American Journal of Bioinformatics 1 (1): 20-29, 2012
ISSN 1948-9862

Bayesian Determination of Disease Associated Differences in Haplotype Blocks

[1]Ivan Kozyryev and [2]Jing Zhang
[1]Department of Physics,
Faculty of Arts and Sciences, Harvard University, Cambridge, MA, USA
[2]Department of Statistics,
Faculty of Arts and Sciences, Yale University, New Haven, CT, USA

Abstract: Problem statement: While experimental ascertainment of haplotype blocks in the genome-scale case-control studies is expensive, accurate computational phasing is still a daunting task for bioinformatics approaches. We used a statistical method to determine differences, potentially associated with a certain disease, in linkage disequilibrium block boundaries in whole-genome Single Nucleotide Polymorphisms (SNPs) data. **Approach:** We utilized a Bayesian model for calculating the posterior probabilities of the block boundaries in the SNPs data and used Metropolis-Hastings algorithm to sample from that posterior distribution. Our method was applied to search for haplotype-block boundary differences associated with two autoimmune diseases: Type I Diabetes (T1D) and Rheumatoid Arthritis (RA). **Results:** We located the regions on chromosome 6 with significant control-case difference in haplotype blocks around the SNPs and genes that were previously known to be associated with T1D and RA (in the HLA complex), as well as around genes whose association with the autoimmune diseases should be further explored in future studies. **Conclusion/Recommendations:** The statistical approach explored in this study provides an efficient and accurate way to study connection of haplotype-block differences to multiple important diseases.

Key words: Linkage disequilibrium, autoimmune diseases, HLA complex, type 1 diabetes, rheumatoid arthritis

INTRODUCTION

Statistical approaches in genetics: Even though it has been a decade since researchers first sequenced the human genome, obvious links between the genes and specific diseases have been much slower to appear than originally expected by everybody. Because original approach based on simple correlation analysis was not successful, now many researchers believe that new advances in genomics will come from a rich statistical understanding of complex interactions of our genetic code (Bansal *et al*., 2010; Zhang and Liu, 2007; Svoboda, 2010). However, it is necessary to perform statistical analysis on a vast amount of data consisting out of the sequences including millions of genomes in order to completely understand how our genetic code interacts with the environment to make us the way we are (Svoboda, 2010).

Whole-genome Single Nucleotide Polymorphisms (SNPs) data from individuals in the case-control studies has potential to help us understand complex interactions among multiple genes (Zhang *et al*., 2011a). SNPs can be thought of as "tiny typos" in the genome with one base being replaced by another. While some SNPs directly contribute to the disease, others can be linked to the genes that do (Carmichael, 2010). However, complications arise when we try to analyze such data, due to the fact that the number of possible interaction combinations among genotype markers is humongous for a large size genetic association study and we are interested in finding very few disease-related interactions (Zhang and Liu, 2007). Additionally, some nearby SNPs are highly correlated due to linkage disequilibrium (Zhang *et al*., 2011b), which further complicates statistical analysis aiming at determining disease related interactions.

Importance of linkage disequilibrium: Linkage Disequilibrium (LD) describes the phenomenon when the genotypes at nearby markers are highly correlated (Zhang *et al*., 2011a). This correlation arises due to shared ancestry of contemporary chromosomes (The International HapMap Consortium, 2005). LD patterns have many important applications in biology and genetics. These patterns can be used for inferring the

Corresponding Author: Jing Zhang, Department of Statistics, Faculty of Arts and Sciences, Yale University, New Haven, CT, USA

distribution of cross-over events at short scales which are hard to study experimentally, studying gene conversion, about which there is only a very limited amount of experimental data, understanding the evolutionary history of humans and detecting natural selection (Wall and Pritchard, 2003).

Patterns of linkage disequilibrium are unpredictable and very noisy (Wall and Pritchard, 2003). Additionally, extent of LD defers from one genomic region to another. Population history, fine-scale heterogeneity in recombination rates and population genetic models all contribute to noisy appearance of spatial structure of LD. However, this complex reality is described by a simple model known as haplotype-block model (Wall and Pritchard, 2003). According to this model, genotype data is divided into discrete blocks with highly correlated SNPs within each block and adjacent blocks are separated by recombination events (Zhang *et al*., 2011b; 2004; Ding *et al*., 2005a; 2005b). Even though described model is simple, experimental data confirms its validity (Wall and Pritchard, 2003; The International HapMap Consortium, 2005).

Background on type 1 diabetes and rheumatoid arthritis: Type 1 Diabetes (T1D) affects 0.5% of the world population and 1.4 million US people (Bottini *et al*., 2004). It is a chronic disease that occurs when not enough insulin is produces to control the sugar levels in blood. It can occur at any age but most frequently it is diagnosed in children and young adults (Devendra *et al*., 2004). Even though the exact cause of the disease is unknown, most researchers suspect that there is some sort of trigger (environmental or viral) that causes an immune reaction in genetically susceptible group of people. As of right now, no cure is available and the outcome for people with this type of diabetes varies (Devendra *et al*., 2004). Previous studies have validated the hypothesis that common genetic variants play important role in the disease formation (Bottini *et al*., 2004; Polychronakos and Li, 2011; Zhang *et al*., 2011a).

Like T1D, Rheumatoid Arthritis (RA) has a significant genetic contribution to its development but the detailed heritability is still not known (Newton *et al*., 2004). RA is a chronic disease that is accompanied by severe pain that arises from destruction of the synovial joints (WTCCC, 2007). Even though recent advances in the Genome-Wide Association Studies (GWAS) are starting to show the directions for new therapies and advancement of understanding the complex relationships between different autoimmune disorders and their genetic causes, complete descriptions of RA and T1D genetic landscapes are still far from being understood (Coenen and Gregersen, 2009; WTCCC, 2007). We chose to focus our study on RA and T1D at the same time because they are already known to share common loci and are both autoimmune diseases (WTCCC, 2007).

