Published in final edited form as:
Genet Epidemiol. 2009 ; 33(Suppl 1): S24–S28. doi:10.1002/gepi.20468.

Haplotype-Based Analysis: A Summary of GAW16 Group 4 Analysis

Elizabeth Hauser[1], **Nadine Cremer**[2], **Rebecca Hein**[2], and **Harshal Deshmukh**[3]

[1]Center for Human Genetics, Duke University, Durham, NC

[2]Cancer Epidemiology, German Cancer Research Center DKFZ, Heidelberg, Germany

[3]Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK

Abstract

In this summary paper, we describe the contributions included in the haplotype-based analysis group (Group 4) at the Genetic Analysis Workshop 16, which was held September 17-20, 2008. Our group applied a large number of haplotype-based methods in the context of genome-wide association studies. Two general approaches were applied: a two-stage approach that selected significant single-nucleotide polymorphisms and then created haplotypes and genome-wide analysis of smaller sets of single-nucleotide polymorphisms selected by sliding windows or estimating haplotype blocks. Genome-wide haplotype analyses performed in these ways were feasible. The presence of the very strong chromosome 6 association in the North American Rheumatoid Arthritis Consortium data was detected by every method, and additional analyses attempted to control for this strong result to allow detection of additional haplotype associations.

Keywords

population stratification; multiple comparisons

Introduction

Genetic Analysis Workshop 16 Group 4 was constituted of groups with a common interest in understanding the role of haplotype methods in the genetic analysis of a common trait. However, once this common interest was acknowledged the hypotheses to be addressed, methods used, method of analysis, and interpretation differed widely among the members. All of the members defined a haplotype similarly as a set of alleles of genetic markers on the same chromosome inherited from a single parent, i.e., the set of alleles ordered in map order. However, only the two-allele genotypes for each individual are observed, not the phased haplotypes themselves. The problem then becomes one of trying to identify the underlying phased haplotype from the observed unphased genotypes and then using this information to perform tests to identify genes for complex disease.

Given the inherent uncertainty in estimating the underlying haplotype, a strong rationale for applying a haplotype analysis is needed. First, haplotypes can be considered to better represent the parental chromosomes that are the defining units. Because they are the inherited units, haplotypes also can contain the allele-specific biological functions. For

Corresponding Author: Elizabeth R. Hauser, Center for Human Genetics, Duke University Medical Center, DUMC Box 3445, Durham, NC 27710 USA, Telephone: 919-684-0603, Fax: 919-684-0913, Elizabeth.Hauser@duke.edu.

example, *cis*-regulatory events would be chromosome-specific and thus knowing the linear order of alleles would be crucial to understanding the relationship between the allele, its function, and its relationship to the disease of interest. Second, haplotypes contain a record of human evolutionary history through the relationship of alleles on inherited chromosomes over time. Populations of humans may be organized through observing these allelic relationships that ultimately make up the haplotypes, as was demonstrated in an evolutionary genetics analysis of populations from Africa [Campbell and Tishkoff, 2008; Tishkoff et al., 2009]. These allelic relationships are summarized by measures of linkage disequilibrium (LD) [Balding, 2006]. This information may be used to further our understanding of relationships between and among populations and also may further our understanding of population similarities and differences in genetic mediators of complex disease [Balding, 2006; Hindorff et al., 2009]. The combination of the potential for a better reporter of biological function as well as a better reporter of population evolutionary history suggests that explicity modeling haplotype relationships can improve power in association analysis.

This improvement in power follows from several strong assumptions about the underlying models for complex disease, which in turn follow directly from the assumption of a strong evolutionary relationship between the surrounding genetic variation and a single disease-susceptibility allele that makes up the haplotype. Ideally, we would hope that the genotyping experiments include and identify a causal variant that causes disease. However, given the large number of genetic variation that distinguishes individuals and populations, it is not realistic to assume that all potential genetic variants can be genotyped. LD relationships in the genome haplotype effects may be stronger than the effects of single-nucleotide polymorphisms (SNPs) if the causal variant is not among the genotyped SNPs and if the causal variant is not in strong LD with one of the genotyped markers [Balding, 2006]. Thus, the haplotyped markers may gain strength from surrounding, weaker LD relationships between the causal variant(s) and the genotyped markers [Liu et al., 2008]. In addition, specifically estimating the phased haplotypes appropriately adjusts for the dependence between multiple significant disease association results due to LD between markers on the same disease haplotype.

