REPORT

Detecting Disease-Causing Mutations in the Human Genome by Haplotype Matching

David H. Spencer, Kerry L. Bubb, and Maynard V. Olson

Comparisons between haplotypes from affected patients and the human reference genome are frequently used to identify candidates for disease-causing mutations, even though these alignments are expected to reveal a high level of background neutral polymorphism. This limits the scope of genetic studies to relatively small genomic intervals, because current methods for distinguishing potential causal mutations from neutral variation are inefficient. Here we describe a new strategy for detecting mutations that is based on comparing affected haplotypes with closely matched control sequences from healthy individuals, rather than with the human reference genome. We use theory, simulation, and a real data set to show that this approach is expected to reduce the number of sequence variants that must be subjected to follow-up analysis by at least a factor of 20 when closely matched control sequences are selected from a reference panel with as few as 100 control genomes. We also define a reference data resource that would allow efficient application of this strategy to large critical intervals across the genome.

Complete resequencing of candidate genomic intervals is emerging as a practical approach for detecting disease-causing mutations in the human genome. Advances in sequencing technology will soon allow human geneticists to move beyond genotyping and limited feature-by-feature sequencing of candidate genes within disease-associated loci and to resequence entire genomic intervals, potentially spanning hundreds of candidate genes, in search of mutations that affect phenotype. As this era of human genetics approaches, better conceptual and computational methods for analyzing these extensive data sets will be necessary. In particular, efficient and systematic methods for distinguishing functional variants against the high level of background neutral variation in sequences from affected individuals will be critically important. In the present study, we explore a method that is based on well-established population-genetics theory that addresses this issue, and we also assess the types of reference data on human genetic variation that would facilitate its application.

Identification of candidates for deleterious mutations relies on sequence comparisons between affected and unaffected individuals. In human genetics, genomic regions implicated in disease are typically compared with the single human reference sequence from the Human Genome Project. These comparisons reveal many discrepancies, most of which are functionally irrelevant. In the absence of obvious loss-of-function mutations, distinguishing potential disease-causing mutations from background variation typically involves surveying large panels of control individuals for the presence or absence of each "private" SNP encountered in the patient. This approach has been practical for small regions with few candidate genes, but larger genomic intervals may contain hundreds of genes and thousands of polymorphisms, making further investigation of potential disease-causing mutations difficult and expensive. Here, we consider a new strategy for identifying potential disease-causing mutations where a local haplotype implicated in disease by genetic linkage or allelic association is compared with a closely matched control sequence, rather than with the human reference sequence, with the idea that the low level of polymorphism between these two closely related sequences will make discovery of functionally important mutations more efficient.

The success of this strategy depends on the level of similarity that is achievable in practice in comparisons between two sequences sampled from real human populations—for example, one from an affected patient and one from a healthy control individual. We will examine this question by first considering the simplest situation, in which a dominant, highly penetrant trait of large effect has been localized by linkage analysis to a well-demarcated region in the genome. We will assume that the sequence of this interval has been determined completely from the appropriate haplotype of the affected patient (i.e., the "query" haplotype), as well as from k chromosomes from control individuals. Initially, we will also assume that these $k+1$ haplotypes have been sampled from a random-mating population of constant historical size and have not experienced recombination or selection since they diverged from a common ancestral sequence. Under these stringent assumptions, coalescent theory predicts that the query will differ from a randomly chosen sequence (i.e., any one of the k control sequences or, for that matter, the human reference sequence) by $\pi = \theta \approx 7 \times 10^{-4}$ per site,

From the Department of Genome Sciences (D.H.S.; K.L.B.; M.V.O.) and University of Washington Genome Center, Department of Medicine (M.V.O.), University of Washington, Seattle

Received May 22, 2006; accepted for publication August 29, 2006; electronically published September 25, 2006.

