

LinkageTracker: A Discriminative Pattern Tracking Approach to Linkage Disequilibrium Mapping

¹Li Lin, ²Limsoon Wong, ³Tzeyun Leong, ⁴Pohsan Lai

^{1,3}School of Computing, National University of Singapore

²Institute for Infocomm Research, Singapore

⁴Dept of Pediatrics, National University Hospital, National University of Singapore

Email: {¹linl,³leongty}@comp.nus.edu.sg; ²limsoon@i2r.a-star.edu.sg; ⁴paelaips@nus.edu.sg

Abstract. Linkage disequilibrium mapping is a process of inferring the disease gene location from observed associations of marker alleles in affected patients and normal controls. In reality, the presence of disease-associated chromosomes in affected population is relatively low (usually 10% or less). Hence, it is a challenge to locate these disease genes on the chromosomes. In this paper, we propose an algorithm known as LinkageTracker for linkage disequilibrium mapping. Comparing with some of the existing work, LinkageTracker is more robust and does not require any population ancestry information. Furthermore our algorithm is shown to find the disease locations more accurately than a closely related existing work, by reducing the average sum-square error by more than half (from 80.71 to 30.83) over one hundred trials. LinkageTracker was also applied to a real dataset of patients affected with haemophilia, and the disease gene locations found were consistent with several studies in genetic prediction.

1 Introduction

Linkage disequilibrium mapping has been used in the finding of disease gene locations in many recent studies [6][13]. The main idea of linkage disequilibrium mapping is to identify chromosomal regions with common molecular marker alleles¹ at a frequency significantly greater than chance. It is based on the assumption that there exists a common founding ancestor carrying the disease alleles, and is inherited by his descendents together with some other marker alleles that are very close to the disease alleles. The same set of marker alleles is detected many generations later in many unrelated individuals who are clinically affected by the same disease. In a realistic setting, the occurrence of such allele patterns is usually very low, and most often consist of errors or noise. For instance, the hereditary mutations of BRCA-1 and BRCA-2 genes only account for about five to ten percent of all breast cancer patients[12]. Assuming that we know that BRCA-1 gene resides somewhere on chromosome 17, the finding of the exact location of BRCA-1 gene on chromosome 17 based on a set of sample sequence collected from breast cancer patients where at most ten percent of the sample sequence exhibit allelic association or linkage disequilibrium is a nontrivial task. To further complicate this task, the linkage disequilibrium patterns also consist of errors due to sample mishandling and contamination.

Due to errors and low occurrence of linkage disequilibrium patterns, existing data mining and artificial intelligence methods involving training and learning will not be applicable. In this paper, we propose a novel method known as *LinkageTracker* for the finding of linkage disequilibrium patterns and inference of disease gene locations. First of all, we identify the set of linkage disequilibrium patterns using a heuristic level-wise neighbourhood search and score each pattern by computing their p -values to ensure high discriminative powers of each pattern. After which, we infer the marker allele that is closest to the disease gene based on the p -value scores of the set of linkage disequilibrium patterns. *LinkageTracker* is a nonparametric method as it is not based on any assumptions about the population structure. The method is robust to cater for missing or erroneous data by allowing gaps in between marker patterns. Comparing our method with Haplotype Pattern Mining (*HPM*) which was reported by Tiovonen et. al. [16],

¹ A molecular marker is an identifiable physical location on the genomic region that either tags a gene or tags a piece of DNA closely associated with the gene. An allele is any one of a series of two or more alternate forms of the marker. From the data mining aspect, we could represent markers as attributes, and alleles as attribute values that each attribute could take on.

LinkageTracker outperforms *HPM* by reducing the average sum-square error by more than half (from 80.71 to 30.83) over one hundred trials.

Organization of this paper. In the next section, related work will be introduced, followed by a technical representation of the problem and a detailed description of the *LinkageTracker* algorithm. Next, the optimal number of gaps to set on *LinkageTracker* to achieve good accuracy will be discussed. We will then evaluate the performance of *LinkageTracker* with a recent work known as Haplotype Pattern Mining (*HPM*). Finally, we conclude our paper with a summary and the directions for future work.