Application of haplotype models to studies of candidate genes for complex diseases has been successful in identifying haplotypes of interest. With so much attention currently focused on genome-wide association (GWA) studies as a means of assaying as much genetic variation as possible in our search for disease susceptibility genes, the question then becomes, what is the place of haplotype association analysis in GWA studies? GWA genotyping panels are created to contain as much genetic variation as possible and, at the same time, increase efficiency by limiting LD between genotyped markers. Power profiles for the different GWA chips can vary dramatically when single-SNP analyses are applied [Spencer et al., 2009]. Thus, it seems that haplotype analysis including multiple SNPs could augment power in certain situations and may serve as a natural complement to the single-SNP analysis [Liu et al., 2008]. It is already clear that the population LD relationships within the human genome can be reliably estimated when hundreds of thousands of SNP genotypes are available with which to estimate population structure induced by human evolutionary relationships, clearing the way for large association studies in which population structure may be estimated [Zhu et al., 2008]. However, the magnitude of the problem of estimating haplotypes and haplotype frequency when performing a genome-wide haplotype analysis is extreme, in terms of both inference and computation. Thus, in our contributions tremendous energy was devoted to defining and limiting the problem to allow for successful computation and accurate inference.

Each participant had a different approach to using the genome-wide data for haplotype analysis. All had access to genome-wide data and most chose to use the North American

Rheumatoid Arthritis Consortium (NARAC) rheumatoid arthritis (RA) dataset because the dataset had the most previous analyses available and a strong signal on chromosome 6 in the HLA region. There was a mixture of analyses of limited candidate regions and application to the whole genome. In addition, Deshmukh et al. [2009] identified a set of regions implicated in a diverse set of autoimmune diseases to evaluate commonalities in genetic signals with the NARAC study. In all cases data reduction of some kind was required, either to avoid intensive computation or to reduce the number of comparisons and to reduce the effect of multiple comparisons on the type 1 error rate.

Estimating Haplotypes

It is not computationally feasible to estimate haplotypes across the genome without limiting the size of the haplotypes. Thus, each group selected different approaches to reducing the computational burden. Several groups chose to focus on genomic regions or single genes and perform the haplotype analysis for a limited number of SNPs. Wormald and Zhou (unpublished) chose to examine haplotype relationships in the *BRD2* gene in the HLA region and Hazlett, Zielinski, and Moslehi (unpublished) chose to examine the *PHTF1*, *RSBN1*, and *PTPN22* genes located in a small region on chromosome 1. Other groups chose SNPs and regions identified by an initial single marker analysis using standard methods and applications such as PLINK [Purcell et al., 2007b] or WHAP [Purcell et al., 2007a] for case-control association or quantitative trait association, respectively. Following the initial SNP identification, a large variety of methods were used to form haplotypes. Schulz, Cremer, Bermejo, Fischer, Hein, Beckmann, and Chang-Claude (unpublished) “grew” haplotypes from seed SNPs in regions identified by initial single-point association results. SNPs were added as long as the multi-locus LD increased with the addition of the new SNPs. Abo, Knight, and Camp (unpublished) used a forward-backward selection routine implemented in HapConstructor to select informative SNPs from the 15 SNPs including and surrounding the significant SNPs [Abo et al., 2008]. Haplotypes were then constructed from the SNP sets defined during the selection process. Park et al. [2009] evaluated haplotype blocks around significant SNPs to guide haplotype estimation [Park et al., 2009]. The most comprehensive methods estimated haplotypes for all SNPs meeting quality control benchmarks but limited the size of the haplotypes. Allen and Satten [2009] used a sliding window of SNPs to create haplotypes with seven SNPs, the maximum window size attempted by any group [Allen and Satten, 2009]. Another whole-genome approach was to limit the haplotype construction to within haplotype blocks defined using the four-gamete rule or the method of Gabriel [Gabriel et al., 2002; Shim et al., 2009; Park et al., 2009]. As demonstrated by Shim et al. [2009] in the genome-wide setting, the definition of a haplotype block can have a strong impact on the results of the subsequent association study because the results could be quite different depending on the algorithm used and the number of SNPs included in the blocks [Shim et al., 2009].

