ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

www.sciencedirect.com/science/journal/16720229

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Article

Scanning for Genomic Regions Subject to Selective Sweeps Using SNP-MaP Strategy

Libin Deng[1,2,3#], Xiaoli Tang[1#], Wei Chen[2,4], Jiari Lin[3], Zhiqing Lai[3], Zuoqi Liu[1],
and Dake Zhang[2*]

[1]*Faculty of Basic Medical Science, Nanchang University, Nanchang 330006, China;*
[2]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;*
[3]*The Institute of Translational Medicine, Nanchang University, Nanchang 330006, China;*
[4]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China.*

Genomics Proteomics Bioinformatics 2010 Dec; 8(4): 256-261 DOI: 10.1016/S1672-0229(10)60027-7

Abstract

Population genomic approaches, which take advantages of high-throughput genotyping, are powerful yet costly methods to scan for selective sweeps. DNA-pooling strategies have been widely used for association studies because it is a cost-effective alternative to large-scale individual genotyping. Here, we performed an SNP-MaP (single nucleotide polymorphism microarrays and pooling) analysis using samples from Eurasia to evaluate the efficiency of pooling strategy in genome-wide scans for selection. By conducting simulations of allelotype data, we first demonstrated that the boxplot with average heterozygosity (HET) is a promising method to detect strong selective sweeps with a moderate level of pooling error. Based on this, we used a sliding window analysis of HET to detect the large contiguous regions (LCRs) putatively under selective sweeps from Eurasia datasets. This survey identified 63 LCRs in a European population. These signals were further supported by the integrated haplotype score (iHS) test using HapMap II data. We also confirmed the European-specific signatures of positive selection from several previously identified genes (*KEL*, *TRPV5*, *TRPV6*, *EPHB6*). In summary, our results not only revealed the high credibility of SNP-MaP strategy in scanning for selective sweeps, but also provided an insight into the population differentiation.

Key words: selective sweep, SNP-MaP, boxplot

Introduction

Natural selection, which influences the patterns of genetic variation and phenotypic diversity in various ways, is the driving force of Darwinian evolution (*1, 2*). Identification of genomic regions under natural selection can improve our understanding of human evolutionary and phenotypic diversity as well as facilitate the detection of functional regions of genome. With the rapid advances in genome-wide genotyping, "population genomics" has been applied to systematic scan for selection signals (*3-10*). Population genomics, with the empirical distributions of a given test statistic (such as F_{ST}, TD_{GEN}, and *LRH*), provides a practical strategy for the precise and accurate identification of

[#] Equal contribution.
*Corresponding author.
E-mail: zhangdk@big.ac.cn

candidates subject to selection. Although several surveys have presented the initial maps of selection using large-scale variation datasets, high cost of genome-wide genotyping limits the widespread application of population genomics in the study of human evolution.

Single nucleotide polymorphism microarrays and pooling (SNP-MaP), an approach broadly applied in gene mapping of complex traits, is a cost-efficient alternative to individual genotyping (IG) in large samples (*11*). This method combines the strength of microarrays to genotype large numbers of SNPs and the strength of DNA pooling to genotype large samples by genotyping pooled DNA on SNP microarrays. Although the pooling variance is the major concern in practical application of SNP-MaP, pooled DNA can be genotyped reliably when a sufficient number of replication arrays are used (*12-14*).

In this report, we first applied SNP-MaP strategy to discover evidence of selective sweep in datasets of the allelotyped SNPs from European samples. We also estimated the performance of the averaged heterozygosity (HET) in detecting selective sweeps from allelotype datasets by simulations with varying levels of pooling error. Then, we further verified the feasibility and accuracy of this SNP-MaP strategy by comparing the allelotype data, genotype data and corresponding coalescent simulation data.

Results

Efficiency of outlier approach in datasets with and without pooling error

Positive selection is expected to affect the site frequency spectrum of the region with the favorable alleles. We hypothesized that regions subject to selective sweeps could be identified as outliers with an extreme low averaged HET by comparing to the empirical distribution of genome. To test this hypothesis, we simulated datasets consisting of 1,000 genomic regions with varying strengths of selection using the program SelSim. Then, we selected a set of SNPs with an SNP density comparable to that of Illumina HumanHap300, and calculated the averaged HET for each 250-kb region. As expected, positive selections led a marked shift of candidate regions toward lower diversity levels (**Figure 1A**). Further statistic test (Mann-Whitney U test, MWU) showed considerable difference among the averaged HETs of selection and neutral datasets ($P<10^{-10}$).

