
Figures and figure supplements

Introgression shapes fruit color convergence in invasive Galápagos tomato

Matthew JS Gibson *et al*

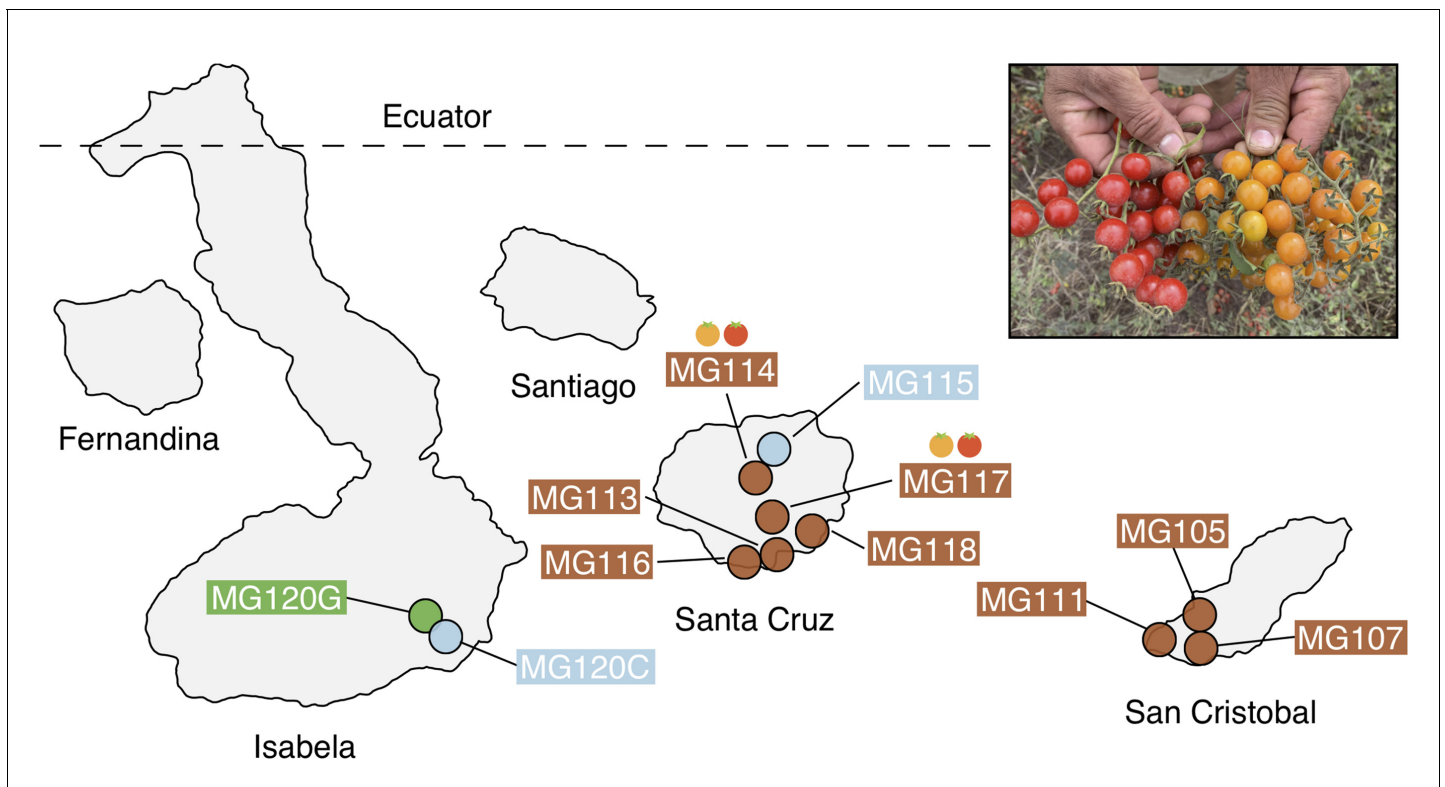


Figure 1. Geographic distribution of focal sampling sites on the Galápagos Islands. Inset: Photograph of polymorphic (red/orange) PIM fruits representative of populations MG114 and MG117. For simplicity, LYC populations as well as sampling sites with <8 individuals are not included here. Refer to **Supplementary file 1a** for a full list of collection localities and sample sizes.

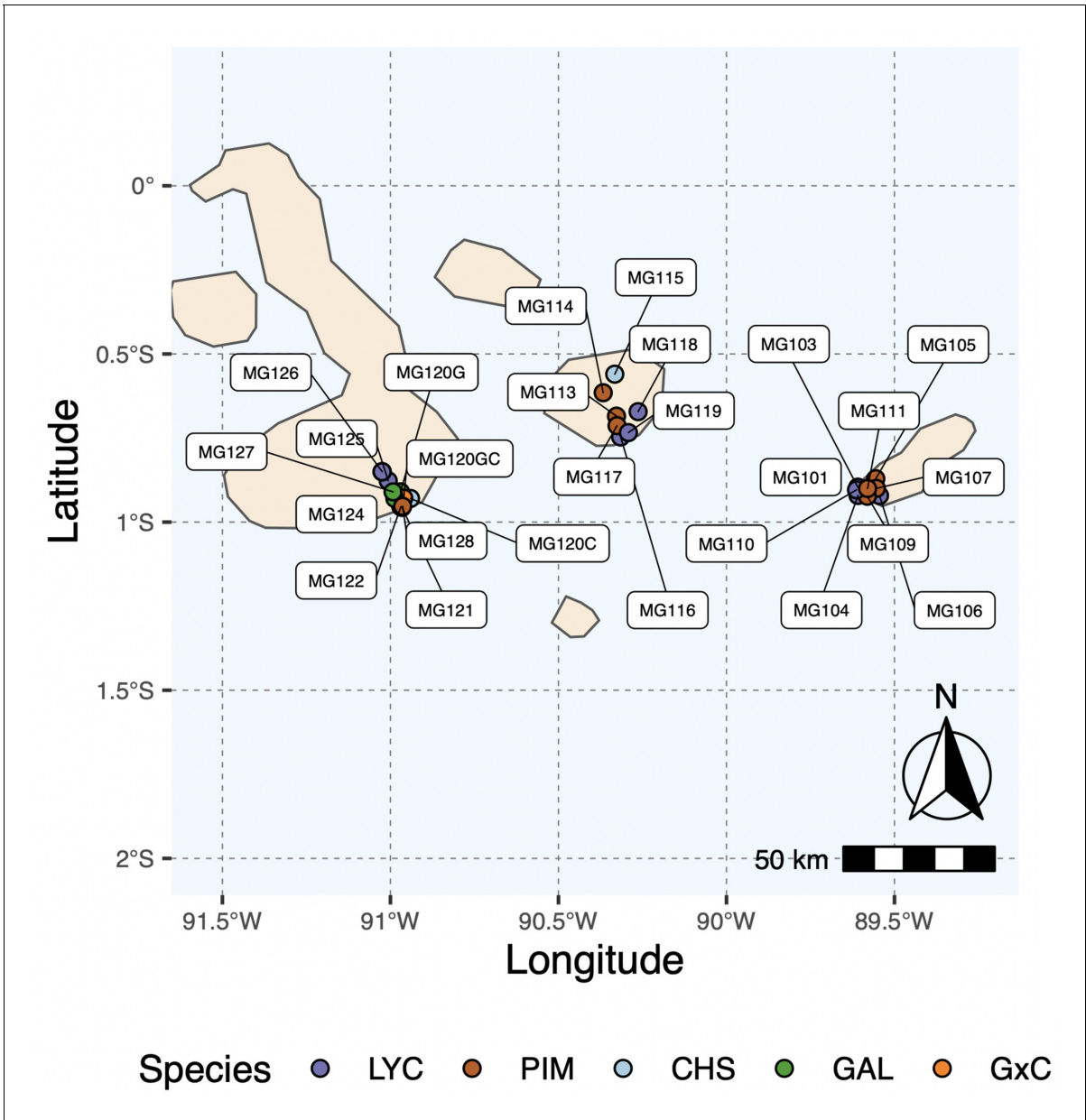


Figure 1—figure supplement 1. Map of all island collection locations. Refer to *Supplementary file 1a* for details.

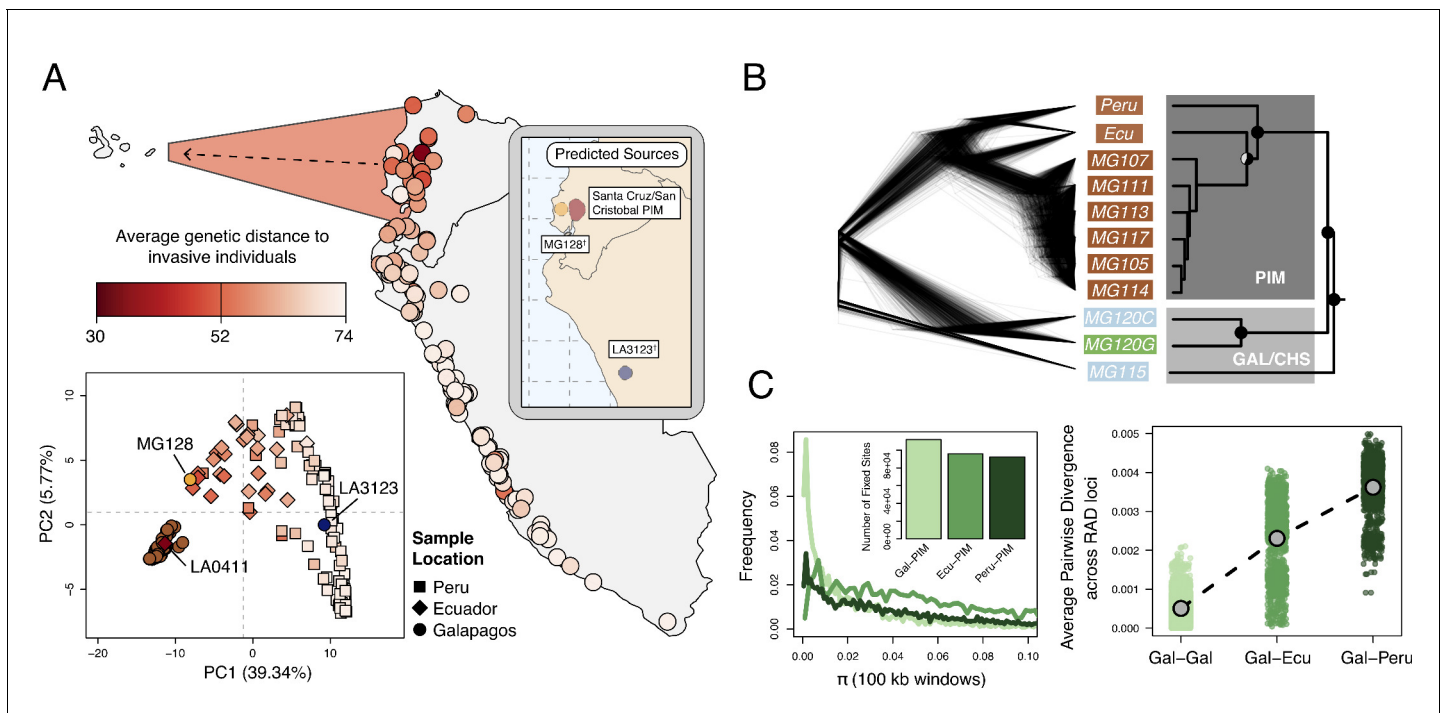


Figure 2. Galápagos PIM is the result of a recent invasion from Ecuador. **(A)** Map: average genetic distance between Galápagos PIM collections and each of the 132 mainland accessions. *Plot*: multi-locus principal components analysis (PCA). Squares, diamonds, and circles indicate Peruvian, Ecuadorian, and Galápagos collections, respectively. *Inset*: Predicted continental origins for Galápagos PIM collections. Colors are same as shown in the multi-locus PCA (*Exact locations vary substantially between runs. Results from a single run are shown). **(B)** Maximum likelihood relationships among focal populations calculated with *Treemix* (allowing no migration). *Left*: inferred trees of 1000 resampled datasets (500 SNPs, with replacement). *Right*: consensus topology. All trees were rate-smoothed ($\lambda = 1$). **(C)** Diversity and divergence metrics. *Left*: nucleotide diversity (π) calculated for Galápagos PIM, Ecuador-PIM, and Peru PIM in overlapping 100 kb windows. Invariant windows ($\pi = 0$) are truncated and are instead shown in the inset bar plot. *Right*: average pairwise sequence divergence for three PIM comparisons: Gal×Gal, Gal×Ecu, and Gal×Peru. Each point represents a comparison between individuals, averaged over all loci.

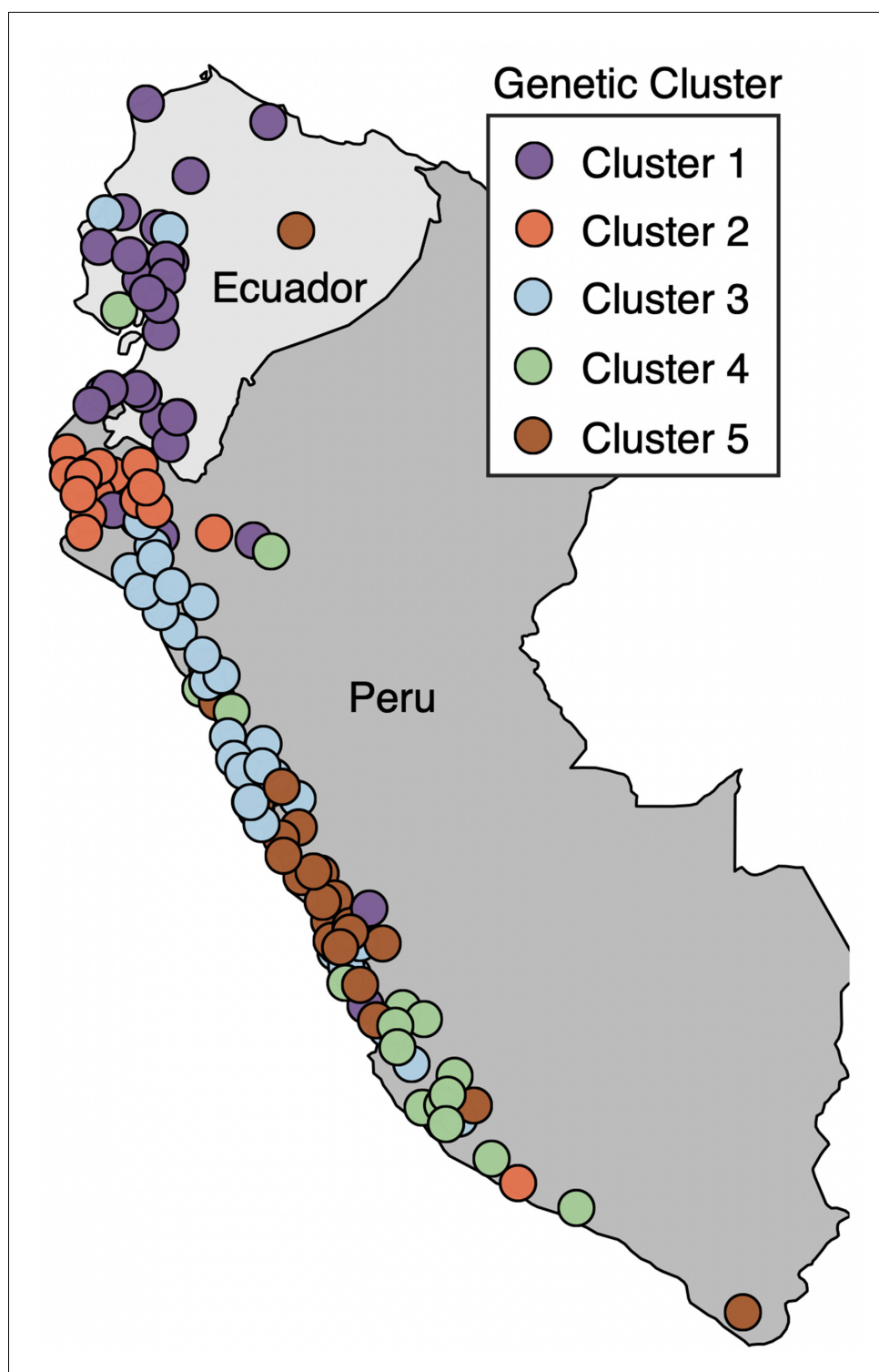


Figure 2—figure supplement 1. Map of mainland collection sites, colored by genetic ancestry cluster as determined in *Gibson and Moyle, 2020*.

