2 Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data

2.1 Introduction

A genome-wide scan of positive selection, in which the entire genome is examined, has been used in several studies. In some scans, such as when non-synonymous amino-acid substitutions showing high levels of population differentiation were chosen[83], there has been a limited prior hypothesis about the target of selection. But genome scans can also be carried out in the absence of any such hypothesis. Such unbiased scans have the attractive feature that they can potentially lead to entirely unsuspected insights into the evolutionary history, but in order to derive full benefit from them, the target of selection must be identified. In practice, most genome scans have been based on SNP genotyping, and methods for detecting potential selection have been primarily based on searching for unusual LD or population differentiation patterns. Such scans have, in some senses, been highly successful. A review summarizing the combined results of nine such genome scans found that 5,110 distinct regions covering 14% of the genome and 4,243 (23%) RefSeq genes showed apparent evidence of positive selection[100]. However, although these findings are impressive for their yield of putatively selected regions, it was notable that there was limited overlap between the individual surveys and only 129 of the regions (2.5%) were identified in four or more studies. This poor concordance was described as "sobering"[101] and pointed to the need for a better understanding of the false positive and false negative rates in such scans. Indeed, other analyses have suggested that the classic selective sweeps detected by these approaches are unlikely to have been frequent enough to dominate overall patterns of human genome diversity[102]. A second feature of some of these scans, particularly those based on LD, is that the candidate regions identified can be very large. For example, the HapMap2 project listed 22 strong candidate regions with a

combined length of ~16.7 Mb and mean size of ~760 kb[66], making it difficult to identify the selected target and further investigate the biological implications of the selection.

We have set out to address three questions raised by genome scans that identify large candidate regions. First, do such candidates show evidence for selection if alternative criteria are used? Second, to what extent can the targets of selection be localized more precisely? And third, if more precise localization is possible, does this lead to increased insights into the possible biological basis of the selection? To achieve these aims, we reasoned that full re-sequence data would provide the most information. Indeed, only technical and cost limitations have previously hindered its use: re-sequencing complete genomes or even hundreds of kilobases (kb) to high accuracy in population samples has not been practical until recently[40]. We have thus explored experimentally the potential for enrichment of such regions followed by next-generation sequencing to generate suitable datasets. We chose for these trials two regions from the HapMap2 survey, which were of intermediate size (~300 kb each) and where there was no obvious target for selection[66]. We show using simulations that alternative tests for selection applied to sequence data from regions identified in such a way should readily distinguish between neutrality and likely selection, and will usually produce a more precise localization of the selected variant. We also show experimentally that suitable high-quality sequence data can be generated using next-gen technology, and finally that plausible biological candidates can then be proposed for these selective events.

This study is published in *Human Genetics*[103]. This chapter is based on this publication, with some modification of the contents. All simulations, statistical and bioinformatic analyses were performed by the author of this thesis, except the CMS calculation, which was done by Irene Gallego Romero. The PCR experiments were done by Qasim Ayub, and the sequencing work was done by the Sanger Institute Sequencing Team.

2.2 Materials and Methods

2.2.1 Simulations

Two-step simulations were performed to model both neutral and positively selected scenarios, and are summarized in Figure 2.1. In the first step, we carried out coalescent simulations using the *cosi* package to generate 1 Mb long haplotypes in a pair of ancestral populations 2,000 generations ago based on the best-fit demographic models for African and Asian populations[26]. These haplotypes were then used as input for the second step - forward simulations, using *mpop*[94]. In some of these forward simulations in the Asian population, one allele with an initial frequency of 0.0006 (default initial frequency for new variants in the package), which would be under selection, was added in the middle of the simulated Asian haplotypes. Four different selection scenarios with selection coefficients (s) of 0.001, 0.004, 0.007 and 0.01 were simulated, and the selection start time was set at 2,000 generations ago. In total, 1,000 independent simulations were performed for each set of conditions. These used the genome-average recombination rate of 1cM/Mb from the HapMap2, a mutation rate of 1.8×10^{-8} per nucleotide per generation calculated from a comparison of human and chimpanzee sequences for the whole of chromosome 4, and a current effective population size of 100,000. The rest of the demographic parameters were as in Schaffner et al.'s best-fit demographic model[26] from the package *cosi*. For the purpose of computational efficiency, we re-scaled the parameters when performing the forward simulations: effective population sizes and times were reduced by a factor of 5, while mutation and recombination rates and selection coefficients were multiplied by 5 (see Appendix A for parameters and commands). Fifty chromosomes were sampled from each simulation. We call this set of data the "simulated re-sequencing data".

The SNPs in the "simulated re-sequencing data" were subsampled to mimic the frequency spectrum of HapMap2 genotype data by matching the proportion of the SNPs of HapMap2 data in each frequency bin (bin size 0.1). We call this set of subsampled simulation data the "simulated genotype data". XP-EHH scores[66] were calculated from the simulated genotype data and normalized using the

mean and variance of the XP-EHH scores from the simulated genotype data in the neutral simulation in the Asian population, using the African as the reference population. We only retained simulations with the XP-EHH score above the 95th neutral percentile continuously for at least 100kb surrounding the selected SNP, which mimics the experimentally-investigated candidate regions from the survey based on the HapMap2 data.

