A Regression-Based Approach to Selection Mapping

PAMELA WIENER AND RICARDO PONG-WONG

From the Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK.

Address correspondence to Dr Pamela Wiener at the address above, or e-mail: pam.wiener@roslin.ed.ac.uk.

Abstract

Selection mapping applies the population genetics theory of hitchhiking to the localization of genomic regions containing genes under selection. This approach predicts that neutral loci linked to genes under positive selection will have reduced diversity due to their shared history with a selected locus, and thus, genome scans of diversity levels can be used to identify regions containing selected loci. Most previous approaches to this problem ignore the spatial genomic pattern of diversity expected under selection. The regression-based approach advocated in this paper takes into account the expected pattern of decreasing genetic diversity with increased proximity to a selected locus. Simulated data are used to examine the patterns of diversity under different scenarios, in order to assess the power of a regression-based approach to the identification of regions under selection. Application of this method to both simulated and empirical data demonstrates its potential to detect selection. In contrast to some other methods, the regression approach described in this paper can be applied to any marker type. Results also suggest that this approach may give more precise estimates of the location of the selected locus than alternative methods, although the power is slightly lower in some cases.

Key words: asymptotic regression, hitchhiking, myostatin gene, selection mapping

The theoretical basis of the “selection mapping” (Kohn et al. 2000) or “hitchhiking mapping” (Harr et al. 2002) approach for detection of genes under selection dates from the 1970s, with the finding that genetic hitchhiking of neutral markers will accompany selection on a linked gene (Maynard Smith and Haigh 1974). As a result, positive selection on a gene will result in reduced variability not only at that gene but also at linked loci. Since then, there have been a number of studies refining the theoretical understanding of this effect (Ohta and Kimura 1975; Kaplan et al. 1989; Stephan et al. 1992; Slatkin 1995; Barton 1998, 2000; Wiehe 1998; Kim and Stephan 2000; Durrett and Schweinsberg 2004; Innan and Kim 2004, 2008; Hermisson and Pennings 2005) and exploiting it for gene mapping (Fay and Wu 2000; Harr et al. 2002; Kim and Stephan 2002; Sabeti et al. 2002; Schlötzer 2002; Li and Stephan 2005; Nielsen et al. 2005; Pollinger et al. 2005; Pennings and Hermisson 2006; Voight et al. 2006; Tang et al. 2007; Wiehe et al. 2007; Grossman et al. 2010; Oleksyk et al. 2010).

The objective of this study is to develop a new approach for detection of genes under strong selection that specifically exploits spatial patterns of diversity within the genome generated by hitchhiking. That is, in addition to a general reduction in the region of a selected locus, theory predicts that diversity will decrease with increasing proximity to this locus. Two sets of simulations are presented: the first is used to evaluate the fit of the predicted heterozygosity relationship to the patterns observed under strong artificial selection such as that imposed on a managed livestock species and the second is used to compare the performance of this approach to a previously published method for identification of selection and determination of the genomic position under selection. The novel method is also applied to 2 data sets from genomic regions known to have been under selection. A protocol for applying the method to genome-wide data is also discussed.

Regression Approach to Selection Mapping

Population genetics theory predicts that under hitchhiking, an allele at a marker linked to a selected locus has a probability of escaping loss during a selective sweep (by recombination onto the selected background) proportional to $1 - e^{\alpha d}$, where $d$ is the distance between the marker and the selected locus and $\alpha$ is a function of population size, strength of selection, and recombination rate (Nielsen et al. 2005). This probability is also proportional to the ratio of post-sweep heterozygosity to pre-sweep heterozygosity at that marker (Durrett and Schweinsberg 2004) and thus, if one assumes that heterozygosity was at similar levels across the genome prior to the selective sweep, we expect to see an asymptotic increase in marker variation (heterozygosity)
with increasing distance from a locus under directional selection. This theoretical relationship was illustrated by Maynard Smith and Haigh (1974) (their Figure 1) and Kim and Stephan (2002) (their Figure 2) and also has been demonstrated with microsatellite data near a gene under selection in cattle (Wiener et al. 2003).

The methodology developed in this paper involves fitting a regression to heterozygosity data as a function of genomic position. A signature of selection is identified if the trend of the relationship between heterozygosity and genomic position follows that predicted by hitchhiking theory. Based on the theoretical predictions mentioned above, heterozygosity levels should increase exponentially with distance from a selected locus due to hitchhiking effects. Therefore, a selected locus should be characterized by such an asymptotic heterozygosity pattern. However, depending on the marker density relative to the strength of selection, the relationship between heterozygosity and distance from the selected locus might appear linear rather than exponential (i.e., if the distances examined only contain points from the exponential phase of the asymptotic relationship such that the asymptote has not yet been reached). Therefore, either asymptotic or linear regression may be appropriate for analyzing such data (see schematic illustration in Figure 1). The theoretical expectation is that heterozygosity ($y$) will increase with respect to distance from a selected locus ($x$) and in the case of the asymptotic regression, be “right directional,” such that the heterozygosity increases exponentially directly away from the selected locus and then plateaus further away (as in Figure 1).

