Chapter 4

Analysis of Graves Disease Association Study Data

4.1 CTLA4 and Autoimmune Disease

Autoimmune disorders, such as Type 1 Diabetes and Graves Disease, together affect up to 3% of the UK population (Vaidya and Pearce, 2004). These diseases result from malfunctions in the immune system, causing intolerance to self. Autoimmune disorders are known to cluster in families, although the specific disease often varies between individuals in the family (Lesage and Goodnow, 2001). This clustering of autoimmune disorders along genetic lines suggests that they share common factors. It is expected that there are multiple genetic factors interacting with enviromental factors that affect susceptibility.

Along with the MHC, the CTLA4 gene has emerged as a convincing susceptibility locus for these diseases (Vaidya and Pearce, 2004). Initially, disease association was mapped to chromosome 2q33, which contains a number of T-lymphocyte regulatory genes, which are logical candidates for disease risk. A subsequent fine-mapping study of association between polymorphisms in 2q33 and Graves Disease (Ueda et al., 2003) showed a strong association between SNPs in the CTLA4 gene and Graves Disease.

In the Ueda et al. (2003) study, a 300kb region (CD28-CTLA4-ICOS) was genotyped for 108 SNPs in 652 control individuals and 384 Graves disease cases. In order to identify novel

SNPs, 32 individuals were resequenced, hence it is reasonable to assume that all common polymorphisms in the region, in the UK population, have been identified. In their analysis, three association peaks were identified; moving from left to right in Figure 4.1, these peaks are at SNPs MH30, CT60 and CTBC217_1. By performing a regression analysis, they concluded that the causative variant is more likely around the CT60 peak than the others. Around the CT60 peak there are three other SNPs, JO31, JO30 and JO27_1, which are also strongly associated, but their analysis was unable to further dissect the signal.

They performed functional studies to follow up the finding, and showed that the associated haplotype at CTLA4 is correlated with lower mRNA levels. In the non-obese diabetic mouse model, it was also shown that there was reduced production of CTLA4.

I applied MARGARITA to the data in order to see whether the CTLA4 signal could be further dissected.

4.2 ARG Analysis of the CTLA4 data

Since the data is unphased and has missing genotypes, I used MARGARITA to infer these (in inferring 100 ARGs, 100 different phase resolutions are obtained, thus marginalising over phase uncertainty when performing the mapping test). However, unphased data will present a hurdle for mapping methods that require phased haplotype sequences. One way to overcome this is to run a phasing algorithm (Marchini et al., 2006) on the data and then pass the result to the mapping method as though it is the true phase resolution (Morris et al., 2004). To examine the effect of this, I also ran MARGARITA on the "best" phase resolution of the data after applying one run of the program PHASE (Stephens et al., 2001), where "best" is defined as the most likely phase resolution found.

Figure 4.1 shows that CT60 has the strongest disease association in my analysis (both when using the PHASEd and unphased data), agreeing with Ueda et al. (2003)'s analysis. All MARGARITA P-values in this chapter were calculated by performing up to 1 million permutations.

MARGARITA on the unphased data gives a stronger association signal at CT60 ($P \approx$