2 Related Works

There are generally two methods used for detecting disease genes, namely, the direct and the indirect methods. Techniques used in the direct method include allele-specific oligonucleotide hybridization analysis, heteroduplex analysis, Southern blot analysis, multiplex polymerase chain reaction analysis, and direct sequencing. A detailed description of these techniques is beyond the scope of this paper but is available in [3] and [10]. Direct method requires that the gene responsible for the disease be identified and specific mutations within the gene characterized. As a result, direct method is frequently not feasible, and, the indirect method is used. The indirect methods such as [7], [14], and [16] involves the detection of marker alleles that are very close to or are within the disease gene, such that they are inherited together with the disease gene generation after generation. Such marker alleles are known as haplotypes. Alleles at these markers often display statistical dependency, a phenomenon known as linkage disequilibrium or allelic association [5]. The identification of linkage disequilibrium patterns allows us to infer the disease gene location. Most commonly, linkage disequilibrium mapping involves the comparison of marker allele frequencies between disease chromosomes and control chromosomes.

Kaplan et. al. [7] developed a maximum likelihood method for linkage disequilibrium mapping which estimates the likelihood for the recombination fraction between marker and disease loci by using a Poisson branching process. The likelihood of the haplotypes observed among a sample of disease chromosomes depends on their underlying genealogical relationships, the rates of recombination among markers, and the time since the mutation arose. Although likelihood methods have many desirable properties when used on data whose population ancestry is well understood, it is difficult to evaluate the likelihood when the data is arising from a huge number of possible ancestries.

DMLE+ proposed by Rannala & Reeve [14] uses Markov Chain Monte Carlo methods to allow Bayesian estimation of the posterior probability density of the position of a disease mutation relative to a set of markers. As similar to the maximum likelihood method, *DMLE+* has many good properties when applied to data whose population ancestry is well understood. However, *DMLE+* requires some prior information such as the fraction of the total population of present-day disease chromosome, growth rate of population and the age of the mutation, which may not be readily available. Furthermore, it is assumed that every sample sequence carries the disease mutation, although the authors claimed that this assumption can be relaxed, details on the extent that this assumption can be relaxed was not discussed.

Recently, Tiovonon et. al. [16] introduced a linkage disequilibrium mapping algorithm known as haplotype pattern mining (*HPM*). Firstly, *HPM* uses the association rule mining algorithm [1] to discover a set of highly associated patterns by setting the *Support* threshold to a certain value. Next, *HPM* uses chi-square test to discriminate disease association from control association. Finally, *HPM* computes the marker frequency for each of the markers. The frequency for each marker is computed by counting the number of associated patterns consisting of that specific marker. The marker with the largest frequency is predicted as closest to the disease gene. The main drawback of this algorithm is that it suffers from the rare item problem. As it uses association rule mining algorithm to discover highly associated patterns, and such patterns are relatively rare in the problem of linkage disequilibrium mapping. As a result the support threshold will need to be set at a very low value in order to discover those highly associated patterns.

Comparing *LinkageTracker* with the maximum likelihood method and *DMLE+*, the two methods require information about the population ancestry and assumes that the disease mutation occurs in most (or all) sample sequences, whereas *LinkageTracker* does not require any population ancestry information and allows for the disease mutation to occur in as low as 10% of the sample sequences. When compared to *HPM*, the *LinkageTracker* does not use *Support* in the assessment of marker patterns, instead *LinkageTracker* uses a statistical method known as *odds ratio* to detect discriminating patterns that are

highly associated within the patient data but not in the control data. Hence, the finding of candidate/potential linkage disequilibrium patterns and scoring their degree of associations are combined into a single step. Also as mentioned by Tamhane & Dunlop [15], chi-square test only indicate whether there exists statistically significant association, but it does not account for the magnitude of association. It is thus possible to have a significant chi-square statistics although the magnitude of association is small. The most common measure of the magnitude of association is the *odds ratio* method. *LinkageTracker* infers the marker closest to the disease gene by combining the *p*-values of association patterns consisting of that marker using a method recommended by Fisher [4], and not based on the marker frequency as in the *HPM* algorithm.

