Lecture 10: Selection scans in humans and domesticated organisms

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

At present, genome-wide scans for genes under recent, or ongoing, selection have been performed on only a modest (but growing) number of species. For natural populations, the most extensive work has been done on humans, *Drosophila*, and *Arabidopsis thaliana*. Given that we know a great deal about the genetics, genomics, and molecular biology of these species, this choice is not surprising. All three groups have undergone major expansions into a wide range of new habitats over the last 100,000 years, and hence harbor the potential for a significant response to evolutionary challenges. For humans, the movement out of Africa into more temperate climates, coupled with the transition from hunting and gathering to agriculture and the resulting increase in population density, generated novel environmental pressures. The commensal *D. melanogaster* and *D. simulans* followed humans into these new environments, while in the northern hemisphere, *Arabidopsis* underwent significant range expansion following the end of the ice age. The environmental challenges faced by these species, as well as demographic changes (such as massive population expansions), leads us to expect a history of recent selection.

Take-home message empirical studies

- Little replication
  - Could be low power
  - Different test detect selection over different time scales
  - Heterogeneity in selection over populations
  - Low power for sweeps other than hard sweeps
  - Hard > soft > polygenic is the power order

An exciting finding is only the start

- No real way of validation
  - Resampling from the same population is not a true replication, as the signal could be simply due to a random, but unusual, coalescent structure in the target region
  - Ideally, show direct marker-fitness association
  - However, failure to do so is not damming
    - Low power (an *s* value of 0.005 leaves a big signal, but is hard to detect with ecological studies)
    - Signal could be correct, but due to previous selection pressures form an environmental change

Table 9.4 Overlap in sweep detection in three early scans (Carlson et al. 2005; Voight et al. 2006; Wang et al. 2006) that used different statistics to infer positive selection in humans. Diagonal elements represent the number of sites declared to be under positive selection in each given study, and off-diagonal elements represent the number shared between studies. See the text for further details. (After Biswas and Akey 2006.)

	Wang (*LDD*)	Voight (*iHS*)	Carlson (*D*)
Wang	1799	125	47
Voight		455	11
Carlson			176

With this last concern in mind, several scans have searched for geographically localized selection by contrasting F_{ST} values among samples of different populations (and hence allowing for population-specific selection). Barreiro et al. (2008) examined the F_{ST} values associated with roughly 3 million SNPs over four populations (Nigerians, Europeans, Chinese, and Japanese). They used a modification of the outlier approach, binning SNPs by functional categories (e.g., synonymous, nonsynonymous, 5′ UTR, etc.). They observed an excess of higher F_{ST} values (relative to the genome-wide distribution) in both nonsynonymous and 5′ UTR SNPs, suggesting that there were around 600 sites under local selection. Further, the excess nonsynonymous SNPs were enriched for long haplotypes, as might be expected under a partial sweep. Pickrell et al. (2009) also found evidence of significant local adaptation (population-specific changes) in a survey of 53 populations, although Hofer et al. (2009) noted that the striking differences in allele frequencies between human populations could have easily arisen as a consequence of population expansion (and the accompanying allelic surfing). One additional concern with these studies is that (as mentioned earlier) F_{ST} values are constrained by the level of heterozygosity (which is influenced by background selection), with SNPs with higher minor-allele frequencies having higher maximal F_{ST} values.

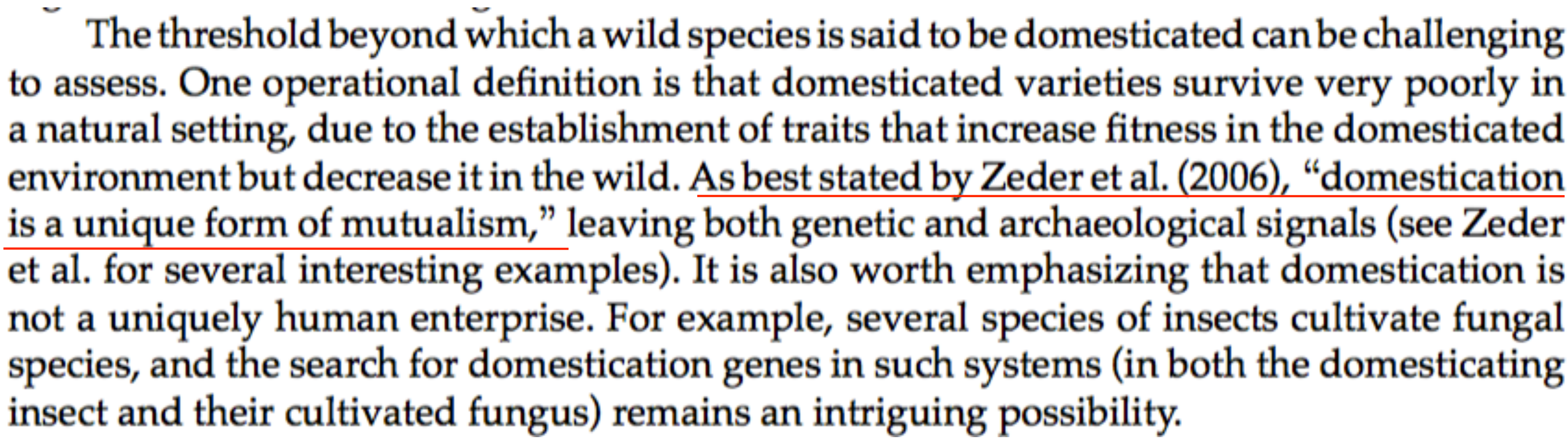
Tempering these results was the declaration by some researchers that classic hard sweeps appear to be rare in humans (Hernandez et al. 2011; Lohmueller et al. 2011; Alves et al. 2012), or "have played a moderate, albeit significant, role" (Fagny et al. 2014). However, as discussed in Chapter 8, Enard et al. (2014) noted that a failure to account for background selection (BGS) can result in a distorted view of the importance of sweeps. After adjusting for this effect, they detected widespread signals for positive selection in humans, which were more correlated with regulatory sequences than amino acid changes. Others have stressed the importance of polygenic sweeps (Hancock et al. 2010a, 2010b; Amato et al. 2011; Fumagalli et al. 2011; Turchin et al. 2012; Daub et al. 2013; Zhang et al. 2013; Berg and Coop 2014; Mathieson et al. 2015, Robinson et al. 2015; Field et al. 2016).

