**Supporting Information**

# Global invasion history of the agricultural pest butterfly *Pieris rapae* revealed with genomics and citizen science

**Authors:**
Sean F. Ryan[1,2], Eric Lombaert[3], Anne Espeset[4], Roger Vila[5], Gerard Talavera[5,6], Vlad Dincă[7], Meredith M. Doellman[8], Mark A. Renshaw[9], Matthew W. Eng[8], Emily A. Hornett[10], Yiyuan Li[8], Michael E. Pfrender[8,11], DeWayne Shoemaker[1]

**Affiliations:**
[1]Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, TN 37996 USA
[2]Ecological and Biological Sciences Practice, Exponent, Inc., Menlo Park, CA 94025
[3]Institut National de la Recherche Agronomique, Université Côte d'Azur, Centre de Recherches de Sophia-Antipolis, Institut Sophia Agrobiotech, 400 Route des Chappes, BP 167, 06 903 Sophia Antipolis, France
[4]Department of Biology, University of Nevada, Reno, NV 89557 USA
[5]Department of Animal Biodiversity and Evolution, Institut de Biologia Evolutiva (Consejo Superior de Investigaciones Científicas and Universitat Pompeu Fabra), Barcelona, 08003 Spain
[6]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA
[7]Department of Ecology and Genetics, PO Box 3000, 90014 University of Oulu, Finland
[8]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556
[9]Shrimp Department, Oceanic Institute of Hawai'i Pacific University, Waimanalo, HI 96795
[10]Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool L69 3BX, United Kingdom
[11] Environmental Change Initiative, University of Notre Dame, South Bend, IN 46556

**Corresponding author:**
Sean F. Ryan,
Email: citscisean@gmail.com



www.pnas.org/cgi/doi/10.1073/pnas.1907492116

## Supplementary Tables

**Table S1.** Contribution of specimens made by citizen scientists.

| Country | Inclusive definition of citizen science | | Restrictive definition of citizen science | | Total specimens collected |
|---|---|---|---|---|---|
| | Specimens collected by citizen scientists | | Specimens collected by citizen scientists | | |
| | # | % | # | % | |
| Algeria | 12 | 100% | 0 | 0% | 12 |
| Australia | 76 | 100% | 62 | 82% | 76 |
| Austria | 3 | 100% | 0 | 0% | 3 |
| Bulgaria | 2 | 22% | 2 | 22% | 9 |
| Canada | 24 | 100% | 11 | 46% | 24 |
| China | 0 | 0% | 0 | 0% | 22 |
| Czech Republic | 43 | 100% | 43 | 100% | 43 |
| England | 3 | 12% | 0 | 0% | 26 |
| Estonia | 0 | 0% | 0 | 0% | 4 |
| Finland | 0 | 0% | 0 | 0% | 7 |
| France | 0 | 0% | 0 | 0% | 12 |
| Georgia | 23 | 100% | 0 | 0% | 23 |
| Gibraltar | 4 | 100% | 4 | 100% | 4 |
| Greece | 2 | 17% | 0 | 0% | 12 |
| Italy | 12 | 86% | 0 | 0% | 14 |
| Japan | 8 | 73% | 0 | 0% | 11 |
| Malta | 10 | 100% | 0 | 0% | 10 |
| Mexico | 0 | 0% | 0 | 0% | 6 |
| Morocco | 2 | 17% | 0 | 0% | 12 |
| New Zealand | 25 | 100% | 17 | 68% | 25 |
| Poland | 12 | 100% | 0 | 0% | 12 |
| Portugal | 2 | 100% | 2 | 100% | 2 |
| Romania | 3 | 25% | 4 | 33% | 12 |
| Russia | 18 | 100% | 16 | 89% | 18 |
| South Korea | 11 | 100% | 11 | 100% | 11 |
| Spain | 21 | 78% | 13 | 48% | 27 |
| Sweden | 0 | 0% | 0 | 0% | 2 |
| Taiwan | 24 | 100% | 0 | 0% | 24 |
| Tunisia | 9 | 75% | 0 | 0% | 12 |
| Turkey | 10 | 100% | 10 | 100% | 10 |
| Ukraine | 2 | 100% | 0 | 0% | 2 |
| USA | 194 | 62% | 150 | 48% | 313 |

**Table S2.** Description of the competing scenarios and results of the six successive ABC analyses to infer the invasion history of *Pieris rapae*.

| Step | Scenario | Prior error rate | | | Random forest votes | | | Posterior probability | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Data 1 | Data 2 | Data 3 | Data 1 | Data 2 | Data 3 | Data 1 | Data 2 | Data 3 |
| | *Analysis 1 – Europe and Asia (west/east) - 18 summary statistics; 13,974 SNPs* | 13.82% | 14.49% | 14.29% | | | | | | |
| | S1: Asia is the source of Europe | | | | 57 | 188 | 207 | - | - | - |
| | S2: Europe is the source of Asia | | | | 132 | 132 | 43 | - | - | - |
| | **S3: Asia and Europe derived from an ancestral population** | | | | **811** | **680** | **750** | **0.8479** | **0.8173** | **0.8353** |
| | *Analysis 2 - Siberia and North Africa - 115 summary statistics; 15,533 SNPs* | 16.26% | 17.22% | 17.07% | | | | | | |
| | **S1: Asia and Europe are respectively the sources of Siberia and Africa** | | | | **602** | **676** | **521** | **0.7020** | **0.7549** | **0.6352** |
| | S2: Asia and Africa are respectively the sources of Siberia and Europe | | | | 162 | 180 | 146 | - | - | - |
| | S3: Africa and Siberia are respectively the sources of Europe and Asia | | | | 56 | 30 | 76 | - | - | - |
| | S4: Europe and Siberia are respectively the sources of Africa and Asia | | | | 180 | 114 | 257 | - | - | - |
| | *Analysis 3 - North America east (NAE) - 51 summary statistics; 16,753 SNPs* | 32.82% | 31.83% | 31.82% | | | | | | |
| | S1: Asia is the source of NAE, 1 introduction | | | | 0 | 9 | 2 | - | - | - |
| | **S2: Europe is the source of NAE, 1 introduction** | | | | **576** | **553** | **559** | **0.5010** | **0.6136** | **0.5064** |
| | S3: Asia is the source of NAE, 2 introductions | | | | 6 | 4 | 2 | - | - | - |
| | S4: Europe is the source of NAE, 2 introductions | | | | 418 | 434 | 437 | - | - | - |
| | *Analysis 4 – North America west (NAW) - 116 summary statistics; 17,049 SNPs* | 11.44% | 11.54% | 10.95% | | | | | | |
| | S1: Asia is the source of NAW | | | | 19 | 35 | 13 | - | - | - |
| | S2: Europe is the source of NAW | | | | 144 | 121 | 41 | - | - | - |
| | **S3: NAE is the source of NAW** | | | | **721** | **720** | **933** | **0.8518** | **0.9288** | **0.9524** |
| | S4: Europe is the source of NAW ~ 1600 CE | | | | 85 | 73 | 7 | - | - | - |
| | S5: Europe is the source of NAW ~ 1600 CE; NAW is the source of NAE | | | | 31 | 51 | 6 | - | - | - |
| | *Analysis 5 – New Zealand - 223 summary statistics; 17,100 SNPs* | 2.18% | 2.30% | 2.18% | | | | | | |
| | S1: Asia is the source of New Zealand | | | | 2 | 6 | 5 | - | - | - |
| | S2: Europe is the source of New Zealand | | | | 14 | 14 | 28 | - | - | - |
| | S3: NAE is the source of New Zealand | | | | 16 | 52 | 130 | - | - | - |
| | **S4: NAW is the source of New Zealand** | | | | **968** | **928** | **837** | **0.9739** | **0.9760** | **0.9802** |
| | *Analysis 6 - Australia - 388 summary statistics; 17,116 SNPs* | 14.94% | 15.00% | 14.81% | | | | | | |
| | **S1: New Zealand is the source of Australia** | | | | **631** | **733** | **613** | **0.7797** | **0.8420** | **0.8124** |
| | S2: NAW is the source of Australia | | | | 63 | 33 | 62 | - | - | - |
| | S3: New Zealand and Europe are the source of Australia (admixture) | | | | 15 | 10 | 18 | - | - | - |
| | S4: New Zealand and Asia are the source of Australia (admixture) | | | | 15 | 7 | 10 | - | - | - |
| | S5: New Zealand and NAW are the source of Australia (admixture) | | | | 276 | 217 | 297 | - | - | - |

Results are provided for all three datasets. For each ABC analysis a forest of 1,000 trees was grown. The lines in bold characters corresponds to the selected (most likely) scenarios.