3 Technical Representation of *LinkageTracker*

The general framework of the *LinkageTracker* can be represented as a quintuple $\langle D, \Omega, L, \Psi, T \rangle$ where

- *D* is a dataset consisting of *M* vectors $\langle x_1, \dots, x_M \rangle$, where each x_i is a vector $\langle d_{i1}, \dots, d_{in} \rangle$ that describes the allele values of *n* genes/markers in a particular biological sample.
- For each position d_{sj} , $\omega_j = \{v_1, \dots, v_n\}$ denotes the set of all possible expression values that d_{sj} could take on, and Ω is a collection of $\{\omega_1, \dots, \omega_n\}$.
- A labelling for *D* is a vector $L = \langle l_1, \dots, l_M \rangle$, where the label l_i associated with x_i is either *abnormal* (a biological sequence derived from an individual exhibiting abnormality) or *normal* (a biological sequence belonging to a normal control).
- Ψ is the neighbourhood definition. The neighbourhood determines the maximum allowable gap size within each pattern. The gap setting is to enable *LinkageTracker* to be robust to noise. In a very noisy environment, larger gap size is required for better accuracy by extending the search space, at the expense of computational speed.
- $T \in \mathcal{R}^+$ is the threshold value for accepting a particular pattern. In statistical terms, *T* is the level of significance of the test. When the pattern score is less than *T*, the pattern is considered as significant, and will be kept for further processing.

The output *P* is a set of linkage disequilibrium patterns with high discriminative powers. A pattern $p = \langle d_{*i}, d_{*j}, \dots, d_{*k} \rangle$ where $p \in P$, such that $i < j < k$. Based on the set of patterns in *P*, we infer the marker allele that is closest to the disease gene. For each marker allele, we combine the *p*-values of all patterns in *P* that consist of that marker allele. The method to combine *p*-values was first introduced by Fisher [4], and will be described in detail in the next section.

4 *LinkageTracker* Algorithm

There are two main steps in the *LinkageTracker* algorithm. Step 1 identifies a set of linkage disequilibrium patterns which are strong in discriminating the abnormal from the normal, and step 2 infers the marker allele that is closest to the disease gene based on the linkage disequilibrium patterns derived in step 1.

4.1 Step 1: Discovery of Linkage Disequilibrium Pattern

LinkageTracker uses a statistical method known as odds ratio to score each potential/candidate pattern. If the *p*-value of a pattern is below the threshold *T*, then it is considered as having a significant discriminative power, and will be kept for further processing. Odds ratio provides a good measure of the magnitude of association between a pattern and the binary label *L*, which is crucial in determining the discriminative power and the allelic associations of a pattern. In this section, we will first of all describe the odds ratio method; follow by the details of level-wise neighbourhood searches for potential/candidate patterns and scoring them.

Odds Ratio. Given a pattern *x*, odds ratio computes the ratio of non-association between *x* and the label *L*, to the association between *x* and *L* based on a set of data. For example, given a pattern, say (1,3), we are

interested in finding out whether the marker pattern (1,3) is strongly associated with the label *abnormal*. Table 1 shows the contingency table for our example; odds ratio is defined as follows:

$$\text{Odds Ratio, } \theta = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}} \quad (1)$$

To test the significance of the magnitude of association, we compute the p -value of each pattern, and compare the p -value against T , if the p -value is less than or equals to T , the pattern is significant and we will use it for marker inference in the later stage. If the p -value is greater than T , the pattern is not significant, and will be discarded. The threshold T that we use has been adjusted using a method called Bonferroni Correction [11] in order to guarantee that the overall significance test is still at level T despite that we have made independent tests on each of the pattern.

Table 1 : 2x2 contingency table

	Abnormal	Normal
not(1,3)	π_{00}	π_{01}
(1,3)	π_{10}	π_{11}

LinkageTracker Algorithm. *LinkageTracker* mines patterns of the form $\langle d_{*i}, d_{*j}, \dots, d_{*k} \rangle$, for example, (3,5,6,*,*,4) is a marker pattern of length 4. The symbol “*” represents missing or erroneous marker allele, and will not be taken into consideration when testing for significance of the pattern. Also the symbol “*” will not be considered when computing the length of a marker pattern. Therefore, marker patterns (1,*,*,3), (1,*,3), and (1,3) are all considered as having length of 2.

A gap is a “*” symbol in between two known marker alleles. For instance, the marker patterns (1,*,*,*3) has three gaps, (1,*,3) has one gap, and (1,3) has no gaps. The maximum number of gaps for this marker pattern (1,*,*,3,*,*,*,*5) is four, as there are at most four gaps in between any two known marker alleles. The user is able to set the maximum number of gaps for the marker patterns. However, we recommend that a maximum allowable gap to be 6, giving the highest accuracy if the markers are spaced at 1 cM². The detail of such a recommendation is given in the later section.