Furthermore, we mimicked the selection and neutral datasets of SNP-MaP with different pooling errors [pooling standard deviation (PSD)=0.08, 0.04, 0.02, and 0.01, respectively]. These simulations demonstrated that the ability of the averaged HETs to detect

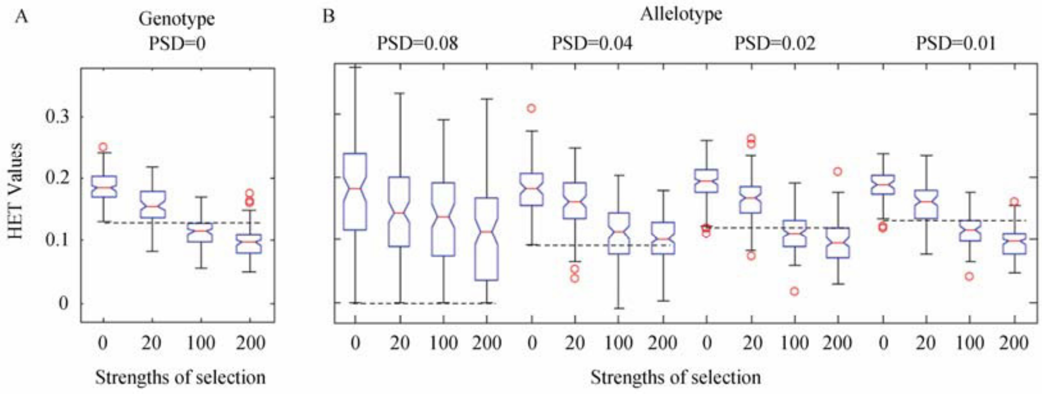

Figure 1 Summary of the averaged HETs in simulated datasets with varying values of pooling errors. The x-axis denotes the values of selection strengths in simulations. The boxplot method shows the most extreme values of averaged HETs (maximum and minimum values), the lower and upper quartiles, and the median in IG (**A**) and SNP-MaP datasets (**B**). The median for each dataset is indicated by the centerline, and the first and third quartiles are represented by the edges of the area, which is known as the inter-quartile range (IQR). The extreme values (within 1.5 times of the IQR from the upper or lower quartile) are represented by the ends of the lines extending from the IQR. Points at a greater distance from the median than 1.5 times of the IQR are plotted individually as circles.

selective sweep increased with the decrease of the pooling error in the dataset (**Figure 1B**). When the pooling error is moderate (PSD=0.02 or 0.01), the boxplot approach resulted in an enrichment of the regions under positive selection. Consistent with the results of simulations without pooling error, boxplot identified most candidate regions (>75%) as outlier when the strength of selection was large than 100 and the context of PSD was less than 0.02. These results suggested that the boxplot with averaged HET could detect strong selective sweeps (s≥100) in SNP-MaP datasets with reasonable power.

Scanning for selection sweeps using allelotype dataset from the Europeans

To evaluate the feasibility of the outlier approach in real SNP-MaP data, we applied a sliding window analysis to allelotype and genotype data of HapMap CEU samples (60 unrelated individuals). We calculated the averaged HETs of 250-kb window across autosomes, and compared the averaged HETs estimated from SNP-MaP and IG data. Although the mean value for the differences of averaged HETs was larger than zero (0.050; t-test, $P<10^{-35}$), a strong correlation between two datasets was observed (R^2=0.895, **Figure 2**). Furthermore, the pooling error estimated from windows containing at least 20 SNPs was less than 0.01 (PSD=0.008), which also suggested that SNP-MaP data could be used to scan for the signal of positive selection.

As a result, we applied 262,837 windows for further analysis, each of which contained at least 20 SNPs. The empirical distribution of the averaged HETs significantly departed from the normal distribution because of the asymmetry tails at low values (**Figure 3A**, Jarque-Bera test, $P<10^{-10}$). In addition, we compared the empirical distribution of allelotype data with those of the best-fitting demographic model (*15*). The quartile-quartile (QQ) plot suggested that CEU allelotype dataset has departed from the neutral assumptions, most likely due to the positive selection (Figure 1B).

