Lecture 09:
Detecting selection with marker data.
3: Polymorphism-based tests: II

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Tests covered here

- SFS tests
  - Very sensitive to the equilibrium population assumption. Lots of false positives
- Haplotype-based tests
  - Perhaps the strongest tests for soft sweeps, ongoing selection
  - A large number of different tests (and approaches!)

SFS tests

- Recall that the full site frequency spectrum under the neutral equilibrium model is simply a function of $\theta = 4N_e u$
- Recall also that there are a number of different ways to estimate θ.
- These estimates should all give the same answer (within sampling error) under the standard neutral model.
- Departures in estimators indicate that this model does not hold

Site frequency spectrum (SFS)

- The distribution of either the minor allele frequency (folded frequency spectrum) or the frequency of a derived allele (unfold frequency spectrum) given by the Watterson distribution (θ/x)
- A sweep inflates the frequency of sites segregating rare alleles (folded spectrum)
- Sweeps inflate the frequency of derived alleles, increasing the number of high-frequency sites in the (unfolded) SFS

Watterson distribution

- Let x = population frequency of all sites with a fraction of x derived alleles

$$\phi(x) = \frac{\theta}{x} \quad \text{for} \quad \frac{1}{2N} \leq x \leq 1 - \frac{1}{2N}$$

Folded Watterson distribution, x = freq of minor allele (x $\leq$ 0.5)

$$\phi(x) = \frac{\theta}{x} + \frac{\theta}{1-x} = \frac{\theta}{x(1-x)} \quad \text{for} \quad \frac{1}{2N} \leq x \leq 1/2$$

Expected number of sites in a sample

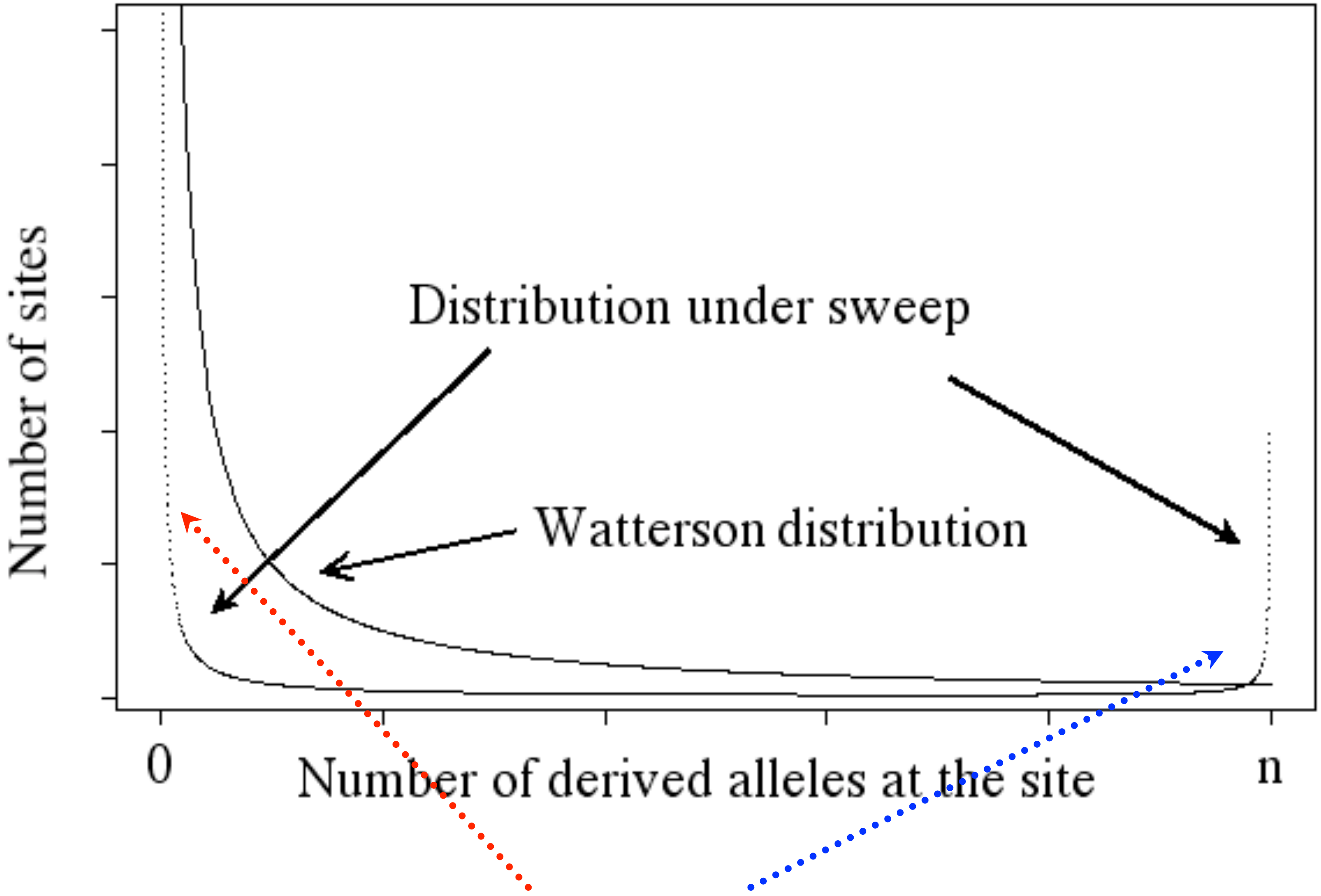
unfolded

$$E(s_i) = \frac{\theta_L}{i}, \quad \text{for} \quad 1 \leq i \leq n-1$$

folded

$$E(s_i) = \frac{\theta_L}{i} + \frac{\theta_L}{n-i} = \frac{\theta_L n}{i(n-i)}, \quad \text{for} \quad 1 \leq i \leq [n/2]$$

Distribution of allele frequencies under a sweep



Excess of low & high frequency sites

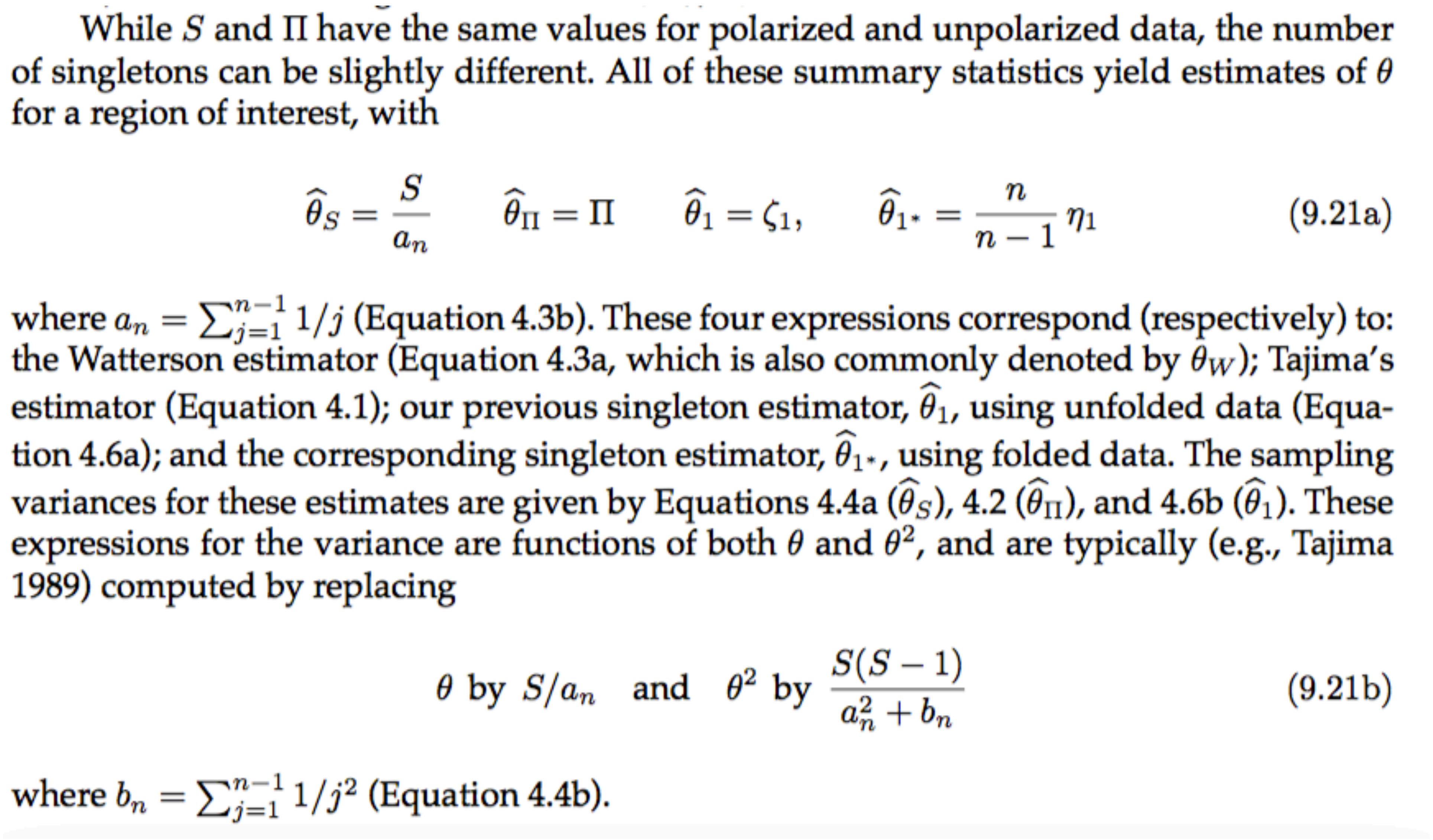
TESTS BASED ON SITE-FREQUENCY SPECTRUM STATISTICS

Under the infinite-sites model, a sequence is treated as a series of L sites, with each new mutation assumed to occur at a new site (Chapter 4). At mutation-drift equilibrium, most features of this model, including the site-frequency spectrum (SFS), are fully specified by the population-size-scaled mutation rate, $\theta = 4N_e\mu$. Depending on the nature of the data, an observed frequency spectrum is viewed as either folded or unfolded (Chapter 2). An unfolded spectrum considers the frequency of the derived allele (Equation 2.35a), and such data are said to be polarized (typically using an outgroup to distinguish between ancestral and derived, or mutant, alleles). The folded spectrum (Equation 2.35b) uses the minor-allele frequency, ignoring whether the rarer allele is ancestral or derived. To distinguish between these different spectra, we use the notation that ζ_i denotes the number of sites that contain exactly i derived alleles ($1 \leq i \leq n-1$), yielding the observed unfolded SFS as the vector $(\zeta_1, \cdots, \zeta_{n-1})$. Similarly, η_i denotes the number of sites with exactly i copies of the minor allele ($1 \leq i \leq [n/2]$), with $(\eta_1, \cdots, \eta_{[n/2]})$ being the observed folded SFS, where

$$[n/2] = \begin{cases} n/2 \text{ for } n \text{ even} \\ (n-1)/2 \text{ for } n \text{ odd} \end{cases}$$

The η_i and ζ_i are simply related by

$$\eta_i = \zeta_i + \zeta_{n-i} \quad \text{for} \quad 1 \leq i \leq [n/2]$$



While S and Π have the same values for polarized and unpolarized data, the number of singletons can be slightly different. All of these summary statistics yield estimates of θ for a region of interest, with

$$\widehat{\theta}_S = \frac{S}{a_n} \qquad \widehat{\theta}_\Pi = \Pi \qquad \widehat{\theta}_1 = \zeta_1, \qquad \widehat{\theta}_{1^*} = \frac{n}{n-1}\,\eta_1 \tag{9.21a}$$

where $a_n = \sum_{j=1}^{n-1} 1/j$ (Equation 4.3b). These four expressions correspond (respectively) to: the Watterson estimator (Equation 4.3a, which is also commonly denoted by θ_W); Tajima's estimator (Equation 4.1); our previous singleton estimator, $\widehat{\theta}_1$, using unfolded data (Equation 4.6a); and the corresponding singleton estimator, $\widehat{\theta}_{1^*}$, using folded data. The sampling variances for these estimates are given by Equations 4.4a ($\widehat{\theta}_S$), 4.2 ($\widehat{\theta}_\Pi$), and 4.6b ($\widehat{\theta}_1$). These expressions for the variance are functions of both θ and θ^2, and are typically (e.g., Tajima 1989) computed by replacing

$$\theta \text{ by } S/a_n \quad \text{and} \quad \theta^2 \text{ by } \frac{S(S-1)}{a_n^2 + b_n} \tag{9.21b}$$

where $b_n = \sum_{j=1}^{n-1} 1/j^2$ (Equation 4.4b).

Test	Contrast	Spectrum	Signal
Tajima's D	$\widehat{\theta}_S$ vs. $\widehat{\theta}_\Pi$	Folded	< 0: Excess of rare alleles Sweep or population bottleneck > 0: Excess of intermediate-frequency alleles Balancing selection or population structure
Achaz's Y^*	$\widehat{\theta}_{S_{-\eta_1}}$ vs. $\widehat{\theta}_{\Pi_{-\eta_1}}$	Folded	Same as for Tajima's D
Achaz's Y	$\widehat{\theta}_{S_{-\zeta_1}}$ vs. $\widehat{\theta}_{\Pi_{-\zeta_1}}$	Unfolded	Same as for Tajima's D
Fu and Li's D	$\widehat{\theta}_S$ vs. $\widehat{\theta}_1$	Unfolded	Same as for Tajima's D
Fu and Li's D^*	$\widehat{\theta}_S$ vs. $\widehat{\theta}_{1*}$	Folded	Same as for Tajima's D
Fu and Li's F	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_1$	Unfolded	Same as for Tajima's D
Fu and Li's F^*	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_{1*}$	Folded	Same as for Tajima's D
Fay and Wu's H	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_H$	Unfolded	< 0: Excess of high-frequency derived alleles. Sweep or allelic surfing
Zeng et al.'s E	$\widehat{\theta}_\Pi$ vs. $\widehat{\theta}_L$	Unfolded	< 0: Excess of low- vs. high-frequency derived alleles. Signal of a recent *past* sweep

The idea behind site-frequency tests of neutrality is to compare two different estimates of θ based on information from *different regions* of the site-frequency spectrum. When the infinite-sites model holds and the population is at mutation-drift equilibrium, these estimates should be within the sampling error of each other, while they can be significantly different when the neutral equilibrium model does not hold. Table 9.1 summarizes the various site-frequency test statistics discussed here, all of which have the form

$$t = \frac{\widehat{\theta}_i - \widehat{\theta}_j}{\sigma(\widehat{\theta}_i - \widehat{\theta}_j)} \tag{9.21c}$$

Example 9.11. As we now illustrate, all of the tests summarized in Table 9.1 follow from a general family of estimators of θ based on the discrete Watterson distribution (Equation 2.35). For a sample of n sequences with L sites, the expected number of segregating sites with i copies of the derived (unfolded, ζ_i) or of the minor (folded, η_i) allele are

$$E(\zeta_i) = \frac{\theta}{i} \qquad \text{for} \qquad 1 \le i \le n-1$$

$$E(\eta_i) = E(\zeta_i) + E(\zeta_{n-i}) = \frac{\theta}{i} + \frac{\theta}{n-i} = \frac{\theta}{i}\,\frac{n}{n-i} \quad \text{for} \quad 1 \le i \le [n/2] \tag{9.22a}$$

where $\theta = 4N_e\mu$ is the scaled mutation rate for the entire region.

