Genetic Epidemiology 33: 599–603 (2009)

VALID: Visualization of Association Study Results and Linkage Disequilibrium

Eric Jorgenson,[1,2]* Mark Kvale,[2] and John S. Witte[2,3]

[1]*Department of Biopharmaceutical Sciences, University of California, San Francisco, San Francisco, California*
[2]*Institute for Human Genetics, University of California, San Francisco, San Francisco, California*
[3]*Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California*

Promising findings from genetic association studies are commonly presented with two distinct figures: one gives the association study results and the other indicates linkage disequilibrium (LD) between genetic markers in the region(s) of interest. Fully interpreting the results of such studies requires synthesizing the information in these figures, which is generally done in a subjective and unsystematic manner. Here we present a method to formally combine association results and LD and display them in the same figure; we have developed a freely available web-based application that can be used to generate figures to display the combined data. To demonstrate this approach we apply it to fine mapping data from the prostate cancer 8q24 loci. Combining these two sources of information in a single figure allows one to more clearly assess patterns of association, facilitating the interpretation of genome-wide and fine mapping data and improving our ability to localize causal variants. *Genet. Epidemiol.* 33:599–603, 2009.

Key words: association; linkage disequilibrium; genetic; genome-wide; fine mapping

Contract grant sponsor: NIH; Contract grant numbers: RR024130, GM061390, CA088164, and CA127298.
*Correspondence to: Eric Jorgenson, Department of Biopharmaceutical Sciences and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA 94143-0794. E-mail: Eric.Jorgenson@ucsf.edu
Received 27 August 2008; Revised 20 November 2008; Accepted 23 December 2008
Published online 5 February 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20411

INTRODUCTION

Genome-wide association studies have recently succeeded in identifying multiple genetic variants that underlie a number of human diseases. A typical genetic locus identified in these studies contains numerous variants which are significantly associated with the disease of interest. It is often difficult to determine which of these variants provides the strongest evidence of association because of correlation between variants due to linkage disequilibrium (LD). To address this problem, authors have reported both association results and LD maps using two adjoining figures; in general, $-\log_{10}$ *P*-values are displayed for the association between disease and each single nucleotide polymorphism (SNP) in the region, and below that a map of pairwise LD between SNPs in the same region [Amos et al., 2008; McPherson et al., 2007; Scott et al., 2007; Weedon et al., 2007; Winkelmann et al., 2007; Yeager et al., 2007]. We show an example of this type of figure using fine mapping data of the prostate cancer 8q24 region from a study by Haiman et al. [2007] in Figure 1. Figure 1a depicts the association *P*-values for SNPs in the region, while Figure 1b shows the LD structure.

Two problems arise with this type of display. First, because association is a single measure while LD is a pairwise measure, it is difficult to line up the results of each figure. Second, it is not clear how the two measures are related. For example, a marginally significant SNP marker may not have a significant association signal once LD with other associated SNPs is taken into account. Both of these problems can be addressed by combining the association study test results with the LD information and displaying them in the same figure. Below we describe a method and a freely available web-based software tool that can be used to combine association and LD data in the same figure.

METHODS

Assume one is interested in the association between two neighboring SNPs (SNP_1 and SNP_2) and a particular phenotype. The standard association approach would simply estimate the marginal association between the genotypes at each SNP and the phenotype (e.g., a χ^2-test statistic and corresponding *P*-value). One can also measure the pairwise LD between SNP_1 and SNP_2 using r^2 [Devlin and Risch, 1995]. It is possible then to calculate the expected χ^2 for SNP_1 conditional on the observed χ^2 for the SNP_2 and the r^2 between the two markers:

$$E(\chi_2^2|\chi_1^2, r_{12}^2).$$

For SNP_2 conditional on SNP_1, this is given by

$$E(\chi_2^2) = \chi_1^2 r_{12}^2.$$

(Formal details of this relationship are given in the Appendix.)

