# Sequencing-era methods for identifying signatures of selection in the genome

Clare Horscroft1,3, Sarah Ennis1,3, Reuben J Pengelly1,3, Timothy J Sluckin2,3, Andrew Collins1,3

1. Genetic Epidemiology and Bioinformatics,

Faculty of Medicine,

University of Southampton,

Duthie Building (808),

Tremona Road, Southampton, SO16 6YD, UK.

2. Mathematical Sciences,

University of Southampton,

Highfield,

Southampton, SO17 1BJ, UK

3. Institute for Life Sciences,

University of Southampton,

Life Sciences Building (85),

Highfield,

Southampton, SO17 1BJ, UK

# Abstract

Insights into genetic loci which are under selection and their functional roles contribute to increased understanding of the patterns of phenotypic variation we observe today. The availability of whole genome sequence data, for humans and other species, provides opportunities to investigate adaptation and evolution at unprecedented resolution. Many analytical methods have been developed to interrogate these large datasets and characterise signatures of selection in the genome. We review here recently developed methods and consider the impact of increased computing power and data availability on the detection of selection signatures. Consideration of demography, recombination and other confounding factors is important, and use of a range of methods in combination is a powerful route to resolving different forms of selection in genome sequence data. Overall, a substantial improvement in methods for application to whole genome sequencing is evident, although further work is required to develop robust and computationally efficient approaches which may increase reproducibility across studies.

# Keywords

Natural selection

Machine learning

Selective Sweep

Genome sequencing

Recombination

# Author profiles

Clare Horscroft is a PhD student in the University of Southampton Genetic Epidemiology and Bioinformatics Research Group and is developing and applying novel approaches for detecting signatures of selection in genome data.

Sarah Ennis is Professor of Genomics at the University of Southampton. Her group specialises in the analysis of next generation sequencing (NGS) data focussed on, but not limited to, understanding the genetic aetiology of human disease.

Reuben J. Pengelly is Lecturer in Computational Genetics at the University of Southampton and works to understand and predict the functional role and impact of variation in genomes, particularly in the context of disease and other phenotypes.

Timothy J Sluckin is Professor of Applied Mathematical Physics within Mathematical Sciences at the University of Southampton. Originally a statistical physicist, he has a wide interest in applying methods of classical applied mathematics and theoretical physics to problems in the biological and social sciences.

Andrew Collins is head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in next generation sequencing-based studies in population genetics and research into the genetic basis of a number of diseases.

# Introduction

It is important to identify genomic regions which have been subject to selection in human populations to understand human history and evolution and the role of selection in the disease genome. Recent studies, for example, Fagny *et al* [1], have shown the adaptive significance of selected regions since they are enriched in, for example, functional variants detected by genome-wide association studies. As such they provide insights into relationships between earlier selection processes and existing benign and disease-related phenotypic variation.

In recent years dramatic increases in computing power have coincided with the availability of abundant genome sequence data and analysis of these vast datasets presents many new opportunities for understanding patterns of selection across the genome. These developments have necessarily accelerated research into novel analytical methods which fully utilize the information to understand the action of natural selection and how it has impacted genome structure and function. An important development is the increasing application of machine learning, which has led to some innovative work to obtain meaningful results from large and complex datasets.

Historically, published evidence has suggested signatures of positive selection in regions of the genome which are not replicated in subsequent studies [2, 3]. While a proportion of these are likely to be false-positives other factors, including under-powered follow-up studies and methodology, may to contribute to low replication rates. There is a need for robust detection methods to produce results which are replicable in other studies and therefore more likely to be biologically real signals [4].

Considering developments since the earlier review by Vitti *et al* in 2013 [5] we focus in particular on recent methods and their applicability for the analysis of genome sequence data. We compare and contrast methodology and evaluate the relative utility of specific methods for detecting selection in different scenarios.

## Overview of Signatures of Selection: Positive, Negative and Balancing Selection

Natural selection, at its simplest, underpins how populations are able to adapt, survive and thrive in their environments. It is the process by which individuals in a population, who possess traits that increase their likelihood of surviving and reproducing, transmit these traits to their offspring such that, over time, beneficial traits increase in frequency throughout the whole population [6]. Under positive selection a beneficial allele is selected for in the population and individuals with this allele have increased fitness to survive and reproduce compared to other individuals who lack the allele [7]. This might reflect the impact of a beneficial mutation arising spontaneously, or some environmental process rendering an existing neutral variant more beneficial than it was previously. Negative selection (also known as “background” or “purifying” selection) has the opposite effect: possession of an allele reduces fitness and the allele may therefore be eliminated from the population by selection. Balancing selection takes place where possession of the heterozygote genotype increases fitness: individuals homozygous for, or lacking, the allele in question have reduced fitness. A classic example is sickle cell anaemia which is found in humans who live in malaria risk areas. Sickle cell is a recessive disease, hence those with two copies of the causal allele have reduced fitness compared to those with one or no copies. However, having at least one copy of this allele provides protection from malaria. In this way, individuals with one copy have a degree of increased fitness through reduced risk of contracting malaria, although there are also disadvantageous traits associated with carrier status [8].

Positive selection is therefore likely to drive an increase in the specific allele frequency in the population, negatively selected alleles should see a decrease in frequency in the population, and alleles with balancing selection should persist at a fairly stable frequency, assuming a constant environment. However, through random chance, any of these alleles could rise to fixation, in which every individual in the population possesses the allele, or be lost completely from the population. This phenomenon, known as “genetic drift”, is a confounding factor when searching for signatures of selection.

## Hard, soft and polygenic sweeps

When a beneficial mutation arises spontaneously in an individual and rapidly increases in frequency in a population, rising to fixation or near fixation, it is called a “hard sweep”. As the mutation arose in one individual, a single haplotype, or small number of haplotypes, will contain the allele in the population upon fixation. However, if the positively selected allele was already present in the population (selection starting from “standing variation”) the sweep is termed “soft”. A soft sweep could involve multiple haplotypes which contain the allele due to recombination and mutation events occurring before the allele in question became beneficial. In addition, soft sweeps can be caused by multiple independent beneficial mutations in a single species sweeping concurrently, also known as “convergent evolution” in which the mutations have the same effect on fitness [9]. An example of this in humans is the way lactose persistence has swept through multiple populations with different adaptive mutations [10, 11]. Thus the term “soft sweep” can refer to multiple types of selection which lead to similar variation patterns in populations, for example, where multiple haplotypes sweep to a moderate frequency through increased fitness from a shared effect on a phenotype [12]. The genomic regions to either side (the “shoulders”) of a hard sweep might be mistaken for a soft sweep [12, 13]. “Incomplete” or “partial” sweeps are either soft or hard sweeps that have not reached fixation at the time of the sample being taken. For some phenotypes, there are multiple alleles which contribute. If these traits are selected for, this is called polygenic selection. A common example of this is human height, which has been found to depend on hundreds of loci throughout the genome [14, 15]. If positive selection occurs for these alleles it is difficult to detect because none of the alleles may reach fixation but all may become slightly more frequent as a group [16, 17]. Of course, these patterns are likely to be confounded by other processes including the impact of migration, multiple origins of alleles and the variable speed of sweeps. These, among other factors, confound efforts to easily classify and identify sweeps [18].

