Genetics and population analysis

Error detection in SNP data by considering the likelihood of recombinational history implied by three-site combinations

Donna M. Toleno, Peter L. Morrell and Michael T. Clegg*

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA

Received on December 13, 2006; revised on April 30, 2007; accepted on May 9, 2007
Advance Access publication May 17, 2007
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Errors in nucleotide sequence and SNP genotyping data are problematic when inferring haplotypes. Previously published methods for error detection in haplotype data make use of pedigree information; however, for many samples, individuals are not related by pedigree. This article describes a method for detecting errors in haplotypes by considering the recombinational history implied by the patterns of variation, three SNPs at a time.

Results: Coalescent simulations provide evidence that the method is robust to high levels of recombination as well as homologous gene conversion, indicating that patterns produced by both proximate and distant SNPs may be useful for detecting unlikely three-site haplotypes.

Availability: The perl script implementing the described method is called EDUT (Error Detection Using Triplets) and is available on request from the authors.

Contact: mclegg@uci.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation. The most rapid method of collecting complete SNP data for a chromosomal segment is through direct nucleotide sequencing of PCR-amplified DNA fragments. SNP data can be used to calculate population-level diversity statistics even when the phase or arrangement of SNPs and insertion/deletion polymorphisms (indels) in the diploid individual remains unknown. The phase of mutations that make up parental haplotypes cannot be determined through direct sequencing of PCR products, but must be resolved through either computational or experimental means (Clark et al., 1998). Correctly inferred haplotypes increase power for inferences and predictions of disease penetrance, drug responses, diagnostic test variation and risk factors for complex diseases (Clark, 2004; Lee et al., 2005).

Accurate inference of haplotypes is also important for population genetics hypothesis testing and parameter estimation. Haplotyping errors have been shown to affect estimation of any measure requiring knowledge of the arrangement of SNP states, such as the population recombination rate and the quantification of associations among polymorphism (Morrell et al., 2006; Ptak et al., 2004; Wall, 2004).

Errors in resequencing data appear to be commonly encountered and seldom reported. For example, Morrell et al. (2005, 2006) report that standard sequencing practices failed to identify a heterozygous individual in 6 out of 11 cases. The impact of these errors on analyses varies, with the greatest impact on estimated rates of recombination and homologous gene conversion (Morrell et al., 2006).

The possible means for error introduction are numerous and include problems related to both genotyping and phasing. Many research projects require the resequencing of a chromosomal region in a large number of diploid individuals. The potential for resequencing errors is high, in part because every nucleotide site in every individual must be screened for heterozygosity. Miscalled genotypes at heterozygous sites are a major source of error (Morrell et al., 2006). The rarest SNP variants are likely to be present in a heterozygous state, therefore, the investigator may not have a priori knowledge of the presence of the SNP during the genotyping process. Even with highly trained technicians and the best possible quality control measures implemented by sequence assembly software, genotypes of individual SNPs are often called incorrectly in directly sequenced diploid regions (Zhang et al., 2005).

Haplotype inference often depends on computational phasing methods which require arranging SNPs into the most likely configurations. Such methods rely heavily on accurate interpretation of raw sequence trace data or SNP genotyping results (Halldórsson et al., 2004). Experimental phasing methods require either cloning of PCR products or the amplification of one haplotype at a time with allele specific PCR (AS-PCR). Inferring haplotypes with AS-PCR is an alternative to computational phasing or labor-intensive cloning methods (Clark et al., 1998). It is important to note that obtaining the correct genotype at every site does not guarantee accurate inference of haplotypes with any phasing method (Lajoie and El-Mabrouk, 2005).

Screening for errors is desirable even when experimental means are used to recover haplotypes in individuals heterozygous for a locus. Both cloning and AS-PCR involve multiple PCR products and templates resulting in an error-prone assembly process. The cloning process is particularly...
error-prone. High rates of polymerase error make it difficult to recover accurate sequence from each clone, forcing the use of a majority rule approach to determine each haplotype. The clones obtained from any given PCR reaction may contain PCR chimeras (Cronn et al., 2002) and biases in the success rate for cloning the two haplotypes (Morrell et al., 2006). Also, AS-PCR can fail to prime specifically, resulting in genotyping errors. Cloning and AS-PCR are also vulnerable to human and software errors; errors may include sequence assembly problems, sample mislabeling, misinterpretation of sequence traces or sample contamination. While careful sequencing protocols are likely to minimize errors, a tool to identify a subset of unlikely base calls that require re-inspection will be extremely valuable when high data quality is essential.