Hence, a method-of-moments estimator for θ using only the number in the ith class from either SFS is simply

$$\widehat{\theta}_i = \begin{cases} i \cdot \zeta_i & i \text{ copies of the derived allele} \quad 1 \le i \le n-1 \\ \\ \dfrac{i \cdot (n-i)}{n}\,\eta_i & i \text{ copies of the minor allele} \quad 1 \le i \le [n/2] \end{cases} \tag{9.22b}$$

Nawa and Tajima (2008) suggested that a plot of $\widehat{\theta}_i$ versus i can be helpful for visualizing departures from the neutral SFS, although values for large i may be more problematic as the variance of $\widehat{\theta}_i$ dramatically increases with i.

Following Zeng et al. (2006), consider any summary statistic, g, of the unfolded site-frequency spectrum of the form

$$g = \sum_{i=1}^{n-1} c_i \, \zeta_i \tag{9.23a}$$

From Equation 9.22a

$$E(g) = \sum_{i=1}^{n-1} c_i \, \frac{\theta}{i} = \theta \, h(n) \quad \text{where} \quad h(n) = \sum_{i=1}^{n-1} \frac{c_i}{i} \tag{9.23b}$$

Thus, a family of estimators for θ based on an arbitrary vector $(c_1, \cdots, c_{n-1})$ of weights is given by

$$\widehat{\theta}_g = \frac{g}{h(n)} \tag{9.23c}$$

where $h(n)$ is a function of the sample size n and the chosen weights c_i, and g is the observed value of the statistic.

The choice of weights allows one to tailor statistics to use different parts of the frequency spectrum when estimating θ. Taking $c_i = 1$ yields $g = S$ and $h(n) = a_n$, recovering the Watterson estimator, $\widehat{\theta}_S = S/a_n$. Taking $c_i = i(n-i)$

$$h(n) = \sum_{i=1}^{n-1} i(n-i)/i = n(n-1)/2$$

yielding

$$\widehat{\theta} = \sum_{i=1}^{n-1} \frac{2i(n-i)}{n(n-1)}$$

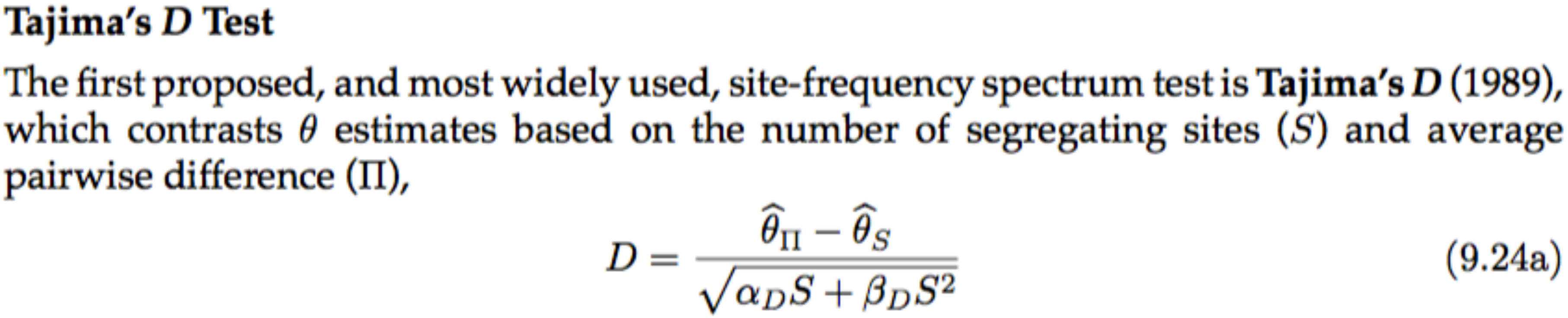
which is simply the average pairwise difference, Π. As with S, Π is symmetric with respect to i and $n-i$, so that both folded and unfolded data return the same estimate. Taking $c_1 = 1, c_{i>1} = 0$ yields $g = \zeta_1$ (the number of derived singletons) and $h(n) = 1$, recovering the $\widehat{\theta}_1$ estimator.

Similarly, for a folded frequency spectrum,

$$g = \sum_{i=1}^{[n/2]} c_i \, \eta_i, \qquad \widehat{\theta}_g = \frac{g}{f(n)}, \qquad f(n) = \sum_{i=1}^{[n/2]} c_i \, \frac{n}{i(n-i)} \tag{9.23d}$$

Consider the estimator using only folded singletons, η_1. Here, $c_1 = 1$, $c_i = 0$ for $i > 1$, and hence $f(n) = n/(n-1)$, giving $\eta_1(n-1)/n$ as an estimator of θ, which recovers $\widehat{\theta}_{1^*}$. Achaz (2009) provided general expressions for the variance of any estimator of the form given by Equations 9.23c or 9.23d, providing all of the machinery to develop general tests in the form of Equation 9.21c using any feature of interest in the SFS.

Fumio Tajima



Tajima's *D* Test

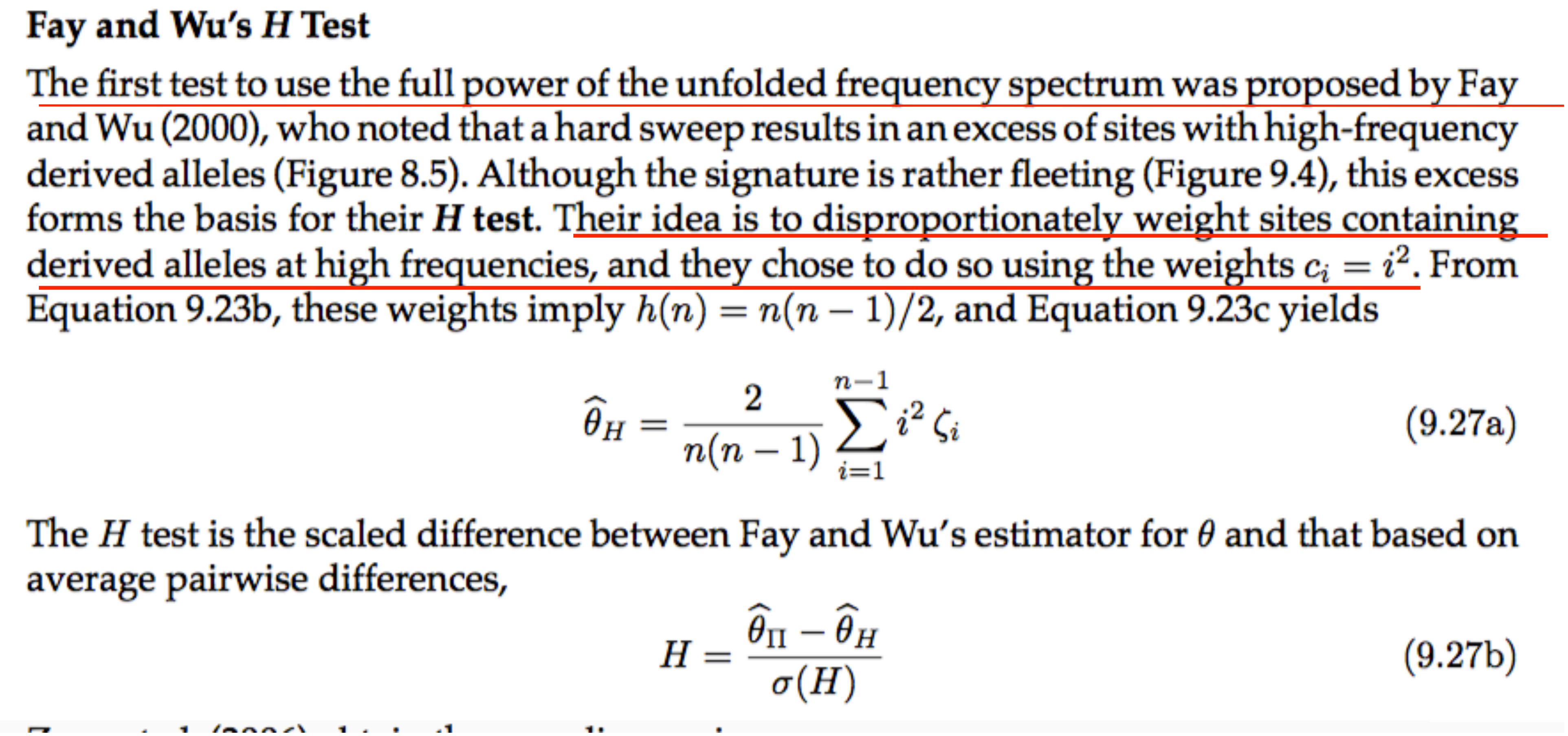
The first proposed, and most widely used, site-frequency spectrum test is **Tajima's *D*** (1989), which contrasts θ estimates based on the number of segregating sites (S) and average pairwise difference (Π),

$$D = \frac{\widehat{\theta}_{\Pi} - \widehat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}} \tag{9.24a}$$

A negative value of D indicates that there are too many low -frequency sites, while a positive value of D indicates that there are too many intermediate-frequency sites.

Expressed another way, D is a test for whether the amount of heterozygosity per site is consistent with the number of polymorphic sites expected under the equilibrium neutral model.

Under selective sweeps (and population expansion), heterozygosity should be significantly less than is predicted from the number of polymorphisms



Fay and Wu's *H* Test

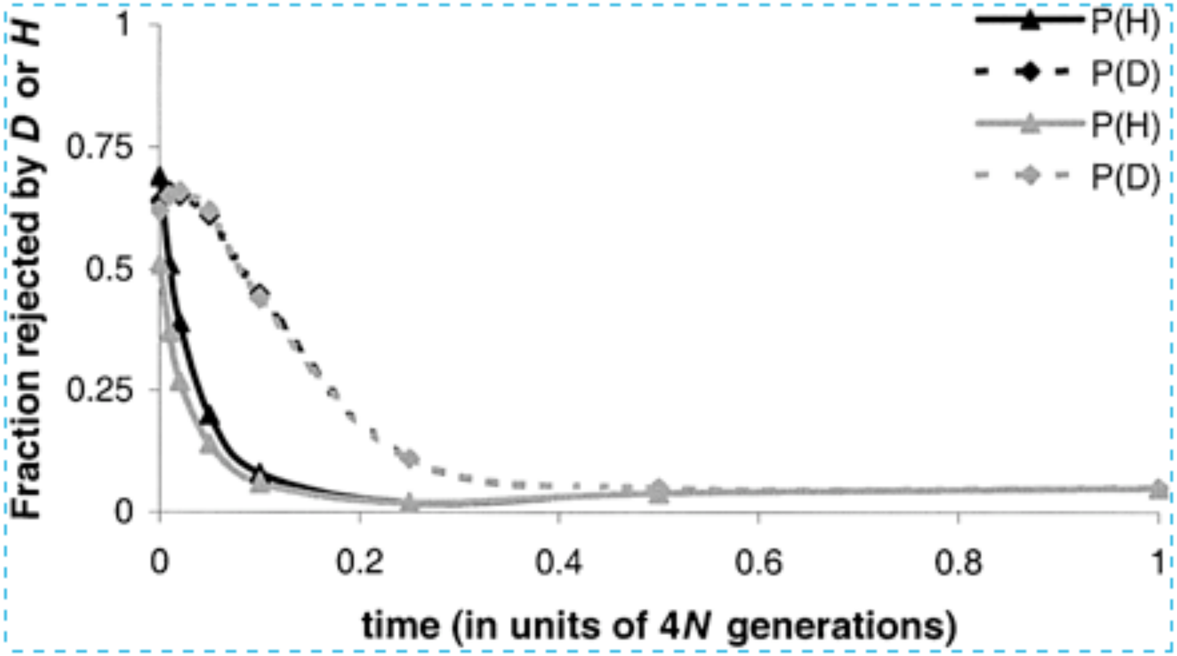
The first test to use the full power of the unfolded frequency spectrum was proposed by Fay and Wu (2000), who noted that a hard sweep results in an excess of sites with high-frequency derived alleles (Figure 8.5). Although the signature is rather fleeting (Figure 9.4), this excess forms the basis for their ***H* test**. Their idea is to disproportionately weight sites containing derived alleles at high frequencies, and they chose to do so using the weights $c_i = i^2$. From Equation 9.23b, these weights imply $h(n) = n(n-1)/2$, and Equation 9.23c yields

$$\widehat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \, \zeta_i \tag{9.27a}$$

The H test is the scaled difference between Fay and Wu's estimator for θ and that based on average pairwise differences,

$$H = \frac{\widehat{\theta}_\Pi - \widehat{\theta}_H}{\sigma(H)} \tag{9.27b}$$

SFS Power quickly dissipates



H loses power very quickly as high-frequency alleles following the sweep are fixed.

