Mutation rate analysis via parent—progeny sequencing of the perennial peach. II. No evidence for recombination-associated mutation

Long Wang¹,†, Yanchun Zhang¹,†, Chao Qin¹, Dacheng Tian¹, Sihai Yang¹ and Laurence D. Hurst²

1. Introduction

Both mutation rates and recombination rates vary between species and between regions within a genome [1,2]. In the accompanying paper, we ask, via parent—progeny sequencing of the peach, whether woody perennials might have low mutation rates [3–5] compared with fast-growing annuals and whether hybrid strains have higher mutation rates [6]. Here, employing the same data, we focus on the possibilities that recombination might be mutagenic [7,8] and whether the recombination rate of domesticated peach is higher than that of wild peach, there commonly being a suggestion that domestication is associated with raised recombination rates [9–11]. If both are true then some variation between genomic regions and between strains in the mutation rate may be attributable to recombination-associated mutation.

The idea that recombination, or meiosis more generally, might be mutagenic stems from the work of Magni [7,8] in which he observed a higher mutation rate in meiotic than mitotic yeasts. From a mechanistic view, a correlation could be expected between mutations raised from double-strand break (DSB)-repairing errors and those DSBs occurring in homologous recombination [12]. If recombination is mutagenic, then we expect domains of high recombination to be domains of high rates of new mutations. The hypothesis has proven highly controversial, with indirect evidence both consistent [13–16] and inconsistent [17–20] with the hypothesis. The best indirect data, however, argue against
the possibility. Notably, in 1000 Genomes data there is no increased variation around recombinogenic hotspot motifs [21]. Moreover, evidence from a correlation between recombination and the rate of putatively neutral evolution [13,14,16], now appears to be better explained as a consequence of biased gene conversion [22]. While then, until recently, convincing direct evidence for recombination being mutagenic has been lacking (for review, see [23]), even more recent direct evidence in humans [24], yeast [25] and bees [1] supports the hypothesis that recombination is mutagenic, although the effect might be very weak. Were recombination mutagenic, we might also predict that species with higher recombination rates should have a higher mutational input. However, higher divergence might in turn lead to reduced recombination rates [20] making prediction harder. There are numerous alternative suggested determinants of intragenomic variation in the mutation rate: for example, it correlates with local sequence context [26], including presence of insertion/deletions (indels) [27], replication timing [28], as well as possibly epigenetic effects such as chromatin organization [29].

The parent–progeny sequencing data that enables us to estimate the mutation rate, also enables us to determine the recombinational landscape of peach. Domesticated species are conjectured to have been indirectly selected for high recombination rates [9–11]. This is because directional selection owing to domestication, might select for modifier alleles that increase the recombination rate; either because drift permits build-up of linkage disequilibrium (especially in smaller populations) or because epistatic effects generate linkage disequilibria among selected loci [11]. Evidence of increased recombination in domesticated plant species, based on the analysis of chiasmata number, is supportive of such a link [30]. A correlation between domestication and high recombination rate could be owing to high recombination prior to domestication, as a form of preadaptation to domestication [31], but current evidence argues against this [30]. However, more recent sequence data-based estimates of recombination rates in mammals contradict the domestication–recombination hypothesis [32]. It is unclear whether this difference in results between analyses reflects a taxonomic (plant–mammal) or methodological (chiasmata counts versus direct recombination inference) difference. Here then we ask whether domesticated peach has a higher recombination rate than a wild close relative and whether mutations occur more often near recombination break points.