## Hitchhiking and recombination

When sweeps occur, they may leave behind evidence in the genome. As an allele approaches fixation other neutral, or even mildly deleterious, alleles near to the selected allele may sweep along with it in a process termed “hitchhiking” by Maynard Smith and Haigh [19]. The process of recombination, where chromosomes cross-over during meiosis, breaks up chromosome segments, shortening haplotypes. Recombination however lags behind the impact of the sweep and the population can be left with highly correlated alleles in regions of the chromosome (linkage disequilibrium or “LD”). The identification of chromosome regions with this structure is the basis for detecting sweeps. However, the impact of other processes may make the detection of these sweeps difficult. A sweep located around a chromosome recombination hotspot might be hard to detect as the hitchhiking effect would be lost quickly due to frequent recombination shortening haplotypes. Recombination intensity is highly variable across the genome and a chromosome region with very low levels of recombination might resemble a candidate sweep. Therefore it is advantageous to correct for recombination rates where known [3]. Population genetic data can be used to make an estimate of the recombination map, for instance the work by Myers *et al* [20]; the increasing availability of whole-genome sequencing data allows for these analyses at unprecedented resolution [21]. However, when compared to recombination maps based on family data, for example trio (mother-father-child) data, which provides sparse but direct data on recombination events [22], it has been shown that LD-based maps may be confounded by selection and other processes to at least some degree [23]. Joint linkage-linkage disequilibrium analysis is a powerful way to combine family-based and population-based maps to resolve relatively recent instances of selection [24, 25].

## Demographic changes

Demographic changes can also make it hard to detect a sweep. If a population size increases rapidly then the amount of genomic variation in the population will be less than expected for a population of that size due to the reduced time to accumulate mutations. Conversely, a population decreasing rapidly could result in more diverse allele frequencies than expected. Thus demographic changes can leave patterns which look similar to sweeps in a population. Bottlenecks, where a subset of a larger population becomes separated and develops into a new population, can leave sweep-like signatures in the genome [26]. Demographic changes can appear to make one type of sweep look like another, for example, if a population experiences a sudden bottleneck whilst undergoing a soft sweep arising from several beneficial mutations. This could result in just one of the mutations surviving, and thus the soft sweep is “hardened” [27]. Therefore, methods which jointly model demographic changes as well as sweeps are required [28-30]. Interestingly, Schrider and Kern [31] suggest that, across the human population, soft sweeps are the overriding process of selection, but there are proportionally more hard sweeps in non-African populations than in African populations.

# Methods used to detect signatures of selection

The ability to sequence genomes cost-effectively provides outstanding new opportunities to identify the signatures of selection genome-wide. We consider here recently developed methods which are able to exploit these data. The methods include eleven novel statistics, including two which are a composite of other statistics, and three machine-learning algorithms, which classify regions based on a range of other statistics. The methods are described in Table 1.

Table 1. Overview of recently published methods

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Method | Input | Output | Software/web link | Key findings / Author benchmarking |
| Zα [32] | Linkage Disequilibrium | Phased SNP data and genetic map | Sweeps | None: statistical  formula | Outperforms Kelly’s ZnS [33] and ωmax [34] and other novel statistics across a range of allele frequencies, with or without recombination rate variation. Better or comparable when considering an “Out of Africa” model and when starting from standing variation. |
| H12 [35, 36] | Haplotype homozygosity | Phased SNP data | Hard and soft sweeps | None: statistical formula | Compared to iHS [37], H12 has increased power for detecting recent soft sweeps. H12 is as good as, or better, at detecting recent hard sweeps. |
| nSL (number of Segregating sites by Length)[38] | Haplotype homozygosity | Phased SNP data including whether each allele is ancestral or derived | Sweeps | <http://www.nielsenlab.org/software/> | Outperforms iHS [37], EHH [39], Tajima’s D [40] and Fay and Wu’s H [41], over a range of scenarios, improved results comparable only to iHS. Method loses power when the selection co-efficient is low, and as the allele frequency nears fixation, especially for soft sweeps. |
| SDS (Singleton Density Score) [42] | Haplotype method | Phased SNP data including ancestral/derived status | Recent sweeps | <https://github.com/yairf/SDS> | Compared to iHS performs better when the selection coefficient is sufficiently strong, and selection began ~100 generations ago or is continuous. iHS performs better when selection stopped 100 generations ago. |
| ΧMD (Comparative Haplotype Identity Statistic) [43] | Haplotype-based method | Two populations, phased SNP data | Sweeps (especially soft and partial sweeps) | <https://github.com/jeremy-lange/CHI-Statistic> | Compared to FST [44] and XP-EHH [45]. XP-EHH outperformed χMD where population bottlenecks or migration were operating, but χMD outperformed XP-EHH in most other scenarios, especially in partial sweeps and soft sweeps, in particular where the effective population size was low. |
| HacDivSel  (Haplotype allelic class - Divergent Selection) [4] | Haplotype or outlier method | Two populations, phased or unphased SNP data | Sweeps | <http://acraaj.webs.uvigo.es/> | nvdFST was compared to SvdM [46] and EOS to BayeScan [47]. Both new methods were more powerful for detecting sweeps over a range of scenarios. |
| TSel (Time to most recent common ancestor Selection) [48] | Pairwise TMRCA and Anomaly Detection | Phased SNP data | Sweeps | <http://blogs.cornell.edu/clarklabblog/clark-lab/software/> | TSel was compared to four other statistics including nSL. In each case TSel was better or equivalent. The difference is especially notable where selection intensity is low where other methods perform no better than random. When compared to HKA [49], TSel was also shown to be good at identifying recent strong balancing selection. |
| SweepFinder2 [50, 51] | Composite Likelihood Method | Allele frequency, recombination map and B-value map [52] | Sweeps | <http://www.personal.psu.edu/mxd60/sf2.html> | When compared to HKA [49] and other composite likelihood methods such as in the original SweepFinder [53], SweepFinder2 is superior given strong background selection, with other methods returning almost 100% false positive rates. |
| CLR  (Composite-Likelihood Ratio test) [54] | Composite likelihood method | Phased SNP data including ancestral/derived status | Incomplete sweeps | None: statistical formula | Vy’s composite likelihood method was compared to iHS, where it outperformed in all scenarios tested, and nSL, where it was almost always better. Correctly identified many locations of sweeps which were not detected by iHS or nSL. |
| DCMS  (De-correlated Composite of Multiple Signals) [55] | Composite Method | Two populations phased SNP data | Sweeps | None: statistical formula | DCMS was compared to CSS and meta-SS [56] as they are all combining methods. DCMS outperforms both over a range of parameters, except where the frequency of the selected allele is low, or the interval distance is very high. |
| CSS (Composite Selection Signals) [57] | Composite Method | Two populations, phased SNP data. Ancestral/derived status if available | Sweeps | None: statistical formula | No comparison to another model. Application to cattle and sheep data showed clusters of extreme CSS values in candidate regions. |
| Deep Learning [58] | Machine Learning – Neural Networks | Phased SNP data | Classification: Neutral, Hard, Soft or Balancing, plus estimate of population sizes | <https://sourceforge.net/projects/evonet/> | No comparison model for the selection part, but when tested on simulations the model is good at specifying the correct class, especially with pre-training. |
| S/HIC  (Soft/Hard  Inference through Classification) [59] | Machine learning, extremely random trees | Phased SNP data, including ancestral/derived status | Classification: Hard, Hard-linked [Near a hard sweep], Soft, Soft-linked, Neutral | <https://github.com/kern-lab/shIC> | Outperforms seven other methods, distinguishing regions with a sweep from those which are neutral or linked to sweeps. |
| Hierarchical Boosting [60] | Machine learning - boosting | Phased SNP data | Classification: complete or incomplete sweeps, ancient or recent | <http://hsb.upf.edu/> | Compared to evolBoosting [61], CMS [62] and SFselect [63] – had the highest sensitivity under every scenario simulated. |

