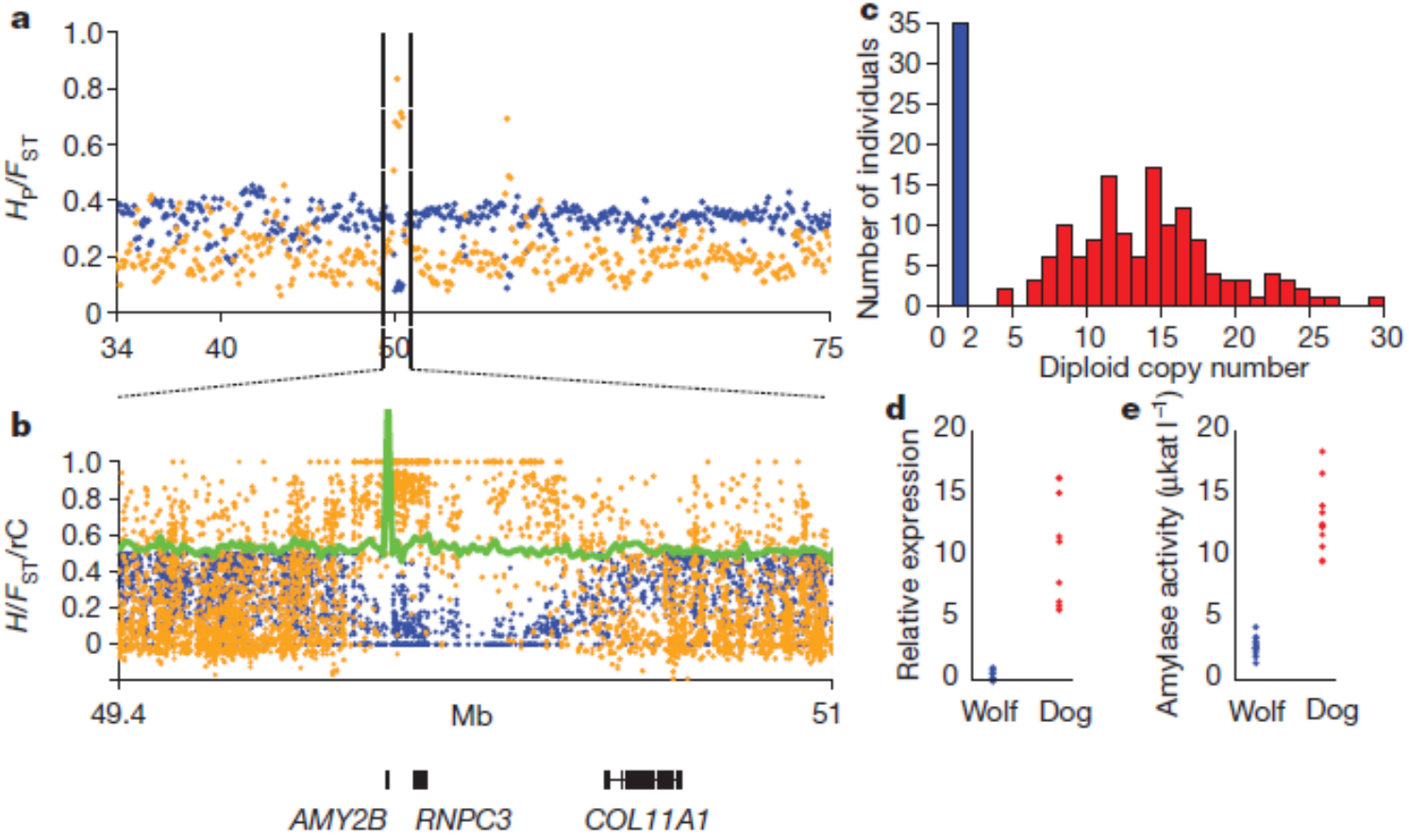


Axelsson et al 2013

Dog adaptation to starch diet

36 candidate domestication loci here → Amylase

Fst



also maltase activity

Axelsson et al 2013

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 ≥2 nt differences

Hellsten et al 2014 PNAS

Samples and analysis sketch

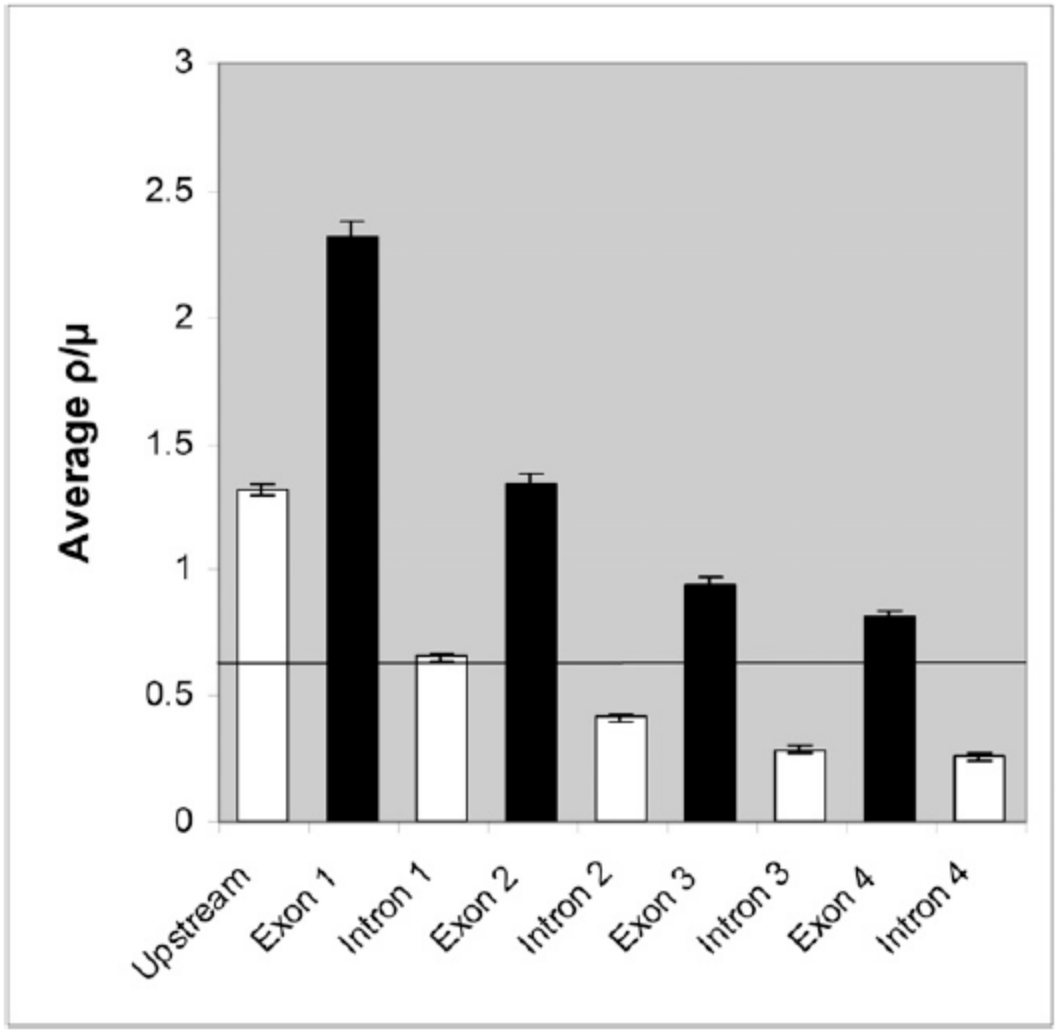
98 samples
4 localities

4 Pool-seq runs
PE 75 bp runs (of ~200 bp fragments)

After QC → 255x coverage

Examined pairs of SNPs within 50 bp

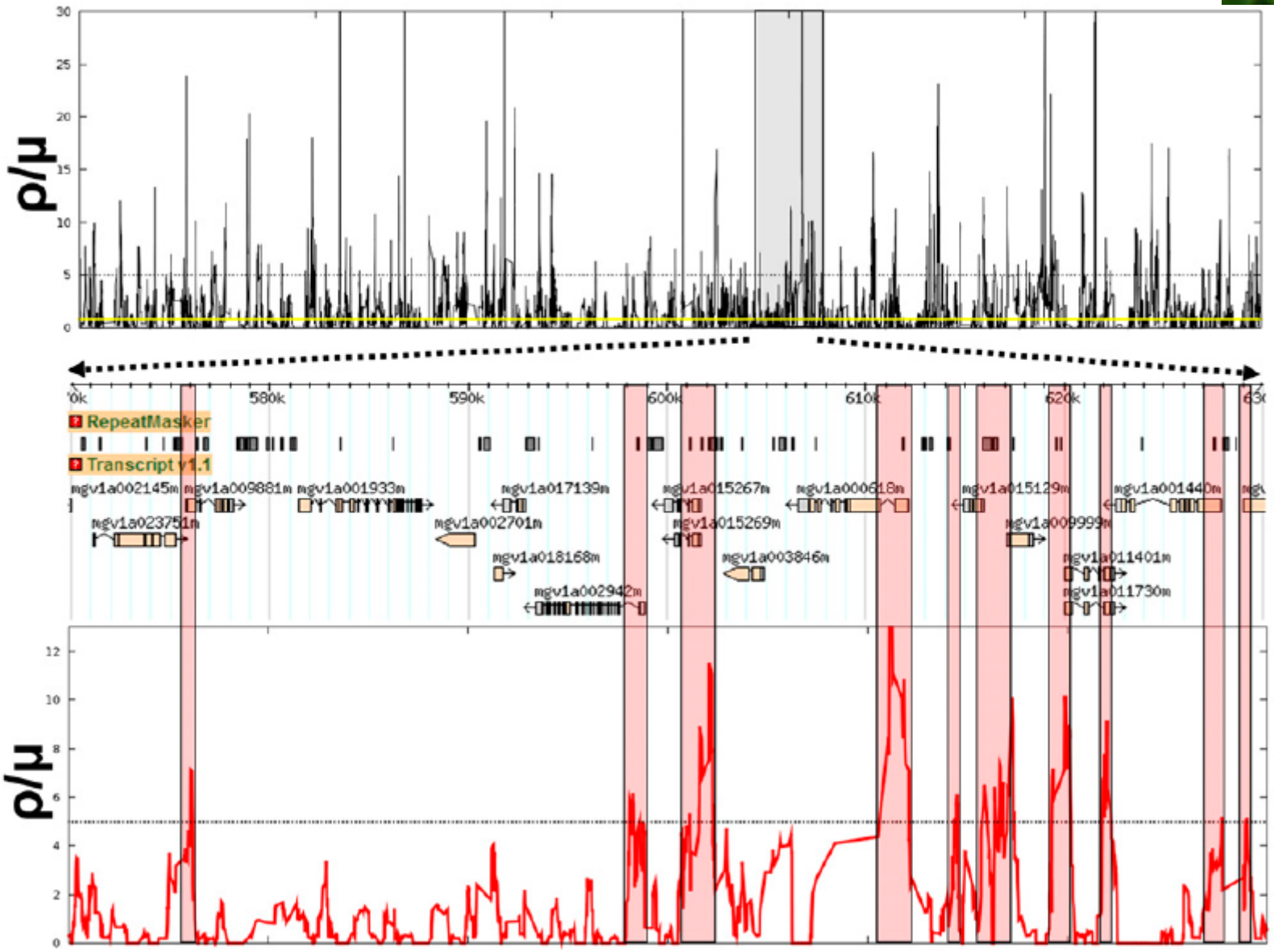
Interesting recombination patterns



- Exons > introns
- Polarity
  5’ > 3’ recomb. rate

Hellsten et al 2014 PNAS

Recombination landscape example

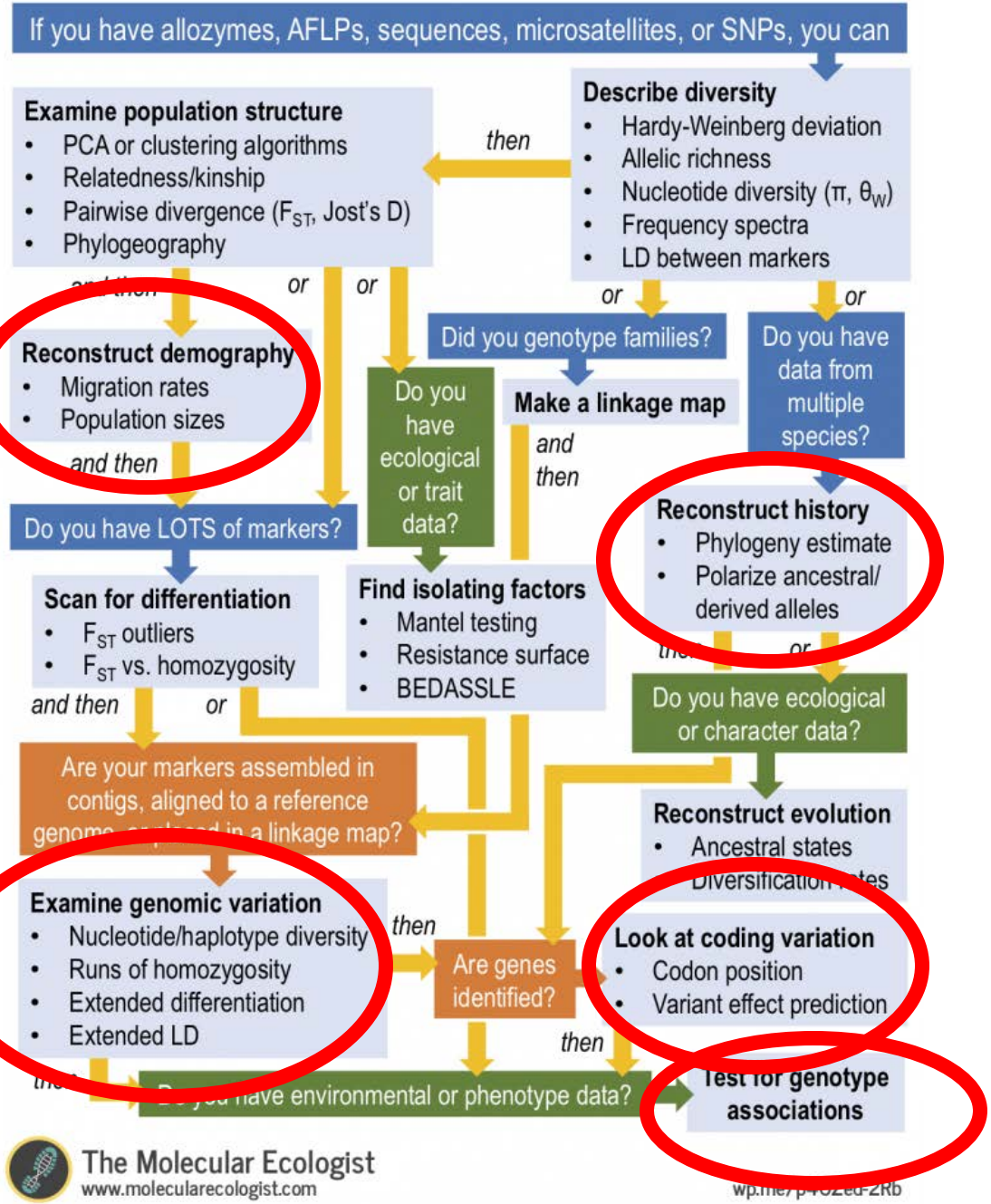


