Bioinformatics, 31(24), 2015, 3946–3952
doi: 10.1093/bioinformatics/btv493
Advance Access Publication Date: 26 August 2015
Original Paper

Genetic and population analysis

Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations

Marc Pybus[1,†], Pierre Luisi[1,2,†], Giovanni Marco Dall'Olio[1,3,†], Manu Uzkudun[1], Hafid Laayouni[1,4], Jaume Bertranpetit[1,*] and Johannes Engelken[1]

[1]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain, [2]Department of Biology, Stanford University, Stanford, CA 94305, USA, [3]Division of Cancer Studies, King's College of London, London SE1 1UL, UK and [4]Departament de Genètica i de Microbiologia, Universitat Autonòma de Barcelona, Bellaterra 8193, Spain

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on May 11, 2015; revised on July 21, 2015; accepted on August 17, 2015

Abstract

Motivation: Detecting positive selection in genomic regions is a recurrent topic in natural population genetic studies. However, there is little consistency among the regions detected in several genome-wide scans using different tests and/or populations. Furthermore, few methods address the challenge of classifying selective events according to specific features such as age, intensity or state (completeness).

Results: We have developed a machine-learning classification framework that exploits the combined ability of some selection tests to uncover different polymorphism features expected under the hard sweep model, while controlling for population-specific demography. As a result, we achieve high sensitivity toward hard selective sweeps while adding insights about their completeness (whether a selected variant is fixed or not) and age of onset. Our method also determines the relevance of the individual methods implemented so far to detect positive selection under specific selective scenarios. We calibrated and applied the method to three reference human populations from The 1000 Genome Project to generate a genome-wide classification map of hard selective sweeps. This study improves detection of selective sweep by overcoming the classical selection versus no-selection classification strategy, and offers an explanation to the lack of consistency observed among selection tests when applied to real data. Very few signals were observed in the African population studied, while our method presents higher sensitivity in this population demography.

Availability and implementation: The genome-wide results for three human populations from The 1000 Genomes Project and an R-package implementing the 'Hierarchical Boosting' framework are available at http://hsb.upf.edu/.

Contact: jaume.bertranpetit@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the past decades, many different methods to detect positive selection from polymorphism data in genomic regions have been developed (for reviews see Biswas and Akey, 2006; Sabeti *et al.* 2006; Vallender and Lahn, 2004; Vitti *et al.*, 2013). Such methods rely on the different genomic patterns left by a hypothetical selection event occurring in an idealized population: a beneficial *de novo* mutation arises and increases in frequency in relatively few generations until eventually it reaches population fixation. This specific mode of positive selection is known as the hard sweep model (Maynard-Smith and Haigh, 1974). Through the effect of genetic hitchhiking, this process leaves some characteristic patterns in the region surrounding the beneficial allele (selective sweep), such as skewed site frequency spectrum toward low-frequency variants (Braverman *et al.*, 1995), strong linkage disequilibrium (LD) patterns leading to extended haplotype homozigosity (EHH) (Stephan *et al.*, 2006), and population differentiation (Beaumont and Balding, 2004; Weir and Cockerham, 1984). Computational methods developed to distinguish such patterns have helped to identify the genetic basis of some examples of human adaptation, such as the lactase persistence allele (Bersaglieri *et al.*, 2004; Tishkoff *et al.*, 2007) or the malaria resistance gene variants (Ayodo *et al.*, 2007; Hamblin and Di Rienzo, 2000; Sabeti *et al.*, 2002a; Tishkoff *et al.*, 2001). However, most of those methods usually lack consistency in reporting the same selective events along the genome (Akey, 2009), causing a loss of confidence in the approach. This disagreement was thought to appear due to specific power of the different methods to uncover selection patterns under some local features of a given genomic region (such as specific recombination map), or due to specific demographic dynamics of the studied populations. Thus, during the last decade, special effort was made to incorporate population-specific demographic models and region-specific recombination maps to approximate the neutral model to more complex and realistic scenarios (Lohmueller *et al.*, 2011; Pickrell *et al.*, 2009; Voight *et al.*, 2006; Zeng *et al.*, 2006, 2007). While these approaches clearly improve the sensitivity to detect positive selection, they did not explain the continued lack of concordance between methods, which raised concerns on false positive and false negative rates (Kelley *et al.*, 2006; Teshima *et al.*, 2006). Recently, a new family of selection tests started to appear: statistics based on composite approaches combining different positive selection tests, and tuned using neutral and selection simulations (Grossman *et al.*, 2010, 2013; Lin *et al.*, 2011; Ronen *et al.*, 2013). Individual tests are sensitive to different modes and tempos of adaptation according to the specific molecular pattern they are aimed to identify. However, individual positive selection tests and most of the composite methods implemented so far address the selection analysis as a binary classification problem (i.e. selection versus no selection), ignoring the specific features of the analyzed selective sweeps, such as the extent of completeness (i.e. final frequency of the selected allele) or the time depth of the event (recent versus ancient selective events). In this study, we developed a hierarchical classification framework based on a *boosting* algorithm (using a similar approach to that presented in Lin *et al.*, 2011) to scan the whole genome for several evolutive scenarios. We trained the algorithm with simulations under different selective scenarios considering different final allele frequencies (FAF) for the selected allele (completeness of a selective sweep) and with different time-spanning selective events (age of a selective sweep). We demonstrate that our approach achieves high sensitivity toward general hard selective sweeps, and can provide information about their relative age of onset. Once calibrated, we applied it to empirical genome-wide data from The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012). We provide the results of our analysis for three continental human populations as UCSC tracks that can be easily loaded in any UCSC Genome Browser server and an R-package implementing the ‘Hierarchical Boosting’ framework (http://hsb.upf.edu/). We detected a ~13-fold and ~15-fold decrease in significant signals in the African population as compared with the European and East Asian populations, respectively, while our method showed better sensitivity in African-ancestry simulations.