Address for correspondence and reprints: David H. Spencer, Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195. E-mail: dhs@u.washington.edu

Am. J. Hum. Genet. 2006;79:958–964. 0002-9297/2006/7905-0018$15.00

or ~700 differences per megabase pair ($\theta = 4N\mu$, where N is the effective population size; μ is the per site, per generation mutation rate; and π is the nucleotide diversity in the population, estimated from surveys of human genetic variation[1,2]). We are interested in how much better a match is likely to be found if all k reference sequences are considered, rather than just one. The tree shown in figure 1, which is based on a simple coalescent simulation,[3] illustrates the basic issues for the case where $k = 10$. The control sequence most similar to the query is the one closest to it on the tree, and the level of divergence between these two sequences is proportional to the time since their most recent common ancestor. This time interval is represented on the tree by the length of the external branch leading to the query. Because the query is equally likely to occupy any position on the tree, the expected length of the external branch leading to it is simply the expected length of any of the $k + 1$ external branches. This quantity was shown by Fu and Li[19] to be $4N/(k + 1)$ generations. Hence, the frequency of neutral mutations between a query and its best-matched comparison sequence is expected to be $8N\mu/(k + 1) = 2\theta/(k + 1)$. As expected, this quantity reduces to θ when $k = 1$, which is equal to π under the prevailing assumptions.

The relationship between the expected frequency of neutral mutations and k is shown in figure 2*A,* along with the average results from 10,000 simulations performed using a standard implementation of sequence evolution based on coalescent theory.[3] As this figure shows, the similarity between best matches varies almost inversely with the size of the reference panel. At a panel size of 100, the expected number of discrepancies between a query and its closest match is 14 per Mbp—a 50-fold reduction in the frequency of background polymorphisms compared with that expected for a single pairwise comparison—and only 5% of matches obtained from panels of this size are expected to differ by >50 discrepancies per Mbp.

The polymorphisms detected from comparisons between a query haplotype and its best match are equally likely to be private to either haplotype. Hence, determining which ones occurred on the lineage giving rise to the query would be expected to reduce the number of potential disease-causing mutations by an additional factor of two. This reduction can be accomplished by comparing the query with any third haplotype, such as the human reference sequence. Sites that remain polymorphic in this comparison are unique to the query and thus should be considered potential disease-causing mutations.

More-realistic predictions about the similarity between closely matched haplotypes must incorporate the effects of population structure, demographic history, and recombination. Population structure has the intuitive effect that closely related sequences will tend to be from the same population subgroup. As a result, the level of matching between a query and the best-matched sequence will depend on the ancestries of the individuals in the reference panel and not just on the panel's overall size, with closer matching expected if a majority of the individuals in the panel are from the same subgroup as the query. Recent studies have shown that signatures of population structure can be detected in human sequences, with subdivision of world populations based on geographic origin.[4,5] Therefore, to maximize the chance of finding appropriate comparison sequences in real data, reference panels should be composed of individuals with geographic origins, at least on a continental scale, that are similar to that of the query. This issue will be revisited later in our analysis of haplotypes sampled from real human populations. For now, we will simply point out that the effects of population structure can be accommodated by taking the geographic ancestry of the query into account when forming reference panels of control individuals.

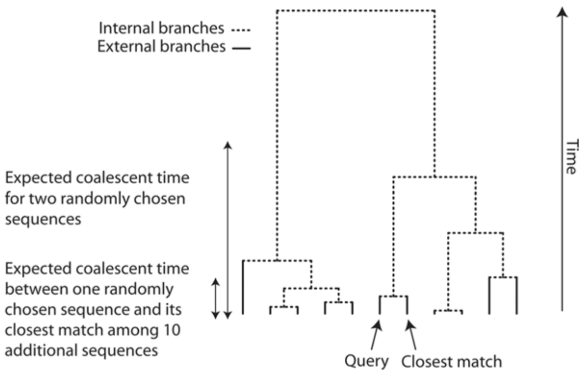
Internal branches
External branches
Expected coalescent time for two randomly chosen sequences
Expected coalescent time between one randomly chosen sequence and its closest match among 10 additional sequences
Query
Closest match
Time

Figure 1. A coalescent tree for one "query" sequence and 10 control sequences. The best-matching control for the particular query sequence indicated by the arrow is the one closest to it on the tree, and the divergence between the two sequences is proportional to the length of the query's external branch. More generally, the vertical arrows on the left show the difference in expected coalescent time for a random query and its closest match among 10 control sequences (*shorter arrow*), compared to the expected coalescent time for the query and an arbitrarily chosen sequence (*taller arrow*).

