# Admixture Aberration Analysis: Application to Mapping in Admixed Population Using Pooled DNA

Sivan Bercovici and Dan Geiger

Computer Science Department
Technion, Haifa 32000, Israel
`{sberco,dang}@cs.technion.ac.il`

**Abstract.** Admixture mapping is a gene mapping approach used for the identification of genomic regions harboring disease susceptibility genes in the case of recently admixed populations such as African Americans. We present a novel method for admixture mapping, called admixture aberration analysis (AAA), that uses a DNA pool of affected admixed individuals. We demonstrate through simulations that AAA is a powerful and economical mapping method under a range of scenarios, capturing complex human diseases such as hypertension and end stage kidney disease. The method has a low false-positive rate and is robust to deviation from model assumptions. Finally, we apply AAA on 600 prostate cancer-affected African Americans, replicating a known risk locus. Simulation results indicate that the method can yield over 96% reduction in genotyping. Our method is implemented as a Java program called *AAAmap* and is freely available.

## 1 Introduction

Many complex disease studies are currently being conducted using population-based genetic association [1]. The premise of this method is that affected individuals carry a common variant of a disease susceptible gene which is in linkage disequilibrium with sampled markers. Hence, the susceptibility locus can be detected via the indirect association between the sampled markers and the disease status. In order to guarantee a sufficiently high power in association studies, thousands of cases and controls are sampled using dense marker panels.

Admixture mapping, also known as Mapping by Admixture Linkage Disequilibrium (MALD), offers a more economical alternative to association studies in certain circumstances without sacrificing the statistical power [2]. MALD is a gene mapping approach used for the identification of genomic regions harboring disease susceptibility genes in the case of recently admixed populations, i.e. populations that are an admixture of several ancestral populations. African Americans are an example of an admixed population, having both European and African ancestries. The method is applicable when the prevalence of a disease is significantly different between the ancestral populations from which the

admixed population was formed. When such a disease is studied, admixed individuals carrying the hereditary disease are expected to show an elevated genomic contribution from the ancestral population that has the higher prevalence of the disease around the disease gene loci. A MALD study is comprised of three main steps. First, a panel of ancestry informative markers (AIM) that differentiate well between ancestral populations is designed. Next, either cases or both cases and controls are individually genotyped using the AIM panel, and the mosaic of ancestries of each individual is inferred. Finally, the inferred ancestral profiles are scanned in search for an aberration towards the ancestral population with the higher risk, as expected to appear near the disease locus.

The MALD method successfully discovered multiple risk alleles for prostate cancer [3,4], a disease with a higher incident rate in Africans compared to Europeans, and a candidate locus for end-stage kidney disease in African Americans [5]. Diseases of similar characteristics include stroke, hypertension and multiple-sclerosis; a more comprehensive list of diseases suitable for admixture mapping appears in the method's review by Smith and O'Brien [2]. In all of these cases, the statistical efficiency of MALD stems from the fact that only a few thousands of ancestry informative markers are required in order to accurately infer the ancestry of the admixed individuals [6,7]. Moreover, only a few hundreds of cases are required for the identification of the ancestral aberration around the disease locus [8].

In this paper we present a novel approach for admixture mapping that considerably reduces the genotyping cost of disease studies by applying admixture aberration analysis (AAA) on pooled DNA of affected admixed individuals. Our analysis detects divergence of allele distribution in a pool of samples near a disease locus without the intermediate step of ancestry inference per individual. The inherent aberration in admixture around the disease locus shifts the sampled allele frequencies towards the distribution of the alleles in the ancestry with the higher risk. It is the examination of this shift, evaluated through the estimation of allele frequencies in the pooled sample, that provides the means for our pooled mapping method. Figure 1 illustrates this idea.

Current MALD studies mainly differ in the informative panel of choice and the method used for ancestry inference. Patterson *et al.* [9] presented a method that employs a hidden Markov model (HMM) for the estimation of ancestry along the genome. The HMM was integrated into a Markov chain Monte Carlo (MCMC) method to account for the uncertainties in model parameters. Tang *et al.* [10] extended previous methods by modeling linkage-disequilibrium in the ancestral populations using a Markov Hidden Markov model (MHMM), namely, dependency between adjacent markers evident in the ancestral populations was modeled. An inference framework developed in [11] enables the incorporation of more complex probability models that account for linkage disequilibrium in the ancestral populations. An earlier work by Chakraborty and Weiss [12] suggested mapping by directly assessing divergence from admixture linkage-disequilibrium, as expected near disease loci.
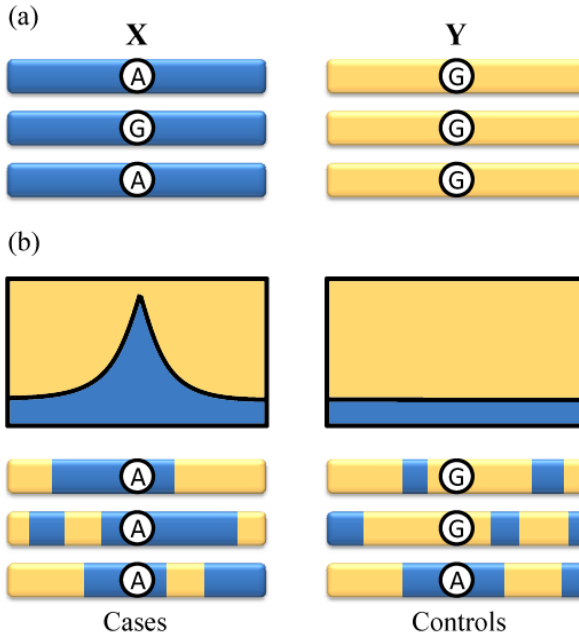
**Fig. 1.** An illustration of admixture aberration. (a) Two distinct ancestral populations, $X$ and $Y$, expressing a different distribution of alleles at a particular location. The greater the distance between allele distributions, the more informative the marker is regarding ancestry. (b) a sample of admixed individuals, descendants of the ancestral populations $X$ and $Y$. In case of a disease with higher prevalence in population $X$, the affected sample will exhibit a higher contribution from population $X$ near the disease locus, as indicated by the graph on the left. Hence, in the affected individuals, the distributions of alleles near the disease locus bears a higher resemblance to that of population $X$. The healthy admixed individuals show a contribution of populations $X$ and $Y$ that corresponds to the admixture process.