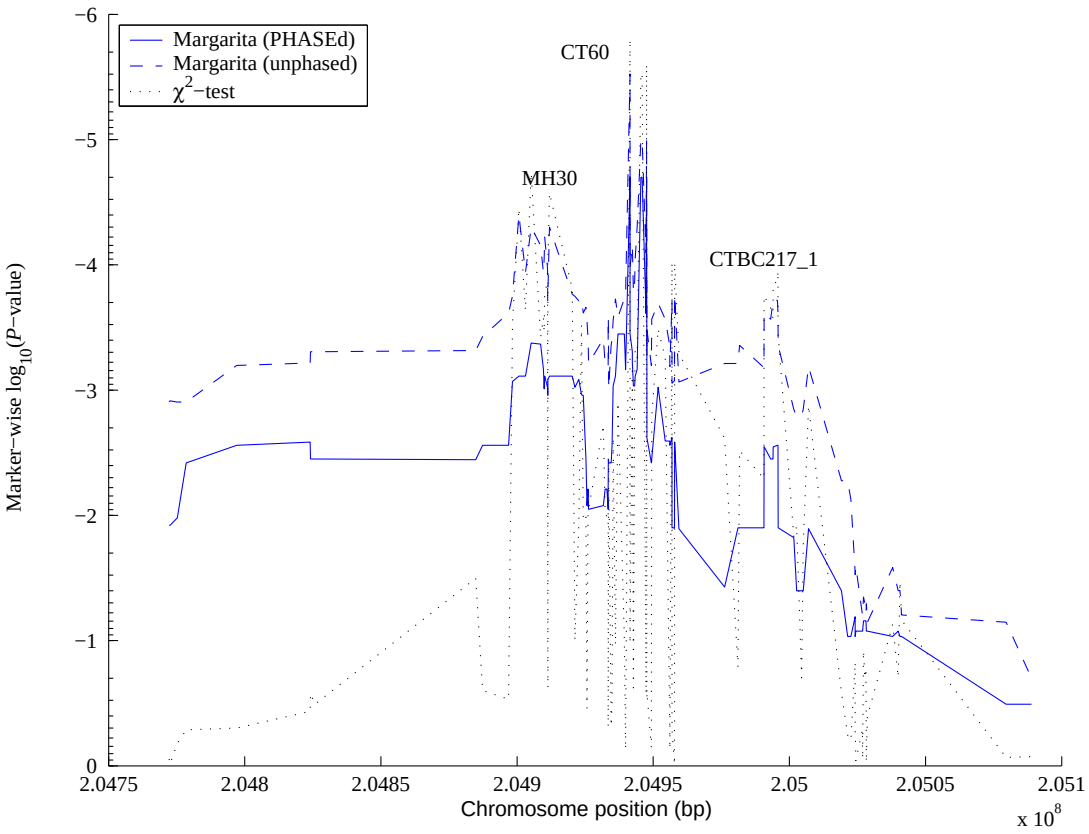

Figure 4.1: Analysis of the CTLA4 data. Association structure of the region.

2×10^{-6}) than does MARGARITA on the PHASEd sequences ($P \approx 2\times10^{-5}$). This result agrees with that of Morris et al. (2004), who similarly show that a two stage approach results in a loss of power compared to handling genotypes directly and marginalising over unknown phase. Both MARGARITA analyses have weaker significance than the chi-square test ($P \approx 1.6\times10^{-6}$) at CT60, which would be expected if CT60 is indeed the causative polymorphism, a hypothesis that can be explored by using the ARGs to further analyse the association signal.

Figure 4.2 gives the distribution of the estimated susceptibility allele frequency in the general population (calculated using the observation that Graves disease has population prevalence 0.5%). The mean estimate for the causative allele in the cases and controls is 65% and 54% respectively, corresponding to the G allele of CT60 (%63 and %52 in cases and controls respectively). This suggests that the bulk of the association signal at CT60 is due to susceptibility caused by CT60, or something extremely tightly linked to it.

However, in 43% of the inferred ARGs for the unphased data, MARGARITA is able to find