The effects of nonstationary demographic history were assessed using coalescent simulations incorporating realistic models of human evolutionary history. Because population expansion results in coalescent trees that are more starlike (i.e., the external branches tend to be longer relative to internal branches), we were most interested in the effects of population growth on the divergence between closely matched sequences. In our analysis, we considered a variety of growth models,[2,6,7] including simple models of population expansion, as well as models featuring expansion following a bottleneck. In all cases, including that of the more complex, "calibrated" model of Schaffner et al.,[7] we emphasized models and parameter values that provide good fits to real data on human-allele-frequency distributions.[2,6,7] Figure 2*B* shows the average divergence between closely matched haplotypes from 10,000 coalescent simulations performed using the set of published bot-

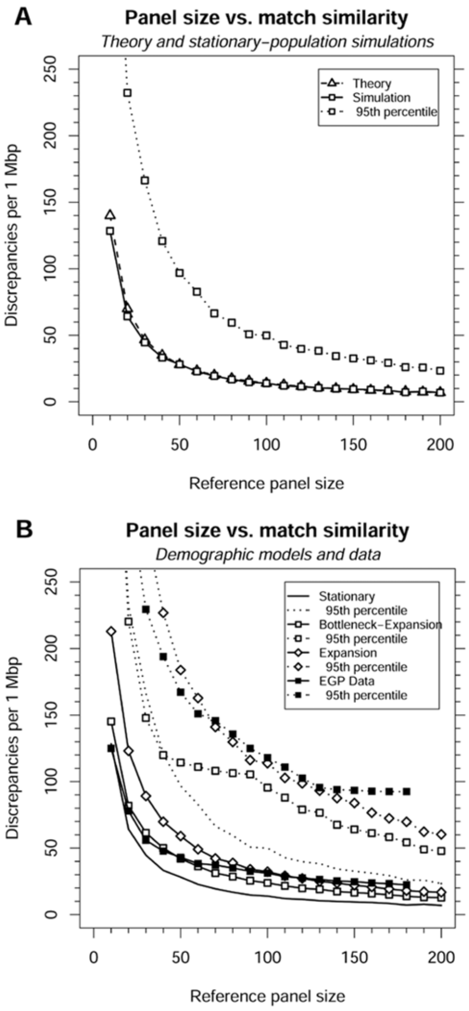

Figure 2. *A,* Average match similarity, in discrepancies per Mbp, from theory and coalescent simulations. The dashed line with unblackened triangles and the solid line with unblackened squares indicate, respectively, the theoretical expectation[19] and averages from 10,000 coalescent simulations[3] with a constant historical population size. The broken line with unblackened squares indicates the 95th percentile match similarity from the simulation. *B,* Average match similarity, in discrepancies per Mbp, from coalescent simulations incorporating demographic models of human evolutionary history and from a real data set. The solid line with no symbols indicates the match similarity from simulations with a stationary population size; lines with unblackened squares and unblackened diamonds, respectively, indicate the results from simulations using bottleneck-expansion and pure-expansion demographic models[6]; the line with blackened squares indicates the match similarity for haplotypes inferred from the sequences of 19 genes from 95 individuals of diverse ancestry.[10] Broken lines indicate the 95th percentile values for all analyses. For both panel A and panel B, θ was assumed to be 7×10^{-4} wherever applicable.

tleneck-growth and growth parameters that results in the greatest departure from the stationary model[6]; all other models we examined produced similar results (data not shown). In general, the demographic models we considered have relatively minor effects on the expected divergence between closely matched sequences. Although population growth results in the greatest departure from the stationary model, realistic growth models still predict that the frequency of discrepancies between closely matched sequences will be quite low: with a panel size of 100, the average divergence between best matches from simulations performed using the three models of population growth we considered never exceeded 32 discrepancies per Mbp, compared with 14 per Mbp for the stationary model.