Balancing Selection in Humans

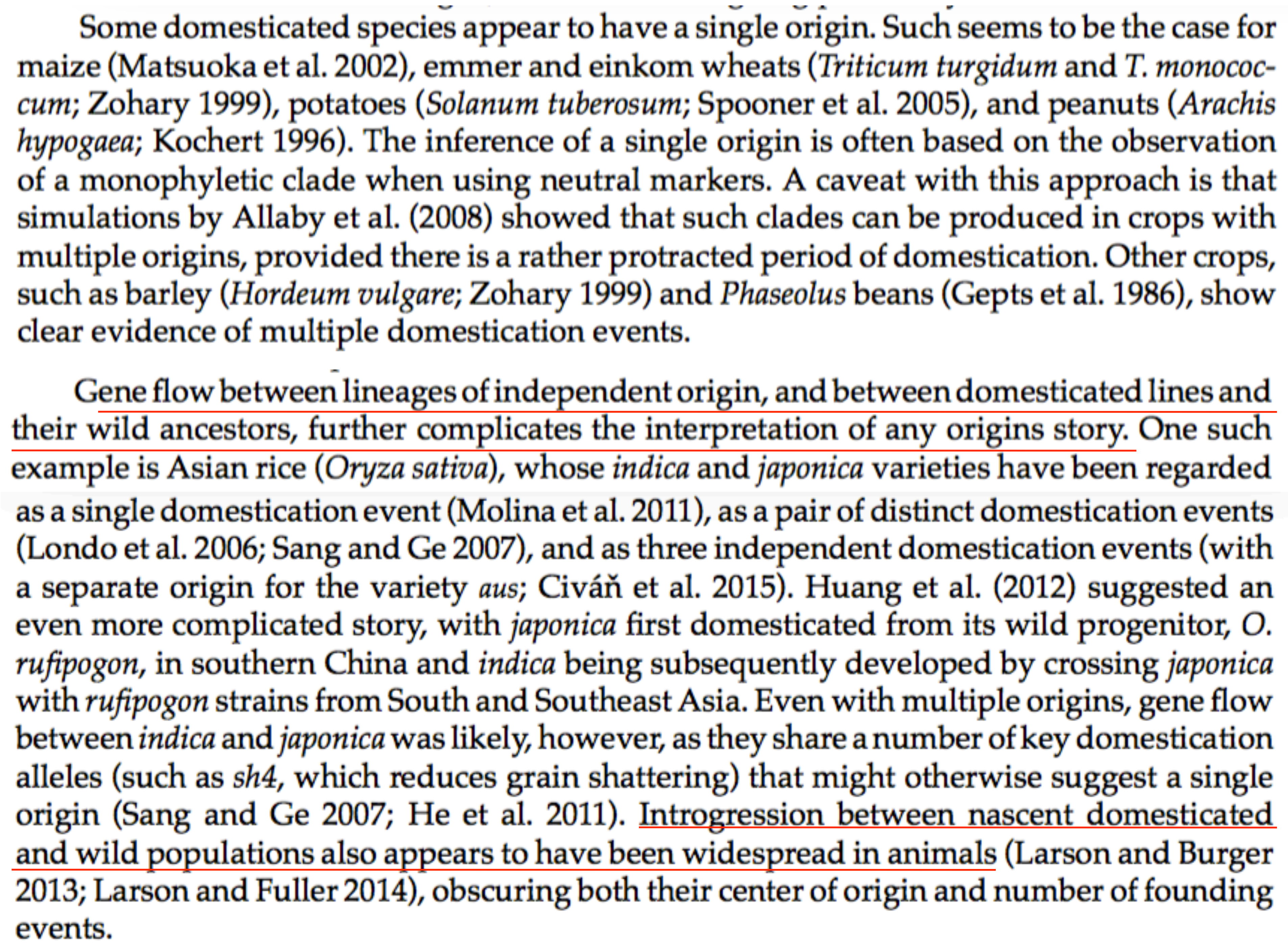
- On short time scales,
  - an excessive amount of diversity
  - excess of intermediate-frequency alleles
- Over longer time scales, the regions of higher diversity gets progressively smaller via recombination
- Trans-species polymorphisms
- Bottom line: little evidence outside of a few known genes (MHC, ABO)

Domestication

- Domestication genes
  - Present in all varieties
- Improvement (or diversification) genes
  - Restricted to subsets
- Darwin noted the process could be due to conscious selection (or methodical selection)
- Or it can be entirely unconscious
  - Simply a byproduct of human-induced changes in the environment



The threshold beyond which a wild species is said to be domesticated can be challenging to assess. One operational definition is that domesticated varieties survive very poorly in a natural setting, due to the establishment of traits that increase fitness in the domesticated environment but decrease it in the wild. As best stated by Zeder et al. (2006), "domestication is a unique form of mutualism," leaving both genetic and archaeological signals (see Zeder et al. for several interesting examples). It is also worth emphasizing that domestication is not a uniquely human enterprise. For example, several species of insects cultivate fungal species, and the search for domestication genes in such systems (in both the domesticating insect and their cultivated fungus) remains an intriguing possibility.



Some domesticated species appear to have a single origin. Such seems to be the case for maize (Matsuoka et al. 2002), emmer and einkom wheats (*Triticum turgidum* and *T. monococcum*; Zohary 1999), potatoes (*Solanum tuberosum*; Spooner et al. 2005), and peanuts (*Arachis hypogaea*; Kochert 1996). The inference of a single origin is often based on the observation of a monophyletic clade when using neutral markers. A caveat with this approach is that simulations by Allaby et al. (2008) showed that such clades can be produced in crops with multiple origins, provided there is a rather protracted period of domestication. Other crops, such as barley (*Hordeum vulgare*; Zohary 1999) and *Phaseolus* beans (Gepts et al. 1986), show clear evidence of multiple domestication events.

Gene flow between lineages of independent origin, and between domesticated lines and their wild ancestors, further complicates the interpretation of any origins story. One such example is Asian rice (*Oryza sativa*), whose *indica* and *japonica* varieties have been regarded as a single domestication event (Molina et al. 2011), as a pair of distinct domestication events (Londo et al. 2006; Sang and Ge 2007), and as three independent domestication events (with a separate origin for the variety *aus*; Civáň et al. 2015). Huang et al. (2012) suggested an even more complicated story, with *japonica* first domesticated from its wild progenitor, *O. rufipogon*, in southern China and *indica* being subsequently developed by crossing *japonica* with *rufipogon* strains from South and Southeast Asia. Even with multiple origins, gene flow between *indica* and *japonica* was likely, however, as they share a number of key domestication alleles (such as *sh4*, which reduces grain shattering) that might otherwise suggest a single origin (Sang and Ge 2007; He et al. 2011). Introgression between nascent domesticated and wild populations also appears to have been widespread in animals (Larson and Burger 2013; Larson and Fuller 2014), obscuring both their center of origin and number of founding events.

