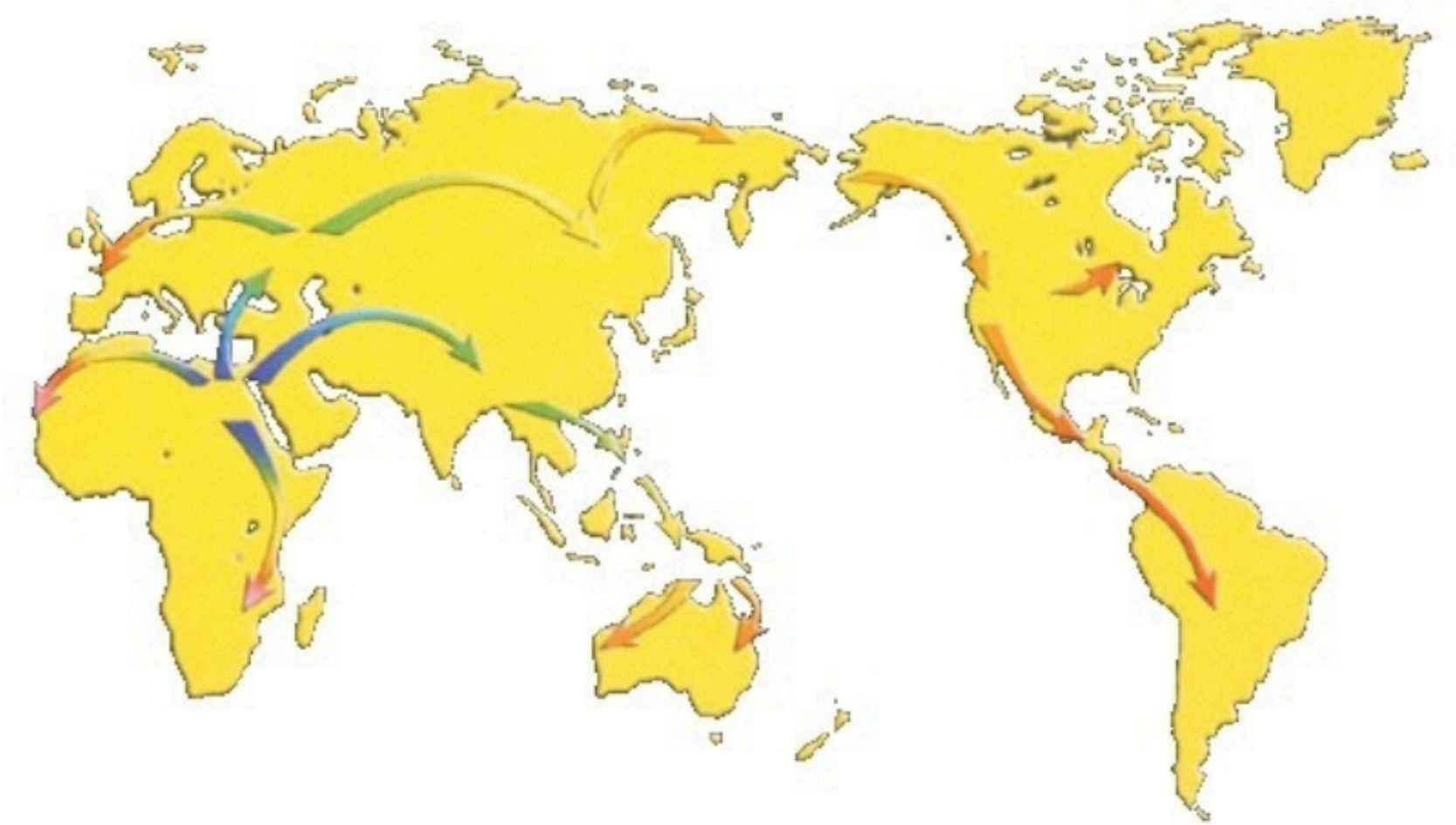
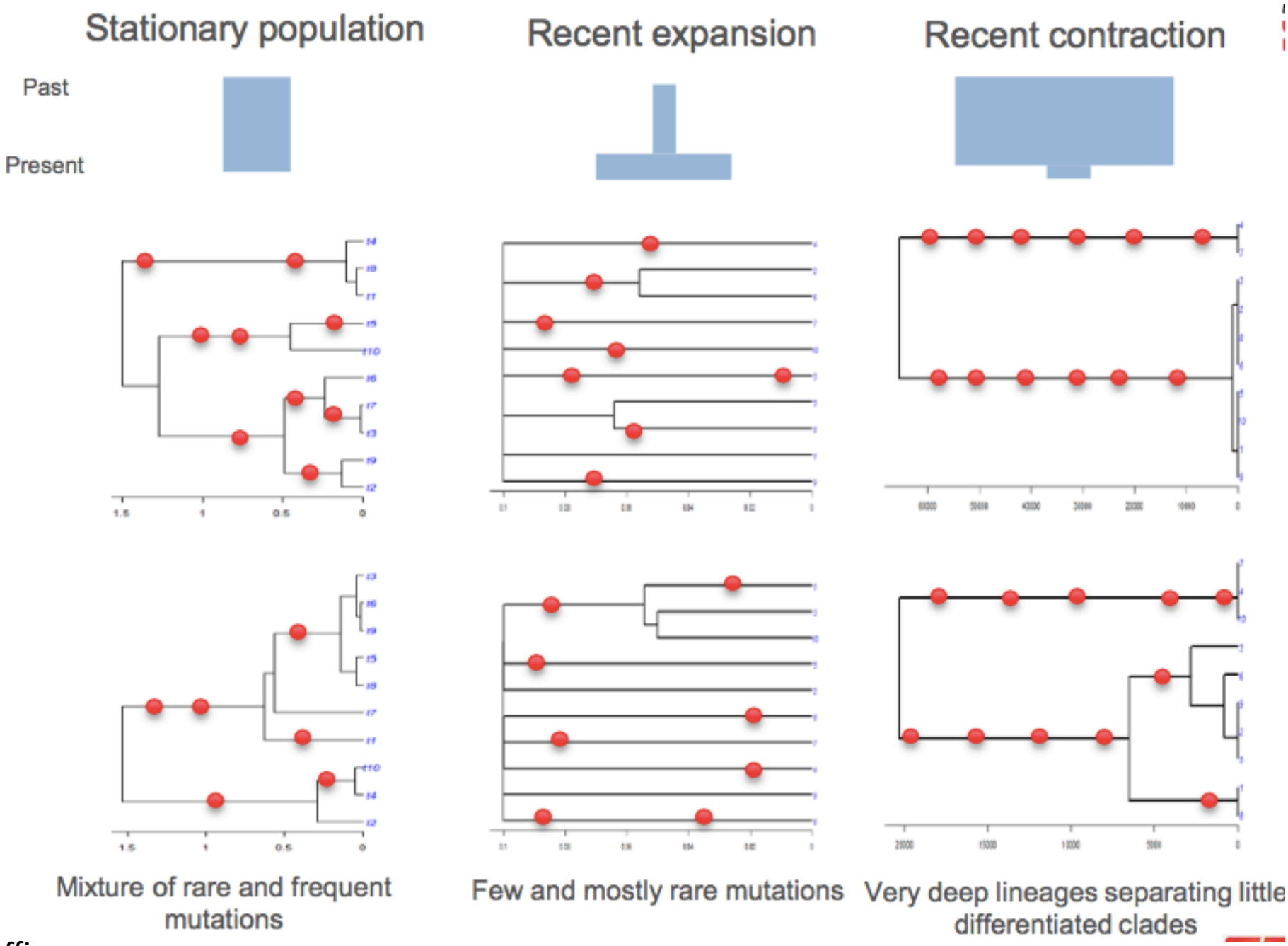
Demography

Why model demographic history?

- Understand population history
  Bottlenecks, gene flow, etc.



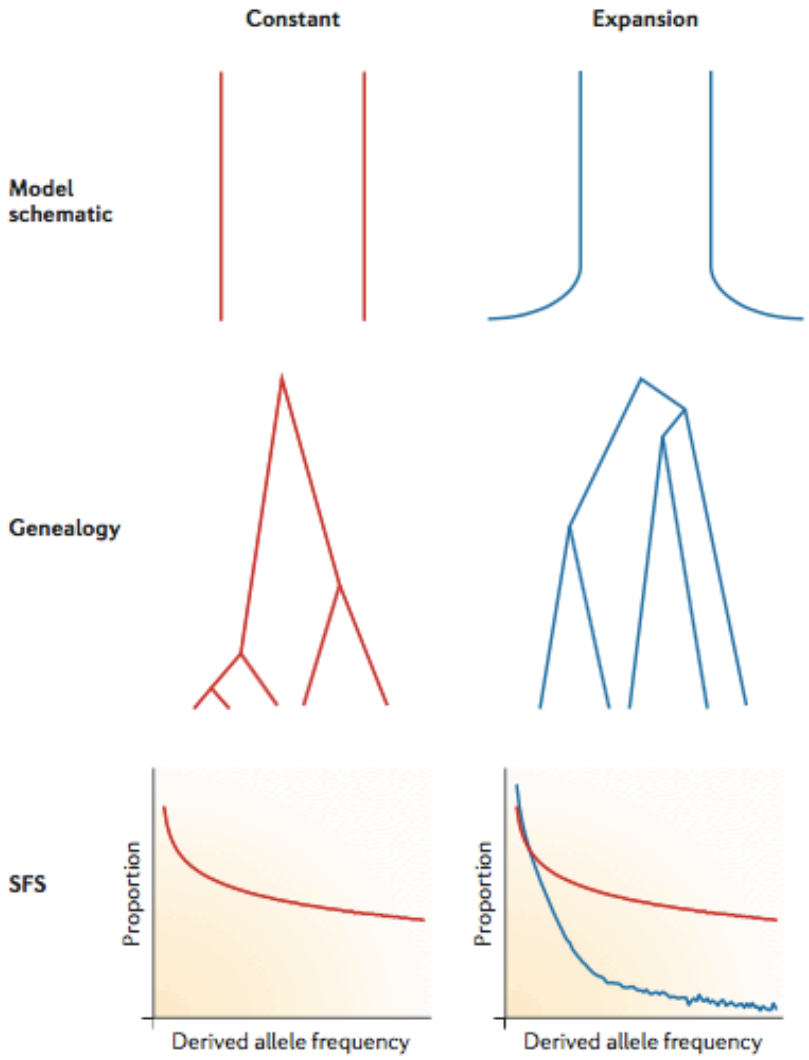
Ryan Gutenkunst


Stationary population
Recent expansion
Recent contraction
Past
Present
Mixture of rare and frequent mutations
Few and mostly rare mutations
Very deep lineages separating little differentiated clades