Goals of the project: The goal of this project was using as a starting point a highly successful approach to classification problems with discrete covariates (Zhang *et al*., 2011b) specifically used for determining block-based epistasis associations, to develop a statistical model to search for disease associated differences in LD-block structure between control and case groups. We determined haplotype-block structure for controls and cases independently and used this information to find regions with genetic variants that could potentially be associated with the disease. Our methods were applied to the actual large data sets consisting of whole-genome single nucleotide polymorphisms data from chromosome 6 (chr6) for T1D and RA patients and control groups.

MATERIALS AND METHODS

Our problem of using haplotype-block structure for whole-genome chr6 SNPs data from patients and controls to determine regions of genetic variants that increase susceptibility to the disease can be divided into two smaller sub-problems. First, since there is no complete experimental data on spatial structure of LD for our regions of interest, we determined LD blocks for controls and cases using computational statistical methods and then extracted the disease relevant information hidden in the background noise. Each of these steps used different modeling and data analysis techniques that are described in detail below.

Statistical model for determining block boundaries due to linkage disequilibrium: The observations consist of genotypes of L SNP markers observed on N patients, combined into final data set D. Each marker can have one of three values. The goal is to partition L markers into blocks in such a way that the correlations between SNPs in different blocks are close to zero, yet the number of observed genotype combinations of SNPs within each block is small. Below we describe a method that uses a Bayesian model and a Monte Carlo algorithm to determine block structure for the given SNPs data set.

General overview: Let B be the block partition variable which has a form of L binary indicators where value of one (1) corresponds to the start of the next

block. Using multiplication law for conditional probability, we can calculate combined probability of the Data (D) and Block structure (B) to be Eq. 1:

$$P(D,B) = P(B \mid D)P(D). \quad (1)$$

Therefore, we find the posterior probability of the block boundaries given by Eq. 2:

$$P(B \mid D) = \frac{P(D,B)}{P(D)} \quad (2)$$

Let's consider numerator and denominator of the above equation separately. Using multiplication law ones more, we can express the numerator as Eq. 3:

$$P(D,B) = P(D \mid B)P(B) \quad (3)$$

Observe that once we impose a prior P(B) on the block boundary distribution, we can calculate P(D,B) explicitly. However, let's now move our attention to the denominator of the expression for the posterior of B. Observe that Eq. 4:

$$P(D) = \sum_{i=1}^{2^{L-1}} P(D \mid B_i)P(B_i) \quad (4)$$

where, we sum over all possible block boundary configurations. For our data we have L ≈ 30, 000. Therefore, calculating P(D) is not computationally feasible and we have to resort to MCMC methods. Using Monte Carlo methods we sample from P(B|D) to obtain probability for each B(i) to be 1 (the start of the next block).

Details of the model: For a prior distribution on B, we assume that indicators are independent and identically distributed Bernoulli random variables. Therefore, the total probability of observing a particular B is given by Eq. 5:

$$P(B) = p^{|B|}(1-p)^{L-|B|}. \quad (5)$$

Another important assumption in our model is that genotype combinations of SNPs (known as "diplotypes") in different blocks are mutually independent, which is a good approximation to reality in biological data (Zhang *et al*., 2011a).

Consider a block of SNPs (s,…,b-1) with the starting marker s and block end marker b-1. Since each SNP can take one of three possible values, there are 3^{b-s} possible diplotypes in the block. Let $\vec{p} = (p_1, p_2, ..., p_{3^{b-s}})$ be a vector with probabilities of a particular diplotype. We model the diplotype counts by the multinomial distribution with its frequency parameters following a Dirichlet prior distribution: P~ Dirichlet(a_1,a_2,…,a_3^{b-s}), where Dirichlet density function is given by (let $K = 3^{b-s}$) Eq. 6:

$$f(\vec{p},\vec{a}) = \frac{1}{B(\vec{a})}\prod_{i=1}^{K} p_i^{a_i-1} \quad (6)$$

and the normalizing constant $B(\vec{a})$ is specified by $\frac{\prod_{i=1}^{K}\Gamma(a_i)}{\Gamma(\sum_{i=1}^{K}a_i)}$ where Γ is the gamma function and $\vec{a}$ is the vector of Dirichlet parameters. Let n_i denote the number of counts of a specific diplotype i out of K in the block under consideration. The joint probability of $D_{[s,b)}$, which denotes the subset of full data for a particular block under consideration and $\vec{p}$ is given by Eq. 7:

$$\begin{aligned} P(D_{[s,b)},\vec{p}) &= P(D_{[s,b)} \mid \vec{p})P(\vec{p}) \\ &= P(D_{[s,b)} \mid \vec{p})\text{Dirichlet}(\vec{a}) \end{aligned} \quad (7)$$

Since we imposed a Dirichlet prior on diplotype counts. Therefore, substituting for the Dirichlet density, we have Eq. 8:

$$P(D_{[s,b)},\vec{p}) = \frac{1}{B(\vec{a})}\prod_{i=1}^{K} p_i^{n_i+a_i-1} \quad (8)$$

However, we are not interested in $\vec{p}$. Thus, in order to find the marginal probability of data for the block, we integrate out $\vec{p}$ Eq. 9:

$$P(D_{[s,b)}) = \int P(D_{[s,b)},\vec{p})d\vec{p} \quad (9)$$

Then, the marginal probability of the data for the block is given as Eq. 10:

$$\begin{aligned} &P(D_{[s,b)} \mid [s,b) \text{ is one block}) = \\ &\left(\prod_{i=1}^{3^{b-s}} \frac{\Gamma(n_i+a_i)}{\Gamma(a_i)}\right)\frac{\Gamma(\sum a_i)}{\Gamma(\sum(n_i+a_i))} \end{aligned} \quad (10)$$

where $a_i = \frac{\rho}{3^{b-s}}$ denotes the Dirichlet parameters for the distribution on the frequency parameters for diplotype counts. Since according to one of our assumptions, the diplotype counts in different blocks are independent, we find the total probability of data to be Eq. 11:

$$P(D \mid B) = \prod_{\text{all blocks}} P(D_{[s,b)} \mid [s,b) \text{ is one block}) \quad (11)$$

