BMC Research Notes

RESEARCH ARTICLE Open Access

Improved haplotype-based detection of ongoing selective sweeps towards an application in Arabidopsis thaliana

Torsten Günther* and Karl J Schmid

Abstract

Background: The increasing amount of genome information allows us to address various questions regarding the molecular evolution and population genetics of different species. Such genome-wide data sets including thousands of individuals genotyped at hundreds of thousands of markers require time-efficient and powerful analysis methods. Demography and sampling introduce a bias into present population genetic tests of natural selection, which may confound results. Thus, a modification of test statistics is necessary to introduce time-efficient and unbiased analysis methods.

Results: We present an improved haplotype-based test of selective sweeps in samples of unequally related individuals. For this purpose, we modified existing tests by weighting the contribution of each individual based on its uniqueness in the entire sample. In contrast to previous tests, this modified test is feasible even for large genome-wide data sets of multiple individuals. We utilize coalescent simulations to estimate the sensitivity of such haplotype-based test statistics to complex demographic scenarios, such as population structure, population growth and bottlenecks. The analysis of empirical data from humans reveals different results compared to previous tests. Additionally, we show that our statistic is applicable to empirical data from *Arabidopsis thaliana*. Overall, the modified test leads to a slight but significant increase of power to detect selective sweeps among all demographic scenarios.

Conclusions: The concept of this modification might be applied to other statistics in population genetics to reduce the intrinsic bias of demography and sampling. Additionally, the combination of different test statistics may further improve the performance of tests for natural selection.

Background

The recent advent of genome-wide surveys of genetic variation provides the opportunity to study genome-wide patterns of selection in model species. Such genome-wide scans detected new candidate regions for positive selection as well as previously identified target genes for selection, which included the lactase gene in European humans [1] or *FRIGIDA* in *Arabidopsis thaliana* [2].

Based on the assumption that the frequency of a new advantageous allele increases rapidly and that extended linkage disequilibrium (LD) around the selected site is expected [3,4], several tests for selective sweeps were designed in the last years [1,2,5-9]. The power of detecting selection with these haplotype-based tests was estimated to be higher than with frequency-based statistics as Tajima's D [1]. Although it is known that demographic history may cause a similar departure from the neutral model than selective sweeps and that test statistics are highly sensitive to these scenarios [10-14], only the pairwise haplotype sharing score (PHS, [2]) corrects for demographic history and relatedness. Unfortunately, because of pairwise comparisons between individuals for each allele, calculating the PHS has a complexity of $O(n^2)$ and is infeasible for large present and future data sets. However, since demography and unequally related individuals introduce a bias and potentially cause flawed results in sweep detection, a correction is required. Population structure also confounds genome-wide

* Correspondence: torsten.guenther@uni-hohenheim.de
Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

association studies and several approaches were developed to circumvent these problems [15]. The ideal sample for an association study as well as for scans for selective sweeps consists of equally related individuals with a star-like phylogeny. For samples from natural populations this assumption is unrealistic.

In order to correct for demographic effects in haplotype-based detection of ongoing selective sweeps, we modified the integrated haplotype score (iHS) statistic introduced by Voight et al. [1] by weighting the contribution of each individual according to its genetic similarity to all other individuals in the sample. Closely related individuals generally share more alleles and haplotypes because of common ancestry. The concept of weighting to account for an unequally related sample is already established in other fields of evolutionary analysis. It was introduced as branch-proportional sequence weighting in the construction of sequence profiles from homologous proteins [16] and also has been shown to improve the accuracy of multiple sequence alignments in CLUSTALW [17]. Here, we describe the weighted iHS (WiHS) method as an improved test statistic to detect ongoing selective sweeps. We utilize coalescent simulations of different complex demographic scenarios to estimate the detection power and the false discovery rate of the new method and compare it to existing methods. Finally, we apply the modified test statistic to empirical data from *Arabidopsis thaliana* and humans.

Materials and methods

Test statistic to detect selective sweeps

The new test statistic is based on the integrated haplotype score (iHS, [1]). The iHS is derived from the extended haplotype homozygosity (EHH, [4]) and assumes that selected haplotypes will be longer than the haplotypes around non-selected alleles in the same region because of hitchhiking of linked variation with the selected mutation. The EHH is defined as the probability that two haplotypes with the same core allele at position x are identical over the complete interval between the core site and a position y. The original EHH considers all individuals as equally weighted in the computation of the score.

We modified the EHH to account for unequally related individuals or population structure in the sample by utilizing a matrix of pairwise distances between all individuals. For the present paper, we calculated the squared genome-wide Hamming distance inferred from the genotypes, which performed well in accounting for relatedness in genome-wide association studies [18], but in general any distance metric is applicable. From pairwise distances we derive a measurement of the uniqueness, U, of each individual, I, to characterize the differences of an individual to a set of other individuals and then the contribution of each individual to the test statistic is weighted based on its uniqueness. We define U as

$$U_x(I) = \frac{D_x(I)}{\sum\limits_{I_i \in X} D_x(I_i)},$$

where $D_x(I)$ is the average pairwise distance of individual I to all other individuals carrying the same core allele at position x and X is the set of these individuals. Note that the sum of all uniquenesses for a certain allele is always $\sum_{i=1}^{m} U_x(I_i) = 1$, therefore only the relative weighting between individuals changes, which depends on the set of individuals carrying the same core allele at position x. Such weighting leads to a higher effect of less close related individuals on the test score and thus aims to reduce the bias in the sweep detection caused by unequal relatedness in the sample. The weighted EHH (wEHH) at position y is then computed for all sites with a minor allele frequency of more than 5% as

$$\mathrm{wEHH}_x(y) = \sum_{h \in H} \frac{\sum_{I_i \in h} U(I_i) \times m}{n} \times \frac{\sum_{I_i \in h} U(I_i) \times (m-1)}{(n-1)},$$

where h is a set of individuals carrying the same haplotype between x and y, H is the set of all haplotypes, m is the number of individuals carrying the same core allele at position x and n is the total sample size. For the classical EHH calculation, $\Sigma U(I_i)$ is replaced by the constant 1.