## Linkage disequilibrium based methods

Linkage Disequilibrium (LD) methods identify genomic regions in which hitchhiking has taken place by finding correlations between single nucleotide polymorphisms (SNPs). The Zα method by Jacobs *et al* [32] uses this relationship by evaluating the average correlation between SNPs on each side of a target locus (Figure 1). In a hard selective sweep, where the beneficial mutation rises rapidly in frequency, there is a reduction in population genetic variation in the region and a breakdown in LD across the selected mutation, but not in the ‘shoulders’, leaving a pattern which resembles a recombination hot-spot [64]. The distinction between hard sweeps, which display this pattern, and narrow regions of intense recombination might be facilitated using fine-scale patterns of recombination determined from population data [20]. However the precise relationship between the ‘true’ linkage map made from family data (which has substantially lower resolution) and the population LD approximation to the linkage map (but at much higher resolution) is unclear [23].



Figure 1. The average pattern of pairwise LD and SNP diversity created by a selective sweep (from Jacobs et al [32]).  
The average pattern of pairwise LD (quantified as r2, in the lower triangles) and SNP diversity (SNP count: upper triangles) created by a simulated selective sweep. During simulation the sweep begins at time t1 generations before present, using an initial selected allele frequency of 0.0005 and an additive selection model with the homozygous state corresponding to s=0.04. The frequency of the selected allele in the present is, on average, ∼0.7 and > 0.99 assuming selection started, respectively, 400 and 1600 generations ago. In the rightmost plot, the “along” (α) and “over” (β) regions are indicated. The rightmost figure shows an increase in LD in the along region, but a decrease in the over region, compared to the central figure which shows increased LD in both regions as found in a partial selective sweep.

## Haplotype methods

Haplotype methods depend on the same concept as LD-based methods in that one of the impacts of hitchhiking is the presence of relatively high frequency longer haplotypes around the locus subject to selection. As with other methods the variable pattern of recombination and the similarity between regions subject to hitchhiking and regions with reduced recombination can present difficulties [65, 66].

H12 is a method which examines haplotype homozygosity in the population to locate sweeps [35]. It relies on the expectation that in a hard sweep one haplotype (the one containing the selected allele) will increase in frequency. In a soft sweep, the selected allele may be on multiple haplotypes and thus H12 considers both the most frequent haplotype and the second most frequent. The method also provides a statistic to distinguish between hard and soft sweeps. To address areas of low recombination rates, which will look very similar to a sweep using this method, the authors removed these areas from their *Drosophila melanogaster* analysis completely, using maps created by Comeron *et al* [67]. Clearly, this method of avoiding false positives is only viable if a reliable genetic recombination map is available for the species in question.

The nSL model is designed to detect the impact of positive selection acting to increase haplotype homozygosity. The model uses haplotype length, rather than frequency, to detect sweeps [38]. The method considers the ratio of haplotype homozygosity for haplotypes containing the derived and ancestral allele in a candidate site. The rationale is that if there is a sweep present there will be more segregating sites in the rest of the sample, due to the long haplotype length in the derived haplotypes compared to the ancestral haplotypes. By considering the distribution of the number of segregating sites between all pairs of chromosomes, hard and soft sweeps are identified. The method is considered robust to misspecification of the recombination rate and population parameters.

SDS (Singleton Density Score) employs a similar rationale, where the distances between the nearest singleton mutations either side of a test SNP are calculated for each individual, as the haplotypes which contain the selected allele tend to carry fewer singleton mutations [42]. See Figure 2 for an example of the rationale for these methods when applied to a hard sweep. The χMD method again draws on the notion that unusually long haplotypes observed frequently in a population could be evidence for a sweep [43]. In this case, the lengths of especially long identical haplotypes in a population are summed and compared to a reference population. HacDivSel implements two different methods, the first, *nvd*FST (normalized variance difference (nvd) combined with the FST index), is a haplotype-based method combined with an FST calculation [4]. The haplotype method compares the variances of minor variant frequencies between haplotypes containing the major and minor allele, developing the idea that if an allele is subject to a sweep then the variances in the haplotypes of the major allele will be lower due to other major alleles sweeping along with it.