To reveal the signals of positive selection, we applied the boxplot method to calculate chromosome-specific thresholds for the outlier detection, and identified a large proportion of low averaged HET windows (4,671, 1.78%). These widows frequently occurred in clusters, being defined as two or more contiguous outliers. In total, there were 467 unique clusters, encompassing almost all outliers (4,540/ 4,671). In particular, although no cluster contained more than 20 outliers in the simulated dataset, we did identify 63 large contiguous regions (LCRs) in the CEU samples (Table S1), as described in Materials and Methods. One potential explanation for clustering of

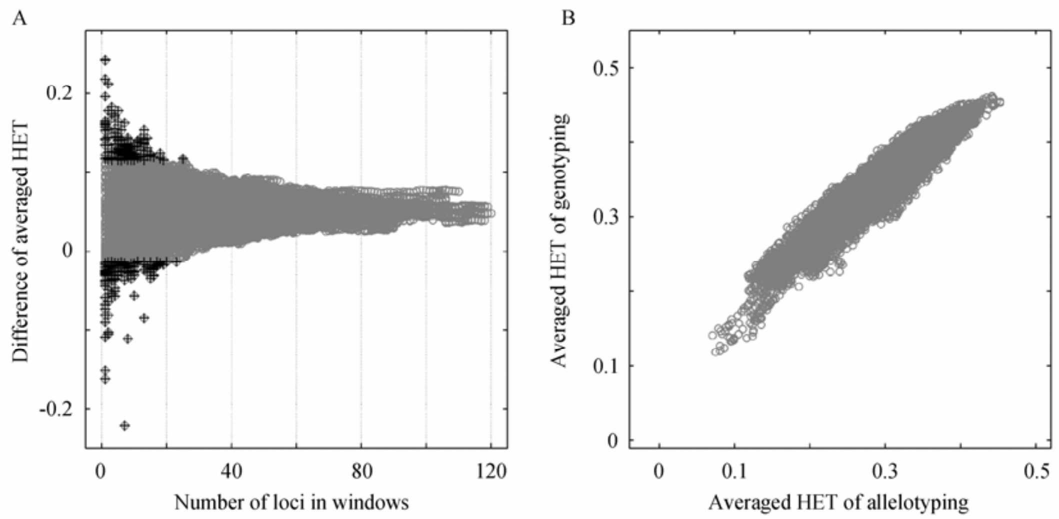

Figure 2 Comparison of the averaged HETs of 250-kb window between allelotype data and IG data from HapMap CEU samples. **A**. The difference of the averaged HET between allelotype and IG datasets *vs.* the number of loci located in the corresponding window. **B**. Comparison of the averaged HETs for the genotype (x-axis) and the allelotype datasets (y-axis). The black crosses represent the 500 top-ranked SNPs with the largest difference of the averaged HETs.

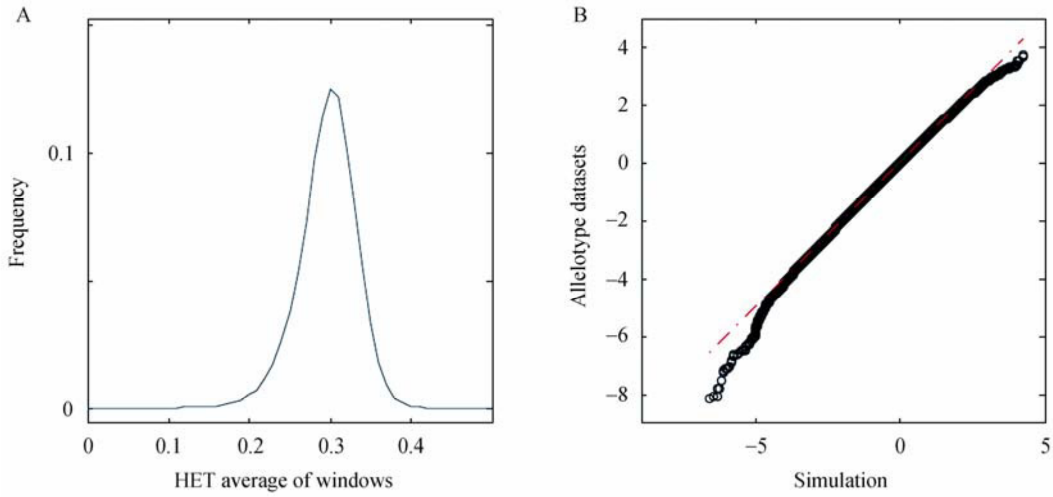

Figure 3 Genome-wide empirical distribution of the averaged HETs from CEU samples and its comparison with that of neutral simulation data. **A**. The genome-wide empirical distribution of the 250-kb window-averaged HETs departed from a normal distribution. **B**. Comparison of the distributions between empirical data and neutral simulation data performed by QQ plot. Quartile points of the neutral simulation data were plotted on the x-axis in the QQ plot.

