Panel Construction for Mapping in Admixed Populations Via Expected Mutual Information

Sivan Bercovici[1], Dan Geiger[1], Liran Shlush[2],
Karl Skorecki[2], and Alan Templeton[3]

[1] Computer Science Department
Technion, Haifa 32000, Israel
{sberco,dang}@cs.technion.ac.il
[2] Faculty of Medicine
Technion, Haifa 31096, Israel
{lshlush,skorecki}@tx.technion.ac.il
[3] Department of Biology
Washington University , St. Louis, Mo. 63130, USA
templeton@biology2.wustl.edu

Abstract. Mapping by Admixture Linkage Disequilibrium (MALD) is
an economical and powerful approach for the identification of genomic
regions harboring disease susceptibility genes in recently admixed pop-
ulations. We develop an information-theory based measure, called EMI
(expected mutual information), that computes the impact of a set of
markers on the ability to infer ancestry at each chromosomal location.
We then present a simple and effective algorithm for the selection of
panels that strives to maximize the EMI score. Finally, we demonstrate
via well established simulation tools that our panels provide considerably
more power and accuracy for inferring disease gene loci via the MALD
method in comparison to previous methods.

1 Introduction

Mapping by admixture linkage disequilibrium (MALD) is an economical and
powerful approach for the identification of genomic regions harboring disease
susceptibility genes in recently admixed populations [11,9]. For the method to
be useful, the prevalence of the disease under study should be considerably dif-
ferent between the ancestral populations from which the admixed population
was formed.

Myeloma, for example, is a type of cancer that is approximately three times
more prevalent in Africans than in Europeans [11]. Hepatitis C clearance is ap-
proximately five times more prevalent in Europeans than in Africans. Stroke, lung
cancer, prostate cancer, dementia, end-stage renal disease, multiple-sclerosis, hy-
pertension and many more diseases all exhibit a higher morbidity in either Africans
or Europeans, when the two ethnically-different populations are compared [11].
This difference in susceptibility to a specific disease is also evident in other ethnic
populations. Native Americans suffer from a high prevalence of type 2 diabetes,

obesity and gallbladder disease, while showing a lower prevalence of asthma, relative to Europeans [8].

When examining an individual that originated from several ancestral populations, such as African-Americans, the likelihood that this individual will carry a given disease is influenced by the susceptibility to the disease in the ancestral populations. When such an admixed individual carries a hereditary disease, the chances are higher that the disease gene or genes are harbored in chromosomal segments that originated from the ancestral population with the higher risk.

The MALD method screens through the genome of either affected or both affected and healthy admixed individuals, looking for chromosomal segments with an unusually high representation of the high-risk ancestry population for the disease. MALD requires 200 to 500-fold fewer markers, in comparison to genome-wide association mapping, while offering the same power [9]. Consequently, the method has an economical advantage over alternative methods. Lately, successful results from admixture mapping have begun to emerge. For example, the usage of MALD led to the discovery of multiple risk alleles (gene variants) for prostate cancer [2].

In this paper, we develop an information-theory based measure, called *EMI* (expected mutual information), to select an effective panel of markers to be used in MALD. Our measure, presented in Section 4, computes the total impact of a set of markers on the ability to infer ancestry at each chromosomal location, averaged over all possible admixture related recombinations. This method improves previous measures such as the Shannon Information Content (SIC)[10], and Fisher Information Content (FIC)[6]. We then present, in Section 5, a simple and effective algorithm for the selection of panels that strives to maximize the EMI score. In Section 6, we demonstrate via well established simulation tools used in previous studies, that our panels provide considerably more power for inferring disease gene loci. For example, our simulations show that in the challenging case of a disease with an ethnicity risk ratio of 1.6 between two ancestral populations, the power increased from 50% to 68%, namely, an increase of about 36% in the ability to detect the loci of disease susceptibility genes. The detection accuracy has also significantly improved with the use of our new panels. The increase in power is particularly important in the detection of weak signals that underlie complex diseases. Section 7 concludes with extensions and discussion.