2 Methods

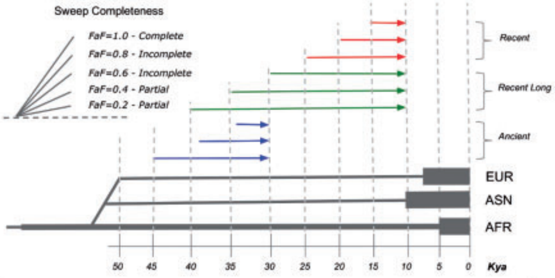
The classification method described in this study is based on a machine-learning algorithm called *boosting* (from the *mboost* R package—Bühlmann and Hothorn, 2008). *Boosting* is a supervised algorithm that estimates linear regressions (hereafter referred as *boosting* functions) of input variables (summary statistics of selection tests) to maximize the differences between two competing scenarios (e.g. complete versus incomplete selective sweeps). Our method sequentially applies different *boosting* functions into a hierarchical classification scheme to optimally classify genomic regions into different evolutive regimes.

2.1 Reference empirical data

We downloaded genome-wide single nucleotide variant (SNV) data representing three continental human populations—Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing, China (CHB) and Utah residents with Northern and Western European ancestry, USA (CEU)—from the low-coverage Phase I release (April 2012) of The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012). To avoid biases in our analysis due to heterogeneous power to detect positive selection across the genome (Fagny *et al.*, 2014), we filtered out any indel and high-coverage SNVs annotated in this release. We also took into account the well-described low SNP calling sensitivity toward rare variants (The 1000 Genomes Project Consortium, 2012) by applying a ‘singleton thinning’ strategy to the simulated dataset as explained in Extended Methods (Supplementary File S1). The SNV data were already phased by The 1000 Genomes Consortium and its phasing state was kept in order to apply haplotype-based statistics. We also used both the ancestral allele state genome and the global genetic map provided by the consortium.

2.2 Coalescent simulations

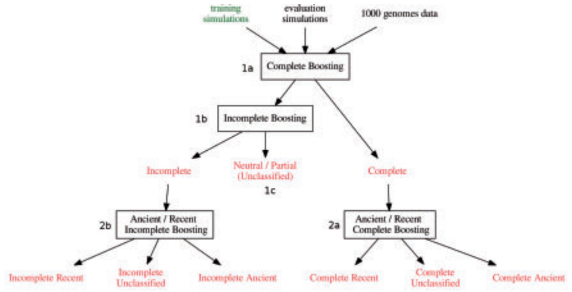
We used the coalescent simulator *cosi* (version 1.2.1; see the initial description in Schaffner *et al.*, 2005) which includes a tuned human demography for three continental populations of Northern Europe, East Asian and African ancestry (CEU, JPT/CHB and YRI, respectively). In addition to the neutral scenario, *cosi* can simulate classic selective sweeps (i.e. under the hard sweep model) under specific constraints (Grossman *et al.*, 2010): *cosi* does not allow any population effective size change or migration between populations while selection is occurring. Accordingly, selective sweeps were simulated in a period when population effective sizes do not change in any population (between 10 and 45 Kya; Fig. 1). Nine different time-spanning selective sweeps were simulated covering different time periods between 10 and 45 Kya (thereafter grouped as Recent, Recent Long or Ancient selective sweeps). And within each time-spanning selective sweep, we simulated five different FAF for the selected allele (and grouped them as Complete, Incomplete or Partial selective sweeps, as explained below). Mutation rate and generation