ABBA-BABA test

ABBA

Shohei BABA

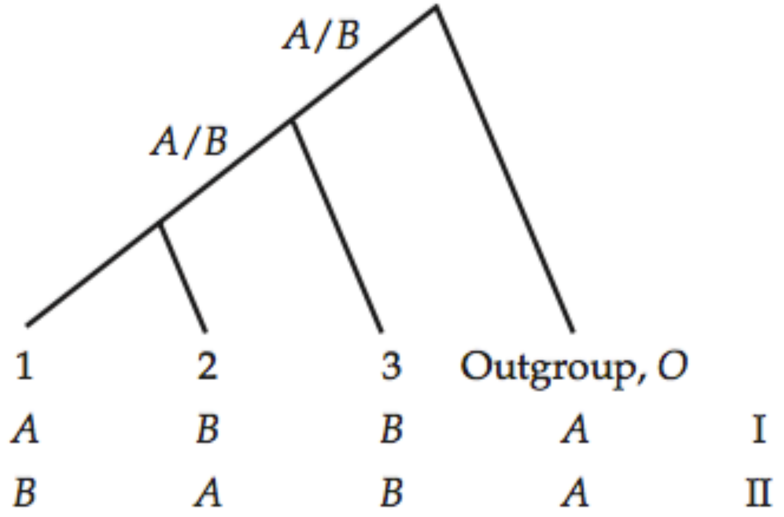


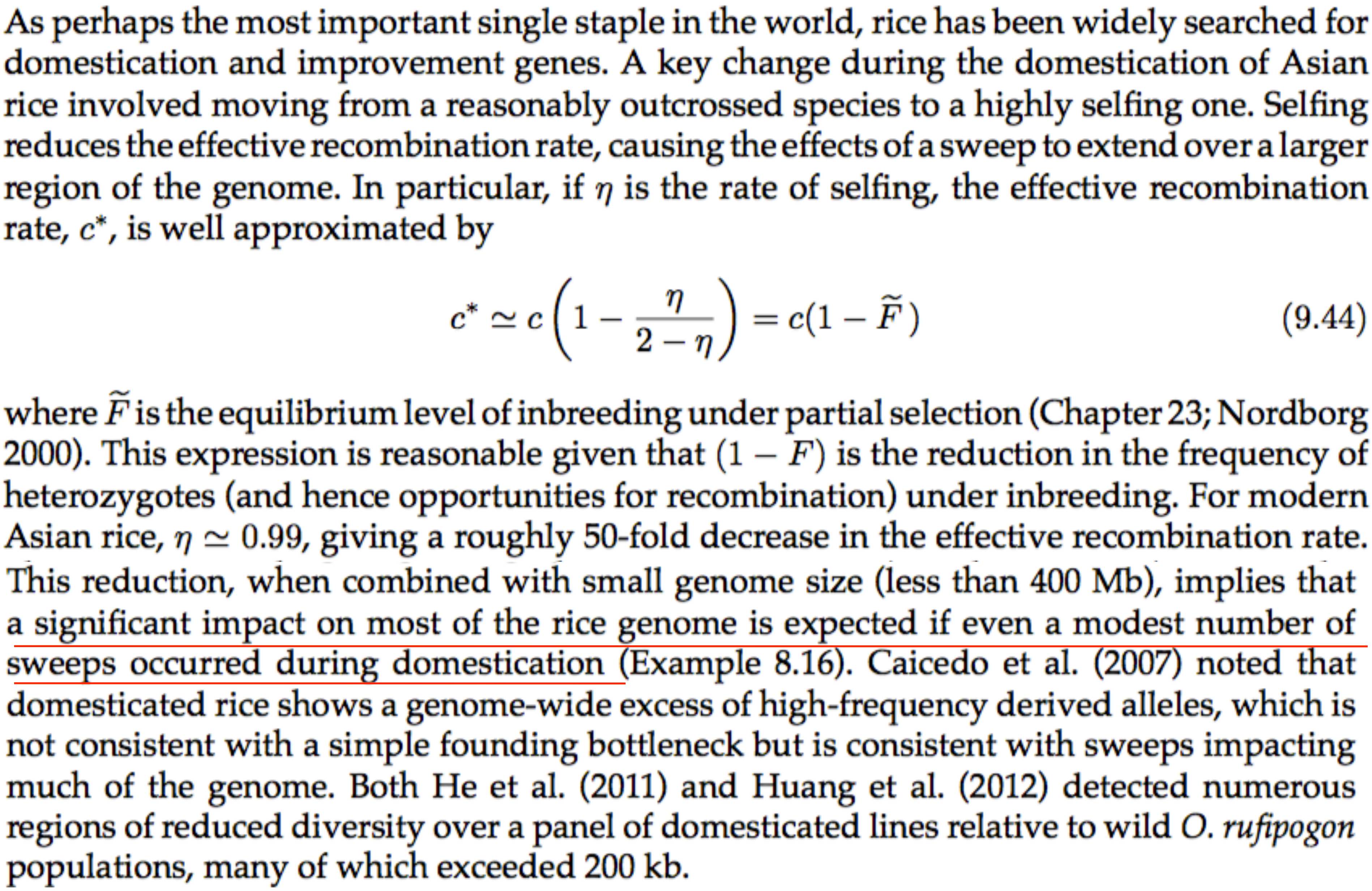
Figure 9.8 The ***ABBA-BABA*** test for detecting the introgression of genes from taxon 3 into either taxon 1 or 2; see Example 9.16 for details. Here *A* and *B* denote the ancestral and derived alleles, with the ancestral allele present in the outgroup, *O*. The test compares the distribution of *A* and *B* in taxa 1 and 2, conditioned on taxon 3 containing the derived allele, *B*. If there is simply neutral lineage sorting between the outgroup and the three resulting taxa, then configurations I (*ABBA*) and II (*BABA*) should be equally frequent. However, if there has been symmetric introgression of alleles from taxon 3 into one of these populations (but not the other), this pattern will be skewed, with one configuration being in excess of 50%.

Define a string of length four and with elements *A* (for the ancestral allele present in *O*) and *B* (a derived allele present in taxon 3), with the positions in this sequence corresponding to species 1, 2, 3, and the outgroup. For example, the configuration given by I in Figure 9.8 is denoted ABBA. Suppose that at a given locus, the ancestral population of 1, 2, and 3 was segregating for *A* and *B*, and that in taxon 3, the lineage was sorted such that *B* was fixed. Conditioned on taxon 3 containing the derived allele, when 1 and 2 carry different alleles (one ancestral and the other derived), the direction should be entirely random (as the sorting would be random for neutral alleles), and hence both cases (*AB* vs. *BA* in these two species, translating into *ABBA* vs. *BABA* for the four-species comparison) should be equally likely. A systematic departure in one direction (i.e., far more *ABBA* than *BABA*) implies introgression from 3 into either 2 or 1 (respectively). Green et al. found a significant skew in favor of introgression from Neandertal into non-African humans. Their *D* statistic is given by

$$D_{ABBA-BABA} = \frac{N_{ABBA} - N_{BABA}}{N_{tot}} \tag{9.43}$$

where N_x is the number of events in class x and $N_{tot} = N_{ABBA} + N_{BABA}$ is the total number of the two events. Significance ($D \neq 0$) is assessed using a jackknife approach. See Durand et al. (2011) for a detailed discussion and development.