Hellsten et al 2014 PNAS

Whole genome (re)sequencing

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

What can I do with this genetic data?



Panda demographic history

34 wild panda + ref seq

Pairwise sequentially Markovian coalescent (PSMC)

Can infer past history from 1 or few individuals

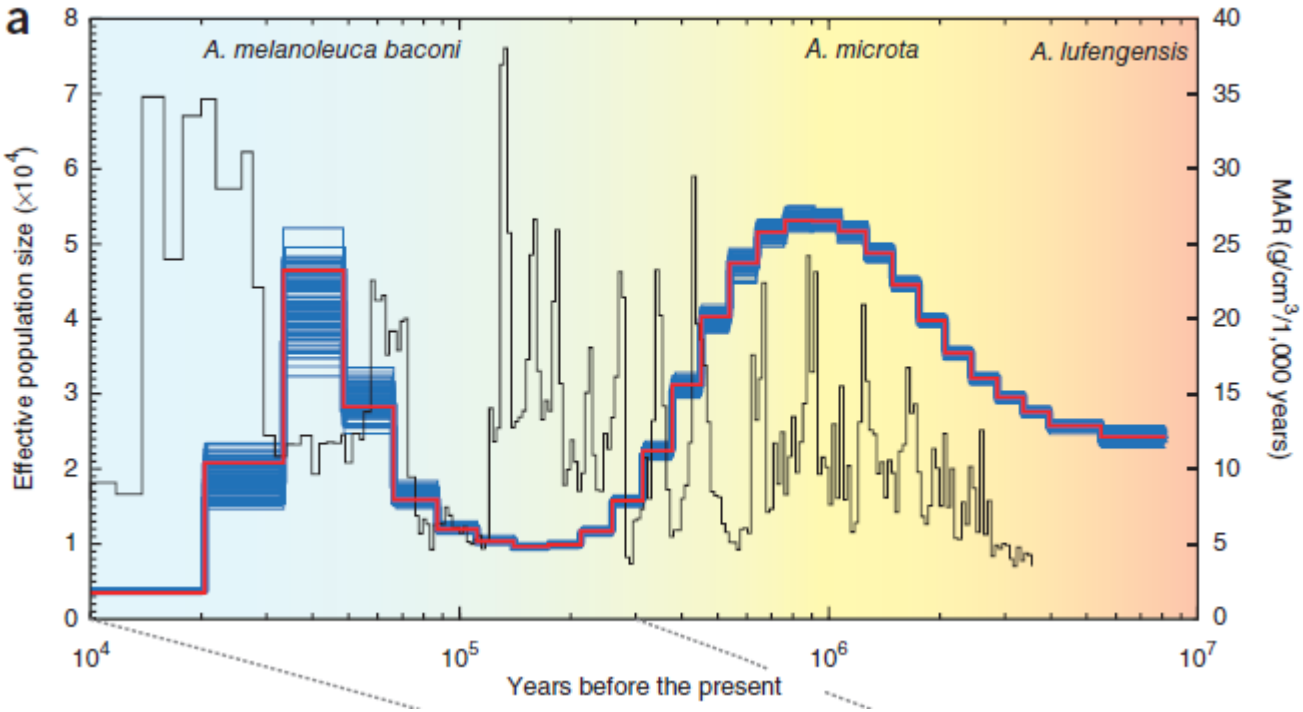
Zhao et al 2013
Li & Durbin 2011

Panda demographic history

34 wild panda + ref seq

Pairwise sequentially Markovian coalescent (PSMC)

Can infer past history from 1 or few individuals

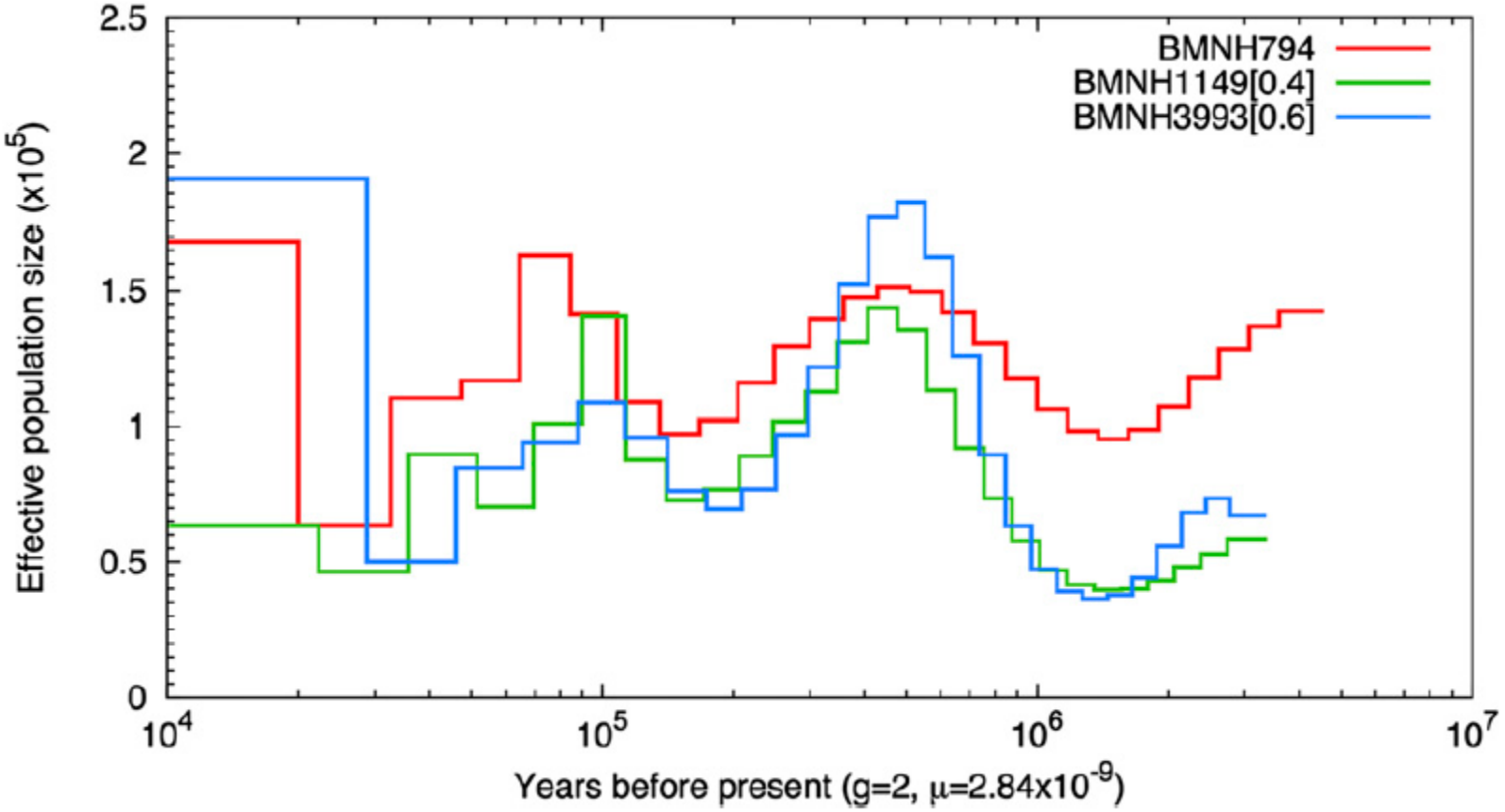


Zhao et al 2013
Li & Durbin 2011

Passenger pigeon demographic history

Passenger pigeon demographic history (+ museum DNA)