**Table S3.** Prior and posterior distributions of all parameters and several composite parameters of the full final complete scenario (Fig 2d) performed with dataset 1.

| | Prior distributions | | | | Posterior distributions | | | |
|---|---|---|---|---|---|---|---|---|
| Parameters | Q 5% | median | mean | Q 95% | Q 5% | median | mean | Q 95% |
| Raw parameters | | | | | | | | |
| $N_1$ | 159 | 10,110 | 107,800 | 629,178 | 2,630 | 13,662 | 80,180 | 592,324 |
| $N_2$ | 158 | 9,872 | 108,500 | 636,286 | 8,217 | 52,803 | 177,076 | 710,940 |
| $N_3$ | 156 | 9,942 | 107,600 | 626,685 | 10,994 | 184,625 | 300,675 | 881,015 |
| $N_4$ | 159 | 10,010 | 108,200 | 630,976 | 2,305 | 88,066 | 245,186 | 914,933 |
| $N_5$ | 158 | 10,050 | 108,900 | 632,660 | 3,477 | 226,668 | 303,680 | 860,610 |
| $N_6$ | 160 | 10,350 | 108,100 | 628,855 | 1,863 | 119,285 | 294,245 | 884,571 |
| $N_7$ | 158 | 10,140 | 108,900 | 630,040 | 2,033 | 98,892 | 239,822 | 872,549 |
| $N_8$ | 160 | 9,863 | 108,100 | 629,561 | 1,038 | 39,466 | 187,263 | 794,756 |
| $N_A$ | 799 | 68,110 | 193,500 | 792,653 | 26,142 | 167,928 | 247,860 | 829,410 |
| $N_D$ | 126 | 1,469 | 23,160 | 123,273 | 269 | 14,173 | 22,626 | 75,263 |
| $NF_3$ | 3 | 20 | 43 | 159 | 13 | 90 | 98 | 191 |
| $NF_4$ | 3 | 20 | 43 | 159 | 11 | 52 | 63 | 152 |
| $NF_5$ | 3 | 20 | 43 | 159 | 10 | 68 | 77 | 164 |
| $NF_6$ | 3 | 20 | 43 | 158 | 11 | 88 | 91 | 179 |
| $NF_7$ | 3 | 20 | 43 | 159 | 13 | 87 | 89 | 179 |
| $NF_8$ | 3 | 20 | 43 | 159 | 6 | 64 | 71 | 172 |
| $DB_3$ | 2 | 15 | 16 | 29 | 1 | 5 | 6 | 15 |
| $DB_4$ | 2 | 15 | 15 | 29 | 5 | 15 | 16 | 28 |
| $DB_5$ | 2 | 16 | 16 | 29 | 2 | 15 | 15 | 29 |
| $DB_6$ | 2 | 15 | 15 | 29 | 1 | 7 | 9 | 21 |
| $DB_7$ | 2 | 16 | 16 | 29 | 2 | 9 | 11 | 26 |
| $DB_8$ | 2 | 16 | 16 | 29 | 3 | 15 | 15 | 29 |
| $t_1$ | 974 | 4,162 | 4,566 | 9,260 | 908 | 3,576 | 4,159 | 8,849 |
| $t_2$ | 973 | 4,165 | 4,562 | 9,265 | 850 | 3,011 | 3,656 | 8,510 |
| $t_3$ | 467 | 480 | 480 | 494 | 467 | 481 | 480 | 494 |
| $t_4$ | 398 | 411 | 411 | 425 | 397 | 408 | 409 | 424 |
| $t_5$ | 260 | 273 | 273 | 287 | 261 | 274 | 274 | 287 |
| $t_6$ | 235 | 249 | 249 | 262 | 235 | 248 | 248 | 262 |
| $t_7$ | 540 | 1,205 | 1,788 | 5,121 | 511 | 674 | 880 | 1,814 |
| $t_8$ | 539 | 1,200 | 1,783 | 5,119 | 529 | 859 | 1,057 | 2,312 |
| $t_a$ | 14,547 | 55,320 | 55,170 | 95,499 | 14,751 | 60,482 | 58,546 | 96,484 |
| Composite parameters | | | | | | | | |
| $BNsev_3$ | 42 | 6,208 | 180,900 | 900,381 | 203 | 1,062 | 11,425 | 51,050 |
| $BNsev_4$ | 41 | 6,155 | 181,600 | 907,241 | 2,513 | 57,450 | 153,718 | 635,627 |
| $BNsev_5$ | 41 | 6,186 | 181,700 | 913,085 | 618 | 23,343 | 99,199 | 459,562 |
| $BNsev_6$ | 42 | 6,290 | 182,500 | 924,448 | 151 | 12,860 | 40,460 | 134,639 |
| $BNsev_7$ | 41 | 6,053 | 183,900 | 930,965 | 539 | 6,793 | 34,863 | 167,428 |
| $BNsev_8$ | 41 | 6,162 | 179,500 | 886,500 | 309 | 5,580 | 13,628 | 46,122 |

Note: $BNsev_i$ = bottleneck severity of population $i$ computed as $[BD_i \times N_{\text{parental population of population } i}) / NF_i]$, with parental populations being populations 2, 3, 4, 5, 2 and 1 for populations 3, 4, 5, 6, 7 and 8 respectively.

**Table S4.** Prior distributions of demographic and historical parameters used in ABC analyses processed to retrace the worldwide invasion routes of *Pieris rapae*.

| Parameters | Distribution | Quantile 5% | Median | Mean | Quantile 95% |
|---|---|---|---|---|---|
| $N_D$ | Log-Uniform [100 – 1,000,000] | 126 | 1,482 | 23,610 | 128,053 |
| $N_A$ | Log-Uniform [100 – 1,000,000] | 774 | 67,560 | 192,800 | 790,045 |
| $N_j$, $N_i$, $N_{ia}$, $N_{ib}$ | Log-Uniform [100 – 1,000,000] | 159 | 10,280 | 108,600 | 629,089 |
| $NF_j$, $NF_{i,}$, $NF_{ia}$, $NF_{ib}$ | Log-Uniform [2 – 200] | 3 | 20 | 43 | 159 |
| $BD_j$, $BD_i$, $BD_{ia}$, $BD_{ib}$ | Uniform [1 – 30] | 2 | 16 | 15 | 29 |
| $ta$ | Uniform [10,000 – 100,000] | 14,547 | 54,930 | 54,990 | 95,453 |
| $t_j$ | Log-Uniform [500 – 10,000] | 585 | 2,265 | 3,197 | 8,617 |
| $t_i$, $t_{ia}$ | Uniform [$x_i$ – $x_i$+30] | DV | DV | DV | DV |
| $t_{mix}$, $t_{ib}$ | Uniform [165 – $x_i$+30] | DV | DV | DV | DV |
| $t_{4old}$ | Uniform [1245 – 1275] | 1,246 | 1,260 | 1,260 | 1,274 |
| $ar_i$ | Uniform [0.1 – 0.9] | 0.14 | 0.50 | 0.50 | 0.86 |