2   Background

The MALD method consists of three steps. First, an admixed population with a significantly higher risk for a specific disease in one of the ancestral populations is identified. Ancestry informative markers that effectively distinguish between the relevant ancestral populations are selected, and either cases or both cases and controls are genotyped. Second, the ancestry along the chromosomes of every individual is computed based on the sampled genotypes. Third, chromosomal regions with an elevated frequency of the ancestral population with the higher
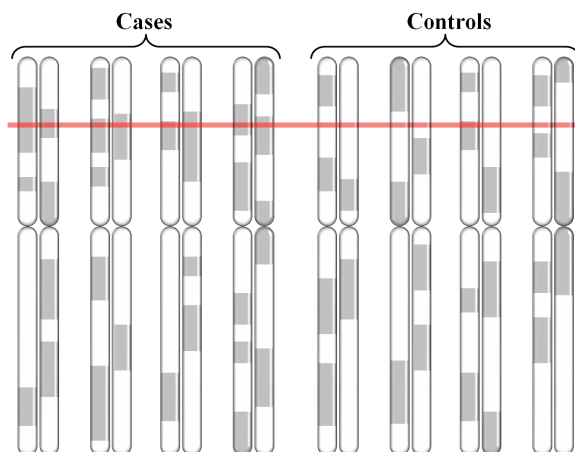
Fig. 1. Ancsetry informative markers are used to compute the ancestry across the chromosomes of cases and controls. The region indicated by the bar shows elevated frequency of one ancestral population in the cases versus the expected distribution of ancestry in the controls, suggesting a disease susceptibility locus.

disease prevalence are identified. Figure 1 shows the ancestral profile of eight individuals, of which half are cases and half controls. The ancestral profiles are indicated as dark and light segments along the chromosomes. The excess of one ancestral population in the cases at the locus marked by the bar suggests that the locus contains the disease susceptibility gene. In the controls, the ancestry at the same locus matches the expected distribution of ancestry, strengthening the hypothesis. The detection of suspected regions can be followed by methods such as high density SNP-based association studies, or a study of nearby candidate genomic regions.

Choosing ancestry informative markers (AIM) for the construction of MALD panels has been pursued in several studies. AIM panels were constructed for African-American [12,15], Mexican-American [14] and Hispanic/Latino [4,8] populations. The construction of such panels requires three ingredients: a database of markers, a measure for the informativeness of a set of markers regarding ancestry, and an algorithm that selects informative markers for the MALD panel.

The work of Rosenberg et. al. [10] introduced a measurement for the information multialleleic markers provide on ancestry, based on the Shannon Information Content. Pfaff et. al. [6] based their measurement on the Fisher Information Content.

The algorithms used for panel construction in the studies that followed were driven by two prime objectives: (1) choose markers with the highest ancestry-informativeness (2) choose evenly spread markers. These guidelines were set to provide informative panels for the estimation of ancestry at each point along the genome. Current panel construction algorithms are "greedy", attempting to locally maximize an informativeness criterion, whilst investing less effort in

ensuring that the chosen markers are evenly spaced or that the informativeness along the genome is well balanced. Smith et. al. used a purely greedy algorithm for marker selection [12]. Tian et. al. divided the chromosome into windows, choosing multiple highly-informative markers within every such window [15].

When considering the informativeness of a set of markers regarding the ancestry at an arbitrary point, previous work offered rough approximations. Smith et. al. considered the informativeness of a set of markers within a constant-size window centered on the point examined as an approximation to the informativeness at that point. Tian et. al. used the mean informativeness between two adjacent markers bounding the point examined. It is this deficiency that is addressed in the current paper. In the next section we develop an improved measure and demonstrate through simulations that panels constructed using our measure provides increased power in the detection of disease susceptibility gene loci.

3   Admixed Individuals Model

The genome of a recently admixed individual is a mosaic of large chromosomal segments, where each segment originated from a single ancestral population. We use the following definitions to describe these segments in admixed individuals.

Definition 1. *An **admixed chromosome** is a chromosome that originated from more than one ancestral population.*

Definition 2. *A **Post Admixture Recombination** point (abbreviated PAR) is a recombination point in which either two chromosomes from different populations crossed, or two chromosomes crossed when at least one of the chromosomes is an admixed chromosome.*

Definition 3. *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.

Figure 2 illustrates the propagation of PAR points along three generations of admixture, and the PAR blocks they induce. In particular, Figure 2 shows a grandmother originating from one population, and a grandfather originating from two populations, yielding a parent with one admixed chromosome (with one PAR point) and one non-admixed chromosome. As the parent's chromosomes recombine to produce the child's admixed chromosome, a second PAR point is added. Hence, three recombination points reside on the child's chromosome, of which only two are PAR points (colored black). Three PAR blocks are defined rather than four as the leftmost recombination point is not a PAR point.