Excoffier

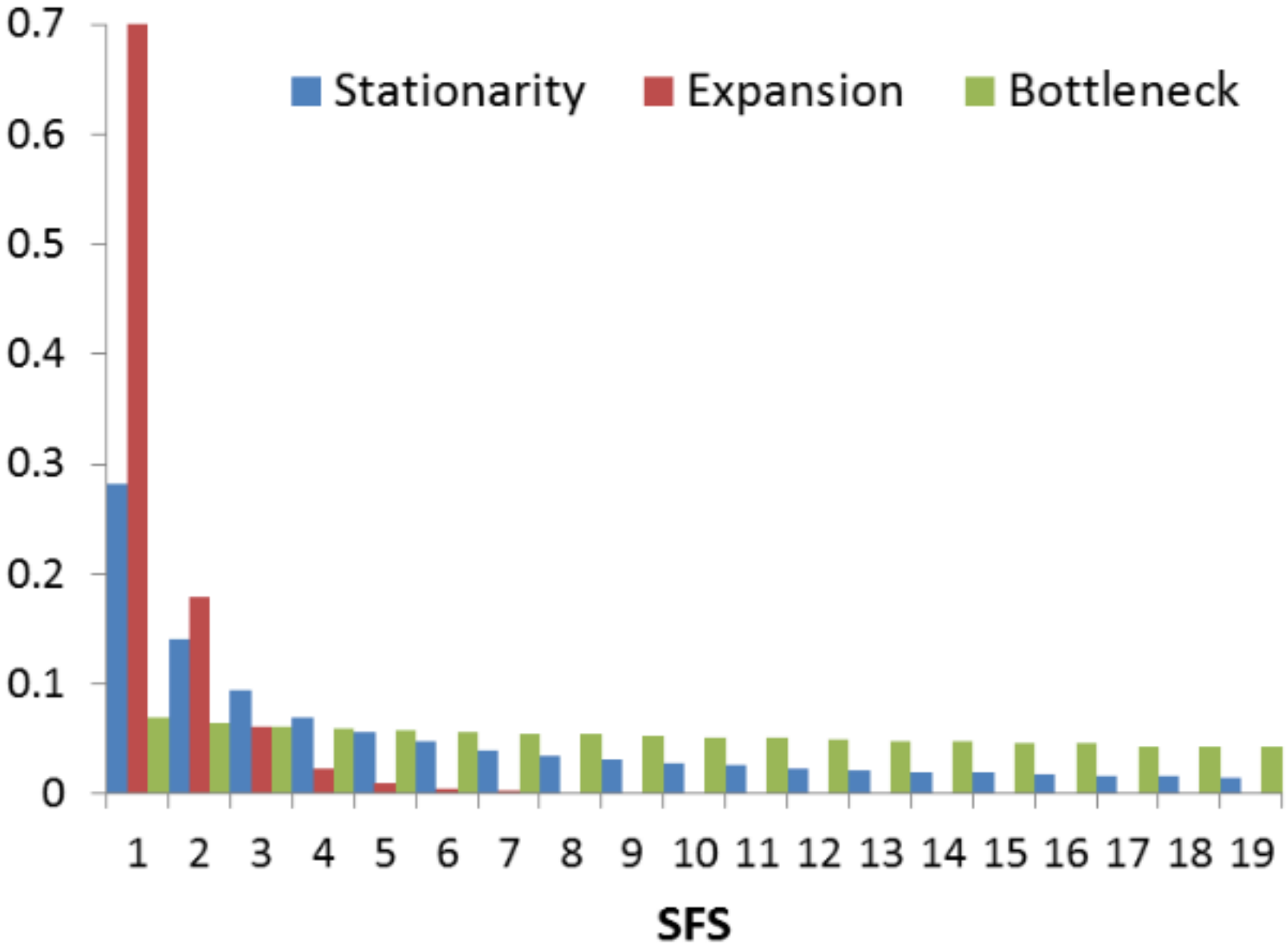
Allele Frequency Spectrum

Demography DISTORTS genealogies



Schraiber and Akey (2015) *Nat Rev Genet*

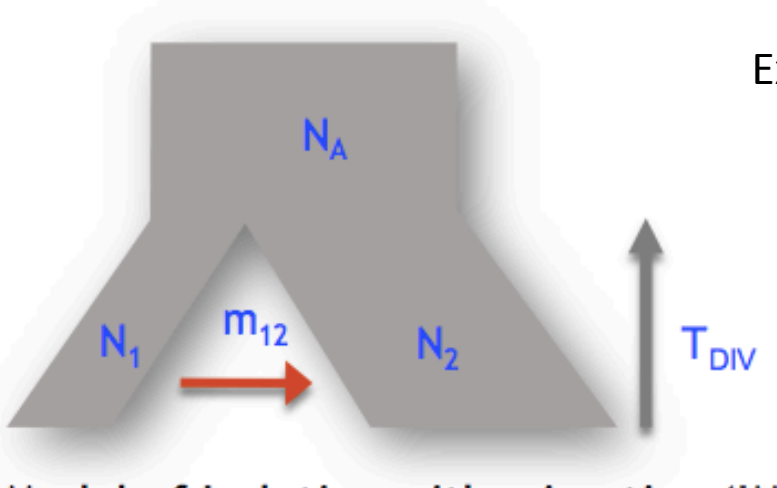
Ryan Gutenkunst



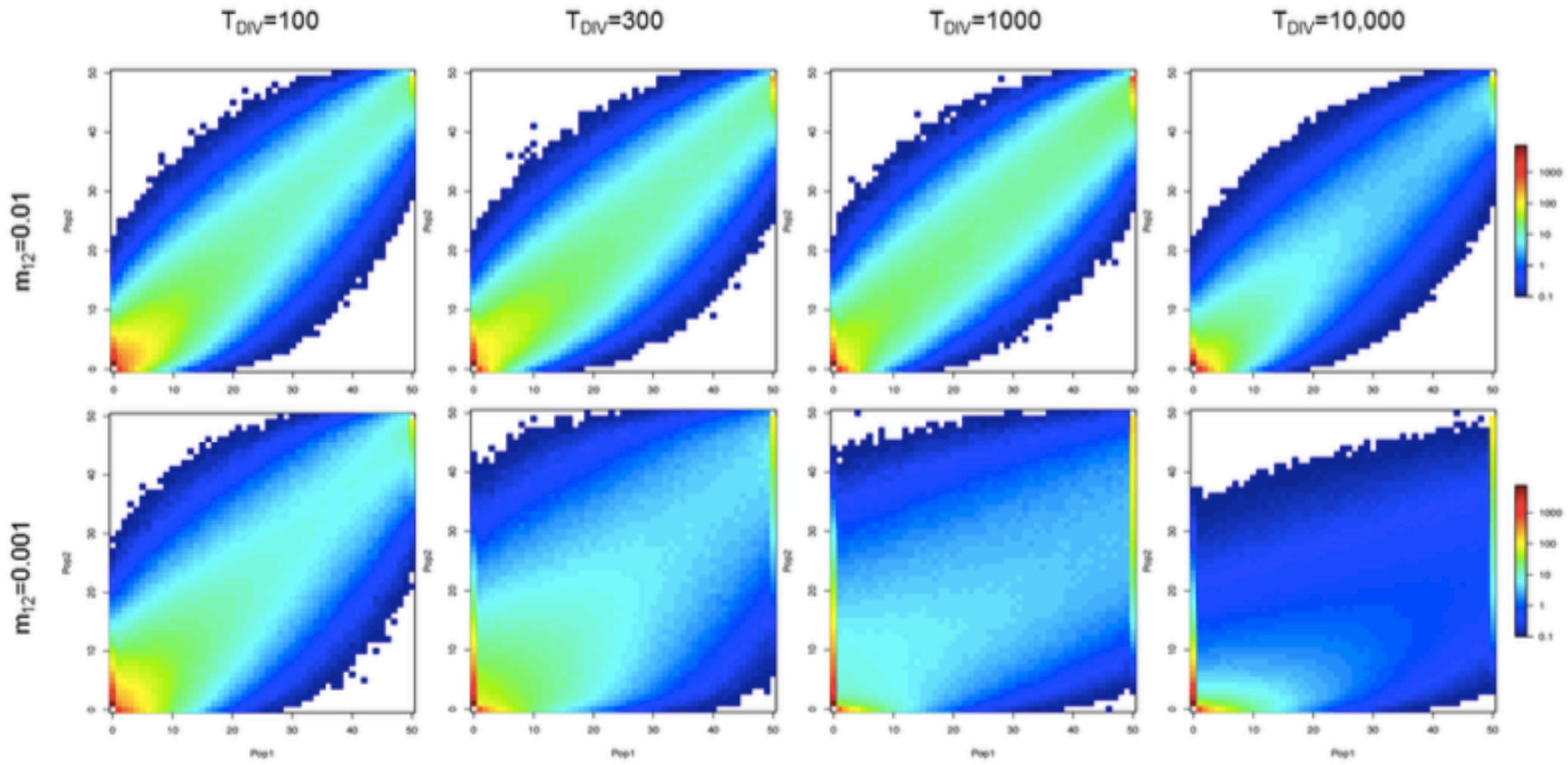
Ryan Gutenkunst

Joint SFS (2D-SFS)

Excoffier



Model of Isolation with migration (IM)



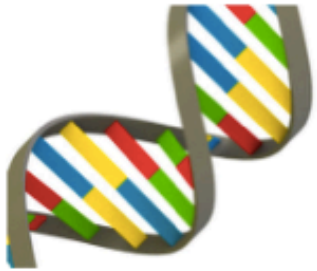
Using estimates of theta

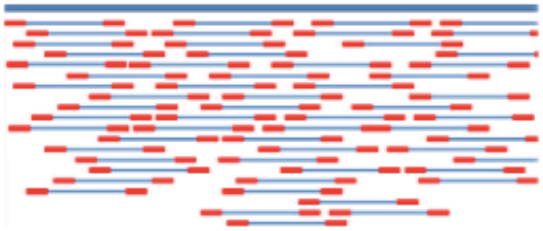
$$E(\pi) = \theta \qquad E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

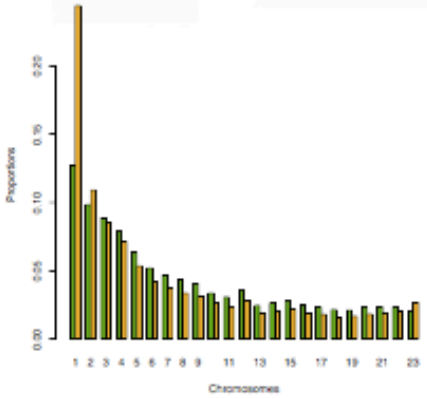
Tajima (1989)

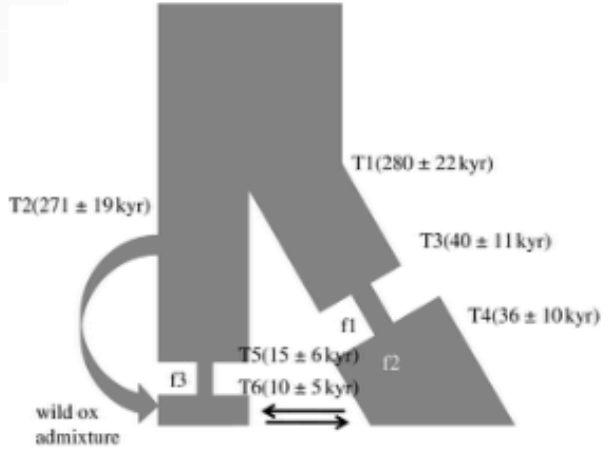
Tajima's D = $(\pi - \theta_W)/\text{stdev}(\pi - \theta_W)$

Workflow









Ryan Gutenkunst

Modeling workflow

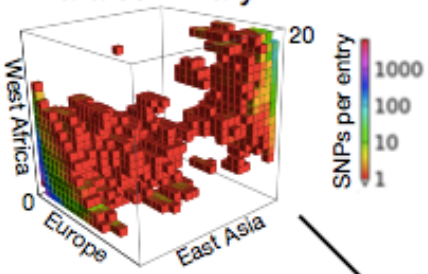
Data


TGGTCACTCTTAT
TGGTCACTCTTAT


TGGTCACTCTTAT
TGGTCACTCTTAT


TGGTCACTCTTAT
TGGTCACTCTTAT

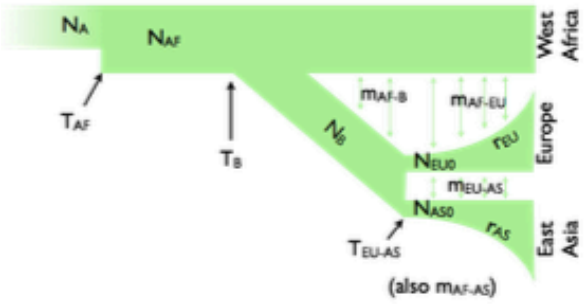
Data summary



Compare simulations with data.

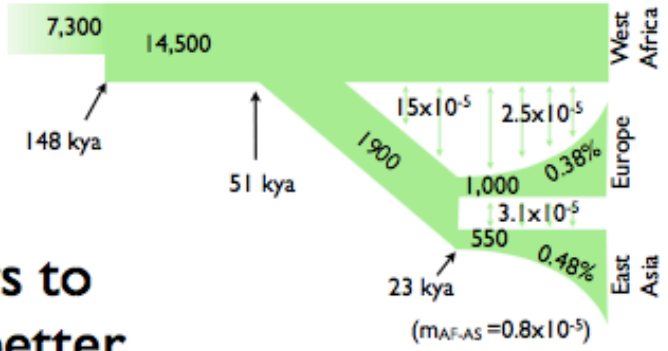
Choose candidate parameter values, simulate data.

Input parameterized model



Update parameters to (hopefully) fit data better.

Converged output model



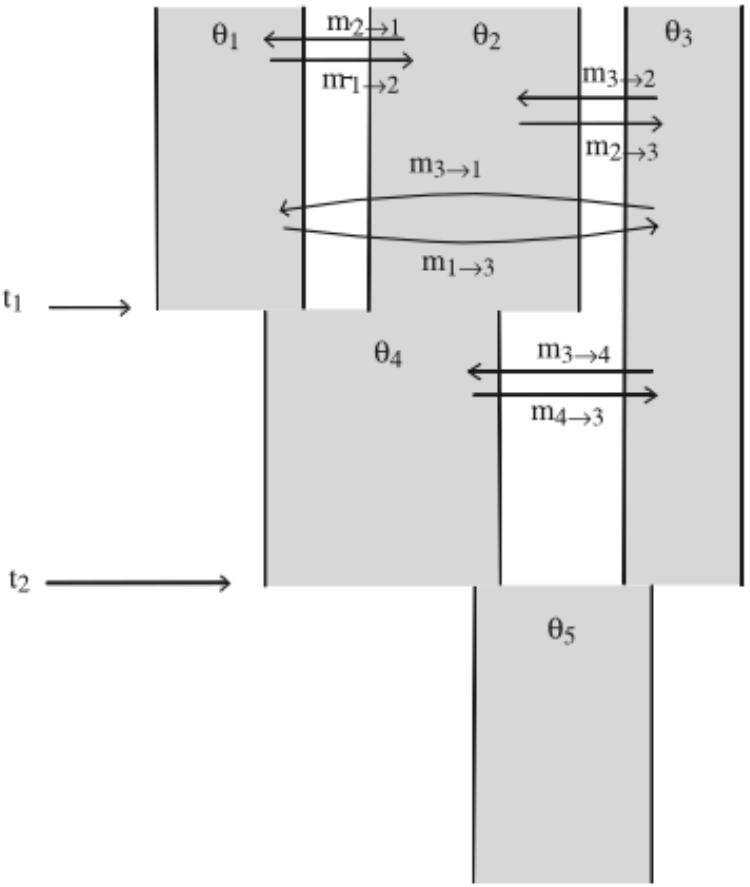
Ryan Gutenkunst

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander et al.[40] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search: redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[111]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

Ryan Gutenkunst

IM/IMa/IMa2

- Uses coalescent simulation to calculate the full likelihood of the data given the model, for non-recombining regions (mitochondria, Y chromosome, small autosomal regions).
- Bayesian inference based on MCMC walk through parameter space, can be computationally expensive.
- Handles arbitrary number of populations.

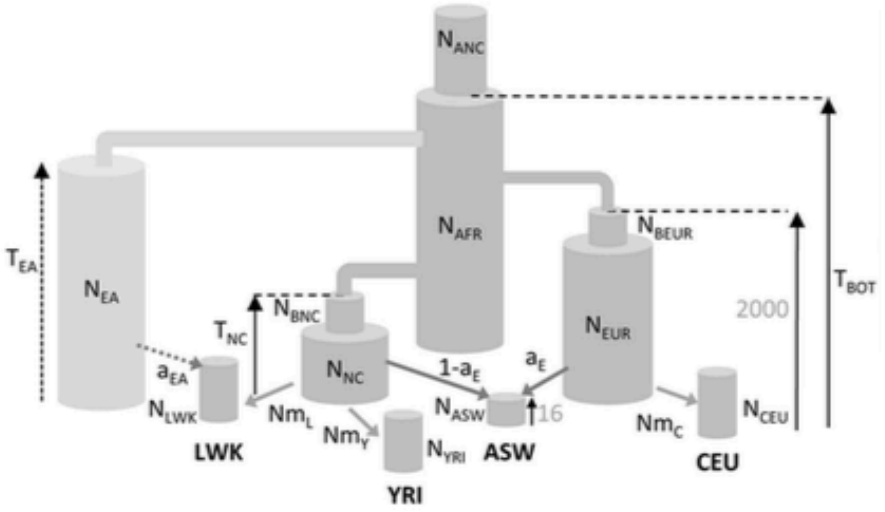


Hey and Nielsen (2004) *Genetics*
Hey (2010) *Mol Biol Evol*

Ryan Gutenkunst

fastsimcoal2

- Estimate pairwise joint frequency spectra using coalescent simulations.
- Scales to arbitrary number of populations.
- Estimate parameters by maximum composite likelihood.
- Optimization may be more robust than ∂a∂i.



Excoffier et al.
PLoS Genet (2013)

Ryan Gutenkunst

∂a∂i: Diffusion Approximations for Demographic Inference

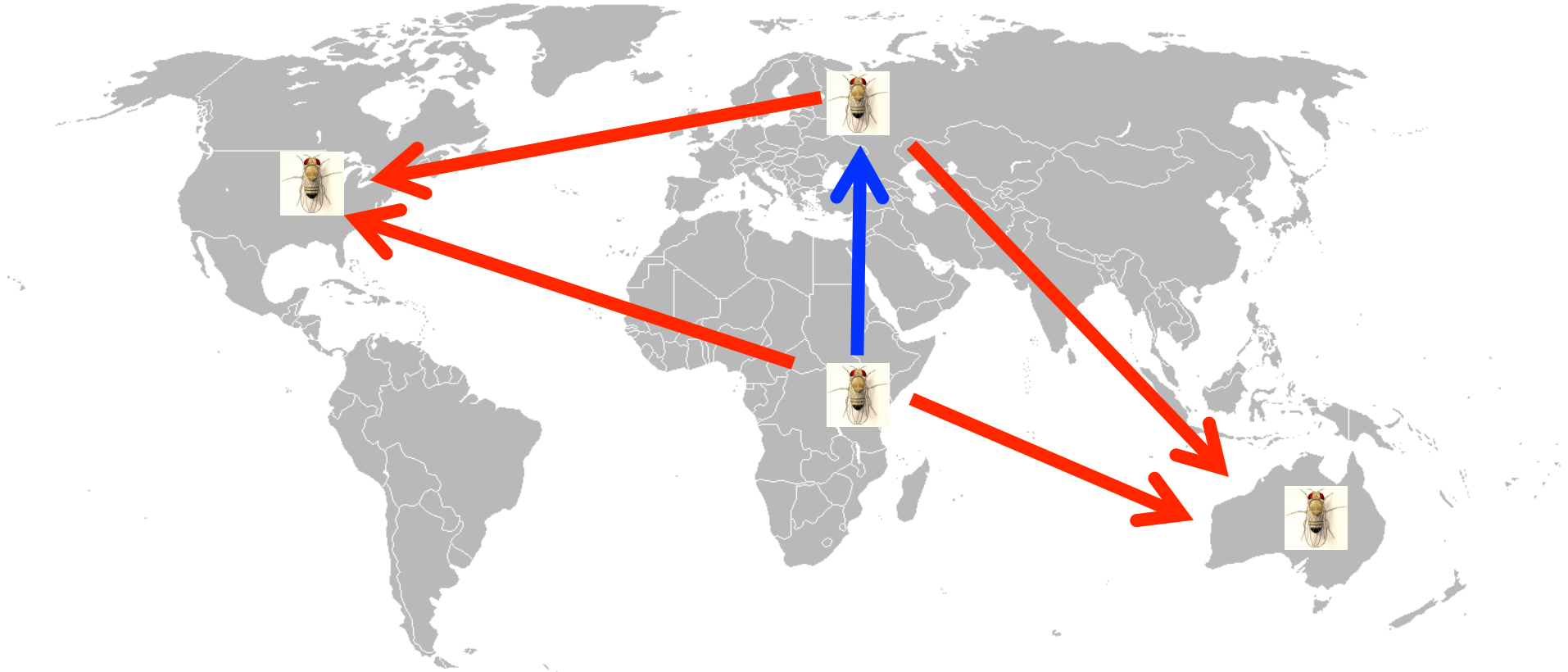
- Up to three interacting populations, with arbitrary parameter time courses
- 1 pop, 20 samples, ~3 params: ~1 minute to fit
  2 pops, 20 samples each, ~6 params: ~10 minutes to fit
  3 pops, 20 samples each, ~12 params: ~3 hours to fit
- Computational cost independent of SNP count, but exponential in number of populations.