Zeng et al.'s *E* Test

A variant of the H test was proposed by Zeng et al. (2006), who noted that the most powerful contrasts between regions of the unfolded frequency spectrum following selection should involve high- versus low-frequency sites. However, most contrasts involve a comparison with θ_Π, which is a measure of intermediate-frequency alleles. To rectify this, Zeng et al. introduced the estimator, θ_L, based on a weight, $c_i = i$, that places more emphasis on high-frequency sites than θ_S (but not as much as θ_H). For these weights, Equation 9.23b implies $h(n) = n - 1$, and hence Equation 9.23c yields

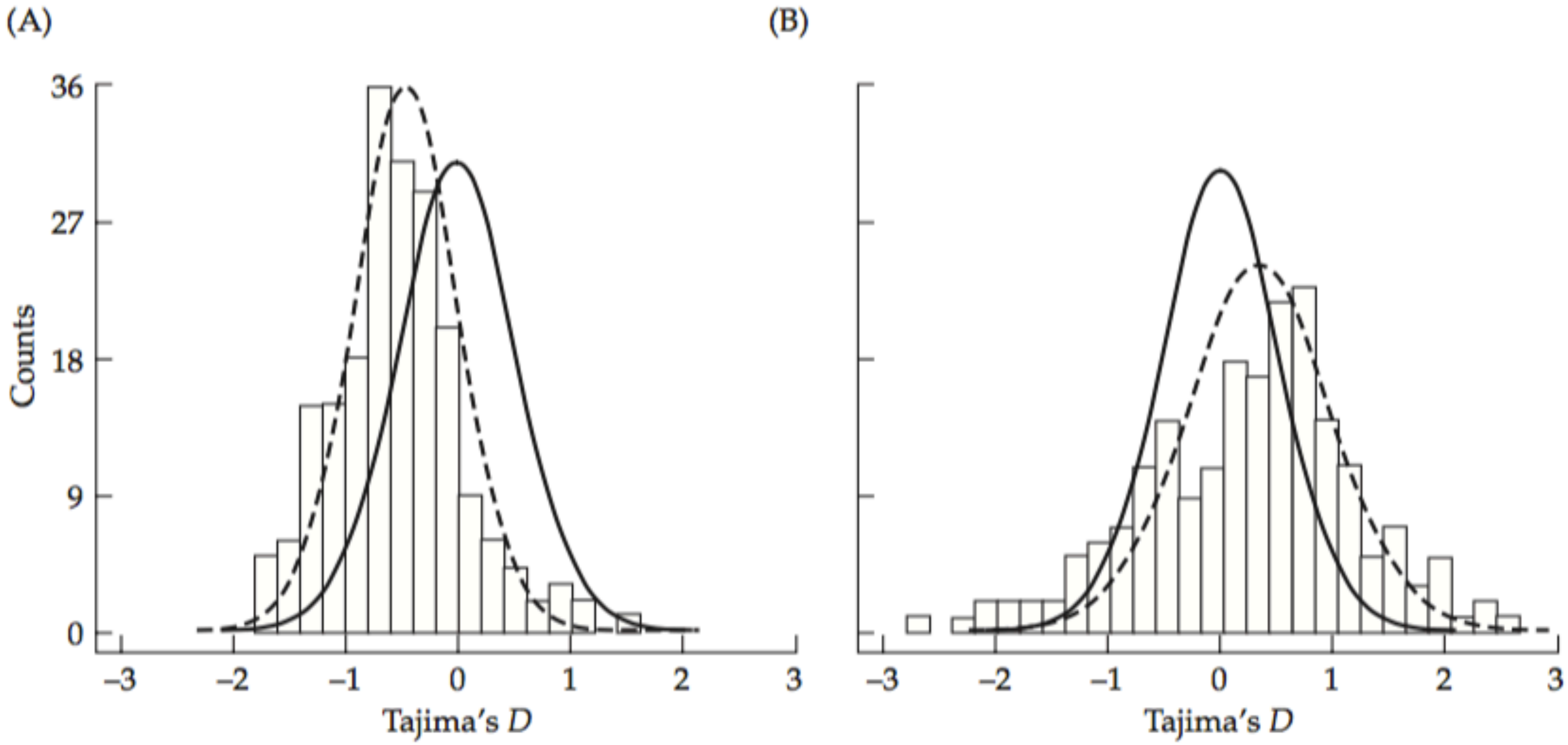
$$\widehat{\theta}_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i\,\zeta_i \tag{9.28a}$$

Zeng et al.'s ***E* test** contrasts the high- and low-frequency regions of the frequency spectrum,

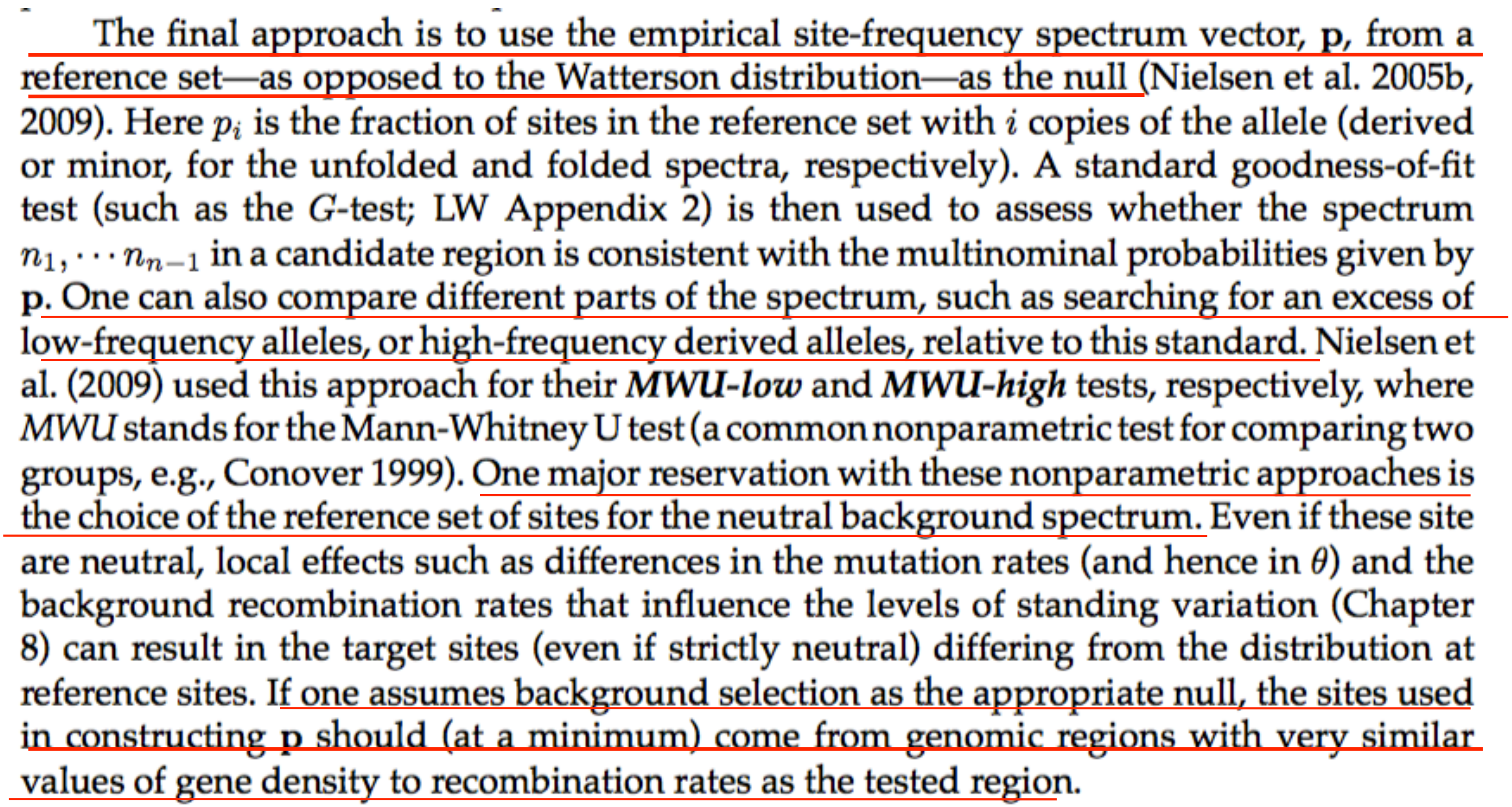
$$E = \frac{\widehat{\theta}_L - \widehat{\theta}_S}{\sigma(E)} \tag{9.28b}$$

Adjusting the Null to Account for Nonequilibrium Populations

- Using the empirical distribution of test statistics from a set of genes in the sample (the outlier approach)
- Using coalescent simulations with marker -based estimates of demographic parameters
- Using the empirical site-frequency spectrum at reference locations as the null.
- Support via a preponderance of evidence, considers the joint signatures from a number of different tests



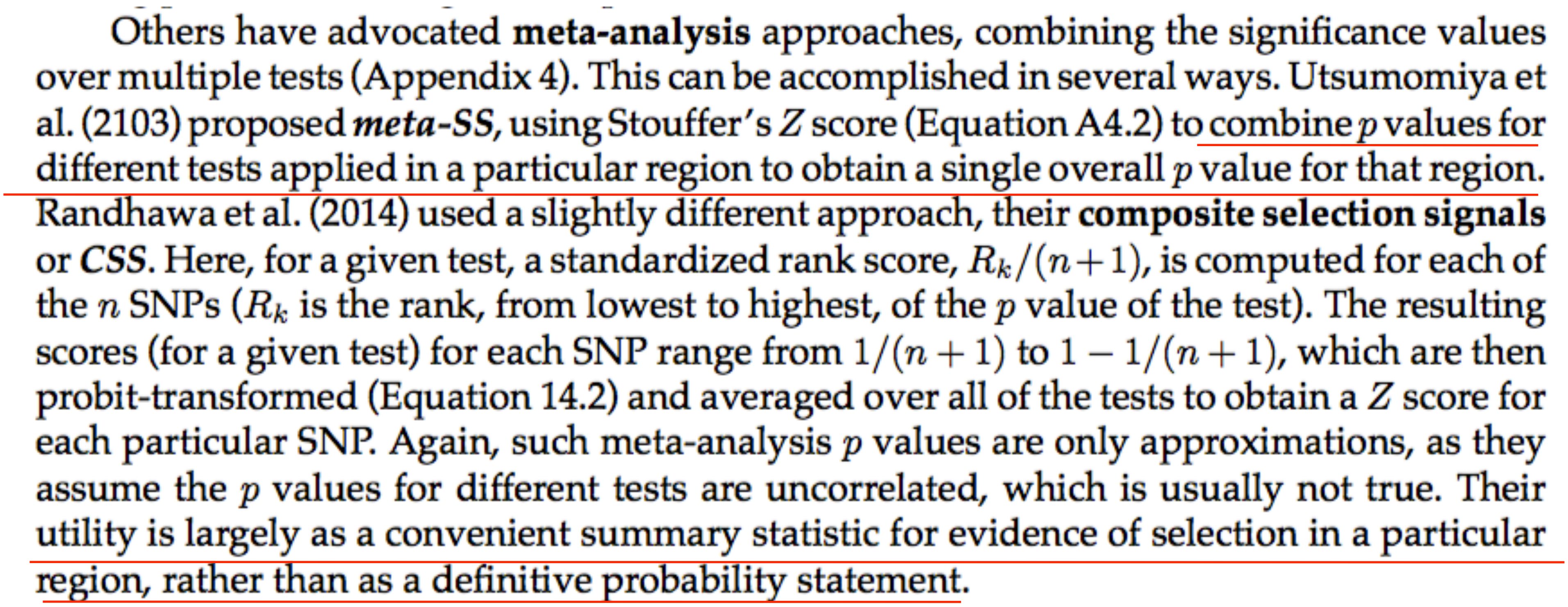
For African-Americans, the mean D is negative, while it is positive for European-Americans. A gene whose negative D value is significant under the equilibrium neutral model is likely to be even more significant in this European-American population (given this population's trend toward a positive D), but is problematic in this sample of African-Americans



The final approach is to use the empirical site-frequency spectrum vector, **p**, from a reference set—as opposed to the Watterson distribution—as the null (Nielsen et al. 2005b, 2009). Here p_i is the fraction of sites in the reference set with i copies of the allele (derived or minor, for the unfolded and folded spectra, respectively). A standard goodness-of-fit test (such as the *G*-test; LW Appendix 2) is then used to assess whether the spectrum $n_1, \cdots n_{n-1}$ in a candidate region is consistent with the multinominal probabilities given by **p**. One can also compare different parts of the spectrum, such as searching for an excess of low-frequency alleles, or high-frequency derived alleles, relative to this standard. Nielsen et al. (2009) used this approach for their ***MWU-low*** and ***MWU-high*** tests, respectively, where *MWU* stands for the Mann-Whitney U test (a common nonparametric test for comparing two groups, e.g., Conover 1999). One major reservation with these nonparametric approaches is the choice of the reference set of sites for the neutral background spectrum. Even if these site are neutral, local effects such as differences in the mutation rates (and hence in θ) and the background recombination rates that influence the levels of standing variation (Chapter 8) can result in the target sites (even if strictly neutral) differing from the distribution at reference sites. If one assumes background selection as the appropriate null, the sites used in constructing **p** should (at a minimum) come from genomic regions with very similar values of gene density to recombination rates as the tested region.

Support via a preponderance of evidence

- Composite of multiple signals, *CMS* (Grossman et al. 2010, 2013)
- Many test are correlated, not independent
- An unusual (but random under drift) genealogy will give rise to a number of different signals for that region



Others have advocated **meta-analysis** approaches, combining the significance values over multiple tests (Appendix 4). This can be accomplished in several ways. Utsumomiya et al. (2103) proposed ***meta-SS***, using Stouffer's Z score (Equation A4.2) to combine p values for different tests applied in a particular region to obtain a single overall p value for that region. Randhawa et al. (2014) used a slightly different approach, their **composite selection signals** or ***CSS***. Here, for a given test, a standardized rank score, $R_k/(n+1)$, is computed for each of the n SNPs (R_k is the rank, from lowest to highest, of the p value of the test). The resulting scores (for a given test) for each SNP range from $1/(n+1)$ to $1 - 1/(n+1)$, which are then probit-transformed (Equation 14.2) and averaged over all of the tests to obtain a Z score for each particular SNP. Again, such meta-analysis p values are only approximations, as they assume the p values for different tests are uncorrelated, which is usually not true. Their utility is largely as a convenient summary statistic for evidence of selection in a particular region, rather than as a definitive probability statement.

Ma et al. (2015) proposed a simple measure to deal with test correlations, their **de-correlated composite of multiple signals**, or *DCMS* statistic. Let $p_{i,k}$ denote the p value for test k for site i, and let r_{kj} be the empirical correlation among the values of the test statistics for tests k and j over all of the scored sites, so that $r_{kk} = 1$ and $r_{kj} = 0$ when tests k and j are uncorrelated. Ma's *DCMS* statistic for site i is given by

$$DCMS_i = \left(\frac{1}{W}\right)\sum_{k=1}^{t} \ln\left(\frac{1-p_{i,k}}{p_{i,k}}\right), \quad \text{where} \quad W = \sum_{k=1}^{t} |\, r_{kt}| \tag{9.29a}$$

The terms in the sum are the odds ratio for each test (which Ma et al. used in place of Bayes factors with equal prior weight on the null and alternative; see Equation A2.10b). The weighting term (W) ranges from 1 (none of the tests are correlated, so that $W = r_{tt} + 0 = 1$), to the case were all of the tests are perfectly correlated, so that $W = t$. In the former case, the composite measure is simply the sum of the odds ratios, while in the latter it is the average of the odds ratio. Ma at el. found in their simulations that *DCMS* had higher power than either *meta-SS* or *CSS* under most settings.

