Published in final edited form as:
Diabetes Obes Metab. 2009 February ; 11(Suppl 1): 92–100. doi:10.1111/j.1463-1326.2008.01009.x.

Association of MHC SNP genotype with susceptibility to type 1 diabetes: a modified survival approach

E. McKinnon[1], **G. Morahan**[2], **D. Nolan**[1], **I. James**[1], and **the Type 1 Diabetes Genetics Consortium**

[1]Centre for Clinical Immunology and Biomedical Statistics, Murdoch University and Royal Perth Hospital, Perth, Western Australia, Australia

[2]Western Australian Institute for Medical Research and Centre for Medical Research, University of Western Australia, Perth, Western Australia, Australia

Abstract

Aim—The Major Histocompatibility Complex (MHC) is a highly polymorphic region on chromosome 6 encompassing the human leucocyte antigen (*HLA*)-*DQ*/*DR* loci most predictive of susceptibility to type 1 diabetes (T1D). To assess the contribution of other MHC genes, in this exploratory analysis of Type 1 Diabetes Genetics Consortium (T1DGC) family data we characterize association between susceptibility and MHC single nucleotide polymorphism (SNP) genotype, with an emphasis on effects of genetic variation additional to carriage of predisposing or protective MHC haplotypes.

Methods—We use Cox regression analyses of age of onset, stratified by family, to jointly test both linkage and association. Analysis is restricted to children from families having both affected and unaffected siblings and is conducted with and without adjustment for known HLA class I and II effects. Model fits provide scores for each individual that are based on estimates of the probability of being affected by the age of 35, given the individual's SNP genotype. The mean within-family variation in these scores provides a measure that closely reflects the relative size of the likelihood ratio test statistics, and their covariation provides a means of mapping patterns of association that incorporate both effect size and commonality of effect that is attributable to the strong linkage disequilibrium (LD) extending across the region.

Results—Univariate analyses yielded strong associations with T1D susceptibility that are dominated by SNPs in the class II *HLA-DR*/*DQ* region but extend across the MHC. Similar effects are frequently observed across SNPs within multiple genes, sometimes spanning hundreds of kilobases. SNPs within a region at the telomeric end of the class II gene *HLA-DRA* yielded significant associations with and without adjustment for carriage of the predictive DR3, DR4, DR2 and DR7 HLA haplotypes, and remained highly prominent in a secondary analysis that was restricted to 66 families in whom at least one of the affected siblings carried neither the DR3 nor DR4 haplotype.

Conclusions—While many of the associations can be attributed to LD between the SNPs and the dominant *HLA-DRB*/*DQA*/*DQB* loci, there is also evidence of additional modifying effects.

Keywords

MHC fine mapping; stratified Cox survival model; type 1 diabetes

Correspondence: Dr Elizabeth McKinnon, Mathematics and Statistics, Murdoch University, Murdoch, WA 6150, Australia. e.mckinnon@murdoch.edu.au.

Conflict of interest: The authors do not declare any conflict of interest relevant to this manuscript.

Introduction

Type I diabetes (T1D) is a polygenic autoimmune disease with genes residing within the Major Histocompatibility Complex (MHC) ascribed to be the major determinants [1–4]. In particular, alleles of the human leucocyte antigen (HLA) class II genes, *HLA-DRB1* and *-DQA1*/*B1* are associated with the greatest risk of developing disease. Other susceptibility loci within the MHC have also been identified by both association and linkage studies as potentially contributing to T1D aetiology [5–9]; however, the heterogeneous nature of the disease and the strong linkage disequilibrium (LD) across the HLA region makes determination of causal variants very difficult.

In this study of data collected by the Type 1 Diabetes Genetics Consortium (T1DGC) [10], we have undertaken a series of analyses to investigate the degree to which observed effects simply reflect those of the known class II susceptibility genes because of LD or whether they identify regions with additional or modifying effects. As we considered these analyses to be of an exploratory nature, we have focused specifically on available single nucleotide polymorphism (SNP) genotype data rather than undertaking more sophisticated analysis that might require some degree of data imputation.

We have used Cox regression modelling [11,12] of age of onset as the basis for these exploratory analyses. There are several reasons for choosing to use this survival approach. First, the notion of censorship accommodates the uncertainty of event definition, in this instance the occurrence of disease. We do not explicitly model disease status but, instead, consider unaffected individuals as simply ‘not-yet-affected’ and right censor the age of onset for these individuals at the age of data collection. Second, the model framework provides a range of formulations, choice of which depends on the hypothesis being addressed and is flexible enough to account for inherent structures within the data. For example, by stratifying on family groups, we effectively conduct a test of both linkage and association. The approach also facilitates a number of outcome measures on which one could focus. These include relative rates of disease onset; expected number of years disease free and the proportions remaining disease free indefinitely. Restriction of analyses to only those families having both affected and unaffected children provide estimates of the effect that reflect the relative proportions in the sample remaining disease free (the notion of ‘immune proportions’ [13]) rather than relative rates of disease onset among the affected siblings. Finally, visual displays in the form of Kaplan–Meier survival plots provide particularly useful group summaries of genotype effect.