The subsequent steps are then identical to the original iHS approach [1]. We integrate under the wEHH decay around the specified core allele until wEHH reaches 0.05 using the trapezoidal rule. The integrated wEHH (iwEHH) is the sum of this integral in both directions from the core allele using distances on a genetic map to the core site to correct for different local recombination rates. The iwEHH is computed for both, the ancestral and derived allele, at position x, resulting in $\mathrm{iwEHH_A}$ and $\mathrm{iwEHH_D}$, respectively. The unstandardized test statistic of the weighted integrated haplotype score, WiHS hereafter, is then computed as

$$\text{unstandardized WiHS} = \log\left(\frac{\mathrm{iwEHH_A}}{\mathrm{iwEHH_D}}\right).$$

This score is negative if the derived haplotype is larger than the ancestral haplotype and positive if the ancestral haplotype is larger. Since young, low frequency haplotypes are generally longer than old, high frequency haplotypes, we obtain a standardized score for the allele frequency f as follows

$$\mathrm{WiHS} = \frac{\text{unstd. WiHS - mean}_f\text{(unstd. WiHS)}}{SD_f\text{(unstd. WiHS)}},$$

where mean_f is the mean score of all sites with the frequency f and SD_f is the associated standard deviation.

Python scripts used for the tests are available from http://evoplant.uni-hohenheim.de

Simulation of selective sweeps

To assess the power of our method to detect selective sweeps, we applied it to simulated data sets. We simulated populations using the coalescent simulator *msms* [19] and sampled 100 chromosomes of 2 Mbp from the data. 4,000 SNPs with a minor allele frequency ≥ 0.05 were randomly selected from all simulated mutations. This sampling scheme corresponds roughly to the SNP density analyzed with SNP arrays in *A. thaliana* [20]. For each simulation run, a single site under positive selection without recurrent mutations was simulated and realistic mutation and recombination rates from *A. thaliana* were used [21,22]. The simulation parameters are summarized in Table 1. To compare the new method to other haplotype based tests for selective sweeps, we additionally computed the unweighted iHS [1] and the pairwise haplotype sharing score (PHS, [2]) for the simulated data sets, using the same standardization for allele frequency in all tests. For all simulations, a constant recombination rate without recombination hotspots was assumed. The selection coefficient was set to $2N_es = 200$, other values are mentioned in the corresponding sections of the paper.

To evaluate the performance of the modified test statistic on different demographic and selection scenarios, we simulated four different models: a panmictic population, an island model of two subpopulations with migration, an exponential population growth model which represents a realistic model for the European metapopulation of *A. thaliana* (growth model C from [23] with parameters scaled according to our population size), and a recent bottleneck (see Figure 1).

To assess whether high scoring SNPs cluster around the selected site, the absolute values of the scores were averaged in a window of ±25 SNPs around the selected site. These values were then used as final test statistic and compared to a null distribution estimated from neutral simulations of the panmictic model.

Table 1 Parameters for the msms simulations

Parameter		Value
Sequence length	l	2,000,000 bp
Sample size	n	100
Population scaled mutation rate (per site)	θ	$6 \cdot 10^{-3}$
Population scaled recombination rate (per site)	ρ	$8 \cdot 10^{-4}$
Effective population size	N_e	1,000
Number of sampled SNPs		4,000

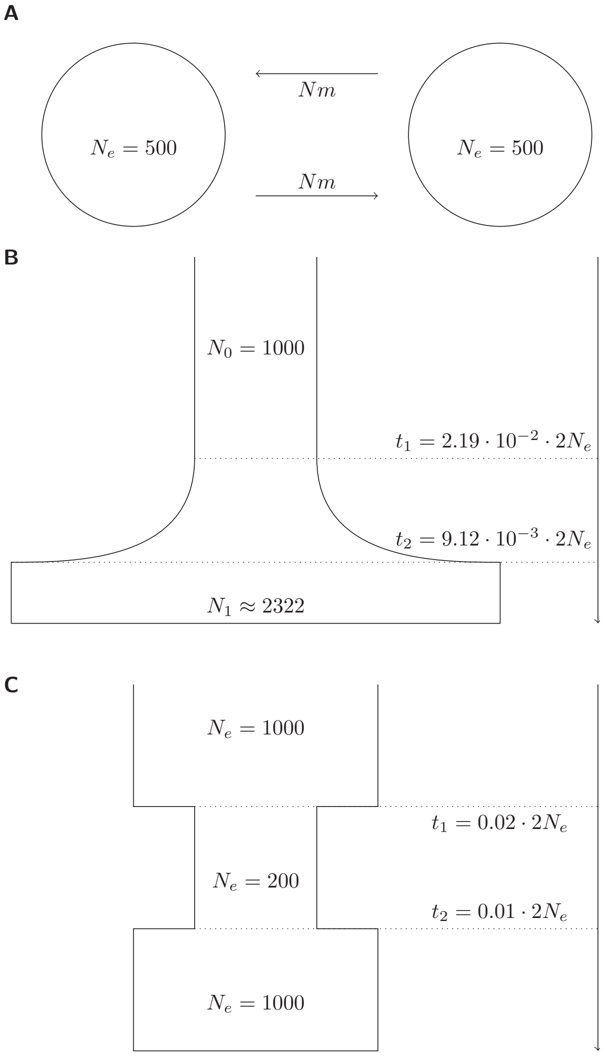

Figure 1 Simulation models. The demographic models used in the simulations included an island model (A), exponential population growth (B) and population bottleneck (C). The parameters of (B) are adjusted to a model by [23]. Time is given in units of $2N_e$ generations.

Application to empirical data sets

We applied our new test statistic to two empirical data sets. The first data set was HapMap 2 [24] of the East Asian (JPT+CHB), European (CEU) and Yoruba (YRI) populations consisting of 120 chromosomes from each population. We included all SNPs for which an ancestral state was available from dbSNP 130 [25]. The estimated recombination rates were downloaded from the HapMap project and a polynomial curve was fitted to the markers for conversion between physical and genetic distances. Additionally, we analyzed SNP data from 199 *A. thaliana* accessions genotyped at approximately 220,000 SNP sites [20]. The alleles were polarized using the genome of the related species *Arabidopsis lyrata*

[26]. For conversion from physical to genetic distances, we fitted a polynomial curve to 253 markers, for which physical and genetic positions are known [12]. All gene annotations were obtained from TAIR version 8 [27].