A final class of composite measures are **multivariate outlier metrics**. Just as the outlier approach is widely used to highlight sites that have exceptional values in a given single test statistics, one can also consider outliers from a *collection* of test statistics. Assuming all the tests have a mean of zero under the null, the total Euclidean distance of a vector of test statistics from the mean value under the null (**0**) would be one approach. However, different test statistics have different variances, and further they are correlated. One standard approach in such cases is to transform all of the tests statistics to have the same variance and to be uncorrelated, which leads to the **Mahalanobis distance** (Equation A5.19),

$$D_i^2 = \mathbf{t}_i^T \, \boldsymbol{\Sigma}_{\mathbf{t}}^{-1} \, \mathbf{t}_i \tag{9.29b}$$

where **t** is the vector of test statistics for site i and $\boldsymbol{\Sigma}_{\mathbf{t}}$ is the empirical variance-covariance matrix for the vector of test scores over all of the sites. Lotterhos et al. (2017) used this metric and a variant replacing the vector (**t**) of test statistics with a vector whose elements were based on the ranks of the p values for a given site (along the lines of Randhawa et al. 2014). They then took the negative log of these rank-based p values as the elements of **t** for the Mahalanobis distance. This approach goes by the compact name of **Mahalanobis distance based on negative-log rank-based p-values**, or ***Md-rank-P***. They found that this approach worked the best of the composite measures they tested, followed by *DCMS*.

Recombination Makes Site-frequency Tests Conservative

A final comment on frequency-spectrum tests is that, ignoring demographic concerns, they are likely conservative in many settings. In particular, Wall (1999) noted that site-frequency spectrum tests all assume that there is no recombination within the region of interest. While recombination does not bias the expected values for various statistics, it does *reduce* their variances (Rozas et al. 1999: Wall 1999), as the observed values represent the average across several genealogies (Depaulis et al. 2003). As a result, when recombination *does* occur within a region, tests are *conservative*, with the true p value being smaller than the zero-recombination values tabulated by the original authors of the various tests. As a result of this conservative nature of SFS tests under recombination, they are often significantly *underpowered*, using more stringent critical values than necessary. Wall found this effect to be significant when the rate of recombination is on the order of the total regional mutation rate, as is often the case (Table 4.1). Coalescent simulations allowing for recombination can significantly improve the power of tests by obtaining more accurate p values. As discussed in Chapter 4, the four-gamete test (Hudson and Kaplan 1985) can be used to detect recombination in the coalescence history of the sample, and the R_M statistic suggested by these authors estimates the minimal number of recombinants in the sample, which can then be incorporated into an appropriate coalescent simulation (e.g., Depaulis et al. 2005).

SFS tests: Summary

- Tajima’s D
  - Negative values (excess of rare alleles) following a sweep
  - Problem: changes in population size also generates this
- Fay and Wu’s H
  - Test for excess of high frequency derived alleles

Haplotype-based tests

- Allele frequency spectrum
  - Ewens sampling formula
  - Number of alleles, heterozygosity
- LD signals (soft and hard sweeps different signals)
- Age of alleles
  - Long haplotypes
  - Inconsistent estimates

Defining and Inferring Haplotypes

If one considers a sufficiently long stretch of DNA, every sequence is a unique haplotype, so just how are haplotypes defined? The answer depends on both the test being used and the features of LD that are of interest. If we are interested in number and diversity of haplotypes in an infinite-alleles framework, the unit of analysis is a sufficiently small region, ideally with no recombination observed in the sample. The four-gamete test of Hudson and Kaplan (1985) can be used to detect recombination in the sample (Chapter 4), helping to define the size of a region (for example, by setting the size of a sliding window moving through a larger region). Practically, one may be constrained to find regions with sufficient haplotype diversity for analysis given either the marker density or background levels of variation, so that small amounts of recombination within the defined region may appear in the sample. For tests based on the average pairwise disequilibrium among all sites within a region, one actually wants some (but not too much) recombination. Finally, tests based on long haplotypes require a **core haplotype** (either a single SNP or a set of a few tightly linked SNPs) to define distinct allelic classes, with the disequilibrium patterns within each class (i.e., as one moves away from the core) forming the basis of tests. Again, recombination (outside of the core) is critical to these tests.

Overview of Haplotype-based Tests

As reviewed in Table 9.2, a number of haplotype features can be used as the basis for tests of ongoing selection. **Strong haplotype structure** occurs when there are fewer haplotypes than expected given the number, S, of segregating sites within a region. This *underdispersion* of haplotypes is a signature of excessive LD within a region. Strong haplotype structure also results in a deficiency in **haplotype diversity**, H (the probability that two random haplotypes from the sample are different, analogous to Π under the infinite-sites model), and an excess of high-frequency haplotypes (roughly analogous to Fay and Wu's H test; Equation 9.27b). Such signatures are created by any process generating a coalescent with long internal branches (relative to the equilibrium neutral model; see Figure 8.3), such as a partial sweep (the favorable allele is not yet fixed), recovery from a moderate bottleneck, balancing selection, or population structure. Conversely, we can have the opposite pattern (*overdispersion* of haplotypes), with an excess of haplotypes, excess haplotype diversity, and an excess of rare-frequency haplotypes. Such signals are generated by a star-like coalescent genealogy, as would occur near the conclusion of a hard sweep, or the recovery from an extreme population bottleneck. However, in these overdispersed settings, LD summary statistics typically have low power, as S is small (most of the variation is removed), so that while haplotype overdispersion occurs, its signal is often weak.

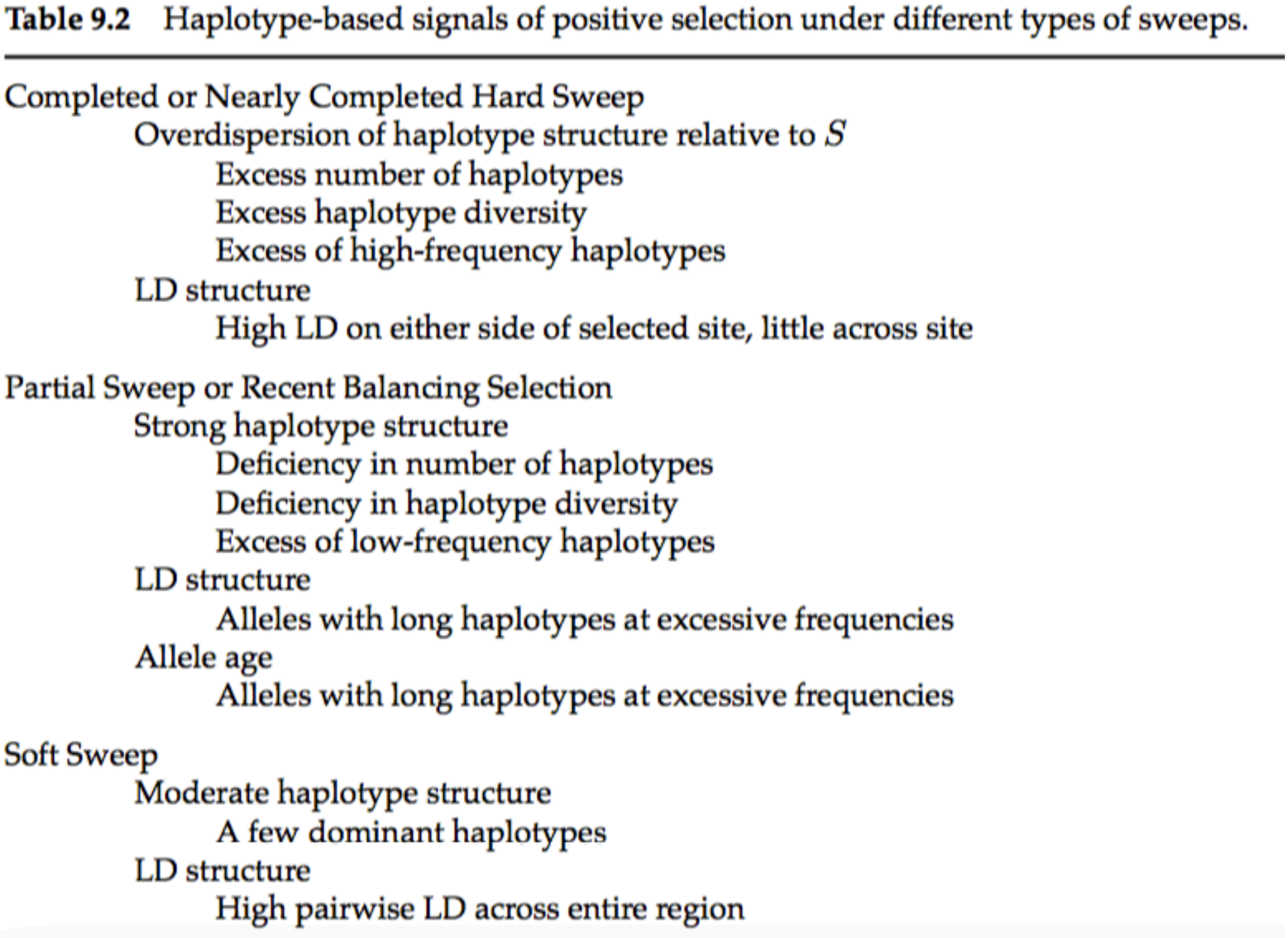


Table 9.2 Haplotype-based signals of positive selection under different types of sweeps.

- Completed or Nearly Completed Hard Sweep
  - Overdispersion of haplotype structure relative to S
    - Excess number of haplotypes
    - Excess haplotype diversity
    - Excess of high-frequency haplotypes
  - LD structure
    - High LD on either side of selected site, little across site
- Partial Sweep or Recent Balancing Selection
  - Strong haplotype structure
    - Deficiency in number of haplotypes
    - Deficiency in haplotype diversity
    - Excess of low-frequency haplotypes
  - LD structure
    - Alleles with long haplotypes at excessive frequencies
  - Allele age
    - Alleles with long haplotypes at excessive frequencies
- Soft Sweep
  - Moderate haplotype structure
    - A few dominant haplotypes
  - LD structure
    - High pairwise LD across entire region

Infinite alleles: Ewen's sampling formula

- Number of alleles, k, in a sample of size n

$$\Pr(k \,|\, \theta_L, n) = \frac{S_n^k \, \theta_L^k}{S_n(\theta_L)}$$

$$S_n(\theta_L) = \theta_L(\theta_L + 1)(\theta_L + 2)\cdots(\theta_L + n - 1)$$

Ewen’s (cont)

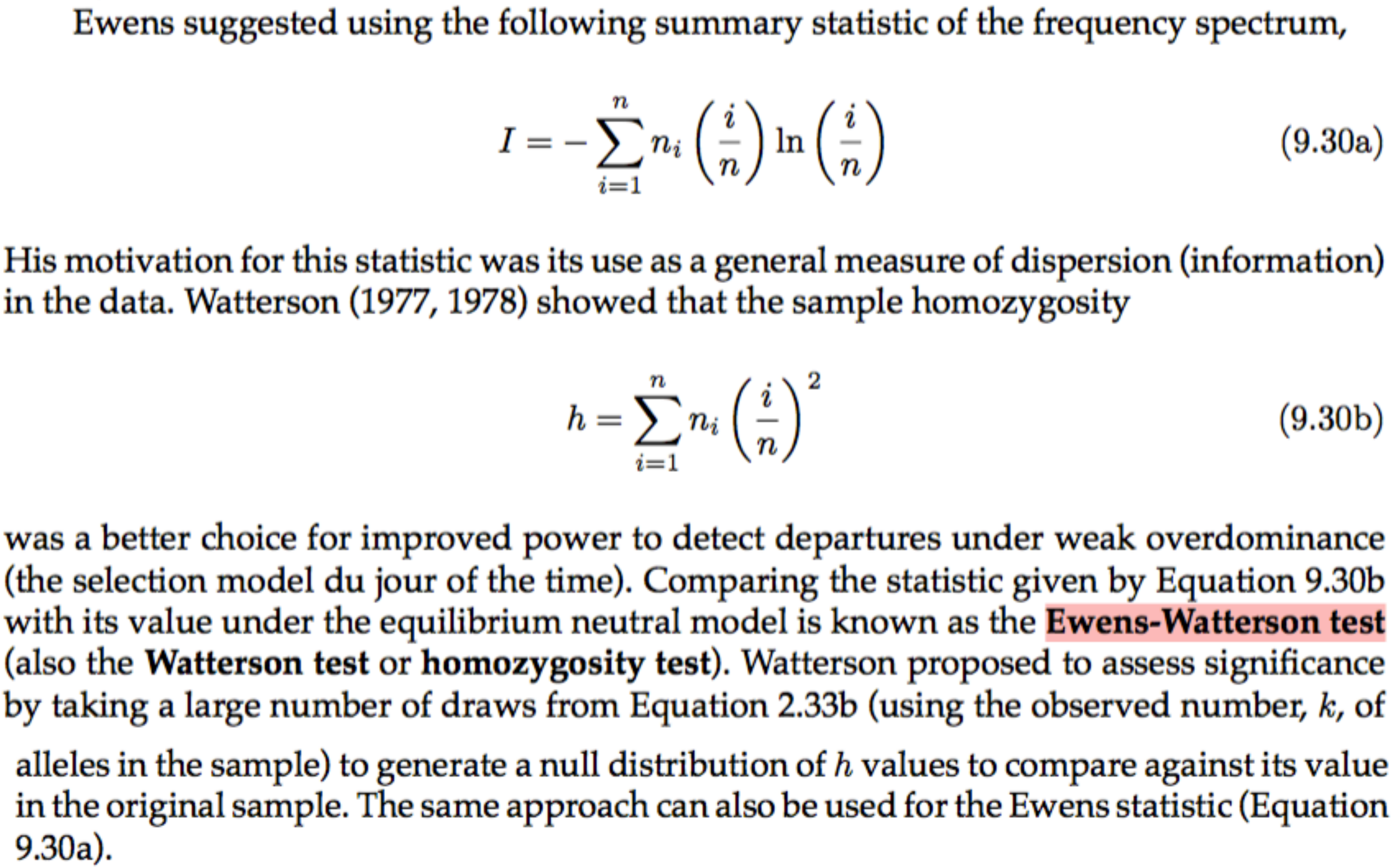
- Prob. Monomorphic

$$\Pr(k=1) = \frac{(n-1)!}{(\theta_L+1)(\theta_L+2)\cdots(\theta_L+n-1)}$$

- Mean and variance in k

$$E(k) = 1 + \theta_L \cdot \sum_{j=2}^{n} \frac{1}{\theta_L + j - 1}, \qquad \sigma^2(k) = \theta_L \cdot \sum_{j=1}^{n-1} \frac{j}{(\theta_L + j)^2}$$

Ewens-Watterson test



Ewens suggested using the following summary statistic of the frequency spectrum,

$$I = -\sum_{i=1}^{n} n_i \left(\frac{i}{n}\right) \ln \left(\frac{i}{n}\right) \tag{9.30a}$$

His motivation for this statistic was its use as a general measure of dispersion (information) in the data. Watterson (1977, 1978) showed that the sample homozygosity

$$h = \sum_{i=1}^{n} n_i \left(\frac{i}{n}\right)^2 \tag{9.30b}$$

was a better choice for improved power to detect departures under weak overdominance (the selection model du jour of the time). Comparing the statistic given by Equation 9.30b with its value under the equilibrium neutral model is known as the **Ewens-Watterson test** (also the **Watterson test** or **homozygosity test**). Watterson proposed to assess significance by taking a large number of draws from Equation 2.33b (using the observed number, k, of alleles in the sample) to generate a null distribution of h values to compare against its value in the original sample. The same approach can also be used for the Ewens statistic (Equation 9.30a).