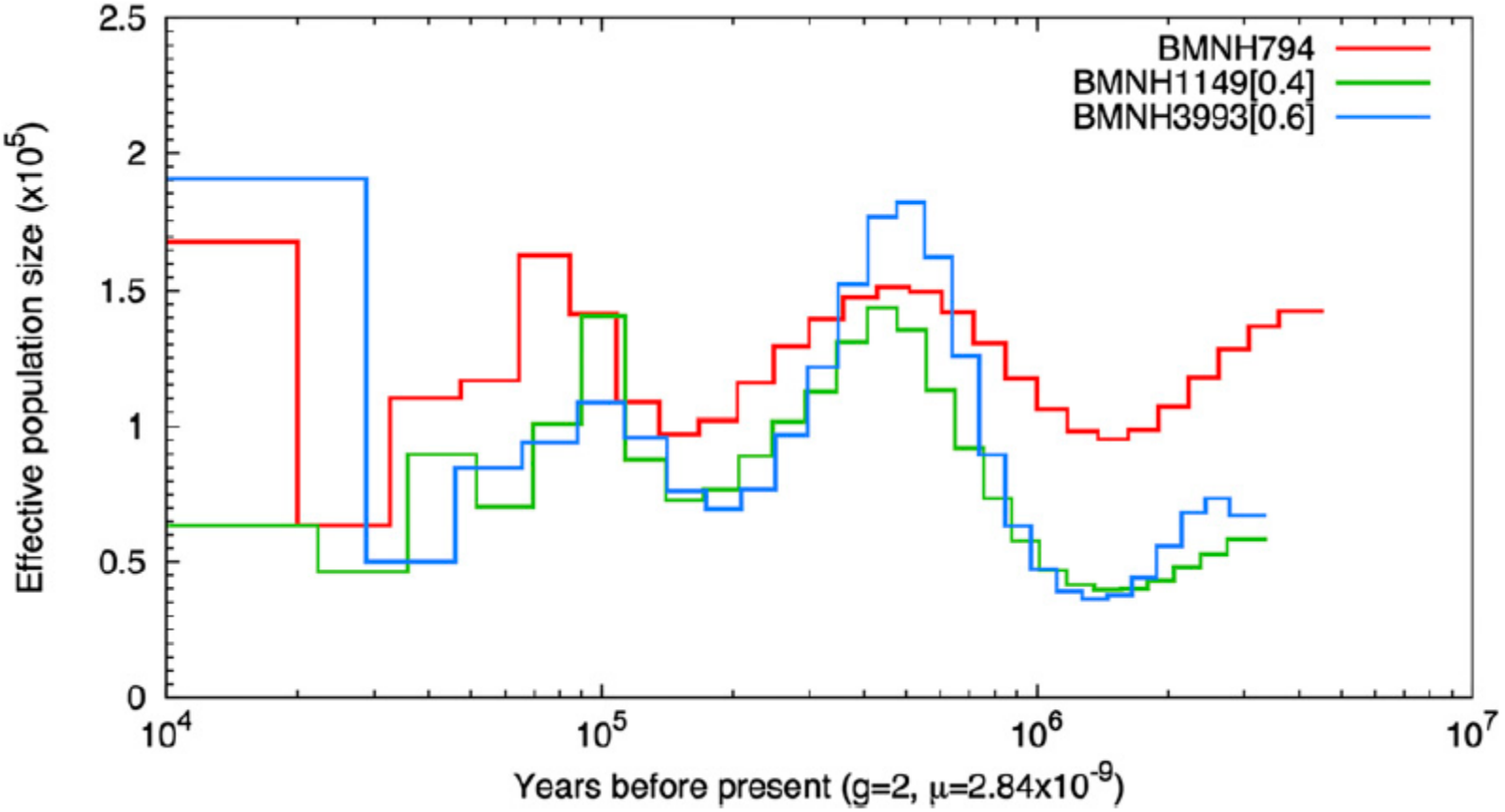
N=3 from museum, PSMC



Hung et al 2014

Passenger pigeon demographic history (+ museum DNA)

N=3 from museum, PSMC



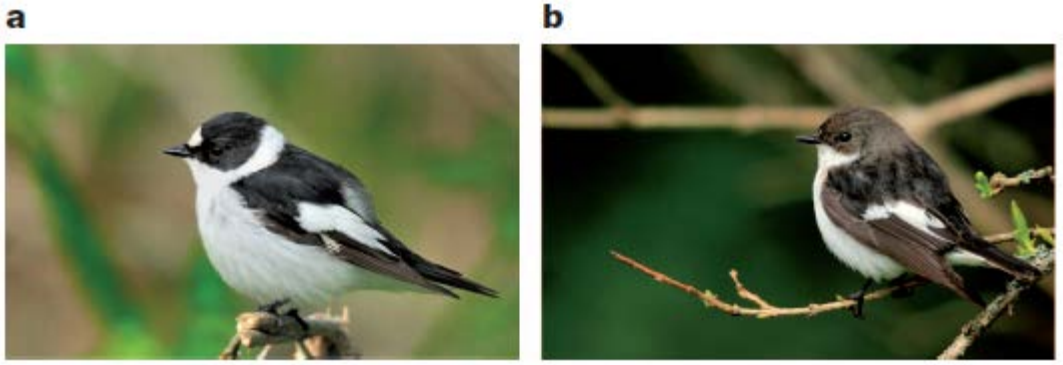
Ne (effective population size; 100,000's) << census N (billions)

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

Increased extinction risk

Hung et al 2014

Speciation islands? (divergence islands)



collared pied

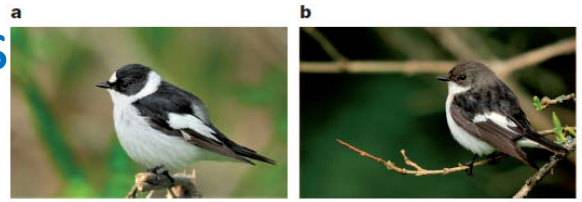
flycatchers
diverged < 2Mya
separated, 2^{o} contact
hybridize

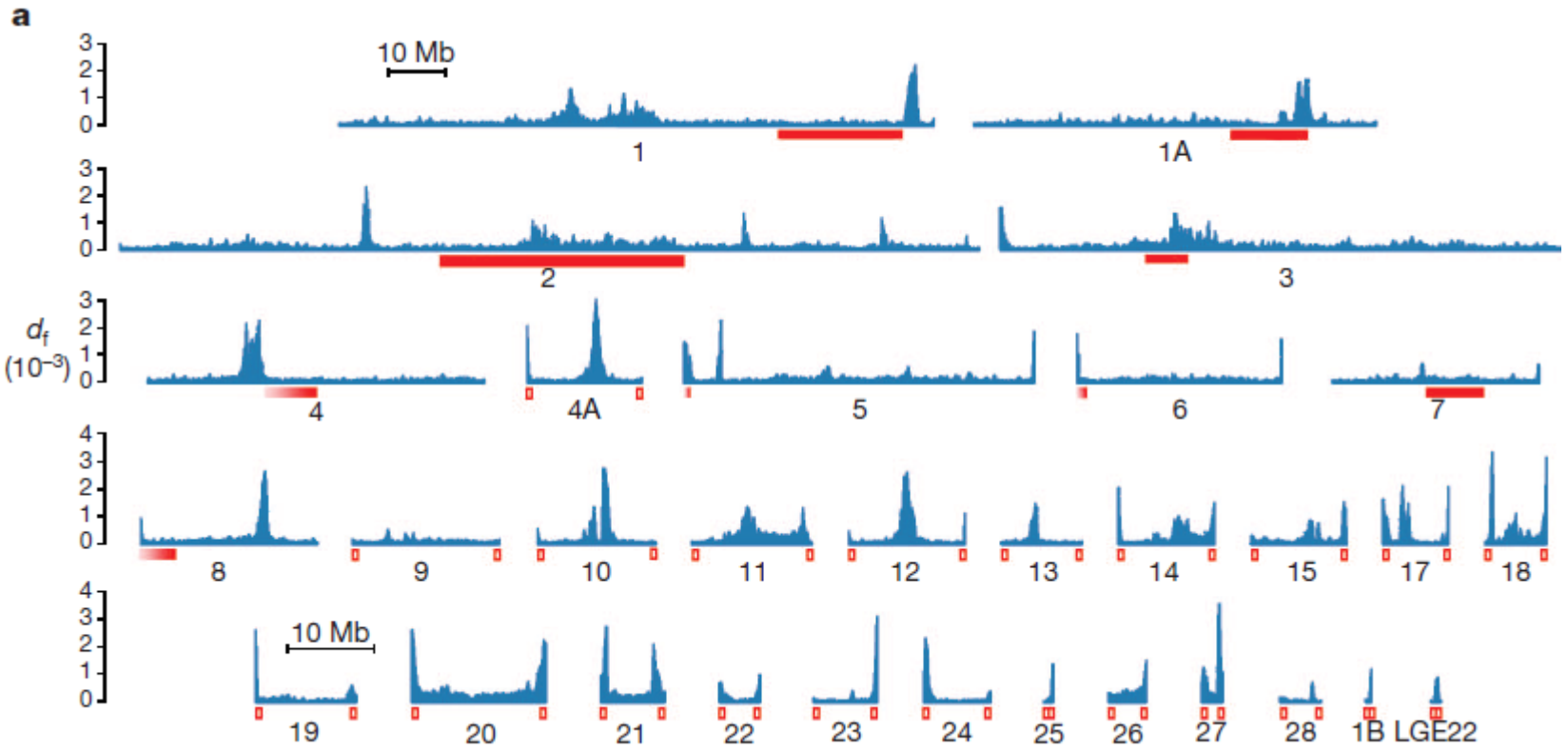
10 re-sequenced individuals + 1 ref seqs (collared)

Compared genomewide divergence to each other

Ellegren et al 2012

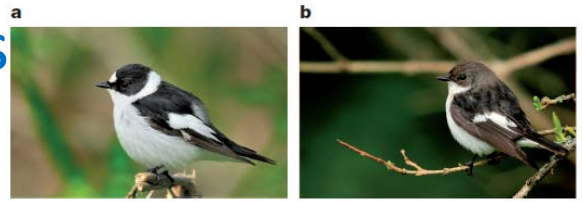
Divergence islands often near centromeres and telomeres