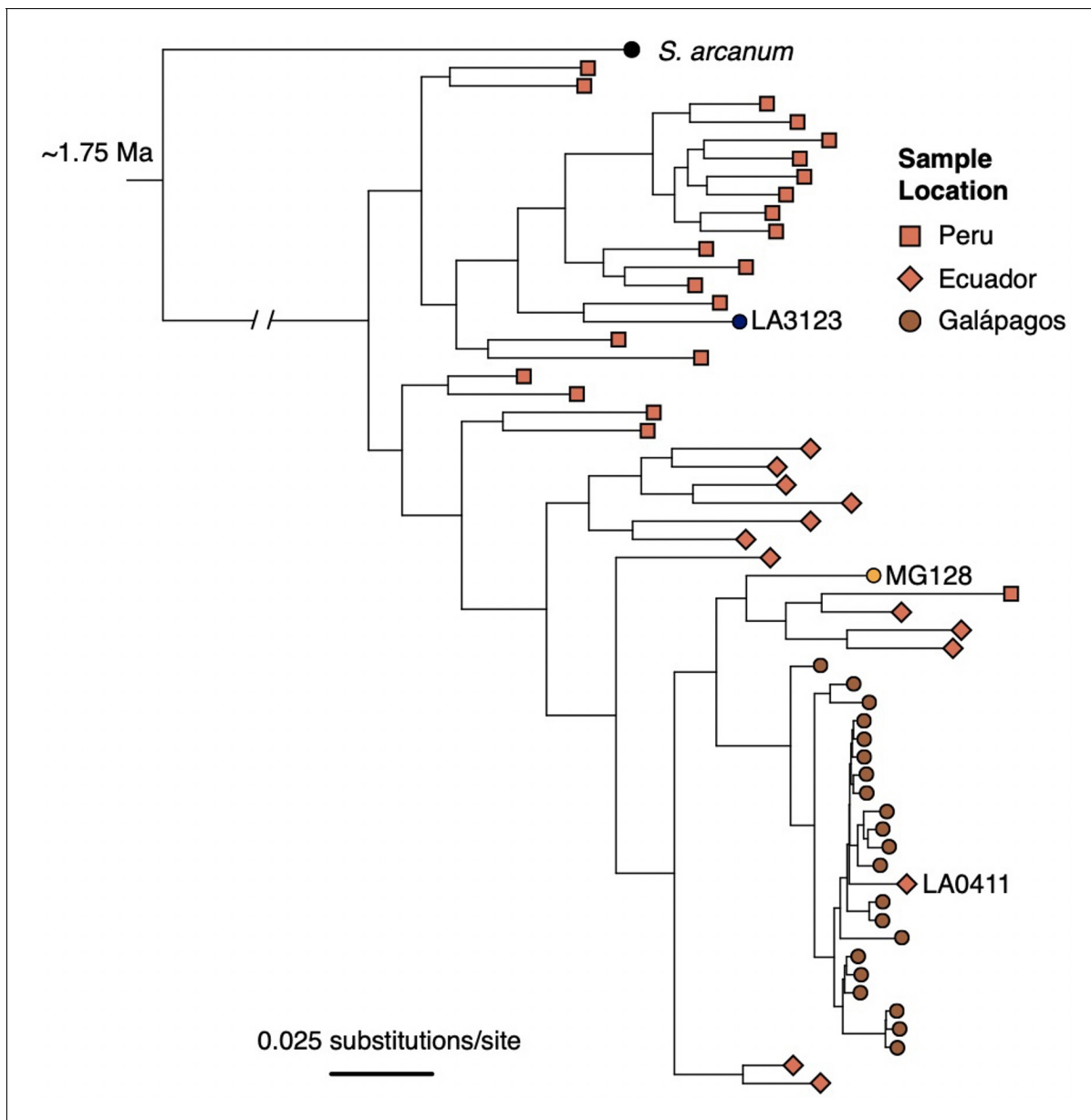


Figure 2—figure supplement 2. ML tree of individual samples inferred with RAxML, using data concatenated across all RAD loci. Samples were subset by population (for Galápagos collections, 1–2 individuals/population) and by geographic region (for mainland accessions; 20 individuals from Peru, 14 individuals from Ecuador) to limit redundancy and increase computation speed.

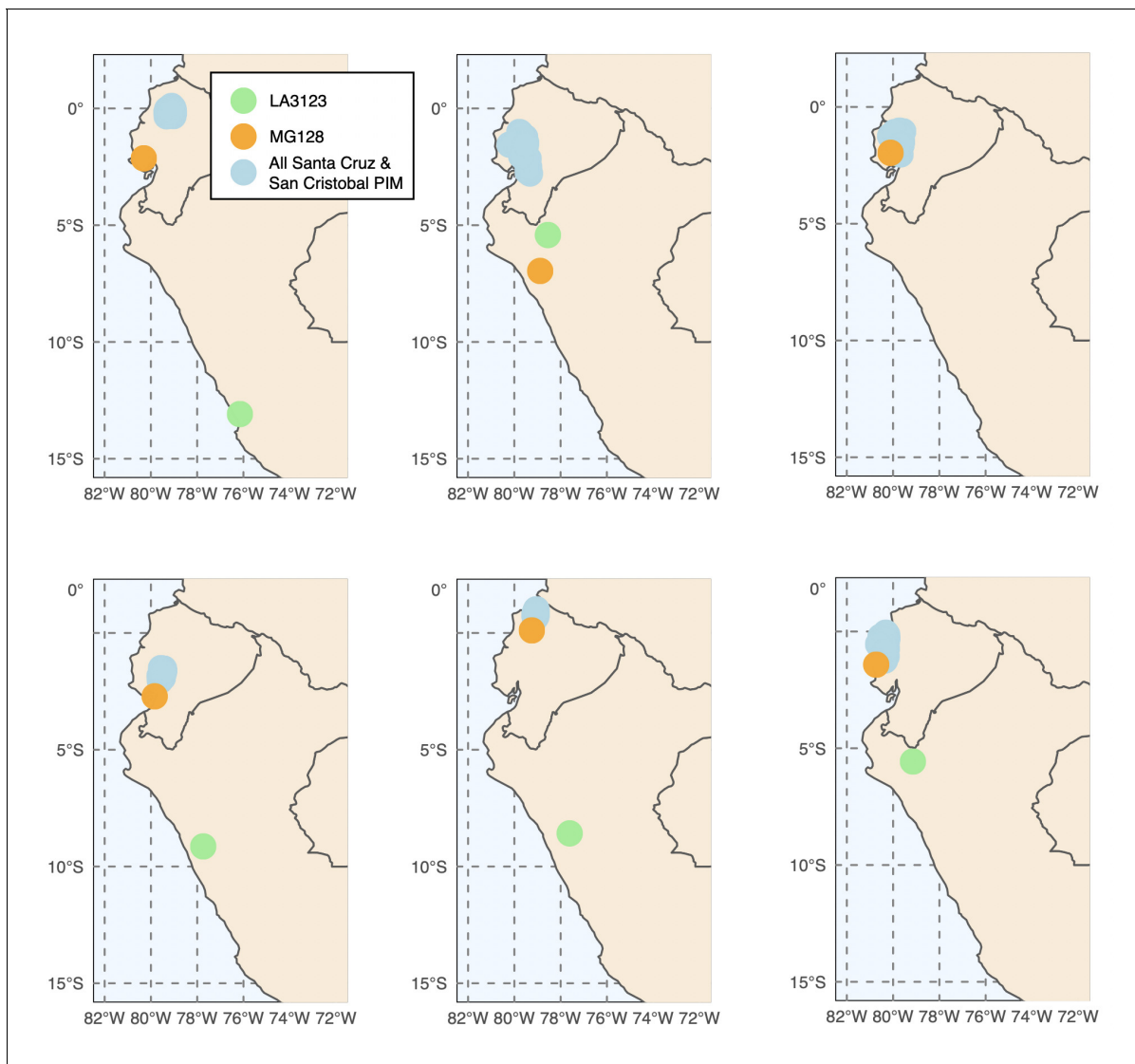


Figure 2—figure supplement 3. Six runs of *Locator* (Battey et al., 2020) generally support a three-invasion scenario. Exact source localities for MG128-1 and LA3123 varied substantially across run.

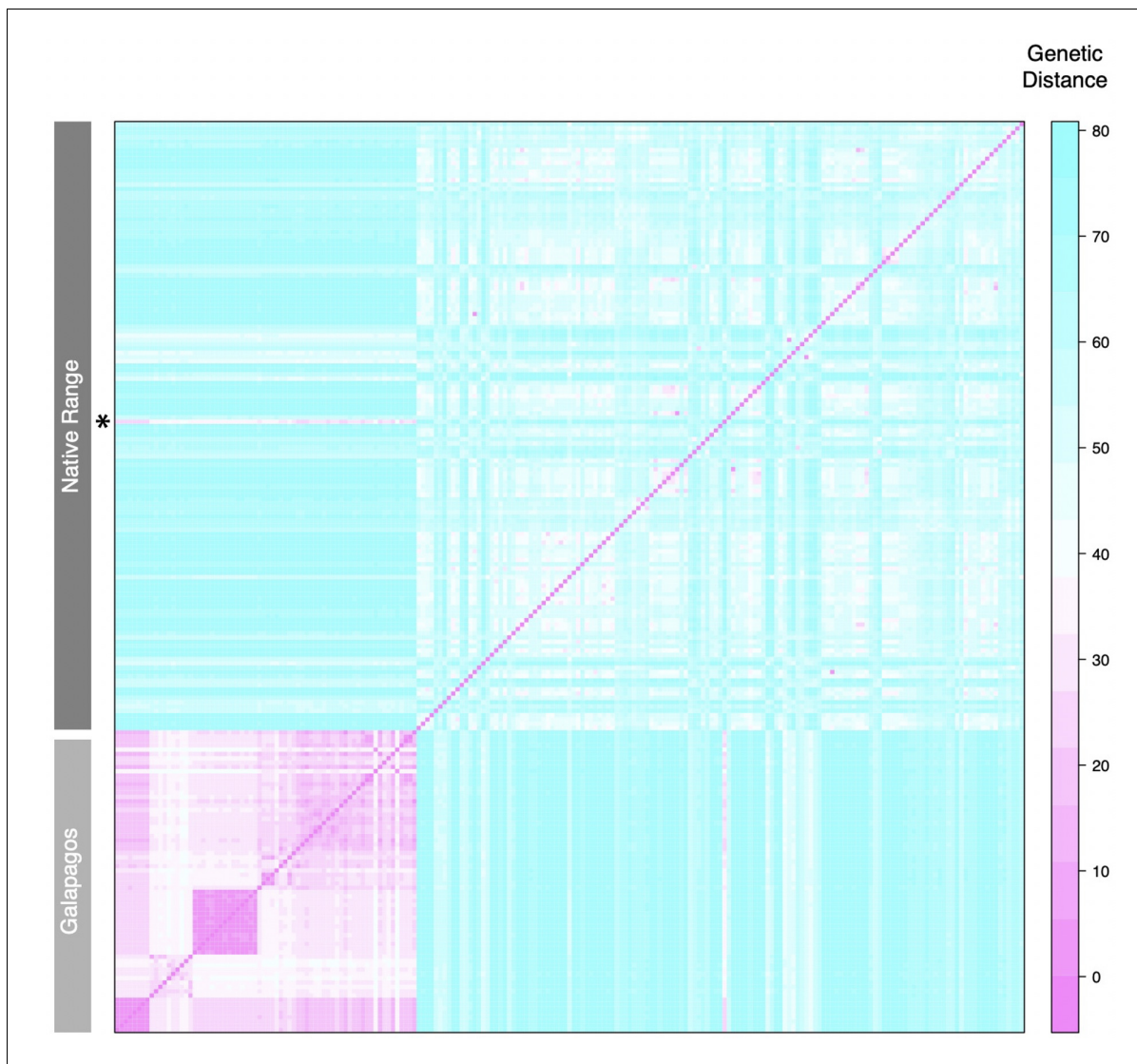


Figure 2—figure supplement 4. Pairwise genetic distances at all polymorphic loci. Row indicated with an asterisk is LA0411, a sample putatively reintroduced to mainland Ecuador from Galápagos.

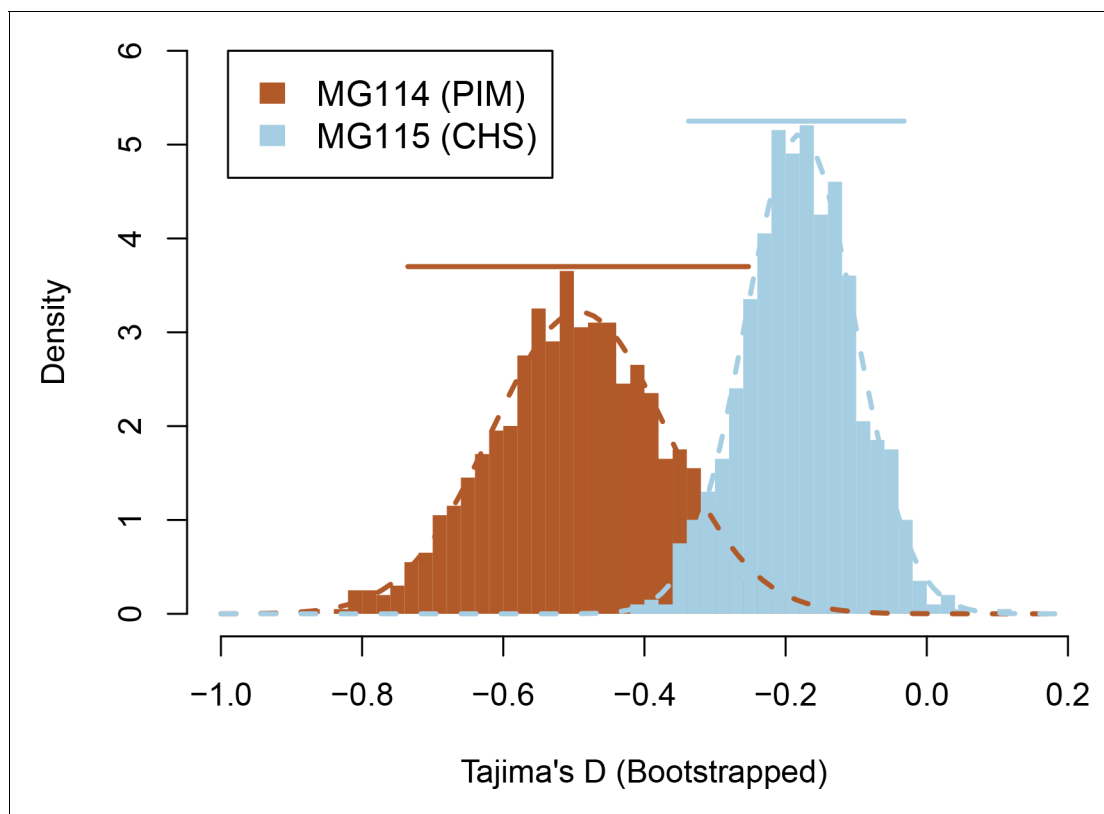


Figure 2—figure supplement 5. Bootstrapped sampling distributions of Tajima's D for populations MG114 and MG115. Colored bars represent 95% CIs. Data were generated from 1000 bootstrap replicates of the site frequency spectrum (SFS).