For linear regression

$$ y = A + Bx, $$

where $B$ is the slope of the heterozygosity relationship.

For asymptotic regression

$$ y = A + BR^x \text{(equivalent to } y = A + B \times e^{-kx}, \text{ for } k = -\log R), $$

where the $R$ parameter measures the rate of exponential increase of the curve, $A$ is the value of the asymptote (an estimate of the background level of heterozygosity in the genome) and $B$ is the difference between the level of heterozygosity at the selected locus ($x = 0$) and the background (asymptotic) level.

We will subsequently use the terms “predicted” relationship, direction, or shape to refer to heterozygosity relationships that are indicative of positive selection. In the case of linear regression, this refers to a positive slope ($B$) and in the case of asymptotic regression, this refers to a right directional and increasing asymptotic relationship (requiring both $0 < R < 1$ and $B < 0$).

The statistical significance of the regressions is based on a test in which the predicted relationship is also satisfied. Regressions that are not in the predicted direction (i.e., any parameters different from those described above) are treated as nonsignificant regardless of their $P$ values as these relationships do not follow the pattern consistent with directional selection. To account for this additional restriction, the following protocol is used to determine significance at the $P < 0.05$ level: for the linear relationship, the threshold for a one-sided test is used to account for the fact that the slope must be positive (equivalent to $P < 0.10$ for a standard two-sided test) and for the asymptotic regression, where there are 4 possible parameter sets ($R < 1, B < 0; R < 1, B > 0; R > 1, B < 0; R > 1, B > 0$), a threshold is used to account for the fact that the curve must be right directional and increasing (equivalent to $P < 0.20$ for a standard multiscided test). This adjustment was shown to give correct Type I error rates under the null hypothesis (see Results for No-Selection Scenarios, below).

**Application to Simulated Test Data**

The purpose of this section was to assess whether the predicted heterozygosity relationships were observed using data simulated under a variety of scenarios including and excluding selection.

**Data Generation**

Monte Carlo simulations were conducted to follow a population of 50 000 diploid individuals (5000 males and 45 000 females). In each generation, male and female parents were chosen with replacement at random from the population and a single offspring was produced for that mating pair. An offspring was chosen to enter the breeding pool of the next generation based on its phenotype, under a truncation selection model, until the population was filled with the appropriate number of males and females. An individual’s phenotype was determined by a single major gene with an additive effect of one unit and an environmental noise term, which took values from a Normal distribution (mean = zero units, standard deviation = one unit). The truncation point was set at a fixed level 0.3 standard deviation units above the initial mean of the population, resulting in a selection intensity of 1.00 (Falconnor and Mackay 1996). Because the truncation point was set at a fixed point, the numbers of parents producing offspring varied across generations although the total number of offspring produced did not change.
The simulations included a fixed number of neutral loci, all with the same number of alleles, linked to the major gene under selection. The number of alleles (n) at these loci was set at a value between 2 and 20 (2, 4, 6, 8, 10, and 20). Distributions of initial allele frequencies are described below. Two genetic maps were considered: a map of 5 linked loci ("sparse" map) and one of 20 loci ("dense" map). For both maps, the loci were equally spaced, within a 10 cM distance of the selected locus. The favored allele at the selected locus was introduced as a single copy in the first generation. At the point at which an offspring was produced, recombination and mutation (at a rate of 0.0001 per site per offspring) were applied. The probability of recombination between loci with map distance m (measured in Morgans) was determined by Haldane’s mapping function, c = [1 - exp(-2m)]/2 (Lynch and Walsh 1998).

For each parameter set (see below), 100 independent simulations were run for 60 generations each, such that the favored allele at the selected locus was fixed or nearly fixed by the last generation (if the favored allele was lost, that simulation was ended and restarted). In order to assess Type 1 error rates, 100 simulations were also run for 60 generations without selection under the same parameter sets. As a special case, 100 simulations incorporating both a population bottleneck and no selection were also carried out in order to test whether demographic changes could produce patterns that would be detected as selection by this regression approach. The bottleneck scenario was implemented by imposing a 75% reduction in population size at generation 46 that lasted for 5 generations. This was followed by a constant increase in each generation until the population had returned to its original size (at generation 60).

After generation 60, asymptotic and linear regression were carried out on the overall population heterozygosity levels as a function of genomic distance from the selected locus (GenStat 2007). The slope and F ratio (linear regression) and the R^2, A, and B parameters and F ratio (asymptotic regression) for predicted relationships were recorded. The proportion of the variance explained by each regression (R^2) was calculated, and from the F ratios, P values for the regression relationships were also calculated. In addition, the statistical power (the proportion of regressions that had the predicted shape and were significant based on the adjusted P < 0.05 threshold described above) was recorded.