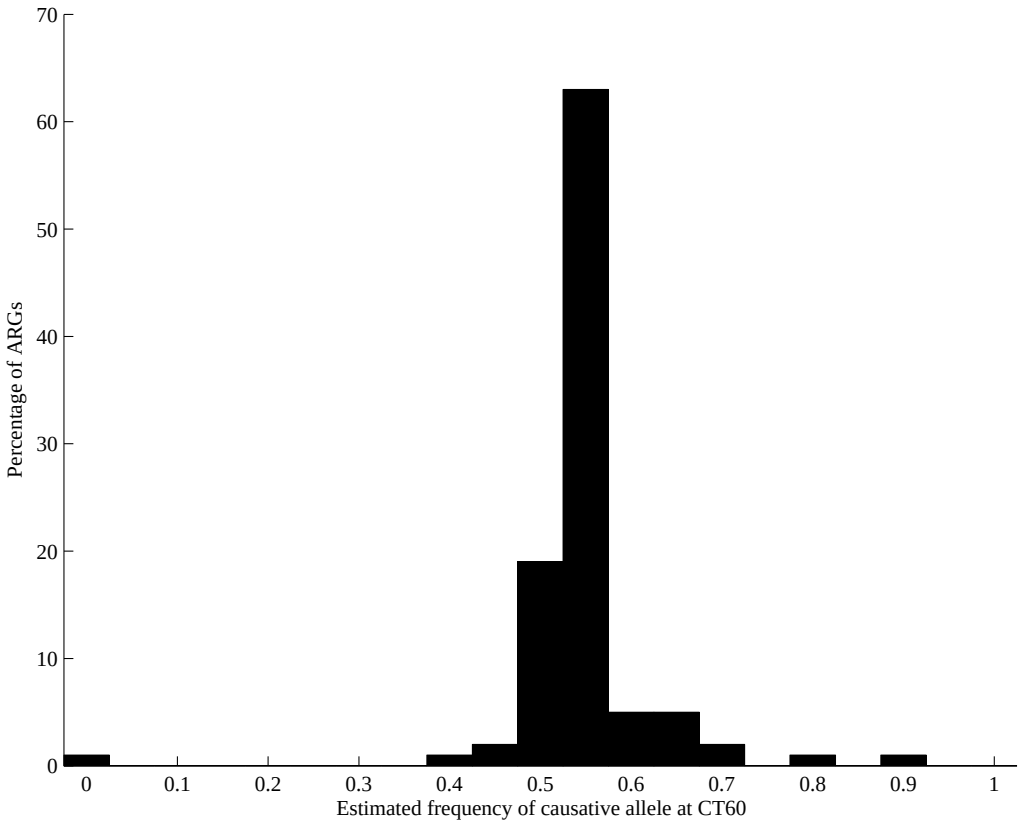

Figure 4.2: Distribution of estimated causative allele frequency in the general population, using marginal trees at CT60.

internal branches of the marginal tree at CT60 that segregate the cases and controls with chi-square test P-values of the order of 10^{-7} or less, with the strongest being of the order of 10^{-9}; this may suggest a second causative polymorphism. I therefore used the inferred ARGs to test explicitly for allelic heterogeneity. I took the 100 marginal trees inferred for each marker and counted the number of times each chromosome appeared under the branch corresponding to the best partitioning of cases and controls—the "best cut" branch. When a chromosome is under the best cut branch it means that if there is a disease causing allele at that position, then it is likely that the chromosome possesses it. Figure 4.3 shows this analysis for an illustrative sample of 167 case chromosomes (with phase inferred on the ARGs). For each marker and chromosome, the intensity of the plot represents the proportion of trees for which the chromosome is under the best cut. Case chromosomes 131-167 show a different pattern to the others. They occur less frequently under the best cut at CT60, and more frequently under