To find linkage disequilibrium patterns using *odds ratio*, one of the way is to use the brute force method. That is, we could enumerate all possible marker patterns of length one, two, and three etc, and compute the *odds ratio* of each of the pattern and select those patterns that are significant. However, there are some practical difficulties to this approach: for n markers each with m alleles, there are $\binom{n}{k} m^k$ marker patterns of

length k , which we need to test for significance. Combinatorial explosion occurs as the length of marker patterns increases.

The enumeration of all possible marker patterns is in fact unnecessary. This is because, base on studies by Long & Langley [9], allelic associations are detectable within a genomic region of 20cM, allelic associations beyond 20cM are weak and are not easily detectable. Therefore, enumerating marker patterns whose marker alleles are more than 20cM apart is unlikely to yield significant results. Based on this observation, *LinkageTracker* uses a heuristic search method which allows the user to restrict its search space by controlling the maximum allowable gap size between two marker alleles. As described in section 3, the gap size setting Ψ helps to define the search space of *LinkageTracker* as well as to enable its robustness to noise. For simplicity of illustration, all examples in this paper assume that the markers are spaced at 1cM apart, furthermore, the markers in the simulated datasets (generated by Toivonen et. al. [16]) applied by *LinkageTracker* are all spaced at 1cM apart.

LinkageTracker is a heuristic level-wise search method which allows only significant marker patterns (or linkage disequilibrium patterns) of length $i-1$ at level i to join with their neighbors (of length 1) whose join would satisfy the maximum gap constraint Ψ to form candidate/potential marker patterns of length i , where $1 \leq i \leq n$ and n is the number of markers. We call the procedure of joining linkage disequilibrium patterns at each level to form longer patterns the *neighborhood join*. Note that in *neighborhood join*, only

² cM stands for centimorgan. It is the unit of measurement for genomic distance. In human genome, 1 centimorgan is approximately equivalent, to 1 million base pairs.

the marker patterns of length $i-1$ need to be significant, the neighbors that they join with need not be significant and may be several markers apart.

<i>Mx</i>	<i>Mh</i>	<i>Mg</i>	<i>Mf</i>	<i>Me</i>	<i>Md</i>	<i>Mc</i>	<i>Mb</i>	<i>Ma</i>	<i>Mz</i>
Level	Join Patterns								Significant Patterns
1									(<i>Mz</i>)
2	(<i>Mz, Ma</i>), (<i>Mz, Mb</i>), (<i>Mz, Mc</i>)								(<i>Mz, Mc</i>)
3	(<i>Mz, Mc, Md</i>), (<i>Mz, Mc, Me</i>), (<i>Mz, Mc, Mf</i>)								(<i>Mz, Mc, Mf</i>)
4	(<i>Mz, Mc, Mf, Mg</i>), (<i>Mz, Mc, Mf, Mb</i>), (<i>Mz, Mc, Mf, Mx</i>)								(<i>Mz, Mc, Mf, Mx</i>)

Fig. 1. Illustration of marker positions

A marker allele exhibits significant allelic association with the disease gene under two conditions. Firstly, it is significant on its own when tested (i.e. at level 1). Secondly, when combine with other marker alleles that exhibit allelic associations with the disease gene, it become significant when tested.

The former condition is trivial to detect, the latter condition is concerned with a marker allele who shows significant allelic association with the disease gene when combine with other significant marker alleles but is insignificant when assessed alone. Let us denote this maker allele as Mx . This problem can be further divided into 2 cases. The first case is that Mx is close to a neighbor Mi that is significant when tested alone. The term “close” here means that Mx will be selected to join with Mi directly to form marker patterns for the immediate next level. For example, two markers say Mx and My are both not significant at level 1, hence they will be discarded when forming marker patterns for level 2. Now, we have Mi which is an immediate neighbor of My showing significant allelic association in level 1 (assuming that the markers are ordered as follows: Mi , My and Mx). Hence, in level 2, Mi will be made to combine with its neighbors to form marker patterns of length 2. Since My is the immediate neighbor of Mi , My will be selected to form pattern with Mi . Although Mx is one marker away from Mi , Mx will also be selected, because *LinkageTracker* allows joining with markers that are some gaps away as described above. Hence, in level 2, both My and Mx are included in the marker patterns.