For the scenarios without selection, the simulation results were assessed to evaluate the rates of Type 1 error (the proportion of significant results) involved in using the fit of the heterozygosity relationship (heterozygosity regressed onto distance from a fixed locus) as a test statistic. Type I error was calculated for both linear and asymptotic regression under the 2 map densities (sparse and dense) and assuming different numbers of alleles per locus, as described above.

For the scenarios incorporating selection, various comparisons were made. The performance of linear and asymptotic regression (heterozygosity regressed onto distance from the selected locus) were assessed by comparing the proportion of variance explained (R^2 values) and the statistical power (the proportion of significant regressions). Comparisons were also made to assess the effects of the following factors considered in the simulations: neutral marker density, allelic diversity at neutral markers, and initial allele distribution at neutral markers. The effect of marker density was assessed by comparing the sparse (5 markers per 10 cM) and dense (20 markers per 10 cM) maps. The effect of allelic diversity was assessed by comparing the different numbers of alleles per locus (2–20). The effect of initial allele frequency distribution was examined by comparing the basic selection scenario (equal allele frequencies across loci) to one in which initial allele frequencies were allowed to vary according to an n-order Dirichlet distribution with mean 1/n for each allele and shape parameters 1 (broad distribution) or 10 (spiky distribution), constructed by renormalizing n gamma variates, each with shape parameter 1 or 10, respectively. The coefficients of variation in initial allele frequencies ranged from 0.58 (2 alleles) to 0.96 (20 alleles) for Dirichlet shape parameter 1 and from 0.22 (2 alleles) to 0.30 (20 alleles) for Dirichlet shape parameter 10.

Results for No-Selection Scenarios
For linear regression, Type I errors ranged between 0.02 and 0.08 (average = 0.047) (Table 1). For asymptotic regression, Type I errors ranged between 0.01 and 0.09 (average = 0.063) (Table 1). Thus, the adjusted thresholds described above gave Type I error rates close to the theoretical expectation of 0.05.

The Type I errors for the bottleneck situation ranged between 0.01 and 0.08 (average = 0.05) for linear regression and between 0.03 and 0.10 (average = 0.05) for asymptotic regression, again at the theoretical expectation (Table 1). Thus, a bottleneck did not produce heterozygosity patterns that would be detected as selection by the regression approach.

Results for Selection Scenarios
Under the basic selection scenario (equal initial allele frequencies across loci), nearly all linear regressions had positive slope and slightly fewer but still the vast majority of asymptotic regressions were right directional for all numbers of alleles (results not shown). More than 85% of the linear regression replicates were significant, and more than 88% of the asymptotic regression replicates were significant (Table 2). The proportion of variance explained (R^2) by significant linear regressions was 0.77–0.88 for the basic selection scenario and that explained by significant asymptotic regressions was 0.90–0.98 (Table 2). These values indicate that the method has substantial power to detect data generated under selection.

Linear Versus Asymptotic Regression
Comparisons of fitting linear and asymptotic regression to the simulated data revealed that the proportion of variance explained (R^2) for significant regressions was generally
greater for asymptotic than linear regression (by 10–17% for the basic scenario), indicating that for the simulated data, heterozygosity levels tended to increase asymptotically with distance from the selected locus. The only exception to this pattern was found for the case where there was variation in initial allele frequency, a dense map and low allelic diversity (2 alleles per locus) (Table 2). The advantage of asymptotic over linear regression was more pronounced for the dense map. Under the basic selection scenario, the statistical power was similar for linear and asymptotic regression, however, for the scenario with variation in initial allele frequency, power was higher for linear regression in some cases (sparse map, low allelic diversity).

Effects of Marker Density

Statistical power was greater or equal for the dense map (Table 2), applying either linear or asymptotic regression. However, the average proportion of variance explained by the significant regressions was greater or equal for the sparse map (Table 2).

Effects of Allelic Diversity

For the sparse map, there was a substantial increase in statistical power and the proportion of variation explained for both linear and asymptotic regression (Table 2) as allele number increased from 2 to 4, however, there was little change as allele number increased beyond 4 (and in some cases, power decreased slightly). For the dense map, the statistical power of both linear and asymptotic regressions was nearly 100% for allele number greater than 2 while the proportion of variance explained increased with numbers of alleles, again with the biggest difference between 2 and 4 alleles.

Effects of Initial Allele Frequency Distribution

Adding variation in initial allele frequencies (Dirichlet parameter = 1) to the simulation reduced statistical power and the proportion of variance explained by the regressions (Table 2); the greatest effects were seen for the fewest alleles per locus. However, the proportion of variance explained was still substantial, especially for high allelic diversity.