For SNP_1 and SNP_2, there are a total of four conditional values, one for SNP_1 conditional on SNP_2, one for SNP_2 conditional on SNP_1, one for SNP_1 conditional on itself, and one for SNP_2 conditional on itself. The latter two

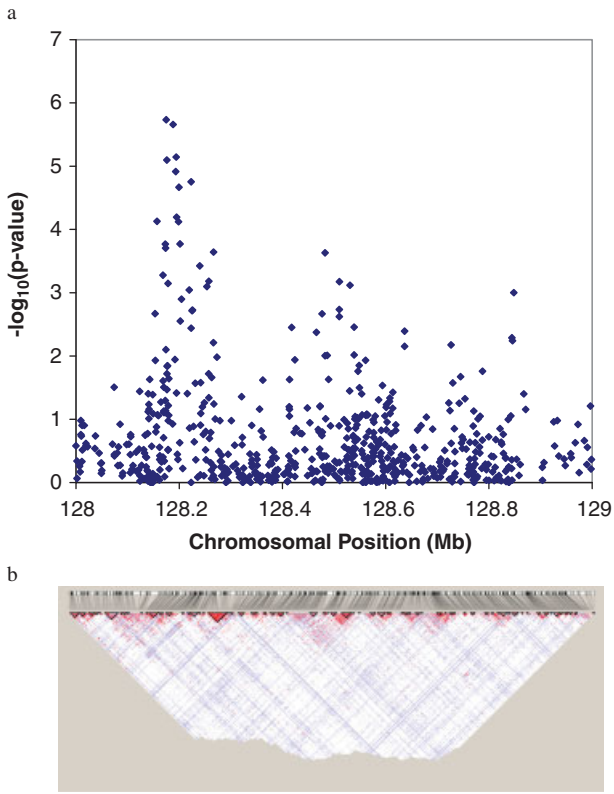

Fig. 1. Association results and corresponding linkage disequilibrium map. (a) Results from the Haiman et al. [2007] study of prostate cancer on chromosome 8q24 in African-Americans. (b) A Haploview plot of pairwise linkage disequilibrium between SNPs in the region of interest from the HapMap YRI population. SNP, single nucleotide polymorphism.

values are simply the observed association signals for SNP_1 and SNP_2, because r^2 is 1 for any SNP and itself. One can display the four corresponding $-\log_{10} P$-values in a two-dimensional heat map. This method can be expanded to display a large number of SNPs in a genomic region. The resulting figure provides information about the strength of association signals in the region (on the diagonal) and correlation of association signals due to LD (off-diagonal elements).

APPPLICATION

We illustrate the strength of this approach with an example plot that utilizes the data from the fine mapping study of Haiman et al. [2007] of the prostate cancer 8q24 region in an African-American sample, along with information on LD in the YRI population from the HapMap (Fig. 2). The location of SNPs in the sample data is represented from most proximal to most distal from left to right on the x-axis and from top to bottom on the y-axis. In this example, there are several strong association signals or ''hotspots'' that can be readily identified by the intensity of warmer colors on the diagonal of the matrix (Fig. 2a). The SNPs with the strongest association signals are labeled to the right of the figure, and the SNP with the strongest overall signal is highlighted in red.

This method provides a way to display association signals and LD together, and it also allows for the visual identification of association signals that can be explained entirely due to their correlation with other associated loci. A second locus provides an example of a signal at one location that may be driven by its correlation with neighboring SNPs (Fig. 2b). Focusing on the top left quadrant, a cluster of positive association signals can be identified by visual inspection of the heat map; two local peaks of association are indicated by hotspots on the diagonal. In addition, warm colors appearing off the diagonal indicate that the SNP in that row is expected to have a positive association signal due to its correlation with SNPs in the corresponding columns.

For the first association peak, the off-diagonal elements in the corresponding rows (to its right in the figure) reach a tepid green in the columns that correspond to the second association peak. This indicates that a modest association signal with those SNPs would be expected solely due to their correlation with SNPs in the second peak. For the second hotspot, the off-diagonal elements in those rows (here, to the left) reach a hotter orange, indicating that a strong association signal would be expected solely due to the correlation of those SNPs with SNPs in the first hotspot. Because the first association signal appears hotter than its expected conditional association signal and the second association signal does not appear hotter than its expected association signal, this suggests that the association signals seen here are due to SNPs in the first hotspot, and that the signal at the second hotspot is most likely due to correlation with SNPs in the first hotspot. The algorithm used for labeling top SNP signals utilizes this information to display the rs numbers of the top SNP signals (see the Appendix).

DISCUSSION

Our method for displaying genome-wide association results and LD in the same figure has several advantages over other methods of display. First, it provides a clear picture of the relative strength of association signals in a specific region of the genome. Second, it allows for the visualization of the effect of LD between neighboring SNPs on their observed association signals. This can help localize the source of a particular association as well as identify multiple independent associations in a single region such as those that exist in the prostate cancer 8q24 region.