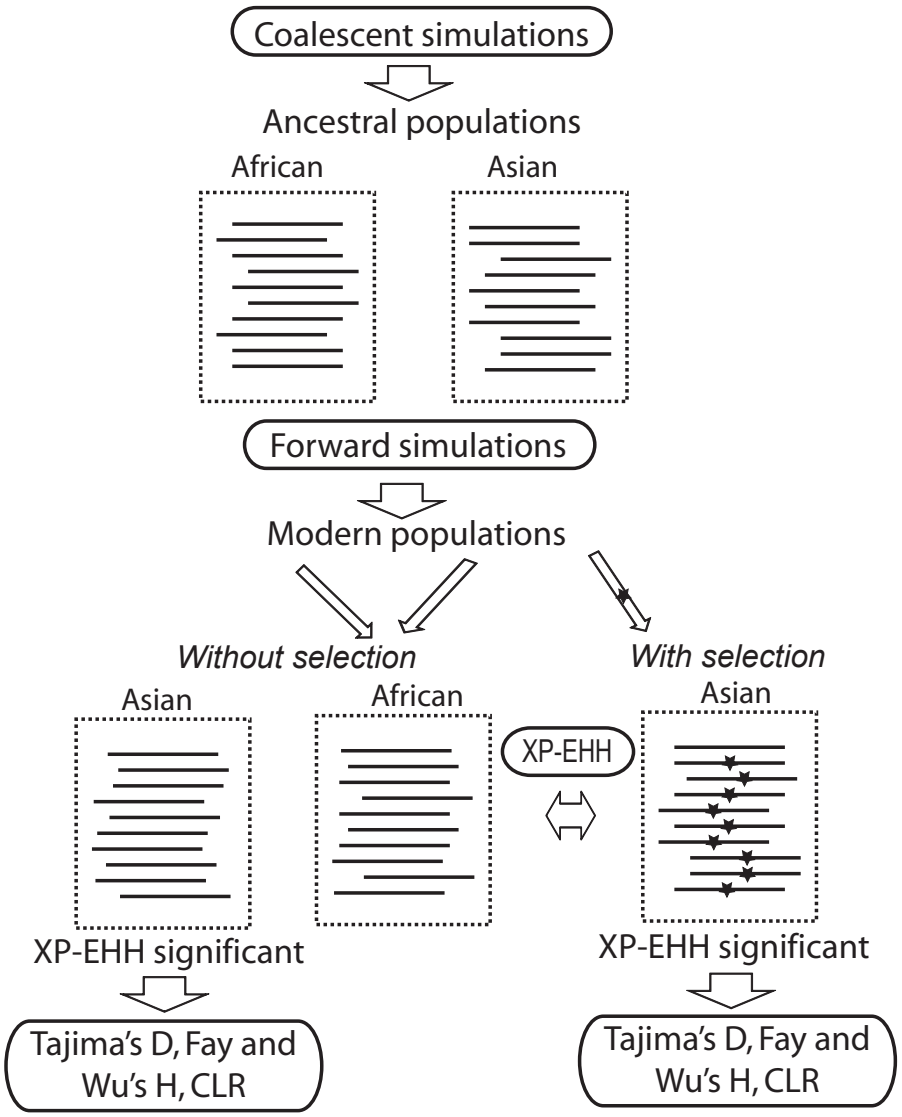

Figure 2.1 Simulation design. Dotted boxes represent simulated haplotype samples; the star indicates the presence of a positively selected SNP. Arrows show the performance of the analyses described in the oval boxes.

We then returned to the corresponding simulated re-sequencing data for the retained simulations and calculated Tajima's D[71], Fay and Wu's H[72] and Nielsen et al.'s CLR[76] statistics. These were calculated in 10 kb non-overlapping windows

across the whole 300 kb region centered on the selected SNP (or equivalent location in neutral simulations) in each individual set. The significance levels for each of the neutrality tests were estimated based on the percentile of the test values in the null distribution from 1,000 neutral simulations with the same demographic model. The background frequency spectrum required by the CLR analysis was calculated on the 1,000 independent neutral simulations with the same recombination and mutation rates. In order to combine signals from the three tests, we assessed the correlation coefficient between Tajima's *D* and Fay and Wu's *H* p values on the neutral simulated data, and found no correlation (r = 0.06); therefore, these two tests were treated as independent, and a combined p value from Tajima's *D* and Fay and Wu's *H* for each 10kb window was calculated using Fisher's method[104], and we use this combined p value to present the results below.

2.2.2 Target region resequencing

Two regions were picked from the HapMap2 list of 22 regions showing strong evidence of selection[66] using the following criteria: no obvious candidate for the selected SNP or gene; selection at least in the CHB+JPT population; moderate size (0.2-1 Mb). The coordinates of the chosen regions were (March 2006, NCBI 36 assembly; all genomic coordinates in Chapter 2 are based on this assembly) chromosome 4: 158,702,285-159,016,211 (314 kb, called chr4:158Mb) and chromosome 10: 22,587,453-22,850,110 (263 kb, called chr10:22Mb). We also included a set of control regions, including CASP12 (13 kb) for which we had the Sanger capillary sequencing data from a subset of the samples for the resequencing of this study[105] and 20 kb of unique sequence from the Y chromosome, where there should be no reads mapped in females and no heterozygote calls in males.

The target regions were then amplified from 28 CHB (Han Chinese in Beijing, China) and 2 YRI (Yoruba in Ibadan, Nigeria) samples from the HapMap collection in a series of long-range PCRs. In total, 49 pairs of PCR primers were designed for chr4:158Mb, 42 for chr10:22Mb and 4 pairs for the Y chromosome to amplify 5-11 kb PCR products with overlap of > 500 bp, using a Perl script

(http://droog.gs.washington.edu/PCR-Overlap.html). Two previous pairs for *CASP12*[105] were also used. The three base pairs at the 3' end of all primers were confirmed not to overlap with any SNP in dbSNP127 (http://www.ncbi.nlm.nih.gov/projects/SNP/). The primer sequences and PCR conditions are listed in Appendix B, PCR primers and protocols. 44 out of 49 fragments from chr4:158Mb, 37 out of 42 from chr10:22Mb and all from the Y chromosome and *CASP12* were successfully amplified in initial tests. These fragments were subsequently amplified in all samples. Three CHB provided poor quality data for chr4:158Mb, and four for chr10:22Mb, and were excluded from all subsequent analyses. Amplification was tested by agarose gel electrophoresis followed by ethidium bromide staining, and approximate quantification was performed from the band intensity. 39 out of 49 (~80%) long PCR primer pairs worked well for 22 or more samples for chr4:158Mb, and 32 out of 42 (~75%) for 20 or more samples for chr10:22Mb. The lab work of PCR enrichment was done by colleague Qasim Ayub. The PCR products from each individual sample were pooled, approximately equalizing the molar yield for the Illumina sequencing paired end library construction.