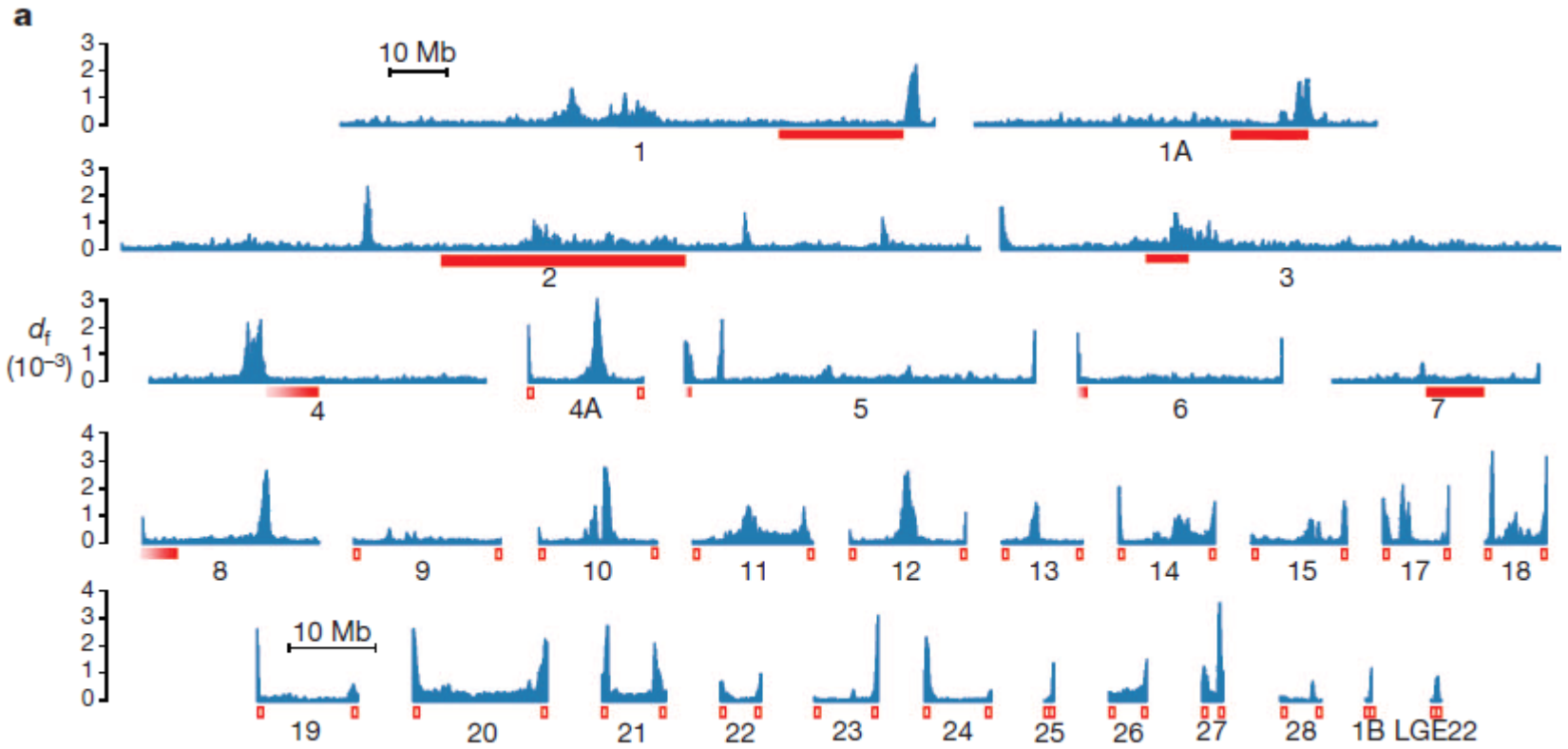




Ellegren et al 2012

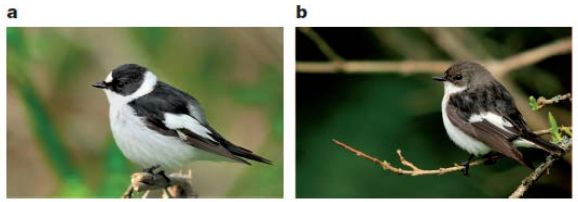
Divergence islands often near centromeres and telomeres





Authors postulate meiotic drive

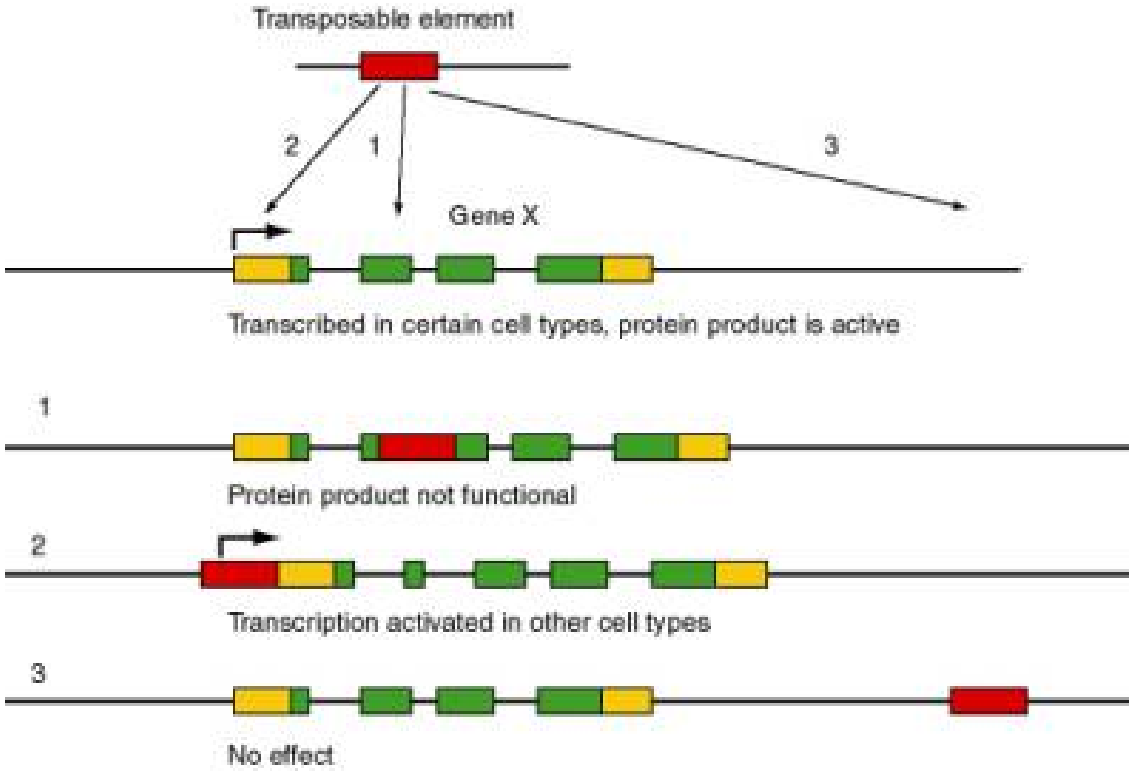
Ellegren et al 2012



Inspect for transposable elements

Transposable elements (TEs) are selfish “jumping genes”

Neutral marker or evolutionary important



Also:

Alternative splicing, methylation, ectopic recombination, etc

http://www.bx.psu.edu/~ross/workmg/TranspositionCh9.htm

TEs are an important source of genomic variation

Humans

Alu element → ~800 insertion differences between 2 specific individuals

Estimated new insertion rates:

Alu → 1 in 21 people

L1 → 1 in 108 to 212

Maize

>177,000 TEs in maize (B73 reference strain)

B73 versus W22 strain → >22,000 TEs at different locations

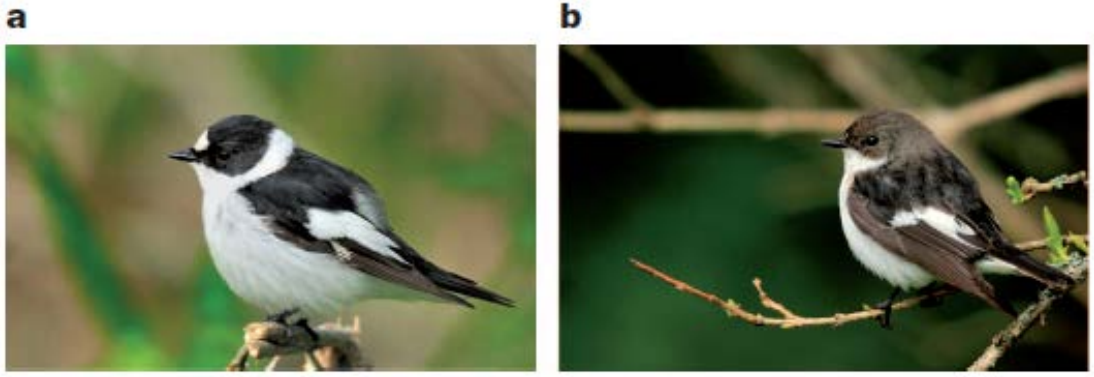
In fact, >10,000 strain-specific subfamilies of TEs

Burns and Boeke 2012 Cell

Springer et al 2018 Nat Genet

Birds have been suggested to have low TE densities

Implying that TEs have not transposed much in bird history

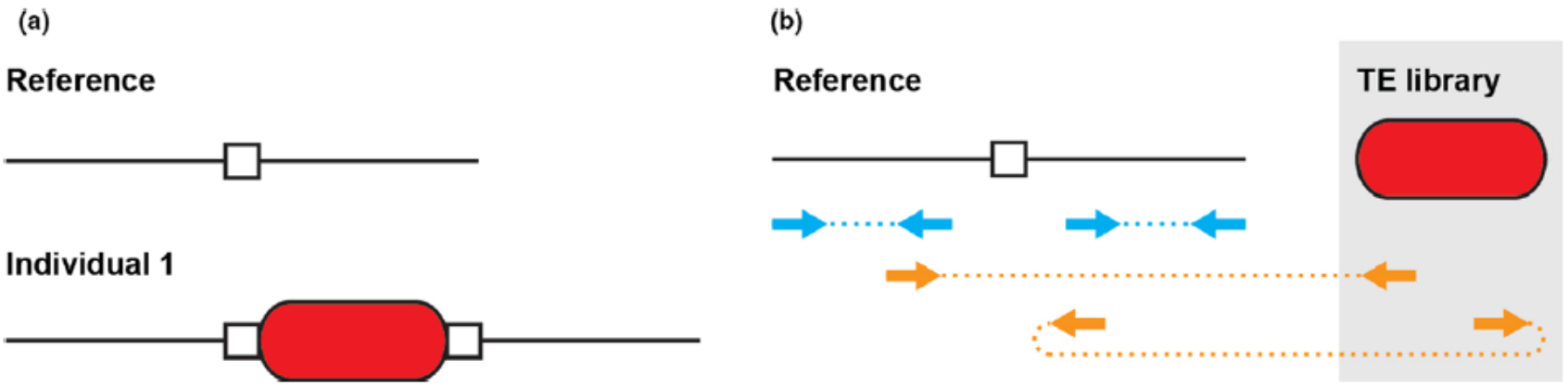


Suh et al looked at:

- TE repertoire in flycatcher genome
- Then polymorphism in 200 resequenced genomes

Suh et al 2018 Mol Ecol (flycatcher)

NGS to locate TEs

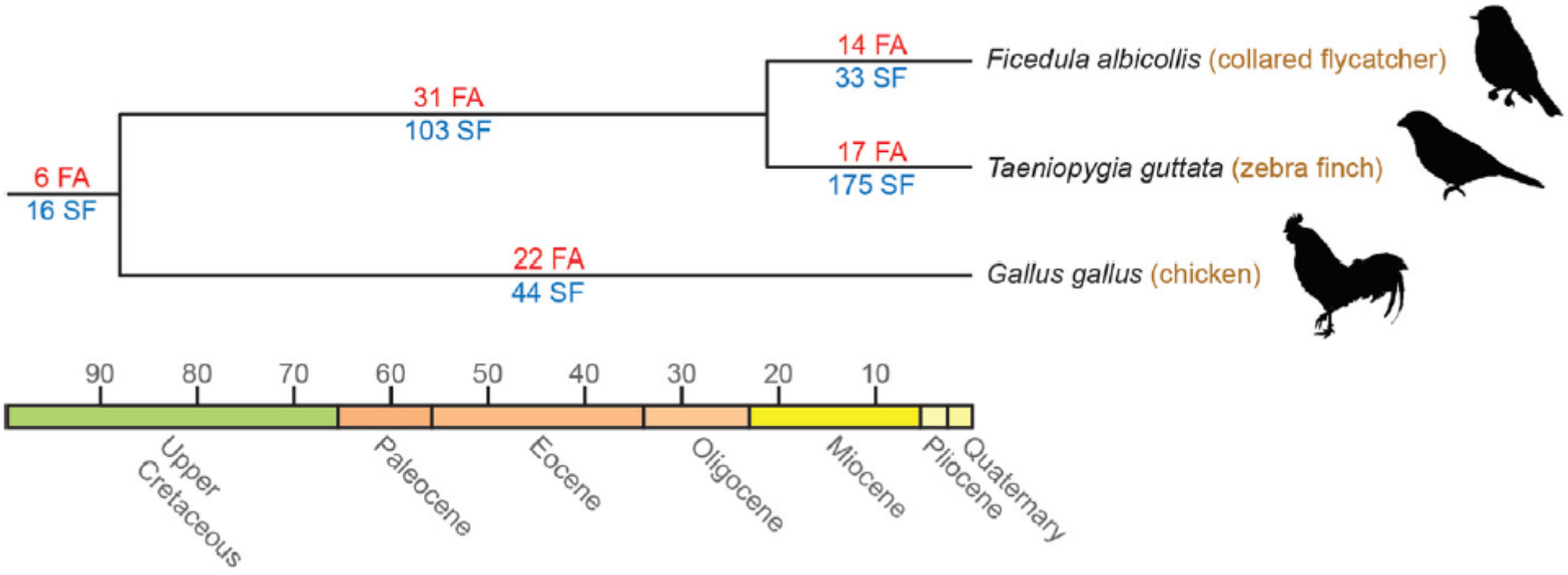


(There's also a "reduced" method using primers from the TEs)