Materials and Methods

Study Design

The study comprises a series of family-based analyses of the associations of SNP genotype with susceptibility to T1D. Data are obtained from the T1DGC, which has recruited over 2300 families from nine cohorts worldwide. The analyses are based on Cox regression modelling of the effects of genotype on age of onset. We analyse data from affected sib-pair families, that also have at least one unaffected child (with all children having the same parentage), and restrict to the six cohorts reporting age of onset for affected siblings and age of collection for the unaffected siblings. Families are considered informative for a marker if not all siblings in the one family are identical by state.

Data and Analyses

The fine mapping data used in this study spanned the MHC (approximately 4 Mb, from Build 34 positions 29299430 to 33811013) and was genotyped using two 1536 SNP chip panels (with 115 SNPs in common) by the Wellcome Trust Sanger Centre. SNPs included for analysis

required at least 15 informative families and a call rate of at least 95% (calculated as the proportion of non-missing values within cohorts with the SNP typed), and individuals having less than 90% of SNPs were not considered eligible for analysis. The final data set comprised a total of 2497 SNPs, with data from 1839 individuals over 544 families containing both affected and unaffected siblings (table 1). For inclusion in analysis of any single marker, families were required to have at least three siblings with non-missing data. HLA-DR haplotypes are defined according to carriage of *DRB1*, *DQA1* and *DQB1* alleles, using nomenclature for class II MHC haplotype assignment described by Pugliese and Eisenbarth [9].

Protective haplotypes – DR2: DRB1*1501/DQA1*0102/DQB1*0602; DR7: DRB1*0701/DQA1*0201/DQB1*0303. Predisposing haplotypes – DR2: DRB1*1601/DQA1*0102/DQB1*0502; DR3: DRB1*0301/DQA1*0501/DQB1*0201; DR4: DRB1*04/DQA1*0301/DQB1*0302.

For illustrative purposes, Kaplan–Meier plots are derived from siblings of informative families and grouped according to genotype with no account taken of within-family correlation. Quantitative analysis was based on Cox models, stratified by family, with genotype effects incorporated under the proportional hazards assumption. We are modelling within-family shifts from underlying family-specific hazards. We note that, as all family members belong to the same cohort, there are no adjustments for cohort effects although these may be subsumed into the shape of the baseline hazards. Significance of an SNP results from sufficient families having discordant alleles at this SNP and consistency of any effect across these families.

Methodology

Under the family-stratified Cox model, the probability of remaining disease-free beyond age *t* for the *i*th individual of the *j*th family is given by:

$$S_{ij}\left(t\middle|x_{ij}\right)=\exp\left(-\exp\left(\beta_1 x_{ij1}+\beta_2 x_{ij2}+\mathrm{K}\right)\Lambda_j(t)\right),$$

where $\Lambda_j(t)$ is the baseline cumulative hazard function for that family and $x_{ij} = (x_{ij1}, x_{ij2}, \ldots)$ is a vector of covariate values incorporating SNP genotype [12]. For the univariable SNP analyses, '*x*' takes a single value according to the individual's genotype: a count of 0, 1 or 2 major alleles as defined in the database documentation. The baseline hazard is not specified in analyses but can be estimated *post hoc*. We use the (partial) likelihood ratio statistic as the measure of test significance. Given the family stratification, we note that only families with non-identical genotype across siblings contribute to this statistic. Therefore, the significance of the test will reflect a combination of both the number of these informative families and the consistency and size of any effect observed across them.

To focus on susceptibility rather than relative rates of disease onset, as is the convention in survival modelling, we obtain estimates of the probability of remaining disease-free beyond the age of 35 (given the sampling scheme). In particular, we estimate $S_{ij}(35|\text{genotype})$ for each SNP and for each individual, including those from non-informative families, to obtain a set of scores $\{S_{ijk}\}$ where *k* refers to the *k*th SNP. These scores are then used as a measure of SNP (co)-effect on susceptibility by defining V_{kl} as the mean over the n_F families, of the within-family covariances between the *k*th and *l*th SNP scores:

$$V_{kl}=n_{\mathrm{F}}^{-1}\sum_{j}\mathrm{Cov}(S_{jk},S_{jl}).$$

When $k = l$, we obtain the average variance of the scores of the kth SNP, a measure that is highly correlated with the corresponding likelihood ratio statistic of the model from which the scores were derived. Furthermore, we note that for $k \neq l$, Cov(S_{jk},S_{jl}) can be written as the product of the square roots of the respective variances and the correlation between the scores of the two SNPs. The first of these terms captures the effect sizes of the respective SNPs and the second term reflects the correlation between their genotype that will be greater for SNPs in tighter LD. The covariance score therefore acts as a measure of commonality of effect that is useful for defining significant markers into haplotype blocks.

Results

Initial analysis examined the relative effects of HLA-DR genotype and yielded results consistent with DR3/DR4 heterozygosity conferring greatest risk of developing diabetes followed by DR4 and DR3 homozygosity and with decreased susceptibility associated with DR2 or DR7 carriage (table 2). These effects are evident in the survival plots (figure 1) derived from the 417 families with non-identical DR genotype. Among these individuals, diabetes developed by the age of 35 in an estimated 94% of those carrying DR3/DR4 compared with approximately 10% of those carrying DR2 or DR7 and 49% of those individuals carrying neither the predisposing nor protective haplotypes.