Using Eq. 2 and 3, we find that the posterior of the block partition variable is proportional to Eq. 12:

$$P(B \mid D) \propto P(D \mid B)P(B) \quad (12)$$

However, as was shown above, because it is computationally unfeasible to sum over all possible block boundaries to determine the proportionality constant we need to use Monte Carlo method to sample from the posterior. Particularly, we use Metropolis Hastings (MH) algorithm; see (Wai-Yuan and Tan, 2002; Liu, 2001) for details. We sample from P (B | D) in order to obtain P (B (i) = 1) for each i = 1, 2,…,L using MH algorithm in the following way:

- Initialization step: initialize B_0 by randomly assigning values to the block partition variable
- Proposal step: given the current block partition state B generate the proposal block partition B'of the data by randomly choosing a block and performing one of the following moves: (1). split the block into two at the chosen index position; (2). merge two sequential blocks together at the chosen index; or (3). shift the block boundary
- Evaluation step: let q(B→B') represent the probability of changing from B to B', then the acceptance probability is given by Eq. 13:

$$r = \min\{1, \frac{P(B' \mid D)q(B' \to B)}{P(B \mid D)q(B \to B')}\} \quad (13)$$

- Movement step: generate u~ Uniform (0, 1); accept B' if u ≤ r and keep B otherwise
- Stop if the number of iterations of the algorithm is ≥ N and go to step 2 otherwise

After the iterations finished, we calculate for each marker i = 1, 2,..,L the probability of the block start at that position P (B (i) = 1). For this calculation we use only the portion of the samples after the chain converged to the target distribution.

Additional methods of data analysis used: The output of the procedure described above is the probability for each marker to be the beginning of the next block: P (B (i) = 1), for each i = 1, 2,..,L. Since the MH algorithm was performed on both controls and cases independently, we obtained P_{cases} (B (i) = 1) and P_{contr} (B (i) = 1) for each marker i. In order to find the regions of genetic variants associated with the case status, for each marker i we considered P_{cases} (B (i) = 1)-P_{contr} (B (i) = 1), which is the difference in the posterior probability of block start and the absolute value of that quantity. We employed various data analysis methods described below to find the case status associated regions of genetic variants.

RESULTS

General form of the data was already mentioned previously. Specifically, data used included genotype data for 2-allelic SNP markers giving for each marker probabilities (p_1, p_2, p_3) to observe a particular combination of two alleles (ignoring the order). Data we used consisted of one data set for 2000 patients with Type I Diabetes (T1D data set), one data set for 1999 patients with rheumatoid arthritis (RA data set) and two data sets with control group data for a total of 3004 individuals (control data set). First, we filtered out SNPs with number of patients with max (p_1, p_2, p_3)<0.9 exceeding 3%. After we filtered out such SNPs simultaneously for both controls and T1D patients 29,483 good SNPs were left for final analysis from original 31,470. Similarly, simultaneous filtering of RA and controls data resulted in 29,468 good SNPs for farther analysis.

We already outlined the general approach to analyze the output results of the MH algorithm used to sample from the posterior P(B|D). We considered differences in posterior probabilities for haplotype-block boundary start between controls and cases at specific SNPs locations. In order to determine SNP markers and genes on chr6 associated with T1D and RA, we analyzed the distribution of the markers for which difference in probability of the block start between controls and cases was larger than 0.5, implying that the specific marker was determined with high probability to signify the beginning of the next haplotype block only for either controls or cases (but not both). Figure 1 and 2 show the histograms of the positions of the markers with such high posterior probability difference for T1D-controls and RA-controls data sets, correspondingly. The data is plotted for various bin widths, starting with 5kb and ending with 5Mb bins. Observe that in plots A-B for Fig. 1 and 2, the spread of the determined disease related haplotype-block differences across the short arm of the chromosome 6 is very noisy and roughly uniform over the whole region for small bin widths (5 and 50kb) for RA and T1D alike. Similarly, looking at the Fig. 1D and 2D we conclude that for bin width equal to 5Mb the pattern of the LD differences is smoothed out, because we are effectively averaging over many marker locations.