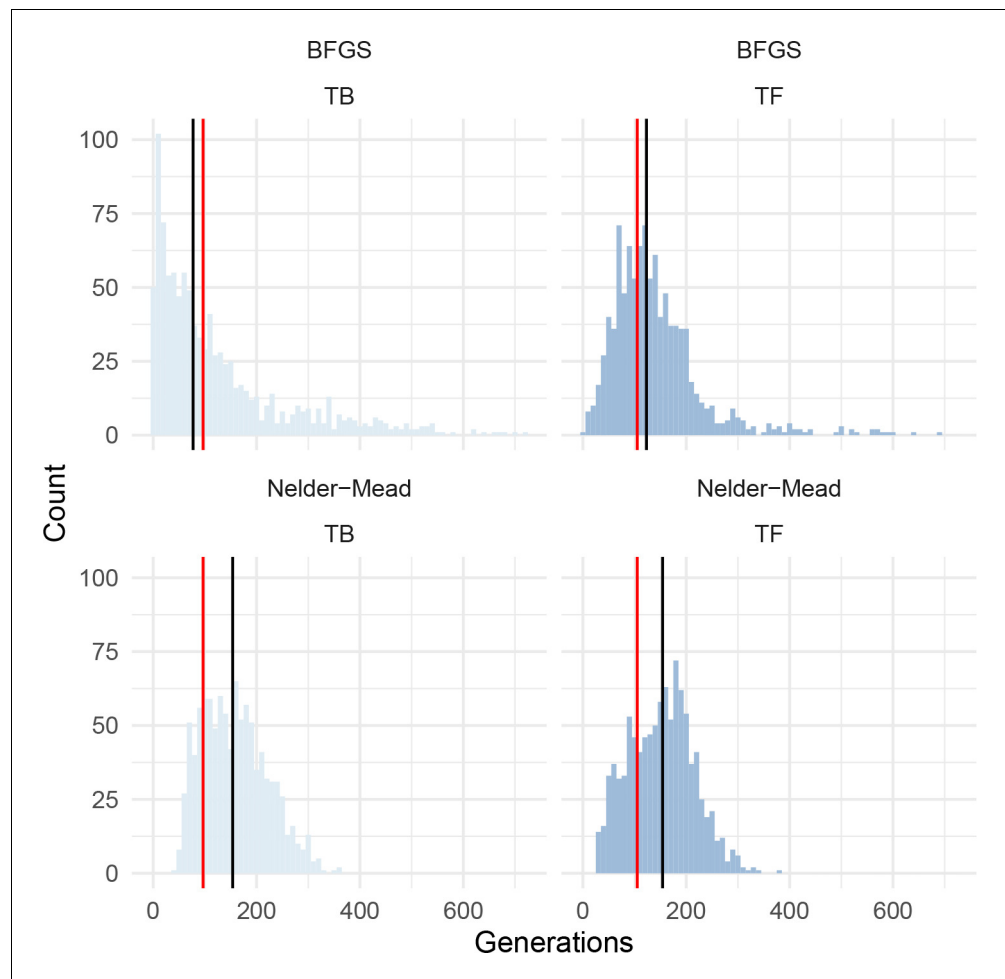


Figure 2—figure supplement 6. Histograms of bootstrapped parameter estimates from dadi for PIM population MG114, using the introgression-masked site frequency spectrum. Red bars indicate the optimum value inferred. Black bars indicate the bootstrapped median value. Results from two optimization algorithms are shown; BFGS (top panels) and Nelder-Mead (bottom panels). TB = length of bottleneck; TF = time since end of bottleneck.

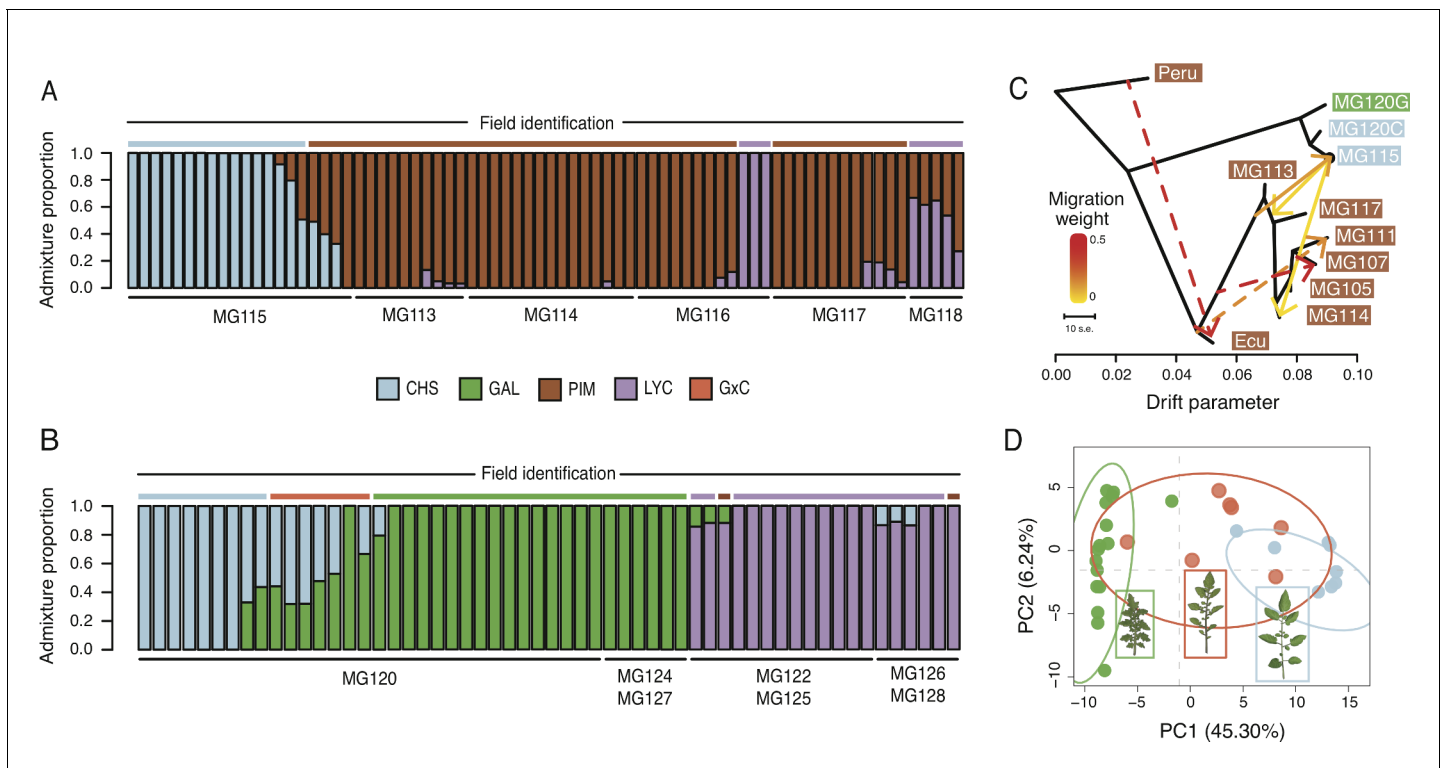


Figure 3. Patterns of population genetic structure and admixture on Santa Cruz and Isabela. (A) *fastStructure* inference for all Santa Cruz samples (N = 74). K = 3. (B) *fastStructure* inference for all Isabela samples (N = 57). K = 3. (C) *Treemix* analysis summary (m = 6; ln[L]=395.08). Solid lines indicate interspecific events and dashed lines indicate intraspecific events. (D) Principal components analysis for samples at site MG120, a hybrid zone between CHS and GAL.

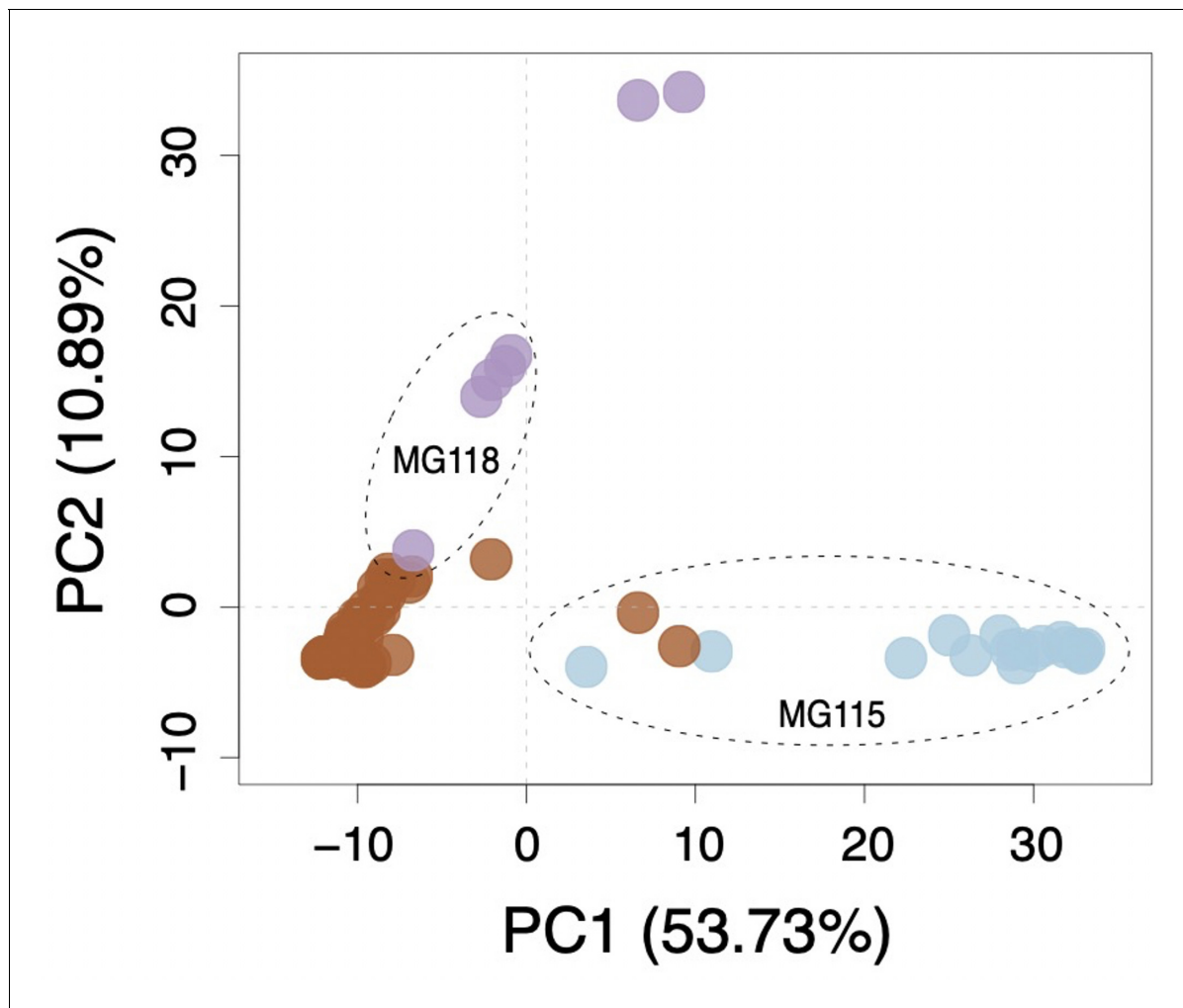


Figure 3—figure supplement 1. Multi-locus principal components analysis (PCA) for Santa Cruz collections. Brown, purple, and blue points correspond to PIM, LYC, and CHS, respectively.

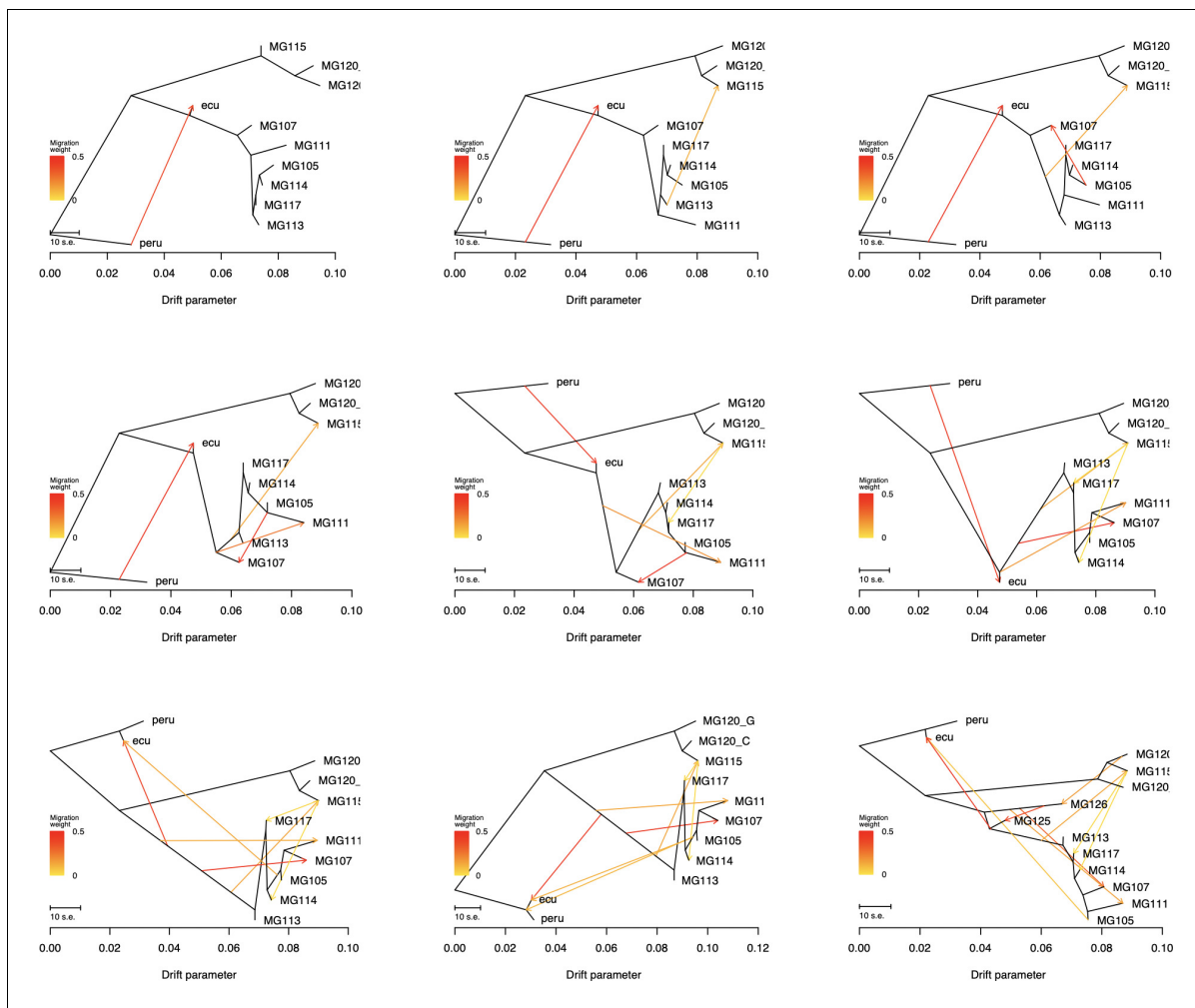


Figure 3—figure supplement 2. Treemix summary figures for all tested values for m (migration events: 1–8).