Results

Comparison of sweep statistics

We restricted the comparison of our statistic to its closest relatives, the iHS and the PHS statistics. To our knowledge, the PHS is the only test with a correction for relatedness. As a basic model, we simulated a panmictic population. First, we checked the ranking of the selected sites based on their absolute scores. The mean rank of the causal SNP out of all 4000 SNPs was 195.9 (±17.6), 193.3 (±17.4) and 455.68 (±27.4) for iHS, WiHS and PHS, respectively. The difference between iHS and WiHS was not significant (pairwise Wilcoxon-test; p = 0.85). The relatively poor ranks show that a single SNP's score may be a bad identifier for a selective sweep. Therefore, we use the averaged absolute scores in a window of ±25 SNPs around the selected site as test statistic. This is similar to the approach chosen by Voight et al. [1]. It takes the hitchhiking variation into account, which is a important advantage if the causal site is not genotyped [1]. The power to detect a sweep using either of the three tests is highly variable across different allele frequencies (Figure 2). None of them is able to distinguish low frequency sweeps from neutral variation. The change in power is similar across different allele frequencies: both weighted and unweighted integrated scores show nearly identical graphs with a maximum power at an allele frequency between 60% and 80%, whereas the PHS test generally has a lower power for all frequencies and achieves its maximum between 80% and 90%. While the maximum power clearly differs at a significance level of α = 0.01 with WiHS having the highest and PHS having the lowest power, it is nearly identical for all three tests at α = 0.05 (data not shown), which is consistent with previous findings that the iHS has a high specificity [28].

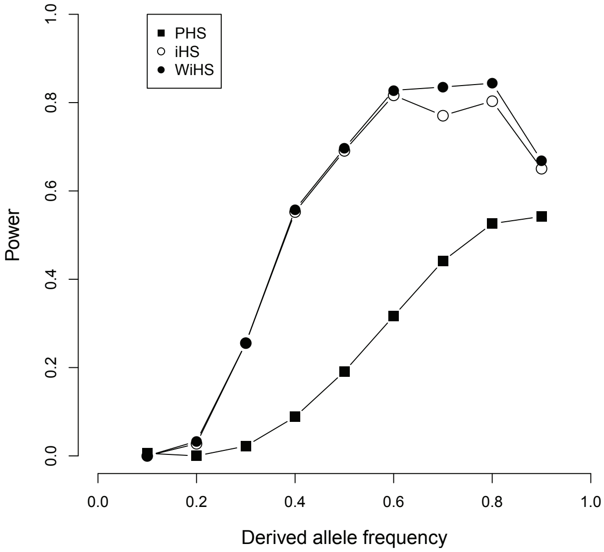

Figure 2 Power to detect selective sweeps. Detection power at a significance level of α = 001 based on the average score of all SNPs in a ±25 SNPs window around the selected site. For each allele frequency, 200 data sets were simulated and analyzed with all three tests. The null distribution was estimated in 1000 simulations without a selected site.

A comparison of the iHS and WiHS tests shows that WiHS performs better than iHS for allele frequencies > 40% even in panmictic populations. As the power itself is based on a single stringent threshold for the test score based on a significance level, we compared the normalized test scores between iHS and WiHS directly and found that WiHS assigns higher absolute scores to the SNPs surrounding the selected site (pairwise Wilcoxon-test, $p < 10^{-15}$). The scores around neutral sites are essentially identical for both tests (Additional File 1 Figure s1), which is expected for normalized scores. Thus, this difference demonstrates a better performance of WiHS in the detection of selective sweeps. While the absolute power decreased for selection weaker than $2N_es$ = 200 (Figure 4), a difference between iHS and WiHS was still observed and significant (pairwise Wilcoxon-test; $p < 10^{-6}$, $p < 10^{-10}$ and $p < 10^{-15}$ for $2N_es$ = 50, $2N_es$ = 100 and $2N_es$ = 150, respectively). This difference is a consequence of the sampling process, because it is impossible to sample genetically equidistant individuals and therefore even random samples of a panmictic population exhibit a certain degree of structure. The weighting corrects for this bias and improves the power of selection tests.

As the number of markers and individuals commonly used in sweep detection is rapidly growing, the running time of algorithms is becoming a limiting factor. We compared running times of all tests on simulated data sets and stepwise increased the number of analyzed chromosomes. Both integrated scores scale linearly, whereas PHS scales quadratically with the number of chromosomes (Figure 3). Since the PHS test is based on pairwise comparisons between individuals for each site, it is inefficient in running time and memory usage (not shown) for large data sets, while iHS and WiHS still have reasonable running times for data sets with thousands of individuals genotyped at hundreds of thousands of sites. However, sample sizes around 100 seem to be adequate for a reasonable power under panmictic scenarios, at least for the detection of strong selection (Additional File 1 Figure s2).