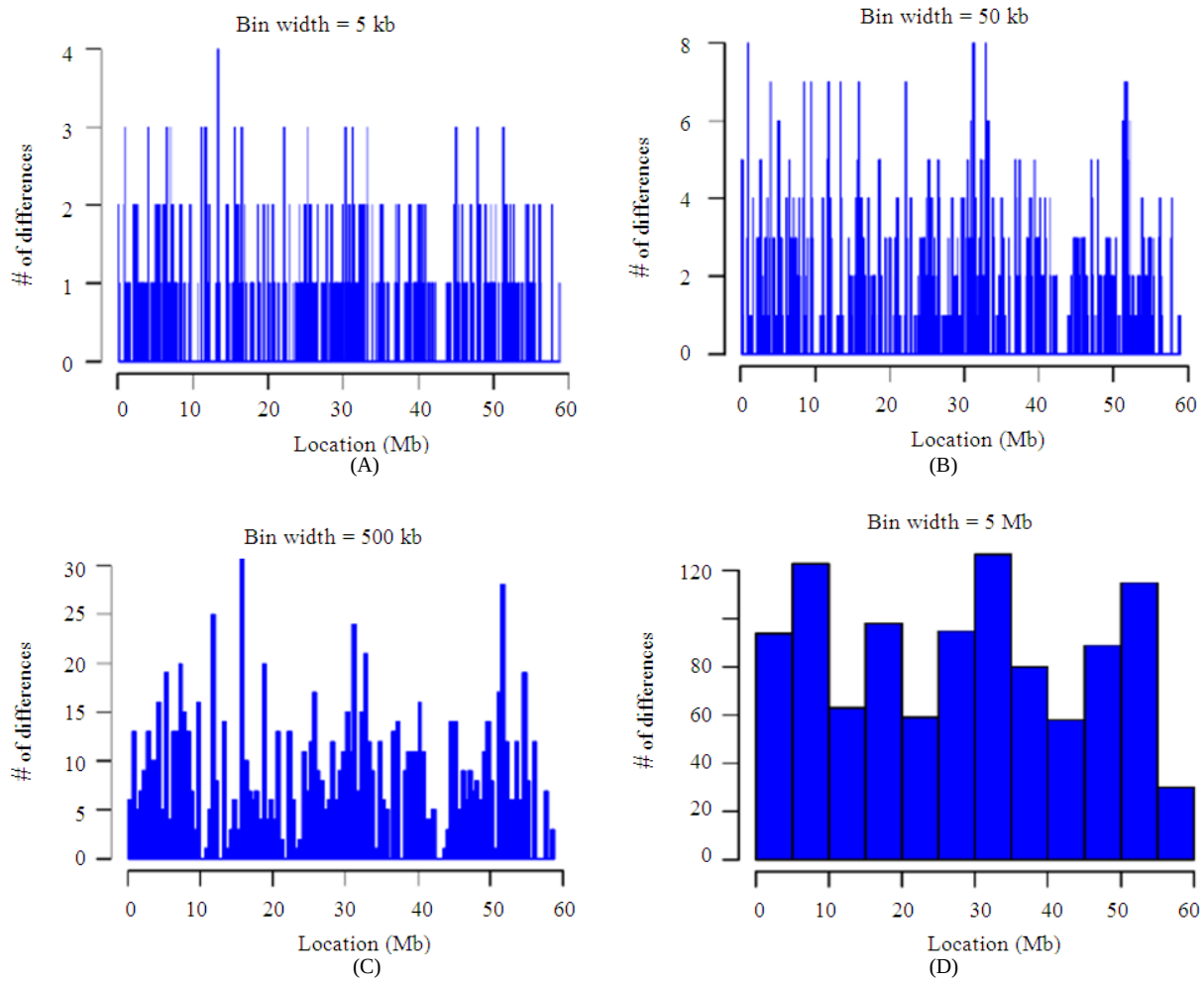

Fig. 1: T1D associated LD-block boundary differences across the short arm of the chromosome 6. Histogram plots of the type I diabetes associated haplotype-block differences on chromosome 6 (position <60Mb) for various bin widths: 5kb (A), 50kb (B), 500kb (C) and 5Mb (D). The block boundaries are considered distinct for controls and T1D patients at a specific SNP marker location if the difference in the determined posterior probability of block start at that positions is larger than 0.5.

Table 1: Regions with the largest number of block differences between T1D and control groups. This table summarizes locations of regions (bin width = 500 kb) on the short part of chromosome 6 (position < 60 Mb) with the number of determined linkage disequilibrium block boundary differences between type 1 diabetes patients and controls larger than or equal to 20 (top 7 regions out of 118 total). We also note whether the regions have been previously identified to be connected with the type 1 diabetes

Location (Mb)	# of diff.	Known T1D loci[a]	RefSeq genes[b]
7.0-7.5	20	None	CAGE1, DSP, RIOK1, RREB1, SSR1
11.5-12.0	25	None	TMEM170B, ADTRP
15.5-16.0	31	None	DTNBP1, JARID2
18.5-19.0	20	None	MIR548A1, RNF144B
31.0-31.5	24	MHC; HLA-B	Multiple[c]
32.5-33.0	21	MHC; HLA-DRB1; BAT1	Multiple[d]
51.5-52.0	28	PKHD1; rs9296661	PKHD1

[a]: Either single or two-SNP strong association with T1D previously determined in (WTCCC, 2007; Zhang *et al*., 2011b) [b]: RefSeq genes from the UCSC genomic database (genome.ucsc.edu) [c]: Complete list of the known RefSeq genes in the region in chromosomal order: VARS2, SFTA2, DPCR1, MUC21, MUC22, HCG22, C6orf15, CDSN, PSORS1C1, PSORS1C2, CCHCR1, TCF19, POU5F1, PSORS1C3, HCG27, HLA-C, HLA-B, MICA. [d]: Complete list of the known RefSeq genes in the region in chromosomal order: HLA-DRA, HLA-DRB5, HLA-DRB6, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DQA2, HLA-DQB2, HLA-DOB, TAP2, PSMB8, LOC100507463, TAP1, PSMB9, LOC100294145