We denote the set of all observed markers by J, and the vector of an individual's PAR-blocks ancestries as Q. The set of an individual's PAR points defines a partition (denoted π) of the chromosomes into blocks. We use the random
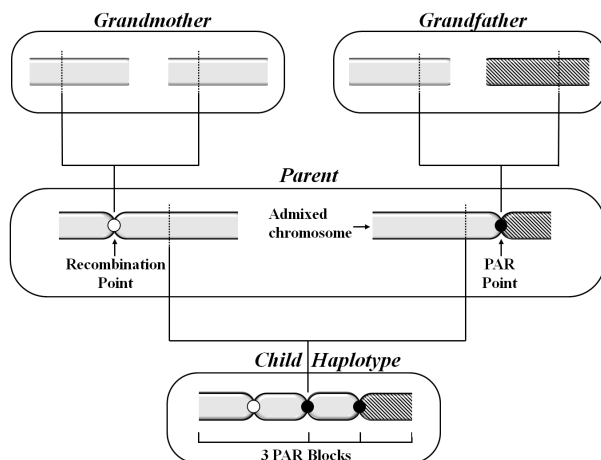
Fig. 2. Three generations admixture example. PAR blocks are limited only by PAR points and the chromosomes' ends.

variable Q_π to denote the vector of ancestries corresponding to the PAR-blocks determined by π, $Q_{\pi,i}$ to denote the ancestry (out of K possible ancestral populations) of the ith PAR block in the given partition π, and the random vector $J_{\pi,i} = \{J_{\pi,i,1}, J_{\pi,i,2}, ..., J_{\pi,i,m_i}\}$ to denote the haplotype assignment of the set of $m_{\pi,i}$ observed markers within this block. Reference to subscript π will be omitted whenever π is clear from the context.

Markers within a PAR block are assigned according to the probability function of the corresponding ancestral population. We further assume that the ancestries of all PAR blocks of a given partition π are mutually independent. A graphical model showing these assumptions is given in Figure 3.
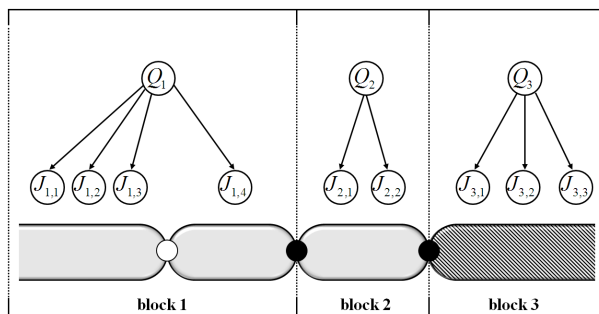

Fig. 3. Graphical model for $P(Q, J)$ assuming markers J_i within a PAR block are independent conditioned on ancestry Q_i. Ancestries of PAR blocks are mutually independent.

The joint probability distribution described via the graphical model is given by

$$P(Q, J) = \sum_{\pi} P(\pi) \cdot \prod_{i=1}^{|Q_{\pi}|} P(Q_{\pi,i}) \prod_{j=1}^{m_{\pi,i}} P(J_{\pi,i,j}|Q_{\pi,i}) \qquad (1)$$

In particular, when considering a specific point x on the genome, the joint probability for the ancestry Q_x at that point is given by

$$P(Q_x, J) = \sum_{\pi} P(\pi) \cdot P(Q_x, J_{\pi,x}) \cdot P(\overline{J}_{\pi,x}) \qquad (2)$$

where $J_{\pi,x}$ are the markers within the same PAR-block as location x, and $\overline{J}_{\pi,x}$ is the complementary set of markers outside this block. We use this joint distribution to derive our panel informativeness measure.

4   Informativeness of Panels

In this section we develop a measure for the contribution of a set of observed markers to the ability to infer the ancestry of a block conditioned on a partition π. We then extend this measure to account for the fact that π is unknown by computing the expectation over all possible partitions, while focusing on the inference of a single location x.