Other Infinite-alleles Tests: Conditioning on $\widehat{\theta}$

Watterson-type tests use the *conditional* allele-frequency spectrum, where the observed number of alleles, k, is used in Equation 2.33b to generate the null distribution. What about tests based on k itself, such as whether there are too many, or too few, alleles based on some other diversity measure? Such tests use the sampling distributions given by either Equation 2.30a or Equation 2.33a, and require an estimate of θ. Fu (1996, 1997) used this approach to test whether a sample contains too many, or too few, alleles (haplotypes) relative to the neutral equilibrium model. His **W test** (1996) used the Ewens sampling formula (Equation 2.30a) with θ replaced by the Watterson estimator, $\widehat{\theta}_S$ (Equation 4.3a), and it returns the probability of seeing k (or fewer) alleles in the sample as

$$W = \Pr(K \leq k) = \sum_{i=1}^{k} \Pr(K = i \,|\, \widehat{\theta}_S, n) = \sum_{i=1}^{k} \frac{S_n^i \cdot [\widehat{\theta}_S\,]^i}{S_n(\widehat{\theta}_S)} \tag{9.32}$$

where S_n^i is the coefficient on $(\widehat{\theta}_S)^i$ in the polynomial

$$S_n(\widehat{\theta}_S) = \widehat{\theta}_S\,(\widehat{\theta}_S + 1)(\widehat{\theta}_S + 2)\cdots(\widehat{\theta}_S + n - 1)$$

Fu's $\boldsymbol{F_S}$ **test** (1997) is the compliment of W, as it tests for an *excess of rare alleles/haplotypes*. It starts by computing the probability of seeing k *or more* alleles/haplotypes in a sample,

$$S' = \Pr(K \geq k) = \sum_{i=k}^{n} \frac{S_n^i \cdot [\,\widehat{\theta}_\Pi\,]^i}{S_n(\widehat{\theta}_\Pi)} \tag{9.33a}$$

but now using $\widehat{\theta}_\Pi$, the estimator of θ based on average number of pairwise differences (which is more sensitive to sites with intermediate allele frequencies). Fu noted that S' is not an optimal test statistic because its critical values are often too close to zero. Because of this, the test uses the transformation

$$F_S = \ln\left(\frac{S'}{1 - S'}\right) \tag{9.33b}$$

As with W, this is also a one-sided test. F_S is negative when there is an excess of rare alleles/haplotypes (as would occur with a selective sweep or population expansion), with a sufficiently large negative value serving as evidence for selection or population expansion. Fu (1997) showed that F_S is more powerful that Tajima's D and the Fu-Li D^* and F^* tests (Table 9.1) for detecting selective sweeps or population expansion following a bottleneck.

Depaulis and Veuille (1998) also used conditioning on S and developed two tests. Their **haplotype number**, or ***K*, test** is essentially Fu's W test (Equation 9.32), but using $\widehat{\theta}_S$ (and hence conditioning on S) rather than $\widehat{\theta}_\Pi$. Their **haplotype diversity**, or ***H*, test**, uses the statistic

$$H = 1 - \sum_{i=1}^{k} p_i^2 \quad \text{with} \quad p_i = \text{ frequency of the } i\text{th haplotype} \tag{9.34a}$$

namely, the haplotype heterozygosity, which is compared to its expected neutral equilibrium value given S. A comparison with Equation 9.30b shows that the H test is essentially the Ewens-Watterson test, but with its significance assessed by conditioning on S rather than k. Note that the range on H is

$$\frac{2(n-1)}{n^2} \leq H \leq 1 - \frac{1}{n} \tag{9.34b}$$

with the lower bound set by the sample consisting of just two haplotypes, one with $n-1$ copies and the other a singleton ($n_{n-1} = 1, n_1 = 1$), while the upper range is set by all of the haplotypes being present as singletons ($n_1 = n$). Critical values for these statistics (conditioned on n and S) generated from coalescent simulations were tabulated by Depaulis and Veuille (1998).

Garud et al.'s H_{12} and H_2 Tests

A number of tests are built around **haplotype homozygosity (*HH*)**, the probability that two randomly chosen haplotypes are identical. This is given by the complement of the Depaulis-Veuille H (haplotype heterozygosity) statistic (Equation 9.34a),

$$H_1 = 1 - H = \sum_{i=1}^{k} p_i^2 \tag{9.35a}$$

where p_i is the frequency of the ith haplotype in the sample. To adjust for sampling, some variants of this statistic replace p_i^2 with $[\,p_i + (1/k)]^2$, where k is the number of haplotypes (e.g., Kemper et al. 2014). As mentioned in Chapter 8, Garud et al. (2015) showed that a simple modification of this statistic results in a test that can detect *both* hard and soft sweeps. Their $\boldsymbol{H_{12}}$ test statistic combines the two largest haplotype classes into a single one,

$$H_{12} = (p_1 + p_2)^2 + \sum_{i>3} p_i^2 = H_1 + 2p_1p_2 \tag{9.35b}$$

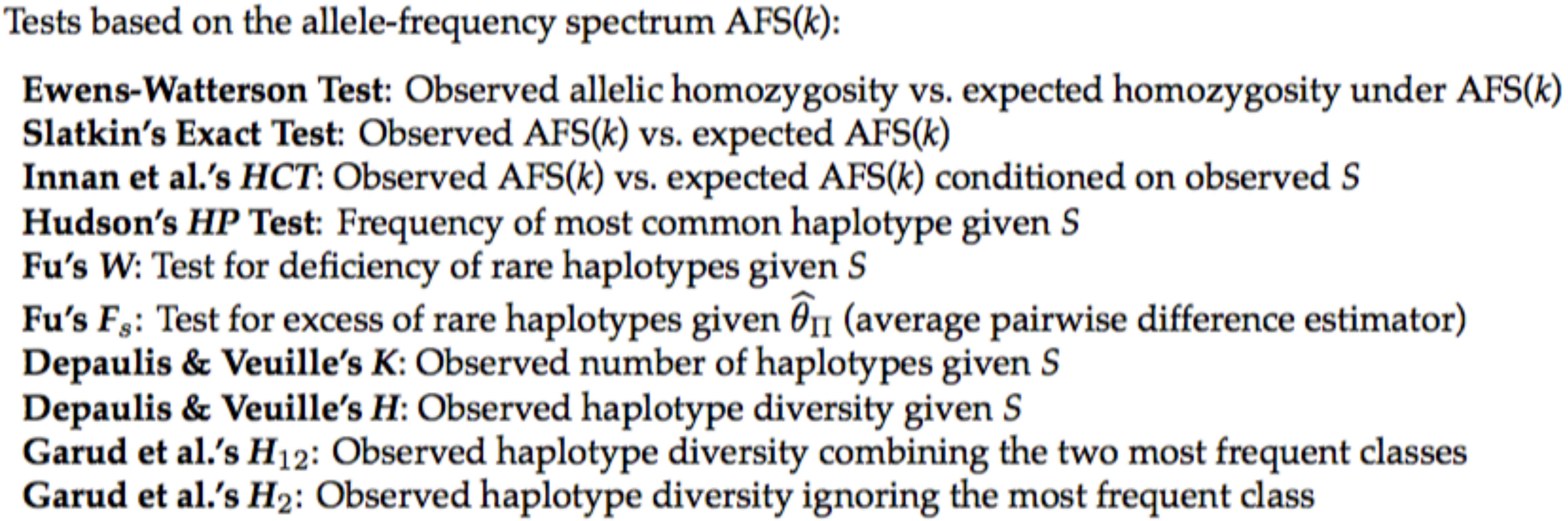
The logic is that a soft sweep results in not one, but several, dominant haplotypes. If the sweep is not too soft, then the first two haplotypes, both presumably harboring the favored allele, will together comprise most of the haplotype variation. In the case of a hard sweep, the second-most frequent haplotype will be sufficiently rare that $H_{12} \simeq H_1$. The authors applied this approach to *Drosophila*, looking at windows with a fixed number of SNPs and adjusting for the local recombination rate and then used coalescent simulations to generate values under the null of neutrality.

Garud et al. considered a second modified *HH* statistic, namely, the homozygosity with the largest class removed

$$H_2 = \sum_{i>1} p_i^2 \tag{9.35c}$$

Under a hard sweep with its single dominant haplotype, H_2 should be considerably smaller than H_1, while under a soft sweep the drop-off in value from H_1 to H_2 should be much less dramatic. Based on this observation, the ratio H_2/H_1 forms the basis of a test as to whether a detected sweep is hard or soft, with moderate values suggesting soft sweeps and very small values suggesting hard sweeps (Garud et al. 2015; Garud and Rosenberg 2015).

Summary of tests based on the allele-frequency spectrum (AFS)

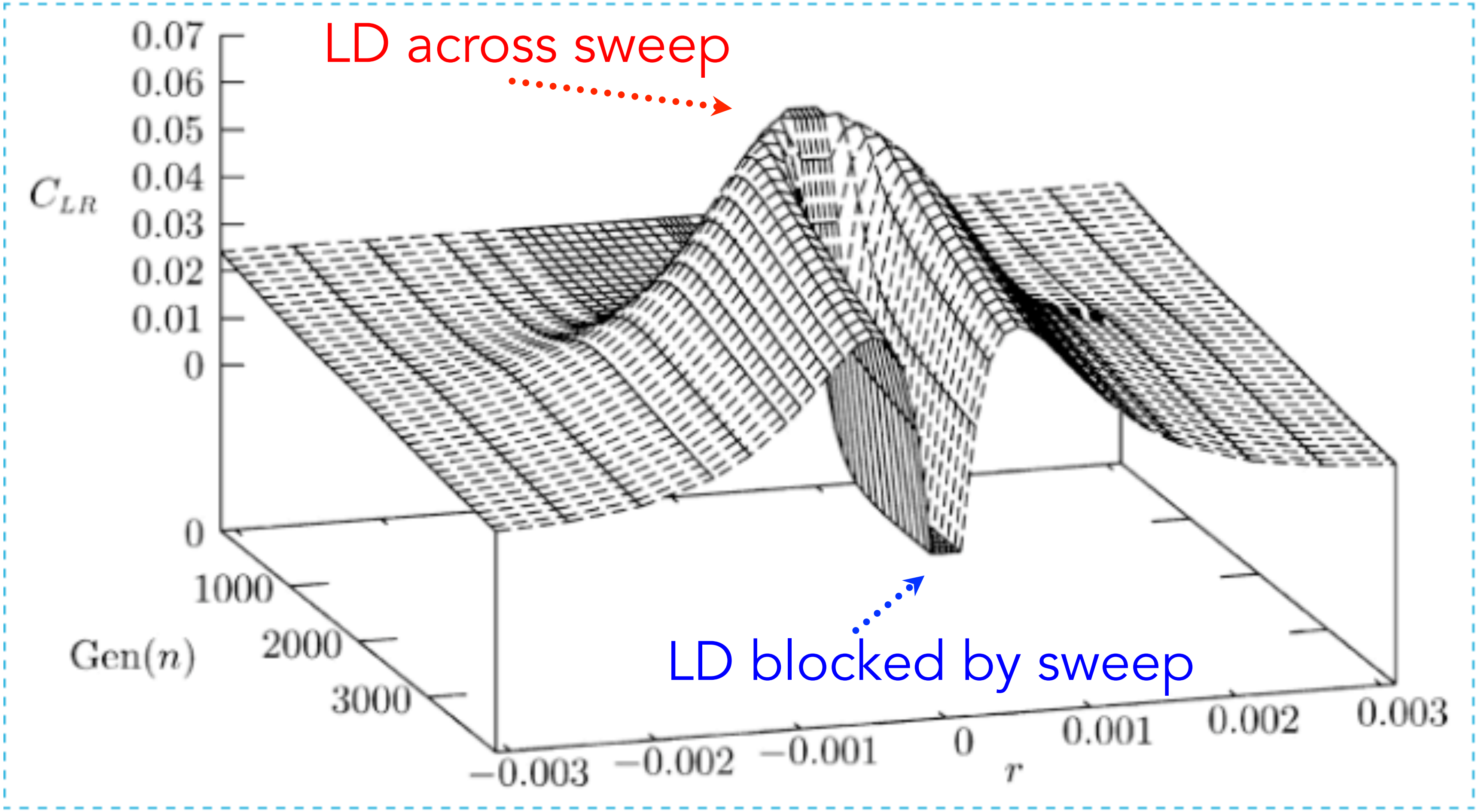


Tests based on the allele-frequency spectrum AFS(k):

Ewens-Watterson Test: Observed allelic homozygosity vs. expected homozygosity under AFS(k)
Slatkin's Exact Test: Observed AFS(k) vs. expected AFS(k)
Innan et al.'s *HCT*: Observed AFS(k) vs. expected AFS(k) conditioned on observed S
Hudson's *HP* Test: Frequency of most common haplotype given S
Fu's *W*: Test for deficiency of rare haplotypes given S
Fu's F_s: Test for excess of rare haplotypes given $\widehat{\theta}_{\Pi}$ (average pairwise difference estimator)
Depaulis & Veuille's *K*: Observed number of haplotypes given S
Depaulis & Veuille's *H*: Observed haplotype diversity given S
Garud et al.'s H_{12}: Observed haplotype diversity combining the two most frequent classes
Garud et al.'s H_2: Observed haplotype diversity ignoring the most frequent class

Test based upon LD measures

Distribution of LD around a hard sweep



LD across sweep

LD blocked by sweep

At start, LD through sweep site. At fixation, strong LD on either side, but not through, a sweep site