Notes: Index *i* stands for the number of the invasive population, i.e. 3, 4, 5 or 6 for North America (east), North America (west), New Zealand or Australia respectively. Index *j* stands for the number of the ancient putative native population, i.e. 1, 2, 7 or 8 for Asia (west/east), Europe, Africa or Siberia respectively. $N_D$ and $N_A$ = stable effective population size (number of diploid individuals) of the ancestral native population respectively before and after a demographic expansion event ($N_G < N_A$); $N_j$, $N_i$ = stable effective population size (number of diploid individuals) of the putative native and invasive populations; $NF_i$ = effective number of founders during a bottleneck lasting $BD_i$ generation(s) for population *i*; $ta$ = time of the demographic expansion in the ancestral native population; $t_j$ = merging time of the putative native populations into the ancestral one; $t_i$ = introduction time of invasive populations *i* with bounds $x_i$ fixed from dates of first observation of established population; $t_{4old}$ corresponds to the particular case of an old introduction hypothesis of the North American (west) population in ABC analysis 4; $N_{ia}$, $N_{ib}$, $NF_{ia}$, $NF_{ib}$, $BD_{ia}$ and $BD_{ib}$, $t_{ia}$, $t_{ib}$ and $t_{mix}$ are the parameters associated to an admixture event leading to the formation of invasive population *i*; $ar_i$ = admixture rate. Depending on the scenarios considered, various conditions were applied to times so that coalescent times fit with each scenario's topology. All times are expressed in number of generations assuming 3 generations per year, and running back in time from time 0 which corresponds to year 2015. All prior quantities presented were computed from $10^5$ values. DV = different values were possible. See Figure S3 for a graphical representation of the evolutionary scenarios with associated historical and demographic parameters considered in the ABC analyses.

**Table S5.** Summary statistics used in all DIYABC simulations (1).

| DIYABC abbreviation | Description |
|---|---|
| *Single sample statistics for each sampled population* | |
| HP0 | Proportion of loci with zero gene diversity |
| HM1 | Mean gene diversity across polymorphic loci (Nei, 1987) |
| HV1 | Variance of gene diversity across polymorphic loci |
| HMO | Mean gene diversity across all loci |
| *Two sample statistics for each pairwise sample combination* | |
| FP0 | Proportion of loci with zero F ST distance (Weir & Cockerham, 1984) |
| FM1 | Mean across loci of non-zero F ST distances |
| FV1 | Variance across loci of non-zero F ST distances |
| FMO | Mean across loci of F ST distances |
| NP0 | Proportion of loci with zero Nei's distance (Nei, 1972) |
| NM1 | Mean across loci of non-zero Nei's distances |
| NV1 | Variance across loci of non-zero Nei's distances |
| NMO | Mean across loci of Nei's distances |
| *Admixture statistics (Choisy et al., 2004) for each combination of parental and admixed populations* | |
| AP0 | Proportion of loci with zero admixture estimates |
| AM1 | Mean across loci of non-zero admixture estimate |
| AV1 | Variance across loci of non-zero admixture estimated |
| AMO | Mean across all locus admixture estimates |