We start by exploring the relationship between observed markers and the ancestries of PAR blocks under the assumption that the partition is known. Using information-theory, we estimate the extent to which a set of markers contribute to the ability to infer ancestry by measuring the informativeness of a set of markers regarding ancestry. The information gain for ancestry due to observing a set of markers can be described by the well known Shannon Information Content (SIC)

$$I(Q_i; J_i) = H(Q_i) - H(Q_i|J_i) \qquad (3)$$
$$= \sum_{Q_i} \sum_{J_i} P(J_i|Q_i) \cdot P(Q_i) \cdot \log \frac{P(J_i|Q_i)}{P(J_i)}$$

where $H(Q_i)$ is the entropy (or the amount of uncertainty) of the PAR block's ancestry, given by

$$H(Q_i) = - \sum_{Q_i=1}^{K} P(Q_i) \cdot \log P(Q_i)$$

and $H(Q_i|J_i)$ is the conditional entropy on ancestry once the markers observations are accounted for, given by

$$H(Q_i|J_i) = - \sum_{Q_i} \sum_{J_i} P(J_i, Q_i) \cdot \log P(Q_i|J_i).$$

In other words, the markers' informativeness is measured by the reduction in uncertainty regarding the ancestry of a given location due to observing these markers. This reduction in uncertainty originates from the fact that each ancestral population has a distinct distribution over the haplotype. The information gain in each PAR block is computed separately through Equation 3 due to our assumption of mutual independence.

The possible presence of linkage disequilibrium between markers within a block raises difficulties partially stemming from the need to estimate the joint probability of a haplotype J_i that contains multiple markers conditioned on the ancestry (i.e., $P(J_i|Q_i)$). To reduce computational cost, we assume conditional independence between all markers given ancestry, yielding a simpler form of mutual-information $I_{ind}(Q_i; J_i)$, explicated in Lemma 1. The relaxation of this assumption is pursued in Section 7.

Lemma 1. *For a given PAR Block, let Q_i be its ancestry, and $J_{i,j}$ be its jth marker (out of m_i markers). Under the assumption that the markers are conditionally independent given Q_i, the mutual information of Q_i and J_i is:*

$$I_{ind}(Q_i; J_i) = H(J_i) - \sum_{j=1}^{m_i} H(J_{i,j}|Q_i) \tag{4}$$

Given a partition π, all PAR blocks are determined, and the informativeness of markers regarding ancestry Q_i, and in particular regarding ancestry Q_x of an arbitrary location x within the ith PAR block, is the informativeness of the markers in J_i alone. All other markers are not informative regarding Q_x. However, π is not known, and for every π a different block contains a location x. Hence, for each π, a different set of markers is informative. The expected informativeness of all markers regarding ancestry at location x is given, in principle, by

$$EMI(Q_x; J) = \sum_{\pi} P(\pi) \cdot I(Q_x; J|\pi) \tag{5}$$

We call this measure EMI for Expected Mutual Information. Since summing over all possible partitions is not feasible, the rest of this section rewrites Equation 5 and explicates how to compute it.

Observe that for any two partitions π_1 and π_2 such that the PAR block that contains location x also contains the same set of markers $J_{\pi,x} \subseteq J$, the term $I(Q_x; J|\pi)$ in Equation 5 is equal. The probability for a partition π to contain a block that contains both location x and markers $J_{\pi,x}$ is defined by three events:

1. The minimal segment $[l, r]$ that spans over $J_{\pi,x}$ and x does not contain a PAR point.
2. The segment between l and the marker to its left (at l'), if such exists, contains a PAR point.
3. The segment between r and the marker to its right (at r'), if such exists, contains a PAR point.

Assuming PAR points are distributed independently, the aforementioned three events are independent as well. This holds because the corresponding three segments are mutually exclusive. Hence, the probability of a partition π to contain a PAR block containing location x and markers $J_{\pi,x}$ alone is given by the product

$$P_{(l,r)} = P(N_{[l',l]} \neq 0) \cdot P(N_{[l,r]} = 0) \cdot P(N_{[r,r']} \neq 0) \tag{6}$$

where $N_{[a,b]}$ is a random variable of the number of PAR points in segment $[a,b]$, and $[l,r]$ is the minimal segment containing location x and markers $J_{\pi,x}$.

The term $P(N_{[l',l]} \neq 0)$ depends on the existence of a marker at l', hence the term will not appear in Equation 6 in case there is no marker to the left of l. Similarly, $P(N_{[r,r']} \neq 0)$ will not appear in Equation 6 if there is no marker to the right of r.

Let $J_{[l,r]}$ denote a sequence of markers within a segment $[l,r]$, and $location(j)$ denote the location of a marker $j \in J$. To compute EMI, we weight the potential contribution $I(Q_x; J_{[l,r]})$ by the probability of such a contribution, namely the probability $P_{(l,r)}$ of a partition π to contain location x and markers $J_{[l,r]}$ within the same block.