We next undertook a series of SNP analyses based on univariate Cox models with concurrent examination of the Kaplan–Meier survival plots (examples in figure 2). The dominant effects were spread over the HLA class II region (figures 2c and 3a), with the highest peaks corresponding to SNPs flanking *HLA-DQB1* (e.g., rs6927022, rs2157051, rs9275184 and rs7744001) and within and telomeric to *HLA-DRA* (e.g., rs3135335, rs2395178 and rs3129871) (figure 2b). While not quite as significant because of the relatively few informative families (n = 70), also visually striking effects occurred at SNPs located between *DRB1* and *DQA1* (rs3135005 and rs9271366), with presence of the minor allele marking the protective DR2 haplotype. SNPs across the class III region were also very significant, particularly in and close to the genes *C6orf10* and *NOTCH4*.

Examination of the survival plots revealed marked similarities in significant effects across SNPs. Sometimes the similarities occurred between adjacent SNPs, but often they were more interspersed. Notably, observed patterns frequently extended over several genes and sometimes even over hundreds of kilobases (figure 2a). These observations motivated the derivation of the score based on the covariances of model-predicted values: SNPs with a larger effect would be expected to yield predicted values with a larger spread (away from the baseline family estimate) than those obtained from a non-significant SNP, and for SNP pairs with correlated allelic values (i.e. for SNPs in LD) the respective predicted values would also be similarly correlated.

We used these scores to obtain a measure of the independence of observed effects. For readability, the heat map of figure 3c has been restricted to those SNPs with at least 50 informative families and either a p value less than 10^{-25} in the univariate analysis or less than 10^{-3} in both the univariate analysis and in analysis with adjustment for DR3/DR4 carriage (figure 4a). Noting the high correlation between the variance scores and the test statistics (figure 3a, b; R > 0.995), we have categorized the scores with cut-offs determined by estimates obtained from regressing the variance scores on the test statistics. The heat map reveals high covariance between many of the class III SNPs reflecting similarly strong effects seen across

multiple genes in strong LD as illustrated in figure 2a. Maintenance of covariance size is seen to extend well into the class II region, in line with these effects being largely haplotypic.

It would appear that the strong effects in the *DRA* region are not simply the result of LD with the *DRB1/DQ* region. This is supported by the results from analyses adjusted for DR3/DR4 carriage (figure 4a, b, vii) where, despite an overall reduction in significance (as expected), there remains a dominant peak around *DRA*. Moreover, there is little abrogation of the effects at the telomeric end of this region when adjustment is made for the other predictive DR haplotypes (figure 4b, ii–iv). This is in contrast to many of the other SNPs centromeric to *TXNB* that lose significance when taking account of the protective DR2 haplotype in particular (figure 4b, iii). Although a number of the class III SNPs telomeric of *TXNB* remain significant with adjustment, others appear to be markers of *HLA-B* alleles.

To further focus on effects additional to those conferred by carriage of the dominant DR haplotypes, we then restricted analysis to those families in whom at least one of the affected siblings carried neither DR3 nor DR4 (figure 5). While the relatively few families in this analysis (n = 66) results in a general reduction of power, there are a number of regions across the MHC yielding significant associations (figure 5a), the most striking of these forming a peak occurring at the telomeric end of *HLA-DRA* that coincides with that observed in the unrestricted analysis. There is also again evidence of block effects across class III SNPs reflecting strong LD, and we note that a number of genes contain SNPs with significant associations in both sets of analyses. Although many of the class I and class II SNPs appear to be markers of at least one of the two HLA class I alleles that were found to be significant in univariate analyses, *-A*24* and *-B*39*, there were others where the association held with these adjustments as well as in models that took account of DR2 or DR7 carriage. Notably, there was only moderate abrogation of the dominant *DRA* effects in any of the adjusted models.

Discussion

Properly regulated expression of class II molecules is vital for maintenance of sound immune function that requires that selective immunity to foreign pathogens and antigens be balanced against tolerance to self-antigen [14]. Hence, class II genes have been implicated in the occurrence of various cancers and autoimmune diseases, including T1D. While the *HLA-DRB1*/*DQ* loci have been most highly associated with onset of T1D, both the primary and family-restricted analyses presented here indicated HLA-DRA may play a secondary role. Despite the close physical proximity to the dominant genes and the strong LD in the class II region, we have shown the associated risk of *DRA* is at least partly additional to that conferred by carriage of the known high-risk and protective DR haplotypes defined by the *DRB1*/*DQ* loci.

Other modifying effects are also evident across the MHC, particularly in the class III region. However, the observation of very similar effects spanning multiple genes highlights the difficulties of isolating a single gene as the prime determinant. Examination of the survival plots also serves as a reminder that, as in any test of this nature, ranking the strength of associations on the basis of p values means large absolute effect sizes will be down-weighted if there are fewer informative families.

By conditioning on family membership, we are able to ignore higher level cohort effects. Haplotypic trends will emerge if the common effects across the component SNPs are evident in enough families to be deemed significant, irrespective of the demographic origins of the contributing families. Further characterization could examine the across-cohort distribution of the observed effects, both at the single and multi-SNP levels.