**Supplementary Figures**

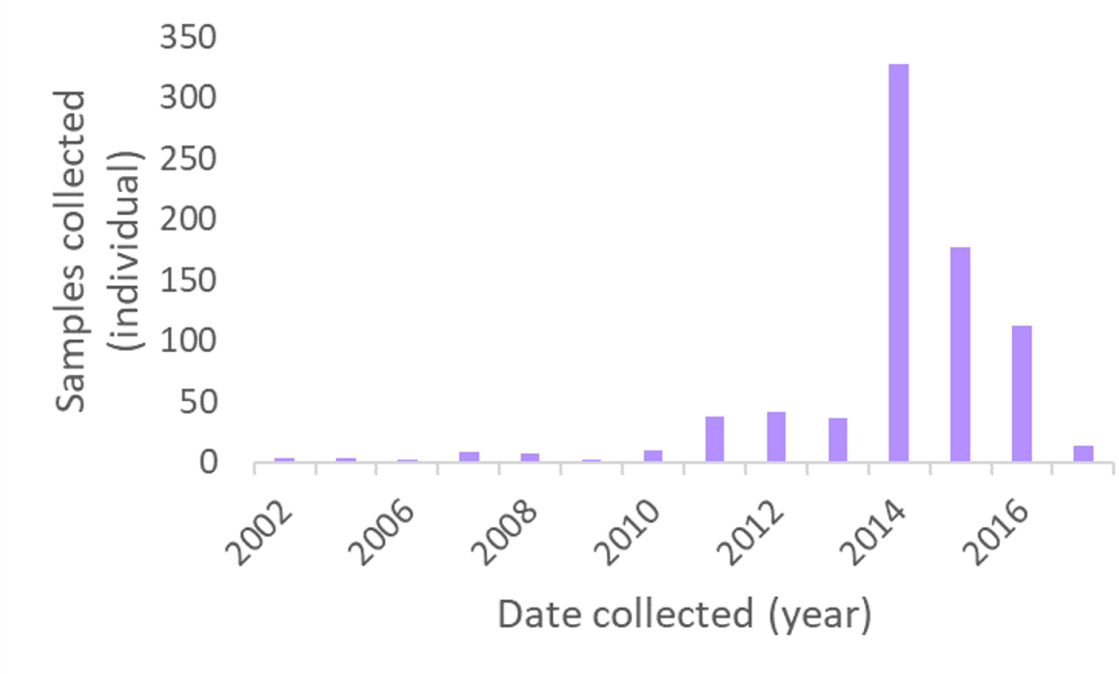



**Fig S1.** Sample sizes of *Pieris rapae* specimens by year collected.

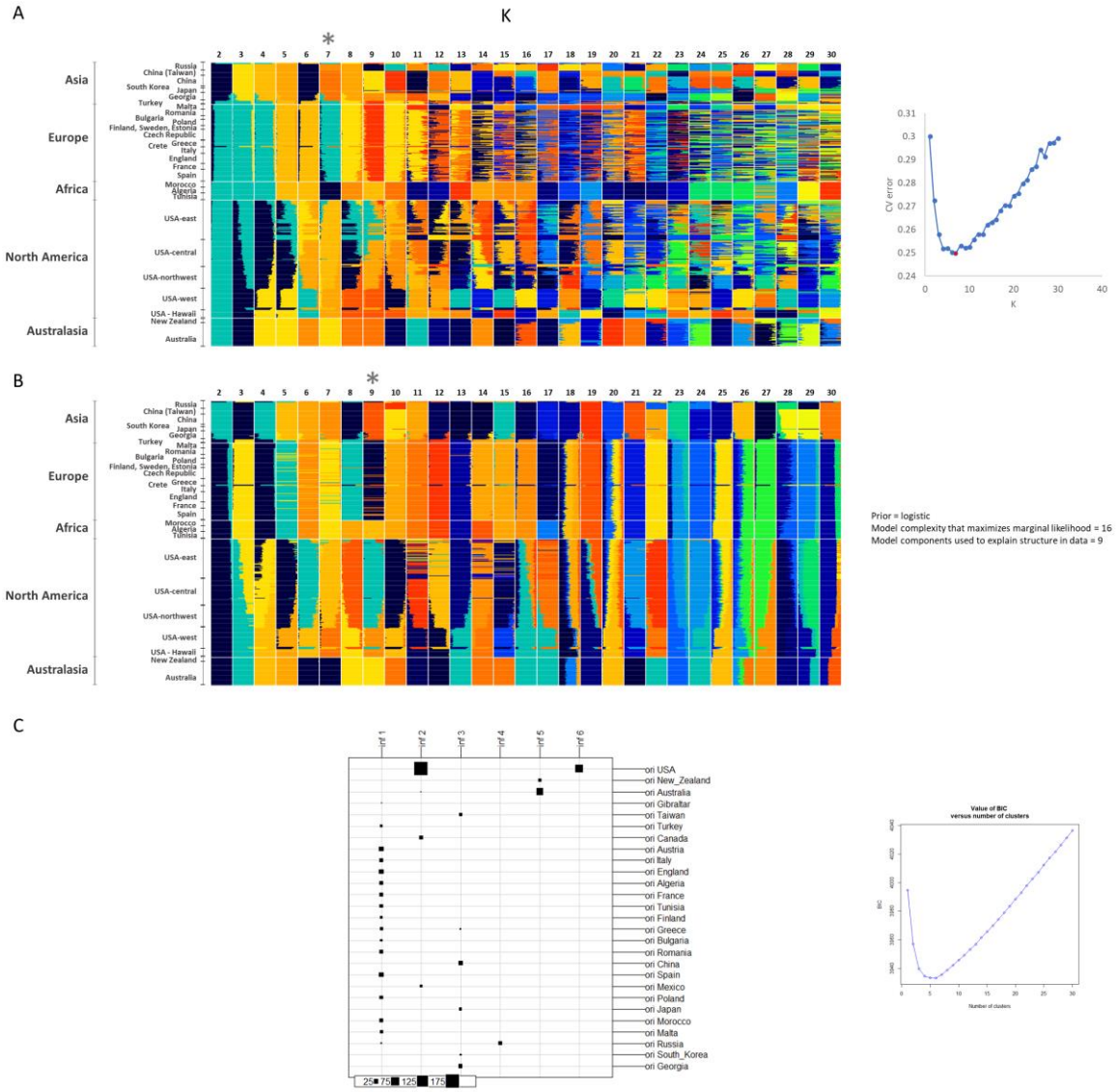
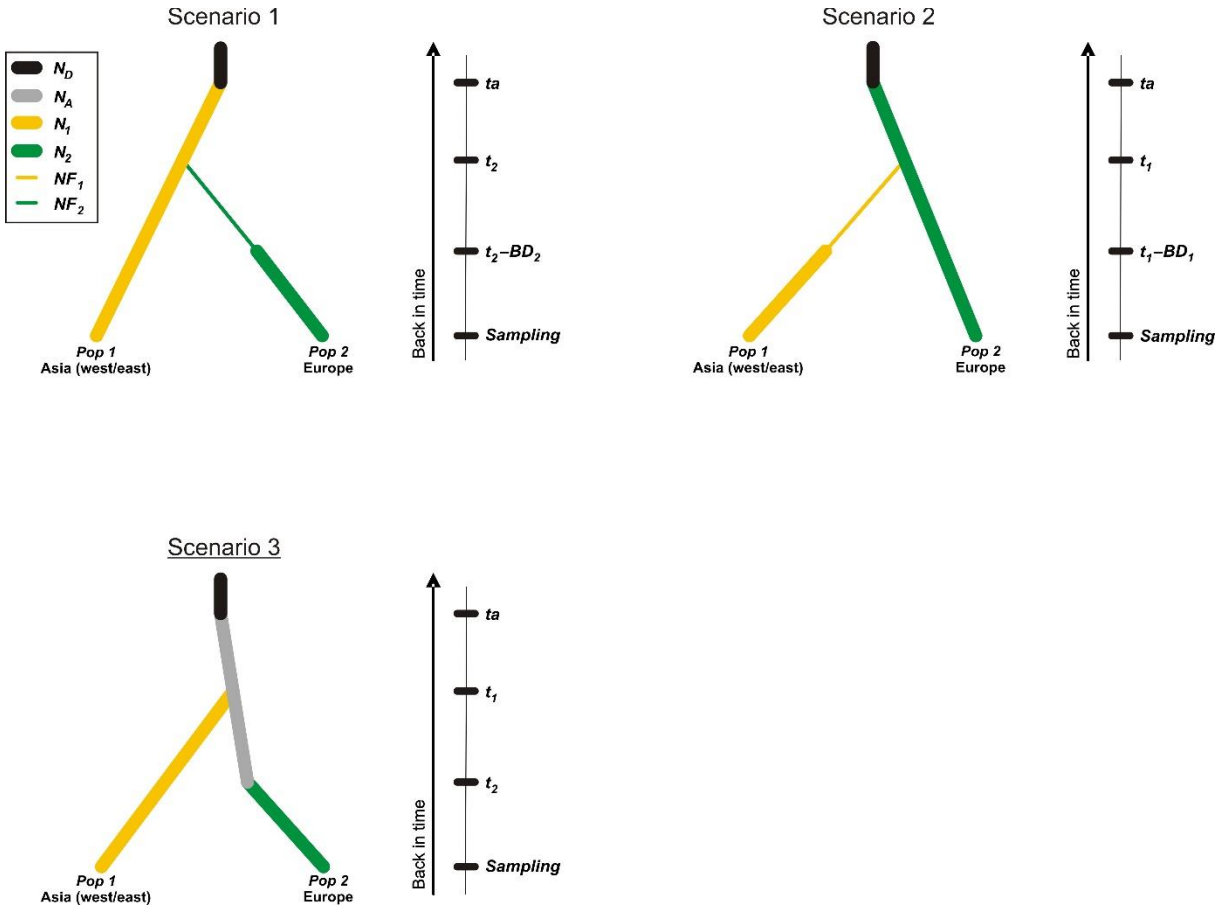
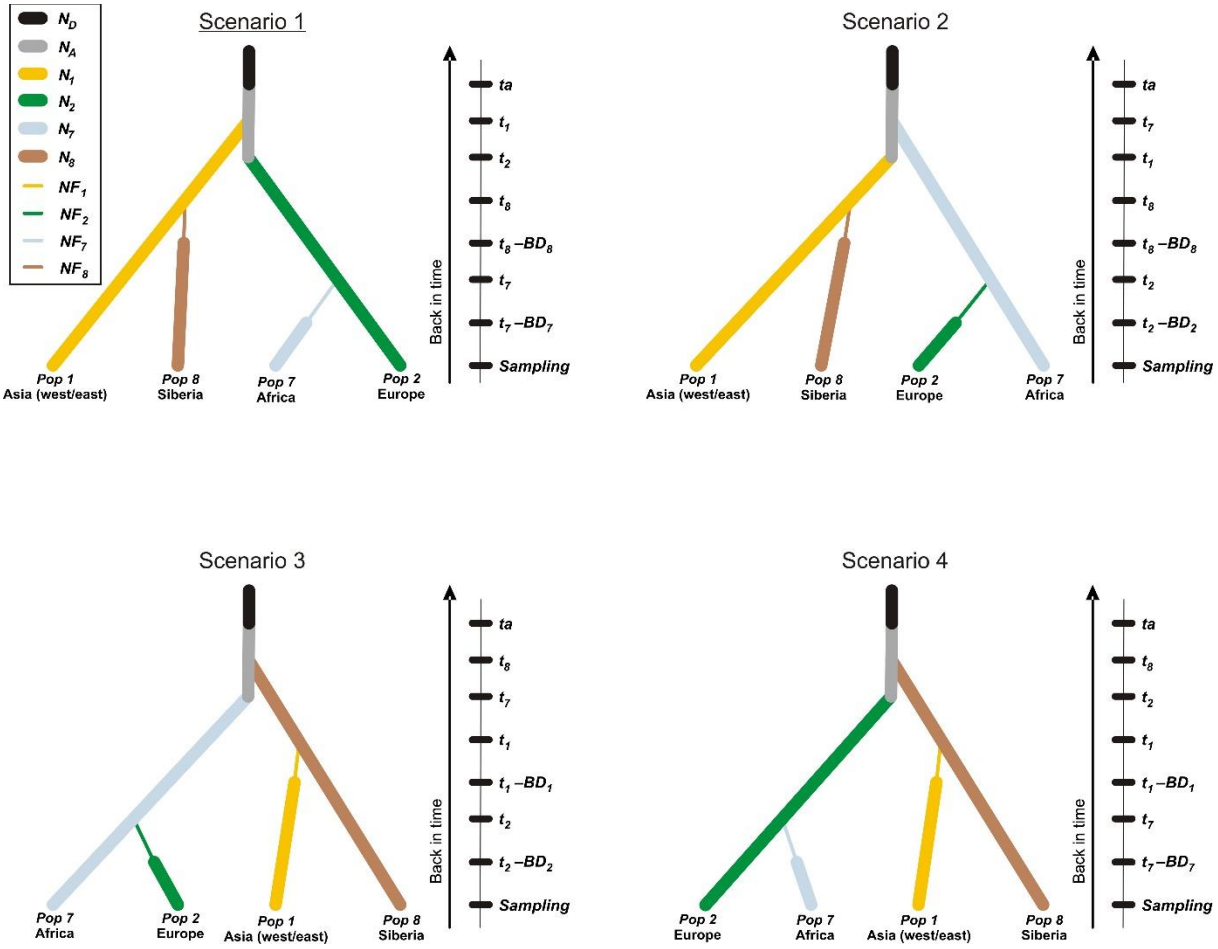