Performance under different demographic scenarios

The recent inclusion of selection in coalescent simulation software [19] permitted us to test WiHS under

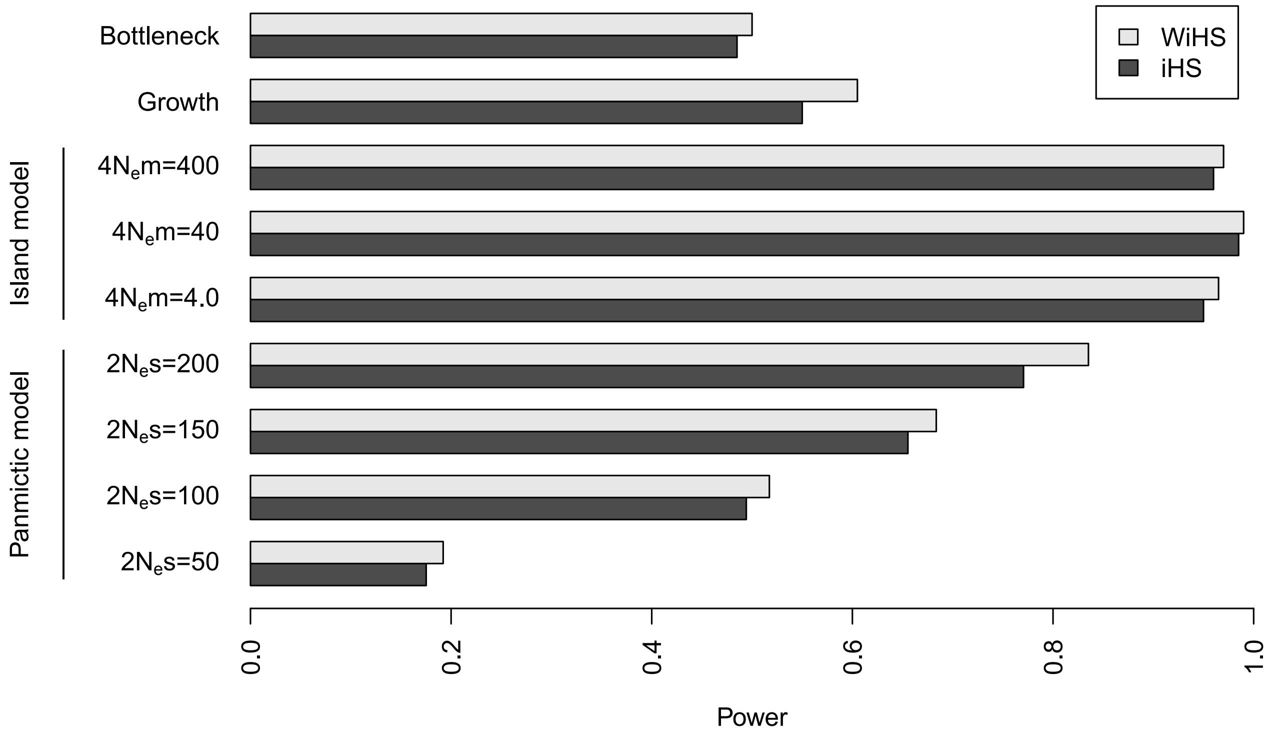

Figure 4 Detection power under different demographic scenarios and parameters. The power was estimated as the number of detected sweeps divided by the number of simulated sweeps. For each scenario, 200 simulations of sweeps with a present derived allele frequency between 60% and 80% were conducted.

different demographic scenarios. First, an island model of two equally sized populations with varying migration rates was simulated (Figure 1A). Three different migration rates of $4N_em \in \{4, 40, 400\}$ corresponding to a population differentiation of F_{ST} between 0.0025 and 0.2 were simulated [29]. Higher levels of differentiation between populations are also possible, but in these cases a cross-population test (e.g. [8,9]) is more practical. The results suggest a marginally higher power of WiHS for all three migration rates (Figure 4) with significantly higher scores around the selected site (pairwise Wilcoxon-test; $p < 10^{-11}$, $p < 10^{-15}$ and $p < 10^{-15}$ for $4N_em = 4.0$, $4N_es = 40$ and $4N_em = 400$, respectively). The absolute power for all three migration rates is higher than observed under panmixia (Figure 4). Since there is no reason to expect such pattern, this may hint at an artifact in the simulations.

Additionally, a model of exponential population growth followed by a constant population size was simulated (Figure 1B). The model by [23] resembles the population history of European *A. thaliana* accessions. Therefore, we regard these simulations as a test case for the analysis of empirical data from *A. thaliana*. Compared to the panmictic model, the detection power was decreased by more than 20% (Figure 4). Nevertheless, WiHS had a power 5.5% higher than the power of iHS and the scores around the selected site were significantly higher for WiHS (pairwise Wilcoxon-test; $p < 10^{-9}$). For the bottleneck model, a previously panmictic population was reduced to one fifth of its size with a later recovery to the original population size (Figure 1C). The bottleneck led to the strongest decrease in detection power (Figure 4), but WiHS still performed better and scored the SNPs in the sweep region higher (pairwise Wilcoxon-test; $p < 10^{-8}$). For models with a non-constant population size, which is the case in the growth and bottleneck model, *msms* requires a defined start time of the selective sweep. The sweeps were initiated directly before the bottleneck or the start of population growth for the simulations above. Simulating different starts for the sweep showed no trend in the relation between time and detection power in both scenarios (Additional File 1 Figures s3, s4).

Biases introduced by demography are supposed to affect both the detection power and the number of false positives. To check for such biases, we used neutral simulations of all demographic models and calculated the false discovery rate (FDR) if the cutoffs were estimated from a panmictic model. The FDRs differ only marginally between iHS and WiHS (Figure 5). In general, the FDR at a nominal significance level of 0.01 is only slightly elevated, ranging form 0.010 to 0.018 for the island and bottleneck model, respectively (Figure 5).

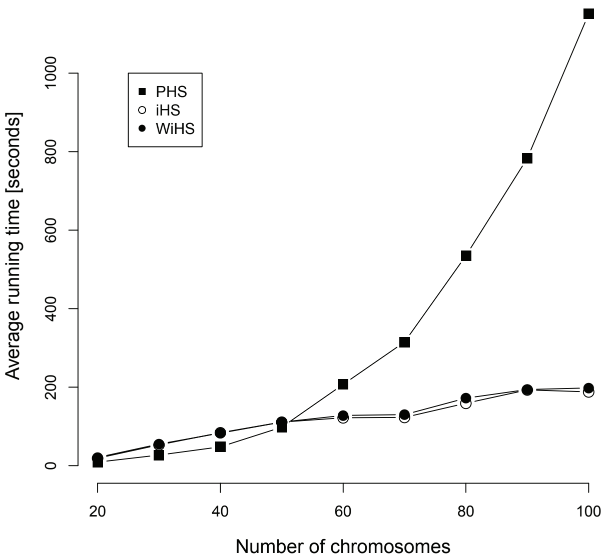

Figure 3 Running time. Average running times for the analysis of ten simulated data sets.