The second method included in HacDivSel is the Extreme Outlier Set test (EOS), which is used when the phased data is not available, and uses FST combined with a k-means classification algorithm to find outliers of interest.



Figure 2. The impact of selection on a “population” of nine chromosomes.  
Step 1 shows a new beneficial mutation (black square) occurring spontaneously on the third chromosome. Step 2 shows the mutation beginning to spread through the population, with nearby variants hitchhiking along with it (grey squares). Recombination events are indicated by a double line. Step 3 shows the population just over half-way to fixation. nSL and SDS are most effective at identifying sweeps when they reach this stage. The rationale for both is that haplotypes containing the beneficial allele will be longer than those carrying the ancestral allele. For chromosomes containing the allele under selection, the variants either side of the target locus which are not found on any other chromosome (the first and fourth chromosomes) are further apart than those without (the sixth and eighth chromosomes). Step 4 shows the beneficial allele reaching fixation. H12 uses the squared frequency of the two most common haplotypes to identify sweeps. For this example H12 = 0.48 in Step 4 compared to H12 = 0.14 in Step 1 where there was no sweep. The rationale for the Zα model is illustrated: the average correlation between the variants on each side of the selected locus (36%) is higher than the correlation between the variants across the locus (22%).

## TMRCA

One method, TSel [48], uses pairwise Time to Most Recent Common Ancestor (TMRCA) distributions at each locus to find deviations from the norm by then applying an anomaly detection algorithm. TMRCA methods estimate the number of years that have passed since the most recent common ancestor of two alleles in a diploid individual. This is done by comparing the number of heterozygotes with the number of homozygotes in a section of the genome: areas of the genome with more heterozygotes will have a longer TMCRA than areas with fewer heterozygotes. One way of calculating this is to use the pairwise sequentially Markovian coalescent (PSMC) method as described by Li and Durbin [68]. Before TSel can be used, the distribution of TMRCA values is found across the population sample for each locus and many features are calculated, including the average, maximum, median and variance. TSel then inputs these vectors of factors for each locus into an anomaly detection algorithm to find areas in the genome which are unusual and therefore could indicate signs of selection.

## Composite likelihood methods

Composite likelihood methods rely on the idea that the site frequency spectrum (SFS) will be distorted around the site of a sweep, compared to the rest of the genome. An allele close to, or at, fixation will be surrounded by other loci at which the frequency of the derived alleles will either be very high or very low depending on whichever was associated with the beneficial allele. The composite-likelihood ratio (CLR) test proposed by Vy and Kim [54] exploits this pattern, as does the SweepFinder2 method, although the latter is also adjusted to account for “background” selection [50]. This is achieved by using a “B-value map”, a map of the genome which provides an estimate of the effect of background selection at each position [52]. Correcting for background selection should reduce false positive rates since background selection can cause a loss of diversity which can resemble a positive sweep [69].

## Composite Methods

Composite methods utilise a range of statistics to combine and create a new statistic. The DCMS method combines eight selection statistics, including the CLR method from the original SweepFinder [53, 55]. DCMS finds the p-values of all the statistics and combines them, weighting for the correlations between statistics, to create a new final statistic. The CSS method combines three separate statistics by ranking each statistic across all SNPs, then identifying clusters where SNPs have extreme test scores due to hitchhiking [57].

## Machine learning

Three machine learning methods have been applied for detecting signatures of selection: neural networks, extremely random trees and boosting. All three of these methods are ‘supervised’ in that they utilise a ‘training’ dataset where the classification is known *a priori* (e.g. regions known to be under selection) to develop a classifier and this trained model is applied to classify new test data. Deep learning has its roots in neural networks which have been studied for decades. The Deep Learning method applied by Sheehan and Song [58] uses a neural network to classify regions by selection type, and also returns a genome-wide estimate of demographic history. Simply, neural networks take data through a network of layers containing nodes where at each node the data get transformed in some way. The method is trained using simulated datasets where 345 summary statistics are calculated, including the H12 statistic described above. It is during this training that the transformations each node will perform are established along with the corresponding weightings between each node. The S/HIC method [59] similarly uses simulated datasets for training. It uses an extremely randomized tree method to classify regions into the type of sweep it may or may not have. Briefly, this method works by creating decision trees semi-randomly which take the input data, in this case multiple statistics, including H12 and H2/H1 as described above. Data are split in two at each node of the tree by examining one of the statistics, until it gets to a “leaf” at the end of the tree with a final classification. Many trees are created and each tree “votes” as to which classification the data should be, with the majority being the final classification given to the data. Finally, the hierarchical boosting method [60] classifies regions by calculating statistics and then putting them through a series of linear regression functions. Again, this method is trained using simulated datasets.

# Examples of four methods

We conducted a comparison of four methods: Zα, H12, nSL and TSel. Two populations of “chromosomes” were simulated: one with a site subject to selection and one without and results compared across methods. Details of methodology are given below in the Methods section. Figure 3 shows the average of the neutral and selected simulations for each of the four methods, focusing on the middle 200 Kb region of the simulation. There are clear differences in the shapes of the graphs between neutral and selected simulations, and between the methods. All of the methods look for the most extreme values as potential areas of selection, except TSel, which considers high median values of regions.



Figure 3. The average values for the statistics across simulations for four methods.  
The plots show the average value of each statistic in 1000 base-pair bins across the simulated region, for both neutral and selected scenarios. Bars on each plot represent one standard deviation away from the mean in each bin. The 100 scores closest to the site of selection for each test are highly significantly different from the same points in the neutral simulations by Mann-Whitney U test.

The four methods were compared using ROC curves [70], highlighting methods which minimise the False Positive Rate (FPR) of 5% or less [71]. The ROC curves are shown in Figure 4 with the areas under the curve given for the whole curve and for the area where FPR < 5%. The curves show that, under this particular selection scenario, Zα and H12 are most effective at detecting selection compared to neutral, followed by TSel with nSL the least effective.



Figure 4. ROC Curves for the four statistics.  
Receiver Operating Characteristic (ROC) curves created for each of the four methods using neutral and selected simulations. For each method the Area Under the Curve (AUC) is given, as well as the partial AUC (pAUC) where the specificity is greater than 95% (equivalent to the false positive rate being less than 5%).