The size of the T1DGC data set from which we drew our samples has meant that we have been able to undertake analyses on restricted subsets of the data and still retain sufficient power to detect meaningful effects. The adoption of a survival approach, which imposed restrictions on the cohorts and families included in the primary analysis, was motivated by the accompanying survival plots that provided such a useful visual display of effects. The analyses should be viewed as complementary to more traditional linkage/family-based approaches that use more of the available data. In contrast to conventional survival analysis where inferences about relative risks and survival probabilities are extrapolated back to a more general population, we have restricted to families having both affected and unaffected siblings as a means of identifying SNPs with good discrimination. This is analogous to the well-known sib-disequilibrium test but takes account of the uncertainty of affection status. Furthermore, restricting to smaller but more homogeneous samples aids with interpretation of results. In particular, here we restricted to families where disease status is not directly attributable to carriage of the high-risk DR3 or DR4 haplotypes to facilitate the focus on this select subgroup. Loss of power because of the relatively small number of families fitting the criteria does not compromise veracity of the findings which were quite strong, given the small sample size, and are more readily interpretable than those obtained from alternative more complex approaches that incorporate model-dependent adjustments or conditioning.

The covariance score presented here has been developed with the aim of jointly identifying discriminatory markers and blocks of SNPs with linked effects on outcome. It is readily applicable to other regression-based methods that have the ability to provide model-based individual-specific fitted values such as, for example, conditional logistic regression models. The score we have derived provides a measure that captures both effect size and commonality of effect across a region of interest. As it has a set of univariable analyses as its basis, it provides a means of significantly reducing computational effort in haplotype and model-building exercises. We therefore believe the approach can be further developed as a useful tool for enhancing understanding of the complex interplay of genetic factors contributing to susceptibility to multifactorial diseases such as T1D.

Acknowledgments

This research used resources provided by the T1DGC, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418.

References

1. Nerup J, Platz P, Anderson OO, et al. HL-A antigens and diabetes mellitus. Lancet 1974;2:864–866. [PubMed: 4137711]
2. Cudworth AG, Woodrow JC. Evidence for HL-A-linked genes in ‘juvenile’ diabetes mellitus. Br Med J 1975;3:133–135. [PubMed: 1139259]
3. Morton NE, Green A, Dunsworth T, et al. Heterozygous expression of insulin-dependent diabetes mellitus (IDDM) determinants in the HLA system. Am J Hum Genet 1983;35:201–213. [PubMed: 6573128]
4. Todd JA. Genetic analysis of type 1 diabetes using whole genome approaches. Proc Natl Acad Sci U S A 1995;92:8560–8565. [PubMed: 7567975]
5. Todd JA, Mikovic C, Fletcher J, et al. Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. Nature 1989;338:587–589. [PubMed: 2494458]
6. Cucca F, Muntoni F, Lampis R, et al. Combinations of specific DRB1, DQA1, DQB1 haplotypes are associated with insulin-dependent diabetes mellitus in Sardinia. Hum Immunol 1993;37:85–94. [PubMed: 8226139]

7. Erlich HA, Zeidler A, Chang J, et al. HLA class II alleles and susceptibility and resistance to insulin dependent diabetes mellitus in Mexican-American families. Nat Genet 1993;3:358–364. [PubMed: 7981758]

8. Noble JA, Valdes AM, Cook M, Klitz W, Thomson G, Erlich HA. The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. Am J Hum Genet 1996;59:1134–1148. [PubMed: 8900244]

9. Pugliese, A.; Eisenbarth, GS. Type 1 Diabetes mellitus of man: genetic susceptibility and resistance. In: Eisenbarth, GS., editor. Type 1 Diabetes: Molecular, Cellular and Clinical Immunology. Vol. Chapter 7. Springer; New York: 2004. p. 170-203.

10. Rich S, Concannon P, Erlich H, et al. The Type 1 Diabetes genetics consortium. Ann N Y Acad Sci 2006;1079:1–8. [PubMed: 17130525]