Recombination does not fundamentally alter the predictions we have made, but it has significant practical effects, because it limits the distance over which a single reference sequence will be optimally matched with a query. Thus, in the presence of recombination, it may be necessary to create a tiling path of multiple, closely matched references to minimize the number of discrepancies between a query haplotype and a set of reference haplotypes. The length of a match between a query and a particular reference depends on the number of historical recombination events that have occurred on the two sequences since they diverged from a common ancestral sequence. As we have shown, theory predicts this time interval will become shorter as the size of the reference panel increases; therefore, fewer recombination events are expected to have occurred when closely matched comparison sequences are obtained from large panels, resulting in matches that extend over greater distances. This relationship was explored further by use of coalescent simulations incorporating a uniform recombination rate. We simulated 10-Mbp sequences by use of the standard coalescent model with a recombination rate of $4Nc = 3.5 \times 10^{-4}$ (where c is the per site, per generation recombination rate) and identified the best match, at the center of the simulated region, between a randomly selected query sequence and the other simulated haplotypes. We then determined the distance over which close matching between the query and this sequence extended. This simulation corresponds to a realistic scenario in which a match to an affected haplotype is found at a particular position in the genome (e.g., at the center of a linkage interval or candidate gene). Figure 3 shows the average length of matches from 1,000 iterations of this simulation procedure, for a range of panel sizes. These results demonstrate the expected relationship between panel size and match length and also suggest that matches obtained from panels with >100 sequences will persist for ~1 Mbp, on average.

Local variation in recombination rate and recombinational hotspots within a candidate region will not adversely affect match length, as long as the overall recombination rate of the candidate region is not unusually high. Recombinational hotspots will have the obvious effect that tiling paths of closely matched sequences will tend to break at hotspots. Of course, matching within the boundaries of a particularly hot region will be difficult; however, because most such regions have been shown to be narrow,[8] comparisons between an affected haplotype

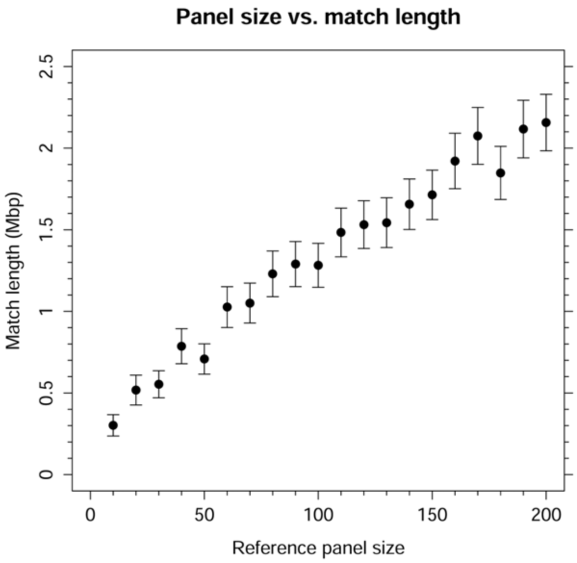

Figure 3. Match length, in Mbp, for different panel sizes. Each point is the average distance over which a query sequence is closely matched to a single reference sequence from 10,000 coalescent simulations with a uniform recombination rate of $4Nc = 3.5 \times 10^{-4}$ (where c is the per site, per generation recombination rate). The error bars indicate the 95% CI.

and a nonoptimal control sequence (e.g., one of the controls that matches closely to the sequence adjacent to the hotspot) will not result in many additional candidates for causal mutations. Genomewide recombination-rate maps[8] could be used to predict where tiling path breaks should be expected and to determine the likelihood that potential causal mutations are, instead, the result of poor matching in a hot region. In addition, because the majority of recombination events appear to occur in these narrow, highly recombinogenic regions,[9,8] the lower background recombination rate will mean that matches outside of hotspots will typically extend for even longer distances than those shown in figure 3. For example, a simulation similar to those described that used a coalescent simulator incorporating a model of recombinational hotspots[7] resulted in an average match length of 3.5 Mbp for a panel size of 100, compared with 1.3 Mbp when a uniform recombination rate was assumed.