Gutenkunst et al.
PLoS Genet (2009)

Ryan Gutenkunst

Demographic History of *Drosophila melanogaster*

D. melanogaster demography



What is the demographic history of East African *D. melanogaster*?

Approach: Data collection

- Sample: 20 strains from Uganda
- Target Region: 2 Mb X chromosome region
- Sequencing strategy: Illumina
  - Barcoded genomic library preparation
  - Multiplex selective enrichment
    - Nimblegen chip-capture
    - 385,000 oligo array
  - Single end (86 bp reads)
    - 2.5 million reads/strain

Approach: Bioinformatics

- Reads mapped with BWA
  - ~72% reads map uniquely
    - ~90% map to target
- Alignments processed in SAMtools
  - Final coverage: 32.4X
  - ~89% sites with ≥ 2X
- SNP calls:

Approach: Bioinformatics

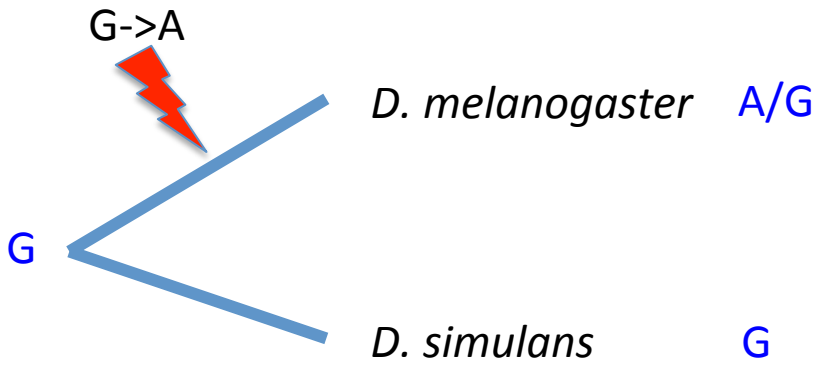
- Reads mapped with BWA
  - ~72% reads map uniquely
    - ~90% map to target
- Alignments processed in SAMtools
  - Final coverage: 32.4X
  - ~89% sites with ≥ 2X
- SNP calls: Joint Genotyper for Inbred Lines
  - Simultaneously considers all reads (per site) across lines
  - Assumes shared error profile across lines
  - Line genotypes depend on population frequency, error

Demographic modeling: dadi

- Restrict to third codon positions

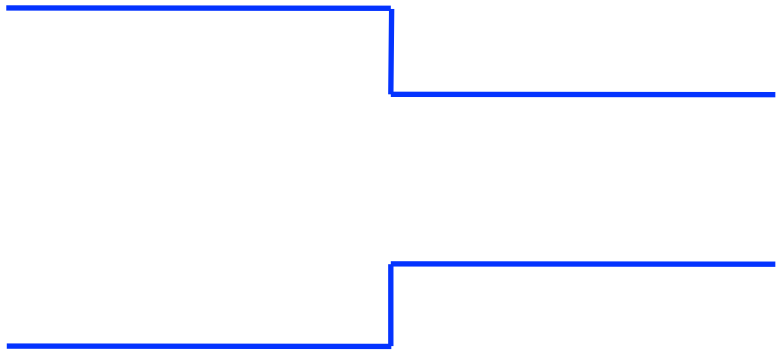
Demographic modeling: dadi

- Restrict to third codon positions
- Polarize polymorphisms (*D. simulans*)



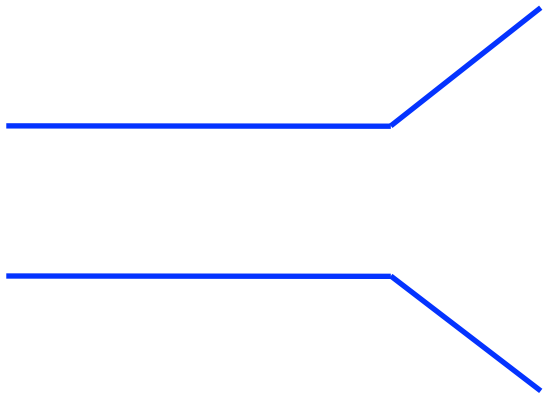
Demographic modeling: dadi

- Restrict to third codon positions
- Polarize polymorphisms (*D. simulans*)
- Five models
  - Neutral
  - Two epoch
  - Growth
  - Bottlegrowth
  - Three Epoch



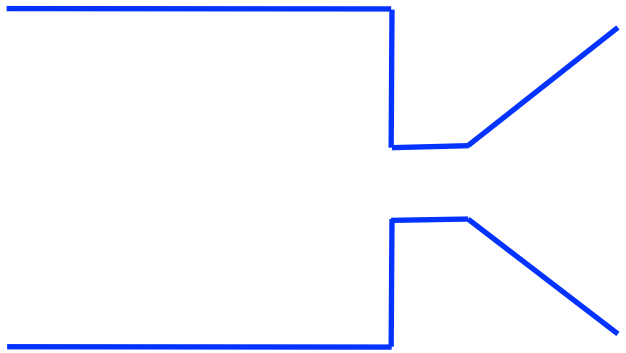
Demographic modeling: dadi

- Restrict to third codon positions
- Polarize polymorphisms (*D. simulans*)
- Five models
  - Neutral
  - Two epoch
  - Growth
  - Bottlegrowth
  - Three Epoch



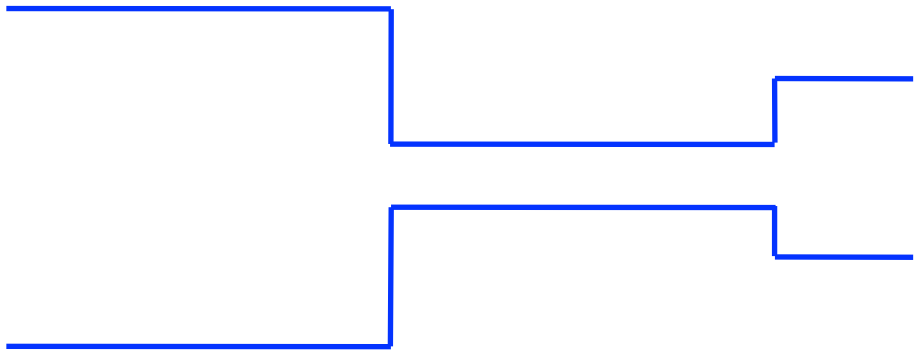
Demographic modeling: dadi

- Restrict to third codon positions
- Polarize polymorphisms (*D. simulans*)
- Five models
  - Neutral
  - Two epoch
  - Growth
  - Bottlegrowth
  - Three Epoch



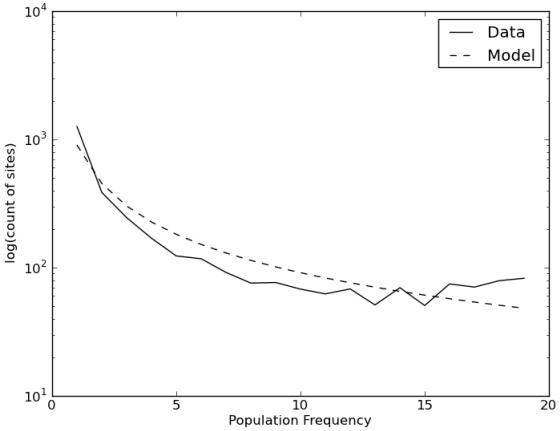
Demographic modeling: dadi

- Restrict to third codon positions
- Polarize polymorphisms (*D. simulans*)
- Five models
  - Neutral
  - Two epoch
  - Growth
  - Bottlegrowth
  - Three epoch

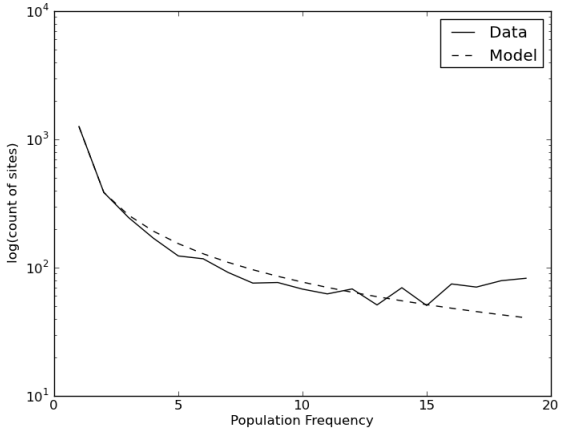


Results: Demography

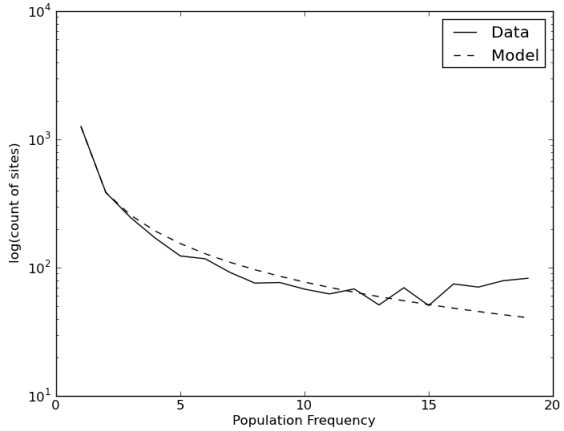
Neutral; LL = -206.63



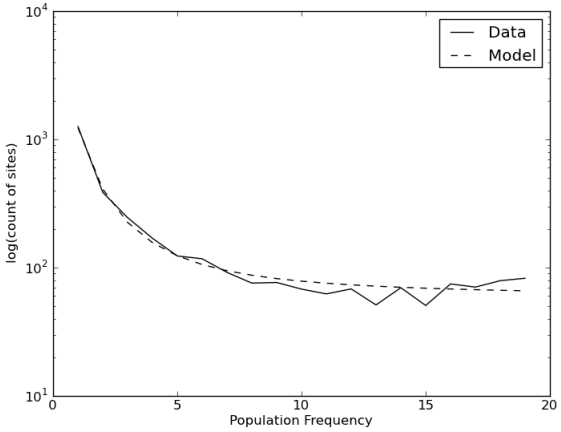
Two Epoch; LL = -116.00



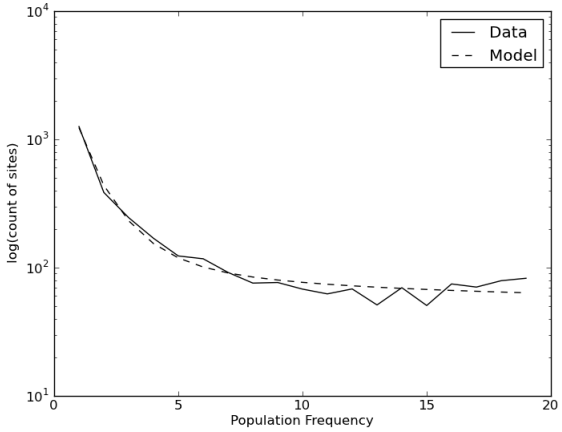
Growth; LL = -116.00



BottleGrowth; LL = -76.64

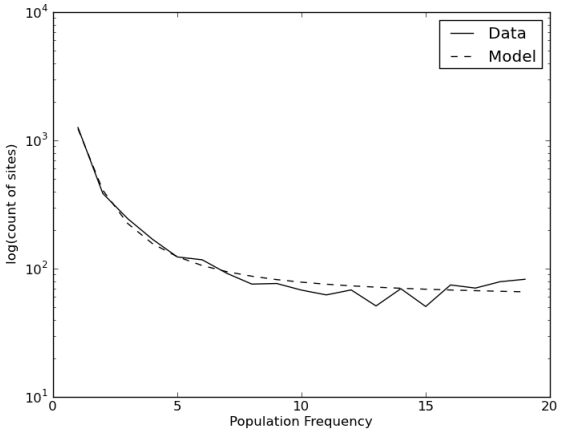


Three Epoch; LL = -79.89

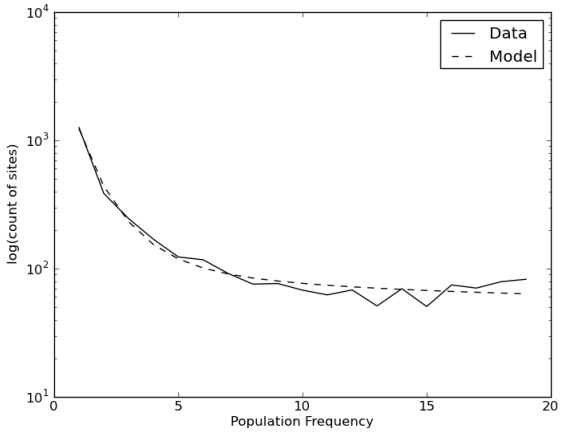


Results: Demography

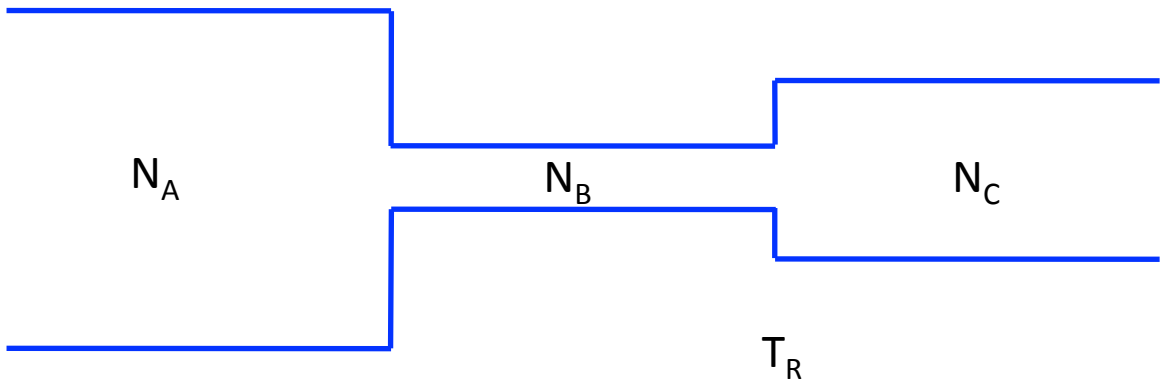
BottleGrowth; LL = -76.64



Three Epoch; LL = -79.89



Results: Demography



N_B/N_A = 0.016- 0.064

N_C/N_A = 0.47- 0.76

T_R= 0.14 (~28,000 ybp)