11. Cox DR. Regression models and life-tables. J R Stat Soc Ser B 1972;34:187–220.

12. Kalbfleisch, D.; Prentice, RL. The Statistical Analysis of Failure Time Data. Wiley & Sons; USA: Hoboken, NJ: 1980.

13. Maller, RA.; Zhou, X. Survival Analysis with Long-term Survivors. New York: Wiley; 1996.

14. Ting J, Trowsdale J. Genetic control of MHC Class II expression. Cell 2002;109:S21–S33. [PubMed: 11983150]

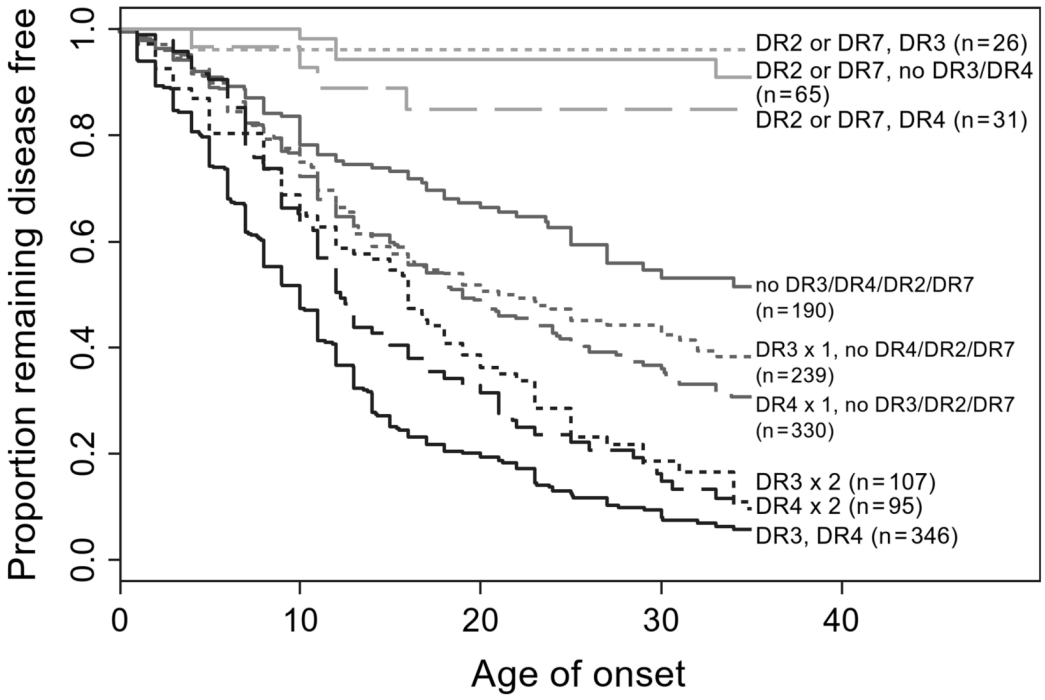

Fig. 1.

Kaplan–Meier survival plots of age of onset derived from siblings of informative families (n = 417) with stratification according to carriage of DR3, DR4 or DR2 (protective) or DR7 haplotypes.

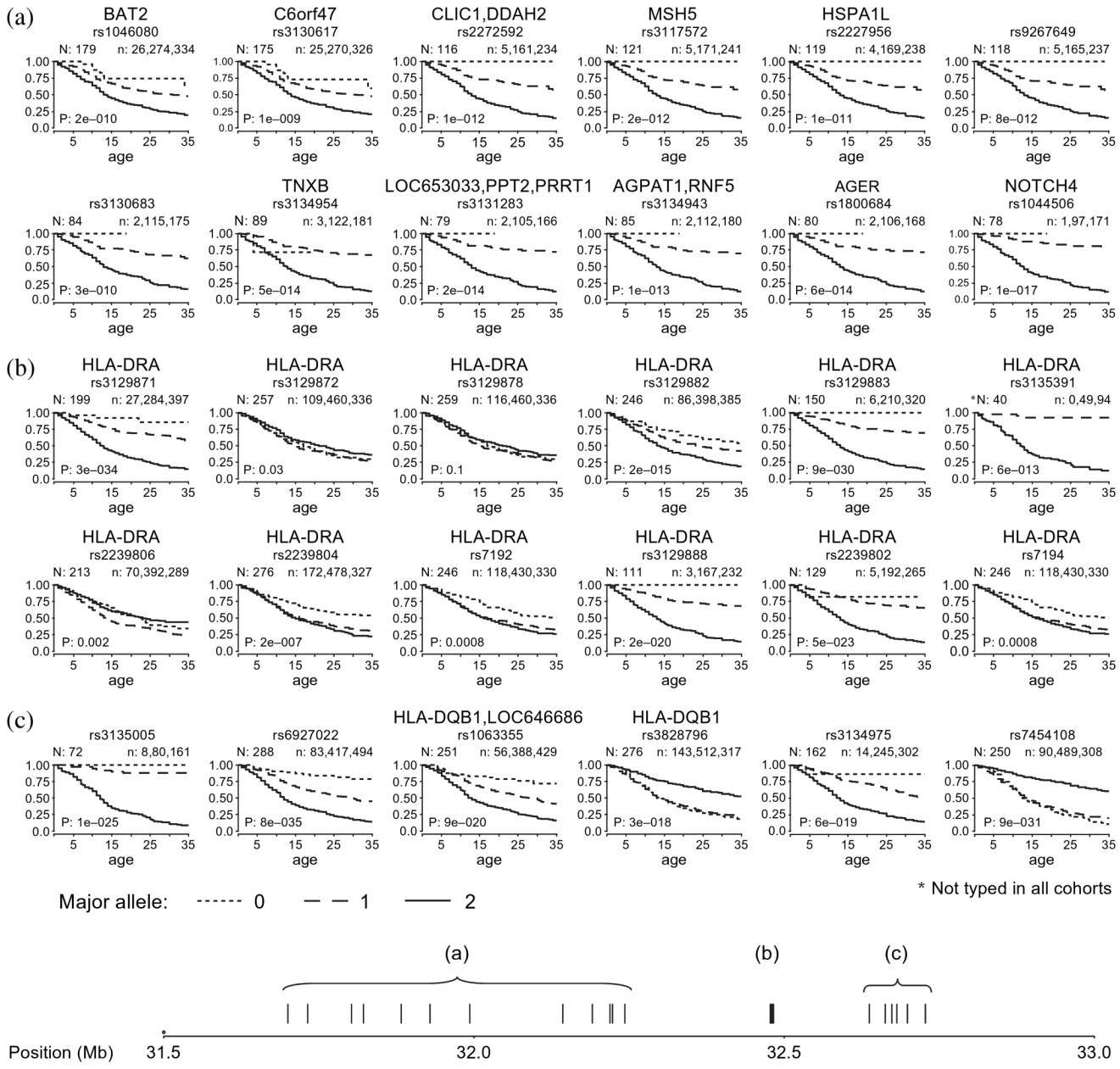

Fig. 2.