In the present article we present a method for error detection in which haplotypes are examined three SNPs at a time. We refer to the method as EDUT (error detection using triplets) and we present and test a Perl script implementing the method. EDUT is most sensitive to ‘switch errors’, SNPs occurring on an incorrect haplotype background. Detecting switch errors is difficult because the base calls that lead to incorrect inference of haplotypes frequently occur at known SNPs. In the absence of a method evaluating the likelihood of SNP arrangements, an error can appear to be a likely base call. Switch errors are often repeatable, and can result from machine scored data. There is currently no routine procedure for detecting this type of error. In a previous paper, the EDUT method was used to identify switch errors in haplotype data from a resequencing project (Morrell et al., 2006).

Researchers have avoided the problems associated with direct sequencing of nuclear genes in diploid organisms by focusing on mitochondrial and chloroplast loci not subject to sequencing difficulties caused by heterozygosity, or on organisms where inbreeding or genetic manipulation result in homozygosity at nuclear genes (Wright and Gaut, 2005). As population studies are expanded to a wider range of diploid organisms (including humans), a method with the broad applicability demonstrated by EDUT will be extremely useful.

Published methods for detecting errors in sequencing and phasing, particularly for data from humans, have made use of the relatedness of haplotypes sampled in parents, progeny and siblings (Abecasis et al., 2002; Becker et al., 2006; Douglas et al., 2000; Heath, 1998; Lange et al., 1988; O’Connell and Weeks, 1998; Sobel et al., 2002; Thomas, 2005). Because haplotypes are passed from parents to progeny with a very small probability of being altered by mutation or rearranged by recombination, nucleotide sequence information from related individuals with a known pedigree is useful for imputation of missing data and haplotype reconstruction (Li and Jiang, 2005), and for error detection (Abecasis et al., 2002; Becker et al., 2006). In samples from most organisms, family relationships are difficult or impossible to determine, thus creating the need for a non-pedigree-based method for error detection.

Evaluation of haplotype likelihood has been used for both phase estimation and missing genotype imputation in samples of unrelated individuals (Schet and Stephens, 2006; Stephens and Scheet, 2005). The detection of genotyping errors is a natural extension of haplotype reconstruction methods; both efforts rely on the SNP associations created by mutation and a finite amount of historical recombination.

EDUT can be applied to resequencing or SNP genotyping data and any combination of computationally or experimentally phased haplotypes. The error detection procedure exploits an important property of haplotype data; some haplotypes are more likely than others. Assuming one mutation per site, 5 SNPs can yield $2^{5} - 1$ unique haplotype configurations, only a subset of which are generally observed. EDUT flags a SNP within a specific haplotype whenever an unusual haplotype arrangement occurs for a SNP triplet. A flagged site will be referred to as a ‘flagged base’ to emphasize the role of the flagged site in a single haplotype. The phrase is not meant to limit the discussion to resequencing data. EDUT identifies a subset of base calls that create unlikely arrangements within the inferred haplotype; thus it finds base calls likely to have been inferred incorrectly. The subset consists of the base calls contributing most directly to evidence for a double recombination event. In principle, the method is analogous to the error detection approach of Lincoln and Lander (1992) routinely used in genetic mapping; in practice, it is analogous to the routine process of singleton SNP verification.

2 METHODS

2.1 The algorithm

EDUT processes population data by first identifying all of the SNPs in the data set. The algorithm initially includes all sites, including those in alignment gaps because such sites can include SNPs and may be informative about the evolutionary history of the genomic region. As illustrated in Figure 1, SNP data is converted to a binary format. By convention, the majority state is coded as 0 and the minority state is coded as 1. EDUT accepts data input in standard formats, including aligned fasta files.