To implement this method, we have developed software that can effectively handle dense genotype data using a simple file input format, and we have made it available on our web site. Investigators can tailor their individual plots by adjusting a number of options to best display their data. In addition to utilizing LD information from the HapMap, one can also use LD information from the study itself. This can provide potentially more accurate estimates of LD, especially for populations that are not well represented by the HapMap populations. Finally, users can also upload the results of association analyses that adjust for the effect of neighboring SNPs. Association results for individual SNPs can again be represented on the diagonal of the matrix, while adjusted results can be represented in the off-diagonal. These types of analyses can address concerns about potential bias in LD estimates in small samples

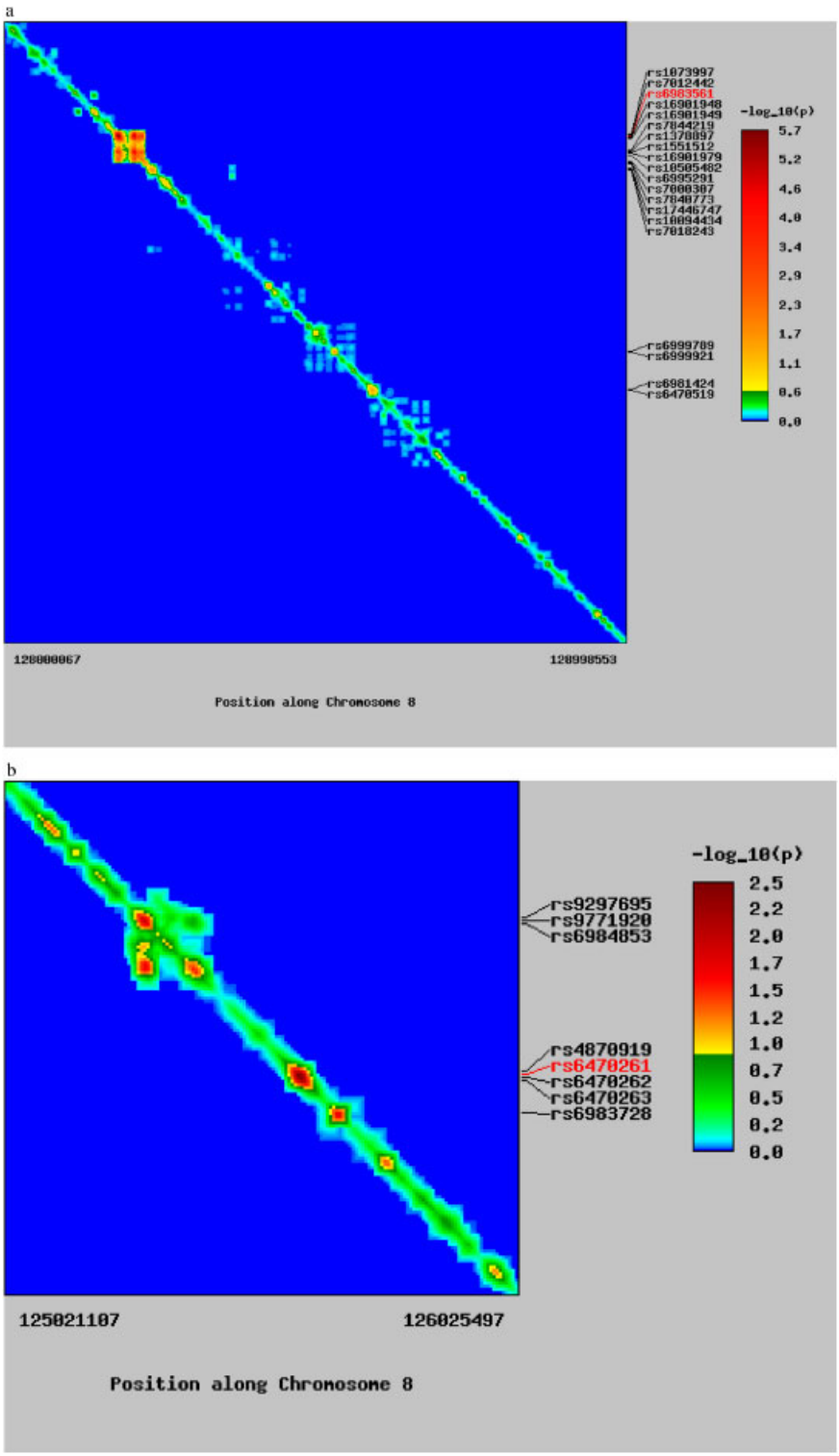

Fig. 2. Heat map for association and LD on chromosome 8q24. (a) Heat map for the 8q24 region. Chromosomal position is labeled on the bottom. SNPs with strong association signals are identified to the right of the figure, with the strongest signal labeled in red. Both the rows and columns represent individual SNPs. Colors in the matrix represent the strength of the association signal for the SNP in that row. On the diagonal, the association signal is simply the observed value for that particular SNP. Off-diagonal elements represent the expected association signal for the SNP in that row due to its correlation with the SNP in the corresponding column. (b) View of additional 8q24 loci illustrating an asymmetric association pattern. SNP, single nucleotide polymorphism.

[Terwilliger and Hiekkalinna, 2006; Thomas and Stram, 2006]. For meta-analyses, such data are often unavailable, and the method described in this paper can be applied.

As the volume of genomic data is rapidly increasing, tools that can help researchers visualize, understand, and interpret this large-scale association data are needed. Our method addresses this need by synthesizing information on association analyses and LD. Ultimately, this tool can help clarify the patterns of association observed, facilitate the interpretation of genomic data, and improve our ability to localize causal variants.

WEB RESOURCES

VALID Visualization Tool: http://www.humgen.medschool.ucsf.edu/Research/Software.aspx.

HapMap: http://www.hapmap.org.

Haploview: http://www.broad.mit.edu/mpg/haploview.

ACKNOWLEDGMENTS