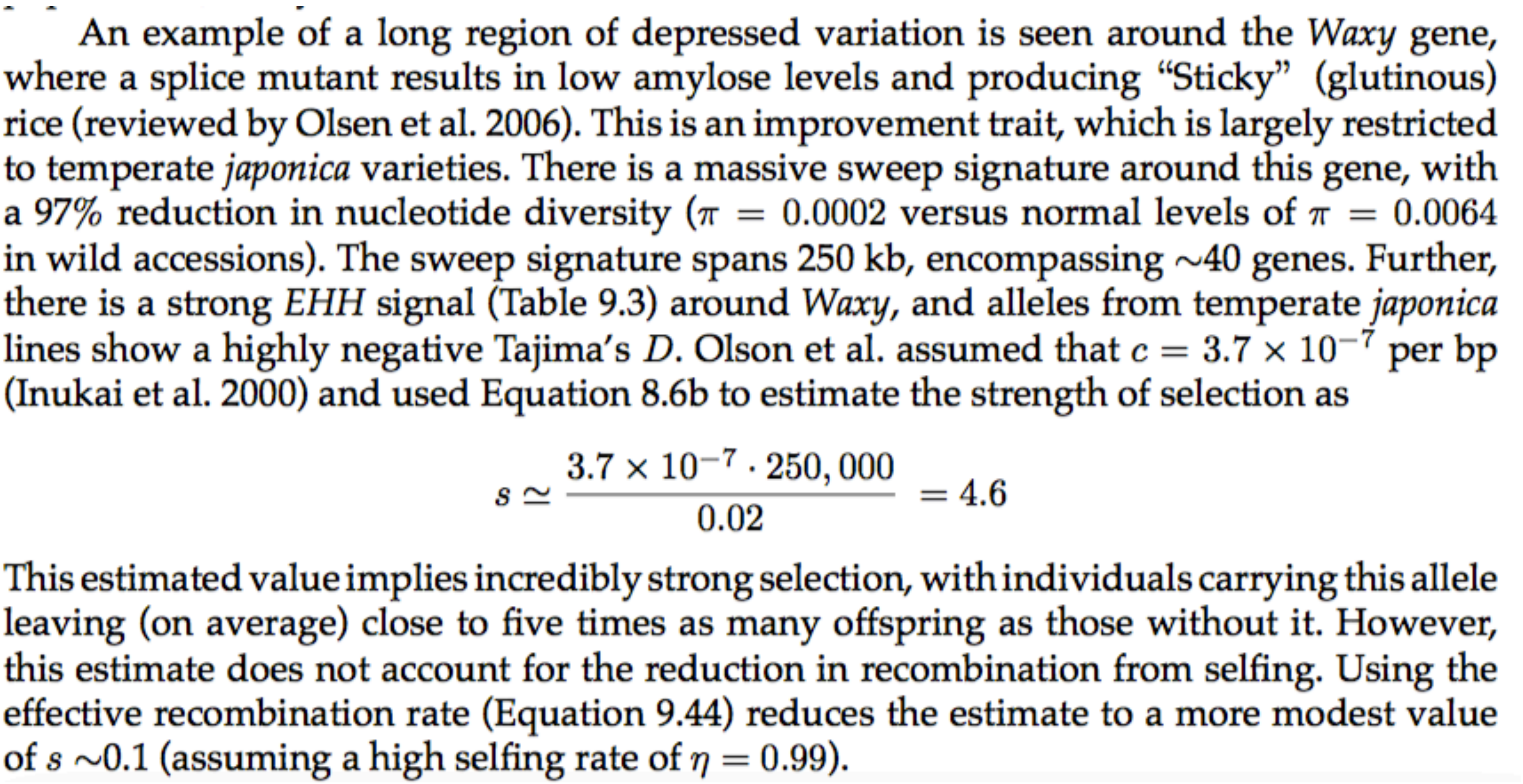
Modifying sweep-theory results for inbred crops



As perhaps the most important single staple in the world, rice has been widely searched for domestication and improvement genes. A key change during the domestication of Asian rice involved moving from a reasonably outcrossed species to a highly selfing one. Selfing reduces the effective recombination rate, causing the effects of a sweep to extend over a larger region of the genome. In particular, if η is the rate of selfing, the effective recombination rate, c^*, is well approximated by

$$c^* \simeq c\left(1 - \frac{\eta}{2-\eta}\right) = c(1 - \widetilde{F}\,) \tag{9.44}$$

where $\widetilde{F}$ is the equilibrium level of inbreeding under partial selection (Chapter 23; Nordborg 2000). This expression is reasonable given that $(1 - F)$ is the reduction in the frequency of heterozygotes (and hence opportunities for recombination) under inbreeding. For modern Asian rice, $\eta \simeq 0.99$, giving a roughly 50-fold decrease in the effective recombination rate. This reduction, when combined with small genome size (less than 400 Mb), implies that a significant impact on most of the rice genome is expected if even a modest number of sweeps occurred during domestication (Example 8.16). Caicedo et al. (2007) noted that domesticated rice shows a genome-wide excess of high-frequency derived alleles, which is not consistent with a simple founding bottleneck but is consistent with sweeps impacting much of the genome. Both He et al. (2011) and Huang et al. (2012) detected numerous regions of reduced diversity over a panel of domesticated lines relative to wild *O. rufipogon* populations, many of which exceeded 200 kb.

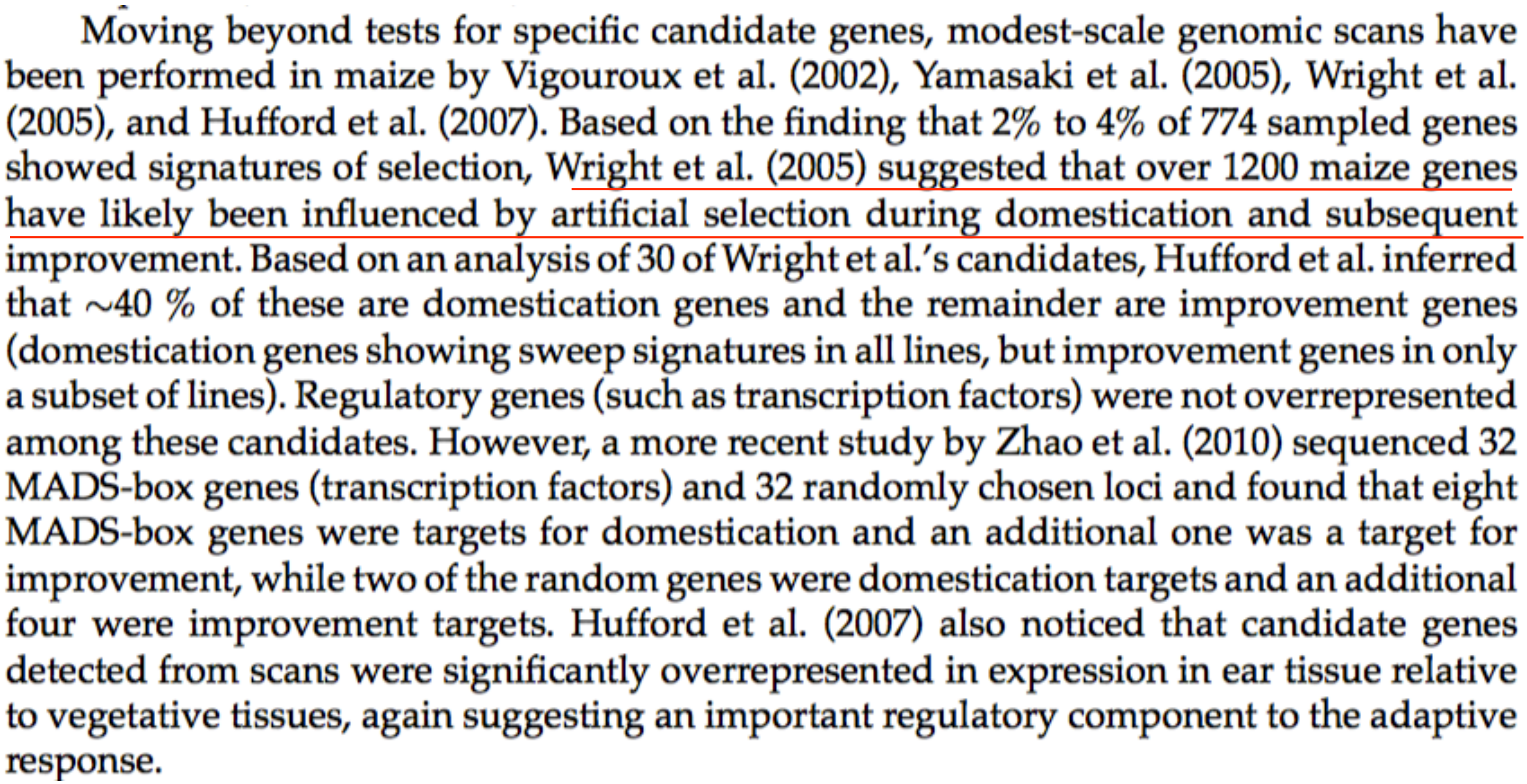


An example of a long region of depressed variation is seen around the *Waxy* gene, where a splice mutant results in low amylose levels and producing “Sticky” (glutinous) rice (reviewed by Olsen et al. 2006). This is an improvement trait, which is largely restricted to temperate *japonica* varieties. There is a massive sweep signature around this gene, with a 97% reduction in nucleotide diversity ($\pi = 0.0002$ versus normal levels of $\pi = 0.0064$ in wild accessions). The sweep signature spans 250 kb, encompassing ~40 genes. Further, there is a strong *EHH* signal (Table 9.3) around *Waxy*, and alleles from temperate *japonica* lines show a highly negative Tajima’s D. Olson et al. assumed that $c = 3.7 \times 10^{-7}$ per bp (Inukai et al. 2000) and used Equation 8.6b to estimate the strength of selection as