The second case is that Mx is very far from a marker allele Mz that is significant when tested alone. The term “far” here means that Mx is less than 20 markers away from Mz , but is far enough such that Mx will not be selected by Mz to form marker pattern for the immediate next level. For example, from Figure 1, Mx and Mz is 8 markers apart. Assuming that the maximum allowable gap size is set to 2, Mz is made to combine with Ma , Mb , and Mc to form patterns of length 2. Assuming that (Mz, Mc) is tested significant, then (Mz, Mc) will combine with Md , Me , and Mf to form patterns of length 3. Assuming that (Mz, Mc, Mf) is tested significant, then (Mz, Mc, Mf) will combine with Mg , Mh , and Mx to form patterns of length 4. Hence, Mx will ultimately be detected to form marker patterns under the condition that there are sufficient significant “intermediate” allele markers such as Mc and Mf , to facilitate the detection of allelic associative marker alleles that are much further away (i.e. Mx). Nevertheless, as in accordance with the studies by Long & Langley [9], most marker alleles exhibiting allelic associations with the disease gene will occur within a distance of 20cM from the disease gene, which means that marker alleles exhibiting allelic associations with the disease gene are quite densely packed within the 20 makers region. Hence, the chances of *LinkageTracker* detecting significant marker alleles within the range of 20 markers are relatively high even though *LinkageTracker* is a heuristic method.

4.2 Step 2: Marker Inference

As mentioned in the earlier section, we infer the marker closest to the disease gene by combining the p -values of the highly associated patterns. Now, let us describe how we could combine p -values from n patterns to form a single p -value. R.A. Fisher’s method [4] specifies that one should transform each p -value into $c = -2 * LN(P)$, where $LN(P)$ represents the natural logarithm of the p -value. The resulting n c -values are added together, and their sum, $\sum(c)$, represents a chi-square variable with $2n$ degree of freedom. For example, to find the marker closest to the disease gene, we compute the combine p -value and the frequency for each marker allele. In Figure 2a, Marker 2 has allele 4 occurring four times, its combined p -value is $1.4 * 10^{-6}$, which is the chi-square distribution of $\sum(c) = 9.4211 + 10.0719 + 11.6183 + 10.8074 = 41.9186$ with 8 degree of freedom. Figure 2b depicts the combined p -value for each of the marker alleles from Figure 2a. As we can see Marker 2 allele 4 has the lowest combined p -value, and hence we infer that Marker 2 is

closest to the disease gene. If more than one marker alleles have the same lowest p -value, then the marker with the highest frequency is selected as the marker closest to the disease gene.

Marker	1 2 3 4 5 6	P-Value	c = -2 * ln(P)
Pattern01	* 4 3 * * *	0.0090	9.4211
Pattern02	2 4 * * 6 1	0.0065	10.0719
Pattern03	2 4 3 5 * *	0.0030	11.6183
Pattern04	* * 3 5 * 1	0.0100	9.2103
Pattern05	2 4 * 5 6 *	0.0045	10.8074

(a)

	Freq	Σ(c)	Combine P-Value
Marker 1 allele 2	3	32.4975	1.3098E-05
Marker 2 allele 4	4	41.9186	1.4027E-06
Marker 3 allele 3	3	30.2497	3.5236E-05
Marker 4 allele 5	3	31.6390	1.9160E-05
Marker 5 allele 6	2	10.0719	0.0392
Marker 6 allele 1	2	19.2822	0.007

(b)

Fig. 2. a) Example of 5 linkage disequilibrium patterns. b) Combine p-value of each marker allele from (a).

5 Setting the Optimal Number of Gaps

To accurately find the marker closest to the disease gene, it is important to determine the optimal number of gaps to use. The marker alleles that show significant allelic associations with the disease gene (within 20 markers region) should minimize the number of joins with neighbors beyond the 20 markers region. This is because the joining of a significant marker allele with some neighbors that are beyond the 20 markers region will inevitably introduce some false positive marker patterns or noise. Such false positive marker patterns will result in the reduction in accuracy during marker inference. On the other hand, we want to be as robust as possible, that is, to maximize the total possible gaps so as to cater for erroneous marker alleles. Based on these two conditions, we compute the *Score* for each gap setting g as follows for patterns of length 2:

$$Score(g) = \frac{\sum_{i=0}^g Robustness_i}{\sum_{i=0}^g Noise_i} \quad (2)$$

Figure 3 shows the *Score* values for gap settings between 0 to 20. Different gap settings will result in different values for *Noise* and *Robustness*. We shall now illustrate how the values of *Noise* and *Robustness* were computed with examples.