CLASSES	Final Allele Frequency	TYPES		
		Recent (25 to 10 Kya)	Recent Long (40 to 10 Kya)	Ancient (45 to 30 Kya)
Complete Sweep	1.0	300 replicates	300 replicates	300 replicates
Incomplete Sweep	0.8 0.6	300 replicates 300 replicates	300 replicates 300 replicates	300 replicates 300 replicates
Partial Sweep	0.4 0.2	300 replicates 300 replicates	300 replicates 300 replicates	300 replicates 300 replicates
Neutral		3000 replicates		

* This simulation scheme was duplicated for evaluation purposes.
* Selection simulation were run for each population demography independently

Fig. 1. Coalescent simulations were run following a calibrated human demographic model (Schaffner *et al.*, 2005) mimicking population genetic data from three reference continental populations (YRI, CEU and JPT/CHB). Nine different time-spanning selective sweeps were simulated (grouped as Neutral, Recent, Recent Long and Ancient) allowing for five different FAF (FAF = 0.2, 0.4, 0.6, 0.8 and 1.0)



Complete Boosting	Complete	*vs*	Incomplete Partial Neutral
Incomplete Boosting	Incomplete	*vs*	Partial Neutral
Ancient/Recent Complete Boosting	Complete Ancient	*vs*	Complete Recent
Ancient/Recent Incomplete Boosting	Incomplete Ancient	*vs*	Incomplete Recent

Fig. 2. The implemented 'Hierarchical Boosting' classification tree

time was set to 1.5×10^{-8} mutations/year/bp and 25 years, respectively. We also used the hotspot recombination model implemented in the simulation package (*recosim*) in order to obtain more realistic genome-wide LD patterns. More details of the demographic model and selection parameters used in our simulations are provided in Supplementary File S2 and Supplementary Table S5. We computed 3000 replicates for the neutral scenario and 100 replicates for each one of the 45 selection scenarios. For each replicate, we simulated regions of 600 kbp to allow extended homozygosity statistics to calculate properly the EHH decay. Sample size in our simulations was 97, 85 and 88 diploid individuals for CEU, CHB and YRI populations, respectively, matching the sample size of the reference dataset.

2.3 Implemented positive selection tests

We previously had implemented a bioinformatic pipeline including 21 different positive selection statistics (Supplementary Table S7). The pipeline design and detailed descriptions for the selection tests can be found in Pybus *et al.* (2014). We ran our positive selection tests to the whole 600 kbp simulated sequence after applying a 'singleton thinning' strategy, as explained in Extended Methods (Supplementary File S1). Then, we used the results from the central 25 kbp region containing the selected allele (and driving the signal of selection) to train our *boosting* algorithms and evaluate the method. Since we needed a unique score for every 25 kbp region to apply the *boosting* algorithm, we used a specific summary statistic for each test (minimum, maximum or mean), as shown in Extended Methods (Supplementary File S1) and Supplementary Table S2. Next, we had to remove some selection tests from the analysis. Some cross population tests were not suitable to be combined under our framework, such as F_{ST}, and others needed some corrections before using them, as with XPEHH or dDAF. In addition, some correlated tests were removed to achieve coefficient convergence and avoid over-fitting during the algorithm training process. The details of this selection process is explained in Extended Methods (Supplementary File S1) and summarized in Supplementary Table S4. The final list of selection tests used in the training process were: CLR (Nielsen *et al.*, 2005), iHS (Voight *et al.*, 2006), XP-CLR (Chen *et al.*, 2010), XP-EHH (Sabeti *et al.*, 2002b), dDAF (Hofer *et al.*, 2009), diHH (Voight *et al.*, 2006), Fay and Wu's *H* (Fay and Wu, 2000), Omega (Pavlidis *et al.*, 2010), EHH Av (Sabeti *et al.*, 2002b), Fu and Li's *D* (Fu and Li, 1993) and Tajima's *D* (Tajima, 1989).

2.4 The hierarchical boosting framework

We define a *boosting* function as a linear regression function of the scores of individual positive selection tests. This function is estimated through a *boosting* algorithm (Bühlmann and Hothorn, 2008; Hothorn *et al.*, 2010), and in turn, can be used as a classification method by setting up a significance threshold. In our framework, four different *boosting* functions were sequentially considered within a hierarchical decision tree implementation (Fig. 2).