2. Material and methods

We constructed three parent–progeny groups (groups I–III). Each group has an F1 parent tree together with its selfed F2 progeny. Groups I and II are low heterozygosity intraspecific crosses employing young (group I) and old (group II) F1s, while group III F1 is an interspecific cross. Group I included one F1 (Prunus persica) and 24 selfed F2 samples (14F2-1 to 24 in the electronic supplementary material, table S1). Group II included one weakly heterozygous F1 (Prunus mima, a wild peach) and nine selfed F2 samples (CZTH-S1 to –S5, –S7 to –S9 and CZTH-5). The interspecific crossing group (group III) included four ancestral parents, one heterozygous F1 (Prunus davidiana × P. persica) and 30 F1 selfed F2 samples (NE1–NE30 in the electronic supplementary material, table S1). In total, 70 peach samples, including four ancestral parents from group I, three F1 parents (i.e. each group with one F1 sample) and 63 F2s were selected for whole-genome resequencing. This was done with high sequence quality (base quality Q20 ≥ 95%), high depth (51.3 × on average and ranging from 38.3 × to 65.8 ×) and relatively long reads (150 bp × 2, paired end sequencing strategy by HiSeq4000 platform; electronic supplementary material, table S1).

For further methods pertinent to sampling, sequencing and alignment, variant calling, de novo mutation identification, Sanger validation of mutation calls, estimation of mutation rate and estimation of heterozygosity, we refer the reader to the prior paper. A full methodology pertinent for both papers is also presented as the electronic supplementary material.

(a) Variant calling and marker identification

Raw variants for each sample were called using GATK Haplotype Caller (HC) in GVCF mode [33]. For recombination analysis, markers with low confidence could hamper the identification of true recombinant blocks; therefore, it is important to exclude false variant calls as thoroughly as possible. To generate a high-confidence variant set, we only use bi-allelic variant loci with: (i) quality greater than or equal to 50; (ii) a depth no less than 10 and not exceeding 80; and (iii) more than half of samples contain informative calls in each group. To reduce the genotyping errors, we also required a reference allelic ratio of 0–5% or 95–100% to be considered as a confident homozygote, while 30–70% was required to make a confident heterozygous call. A confident marker was thus identified where the F1 samples were present in a confident heterozygous status. This allele-balance filter is efficient for removing genotyping errors owing to sequencing errors or possible contaminates, as those errors were most likely at a low frequency. However, mapping errors owing to highly similar paralogous sequences could also result in pseudo-heterozygosity. To minimize these errors, we remove those markers residing in large structural variant (SV) regions of F1 samples compared with the reference genome in each group. The SVs were detected by combining three different algorithms: a read-depth approach (CNVNator) [34], a split-read approach (PINDEL) [35] and from the analysis of discordant pairs (BREAKDANCER) [36]. CNVNator (v. 0.3) was run with a bin size of 100 bp, which predicts large deletions and duplications. PINDEL (v. 0.2.5b6) was run with default options. Results were collected for large deletions (greater than or equal to 100 bp), inversions and translocations. Deletion, duplication and inversion results were also collected from BREAKDANCER (v. 1.1.2) with default settings. We generated a union set of results collected from all three approaches without further filtering. SVs with a size smaller than 100 kbp were directly used. We also include 200 bp flanking regions of all inversion events. For SVs larger than 100 kbp, we use the 400 bp flanking regions around each predicted SV breakpoint.

(b) Detection of crossover events

For interspecific F2 samples, we first genotyped each marker as P. persica-homozygous, P. davidiana-homozygous or heterozygous, by comparing against these parents. The markers were then clustered using a ‘seeding and extension’ approach to form the original inherited blocks. First, fragments with 25 consecutive markers of the same genotype and a length over 10 kbp were chosen as a seed; adjacent seeds with same genotype were then merged into larger fragments (blocks) until all adjacent fragments were of different genotypes. Each block was further extended to the furthest marker of the same genotype where the overall proportion of this genotype started to decline. This algorithm has been implemented in the script vcf_process.pl and is available from https://github.com/wl13/BioScripts. Finally, all boundaries of blocks were manually inspected and revised. The location of crossover (CO) events was determined as the location where block genotype switched.
For the intraspecific *P. persica* group, it is difficult to first genotype each marker as neither of the parental individuals were available. Thus, we only genotyped those markers as homozygous or heterozygous at first, and formed the blocks using the same clustering method mentioned earlier. This rests on the assumption of there being only a negligible chance for two CO events to be observed in a very narrow region (i.e. within two adjacent markers) from a single F2 genome. This is reasonable as the two haplotypes of the same F2 genome came from independent meiotic processes. Once the initial blocks were formed, the F1 and other F2 chromosomes could then be phased according to those homozygous blocks (electronic supplementary material, figure S1a). For each chromosome, we picked out a sample in which only a homozygous genotype was observed. As the selected sample consists of two identical haplotypes (defined as ‘Haplotype1’), the F1 chromosome as well as other F2 chromosomes could thus be phased through comparison with this haplotype (electronic supplementary material, figure S1b). This process also relaxed the previous assumption and was robust to possible phaseing errors (electronic supplementary material, figure S1c). The final phased blocks were used to detect CO events as described before.

