BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 6 2011, pages 822–828
doi:10.1093/bioinformatics/btr007

Genetics and population analysis Advance Access publication January 6, 2011

A method for identifying haplotypes carrying the causative allele in positive natural selection and genome-wide association studies

Rick Twee-Hee Ong[1,2,*], Xuanyao Liu[1], Wan-Ting Poh[3], Xueling Sim[2], Kee-Seng Chia[2,3] and Yik-Ying Teo[1,3,4,5,*]

[1]NUS Graduate School for Integrative Science and Engineering, [2]Centre for Molecular Epidemiology, [3]Department of Epidemiology and Public Health, [4]Department of Statistics and Applied Probability, National University of Singapore and [5]Genome Institute of Singapore, Singapore

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Methods for detecting positive selection relied on finding evidence of long haplotypes to identify candidate regions under selection. However, these methods generally do not identify the length and form of the selected haplotype.

Results: We present HapFinder, a method which can find the common longest haplotype under three different settings from a database, which is relevant in the analysis of positive selection in population genetics and also in medical genetics for finding the likely haplotype form carrying the causal allele at the functional polymorphism.

Availability: A java program, implementing the methods described in HapFinder, together with *R* scripts and datasets for producing the figures presented in this article are publicly available at http://www.nus-cme.org.sg/sgvp/software/hapfinder.html. The site also hosts an online browser for finding haplotypes from the International HapMap Project and the Singapore Genome Variation Project.

Contact: g0801900@nus.edu.sg; statyy@nus.edu.sg

Received on September 30, 2010; revised on December 21, 2010; accepted on December 22, 2010

1 INTRODUCTION

Haplotypes refer to the specific combinations of alleles at different locations on a chromosome. Diploid organisms such as humans carry two copies of chromosomes, and thus two haplotypes are present for each individual when considering the chromosomal arrangement of the alleles at several variant sites, such as single nucleotide polymorphisms (SNPs). At the fundamental level, the haplotype carries most of the genetic information, particularly for the assessment of allelic correlation in the genome and in situations where the exact sequence arrangement on the chromosome is important. However, the technical ease of assaying a single base position in the genome means it is more common to obtain the aggregate information across the two chromosomes of an individual, also known as the genotype. The advent of affordable large-scale genotyping technologies has permitted the genotypes of up to a million positions across the human genome to be assayed simultaneously, facilitating the studies on the genetic etiology of common diseases and complex traits across thousands of samples (Donnelly, 2008; McCarthy *et al*., 2008). When only genotype information is available, resolving the exact arrangements of the alleles on the two haplotypes for an individual requires sophisticated statistical machinery in a process known as *haplotype phasing*. A number of statistical procedures have been formulated for inferring the haplotype phases from genotype data, such as PHASE (Stephens and Scheet, 2005), fastPHASE (Scheet and Stephens, 2006) and Beagle (Browning and Browning, 2007). Public databases like the International HapMap Project (Consortium, 2007, 2010), Human Genome Diversity Project (Jakobsson *et al*., 2008) and Singapore Genome Variation Project (Teo *et al*., 2009) have generated reference genotypes for a considerable number of populations globally, with statistically inferred haplotypes also available for a number of these populations.

Generally, the lengths of common shared ancestral segments of chromosomes in a population are short since recombination acts over time to break down long haplotypes. An exception is in genomic regions experiencing positive evolutionary pressure of natural selection (Sabeti *et al*., 2002), where greater fitness in survival and procreation results in a higher propensity that offsprings in subsequent generations will increasingly carry the advantageous mutations. This can increase the frequency of an advantageous allele and, due to the hitch-hiking effects of neighbouring alleles and insufficient time for recombination to occur, result in haplotypes that are uncharacteristically long for a given haplotype frequency. A number of sophisticated statistical methods, for example, XP-EHH (Sabeti *et al*., 2007) and iHS (Voight *et al*., 2006), have relied on finding such genomic signatures of long haplotypes for identifying candidate regions that are experiencing positive selection. However, these methods generally do not identify the length and form of the selected haplotype.

The presence of linkage disequilibrium (LD) in the human genome implies that there will be numerous SNPs that are correlated with each other. In large-scale genotype–phenotype association studies, the discovery of a trait-associated region is often accompanied by a list of neighbouring SNPs displaying similar degrees of statistical evidence. Due to the nature of LD, most, if not all of the identified alleles carried on these SNPs are likely to be found on a haplotype that also carries the functional allele at the unknown causal variant. This would be useful for narrowing the likely candidate region where the functional variant may be located, particularly when evidence from multiple

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025

genetically diverse populations is available. We can then localize the candidate region to genomic segments where these population-specific implicated haplotypes are consistent across the various populations.