Under an infinite sites model, segregating sites will only be altered by mutation one time in the history of the population (Kimura, 1969). This assumption holds for the vast majority of SNPs within a sample from a randomly mating population of a single species. For example, <2% of all SNPs in wild barley resequencing data of Morrell et al. (2006) show evidence of multiple hits by mutation. When there are more than two nucleotide states at a site, EDUT codes the majority state as a 0 and the two or three minor states as 1. Homoplasies events (multiple origins of the same SNP state) are less likely than mutation to a unique 3rd state. When the data shows only minimal violations of the infinite sites assumption, homoplasies have a negligible effect on the effectiveness of this algorithm and are therefore not considered during processing.

All $(S(S-1))/2$ pairs of SNPs are evaluated with the four-gamete test. For any pair of nucleotide sites, only three configurations are possible on the basis of unique mutations. The presence of all four gametes indicates a historical recombination event between the pair of sites (Hudson and Kaplan, 1985). EDUT creates a table storing the results of the four-gamete test. In Figure 1, four gamete tests are true for the site pairs labeled AB and BC. The number of sets of three site combinations is denoted $N_{T}$ where $N_{T} = (S(S-1)(S-2))/6$. All $N_{T}$ triplets are evaluated using stored four-gamete test results so that ‘pattern a’ is identified when no recombination is inferred for the outer pair of SNPs with all four arrangements present at each of the SNP pairs involving the middle SNP (Fig. 1). For all pattern a triplets detected, the rare types contributing to the two diagnostic four-gamete tests are flagged and reported. Rare types are defined by default.
Error detection in SNP data

2.2 Applying the method

The EDUT program was tested against resequencing data from wild barley and avocado where all sequence traces were available. A total of 22 data sets of inferred haplotypes from 25 individuals for wild barley and 41–53 individuals for avocado were used as input into EDUT. In the majority of the data sets no sites were flagged, but in one case the weighted scores drew attention to an assembly error. Inspection of EDUT-flagged base calls in the trace data and consideration of peak heights and quality scores in Consed (Gordon et al., 1998) aid in the assessment of base call accuracy.

EDUT includes an option for processing data directly from the coalescent simulation package, ms (Hudson, 2002). Simulations permitted rigorous testing of EDUT (see Table 1). In each simulation replicate, an error was introduced by changing the 10th site in the first chromosome from one binary state to the other (1 changed to 0, 0 changed to 1).

The HU simulations were designed to resemble human resequencing data. The GE (genotyping error) simulations are designed to represent data as it is frequently collected, i.e. from diploid individuals where the phase of each SNP is unknown. Paired haplotypes in the simulations served as the simulated genotypes. A single error in a single genotype was introduced at the 10th SNP in each replicate simulation; pairs of the simulated chromosomes were subsequently passed as input to Phase 2.1 (Stephens et al., 2001). The resulting inferred haplotypes were used as input for EDUT. For 100 replicate simulations, the process was repeated a total of three times so that the single error was introduced into each of the three possible genotypes (homozygous for the SNP state with major frequency, heterozygous or homozygous for minor SNP state). LS (large sample) simulations included an error at every 100th simulated chromosome in the sample. This represents a low frequency but systematic occurrence of haplotype error.

PS simulations were designed to include population structure. Population structure is common in samples from multiple geographic regions and in comparative studies of populations. Four populations with 25 chromosomes per population were simulated with populations exchanging 1 migrant per generation. The SF (SNP frequency) simulations were designed to test the sensitivity of EDUT to various allele frequencies and states. For each simulation, allele frequencies at the SNP where the error was introduced were categorized as major or minor and either ancestral or derived. We evaluate performance of EDUT for all four combinations (Tables 1 and 2). Unlike the GE simulations above, the SNP frequencies were tabulated after the simulation analysis. For SF simulations, 10 000 replicates were used to increase the incidence of extreme allele frequencies.