identified as sweep regions and discussed in iHS studies of the particular populations [1,8,31]. Some of the candidate regions were regarded as sweep candidates in more than one population, therefore we investigated 34 regions in total (see Table 2). Scores were computed for all SNPs and then the average absolute score was estimated in a sliding window approach (50 SNPs window size, 20 SNPs offset between windows). Twenty out of the 34 regions were ranked among the genome-wide 5% highest scoring windows by both tests. We expected no complete overlap, since these candidates were identified in different data sets using different methods [30]. The WiHS test ranked 17 candidates better than the iHS while the latter test ranks eleven regions higher. The remaining six candidates were ranked identical. Summing up for all candidate regions, the ranking by WiHS was marginally improved in comparison to iHS (pairwise one-sided Wilcoxon-test, $p < 0.1$). The top 100 ranked windows differ only slightly between both tests (data not shown).

Selective sweeps in the HapMap data

To apply our new test statistic to empirical data, we reanalyzed 690,566, 748,881 and 709,542 HapMap2 SNPs [24] from the JPT+CHB, YRI and CEU populations, respectively. As numerous maps for positive selection in humans have been published earlier (reviewed by [30]), we were mainly interested in differences between the iHS and the WiHS tests instead of presenting an additional map of sweeps. We selected the 27 most relevant candidate genes or gene clusters that were previously

Selective sweeps in the A. thaliana data

As the one of the highest power differences was observed for the growth model, the simulations indicate that WiHS offers an increased power for the analysis of data from *A. thaliana*. As a showcase for a genome-wide scan for selection in *A. thaliana*, we analyzed a genome-wide SNP data set of 220,000 SNPs from 199 accessions [20]. After WiHS was calculated for all SNPs, the genome was divided into non-overlapping windows of 50 SNPs and the absolute scores in these windows were averaged (Figure 6). A co-occurrence analysis of

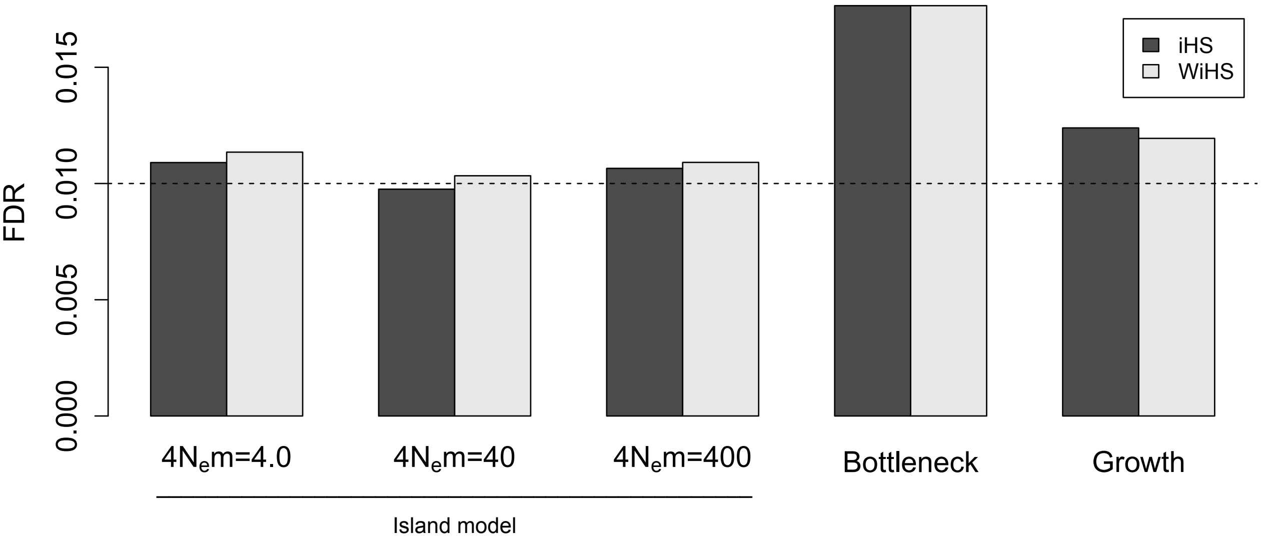

Figure 5 False discovery rate under different demographic scenarios and parameters. The FDR was estimated as the number of detected sweeps divided by the number of analyzed windows in simulations without selection. The dashed line represent the nominal significance level of 0.01. For each scenario, 200 simulations were conducted.

Table 2 Ranking of previously reported candidate genes in Human HapMap2 data

Gene(s)	Population	p_{iHS}	p_{WiHS}
NCOA1, ADCY3	YRI	0.003275	0.003189
SNTG1	YRI	0.023862	0.024517
ITGB4BP, CEP2, SPAG4	YRI	0.002392	0.002392
SYT1	YRI	0.115920	0.115265
RSBN1	YRI	0.062787	0.063043
CPEB2	YRI	0.252285	0.250491
FZD6	YRI	0.032490	0.032546
CHST5, ADAT1, KARS	YRI	0.146274	0.145904
LARGE	YRI	0.000598	0.000598
NCDN, TEKT2	CEU	0.001258	0.001228
LCT	CEU	0.000491	0.000552
SNTG1	CEU	0.001136	0.001136
ITGB4BP, CEP2, SPAG4	CEU	0.000583	0.000491
CYP3A5	CEU	0.597489	0.594758
SLC24A5	CEU	0.718933	0.714145
OCA2	CEU	0.091950	0.092533
TYRP1	CEU	0.008962	0.008900
ERBB4	CEU	0.117822	0.115827
NRG3	CEU	0.072522	0.072308
ODF2	CEU	0.874075	0.869595
ACVR1	CEU	0.067336	0.065095
PDE11A	CEU	0.026640	0.027438
SNTG1	JPT+CHB	0.006990	0.007055
ITGB4BP, CEP2, SPAG4	JPT+CHB	0.000755	0.000755
CHST5, ADAT1, KARS	JPT+CHB	0.095658	0.095790
PDE11A	JPT+CHB	0.081318	0.084665
ERBB4	JPT+CHB	0.020674	0.020641
BLZF1, SLC19A2	JPT+CHB	0.007613	0.007416
SLC30A9	JPT+CHB	0.012404	0.012929
PCDH15	JPT+CHB	0.001477	0.001477
SLC44A5	JPT+CHB	0.002658	0.002789
SULT1C	JPT+CHB	0.000525	0.000525
ADH cluster	JPT+CHB	0.023037	0.022938
FLJ32745, EDAR	JPT+CHB	0.765563	0.756571