Results for the scenario with less extreme variation in initial allele frequencies (Dirichlet parameter = 10) were broadly similar to those of the basic selection scenario (results not shown). Statistical power was reduced for the sparse (but not dense) map for both linear and asymptotic regressions, but this was only pronounced for low allelic diversity (≤4 alleles/marker) and $r^2$ values were slightly lower across all scenarios.

Comparison with Extended Homozygosity Approach

The purpose of this section was to compare the proposed regression approach with a standard alternative method, using simulated data. One group of alternative approaches includes methods based on the decay of homozygosity surrounding a single nucleotide polymorphism (SNP) locus (Sabeti et al. 2002; Voight et al. 2006), which are conceptually the closest to the regression approach in that they also exploit the hitchhiking effect and are based on multiple markers. One of the standard extended homozygosity-based methods is that of Voight et al. (2006), in which the test statistic (iHS) for a given marker is a function of the relative decay of extended haplotype homozygosity.
surrounding the “ancestral” allele compared with that of the “derived” allele. The iHS values are standardized based on the genome-wide distribution of the test statistic for markers of similar allelic diversity (for further details, see Voight et al. 2006). In order to compare the regression method with iHS, simulated data was analyzed using both approaches. As iHS is restricted to use with SNP data, a 2-allele scenario was simulated to allow direct comparison. Other assumptions and manipulations to the data to allow analysis with iHS are described below.

### Scenarios Analyzed

Data were simulated as described above with a few modifications. The position of the selected locus was placed in the center of a 25-cM interval (at position 12.5), with neutral markers on either side, positioned every 0.25 cM. Each of the neutral markers carried 2 alleles, as in the low allele diversity case above. In the basic scenario for these simulations, the initial frequency of the first allele was chosen from a uniform distribution (0,1), comparable to the variable allele frequency scenario (as described above) for 2 alleles. In contrast to the scenarios described above, no mutation was allowed. Sampling involved random selection of 100 individuals from the population after 80 generations.

Several variants on the basic scenario were also explored: equal initial allele frequency (frequency of each allele = 0.5), recent selection (sampling at 20 generations), and weak selection. The latter scenario was run with lower selection intensity than the other scenarios (with truncation point 0.4 units below the phenotypic mean of the original population) and was sampled earlier (at 50 generations) to ensure that the favored allele had not yet fixed. Although this case was implemented in the same way as the other scenarios, it was not true truncation selection since the truncation point was less than the phenotypic means of all 3 genotype classes.

In order to allow iHS standardization, additional simulations were run under a neutral scheme. Appropriate sampling, at 20, 50, or 80 generations, allowed standardization of each of the selection scenarios (1000 simulations were run per standardization set).

### Regression Analysis

For each population sample (100 in total) of 100 individuals generated under the selection scenarios, heterozygosity was calculated for all neutral markers. Based on the above results comparing linear and asymptotic regression, asymptotic regression was then implemented in the following way. A regression was fitted to marker heterozygosity as a function of distance from a fixed (test) location (positioned every 0.25 cM) (i.e., $x = \text{abs}(\text{marker position} - \text{test position})$). The formulation of the regression equation was

$$\text{heterozygosity} = A + BRx,$$

as described above. The “best” position was identified as that for which the maximum proportion of the variance ($r^2$) was explained by the regression relationship, restricted to cases in which the response was in the predicted direction ($R < 1, B < 0$). As described for the previous simulations, significance was assessed using an adjusted $P < 0.05$ threshold.

### Regression Equation

**Table 2** Summary of analysis of simulation results (100 replicates per parameter set) applying regression of heterozygosity onto distance from the selected locus, showing the proportion of significant positive-slope linear regression and right-directional asymptotic regressions and the average proportion of variance explained ($r^2$) for the significant regressions

<table>
<thead>
<tr>
<th>No. of alleles</th>
<th>Basic selection</th>
<th>Basic selection + variable initial allele frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear regression</td>
<td>Asymptotic regression</td>
</tr>
<tr>
<td>Sparse map</td>
<td>Proportion of significant regressions</td>
<td>Mean $r^2$ (significant regressions)</td>
</tr>
<tr>
<td>2</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>0.94</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>0.97</td>
<td>0.87</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>10</td>
<td>0.98</td>
<td>0.88</td>
</tr>
<tr>
<td>20</td>
<td>0.97</td>
<td>0.87</td>
</tr>
<tr>
<td>Dense map</td>
<td>Proportion of significant regressions</td>
<td>Mean $r^2$ (significant regressions)</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>0.81</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Significance was determined using an adjusted $P < 0.05$ threshold, as described in the text.
iHS Analysis

Similar to the approach of Sikora et al. (2008), alleles were designated at each neutral locus as either “more common” (0) or “less common” (1), based on their allele frequency in the overall population, since in the simulations, there is no sense of “ancestral” or “derived” alleles as assumed for the iHS method (Voight et al. 2006). There is also no family structure in the population sample, in order to mimic random sampling from a population, and therefore haplotypes were inferred based on genotype data using fastPHASE (Scheet and Stephens 2006), using default options.