DNA Pooling has been suggested as a practical way to reduce the cost of large-scale association studies [13]. Rather than analyzing thousands of cases and controls that were sampled separately, association analysis was first applied on pooled cases and pooled controls in the work of Arnheim *et al* [14]. Steer *et al.* [15] have recently demonstrated the feasibility of pooled association studies using high resolution microarrays for rheumatoid arthritis. Zeng and Lin [16] examined the analysis of pooled DNA, extending the single-marker association methods to haplotype association using a likelihood-based approach. Kirov *et al.* [17] investigated the accuracy by which the allele frequency difference between pools can be estimated. This work was extended by Wilkening *et al.* [18] for higher resolution SNP microarrays of 250K. Pooling was also used in QTL studies. For example, Darvasi and Soller [19] presented a statistical test of marker-QTL linkage based on selective pools of individuals with extreme quantitative trait values.

The main contribution of this paper is the introduction of pooling to admixture mapping, and the demonstration of its power to the mapping of disease susceptibility loci. Pooling is a far more effective tool for admixture mapping in comparison to association studies. In the case of a recently formed admixed population, the linkage-disequilibrium patterns generated by the admixture process stretch over regions of several centimorgans, resulting in a wider effect which is easier to detect. In addition, using ancestry informative markers improves the ability to locate deviations of LD and marker distribution from those expected by the admixture process alone. The efficiency of our pooled AAA method has been established through simulation and via analysis of diseases that are currently being studied using the non pooled MALD approach. Specifically, we first develop the aberration analysis method based on a window of markers while accounting for linkage disequilibrium in the ancestral populations. We then determine the method's power through simulations. We show, for example, that a power of over 70% is achieved in a simulated study of an African American population carrying a disease with ethnicity relative risk of 1.3, comparable with end-stage kidney disease, using 7 pools of 200 individuals with 4 repetitions. The results in this case indicate a more than 25-fold decrease in genotyping versus a non-pooled MALD method. We also demonstrate the strength of our pooled method on a sample of African American cases of prostate cancer, replacing 600 independently measured individuals with a single simulated pool. The result demonstrate that a significant signal (LOD 7.2) is obtained near the risk locus found by Amundadottir *et al.* [20] and Freedman *et al.* [3]. Finally, we discuss the robustness of our method to measurement errors and to deviation from model assumptions.

## 2   Material and Methods

### 2.1   Definitions and Model Assumptions

The genome of a recently admixed individual is a mosaic of long, single ancestry, chromosomal segments. We use the following definitions to describe these segments in admixed individuals. An *admixed chromosome* is a chromosome that originated from more than one ancestral population. A *Post Admixture Recombination* point (abbreviated PAR) is a recombination point in which either two chromosomes from different populations crossed, or two chromosomes crossed where at least one of the chromosomes is an admixed chromosome. A *(PAR) block* is a chromosomal segment limited by two consecutive PAR points, or by a chromosome edge and its closest PAR point. An immediate implication of these definitions is that every PAR block originated from a single ancestral population, designated as the ancestry of the block, for otherwise the block would have been further divided. In our model, we assume that the ancestry of PAR blocks are mutually independent. We further assume that given the ancestry of a PAR block, the markers within that PAR block are independent of the markers outside the PAR block and are determined strictly according to the distribution that corresponds to the ancestry of that PAR block [7]. The markers within a

PAR block are assumed to be dependent, accounting for the background linkage-disequilibrium in the ancestral populations.

Consider an admixed population that originated from two ancestral populations $X$ and $Y$. Each ancestral population may have a different prevalence for a disease. A common way to characterize the disease risk attributed to the ancestral profile is by the ethnicity relative risk (ERR) which measures the increased risk due to an additional allele from population $Y$. Under a multiplicative disease model, ERR is defined as

$$r = \frac{\psi(XY)}{\psi(XX)} = \frac{\psi(YY)}{\psi(XY)} \tag{1}$$

where $\psi(\cdot)$ is the probability of the disease given that the ancestry pair at the disease susceptibility locus is either $XX$, $XY$, or $YY$.

When studying an admixed population with an hereditary disease characterized by an $ERR \neq 1$, the regions around the disease loci are expected to show an aberration towards the ancestry with the higher risk, shifting the distribution of nearby allele frequencies. Our method scans through the genome, computing for each examined location the ratio between the likelihood of the measured allele frequencies under the assumption of a close disease locus and the likelihood of the measured frequencies under the null assumption of no disease:

$$\Lambda_0 = \frac{P(S|\text{nearby disease locus})}{P(S|\text{no disease})} \tag{2}$$

where $S$ are the observed allele frequencies. Since the computation of this likelihood becomes intractable as the number of samples and markers grow, we approximate these probabilities via the multivariate central limit theorem over a window of markers. This approximate measure, denoted $\Lambda$, is used in the reported results. In the remaining method section, we derive the distribution of alleles under the two hypothesis, and the $\Lambda$ score. We first assume a window with a single marker and then extend the results to multi-marker windows.

## 2.2   Single Marker Analysis

We first compute the probability $P(J|d)$ of a bi-allelic marker $J \in \{0, 1\}$ of an individual, given the individual is affected (denoted by $d$). This probability is given by

$$P(J|d) = P(J|\bar{r}, d) \cdot P(\bar{r}|d) + P(J|r, d) \cdot P(r|d) \tag{3}$$

where $r$ indicates that at least one recombination has occurred between the disease locus and the location of allele $J$ since the first admixture event, and $\bar{r}$ is the complementary event.

The occurrence of post-admixture recombination points (PAR) can be modeled as a Poisson process with rate $\lambda$ which is derived from the admixture dynamics. In the case of a hybrid-isolated admixture model [21], $\lambda$ roughly corresponds

to the number of generations since the admixture began. Hence, under the assumption that the event of a recombination is independent of the disease status, the probability of at least one PAR point between location $l_1$ and $l_2$ is

$$P(r|d) = P(r) = 1 - e^{-\lambda \cdot |l_1 - l_2|} \qquad (4)$$

To compute $P(J|r, d)$ in Equation 3, we note that given $r$, namely that at least one PAR point occurred between sampled allele $J$ and the disease locus, the distribution of the allele is determined solely by the ancestry at the location and the admixture coefficient $P(Q)$:

$$P(J|r, d) = \sum_Q P(J|Q, r, d) \cdot P(Q|r, d) \qquad (5)$$

$$= \sum_Q P(J|Q) \cdot P(Q)$$

where $Q$ is the ancestry at the marker location.