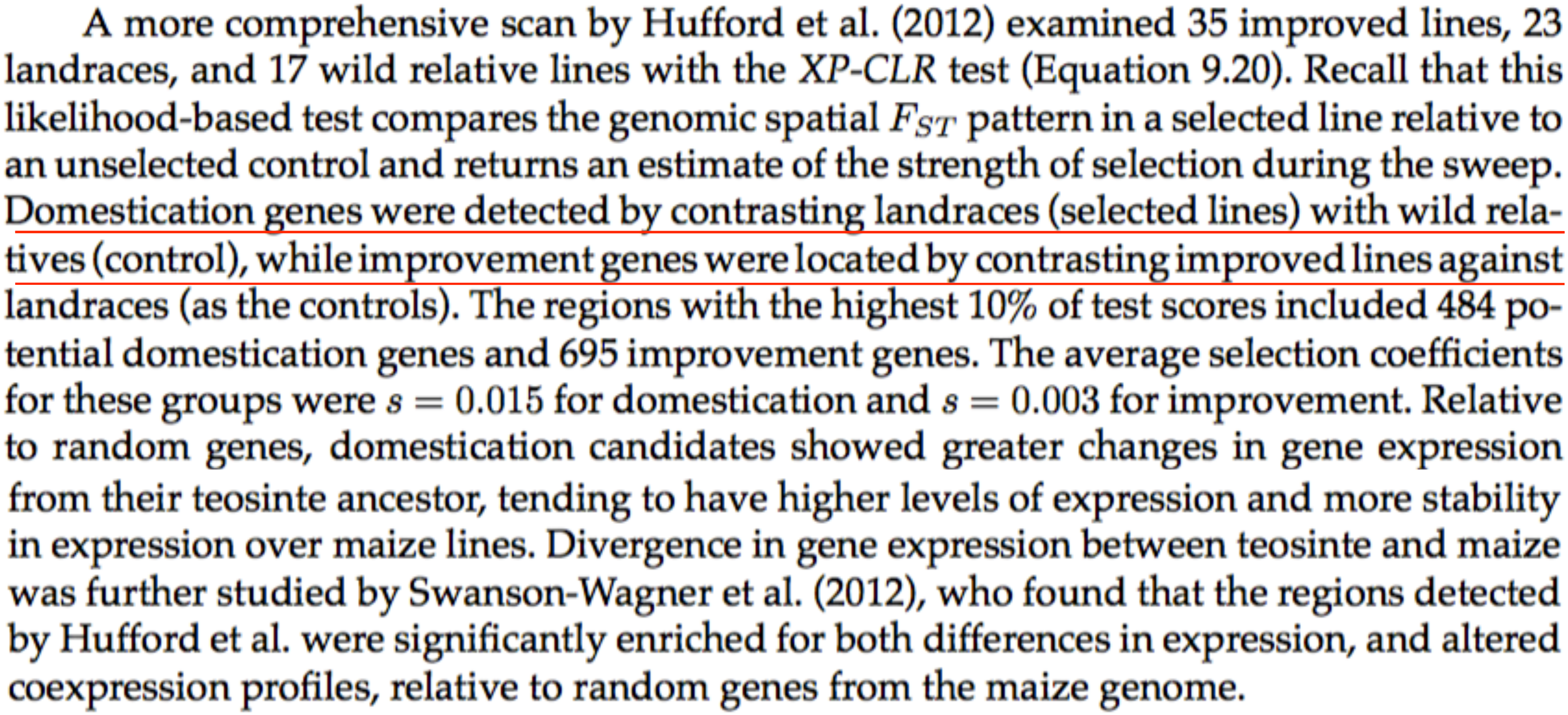
$$s \simeq \frac{3.7 \times 10^{-7} \cdot 250,000}{0.02} = 4.6$$

This estimated value implies incredibly strong selection, with individuals carrying this allele leaving (on average) close to five times as many offspring as those without it. However, this estimate does not account for the reduction in recombination from selfing. Using the effective recombination rate (Equation 9.44) reduces the estimate to a more modest value of s ~0.1 (assuming a high selfing rate of $\eta = 0.99$).