Raw iHS values were calculated using the phased genotype data for the neutral markers and the map information for both the selection (100 replicates) and neutral (1000 replicates) scenarios. In order to account for variation in allele frequency between markers, a standardization procedure was used, based on the iHS values calculated under the neutral scenario since genome-wide values (as used by Voight et al. 2006) were not available for the simulated data. After binning iHS values by allele frequency (into bins of 0.1 frequency units), standardized iHS values (Voight et al. 2006, equation (2)) were calculated for each selection scenario using means and standard deviations calculated from the appropriate neutral scenario. The position of the marker with maximum |standardized iHS| value (max |iHS|) was identified for each replicate. Values of max |iHS| > 1.96 were considered as indicative of selection as that is the 5% threshold for a folded standard normal distribution (the absolute value of iHS rather than raw iHS is the test statistic). Assessment of standardized iHS values generated by the no-selection scenario confirmed the normality of the distribution (results not shown).

Comparison of Regression and Extended Homozygosity Analyses

Basic Scenario

Under this scenario, the favored allele had always fixed by the time of sampling. Both the regression and iHS analyses identified signatures of selection for all replicates. For regression, the $P$ values calculated at the best position under the regression method were all less than 0.001 and the variance explained was at least 10%. For the iHS analysis, max |iHS| was greater than 6 in all cases ($P = 0.000$).

The position of the strongest selection signal differed, however, between the methods. The distance between the best position identified by the regression method and the true position of the selected locus varied between 0 and 1.50 cM (average distance = 0.28 cM), whereas the distance between the position of max |iHS| values and the true position varied between 1.75 and 7.50 (average = 4.23 cM) (Table 3), thus there were no max |iHS| values at the markers directly surrounding (i.e., within 1.5 cM of) the selected locus (Figure 2). This may be partly due to a deficit of iHS values calculated for markers surrounding the selected locus (over all simulations, iHS values could not be calculated for 32% of markers within 1.5 cM of the selected locus, presumably because the homozygous haplotypes associated with these markers extended to the edge of the range of markers analyzed), however, even for the runs in which iHS values were calculated for the 2 markers flanking the selected locus, these markers never had the max |iHS| values.

Effects of Equal Allele Frequency

When initial allele frequencies were set at 0.5 (equal frequency scenario), the favored allele fixed again in all cases. In addition to detecting selection in all cases, both regression and iHS generally performed better than they did for the basic scenario in that the variance explained was greater (regression) and max |iHS| values were higher (iHS). The distances between the best position and the true position of the selected locus for both methods were similar to that seen for the basic scenario (Table 3).

Effects of Recent Selection

When sampling occurred at 20 rather than 80 generations (recent selection scenario), such that the average frequency of the favored allele was 0.65 at sampling, regression was slightly less powerful than iHS: significant asymptotic regressions were detected in 96/100 cases while max |iHS| values exceeded the threshold in all cases. For those cases where there was a significant asymptotic regression (“sig reg”), the best position was again close to the true position of the selected locus in most cases (average distance = 0.72 cM), but the range of distances was greater (0–8.5 cM) than in the basic scenario (Table 3). The distance between the position of max |iHS| and the true position, which ranged from 0.5 to 9.0 cM (average = 3.41 cM), was similar to the basic scenario. Results using iHS were similar for sig reg and for those cases in which max |iHS| > 1.96 (sig iHS) (Table 3).
Effects of Weak Selection

Similar results to the recent selection scenario were seen with weak selection intensity (sampled at 50 generations, average frequency of favored allele = 0.57 at sampling), where asymptotic regression was significant in 95/100 cases whereas max $|\text{iHS}|$ exceeded 1.96 in 98/100 cases. Again, the average distance between the best position and the true position of the selected locus was lower for regression (1.0 cM) than iHS (2.2 cM for sig reg or 2.1 for sig iHS, Table 3). There were, however, more outliers in the regression analysis than was seen in the recent selection scenario: there were 6/95 significant regressions in which the best position was $>8$ cM from the true position (resulting in higher upper 95% confidence interval and standard deviation of distance between the best and true positions, Table 3), compared with only 2/96 for the recent selection scenario. The deficit of calculated iHS values was less for the equal frequency, recent selection, and weak selection scenarios than for the basic scenario (5%, 7%, and 4% missing iHS values, respectively, for markers within 1.5 cM of the selected locus).

In all 4 scenarios, when the asymptotic regression was significant, the estimation of best position was more precise than seen with iHS such that the average distance between best position and true position was lower and the proportion of best positions within 2 cM of the true position was higher (Table 3).