Kaplan–Meier survival plots of age of onset over informative families grouped by single nucleotide polymorphism (SNP) genotype. N, number of families included in analysis; n, number of individuals with 0, 1 and 2 copies of the major allele, respectively; P, p value of likelihood ratio statistic of Cox model, stratified by family. The plots illustrate (a) similarities in effects extending across multiple SNPs and multiple genes; (b) both similarities and variation between SNPs within a single gene; (c) strong effects evident within the *DRB1*/*DQA1*/*DQB1* region.

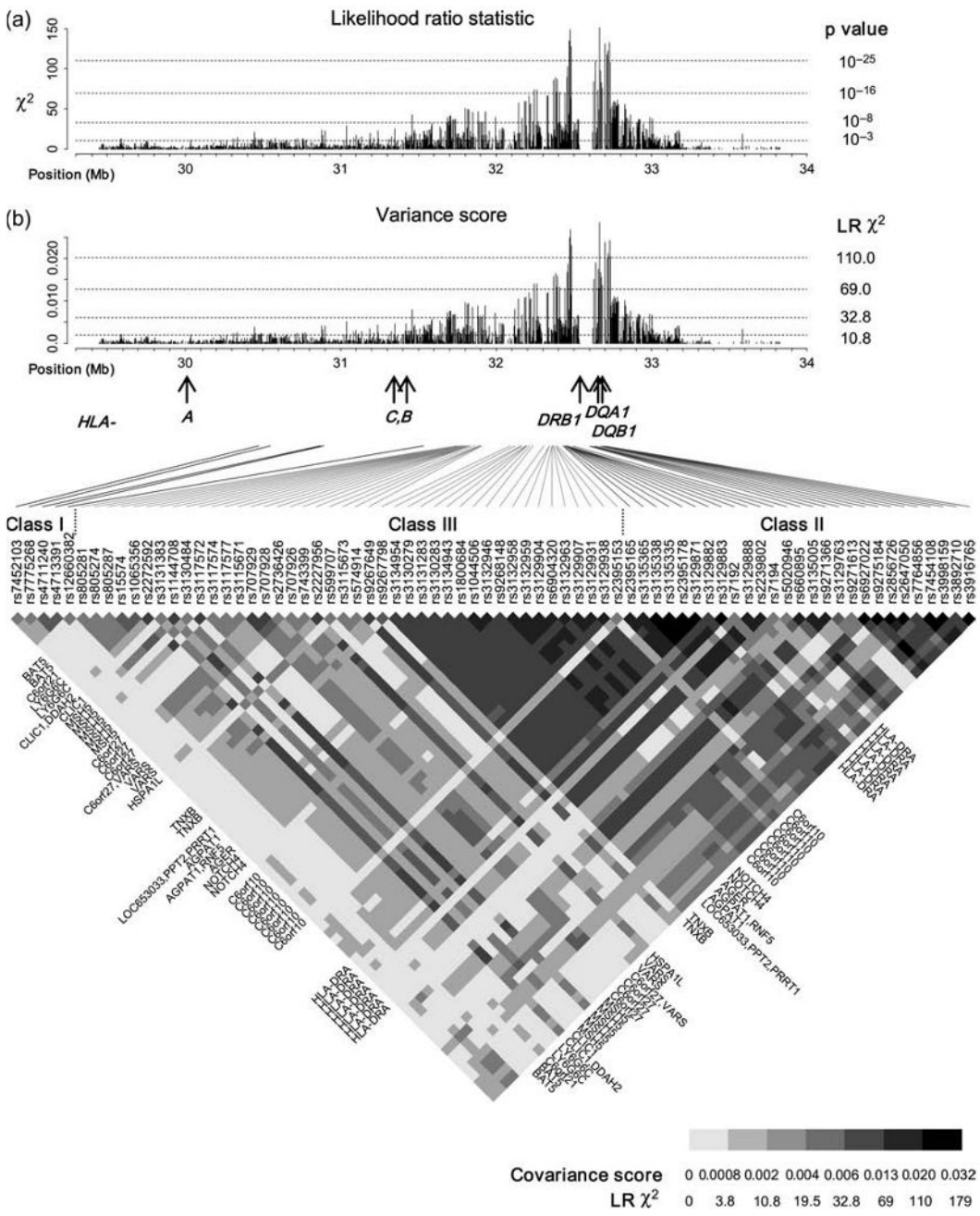

Fig. 3.

Results from Cox models fitting single nucleotide polymorphism (SNP) genotype, stratified by family: (a) the likelihood ratio statistics; (b) the corresponding variance scores; (c) heat map of covariance scores over SNPs either having $p < 10^{-25}$ or $<10^{-3}$ in both the univariate analysis and with adjustment for DR3/DR4 carriage. The variance/covariance scores are calculated as the across-family mean covariances of model-predicted values of *S*(35|genotype).