2.4.1 Competing scenarios

We grouped the simulated datasets according to common selective sweep features. Hence, we created classes of different selection scenarios according to the FAF of the selected variant as main property or Class (Complete: FAF = 1.0, Incomplete: FAF = 0.8 and 0.6, Partial: FAF = 0.4 and 0.2 and Neutral scenarios). Concurrently, the scenarios were grouped again according to the number of generations elapsed since the end of the simulated selective sweep or Types (Recent: from 25 to 10 Kya, Ancient: from 45 to 30 Kya) (Fig. 1). Once our competing scenarios (Fig. 2) were defined, and the 11 positive selection tests were calculated, we ran the machine-learning algorithm to train four *boosting* functions which allowed us to distinguish between competing groups of scenarios (Supplementary Table S3). We systematically verified coefficient convergence for every estimated *boosting* function (Supplementary Fig. S14). To circumvent a putative convergence to local instead of global optimal, and thus, to obtain a more robust regression, we developed a bootstrapping strategy explained in Extended Methods (Supplementary File S1). We used the mean coefficient values for each positive selection test obtained across the bootstraps to build our final *boosting* functions. Then, we calculated the thresholds for the estimated regression scores that were needed to classify the evaluation datasets allowing up to a 1% false positive rate (FPR).

2.4.2 Algorithm description

After noticing that the main feature driving selection signals in almost all the selection tests was FAF, we decided to classify complete and incomplete selective sweeps as a first step (or first level) in our hierarchical classification framework (Fig. 2). Thus, a first *boosting* function (Complete Boosting—function 1a) comparing complete sweep cases against incomplete, partial and neutral cases was trained and calibrated (at 1% FPR) to be sensible exclusively to complete selective sweeps. Once complete sweep signals were classified/discarded, a second *boosting* function (Incomplete Boosting—function 1b) was trained and calibrated (1% FPR) to distinguish between incomplete sweep cases, and partial and neutral cases. At this point, the cases classified as partial or neutral were left unclassified, as the sensitivity of the selection tests to partial sweeps was extremely low and hardly distinguishable from neutral cases. Once this first 'classification level' was finished, we trained specific *boosting* functions (Recent/Ancient Boosting—functions 2a/2b) for each one of the categories obtained in the previous step (Complete/Incomplete scenarios) to distinguish between recent and ancient cases. Again, thresholds of 1% FPR were used, so only true cases of each category could be correctly classified. Because of the difficulty of assigning the age category with accuracy, many cases remained with age category undefined at this second step. According to the hierarchical nature of the algorithm, a given genomic region (either empirical or simulated) is sequentially classified into the nine categories defined in the decision tree (Fig. 2).

2.4.3 Alternative classification trees

Two more alternative classification trees were tested but presented lower performance: an 'independent' configuration showed higher misclassification rates and a 'non-hierarchical' configuration produced a lower accuracy overall. See the details for the alternative classification trees in Supplementary Figure S2 and their classification power in Supplementary Table S6.

3 Results

3.1 Method performance

Using an independent set of evaluation simulations with the same parameters as the training dataset, we evaluated our framework performance. We calculated population-average classification power for each scenario in the evaluation dataset (Supplementary Fig. 3A and B). Similar population-specific tables are found in Supplementary Table S1. Population-averaged, our Hierarchical Boosting method was able to classify the evaluation scenarios with low number of misclassified replicates (5.37% of Complete sweeps classified as incomplete or vice versa), and even lower false positive cases (Neutral and Partial scenarios classified either as Complete or Incomplete; 1.06%). Nonetheless many cases of Incomplete sweeps were left unclassified (47.95%), making Hierarchical Boosting a conservative method for Incomplete sweeps. It also showed different correct classification rates depending on the scenario: Complete sweeps were easier to classify (89.58%) than Incomplete sweeps (43.41%), probably because most of the positive selection tests were implemented to detect hard sweeps that already reached fixation. Concerning the Ancient/Recent classification, our method is again conservative, leaving 66.28% evaluation replicates unclassified while showing an extremely low misclassification rate (1.26%). Figure 3B shows the classification power of the time-frame *boosting* functions (Ancient/Recent) for those replicates that have passed the Complete/Incomplete classification step. In this case, Ancient sweeps (both Complete and Incomplete) were slightly more well classified (33.35%) than Recent sweeps (31.56%). When looking at population-specific performance (Fig. 3C; Supplementary Table S1) we noted that Hierarchical Boosting performed better in correctly classifying selective sweeps for the simulated African-ancestry population (93.44% for Complete and 52.27% for Incomplete sweeps) than in the simulated Out-of-Africa populations (89.42% for Complete and 37.44% for Incomplete sweeps in the European population, 85.87% for Complete and 40.50% for Incomplete sweeps in the East Asian population).