Here, we introduce a novel method for finding haplotypes in three scenarios from a haplotype database: (i) identifying the longest haplotype for a user-defined core haplotype frequency that is carrying a specific allele at a focal SNP; (ii) around a focal position with a given core haplotype frequency; (iii) matching a specific combination of alleles from a set of user-defined SNPs. Type 1 is particularly useful when the functional allele at a causal or positively selected SNP is known a priori. Type 2 is relevant when only the approximate genomic region of the functional variant is known, such as the situation where there is preliminary evidence of positive natural selection in the region from iHS or XP-EHH without explicit knowledge of the exact focal SNP and causal allele. Type 3 can be used to find the haplotype form that carries most, if not all, of the implicated alleles associated with disease onset or increased severity at SNPs that are identified from genome-wide association studies (GWAS).

In Section 2, we will describe in details how the method works in the three settings. Section 3 demonstrates the utility of HapFinder on known positively selected genomic regions in diverse population groups from the HapMap, and also via simulated case–control studies. Finally, in Section 4, we discuss how the method will be useful in both population and medical genetics studies.

2 METHODS

All three applications of HapFinder require phased haplotype data in the format of a $N \times L$ matrix which we denote as H, where each row represents a phased haplotype chromosome of an individual and each column represents a unique biallelic SNP. Thus, $N = 2n$ with n denoting the number of individuals in the dataset, since humans are diploid and each individual possesses two chromosomes. Let h_{il} denote the (i, l) entry of the matrix H, where $h_{il} \in \{0, 1\}$, representing the two possible alleles for each SNP. Note that we assume there is no missing allele information for any haplotype after the phasing procedure. As HapFinder searches for haplotypes carrying specific alleles, it is important to define accurately what the '0' and '1'alleles for each h_{il} map to. In all our example applications, we have assumed the alleles are mapped to the positive strand while following the definitions of the '0' and '1' alleles according to Phase 2 of the HapMap as encoded in the legend files. A schematic overview of the three applications in HapFinder can be seen in Figure 1.

2.1 Algorithm for Type 1

In searching for the longest haplotype that is at a user-specified core haplotype frequency f in the haplotype database and is specifically carrying a particular target allele at the focal SNP, we first determine the critical number of chromosomes $c = \text{floor}[f \times N]$. The algorithm first assesses whether the allele frequency of the target allele is at least f, and returns an error message when the target allele frequency is less than f. When the number of chromosomes carrying the target allele is at least c, the SNP on the immediate left of the focal SNP is appended to the haplotype form. This means there are at most two possible haplotypes for these two SNPs that carry the target allele at the focal SNP, if we assume the neighbouring SNP is either monomorphic or biallelic. When the number of chromosomes carrying the more common haplotype is at least c, the next SNP on the left is appended. The algorithm iterates between adding another SNP on the left and checking whether the number of chromosomes carrying the most common haplotype is at least c. When the number of chromosomes carrying the most common haplotype falls below c, the most recently appended SNP is removed from the haplotype, and the SNP to the immediate right of the focal SNP is appended. The algorithm now proceeds to append SNPs on the right until the number of chromosomes carrying the most common haplotype is less than c, where the most recently appended SNP on the right is then removed. The longest haplotype spanned is then returned as the output.

2.2 Algorithm for Type 2

The algorithm for Type 2 is similar to that for Type 1, except that a focal position is specified instead of a focal SNP and there is no pre-determined target allele. In this instance, the SNP on the chromosome that is closest to the specified focal position is chosen as the focal SNP, while either of the two alleles is allowed to be the target allele. The algorithm proceeds according to the procedure for Type 1, effectively running two separate operations for each of the two target alleles and identifying the longest haplotype form out of the two analyses. In this instance, an error message is produced when the user-specified core haplotype frequency f is larger than the major allele frequency.

2.3 Fuzzy matching in Type 1 and Type 2

Large-scale genotyping inevitably introduces errors in the called genotypes that propagate downstream to the haplotype phasing, thus generating phased haplotypes that are more likely to be dissimilar at genomic sites affected by genotyping errors. To allow for such spurious errors in the phased haplotypes, we permit a small proportion of mismatches when counting the number of chromosomes carrying the most common haplotype. As before, the most common haplotype form is first identified, and the similarity score between this haplotype form and each of the N chromosomes is calculated. The similarity score between two haplotypes is calculated as the proportion of SNPs where the alleles are identical across the two haplotypes. When the similarity score between the most common haplotype form and a sample chromosome is greater than the user-specified threshold $s*$, the sample chromosome is considered to be sufficiently similar to the most common haplotype. Thus, instead of counting the number of chromosomes carrying the most common haplotype, fuzzy matching with $s* < 1$ counts the number of chromosomes that are similar to the most common haplotype.