Suh et al 2018 Mol Ecol (flycatcher)

Novel transposon LTR families in 3 birds

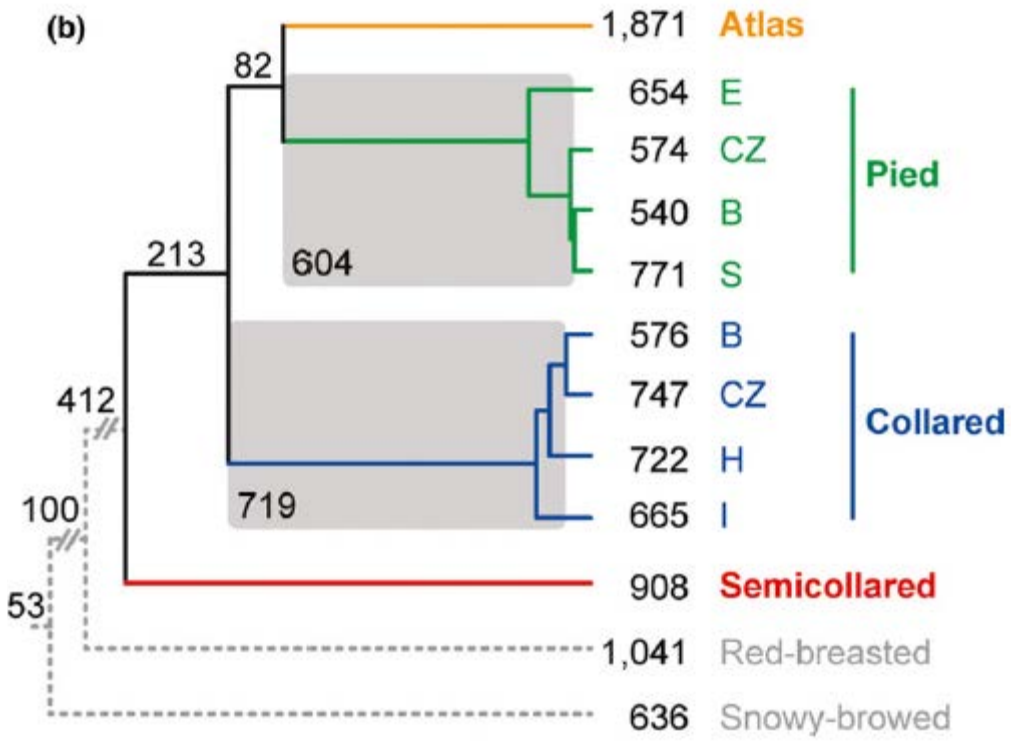
i.e., These are NEW to science



Suh et al 2018 Mol Ecol (flycatcher)

TE variation in flycatchers

Retrotransposition events per lineage



Suh et al 2018 Mol Ecol (flycatcher)

Flycatcher case study conclusions

Suh et al 2018 Mol Ecol (flycatcher)

Flycatcher case study conclusions

Flycatchers (songbirds) actually have a lot of transposition events

Suh et al 2018 Mol Ecol (flycatcher)

Flycatcher case study conclusions

Flycatchers (songbirds) actually have a lot of transposition events

New TE families
and current polymorphism

Suh et al 2018 Mol Ecol (flycatcher)

Flycatcher case study conclusions

Flycatchers (songbirds) actually have a lot of transposition events

New TE families
and current polymorphism

These may have phenotypic consequences for their diversification

Suh et al 2018 Mol Ecol (flycatcher)

Flycatcher case study conclusions

Flycatchers (songbirds) actually have a lot of transposition events

New TE families
and current polymorphism

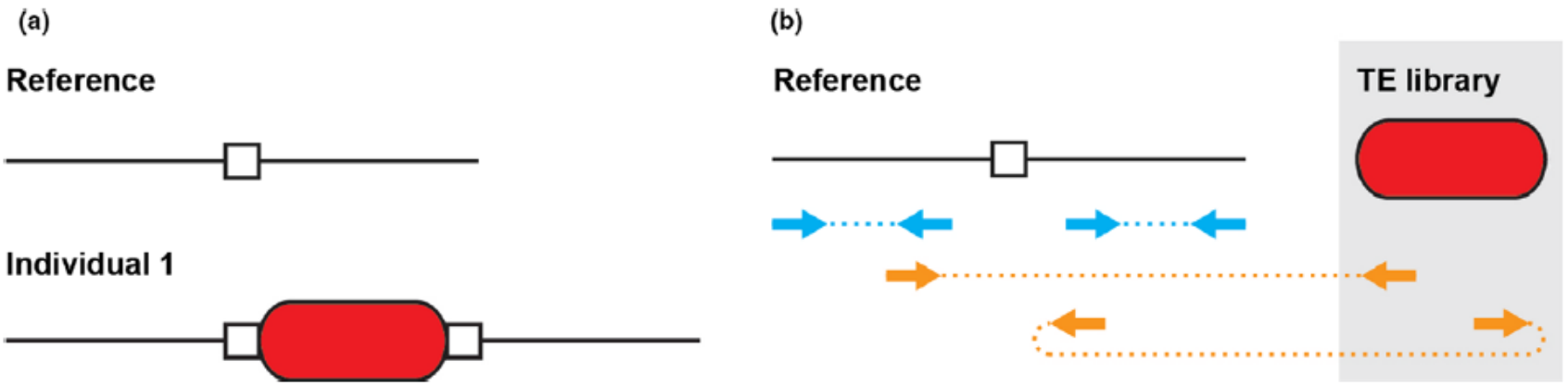
These may have phenotypic consequences for their diversification

In general TEs seem to be underappreciated for their impact

Suh et al 2018 Mol Ecol (flycatcher)

Paired-end method has drawbacks

TEs can be 1000's bp long

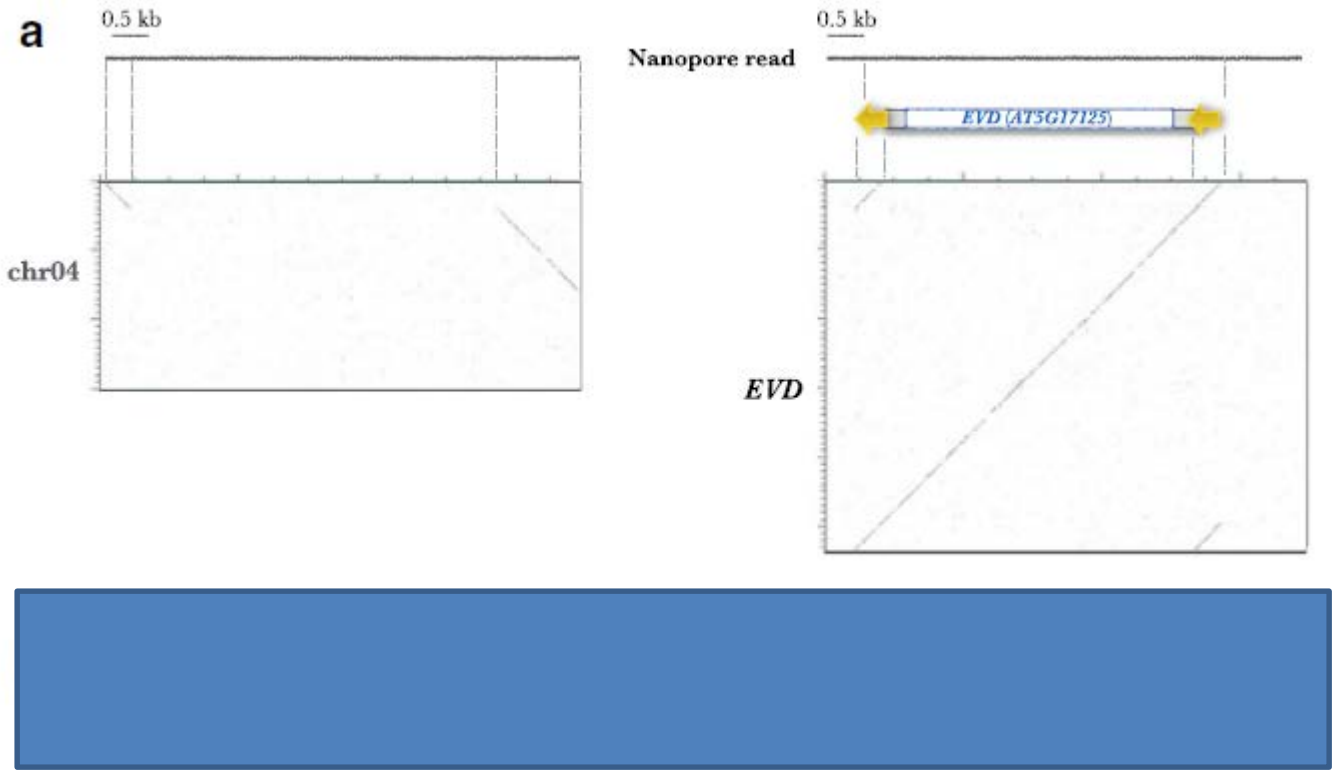


Short reads have higher False Discovery Rates
Need lots of depth
Somewhat inferred (e.g., don't know the internal sequence)
etc

