BMC Proceedings

Inference of disease associations with unmeasured genetic variants by combining results from genome-wide association studies with linkage disequilibrium patterns in a reference data set

David Hadley*† and David P Strachan†

Address: Division of Community Health Sciences, St George's, University of London, London, SW17 0RE, UK

E-mail: David Hadley* - d.hadley@sgul.ac.uk; David P Strachan - d.strachan@sgul.ac.uk
*Corresponding author   †Equal contributors

Abstract

Results from whole-genome association studies of many common diseases are now available. Increasingly, these are being incorporated into meta-analyses to increase the power to detect weak associations with measured single-nucleotide polymorphisms (SNPs). Imputation of genotypes at unmeasured loci has been widely applied using patterns of linkage disequilibrium (LD) observed in the HapMap panels, but there is a need for alternative methods that can utilize the pooled effect estimates from meta-analyses and explore possible associations with SNPs and haplotypes that are not included in HapMap.

By a weighted average technique, we show that association results for common SNPs in an observed data set can be scaled and combined to infer the effect of a genetic variant that has been measured only in an independent reference data set. We show that the ratio $p(R\text{-}1)/[1 + p(R\text{-}1)]$, where R is the relative risk associated with a measured or unmeasured allele of frequency p, is appropriately scaled by $1/D'$ and weighted in proportion to r^2, both common measures of LD being derived from the reference data set.

We illustrate this computationally simple method by combining the results of a genome-wide association screen from the North American Rheumatoid Arthritis Consortium with LD measures from the British 1958 Birth Cohort, and explore the validity of underlying assumptions about the generalizability of LD from one population to another, and from healthy subjects to subjects with clinical disease.

Background

The HLA allele *DRB1*04* has been shown to be more strongly associated with rheumatoid arthritis than nearby tagging single-nucleotide polymorphisms (SNPs) [1]. We propose here a method of inferring the effects of an unmeasured genetic variant (such as *DRB1*04*) using linkage disequilibrium (LD) measures from an independent reference data set (the British 1958

Birth Cohort) to scale and weight the associations of rheumatoid arthritis with tagSNPs in the North American Rheumatoid Arthritis Consortium (NARAC) data set, supplied as Problem 1 for the Genetic Analysis Workshop 16.

Methods

The British 1958 Birth Cohort (B58C) compromises all infants in England, Wales, and Scotland born in one week in 1958. During a follow-up in 2002 to 2004 [2], a cell-line-backed DNA collection was established as a nationally representative reference set for genetic case-control studies. Field protocols and consent forms were approved by the South East England Multi-Centre Research Ethics Committee. Genome-wide data from the Illumina HumanHap550 Beadarray on 1430 members of the B58C was deposited by the Wellcome Trust Sanger Institute [3]. In addition, data on HLA typing using Dynal technologies were deposited by the Diabetes & Inflammation Laboratory, Cambridge [4]. Further details about these deposits and the B58C DNA Collection are published online [5].

The NARAC data set [6], provided as Problem 1 for the Genetic Analysis Workshop 16, consists of individual-level genotype data for 550,000 SNPs tested on the Illumina HumanHap550 Bead Array linked to case/control status for rheumatoid arthritis as well as the HLA alleles at the *DRB1* locus. We derived a numerical score to represent the number of *DRB1*04* alleles for each individual in both the NARAC and B58C data sets. Association tests between the rheumatoid arthritis case/control status and each tagSNP in the MHC region (chromosome 6; 27-33 Mb [7]) were generated using Stata™ 9.2.

We tested whether combining the LD patterns from the B58C and the association patterns from NARAC provide an unbiased estimate the effect of the *DRB1*04* allele, using only neighboring tagSNPs (± 300 kb) that are common to both data sets. This application has the advantage that the *DRB1*04* allele (our target variant) has been measured directly in both data sets, permitting validity checks between our inferred relative risk estimate and the observed effect in the NARAC case-control study.