2.4 Algorithm for Type 3

Type 3 of HapFinder allows multiple focal SNPs with corresponding target alleles to be specified, and the algorithm aims to identify the haplotype forms that carry most, if not all, the target alleles. The importance of matching the allele at each of the K target SNPs is defined by either the SNP weightings or the genetic distances between the SNPs, or as a composite function of both. When attempting to identify the haplotype that is carrying the high-risk alleles at implicated SNPs, the weightings $(q_1, q_2, \ldots, q_K)$ can be the statistical evidence of phenotypic association (e.g. $-\log_{10}P$-values or log Bayes factors). This effectively prioritizes the matching of the target alleles on SNPs with strong evidence of phenotypic association, and the SNP possessing the strongest weighting is defined as the central SNP. When SNP weightings are not provided, the genetic distances of the SNPs $(d_1, d_2, \ldots, d_K)$ must then be provided and the focal position is defined as the average of the genetic distances of the first and last focal SNP d_{centre}. The contribution of each SNP is thus defined as

$$W_k = \begin{cases} \frac{q_k}{\sum_l q_l}, & \text{when only SNP weightings are provided} \\ \frac{\exp(-|d_k - d_{\text{centre}}|)}{\sum_l \exp(-|d_l - d_{\text{centre}}|)}, & \text{when only genetic distances are provided} \\ 0.5\left(\frac{q_k}{\sum_i q_l} + \frac{\exp(-|d_k - d_{\text{centre}}|)}{\sum_l \exp(-|d_l - d_{\text{centre}}|)}\right), & \text{when both SNP weightings and genetic distances are provided} \end{cases}$$

with the summation notations in the expressions performed over the set of K SNPs. This allows a match score (between zero and one inclusively) to