In order to avoid artifacts in tests results due to the missing data in PCR gaps, we used another set of data from a hybridization enrichment experiment based on a Nimblegen custom array or solution pulldown approach[106] on the same two regions in a subset of samples (19 CHB) to fill the missing data. For chr4:158Mb region, six gaps were filled:158,702,285-158,708,035, 158,770,931-158,783,816, 158,827,935-158,840,376, 158,880,521-158,900,211, 158,906,161-158,913,233 and 158,985,263-158,992,841. For chr10:22Mb region, six gaps were also filled: 22,624,537-22,630,034, 22,643,514-22,656,292, 22,662,169-22,675,644, 22,689,042-22,696,558, 22,761,012-22,769,435 and 22,801,106-22,813,376.

Illumina paired-end libraries of ~200 bp fragments were then constructed on the enriched regions, and 37 bp from each end sequenced on an Illumina GAII[107] platform, with one sample per lane. After filtering out duplicate reads, the amount of mapped data ranged from 322 Mb to 572 Mb, leading to a mean coverage per individual of ~500x to >1000x for the parts which PCR amplified and ~ 35x to ~ 250x for pulldown regions. The paired-end sequence reads were

mapped back to the target reference sequences or the whole genome by SSAHA2 and candidate SNPs were called by SSAHASNP[48] for the PCR amplified regions, while MAQ[46] and SAMtools[108] were used for the data from the pulldown-enriched regions. By comparing the SNP calls based on Illumina data from *CASP12* with the existing capillary sequence data and avoiding heterozygous Y chromosome SNP calls, we set filtering criteria to filter out unreliable calls. For the SSAHA2 candidate SNPs from PCR enrichment, we filtered out all SNPs which lay within the primers or SSAHASNP indel calls, had coverage less than 30, or showed a ratio of the second-highest:total read depth of < 0.30:1 for a heterozygous SNP call. We only consider SNPs since indel variants are not reliably identified by this approach. For the MAQ and SAMtools candidate SNPs from the pulldown enrichment, SNP calls were filtered individually based on coverage, SNP score and mapping quality using criteria set based on the *CASP12* and Y chromosome data. The quality of the filtered SNP data was assessed by comparing the overlapping calls from our data with the HapMap2 genotypes from the same individuals. There were 43 discrepancies out of 2,981 comparisons for the chr4:158Mb and 5 out of 857 for the chr10:22Mb region, which suggested a low error rate for both regions (98.6% and 99.4% concordance, respectively). To assess whether such error rates affect the quality of subsequent statistical analyses, random errors were introduced into the simulations described above, matching the error rates, and results were compared with simulations without errors. This analysis showed that such error rates would not affect the power of the sequence analyses (results shown in Section 2.3.1).

We inferred haplotypes and occasional missing data using PHASE 2.1[37]. Then the neutrality tests and Nielsen et al.'s CLR test were performed on non-overlapping 10-20kb regions containing two or three PCR fragments chosen based on the size of each PCR fragment and the SNP densities.

2.2.3 Bioinformatic analysis

All miRBase (Release 13) mature miRNA sequences were scanned against the selected regions of the human genome using the MapMi algorithm[109]. This

approach involves first scanning the regions for matches to mature miRNA sequences; regions with matches to known miRNAs (allowing one mismatch) were then excised and folded using RNAfold from the ViennaRNA package[110]. These candidate regions were scored and filtered according to how well they fitted the stem-loop precursor structure common to miRNAs. We ran the pipeline in stand-alone mode, using non-repeat masked genomic sequence for increased sensitivity. The chr10:22Mb region had no significant hits for any known miRNA; however, the chr4:158Mb region had two hits to the miR-548 family of miRNAs, discussed below. This analysis was done in collaboration with José Afonso Guerra-Assunção from the European Bioinformatics Institute.

2.3 Results

2.3.1 Simulation of the power to detect and localize positive selection using genotype-based and sequence-based tests

In order to understand whether sequencing data provide more power in detecting and localizing selection signals, we started by comparing the power of genotype-based and sequence-based analyses using simulations. We first modeled the genotype-based tests mimicking those in the HapMap2 study, and in particular, the selective events seen in the CHB+JPT by comparison with the YRI population. To do this, we performed forward simulations under neutrality using the YRI and CHB demographic models, and with selection coefficients of 0.001, 0.004, 0.007 and 0.01 using the CHB demographic model, as described in section 2.2.1. Of the 1,000 simulations in each neutral and selected CHB set, there were 16, 16, 233, 724 and 779, respectively, that met the XP-EHH filtering criteria. These were combined into 16 significant XP-EHH results under neutrality and 1,752 under a range of selective conditions that would reflect the data that might be obtained from a population experiencing a variety of selective pressures.

We next applied the sequence-based tests to the 16 neutral and 1,752 selected datasets. There were 2 simulations among the 16 retained neutral ones that showed at least one significant window for the combined p value (≤ 0.01), and 7

for Nielsen et al.'s CLR. These numbers represent the false positive rates for the two methods, and are significantly higher for the CLR (p = 0.048, Fisher exact test). In the retained selected simulations, 84% (1,469 out of 1,752) for combined p value and 85% (1,494 out of 1,752) for Nielsen et al.'s CLR showed at least one significant window. Thus there is good power to detect this form of selection using sequence-based tests.

To investigate the ability to localize the causal SNP using the sequence-based tests, we first examined the test statistics averaged over all retained simulations. The average values of both showed no pattern along the DNA in the neutral simulations, but a strong peak centered on the window containing the selected site in the selected set, with a gradual decrease on either side (Figure 2.2 A and B). This indicates that, on average, the frequency spectrum-based neutrality tests can correctly identify the location of the causal SNP, but that there is considerable variation between simulations.