Noise. *Noise* is defined as the maximum possible number of patterns consisting of markers beyond the 20 markers region. Figure 4 shows a disease gene that is very close to marker $M1$, markers $M21$ and $M22$ are in dotted boxes as they are beyond the 20 markers region from the disease gene. Assuming that marker $M2$ shows significant association with the disease gene, and we set the maximum allowable gaps to 1, then $M2$ can join with its neighbors $M3$ and $M4$ to form patterns of length 2, i.e. $(M2, M3)$ and $(M2, M4)$. Recall that the joining of a significant marker with some neighbors that are beyond the 20 markers region will introduce *Noise*. In this case, if markers $M19$ and $M20$ are significant, they will join with $M21$ and $M22$ to form patterns of length 2. We can see from Figure 4 that $M19$ and $M20$ will join with $M21$ and $M22$ in three ways, as illustrated by the dotted arrows. Hence, the maximum possible number of patterns consisting of

markers beyond the 20 markers region (i.e. $\sum_{i=0}^1 Noise_i$) is 3 when the gap setting is 1. The *Noise* values for gap settings from 2 to 20 were computed similarly.

Num. of Gaps (g)	Noise	Num. Of patterns p form with g gaps	Robustness = p × g	Score(g)
0	1	19	0	0
1	2	18	18	6
2	3	17	34	8.67
3	4	16	48	10
4	5	15	60	10.67
5	6	14	70	10.95
6	7	13	78	11
7	8	12	84	10.89
8	9	11	88	10.67
9	10	10	90	10.36
10	11	9	90	10
11	12	8	88	9.59
12	13	7	84	9.14
13	14	6	78	8.67
14	15	5	70	8.17
15	16	4	60	7.65
16	17	3	48	7.11
17	18	2	34	6.56
18	19	1	18	6
19	20	0	0	0
20	21	0	0	0

Fig. 3. Score values for 0 to 20 gaps

Robustness. Before computing the *Robustness* values, we need to compute the maximum possible number of patterns p formed within the 20 markers region when the gap setting is g . When the gap setting g is set to 1, we can have at most 18 patterns (i.e. $p = 18$) as illustrated by the arrows in Figure 5. With the values of p for different values of g , we define *Robustness* as the maximum number of patterns formed within the 20 markers region weighted by the gap setting g itself:

$$Robustness = p \times g. \quad (3)$$

Recall that it is desirable to have wider gaps so as to cater for erroneous marker alleles, hence the value of *Robustness* increases as the value of g increases. As we can see from Figure 3 that the gap setting of 6 has the highest *Score* value, hence we recommend that for a dataset with more than 20 markers to each chromosome (i.e. more than 20 attributes to each record) and each marker is spaced at 1cM apart, the optimal allowable gap setting should be 6.



Fig. 4. The darkened circle indicates the disease gene location

To verify our above recommendation, we evaluated the performance of *LinkageTracker* by varying the gap settings from 2 to 10 on 100 realistically simulated datasets generated by Tiovonon et. al. [16] (details in the next section). The sum-square errors were computed for different gap settings g when applied to the 100 datasets. We found that the gap setting of 6 has the lowest sum-square error, which means that it has the highest accuracy. This is in compliance with our above recommendation.

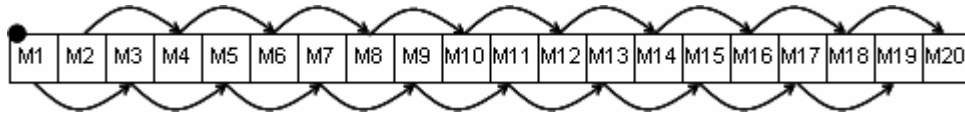


Fig. 5. Joining of markers when gap setting g is 1

6. Evaluation

6.1 Generated datasets

The datasets used in our experiments are generated by Tiovonon et. al. [16] and are downloadable from the following URL: <http://www.genome.helsinki.fi/eng/research/projects/DM/index-ajhg.html>. The simulated datasets correspond with the realistic isolated founder populations which grow from 300 to about 100,000 individuals over a period of 500 years. The simulation of isolated population is suited to linkage disequilibrium studies as recommended by Wright et. al. [17].