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025

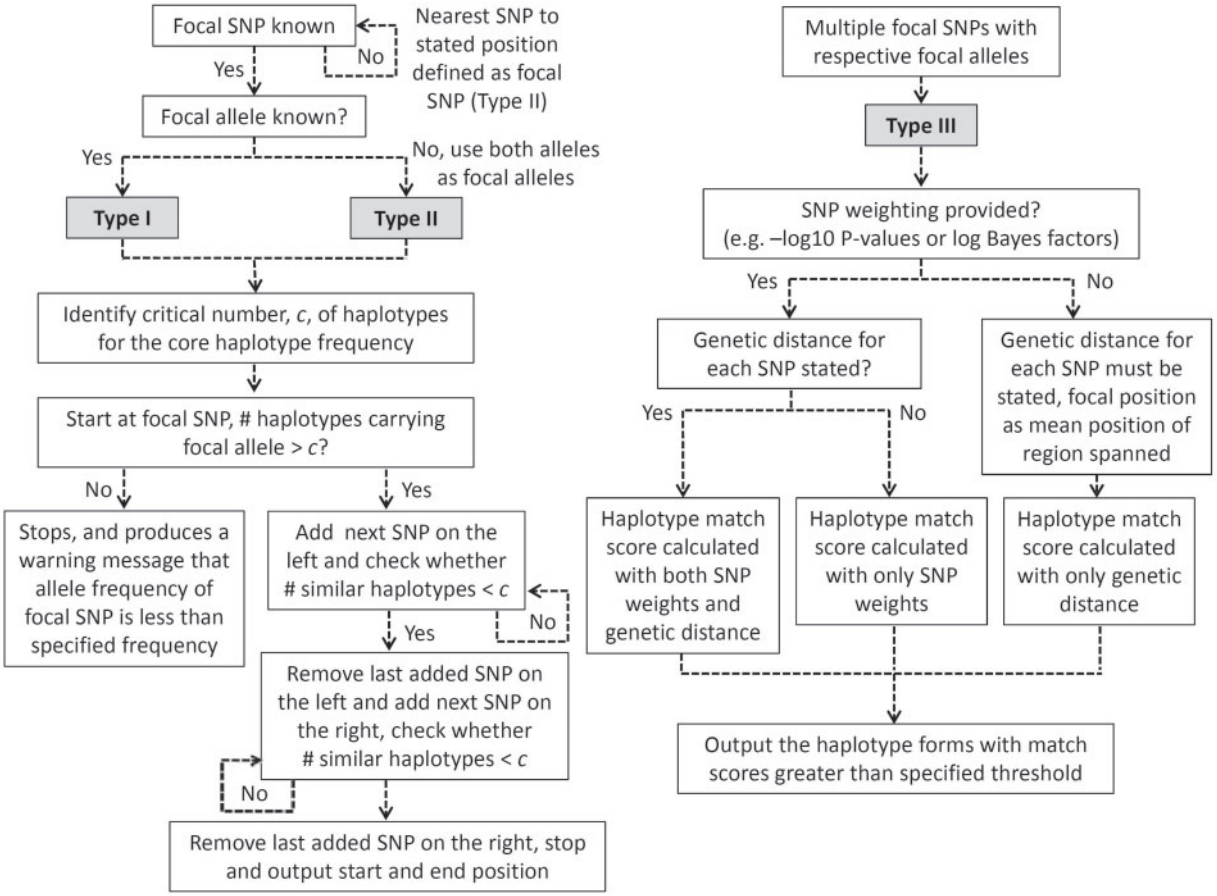

Fig. 1. Schematic overview of the algorithm behind the three applications of HapFinder.

be calculated for each chromosome in the database, with the match score for the i-th chromosome defined as the sum of W_k over the set of focal SNPs where the chromosome carries the exact target alleles. HapFinder locates the chromosomes with scores above a user-specified threshold: when the threshold is one, only the haplotype forms which carry all the target alleles at the focal SNPs are identified; when the threshold is less than one, mismatches with the target alleles for some focal SNPs may be permissible as long as the match score for the chromosome is greater than the threshold.

2.5 Software availability and output

A JAVA program for HapFinder is freely available for download from http://www.nus-cme.org.sg/sgvp/software/hapfinder.html, along with scripts for producing graphical displays in *R*. An online web application is also available for finding haplotypes in the populations in Phase 2 of the International HapMap Project (Consortium, 2007) and in the Singapore Genome Variation Project (Teo *et al*., 2009). Each analysis generates four files: (i) a .haps file that contains the identified haplotypes in 0/1 format; (ii) a .legend file containing the rs identifiers, coordinates and the 0/1 allele maps for the SNPs on the identified haplotypes; (iii) a .sample file that indicates which samples and which haplotype of the sample (suffixed by −1 and −2 to the sample ids) do the identified haplotypes correspond to; (iv) a .log file containing all the haplotype forms that are identified for that analysis (which is not restricted to only the longest haplotypes for Types 1 and 2). An additional fifth file is also output for Type 3: (v) a .snp file containing the rs identifier, coordinate, the allele that is tagging the identified haplotype and the LD measured in r^2 between each SNP and the identified haplotype.

3 APPLICATIONS AND RESULTS

To illustrate the utility of HapFinder, we applied the method for the three settings described on publicly available haplotype data from the International HapMap Project (Consortium 2007). All genomic coordinates quoted are on NCBI build 36.

The sickle cell allele (adenine allele) at rs334 on the p15.5 arm of chromosome 11 has been well established to be under balancing selection in the Yoruba population of HapMap (YRI) (Feng *et al*., 2004; Hedrick, 2004), conferring up to 10-fold protection against malaria (Ackerman *et al*., 2005; Hill *et al*., 1991; Jallow *et al*., 2009) while providing a recessive Mendelian risk of sickle cell anaemia. Due to positive selection, the sickle cell allele is expected to reside on an uncharacteristically long haplotype compared with other alleles in the genome at the same frequency. We test this hypothesis by running HapFinder Type 1 on phased haplotypes of Hapmap Phase 2 YRI data, which included a direct assay of rs334 (Fig. 2). By specifying rs334 as the focal SNP and performing two Type 1 analyses on the wild-type allele (thymine, allele T) and the sickle allele (allele A) at a core haplotype frequency of 10%, HapFinder identified a