We therefore investigated this variation further by counting the occurrence of the most significant signals in each window in different simulations. For the combined p value, the most significant window lay within the 40 kb region (i.e. ±20 kb) surrounding the selected allele in 46% of the simulations, compared with 68% for the CLR (Figure 2.2 C). These results show that Nielsen et al.'s CLR performs better for localizing the selection signal, as previously reported[76]. Although the combined p value of Tajima's *D* and Fay and Wu's *H* and Nielsen et al.'s CLR have similar power for detecting selection (84% and 85%), we saw a lower false positive rate on the combined p value but a better localization power in Nielsen et al.'s CLR. Therefore, we investigated the benefits of further combining these signals. We tried using the combined p value to detect selection and then the CLR to localize the signal. This approach did systematically increase the accuracy of localization, although only by a small amount (Figure 2.2 D). We also considered the subset of simulations where the combined p value and Nielsen et al.'s CLR signals lie within the same 10kb window. Although the proportion is low (11.3%, or 198 out of 1,752 simulations), these might represent a favorable situation with the best chance to localize the selection signal. Indeed, this subset of simulations has about 90% chance to localize the

selection to a 40kb region and 80% to 20kb. These results provide an overall view of the power for localizing the signals in different scenarios and can guide the search for the biological basis of the selection.

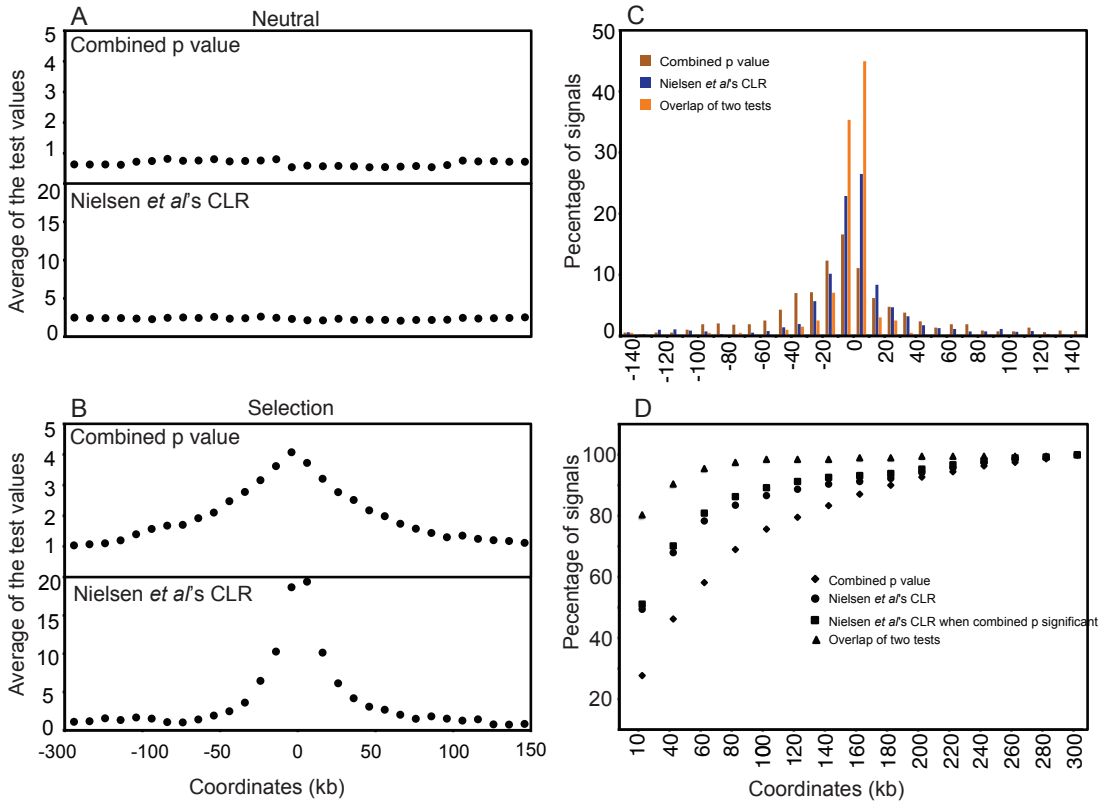

Figure 2.2 Simulation results. A. Simulations were carried out under neutrality, and tests for selection (-ln combined p values for Tajima's *D* and Fay and Wu's *H* (top) or Nielsen et al.'s CLR (bottom)) were calculated in non-overlapping 10 kb windows across 300 kb. Values of the test were averaged over 1,000 independent simulations. No departures from neutrality were seen. B. Simulations were carried out with selection (selection coefficient 0.007) and neutrality tests applied as in A. Departures from neutrality are seen most strongly in the window containing the selected SNP. C. The distribution of the top signal (lowest p value) in each simulation is shown across the 300 kb region. D. Probability that the known selected variant is found at each distance from the peak test value.

As mentioned above, there is ~1% error in the SNP calls from our sequencing data. In order to evaluate the effect of these errors on our analyses, we added 1% random base substitution errors to one of the datasets simulated with selection (s = 0.007) and recalculated Tajima's *D* and Fay and Wu's *H* on the data with errors. Signals were overall slightly lower, but the pattern of signals was not affected (Figure 2.3). We therefore conclude that sequence errors at this level would not significantly influence our conclusions.