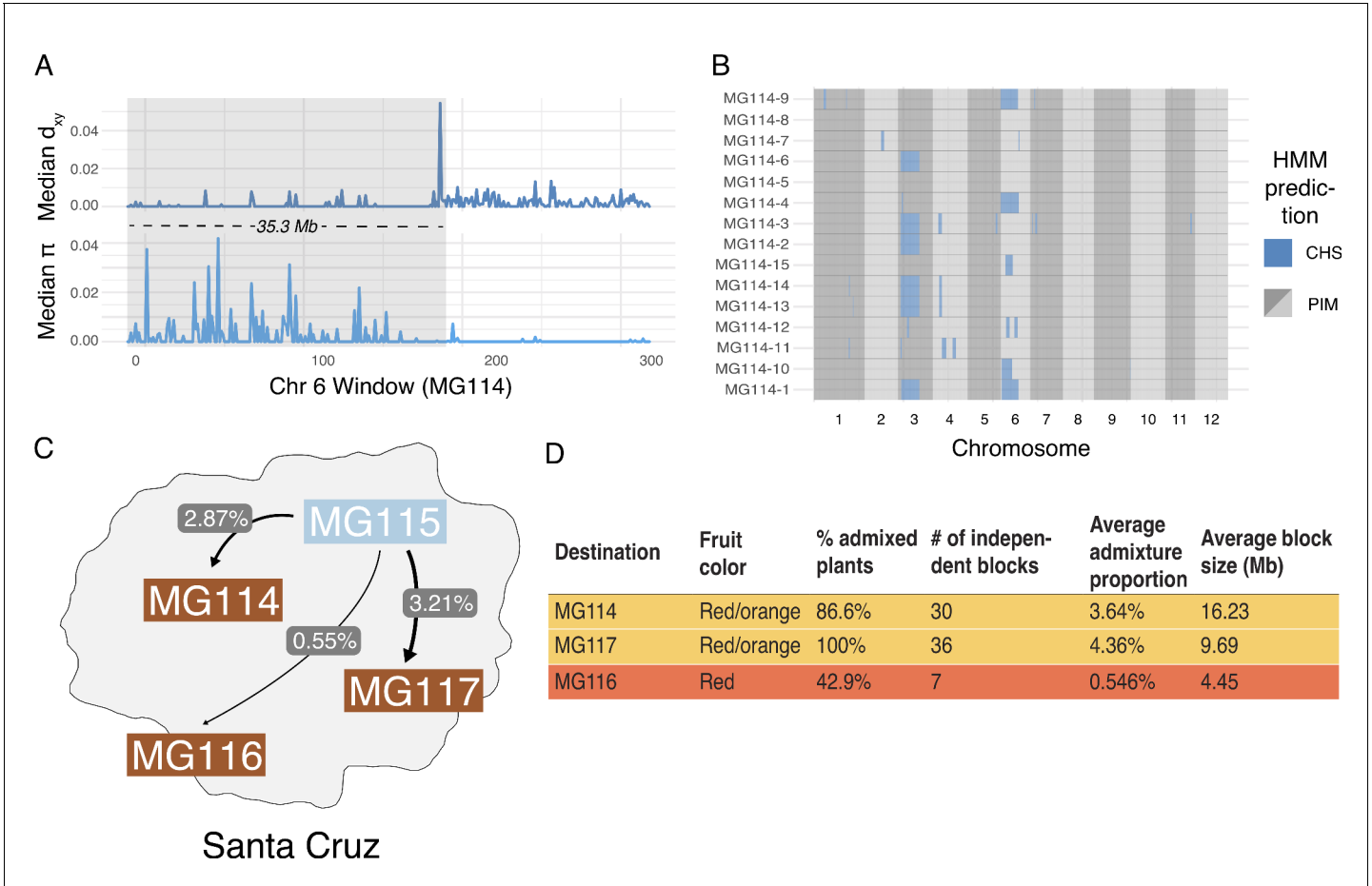


Figure 4. Local ancestry assignment using hidden Markov model (HMM) characterizes a history of endemic \times invasive introgression. **(A)** Patterns of diversity and divergence along chromosome 6 for an MG114 individual. The region of recent coalescence (low divergence; high diversity) with CHS is annotated in gray. This 20.2 kb block segregates at 20% in MG114. **(B)** Genome-wide HMM predictions for all individuals in MG114. The x-axis is ordered by chromosome and y-axis is ordered by individual. Two large CHS haplotypes segregate at high frequency on chromosomes 3 (40%) and 6 (20%). **(C)** Visual summary of admixture proportions from CHS into three PIM populations. **(D)** Summary of HMM assignment for each PIM population. Populations displaying variation in fruit color (MG114 and MG117) have more CHS ancestry than those which are fixed for the ancestral red state (MG116).

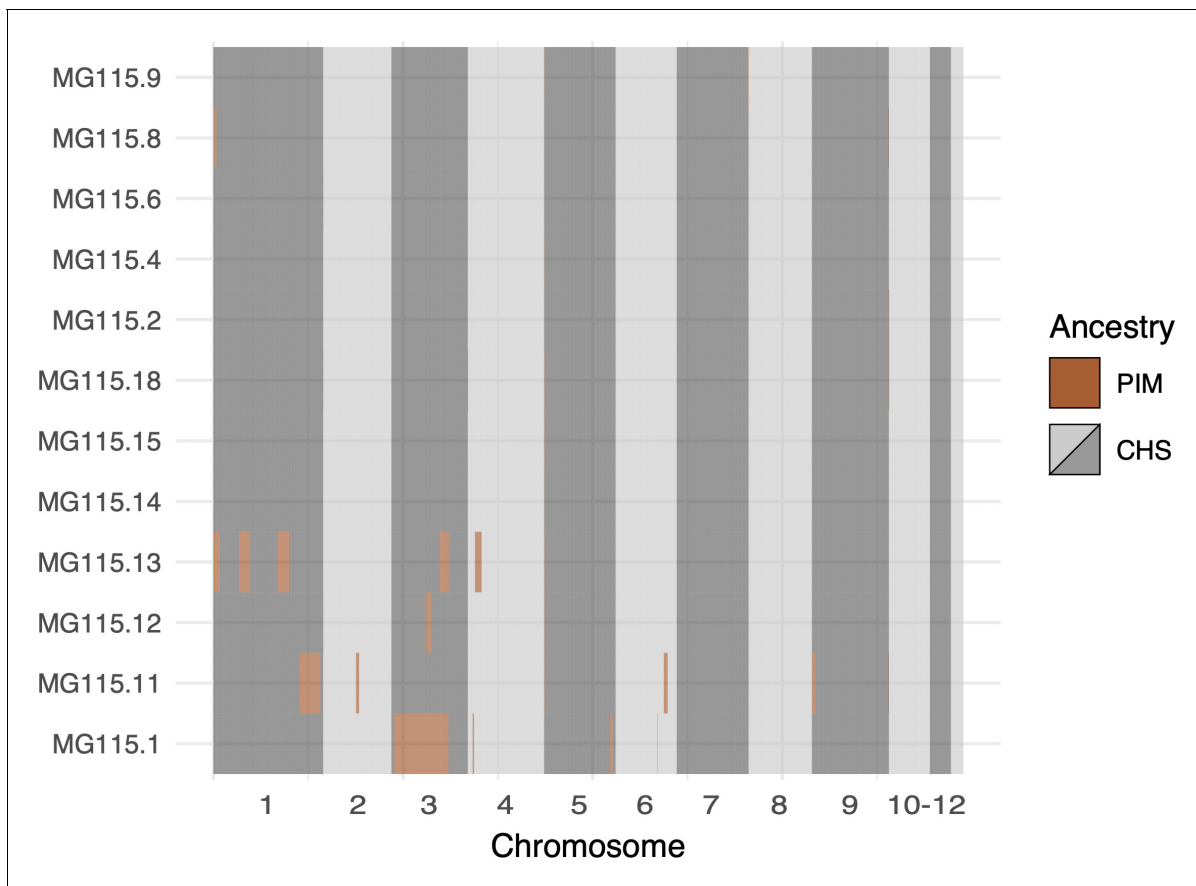


Figure 4—figure supplement 1. Local ancestry assignment throughout the genomes of CHS plants from population MG115 (Santa Cruz), using MG114 as the PIM reference population. These data represent the opposite direction of gene flow (PIM → CHS) from that shown in **Figure 5B** (CHS → PIM). Admixture in the PIM → CHS direction was markedly lower than CHS → PIM.

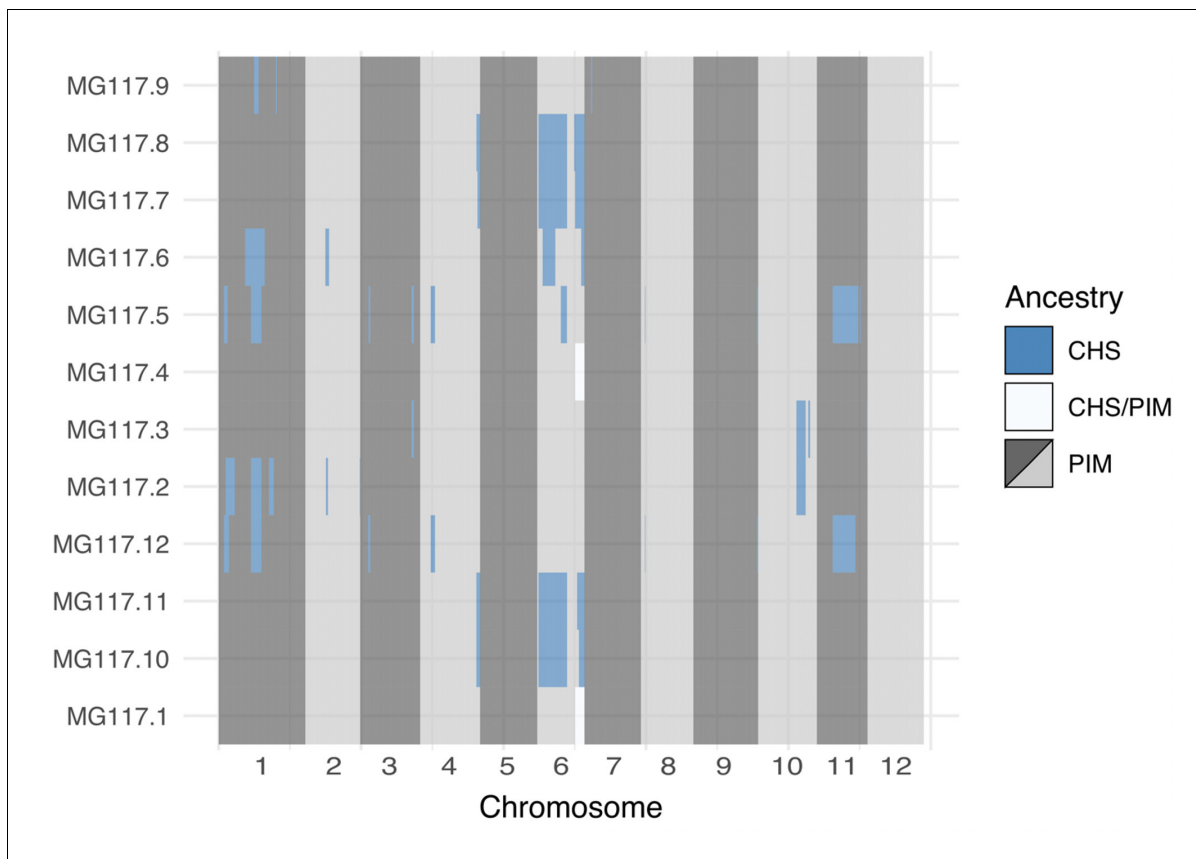


Figure 4—figure supplement 2. Genome-wide local ancestry in population MG117 as inferred by the hidden Markov model (HMM).

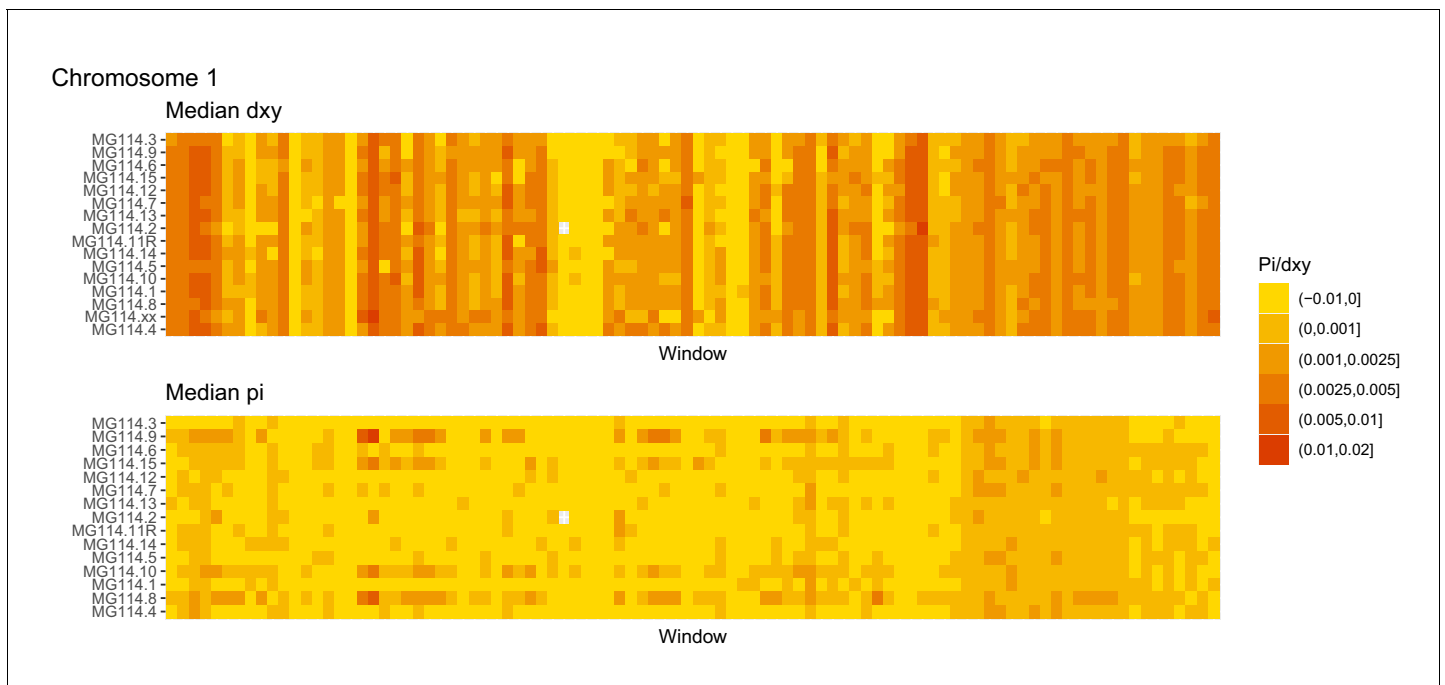


Figure 4—figure supplement 3. Diversity and divergence across the genomes of MG114 individuals (Chromosome 1). Each cell represents 1 Mb and colors correspond to median π or d_{xy} (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