In order to make sure the stringent filtering steps did not remove many true variants and lead to an underestimation of CO events, we also identified inherited blocks and CO events before each filtering step was implemented. Through comparison of the CO events identified in those intermediate steps with the final results, we identified those filtered CO events that were always shared among many different individuals, which was not likely to happen in the randomly sampled F2 samples. Manual inspection of those regions also confirmed the non-proper mapping status and artefactual clustering of markers (standard error of distances between each two adjacent markers).

To test whether the CO rate was correlated with the intraspecific population diversity, 70 *P. persica* individuals were collected from published data [37]. All reads were mapped to the reference genome using BWA-backtrack algorithms [38], followed by marking of PCR duplicates (i.e. probably PCR amplification artefacts) and realignment processes as described before. Both variants and non-variant sites were called with HC in GVCF mode. Variant sites with more than half missing alleles or with a non-reference allele frequency of less than 7 (e.g. 5% of all 70 diploid individuals) were excluded to reduce false positive calls.

The population diversity was calculated as the average pairwise differences among all possible pairs. The pairwise difference was defined as the per site nucleotide difference between each of the two compared individuals, e.g. 1 would be counted for a difference between two different homozygous genotypes while 0.5 would be counted for a difference between a homozygous genotype and a heterozygous genotype. The pairwise differences were obtained by first summing up all nucleotide differences in a window, then dividing by the number of informative sites (sites genotyped in both individuals) in the same window. For each pair, only windows with more than 50% informative sites were considered as an informative pair in this window. Windows with less than 1208 informative pairs (e.g. 50% of all total 2415 pairs) were discarded from the correlation test. Statistics and correlation tests were performed in R [39].

### 3. Results

(a) Identification of accurate markers in each parent—progeny peach group

To ensure the accuracy of the called markers used in recombination analysis in each parent—progeny group, several strategies were employed (see Material and methods for details). In total, 302,164, 132,572 and 1,110,854 reliable single nucleotide polymorphisms, as well as 37,856, 21,426 and 115,874 single insertion/deletion (indel) markers were called for groups I, II and III, respectively. This corresponds to an average of 1.51, 0.68 and 5.44 variant markers per kilo base pair for groups I, II and III, respectively. These markers were used to identify the genotypes of heterozygous or homozygous regions in these F2 genomes. In our three parent—progeny groups, the average nucleotide diversity (number of nucleotide differences per site) were approximately 0.29%, 0.27% and 1.24% at the whole-genome level between the two haplotypes derived from a single F1 in group I, II and III, respectively. As expected, an approximate 4.4-fold higher diversity was detected in the interspecific crossing group compared with the intraspecific groups.

(b) The recombination rate is consistent with low rates in woody perennials

Before addressing the question of whether peach has high recombination rates compared with wild relatives and whether mutation and recombination are coupled, we first sought to determine aspects of the basic biology of recombination in peach. For example, for benchmarking, we ask whether our rate estimation is consistent with prior estimates [40, 41] and with the suggestion that woody perennials have overall low rates [5, 42, 43].