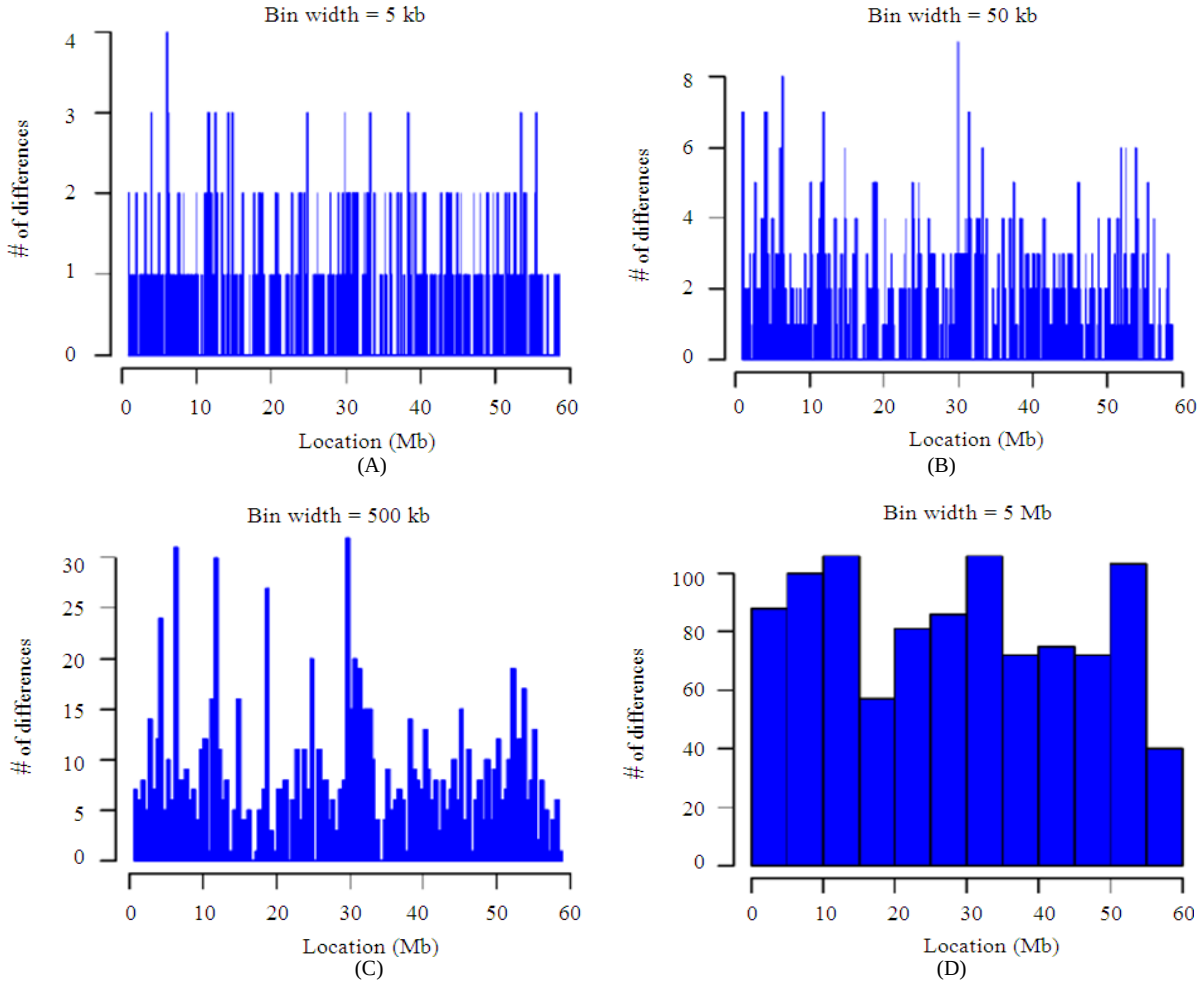

Fig. 2: RA associated LD-block boundary differences across the short arm of the chromosome 6. Histogram plots of the rheumatoid arthritis associated haplotype-block differences on chromosome 6 (position < 60Mb) for various bin widths: 5kb (A), 50kb (B), 500kb (C) and 5Mb (D). The block boundaries are considered different for controls and RA patients at a specific SNP marker location if the difference in the determined posterior probability of block start at that positions is larger than 0.5

Table 2: Regions with the largest number of block differences between RA and control groups. This table summarizes locations of regions (bin width = 500 kb) on the short part of chromosome 6 (position < 60 Mb) with the number of determined linkage disequilibrium block boundary differences between rheumatoid arthritis patients and controls larger than or equal to 20 (top 7 regions out of 117 total). We also note whether the regions have been previously identified to be connected with the rheumatoid arthritis

Location (Mb)	# of diff.	Known RA loci[a]	RefSeq genes[b]
4.0-4.5	24	None	C6orf146, C6orf201, ECI2, PRPF4B
6.0-6.5	31	None	F13A1, LY86-AS1
11.5-12.0	30	None	TMEM170B, ADTRP
18.5-19.0	27	None	MIR548A1, RNF144B
24.5-25.0	20	None	Multiple[c]
29.5-30.0	32	MHC; rs1233400	Multiple[d]
30.5-31.0	20	MHC; rs1075496	Multiple[e]

[a]: Either strong or moderate single SNP association with RA previously determined in (WTCCC, 2007) [b]: RefSeq genes from the UCSC genomic database (genome.ucsc.edu) [c]: Complete list of the known RefSeq genes in the region: ACOT13, ALDH5A1, C6orf62, FAM65B, GMNN, GPLD1, KIAA0319, MRS2, TDP2. [d]: Complete list of the known RefSeq genes in the region: GABBR1, HCG4, HLA-F, HLA-F-AS1, HLA-G, HLA-H, IFITM4P, LOC100507362, LOC554223, MAS1L, MOG, OR10C1, OR11A1, OR2H1, OR2H2, SNORD32B, UBD, ZFP57. [e]: Complete list of the known RefSeq genes in the region: ABCF1, ATAT1, C6orf136, DDR1, DHX16, FLOT1, GNL1, GTF2H4, HLA-E, IER3, MDC1, MIR4640, MIR877, MRPS18B, NRM, PPP1R10, PPP1R18, PRR3, TUBB, VARS2