The nSL method was specifically designed to detect incomplete sweeps as was simulated here. The plot in Figure 3 shows that nSL is able to differentiate selection from neutral, however discrimination is impacted by many high values for absolute nSL in the neutral simulations. Therefore differentiating using the maximum value across the region is less effective than using for example, the median value, which increases the AUC to 100%. The success of individual methods is likely to be highly dependent on the selection scenario (either simulated or real) and comprehensive understanding of specific selection signatures is likely to be enhanced through the application of a series of methods which have different strengths and weaknesses.

# Discussion

The diverse set of methods presented cover a range of different approaches and rationales. It is interesting to see the development of methods over time, where previously simply looking for a “sweep” has evolved into the identification of particular modes and types of sweep. Refinements include for example nSL [38], which is designed to find incomplete sweeps, and TSel [48] which targets hard sweeps, and SDS [42] for recent sweeps. Some methods, such as HacDivSel [4], take classical methods and adapt them with the goal of reducing false positive rates.

These methods are not applied without difficulties however. Many methods require detailed genomic information from multiple populations with the detection of regions under selection depending on cross-population analysis. These methods obviously will not be appropriate to use when only a single population is available. Multiple methods, including all those which rely on having defined haplotypes, require phased genotype data which is not always readily available. HacDivSel avoids this problem by offering two different methods, one for phased and one for phase-unknown data. Other methods require information about the ancestral and derived status of SNP alleles which is not always available. The CSS method [57] allows for this by offering an alternative statistic when this information is not available. A particular difficulty is the reliance of some methods on a genetic linkage map. For example, Zα [32] does not perform as well in the presence of recombination rate variation without correction from a genetic map. However, approximations to the high-resolution linkage map using population-data based methods such as LDhat [72] might provide a basis for discriminating between genomic regions subject to sweeps and those with reduced recombination if the recombination signatures captured are not confounded by other processes which impact LD.

One of the biggest problems with the application of most statistics to identify signatures of selection is that there is no neutral distribution with which to compare the statistic, so there is no way of knowing whether a finding is statistically significant or not [73]. Some of the methods here use simulations under neutrality to approximate the distribution, others just use the most extreme values, for example for TSel [48] the authors examined the top 1% of results.

The impact of background selection can be a problem for some methods, for example TSel [48] may not perform well in presence of strong background selection. Jensen *et al* [74] has argued that the neutral model should be corrected for background selection and should not be ignored. In this way, SweepFinder2 requires a B-value map to correct for background selection, however it is currently only available for humans [52] and D. melanogaster [75]. B-value maps themselves are not infallible, and can over-correct in some areas and under-correct in others.

Frequently methods focus on selection at SNPs specifically, however there are other changes in the genome which could be adaptive, for example transposable element insertions, where sections of DNA move from one location to another. Villenueva-Cañas *et al* [76] provide information of the use of nSL and H12 to help find this kind of adaptation, and suggest that χMD may also be an appropriate model [77].

Polygenic sweeps are a distinct category of sweeps which we do not consider in detail here. They are notoriously hard to detect as they do not share the properties of typical hard and soft sweeps. However, some methods described here have shown promise at identifying polygenic sweeps, such as CSS [78] and HacDivSel [4]. Other methods have been developed or are in development with the purpose of finding polygenic sweeps [16, 79].

Populations can go through significant changes rapidly, for example a bottleneck where a larger population shrinks suddenly and then grows from this smaller sample, and this will itself leave signatures in the genome. An example of this is the founder effect where a sub-sample of a population migrates to a new location and then the population grows from there [80]. Even if the demography is known, some methods perform badly under these conditions and even worse if the demography is misspecified. Some methods are considered robust to uncertain demography, for example S/HIC [59], TSel [48] and nSL [38]. The Deep Learning method [58] was designed to detect demography across the genome, along with finding sweep sites in individual regions. It is important that statistics are tested for robustness across different demographic scenarios, as populations will rarely behave according to the assumptions of most models, for example that the population is always a stable size and that each member of the population has an equal chance of mating with another [81].

Machine learning methods have some unique problems. Often, the processing power and time needed to train a model is not inconsiderable, and access to high performance computing is becoming necessary [82]. Sheehan and Song report hundreds of hours to simulate data, compute statistics, and train and test their Deep Learning model [58]. Often machine learning methods have a reputation for being a “black box” in that data go in and results come out with the internal workings being difficult to discern. The methods in this case however seem to be fairly transparent, with S/HIC [59] and hierarchical boosting [60] reporting the relative ranks and contributions of each statistic used. One of the benefits of these machine learning methods, and indeed the combining methods in general, is that they are flexible and different statistics could be loaded in with very little change to the structure of the methods. Caution should be used however, as machine learning methods can only be as good as the training set that they are given [83].

Irreproducibility of published selection signals historically is a problem, as it is important that results can be verified by other researchers to tease false positives from true. This is often not possible for many reasons, including: parameter values not defined, or assigned arbitrary values with little or no justification, bugs in computer software or changes to software of which the authors may not even be aware, or different computer architectures returning different results, for example because of floating point errors [84, 85]. It is also tempting to validate results by creating a biological narrative, by searching the literature for evidence that the results may have selective importance. Somewhat disturbingly, Pavlidis *et al* [85] show that a plausible account can be constructed for genes identified using a neutral simulation, and it is therefore important for researchers not to rely on “storytelling” as a method of verification. This problem of irreproducibility highlights the need for distinct, orthogonal methods that can be applied to increase robustness.

Many of the papers suggest using statistics in parallel with others, for example, Vy suggests using their CLR with nSL, and Lange suggests using FST in combination with a haplotype based statistic [43, 54]. This suggestion is because different methods identify different signals in the data and therefore when used together more signals can be captured more accurately. The Vy paper demonstrates this clearly in their comparison of their CLR with nSL where they correctly detect a similar number of sweeping sites each but only a fraction of those overlapped [54]. Indeed this is the rationale for many combining methods and machine learning approaches: the more complementary information available, the better to identify a true sweep. Furthermore, many papers suggest that combining methods is the best way to find and classify sweeps [86-88].

The development of a vast array of different methods, from simple statistics to complex machine learning methods, shows that this area of research is particularly important and exciting. While it is acknowledged that demography can be a confounding factor and steps are being taken to check for robustness to different demographic scenarios, there is still much work to be done to tease apart the effects of demography from sweep signals. Applying these methods to other species and organisms will also be important future work in understanding natural selection and the genome. It has already been shown that chimpanzees, despite the close evolutionary relationship, have recombination hotspots in different places to humans [72], and so there is a great need for accurate recombination maps to be constructed for species of note. As new methods are developed and refined and made statistically robust, it is fascinating to see what new discoveries will be made in natural history and evolution.