(c) Statistical analysis

The CO coldspot and hotspot regions were detected by first dividing the whole genome in non-overlapping 500 kbp windows. Midpoints of CO breaks were used as the location of CO events and were counted for each window. Windows with similar CO numbers were merged. All windows after merging were tested using a Monte Carlo process, with 10,000 randomizations of shuffling all CO events across the whole genome to derive the p-values. Regions with observed CO events significantly deviating (p < 0.05) from the expectation of randomizations were defined as hot-spot regions (more than expectation) or coldspot regions (less than expectation), respectively.

To test whether the CO rate was correlated with the mutation rate, we binned the genome into 500 kbp, 1 Mbp, 2 Mbp and 5 Mbp domains. CO events and mutations were collected from both intraspecific *P. persica* group I and interspecific group III.
To identify CO events, we searched for the genotype switching point, e.g. from homozygosity to homozygosity or from homozygosity to heterozygosity, along the chromosome pairs in each F2 genome [44,45] (see Material and methods for details). A total of 286 COs were detected in 24 F2 samples from intraspecific group I, corresponding to 11.92 COs on average or 2.64 cM Mbp

The CO rate of 2.6 cM Mbp

(c) Larger chromosomes have fewer recombination events per base pair

Among all eight chromosomes, chromosome 5 had the highest CO rate, whereas chromosome 6 had the lowest CO rate (table 1). At least in some taxa, CO rates scale inversely with chromosome size [47,48]. Consistent with this observation, a significant negative correlation was obtained between chromosome physical length and the CO rate per mega base pair (Spearman’s $r = -0.857, p = 0.0107$; electronic supplementary material, figure S2). Unlike some species whose number of CO events per unit physical distance is approximately a constant [44], no positive correlation between chromosome physical length and number of CO events per chromosome was detected (Spearman’s $r = 0.286, p = 0.501$).

(d) Recombination profile is repeatable

Is the profile of recombination rate variation specific to a particular cross or repeatable between crosses? To address this we compare the variation in the recombination rate between the intra- and interspecifics (I and III, respectively). Despite the fact that CO number and rate varied across each chromosone, they were well correlated (Spearman’s $r = 0.952, p = 0.001$) between intra- and interspecific groups (table 1; electronic supplementary material, figure S3). While the above trend reflects a between-chromosome correlation, the trend remains even if we use a small bin size of 500 kbp along each chromosome (Spearman’s $r = 0.286, p = 0.473$). The CO events in peach were unevenly distributed on the chromosomes. The CO rate varied between 0 and 4.53 cM Mbp

(e) Peach has hotspots and coldspots of recombination

The CO events in peach were unevenly distributed on the chromosomes. The CO rate varied between 0 and 16.67 cM Mbp

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{Distribution of de novo mutations and CO events on the chromosomes. The vertical red lines show the mutation distributions of intraspecific samples (P. persica and P. mira) along each chromosome, while the blue lines reflect interspecific samples (P. persica × P. davidiana). The plots above show the number of CO events within intra- (P. persica) and interspecific samples counted in non-overlapping 500 kbp windows. The number of COs in each window was well conserved between intra (P. persica) and interspecific samples (Spearman’s $r = 0.150, p = 0.00137$), therefore both events were counted together to generate the overall distributions. The left vertical bars show the CO numbers. CO hotspots are marked as cyan, while coldspots are marked as sky blue.}
\end{figure}
material, table S5) with a combined span of approximately 19-Mbp and 14 CO coldspots (10,000 randomizations, \(p < 0.05\); electronic supplementary material, table S6), with a combined length of 53.8 Mbp. In other words, approximately 29% of CO events are clustered within approximately 8.6% of the entire genome (electronic supplementary material, table S5), and approximately 23.9% of the genome is devoid of the CO events (electronic supplementary material, table S6). The average recombination rate in hotspot regions (8.04 cM Mbp\(^{-1}\)) is about 16.8-fold higher than that in the cold-spot regions \((0.48 \text{ cM Mbp}^{-1}\); \(t\)-test, \(p = 1.58 \times 10^{-17}\)). Gene ontology analysis reveals a slight enrichment in serine-type endopeptidase activity under molecular function near hotspots (electronic supplementary material, figure S4), while coldspots are enriched for cysteine-type peptidase activity or other various binding activities, and most genes were related to the macromolecule metabolic process (electronic supplementary material, figure S5).