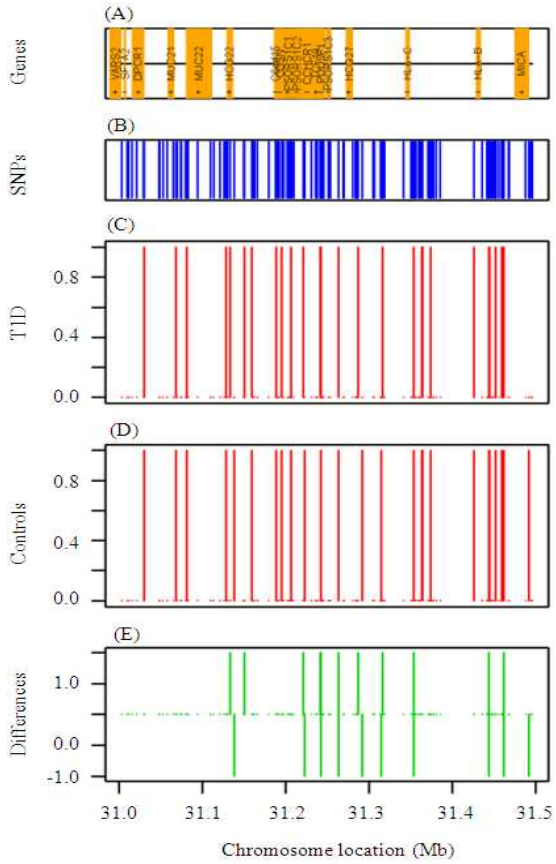

Fig. 3: Positions of T1D associated block differences on the specific genes in the 31.0-31.5 Mb region. Plots of the known RefSeq genes in the part of the MHC region (A), used SNPs locations (B), posterior probabilities of the block boundary start for T1D patients (C) and controls (D), as well as the difference in the posterior probabilities between the case and control data sets (E). For plots A-E the common x-axis (location) is shown at the bottom of plot E. In plot A the strand is marked with "+" or "-" below the gene name. Positions of the RefSeq genes were obtained from the UCSC genomic data base (genome.ucsc.edu). Table 1 for a complete list of RefSeq genes in the region in their chromosomal order

Thus, most of our further analysis concentrated on the data for bin width of 500kb for both T1D-controls and RA -controls data sets. It is noticeable that plots C in Fig. 1 and 2 for such bins show interesting substructure in the distribution of the haplotype-block differences across chromosome 6 that was further explored in detail (independently for both T1D and RA associations).

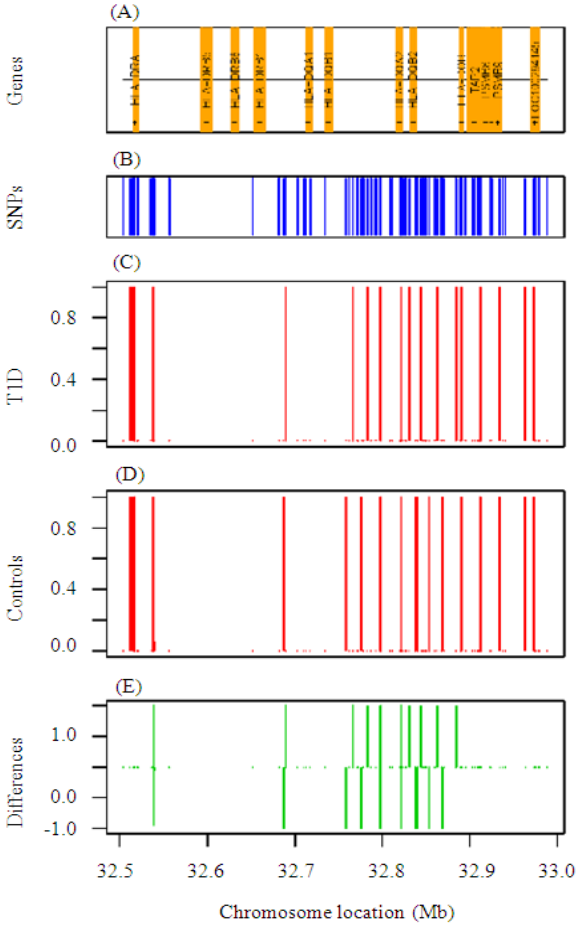

Fig. 4: Positions of T1D associated block differences on the specific genes in the 32.5-33.0 Mb region. Plots of the known RefSeq genes in the part of the MHC region (A), used SNPs locations (B), posterior probabilities of the block boundary start for T1D patients (C) and controls (D), as well as the difference in the posterior probabilities between the case and control data sets (E). For plots A-E the common x-axis (location) is shown at the bottom of plot E. In plot A the strand is marked with "+" or "-" below the gene name. Positions of the RefSeq genes were obtained from the UCSC genomic data base (genome.ucsc.edu). To avoid overlapping names, genes TAP1 and LOC100507463 were not labeled in the plot A. Table 1 for a complete list of RefSeq genes in the region in their chromosomal order

Previous studies (WTCCC, 2007) indicated that because of their autoimmune background, both T1D and RA are known to share same disease loci in the genome.

