Chapter 5

Analysis of Prostate Cancer Association Study Data

5.1 Risk factors for Prostate Cancer

Globally, about 9.7% of cancers in men are prostate cancers, and the risk of developing the disease has been correlated with age, family history and ethnicity. The highest rates are reported in the Caribbean and Scandinavia and the lowest rates in China and Japan (Crawford, 2003). Within North America, the incidence of prostate cancer is 1.6 times greater among African American men than European American men (Freedman et al., 2006), suggesting a genetic component to the disease. However, it has also been shown that Japanese men who migrate to North America have an increased incidence of prostate cancer, suggesting that enviromental factors also have a role (Crawford, 2003).

Five recent association studies have sought to investigate the genetic basis of the disease (Amundadottir et al., 2006; Freedman et al., 2006; Gudmundsson et al., 2007; Haiman et al., 2007; Yeager et al., 2007), and all of them show strong association between variants in the 8q24 region and disease risk.

The first study, Amundadottir et al. (2006), identified a SNP and a microsatellite in 8q24 associated with risk in an Icelandic population, a signal which they found to be replicated in European American and Swedish populations. In the second study, Freedman et al. (2006),

admixture mapping (Patterson et al., 2004; Smith et al., 2004) was applied to a data set of 1,597 African Americans, identifying a 3.8Mb region of 8q24. The two polymorphisms identified in Amundadottir et al. (2006) were found to explain very little of the admixture signal in the African American population, suggesting additional causative variants in the region. This suggestion partly motivated further studies (Gudmundsson et al., 2007; Haiman et al., 2007; Yeager et al., 2007), although two of these (Gudmundsson et al., 2007; Yeager et al., 2007) were whole genome scans, looking beyond 8q24. In these subsequent studies independent risk alleles were identified to those in Amundadottir et al. (2006), however, these are all in 8q24. Of particular interest to us is Yeager et al. (2007), in which MARGARITA was applied to dissect the 8q24 association signal, aiding the identification of two independent association peaks. I now discuss results from these five studies in more detail, including a description of the MARGARITA analysis.

5.2 First Identification of the 8q24 Association Signal

In the first study (Amundadottir et al., 2006), a genome-wide linkage scan was performed, using 1,068 microsatellite markers typed in 871 Icelandic men with prostate cancer from 323 extended families. A second stage case-control association study in 8q24 identified the strongest associations to be a microsatellite DG8S737 ($P = 2.3 \times 10^{-8}$) and a SNP rs1447295 ($P = 1.7 \times 10^{-9}$). They replicated these signals in two additional case-control studies: one Swedish (1,435 cases and 779 controls) and one European American (458 cases and 247 controls).

They then investigated the independence of the two association signals, because in the Icelandic sample, allele -8 of DG8S737 and the A allele of rs1447295 have LD of $r^2 \approx 0.5$, and could therefore correspond to the same causative variant. They found that individuals with both risk alleles have greater risk than those with only one; and those with one have greater risk than those with neither. This suggests that neither of the alleles by themselves explains the risk, so either there are multiple causative variants in the region, or the two risk alleles are in strong, but imperfect, LD with an unknown risk variant.

They then undertook a study in African Americans in order to map more finely the risk variants; the basis for this being that African populations tend to have greater genetic diversity and weaker LD. Specifically, consider the 92kb LD block around DG8S737. In the CEU HapMap sample, there are 19 SNPs, including rs1447295, that have $r^2 = 1$ with each other; whereas in the YRI HapMap sample, only 2 SNPs have $r^2 = 1$ with rs1447295. In the African American sample, DG8S737 showed association of $P = 0.0022$ and rs1447295 was nonsignificant.

In the second study (Freedman et al., 2006), a technique known as admixture mapping was applied genome-wide to a data set of 1,597 African Americans. Admixture mapping (Patterson et al., 2004; Smith et al., 2004) is based on the observations that:

1. Some disease causing variants have significantly different frequencies in different populations; and

2. There are some diseases where the incidence of disease is also significantly different between populations. For example, in Americans, autoimmune diseases are more common in those of European descent (Patterson et al., 2004); whereas prostate cancer is more common in those of African decent (Amundadottir et al., 2006).