candidate windows is genetic hitchhiking/selective sweep (*i.e.*, the effect of positive selection on linked neutral variation). This assumption was further supported by the integrated haplotype score (iHS) test using the HapMap II dataset. Large proportion of LCRs (47, 74.6%) contained a significant excess of SNPs with |iHS|>2 (Fisher's exact test, *P*<0.01), which suggested that most of the selection events detected by our approach occurred recently.

Discussion

As suggested by previous research, the magnitude of the PSD is an important factor affecting the efficiency of pooling compared with IG (*13*). Our results showed that the boxplot analysis could help to detect strong selective sweeps (s≥100) with a reasonable sensitivity when PSD was close to 0.01. In this study, the PSD of HET values was estimated using both pooling and IG data from CEU samples. The PSD of single-locus HET was 0.06, and this value could be reduced by combining data from several adjacent markers. Our result showed that the PSD of the averaged HET dropped substantially with the increase of window sizes; therefore, a 250-kb window (PSD=0.012) was chosen to scan for candidates under selective sweep.

The efficiency of the boxplot strategy with averaged HET was also validated by the empirical distributions of CEU population. Not only a high proportion of outliers but also large candidate regions were detected in the allelotype dataset than in its neutral simulation data (t-test, *P*<0.01). Most of these regions were further identified as candidates under recent positive selection supported by HapMap II dataset (Fisher's exact test, chi-square test, *P*<0.01) (*9*).

Identifying candidate regions in the human genome can provide important clues for genotype/phenotype research. For example, *TRPV6* and *TRPV5* play an important role in the route of dietary calcium uptake, lying within an LCR in allelotype datasets. This cluster of genes across chromosome 7q34-35 (*KEL*, *TRPV5*, *TRPV6*, and *EPHB6*) also exhibited complex signals of natural selection in previous studies (*9*). Our study provides a comprehensive assessment of scanning for selection sweeps using SNP-MaP data, and demonstrates the feasibility and replication of the boxplot approach in detecting large genomic regions subject to natural selection using different population samples. On the other hand, although our map gave insights into the functionally important polymorphisms, several well-known genes under selection in Europe were not identified in candidate regions due to lack of enough genomic coverage. For example, the *LCT* gene was located within "blank" regions with the extreme low density of allelotyping SNPs. These results suggest that the power of our analyses is limited by the coverage of genomic scans, and that we may need to increase the SNP number in the future population genomic research.

Materials and Methods

Data collection

Publicly available allelotype data of Caucasians of European descent were downloaded from Sebastiani *et al* (*16*). Eventually, 303,894 autosomal SNPs were obtained from this dataset. The genotype data of the corresponding markers were downloaded from the HapMap website (HapMap Phase II/rel#21a; http://www.hapmap.org/) (*8*). In addition, the genome-wide iHS for the HapMap Phase II were downloaded from the Pritchard Lab website (http://haplotter.uchicago.edu/) (*9*), and then mapped to the corresponding locations on NCBI build 35.

Population genetic analysis

HET was applied to measure the levels of the nucleotide diversity (*17*). For the sliding-window analysis, the averaged HET of all segregate loci for certain window was calculated. A series of 250-kb sliding windows, with a step of 10 kb, were set to analyze across all autosomal regions. That is, the first window evaluated on chromosome 1 was genome coordinates chr. 1, 1-250,000; the second window was genome coordinates chr. 1, 10,001-260,000; and so on. In addition, windows containing less than 20 polymorphic loci were excluded from our analyses to eliminate the biases caused by the differences in SNP number among windows. Based on the empirically distribution of the number of SNPs, 262,837 sliding windows were obtained in the Europe dataset.