We define $\Phi_{obs} = q(R_{obs}-1)/[1 + q(R_{obs}-1)]$, where R_{obs} is the relative risk of disease associated with each copy of a measured allele of frequency q. The expression for Φ is similar to that for the population-attributable risk fraction (PARF) in epidemiological studies. However, in this genetic application, because each individual has two chromosomes, the PARF of a variant is $\Phi(2-\Phi)$. In the Appendix we show that when the tagSNP is more

common than the target variant, Φ_{obs}/D' is an unbiased estimator of Φ_{true}, where Φ_{true} is the equivalent parameter at the unmeasured target locus and D' is the conventional measure of LD between the measured tagSNP and the target variant in an undiseased population. We also show (see Appendix) that, for any given target variant and study design, the variance of Φ_{obs} for each tagSNP under the null hypothesis is inversely proportional to the LD measure r^2, relating the measured tagSNP to the unmeasured target variant in the undiseased population.

We calculated a weighted average of the values of $\Phi_{infer} = \Phi_{obs}/D'$ across all tagSNPs in the selected 600-kb region, and derived an inferred relative risk from this pooled estimate of Φ. In this paper we explore the effect of different assumptions about the generalizability of LD measures and selection of tagSNPs upon this point estimate. An empirical variance of this pooled estimate under the null hypothesis was derived from multiple random permutations of the B58C data set.

Results

Individually linked *HLA-DRB1* diplotypes and Illumina HumanHap550 genotypes were available for 1217 members of B58C and for 1187 controls and 799 cases from NARAC. The frequency of *DRB1*04* was 21.1% in the B58C reference set, 16.5% in the NARAC controls and 52.2% in the NARAC cases. Counting chromosomes among cases and controls, the observed odds ratio of rheumatoid arthritis per copy of the *DRB1*04* variant was 5.54 (95%CI 4.78-6.41) corresponding to a Φ of 0.4280.

All subsequent data relates to 156 SNPs common to both data sets where the minor allele frequency (MAF) of the tag is greater than that of *DRB1*04*, also less than 44% in the reference data set and $|D| > 0.01$, where D is the covariance measure of LD. These restrictions are imposed to remove SNPs where the effect allele could be inconsistent between the B58C data and the NARAC controls data.

