

Supplementary Text S2

The Hidden Markov Model for assembly painting

We here describe the one-order Hidden Markov Model (HMM) we developed to model the *Drosophila suzukii* assembly WT3-2.0 as a mosaic of chromosomal segments from either Chinese ("C") or Hawaiian ("H") origins.

Let s_i represent the (local) ancestral origin ($s_i = C$ or $s_i = H$) of the allele in the reference assembly at SNP position i and d_i the physical distance (in bp) between SNP position $i+1$ and i . The four different transition probabilities between the (hidden) ancestral origin s_i and s_{i+1} of two adjacent SNPs then read:

$$\begin{cases} \mathbb{P}[s_{i+1} = C \mid s_i = C] &= e^{-\rho d_i} + (1 - e^{-\rho d_i})\alpha \\ \mathbb{P}[s_{i+1} = H \mid s_i = C] &= (1 - e^{-\rho d_i})(1 - \alpha) \\ \mathbb{P}[s_{i+1} = C \mid s_i = H] &= (1 - e^{-\rho d_i})\alpha \\ \mathbb{P}[s_{i+1} = H \mid s_i = H] &= e^{-\rho d_i} + (1 - e^{-\rho d_i})(1 - \alpha) \end{cases} \quad (1)$$

The parameter ρ corresponds to the per bp ancestry switching rate along the sequence assembly and the parameter α to the proportion of the genome assembly of Chinese ("C") origin. Hence, the term $(1 - e^{-\rho d_i})$ is the probability that a change of ancestry occurred between SNP position i and $i+1$, and in that case the reference allele at position $i+1$ is of Chinese (respectively Hawaiian) origin with probability α (respectively $1 - \alpha$). It might be tempting to interpret ρ as a historical recombination rate with $\rho = \rho_0(t_a + t_i)$ where ρ_0 is the recombination rate per bp and generation; t_a is the time (in generations) separating the WT3 founding population (Watsonville, USA) from the admixture between Chinese and Hawaiian populations; and t_i is the number of generations of inbreeding to create the sequenced WT3 strain. However, because inbreeding in the isofemale strains remains incomplete in several genomic regions, the sequence assembly cannot strictly be viewed as a single haplotype and may rather correspond locally to a consensus of different chromosomal segments (i.e., *in silico* recombinant). As a result, the estimated ρ is expected to provide an upwardly biased estimate of the historical recombination rate. It should also be noticed that the overall length of tracts of alleles from Chinese (resp. Hawaiian) ancestry has an expected mean equal to $\frac{1}{(1-\alpha)\rho}$ (resp. $\frac{1}{\alpha\rho}$) as consecutive segments in the sequence assembly might belong to the same ancestry.

To complete the specification of the HMM, the emission probability for the observed reference allele a_i at SNP position i given the ancestral origin $s_i = j$ was set equal to its frequency in the corresponding ancestry j (i.e., $j = C$ or $j = H$) as estimated from the

Pool-Seq data, i.e. $\mathbb{P}[a_i | s_i = j] = \widehat{f}_{ij}$. The estimated frequencies \widehat{f}_{ij} of the reference allele (i.e., the one of the assembly) were computed using a Laplace estimator as: $\widehat{f}_{ij} = \frac{r_{ij} + \epsilon}{c_{ij} + \epsilon}$ where r_{ij} and c_{ij} represent the observed read count for the reference allele and overall coverage for SNP i in population j . The parameter ϵ controls the prior Beta distribution for allele frequency (namely $\beta(\epsilon, \epsilon)$ which is the expected distribution at the mutation-drift equilibrium for a constant size population with $\theta = 4N_e\mu = \epsilon$) and was set to 0.01. Note that the main rationale for using such an estimator was to avoid fixed allele frequencies estimates, i.e., equal to 0 or 1 (the minimal possible minor allele frequency being equal to 0.005) which would lead to absorbing states in the painting model. In other words, an allele unobserved in the Pool-Seq reads was considered as extremely rare rather than completely absent from the corresponding population.