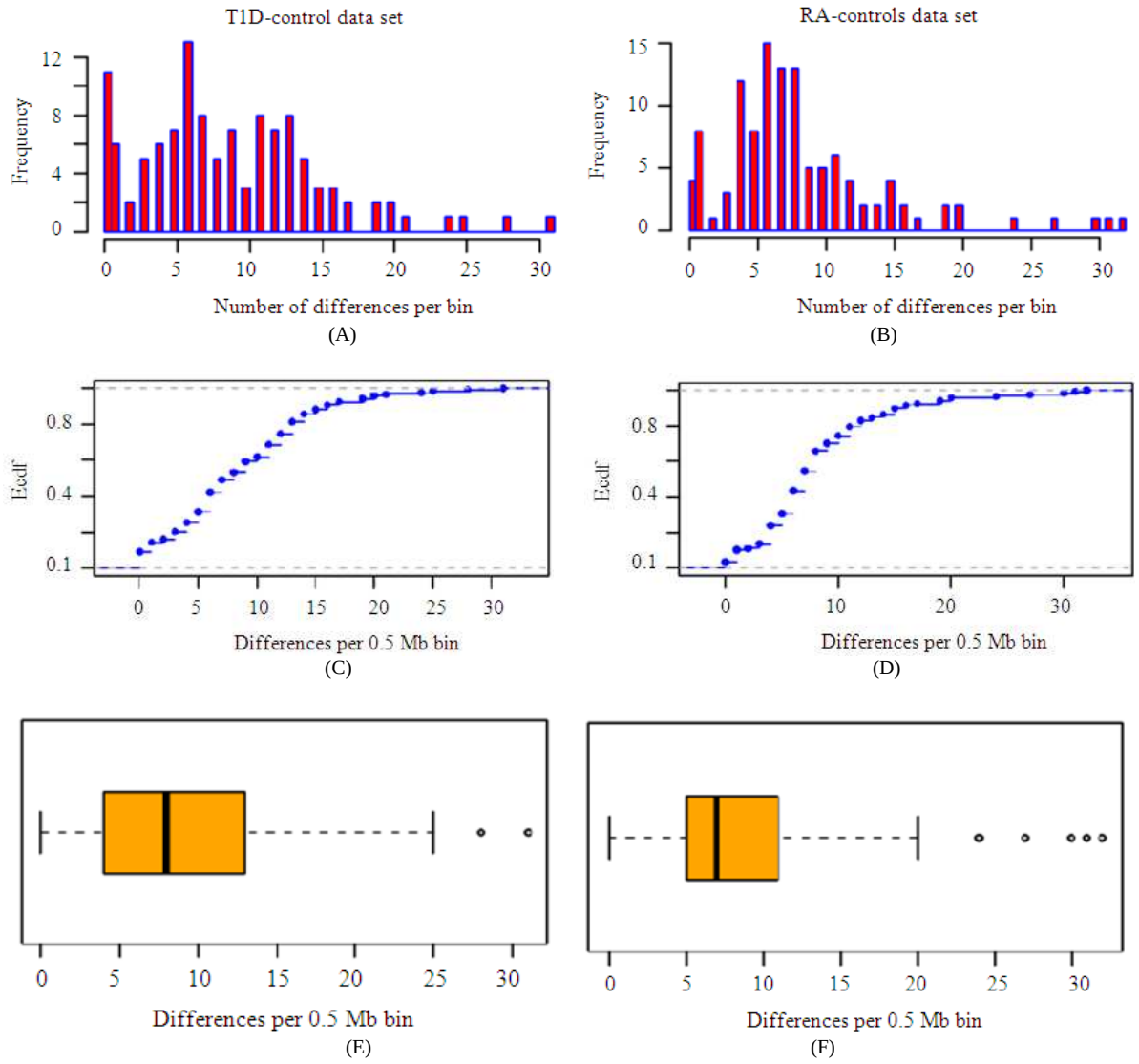

Fig. 5: Distribution characteristics for the determined haplotype-block differences in the 0.5Mb bins. Graphical summary of the LD-block difference distributions for T1D-controls and RA-controls data sets using histograms (A-B), empirical cumulative distribution functions (C-D) and boxplots (E-F) for evaluating central location, dispersion and outliers in the data

More specifically, it has been observed (WTCCC, 2007; Johnson and O'Donnell, 2009; Zhang *et al*., 2011a) that T1D and RA share same loci with the strongest association signals in the Major Histocompatibility Complex (MHC, also known as HLA) region on the chromosome 6. That is why even though we determined the haplotype-block boundaries for the whole chromosome 6, in the analysis part we focused our attention mainly on the short arm of the chromosome 6 (position <60Mb) that contained the MHC and its surrounding regions.

In order to analyze the locations of the disease associated regions on the short arm of the chromosome we looked more closely at the 500kb regions plotted in the Fig. 1C and 2C that were discriminated from the rest by the large number of LD-block differences. Specifically, we looked at the regions for which the number of block differences per bin was larger than 20 (96th-quantile for both data sets).

Table 1 summarizes the regions of interest determined for T1D-controls data set. Specifically, we note whether the region found to have a large number

of LD-block differences between T1D patients and controls has previosly been associated with the T1D. We observed that three such regions that stand out in our study had previously been linked to T1D (Zhang *et al*., 2011b; WTCCC, 2007) either through single or two-SNP strong disease association.

In Table 1 we also note what specific RefSeq genes are present in all the regions with significant differences in the LD-block boundaries identified in this study. Figures 3 and 4 look in detail at the regions of interest in the MHC complex that were known before to be connected to T1D. We showed the known genes in the regions (3A and 4A) and the locations of the determined block boundary differences (3E and 4E).

Table 2 summarizes the regions with the largest block boundary differences that are potentially associated with the RA status of the patients. Two of the determined regions in this study posses the previously known loci of RA, rs1233400 and rs1075496 (WTCCC, 2007) and are located in the MHC region of the chromosome 6. For the rest of the determined regions of high LD block differences we note the known RefSeq genes located on those parts of the chromosome.

DISCUSSION

It is important to note that in both instances we located regions with high hyplotype-block differences for both RA and T1D on the HLA complex that is associated with the autoimmunity and infections (The International HapMap Consortium, 2005). Additionally, two common regions outside of the MHC complex show high differences in the haplotype blocks for both T1D and RA at the same time (Table 1 and 2).

Even though we do not have a definite stochastic model for the distribution of the LD-block differences, plots in Fig. 5 reveal the structure of the data rather strikingly. Despite the fact that there is a background of the differences across the chromosome with the average per 0.5Mb bin between 5 and 10, clearly a few regions for both T1D-controls and RA-controls data sets are outliers of their distributions (Fig. 5E and F). Careful examinations of such extreme observations that are far from the rest of the regions are presented in Table 1 and 2 since the disease associated LD-block differences are most likely to be located there. Finally, the background differences are probably arising from the MCMC step of our approach.