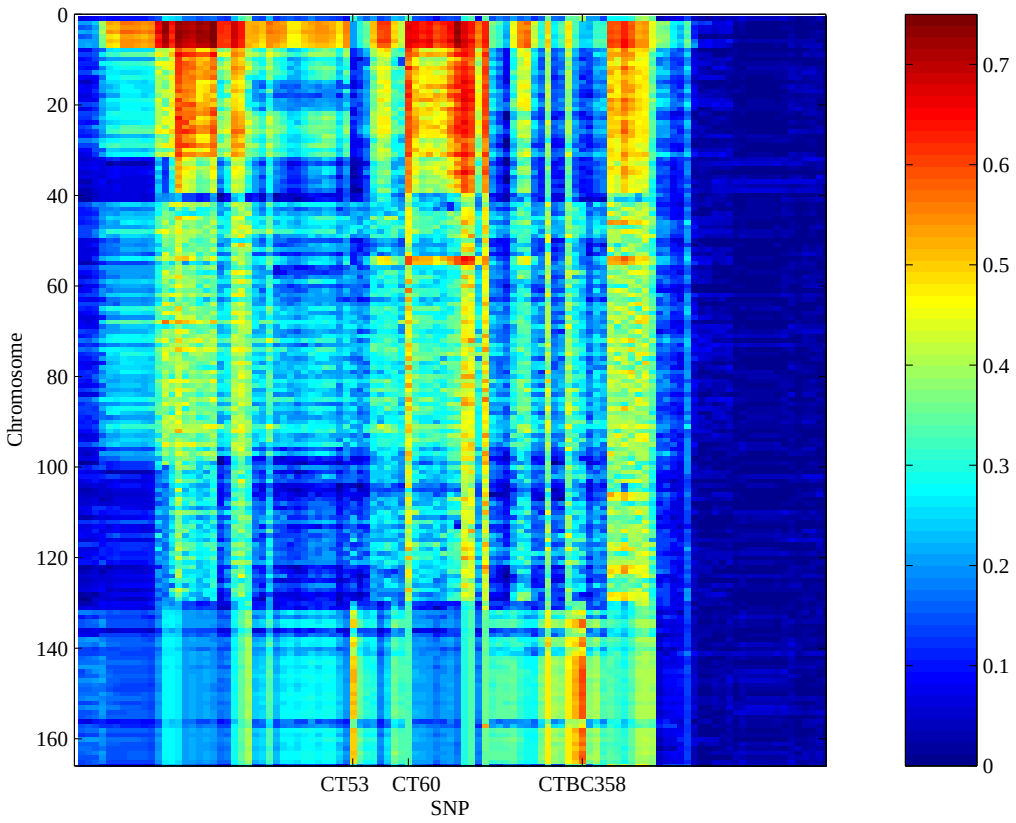

Figure 4.3: Test for allelic heterogeneity. The proportion of inferred marginal trees at each position for which a chromosome appears under the branch that best segregates the cases and controls.

the best cut at CT53 and CTBC358, whereas case chromosomes 1-130 appear frequently under the best cut at CT60, but infrequently under the best cut at CT53 and CTBC358. Although not shown in the figure, there are other case chromosomes not associated with any of these loci.

To test whether CT53 or CTBC358 are also susceptibility loci (or linked to susceptibility loci), I stratified the case-control population in three ways:

Only those chromosomes with the protective allele at CT60.

I took the PHASEd chromosomes and removed all those with the CT60 susceptibility allele, running the analysis on the remaining 282 case chromosomes and 620 controls with the protective allele (Figure 4.4). When the population is stratified in this way, the association signals at MH30 and CTBC217_1 collapse into the background, suggesting that the association signals at those locations are due to LD with CT60. Furthermore, there is an association