To test our results based on theory and simulations against real data, we used a set of 19 genes sequenced in 95 individuals of African American (15 individuals), Yoruban (12 individuals), Asian (24 individuals), Hispanic (22 individuals), or European (22 individuals) ancestry.[10] Computational phasing of these sequences resulted in 190 inferred haplotypes for each gene,[11,12] with an average of 311 segregating sites per locus and a π of 7.0×10^{-4}. Because matching in this data set is likely affected by historical recombination events, a hidden Markov model (HMM) was used to identify "composite matches" comprising the most probable tiling path of closely matched reference sequences. Applying this matching algorithm to each of the 190 sequences per gene × 19 genes = 3,610 sequences and using the remaining 189 sequences from the same locus as a reference panel (i.e., $k = 189$) resulted in an average divergence between each sequence and its best-matched tiling path of 22 discrepancies per Mbp, which is comparable to theoretical predictions and simulations for this panel size. Most best-matched tiling paths identified by the HMM comprise a single sequence, with only 28% having >1 sequence and a mean number of reference sequences, for all tiling paths, of 1.4.

The level of matching in these data also follows theoretical expectations at smaller panel sizes. This was assessed by performing 1,000 iterations of a resampling procedure where the HMM was used to compare randomly selected query sequences with panels of various size, sampled from the remaining 189 haplotypes for each of the 19 genes. The relationship between panel size and average divergence between matches follows the trend predicted by theory and, for most panel sizes, is within the range of values predicted by simulations with realistic demographic models (fig. 2*B*). For example, with a panel size of 100, the average divergence between closely matched sequences is 31 per Mbp, with only 5% of the matches differing by >118 discrepancies per Mbp. The average number of sequences per best-matched tiling path, reflecting the level of recombination between close matches, decreased from 1.8 sequences at a panel size of 10 to 1.4 at a panel size of 180, which is consistent with the relationship between panel size and match length predicted by theory.

We used the same data set to explore the effects of population structure on the divergence between closely matched haplotypes. To this end, we divided the haplotypes into four groups (with the haplotypes of African American and Yoruban ancestries combined into a single group) and performed 1,000 iterations of a resampling procedure where query sequences were compared with randomly chosen 40-member reference panels in all pairwise combinations of population groups. As expected, the geographic origin of the query and reference panel has a significant effect on the level of matching (fig. 4). The average divergence between best matches varies from 25 discrepancies per Mbp, when both the query and reference panel are of European ancestry, to >200 discrepancies per Mbp, when the query sequences of African ancestry are matched to reference panels of Asian ancestry. Matching is best when the query sequence and reference panel are selected from the same population groups, although within-group matching varies from ~25 discrepancies per Mbp between sequences of European and Asian ancestry to almost 60 discrepancies per Mbp between sequences of African ancestry. These results confirm the importance of taking geographic origin into account when finding closely matched comparison sequences in real data.

The ease of applying haplotype matching as a strategy for detecting causal mutations in human genetics could

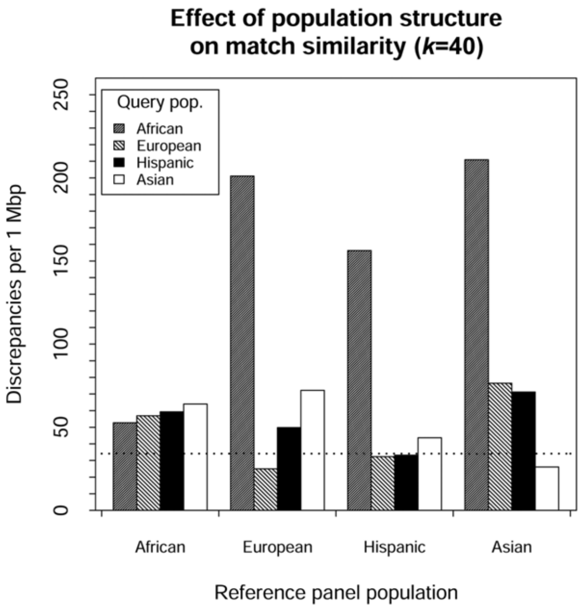

Figure 4. Average match similarity, in discrepancies per Mbp, in a real data set sampled from diverse human populations[10]; closely matched sequences are from either the same or a different subpopulation as the query sequence. Each bar represents an average of 1,000 iterations of a resampling procedure in which a query haplotype was sampled from one population (indicated by the pattern on the bar) and then was compared with a randomly chosen, 40-member reference panel comprising only sequences from one of the four populations (indicated on the *X*-axis). The dotted line represents the theoretical match similarity, under the assumption of a single population of constant historical size, when $k = 40$.