The GC simulations were designed to test for the effect of gene conversion on the sensitivity of error detection. Simulation parameters included the mutation parameter, \( \theta = 4N_{e} \mu \) and the recombination parameter, \( \rho = 4N_{e}r \) where \( N_{e} \) is effective population size, \( \mu \) is per base rate of mutation and \( r \) is the per base pair rate of recombination. The gene conversion parameter, \( f \), was defined as the ratio of the rate of gene conversion relative to the rate of crossover. The values of \( f \) used in simulations were 0, 2, 5 and 10; spanning the range of values estimated from empirical data (Padhukasahasram et al., 2006; Wall, 2004). All data analysis, including graphics, was performed with the R statistical environment and programming language (R Development Core Team, 2005).

3 RESULTS

3.1 The EDUT method

The inspiration for the EDUT methodology resulted from careful evaluation of resequencing data from wild barley (Morrell et al., 2005, 2006). All sequences reported were inferred from high quality sequence trace data with a minimum quality criterion of a Phred score \( \geq 20 \) on both forward and reverse reads. As is the customary practice in population genetics resequencing projects, singleton mutations were verified with a new PCR amplification followed by sequencing with coverage of the SNP on the forward and reverse strand. While surveying the data for patterns indicating gene conversion, it was clear that typing errors at key SNP sites could be responsible for producing a triplet pattern resembling the effect of gene conversion. Figure 1 illustrates how EDUT detects
errors using three SNPs and haplotype frequencies typical for a sample of 30 chromosomes. Pattern a, utilized in error detection as illustrated in Figure 1, occurs at a consistently low frequency under many biologically realistic circumstances (See Fig. 2 above and discussion of Fig. 2 below).

3.2 Analyses performed on empirical data

An example of the utility of EDUT on empirical data was presented in Morrell et al. (2006). For example, in the Chf3 locus (Morrell et al., 2005, 2006), EDUT assisted in the detection and correction of four incorrect haplotypes. Homozygosity was incorrectly assumed for 13 SNPs in one sample and 5 SNPs in another. In each case, two switch errors lead to the inference that base calls from these two haplotypes were incorrectly combined. Table S1 (found online) includes the example EDUT output from the wild barley Chf3 locus. Figure S1 demonstrates the resolution of haplotypes at Chf3.

Inspection of the trace files associated with possible errors flagged by EDUT lead to the identification of localized drops in Phred quality scores for incorrectly genotyped heterozygous sites. However, quality scores for these base calls were greater than Phred 20, the quality criterion used in the assembly of the data (Fig. S2). A partial sequence trace is shown in Figure S3. These results demonstrate that inspection of sequence quality can provide supporting evidence for a switch error in any inferred haplotype, but searching for a low quality score is not necessary or sufficient for detecting a sequencing error.

3.3 Analyses performed on simulated data

As discussed above, pattern a occurs at low and relatively consistent frequency under many biologically realistic circumstances (Wiuf and Hein, 2000). Figure 2 illustrates the proportion of pattern a for groups of 100 simulations across \( \rho \) values ranging from 5 to 90. The total number of triplets considered in the proportion includes SNPs for which the minority state occurs at least twice in the sample. Singleton SNPs, occurring only once in the sample, can never participate in pattern a. Simulations represented in Figure 2 are based on \( \rho/\theta = 1.5 \). At lower rates of recombination, holding \( \rho/\theta \) constant causes the simulated data to have few SNPs and thus the SD for