Haplotype Association Methods

Once haplotypes were defined, whether in a well defined region or genome-wide, a variety of haplotype association methods were applied. In most cases these were tests of a global hypothesis of no differences in haplotype frequencies between cases and controls, as opposed to identifying specific haplotypes that are different in cases and controls. Guo et al. [2009] applied regularized generalized linear models (rGLM) analysis, which implements a LASSO penalty in the logistic regression to limit the effect of rare haplotypes on the analysis. In this analysis, Guo et al. [2009] examined only chromosomes 6 and 18 and thus were able to replicate previous association results in these regions. Allen and Satten [2009] and Schulz, Cremer, Bermejo, Fischer, Hein, Beckmann, and Chang-Claude (unpublished) employed haplotype-sharing metrics to evaluate evidence for association exploiting the idea

that under models of a recent disease mutation, cases should exhibit increased genetic similarity in the region of the disease variant compared to the controls. Allen and Satten [2009] detected two novel loci for RA, one near the pleiotrophin (*PTN*) gene on chromosome 7q33 and one near glypican 6 (*GPC 6*) on chromosome 13q31.3. In a very ambitious effort, Park et al. [2009] not only performed haplotype association analysis but also evaluated interactions between all significant haplotype blocks to evaluate pathways and haplotype networks. The haplotypes participating in the most interactions were termed hub haplotypes. The majority of the hub haplotypes were located on chromosome 6, but one hub haplotype was on chromosome 7q36.2. In summary, all of these methods successfully identified the strong association with the chromosome 6 region but beyond that there was no agreement in results for other chromosomal regions. The reasons for this lack of replication remain obscure and comparisons through simulation studies are likely necessary to identify the sources of these differences.

Conditioning on Chromosome 6

One of the most challenging aspects of applying haplotype analyses to the NARAC dataset was how to handle the very large effects of the DR alleles in the HLA region at chromosome 6p21. When analyzed, whether by haplotype analysis or single-point analysis, chromosome 6 always provided extremely strong evidence of association. Thus, it is difficult to draw comparisons of different approaches when analyzing the chromosome 6 data. Perhaps a more interesting question is what else can be found, either within the HLA region or in other regions across the genome. The problem is to evaluate evidence for additional susceptibility alleles after accounting for the strong DR effect. This theme appeared in several of the contributions when conditioning on the shared epitope (SE). Lemire [2009] used SNP genotypes and haplotypes to predict or tag the class I and class II HLA alleles and then performed association analysis on RA using the SNP predictors and surrogates. This analysis also conditioned on SE, which showed that taking SE into account can change the evidence for association at other HLA loci and identified components of the DQ8 serotype as associated with RA, independent of LD with DRB1. Taylor and Criswell [2009] also identified additional loci within the HLA region when conditioning on SE and then estimated haplotypes for the new markers along with SE. Within the MHC region they identified two HLA class I markers and one HLA class III marker that do not appear to be associated due to strong LD with DRB1. It is possible that this conditioning approach could be useful for identifying additional regions outside of HLA associated with RA.

Effect of Population Stratification on Haplotype-Based Approaches

All participants recognized the need for some type of correction for population stratification. The haplotype-based methods of case-control association suffer from the same problems due to population stratification as those seen for single SNPs. The haplotype frequencies in any given population likely depend strongly on the evolutionary history of the population. In fact, the haplotype-based methods, by virtue of the increased information in haplotypes, may suffer even more strongly from undetected or uncorrected population stratification. The large variance inflation factors (VIF approximately 1.4) calculated for this study imply false associations and lack of control of type I error. Thus, most participants made some correction for population stratification before beginning the haplotype analysis. The majority used a principal-components (PC) approach [Price et al., 2006] but there were differences in the sets of markers used in the analysis, the number of PCs used, and the definition and removal of outliers. Clearly no correction, no matter how extreme, will materially change the qualitative importance of the general HLA association. However, association with specific HLA alleles may be influenced by population stratification, and large changes in the *p*-values can occur after adjustment for population stratification.