LD-based sweep tests

Kim and Nielson's ω statistic: LW with vs. across a test region

$$\omega = C_{S,\ell} \frac{\sum_{i,j \in L} r_{ij}^2 + \sum_{i,j \in L} r_{ij}^2}{\sum_{i \in L, j \in R} r_{ij}^2}, \qquad C_{S,\ell} = \frac{1/(\ell(S-\ell)}{\binom{\ell}{2} + \binom{S-\ell}{2}}$$

Strong test for hard sweeps, little power for soft sweeps

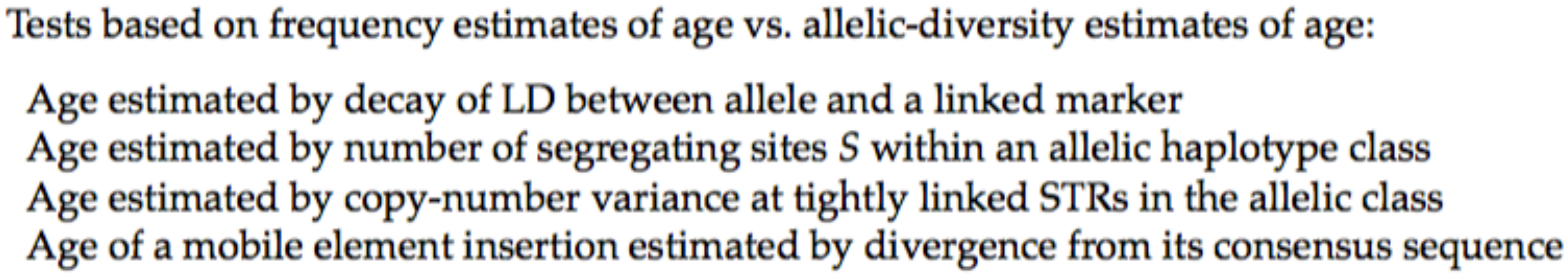
Kelly's Z_{nS} Statistic: Average pairwise LD through a region

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{ij}^2$$

Strong test for soft sweeps, little power for hard sweeps

Tests based upon allelic age

- The frequency of a neutral allele can be used to estimate its age
  - Can contrast this age estimator with others (STR variation, recombination, etc)
- Key idea: under drift, a common allele is an old allele
  - Hence, "long haplotypes" should not be found for common alleles

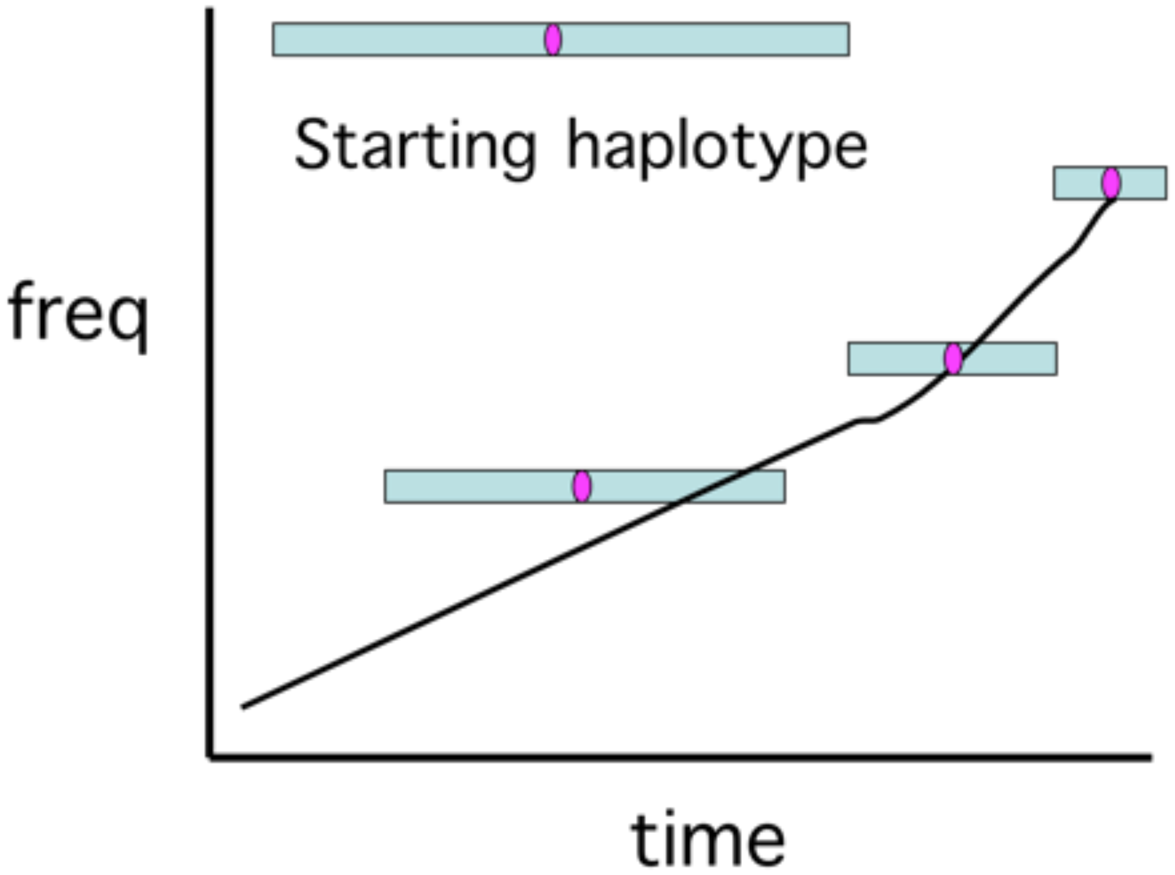


Tests based on frequency estimates of age vs. allelic-diversity estimates of age:

- Age estimated by decay of LD between allele and a linked marker
- Age estimated by number of segregating sites *S* within an allelic haplotype class
- Age estimated by copy-number variance at tightly linked STRs in the allelic class
- Age of a mobile element insertion estimated by divergence from its consensus sequence

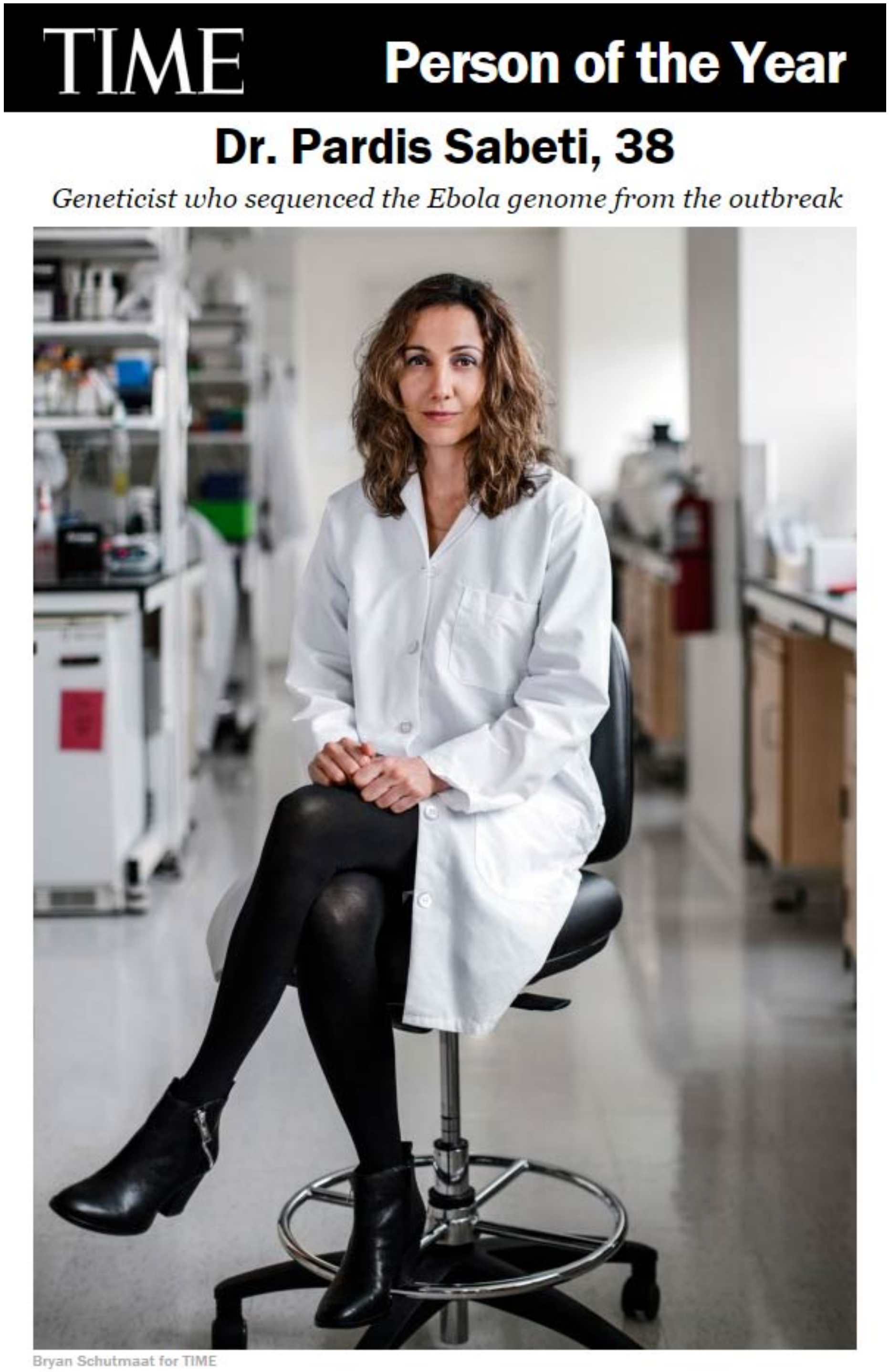
See Chapter 9 for examples and methods

Long haplotype tests



Common alleles should have short haplotypes under drift -- longer time for recombination to act

Common alleles with long haplotypes --- good signal for selection, rather robust to demography



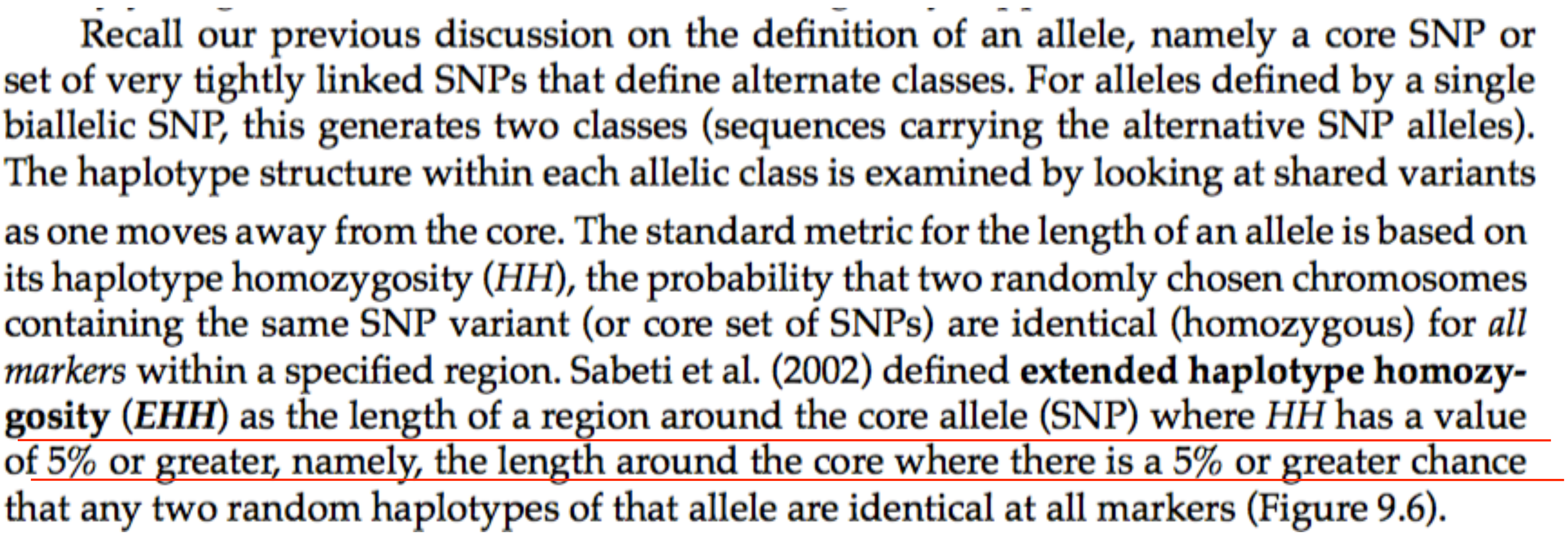
TIME Person of the Year

Dr. Pardis Sabeti, 38

Geneticist who sequenced the Ebola genome from the outbreak

Bryan Schutmaat for TIME

Extended haplotype homozygosity (*EHH)*



Recall our previous discussion on the definition of an allele, namely a core SNP or set of very tightly linked SNPs that define alternate classes. For alleles defined by a single biallelic SNP, this generates two classes (sequences carrying the alternative SNP alleles). The haplotype structure within each allelic class is examined by looking at shared variants as one moves away from the core. The standard metric for the length of an allele is based on its haplotype homozygosity (*HH*), the probability that two randomly chosen chromosomes containing the same SNP variant (or core set of SNPs) are identical (homozygous) for *all markers* within a specified region. Sabeti et al. (2002) defined **extended haplotype homozygosity (*EHH*)** as the length of a region around the core allele (SNP) where *HH* has a value of 5% or greater, namely, the length around the core where there is a 5% or greater chance that any two random haplotypes of that allele are identical at all markers (Figure 9.6).