To compute $P(J|\bar{r}, d)$ in Equation 3, namely when assuming no PAR point exist between the disease locus and the sampled allele, the distribution of the allele is given by

$$P(J|\bar{r}, d) = \sum_{Q'} P(J|Q') \cdot P(Q'|d) \qquad (6)$$

where $Q'$ is the ancestry at the disease locus. The above equality relies on the assumption that given the ancestry of the chromosomal segment containing marker $J$, the affection status and the allele are independent, an assumption that is common in admixture mapping models [9]. The probability $P(Q'|d)$ of the ancestry of an affect individual at disease locus $Q'$ is formalized in terms of the multiplicative disease model. Let $Z' \in \{XX, XY, YY\}$ denote the ancestry pair at the disease locus. The probability of ancestry $Q'$ given the disease can be written as

$$P(Q' = X|d) = \sum_{Z'} P(Q' = X|Z', d) \cdot P(Z'|d) \qquad (7)$$

$$= P(Z' = XX|d) + \frac{P(Z' = XY|d)}{2}$$

The probability $P(Z' = XX|d)$ is computed from $\psi(\cdot)$ as follows:

$$P(Z' = XX|d) = \frac{P(D|Z' = XX) \cdot P(Z' = XX)}{\sum_{Z'} P(d|Z') \cdot P(Z')}$$

$$= \frac{\psi(XX) \cdot p_X^2}{\psi(XX) \cdot p_X^2 + 2\psi(XY)p_X(1 - p_X) + \psi(YY)(1 - p_X)^2}$$

where $p_X$ is the a priori probability of ancestry $X$ in an admixed individual. The probabilities $P(Z' = XY|d)$ and $P(Z' = YY|d)$ are derived in a similar fashion. This completes the derivation of all terms of Equation 3.

We continue by considering a set of independent marker observations $J_1, J_2, ...$ $J_n$ sampled from $n$ affected admixed individuals. We need to compute the likelihood ratio $\mathcal{L}$ of these observations, namely the probability of the observations under the hypothesis of a nearby disease susceptibility locus divided by the probability under the null hypothesis of no disease

$$\mathcal{L} = \frac{P(J_1, .., J_n | H_1)}{P(J_1, .., J_n | H_0)}$$

As we assume independent and identically distributed $J_i$, we conclude that

$$\binom{n}{|\{J_i | J_i = 1\}|} \cdot P(J_1, .., J_n) = P(S_n)$$

where $S_n = \sum_i J_i$. Hence, the likelihood ratio can be rewritten as follows

$$\mathcal{L} = \frac{P(J_1, .., J_n | H_1)}{P(J_1, .., J_n | H_0)} = \frac{P(S_n | H_1)}{P(S_n | H_0)}$$

We now explicate how to approximate the probabilities $P(S_n | H_0)$ and $P(S_n | H_1)$.

According to the central limit theorem, the standardized sum of $n$ observations converges to the standard normal distribution $N(0, 1)$ as $n$ grows

$$S_n^* = \frac{\sum J_i - n \cdot \mu}{\sigma \sqrt{n}} \to N(0, 1)$$

where $\mu$ and $\sigma$ are determined by the distribution of $J$. For the two hypotheses, we use the following means and variances:

$$\mu_0 = P(J | r, d), \ \sigma_0 = \sqrt{P(J | r, d) \cdot (1 - P(J | r, d))}$$
$$\mu_1 = P(J | d), \ \ \sigma_1 = \sqrt{P(J | d) \cdot (1 - P(J | d))}$$

Note that $P(J | d)$ is given by Equation 3, and that $P(J | r, d)$ is given by Equation 5. The use of $P(J | r, d)$ for the null hypothesis is justified because this case is equivalent to an infinitely distant disease locus. Each hypothesis yields a different distribution of the markers hence a different standardization, and in turn, a corresponding probability for the sum of observations. We denote the standardized sums of $S_n$ according to hypotheses $H_0$ and $H_1$ by $S_n^{H_0}$ and $S_n^{H_1}$, respectively. The likelihood ratio of the observations under the two hypothesis can now be approximated as follows

$$\mathcal{L} = \frac{P(J_1, .., J_n | H_1)}{P(J_1, .., J_n | H_0)} \to \frac{P(S_n^{H_1})}{P(S_n^{H_0})} = \Lambda \tag{8}$$

The $\log_{10}$ of $\Lambda$ is called the LOD score; high LOD scores are indicative of a nearby disease locus.

In the above derivation, we assumed that the $n$ marker observations are independent even though each affected individual contributes two observations to the sample. The effect of this discrepancy weakens as the sample size increases.

As a final note consider the case of fully informative markers. Such markers have one allele with probability 1 in the first ancestral population and the other allele with probability 1 in the second ancestral population. When using fully informative markers and assuming no errors reading them, the ancestry at each marker location is known with certainty using a single marker readings. In this case, a non-pooled MALD locus statistic such as the one described in [9], reduces to the ratio between the probability of ancestry given a nearby disease locus and the a priori probability of ancestry (rather than a ratio between probabilities of marker data). This ratio exactly equals, in the limit of sufficiently large samples, to our $\Lambda$ statistic under fully informative markers. Consequently, for sample sizes that one normally deals with in MALD studies ($> 500$ samples), our AAA method retains the same statistical power as non-pooled MALD but at orders of magnitude less genotyping under this scenario. A comparison of the power of the two methods is further studied in Section 3 without assuming fully informative markers.

## 2.3   Multi-marker Analysis

We now extend our analysis from a single marker to the case of haplotypes where $m$ bi-allelic markers are sampled. First, we derive the probability $P(J|d)$ of an individual to carry haplotype $J \in \{0, .., 2^m - 1\}$ given that the individual is affected (denoted by $d$). This probability can be written via

$$P(J|d) = \sum_{\pi} P(J|\pi, d) \cdot P(\pi) \tag{9}$$

where $\pi$ is a partition of the haplotype into PAR blocks. The probability of a partition $p(\pi)$ is determined by the independent PAR points that either occurred or did not occur between sampled markers $\prod_{i=1}^{m-1} P(R_i)$, where the variable $R_i \in \{0, 1\}$ denotes whether a PAR point occurred between markers $i$ and $i+1$, and the probability $P(R_i = 1)$ is given by Equation 4.