**Fig S2.** Population ancestry assignment plots for K:2-30, using **a**, ADMIXTURE, **b**, fastSTRUCTURE, and **c**, Discriminant Analysis of Principal Components (DAPC). For each analysis the evaluation for optimal K is included.

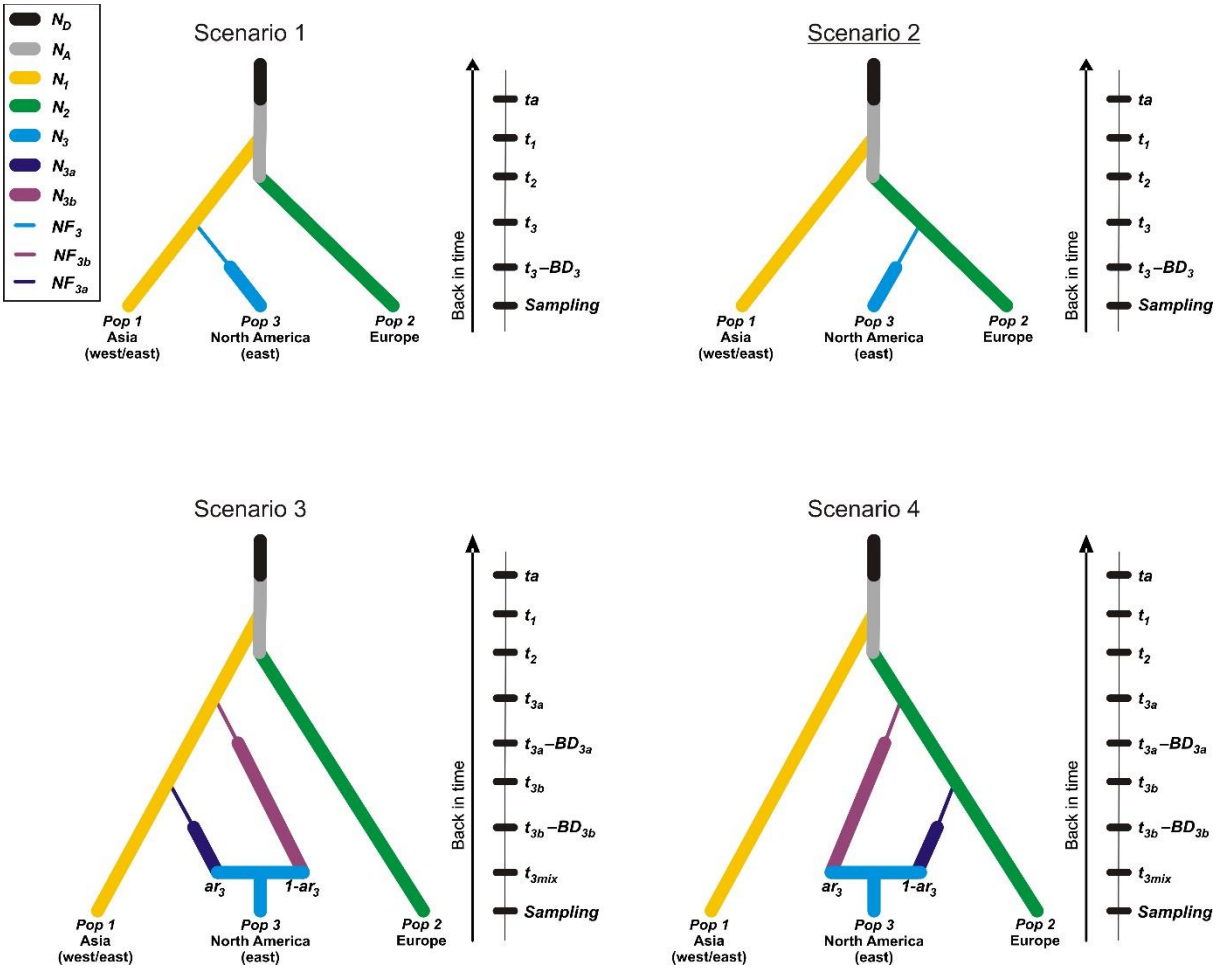
## A. Analysis 1 – Eurasia



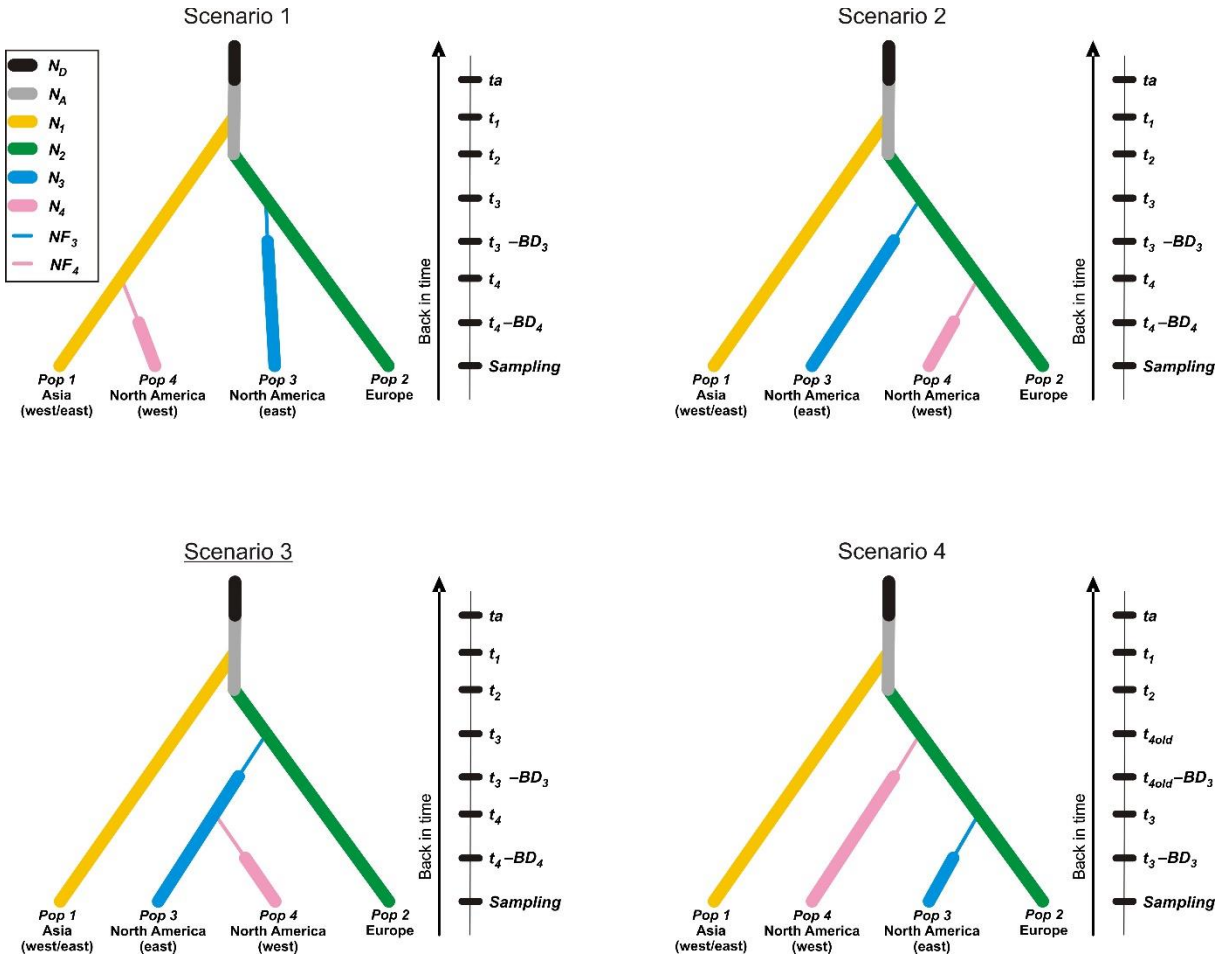
B. Analysis 2 – Siberia and North Africa