We thank Drs. Gary Chen, Iona Cheng, Vincent Fradet, Sung Kim, Audrey Schnell, and Yu Chuan Tai for their helpful comments on the web application.

REFERENCES

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS. 2008. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 40:616–622.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, Greenway SC, Stram DO, Le Marchand L, Kolonel LN, Frasco M, Wong D, Pooler LC, Ardlie K, Oakley-Girvan I, Whittemore AS, Cooney KA, John EM, Ingles SA, Altshuler D, Henderson BE, Reich D. 2007. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 39:638–644.

Matthews AG, Haynes C, Liu C, Ott J. 2008. Collapsing SNP genotypes in case-control genome-wide association studies increases the type I error rate and power. Stat Appl Genet Mol Biol 7:Article23.

McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. 2007. A common allele on chromosome 9 associated with coronary heart disease. Science 316:1488–1491.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14.

Sasieni PD. 1997. From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345.

Terwilliger JD, Hiekkalinna T. 2006. An utter refutation of the "Fundamental Theorem of the HapMap." Eur J Hum Genet 14:426–437.

Thomas DC, Stram DO. 2006. An utter refutation of the "Fundamental Theorem of the HapMap" by Terwilliger and Hiekkalinna. Eur J Hum Genet 14:1238–1239.

Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B et al. 2007. A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39:1245–1250.

Winkelmann J, Schormair B, Lichtner P, Ripke S, Xiong L, Jalilzadeh S, Fulda S, Putz B, Eckstein G, Hauk S, Trenkwalder C, Zimprich A, Stiasny-Kolster K, Oertel W, Bachmann CG, Paulus W, Peglau I, Eisensehr I, Montplaisir J, Turecki G, Rouleau G, Gieger C, Illig T, Wichmann HE, Holsboer F, Muller-Myhsok B, Meitinger T. 2007. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. Nat Genet 39:1000–1006.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni Jr. JF, Hoover R, Hunter DJ, Chanock SJ, Thomas G. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645–649.

Zhang Q, Wang S, Ott J. 2008. Combining identity by descent and association in genetic case-control studies. BMC Genet 9:42.

APPENDIX

We describe a method to combine information on association and LD and display them together in the same plot. The goal of displaying information in this way is to determine which SNPs have the strongest evidence for association and which association signals exceed the signal expected due to the association of surrounding markers. To do this, we can use the test of association for each marker and a pairwise measure of LD between markers.