Application to Empirical Data

In order to further assess the performance of the regression-based approach, it was applied to 2 data sets for which there is independent evidence of selection acting on the genomic region. The first data set encompassed SNP genotypes near the myostatin locus in cattle and the second included microsatellite genotypes near an anticoagulant-resistance locus (VKORC1) in rats.

The Myostatin Gene Region in Cattle

A gene for which there is strong evidence of selection in several cattle breeds is myostatin (Rehfeldt et al. 2000), located at the centromeric end of bovine chromosome 2 (BTA2). Variation in this gene is associated with muscle-related traits in various mammals and, in particular, with the “double-muscled” phenotype (exaggerated muscle conformation) seen in a number of cattle breeds, including Belgian Blue, Piedmontese, Asturiana de los Valles, and South Devon. Animals with this phenotype have a significantly greater percentage of muscle than those with the wild-type genotype, and thus produce more meat at slaughter (Wiener et al. 2009), suggesting that this characteristic has been strongly selected in these breeds due to its positive association with meat production capacity. The Piedmontese breed carries at high frequency a nonsynonymous mutation in myostatin (Kambadur et al. 1997; McPherron and Lee 1997; Pozzi 2009), rather than the 11-bp deletion found in the Belgian Blue, Asturiana de los Valles, and South Devon breeds.

The Bovine HapMap Project (Gibbs et al. 2009) genotyped more than 37,000 SNPs across the bovine genome in 19 cattle breeds, including Piedmontese (the only double-muscled breed in that study). Using this data in the current study, heterozygosity levels were calculated for the Piedmontese breed at all markers across the centromeric end of BTA2. Sequence positions of the markers were based on bovine assembly Btau4.0 (http://www.hgsc.bcm.tmc.edu/project-species-m-Bovine.hgsc?pageLocation=Bovine, Elskie et al. 2009). A region was analyzed encompassing 7 Mb (3–10 Mb), centered at 6.5 Mb, the approximate position of the myostatin gene (6 534 893). This region included 59 markers.

The iHS method was previously applied to a subset of the Bovine HapMap data set (Gibbs et al. 2009). The analysis
included 23 markers in the region under study, the subset where the ancestral and derived alleles could be distinguished, and standardization of iHS was based on genome-wide SNP values. For the Piedmontese breed, extreme values of the test statistic, |standardized iHS|, were found throughout the *myostatin* region, with the peak value (3.909) at a marker positioned at ~9.6 Mb (Gill CA, personal communication), 3.1 Mb from *myostatin*. Other markers closer to the *myostatin* gene were included in this data set but these had lower |standardized iHS| values (the 2 markers closest to the gene had values <1.0).

As described above for simulated data, tests of a selection signal were performed by moving the putative position of a selected gene every 0.10 Mb across the genomic region and testing for the predicted heterozygosity relationship. As with the simulated data, the significance was assessed using an adjusted *P* < 0.05 threshold.

Analysis of the Piedmontese data revealed significant asymptotic regression (right-sense and increasing) relationships at test positions between 6.6 and 7.3 Mb, with the best position at 7.3 Mb (*r*² = 0.16; *P* = 0.0025; estimates: *R* = 0.949, standard error [SE] = 0.519; *B* = −1.9, SE = 18.5, *A* = 2.0, SE = 18.6), 0.8 Mb from *myostatin*. The asymptotic regression relationship fitted with distance from 7.3 Mb is nearly linear in this region (Figure 3A), with a large (negative) value of *B*, the parameter measuring the difference between the value of the curve when distance from the test position is 0 and the asymptotic value of the curve. In contrast, the curve fitted with respect to distance from 6.5 Mb (the approximate position of the *myostatin* gene) tapers off within the graphical region (Figure 3B), with a smaller absolute value of *B*; however, the regression of heterozygosity against distance from 6.5 Mb was not statistically significant. Because the relationship with respect to distance from the best position appeared linear (Figure 3A), linear regression was also fitted to the data. There were significant and positive linear regression relationships for test positions between 6.8 and 9.9 Mb, with the best position at 7.8 Mb (*r*² = 0.21, *P* = 0.00017; estimate: *B* = 0.085, SE = 0.021), 1.3 Mb from *myostatin*. When the regression approach was applied to the subset of the Piedmontese HapMap data previously analyzed with iHS, the asymptotic relationship was not significant but the linear regression was significant and positive for positions between 7.1 and 7.9 Mb, with the best position at 7.4 Mb (*r*² = 0.18, *P* = 0.027; estimate: *B* = 0.102, SE = 0.043), 0.9 Mb from *myostatin*.