# Methods

## Simulations

Two simulations were performed: a neutral simulation and a simulation with selection at one site, both using the “MSMS” simulation software [89]. Simulated data comprised a population sample of 100 ‘chromosomes’ represented as a 500 Kb region. For both scenarios 100 simulations were generated. For the whole 500 Kb region a mutation rate of 10-9 per site and a recombination rate of 10-8 between adjacent base pairs per generation was assumed. The selective advantage for the homozygote was set at 0.012 and an additive model was applied with the degree of dominance set to 0.5. In the simulations which included selection the selected site was set in the centre of the region achieving a final allele frequency of 0.5 consistent with a partial sweep.

## Application of Sweep Software

The Zα and H12 models were implemented using R (v3.4.2) [90]. The Zα statistic was calculated using SNPs on average 1 Kb apart and a window size of 200 Kb. H12 was calculated using a window size of 400 SNPs in intervals of 20 SNPs.

nSL was applied using the software described in the article by Ferrer-Admetlla *et al* [38] downloaded on April 19th 2018. The default maximum window length of 200 SNPs was used, and the output was normalised by allele frequency bins with the frequency bin increments set at 1%.

TSel was calculated using the “tsel” package (v0.5) in R [91]. The coalescent tress were extracted from the MSMS simulation output and manipulated using the “ape” R package (v4.1) to extract the features [92]. Features were calculated using the pairwise TMRCA for non-recombining loci, the features used were the average, maximum, minimum, median, variance, fraction of pairs equalling the maximum, the first quartile and the third quartile. TSel scores were calculated for every 20th SNP in the simulated dataset.

## ROC Curve analysis

ROC curves were created by taking the statistics for the middle 200 Kb section of each simulation and applying the “pROC” package (v1.11.0) in R [93]. The area under the curve (AUC) and partial AUC, where the false positive rate was less than 5%, were calculated using the same package.

# Key points

1. Whole genome sequencing in many populations provides outstanding opportunities to screen for signatures of selection. These data, combined with powerful new methodology such as machine learning, provide good prospects for improved understanding.
2. Detecting selection is very difficult as signatures are confounded by processes such as recombination and drift and the effects of changing demography over time.
3. The mechanisms of selection, along with variable time since selection began and/or ended, in combination leave different signatures on the genome. Therefore methods have been developed which focus on a specific type of selection along with methods designed to detect all selection sub-types.
4. The challenge of distinguishing between signatures of selection and other processes which leave similar looking patterns in the genome remain. Deeper annotation of genomes through sequencing and the application of multiple analytical approaches will facilitate more robust findings in the future.

# Acknowledgements

This work was supported by funding from the University of Southampton Institute for Life Sciences, the Faculty of Medicine and the Department of Mathematics. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

# References

1. Fagny, M., E. Patin, D. Enard, et al., *Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets.* Molecular Biology and Evolution, 2014. **31**(7): p. 1850-1868.

2. Nei, M., Y. Suzuki, and M. Nozawa, *The Neutral Theory of Molecular Evolution in the Genomic Era*, in *Annual Review of Genomics and Human Genetics, Vol 11*, A. Chakravarti and E. Green, Editors. 2010, Annual Reviews: Palo Alto. p. 265-289.

3. Haasl, R.J. and B.A. Payseur, *Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication.* Molecular Ecology, 2016. **25**(1): p. 5-23.

4. Carvajal-Rodríguez, A., *HacDivSel: Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations.* PLOS ONE, 2017. **12**(4): p. e0175944.

5. Vitti, J.J., S.R. Grossman, and P.C. Sabeti, *Detecting Natural Selection in Genomic Data*, in *Annual Review of Genetics, Vol 47*, B.L. Bassler, M. Lichten, and G. Schupbach, Editors. 2013, Annual Reviews: Palo Alto. p. 97-120.

6. Alachiotis, N. and P. Pavlidis, *Scalable linkage-disequilibrium-based selective sweep detection: a performance guide.* GigaScience, 2016. **5**(1): p. 1-22.

7. Nielsen, R., *Molecular signatures of natural selection*, in *Annual Review of Genetics*. 2005, Annual Reviews: Palo Alto. p. 197-218.

8. Rees, D.C., T.N. Williams, and M.T. Gladwin, *Sickle-cell disease.* Lancet, 2010. **376**(9757): p. 2018-2031.

9. Pennings, P.S. and J. Hermisson, *Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation.* PLOS Genetics, 2006. **2**(12): p. e186.

10. Tishkoff, S.A., F.A. Reed, A. Ranciaro, et al., *Convergent adaptation of human lactase persistence in Africa and Europe.* Nat Genet, 2007. **39**(1): p. 31-40.

11. Enattah, N.S., T.G.K. Jensen, M. Nielsen, et al., *Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture.* American Journal of Human Genetics, 2008. **82**(1): p. 57-72.

12. Jensen, J.D., *On the unfounded enthusiasm for soft selective sweeps.* 2014. **5**: p. 5281.

13. Schrider, D.R., F.K. Mendes, M.W. Hahn, et al., *Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps.* Genetics, 2015. **200**(1): p. 267-284.

14. Turchin, M.C., C.W.K. Chiang, C.D. Palmer, et al., *Evidence of widespread selection on standing variation in Europe at height-associated SNPs.* Nat Genet, 2012. **44**(9): p. 1015-1019.

15. Lettre, G., *Recent progress in the study of the genetics of height.* Human Genetics, 2011. **129**(5): p. 465-472.

16. Berg, J.J. and G. Coop, *A Population Genetic Signal of Polygenic Adaptation.* PLOS Genetics, 2014. **10**(8): p. e1004412.

17. Fan, S., M.E.B. Hansen, Y. Lo, et al., *Going global by adapting local: A review of recent human adaptation.* Science, 2016. **354**(6308): p. 54-59.

18. Hermisson, J. and P.S. Pennings, *Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation.* Methods in Ecology and Evolution, 2017. **8**(6): p. 700-716.

19. Maynard Smith, J. and J. Haigh, *The hitch-hiking effect of a favourable gene.* Genet Res, 1974. **23**(1): p. 23-35.

20. Myers, S., L. Bottolo, C. Freeman, et al., *A fine-scale map of recombination rates and hotspots across the human genome.* Science, 2005. **310**(5746): p. 321-324.