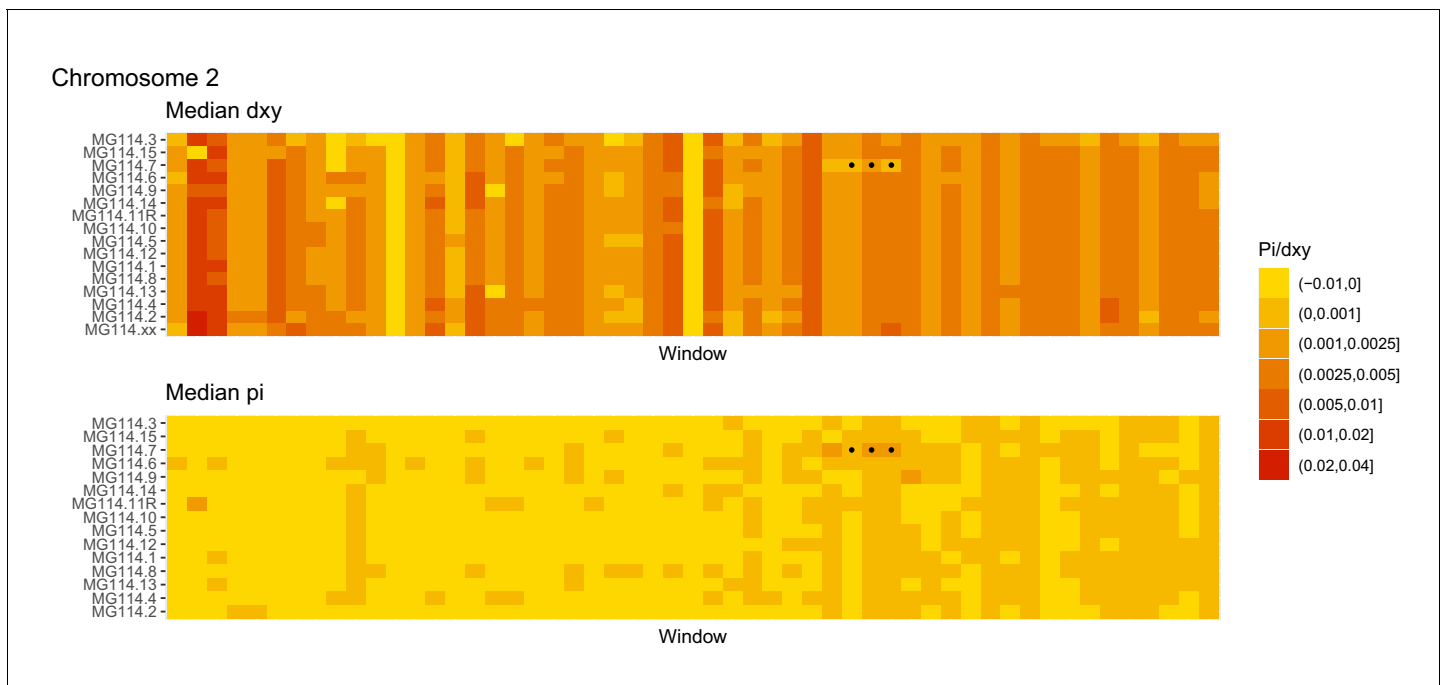


Figure 4—figure supplement 4. Diversity and divergence across the genomes of MG114 individuals (Chromosome 2). Each cell represents 1 Mb and colors correspond to median π or d_{xy} (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

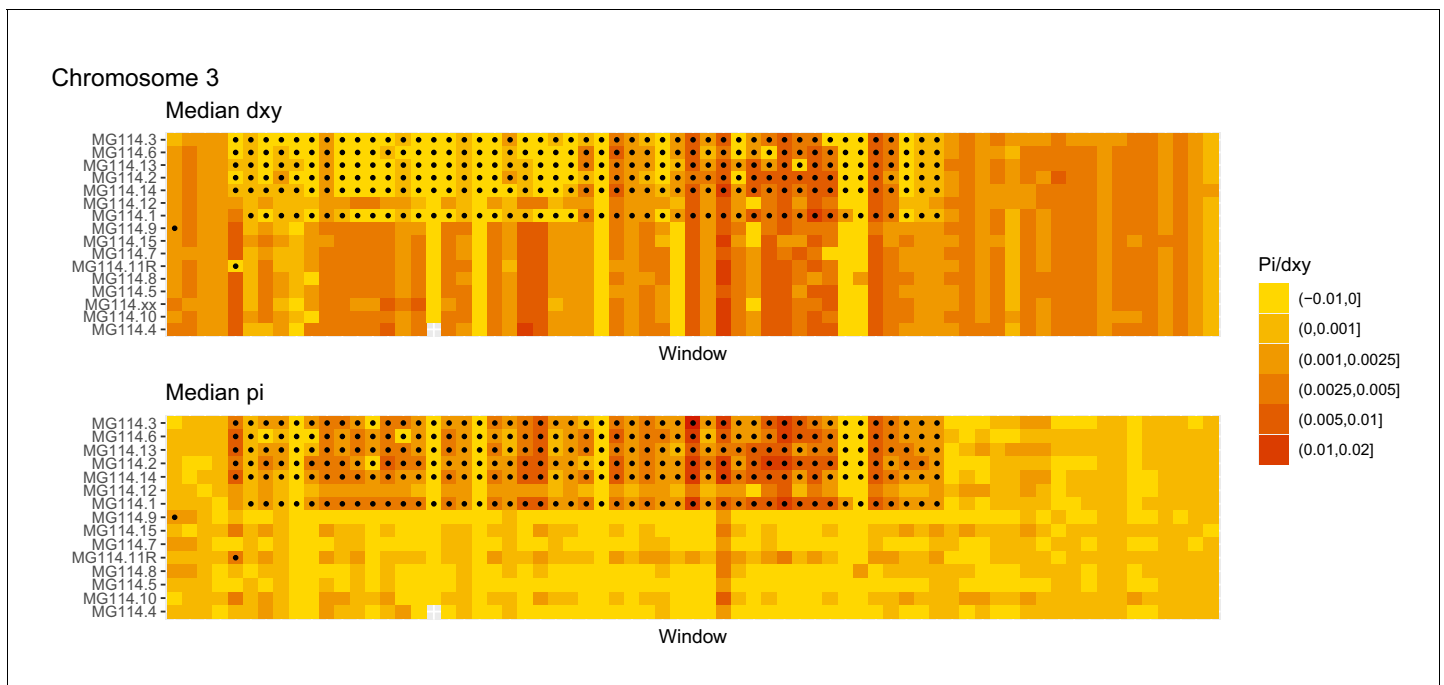


Figure 4—figure supplement 5. Diversity and divergence across the genomes of MG114 individuals (Chromosome 3). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

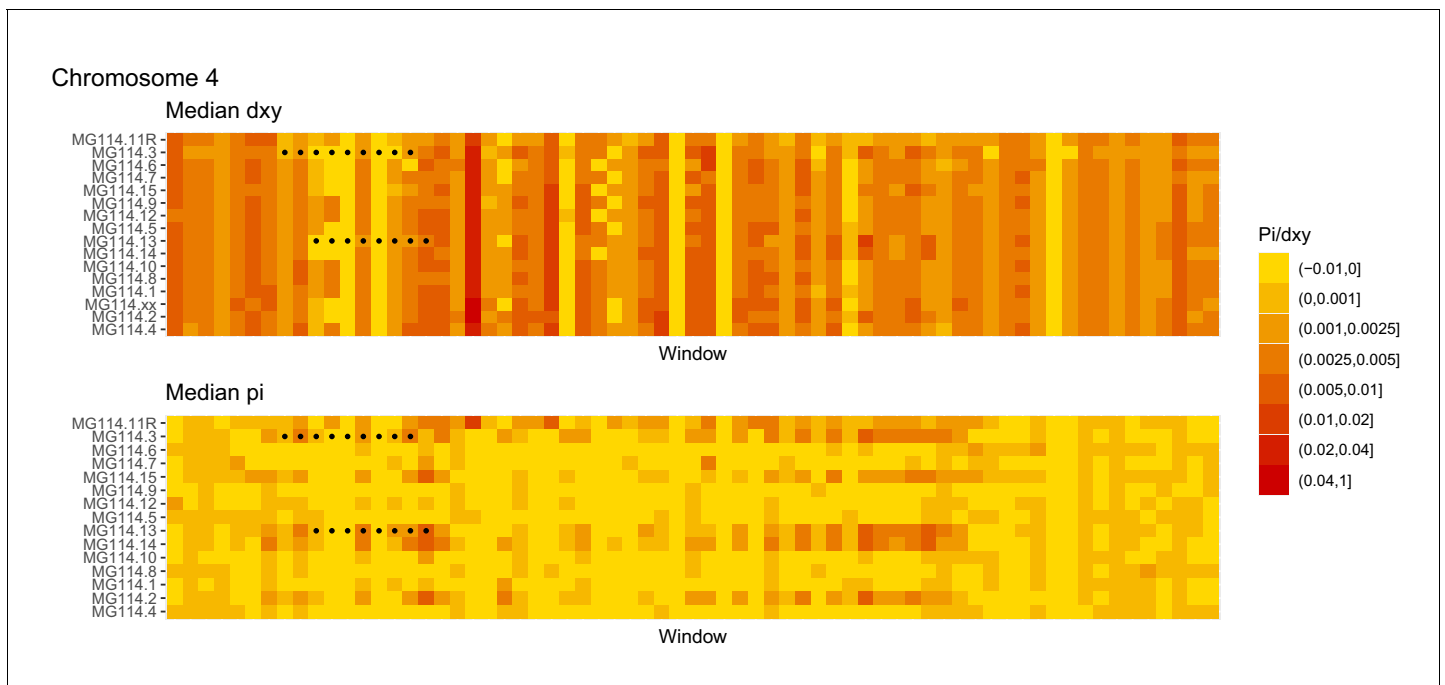


Figure 4—figure supplement 6. Diversity and divergence across the genomes of MG114 individuals (Chromosome 4). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

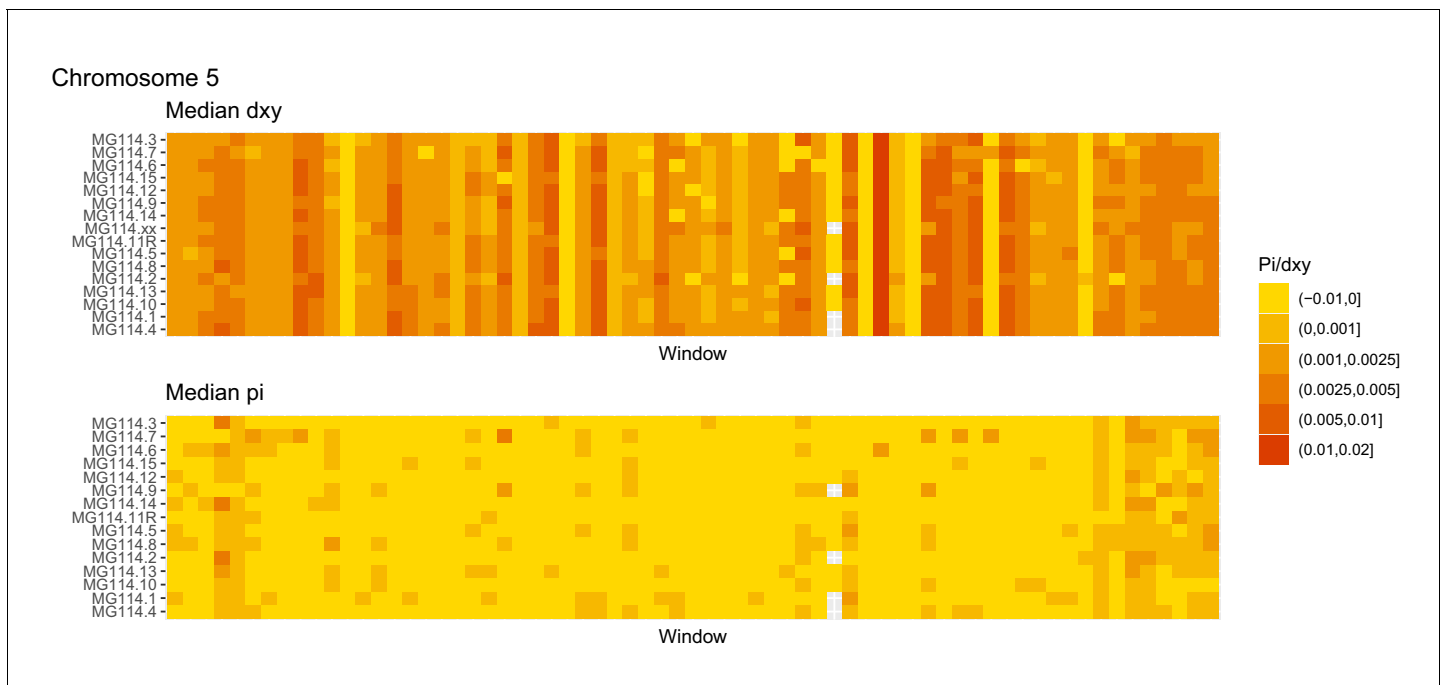


Figure 4—figure supplement 7. Diversity and divergence across the genomes of MG114 individuals (Chromosome 5). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