Long reads (PacBio, ONT) can just read across TE insertions

TEs can be 1000's bp long

ONT analysis in Arabidopsis
Example of a ~7kb that crosses 1 TE

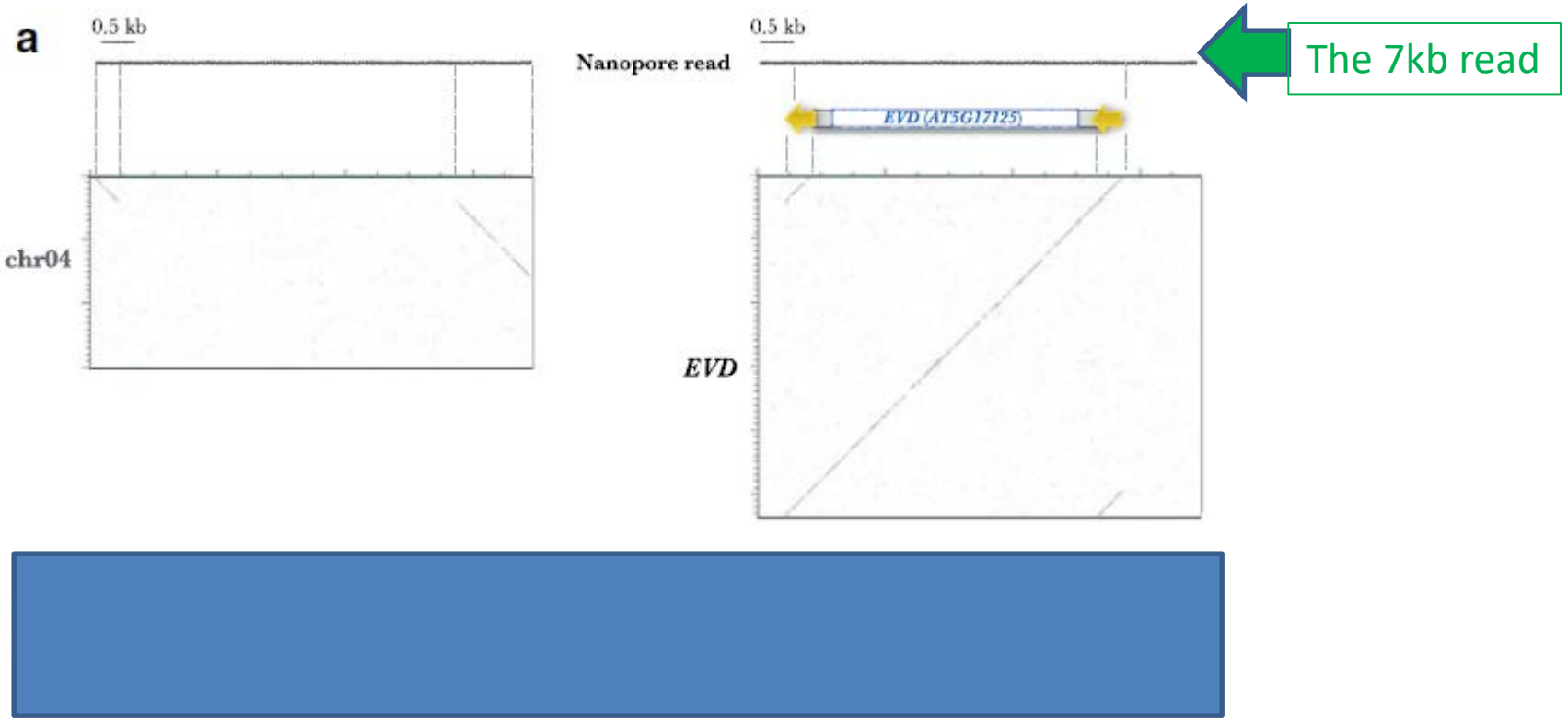


Debladis et al 2017 BMC Genomics

Long reads (PacBio, ONT) can just read across TE insertions

TEs can be 1000's bp long

ONT analysis in Arabidopsis
Example of a ~7kb that crosses 1 TE

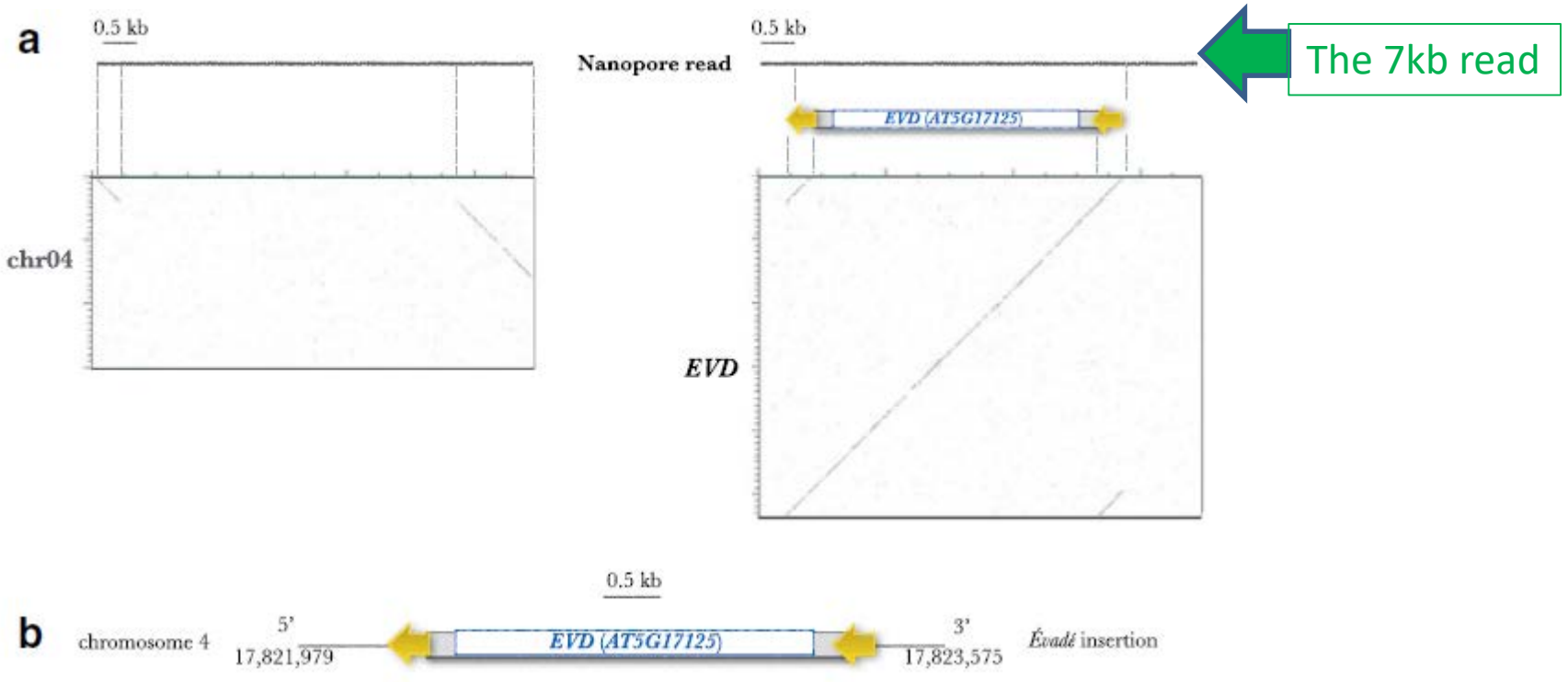


Debladis et al 2017 BMC Genomics

Long reads (PacBio, ONT) can just read across TE insertions

TEs can be 1000's bp long

ONT analysis in Arabidopsis
Example of a ~7kb that crosses 1 TE

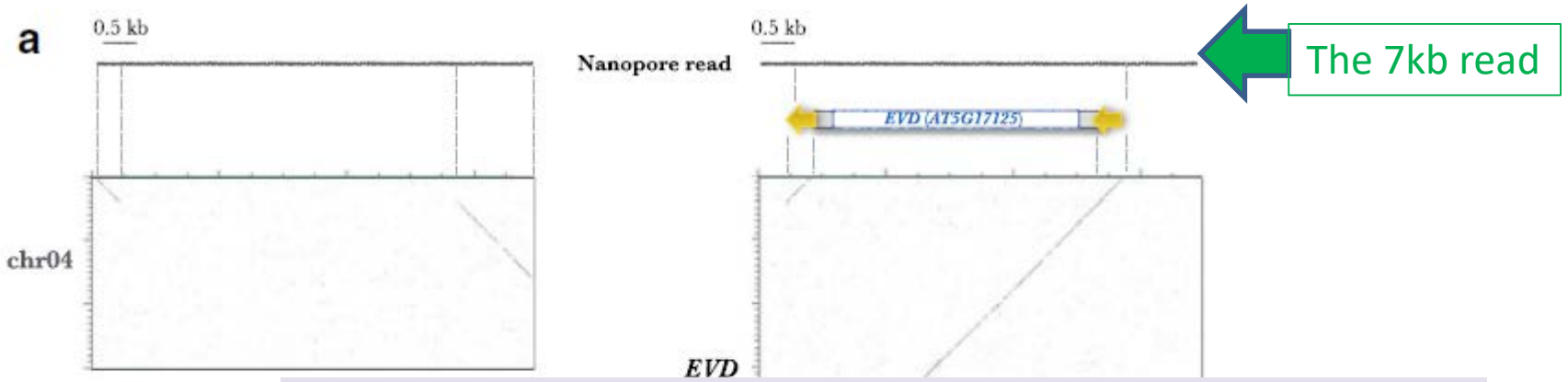


Debladis et al 2017 BMC Genomics

Long reads (PacBio, ONT) can just read across TE insertions

TEs can be 1000's bp long

ONT analysis in Arabidopsis
Example of a ~7kb that crosses 1 TE



Less depth needed, sort of

Does tell you if "whole/intact" or "partial" copy

Goes from algorithm to just "inspecting"