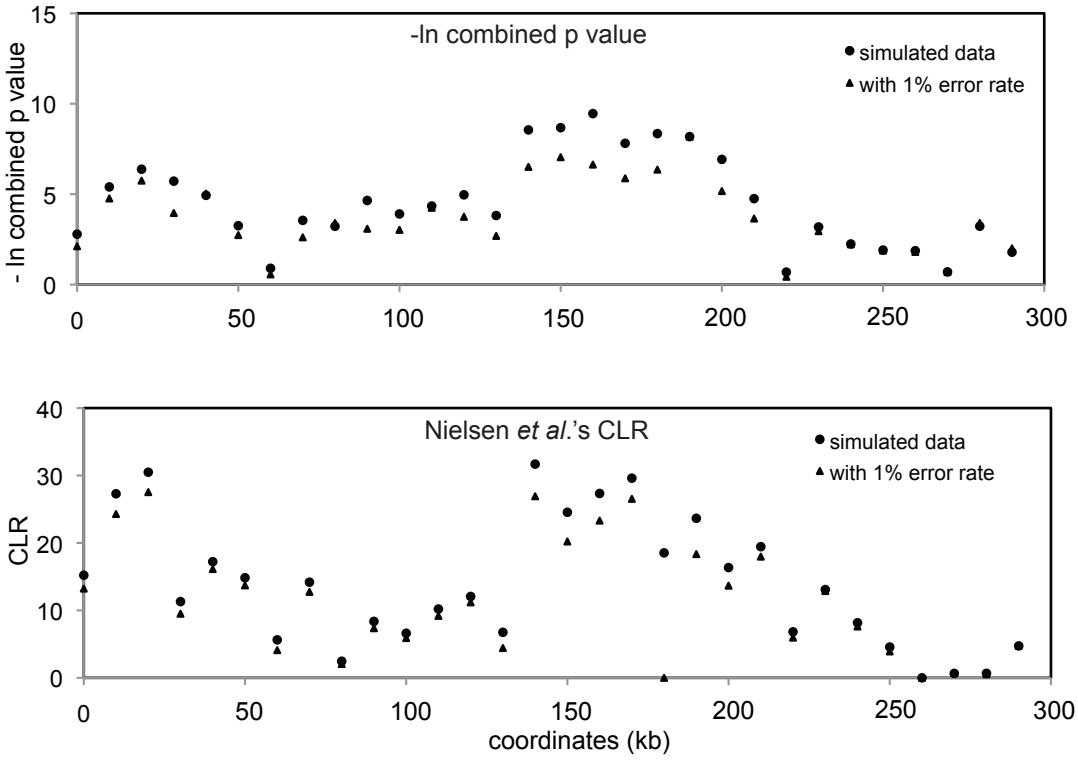

Figure 2.3 Test results on simulated data with 1% sequencing error rate versus no error. Dots represent results with no error in the simulated data, and triangles represent results with 1% random substitution errors introduced in simulated data.

2.3.2 Detection and localization of positive selection signals in experimental data

We re-sequenced two ~300 kb regions that had shown strong signals of positive selection in the HapMap2 study in 25 (chr4:158Mb) or 24 (chr10:22Mb) CHB individuals. The combined p value and Nielsen et al.'s CLR were calculated in chunks spanning either two or three PCR fragments, and are plotted in Figure 2.4 A and B. In both cases, a single window carries the most significant signal from each test: a combined p value of 0.00036 for chr4:158Mb, and 0.000015 for chr10:22Mb, and corresponding CLR values of 47 and 62. The two windows are located at 158,971,591-158,985,262 of chr4, and 22,755,918-22,776,116 of chr10, with sizes of ~13 kb and ~20 kb, respectively. Based on the simulations, this is a particularly favorable situation for localizing the selected variant, and we have 80% confidence that the target of selection should lie in a 20 kb region centered on these windows.