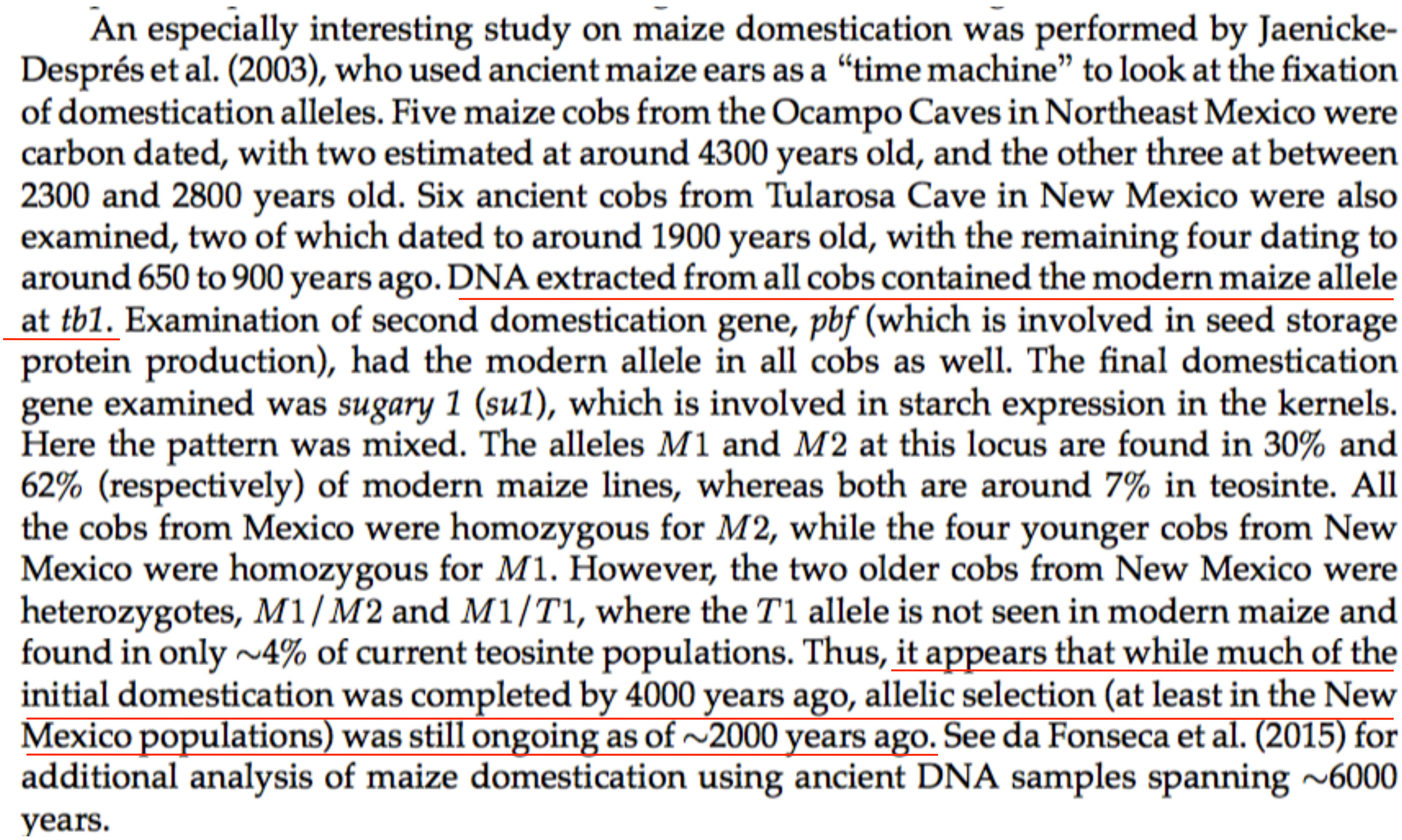
Maize



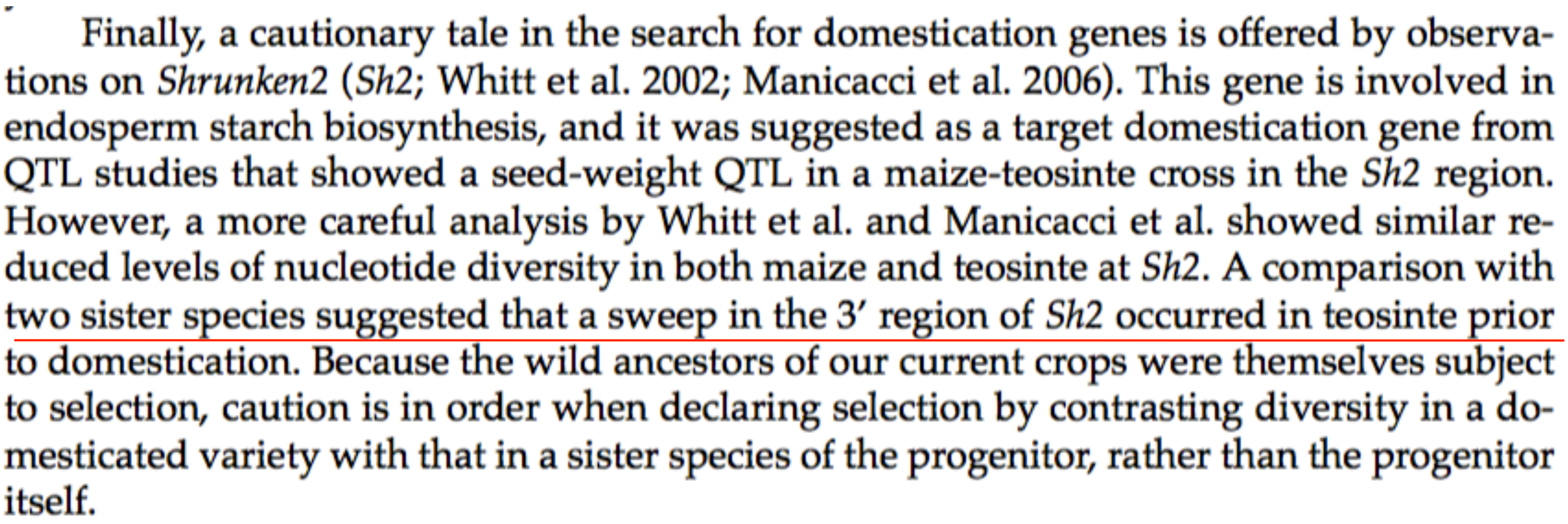
Moving beyond tests for specific candidate genes, modest-scale genomic scans have been performed in maize by Vigouroux et al. (2002), Yamasaki et al. (2005), Wright et al. (2005), and Hufford et al. (2007). Based on the finding that 2% to 4% of 774 sampled genes showed signatures of selection, Wright et al. (2005) suggested that over 1200 maize genes have likely been influenced by artificial selection during domestication and subsequent improvement. Based on an analysis of 30 of Wright et al.'s candidates, Hufford et al. inferred that $\sim$40 % of these are domestication genes and the remainder are improvement genes (domestication genes showing sweep signatures in all lines, but improvement genes in only a subset of lines). Regulatory genes (such as transcription factors) were not overrepresented among these candidates. However, a more recent study by Zhao et al. (2010) sequenced 32 MADS-box genes (transcription factors) and 32 randomly chosen loci and found that eight MADS-box genes were targets for domestication and an additional one was a target for improvement, while two of the random genes were domestication targets and an additional four were improvement targets. Hufford et al. (2007) also noticed that candidate genes detected from scans were significantly overrepresented in expression in ear tissue relative to vegetative tissues, again suggesting an important regulatory component to the adaptive response.



A more comprehensive scan by Hufford et al. (2012) examined 35 improved lines, 23 landraces, and 17 wild relative lines with the *XP-CLR* test (Equation 9.20). Recall that this likelihood-based test compares the genomic spatial F_{ST} pattern in a selected line relative to an unselected control and returns an estimate of the strength of selection during the sweep. Domestication genes were detected by contrasting landraces (selected lines) with wild relatives (control), while improvement genes were located by contrasting improved lines against landraces (as the controls). The regions with the highest 10% of test scores included 484 potential domestication genes and 695 improvement genes. The average selection coefficients for these groups were $s = 0.015$ for domestication and $s = 0.003$ for improvement. Relative to random genes, domestication candidates showed greater changes in gene expression from their teosinte ancestor, tending to have higher levels of expression and more stability in expression over maize lines. Divergence in gene expression between teosinte and maize was further studied by Swanson-Wagner et al. (2012), who found that the regions detected by Hufford et al. were significantly enriched for both differences in expression, and altered coexpression profiles, relative to random genes from the maize genome.