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025

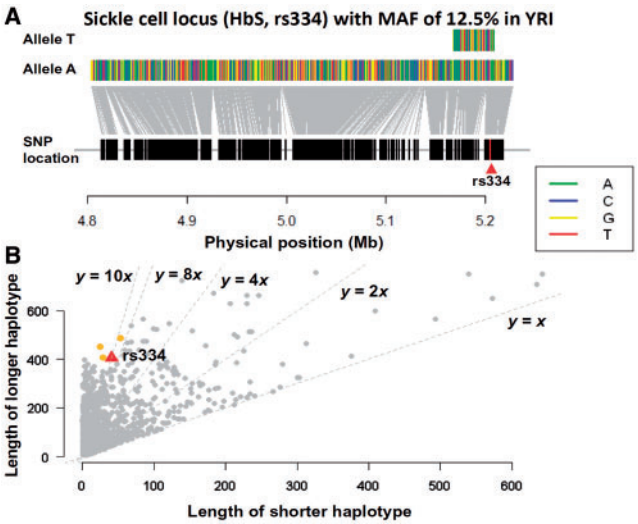

Fig. 2. (**A**) An example application of the Type 1 application of HapFinder for finding the longest haplotypes carrying the two alleles at the sickle cell locus (rs334) in YRI, which has a MAF of 12.5%. The alleles on each haplotype have been coloured accordingly for each of the four possible bases (A, green; C, blue; G, yellow; T, red). The vertical black lines on the horizontal grey bar represents where each of the SNPs is located, with the vertical red line indicating the focal SNP. (**B**) Comparing the lengths of the two haplotypes identified by Type 1 HapFinder for each of the 1000 randomly chosen SNPs with MAFs of 12.5% in YRI, where the red triangle represents rs334. The three random SNPs that also display the largest extent of differences between lengths of haplotypes are shaded in orange. The grey dashed diagonal lines represent the boundaries for the various sizes of the ratio of the haplotype lengths. All examples were run with parameters core haplotype freq, $f = 0.10$, and similarity score, $s* = 0.98$.

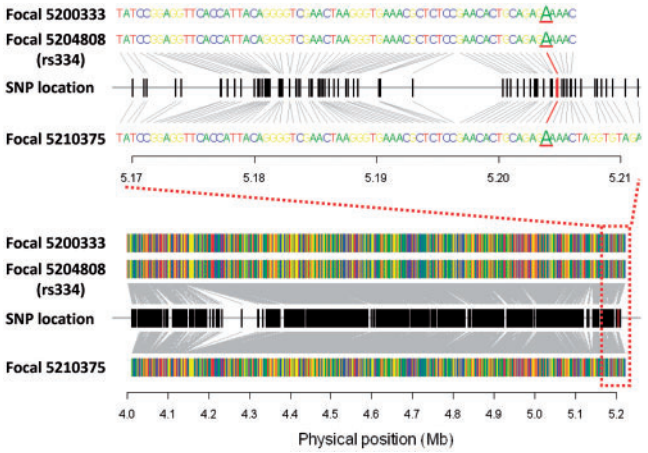

Fig. 3. Visual representations of the haplotypes identified by the Type 2 application in HapFinder, searching at three separate locations around the sickle cell locus (rs334, represented as the vertical red line in the horizontal bars indicating the SNP location) at a core haplotype frequency of 5%. The alleles on each haplotype have been coloured accordingly for each of the four possible bases (A, green; C, blue; G, yellow; T, red). The outcome from the analysis with the focal position of 5 204 808 corresponds to searching for the longest haplotype with rs334 as the focal SNP. The bottom panel shows that the three identified haplotypes are almost identical regardless of the focal position used in the Hapfinder analysis, while the top panel zooms into a 40 kb window around rs334 and illustrates that all three haplotypes carry the A allele at rs334 that is experiencing positive selection.

haplotype form carrying the sickle allele that spans nearly 0.4 Mb, while the haplotype form carrying the wild-type T allele spans <40 kb (Fig. 2A). By comparison, there were only 3 SNPs out of 1000 randomly selected SNPs [each with minor allele frequency (MAF) of 12.5% in YRI] that displayed similar extent of differences between the lengths of the shorter and longer haplotypes (Fig. 2B).

The sickle cell locus provides a convenient example where the functional polymorphism that is experiencing evolutionary pressure of positive selection is actually known. In practice, imperfect SNP coverage in genetic databases like the HapMap means recent discoveries on candidate regions undergoing positive selection generally do not identify the selected functional polymorphisms. Instead, discoveries using selection metrics like iHS and XP-EHH tend to highlight candidate regions that are displaying putative evidence of selection. In such situations where only the broad region is known, without specific knowledge of the focal SNP or the selected allele, we can rely on Type 2 of HapFinder to identify the likely haplotype where the unknown functional allele sits on. We use the sickle cell example to show that similar haplotype forms are identified whether the analysis is performed at the functional polymorphism with the known selected allele (at position 5 204 808 on chromosome 11), or when the analysis is performed at neighbouring focal positions of 5 200 000 and 5 210 000 (Fig. 3). This is directly relevant for extracting the selected haplotype underpinning a candidate signature of positive selection identified by iHS or XP-EHH.