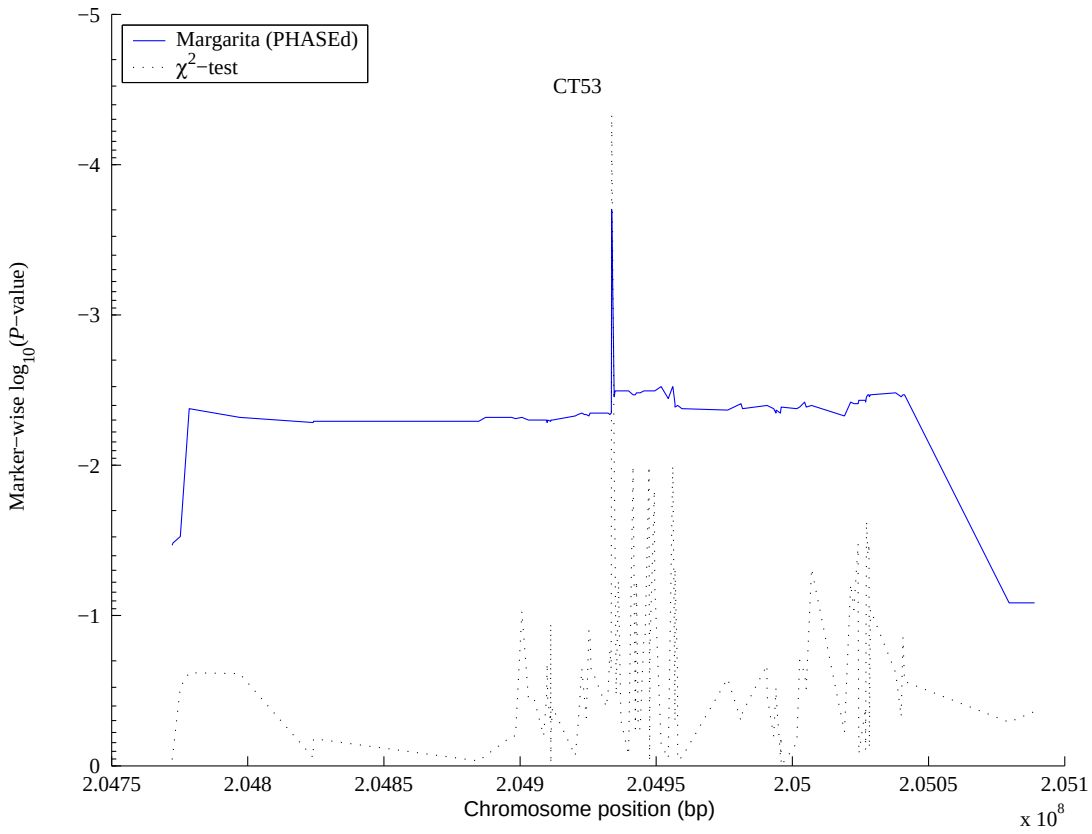

Figure 4.4: Association structure for a subset of the CTLA4 data—only those chromosomes with the protective CT60 allele.

peak at CT53 (marker-wise chi-square test $P \approx 5 \times 10^{-5}$; MARGARITA $P \approx 2 \times 10^{-4}$). Using MARGARITA, the estimated frequency of the causative allele (92% in the cases and 82% in the controls) matches that of the A allele at CT53 in this subpopulation (93% in the cases and 83% in the controls), suggesting that the A allele confers susceptibility on this CT60 background.

Only those chromosomes with the susceptibility allele at CT60.

After conditioning on the CT60 susceptibility allele there are 486 case chromosomes and 684 control chromosomes. In this subpopulation, CT53 has a weak signal of association with the disease (marker-wise chi-square test $P \approx 0.023$; MARGARITA $P \approx 0.016$). In contrast to the previous stratification, the A allele at CT53 is less frequent in the cases (2%) than in the controls (5%), suggesting that A may be protective on this haplotypic background. This reversal of the effect of CT53 dependent on CT60 status may explain why CT53 is not

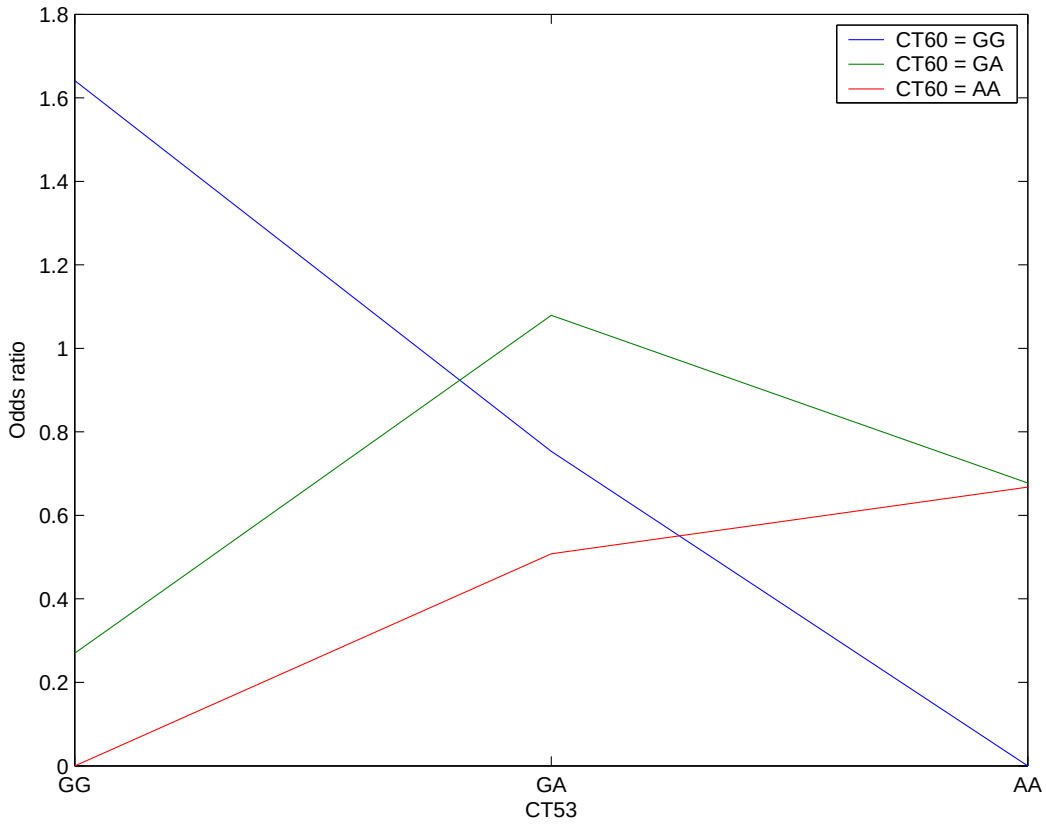

Figure 4.5: Epistasis between CT60 and CT53. The chart gives the odds ratios for genotypes at CT60-CT53.