For general application, the method needs to apply measures of LD derived from a reference data set to the observed data set where the target variant has not been measured. Figure 1 shows good correspondence for all measures of LD between B58C and NARAC controls. Our inference method assumes that the *s* measure of LD (see Appendix for definition) is constant between cases and controls. In this example, even where there is a strong association between the target variant and disease, this assumption appears valid, as shown in Figure 2. In most situations, the association of the target variant with
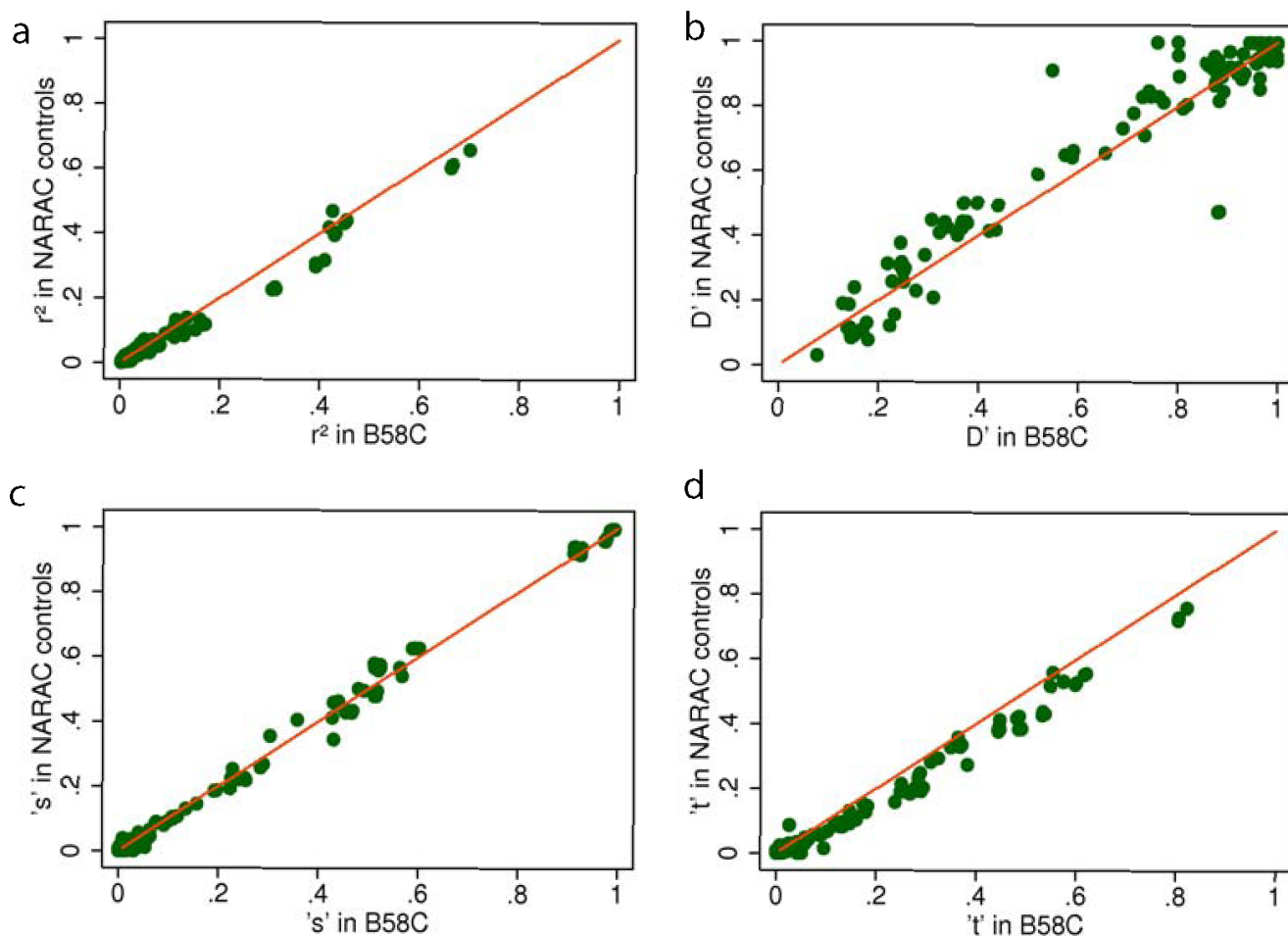
Figure 1
LD measures in NARAC controls versus B58C. Clockwise from top left: a) r^2, b) D', c) s, and d) t. See Appendix for description of s and t.

tagSNPs in the observed case-control data would be unknown and estimated from the reference data set (e.g., B58C). Using D' and r^2 from the B58C data, the pooled estimate of Φ is 0.4530, which is very close to the value of 0.4280 derived directly in the NARAC case-control data.

The variance of the pooled estimate of Φ is greater than the sum of the weights, because many of the neighboring tagSNPs are intercorrelated. The precision of this pooled estimate was estimated by performing 1000 random permutations of the B58C data set to derive the variance of the pooled estimate under the null hypothesis. These results are summarized in Table 1. Inclusion of all 156 SNPs generated a pooled estimate of similar magnitude and precision to that from the three most closely associated SNPs. On the other hand, if no closely associated tagSNPs had been available, results with similar magnitude and only slight loss of precision could be obtained from the majority of SNPs with

$r^2 < 0.3$. Indeed, in this example, from a region of high LD ($D' > 0.5$), useful information could be derived even from tagSNPs with $r^2 < 0.1$.