To compute the remaining term $p(J|\pi, d)$ in Equation 9, recall that our admixture model assumes that markers within a PAR block are independent of markers outside the PAR block given the ancestry of the block. Hence, given partition $\pi$, the probability of haplotype $J$ is given by

$$P(J|\pi, d) = \prod_{b} P(J_b|d) \tag{10}$$

$$= \prod_{b} \sum_{Q_b} P(J_b|Q_b, d) \cdot P(Q_b|d)$$

where $b$ is a block in partition $\pi$, $J_b$ are the markers within block $b$, and $Q_b$ is the ancestry of that block. The probability of a block's ancestry given an affected individual is determined by whether or not the disease locus is within the PAR block in question, hence is

$$P(Q_b|d) = \begin{cases} P(Q'|d) & l_d \in b \\ \pi_Q & otherwise \end{cases}$$

where $l_d$ is the tentative disease locus, $\pi_Q$ is the a prior probability of ancestry $Q$, and $P(Q'|d)$ is given in Equation 7.

Our model assumes that given the ancestry of a block, the haplotype distribution is independent of the disease status. Hence, the term $P(J_b|Q_b, d)$ in Equation 10 is equals the probability $P(J_b|Q_b)$ which can be computed via samples taken from the ancestral populations. For example, European and West African individuals phased in the HapMap project [22] were used in Section 3 to construct the ancestral haplotype distribution $P(J|Q)$ for the analysis of African American. This concludes the derivation of all the terms used in the computation of Equation 9.

Finally, we consider a set of independent haplotype observations $J_1, J_2, ...J_n$ sampled from $n$ affected admixed individuals. We compute the likelihood ratio of the pooled observations, dividing the probability under the hypothesis of a nearby disease susceptibility locus by the probability under the null hypothesis of no disease:

$$\mathcal{L} = \frac{P(S_n|H_1)}{P(S_n|H_0)}$$

where $S_n$ is the sum of observations $J_i$.

We continue by explicating the computation of the probabilities $P(S_n|H_0)$ and $P(S_n|H_1)$. According to the multivariate central limit theorem, under the assumption that the covariance matrix of $J$ is positive-definite, the standardized sum of $n$ observations converges towards the standard normal distribution $N(0, \Sigma)$ as $n$ grows

$$S_n^* = \frac{\sum J_i - n \cdot \mu}{\sqrt{n}} \rightarrow N(0, \Sigma)$$

where $\mu$ and $\Sigma$ are determined by the distribution of $J$ assuming an affected admixed individual. For the two hypotheses, we use the following means and covariance matrices:

$$\mu_0 = \sum_J J \cdot P(J|d, l_d = \infty)$$

$$\mu_1 = \sum_J J \cdot P(J|d, l_d = l)$$

$$\Sigma_{i,j}^0 = E\left((J^i - \bar{J}^i)(J^j - \bar{J}^j)\,\Big|\, l_d = \infty\right)$$

$$\Sigma_{i,j}^1 = E\left((J^i - \bar{J}^i)(J^j - \bar{J}^j)\,\Big|\, l_d\right)$$

where $J^i$ indicates the $i^{th}$ component of haplotype $J$. Under the alternative hypothesis, the distribution $P(J|d, l_d = l)$ equals $P(J|d)$ given by Equation 9, setting $l_d$ to equal the suspected locus $l$. When assuming no disease locus, the distribution $P(J|d, l_d = \infty)$ equals $P(J|d)$ from Equation 9 under the assumption $l_d = \infty$.

We denote the standardized sums of $S_n$ according to hypotheses $H_0$ and $H_1$ by $S_n^{H_0}$ and $S_n^{H_1}$, respectively. The likelihood ratio under the two hypothesis can now be approximated as follows

$$\mathcal{L} = \frac{P(S_n|H_1)}{P(S_n|H_0)} \rightarrow \frac{P(S_n^{H_1})}{P(S_n^{H_0})} = \Lambda \tag{11}$$

The AAA method is defined to be the process of computing the LOD score $\log_{10} \Lambda$ via Equation 11 at examined locations along the genome, declaring a region that shows a LOD above 3.3 as a suspect area that may contain a disease locus. Subsequently, significant peaks serve as candidates for fine-mapping. Section 3 details the process of selecting the LOD threshold.

### 2.4  Pooling Strategies

In the case of DNA pooling, two parameters affect the number of panels used, namely the pool size $k$ and the number of pool repetitions $l$. It was shown that these two parameters can increase the accuracy of allele frequency estimation in the pooled sample which affects the method's statistical power [13]. Based on previous studies, when using a high-throughput platform for genotyping, pooling is recommended to be applied in quadruplets ($l = 4$). An empirical study of pooling examined the efficiency of this approach in association studies, using pools of $k = 250$ individuals [15]. We report our results with $l = 4$ and $k = 200$.

### 2.5  Leave-One-Out Filter

The leave-one-out (LOO) approach is a common filtering method that can be used in this context to discard false-positive signals originating from markers with erroneous frequencies. One potential source for bias is the inaccurate estimation of the allele frequencies in the ancestral populations. Biased genotyping errors can also result in false signals. Both error sources are assumed to occur independently between the markers and with low probability. When applying the AAA method, the robustness of a high LOD signal is examined via LOO by repeatedly removing markers and evaluating the effect on the LOD; the minimal LOD is reported, conferring with a conservative approach. A significant signal that persists after the removal of the marker with the highest contribution to the LOD is less likely to be false. LOO is especially effective in admixture mapping because suspected regions are usually supported by multiple SNP markers, retaining the method's power throughout the filterring phase as opposed to association studies, which often pinpoints a small suspected region with a single SNP marker.

## 3   Results

In this section we evaluate the performance of AAA through simulations, showing that the method has high statistical power and can detect loci of disease genes with even modest ethnicity relative risk. We investigate our statistics in the absence of a disease, bounding the false-positive rate to 5% genome-wide. We examine the effect of deviation from model assumptions, showing that for many

realistic disease models (ERR $> 1.4$) the method is robust to the inaccuracies expected in real data, and for milder ERR, the power can be retained through additional samples. We compare our AAA method to non-pooled MALD, demonstrating significant reduction in panel assays due to pooling at the cost of an increase in sample size. Finally, we validate our method by replicating the result of a prostate cancer risk locus using real data.