detected in analyses using the full data.

Only those individuals that are homozygous for the CT60 protective allele.

To check that the CT53 association is not due to some spurious signal resulting from selecting chromosomes on the basis of their inferred haplotype phase, I took the genotype sequences homozygous for the CT60 protective allele and ran MARGARITA on these unphased sequences. There are 102 case chromosomes and 300 controls, giving a weaker but still significant signal of association (marker-wise chi-square test $P \approx 0.012$; MARGARITA $P \approx 0.013$). As expected, on this background the A allele of CT53 is the susceptibility allele.

These results suggest epistasis between CT60 and CT53, with the A allele at CT53 conferring susceptibility on a CT60 protective background, but being protective on a CT60 susceptibility background. Figure 4.5 shows how the disease effect of one locus is modified by the genotype at the other.

To test explicitly for epistasis between CT60 and CT53 I performed a logistic regression test for interaction (Cordell, 2002; Macgregor and Khan, 2006) and obtained $P \approx 0.004$ for interaction effects over and above single marker effects.

The reversal of the effect of alleles at CT53, conditional on the allele at CT60, would tend to reduce any significance in a logistic regression test for additional effects, as used in Ueda et al. (2003).

4.3 Replication of Result

Given the small samples sizes of the data after stratification into CT60 allele subpopulations, further genotyping in more samples is required in order to determine whether the observed signal at CT53 is a true positive or an artefact of the data.

In response to this ARG analysis, further genotyping was undertaken by my collaborators in the Diabetes and Inflammation Laboratory at the Cambridge Institute for Medical Research. They typed an additional 1,593 Graves cases and 4,055 controls at CT53 and CT60. In this larger sample, the previously reported effect of CT53 did not achieve significance when performing the same subgroup analysis.

However, an independently developed method (Dawy et al., 2006) applied to the original data (Ueda et al., 2003) did show the same signal as the ARG analysis. In Dawy et al. (2006) an information theoretic method was developed for disease mapping. Their method has the feature that it is entirely general regarding epistatic and risk model (for example, it does not assume a multiplicative model of genotype risk). It is designed to identify disease associated markers and to cluster them according to their pattern of variability, with the motivation being that markers with similar variability (i.e. in strong LD) are likely to have the same genealogical history and should be interpreted together. With this method, the degree of association between a marker and disease is measured as the quantity of information contained in the marker about the disease. They then search groups of jointly associated markers by using a "relevance chain" technique, where the reduction in uncertainty on the disease state given a genotype observation is measured conditional on observing the genotypes at other

positions.

Their approach identified the same peak at CT60 as in Ueda et al. (2003) and in the analysis above, and they also found an additional experiment-wise significant signal at CT53, in agreement with my analysis. However, it should be noted that since this is on the same data set, it is not a proper replication of the association signal; it merely shows that two independently developed methods identify the same signal (albeit not previously identified) in the same data set.

Dawy et al. (2006) give the following reason for why the CT53 signal was not detected in the original Ueda et al. (2003) analysis:

> "In the original article, the effect of secondary loci in addition to the main associated loci was tested, assuming a multiplicative model for the allele effects. Such trend regression approaches, however, imply a continually increasing or decreasing causality scheme across genotypes, which is possibly not always an accurate assumption. The slight difference between the original and our results might be attributed to the fact that the use of mutual information does not assume any particular mode of allelic risk."

In conclusion, since the signal has not been replicated in an independent population, the results suggesting epistatic interaction should be treated with caution; nevertheless, the analyses in this chapter show how the ARG approach can be used to dissect disease association signals, arriving at potentially interesting additional conclusions.