A number of important questions need to be addressed in order to ensure wider applicability of our method as well as to make discrimination of the disease associated block differences easier. Particularly, care must be taken to differentiate the LD-block differences coming from the disease association against those arising from the computational step of our approach: primarily from the convergence of the MCMC chains to different local modes of the distribution. Even though the chains converged to roughly the same final posterior probability of the data set, there were slight differences in the computationally determined block differences; for example, in two MCMC chains for the controls data there were 2,275 differences in the determined blocks out of total of 29,483 SNPs in the chromosome 6 data set (7.7 percent differences). One of the possible solutions would be to employ simulated annealing to ensure the convergence of the different chains to the global mode using the concept of “temperature” (Liu, 2001).

CONCLUSION

In conclusion, we proposed and explored a new method to determine disease associated differences in haplotype-block boundaries based on the Bayesian model and Markov Chain Monte Carlo method. We applied our method to the WTCCC data to search for block differences associated with the autoimmune diseases (T1D and RA). Among the determined regions of high differences lie known loci of the RA and T1D (HLA genes). Additionally, we point to the chromosome 6 regions that should be further tested for T1D and RA associations. Over small spatial scales of 5 and 50kb we did not see any regions containing unusually large numbers of block differences. However, on the spatial scales of 500kb we pointed out the regions of haplotype-block differences that are potentially associated with the RA and T1D diseases. For example, common regions for both diseases of high haplotype-block differences appeared around 11.5-12.0 Mb and 18.5-19.0 Mb and their connections with the autoimmune disorders should be further explored in future studies.

ACKNOWLEDGEMENT

Zhang was supported by start-up funding and Sesseel Award from Yale University. The computation was done with the help from the Yale University Biomedical High Performance Computing Center, which was supported by the NIH grant RR19895. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. All

the chromosomal positions are in NCBI build 35 coordinates. We would like to acknowledge the use of the source code from www.stat.psu.edu/~yuzhang/ which was modified in order to implement the statistical model used during this project.

REFERENCES

Bansal, V., O. Libiger, A. Torkamani and N.J. Schork, 2010. Statistical analysis strategies for association studies involving rare variants. Nature Rev. Genetics, 11: 773-785. DOI: 10.1038/nrg2867

Bottini, N., L. Musumeci, A. Alonso, S. Rahmouni, K. Nika *et al*., 2004. A functional variant of lymphoid tyrosine phosphatase is associated with Type I Diabetes. Nature Genetics, 36: 337-338. DOI: 10.1038/ng1323

Carmichael, M., 2010. One hundred tests. Sci. Am. 303: 50-50. DOI: 10.1038/scientificamerican1210-50

Coenen, M.J.H. and P.K. Gregersen, 2009. Rheumatoid arthritis: A view of the current genetic landscapeRA genetics. Genes Immunity, 10: 101-111. DOI: 10.1038/gene.2008.77

Devendra, D., E. Liu and G. S. Eisenbarth, 2004. Type 1 diabetes: Recent developments. BMJ, 328: 750-750. DOI: 10.1136/bmj.328.7442.750

Ding, K., J. Zhang, K. Zhou, Y. Shen and X. Zhang, 2005a. htSNPer1.0: Software for haplotype block partition and htSNPs selection, BMC Bioinform., 6: 38-38. DOI: 10.1186/1471-2105-6-38

Ding, K., K. Zhou, J. Zhang, J. Knight and X. Zhang *et al*., 2005b. The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. Mol. Biol. Evolut., 22: 148-159. DOI: 10.1093/molbev/msh266

Johnson, A.D. and C.J. O'Donnell, 2009. An open access database of genome-wide association results. BMC Med. Genetics, 10: 6-6. DOI: 10.1186/1471-2350-10-6

Liu, J.S., 2001. Monte Carlo Strategies in Scientific Computing. 1st Edn., Springer, New York, ISBN: 0387952306, pp: 343.

Newton, J.L., S.M.J. Harney, B.P. Wordsworth and M.A. Brown, 2004. A review of the MHC genetics of rheumatoid arthritis. Genes Immunity, 5: 151-157. DOI: 10.1038/sj.gene.6364045

Polychronakos, C. and Q. Li, 2011. Understanding type 1 diabetes through genetics: Advances and prospects. Nature Rev. Genetics, 12: 781-792. DOI: 10.1038/nrg3069

Svoboda, E., 2010. The DNA transistor. Sci. Am., 303: 46-46. DOI: 10.1038/scientificamerican1210-46

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature, 437: 1299-1320. DOI: 10.1038/nature04226

Wai-Yuan, T. and W.Y. Tan,, 2002. Stochastic Models with Applications to Genetics, Cancers, AIDS and Other Biomedical Systems. 1st Edn., World Scientific, River Edge, ISBN: 9810248695, pp: 441.

Wall, J.D. and J.K. Pritchard, 2003. Haplotype blocks and linkage disequilibrium in the human genome. Nature Rev. Genetics, 4: 587-597. DOI: 10.1038/nrg1123

WTCCC, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447: 661-678. DOI: 10.1038/nature05911

Zhang, J., F. Li, J. Li, M.Q. Zhang and X. Zhang, 2004. Evidence and characteristics of putative human α recombination hotspots. Hum. Mol. Genet., 13: 2823-2828. DOI: 10.1093/hmg/ddh310

Zhang, Y. and J.S. Liu, 2007. Bayesian inference of epistatic interactions in case-control studies. Nature Genet., 39: 1167-1173. DOI: 10.1038/ng2110

Zhang, J., Q. Zhang, D. Lewis and M. Zhang, 2011. A Bayesian method for disentangling dependent structure of epistatic interaction. Am. J. Biostat., 2: 1-10. DOI: 10.3844/amjbsp.2011.1.10

Zhang, Y., J. Zhang and J.S. Liu, 2011. Block-based bayesian epistasis association mapping with application to WTCCC Type 1 Diabetes data. Ann. Applied Stat., 5: 2052-2077. DOI:10.1214/11-AOAS469