To evaluate the performance of our proposed method, we simulated data following the characteristics of recent MALD studies. We examined a range of disease models, including a mild value of ERR $= 1.3$ (corresponding to end-stage kidney diseases) which produces signals that are harder to detect in comparison to diseases with higher ERR values such as hypertension (ERR 1.6) [2]. The population of African Americans was simulated using the haplotypes of 60 unrelated European and 60 unrelated West African individuals phased in the HapMap project [22].

The simulation assumed a Hybrid-Isolated admixture model with 0.2 European contribution, 0.8 African contribution, and 8 generations of admixture. The simulated individuals were sampled according to a published panel of 1955 ancestry informative SNP markers [7], of which approximately 150 SNPs are on chromosome 1.

Figure 2 illustrates the output of the AAA method, using pools of 500,1000, 1500 and 2000 affected individuals. The disease susceptibility locus was set to $50cM$ and the simulated disease ERR was 1.3. A 3-marker sliding window was used to examine chromosome 1. One can clearly note that the evident peak, co-located with the disease locus, becomes significantly differentiated from distant locations with every increase in sample size.

We evaluated the distribution of our LOD statistic in the absence of a disease by performing simulations of pools of 500, 1000 and 2000 admixed controls, analyzing the sample using a window of 2, 3 and 4 markers at $1cM$ steps. We assume an ERR between 1.3 and 1.8 using a multiplicative increase risk model (Equation 1) with a higher prevalence in Africans. Each configuration was repeated 2500 times. The results illustrate that the gap between random and significant signals increases markedly with both the sample size and the window size (see Table 2 in the appendix for more details). The 95th percentile was approximately LOD $= 3.3$ when a pool of 1000 individuals was analyzed using a window of 2-4 markers over the entire genome, assuming an ERR of 1.3. This means that by defining the significance threshold to be a LOD $> 3.3$, we consequently confer a less than 5% type I error under the unfavorable condition of a hard to detect disease. Our recommended threshold of LOD $> 3.3$ is applicable for a wide range of parameters, as seen in Table 2, but can be relaxed depending on the admixture model and sample size, as can be determined through appropriate simulations.

To establish the statistical power of AAA we simulated a range of models with ERR values ranging from 1.2 to 1.8. For each disease model, we evaluated the performance for a single pool of 500, 750, 1000, 1500 and 2000 cases. In each simulation, a uniformly random locus along chromosome 1 was chosen as the disease locus. Each configuration, consisting of a specific sample size and
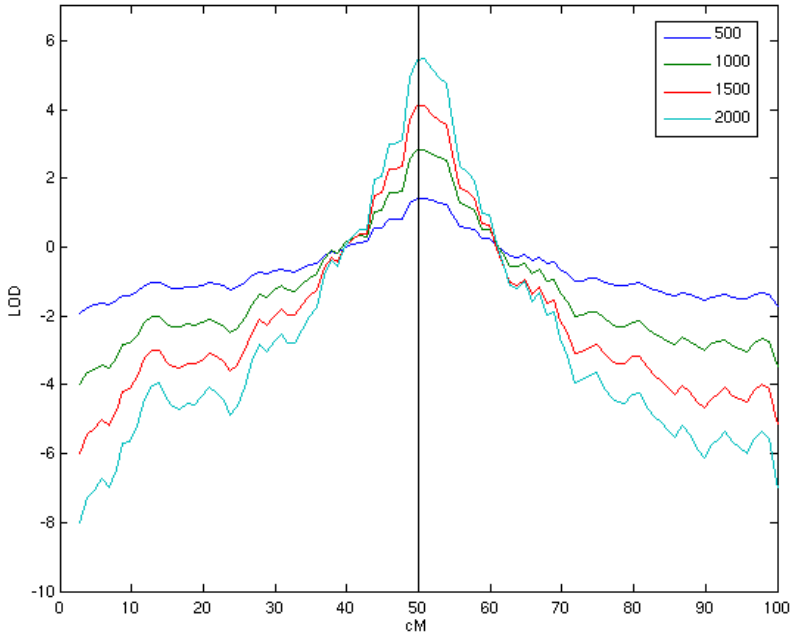
**Fig. 2.** LOD score along chromosome 1 showing a peak co-located with disease locus at 50 cM. The significant signal is enhanced with the increase of sample size while nearby LOD scores drop. The simulated disease ERR (1.3) is comparable to end-stage kidney disease. Chromosome 1 was sampled using 147 ancestry informative markers.

an ethnicity relative risk, was repeated 2500 times. Figure 3 summarizes the results of applying AAA using a window of 4 markers. A successful detection was defined as a peak with LOD $> 3.3$ within $5cM$ of the actual disease locus. The results indicate high statistical power (over 80%) under disease models that are considered difficult to detect (e.g., ERR of 1.3) when a pool of 1500 affected individual is used. We further found that 500 cases suffice to detect a disease of ERR $\geq 1.6$ with a power of approximately 80%, and 1000 cases yield a power of over 83% in the analysis of a disease with ERR $\geq 1.4$.

To evaluate the robustness of AAA to deviation from model assumptions, we examined the performance under inaccuracies in the admixture parameters. Namely, the inaccurate estimate $\lambda$ of the number of generations since first admixture, and the inaccurate estimate of the ancestral distribution $P(Q)$. Using a simulated population with African American admixture characteristics we conclude that the statistical power is insensitive (less than 1% decrease in power) to an inaccuracy of up to 5% in $\lambda$. Error in the estimate of $P(Q)$ has a greater effect on power. In particular, a 5% overestimation of the contribution of the ancestry with the higher risk yields a 4.8% drop in power for a study with 2000 cases and ERR $\geq 1.5$, and a 1.8% drop for a study with 1000 cases and ERR 1.8. When only 1000 cases are used to study a disease with a milder ERR of 1.5, the
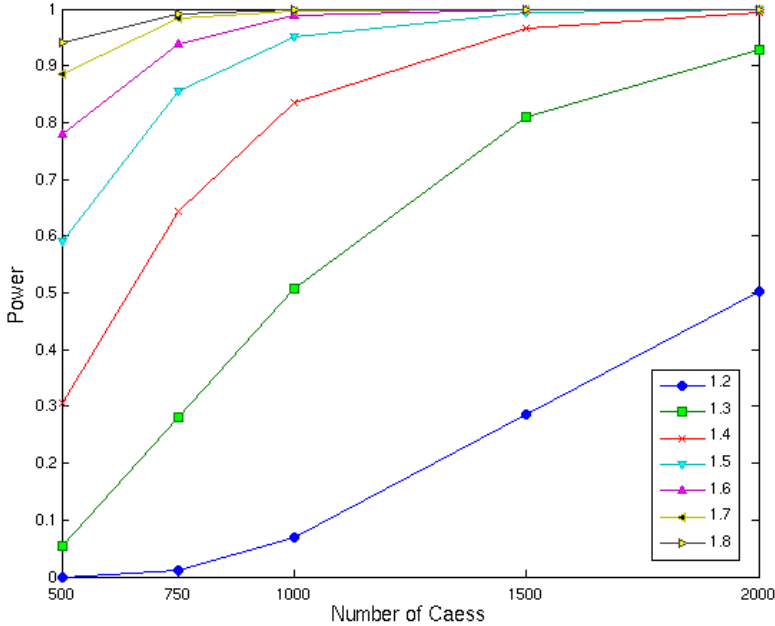
**Fig. 3.** Statistical power of 4-marker window analysis under different disease models and sample sizes