The technique involves scanning case individuals from populations of mixed ancestry. When a chromosomal region contains causative variants, it may show an over-representation of ancestry from the population with more risk alleles at that locus.

The advantage of admixture mapping is that it requires around 1% of the markers required in a LD based scan (Patterson et al., 2004); for example, in African Americans, admixture has occurred within the past 15 generations (Smith et al., 2004), hence there has been little time for recombination to break up the tracts of ancestral material, which means that the regions of excess ancestry around causative variants are likely to extend for tens of Mb, requiring far fewer markers to tag. In Freedman et al. (2006) only 1,365 SNPs were used for a genome-wide scan.

From their study of 1,597 cases, they located the admixture peak to a 3.8Mb region of 8q24. By also genotyping 873 African-American controls (controls are not required for

admixture mapping, but are useful for subsequent analyses) they estimated that the fraction of all prostate cancer incidence for African Americans below 72 years of age that could be explained by ancestry at this locus is 49%. This suggests that if the region of 8q24 were replaced with that from European ancestors, the rate of prostate cancer in African Americans would decrease by approximately 49%. However, it should be noted that such population attributable risk estimates should be treated with caution; they often sum to greater than 100%.

Freedman et al. (2006) also compared their results to Amundadottir et al. (2006). Because of the systematic differences in ancestry between cases and controls across 8q24, Freedman et al. (2006) tested whether the association at the DG8S737 microsatellite, detected in Amundadottir et al. (2006), corresponds to a fine mapping signal or the admixture signal of the larger region. (Amundadottir et al. (2006), tested for mismatching of cases and controls in overall ancestry, but not for a local rise in African ancestry at 8q24 in the cases.) Freedman et al. (2006) corrected for this local effect in their African-American cases and controls by testing whether the differences in allele frequencies between cases and controls could be explained just from the enrichment of African ancestry in the cases. After correction, they found that the contribution of the microsatellite to disease risk was nonsignificant. However, when typing rs1447295 in 1,614 cases and 1,547 controls from four non-African populations (Japanese Americans, Native Hawaiians, Latino Americans and European Americans) they replicated a strong association signal ($P < 4.2 \times 10^{-9}$).

Together, these two studies support the hypothesis of rs1447295 being associated with prostate cancer risk in non-Africans, while also suggesting a higher proportion of as yet unidentified risk alleles at 8q24 in the African American population.

5.3 ARG Analysis of 8q24 data

In the study of Yeager et al. (2007), 550,000 SNPs were genotyped in 1,172 cases and 1,157 controls of European origin. The association signal at rs1447295 was replicated with $P = 9.75 \times 10^{-5}$ (using a four degree of freedom logistic regression test), with seven SNPs near to