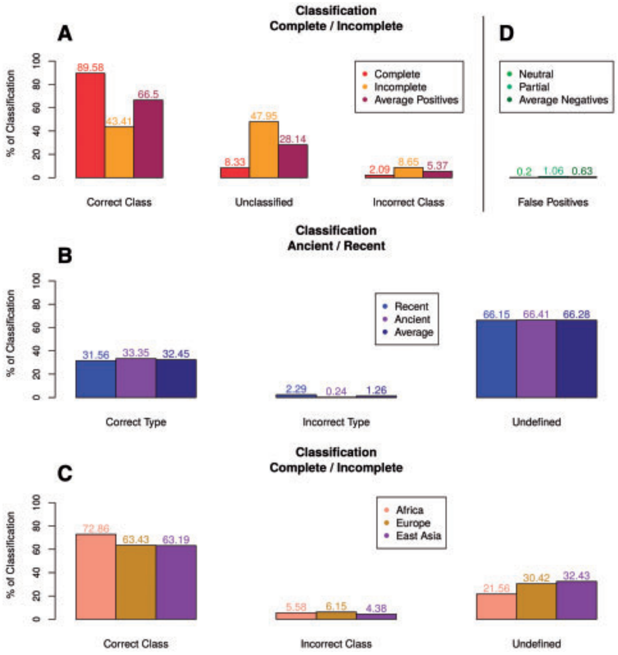


Fig. 3. Population-averaged classification power of Complete and Incomplete scenarios (**A**) Population-averaged classification power for Recent and Ancient scenarios (within correctly classified complete and incomplete scenarios) (**B**) Population-specific classification power for Incomplete and Complete scenarios in our demography-specific simulations (**C**) FPR at 'negative' scenarios (Neutral/Partial scenarios) (**D**)

3.2 Comparison with other composite methods

We used three state-of-the-art composite methods—evolBoosting (Lin *et al.*, 2011), CMS (Grossman *et al.*, 2010, 2013) and SFselect (Ronen *et al.*, 2013)—to detect positive selection to compare their performance with our method. These methods are also tuned using neutral and selection simulations, thus we consider them ideal for a comparison analysis. However, all of them confront the problem of detecting selection as a binary outcome (i.e. selection versus no selection). Instead, our method tries to give insights about the nature of an observed sweep, thus it classifies it in more than two categories. We believe that this approach increases the value of our method regardless of its sensitivity to general hard selective sweeps. To allow a fair comparison between the methods evaluated, we used the Complete and Incomplete *boosting* functions independently (outside of the classification framework). We applied those methods to the evaluation simulations for the European population, and used the distribution of the central 25 kbp region on neutral simulations to calculate the 1% FPR threshold to, in turn, calculate their sensitivity at our selection simulations (always using the central 25 kbp region). A full description of the methods implementation can be found in Supplementary File S4. Table 1 shows the resulting sensitivity of each composite method for the selective sweep categories defined in this study. We also evaluated the FPR at Neutral simulations for all the methods as well. Our method (more specifically, the Incomplete

Table 1. Sensitivity analysis (in percentage) for different composite methods in European-ancestry simulations

	Neutral	Complete		Incomplete	
		Recent	Ancient	Recent	Ancient
'Complete' Boosting[a]	0.80	98.50	98.23	8.08	17.84
'Incomplete' Boosting[a]	1.23	98.13	96.46	91.91	59.25
evolBoosting (1% FPR)	1.47	97.67	88.67	81.00	57.33
SFselect (general)	1.00	63.63	71.38	0.00	2.02
CMS-GW	1.00	33.06	12.45	78.16	49.83
CMS-local	1.00	8.87	1.01	70.17	6.67

[a] In our classification framework both Complete and Incomplete *boosting* functions are considered together. This way, the low sensitivity shown by the Complete *boosting* function toward incomplete sweeps is masked by the Incomplete *boosting* function. In this sensitivity analysis we have used them separately to be comparable to the other methods.

boosting function) showed the highest sensitivities at all the simulated selection scenarios compared with the other implemented methods, with the lowest sensitivity for the Incomplete Ancient scenario (59.25%). Our Complete *boosting* function shows lower sensitivity at Incomplete scenarios because it was trained not to be sensitive to incomplete selection sweeps. Curiously, the CMS genome-wide and CMS local implementations failed to detect complete sweep signals, most probably because both CMS scores combine results from all the applied tests: iHS and diHH (population-specific tests) cannot be calculated in alleles that have reached fixation, leading to a lack of polymorphism in the central 25 kbp region of our complete sweep simulations. In Supplementary File S4, we discuss why, in our opinion, the CMS functions exhibit much lower sensitivity in our evaluation sets than described in the original articles.