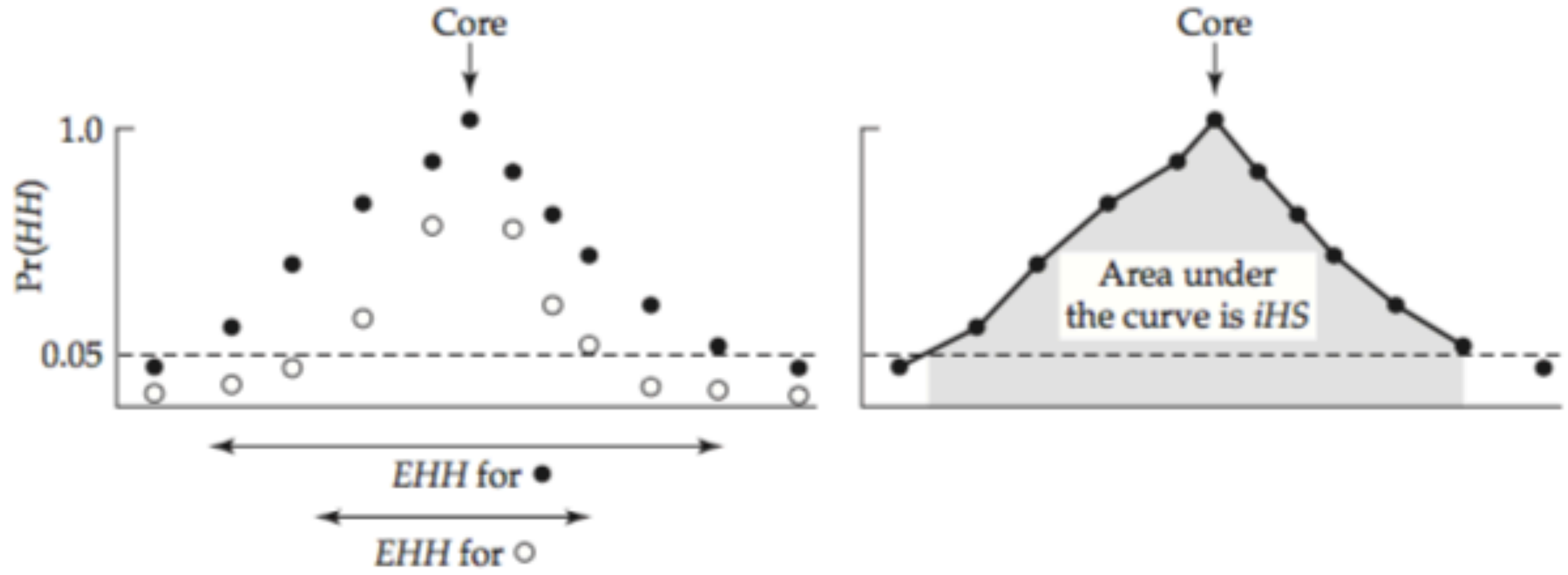


Figure 9.6 Haplotype homozygosity (*HH*) is defined as the probability that two randomly chosen chromosomes containing the same core SNP variant (used to define allelic classes) are identical (homozygous) at all markers within a defined window. In the figure, *HH* is computed at a series of SNP markers moving away from the core (allelic-defining) SNP. The open and filled circles correspond to the *HH* values at a given SNP in the two allelic classes, namely, the probability that random draws of chromosomes from the same allelic class are identical within the region between the core SNP and the marker SNP. The relationship between *HH* and distance from the core is usually summarized using one of two statistics. (**Left**) The extended haplotype homozygosity (*EHH*) for an allelic class is the length of the region around the core where the *HH* value is ≥ 5% (above the dashed line). The allele corresponding to the filled circles has a larger *EHH* value, and thus a longer haplotype. (**Right**) A potentially more informative measure is given by the **integrated EHH score**, *iHS*, the total area under the *HH* curve over the region spanned by the *EHH* for that allele. For ease of presentation, only the values corresponding to the allele with the larger *EHH* value (filled circles) are plotted.

While alleles with excessive values of *EHH* are produced by partial sweeps, simply scanning for sites with large *EHH* values will not serve as a sufficient indicator of selection, as a localized decrease in the recombination rate inflates the *EHH* value. The formal use of *EHH* as a selection-detecting statistic thus requires an internal control. Sabeti et al. (2002) proposed considering the **relative extended haplotype homozygosity (*rEHH*)** of a particular allele (SNP variant), defined as the ratio of the *EHH* value for that allele divided by the average *EHH* value for all other core alleles at the focal locus. For allele i, this is given by

$$rEHH_i = \frac{EHH_i}{\text{ave}(EHH_j) \text{ for } j \neq i} \tag{9.40}$$

where ave(EHH_j) denotes the average *EHH* values for all SNPs at the allelic-defining site. For the biallelic case (an allele defined by a single SNP, as opposed to a collection of tightly linked SNPs), *rEHH* is simply the ratio of the *EHH* values for the two alleles. By contrasting different alleles at the same site, most concerns about local variation in the recombination rates are ameliorated. However, if there are haplotype-specific recombination rates (e.g., the insertion of a mobile element reducing local recombination rates; Macpherson et al. 2008), then this test may be compromised. One consequence of comparing different alleles at a site is that as one allele approaches fixation, the power of the test disappears, as there are too few individuals in the comparison class to produce a meaningful statistic. As a result, the *rEHH* test has a rather narrow time window for the detection of a sweep: a rough rule is that the frequency of the favored allele must be 0.7 or less. Within such a time window, this test is among the most powerful for detecting selection. Nonetheless, a large *rEHH* value is not sufficient for suggesting selection, as some rare alleles (potentially being very young, and hence with reduced time for recombination) are expected to have large *rEHH* values. To

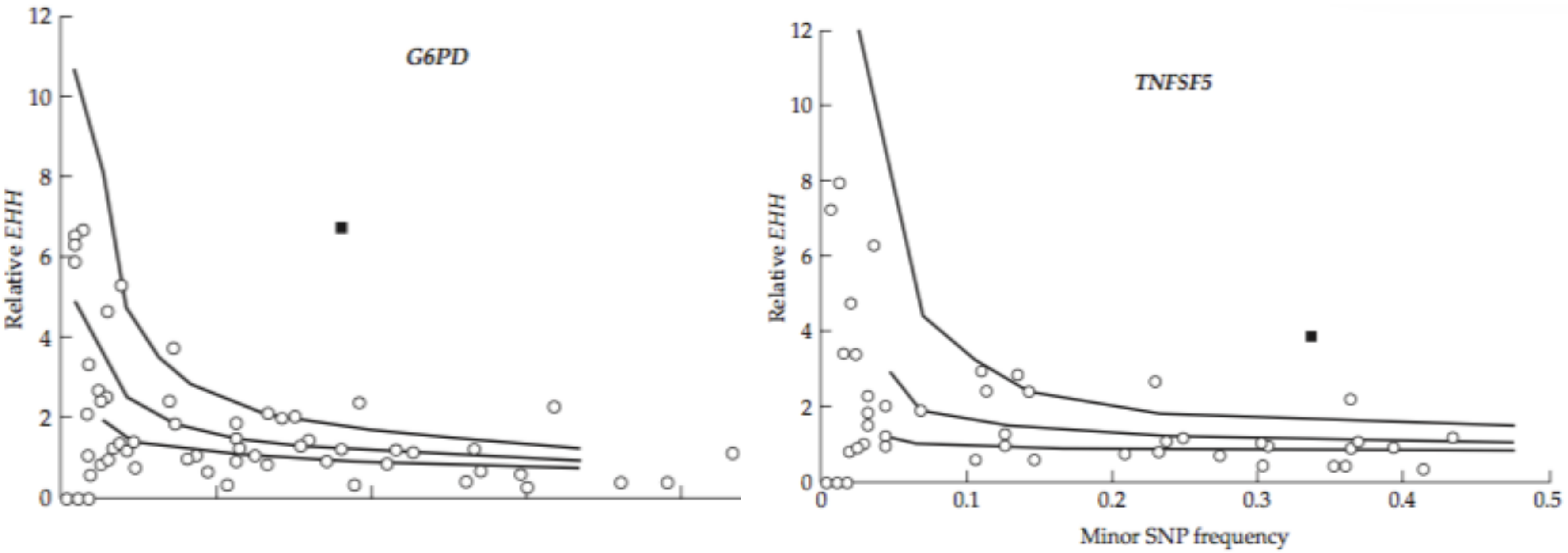
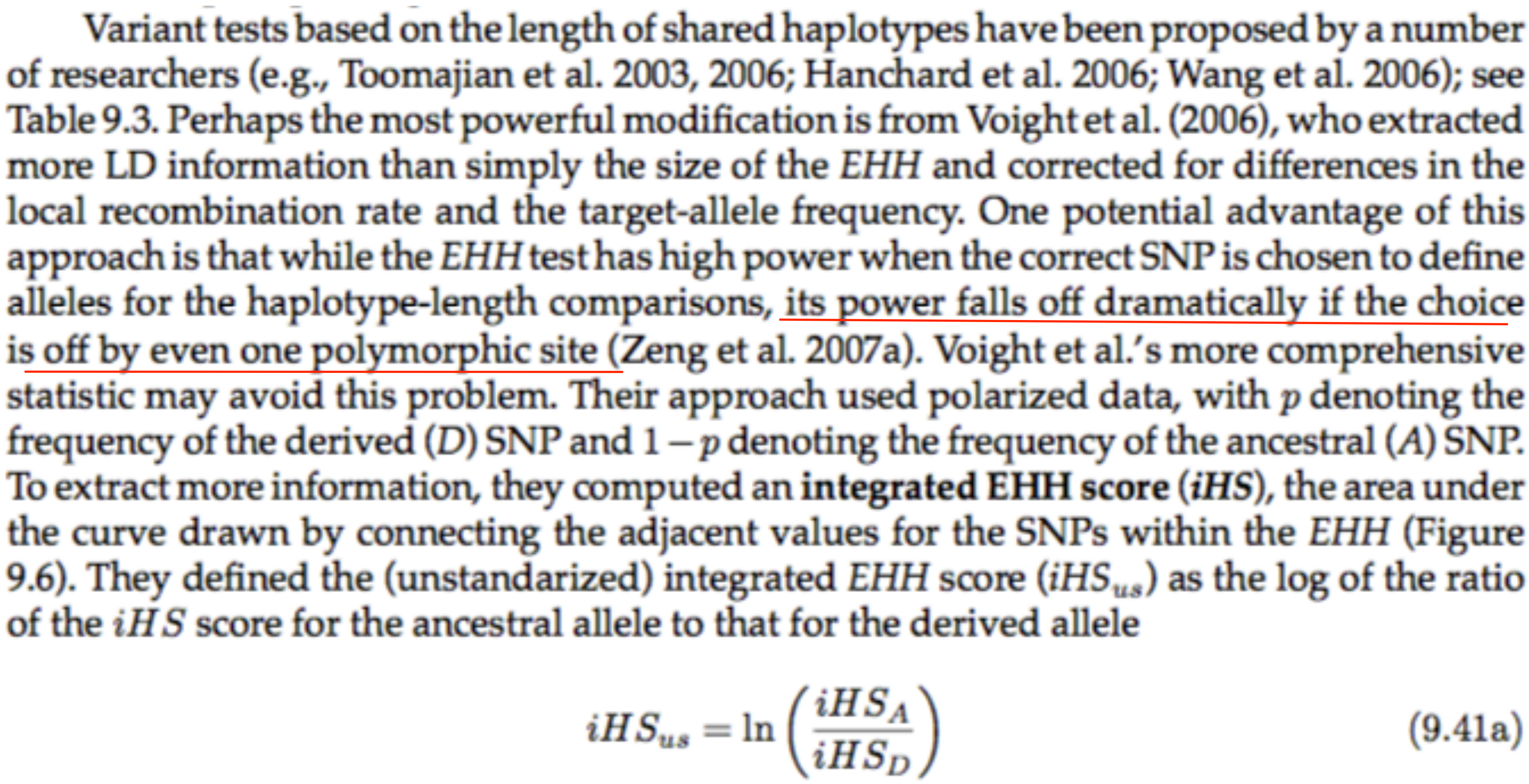


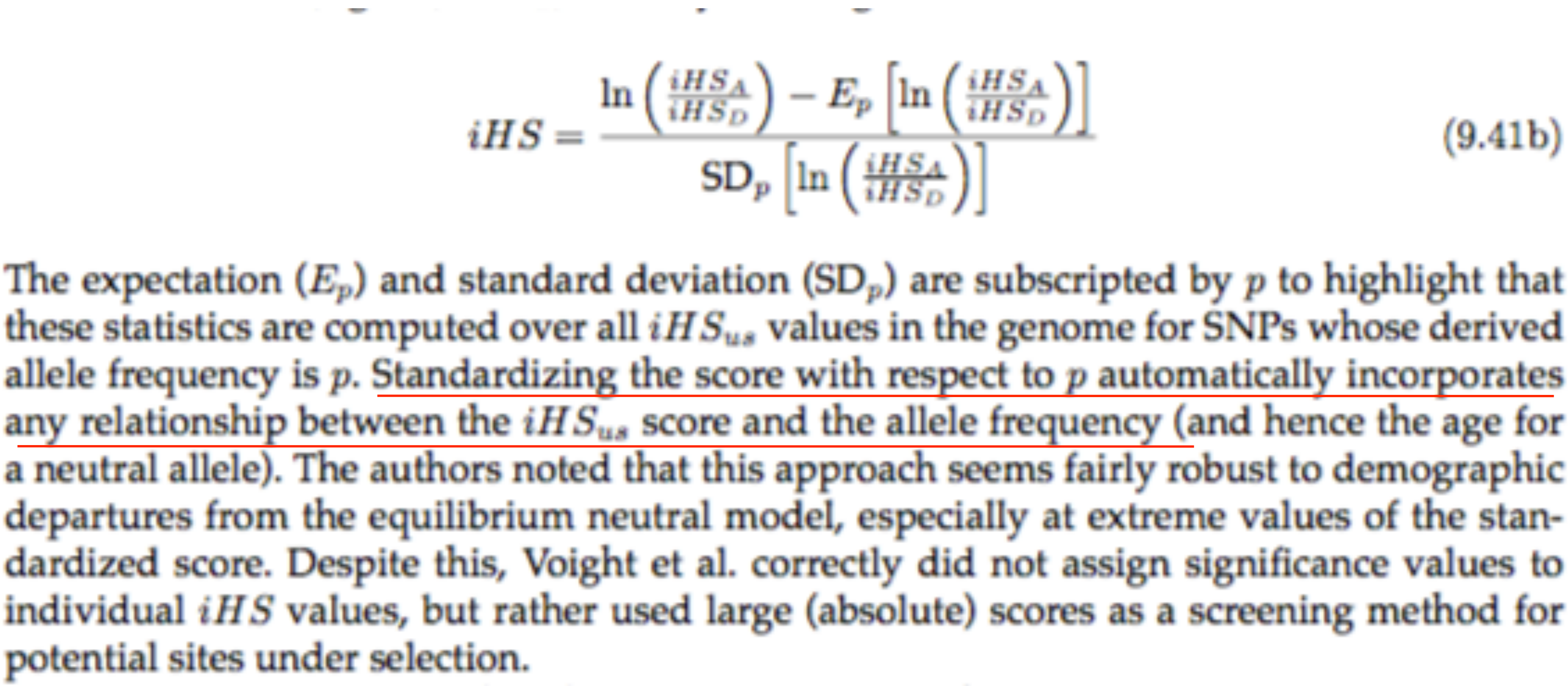
Figure 9.7 As a proof-of-concept of the *rEHH* method, Sabeti et al. (2002) looked for signatures of selection at two loci, *G6PD* and the *CD40* ligand gene (*TNFSF5*), that carry segregating alleles that are strongly suspected of increasing resistance to malaria. Standard site-frequency tests (Taijma's *D*, Fu and Li's *D**, and Fay and Wu's *H*; see Table 9.1) were all nonsignificant. However, recall from Chapter 8 that site-frequency spectrum signals are weak when the favored allele is at a modest frequency. The figure displays *rEHH* versus allele frequency for the candidate alleles (solid squares) along with values for alleles at other randomly chosen autosomal loci (open circles). The curves (from top to bottom) correspond to the empirical 95th, 75th, and 50th percentiles, respectively, of the cumulative distribution. (After Sabeti et al. 2002.)