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Number of chromosomes</th>
<th>Number of SNPs</th>
<th>Recombination (distance) in units of kb of human DNA</th>
<th>Purpose</th>
<th>Total number of simulated data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>HU</td>
<td>100</td>
<td>Variable, mean = 208 SNPs (expected for 100 chromosomes and ~50 kb human DNA)</td>
<td>50</td>
<td>Testing the effectiveness of EDUT on simulated human data.</td>
<td>100</td>
</tr>
<tr>
<td>GE</td>
<td>100</td>
<td>Fixed, 50 SNPs</td>
<td>100</td>
<td>Testing the effects of various types of genotyping error: heterozygous SNPs, homoyzygous common, homozoygous rare. See Methods section for details.</td>
<td>100</td>
</tr>
<tr>
<td>LS</td>
<td>200</td>
<td>Fixed, 50 SNPs</td>
<td>100</td>
<td>Testing the effects of a very large sample. For one SNP an error was introduced every 100th haplotype for up to eight total errors at one SNP.</td>
<td>100 for each sample size</td>
</tr>
<tr>
<td>PS</td>
<td>4 groups of 25</td>
<td>Fixed, 50 SNPs</td>
<td>100</td>
<td>Testing the effect of population subdivision. Simulated chromosomes are split evenly into four subpopulations with one migration per generation.</td>
<td>100</td>
</tr>
<tr>
<td>SF</td>
<td>100</td>
<td>Fixed, 50 SNPs</td>
<td>100</td>
<td>Testing the effects of allele frequency on error flagging and priority scores. Exactly the same as ‘GE’ simulations.</td>
<td>10 000</td>
</tr>
<tr>
<td>GC</td>
<td>25, 50, 100</td>
<td>Variable, mean = 20.9 SNPs</td>
<td>10</td>
<td>Testing the effects of gene conversion (( f = 0, 2, 5, 10 )) and number of sampled haplotypes. (25, 50, 100). Error introduced into one haplotype at one SNP, and two SNPs.</td>
<td>100 each</td>
</tr>
</tbody>
</table>
The proportion of triplets in pattern a is elevated. As the simulated recombination rate increases, the mean of a (proportion of all triplets in pattern a) remains relatively constant and small (≤10%).

The results of EDUT with a variety of simulated data are summarized with the rightmost three columns of Table 2. The percentage of all replicate simulations flagging the known error is reported. For the simulated human data, the known error was flagged in 93% of simulations. In the SF simulations, the error was flagged in 76% of simulations. Detection is decreased when the errors are introduced into unphased genotypes as in the GE simulations, where the error was flagged in ~40% of replicate simulations for the two genotype classes including at least one majority SNP and only 10% of the time for the homozygous minority SNP genotype. There is also a trend toward lower detection levels when gene conversion is included and when allowing more than one error at a SNP (LS simulations).

However, there is no clear impact of population structure on error flagging (PS simulations).

The second important summary of the EDUT results is the mean number of sites appearing in the flagged site lists. A shorter list is more desirable because it will be easier to verify

The sensitivity indicated in the last column is defined as follows: [(mean error score)—(mean score for all results)]/(SD of all scores).

Fig. 2. The results of coalescent simulations designed to test the impact of recombination rate on the proportion of pattern a (A) in a data set. A and 95% confidence intervals (defined by ±1.96 standard deviations) are plotted against the ρ per locus indicated on the x-axis. Simulations are based on ρθ=1.5 and with a gene conversion rate of f=5.
the flagged genotypes in the raw data. The number of flagged bases increases along with the number of SNPs in the data used for input. There is a trade-off between the percentage of errors flagged and the total number of sites flagged. For the HU simulations, which represent 50 kb of human data, there were an average of 208 SNPs per simulation (Table 1). This resulted in 895 flagged bases or ~4% of all the (208 × 100) possible sites. The trade-off is more favorable for the smaller data sets where the total sites flagged represents <1% of the possible sites.

The third piece of information in Table 2 describes the ability of the priority score to distinguish between true errors and the background sites flagged. A sensitivity score is calculated in which the mean score for the known errors is compared to the priority scores for all sites. In all but one error category in Table 2 the score was positive, indicating the difference is in the expected direction, i.e. known errors have higher priority scores. The magnitude of the difference is highest for the two categories of simulation in which fixed and realistic ratios of ρ/θ were considered (HU and GC). Considering the standard deviation (SD) of the overall score, the differences were ~2 SDs for the GC simulation with f = 0 and 3.1 SDs for the simulated human data. The GE, LS, PS and SF simulations are more conservative tests of the method because they use a high recombination rate with the number of SNPs fixed at 50 per simulation.