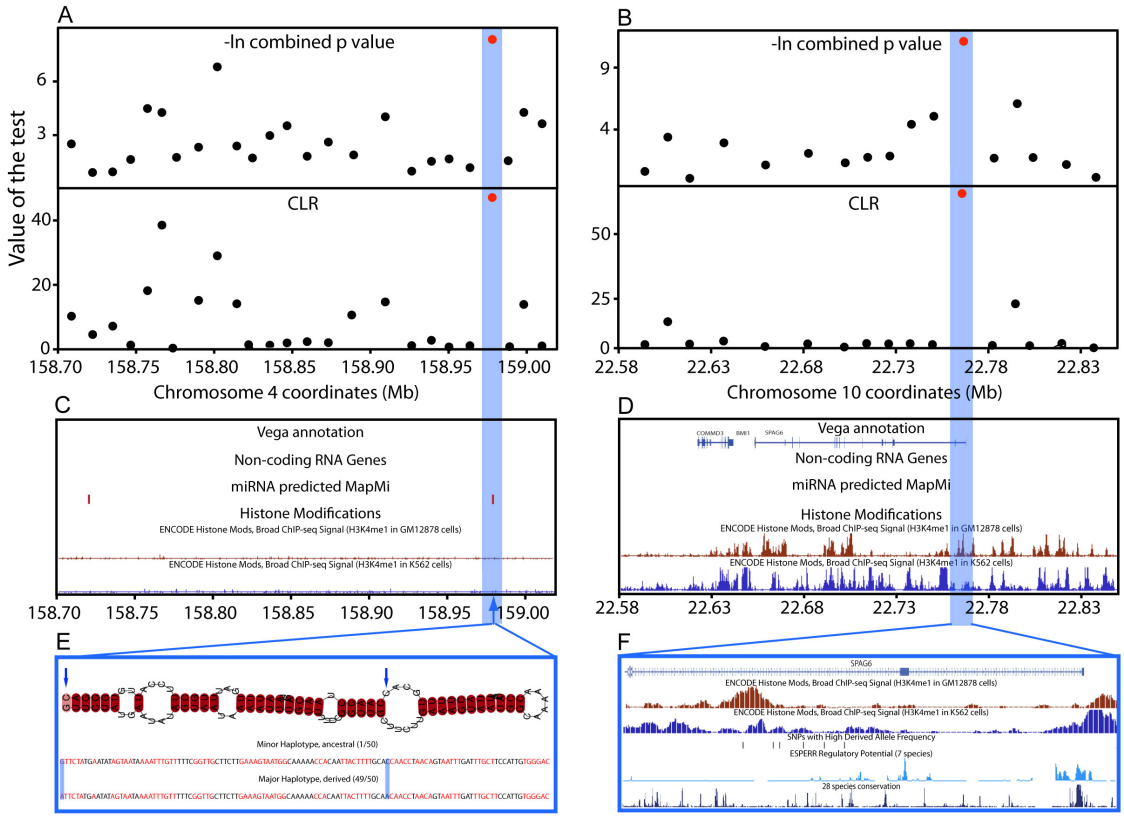

Figure 2.4 Experimental results. These figures show localization of likely selection targets in the chr4 and chr10 regions. A. -log e of combined p values from Tajima's *D* and Fay and Wu's *H* (top) and Nielsen et al.'s CLR (bottom) calculated from re-sequencing data in windows corresponding to two or three PCR fragments (10-20 kb). The most significant statistics are shown in red, and fall into the same window overlap at ~158.98 Mb (blue highlight). B. Corresponding analysis of the chr10:22Mb region, where the most significant signals again fall into the same window, this time at ~22.78 Mb. C, D. Protein-coding genes from the Vega annotation, non-coding RNA and miRNA genes, and relevant ENCODE chromatin modifications in the two regions. E. Predicted miRNA in the chr4:22Mb target region. Two SNPs are present, including a G>A at the end of the miRNA carried on the major haplotype (49/50 chromosomes, selected in CHB) that may influence the strand forming the mature miRNA. F. H3K4me1 chromatin modifications indicating enhancer regions in GM12878 (second) and K562 (third) cells, SNPs with high derived allele frequencies (fourth), predicted regulatory potential (fifth) and 28 species conservation (bottom). Three high-frequency derived SNPs lie within candidate enhancers in one or other of the cell lines, but high-frequency derived SNPs do not lie within regions with high predicted regulatory potential or conservation.

2.3.3 Biological targets of selection

The final stage of our analysis was to search for possible biological targets of selection. Such targets should most likely lie within the narrowed interval, and carry a biologically relevant difference between the selected and non-selected haplotypes. The 314 kb region on chromosome 4 consists entirely of intergenic sequence, and the nearest annotated protein-coding gene is located more than 50 kb outside this region. No histone modifications indicative of promoters,

insulators or enhancers were apparent in publically available data (Figure 2.4 C). However, using the MapMi approach[109], we found two predicted microRNAs (miRNAs) belonging to the mir-548 family (Figure 2.4 C). One of these lay far from the selection signal but the other, hsa-miR-548c, lay at 158.982 Mb, within the narrowed region (Figure 2.4 C). Strikingly, two SNPs are present within this predicted miRNA and both show high derived-allele frequencies in the CHB population. One of these SNPs lies within a loop in the predicted RNA and is not predicted to have functional consequences. However, the other is the first nucleotide of the miRNA precursor and could therefore determine which strand is processed to form the mature miRNA and consequently change the set of target genes (Figure 2.4 E).

The chromosome 10 region contains three annotated protein-coding genes, *COMMD3*, *BMI1* and *SPAG6*, and no miRNA genes (Figure 2.4 D). *SPAG6* transcripts (e.g. SPAG6-002, OTTHUMT00000047185: http://vega.sanger.ac.uk/Homo_sapiens/index.html) extend into the narrowed region (Figure 2.4 D). ChIP-seq experiments reveal extensive chromatin modification within the 263 kb region, including H3K4me1, H3K4me2, H3K4me3, H3K9ac, H327me3 and H3K27ac (http://genome.ucsc.edu/ ENCODE Histone Mods, Broad ChIP-seq; Figure 2.4 D), as would be expected for a region containing several protein-coding genes. The narrowed region contains two peaks of H3K4me1, which could indicate an enhancer[111]. Thus *SPAG6* provides a good candidate on the basis of its location relative to the signal of selection. Although *SPAG6* contains a relatively high-frequency derived non-synonymous SNP (rs7074847) in the YRI[66], there are no non-synonymous differences between the selected and non-selected CHB haplotypes, suggesting that selection is more likely to be acting on an aspect of transcription than on a change in the protein sequence.

In conclusion, based on our analyses, possible targets for selection can be identified in both regions and there is strong functional evidence for selection.

2.4 Discussion

2.4.1 Power of detection and localization

The first question we addressed was whether or not candidate regions identified in genome scans for positive selection using LD-based tests on genotype data, such as that performed by the HapMap2 project, would show supporting evidence for selection when frequency spectrum-based neutrality tests were applied to re-sequencing data. Such tests are sometimes considered most suitable for detecting complete sweeps, in contrast to the partial sweeps detected by LD-based methods, but are also highly effective in detecting partial sweeps[112]. The answer to this question, from both our simulations and the two experimental examples investigated, was a clear "yes". Significant departures from neutrality (combined p value from Tajima's *D* and Fay and Wu's *H*) were seen in 84% of the 1,752 simulations that passed the XP-EHH threshold, contrasted with just 2 out of the 16 neutral simulations that by chance passed (not significantly different from 0 out of 16, Fisher exact test). A similar result was seen with Nielsen et al.'s CLR, although the false positive rate was higher. This correspondence is unsurprising, given the similar underlying basis for the two tests, but there was value in combining the two (see Section 2.2). In the two regions investigated experimentally, significant values were seen in both with all the tests applied.

The second question was the extent to which targets of selection could be localized more precisely when using re-sequencing data. From the simulations, we found that re-sequencing data do provide valuable additional information about the localization of selection targets. Higher SNP density and the presence of more rare variants make a higher resolution of signals possible. One of the disadvantages of LD-based test is that they detect large LD blocks, which are often several hundred kb in length. Although some frequency spectrum-based tests can also be used on genotype data, for example Nielsen et al.'s CLR, the window size often has to be relatively large because information from many SNPs needs to be combined to get enough power. We applied Nielsen et al.'s CLR using HapMap2 genotype data, and for chr4:158Mb, it detected a signal of

selection localized to ~40 kb, while for the chr10:22Mb region, where the HapMap2 SNP density was high in the critical interval, the selected region was narrowed down to a similar length to the sequencing data (Figure 2.5). A method for combining multiple signals derived from genotype data has been described[113], which provides a median localization to a 55 kb interval. This method identified a chr4:158Mb interval spanning ~60kb (158,862,019-158,921,890, with top SNP at 158,904,521), but failed to find any significant signal at chr10:22Mb[68] (Figure 2.5). We repeated the CMS analysis using the HapMap2 genotype data and localized the chr4:158Mb signal to a similar ~58kb interval, although with a different peak SNP (158,862,019- 158,920,326, with top SNP at 158,920,326), and also found no signal in the chr10:22Mb interval (Figure 2.5 and Figure 2.6). In contrast, re-sequencing followed by the application of the tests used here provided localization to a < 20 kb interval in both cases.