3.3 Application to the 1000 Genomes data

We applied our population-specific 'Hierarchical Boosting' implementations to the reference empirical genome-wide data that was used to calibrate our simulations. We obtained a list of 25 kbp windows per population that were classified according to the different *boosting* functions described above. Then, we implemented an algorithm that estimates the number of selective events by concatenating consecutive 25 kbp windows, allowing for a valley of non-significant scores as long as they do not contain any recombination hotspot. The algorithm implementation is explained in detail in Supplementary File S3 and Supplementary Figures S3–S6. After applying the algorithm we counted 27, 355 and 424 selective events in YRI, CEU and CHB populations, respectively (Supplementary Table S8). A ∼13-fold and ∼15-fold difference in the number of selective events were detected in CEU and CHB populations in relation to YRI population, respectively. In addition, we classified the selective events according to the *boosting* function scores showing significance in the genomic region encompassing the selective sweep signal (Supplementary Table S8). We observed few signals with any ambiguity for the sweep Class (Complete or Incomplete): only 10.2% (0, 7.6 and 12.1% in YRI, CEU and CHB, respectively) of the identified selective sweeps in any of the three populations show significant scores for both Complete and Incomplete *boosting* functions (Supplementary Table S8). Moreover, these ambiguous signals exhibit a much longer size and lower proportion of significant scores as compared with unambiguous (discriminant) signals (Supplementary Figures S4 and S5); hence, most of the ambiguous signals may actually arise from different adjacent independent selective events. Conversely, most of the selective events could not be assigned to a sweep age (Ancient or Recent): only 59.4% of the identified selective sweeps in any of the three populations (44.4, 71.5 and 50.2% in YRI, CEU and CHB, respectively) could be assigned a given age. This demonstrates the difficulty to assess the age of a sweep even considering a large amount of tests. Here, we have designed a very conservative framework for that purpose, as demonstrated by the very low number of regions with a signal that have been assigned to both Recent and Ancient selective sweeps (0, 0.6, 2.4 and 1.5 in YRI, CEU, CHB and any of the three populations, respectively). Moreover, although the classification power of our method is lower for Incomplete scenario we detected more Incomplete than Complete sweeps in the CEU population (58.9 and 33.5%, respectively), and even numbers of Complete and Incomplete sweeps detected in YRI and CHB. Finally, we generated UCSC supertracks to easily visualize our Hierarchical Boosting results in a UCSC Genome Browser server (Kent *et al.*, 2002; Raney *et al.*, 2014). Visualizing selective sweeps in a genome browser helps to properly evaluate their genomic context, and to propose candidate genes under putative positive selection (Supplementary File S5). The provided UCSC tracks represent a novel and unified view of the different types of selection analyzed in this study (complete, incomplete, ancient and recent selective sweeps). Using individual selection tests will only highlight the types of selective sweeps that a given test is sensitive to. For example, iHS alone may be able to detect incomplete selective sweeps, but it will never detect complete ones. Furthermore, our method provides information about the nature of the selective sweep detected (age of onset), which can help to elucidate the biological and historical context of a given selective event. The supertracks and the raw Hierarchical Boosting results can be found in a dedicated server (http://hsb.upf.edu/), as well as a detailed explanation on how to interpret them.

4 Discussion

4.1 Detection and classification of hard selective sweeps

The inference of positive selection using population genetics data usually focuses in determining whether a genomic region exhibits signals of positive selection or not. However, while in simulation studies most of those tests show similar power to detect hard selective sweeps, in real data few of them coincide to point to common selection signals. We believe that this lack of consistency among tests is explained by their different power to detect different types of selective sweeps. Because of that, we think that the strategy of combining shared signals from different selection tests, as implemented in other composite methods, will highlight only those selective sweeps for which all the implemented tests show high detection power, which generally means the 'strong recent hard' selective sweep. We showed that the implementation of a hierarchical classification framework enables classification of a given genomic region into specific adaptive regimes and can explain the lack of consistency of different selection tests in empirical data. We have compared Hierarchical Boosting with other composite methods (SFselect, CMS and evolBoosting) and we found that our method shows higher sensitivity to the different selection scenarios evaluated. In addition, Hierarchical Boosting is the only unifying method that gives information about the nature of the detected selective sweep while preserving great sensitivity toward general hard selective sweeps.

4.2 Interpretation of the estimated boosting functions

We standardized the regression coefficients assigned to each positive selection statistic within a given *boosting* function, as in Lin *et al.*