21. Pengelly, R.J., W. Tapper, J. Gibson, et al., *Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations.* BMC Genomics, 2015. **16**(1): p. 666.

22. Fan, H.C., J.B. Wang, A. Potanina, et al., *Whole-genome molecular haplotyping of single cells.* Nature Biotechnology, 2011. **29**(1): p. 51-+.

23. Kong, A., G. Thorleifsson, D.F. Gudbjartsson, et al., *Fine-scale recombination rate differences between sexes, populations and individuals.* Nature, 2010. **467**(7319): p. 1099-1103.

24. Wu, R. and Z.-B. Zeng, *Joint Linkage and Linkage Disequilibrium Mapping in Natural Populations.* Genetics, 2001. **157**(2): p. 899-909.

25. Sun, L., X. Zhu, Q. Zhang, et al., *A unifying experimental design for dissecting tree genomes.* Trends in Plant Science, 2015. **20**(8): p. 473-476.

26. Crisci, J., Y.-P. Poh, S. Mahajan, et al., *The impact of equilibrium assumptions on tests of selection.* Frontiers in Genetics, 2013. **4**(235).

27. Wilson, B.A., D.A. Petrov, and P.W. Messer, *Soft Selective Sweeps in Complex Demographic Scenarios.* Genetics, 2014. **198**(2): p. 669-684.

28. Bank, C., G.B. Ewing, A. Ferrer-Admettla, et al., *Thinking too positive? Revisiting current methods of population genetic selection inference.* Trends in Genetics, 2014. **30**(12): p. 540-546.

29. Schrider, D.R., A.G. Shanku, and A.D. Kern, *Effects of Linked Selective Sweeps on Demographic Inference and Model Selection.* Genetics, 2016. **204**(3): p. 1207-1223.

30. Mathew, L.A. and J.D. Jensen, *Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography.* Frontiers in Genetics, 2015. **6**: p. 268.

31. Schrider, D.R. and A.D. Kern, *Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome.* Molecular Biology and Evolution, 2017. **34**(8): p. 1863-1877.

32. Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807-+.

33. Kelly, J.K., *A test of neutrality based on interlocus associations.* Genetics, 1997. **146**(3): p. 1197-1206.

34. Alachiotis, N., A. Stamatakis, and P. Pavlidis, *OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets.* Bioinformatics, 2012. **28**(17): p. 2274-2275.

35. Garud, N.R., P.W. Messer, E.O. Buzbas, et al., *Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps.* PLOS Genetics, 2015. **11**(2): p. e1005004.

36. Garud, N.R. and N.A. Rosenberg, *Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps.* Theoretical Population Biology, 2015. **102**(Supplement C): p. 94-101.

37. Voight, B.F., S. Kudaravalli, X. Wen, et al., *A Map of Recent Positive Selection in the Human Genome.* PLoS Biology, 2006. **4**(3): p. e72.

38. Ferrer-Admetlla, A., M. Liang, T. Korneliussen, et al., *On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure.* Molecular Biology and Evolution, 2014. **31**(5): p. 1275-1291.

39. Sabeti, P.C., D.E. Reich, J.M. Higgins, et al., *Detecting recent positive selection in the human genome from haplotype structure.* Nature, 2002. **419**(6909): p. 832-837.

40. Tajima, F., *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics, 1989. **123**(3): p. 585-595.

41. Fay, J.C. and C.I. Wu, *Hitchhiking under positive Darwinian selection.* Genetics, 2000. **155**(3): p. 1405-1413.

42. Field, Y., E.A. Boyle, N. Telis, et al., *Detection of human adaptation during the past 2000 years.* Science, 2016. **354**(6313): p. 760-764.

43. Lange, J.D. and J.E. Pool, *A haplotype method detects diverse scenarios of local adaptation from genomic sequence variation.* Molecular Ecology, 2016. **25**(13): p. 3081-3100.

44. Hudson, R.R., M. Slatkin, and W.P. Maddison, *ESTIMATION OF LEVELS OF GENE FLOW FROM DNA-SEQUENCE DATA.* Genetics, 1992. **132**(2): p. 583-589.

45. Sabeti, P.C., P. Varilly, B. Fry, et al., *Genome-wide detection and characterization of positive selection in human populations.* Nature, 2007. **449**(7164): p. 913-918.

46. Rivas, M.J., S. Dominguez-Garcia, and A. Carvajal-Rodriguez, *Detecting the Genomic Signature of Divergent Selection in Presence of Gene Flow.* Current Genomics, 2015. **16**(3): p. 203-212.

47. Foll, M. and O. Gaggiotti, *A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective.* Genetics, 2008. **180**(2): p. 977-993.

48. Hunter-Zinck, H. and A.G. Clark, *Aberrant Time to Most Recent Common Ancestor as a Signature of Natural Selection.* Molecular Biology and Evolution, 2015. **32**(10): p. 2784-2797.

49. Hudson, R.R., M. Kreitman, and M. Aguade, *A TEST OF NEUTRAL MOLECULAR EVOLUTION BASED ON NUCLEOTIDE DATA.* Genetics, 1987. **116**(1): p. 153-159.

50. DeGiorgio, M., C.D. Huber, M.J. Hubisz, et al., *S weep F inder 2: increased sensitivity, robustness and flexibility.* Bioinformatics, 2016. **32**(12): p. 1895-1897.

51. Huber, C.D., M. Degiorgio, I. Hellmann, et al., *Detecting recent selective sweeps while controlling for mutation rate and background selection.* Molecular Ecology, 2016. **25**(1): p. 142-156.

52. McVicker, G., D. Gordon, C. Davis, et al., *Widespread Genomic Signatures of Natural Selection in Hominid Evolution.* Plos Genetics, 2009. **5**(5): p. 16.

53. Nielsen, R., S. Williamson, Y. Kim, et al., *Genomic scans for selective sweeps using SNP data.* Genome Research, 2005. **15**(11): p. 1566-1575.

54. Vy, H.M.T. and Y. Kim, *A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data.* Genetics, 2015. **200**(2): p. 633-649.

55. Ma, Y., X. Ding, S. Qanbari, et al., *Properties of different selection signature statistics and a new strategy for combining them.* Heredity, 2015. **115**(5): p. 426-436.

56. Utsunomiya, Y.T., A.M.P. O'Brien, T.S. Sonstegard, et al., *Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods.* Plos One, 2013. **8**(5): p. 11.