Conclusion

Comparing B58C to the NARAC controls, there was close correspondence in LD patterns. Within the NARAC data, however, the LD measures r^2 and t in the region of interest were markedly different in cases and controls, reflecting the strong association between the target variant (*DRB1*04*) and rheumatoid arthritis. When compared with imputation approaches, the inference method has the advantage of working only with the LD pattern among controls, although it does assume generalizability of the sensitivity parameter s from controls to cases. This latter assumption was supported in the NARAC data. The inference approach that we describe can, with modification, be applied to continuous outcomes and offers an alternative to imputation
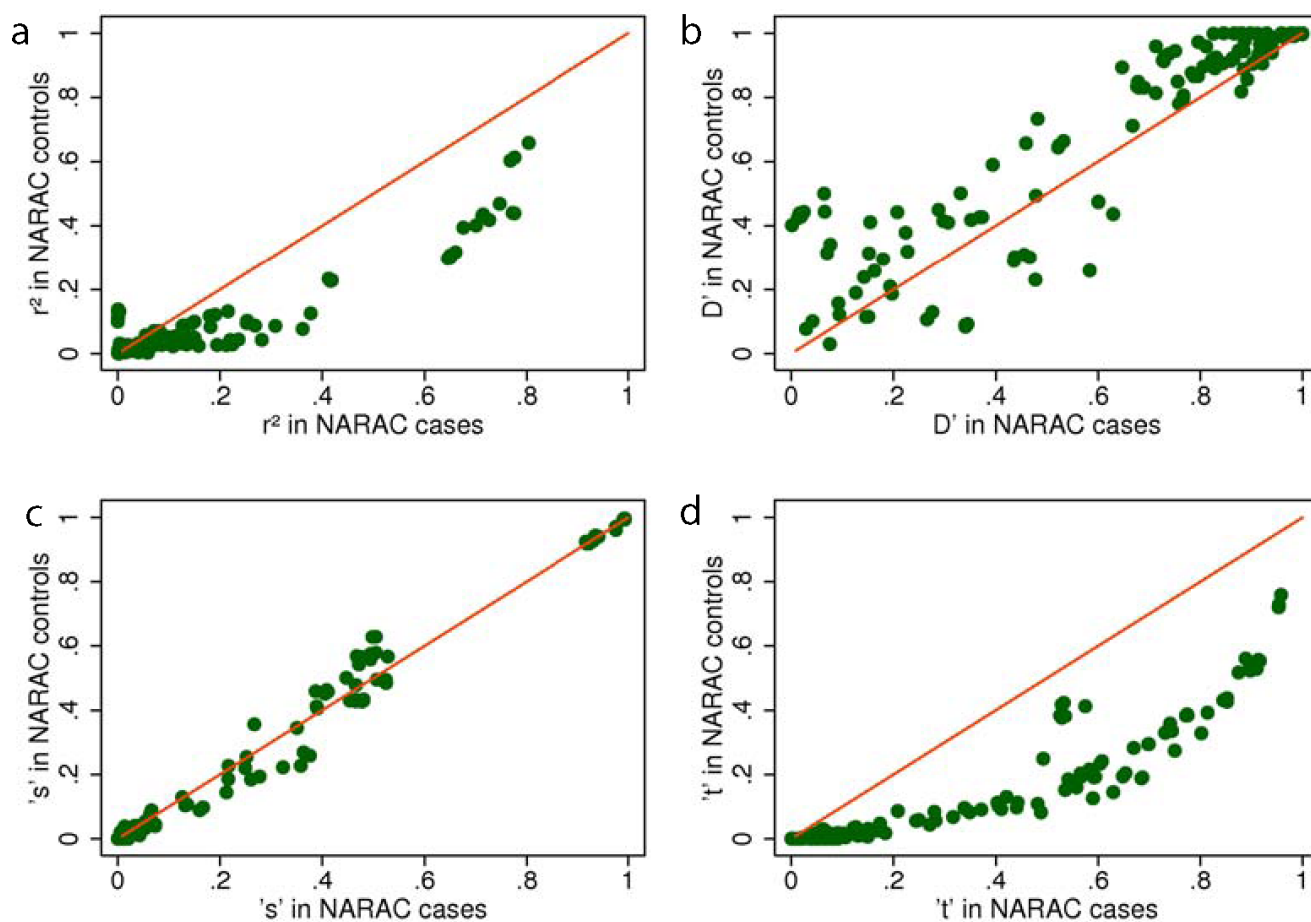
Figure 2
LD measures in NARAC controls versus case. Clockwise from top left: a) r^2, b) D', c) s, and d) t. See Appendix for description of s and t.