Control for Multiple Tests in Haplotype-Based Methods

While the data reduction inherent in haplotype analysis can help ameliorate the burden of multiple tests corrections, even when haplotype methods are applied it is very important to adjust for multiple comparisons. Group 4 used several methods for type I error rate control, incuding Bonferroni adjustment, the use of the false-discovery rate (FDR), and permutation testing with calculation of empirical *p*-values. Allen and Satten [2009] performed permutation testing taking into account both adjustment for population stratification and genome-wide association. The FDR procedures were hampered by the strong correlation of the statistical results in some regions, with the hundreds of significant results on chromosome 6 being a prime example of violation of the independence assumption of the FDR procedures. The appropriate adjustment for multiple comparisons remains under debate.

Implications of Haplotype Analysis

The analyses performed by the participants in Group 4 highlighted different facets of the potential of haplotype analysis, particularly in the setting of GWA studies. One question that must be asked is whether the additional information generated by haplotype analysis outweighed the computational burden? Further, did the haplotype analysis provide more insight than the single-marker analysis? There was not clear concordance between the single-marker analyses and haplotype analyses; SNPs significant in the single-marker analyses were not always included in significant haplotypes and significant haplotypes did not always contain SNPs significant in single-marker tests. These differences may be due to variations in power [Taylor and Criswell 2009], particularly for SNPs of modest effect or may be due to differences in the underlying biological genetic models at different loci. It might be postulated that the haplotype analysis could be closer to the biological mechanism. For example, Park et al. [2009] analyzed haplotype interactions followed by analysis of the most highly interacting genes to identify hub haplotypes. Not surprisingly, most of the hub genes were in the HLA region on chromosome 6. Haplotype association analysis may also provide links to other related conditions whose similarities for both endophenotypes and genetic associations may have implications for functional relationships [Caillat-Zucman, 2009; Deshmukh et al., 2009; Traherne, 2008]. The conditional analysis also provided an opportunity to examine evidence for heterogeneity related to SE [Lemire, 2009; Taylor and Criswell, 2009]. Genes identified in the SE+ group might suggest the presence of interactions while genes identified in the SE− group might suggest genetic heterogeneity.

Haplotype analysis may also have particular implications for follow-up analysis. For example, strong association of a particular haplotype can be used to design sequencing studies to identify the variant on the haplotype. The need for replication, even for haplotype results, remains acute. However, replication in haplotypes is perhaps even more difficult because the LD relationships may vary from population to population and across samples [Crosslin et al., 2009; Gu et al., 2008; Zintzaras et al., 2009]. Finally, the two-stage process in which SNPs are selected in the first stage and haplotypes are constructed in the second stage makes replication very difficult if the first stage does not identify the proper SNPs from which to build the haplotypes. If true genome-wide analysis becomes available without need for an initial data reduction step, then replication should, in theory, become easier.

Conclusions

The group reached a number of general conclusions based on comparisons across all of the projects. First genome-wide haplotype analysis is feasible and was accomplished by several groups. However, all approaches required a data reduction step of some kind, whether

limiting the haplotype estimation to small windows across the genome or to haplotype blocks. The most common data reduction strategy was a two-stage analysis that selected a set of SNPs around which to perform haplotype analysis. Second, the two-stage approach runs the risk of missing haplotypes that do not include single significant SNPs. Further concordance between single-point analysis and haplotype-based analysis can be poor and the interpretation of differences between the two analyses can be extremely difficult, even in the face of well defined LD patterns. These differences between the single SNP and haplotype association results were complex and depended on the different analyses. Thus, general conclusions could not be drawn about the cause of these differences. Finally, as might be expected, haplotypes with more SNPs capture more information which may, in turn, provide a more powerful test and may increase significance. Increasing the number of SNPs in the haplotype was an overwhelming computational burden and in general haplotype applications were limited to seven or fewer SNPs at a time.

The analyses presented in Group 4 suggest that haplotype analysis in all its variation holds promise for increasing information and power for genetic association studies of complex traits. The contributions demonstrated that haplotype analyses are relevant and feasible in genome-wide association studies.