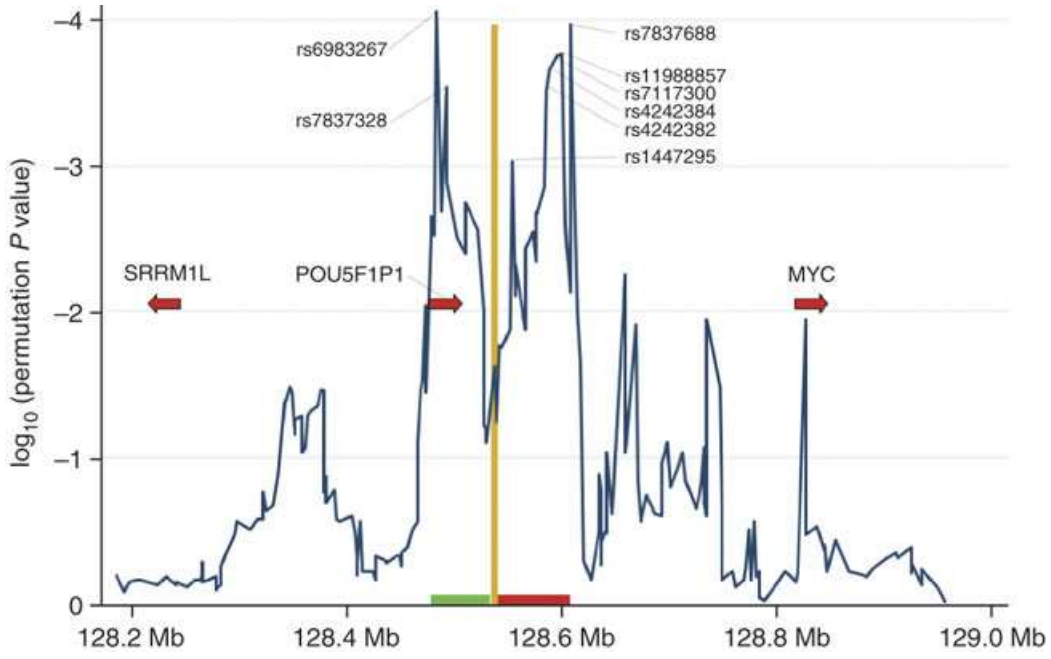

Figure 5.1: MARGARITA associations structure for the 8q24 region of Yeager et al. (2007).

rs1447295 showing still greater association.

In order to dissect the association signal, 197 SNPs in an 800kb region around rs1147295 were analysed using an adapted version of MARGARITA. I made modifications to MARGARITA so as to handle two departures from the standard model:

1. There are three phenotypes: case, non-aggressive prostate cancer and aggressive prostate cancer, with aggression defined by standard clinical phenotypes (Gleason index and disease stage).

2. Genotypes, rather than alleles were used in the chi-square test; which together with the three phenotypes gives a 3×3 contingency table with four degrees of freedom.

The results of the initial MARGARITA analysis are shown in Figure 5.1. MARGARITA gives two association signals: at a "centromeric" region around the association peak of rs6983267 and at a "telomeric" region around the peak of rs7837688.

When MARGARITA was first applied, the run time was significantly longer than expected for a similarly sized dataset with constant recombination rate; from experience this suggests a

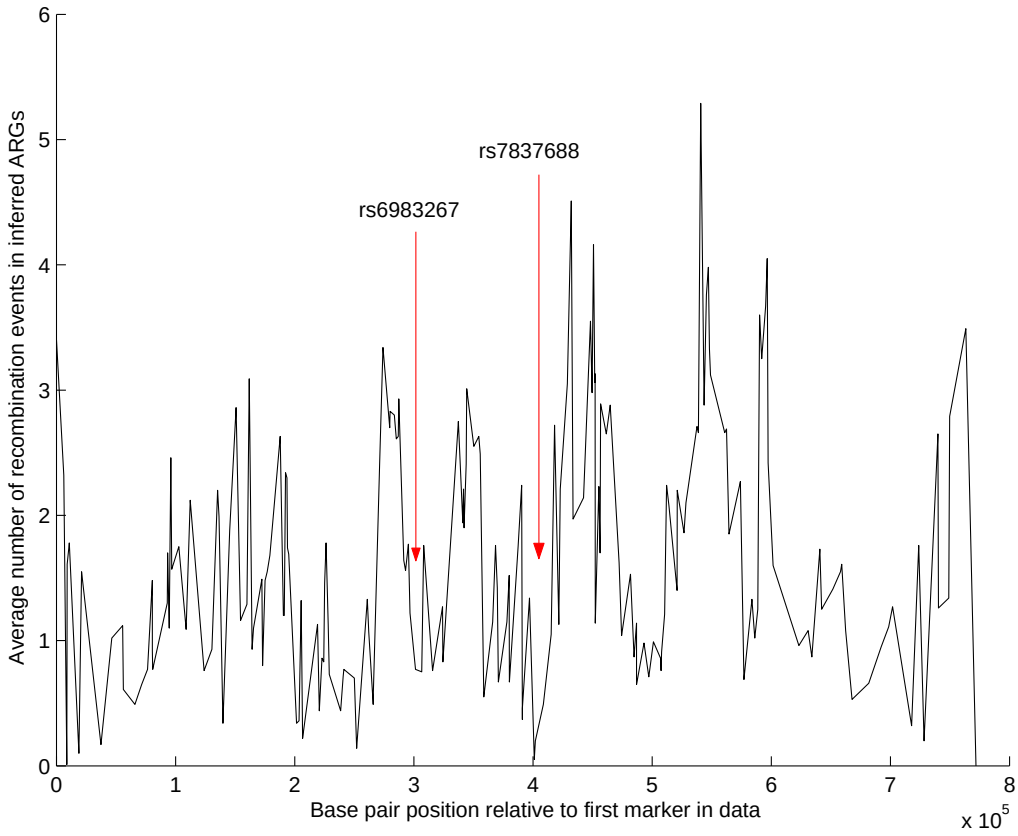

Figure 5.2: Number of inferred recombination events between markers in 8q24.

more complex recombination structure in the region, in particular one or more recombination hotspots.