**Anticoagulant Resistance in Rats**

Warfarin and other anticoagulants have been used routinely as rodenticides over the last 60 years and the selection pressure for resistance to these compounds is extremely strong as a result of high mortality levels (Bonnet and Gross 1953; Greaves and Ayres 1967). In a study by Kohn et al. (2000), rats were sampled from a number of sites across northwestern Germany where rat control problems had been increasing despite use of anticoagulants. Rats were genotyped using microsatellite markers selected from a region on rat chromosome 1 that had previously been associated with blood clotting response. The authors used a selection mapping approach and linkage-disequilibrium mapping to narrow down the position of the resistance locus to a smaller region. Mutations in the *VKORC1* gene (located at 1:187 176 745 of the rat sequence, within this region) were subsequently found to be responsible for warfarin resistance (Rost et al. 2004).

The regression approach was implemented using microsatellite marker data from a ~14-cM region encompassing the *VKORC1* gene. These markers were genotyped previously by Kohn and colleagues for 93 rats from the study sites described above (Kohn et al. 2000, Kohn M, personal communication). Rat genomic sequence positions (http://www.ensembl.org/Rattus_norvegicus/Info/Index) were
identified for 24 markers within 12 Mb of the \textit{VKORC1} gene. Heterozygosity levels were calculated for each marker. Tests of a selection signal were performed as for the Piedmontese data.

Analysis of the rat data revealed significant right-sense and increasing asymptotic regressions relationships at test positions between 185.8 and 187.3 Mb, with the best position at 186.7 Mb ($r^2 = 0.26; P = 0.016$; estimates: $R = 0.475, SE = 0.236; B = -0.2891, SE = 0.0958; A = 0.5069, SE = 0.0403$), 0.5 Mb from \textit{VKORC1}. The regressions fitted with test position at 186.7 Mb and at the \textit{VKORC1} position were both visually asymptotic (Figure 4A,B). The iHS method could not be applied to the Norway rat data because the genotypes are multiallelic.

**Discussion**

This paper describes a novel regression-based approach to diversity-based gene mapping that utilizes the pattern of diversity with respect to genome position. The statistical relationship between heterozygosity and distance from a selected locus was evaluated using simulated data, supporting the use of the regression approach for detection of selection. The regression method was then compared with an alternative extended homozygosity-based method (iHS) using simulated data and shown to have good power and precision for detecting selection. Finally, the regression method was applied to 2 data sets covering genomic regions for which there is independent evidence of selection and was again shown to perform well.

There have been various published approaches that take advantage of the hitchhiking effect for the detection of selection (i.e., selection/hitchhiking mapping) using neutral markers, as mentioned in the Introduction. The main difference between the regression approach introduced here and other published methods is that it exploits the predicted spatial relationship between diversity at linked markers by testing the statistical fit of genetic variation data across a region to the specific relationship predicted from population genetics theory. Moreover, there are practical limitations to many of the alternative approaches. For example, some are designed for bi-allelic markers (i.e., SNP data) and cannot easily be modified to apply to multiallelic markers. This is limiting for species in which dense SNP chips are not yet available. Some methods also require genotype data for an outgroup, which may not be easily available.

Simulations were used to evaluate the regression approach and results indicate that it can detect selection under a variety of assumptions. A clear result from the simulation analysis is that the performance of the regression method to detect selection improved most dramatically when allelic diversity increased from 2 alleles per locus to greater than 2 alleles and in fact, under the scenario with variable initial allele frequencies, the regression performed better under the sparse map with high allelic diversity than under the dense map with low diversity. This suggests that despite the increasing convenience and cost-effectiveness of genome-wide SNP data, it does not necessarily provide the ideal information for detection of selection.

Comparison of the regression and iHS (Voight et al. 2006) approaches for the SNP-based simulation results revealed that both methods always detected selection in the case of strong selection where the favored allele had fixed. Power to detect selection was also high for both methods in the other scenarios examined (e.g., midway through the selective sweep and/or under weaker selection), although the regression method was slightly less powerful than the iHS approach.

The estimation of the location of the selected locus, however, was more precise using the regression approach for all scenarios examined. Voight et al. (2006) suggest that power to detect selection with iHS can be increased by using
multiple extreme scores over a genomic region (their Supplementary Figure S2), which also indicates that the marker with maximum $|iHS|$ value is not necessarily located closest to the selected locus. The advantage in precision using regression may result from the use of spatial diversity pattern by this approach, that is, the expectation that diversity should increase according to an asymptotic function as one moves away from the selected locus. However, this discrepancy could also be related to the modifications to $iHS$ implementation in the current study (i.e., the standardization procedure and the definition of ancestral and derived alleles), which were required to analyze the simulated data.

The utility of the regression approach was further demonstrated by applying the method to the Piedmontese cattle data from BTA2, for which there is independent evidence of selection on the myostatin gene. Like $iHS$, the regression approach detected selection in the Piedmontese HapMap data, but the best position identified using this method was closer to the myostatin gene than was the marker with maximum $|iHS|$ value. However, the $iHS$ method also detected a selection signal with a smaller data set than did the asymptotic (but not linear) regression approach.