Summary: *D. melanogaster*

- Population genetic data from ancestral population
  - Next-generation sequencing
  - Targeted enrichment
  - JGIL
- Nonequilibrium demography in Uganda
  - Population contraction followed by expansion

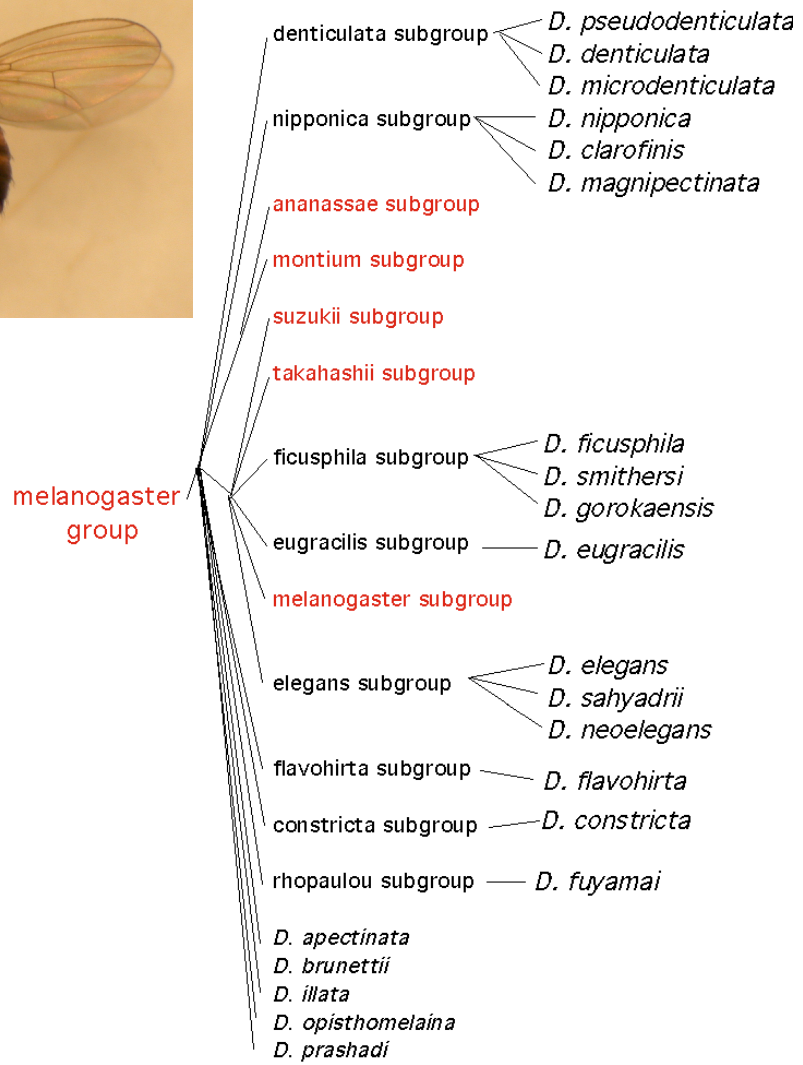
Demography in *Drosophila suzukii*

D. suzukii

- Native to Southeast Asia

D. suzukii

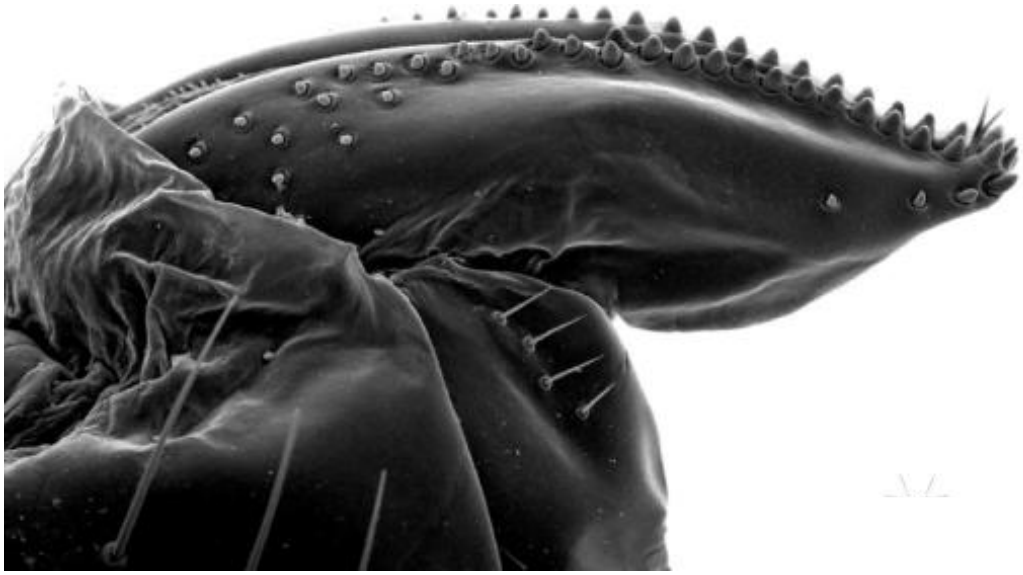
- Native to Southeast Asia
- Closely related to *D. melanogaster*

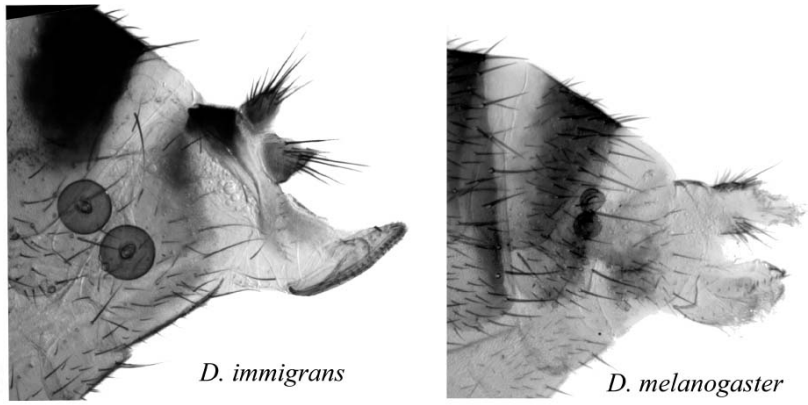


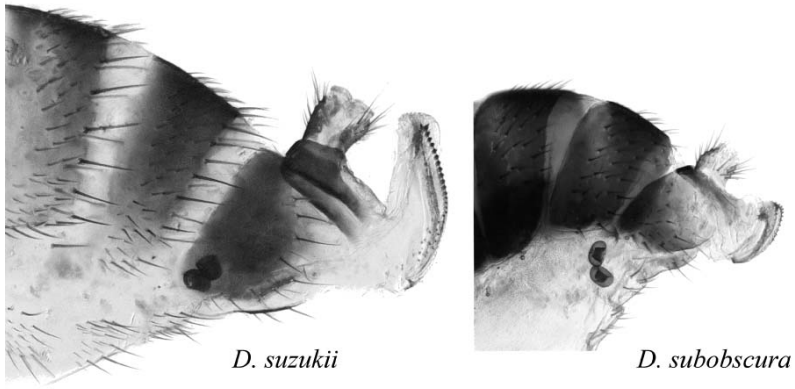
D. suzukii

- Native to Southeast Asia
- Closely related to *D. melanogaster*
- Pest of soft-skinned fruits