Theorem 1. *Let Q_x be the ancestry at location x, and J the set of observed markers. The expected mutual information between Q_x and J is*

$$EMI(Q_x; J) = \sum_{l \in L} \sum_{r \in R} P_{(l,r)} \cdot I(Q_x; J_{[l,r]}) \tag{7}$$

where

$$L = \{location(j) \leq x | j \in J\} \cup \{x\},$$
$$R = \{location(j) \geq x | j \in J\} \cup \{x\}.$$

The common realization of the term $P_{(l,r)}$ in Equation 7 is via the exponential distribution. In particular $P(N_{[a,b]} = 0) = e^{-\lambda \cdot |b-a|}$, where λ is the rate of PAR points in admixed individuals, as derived from the admixture model being used. Consequently,

$$P_{(l,r)} = (1 - e^{-\lambda \cdot |l-l'|}) \cdot e^{-\lambda \cdot |r-l|} \cdot (1 - e^{-\lambda \cdot |r'-r|}). \tag{8}$$

Equation 7 defines the EMI at a specific location x. The average information gain regarding the entire chromosome is given by

$$\text{EMI}_{avg}(J) = \frac{1}{|N|} \cdot \sum_{x \in N} EMI(Q_x, J) \tag{9}$$

which measures the the average EMI along the chromosome. The set N consists of all locations x on an evenly spaced grid with a specific resolution. For example, for chromosome 1, a set N of 280 points means about one location per cM.

In the task of mapping disease genes in admixed populations using the MALD method, panels of high EMI_{avg} are shown in Section 6 to outperform previous panels.

5   Panel Construction

We employ a greedy algorithm that constructs panels of markers for which the EMI_{avg} is high. In principle, the algorithm iterates over the candidate markers, selecting the marker with the highest EMI_{avg} gain given the markers chosen so far. Namely, in each iteration the algorithm chooses a marker j that maximizes

$$\text{EMI}_{avg}(J \cup \{j\}) - \text{EMI}_{avg}(J) \tag{10}$$

where J is the set of markers selected so far.

The evaluation of EMI_{avg} is a computationally intensive task that is repeated with every iteration, and for every candidate marker. To reduce execution time, for each examined candidate we locally evaluate EMI_{avg} on a set of points located in a segment of length w centered on the candidate marker. Equation 11 evaluates the EMI_{avg} on a subset of points $w_j \subseteq N$

$$\text{EMI}_{avg}(J) = \frac{1}{|w_j|} \cdot \sum_{x \in w_j} \text{EMI}(Q_x, J) \tag{11}$$

where

$$w_j = \{p \in N| \quad location(j) - \frac{w}{2} \le p \le location(j) + \frac{w}{2}\}$$

rather than on the entire chromosome. Once a marker j is chosen, the EMI_{avg} gain in the next iteration is computed only for those markers that are within w_j, as the last chosen marker mostly affects their potential gain.

The most computationally dominant factor in EMI is the evaluation of $H(J)$ (Equation 4), as it is exponential in the number markers $|J|$. However, for a given PAR-block, a small number of ancestry informative markers suffice to nearly eliminate the uncertainty regarding its ancestry; the information gain regarding the ancestry of the PAR-block saturates rapidly as the number of markers within the PAR-block increases. Hence, limiting the number of markers used in the evaluation of Equation 4 yields an eligible approximation. In our implementation, we limited the number of markers in the evaluation of Equation 4 to a maximum of 17 markers, offering a plausible tradeoff between performance and approximation accuracy.

6   Evaluation

In this section we demonstrate the power of panels produced by our algorithm and EMI. We compare performance with the works of Smith et. al. [12], and Tian et. al. [15].

Similarly to the panels of Smith et. al. and Tian et. al., we constructed a panel for the African-American admixed population. The International HapMap Project [13] was used as the SNP allele frequencies source for the two ancestral populations, namely the west-African and European populations. HapMap has been shown to reflect these two distinct populations well [1].

We constructed a panel of 148 markers (denoted EMI-148) for chromosome 1, matching the number of corresponding markers in the screening panel of Tian et. al.. The panel of Smith et. al. contains 238 markers. We further constructed a more economical panel of 100 markers for chromosome 1 (denoted EMI-100) which is two thirds the number of markers in the panel constructed by Tian et. al. Based on the admixture-dynamics of African-Americans as described in [3,11,12], we used $\lambda = 6$ (Equation 8), a proportion of 0.8 African/0.2 European contribution to the admixed population, and $w = 45cM$ (Equation 11).