By performing multiple iterations of the Type 2 analysis at the same genomic site but across different core haplotype frequencies, it may be possible to identify a range of frequencies that the frequency of the selected allele may be found in. As the specified core haplotype frequency changes, the identified haplotypes can switch between carrying the non-selected allele to the selected allele, resulting in a significant increase in the length of the identified haplotype (Fig. 4). For example, we performed three separate sets of Type 2 analyses in the *HBB* gene region in YRI, with focal positions specified as the position of rs334 (5 204 808 bp), and two neighbouring positions (5 200 000 bp) and (5 210 000 bp), respectively. We iterate through core haplotype frequencies of 20 to 5% in step-size reduction by 1%, observing significant increases in the haplotype lengths at core haplotype frequencies of 12, 11 and 9% (Fig. 4A, top panel). By measuring the proportion of discordant sites at the unambiguous overlapping regions between the haplotypes identified at two consecutive frequencies, we observed that the discordance was particularly large at the frequency where there was a substantial increase in the spanned distance (Fig. 4A, bottom panel). A large value for discordance typically indicates that different haplotypes have been identified between the two consecutive analyses, and this likely reflects the switch to the selected haplotype that is carrying the advantageous allele. We also performed similar analyses at the *LARGE* gene (Fig. 4B) that has been previously reported to be positively selected for protection against lassa fever in North Central Africa (Sabeti *et al.*, 2007). The putatively selected allele at the SNP

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025

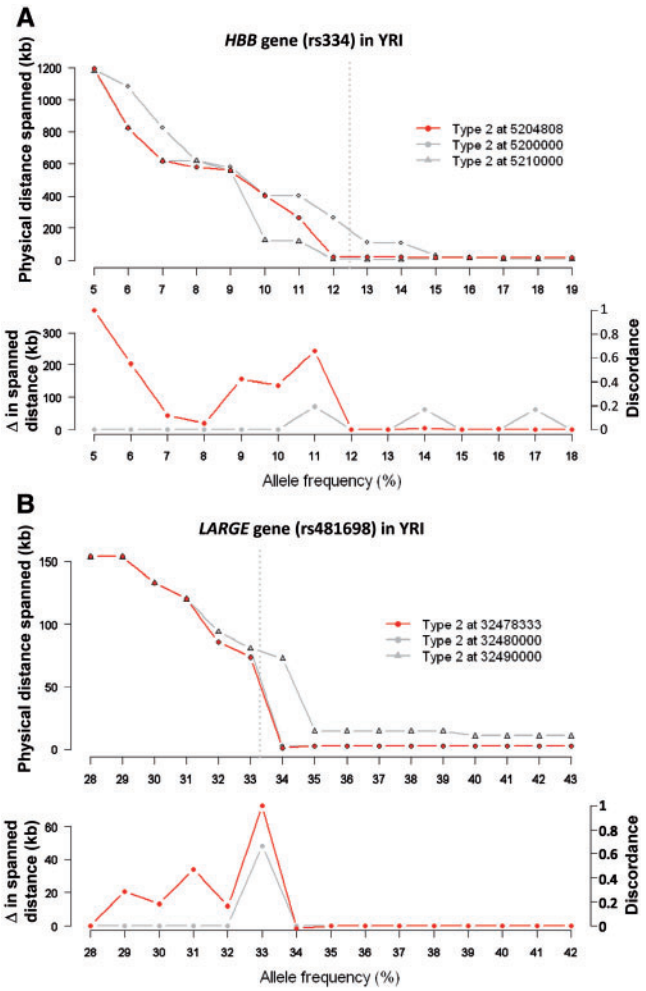

Fig. 4. Two example applications of Type 2 in HapFinder for identifying the longest haplotypes by specifying a focal position, rather than a focal SNP. For each of the two genomic regions, we selected the physical position of the known/putative SNP undergoing positive natural selection (lines and circles in red in top panels) and two other physical positions in the neighbourhood (in grey) as focal positions, and recorded the length of the longest haplotype across a range of core haplotype frequencies. The bottom panel of each region shows the change in haplotype distances (red lines and circles) and the degree of haplotype discordance (grey lines and circles) with a 1% reduction in allele frequency for the known/putative SNP. (**A**) Looks at three positions in the locality of the *HBB* gene in YRI, with the frequency of the selected allele (at rs334) to be known at 12.5% in YRI; (**B**) looks at three positions in the locality of the *LARGE* gene in YRI, with the frequency of the non-synonymous substitution (rs4481698) to be known at 33.3% in YRI.