Female morphology

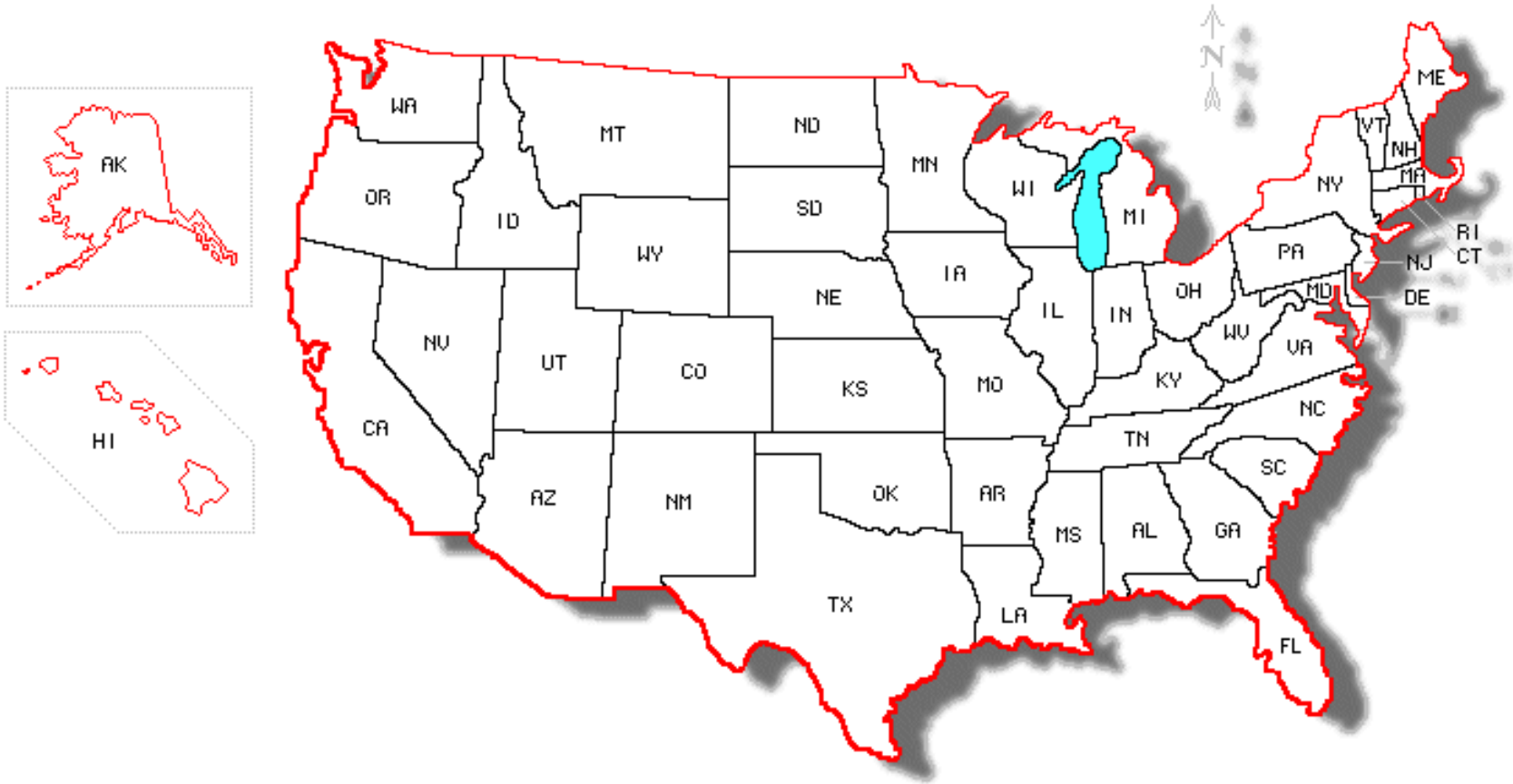


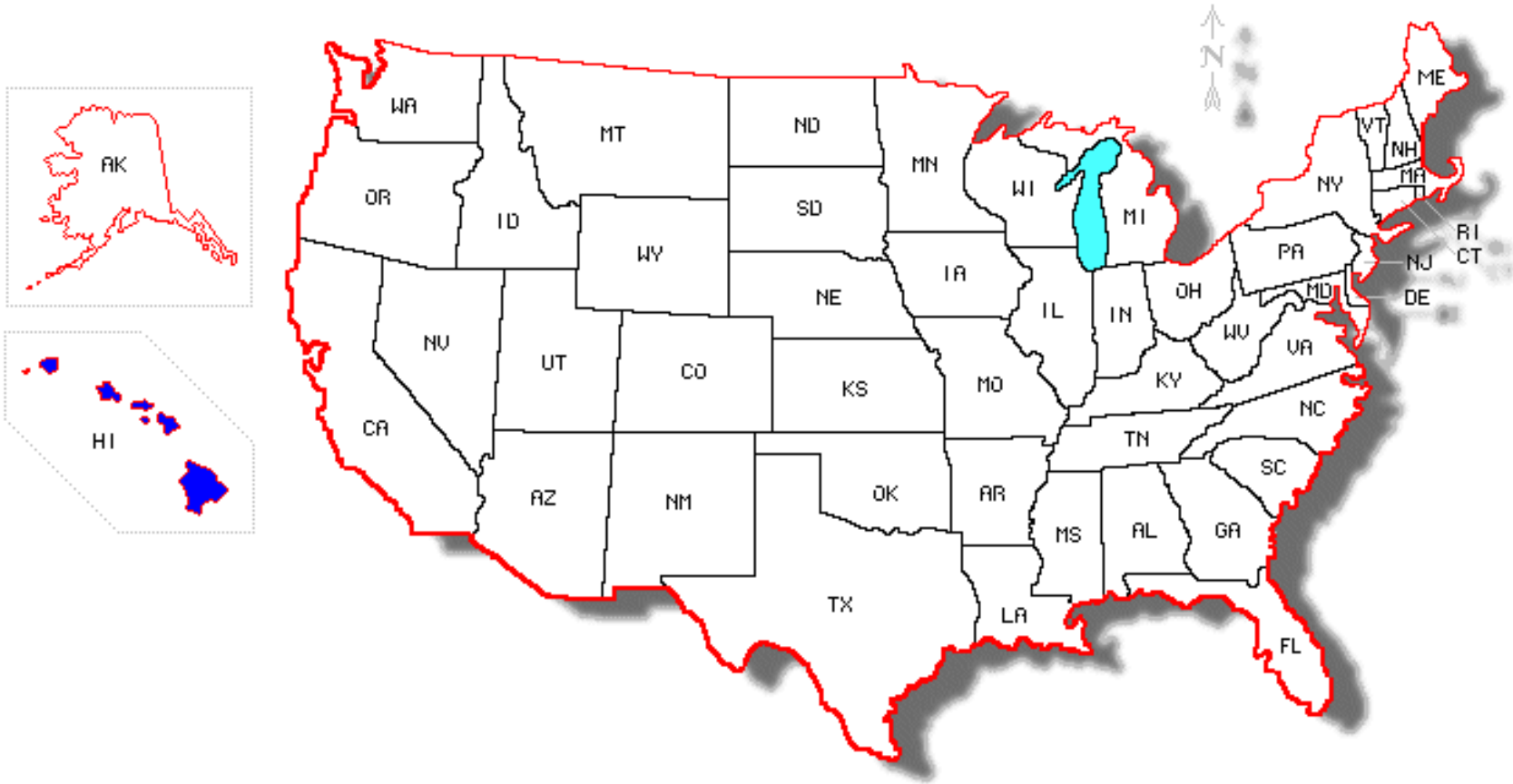


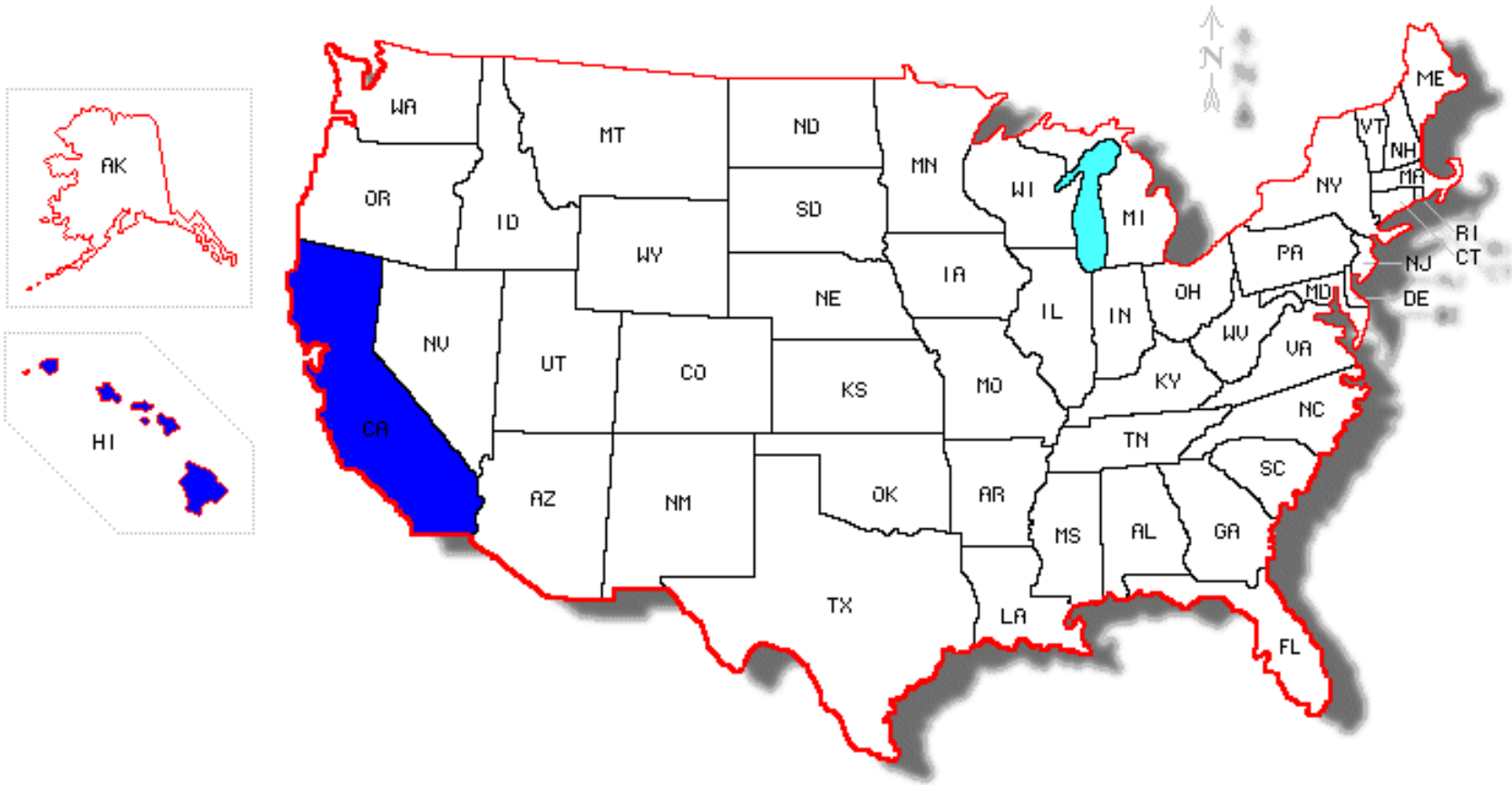


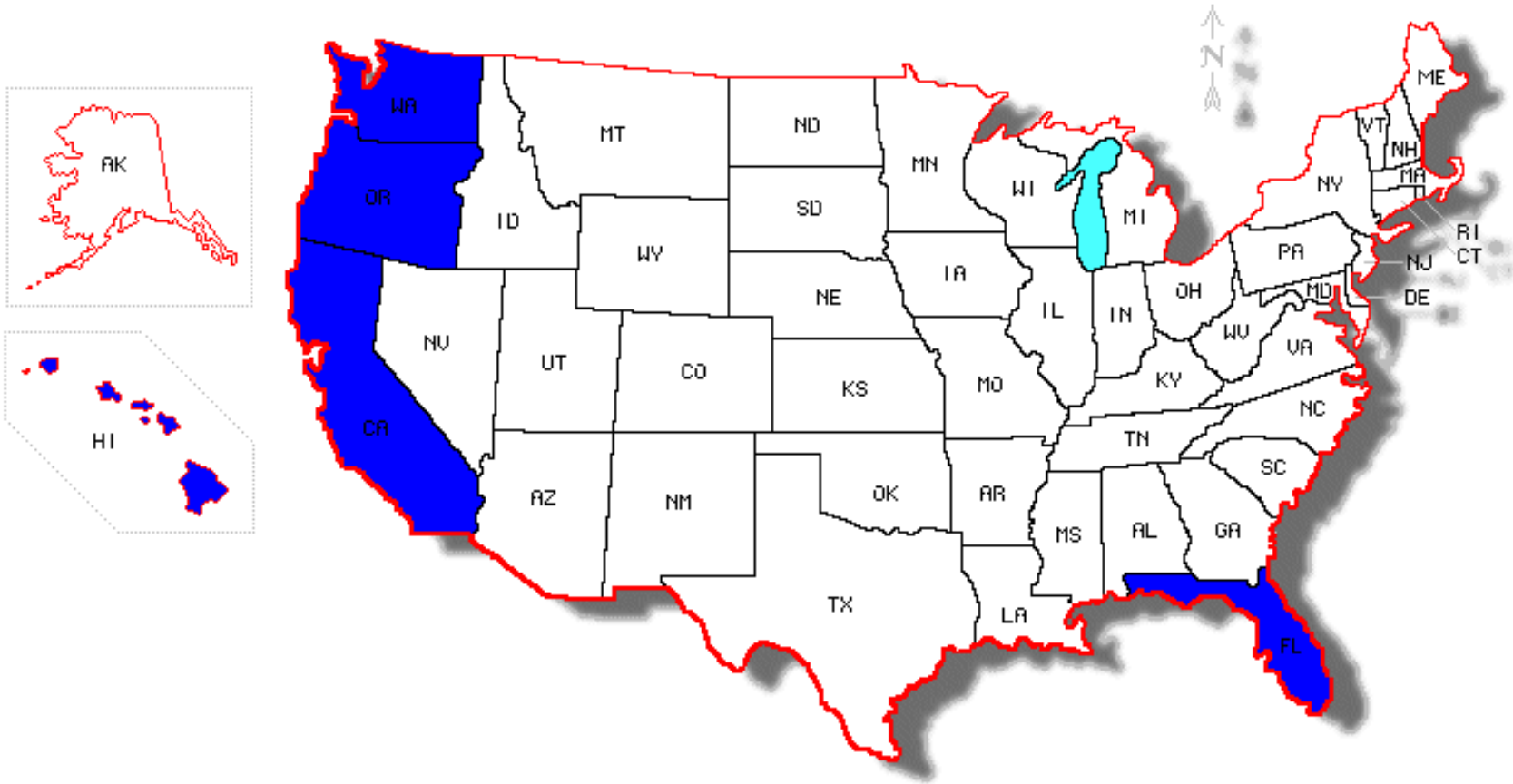
Photos courtesy of H. Burrack

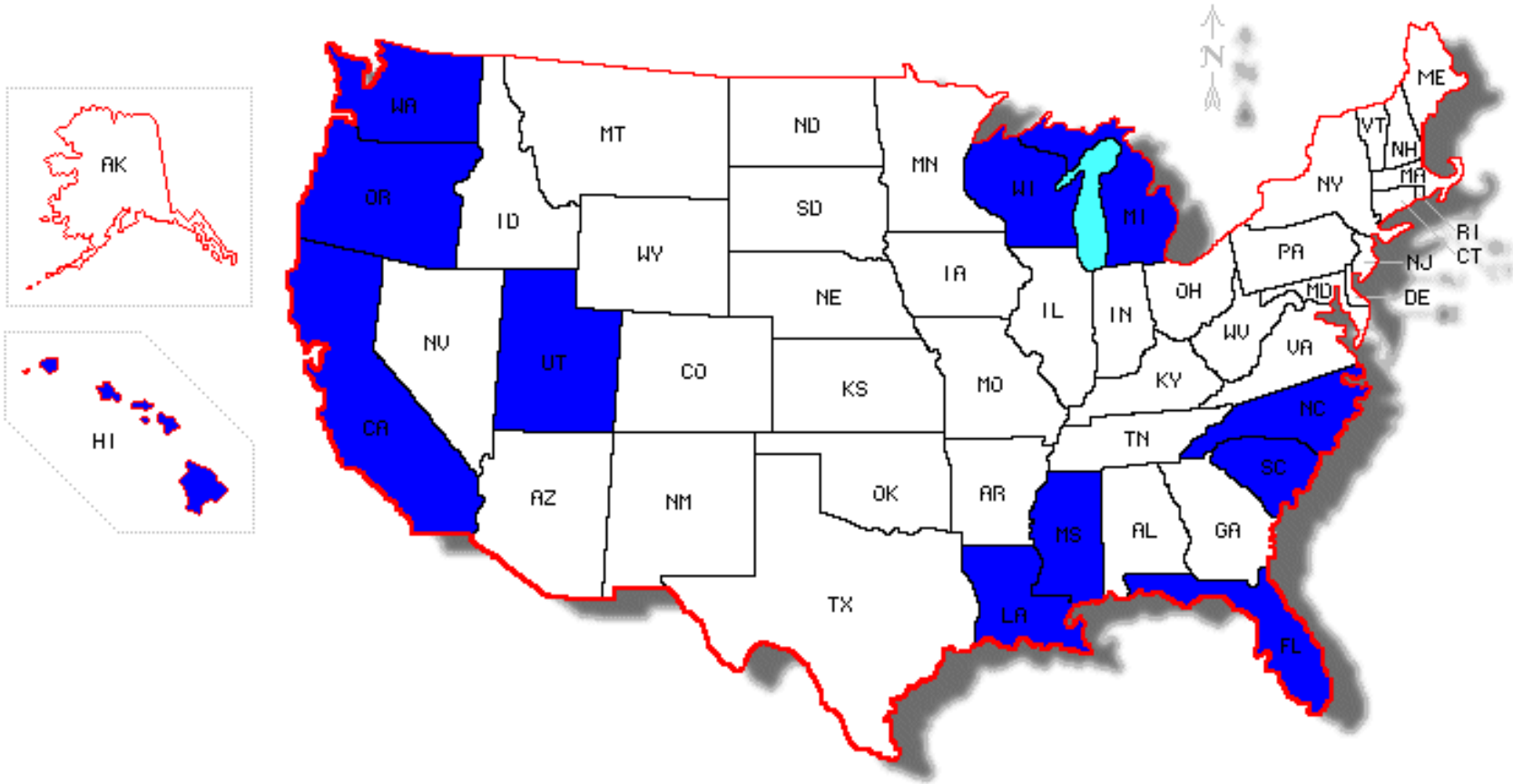
Hauser 2011

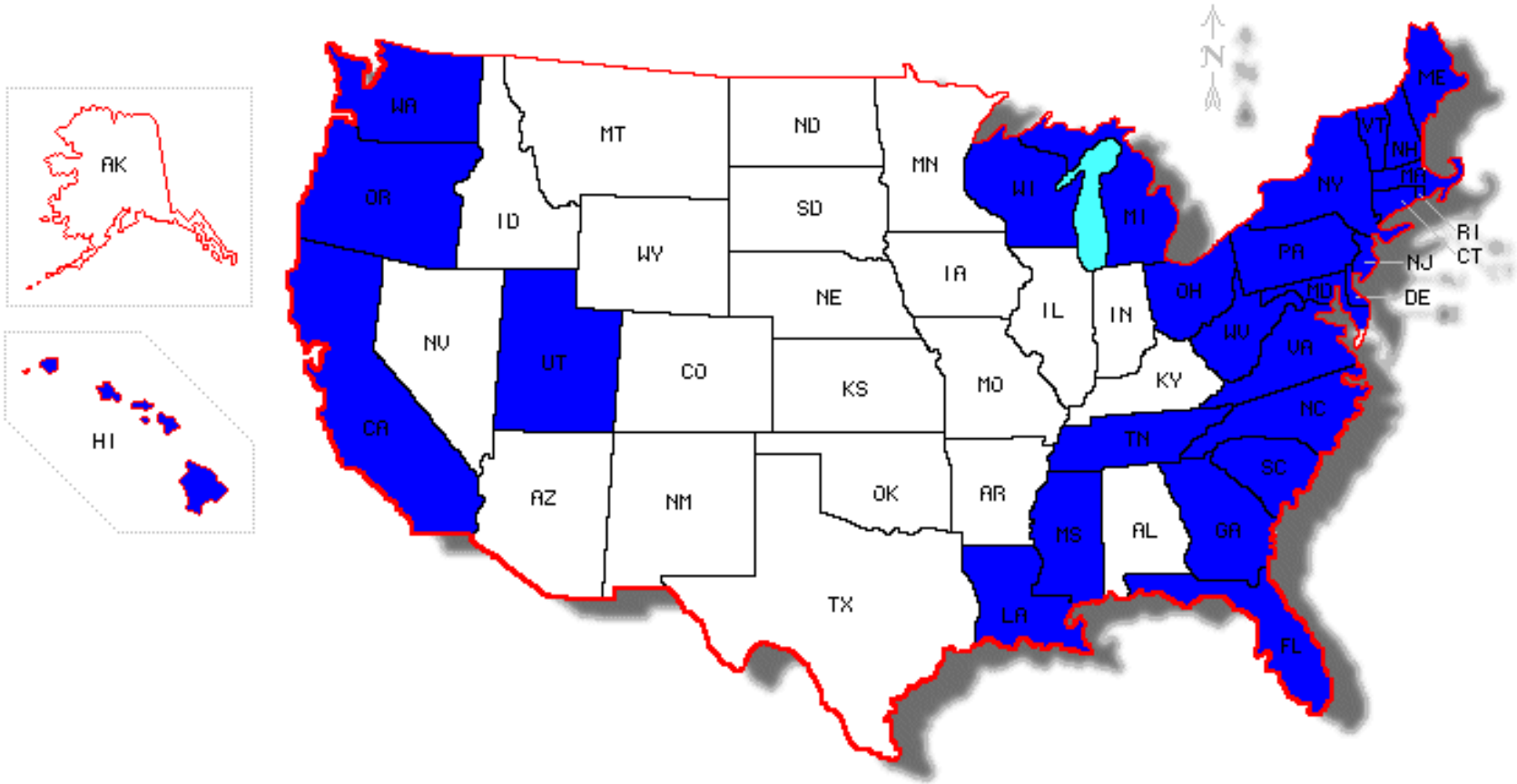

AK
HI
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME


AK
HI
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME


AK
HI
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME
N


AK
HI
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME


AK
HI
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME


AK
HI
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
VT
NH
ME
N

Economic impact: West Coast

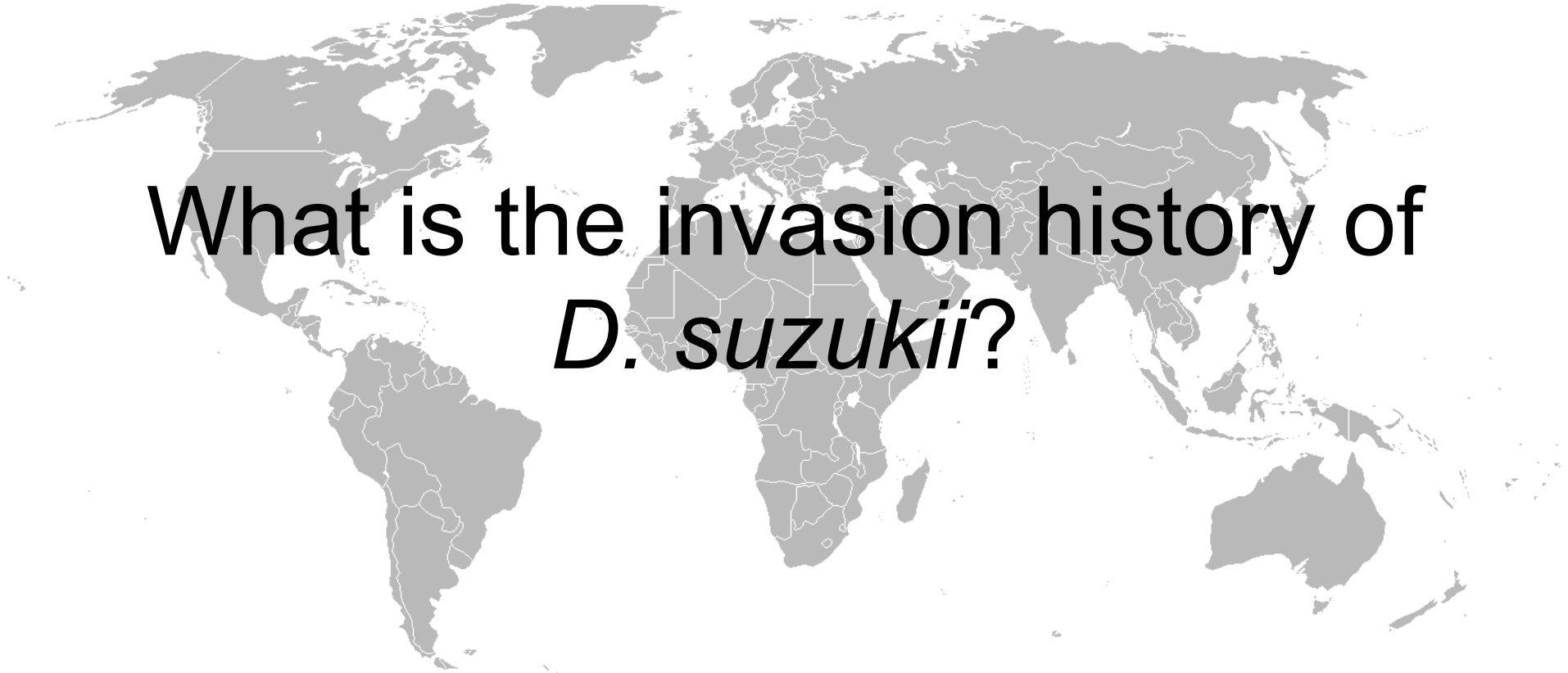
Table 1. Revenue Losses Due to SWD: 20% Yield Loss, 2008 Value of Production				
	California	Oregon	Washington	Three-state Total
Strawberries				
Total farmgate value ($Million)	1,544.7	16.8	10.1	1,571.5
Share of U.S. production (%)	82	1	1	83
Total losses ($Million)	308.9	3.4	2.0	314.3
Blueberries (cultivated)				
Total farmgate value ($Million)	49.1	49.4	43.4	141.9
Share of U.S. production (%)	9	9	8	26
Total losses ($Million)	9.8	9.9	8.7	28.4
Raspberries and Blackberries				
Total farmgate value ($Million)	179.5	41.7	92.1	313.3