There are altogether 100 datasets each consists of 400 biological sequences where 200 sequences were labeled “*abnormal*” and 200 labeled “*normal*”, each biological sequence consists of 101 markers. The datasets were generated such that each dataset has a different disease gene location, and our main task is to predict the marker (or attribute) that is nearest to the disease gene for each dataset.

6.2 Comparison of Performance on Generated Datasets

Figure 6 shows the performance of *HPM* (proposed by Tiovonon et. al. [16]) and *LinkageTracker* when applied to the generated datasets. Each point on the graph depicts the predicted disease gene location by *HPM* if marked “ \diamond ” and the predicted disease gene location by *LinkageTracker* if marked “+”, for the 100 dataset. The straight line depicts that the predicted location is the same as the actual location, the closer the “ \diamond ” or “+” marks to the straight line the more accurate is the prediction. As we can see that the accuracy of *LinkageTracker* is reasonably good with only one significant outlier, whereas *HPM* has two significant outliers. The same outlier was encountered by *LinkageTracker* when tested on different gap settings, which means that there may exist some errors in this dataset such that a “pseudo region” occurs that differentiate itself from the normal population that is much more significant than the true region with the disease gene. The average sum-square error for *HPM* is 80.71, and the average sum square error for *LinkageTracker* is 30.83. Hence, *LinkageTracker* outperforms *HPM* in general with lower sum-square error. Even after we remove the common outlier between *LinkageTracker* and *HPM*, *LinkageTracker* continues to outperform *HPM* with an average sum-square error of 6.40, as compared to *HPM* with an average sum-square error of 15.47.

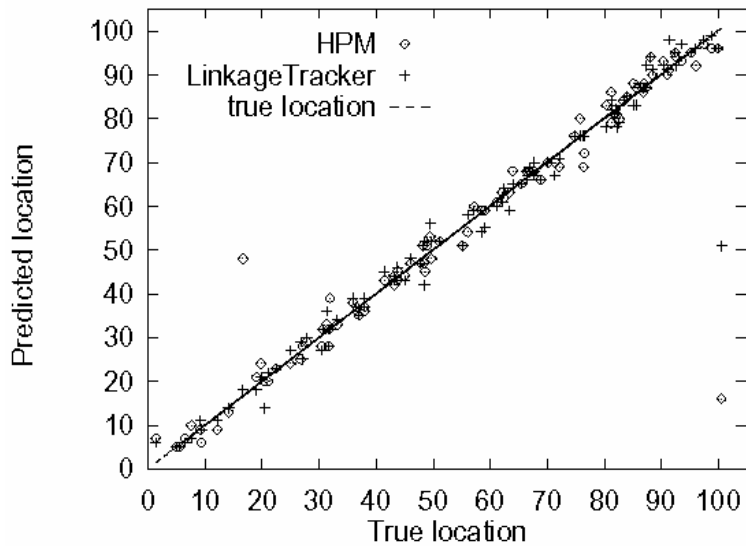


Fig. 6. Comparison of prediction accuracies between *HPM* and *LinkageTracker*

6.3 Performance on Real Dataset

We applied our algorithm on a real dataset, consisting of patients affected by hemophilia from Singapore³, and a set of matching unaffected individuals. Hemophilia A is an X-linked recessive bleeding disorder that results from deficiency and/or abnormality of coagulation factor VIII (FVIII) [2]. The FVIII gene spans 186 kb of DNA and resides on 0.1% of the X chromosome (band Xq28).

A set of markers located on chromosome Xq28 which tags the hemophilia A disease gene were collected and analyzed from 47 patients and 47 matched normal controls. The *LinkageTracker* detected Bcl I RFLP marker as the closest to the disease susceptible gene. Our prediction results showing Bcl I association was found and confirmed through elaborate biological experiments, as Bcl I is an intragenic SNP (single nucleotide polymorphism) in intron 18 of FVII gene and is linked to hemophilia A disease phenotype [8]. *LinkageTracker* is able to guide or narrow the investigation in identifying the polymorphic markers that tag the disease genes.