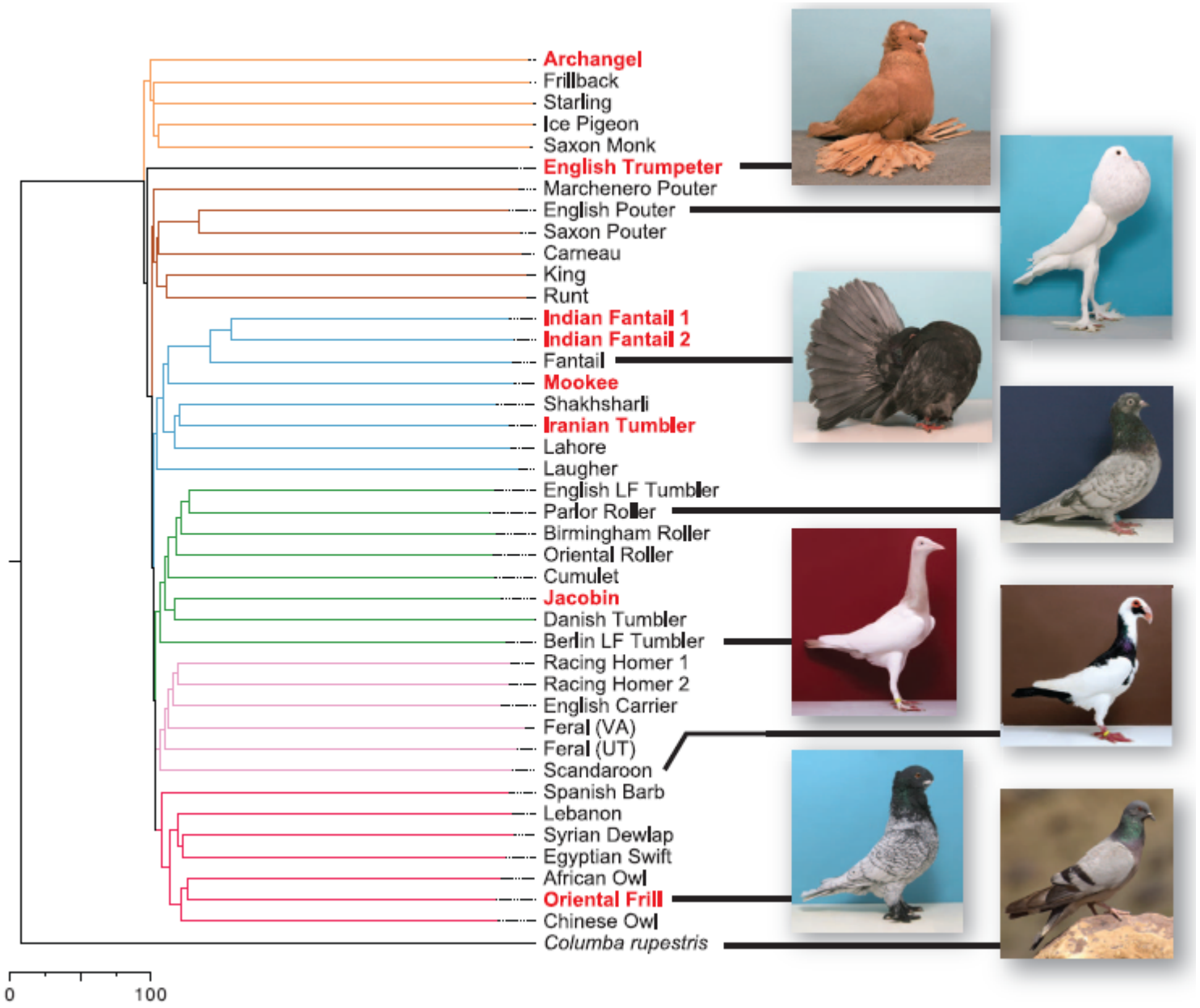
Debladis et al 2017 BMC Genomics

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

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

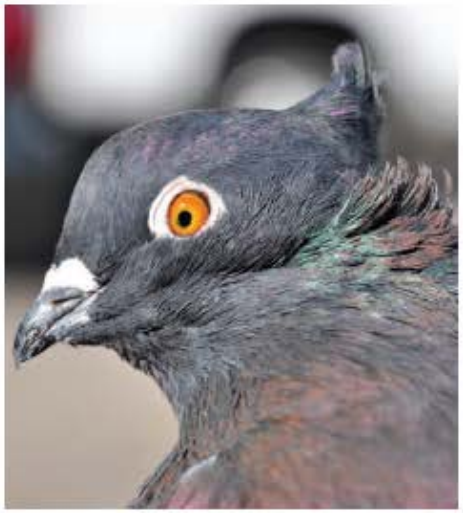
1 ref + 40 additional re-sequenced pigeon genomes

1 ref + 40 additional re-sequenced pigeon genomes



What is the genetic basis for head crests?

Recessive mendelian trait



Peak crest

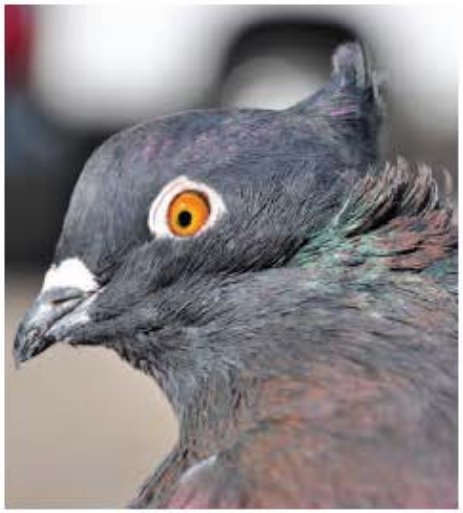
Shell crest

Mane

Hood

What is the genetic basis for head crests?

Recessive mendelian trait

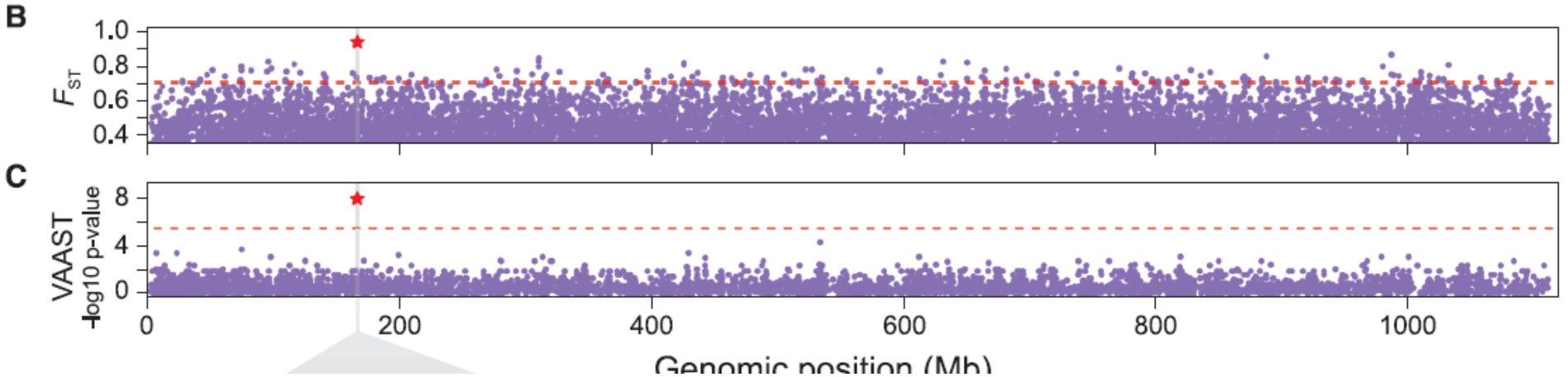


Peak crest

Shell crest

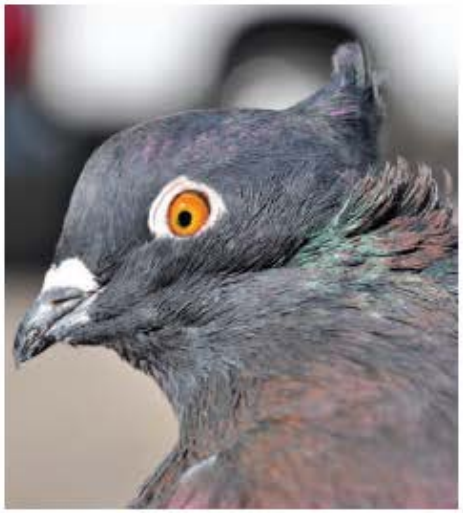
Mane

Hood



What is the genetic basis for head crests?

Recessive mendelian trait



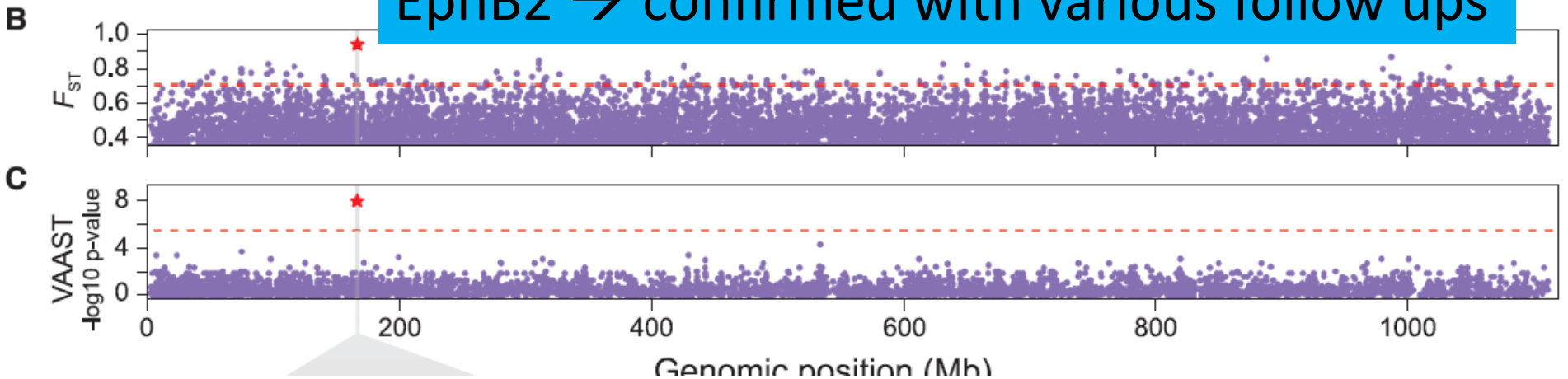
Peak crest

Shell crest

Mane

Hood

EphB2 → confirmed with various follow ups



Human genetics and health

GWAS
using SNP arrays

11 OCTOBER 2018 | VOL 562 | NATURE

ARTICLE

OPEN

https://doi.org/10.1038/s41586-018-0579-z

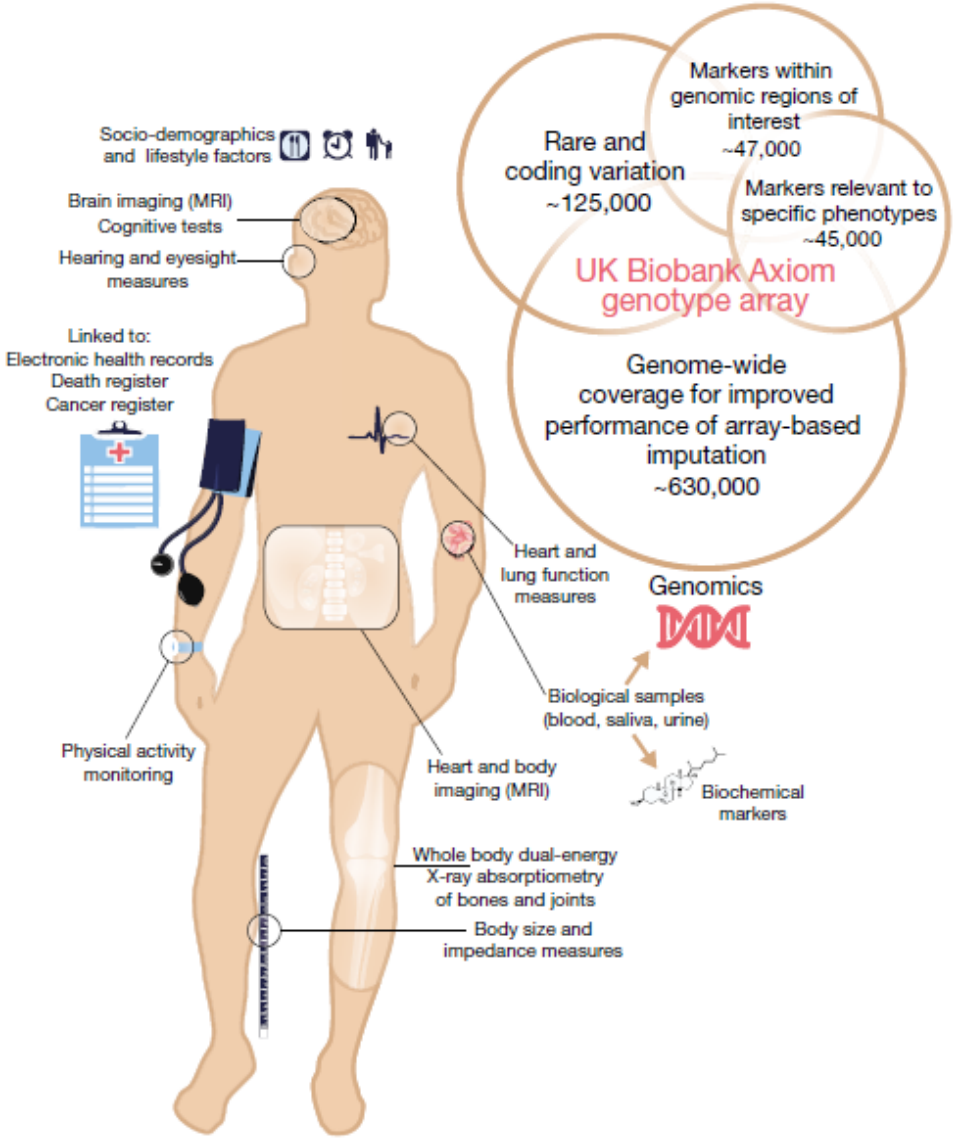
The UK Biobank resource with deep phenotyping and genomic data