Share of U.S. production (%)	57	13	29	100
Total losses ($Million)	35.9	8.3	18.4	62.7
Cherries				
Total farmgate value ($Million)	194.5	58.7	297.1	550.3
Share of U.S. production (%)	30	9	45	84
Total losses ($Million)	$38.3	$9.9	$57.8	$105.9
ALL CROPS				
Total farmgate value ($Million)	1,967.9	166.5	442.6	2,577.0
Share of U.S. production (%)	58	5	13	76
Total losses ($Million)	393.0	31.4	86.9	511.3
Source: Authors' calculations based on data from the National Agricultural Statistics Service (NASS), 2009.				

Bolda *et al.* 2009

Economic impact: East Coast

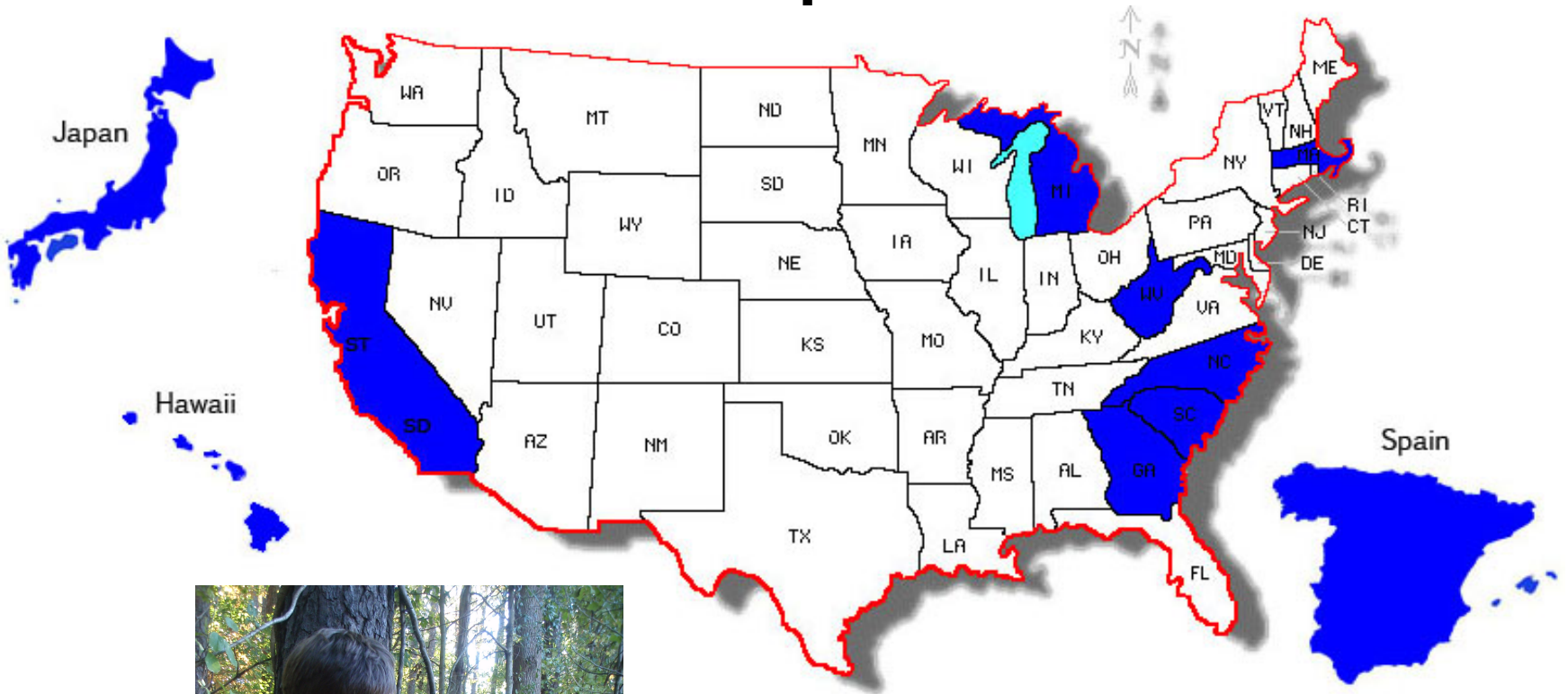
Crop	Total Farmgate Value ($1,000s)	Potential Yield Loss (%)	Potential Losses ($1,000s)
Blueberries	192,859	40	77,144
Caneberries	4,395	50	2,198
Peaches	144,005	20?	28,801
Fresh Strawberries	386,332	0?	0
Total	727,591		108,143

Courtesy of H. Burrack



What is the invasion history of *D. suzukii*?

Samples


Japan
Hawaii
Spain
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME

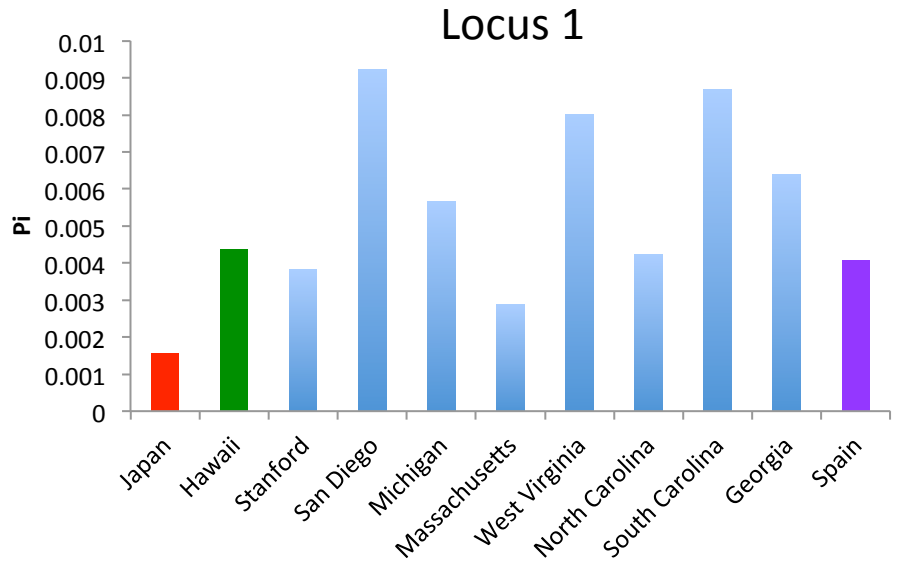
Markers

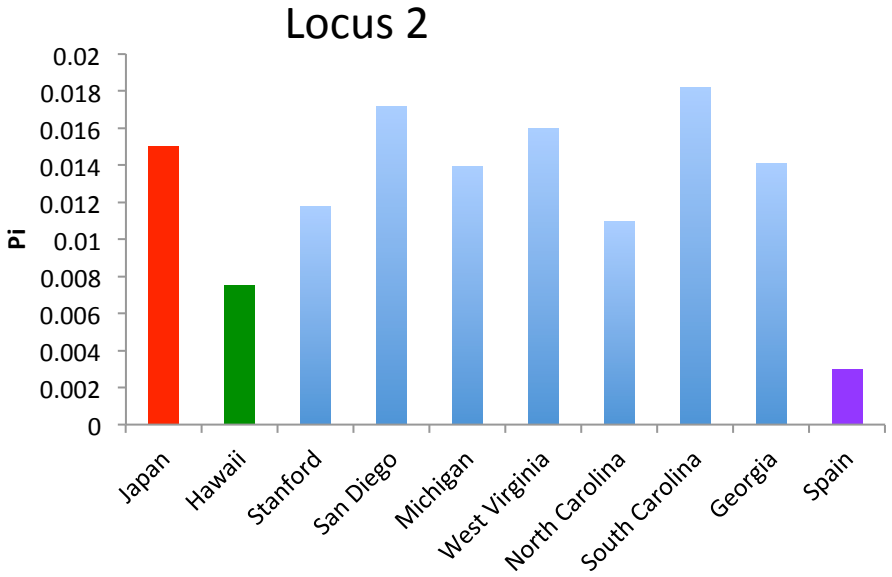
- Draft genome provided by M. B. Eisen
- *D. melanogaster* annotations
  - Gene location
  - Gene model
- 6 X-linked gene fragments
  - Evenly spaced
  - 700 bp (coding + noncoding)