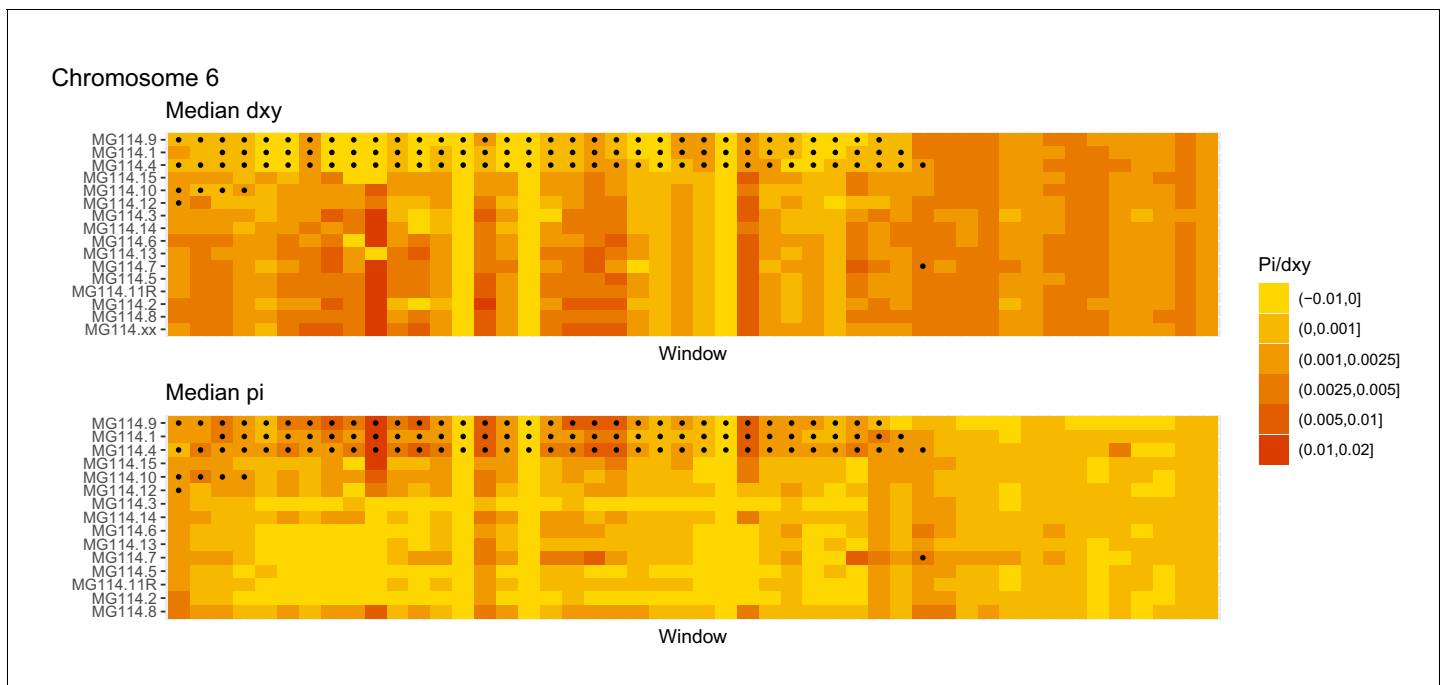


Figure 4—figure supplement 8. Diversity and divergence across the genomes of MG114 individuals (Chromosome 6). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

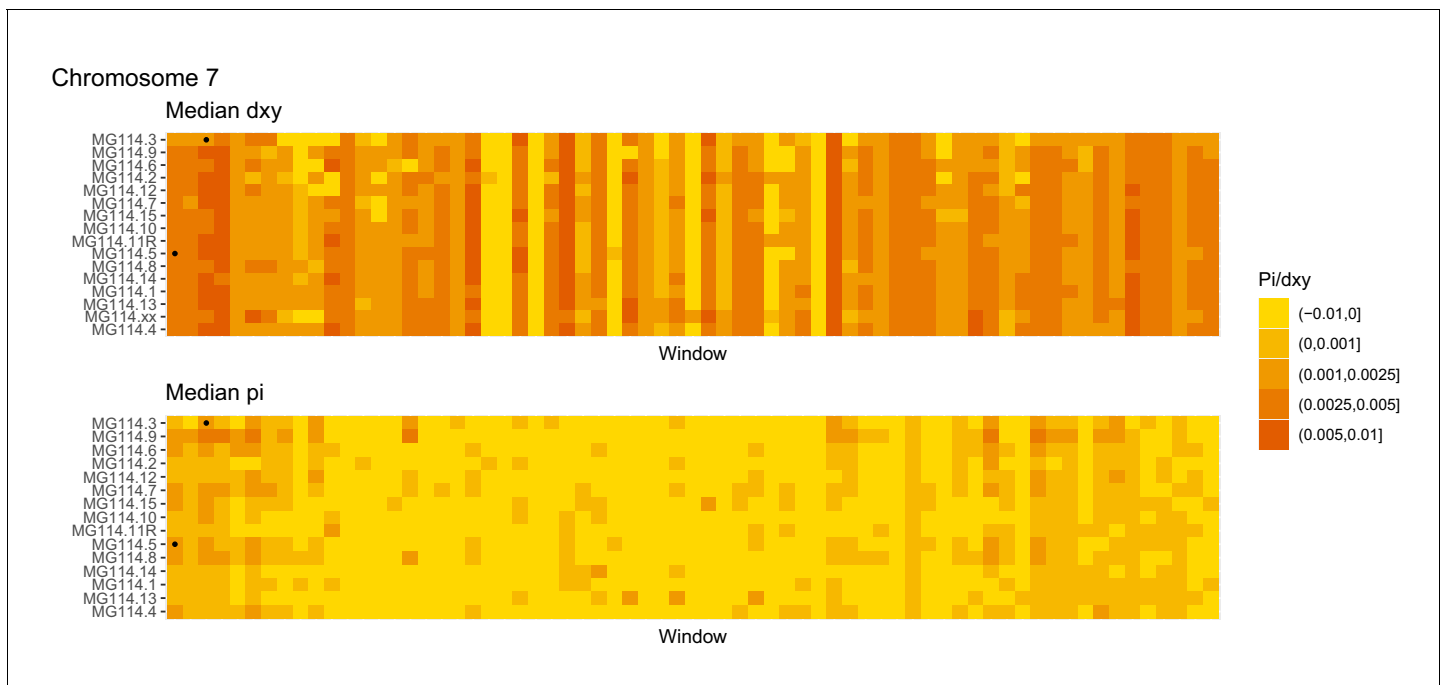


Figure 4—figure supplement 9. Diversity and divergence across the genomes of MG114 individuals (Chromosome 7). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

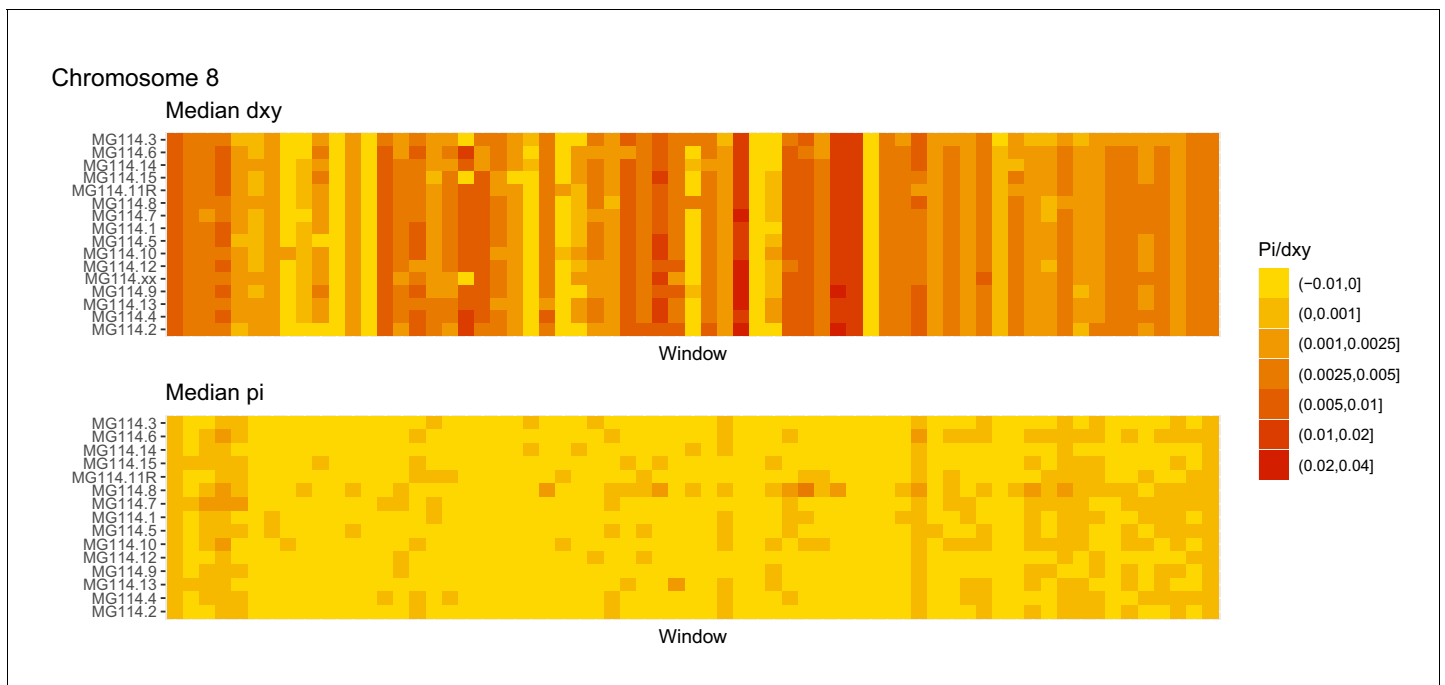


Figure 4—figure supplement 10. Diversity and divergence across the genomes of MG114 individuals (Chromosome 8). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

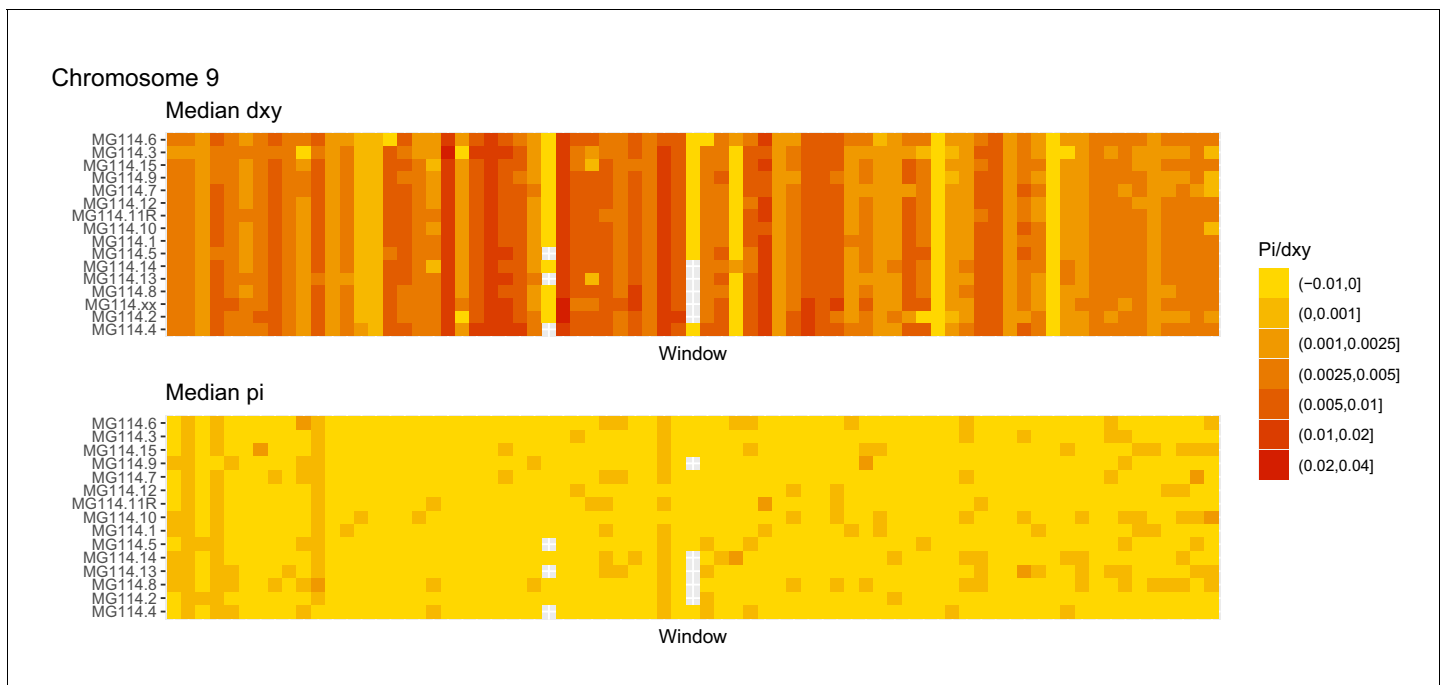


Figure 4—figure supplement 11. Diversity and divergence across the genomes of MG114 individuals (Chromosome 9). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

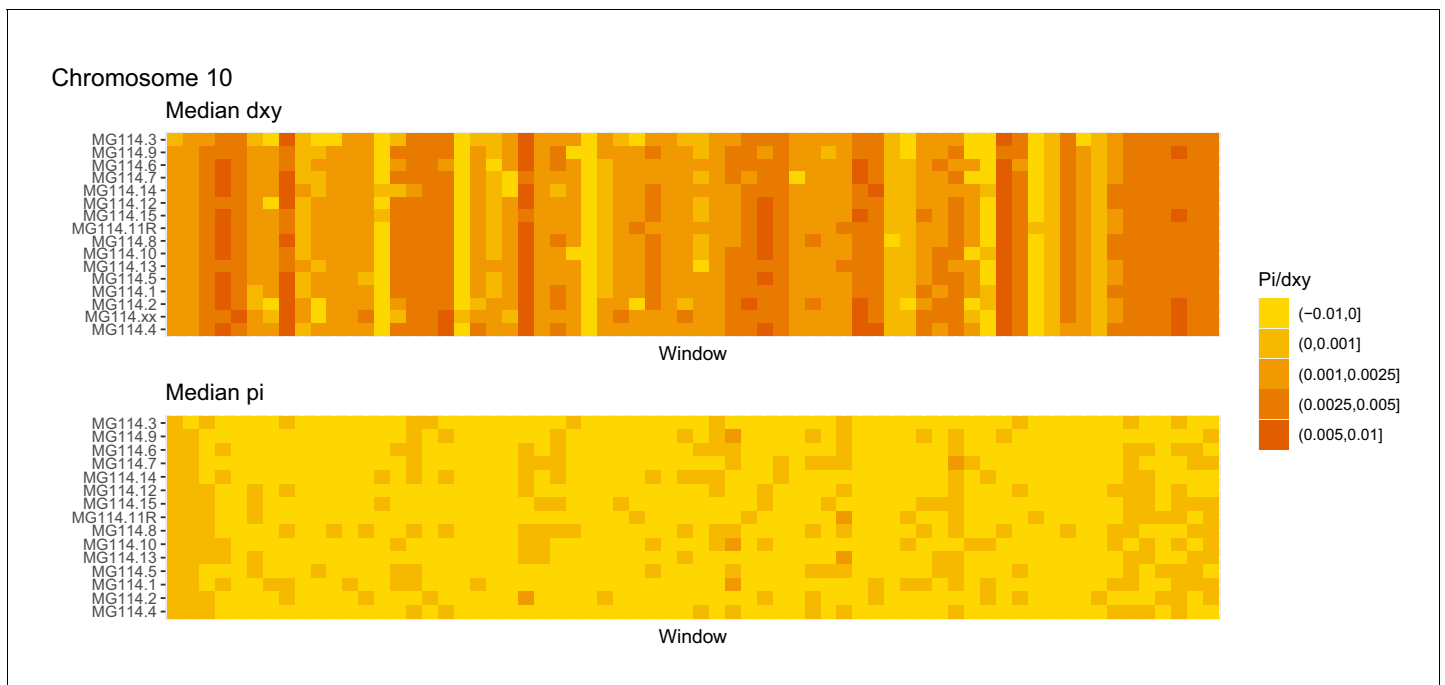


Figure 4—figure supplement 12. Diversity and divergence across the genomes of MG114 individuals (Chromosome 10). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