Table 1: Effect of different choices of tagSNP inclusion criteria on estimation of Φ

	Minimum r^2				Maximum r^2			
	0	**0.2**	**0.4**	**0.6**	**0.1**	**0.3**	**0.5**	**0.7**
Pooled Φ	0.453	0.4046	0.4065	0.4253	0.3768	0.5104	0.4569	0.4530
SE of pooled Φ	0.0258	0.0256	0.0256	0.0256	0.0352	0.0292	0.0263	0.0257
Number of tagSNPs used in estimate	156	21	12	3	102	135	153	156

Different combinations of tagSNPs from the NARAC case-control data set, selected by inclusion based on r^2 between the tagSNPs and the *DRB1*04* allele, give slightly different effect estimates (Φ). Also shown is the standard error of Φ under the null hypothesis, derived empirically from 1000 random permutations of the B58C data set.

methods that will be particularly attractive for target variants that are not part of standard panels such as HapMap, and for exploring further genome-wide sets of association measures that have been derived from meta-analysis of a particular disease or quantitative trait.

List of abbreviations used
B58C: British 1958 Birth Cohort; LD: Linkage disequilibrium; NARAC: North American Rheumatoid Arthritis Consortium; PARF: Population-attributable risk fraction; SNPs: Single-nucleotide polymorphisms.

Table 2: Haplotype frequencies and disease incidence for variants in positive LD

Measured tagSNP	Target variant (e.g., *HLA-DRB1*04*)				
	Minor allele	Major allele	Relative frequency	Incidence	Derivation
Minor allele	sp	$q-sp$	q	I_a	$I_a = I_b R_{obs}$
Major allele	$(1-s)p$	$1-q-(1-s)p$	$1-q$	I_b	$I_b = I_o [R(1-s)p+(1-q)-(1-s)p]/[1-q]$
Relative frequency	p	$1-p$	1		
Incidence	RI_o	I_o		I_{total}	$I_{total} = I_b [1+q(R_{obs}-1)]$ $I_{total} = I_o [1+p(R-1)]$

Appendix

Within a total population, the LD between each tagSNP and an unmeasured target variant may be represented by Table 2, based on relative counts of chromosomes. The parameter s represents the proportion of the target variant that also has the tagSNP variant (akin to sensitivity in screening terminology). Another measure, equivalent to the positive predictive value, is t ($= sp/q$), the proportion of chromosomes with the tagSNP allele that also have the target variant. We seek to infer a value for the true relative risk (R) from the measured relative risk (R_{obs}) associated with each tagSNP.

We define a transformation of R, as follows:

$$\Phi_{\text{true}} = p(R-1)/[1+p(R-1)] \text{ and } \Phi_{obs} = q(R_{obs}-1)/[1+q(R_{obs}-1)].$$

We show below that Φ_{obs}/D' is an unbiased estimator of Φ_{true}.

In Table 2: $I_b = I_o [R (1-s)p+(1-q)-(1-s)p]/[1-q]$,

so : $\begin{aligned} I_b / I_o &= [pR - spR - p + sp + (1-q)]/[1-q] \\ &= [p(R-1)(1-q) - p(R-1)(s-q) + (1-q)]/[1-q] \\ &= p(R-1)(1-D') + 1 \quad (\text{noting that } D' = (s-q)/(1-q)) \end{aligned}$

Also in Table 2: $I_{total} = I_o [1+p(R-1)] = I_b [1+q (R_{ob} -1)]$

so : $\quad I_b / I_o = [1 + q(R_{obs}-1)]/[1+p(R-1)](= p(R-1)(1-D')+1, \text{ above}).$
Thus: $\quad 1+q(R_{obs}-1) = [1+p(R-1)]/[1+p(R-1)(1-D')],$
and : $\quad q(R_{obs}-1) = D'[p(R-1)]/[1+p(R-1)(1-D')],$
So : $\quad \Phi_{\text{obs}} = D'[p(R-1)]/[1+p(R-1)] = D'\Phi_{\text{true}}.$