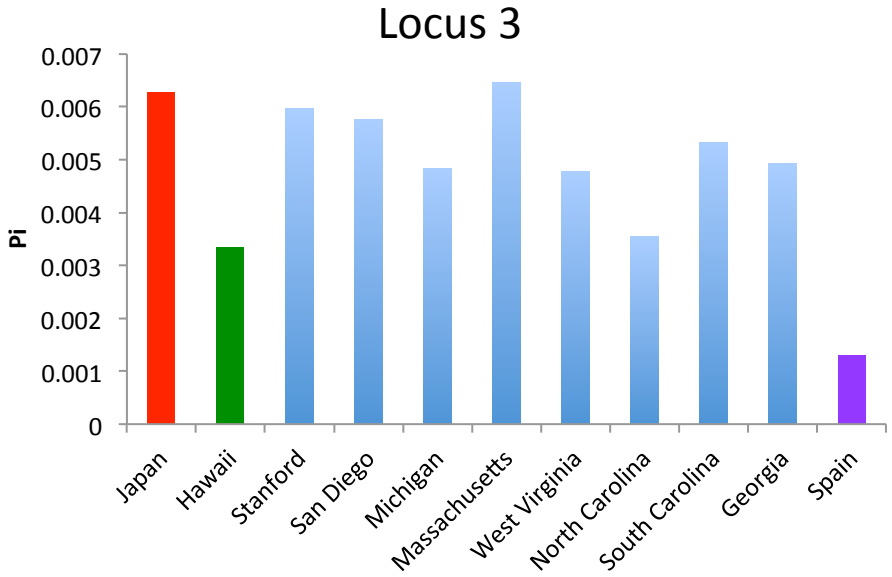
Methods

- Single male DNA extraction
- PCR
- Sanger sequencing

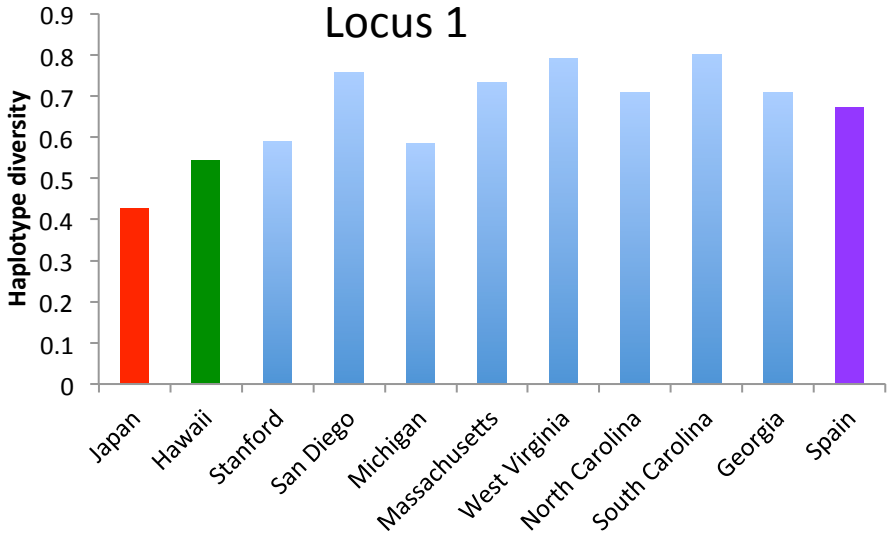
Results: Nucleotide diversity

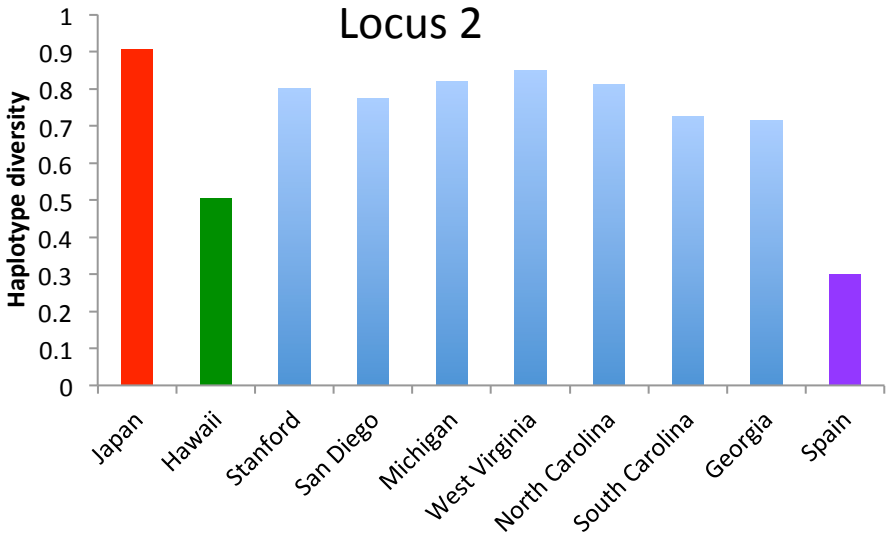


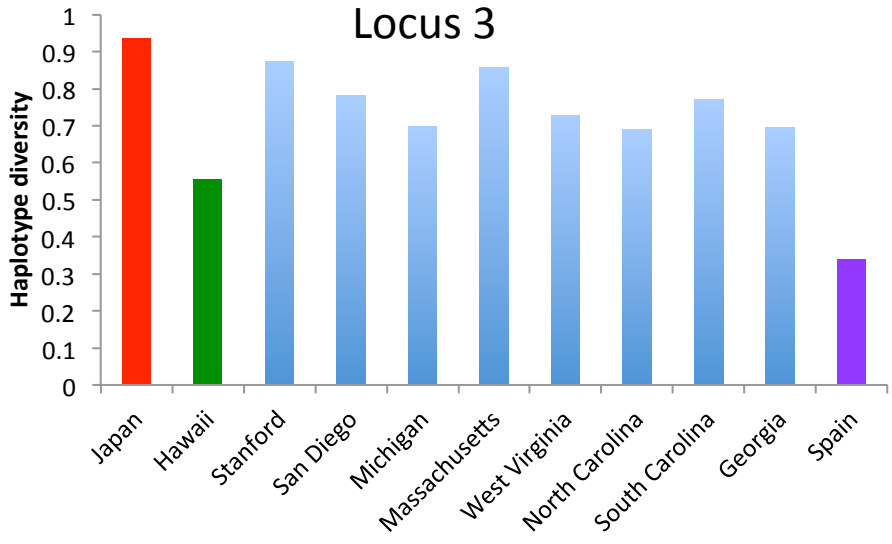




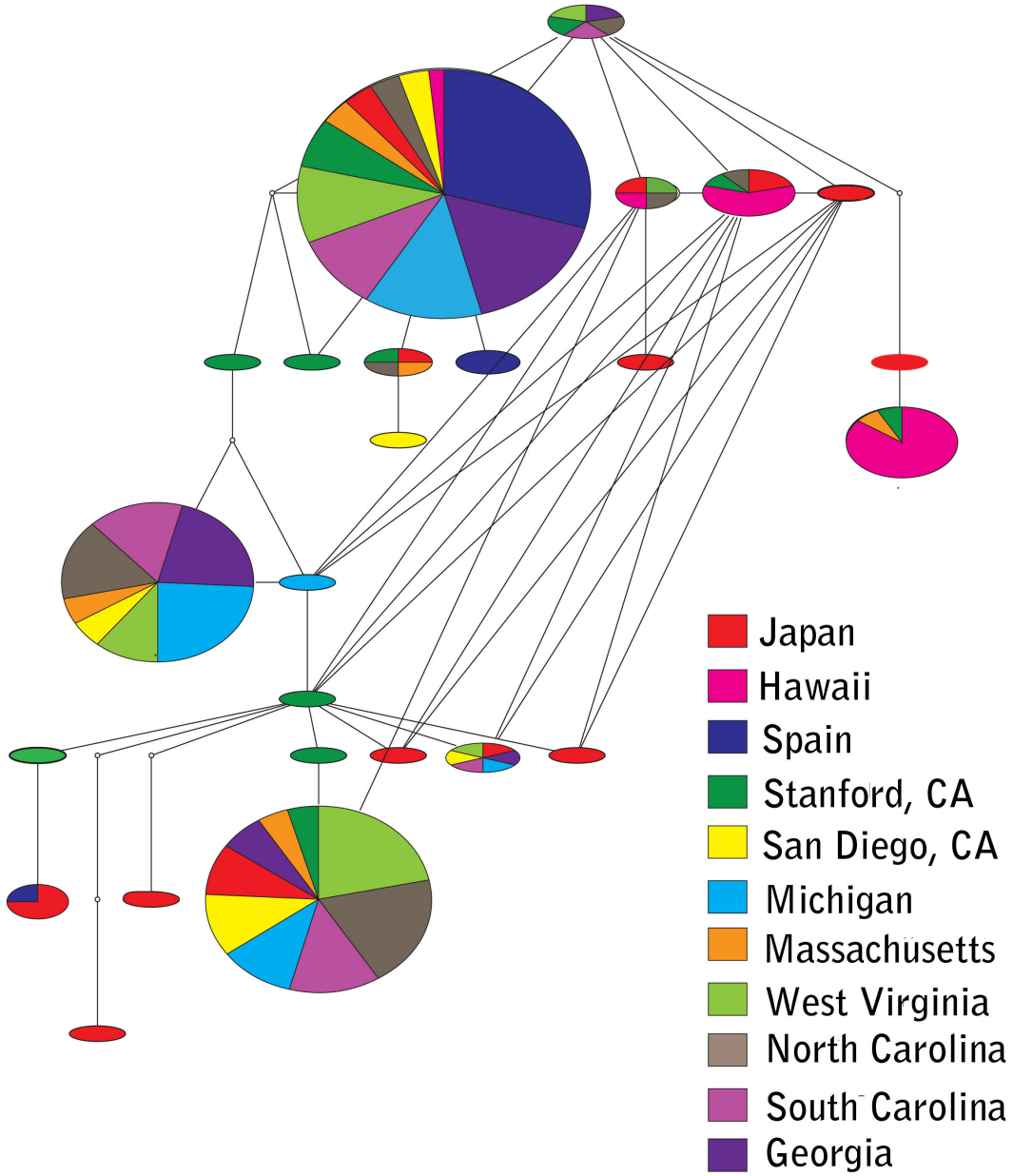
Results: Haplotype diversity



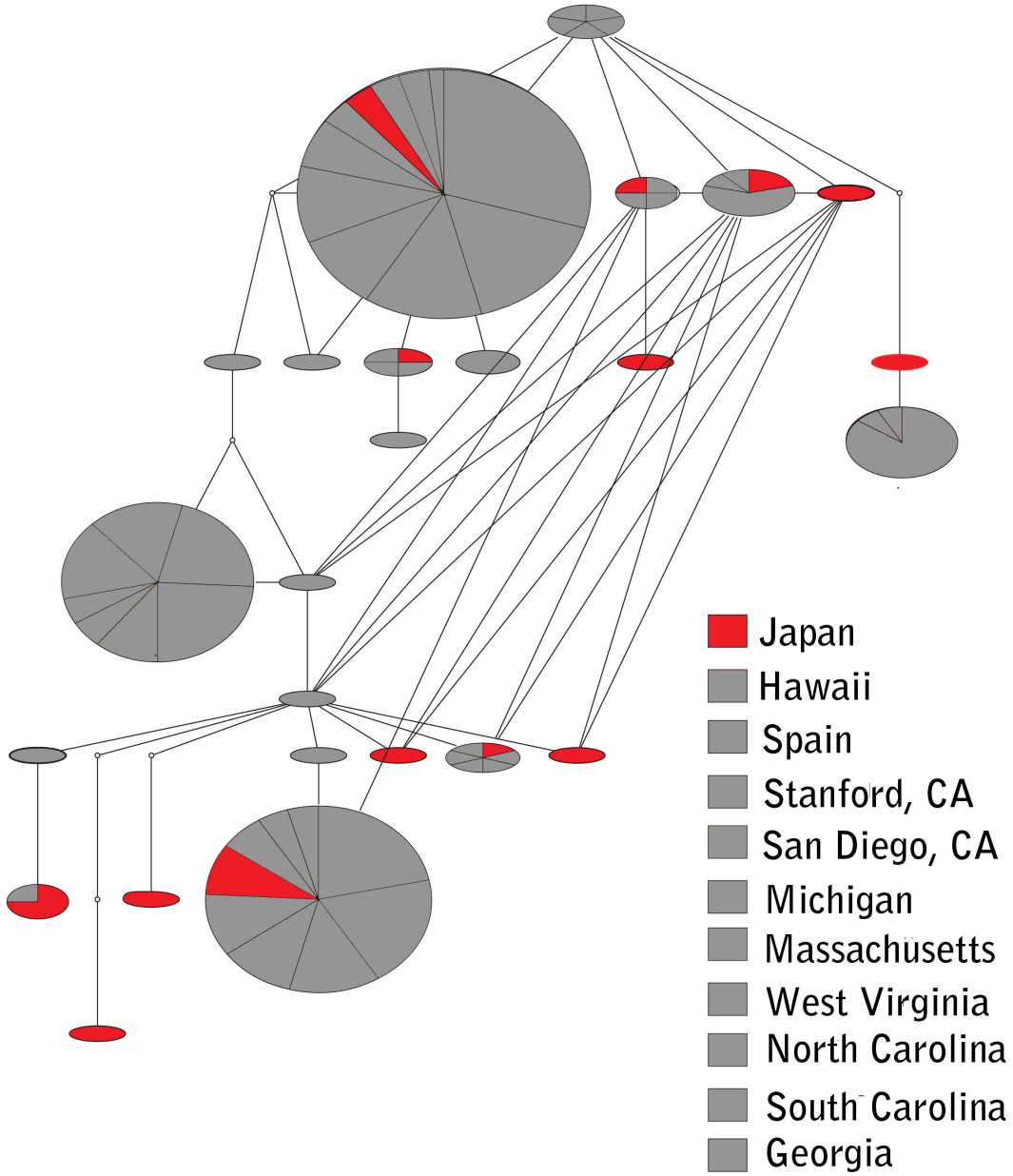




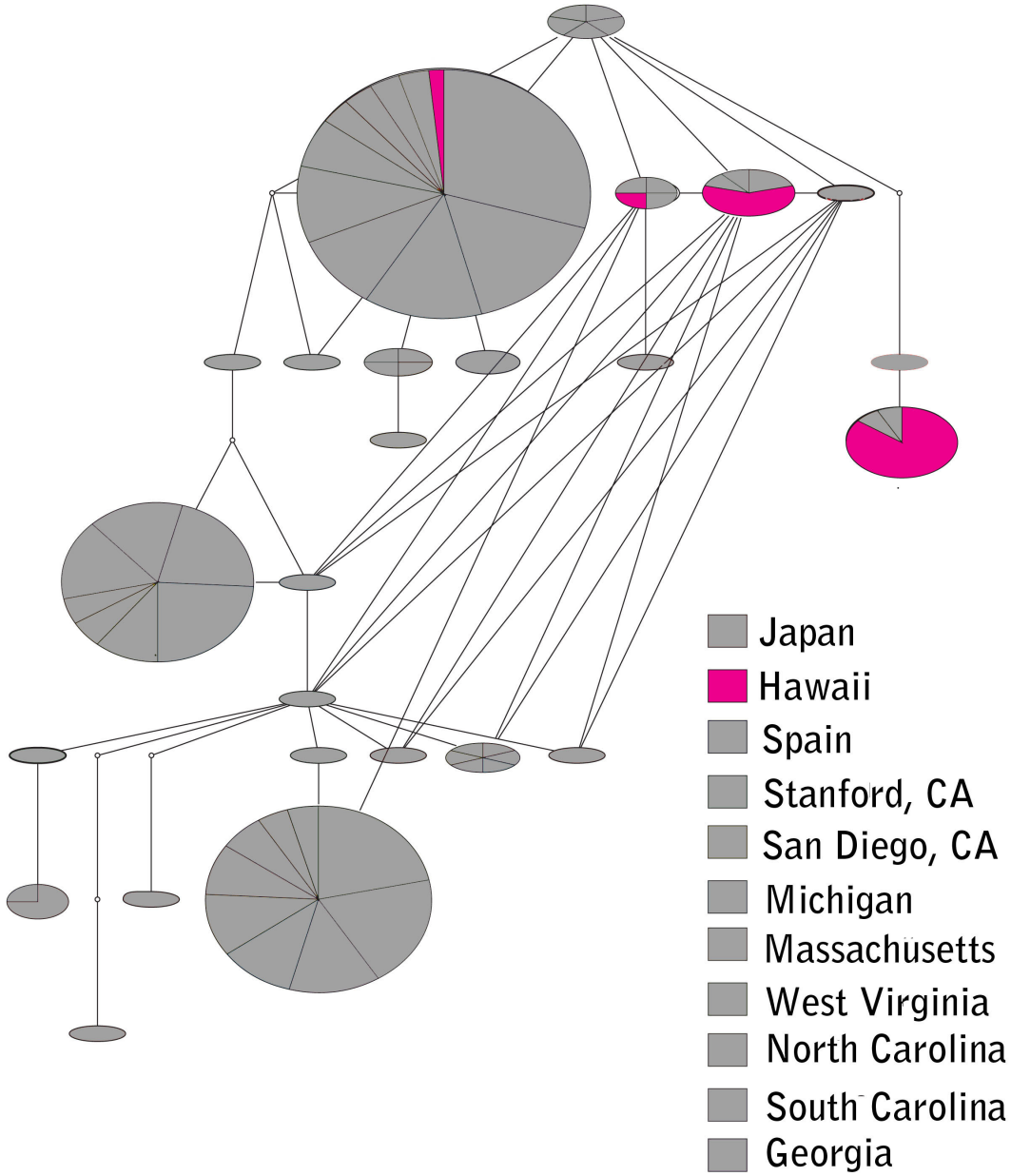
Results: Haplotype network



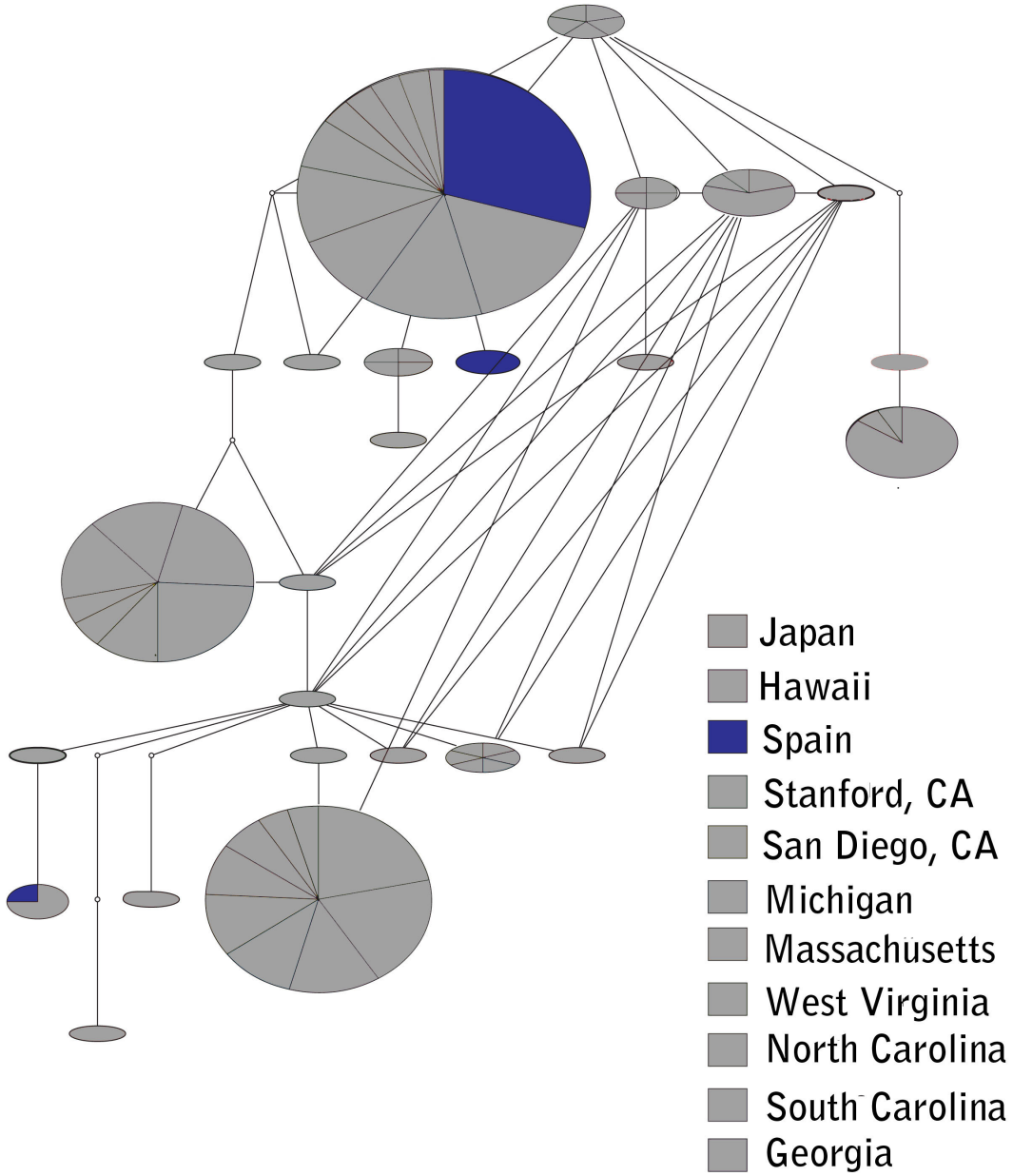
Results: Haplotype network



Results: Haplotype network



Results: Haplotype network



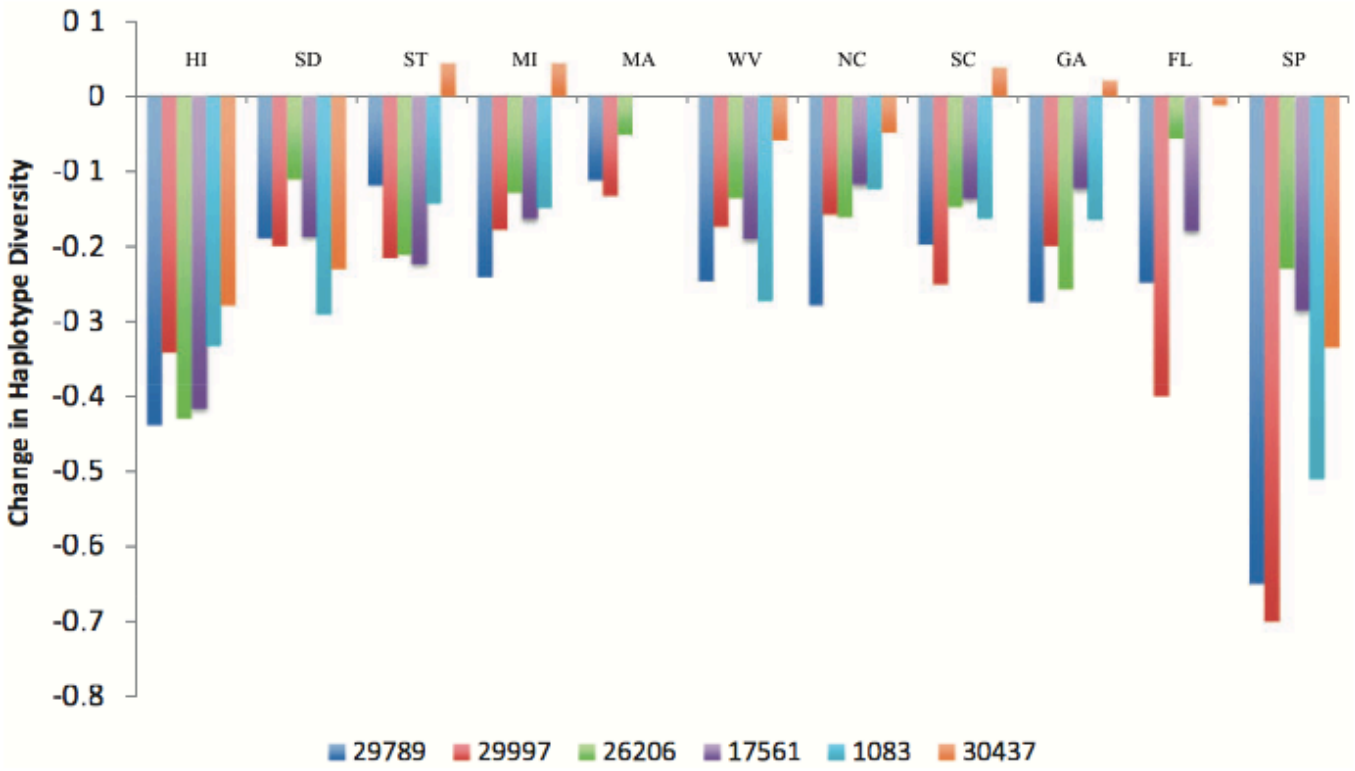


FIG. 2. Change in haplotype diversity at each locus in each population relative to Japan as estimated by the following equation:

$$\frac{Hd_{\text{sample}} - Hd_{\text{Japan}}}{Hd_{\text{Japan}}}.$$

Table 2. Fst Based on Sites for Which at Least Two Individuals Were Sampled Per Population.

Population	JP	HI	ST	SD	FL	GA	MA	MI	NC	SC	WV
HI	0.23*	—									
ST	0.112*	0.042	—								
SD	0.113*	0.118*	0.011	—							
FL	0.032	0.199*	−0.097	−0.352	—						
GA	0.076*	0.195*	0.058	0.037	−0.025	—					
MA	0.027	0.177*	0.05	−0.012	−0.151	0.017	—				
MI	0.036	0.177*	0.075*	0.047	−0.129	−0.028	−0.019	—			
NC	0.061*	0.265*	0.13*	0.061	0.008	0.02	0.042	−0.022	—		
SC	0.111*	0.219*	0.14*	0.02	−0.313	0.061	−0.073	0.052	0.097*	—	
WV	0.076*	0.242*	0.128*	0.021	−0.153	0.024	0.001	0.024	0.006	0.032	—
SP	0.287*	0.491*	0.379*	0.472*	0.535*	0.319*	0.42*	0.285*	0.395*	0.383*	0.363*

NOTE.—Asterisks denote values significant at $P < 0.05$ (permutation test, see Materials and Methods).

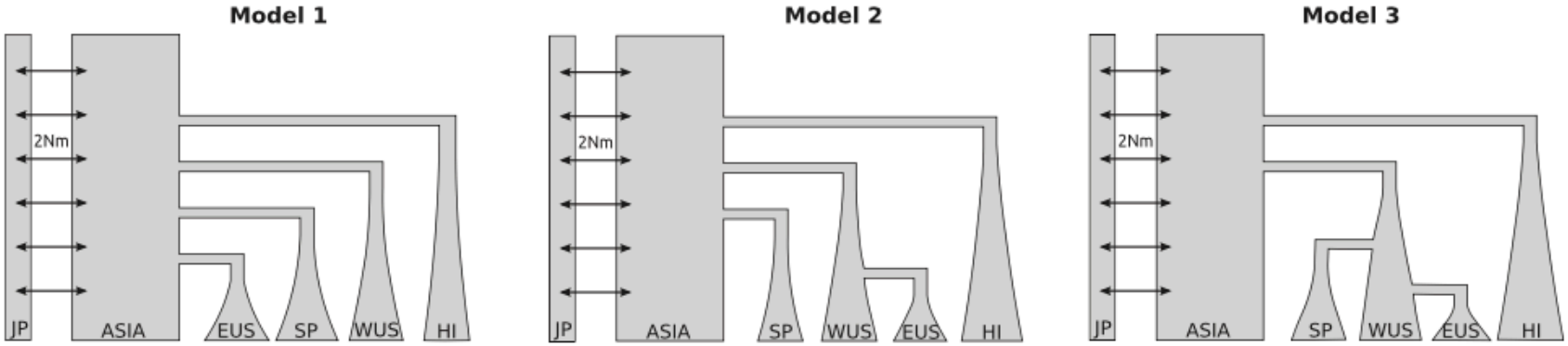


FIG. 4. Invasion models for *Drosophila suzukii*. We denote as ASIA the unsampled source population of the invasions and JP, HI, WUS, EUS, SP the Japanese, Hawaii, Western United States, Eastern United States, and Spanish populations for which we have samples. For Model 1, we assumed independent colonization of all continents from Japan. For Model 2, we assumed that EUS was colonized from WUS. For Model 3, we assumed that both the SP and EUS populations were colonized from WUS. The arrows between ASIA and JP denote migration between those populations at a rate equal to 2 Nm.

Table 3. Model Choice Results.

Colonization History Models	Posterior Probability	Observed *P* Value	Tukey Depth	Tukey *P* Value