The utility of the regression approach was also demonstrated by applying the method to the Norway rat data from chromosome 1, for which there is evidence of very strong selection on the VkorC1 gene for resistance to anticoagulants. The heterozygosity relationship was more clearly asymptotic in the anticoagulant resistance case than the myostatin case, which may result from the fact that the former study used multiallelic markers, whereas the latter was based on SNPs or from the probable stronger selection pressure on anticoagulant resistance than on muscle conformation. A comparison with $iHS$ could not be made because this method cannot be applied to multiallelic data.

The difficulty of discriminating between population bottlenecks and selective sweeps has been highlighted as a limitation of various selection mapping approaches (Teshima et al. 2006; Thornton et al. 2007; Hermisson 2009). Simulations of a bottleneck showed no increase in Type I error for the regression approach, indicating that this method is able to distinguish selection from simple demographic processes.

Further analysis, however, will be required to quantify more generally the power, precision and Type I error rate of the regression method and other tests under a greater range of scenarios, including more complex demographic processes and background selection. Furthermore, methodological issues such as ascertainment bias in the identification of markers (Nielsen et al. 2005) or nonrandom selection of loci (Thornton et al. 2007) may also affect the detection of selection. It is not yet clear whether the problems highlighted for other methods will also apply to the regression approach. Another issue requiring further study is a characterization of the asymptotic regression curve. There was substantial variability in the distributions of $R$ and $B$, the parameters of the regression curve, in data sets simulated under the same assumptions. Further analysis is required to assess whether the magnitudes of these parameters provide information about the process of selection generating the heterozygosity pattern and whether adjusting heterozygosity by allele frequency could improve curve fitting.

Following on from analyses of candidate regions as illustrated by the cattle and rat examples, the next logical step in the implementation of the regression approach introduced here will be to perform genome-wide scans of SNP data using this method. In order to achieve this, we devise the following extension of the approach. The test position for a selected locus will be moved across the genome in fixed length increments (e.g., 0.1 Mb, as in this study) and the asymptotic regression of heterozygosity onto distance from the test position will then be applied within a window of fixed length (centered with respect to the test position, except at the ends of the chromosomes, where the region will be asymmetric with respect to the test position). The size of the window in physical units will depend on the average recombination rate of the species under study. In the 2 cases studied here, we analyzed regions of $\sim 7$ cM (Piedmontese) and $\sim 14$ cM length (rat) and were able to detect selection using the regression approach, although the latter case had a clearer asymptotic shape than the former. An ad hoc starting point is to use window lengths of 5–10 cM (which encompass regions with ca. 5–10% recombination). However, additional examination of the choice of window and increment size will be carried out in further studies, which may lead to subsequent refinement of the protocol.

In implementing a full genome scan or even a candidate region analysis, one must address the issue of multiple testing for determining appropriate significance thresholds. The standard approach in Quantitative Trait Loci mapping and other statistical analyses is to apply permutation testing (Churchill and Doerge 1994), where a null distribution of the test statistic is determined by breaking the relationship between the variables under study. In this case, heterozygosities would be randomly assigned to positions within the genome (or candidate regions) and the regression analysis performed for each permutation. The upper tail of the distribution of maximum $F$ ratios from the permutations would define the significance threshold.

Although further analyses will help clarify features and extend the utility of the regression approach introduced here, results from this study clearly demonstrate that it can detect selection based on the pattern of heterozygosity across the genome. They also suggest that this approach may prove particularly useful for narrowing down a region already identified as showing evidence of a selective sweep.

**Funding**

British Biotechnology and Biological Sciences Research Council (BBS/B/05710).
Acknowledgments

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical programming.

References


Caroline McCorquodale (Roslin) for assistance with statistical program-

Johnson (Queen Mary, University of London) for valuable discussions, and

for assistance with data manipulation, John Woolliams (Roslin) and Toby

HapMap Consortium for access to the Bovine HapMap Project data, Andy

Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby

Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-

Acknowledgments

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-

We are grateful to Clare Gill (Texas A&M University) for providing iHS results from the Bovine HapMap Project, Michael Kohn (Rice University) for providing the Norway rat data, Benjamin Voight (The Broad Institute) for providing code for iHS calculations and helpful advice, the Bovine HapMap Consortium for access to the Bovine HapMap Project data, Andy Law (Roslin) and Mohammad Edris (Isfahan University of Technology) for assistance with data manipulation, John Woolliams (Roslin) and Toby Johnson (Queen Mary, University of London) for valuable discussions, and Caroline McCorquodale (Roslin) for assistance with statistical program-


Received August 16, 2010; Revised February 8, 2011; Accepted February 8, 2011

Corresponding Editor: Howard Ross