Acknowledgments

Additional Group 4 contributing members include Ryan Abo, Lars Beckmann, Anke Schulz, Hanna Wormald and Michelle Zhou, Allison Hazlett, Kristen Zielinski and Roxanna Moslehi. This work was supported by NIH grants R01 GM031575 and R01 MH59528 and the Neurosciences Education and Research Foundation.

References

Abo R, Knight S, Wong J, Cox A, Camp NJ. hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. Bioinformatics. 2008; 24:2105–7. [PubMed: 18653522]

Allen AS, Satten GA. Genome-wide association analysis of rheumatoid arthritis data via haplotype sharing. BMC Proc. 3(Suppl 7):S30. [PubMed: 20018021]

Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006; 7:781–91. [PubMed: 16983374]

Caillat-Zucman S. Molecular mechanisms of HLA association with autoimmune diseases. Tissue Antigens. 2009; 73:1–8. [PubMed: 19017300]

Campbell MC, Tishkoff SA. African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008; 9:403–33. [PubMed: 18593304]

Crosslin DR, Shah SH, Nelson SC, Haynes CS, Connelly JJ, Gadson S, Goldschmidt-Clermont PJ, Vance JM, Rose J, Granger CB, Seo D, Gregory SG, Kraus WE, Hauser ER. Genetic effects in the leukotriene biosynthesis pathway and association with atherosclerosis. Hum Genet. 2009; 125:217–29. [PubMed: 19130089]

Deshmukh H, Kim-Howard X, Nath SK. Replication of recently identified associated single-nucleotide polymorphisms from six autoimmune diseases in Genetic Analysis Workshop 16 rheumatoid arthritis data. BMC Proc. 2009; 3(Suppl 7):S31. [PubMed: 20018022]

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science. 2002; 296:2225–9. [PubMed: 12029063]

Gu CC, Yu K, Rao DC. Characterization of LD structures and the utility of HapMap in genetic association studies. Adv Genet. 2008; 60:407–35. [PubMed: 18358328]

Guo W, Liang Cy, Lin S. Haplotye association analysis of North American Rheumatoid Arthritis Consortium data using a generalized linear model with regularization. BMC Proc. 2009; 3(Suppl 7):S32. [PubMed: 20018023]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–7. [PubMed: 19474294]

Lemire M. On the association between rheumatoid arthritis and classical HLA class I and class II alleles predicted from single-nucleotide polymorphism data. BMC Proc. 2009; 3(Suppl 7):S33. [PubMed: 20018024]

Liu N, Zhang K, Zhao H. Haplotype-association analysis. Adv Genet. 2008; 60:335–405. [PubMed: 18358327]

Park J, Namkung J, Jhun M, Park T. Genome-wide analysis of haplotype interaction for the data from the North American Rheumatoid Arthritis Consortium. BMC Proc. 2009; 3(Suppl 7):S34. [PubMed: 20018025]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]

Purcell S, Daly MJ, Sham PC. WHAP: Haplotype-based association analysis. Bioinformatics. 2007a; 23:255–6. [PubMed: 17118959]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007b; 81:559–75. [PubMed: 17701901]

Shim H, Chun H, Engelman CD, Peyseur BA. Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: An empirical comparison with data from the North American Rheumatoid Arthritis Consortium. BMC Proc. 2009; 3(Suppl 7):S35. [PubMed: 20018026]

Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009; 5:e1000477. [PubMed: 19492015]

Taylor KE, Criswell LA. Conditional analysis of the major histocompatibility complex in rheumatoid arthritis. BMC Proc. 3(Suppl 7):S36. [PubMed: 20018027]

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM. The genetic structure and history of Africans and African Americans. Science. 2009; 324:1035–44. [PubMed: 19407144]

Traherne JA. Human MHC architecture and evolution: Implications for disease association studies. Int J Immunogenet. 2008; 35:179–92. [PubMed: 18397301]

Zhu X, Tang H, Risch N. Admixture mapping and the role of population structure for localizing disease genes. Adv Genet. 2008; 60:547–69. [PubMed: 18358332]

Zintzaras E, Rodopoulou P, Sakellaridis N. Variants of the arachidonate 5-lipoxygenase-activating protein (ALOX5AP) gene and risk of stroke: A HuGE gene-disease association review and meta-analysis. Am J Epidemiol. 2009; 169:523–32. [PubMed: 19126581]