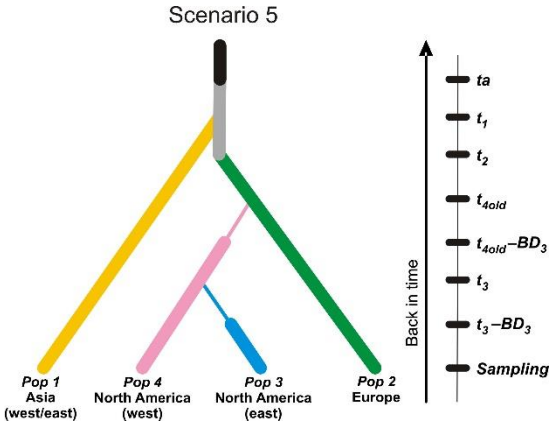


## C. Analysis 3 – North America (east)

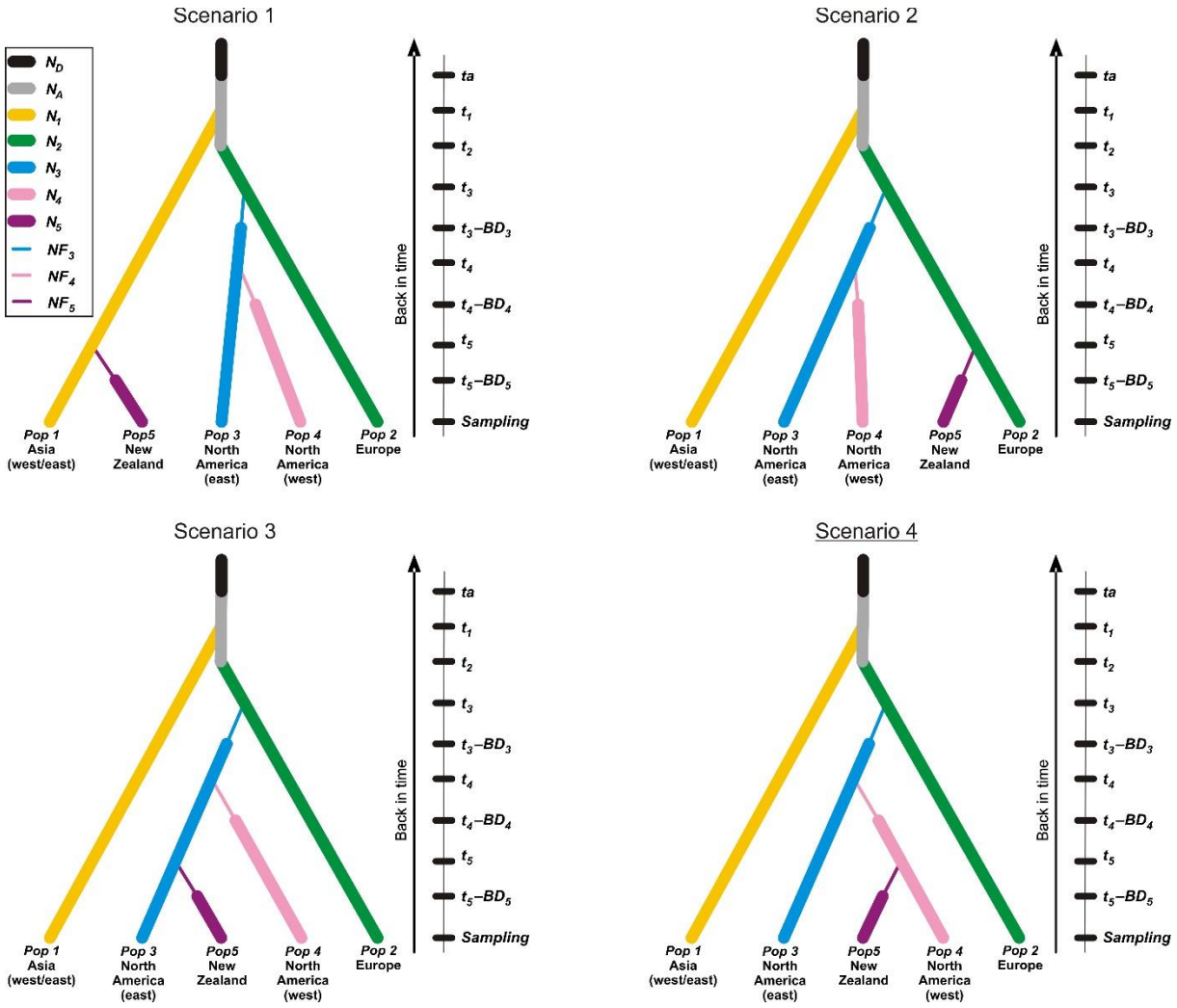


## D. Analysis 4 – North America (west)

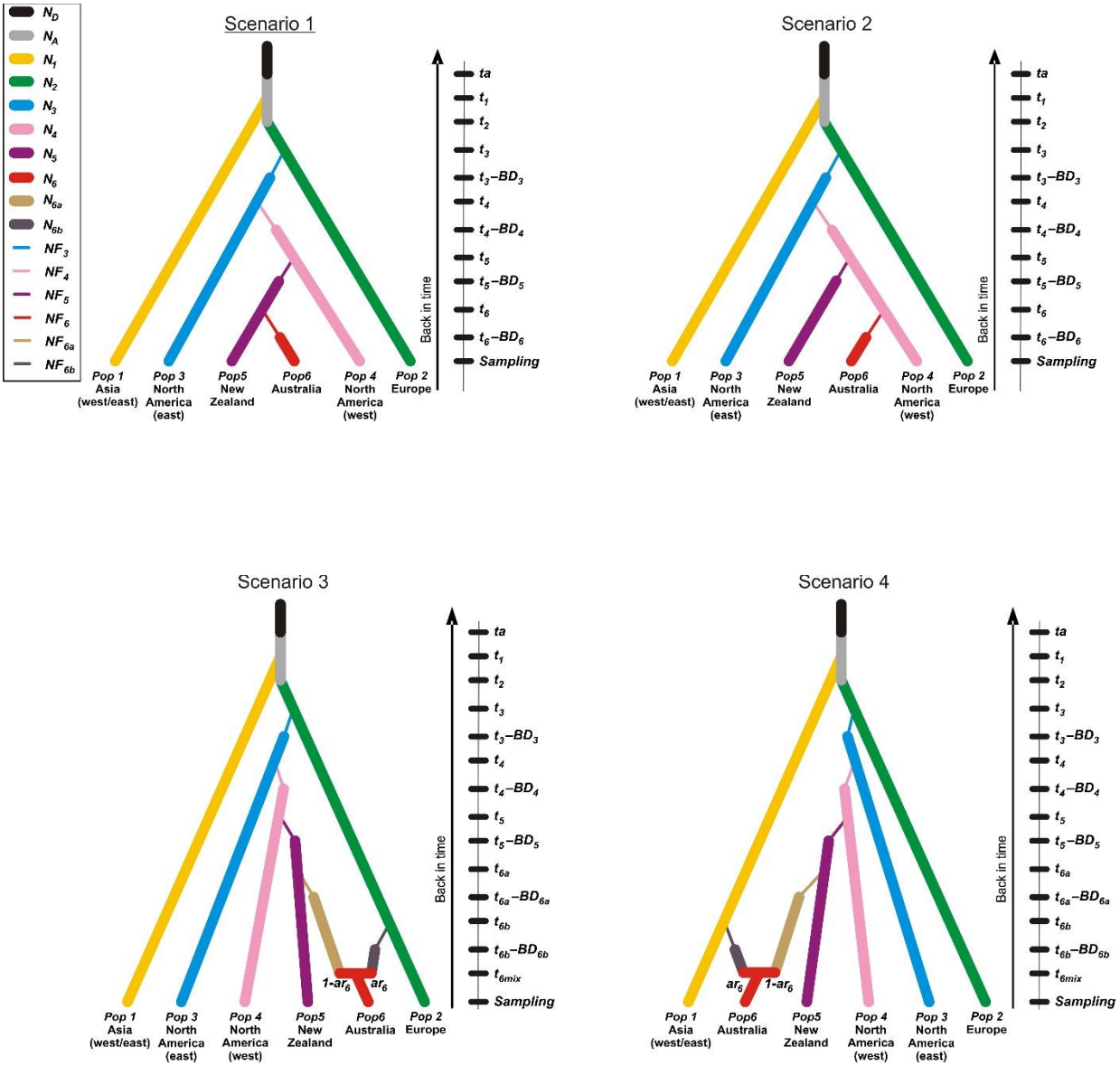



Scenario 5
ta
$t_1$
$t_2$
$t_{4old}$
$t_{4old}$–$BD_3$
$t_3$
$t_3$–$BD_3$
Sampling
Back in time
Pop 1
Asia
(west/east)
Pop 4
North America
(west)
Pop 3
North America
(east)
Pop 2
Europe

## E. Analysis 5 – New Zealand



## F. Analysis 6 – Australia



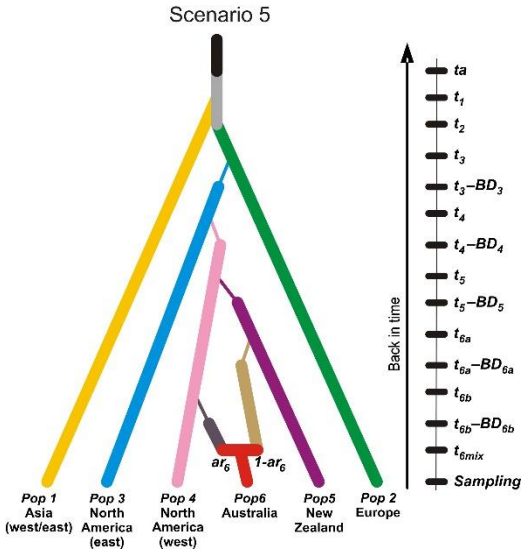



**Fig S3.** Schematic representation of each set of scenarios used in the ABC analyses to decipher the