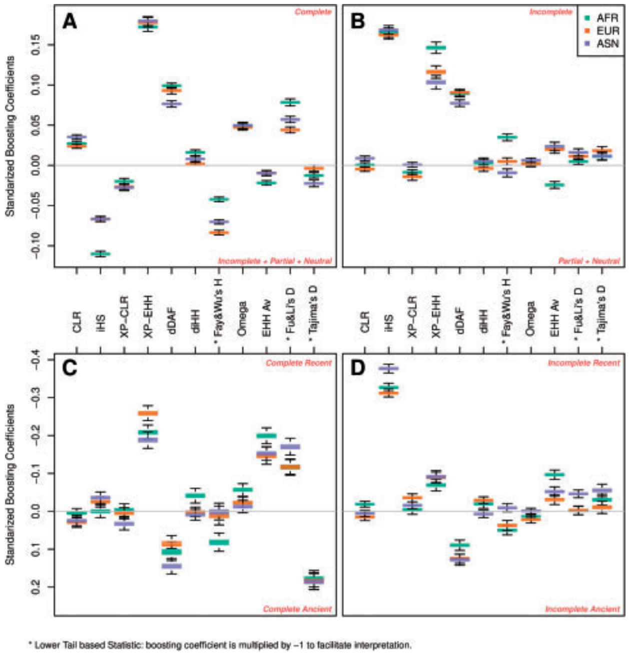


Fig. 4. Standardized coefficients for the three populations and implemented *boosting* functions across bootstraps. Estimated coefficients for each population in the four *boosting* functions used in the classification tree: Complete (**A**), Incomplete (**B**), Complete Recent/Ancient (**C**) and Incomplete Recent/Ancient (**D**). The relevance of the positive selection tests to classify the different scenarios is given by the strength of its standardized coefficient

(2011). These standardized coefficients give an insight into the relevance of a given test to distinguish between two competing scenarios (Fig. 4). We found that to indicate Complete sweeps, the statistics positively correlated are, in order of importance, XP-EHH, dDAF, Fu & Li's *D*, Omega and CLR. Conversely, the statistics negatively correlated to a Complete sweep signal are iHS, Fay & Hu's *H* and XP-CLR. Once Complete sweep signals are classified, the Incomplete *boosting* function is applied to the remaining ones. In this case, the statistics indicating Incomplete sweeps are iHS, XP-EHH and dDAF. For the *boosting* functions uncovering time-frame properties of a selective sweep, we observed that in Complete Recent sweeps, XP-EHH contributed the most along with EHH Average, and Fu & Li's *D*. Instead, for Complete Ancient sweeps, Tajima's *D* and dDAF are the more relevant ones. Within Incomplete sweep cases, iHS highlights recent selection patterns while dDAF defines older selective events. Those results are concordant to the sensitivity described for individual tests given the final frequency of the selected allele and the tempo of selection (e.g. see Sabeti *et al.*, 2006). Finally, we notice that our three population-specific *boosting* functions showed very similar coefficients, indicating that the method is robust to continental human demography (Fig. 4).

4.3 Missing hard sweep signals in Yoruba population

We report fewer selective events or regions under selection (∼13-fold and ∼15-fold reduction) in African-ancestry populations (YRI) than Out-of-Africa populations (CEU and CHB, respectively). Low number of selective events in several African populations was also reported in Granka *et al.* (2012). But the authors acknowledged that this result could be explained by low sensitivity of the implemented tests of selection on genotyping data in African populations. On the contrary, our method shows greater power to uncover selective events for African-like demography than Out-of-Africa ones (CEU and JPT/CHB) as shown in Supplementary Table S1. The dearth of selection footprints in African populations could be explained by selection acting on standing variation in African-ancestry populations (soft sweep), rather than *de novo* mutations (hard sweep). Moreover, the Out-of-Africa human diaspora likely occurred through serial founder effects, a specific case of population bottlenecks. Such a demographic scenario seems to increase the fixation rates of *de novo* favored alleles (Coop *et al.*, 2009). Moreover, Wilson *et al.* (2014) recently showed that population bottlenecks can bring a soft selective sweep to generate molecular footprints that are expected under the hard sweep model (hardening of soft sweeps). In fact, under a serial founder demography, it is likely that a unique haplotype carrying the standing favored mutation is sampled during a strong bottleneck event (Messer and Petrov, 2013). This would imply that complete hard sweep signals (even though they started as soft sweep) should be more frequent in Out-of-Africa populations, as observed in our study. Nonetheless we demonstrate that hard sweeps have definitely not been common in African populations (or at least in the Yoruba population studied here), and underlines the still crucial role of demography in understanding human adaptation (Coop *et al.*, 2009).

4.4 Perspectives

The statistical framework applied in this study is based on the estimation of composite scores of selection tests (*boosting* functions) that maximize the differences between two competing scenarios. Appropriate thresholds are then used to produce a binary outcome. These competing scenarios (and their *boosting* functions) are embedded in a hierarchical classification tree, which is structured according to the relevance of the properties that can or want to be classified. Unlike multinomial logistic regression and other multi-category algorithms, our framework allows to set up significance thresholds at each classified category and hierarchical step. This way, it is possible to minimize the FPR for a given scenario, or relax classification accuracy for other ones. We believe that this flexibility increases the value of our statistical framework when applied to the study of natural selection. A possible improvement in classification accuracy could imply the use of non-linear classifier algorithms (like some variations of SVMs, ANNs or *k*-nearest neighbors algorithms), as the response of the selection tests to different selection scenarios is likely to be non-linear. It is still unknown whether other boosting-like algorithms, like gradient boosting, may improve the strategy outlined here. This study offers a unique and powerful way of detecting candidate regions in the genome that have been evolving under positive selection in a more reliable way than many lists produced by single selection tests or even some other existing composite methods. It also distinguishes, in many cases, the final state (complete/incomplete) and the relative age (ancient/recent) of a given selective event. Our framework implementation emphasizes the minimization of false positive results, even if it implies a number of unclassified results. Thus, we give strong support to the cases in which positive selection is detected.