The number of inferred recombination events between markers, averaged over 100 inferred ARGs, is shown in Figure 5.2. There is a pattern of peaks and valleys suggesting that there is local variation in recombination rate, including a peak between rs6983267 and rs7837688. However, it should be noted that MARGARITA is not a powerful method for estimating relative recombination rates because it only infers obligate recombination events, which are dependent on the frequencies of alleles.

The presence of a recombination hotspot separating those regions was confirmed by my collaborators at the NIH National Cancer Institute by applying the SequenceLDhot program of Fearnhead (2006). In the 130kb region covering the two peaks, SequenceLDhot identified a 5.5kb hotspot region which it estimated to contain 90% of the recombinations. This gives a population scaled recombination rate within the hotspot of 260, and 30 across the remainder

of the region. In Figure 5.1 the position of the recombination hotspot is shown with a yellow line.

Furthermore, the marginal trees from MARGARITA at the two peaks were found to be decorrelated, which, together with the evidence of a recombination hotspot, suggests that each region followed an independent history.

Since rs1447295 and rs7837688 both reside in the telomeric peak, and rs1447295 was identified as significant in previous studies, it was decided to focus on rs1447295 for the telomeric peak, rather than rs7837688, in the next analyses.

The ARGs were then used to estimate the frequencies of the causative alleles at rs1447295 and rs6983267. At rs6983267, the frequency of the inferred causative allele in the controls is 46% and at rs1447295 is 12%. These differences in inferred causative allele frequency further strengthen the hypothesis that the two signals are independent. These inferred causative allele frequencies also match the typed frequencies at rs6983267 and rs1447295, suggesting that the typed SNPs may be causative or essentially in complete association with the causative alleles. In the control populations, the predisposing allele of rs6983267 has frequency 48%, and at rs1447295 it has frequency 14%.

The ARGs were then used to guide a haplotype analysis, performed by my collaborators at the NIH National Cancer Institute. SNPs around each of rs6983267 and rs1447295 were phased using the program PHASE (Stephens and Donnelly, 2003): 20 SNPs around rs6983267, and 27 around rs1447295. The phase resolution with the greatest likelihood was taken, and 100 ARGs were constructed. The frequency with which each haplotype fell under the "best cut" for each region, that is, possessing an imputed causative polymorphism, was determined.

Figure 5.3 shows the results of this analysis. The haplotypes in green are those that tend to fall often under the best cut at the centromeric region (left) and the telomeric region (right), and those in red are those that tend to fall on the protective side of the best cut. The "Hap. freq" is the frequency of the haplotype in the data, and "Prediction" is the frequency with which the haplotype falls under the best cut.

For the centromeric region (around rs6983267), the protective haplotypes were found to be far less diverse than the susceptibility haplotypes, suggesting that the protective allele