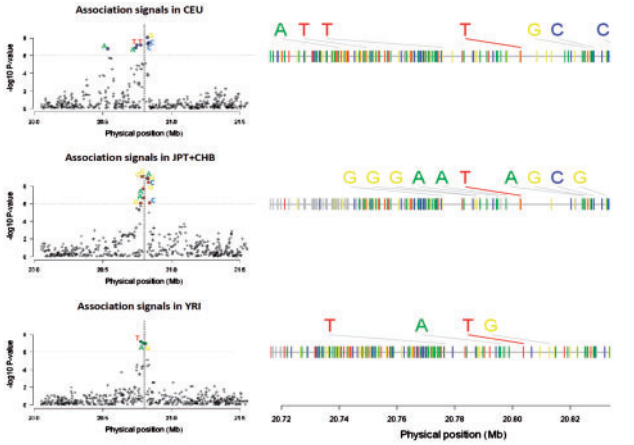

Fig. 5. Example application of Type 3 in HapFinder in identifying the haplotype form that is carrying the implicated risk alleles from the associated SNPs. The vertical dashed line in each of the three panels on the left represents the position of 'causal' SNP (Chr6:rs2206734), where 2000 cases and 2000 controls are simulated with HAPGEN in the three HapMap populations with a multiplicative effect of RR = 1.5 at allele T. Only SNPs on the Illumina 1M array are shown in the region plots, and SNPs with $P < 10^{-6}$ in each population are extracted as input SNPs for Hapfinder. The haplotype spanning a pre-specified start and end position, and is also carrying the implicated alleles at the input SNPs for each population is shown on the right. The alleles on each haplotype have been coloured accordingly for each of the four possible bases (A, green; C, blue; G, yellow; T, red). As there may be multiple haplotypes that carry the implicated alleles, SNPs with non-unique alleles on these haplotypes are represented as vertical grey lines. While different SNPs are identified in the association analyses for the three populations, all three identified haplotypes correctly carry the high-risk allele (allele T) at the simulated SNP.

(chr22: rs481698) with strong evidence of positive selection (via iHS) has a frequency of 33.3% in YRI, and our analyses observed a large haplotype discordance and significant increases in the length of the identified haplotypes when the frequency decreases from 34 to 33%.

To illustrate the application of Type 3 in HapFinder, we simulated 2000 cases and 2000 controls in each of the three HapMap panels from Phase 2 using HAPGEN (Spencer *et al.*, 2009) at an artificial causal SNP (rs2206734) located in the *CDKAL1* gene on chromosome 6, previously established to display significant variation in patterns of LD (Teo *et al.*, 2009). In each of the three simulations, we introduced a multiplicative effect size equivalent to an allelic relative risk of 1.5 at the T allele and limited the association analysis to only SNPs present on the Illumina1M array while masking the causal SNP (Fig. 5). Genetic markers displaying association $P < 10^{-6}$ are extracted as input SNPs to HapFinder, along with the corresponding alleles that are associated with higher risks and the corresponding observed *P*-values as SNP weightings (Refer to Fig. 1). We observed that different implicated SNPs emerged from each of the three simulations. In each population panel, we ran Type 3 of HapFinder to identify the haplotype forms that are carrying most of the high-risk alleles at the associated SNPs, subject to matching scores of at least 98%. While the association analyses discovered different implicated SNPs in the three population panels, the haplotype forms that are identified by HapFinder across all three populations correctly carried the high-risk allele T at the simulated causal SNP (Fig. 5). We also performed the same simulation experiment at 1000 randomly chosen SNPs that are present in all three HapMap population panels, in order to assess how often does the haplotype identified by HapFinder carry the risk allele at the simulated causal variant. Out of these 1000 simulations, we observed that 84.4% of the haplotype forms identified by HapFinder

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025

across all three populations carried the causal allele that is associated with a higher disease risk.

4 DISCUSSION

We have introduced a strategy for finding haplotypes under three scenarios that are relevant to the analysis of positive natural selection and GWAS. The classical example of the sickle cell locus was also used to highlight the utility of the method, both in finding the selected haplotype and at estimating the likely frequency of the selected allele. By simulating case–control data around an artificial causal SNP for each of the three HapMap panels, the method is able to identify the haplotype forms that carry the detrimental alleles at the associated SNPs that had emerged from the association analysis in each population separately. All three haplotype forms identified from the respective HapMap populations correctly carried the functional allele at the simulated causal variant. Further simulations indicate at least 80% power in identifying the haplotype form which the functional allele resides on.