power drops significantly from 95% to 72%. The inaccuracies in the estimation of these admixture parameters are expected to be lower than 5% in the case of African Americans [9].

To investigate the extent of genotyping reduction due to pooling we examined the number of SNP assays needed in order to achieve 70% power using our AAA method versus MALD. The MALD method performance was evaluated under the optimal condition where the ancestries are perfectly inferred by a fully informative single marker (as described in Section 2). The performance of AAA was examined over 2500 uniformly chosen locations along chromosome 1, using a window of 4 markers. In the case of AAA, we report the results under the configuration of $k = 200$ and $l = 4$ which resembles the choice of [13] and [15]. The results are shown in Table 1. For ERR $= 1.3$, MALD requires a sample of 700 affected individuals, with one assay per individual. For the same disease model, AAA uses 28 assays, which suggests a 96% reduction in genotyping. The disadvantage of AAA is the need to collect additional affected individuals. However, for less than doubling the number of individuals, a 25-fold reduction in the number of assays is achieved. The performance of AAA was evaluated using a real panel for admixture mapping. When considering only perfect markers, AAA performance improves even when a single marker window analysis is applied, reducing the number of cases from 1300 to 1200, and the number of assays from 28 to 24. Similar results are obtained for ERR $= 1.4$.

**Table 1.** The number of SNP assays needed to achieve a power of 70% using MALD and AAA. The AAA method yields over 25-fold decrease in the number of SNP assays when using pool size $k = 200$ and number of replicates per pool $l = 4$.

| | Cases | | Assays | |
|---|---|---|---|---|
| ERR | MALD | AAA | MALD | AAA |
| 1.3 | 700 | 1300 | 700 | 28 |
| 1.4 | 470 | 820 | 470 | 20 |

To evaluate the performance of AAA on real data, we examined a sample of 1646 African Americans with prostate cancer that were genotyped using 1985 ancestry informative SNPs. This sample led to the confirmation of prostate cancer risk locus in African American men through admixture mapping [3]. We simulated a pool using 600 cases that were genotyped with the same 1276 markers. The allele frequencies in the ancestral populations were estimated using a sample of 343 Europeans and 183 Africans. An ERR of 1.65 was used for the analysis based on [2]. A European genetic contribution of 0.215 was estimated using a maximum likelihood approach on the pooled sample of affected admixed individuals.

Applying AAA using a window of 4 markers results in a significant signal near a known risk locus (see Figure 4 in the appendix for more details). The peak on chromosome 8 (LOD 7.2) is less than 5Mb from the susceptibility locus reported by [3]. Applying the AAA method genome-wide yielded 2 additional less significant signals on chromosomes 5 and 9 (LOD $3.7 - 3.8$). To evaluate the robustness of the three significant signals, we applied AAA with 4-marker and LOO filtering. The analysis shows that only the known locus on chromosome 8 persist, with a significant LOD of 5.88, while the other two peaks at chromosomes 5 and 9 drop to 0.2 and 1.46, respectively. We attribute the two additional signals to biased markers.

## 4   Discussion

Pool-based methods rely on estimates of the allele frequencies in the pooled sample. It is known that pool-based association analysis is sensitive to errors in these estimates. Previous studies evaluated an error in the estimation of allele frequency difference between pools of less than 1.4% in 10K SNP arrays [17]. We now discuss the effects of these errors on AAA.

The model we used to simulate allele frequencies assumed independent normally-distributed errors with zero mean. Three error levels were tested, adjusting the variance of the error so as to reflect a $95^{\text{th}}$ percentile of 1, 3 and 5 percent error in observed allele frequency. We performed simulations using pools of 500, 1000 and 2000 admixed controls, analyzing a window of 4 markers while using LOO filtering at $1cM$ steps, and assuming an ERR between 1.3 and 1.8. Each configuration was repeated 2500 times. The results are that the selected threshold of LOD $= 3.3$ is still valid for up to 5% error in allele frequencies for the case of ERR $1.3 - 1.5$ and

$500 - 1000$ affected individuals. These results further suggest that the analysis of the prostate cancer sample is robust to 5% allele frequency estimation error. Error in the estimation of allele frequencies has a greater impact on the false-positives rate in the case of a disease with a higher ERR or a larger sample, increasing the needed significance threshold defined by the 95th percentile. One should adjust the significance threshold according to the expected allele frequency error via appropriate simulation.

We also repeated the experiment with cases, evaluating the impact of allele frequency estimation errors on the statistical power of AAA. The power of analyzing a disease with ERR 1.5 using 1000 cases decreases from 95% to 82%. The tested error levels had a smaller effect on the analysis of a larger sample or a disease with a higher ERR value, still retaining a power of over 90%. In the analysis of a smaller sample size or a disease with a lower ERR, that achieved a power between 50% and 60% under accurate allele frequency estimation, the power decreased to $33 - 38\%$ once such errors were introduced. However, in most of these settings, our simulated experiments on pooled controls suggest that a less stringent LOD threshold can be used without sacrificing the low level of false-positives.

The AAA method has an advantage over pooled association studies with respect to allele frequency estimation errors because (1) only a small fraction of SNP markers are required for the analysis, enabling the use of higher accuracy genotyping platforms, and (2) the chosen panel of markers are biased towards a high minor allele frequency in the admixed population, which increases the expected accuracy [18]. The common enhancements applied in pool-based association studies of repeated measures and the subdivision of samples into pools should also increase the robustness of our method considerably.