Funding

This work was supported by Ministerio de Economía y Competitividad (Spain) [grants BFU2010-19443, BFU2013-43726-P]; and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya [GRC 2014 SGR 866] to J.B. M.P. and G.D. have been supported by a grant of the FPI program, Ministerio de Economia y Competitividad; P.L. by a grant from the Instituto de Salud Carlos III; J.E. was supported through a Postdoc scholarship from the Volkswagenstiftung [Az: I/85 198].

Conflict of Interest: none declared.

References

Akey,J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.

Ayodo,G. *et al.* (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am. J. Hum. Genet.*, **81**, 234–242.

Beaumont,M.A. and Balding,D.J. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.*, **13**, 969–980.

Bersaglieri,T. *et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.*, **74**, 1111–1120.

Biswas,S. and Akey,J.M. (2006) Genomic insights into positive selection. *Trends Genet.*, **22**, 437–446.

Braverman,J.M. *et al.* (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**, 783–796.

Bühlmann,P. and Hothorn,T. (2008) Rejoinder: boosting algorithms: regularization, prediction and model fitting. *Stat. Sci.*, **22**, 477–505.

Chen,H. *et al.* (2010) Population differentiation as a test for selective sweeps. *Genome Res.*, **20**, 393–402.

Coop,G. *et al.* (2009) The role of geography in human adaptation. *PLoS Genet.*, **5**, 1000500.

Fagny,M. *et al.* (2014) Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing datasets. *Mol. Biol. Evol.*, **31**, 1850–1868.

Fay,J.C. and Wu,C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Granka,J.M. *et al.* (2012) Limited evidence for classic selective sweeps in African populations. *Genetics*, **192**, 1049–1064.

Grossman,S.R. *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 883–886.

Grossman,S.R. *et al.* (2013) Identifying recent adaptations in large-scale genomic data. *Cell*, **152**, 703–713.

Hamblin,M.T. and Di Rienzo,A. (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.*, **66**, 1669–1679.

Hofer,T. *et al.* (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann. Hum. Genet.*, **73**, 95–108.

Hothorn,T. *et al.* (2010) Model-based Boosting 2.0. *J. Mach. Learn. Res.*, **11**, 2109–2113.

Kelley,J.L. *et al.* (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.*, **16**, 980–989.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Lin,K. *et al.* (2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics*, **187**, 229–244.

Lohmueller,K.E. *et al.* (2011) Detecting directional selection in the presence of recent admixture in African-Americans. *Genetics*, **187**, 823–835.

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

Messer,P.W. and Petrov,D.A. (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.*, **28**, 659–669.

Nielsen,R. *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**, 1566–1575.

Pavlidis,P. *et al.* (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, **185**, 907–922.

Pickrell,J.K. *et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.*, **19**, 826–837.

Pybus,M. *et al.* (2014) 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.*, **42**, 1–7.

Raney,B.J. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

Ronen,R. *et al.* (2013) Learning natural selection from the site frequency spectrum. *Genetics*, **195**, 181–193.

Sabeti,P. *et al.* (2002a) CD40L association with protection from severe malaria. *Genes Immun.*, **3**, 286–291.

Sabeti,P.C. *et al.* (2002b) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Sabeti,P.C. *et al.* (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.

Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.

Stephan,W. *et al.* (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, **172**, 2647–2663.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Teshima,K.M. *et al.* (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Res.*, **16**, 702–712.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1 092 human genomes. *Nature*, **491**, 56–65.

Tishkoff,S. *et al.* (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, **293**, 455–462.

Tishkoff,S. *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, **39**, 31–40.

Vallender,E.J. and Lahn,B.T. (2004) Positive selection on the human genome. *Hum. Mol. Genet.*, **13**, 245–254.

Vitti,J.J. *et al.* (2013) Detecting natural selection in genomic data. *Annu. Rev. Genet.*, **47**, 97–120.

Voight,B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.

Weir,B. and Cockerham,C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wilson,B. *et al.* (2014) Soft selective sweeps in complex demographic scenarios. *Genetics*, **198**, 669–684.

Zeng,K. *et al.* (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.

Zeng,K. *et al.* (2007) Compound tests for the detection of hitchhiking under positive selection. *Mol. Biol. Evol.*, **24**, 1898–1908.