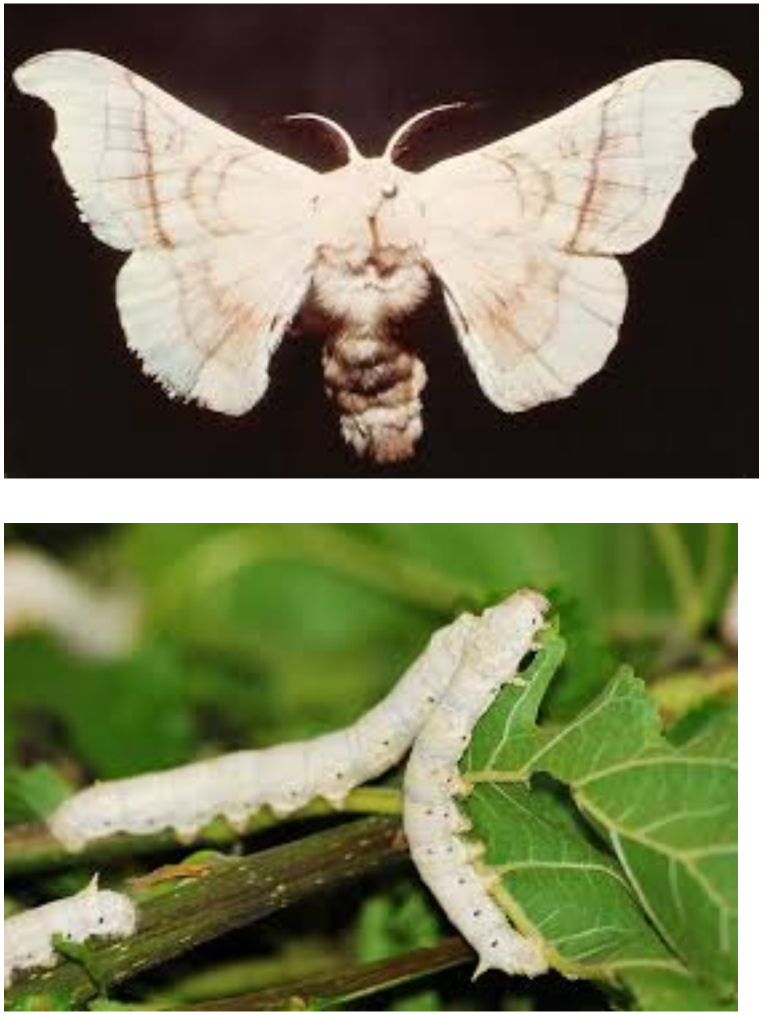


An especially interesting study on maize domestication was performed by Jaenicke-Després et al. (2003), who used ancient maize ears as a “time machine” to look at the fixation of domestication alleles. Five maize cobs from the Ocampo Caves in Northeast Mexico were carbon dated, with two estimated at around 4300 years old, and the other three at between 2300 and 2800 years old. Six ancient cobs from Tularosa Cave in New Mexico were also examined, two of which dated to around 1900 years old, with the remaining four dating to around 650 to 900 years ago. DNA extracted from all cobs contained the modern maize allele at *tb1*. Examination of second domestication gene, *pbf* (which is involved in seed storage protein production), had the modern allele in all cobs as well. The final domestication gene examined was *sugary 1* (*su1*), which is involved in starch expression in the kernels. Here the pattern was mixed. The alleles $M1$ and $M2$ at this locus are found in 30% and 62% (respectively) of modern maize lines, whereas both are around 7% in teosinte. All the cobs from Mexico were homozygous for $M2$, while the four younger cobs from New Mexico were homozygous for $M1$. However, the two older cobs from New Mexico were heterozygotes, $M1/M2$ and $M1/T1$, where the $T1$ allele is not seen in modern maize and found in only $\sim$4% of current teosinte populations. Thus, it appears that while much of the initial domestication was completed by 4000 years ago, allelic selection (at least in the New Mexico populations) was still ongoing as of $\sim$2000 years ago. See da Fonseca et al. (2015) for additional analysis of maize domestication using ancient DNA samples spanning $\sim$6000 years.

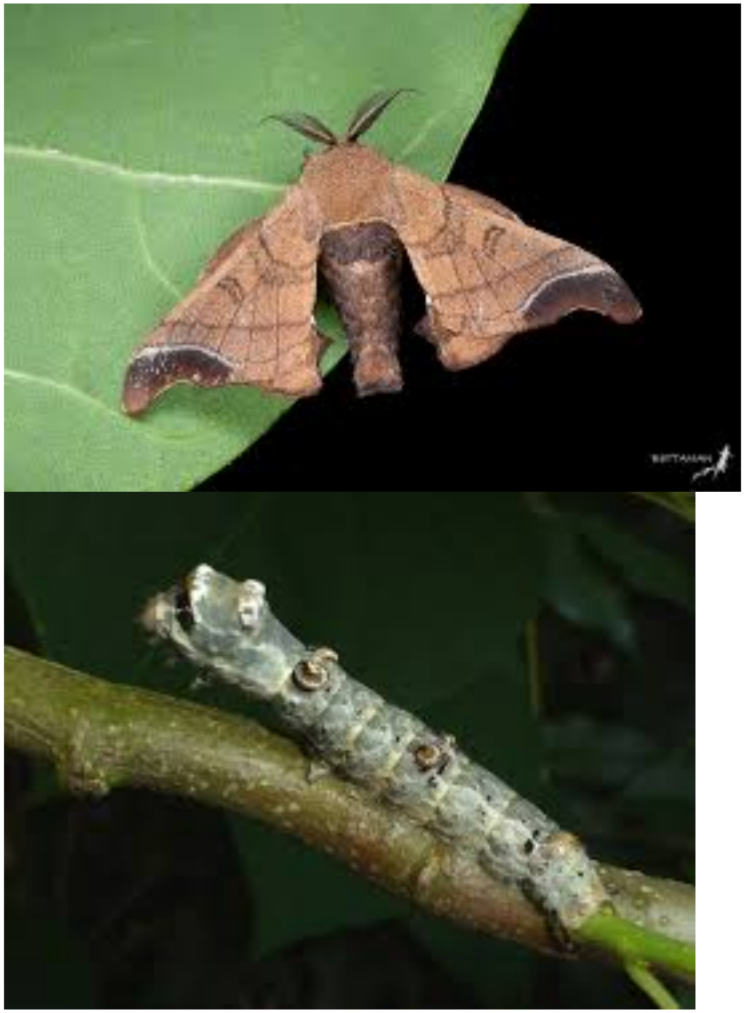


Finally, a cautionary tale in the search for domestication genes is offered by observations on *Shrunken2* (*Sh2*; Whitt et al. 2002; Manicacci et al. 2006). This gene is involved in endosperm starch biosynthesis, and it was suggested as a target domestication gene from QTL studies that showed a seed-weight QTL in a maize-teosinte cross in the *Sh2* region. However, a more careful analysis by Whitt et al. and Manicacci et al. showed similar reduced levels of nucleotide diversity in both maize and teosinte at *Sh2*. A comparison with two sister species suggested that a sweep in the 3′ region of *Sh2* occurred in teosinte prior to domestication. Because the wild ancestors of our current crops were themselves subject to selection, caution is in order when declaring selection by contrasting diversity in a domesticated variety with that in a sister species of the progenitor, rather than the progenitor itself.