Another source of error lies in the inaccurate estimation of allele frequencies of the ancestral populations which may lead to an increase in the number of false-positive signals. Indeed, initial experiments indicate that errors in the ancestral allele distribution increase the false-positive signals as these mimic the effect of a true risk allele. Such results may explain few of the additional suspected regions in the prostate cancer sample that were detected prior to applying LOO.

Our analysis assumes knowledge of the admixture coefficient $P(Q)$, and the number of generations since the first admixture $\lambda$. While reasonable estimates of these parameters exists for some admixed populations, such as the African American and the Latino populations, it is recommended to tune the $\lambda$ and $P(Q)$ estimates using the sampled cases. We evaluated the genetic contribution of Europeans by applying a maximum likelihood approach on our prostate cancer cases pool, computing $P(Q = \text{Europe}) = 0.215$.

One of the properties of admixture mapping is that it can be applied on cases only, a property which holds for AAA as well. Nevertheless, similar to the use of control samples in MALD, healthy admixed individuals can increase the statistical power and decrease the rate of false-positives by providing a more accurate estimation of the allele frequencies in the ancestral population $P(J|Q)$ as well as a more accurate estimation of the admixture parameters. Admixed controls pooled

in several groups, each of similar admixture coefficient, can be used to adjust the estimates of ancestral allele frequencies using a maximum likelihood approach. In particular, measuring a marker's frequency in two African American control groups with a known and different admixture coefficient allows the estimation of the marker's frequencies in the ancestral populations via Equation 5.

The AAA method presented in Section 2 is developed for the case of an admixed population that was formed by two ancestral populations. Supporting admixed populations with more than two ancestral populations, as is the case with the Latino admixed population who are descendants of Native Americans, Europeans, and Africans, can be achieved through an adjustment of Equation 7. Another approach is to model all low risk populations as one ancestral population and the high risk population as the second ancestral population, applying the method as is.

Our multi-marker AAA method takes into account knowledge of linkage-disequilibrium evident in the ancestral population. Such inherent and complete incorporation of LD in the analysis further increases the method's statistical power, whereas other MALD methods do not fully benefit from this information, ranging from partial to no support of background LD. In addition, the analysis we developed is applied on a window of markers, while common MALD statistics employ an analysis of a single locus. Interestingly, the development steps presented in Section 2 imply that non-pooled MALD methods can also benefit from a multi-marker approach by deriving a statistic that evaluates aberration of inferred ancestries in a region, examining a range of marker locations rather than a single marker location at a time.

The goal of this work has been to alleviate the considerable cost of mapping. As the results indicate, a high power of 70% can be achieved for a disease with ethnicity prevalence differences comparable with end-stage kidney disease by pooling 1300 affected individuals, yielding a 25-fold reduction in genotyping in comparison to previous non-pooled MALD methods. We showed that AAA can be used by gene mapping groups as an economical, practical and powerful approach for the initial localization of regions containing disease genes.

## Acknowledgement

## Web Resources

The URL for AAAmap cited in the text is as follows: AAAmap Web site, *http://bioinfo.cs.technion.ac.il/AAAmap.*

# References

1. The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145), 661–678 (2007)
2. Smith, M.W., O'Brien, S.J.: Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat. Rev. Genet. 6(8), 623–632 (2005)
3. Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M., Oakley-Girvan, I., Whittemore, A.S., Cooney, K.A., Ingles, S.A., Altshuler, D., Henderson, B.E., Reich, D.: Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proceedings of the National Academy of Sciences 103(38), 14068–14073 (2006)
4. Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., Mcdonald, G.J., Greenway, S.C., Stram, D.O., Le Marchand, L., Kolonel, L.N., Frasco, M., Wong, D., Pooler, L.C., Ardlie, K., Girvan, O.I., Whittemore, A.S., Cooney, K.A., John, E.M., Ingles, S.A., Altshuler, D., Henderson, B.E., Reich, D.: Multiple regions within 8q24 independently affect risk for prostate cancer. Nat. Genet. 39(5), 638–644 (2007)
5. Kao, W.H.L., Klag, M.J., Meoni, L.A., Reich, D., Berthier-Schaad, Y., Li, M., Coresh, J., Patterson, N., Tandon, A., Powe, N.R., Fink, N.E., Sadler, J.H., Weir, M.R., Abboud, H.E., Adler, S.G., Divers, J., Iyengar, S.K., Freedman, B.I., Kimmel, P.L., Knowler, W.C., Kohn, O.F., Kramp, K., Leehey, D.J., Nicholas, S.B., Pahl, M.V., Schelling, J.R., Sedor, J.R., Thornley-Brown, D., Winkler, C.A., Smith, M.W., Parekh, R.S.: Myh9 is associated with nondiabetic end-stage renal disease in african americans. Nat. Genet. 40(10), 1185–1192 (2008)
6. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., De Jager, P.L., Mignault, A.A., Yi, Z., de Thé, G., Essex, M., Sankalé, J.L., Moore, J.H., Poku, K., Phair, J.P., Goedert, J.J., Vlahov, D., Williams, S.M., Tishkoff, S.A., Winkler, C.A., De La Vega, F.M., Woodage, T., Sninsky, J.J., Hafler, D.A., Altshuler, D., Gilbert, D.A., OBrien, S.J., Reich, D.: A high-density admixture map for disease gene discovery in african americans 74(5), 1001–1013 (May 2004)
7. Bercovici, S., Geiger, D., Shlush, L., Skorecki, K., Templeton, A.: Panel construction for mapping in admixed populations via expected mutual information. Genome Res. 18(4), 661–667 (2008)
8. Reich, D., Patterson, N.: Will admixture mapping work to find disease genes? Philosophical Transactions of the Royal Society B: Biological Sciences 360(1460), 1605–1607 (2005)
9. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., OBrien, S.J., Altshuler, D., Daly, M.J., Reich, D.: Methods for high-density admixture mapping of disease genes 74(5), 979–1000 (May 2004)
10. Tang, H., Coram, M., Wang, P., Zhu, X., Risch, N.: Reconstructing genetic ancestry blocks in admixed individuals (July 2006)
11. Bercovici, S., Geiger, D.: Inferring ancestries efficiently in admixed populations with linkage disequilibrium. Journal of Computational Biology 16(8), 1141–1150 (2009)