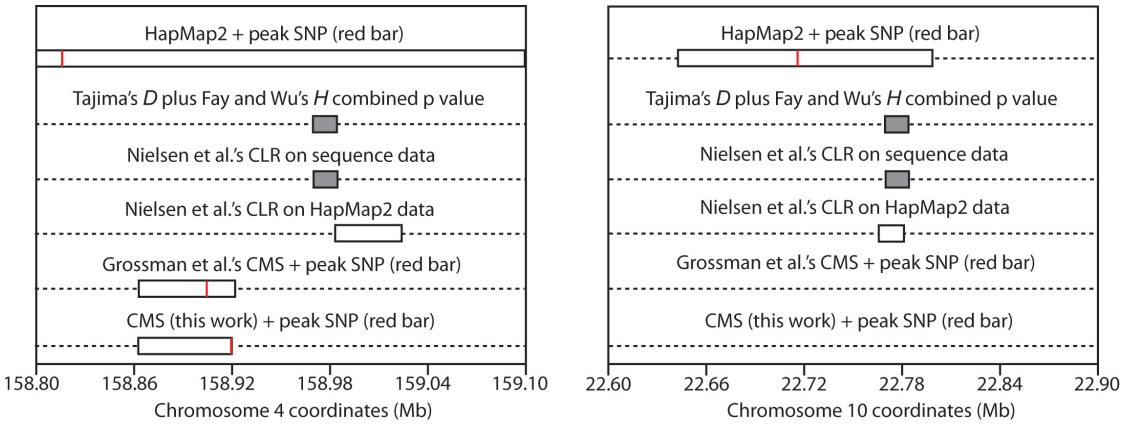

Figure 2.5 Comparison of different approaches of signal localization. These figures show localization of the signal of selection within the chr4 and chr10 regions using different approaches. The two starting regions are shown at the top (Sabeti et al. 2007), localizations using sequence data (grey bars) or HapMap2 genotype data (white bars) by this study in the middle, and the localization by the CMS statistic (Grossman et al. 2010 or this work) at the bottom.

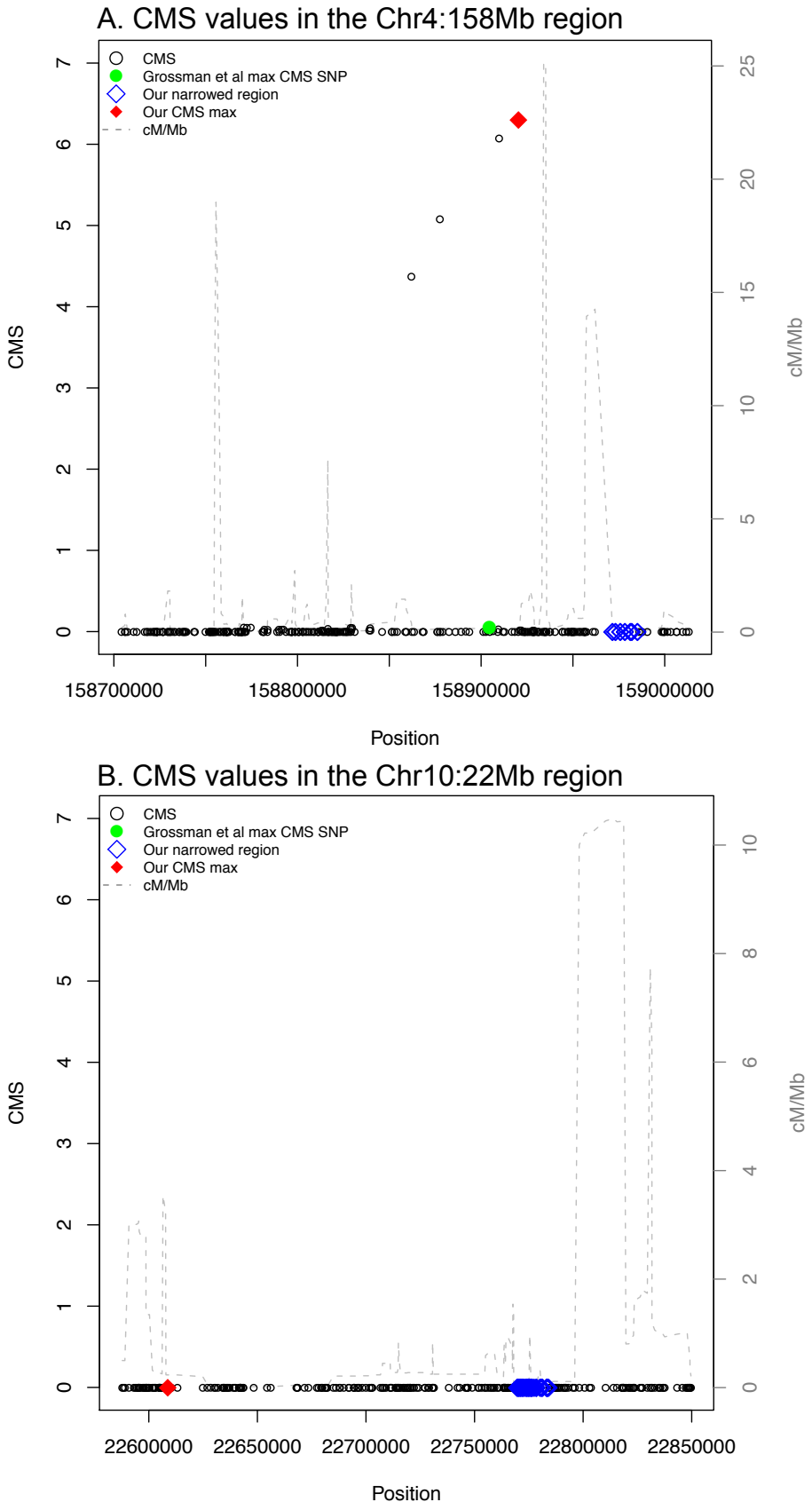

Figure 2.6 CMS results on both regions. Recombination intensities are shown as dashed lines.