worldwide invasion routes of *Pieris rapae* (see also Table 1). Population numbers are as follows: 1 for Asia (west/east); 2 for Europe; 3 for North America (east); 4 for North America (west); 5 for New Zealand; 6 for Australia; 7 for North Africa; 8 for Siberia. For each analysis, the name of the most likely scenario is underlined. Thin lines indicate bottlenecks. For parameters descriptions and priors see Table S3. Time is not to scale.

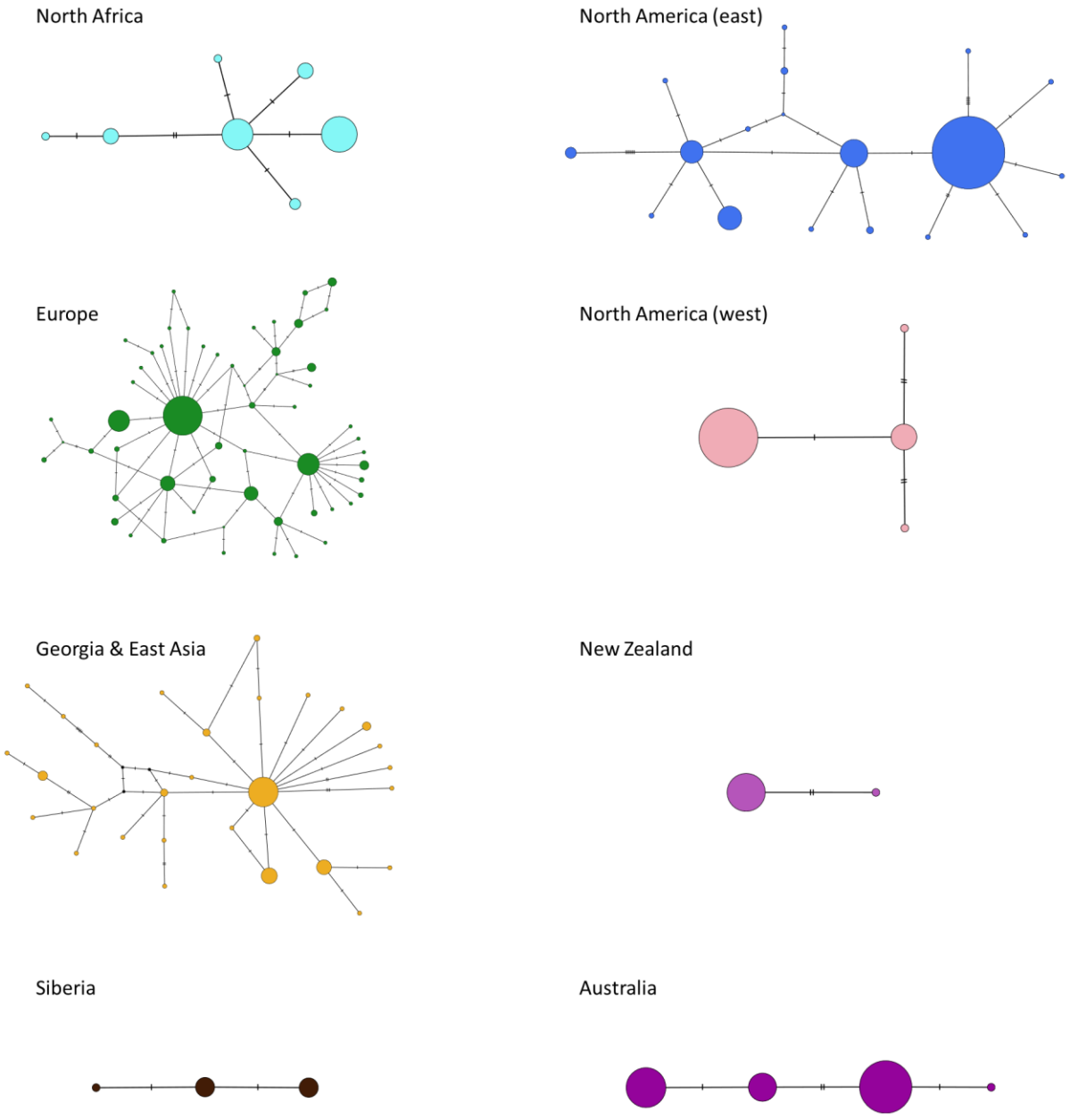



**Fig S4.** Median-joining haplotype networks for each population. Hash marks between haplotypes represent base changes (mutations).