For a case-control study with a sample of size N, we can measure the association between the locus, with alleles A and a, and disease using a χ^2-test:

$$\chi_1^2 = \frac{(\hat{\pi}_{DA} - \hat{\pi}_{CA})^2 2N\phi(1-\phi)}{\hat{\pi}_A(1-\hat{\pi}_A)}$$

where $\hat{\pi}_{DA}$ is the frequency of allele A among cases, $\hat{\pi}_{CA}$ is the frequency of allele A among controls, N is the total sample size, ϕ is the proportion of the same, that is, cases, and $\hat{\pi}_A$ is the frequency of allele A in the total sample.

Similarly, a test between a second locus, with alleles B and b, and disease is

$$\chi_2^2 = \frac{(\hat{\pi}_{DB} - \hat{\pi}_{CB})^2 2N\phi(1-\phi)}{\hat{\pi}_B(1-\hat{\pi}_B)}.$$

The distribution of the χ^2 statistic is approximately the square of that of a normal random variable, Z. The expected value (mean) of Z for locus 1 is

$$E(Z_1) = (\pi_{DA} - \pi_{CA})\left[\frac{2N\phi(1-\phi)}{\bar{\pi}_A(1-\bar{\pi}_A)}\right]^{1/2}$$

and for locus 2:

$$E(Z_2) = (\pi_{DB} - \pi_{CB})\left[\frac{2N\phi(1-\phi)}{\bar{\pi}_B(1-\bar{\pi}_B)}\right]^{1/2}$$

where $\bar{\pi}_A = \phi\pi_{DA} + (1-\phi)\pi_{CA} \approx \pi_A$ and $\bar{\pi}_B = \phi\pi_{DB} + (1-\phi)\pi_{CB} \approx \pi_B$. If q_{AB} is the probability that a chromosome carrying the A allele at the first locus carries the B allele at the second locus and similarly q_{aB} is the probability that a chromosome carrying the a allele at the first locus carries the B allele at the second locus, then

$$\pi_{DB} - \pi_{CB} = (\pi_{DA} - \pi_{CA})(q_{AB} - q_{aB})$$

and therefore

$$E(Z_2) = (\pi_{DB} - \pi_{CB})\left[\frac{2N\phi(1-\phi)}{\pi_B(1-\pi_B)}\right]^{1/2}$$

is equivalent to

$$E(Z_2) = (\pi_{DA} - \pi_{CA})(q_{AB} - q_{aB})\left[\frac{2N\phi(1-\phi)}{\pi_B(1-\pi_B)}\right]^{1/2}.$$

As noted by Pritchard and Przeworski [2001], since the pairwise LD measure r^2 between loci 1 and 2 is

$$r_{12}^2 = \frac{(q_{AB} - q_{aB})^2\pi_A(1-\pi_A)}{\pi_B(1-\pi_B)},$$

this is equivalent to

$$E(Z_2) = (\pi_{DA} - \pi_{CA})\left[\frac{2N\phi(1-\phi)}{\pi_A(1-\pi_A)}\right]^{1/2} \times \left[\frac{(q_{AB} - q_{aB})^2\pi_A(1-\pi_A)}{\pi_B(1-\pi_B)}\right]^{1/2}.$$

Therefore, the expected value of Z for locus 2 due to association at locus 1 is

$$E(Z_2|Z_1, r_{12}^2) = Z_1\sqrt{r_{12}^2}$$

or

$$E(\chi_2^2|\chi_1^2, r_{12}^2) = \chi_1^2 r_{12}^2.$$

Because the top association signals in our example are clear from scanning the figure, and can be easily gleaned by examining the association results themselves, we label a subset of SNPs on the y-axis of the figure when they meet two criteria: first, being in the top strata of association signals of n SNPs in the region of interest, where the default value is

$$\text{Threshold} = \frac{\max_{i=1}^{n}[\chi_n^2] + \min_{i=1}^{n}[\chi_n^2]}{1.5}.$$

We also provide a user option to specify a threshold.

Of the SNPs that pass the first criteria, we label those meeting or exceeding the expected association signal due to any neighboring SNP as

$$\chi_2^2 - \chi_1^2 r_{12}^2 \geq 0.$$

Finally, we note that our example association test compares differences in allele frequencies across groups, as is commonly reported in the literature, and not genotype frequencies. Allelic association tests are confounded with deviations from Hardy-Weinberg equilibrium [Sasieni, 1997], and several other tests have been shown to be more powerful than allelic association tests [Matthews et al., 2008; Zhang et al., 2008].