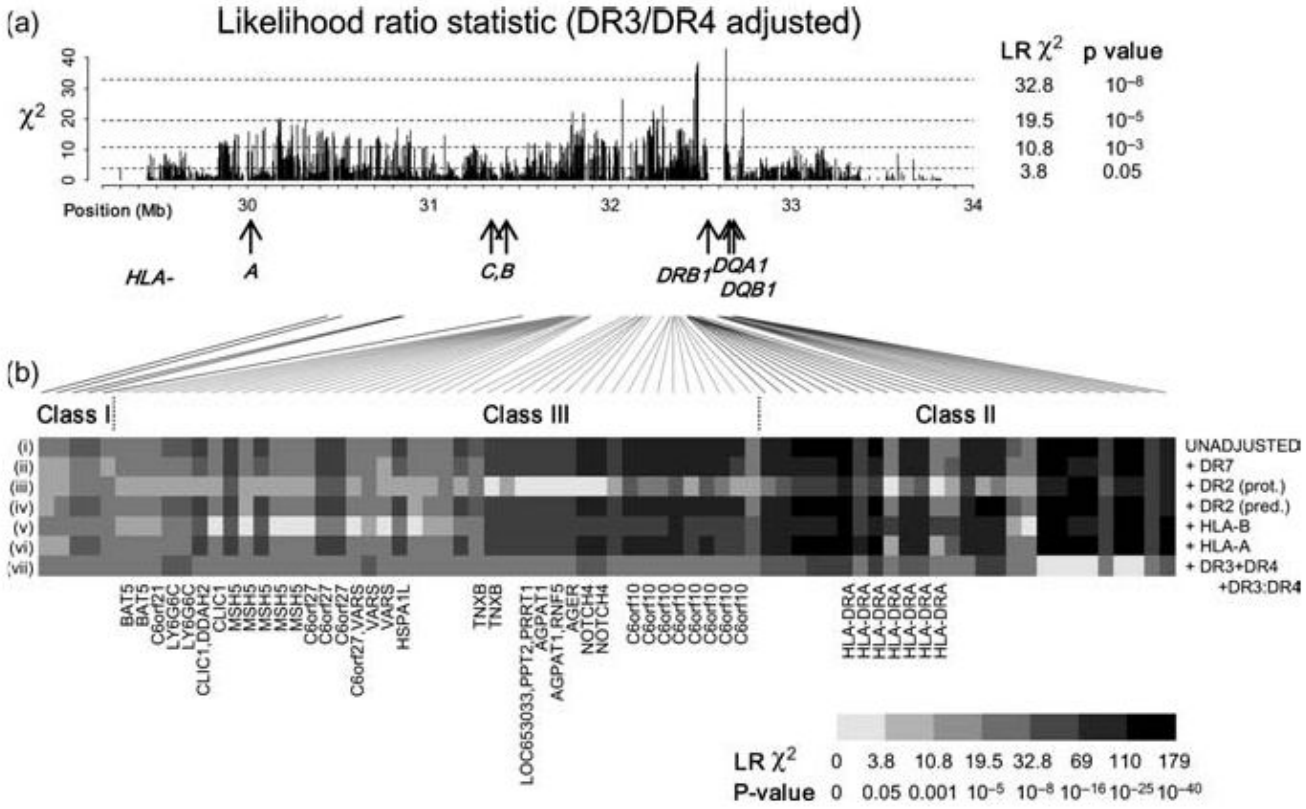

Fig. 4.

Likelihood ratio (LR) test statistics for single nucleotide polymorphism (SNP) genotype effect after human leucocyte antigen (HLA) adjustment. (a) Analyses across all SNPs with adjustment for DR3/DR4 carriage; (b) categorized LR statistics from adjusted analyses over those SNPs with at least 50 informative families and either $p < 10^{-25}$ in univariate analysis or $<10^{-3}$ in both the univariate analysis and after adjustment for DR3/DR4 carriage: (i) unadjusted; (ii)–(vii) with adjustment for carriage of (ii) DR7 (iii) protective and (iv) predisposing DR2 haplotypes, (v) *HLA-A* alleles significant in univariate analyses (A*11 + A*24 + A*30), (vi) significant *HLA-B* alleles (B*07 + B*08 + B*18 + B*39 + B*44 + B*51 + B*57), (vii) DR3/DR4 carriage (DR3 + DR4 + DR3:DR4).