We first examine the performance of the four panels according to the EMI measure. The maximal EMI value is derived from the admixed-population characteristics, namely the number of admixed populations and the admixture coefficient. In the case of African-Americans, the maximal value is approximately 0.5. As illustrated in Figure 4, the proportion of chromosome 1 above most EMI thresholds is higher in the panels constructed in the current work. Moreover, the EMI-148 panel constructed by our algorithm has a low EMI standard-deviation of 0.0041 in comparison to the screening panel of Tian et. al. (0.0142) and the panel of Smith et. al. (0.0178); indeed, our algorithm strives to balance the informativeness of markers across the chromosome. It is interesting to note that our lighter panel, EMI-100, has good performance as well, with a low EMI standard-deviation of 0.0056.

ANCESTRYMAP [5] is a tool we used for the estimation of the ancestral origin of a locus on the basis of sampled genotypes. Given genotypes of cases and controls, the tool can compute the likelihood of each point along the genome to be the disease locus, under a specified disease model. ANCESTRYMAP can also generate admixed-individual genotypes for cases and controls under a given
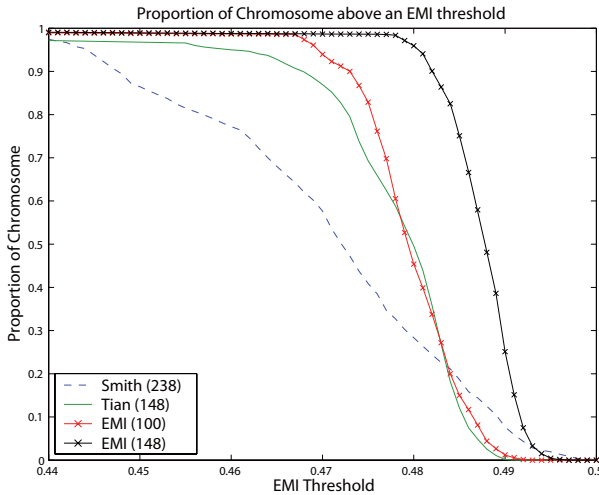

Fig. 4. Proportion of chromosome above an EMI threshold. For most levels of informativeness, our panel covers larger segments of the chromosome.

disease model. This software was used in [12] and [15] to evaluate the power of the Smith and Tian panels, respectively.

In the experiments conducted, we generated 576[1] admixed-individual cases per experiment through the use of ANCESTRYMAP. In each experiment, a single location on chromosome 1 was used as the disease predisposition locus. In order to evaluate the performance of the panel across the entire chromosome, a set of disease predisposition loci were chosen using a resolution of 4 points per cM; above 900 uniformly selected locations across chromosome 1 were used in the experiments. Multiple disease models were examined, all with higher prevalence in the European population. A range of ethnicity relative risk (ERR) factors, between 1.4 and 1.8, were set as the disease model parameters. We focused on this range as it captures diseases such as stroke and lung cancer [11], which are considered mild in their ERR, hence harder to detect. We proceeded by employing ANCESTRYMAP to locate the disease gene. Similar to the threshold used for the evaluation of Tian's panel [15], we used a LOD score above 4.0 as an indicator for successful detection. Figure 5 shows the power, namely, the detection success-rate, in a total of 5500 experiments per panel.
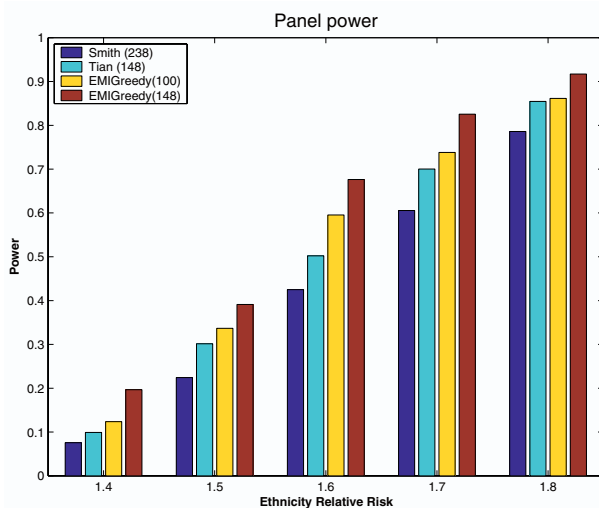

Fig. 5. EMI-148 achieves a significantly higher power in all tested disease models

Measuring the distance between the highest detection signal and the actual disease predisposition locus reveals that our panel also has a high detection accuracy, as illustrated in Figure 6. Approximately 55% of the experiments conducted on the entire range of ERR using our EMI-148 panel detected a signal within a 3 cM distance from the actual disease predisposition locus, whilst the other two panels achieved approximately 42% (Tian et. al.) and 37% (Smith et. al.). The EMI-100 panel achieved a 46% detection-rate within 3 cM.