2.4.2 Functional targets of selection

The final question we set out to address was whether increased insights into the possible biological basis for the selection could be obtained. Due to our inability to predict the phenotypic consequences of most DNA variants, particularly when these lie outside protein-coding regions, it is often still difficult to identify the causal variant. Nevertheless, the narrowed region provides the best starting point for further investigation. It is, in principle, possible that variants in a region could be acting on distant genes, but this in practice seems rare: a study of human eQTLs, for example, found that most lie either within or close to the genes they affect, with only 5% lying > 20 kb away[114]. On this basis, we therefore focus on targets close to the narrowed regions in the following discussion.

For chr4:158Mb, the above considerations and the lack of any annotated protein-coding genes in the vicinity make a direct effect on a protein-coding gene unlikely. Predicted miRNA hsa-miR-548c, however, provides an intriguing candidate. Members of the hsa-miR-548 family are derived from the transposable element Made1, present in multiple (~30) copies in the human genome[115]. Made1 elements are found only in primates, and hsa-miR-548 sequences have been documented only in the human, chimpanzee and macaque genomes, where they appear to be evolving rapidly. Since miRNAs function as post-transcriptional regulators by binding to partially complementary target sites in the 3' untranslated regions of mRNAs and inhibiting their expression, a change in the sequence of a mature miRNA could influence the expression of a large number of genes, and a change in the strand present in the miRNA could have even greater regulatory effects. More than 3,500 genes have been listed as predicted hsa-miR-548 targets, enriched in functions such as cell proliferation[115]. We can thus speculate that a variant hsa-miR-548c might have been selected because of altered target gene regulation, but the large number of hsa-miR-548 family members and potential targets makes it difficult to formulate or test more precise predictions. Nevertheless, a link to changes in gene regulation fits well with general thinking about the importance of regulatory mutations in human evolution[116] and the inference of recent positive selection acting on a miRNA-rich region on chromosome 14 devoid of annotated protein-coding genes[117] .

For chr10:22Mb, similar considerations lead to the suggestion that *SPAG6* is the most likely target of selection, and a change in the level, timing or location of its expression as the most likely mechanism. In support of this possibility, it was notable that a *SPAG6* transcription end site lay within the narrowed region, and Veyrieras et al.[114] had reported a strong enrichment of eQTLs in the 250 bp just upstream of the transcription end site. However, in the *SPAG6* data, the closest SNPs were 2,055 bp upstream and 843 bp downstream of the transcription end site. In contrast, two H3K4me1 signals indicative of enhancers are located within the narrowed region, and three high-frequency derived SNPs (rs16922285 at 22,773,002, rs11012996 at 22,773,902 and rs11012997 at 22,774,094) specific to the selected haplotype overlap with them (Figure 2.4 F). An altered enhancer activity thus provides the most plausible biological mechanism. SPAG6 is a component of sperm[118], and mouse knockout models have been investigated: 50% of *Spag6*$^{-/-}$ mice died within eight weeks due to hydrocephalus (fluid on the brain); males surviving to maturity showed abnormalities of sperm structure and mobility and were infertile[119]. Heterozygous *Spag6*$^{+/-}$ animals showed a much milder phenotype and were fertile, but their sperm swam more slowly, suggesting that a reduced level of SPAG6 protein can have a detectable effect on the sperm phenotype. The hydrocephalus phenotype, however, points towards a wider role of the protein in the function of cilia, and thus other potential modes of selection. Nevertheless, the best candidate remains an effect on reproduction, which would be consistent with both the inference of recent positive selection on another sperm protein gene, *SPAG4*, in the CHB among other populations[67], and the high frequency with which genes linked to reproduction are found more generally in surveys of positive selection[63,67].

There are two other protein-coding genes in the interval, both > 100 kb from the strongest selection signal. Little is known about *COMMD3* itself, but diverse functions have been ascribed to other COMMD family members, including copper metabolism and regulation of the activity of the transcription factor NF-κB and cell proliferation, perhaps through the ubiquitin pathway[120]. *BMI1*, in contrast, has been studied extensively. It is a polycomb protein, involved in DNA repair, chromatin remodeling and stem cell renewal, and its inappropriate over-

expression can lead to tumor formation[121-123]. Knockout mice are viable and homozygotes show hematopoietic, skeletal and neurological abnormalities, but phenotypic effects in the heterozygotes were not noted[124]. In humans, a cysteine to tyrosine substitution at position 18 leads to substantially lower levels of BMI1 protein, and is present in the general population, including in the YRI and CEU (but not CHB) HapMap samples[125]. Since increased expression of BMI1 leads to cancer, and a decreased expression phenotype is present in HapMap populations but has not been positively selected, both *COMMD3* and *BMI1* seem less strong candidates than *SPAG6* for the target of chr10:22Mb selection.

2.4.3 Conclusion

From these examples, we can conclude that the approach used here, of re-sequencing large target regions, refining the target location and making inferences about the biology of the selection events, is fruitful. However, it could be improved in several ways. Re-sequencing technology is still imperfect and data quality needs to be improved. This study required a combination of two enrichment strategies, PCR and pulldown, to generate adequate coverage, and such intensive effort is impractical for large-scale studies. Most urgently, however, better statistics for localizing the target of selection using re-sequencing are needed, and improved methods for interpreting the biological consequences of DNA variants discovered are especially needed. But even with the present tools, specific topics to follow up experimentally can be suggested, e.g. comparison of sperm mobility and other sperm characteristics between carriers of selected and non-selected haplotypes in the chr10:22Mb region. More generally, the availability of population-scale re-sequencing data from both the increasing number of personal genome projects[126] and projects such as the 1000 Genomes Project[40] will make the approach used here applicable across the genome.