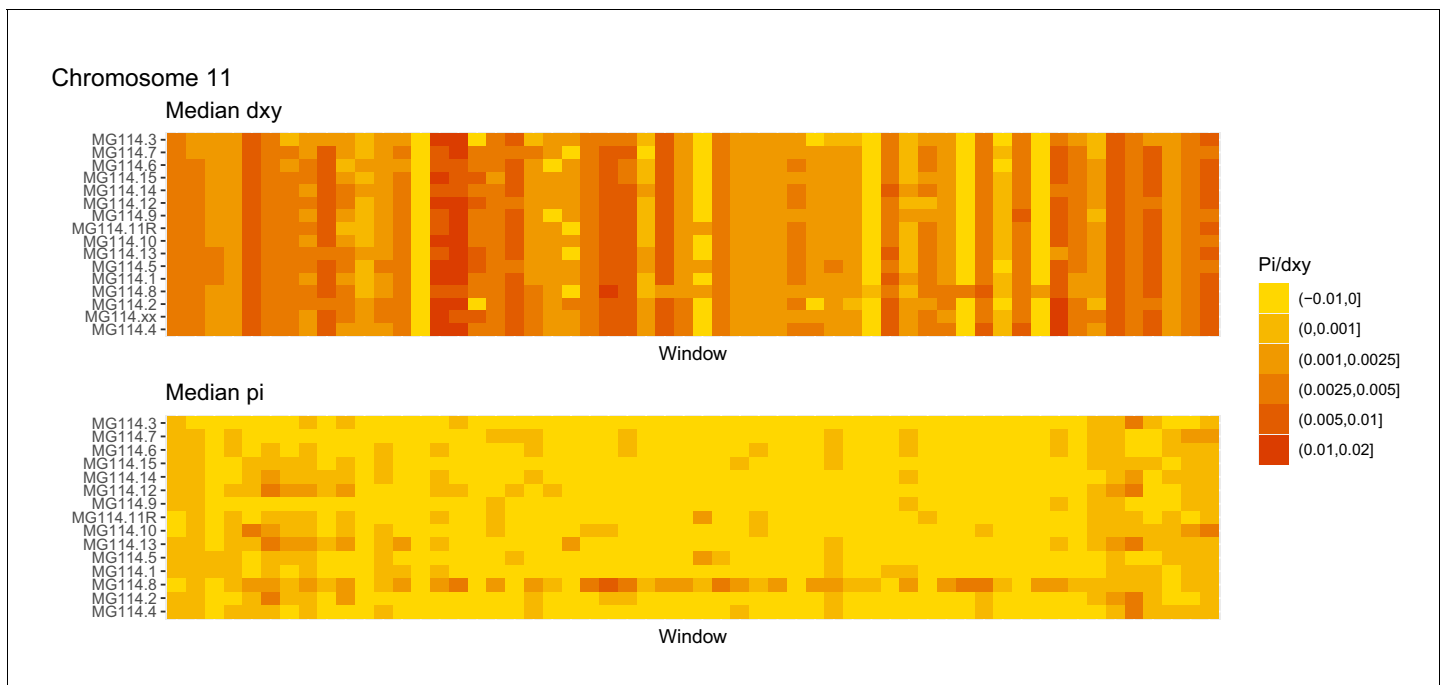


Figure 4—figure supplement 13. Diversity and divergence across the genomes of MG114 individuals (Chromosome 11). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

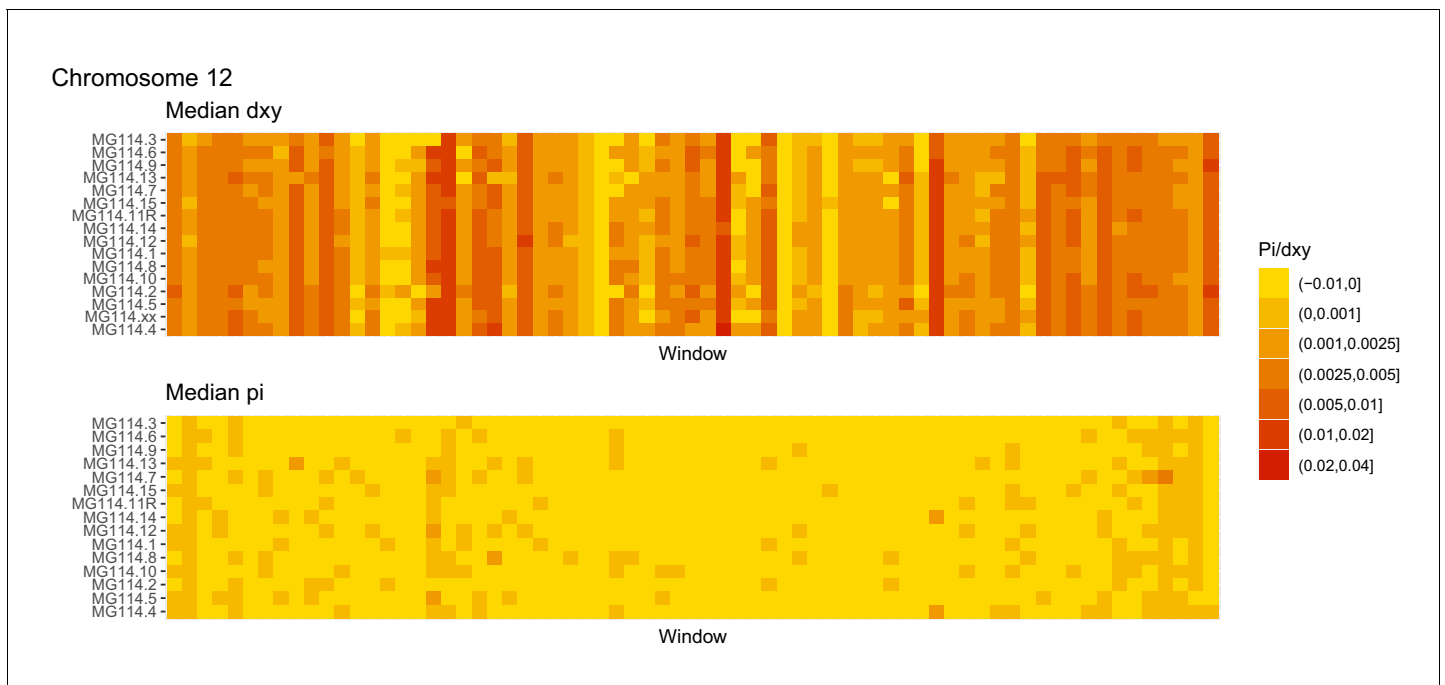


Figure 4—figure supplement 14. Diversity and divergence across the genomes of MG114 individuals (Chromosome 12). Each cell represents 1 Mb and colors correspond to median π or dXY (to CHS population MG115). Black dots show regions of CHS ancestry predicted by the hidden Markov model (HMM). HMM predictions were done in 100 kb windows. Annotations here reflect the consensus prediction of all windows within each 1 Mb cell.

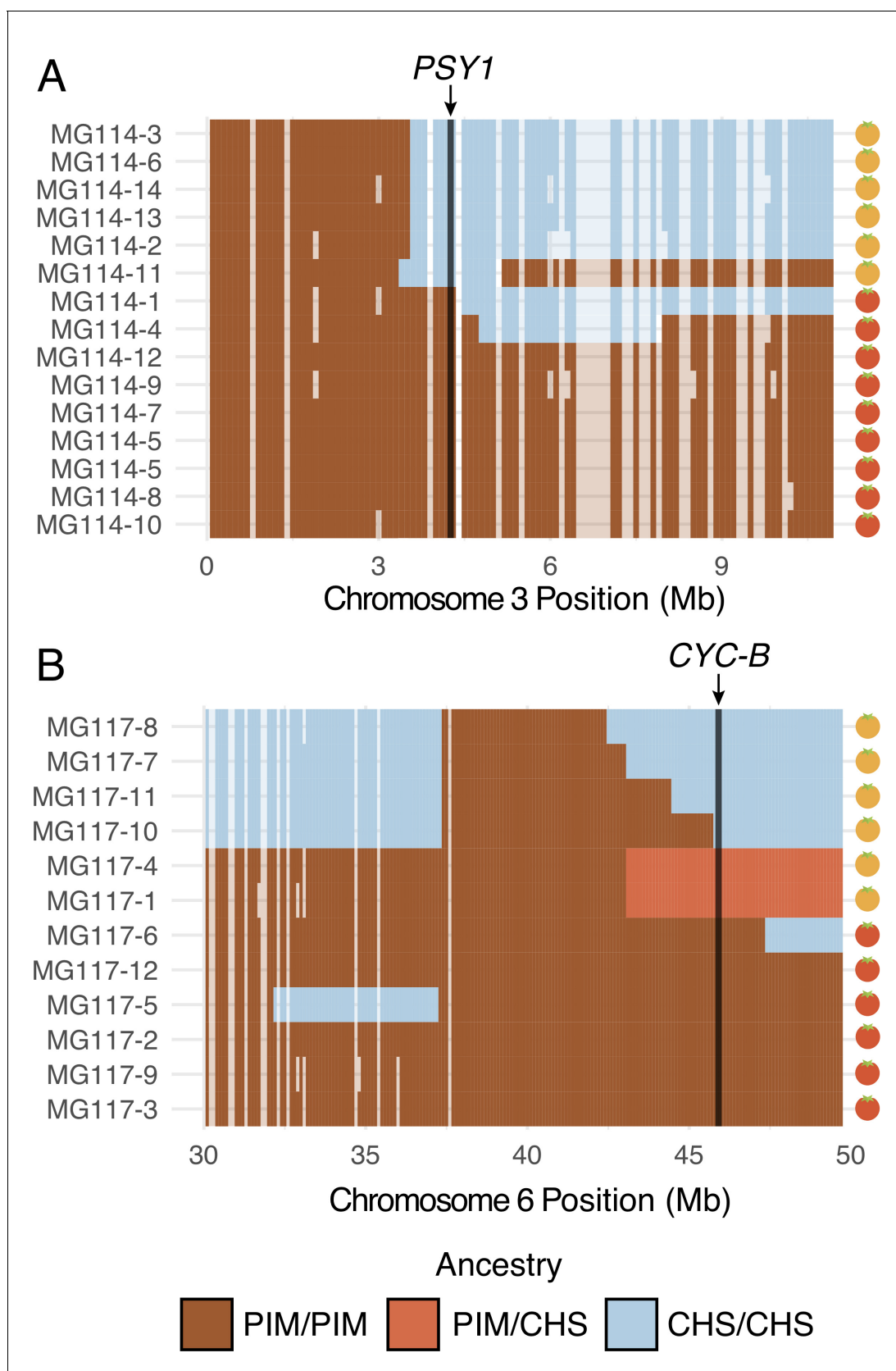


Figure 5. Patterns of local ancestry across focal chromosome regions of MG114 and MG117, enlarged to show variation in introgression block break points at color pathway genes. (A) CHS ancestry at carotenoid biosynthesis gene *PSY1* on chromosome 3 correlates with observed fruit color variation

Figure 5 continued on next page

Figure 5 continued

in MG114. **(B)** CHS ancestry at carotenoid biosynthesis gene *CYC-B* on chromosome 6 correlates with fruit color variation in MG117. Each cell represents 100 kb. Empty cells indicate windows with no sequence data. Empty cells are ghost shaded with each ancestry color based on neighboring assignments.

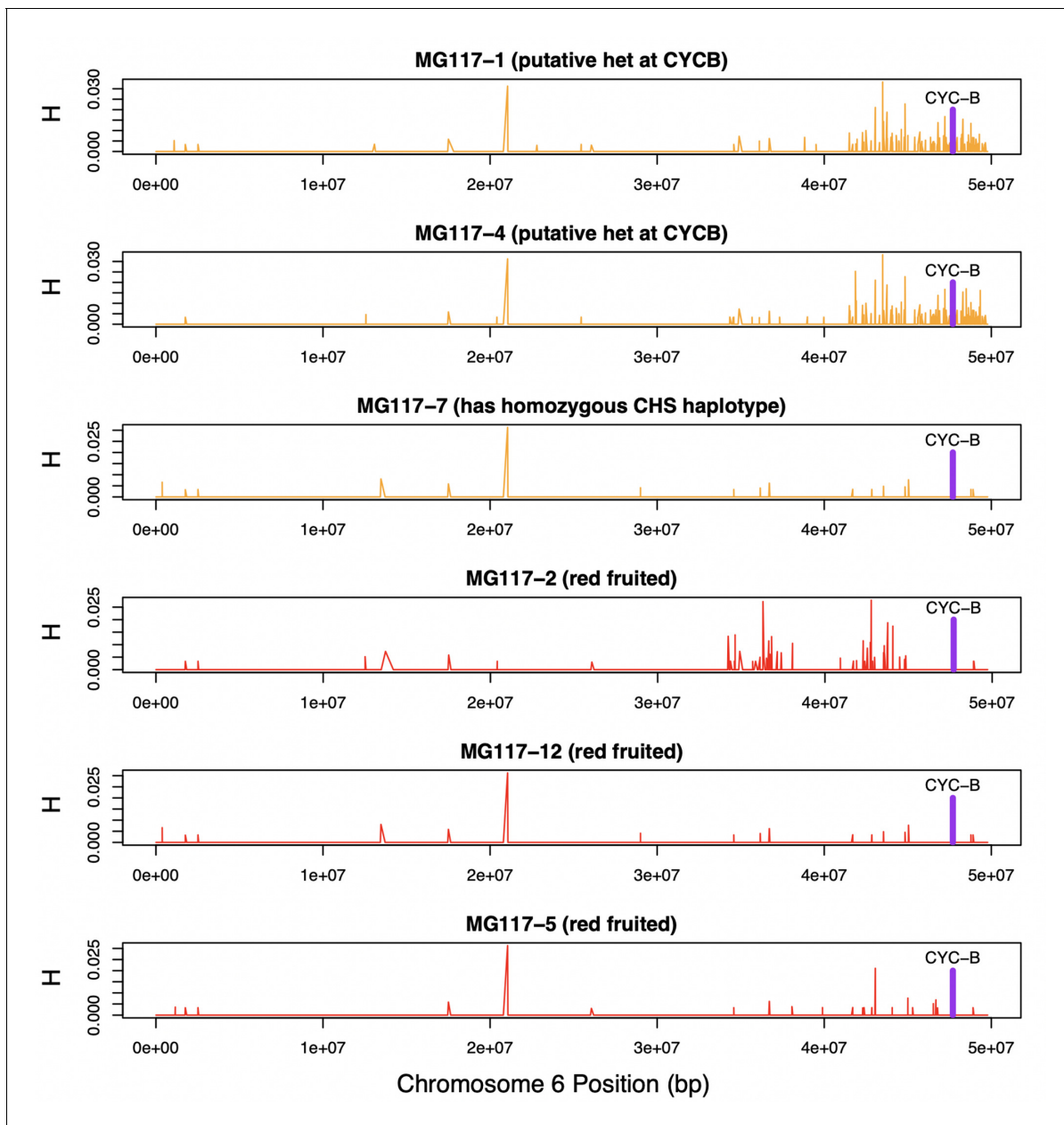


Figure 5—figure supplement 1. Heterozygosity along chromosome 6 of population MG117. CHS ancestry at CYC-B (dashed line) in MG117-1 and MG117-4 is heterozygous, as shown by elevated heterozygosity estimates at that location in these individuals.