1	*0.4181*	*0.999*	*0.227*	*0.999*
2	0.2770	0.960	0.098	0.958
3	0.3049	0.943	0.105	0.937

NOTE.—Reported are the posterior probabilities for models of the colonization history of *Drosophila suzukii*. The model with the highest probability is shown in italic. The *P* value for the observed data and the Tukey depth and the *P* value for a Tukey test are reported.

Table 4. Priors and Weighted Posterior Estimates for Parameters of the Three Models of Colonization for *Drosophila suzukii*.

Parameter	Population (*i*)	Prior	Mode	Mean	Median	Q5%	Q95%
$\text{Log}_{10}(N_i)$	Asia	$U[4, 8]$[a]	6.27	6.26	6.26	5.94	6.56
	Japan (JP)	$U[2, 6]$[a]	5.64	5.31	5.40	3.46	6.88
	Hawaii (HI)		5.18	4.32	4.37	2.78	5.77
	Western United States (WUS)		5.37	4.66	4.77	3.12	5.84
	Eastern United States		3.66	4.25	4.24	2.76	5.74
	Spain		3.03	4.04	3.98	2.68	5.61
$\text{Log}_{10}(f_i)$	Hawaii	$U[0.6, 3]$[a]	2.30	2.10	2.14	1.23	2.84
	Western United States		2.52	2.14	2.25	1.00	2.89
	Eastern United States (EUS)		1.81	1.79	1.79	0.95	2.63
	Spain (SP)		0.87	1.03	0.98	0.65	1.56
τ_i	Hawaii (HI)	$U[100, 750]$[a]	421	431	434	147	704
	Western United States (WUS)	$U[10, 100]$[b]	88	75	78	42	97
	Eastern United States (EUS)	$\tau_{HI} > \tau_{WUS} > \tau_{SP} > \tau_{EUS}$	21	32	29	12	62
	Spain (SP)		60	58	58	24	90
$\text{Log}_{10}(\text{Nm})$	—	$U[-2, 2]$[a]	−0.27	−0.21	−0.23	−1.05	0.69
$\mu \times 10^9$	—	$N(3.46, 0.28)$[c]	3.49	3.49	3.49	2.46	4.52

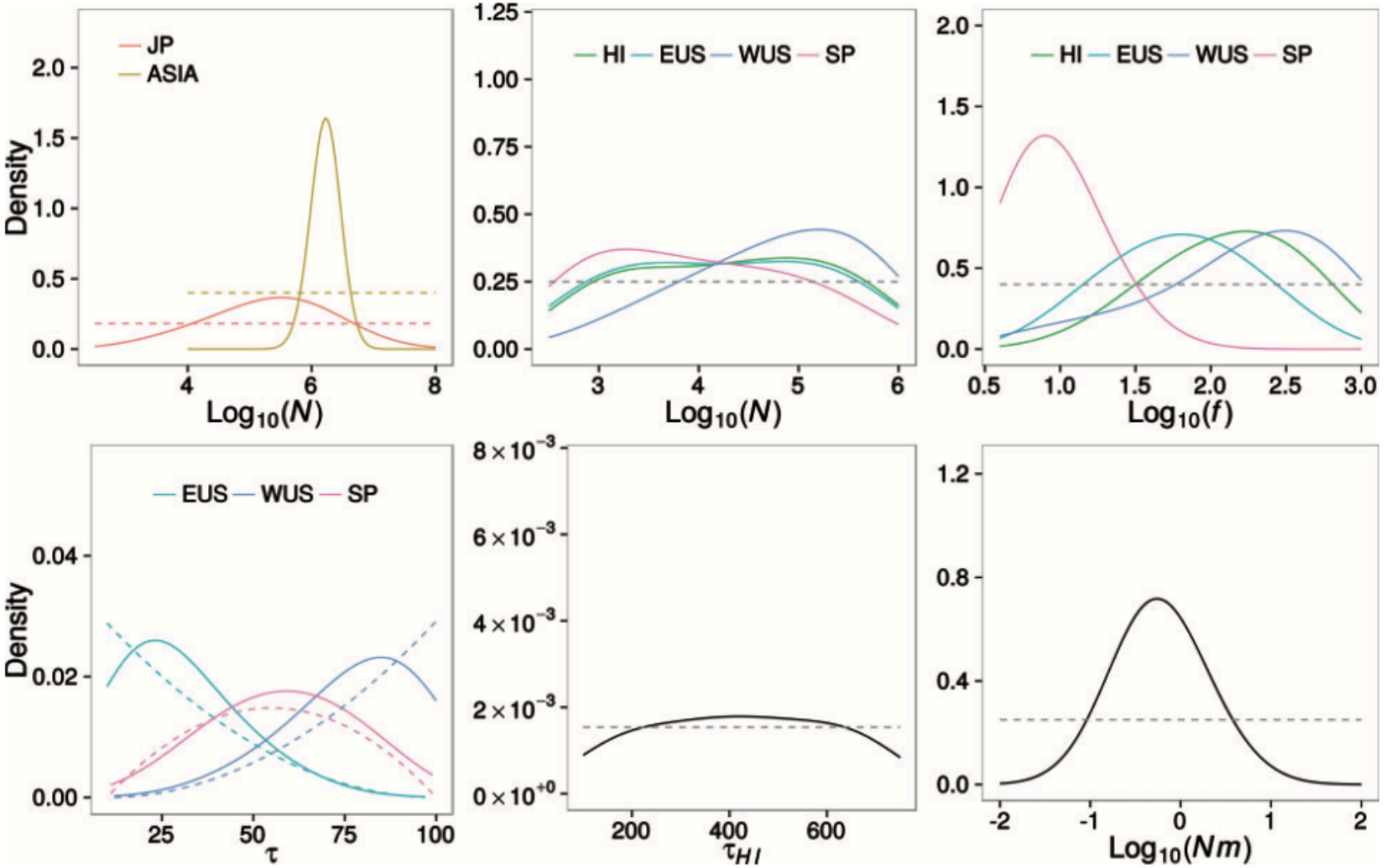


FIG. 6. The posterior probability of the parameters of the three colonization models for *D. suzukii* weighted by the posterior probability of each model, where N is the current population size, f is the number of founding individuals, τ is the colonization time for each population (in generations), and Nm is the migration rate among demes in the structured Japan population model. Note that the posterior distribution of τ_{HI} was plotted separately from the remaining τ estimates due to its unique prior range.

Summary: *D. suzukii*

- Interested in US colonization history
- Sampled worldwide populations
- Used X-linked loci to probe population structure
- Preliminary data inconsistent with one US invasion
- No clear founding population for US invasion
- Europe may reflect single invasion