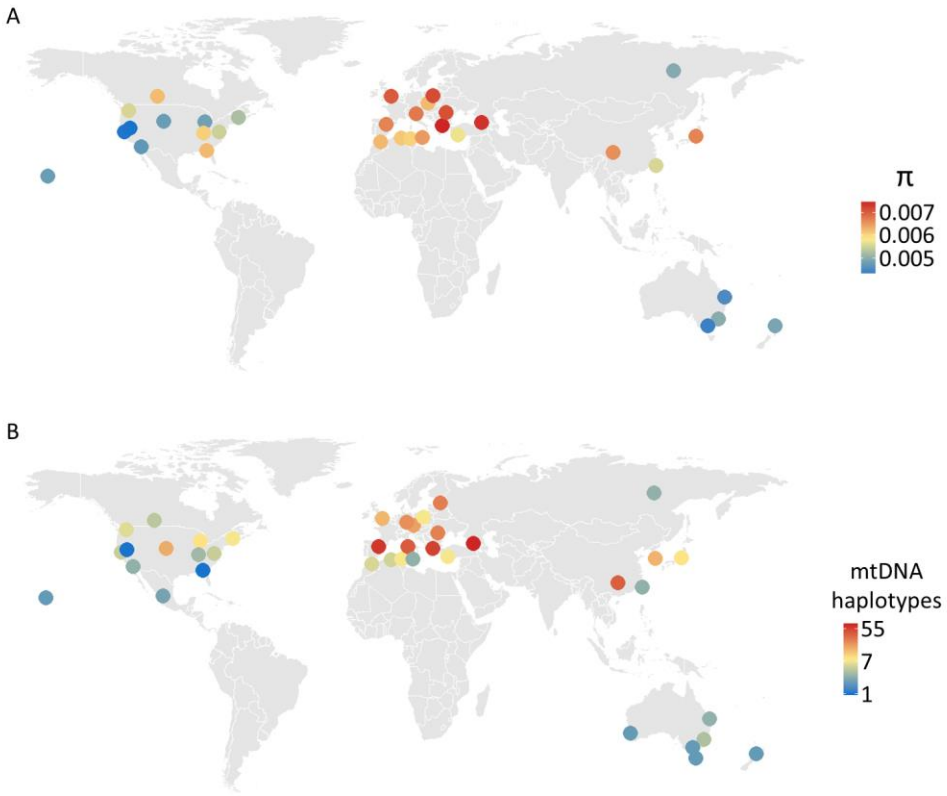



**Fig S5. Global patterns of genetic diversity. a,** Estimate of pairwise nucleotide diversity for each subpopulation based on autosomal ddRADseq data. **b**, mtDNA haplotype diversity estimated from rarefaction curves (note, colors are based on a log scale).

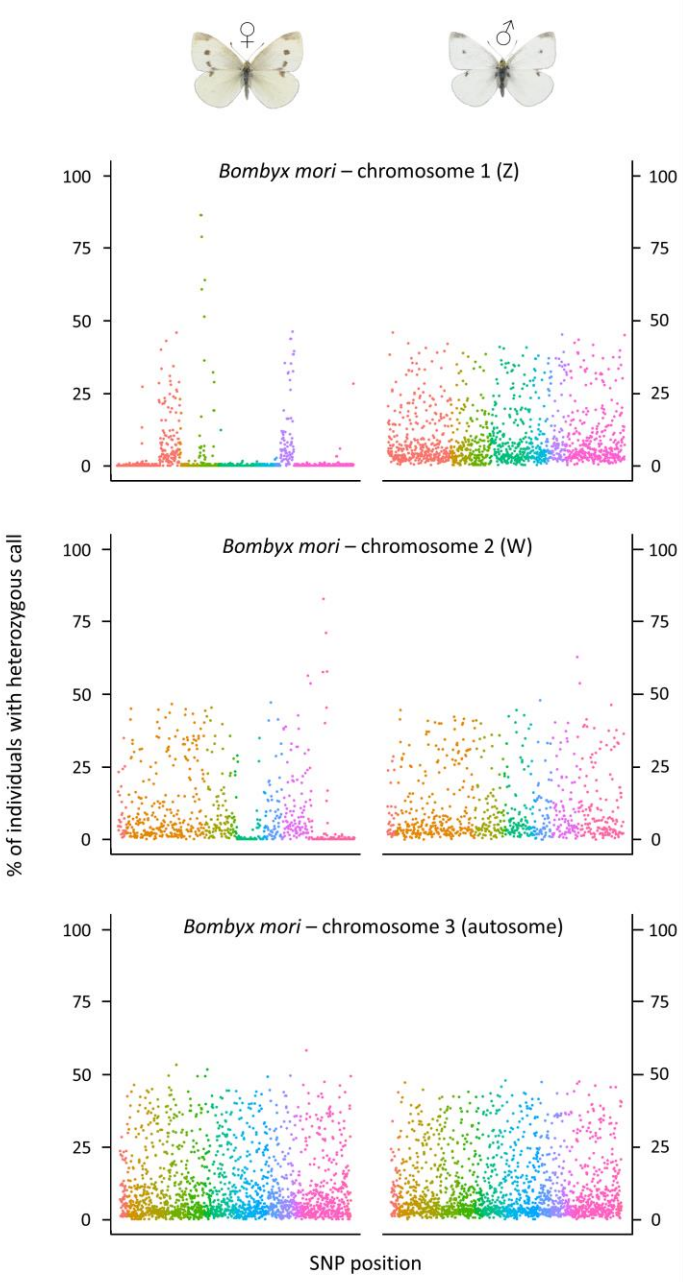



**Fig S6. Percentage of individuals with heterozygous calls for each locus, plotted separately for females and males.** The location of each locus is based on its position within each *P. rapae* scaffold, with each *P. rapae* scaffold then ordered in each *B. mori* chromosome based on its homology to each *B. mori* scaffold (see Methods). Loci are colored by the *B. mori* scaffold to which they are associated. An autosome (chromosome 3) is plotted for reference and the pattern reflects those observed in other autosomes—no discernable difference in heterozygosity between males and females. Note, the W chromosome was not sequenced or assembled in the reference genome used in this study and is thus likely to be made up of portions of other chromosomes, including the Z (regions with no heterozygosity in females).

*Video included as a supplementary file*
**Video S1.** Development of railroad lines in the United States from 1830-1972. Railroad line data were obtained from (2) and plotted by their date of operation. Note the completion of railroad lines connecting eastern and western US in 1872, a few years prior (1879) to when a small population originating from North America (east) was believed to be introduced to that exact region—North America (west) (i.e., central California).

## References


1. Cornuet J-M, et al. (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* 30(8):1187–1189.

2. Atack J (2016) Historical Geographic Information Systems (GIS) database of U.S. Railroads for 1830-1972. Available at: https://my.vanderbilt.edu/jeremyatack/data-downloads/.