be enhanced greatly by new sources of reference data. Of course, investigators will be on their own in determining the sequence of haplotypes from affected individuals. Here, we note only that methods for acquiring these data are steadily improving as it becomes increasingly practical to tile long regions with overlapping PCR amplicons[13] or recombinant-DNA molecules[14,15] and as next-generation DNA-sequencing technologies mature.[16,17]

However, new sources of reference data would facilitate detection of potential disease-causing mutations in these sequences by allowing closely matched comparison haplotypes to be identified rapidly without the need for case-by-case data collection. Although the need for more comprehensive reference data on human variation for purposes of mutation detection will be met ultimately by the emergence of full-reference-genome sequences sampled from multiple human populations, haplotype matching allows much smaller reference data sets to be immediately useful. All that would be required for comparison-haplotype identification is light-shotgun sampling of the genomes of a panel of ~100 phenotypically "normal" humans. Cell lines from these individuals would need to be available as a source of DNA from which investigators could determine complete sequences of a candidate region.

The key variable in implementing this plan is the required depth of sampling. This depth would have to be sufficient to support the discovery of enough rare SNPs to allow effective matching between a haplotype sequenced from an affected individual and one present in the reference panel. Most rare SNPs have arisen recently in human evolution[18] and thus provide "lineage-specific" tags that identify sequences sharing a recent common ancestor with the query. Knowledge of only a handful of these lineage-specific tags would allow selection of comparison haplotypes that are well matched to a given query. Current SNP databases are of little value for this purpose, because their contents are biased toward common SNPs; furthermore, even when rare SNPs are present, their allele frequencies are typically unknown.

To estimate the level of sequence sampling that would be required to produce a useful "rare-SNP" database, we conducted simulations of haplotype matching, using resources developed at different sampling levels. In these simulations, query haplotypes were matched to comparison haplotypes using only the information acquired by

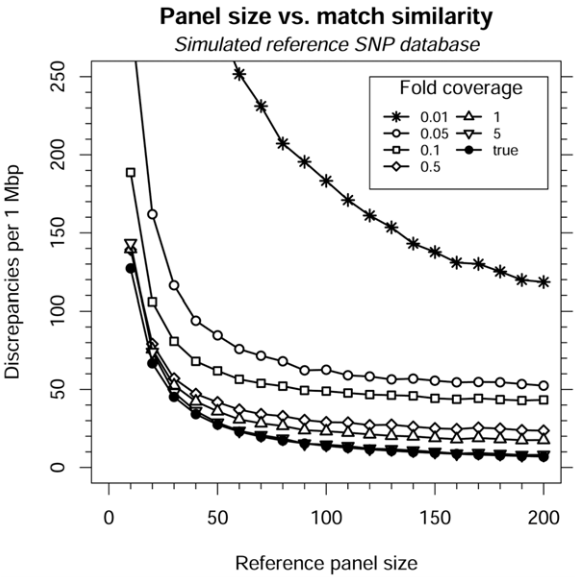

Figure 5. Simulations of haplotype matching peformed using a reference panel database from whole-genome-sequence sampling. Query and reference sequences were generated using standard coalescent simulations with a stationary, randomly mating population ($\theta = 7.0 \times 10^{-4}$). Sequences in the reference panel were sampled to reflect shotgun-sequence sampling to various levels of coverage—that is, under the assumption of a Poisson coverage model, the probability of observing a particular polymorphic site on a reference sequence at least once is $1 - e^{-c}$, where c is the fold coverage of the haploid genome. Sites sampled from the reference panel formed the basis for selecting the closely matched sequences. The line with blackened circles indicates the average match similarity for true best match from the simulations.

the simulated-light-shotgun sampling of reference genomes. The results, shown in figure 5, demonstrate that quite good matching can be achieved even at 0.1-fold coverage of each haploid genome in the reference panel; however, sampling in the 0.5-fold range is required to bring match qualities within a factor of 2–3 of optimal.