3.4 Processing time

For computational efficiency, EDUT initially evaluates each SNP pair one time to create a table of inferred recombination events. The stored information for the SNP pairs is referenced as each triplet is considered. The processing time therefore increases exponentially with the number of SNPs and is much less dependent on the depth of the sample. Small data sets are processed quickly. For example, the Chf3 locus discussed in the text, which includes 26 chromosomes and 25 informative SNPs, requires only ~3 s of processing time on an AMD Opteron 275 (2.2 GHz) with 16 GB of RAM. A simulated human sequence representing ~50 kb of human data (labeled HU in Table 1) was processed in ~6 min on the same processor with processing time increasing ~5-fold for a region including only twice as many SNPs.

4 DISCUSSION

EDUT is useful under many biologically realistic conditions. Much of the utility of the method lies in the potential to detect switch errors in heterozygous individuals. Missed heterozygous SNPs represent a very significant source of error, particularly in highly heterozygous samples. Every site in every individual must be genotyped as either homozygous or heterozygous. We have used coalescent simulations to demonstrate that EDUT can be expected to flag a single base genotyping error with >50% probability, returning a higher priority score for known switch errors than for false positives. Because individual sequence reads tend to contribute switch errors at multiple adjacent SNPs, the probability of identifying individuals that are heterozygous at a locus is high. In combination with other available tools for genotyping and resolving phase of heterozygous individuals, EDUT can be used to dramatically improve the accuracy of haplotype data.

4.1 Evidence in support of the method’s utility

Coalescent simulations (Hudson, 2002; Wiuf and Hein, 2000) demonstrate that a (the proportion of all triplets in pattern a) is small and fairly constant. The variance of a decreases as larger regions of sequence are considered. Small regions often do not include any sites in pattern a (Morrell et al., 2006). Figure S1 illustrates that for the Chf3 locus, despite the lack of pattern a in the resolved haplotypes, the EDUT method assisted in resolving the phase of SNPs for two heterozygous genotypes because the original incorrect resequencing data included erroneous pattern a triplets.

As described in the results, weighted priority scores tend to be higher for the known errors than for other flagged bases. Priority scores can inform user decisions about which SNP genotypes and/or haplotypes to double-check. Whether a user should check the complete list of flagged bases or a subset of the list depends on the likelihood of the presence of an error, the perceived cost of the effort of checking a site that is not an error, and the importance of discovering errors in the data set.

4.2 Errors EDUT is not likely to detect

There are three general types of errors that EDUT is not likely to detect. These include errors occurring in data with no evidence of a recombination event, switch errors at a singleton SNP where the minority state occurs only once in the sample, and any error causing a haplotype to masquerade as one of the other observed haplotypes.

When haplotypes result from computationally phased genotype data, an erroneous base call will rarely be detected if the correct genotype is homozygous rare. Table 2 indicates that detection occurred 10% of the time in a group of 100 simulated data sets and the mean priority score for an error is not greater than the priority scores for all flagged bases.

Simulations also indicate that errors in computationally phased data are much less likely to be flagged than errors in directly obtained haplotypes, although computational phasing seems to have no effect on the sensitivity of the weighted priority scores (compare GE and SF simulation results in Table 2). An error at a minority allele is also less likely to be detected than an error at the majority allele.

Also, very large samples present a challenge for error detection because even a 1% error rate results in eight errors for a sample size of 400 diploid individuals. The presence of numerous errors limits the utility of EDUT because the errors can erode the signal from the rare SNP arrangements that provide the basis for error detection. Thus the potential to detect an error is decreased with each additional error.

4.3 Correctly genotyping each site

Accurate haplotype data is highly dependent on accurate genotype data. For resequencing data, there are a number of issues that complicate accurate genotyping. First, the discovery of all SNPs in a sample is difficult, because the rarest class of
SNPs may be found only in the heterozygous state. Very rare SNPs are difficult for SNP detection software to identify, because the only evidence of the SNP occurs as overlapping peaks in resequencing trace data. Segregating indel polymorphisms also limit the utility of SNP detection tools, such as Polyphred 5.03 (Stephens et al., 2006) and SNPdetector (Zhang et al., 2005) because heterozygous indels create PCR products from a single individual that differ in length, resulting in partially unreadable sequence traces.