Clare Bycroft[1,13], Colin Freeman[1,13], Desislava Petkova[1,12,13], Gavin Band[1], Lloyd T. Elliott[2], Kevin Sharp[2], Allan Motyer[3], Damjan Vukcevic[3,4], Olivier Delaneau[5,6,7], Jared O'Connell[8], Adrian Cortes[1,9], Samantha Welsh[10], Alan Young[11], Mark Effingham[10], Gil McVean[1,11], Stephen Leslie[3,4], Naomi Allen[11], Peter Donnelly[1,2,14] & Jonathan Marchini[1,2,14]*

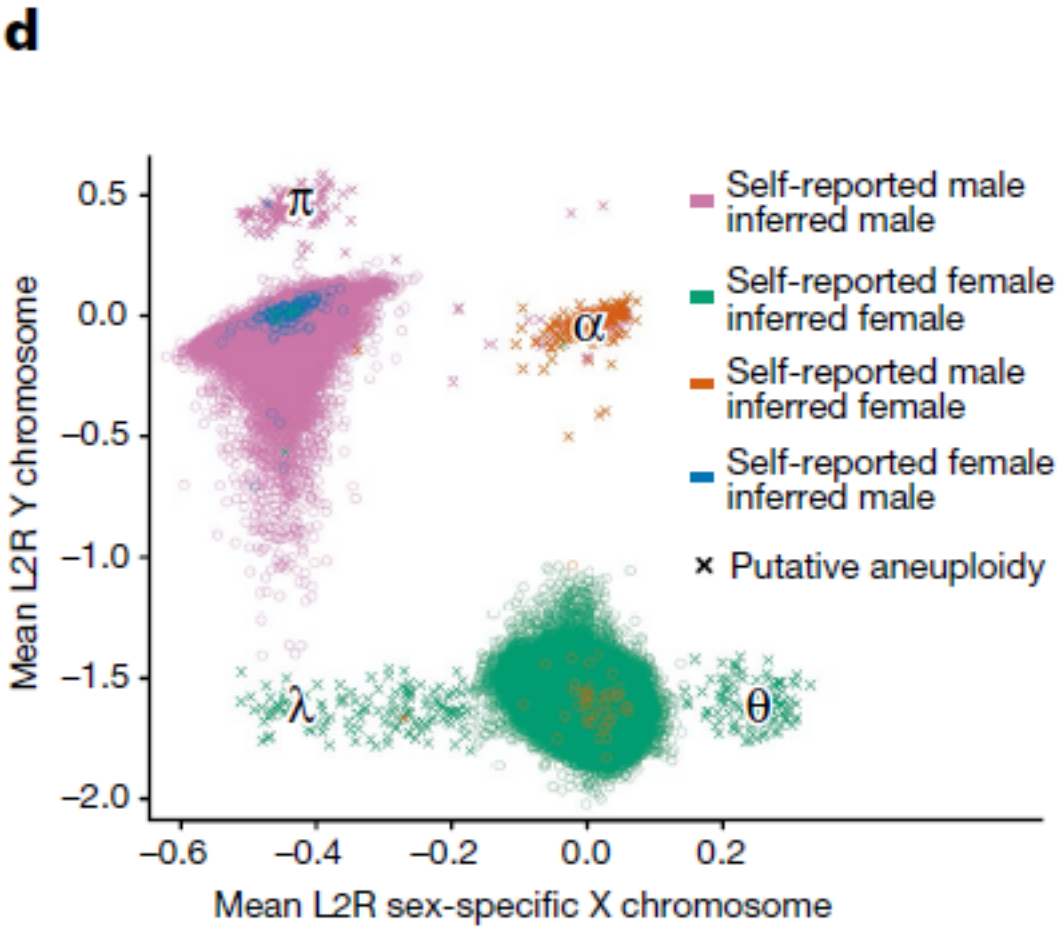
UK Biobank: overview

~500,000 people
Genotyping
Phenotyping

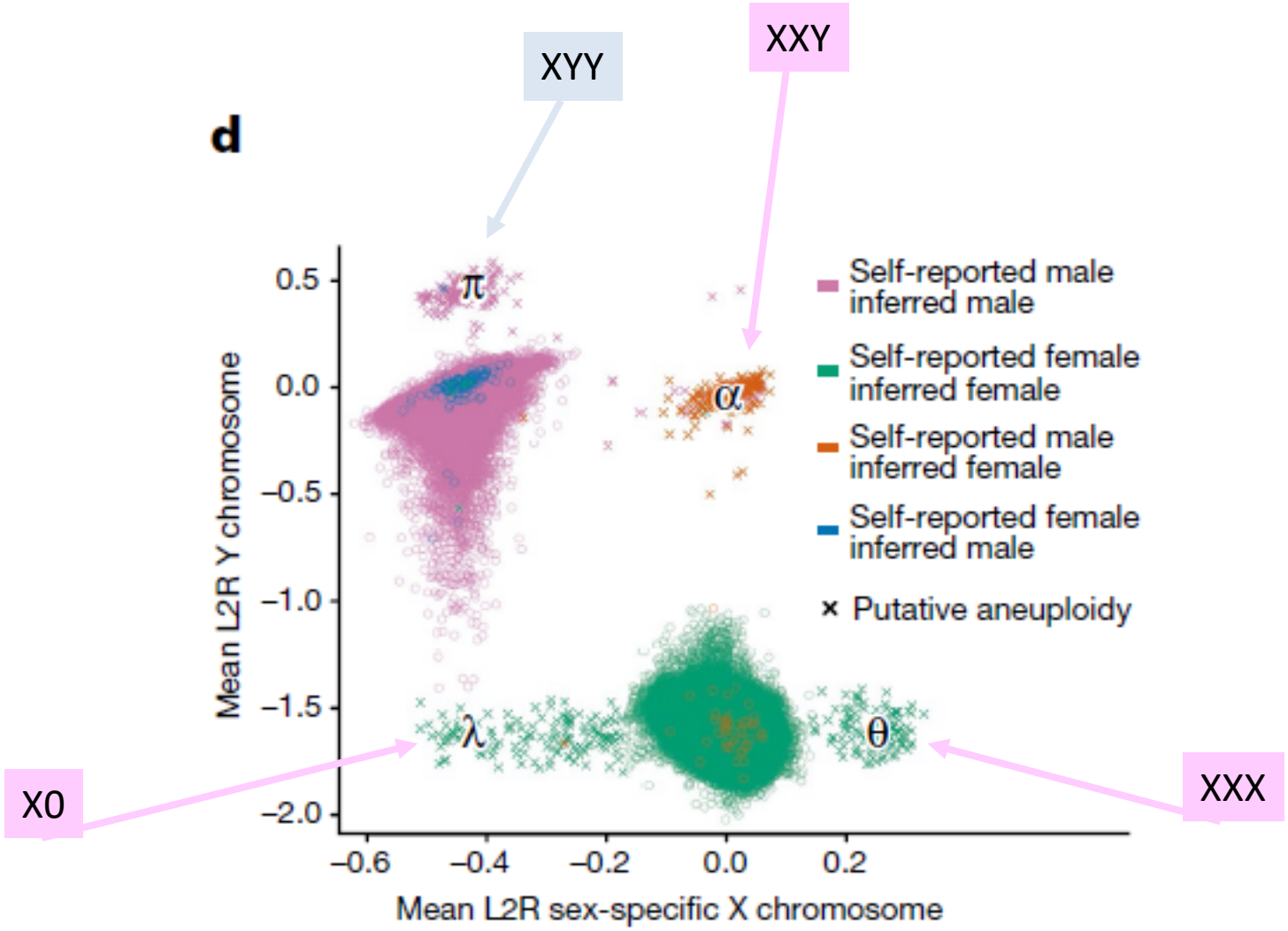
Prospective cohort study



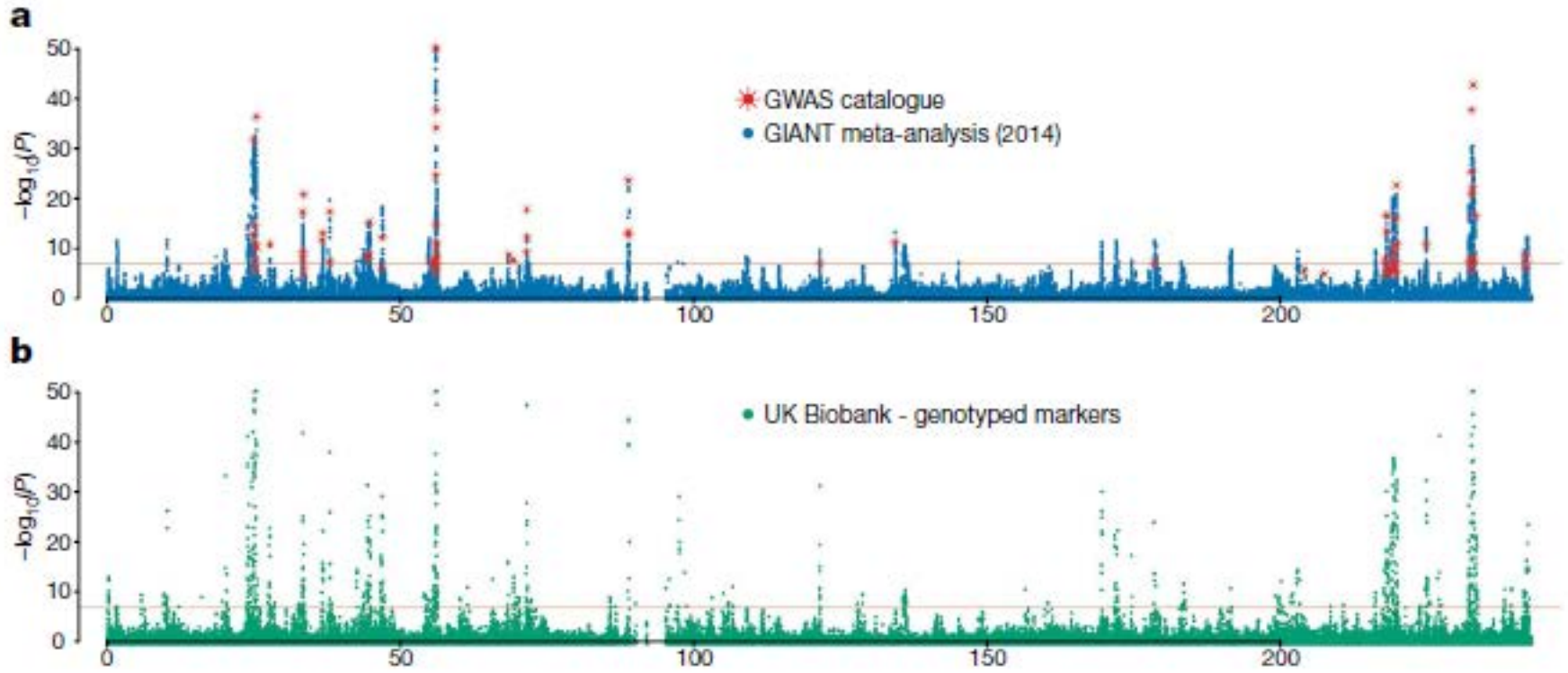
UK Biobank: non-standard XX/XY sex genotypes



UK Biobank: non-standard XX/XY sex genotypes



UK Biobank: height GWAS, more detection power



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

THANKS FOR YOUR ATTENTION

Further reading

Sequencing pools of individuals - mining genome-wide polymorphism data without big funding.
Schlötterer C, Tobler R, Kofler R, Nolte V.
Nat Rev Genet. 2014 Nov;15(11):749-63. doi: 10.1038/nrg3803. Epub 2014 Sep 23. Review.