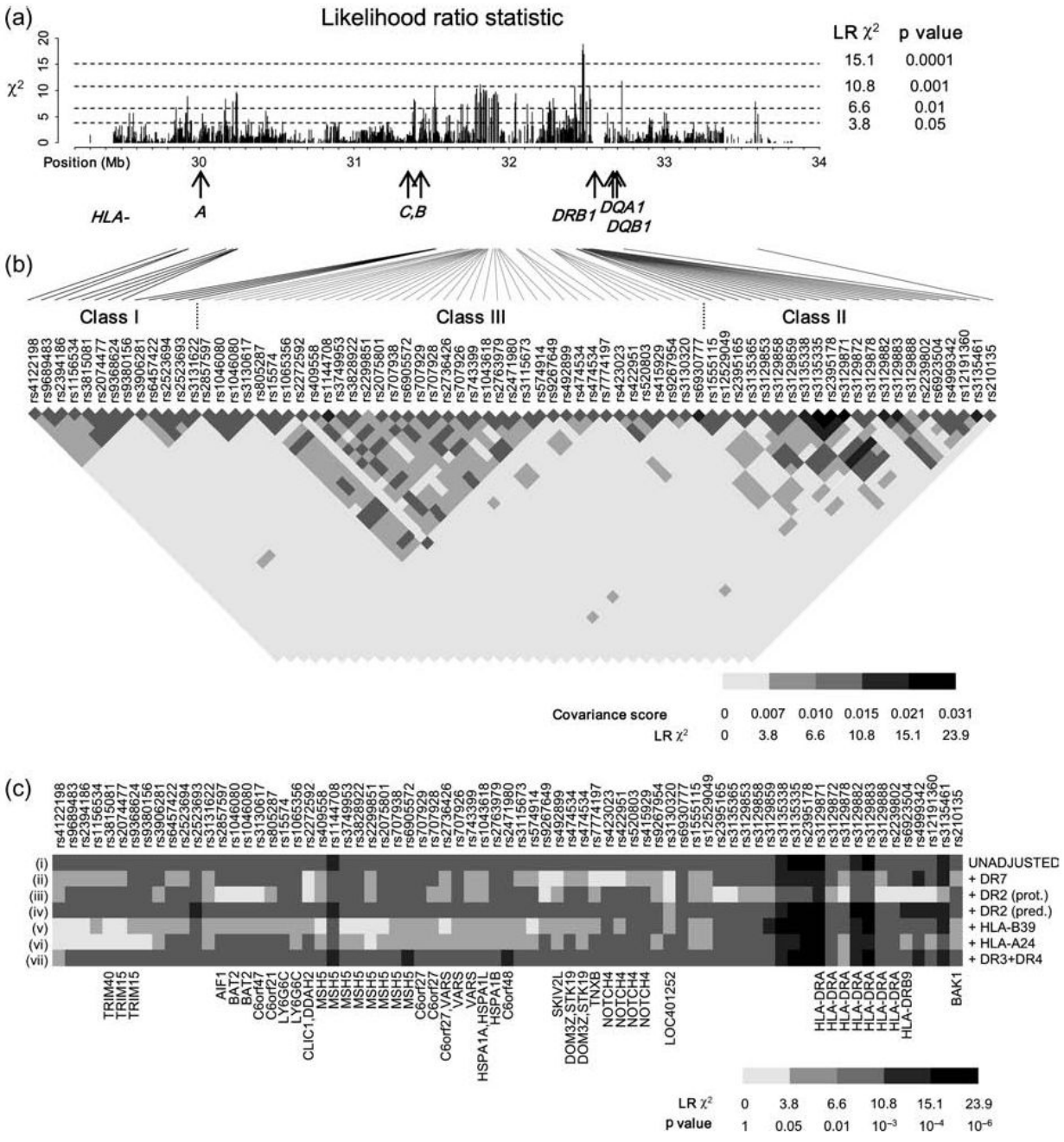

Fig. 5.

Results from analyses restricted to the families (n = 66) in whom at least one affected sibling carried neither the DR3 nor DR4 haplotype. (a) Likelihood ratio (LR) statistics across all single nucleotide polymorphism (SNPs) having at least 15 informative families; (b) heat map of covariance scores across those SNPs with a $p < 0.01$; and (c) categorized LR statistics: (i) unadjusted; (ii)–(vii) with adjustment for carriage of (ii) DR7 (iii) protective and (iv) predisposing DR2 haplotypes, (v) HLA-A*24, (vi) HLA-B*39 and (vii) DR3 or DR4 carriage (DR3 + DR4).

Table 1

Demographics of families satisfying selection criteria

	All families	Families with ≥1 affected sibling carrying neither DR3 nor DR4
Families (n)	544 (84% Caucasian)	66 (77% Caucasian)
Australia-Pacific	85	10
Danish	75	11
European	172	24
North American	134	16
Sardinian	43	2
United Kingdom	35	3
Individuals (n)	1839	220
Affected by age 35	1100	133
Censored before 35	547	60
Diabetes-free beyond 35	192	27
DR haplotype carriage (n)		
DR2 (protective)	83	11
DR2 (predisposing)	17	14
DR3	943	28
DR4	1067	38
DR7	33	9

Table 2

Cox model of age of onset, factored according to human leucocyte antigen DR haplotype

	Relative hazard	95% CI
no DR3/DR4/DR2(prot.)/DR7	1	–
DR2(prot.) or DR7, no DR3/DR4	0.12	0.04–0.35
DR2(prot.) or DR7, DR3	0.18	0.02–1.35
DR2(prot.) or DR7, DR4	0.21	0.04–1.08
DR3 × 1, no DR4/DR2(prot.)/DR7	2.61	1.71–3.98
DR4 × 1, no DR3/DR2(prot.)/DR7	3.30	2.18–4.97
DR3 × 2	7.54	4.09–13.91
DR4 × 2	14.09	7.40–26.82
DR3, DR4	15.94	9.90–25.69