In contrast to prior observations in the peach genome paper [40], we observed suppression of CO in peri-centromeric regions. Among all 14 CO cold regions detected, eight were found to overlap with the putative peri-centromeric regions of all eight chromosomes (electronic supplementary material, table S6).

(f) No evidence for higher recombination rates in domesticated peach compared with wild relatives

If domestication leads to increased recombination rates, we expect that the intraspecific cross of the domesticated peach (group I) to have a higher recombination rate than an intraspecific cross employing wild peach (group II). A conservative estimation method (see Material and methods for details), predicts an average of 3.18 cM Mbp\(^{-1}\) CO rate in wild peach (group II). Importantly, this is higher, not lower, than its domesticated relative \(P. \text{persica}\) (2.64 cM Mbp\(^{-1}\)). The CO rate (3.02 cM Mbp\(^{-1}\)) of a cross between peach and \(P. \text{ferganensis}\), another wild undomesticated peach (virtually undistinguishable from \(P. \text{persica}\) at molecular level), is also higher [40].

One cross has a lower CO rate than the domesticated cross (group I), this being the interspecific cross (group III). A total of 284 COs were detected in 30 interspecific F\(_2\) samples (table 1), corresponding to 9.47 COs on average or
2.10 cM Mbp$^{-1}$ per meiosis per sample, which is significantly lower than that (2.61 cM Mbp$^{-1}$) in the intraspecific samples (Brunner Munzel test, $p = 0.04$), and also is lower than the previous estimation (3.02 cM Mbp$^{-1}$) in the interspecific peach map of P. persica × P. ferganensis BC$_1$ (P × F) [40]. The recombination reduction in interspecifics is seen in all eight chromosomes (table 1). The suppression of recombination could have resulted from decreased DNA mismatch repair activity between two diverged haplotypes [49]. Given the possibility of recombination suppression owing to the nature of the cross, we suggest that it is inappropriate to consider the group III–group I comparison when considering the domestication–recombination hypothesis.

(g) No evidence for a correlation between recombination and mutation

While we find no increased recombination in domesticated peach, it remains interesting to ask whether recombination and mutation are coupled. Despite the abundant intragenomic variation in recombination rate, we observe no significant relationship, regardless of the bin size, between CO rate and mutation rate (500 kbp bin: Spearman’s $r = 0.0231$, $p = 0.636$; 1 Mbp bin: $r = 0.461$, $p = 0.505$; 2 Mbp bin: $r = 0.107$, $p = 0.275$; 5 Mbp: $r = 0.00317$, $p = 0.984$). This mode of analysis however, may well be too crude if recombination-induced mutations are rare. Prior evidence looked for an excess of mutations within 2 kbp of recombination breakpoints [24]. In peach, however, no mutation was observed near the break points, even allowing for a more generous definition of proximity (less than 10 kbp). The nearest mutation was about 12 kbp, and only four mutations were found within 100 kbp (1 within 24 kbp and 2 within approx. 90 kbp). We conclude that we find no evidence for a coupling between mutation and recombination.

4. Discussion

Recent evidence, through sequencing in the vicinity of recombination break points, has found evidence that in humans [24], yeast [25] and bees [1] recombination may well be weakly mutagenic. That we failed to detect any coupling between recombination and mutation, suggests that any effect is modest at best or that peach may be unusual (perhaps domestication somehow affects this).