In this article, we have described a conceptually simple strategy for identifying potential disease-causing mutations that involves comparing complete haplotype sequences from affected individuals with the sequences of closely matched control haplotypes. We have used coalescent theory, simulations, and reference data to show that the expected number of discrepancies between an affected haplotype and its best-matched sequence from a panel with as few as 100 control genomes is at least 20-fold less than the expectation for pairwise comparison with a single, arbitrary reference genome. We have also suggested an efficient method for implementing this strategy, which should make it accessible to individual human genetics laboratories. This implementation requires genotyping only a handful of SNPs in a reference panel to find a closely matched comparison sequence, thus enabling researchers to efficiently identify potential causal mutations. Finally, we have defined a reference-data resource that would facilitate the application of this method to genetic studies across the genome.

We conclude with a brief comment about the range of applicability of the proposed method. As described, the simplest application would be for detecting mutations that cause dominant disorders segregating in multigenerational families. The case of rare recessive disorders detected in inbred families is formally identical, because candidate regions are defined by genome segments that are identical by descent and homozygous in affected individuals. Because the major frontier of Mendelian analysis in humans involves rare dominant and recessive diseases, the causal mutations of which are often poorly localized because of the small number of informative meioses available, these cases alone provide a large range of applications for which haplotype matching could be very useful. In these studies, searching entire linkage intervals containing hundreds of candidate genes could be accomplished without investigators being overwhelmed by large numbers of potential causal mutations.

A still larger area of application involves diseases with complex modes of inheritance. Like all other methods of genetic analysis, the power of haplotype matching will depend on the nature and extent of genetic complexity exhibited by a particular disease. Although this strategy will not be useful in diseases caused by common mutations, it generally will offer significant improvement over current approaches in studies of any complex disease in which the disease-causing mutation is rare. In these instances, controls that are closely matched to an affected haplotype yet do not carry the disease-causing mutation should be present in the population, because the causal mutation typically will have arisen recently.

Modifications to the basic matching strategy we have described here could be introduced to address some of the complications that contribute to the genetic complexity of a trait. For example, in situations where inheritance patterns are complex because of contributions from several individually rare, weakly penetrant alleles, the use of multiple comparison sequences would reduce the risk that any one comparison sequence from an unaffected individual contains one of the relevant mutations. Other situations may benefit from the use of custom panels of “hypernormal” controls, to minimize the risk that control individuals carry mutations that can contribute to the disease phenotype in some environments or genetic backgrounds. With these and other modifications, matching of complete haplotype sequences from affected individuals with well matched controls could serve as an optimum strategy for identifying causal sequence variants in a wide variety of inherited diseases.

Acknowledgments

We thank Arnie Kas and James Ronald for help in developing the matching HMM. This work was supported by a Center for Excellence in Genome Science grant (5 P50 HG002351) to M.V.O.

References

1. Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831
2. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102:18508–18513
3. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338
4. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
5. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
6. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351–372
7. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576–1583
8. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324
9. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584
10. NIEHS SNPs (2005) NIEHS Environmental Genome Project. University of Washington, Seattle, WA

11. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
12. Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169
13. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320
14. Raymond CK, Subramanian S, Paddock M, Qiu R, Deodato C, Palmieri A, Chang J, Radke T, Haugen E, Kas A, Waring D, Bovee D, Stacy R, Kaul R, Olson MV (2005) Targeted, haplotype-resolved resequencing of long segments of the human genome. Genomics 86:759–766
15. Traherne JA, Horton R, Roberts AN, Miretti MM, Hurles ME, Stewart CA, Ashurst JL, Atrazhev AM, Coggill P, Palmer S, Almeida J, Sims S, Wilming LG, Rogers J, de Jong PJ, Carrington M, Elliott JF, Sawcer S, Todd JA, Trowsdale J, Beck S (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. PLoS Genet 2:e9
16. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
17. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732
18. Kimura M, Ota T (1973) The age of a neutral mutant persisting in a finite population. Genetics 75:199–212
19. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709