Integrated EHH score (*iHS*)



Variant tests based on the length of shared haplotypes have been proposed by a number of researchers (e.g., Toomajian et al. 2003, 2006; Hanchard et al. 2006; Wang et al. 2006); see Table 9.3. Perhaps the most powerful modification is from Voight et al. (2006), who extracted more LD information than simply the size of the *EHH* and corrected for differences in the local recombination rate and the target-allele frequency. One potential advantage of this approach is that while the *EHH* test has high power when the correct SNP is chosen to define alleles for the haplotype-length comparisons, its power falls off dramatically if the choice is off by even one polymorphic site (Zeng et al. 2007a). Voight et al.'s more comprehensive statistic may avoid this problem. Their approach used polarized data, with p denoting the frequency of the derived (D) SNP and $1-p$ denoting the frequency of the ancestral (A) SNP. To extract more information, they computed an **integrated EHH score** (***iHS***), the area under the curve drawn by connecting the adjacent values for the SNPs within the *EHH* (Figure 9.6). They defined the (unstandarized) integrated *EHH* score (iHS_{us}) as the log of the ratio of the iHS score for the ancestral allele to that for the derived allele

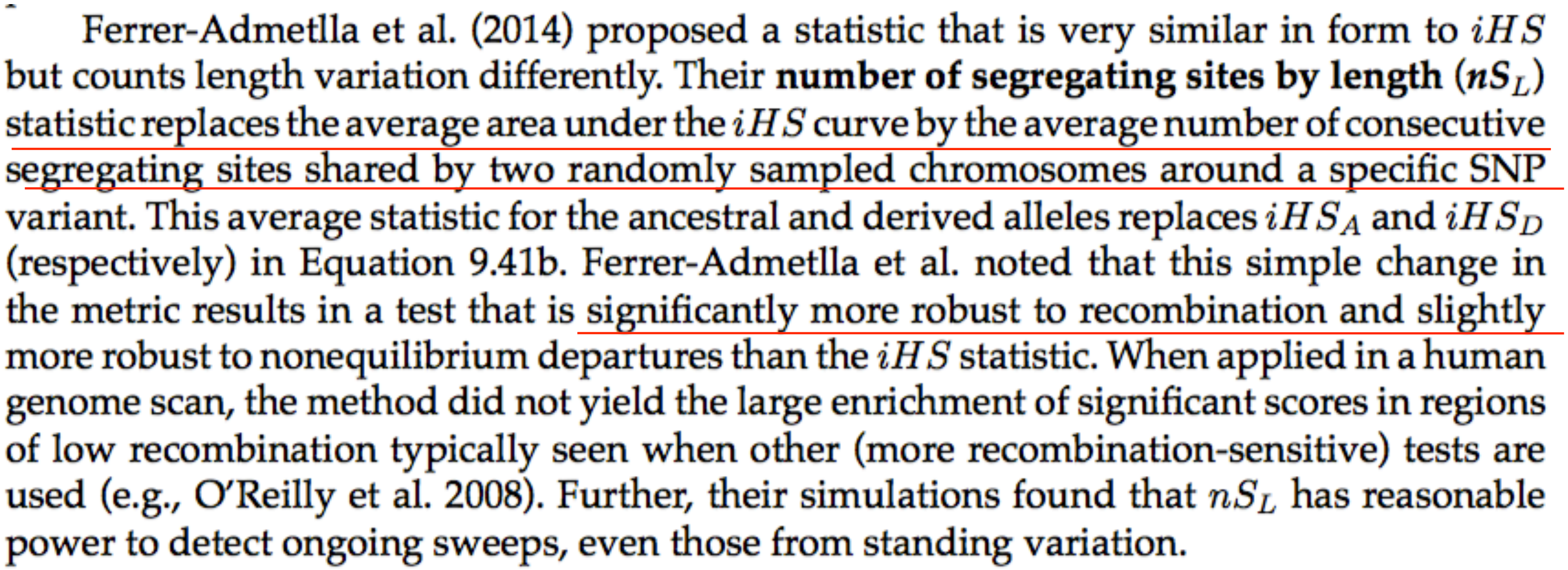
$$iHS_{us} = \ln\left(\frac{iHS_A}{iHS_D}\right) \tag{9.41a}$$



$$iHS = \frac{\ln\left(\frac{iHS_A}{iHS_D}\right) - E_p\left[\ln\left(\frac{iHS_A}{iHS_D}\right)\right]}{\mathrm{SD}_p\left[\ln\left(\frac{iHS_A}{iHS_D}\right)\right]} \tag{9.41b}$$

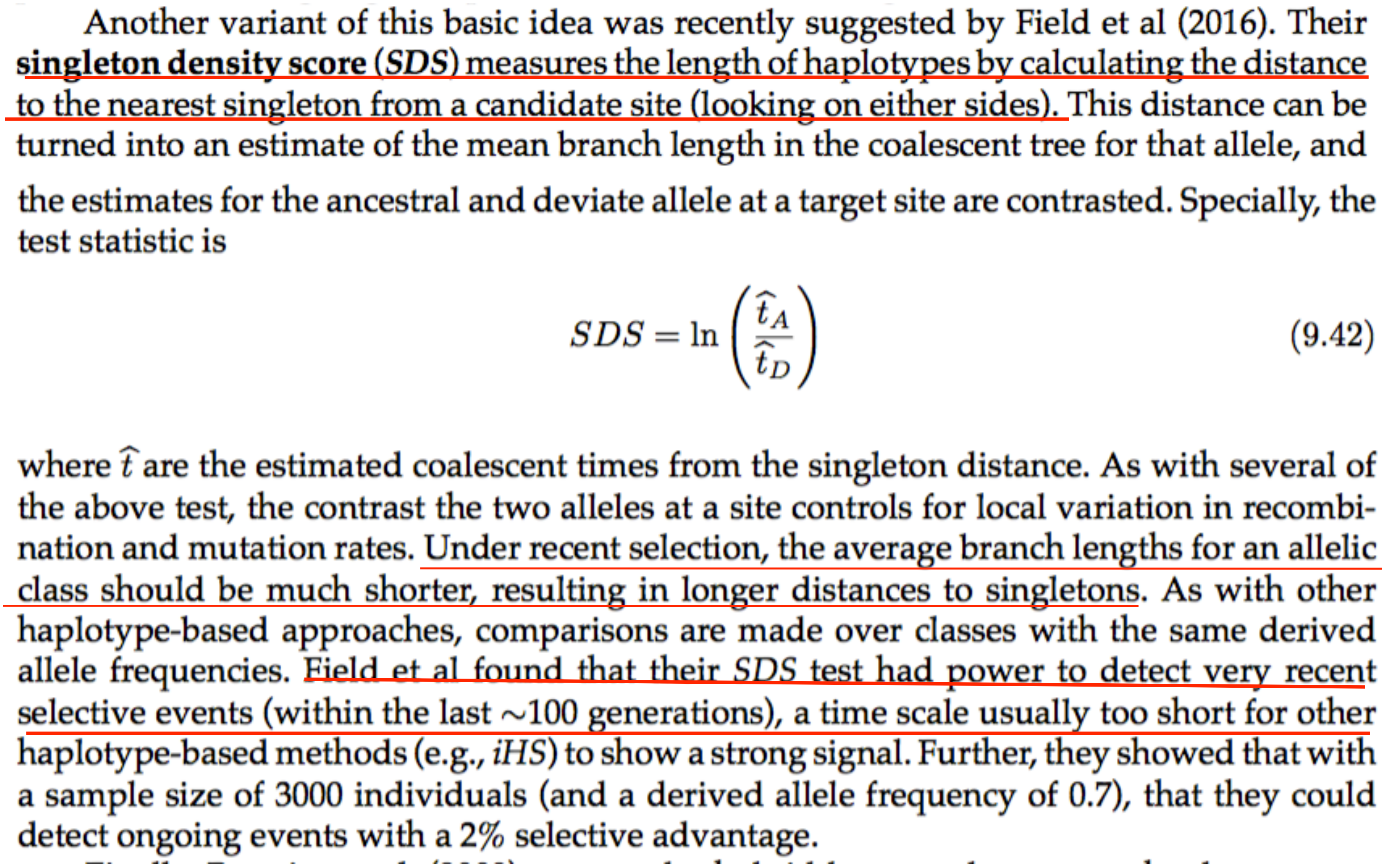
The expectation (E_p) and standard deviation (SD_p) are subscripted by p to highlight that these statistics are computed over all iHS_{us} values in the genome for SNPs whose derived allele frequency is p. Standardizing the score with respect to p automatically incorporates any relationship between the iHS_{us} score and the allele frequency (and hence the age for a neutral allele). The authors noted that this approach seems fairly robust to demographic departures from the equilibrium neutral model, especially at extreme values of the standardized score. Despite this, Voight et al. correctly did not assign significance values to individual iHS values, but rather used large (absolute) scores as a screening method for potential sites under selection.

Number of segregating sites by length (*nSL)*



Ferrer-Admetlla et al. (2014) proposed a statistic that is very similar in form to iHS but counts length variation differently. Their **number of segregating sites by length** ($\boldsymbol{nS_L}$) statistic replaces the average area under the iHS curve by the average number of consecutive segregating sites shared by two randomly sampled chromosomes around a specific SNP variant. This average statistic for the ancestral and derived alleles replaces iHS_A and iHS_D (respectively) in Equation 9.41b. Ferrer-Admetlla et al. noted that this simple change in the metric results in a test that is significantly more robust to recombination and slightly more robust to nonequilibrium departures than the iHS statistic. When applied in a human genome scan, the method did not yield the large enrichment of significant scores in regions of low recombination typically seen when other (more recombination-sensitive) tests are used (e.g., O'Reilly et al. 2008). Further, their simulations found that nS_L has reasonable power to detect ongoing sweeps, even those from standing variation.

Singleton density score (*SDS*)



Another variant of this basic idea was recently suggested by Field et al (2016). Their **singleton density score** (*SDS*) measures the length of haplotypes by calculating the distance to the nearest singleton from a candidate site (looking on either sides). This distance can be turned into an estimate of the mean branch length in the coalescent tree for that allele, and the estimates for the ancestral and deviate allele at a target site are contrasted. Specially, the test statistic is

$$SDS = \ln\left(\frac{\widehat{t}_A}{\widehat{t}_D}\right) \tag{9.42}$$

where $\widehat{t}$ are the estimated coalescent times from the singleton distance. As with several of the above test, the contrast the two alleles at a site controls for local variation in recombination and mutation rates. Under recent selection, the average branch lengths for an allelic class should be much shorter, resulting in longer distances to singletons. As with other haplotype-based approaches, comparisons are made over classes with the same derived allele frequencies. Field et al found that their *SDS* test had power to detect very recent selective events (within the last ∼100 generations), a time scale usually too short for other haplotype-based methods (e.g., *iHS*) to show a strong signal. Further, they showed that with a sample size of 3000 individuals (and a derived allele frequency of 0.7), that they could detect ongoing events with a 2% selective advantage.

Tests contrasting haplotype lengths of alternative alleles in the same population:

Sabeti et al.'s *rEHH*: Ratio of the haplotype lengths (*EHH*) of two alternative alleles
Wang et al.'s *LDD*: Rate of linkage disequilibrium decay, modification of EHH
Hanchard et al.'s *nHS*: Haplotype diversity of the derived allele relative to the ancestral allele
Voight et al.'s *iHS*: Ratio of area under the *EHH* curve for ancestral vs. derived alleles
Ferrer-Admetlla et al.'s nS_L: Very similar to *iHS*, with the number of consecutive shared polymorphic sites replacing the area under the *EHH* curve
Field et al.'s *SDS*: Distance to nearest singleton, yielding an estimated mean allelic branch length
Barreiro et al.'s *DIND*: Ratio of nucleotide diversity in derived vs. ancestral allele

Tests contrasting haplotype lengths of the same allele in two populations:

Sabeti et al.'s *XP-EHH* , Tang et al.'s *ln(Rsb)*: Ratio of area under the *EHH* curve in different populations
Kimura et al.'s *rHH* vs. *rMHH* plot: Ratios of overall *HH* to *HH* based on most frequency haplotype
Lange and Poll's χ_{MD} test: Contrast of pairwise haplotype sharing between populations

Summary: Tests Based on Haplotype/LD Information

As summarized in Table 9.2, different kinds of sweeps (hard, partial, and soft) leave different haplotype signals. Given the diversity of such signals, it is not surprising that there are a number of haplotype-based tests to detect these different features (Table 9.3). LD-based tests are generally regarded as the *most powerful for sweeps that are currently underway.* Site-frequency spectrum tests often perform poorly under a partial sweep, as the distortion in the frequency spectrum is often not sufficiently powerful. Signatures from both a recently completed partial sweep, and a currently ongoing hard sweep, include long haplotypes at excessive frequencies, alleles that are at too high a frequency given other estimates of their age, an excess of one or a few haplotypes, and a reduction in haplotype diversity.

In addition to their unique role in detecting partial sweeps, LD summary statistics can also offer significant power to detect *just-completed sweeps.* Under a hard sweep, the unusual pattern of high LD on either side of, but not across, a selected site can be detected using Kim and Nielson's ω statistic (Equation 9.37). However, this statistic has no power to detect a soft sweep. Conversely, Kelly's Z_{ns} statistic (measuring average pairwise LD throughout a region; Equation 9.36b) can detect a recently completed soft sweep but has no power to detect a just-completed hard sweep.

As with almost all the tests discussed in earlier sections, haplotype-based tests can also generate false positives for neutral alleles in nonequilibrium populations. The standard approach of using outlier analysis to suggest regions of interest and coalescent simulations (using marker-based demographic estimates) can also be used here, with the same caveats. As mentioned, both outlier analysis and coalescent simulations use corrections based on genome-wide patterns and thus do not adjust for allelic surfing. This is especially troublesome, for as outlined below, the species most surveyed for recent selection—humans, the cosmopolitan human commensal *Drosophila* (*melanogaster* and *simulans*), and *Arabidopsis*—are all known to have undergone massive spatial expansions over the last 100,000 years, making them prime candidates for surfing.