We infer values of $\Phi_{\text{infer}} = \Phi_{\text{obs}}/D'$ for each tagSNP and derive a weighted average, with weights inversely proportional to their variance under the null hypothesis. An approximate variance for Φ_{infer} can be derived by the Delta method by considering Φ_{obs} as a function of p, q, D' and $\beta = \ln(R_{obs})$, as follows, noting that:

$\begin{aligned} \Phi_{\text{obs}} &= \text{f}(\beta) &&= q(e^\beta - 1)/[1+q(e^\beta -1)] \\ & &&f'(\beta) = [qe^\beta]/[1+q(e^\beta - 1)]^2 = q \text{ under the null hypothesis } (R = R_{obs} = 1). \end{aligned}$

Also, if $R = R_{obs} = 1$, and c is the proportion of N chromosomes that come from cases:

$$\text{Var}(\beta) = 1/[q(1-q)] * [1/[c(1-c)]] * [1/N],$$

So : $\begin{aligned} \text{Var}(\Phi_{\text{infer}}) &= (1/D')^2 * (f'(\beta))^2 * \text{Var}(\beta) \\ &= (1/D')^2 * [q]^2 * [1/[q(1-q)]] * [1/[c(1-c)]] * [1/N] \text{ if } R = R_{obs} = 1 \\ &= (1/r^2) * [p(1-q)/[q(1-p)]] * [q/(1-q)] * [1/c(1-c)] * [1/N] \\ &= (1/r^2) * [p/(1-p)] * [1/[c(1-c)]] * [1/N]. \end{aligned}$

Inverse-variance weights proportional to the LD measure r^2 are therefore appropriate because p (the minor allele frequency for the target variant) is the same for all tag SNPs and c and N are fixed by the study design.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DH carried out the data analysis. DPS conceived the method. DH and DPS drafted, read, and approved the manuscript.

Acknowledgements

References

1. Newton JL, Harney SM, Wordsworth BP and Brown MA: **A review of the MHC genetics of rheumatoid arthritis.** *Genes Immun* 2004, **5**:151–157.
2. Strachan DP, Rudnicka AR, Power C, Shepherd P, Fuller E, Davis A, Gibb I, Kumari M, Rumley A, Macfarlane GJ, Rahi J, Rodgers B and Stansfeld S: **Lifecourse influences on health among British adults: effects of region of residence in childhood and adulthood.** *Int J Epidemiol* 2007, **36**:522–531.
3. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, Feighery C, Jewell D, Kelleher D, Kumar P, Travis S, Walters JR, Sanders DS, Howdle P, Swift J, Playford RJ, McLaren WM, Mearin ML, Mulder CJ, McManus R, McGinnis R, Cardon LR, Deloukas P and

Wijmenga C: **A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21.** *Nat Genet* 2007, **39:**827–829.

4. Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, Hulme J, Maier LM, Smyth D, Bailey R, Cooper JD, Ribas G, Campbell RD, Clayton DG, Todd JA and Wellcome Trust Case Control Consortium: **Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A.** *Nature* 2007, **450:**887–892.

5. **Genetic information from the British 1958 Birth Cohort.** http://www.b58cgene.sgul.ac.uk.

6. Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Monteiro J, Kern M, Criswell LA, Albani S, Nelson JL, Clegg DO, Pope R, Schroeder HW Jr, Bridges SL Jr, Pisetsky DS, Ward R, Kastner DL, Wilder RL, Pincus T, Callahan LF, Flemming D, Wener MH and Gregersen PK: **A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases.** *Am J Hum Genet* 2001, **68:**927–936.

7. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ, Trowsdale J, Wijmenga C, Vyse TJ, Beck S, Murray SS, Carrington M, Gregory S, Deloukas P and Rioux JD: **A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.** *Nat Genet* 2006, **38:**1166–1172.