As an alternative to resequencing, haplotypes might be derived from a series of SNP genotypes. There are a variety of reasons why such genotyping assays may fail to yield accurate genotypes. Many SNP genotyping platforms rely on carrying out an initial multiplex PCR amplification of specific genomic regions, and use this as template for the subsequent genotyping assay. For individual assays, this amplification can fail, leading to missing data for an individual. More seriously, there might be allele specific amplification (due to a heterozygous site within the binding site for one of the PCR amplification oligos). This could result in a heterozygous SNP being incorrectly called homozygous. After PCR amplification, a potential problem can arise if the subsequent oligo binding step involves an undiscovered SNP. This can lead to a failure of the assay for one of the haplotypes such that a heterozygous query SNP will be called homozygous. The potential for undetected SNPs to disrupt an assay is likely to be dependent on sequence composition, and will be expected to vary with the genotyping platform and its implementation (Koboldt et al., 2006).

Empirical data and simulation experiments indicate that EDUT is helpful for detecting occasional errors in SNP genotyping and less helpful for finding errors at SNPs with high rates of genotyping errors. Since missing data does not interfere with the method, it would be prudent to omit any data resulting from suspect assays.

As with the methods presented by Lincoln and Lander (1992), EDUT lends itself to reexamination of unlikely results. Although the errors are detected through information about SNP arrangement within haplotypes, both resequencing and SNP genotype data verification require a review of the original SNP genotype call.

### 4.4 Determining the phase of the mutations

Even with correctly genotyped data, computational methods may not always phase the genotype data with high confidence. Rare mutations are difficult to phase, because there is little prior information about the haplotypes with which they are associated. Also, current computational phasing methods generally assume that all recombination is due to crossover, and do not account for the potential contribution of homologous gene conversion to haplotype diversity (Lajoie and El-Mabrouk, 2005). This is problematic, because the best available estimates suggest that gene conversion may contribute as much or more than crossover to total recombination (Morrell et al., 2006; Ptak et al., 2004). EDUT can be used in conjunction with computational and experimental phasing in the process of accurate haplotype inference and can be used with phasing methods as part of an iterative quality control procedure.

### 4.5 Importance of accurate haplotype inference

Accurate haplotype data is important for many basic biological questions and practical applications. One fundamental biological question for which accurate haplotype data is essential is the estimation of the rate of homologous gene conversion from resequencing data (Ptak et al., 2004; Wall, 2004). Errors can result in large biases in the gene conversion rate estimate (Morrell et al., 2006; Ptak et al., 2004; Wall, 2004).

Both whole genome scans and candidate gene association studies will benefit from accurate haplotype data. Haplotype associations with phenotypic variation of interest (association mapping) can facilitate the process of finding candidate genes and could lead to elucidating important structural and regulatory gene interactions. Incorrect haplotypes will not only decrease the effectiveness of whole genome scans but will also decrease the power of hypothesis testing in case-control studies.

It has been demonstrated that genotyping errors can have a large impact on association studies. Knapp and Becker (2004) used a simulation approach to determine the impact of genotyping errors. Specifically, they measured the effect on Zhang’s haplotype-sharing transmission disequilibrium test (HS-TDT), a non-parametric test to determine if a haplotype transmission from parent to offspring is associated with a phenotype (Zhang et al., 2003). With only 1% of markers incorrectly genotyped, the error rate is increased from a nominal error rate of 0.05 to around 0.5 (Knapp and Becker, 2004). Clearly genotyping errors can have large effects on haplotype-based association tests.

The method presented is a computationally tractable means of examining haplotype data. Through the use of three site haplotype arrangements, EDUT exhibits considerable utility for reducing the data to the portion most worthy of reexamination. Though the time and resources devoted to verifying haplotype base calls may be significant for large genotyping or resequencing projects, the improvement in data quality can be substantial.

### ACKNOWLEDGEMENTS

The authors thank M. Brandt, C. Torres, L. DeRose, S. MacDonald and two anonymous reviewers for helpful comments on previous versions of the manuscript. This work was supported by National Science Foundation grant DEB-0129247.

Conflict of Interest: none declared.

### REFERENCES