In many species, there is a correlation between heterozygosity and the recombination rate [50,51]. While this is classically considered a consequence of reduced Hill Robertson interference [52] in domains of high recombination, mutagenic recombination [1,24,25] is, at least in theory an alternative possibility [51,53]. In peach, we unusually do not observe a correlation between intraspecific diversity and recombination rate (500 kbp windows, $p = 0.98$, $r = -0.001$; 1 Mbp windows, $p = 0.32$, $r = 0.084$). It might then be tempting to speculate that an absence of this correlation might be coupled to the absence of mutagenic recombination and hence in those taxa with the correlation it could be owing to recombinogetic mutation. We caution against this interpretation. First, in the taxa in which recombination appears to be mutagenic the effect appears to be far too weak to explain the recombination–mutation correlation, although this will require quantitative modelling to confirm. Second, the absence of the heterozygozity–recombination correlation may have a simpler explanation, namely it is a result of domestication. Indeed, all the above results come with the caveat that peach, being a domesticated species, need not be representative and further analysis of different taxa is needed to judge the generalizability of any results.

We also fail to find evidence that domestication in this plant has led to increased recombination rates. This latter result inclines support to the view that the prior discrepancy (indirect estimation in plants supportive [30], direct evidence in mammals not supportive [32]) is owing to methodological limitations of indirect inference of recombination rather than a plant–mammal difference. One might alternatively conjecture that domestication of peach may somehow be atypical. With a sample size of one we do not wish to advocate strongly.

One notable result is the strong agreement on the local recombination rate observed between different crosses. This is not simply owing to stereotypical rates at telomeres and centromeres. This suggests that the recombinational profile of peach is relatively fixed. One might conjecture that this is as expected in a species lacking PRDM9, as hotspots defined by a mechanism dependent on PRDM9 tend to relocate over relatively short time spans, while non-PRDM9 ones do not [54–56].

**Table 1.** Number of COs along each chromosome.

<table>
<thead>
<tr>
<th>samples</th>
<th>Pp01</th>
<th>Pp02</th>
<th>Pp03</th>
<th>Pp04</th>
<th>Pp05</th>
<th>Pp06</th>
<th>Pp07</th>
<th>Pp08</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>intraspecific groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean COs</td>
<td>2.08</td>
<td>1.63</td>
<td>1.54</td>
<td>1.33</td>
<td>1.58</td>
<td>1.08</td>
<td>1.46</td>
<td>1.21</td>
<td>11.92</td>
</tr>
<tr>
<td>CO rate (cM Mbp$^{-1}$)</td>
<td>2.18</td>
<td>2.67</td>
<td>2.82</td>
<td>2.58</td>
<td>4.28</td>
<td>1.76</td>
<td>3.26</td>
<td>2.68</td>
<td>2.64</td>
</tr>
<tr>
<td>interspecific group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean COs</td>
<td>1.60</td>
<td>1.27</td>
<td>1.23</td>
<td>1.10</td>
<td>1.17</td>
<td>1.00</td>
<td>1.07</td>
<td>1.03</td>
<td>9.47</td>
</tr>
<tr>
<td>CO rate (cM Mbp$^{-1}$)</td>
<td>1.67</td>
<td>2.08</td>
<td>2.25</td>
<td>2.13</td>
<td>3.15</td>
<td>1.63</td>
<td>2.38</td>
<td>2.29</td>
<td>2.10</td>
</tr>
</tbody>
</table>

Data accessibility. The sequence data from this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession no. SRP071980.

Authors’ contributions. L.D.H., S.Y. and D.T. designed the project. L.W., L.D.H. and S.Y. wrote the paper. C.Q. and L.D.H. performed the Sanger verification. L.W., L.D.H. and S.Y. performed the experiments and analysed the data. C.Q. performed the Sanger verification. L.W., L.D.H. and S.Y. wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the National Natural Science Foundation of China (31671322, 31601041 and 91331205). L.D.H. is supported by ERC grant EvoGenMed, ERC-2014-ADG 669207 and BBSRC grants BB/N015630/1 and BB/M01035X/1.