57. Randhawa, I.A.S., M.S. Khatkar, P.C. Thomson, et al., *Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep.* Bmc Genetics, 2014. **15**: p. 19.

58. Sheehan, S. and Y.S. Song, *Deep Learning for Population Genetic Inference.* PLOS Computational Biology, 2016. **12**(3): p. e1004845.

59. Schrider, D.R. and A.D. Kern, *S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning.* PLOS Genetics, 2016. **12**(3): p. e1005928.

60. Pybus, M., P. Luisi, G.M. Dall'Olio, et al., *Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations.* Bioinformatics, 2015. **31**(24): p. 3946-3952.

61. Lin, K., H.P. Li, C. Schlotterer, et al., *Distinguishing Positive Selection From Neutral Evolution: Boosting the Performance of Summary Statistics.* Genetics, 2011. **187**(1): p. 229-244.

62. Grossman, S.R., K.G. Andersen, I. Shlyakhter, et al., *Identifying Recent Adaptations in Large-Scale Genomic Data.* Cell, 2013. **152**(4): p. 703-713.

63. Ronen, R., N. Udpa, E. Halperin, et al., *Learning Natural Selection from the Site Frequency Spectrum.* Genetics, 2013. **195**(1): p. 181-+.

64. Kim, Y. and R. Nielsen, *Linkage Disequilibrium as a Signature of Selective Sweeps.* Genetics, 2004. **167**(3): p. 1513-1524.

65. O'Reilly, P.F., E. Birney, and D.J. Balding, *Confounding between recombination and selection, and the Ped/Pop method for detecting selection.* Genome Research, 2008. **18**(8): p. 1304-1313.

66. Liebert, A., S. Lopez, B.L. Jones, et al., *World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection.* Human Genetics, 2017. **136**(11-12): p. 1445-1453.

67. Comeron, J.M., R. Ratnappan, and S. Bailin, *The Many Landscapes of Recombination in Drosophila melanogaster.* Plos Genetics, 2012. **8**(10): p. 21.

68. Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences.* Nature, 2011. **475**(7357): p. 493-U84.

69. Hernandez, R.D., J.L. Kelley, E. Elyashiv, et al., *Classic Selective Sweeps Were Rare in Recent Human Evolution.* Science, 2011. **331**(6019): p. 920-924.

70. Metz, C.E., *BASIC PRINCIPLES OF ROC ANALYSIS.* Seminars in Nuclear Medicine, 1978. **8**(4): p. 283-298.

71. McClish, D.K., *ANALYZING A PORTION OF THE ROC CURVE.* Medical Decision Making, 1989. **9**(3): p. 190-195.

72. Winckler, W., S.R. Myers, D.J. Richter, et al., *Comparison of fine-scale recombination rates in humans and chimpanzees.* Science, 2005. **308**(5718): p. 107-111.

73. Kemper, K.E., S.J. Saxton, S. Bolormaa, et al., *Selection for complex traits leaves little or no classic signatures of selection.* Bmc Genomics, 2014. **15**: p. 14.

74. Jensen, J.D., M. Foll, and L. Bernatchez, *The past, present and future of genomic scans for selection.* Molecular Ecology, 2016. **25**(1): p. 1-4.

75. Comeron, J.M., *Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome.* Plos Genetics, 2014. **10**(6): p. 18.

76. Villanueva-Cañas, J.L., G.E. Rech, M.A.R. de Cara, et al., *Beyond SNPs: how to detect selection on transposable element insertions.* Methods in Ecology and Evolution, 2017. **8**(6): p. 728-737.

77. Merenciano, M., A. Ullastres, M.A.R. de Cara, et al., *Multiple Independent Retroelement Insertions in the Promoter of a Stress Response Gene Have Variable Molecular and Functional Effects in Drosophila.* PLOS Genetics, 2016. **12**(8): p. e1006249.

78. Randhawa, I.A.S., M.S. Khatkar, P.C. Thomson, et al., *Composite Selection Signals for Complex Traits Exemplified Through Bovine Stature Using Multibreed Cohorts of European and African <em>Bos taurus</em>.* G3: Genes|Genomes|Genetics, 2015. **5**(7): p. 1391-1401.

79. Jain, K. and W. Stephan, *Modes of Rapid Polygenic Adaptation.* Molecular Biology and Evolution, 2017. **34**(12): p. 3169-3175.

80. Provine, W.B., *Ernst Mayr.* Genetics and Speciation, 2004. **167**(3): p. 1041-1046.

81. Jensen, J.D., Y. Kim, V.B. DuMont, et al., *Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data.* Genetics, 2005. **170**(3): p. 1401-1410.

82. Pavlidis, P. and N. Alachiotis, *A survey of methods and tools to detect recent and strong positive selection.* Journal of Biological Research-Thessaloniki, 2017. **24**: p. 17.

83. McCoy, R.C. and J.M. Akey, *Selection plays the hand it was dealt: evidence that human adaptation commonly targets standing genetic variation.* Genome Biology, 2017. **18**(1): p. 139.

84. Stamatakis, A. and F. Izquierdo-Carrasco, *Result verification, code verification and computation of support values in phylogenetics.* Briefings in Bioinformatics, 2011. **12**(3): p. 270-279.

85. Pavlidis, P., J.D. Jensen, W. Stephan, et al., *A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans.* Mol Biol Evol, 2012. **29**(10): p. 3237-48.

86. Vatsiou, A.I., E. Bazin, and O.E. Gaggiotti, *Detection of selective sweeps in structured populations: a comparison of recent methods.* Molecular Ecology, 2016. **25**(1): p. 89-103.

87. Fumagalli, M. and M. Sironi, *Human genome variability, natural selection and infectious diseases.* Current Opinion in Immunology, 2014. **30**(Supplement C): p. 9-16.

88. Cadzow, M., J. Boocock, H.T. Nguyen, et al., *A bioinformatics workflow for detecting signatures of selection in genomic data.* Frontiers in Genetics, 2014. **5**: p. 293.

89. Ewing, G. and J. Hermisson, *MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus.* Bioinformatics, 2010. **26**(16): p. 2064-2065.

90. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> 2017

91. Hunter-Zinck, H., *tsel: Time to most recent common ancestor selection inference (TSel)*. 2014

92. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language.* Bioinformatics, 2004. **20**(2): p. 289-290.

93. Robin, X., N. Turck, A. Hainard, et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**(1): p. 77.