Given a set of parameter values, we used the forward-backward algorithm (Rabiner, 1989) to compute the model likelihood (i.e., the conditional probability of the sequence of reference alleles at each and every SNP positions) and the probability to belong to the two ancestries at each SNP position by integrating over all possible sequences of ancestries. Note that the contigs were randomly ordered and the distances between markers belonging to different contigs was set to an arbitrary large value equal to 10^{10} . To estimate the parameter values that maximized the likelihood, we then used the default Nelder and Mead method as implemented in the `optim` function from the R stats package (R Core Team, 2013). In order to work with unconstrained parameters in the optimization procedure, the model was reparameterized with the following two new parameters, $\rho^* = \log(\rho)$ and $\alpha^* = \log(1 - \alpha) - \log(\alpha)$. Standard errors of the resulting Maximum Likelihood Estimates (MLEs) for the two HMM parameters were obtained from the approximate Hessian of minus the log-likelihood (Zucchini & MacDonald, 2009, , p.53).

We further computed estimates for each contig k , with a number of analyzed SNPs $n_k > 2$, both the estimated proportion of Chinese origin $\widehat{\alpha}_k$ and the ancestry change rate $\widehat{\rho}_k$ based on the (posterior) expected number of chunks from Chinese ($\widehat{\eta}_k^{(C)}$) or Hawaiian ($\widehat{\eta}_k^{(H)}$) origin and their corresponding overall expected length ($\widehat{\lambda}_k^{(C)}$ and $\widehat{\lambda}_k^{(H)}$ respectively) as:

$$\widehat{\alpha}_k = \frac{\widehat{\lambda}_k^{(C)}}{\widehat{\lambda}_k^{(C)} + \widehat{\lambda}_k^{(H)}} \quad \text{and} \quad \widehat{\rho}_k = \frac{\widehat{\eta}_k^{(C)} + \widehat{\eta}_k^{(H)}}{\widehat{\lambda}_k^{(C)} + \widehat{\lambda}_k^{(H)}} \quad (2)$$

More precisely, following Lawson *et al* (2012) (eqs. 3 and 4 in Text S1), $\widehat{\lambda}_k^{(C)}$ and $\widehat{\lambda}_k^{(H)}$ were computed as:

$$\widehat{\lambda}_k^{(C)} = \frac{1}{2} \sum_{j=2}^{n_k} (\widehat{p}_{k,j-1} + \widehat{p}_{k,j}) d_{j-1} \quad \text{and} \quad \widehat{\lambda}_k^{(H)} = \frac{1}{2} \sum_{j=2}^{n_k} (2 - \widehat{p}_{k,j-1} + \widehat{p}_{k,j}) d_{j-1} \quad (3)$$

and $\widehat{\eta}_k^{(C)}$ and $\widehat{\eta}_k^{(H)}$ were computed as:

$$\widehat{\eta}_k^{(C)} = \widehat{p}_{k,1} + \sum_{j=2}^{n_k} \widehat{\tau}_j^{(C)} \quad \text{and} \quad \widehat{\eta}_k^{(H)} = 1 - \widehat{p}_{k,1} + \sum_{j=2}^{n_k} \widehat{\tau}_j^{(H)} \quad (4)$$

where, for the j th SNP position of contig k , $\widehat{p}_{k,j}$ is the posterior probability that the reference allele on the assembly is of Chinese origin, and $\widehat{\tau}_j^{(C)}$ (respectively $\widehat{\tau}_j^{(H)}$) represents the posterior probability that the reference assembly allele is of Chinese (resp. Hawaiian) origin given at least one change of ancestry occurred between SNP position $j-1$ and j . The $\widehat{p}_{k,j}$'s, $\widehat{\tau}_j^{(C)}$'s and $\widehat{\tau}_j^{(H)}$'s posterior probabilities were all estimated using the forward-backward algorithm (Rabiner, 1989) setting the genome-wide parameters ρ and α to their MLE values.

References

- Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genet.*, **8**, e1002453.
- R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.
- Zucchini W, MacDonald I (2009) vol. 110 of *Monographs on Statistics and Applied Probability*. CRC Press.