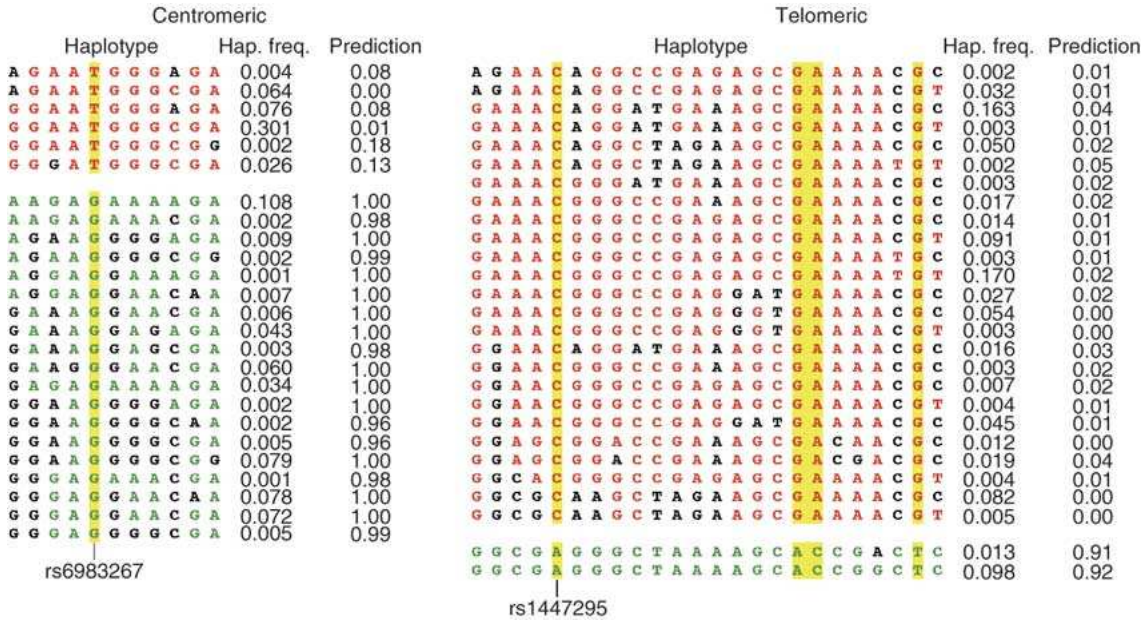

Figure 5.3: Haplotypes in the centromeric and telomeric peaks. Haplotypes coloured in red are those that tend to fall on the protective side of the marginal tree, and those coloured in green tend to fall under the imputed causative mutation. "Hap. freq." is the frequency of the haplotype in the data, and "Prediction" is the frequency with which the haplotype falls under the best cut.

is either more recent or positively selected. Further support for the hypothesis of positive selection comes from the observation that the protective allele at rs6983267 has frequency 52%, a high frequency given its relative haplotype diversity.

Conversely, the protective haplotypes in the telomeric region (around rs1447295) were found to be more diverse than the susceptibility haplotypes, suggesting that in this case the risk allele is either selected or more recent. Since the frequency of the risk allele of rs1447285 is 14%, there is better support here for the hypothesis that this risk allele is a recent mutation.

These conflicting genealogical accounts again support the hypothesis that the two markers correspond to distinct association signals, following independent histories. The deleterious mutation in the telomeric region is a more recent event than the protective mutation in the centromeric region.

5.4 Replication of Result

The signal at rs6983267 has been independently replicated in another case-control study (Haiman et al., 2007).

In Haiman et al. (2007) 2,973 SNPs were typed in up to 4,266 cases and 3,252 controls from five populations, across the 3.8Mb admixture peak of Freedman et al. (2006). They found three clusters of association, which they separated by comparing genetic and physical maps. Two of these clusters correspond to the "centromeric" and "telomeric" regions identified with the aid of MARGARITA in Yeager et al. (2007).

They then performed a stepwise logistic regression to determine the independence of the associated polymorphisms. This was done by incorporating each SNP into the model, in order of strength of association, and then repeating the analysis for the remaining SNPs, conditional on those already in the model. This analysis resulted in the identification of seven independent risk variants across 8q24, including rs6983267. However, the most strongly associated region of Haiman et al. (2007) was found sitting a few hundred kb centromeric from the signals found in the previously discussed case-control studies.

In another study, Gudmundsson et al. (2007) performed a genome-wide scan using 316,515 SNPs typed in 1,453 cases and 3,064 controls from Iceland. By testing each SNP separately, they replicated the previously identified signal at rs1447295, corresponding to the telomeric signal of Yeager et al. (2007). They then performed a haplotype block test, and identified another genome-wide significant signal in the same novel region as Haiman et al. (2007), and replicated these results in three other populations of European descent.

These studies together indicate that there are multiple independent risk alleles in 8q24, confirmed in multiple populations, but which do not lie in known genes. These variants could regulate nearby cancer causing genes, however, although there are genes in 8q24, such as the MYC oncogene, no differences in expression levels for genes in that region have been detected between carries and non-carriers of risk alleles (Gudmundsson et al., 2007). However, Haiman et al. (2007) note that 8q24 is the most frequently gained chromosomal region in prostate tumours, and speculate that the risk alleles make the region more prone to gain.