7. Conclusions and Future Work

We have introduced a new method of inferring the location of disease genes based on observed associations known as *LinkageTracker*. *LinkageTracker* has shown to be highly accurate in both simulation-generated and real genetic datasets. We have also recommended the optimal number of gaps to set on *LinkageTracker* to achieve good accuracy. Comparing with the maximum likelihood method and *DMLE+*, the two methods require information about the population ancestry and assume that the disease mutation occurs in most or all sample sequences, whereas *LinkageTracker* does not require any population ancestry information and allows for the disease mutation to occur in as low as 10% of the sample sequences. Comparing the performance of *LinkageTracker* with a recent work known as *HPM*, *LinkageTracker* outperforms *HPM* with lower average sum-square error. Even after we remove the common outlier, the sum-square error of *LinkageTracker* remains significantly lower than the average sum-square error of *HPM*. In the future, we plan to extend this work to identify boundaries in which all the significant patterns can be bounded and ultimately guarantees that all significant patterns can be found.

³ Data is obtained from Department of Pediatrics, National University Hospital, National University of Singapore.

Acknowledgements

This research is partially supported by a Research Grant No. R-252-000-111-112/303 from the Agency for Science, Technology, and Research (A*Star) and the Ministry of Education in Singapore.

References

- [1] R. Agrawal, and R. Srikant. Fast algorithm for mining association rules. In *Proceedings of the Very Large Data Bases (VLDB) Conference*, 1994.
- [2] S. Antonarakis, H. Kazazian, E. Tuddenham. Molecular etiology of factor VIII deficiency in hemophilia A. *Human Mutation*, 5:1-22,1995.
- [3] A. Beaudet, C. Scriver, W. Sly, D.Valle. Genetics, biochemistry, and molecular basis of variant human phenotypes. In: Scriver CR, Beaudet AL, Sly WS, et al, eds. *The Metabolic and Molecular Basis of Inherited Disease*. 7th ed. New York, NY: McGraw-Hill, Inc; 2351-2369, 1995.
- [4] R. Fisher. *Statistical methods for research workers*, 14th edition. Hafner/MacMillan, New York, 1970.
- [5] D. Goldstein, and M. Weale. Population genomics: Linkage disequilibrium holds the key. *Current Biology*, 11:R576-R579, 2001.
- [6] J. Hastbacka, A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver, and E. Lander. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics*, 2:204-211, 1992.
- [7] N. Kaplan, W. Hill, and B. Weir. Likelihood methods for locating disease genes in non-equilibrium populations. *American Journal of Human Genetics*, 56:18-32, 1995.
- [8] S. Kogan, M. Doherty, J. Gitschier. An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. Application to hemophilia A. *New England Journal of Medicine*, 317: 985-990, 1987.
- [9] A. Long, and C. Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research*, 9: 720-731, 1999.
- [10] S. Malcolm. Molecular methodology. In: Rimoin DL, Connor JM, Pyeritz RE, eds. *Emery and Rimoin's Principles and Practice of Medical Genetics*. 3rd ed. New York, NY: Churchill Livingstone; 67-86, 1997.
- [11] R. Miller. *Simultaneous statistical inference* . 2nd edition. Springer Verlag, 1981.
- [12] National Cancer Institute. Cancer Facts. http://cis.nci.nih.gov/fact/3_62.htm. Date reviewed: 02/06/2002.
- [13] L. Ozelius, P. Kramer, D. de Leon, N. Risch, S. Bressman, D. Schuback et. al. Strong allelic association between the torsion dystonia gene (DYT1) and loci on chromosome 9q34 in Ashkenazi Jews. *American Journal Human Genetics* 50: 619-628, 1992.
- [14] B. Rannala and J. Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *American Journal of Human Genetics* 69:159-178, 2001.
- [15] A. Tamhane, and D. Dunlop. *Statistics and data analysis: from elementary to intermediate*. Prentice Hall, 2000.
- [16] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *American Journal of Human Genetics*, 67:133-145, 2000.
- [17] A. Wright, A. Carothers, and M. Pirastu. Population choice in mapping genes for complex diseases. *Nature Genetics*, 23:397-404, 1999.