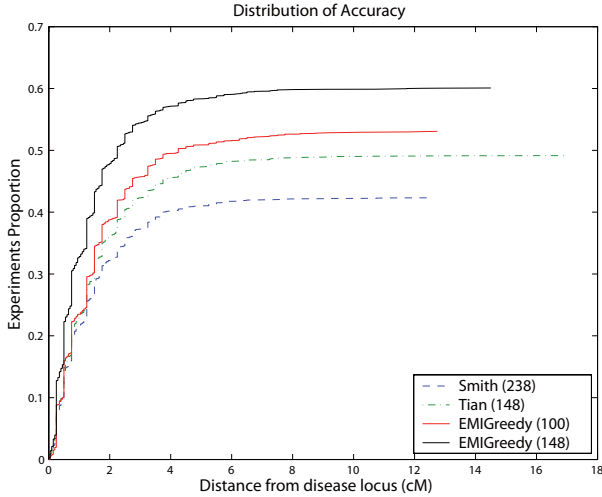[1] Commercial panel infrastructure requires sample size to be a multiplicative of 96.

Fig. 6. Experiments percentage above accuracy threshold for each of the four panels. A higher percentage of the experiments yield higher accuracy for EMI-148, in comparison to the other three panels.

Detailed information regarding our panel can be found at *bioinfo.cs.technion. ac.il/MALD*.

7   Extensions and Discussion

The EMI measure, introduced in Section 4, provides an estimate for the informativeness of a set of markers regarding ancestry at a specific location. It improves upon previous measures as it takes into account the expected informativeness of a set of markers with respect to ancestry, over all possible partitions. The higher accuracy of EMI, especially in regions between markers, enables the creation of panels that are well balanced in terms of the informativeness provided by the set of markers across the genome. Finally, the panels constructed by our algorithm demonstrated significantly higher power and accuracy.

An immediate extension of EMI that we pursued addresses possible dependencies between markers given ancestry. Lemma 1 disregards LD within ancestral population in favor of a lower computational cost. We now use a first-order Markov-chain to model marker dependencies within ancestral populations in order to provide a more accurate model. Under this model, the transition probabilities are derived from the LD present between every two adjacent markers given the ancestry. Such a model still yields a computationally plausible form, as shown in the next lemma.

Lemma 2. *For a given PAR Block, let Q_i be the ancestry, and $J_{i,j}$ be the jth marker (out of m_i markers). Under the assumption that each marker is*

*dependent on the preceding marker and conditionally independent of the rest
of the markers given Q_i, the mutual information of Q_i and J_i is:*

$$I_{chain}(J_i; Q_i) = \qquad H(J_i) - H(J_{i,1}|Q_i) - \sum_{j=2}^{m_i} H(J_{i,j}|Q_i, J_{i,j-1}) \qquad (12)$$

However, we have not employed this extension in the panel constructed because
the public data of LD categorized by ancestral population is still too sparse.
In addition, as our algorithm inherently yields a balanced panel in terms of
EMI_{avg}, the selected markers tend to be evenly spaced (hence with a large inter-
distance), decreasing the probability for LD between two consecutive markers,
given ancestry.

Another extension of EMI relaxes the assumption that the rate of PAR points,
used in Equation 8, is constant across the chromosomes. Recombinational hot-
spots can be taken into account by using a PAR point rate as a function of
location $\lambda(x)$ instead of the constant rate λ. For example, assume that a chro-
mosome is divided into regions of different PAR point rates $\lambda_1, \lambda_2, ..., \lambda_n$. For a
segment $[l, r]$ that spans two consecutive regions with PAR rates λ_i and λ_{i+1},
the term $P(N_{[l,r]})$ in Equation 6 equals $e^{-(\lambda_i \cdot t + \lambda_{i+1} \cdot (1-t)) \cdot |r-l|}$, where t is the
proportion of segment $[l, r]$ with PAR rate λ_i. Furthermore, the framework pro-
vided in this paper can address the effect on EMI of different admixture models,
such as continuous-gene-flow [7], by considering a different realization of Equa-
tion 6.