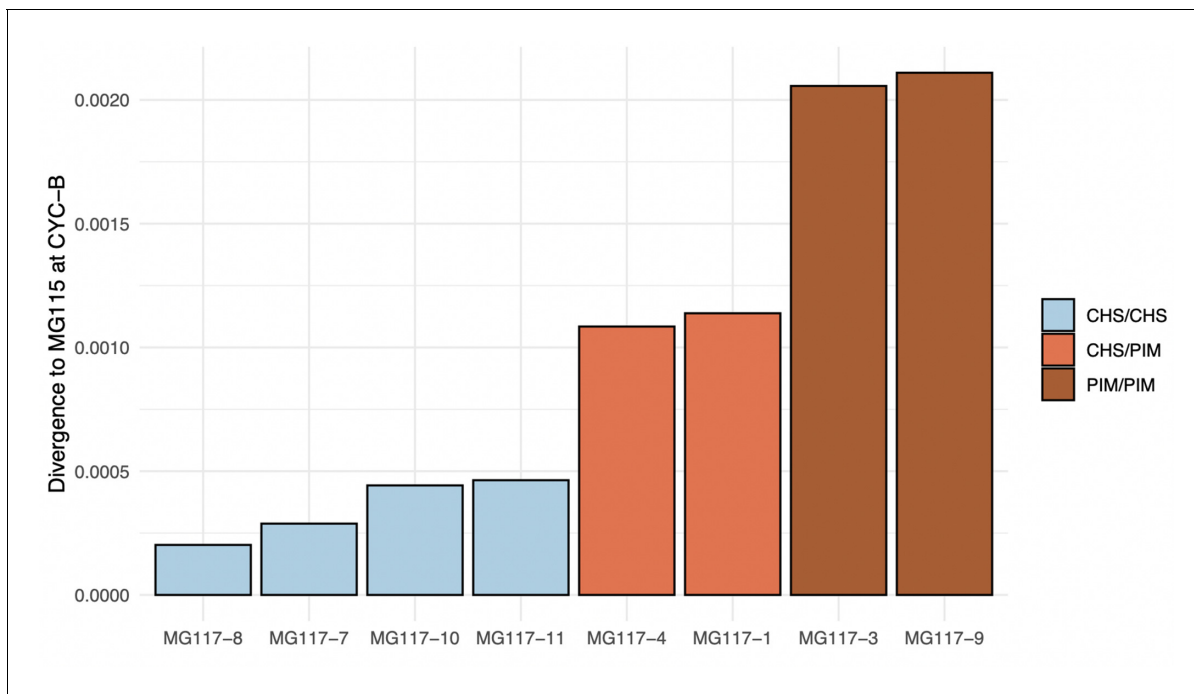
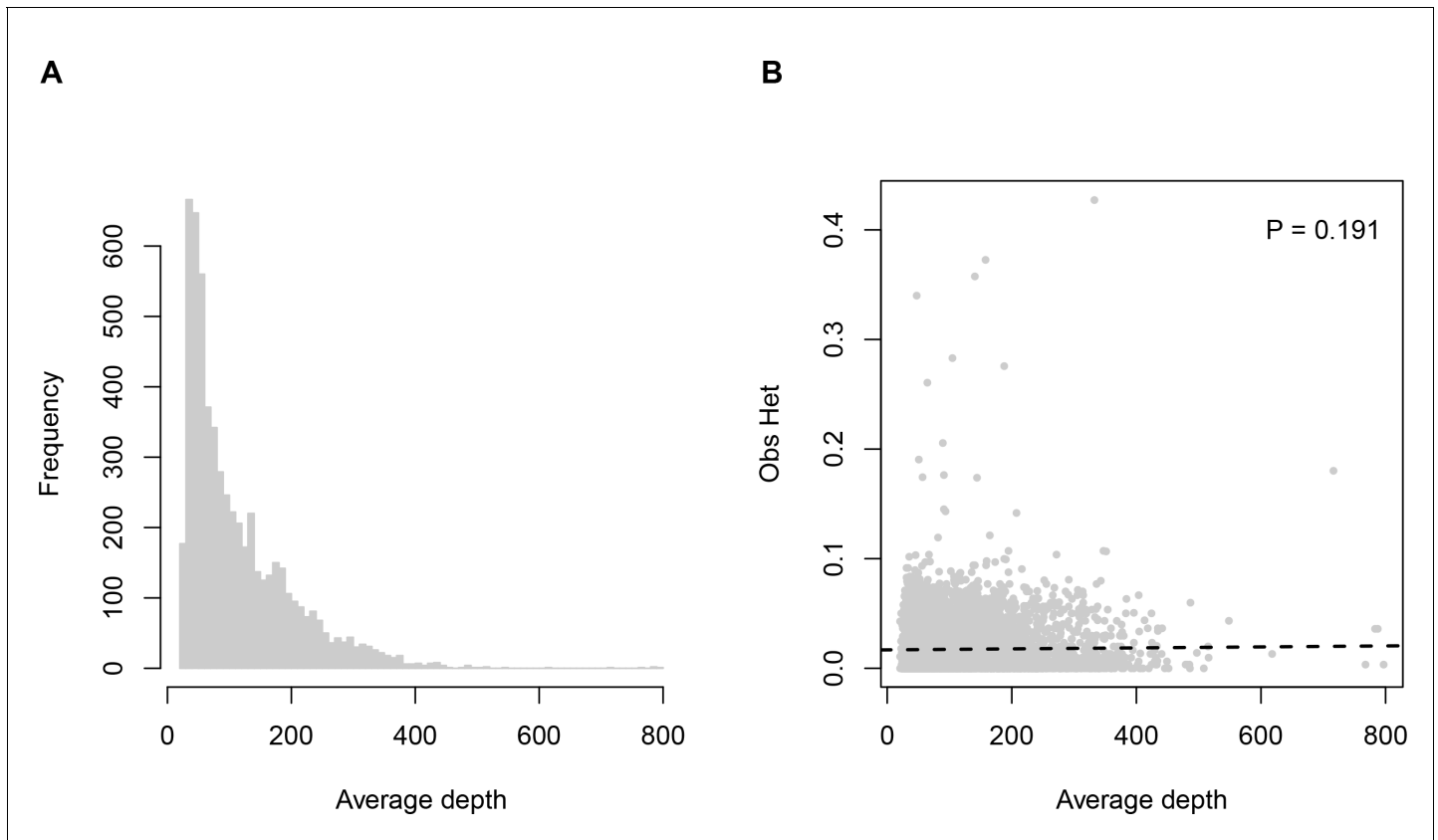


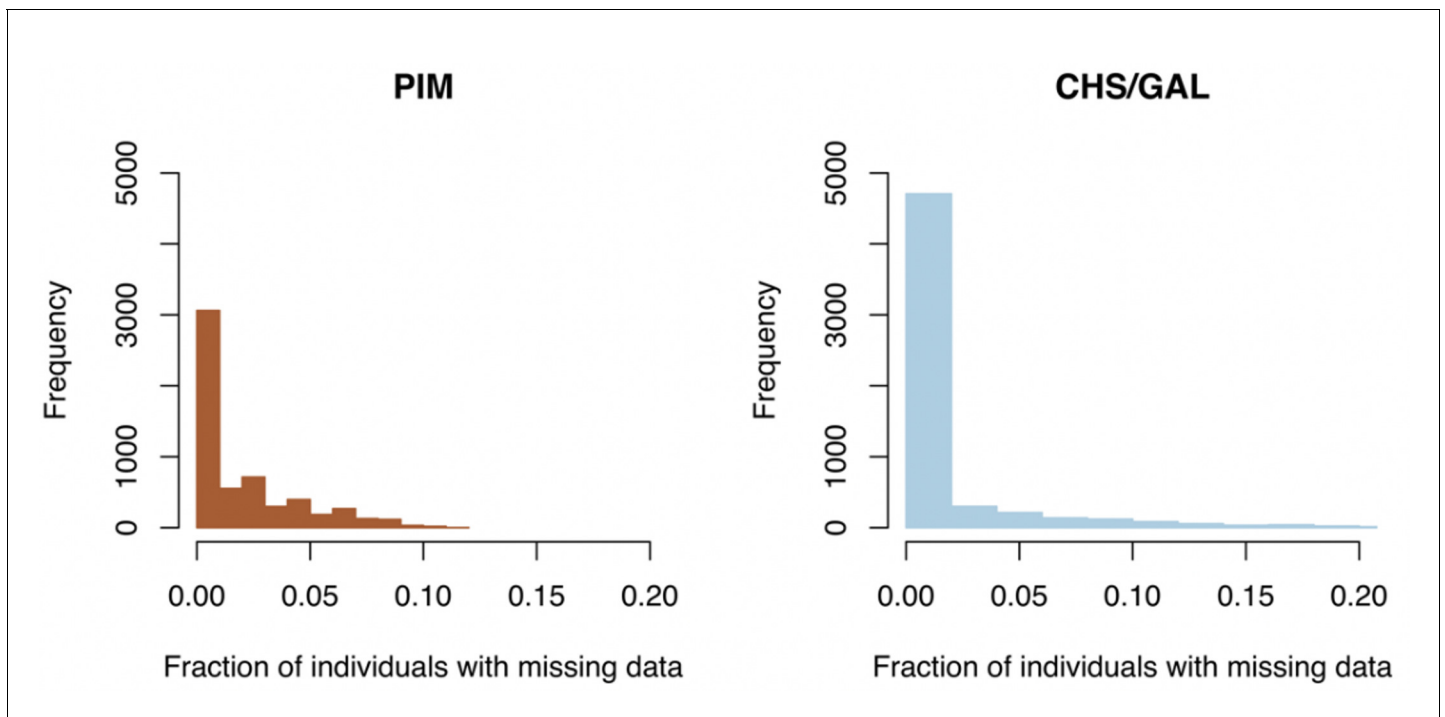
Figure 5—figure supplement 2. Median divergence estimates between MG117 individuals and MG115 at CYC-B. The hidden Markov model (HMM) correctly classifies intermediately diverged regions as heterozygous.

	1	10	20	30	40	50	60
LA1777	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA3778	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA0716	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA4117	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA2172	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA1316	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA3475	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA1589	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA3909	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA0429	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA3124	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
LA0436	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
consensus>70	MSVALLWVSPCDVSN	GTSMESVREGN	RF	FDS	SRHRLVSN	ERINRGGGKQT	NNGRKFS
	70	80	90	100	110	120	
LA1777	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA3778	VRSAILATPSX	ERTMTSEQMVYD	XVIRQAALVKRQLRSTNE	XEVKKPDIP	XIPGNLGLLSEA		
LA0716	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA4117	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA2172	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA1316	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA3475	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA1589	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA3909	VRSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA0429	VWSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA3124	VWSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
LA0436	VWSAILATPSG	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
consensus>70	V.SAILATPSg	ERTMTSEQMVYD	VVIRQAALVKRQLRSTNE	LEVKKPDIP	IPGNLGLLSEA		
	130	140	150	160	170	180	
LA1777	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA3778	YXRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA0716	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA4117	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA2172	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA1316	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA3475	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA1589	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA3909	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA0429	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA3124	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
LA0436	YDRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
consensus>70	YdRCGEVC	AEYAKTFNL	GMTLMTPE	RRRAIWA	IYVWCRR	TDELVDG	PNASYITPAALDRW
	190	200	210	220	230	240	
LA1777	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA3778	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA0716	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA4117	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA2172	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA1316	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA3475	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA1589	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA3909	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA0429	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA3124	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
LA0436	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
consensus>70	ENRLEDVF	NGRPFDM	LDGALSD	TVSNFP	VVDIQP	FRDMIEG	MRMDLRKSRYKNFDELYLYC
	250	260	270	280	290	300	
LA1777	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA3778	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA0716	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA4117	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA2172	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA1316	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA3475	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA1589	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA3909	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA0429	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA3124	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
LA0436	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP
consensus>70	YYVAGTV	GLMSVPI	MGIAPE	SKATTES	VYNAAL	ALGIAN	QLTNILRDVGEDARRGRVYLP

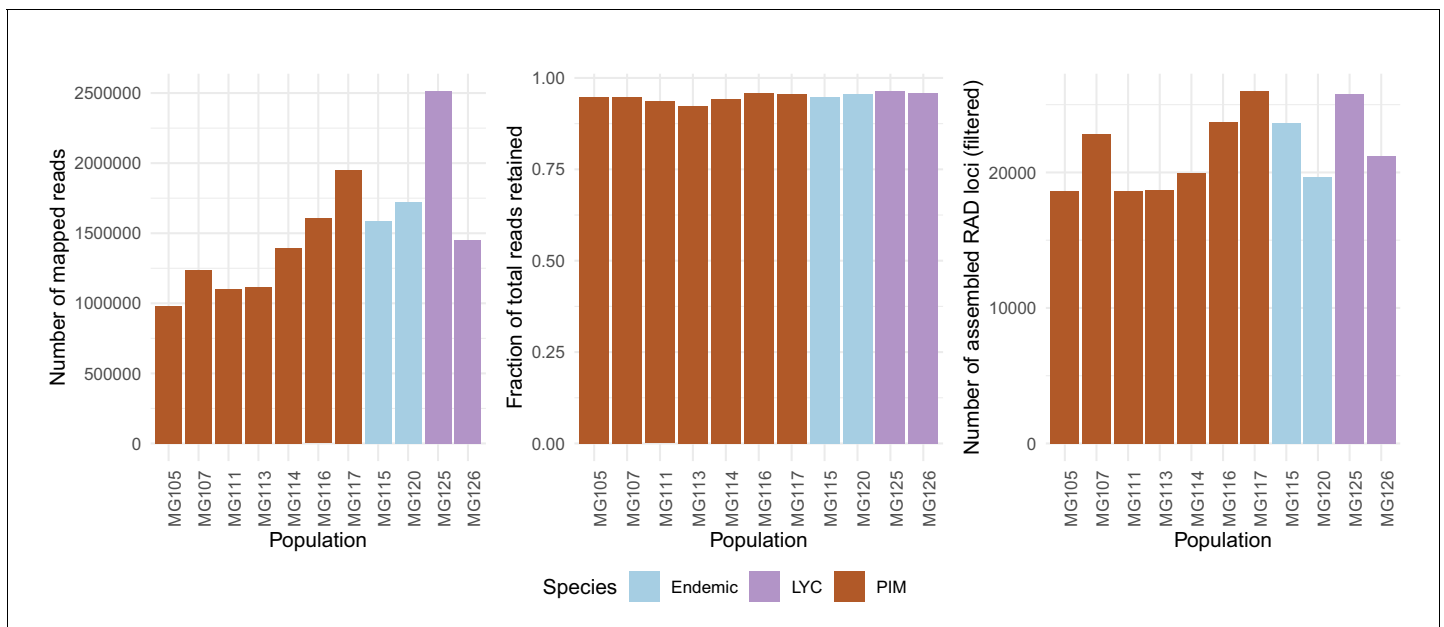
Figure 5—figure supplement 3. Coding sequence alignment of *PSY1* for nine wild tomato species (12 accessions). Endemic accessions are colored orange. The nonsynonymous substitution which defines the endemic clade is indicated with an orange arrow. Data were obtained from **Pease et al., 2016**.



Appendix 1—figure 1. Sequencing depth (average number of reads per individual) across loci. Panel (A) histogram of depth values. Panel (B) relationship between sequencing depth and observed heterozygosity.



Appendix 1—figure 2. Clade-specific distributions of missing data fractions across loci. Left panel: histogram of missing data fractions for all PIM populations. Right panel: histogram of missing data fractions for all endemic (CHS/GAL) populations.



Appendix 1—figure 3. Population-specific estimates of the total number of assembled RAD loci (left panel), fractions of total reads retained after assembly (center panel), and total number of reads mapped with BWA (right panel). The total number of mapped reads was on average lower in PIM, yet the average fractions of mapped reads assembled into loci (or ‘stacks’) and the total number of final RAD loci were roughly equal across populations and species. The differences in numbers of mapped reads point to substantial variation in total sequencing output rather than an inability to map in PIM.