To identify outlier regions or windows with considerable low nucleotide diversity, we firstly determined the IQR (*18*) based on empirical distributions across the particular chromosome:

$$IQR = F^U - F^L \qquad (1)$$

where F^L is lower quartile values and F^U is upper quartile values. Then the threshold values (*UL* and *LL*) were defined as within 1.5 times of the IQR from the upper or lower quartile:

$$UL = F^U + 1.5 \times IQR \qquad (2)$$

$$LL = F^L - 1.5 \times IQR \qquad (3)$$

Low averaged HET values at a greater distance from the median than 1.5 times of the IQR are plotted individually representing potential outliers.

The empirically determined distribution of averaged HET within the sliding windows was used to identify LCRs, defined as a region of more than 20 contiguous windows, where over 75% of the windows were outliers.

Coalescent simulations

To obtain genomic regions under selective sweeps, coalescent simulations were performed using the program SelSim (*19*). We assumed that the advantageous mutation follows a stochastic trajectory, and set the strength of selection as the population selection coefficient (δ=2$N_e s$=0, 20, 100, and 200). Other parameters of the simulations were the population mutation rate (4$N_e\mu$=100) and the population recombination rate (4$N_e r$=100). In these formulas N_e denotes the effective population size, *s* denotes the selective advantage of the beneficial mutation per copy per generation, μ denotes the mutation rate per site per generation, and *r* denotes the recombination rate per site per generation. We set an effective population size of 10^4, a mutation rate of 10^{-8} per base per generation, and a recombination rate between base pairs of 10^{-8} per base per generation. For each simulation, 120 chromosomes were generated.

To obtain mimicked data of the allelotyped SNPs for the CEU samples, coalescent simulations of neutral evolution with the best-fitting demographic model of Schaffner *et al* was firstly performed using the program COSI (*15*). In this simulation program, recombination rate varied within the simulated regions (5 Mb). The simulation runs were replicated 600 times and all the simulated regions were connected to make chromosomes with the length of ~3 Gb. Then, in order to imitate the SNP selection in Illumina HumanHap300K, we determined the probabilities that SNPs were "allelotyped" according to the minor allele frequency spectra in the simulation and the CEU data from Illumina HumanHap300K chip. Finally, we obtained a set of SNP data from the simulation, and constructed allelotyped data from the haploid data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 30871348 and 30700470), Educational Department of Jiangxi Province (No. GJJ10303) and National Key Laboratory Specific Fund (No. 2060204).

Authors' contributions

LD and XT performed major data analyses and drafted the manuscript. WC, JL, Z Lai and Z Liu collected the dataset and performed part of data analyses. LD and DZ designed the study, supervised the project and co-wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Ronald, J. and Akey, J.M. 2005. Genome-wide scans for loci under selection in humans. *Hum. Genomics* 2: 113-125.

2 Sabeti, P.C., *et al.* 2006. Positive natural selection in the human lineage. *Science* 312: 1614-1620.

3 Akey, J.M., *et al.* 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12: 1805-1814.

4 Carlson, C.S., *et al.* 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15: 1553-1565.

5 Sabeti, P.C., *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913-918.

6 Tang, K., *et al.* 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5: e171.

7 International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437: 1299-1320.

8 International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

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

10 Wang, E.T., *et al.* 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci. USA* 103: 135-140.

11 Davis, O.S., *et al.* 2009. The SNPMaP package for R: a framework for genome-wide association using DNA pooling on microarrays. *Bioinformatics* 25: 281-283.

12 Macgregor, S. 2007. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *Eur. J. Hum. Genet.* 15: 501-504.

13 Macgregor, S., *et al.* 2008. Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Res.* 36: e35.

14 Melquist, S., *et al.* 2007. Identification of a novel risk locus for progressive supranuclear palsy by a pooled genomewide scan of 500,288 single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 80: 769-778.

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

16 Sebastiani, P., *et al.* 2008. A hierarchical and modular approach to the discovery of robust associations in genome-wide association studies from pooled DNA samples. *BMC Genet.* 9: 6.

17 Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York, USA.

18 McGill, R., *et al.* 1978. Variations of Box Plots. *The American Statistician* 32: 12-16.

19 Spencer, C.C. and Coop, G. 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* 20: 3673-3675.

Supplementary Material

Table S1

DOI: 10.1016/S1672-0229(10)60027-7