We note that EMI assumes a model for haplotypes rather than genotypes,
and that the allele frequencies $P(J|Q)$ are definite. In reality these frequencies
are derived from a small set of samples (60, barring missing data, in the case of
HapMap). In its current form, EMI lacks an appropriate treatment for the uncer-
tainty involving allele frequencies. It is advisable to validate the allele frequencies
by taking more samples for candidate markers, as done in [15].

The approach presented in this paper for panel construction also applies to
the second phase of the MALD method. This second step currently employs a
Markov chain model that assigns the most probable ancestry for each location,
given the model and marker data [5,3]. By conditioning on possible partitions
π, one can compute the expected ancestry $P(Q_x|J = j)$ of a point x given mea-
surements $J = j$ via Equation 2, similarly to our computation of the expected
informativeness. It would be interesting to see whether this approach yields
higher accuracy in ancestry inference.

Finally, further research could determine the relative contribution of each of
the following three ingredients to the reported increase in power: the richer set of
SNP markers available in the current HapMap project, the validity of the EMI
measure, and the success of our proposed algorithm. Nevertheless, we clearly
showed that the panel produced using EMI has a well balanced high score in
terms of informativeness of markers, yielding a significant improvement in both
power and accuracy, compared to previous work.

Acknowledgements

References

1. Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., Pritchard, J.K.: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics 38, 1251–1260 (2006)
2. 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., Marchand, L.L., Kolonel, L.N., Frasco, M., Wong, D., Pooler, L.C., Ardlie, K., Oakley-Girvan, 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. Nature Genetics 39, 638–644 (2007)
3. Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., McKeigue, P.M.: Design and analysis of admixture mapping studies. The American Journal of Human Genetics 74, 965–978 (2004)
4. Mao, X., Bigham, A., Mei, R., Gutierrez, G., Weiss, K., Brutsaert, T., Leon-Velarde, F., Moore, L., Vargas, E., McKeigue, P., Shriver, M., Parra, E.: A genome-wide admixture mapping panel for hispanic/latino populations. The American Journal of Human Genetics 80, 1171–1178 (2007)
5. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., Daly, M.J., Reich, D.: Methods for high-density admixture mapping of disease genes. The American Journal of Human Genetics 74, 979–1000 (2004)
6. Pfaff, C.L., Barnholtz-Sloan, J., Wagner, J.K., Long, J.C.: Information on ancestry from genetic markers. Genetic Epidemiology 26, 305–315 (2004)
7. Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E., Shriver, M.D.: Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. The American Journal of Human Genetics 68, 198–207 (2001)
8. Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., Duque, C., Villegas, A., Bortolini, M.C., Salzano, F.M., Gallo, C., Mazzotti, G., Tello-Ruiz, M., Riba, L., Aguilar-Salinas, C.A., Canizales-Quinteros, S., Menjivar, M., Klitz, W., Henderson, B., Haiman, C.A., Winkler, C., Tusie-Luna, T., Ruiz-Linares, A., Reich, D.: A genomewide admixture map for latino populations. The American Journal of Human Genetics 80, 1024–1036 (2007)
9. Reich, D., Patterson, N.: Will admixture mapping work to find disease genes? Philosophical Transcations of the Royal Society B 360, 1605–1607 (2005)
10. Rosenberg, N.A., Li, L.M., Ward, R., Pritchard, J.K.: Informativeness of genetic markers for inference of ancestry. The American Journal of Human Genetics 73, 1402–1422 (2003)

11. Smith, M.W., O'Brien, S.J.: Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nature Review Genetics 6, 623–632 (2005)
12. 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 The, G., Essex, M., Sankale, 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., O'Brien, S.J., Reich, D.: A high-density admixture map for disease gene discovery in african americans. The American Journal of Human Genetics 74(5), 1001–1013 (2004)
13. The International HapMap Project. A haplotype map of the human genome. Nature Publishing Group 437, 1299–1319 (2005)
14. Tian, C., Hinds, D.A., Shigeta, R., Adler, S.G., Lee, A., Pahl, M.V., Silva, G., Belmont, J.W., Hanson, R.L., Knowler, W.C., Gregersen, P.K., Ballinger, D.G., Seldin, M.F.: A genomewide single-nucleotidepolymorphism panel for mexican american admixture mapping. The American Journal of Human Genetics 80, 1014–1023 (2007)
15. Tian, C., Hinds, D.A., Shigeta, R., Kittles, R., Ballinger, D.G., Seldin, M.F.: A genomewide single-nucleotide–polymorphism panel with high ancestry information for african american admixture mapping. The American Journal of Human Genetics 79, 640–649 (2006)