The data was analyzed with both tests and the value represents the proportion of windows with a higher percentage of high scoring SNPs than a window overlapping with the candidate region.

molecular function and biological process GO terms among the top 100 windows using GeneCoDis [32] revealed several over-represented categories (Additional File 1 Table s1). They include some categories that are of particular interest when looking for selection candidates. These categories comprise response to external and internal stimulus (e.g. auxin, light, salt stress) and flower development. The highest ranked term is the molecular function 'chitinase activity', some chitinases have been associated with pathogen response in *A. thaliana* [33].

In addition, a more detailed look was taken at the genes among the top 6 windows (Figure 6). The top ranked window overlaps with a region on chromosome 3 that was previously suggested as a sweep candidate [34]. This window includes *ARR5*, a gene involved in the cytokin signaling pathway, whose mutant shows a reduced rosette size and an increased sensitivity to red light. The windows ranked second and third contain *FKF1*, an F-box protein which is involved in the regulation of flower development and response to blue light, and *ANN5*, which is contributing in the response to heat, cold, salt stress, red light and water deprivation. Finally, the fourth and sixth ranked regions on chromosome 4 comprise *LUG1*, a regulator of *AGAMOUS* involved in the flower development.

Discussion

Detection of selective sweeps

Even in unstructured populations, sampling and relatedness introduce a bias into the sample. We improved the accuracy of detecting selective sweeps with haplotype based methods by weighting the contribution of each individual to the statistic according to its uniqueness in the sample. The improvement was observed in all simulated demographic scenarios including a panmictic population, a model of two subpopulations, exponential population growth and a population bottleneck. The increase of detection power of WiHS compared to iHS was less than expected but significant, reaching a maximum of 6.5%, 1%, 5.5% and 1.5% in the panmictic, island, population growth and bottleneck models, respectively. Simulation of different models and different model parameters, such as more severe bottlenecks, may give different results than the simulations in this study. The highest improvement was achieved in panmictic and growing populations. As the latter scenario was previously fitted to European accessions of *A. thaliana* [23], our improvement can result in additional sweep candidates for this species. While the detection power decreased in the more complex models, there was no significant increase of FDR if the sample was incorrectly assumed to arise from a panmictic population. As iHS and WiHS are genome-wide normalized scores, an excess of extreme scores and false positives under different demographic models is avoided.

The presented approach corrects for genome-wide IBD by upweighting more unique individuals in the sample. Since selective sweeps generate locally elevated IBD, which was suggested as a test for selection [35], one could also think of an opposite weighting based on local IBD. Local weighting would require the calculation of an IBD matrix for every single region, causing numerous pairwise comparisons between individuals and inflating the running time, which is beyond the scope of this paper.