Our method is able to identify the longest haplotype at a specified core frequency around a focal position. Theoretically, this can be applied across the genome at every available SNP and across a range of core haplotype frequencies. For a particular core frequency, one may expect the lengths of the identified haplotypes to be distributed within a specific range, and haplotypes that are uncharacteristically long for a particular core frequency could be the result of positive natural selection happening within those genomic regions. This may provide another strategy in searching for candidate signals of positive selection, although there is a need to account for local effects of LD that may potentially bias the analysis (work in progress).

The field of medical genetics is now focusing its attention towards the fine-mapping of the functional polymorphisms that explain the biological mechanisms and underpin the genotype–phenotype association signals emerging from GWAS. However, while LD has benefitted the initial examination for implicated regions in the genome, long stretches of high LD have paradoxically confounded the fine-mapping process by yielding numerous near-perfect surrogates of the unknown causal variant. This complicates the process of distinguishing the causal variant from neighbouring correlated markers. When GWAS data and high-resolution haplotypes (e.g. those from the 1000 Genomes Project, http://www.1000genomes.org) for multiple populations are available, our method can identify the haplotype forms that are carrying most of the implicated detrimental alleles in the different populations. By framing such *trans*-population analysis within a rigorous statistical framework, it may be possible to identify the genomic regions that are consistent with the association findings and the population-specific reference haplotype structure. This could subsequently be developed to localize the candidate positions of the functional polymorphism (Teo *et al*., 2010) (work in progress).

The popularity of genome-wide studies coupled with fast haplotype phasing software like fastPHASE (Stephens and Scheet, 2005) and Beagle (Browning and Browning, 2007) means it is realistically possible to statistically construct the haplotypes from the genotype data of thousands of samples. This is likely to be extremely useful in both population and medical genetics, as second-order information involving the arrangement of alleles on a chromosome can be more informative than first-order information like the allele frequencies from genotypes of individual SNPs, particularly in understanding genomic diversity across multiple populations. In fact, several major population genetics studies have relied on the haplotype diversity plots as empirical evidence on the extent of inter-population dissimilarities (Conrad *et al*., 2006; Jakobsson *et al*., 2008). We thus introduce HapFinder, a novel methodological development specifically designed to find haplotypes within a population setting, which complements statistical tools for detecting positive natural selection and facilitates progress in medical genetics by locating the likely haplotype structure that the functional allele will sit on. This method has been implemented in a Java program which is packaged together with scripts for producing graphical displays in *R* and is freely available from http://www.nus-cme.org.sg/sgvp/software/hapfinder.html. This URL also contains an interactive application for submitting online queries to find haplotypes from populations in Phase 2 of the HapMap and the Singapore Genome Variation Project.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their insightful comments and suggestions, which greatly improved the manuscript and the software.

Funding: NUS Graduate School for Integrative Science and Engineering (to R.T.-H.O. and X.L.); the Yong Loo Lin School of Medicine from the National University of Singapore (to X.S. and K.-S.C.); the Singapore National Research Foundation (NRF-RF-2010–05 to Y.-Y.T. and W.-T.P.).

Conflict of Interest: none declared.

REFERENCES

Ackerman,H. *et al.* (2005) A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. *Ann. Hum. Genet.,* **69**, 559–565.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.,* **81**, 1084–1097.

Conrad,D.F. *et al.* (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.,* **38**, 1251–1260.

Donnelly,P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature,* **456**, 728–731.

Feng,Z. *et al.* (2004) Coupling ecology and evolution: malaria and the S-gene across time scales. *Math. Biosci.,* **189**, 1–19.

Hedrick,P. (2004) Estimation of relative fitnesses from relative risk data and the predicted future of haemoglobin alleles S and C. *J. Evol. Biol.,* **17**, 221–224.

Hill,A.V.S. *et al.* (1991) Common West African HLA antigens are associated with protection from severe malaria. *Nature,* **352**, 595–600.

Jakobsson,M. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature,* **451**, 998–1003.

Jallow,M. *et al.* (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.,* **41**, 657–665.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.,* **9**, 356–369.

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

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

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.,* **78**, 629–644.

Spencer,C.C.A. *et al.* (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.,* **5**, e1000477.

Stephens,M. and Scheet,P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.,* **76**, 449–462.

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025

Teo,Y.Y. *et al.* (2009) Singapore Genome Variation Project: ahaplotype map of three Southeast Asian populations. *Genome Res.,* **19**, 2154–2162.

Teo,Y.Y. *et al.* (2010) Identifying candidate causal variants via trans-population fine-mapping. *Genet. Epidemiol.,* **34**, 653–664.

Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature,* **449**, 851–861.

Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature,* **467**, 52–58.

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

Downloaded from https://academic.oup.com/bioinformatics/article/27/6/822/234428 by guest on 25 March 2025