12. Chakraborty, R., Weiss, K.M.: Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences of the United States of America 85(23), 9119–9123 (1988)
13. Sham, P., Bader, J.S., Craig, I., O'Donovan, M., Owen, M.: Dna pooling: a tool for large-scale association studies. Nat. Rev. Genet. 3(11), 862–871 (2002)
14. Arnheim, N., Strange, C., Erlich, H.: Use of pooled dna samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: Studies of the hla class ii loci. Proc. Natl. Acad. Sci. 82(20), 6970–6974 (1985)
15. Steer, S., Abkevich, V., Gutin, A., Cordell, H.J., Gendall, K.L., Merriman, M.E., Rodger, R.A., Rowley, K.A., Chapman, P., Gow, P., Harrison, A.A., Highton, J., Jones, P.B.B., O'Donnell, J., Stamp, L., Fitzgerald, L., Iliev, D., Kouzmine, A., Tran, T., Skolnick, M.H., Timms, K.M., Lanchbury, J.S., Merriman, T.R.: Genomic dna pooling for whole-genome association scans in complex disease: empirical demonstration of efficacy in rheumatoid arthritis. Genes Immun. 8(1), 57–68 (2006)
16. Zeng, D., Lin, D.Y.: Estimating haplotype-disease associations with pooled genotype data. Genetic epidemiology 28(1), 70–82 (2005)
17. Kirov, G., Nikolov, I., Georgieva, L., Moskvina, V., Owen, M.J., O'donovan, M.C.: Pooled dna genotyping on affymetrix snp genotyping arrays. BMC Genomics 7(1) (February 2006)
18. Wilkening, S., Chen, B., Wirtenberger, M., Burwinkel, B., Forsti, A., Hemminki, K., Canzian, F.: Allelotyping of pooled dna with 250k snp microarrays. BMC Genomics 8, 77 (2007)
19. Darvasi, A., Soller, M.: Selective DNA Pooling for Determination of Linkage Between a Molecular Marker and a Quantitative Trait Locus. Genetics 138(4), 1365–1373 (1994)
20. Amundadottir, L.T., Sulem, P., Gudmundsson, J., Helgason, A., Baker, A., Agnarsson, B.A., Sigurdsson, A., Benediktsdottir, K.R., Cazier, J.B., Sainz, J., Jakobsdottir, M., Kostic, J., Magnusdottir, D.N., Ghosh, S., Agnarsson, K., Birgisdottir, B., Le Roux, L., Olafsdottir, A., Blondal, T., Andresdottir, M., Gretarsdottir, O.S., Bergthorsson, J.T., Gudbjartsson, D., Gylfason, A., Thorleifsson, G., Manolescu, A., Kristjansson, K., Geirsson, G., Isaksson, H., Douglas, J., Johansson, J.E., Bälter, K., Wiklund, F., Montie, J.E., Yu, X., Suarez, B.K., Ober, C., Cooney, K.A., Gronberg, H., Catalona, W.J., Einarsson, G.V., Barkardottir, R.B., Gulcher, J.R., Kong, A., Thorsteinsdottir, U., Stefansson, K.: A common variant associated with prostate cancer in european and african populations. Nature Genetics 38(6), 652–658 (2006)
21. Long, J.C.: The Genetic Structure of Admixed Populations. Genetics 127(2), 417–428 (1991)
22. The International HapMap Project.: A haplotype map of the human genome. Nature 437(7063), 1299–1320 (2005)

# Appendix

**Table 2.** The $95^{th}$ percentile of LOD scores using pools of 500, 1000 and 2000 simulated controls analyzed using a window of 2, 3 and 4 markers under the false assumption of ERR between 1.3 and 1.8. All tested configurations exhibit a score lower than 3.3 in the $95^{th}$ percentile. The simulations demonstrate that in most cases an increase in either sample size or in the size of the sliding window results in a reduction of the threshold.

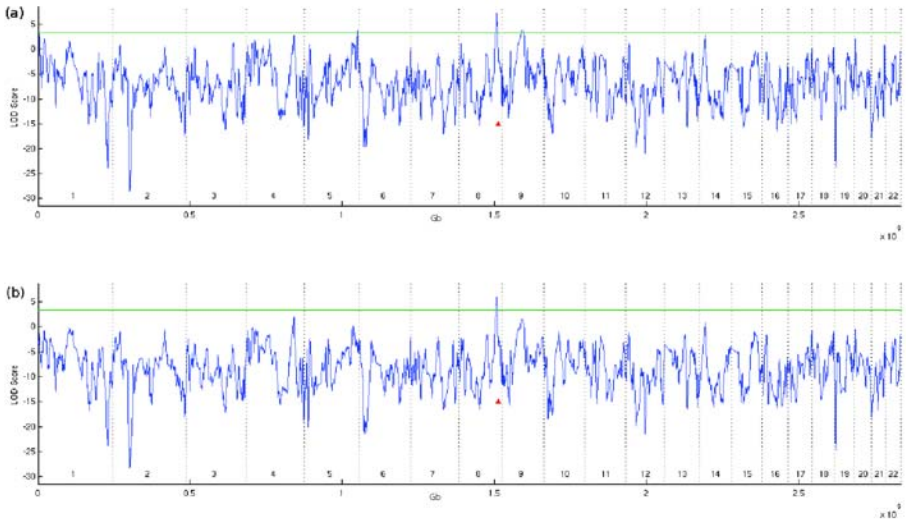| | 2 Markers Window | | | 3 Markers Window | | | 4 Markers Window | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample Size | 1.3 | 1.5 | 1.8 | 1.3 | 1.5 | 1.8 | 1.3 | 1.5 | 1.8 |
| 500 | 2.8 | 3.28 | 3.12 | 2.83 | 3.28 | 2.99 | 2.84 | 3.26 | 2.72 |
| 1000 | **3.29** | 3.14 | 1.72 | **3.3** | 2.89 | 0.78 | **3.28** | 2.75 | 0.26 |
| 2000 | 3.28 | 1.56 | -1.78 | 3.06 | 0.65 | -4.4 | 2.93 | 0.05 | -6.39 |



**Fig. 4.** The analysis of 600 prostate cancer cases using AAA and a 4 markers window. (a) The significant peak of 7.2 LOD is evident in close proximity to a validated prostate cancer risk locus at 129 *Mb* (marked by a triangle) that was previously discovered through a linkage scan by Amundadottir *et al.* [20] and later reported by Freedman *et al.* [3] using admixture mapping. Two additional significant signals are evident on chromosome 5 and 9. (b) Only the validated locus passes the LOO filter with a significant LOD of 5.88.