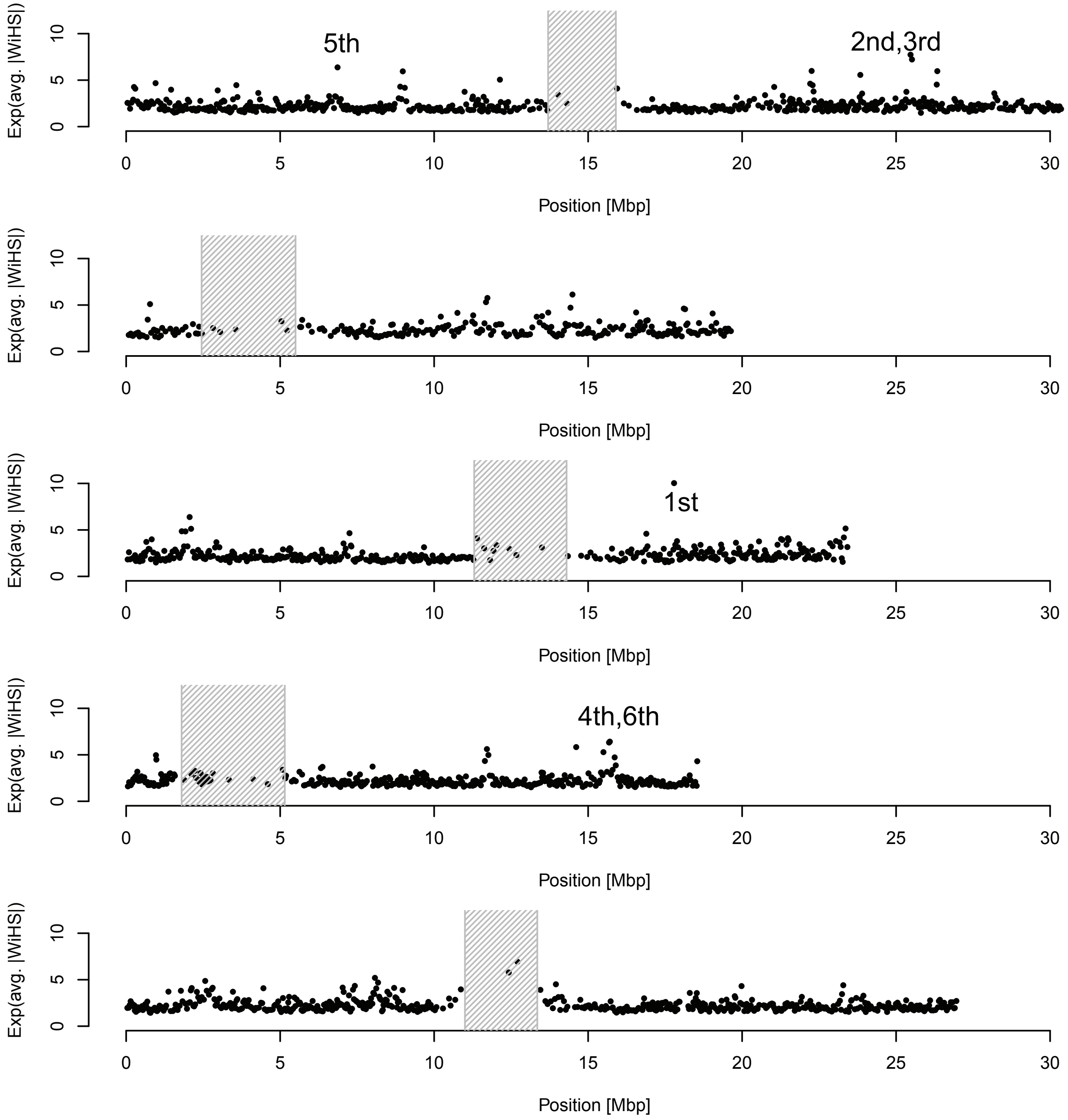

Figure 6 WiHS results for the A. thaliana data. WiHS was calculated for all SNPs and then the absolute scores were averaged in windows of 50 SNPs. To highlight the outstanding windows, the values were then exponentiated for this figure. The top six windows are labeled in the figure. The shaded regions denote the centromeric regions.

Our simulation results extend the findings from previous studies for other test statistics [13,14,36-38] and show that haplotype-based tests are sensitive to demographic scenarios such as population structure and exponential growth. To identify candidates for selective sweeps, the search for outlier regions is commonly used, although they may represent the outliers of a neutral distribution [30]. Therefore, additional validation using tests based on other characters than haplotype length, such as site frequency spectrum [28,39-44] or population differentiation [44-46], will increase the reliability of sweep detections. Recently, compositions of different statistics have been shown to perform better in the detection of causal variants than each statistic separately

[47-50] and the WiHS statistic might be included in such composite approaches as well and lead to a further improvement of these methods.

Recent selection in empirical data sets

The analysis of empirical data sets provides an insight into the effect of the modification under real conditions. Among the top scoring windows of the HapMap data, some prominent candidate regions were found, such as *LCT* for lactose metabolism, *TYRP1* for skin pigmentation and *SPAG4* for sperm motility. Most but not all of these genes ranked better by WiHS, so we found only a weak significance. We are aware of the fact that some of these genes represent only candidates for positive selection that have not been validated. The trend suggests that general long-haplotype pattern in these regions is better detected by the WiHS and it is still possible that the ranking generated by WiHS is more accurate in the identification of selective sweeps.

The *A. thaliana* results revealed some promising candidates for selective sweeps. As the windows are still quite big, looking for particular candidate genes in these regions remains some kind of fishing in murky waters. Therefore, we leave the identification of sweep candidates to further studies, which employ a combination of different tests and use a more precise estimation of the genetic map. However, the simulations and the detection of some interesting genes in our preliminary scan suggest that WiHS is useful for the detection of selective sweeps in *A. thaliana*.

Conclusions

Next-generation sequencing projects will provide sufficiently large data sets for the genome-wide detection of natural selection in many species (e.g. 1000genomes.org, 1001genomes.org, The *Drosophila* Genetic Resource Panel). The upcoming flood of data demands for time efficient and accurate analysis methods. Several methods operate with an equal contribution of individuals, which means that all individuals in the sample are assumed to be statistically independent. As it is very likely that not all pairs of individuals share the same most recent common ancestor, the assumption of independence should be violated in most biological samples. Thereby, unequally related individuals introduce a minor but significant bias into analyses, because the contribution of closely related individuals is overestimated while the contribution of others is underestimated. Such bias may be increased by demographic history and population structure. Genome-wide marker data allow to assess the relationship between individuals. This information can be used to cope with the dependency and to reduce the bias in estimates by differentially weighting the contribution of each individual. This concept could be extended to other unweighted statistics in population genetics. The consistent improvement across all simulated scenarios shows the general positive effect of differential weighting. Nevertheless, the slight increase of power leaves room for further improvement in the calculation of weights for each individual and the incorporation of these weights in test statistics, and for the detection of selective sweeps in general.

Additional material

Additional file 1: Supplementary Information. The Supplementary Informations include additional figures and tables.

Acknowledgements
We thank the High Performance Computing Center Stuttgart (HLRS) for assistance with the bwGRID, and Sariel Hübner and Inka Gawenda as well as two anonymous reviewers for discussion and comments on the manuscript. This work was funded by a Volkswagen Foundation Evolutionary Biology scholarship to TG (I/84 225).

Authors' contributions
TG conceived the initial approach. TG and KS designed the experiments. TG conducted the experiments. Both authors interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 11 February 2011 Accepted: 5 July 2011
Published: 5 July 2011

References

1. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**:e72.
2. Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, Zhao K, Calabrese P, Dean C, Nordborg M: **A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome.** *PLoS Biol* 2006, **4**:e137.
3. Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ: **Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster.** *Genetics* 1994, **136**:1329-1340.
4. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: **Detecting recent positive selection in the human genome from haplotype structure.** *Nature* 2002, **419**:832-837.
5. Wang ET, Kodama G, Baldi P, Moyzis RK: **Global landscape of recent inferred Darwinian selection for *Homo sapiens*.** *Proc Natl Acad Sci USA* 2006, **103**:135-140.
6. Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, Kennedy GC: **A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations.** *Bioinformatics* 2006, **22**:2122-2128.
7. Kimura R, Fujimoto A, Tokunaga K, Ohashi J: **A practical genome scan for population-specific strong selective sweeps that have reached flxation.** *PLoS ONE* 2007, **2**:e286.
8. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W,

Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JTF, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Sham PC, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449**:913-918.