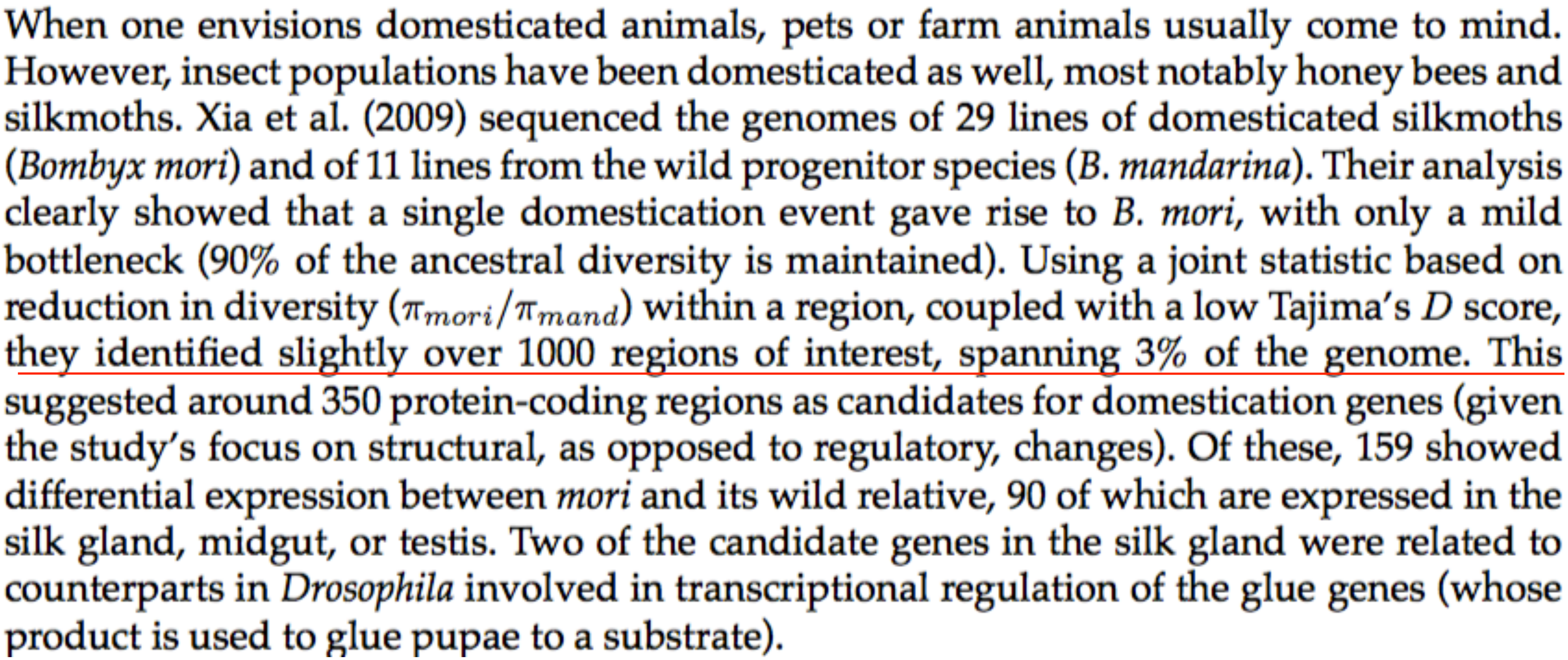
Domesticated Insects: Silkworms



Bombyx mori

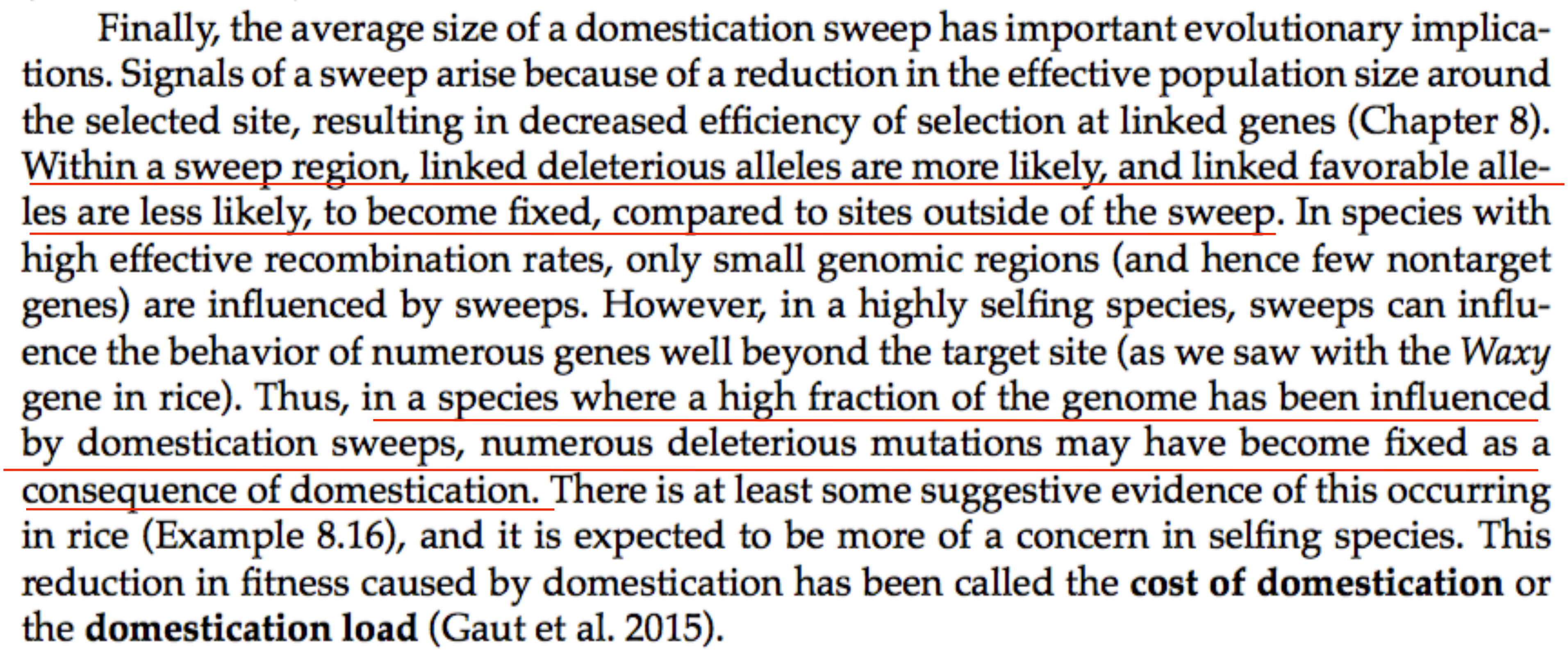


Bombyx mandarina



When one envisions domesticated animals, pets or farm animals usually come to mind. However, insect populations have been domesticated as well, most notably honey bees and silkmoths. Xia et al. (2009) sequenced the genomes of 29 lines of domesticated silkmoths (*Bombyx mori*) and of 11 lines from the wild progenitor species (*B. mandarina*). Their analysis clearly showed that a single domestication event gave rise to *B. mori*, with only a mild bottleneck (90% of the ancestral diversity is maintained). Using a joint statistic based on reduction in diversity (π_{mori}/π_{mand}) within a region, coupled with a low Tajima's D score, they identified slightly over 1000 regions of interest, spanning 3% of the genome. This suggested around 350 protein-coding regions as candidates for domestication genes (given the study's focus on structural, as opposed to regulatory, changes). Of these, 159 showed differential expression between *mori* and its wild relative, 90 of which are expressed in the silk gland, midgut, or testis. Two of the candidate genes in the silk gland were related to counterparts in *Drosophila* involved in transcriptional regulation of the glue genes (whose product is used to glue pupae to a substrate).

The cost of domestication



Finally, the average size of a domestication sweep has important evolutionary implications. Signals of a sweep arise because of a reduction in the effective population size around the selected site, resulting in decreased efficiency of selection at linked genes (Chapter 8). Within a sweep region, linked deleterious alleles are more likely, and linked favorable alleles are less likely, to become fixed, compared to sites outside of the sweep. In species with high effective recombination rates, only small genomic regions (and hence few nontarget genes) are influenced by sweeps. However, in a highly selfing species, sweeps can influence the behavior of numerous genes well beyond the target site (as we saw with the *Waxy* gene in rice). Thus, in a species where a high fraction of the genome has been influenced by domestication sweeps, numerous deleterious mutations may have become fixed as a consequence of domestication. There is at least some suggestive evidence of this occurring in rice (Example 8.16), and it is expected to be more of a concern in selfing species. This reduction in fitness caused by domestication has been called the **cost of domestication** or the **domestication load** (Gaut et al. 2015).