9. Tang K, Thornton KR, Stoneking M: **A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome.** *PLoS Biol* 2007, **5**:e171.
10. Wakeley J, Aliacar N: **Gene genealogies in a metapopulation.** *Genetics* 2001, **159**:893-905.
11. Przeworski M: **The signature of positive selection at randomly chosen loci.** *Genetics* 2002, **160**:1179-1189.
12. Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T: **A multilocus sequence survey in *Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism*.** *Genetics* 2005, **169(3)**:1601-15.
13. Teshima KM, Coop G, Przeworski M: **How reliable are empirical genomic scans for selective sweeps?** *Genome Res* 2006, **16**:702-712.
14. Zeng K, Mano S, Shi S, Wu CI: **Comparisons of site- and haplotype-frequency methods for detecting positive selection.** *Molecular Biology and Evolution* 2007, **24(7)**:1562-74.
15. Price AL, Zaitlen NA, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies.** *Nature reviews. Genetics* 2010, **11(7)**:459-463.
16. Thompson JD, Higgins DG, Gibson TJ: **Improved sensitivity of profile searches through the use of sequence weights and gap excision.** *Comput Appl Biosci* 1994, **10**:19-29.
17. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
18. Kang HM, Zaitlen Na, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: **Efficient control of population structure in model organism association mapping.** *Genetics* 2008, **178(3)**:1709-23.
19. Ewing G, Hermisson J: **MSMS: A Coalescent Simulation Program Including Recombination, Demographic Structure, and Selection at a Single Locus.** *Bioinformatics* 2010, **26(16)**:2064-2065.
20. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M: **Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana inbred lines*.** *Nature* 2010, **465(7298)**:627-31.
21. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbeisch T, Schulz V, Kreitman M, Bergelson J: **The Pattern of Polymorphism in *Arabidopsis thaliana*.** *PLoS Biology* 2005, **3**:e196.
22. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M: **Recombination and linkage disequilibrium in *Arabidopsis thaliana*.** *Nature genetics* 2007, **39(9)**:1151-5.
23. François O, Blum MGB, Jakobsson M, Rosenberg NA: **Demographic history of european populations of *Arabidopsis thaliana*.** *PLoS genetics* 2008, **4(5)**: e1000075.
24. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JTF, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851-861.
25. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Research* 2001, **29**:308-311.
26. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett Ja, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov Aa, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL: ***The Arabidopsis lyrata genome sequence and the basis of rapid genome size change*.** *Nature genetics* 2011, **43(5)**.
27. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation.** *Nucleic acids research* 2008, , 36 **Database:** D1009-14.
28. Hussin J, Nadeau P, Lefebvre JF, Labuda D: **Haplotype allelic classes for detecting ongoing positive selection.** *BMC Bioinformatics* 2010, **11**:65.
29. Wright S: **Isolation by distance.** *Genetics* 1943.
30. Akey JM: **Constructing genomic maps of positive selection in humans: where do we go from here?** *Genome research* 2009, **19(5)**:711-22.
31. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK: **Signals of**

recent positive selection in a worldwide sample of human populations. *Genome Res* 2009, **19**:826-837.
32. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: **GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists.** *Genome biology* 2007, **8**:R3.
33. Verburg JG, Huynh QK: **Purification and Characterization of an Antifungal Chitinase from *Arabidopsis thaliana*.** *Plant Physiology* 1991, **95(2)**:450-5.
34. Childs LH, Witucka-Wall H, Günther T, Sulpice R, V Korff M, Stitt M, Walther D, Schmid KJ, Altmann T: **Single feature polymorphism (SFP)-based selective sweep identification and association mapping of growth-related metabolic traits in *Arabidopsis thaliana*.** *BMC Genomics* 2010, **11**:188.
35. Albrechtsen A, Moltke I, Nielsen R: **Natural selection and the distribution of identity-by-descent in the human genome.** *Genetics* 2010, **186**:295-308.
36. Santiago E, Caballero A: **Variation after a selective sweep in a subdivided population.** *Genetics* 2005, **169**:475-483.
37. Chevin LM, Billiard S, Hospital F: **Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation.** *Genetics* 2008, **180**:301-316.
38. Huff CD, Harpending HC, Rogers AR: **Detecting positive selection from genome scans of linkage disequilibrium.** *BMC Genomics* 2010, **11**:8.
39. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123(3)**:585-95.
40. Kim Y, Stephan W: **Detecting a local signature of genetic hitchhiking along a recombining chromosome.** *Genetics* 2002, **160(2)**:765-77.
41. Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD: **Distinguishing between selective sweeps and demography using DNA polymorphism data.** *Genetics* 2005, **170(3)**:1401-10.
42. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data.** *Genome Research* 2005, **15(11)**:1566-75.
43. Zhu L, Bustamante CD: **A composite-likelihood approach for detecting directional selection from DNA sequence data.** *Genetics* 2005, **170(3)**:1411-21.
44. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps.** *Genome research* 2010, **20(3)**:393-402.
45. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG: **Measures of human population structure show heterogeneity among genomic regions.** *Genome Research* 2005, **15**:1468-1476.
46. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW: **Tracking footprints of artificial selection in the dog genome.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107(3)**:1160-5.
47. Zeng K, Shi S, Wu CI: **Compound tests for the detection of hitchhiking under positive selection.** *Molecular Biology and Evolution* 2007, **24(8)**:1898-908.
48. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, Lander ES, Schaffner SF, Sabeti PC: **A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection.** *Science* 2010, **166(February)**:2008-2011.
49. Pavlidis P, Jensen JD, Stephan W: **Searching for Footprints of Positive Selection in Whole-genome SNP Data from Non-equilibrium Populations.** *Genetics* 2010.
50. Lin K, Li H, Schlötterer C, Futschik A: **Distinguishing Positive Selection from Neutral Evolution: Boosting the Performance of Summary Statistics.** *Genetics* 2010, 1-39.

doi:10.1186/1756-0500-4-232
Cite this article as: Günther and Schmid: Improved haplotype-based detection of ongoing selective sweeps towards an application in Arabidopsis thaliana. *BMC Research Notes* 2011 **4**:232.