American Journal of Bioinformatics 1 (1): 70-78, 2012
ISSN 1948-9862

Bayesian Exploration of Multilocus Interactions on the Genome-Wide Scale

[1]Ivan Kozyryev and [2,3]Jing Zhang
[1]Department of Physics, Harvard University, Cambridge, MA, USA
[2]Department of Statistics, Yale University, New Haven, CT, USA
[3]Program in Computational Biology and Bioinformatics,
Yale University, New Haven, CT, USA

Abstract: Problem statement: Recent technological and scientific advances propelled the field of Genome-Wide Association Study (GWAS), which promises to be instrumental in linking many common complex diseases to their genetic origin. While so far such large-scale surveys have been moderately successful in identifying disease related genetic variants, much of disease heritability is still not accounted for by the discovered loci. There is an urgent need for advanced statistical methods for efficient automatic detection of complicated multilocus interactions on significant scales. **Approach:** Novel statistical methods based on Bayesian data analysis ideas, specifically Bayesian modeling, Bayesian variable partitioning, graphical and network models are promising to aid in search for missing disease heritability and shed light on complex biological processes involved in disease development. First crucial difference setting these methods apart from all the mainstream previous approaches (hypothesis testing methods) is their joint disease mapping capability via the simultaneous fitting of a statistical model for the whole case-control data set. Additionally, such Bayesian methods allow for the construction of complicated data models and quantitative incorporation of diverse prior information into the final statistical model. **Results:** The use of Bayesian techniques has already yielded new insights into the details of epistatic interactions across the genome associated with various important diseases. **Conclusion/Recommendations:** Bayesian approaches provide a way to detect and understand complicated multilocus interactions that already started to elucidate important disease pathways. As the field of GWAS matures, Bayesian strategies can surely aid in converting such multiple surveys into useful biomedical information.

Key words: Genome-wide association study, Bayesian model selection, epistasis, linkage disequilibrium

INTRODUCTION

The promise of personalized medicine and genomics: Improved disease prevention and diagnosis as well as novel routes to therapies are the main motivations for extensive studies aimed at finding disease related genes and variants. Particularly, genetic tests capable of showing individual's risks to develop certain diseases would help to tailor preventive and therapeutic treatments to every single patient in order to achieve best possible results (Hall, 2010; McCarthy *et al*., 2008). While there are already a few companies offering 'consumer genomics' services to provide estimated disease risks via characterization of known genetic risk factors (Donnelly, 2008; Carmichael, 2010), currently this kind of information on genetic markers can give only a limited help in common illness propensity risk assessment (Donnelly, 2008; Hall, 2010). Even though a plethora of resources has been directed in this direction in the past dozen years, the genetic basis of common human diseases has not been identified for the most part (WTCCC, 2007). Recent emergence of successful strategies for the genome-wide association studies was supposed to provide the necessary tools for deciphering genetic causes of complex human illnesses like type 1 and 2 diabetes (Todd *et al*., 2007), rheumatoid arthritis and bipolar disorder (Hirschhorn and Daly, 2005; WTCCC, 2007).

An emergence and development of GWAS*:* An examination of an immense number of genetic markers across the whole genome for multiple individuals with the goal of identifying variants-disease associations is known as Genome-Wide Association Study (GWAS). Novel scientific and technological advances (Metzker, 2010; Branton *et al*., 2008; Schaffer, 2012) made GWAS fully capable of unlocking the basis of complex diseases. Particularly, development of the International

Corresponding Author: Jing Zhang, Department of Statistics, Yale University, New Haven, CT, USA

HapMap resource (IHMC, 2005) that simplified design and analysis of association studies, emergence of dense genotyping chips (Metzker, 2010; Svoboda, 2010) and assembly of large and characterized clinical samples (WTCCC, 2007) should be singled out as important factors in GWAS recent successful progress. While many disease loci have been identified in such surveys (WTCCC, 2007; Johnson and O'Donnell, 2009), discovered variants explain only a small proportion of the observed familial aggregation (McCarthy *et al*., 2008; Altshuler and Daly, 2007). This is known as a 'missing heritability problem' (Gibson, 2012). Currently there are three alternative mainstream ideas for the genetic architecture of complex diseases: the infinitesimal model, the rare allele model and the broad sense heritability model (Gibson, 2012). Thus, the most urgent contemporary debate that needs to be solved is regarding the architecture of complex human traits. While, 'common variant' hypothesis has come under a lot of criticism lately (Hall, 2010; Gibson, 2012), it is now necessary to dig deeper and choose which one of the alternative proposed architectures is closer to reality in order to help develop future studies efficiently (Gibson, 2012; Hall, 2010; Donnelly, 2008).

Beyond single-locus analysis: Despite striking success in the 20th century in pinpointing genes responsible for mendelian diseases, genetic origins of common complex diseases are, in fact, non-mendelian in nature (Zhang and Liu, 2007; Jiang *et al*., 2011). Particularly, gene-gene interactions are involved in many complex biological processes like metabolism, signal transduction and gene regulations and, thus, genetic variants in multiple loci may contribute to the disease formation together (Moore, 2003; Chen *et al*., 2011a). For example, breast cancer and type 2 diabetes have been linked to multi-SNP interactions (Chen *et al*., 2011a; Ritchie *et al*., 2001; Wiltshire *et al*., 2006). While most current bioinformatics approaches focus on detecting single-SNP associations, advanced statistical methods are necessary for multi-SNP association mapping because single-variant methods not only loose power when interactions exist but are, in fact, helpless in detecting rare mutations (Zhang, 2012). Also, the number of possible interactions is so vast that it is computationally unrealistic to search though all possible interactions in the genome for a large scale case-control study (Zhang *et al*., 2011a; Cordell, 2009).

Additional challenge for disease origin discovery comes from the statistical correlation between nearby variants known as linkage disequilibrium or LD (Zhang *et al*., 2011a; Kozyryev and Zhang, 2012). LD patterns have many important applications in genetics and biology (Wall and Pritchard, 2003) and arise due to shared ancestry for contemporary chromosomes (IHMC, 2005). Due to LD patterns, it is likely that there will be a lot of redundant positive signals in dense studies (Zhang, 2012). Later on we address in detail how Bayesian strategies can address the burning problems in genetics while dealing with epistasis and linkage disequilibrium.

Statistical approaches for GWAS: Currently, most of the approaches to disease association mapping employ the standard 'frequentist' attitude to the evaluation of significance (McCarthy *et al*., 2008). Particularly, such algorithms use hypothesis testing procedures to deal with one variant at a time (Zhang, 2012). The accepted threshold for the p-value is $\sim 5\times10^{-8}$ (Risch and Merikangas, 1996; Hoggart *et al*., 2008; McCarthy *et al*., 2008). However, failures of such 'frequentist' methods to account for the power of a study and the number of likely true positives (McCarthy *et al*., 2008) combined with the increased likelihood to report a multitude of redundant associations (Zhang, 2012) sparked a wide interest in the Bayesian procedures. In this review we survey the challenges facing statistical geneticists while analyzing the GWAS data and outline how recently emerged Bayesian methods can help with the process. In addition to outlining the main differences between various proposed approaches, we highlight limitations and advantages of each method and describe future prospects in the field and how Bayesian approaches can aid in answering outstanding questions in biomedicine.

MATERIALS AND METHODS

Previously, we mentioned multiple complicated interactions that have to be considered while developing statistical models for understanding of the multilocus interactions. In Fig. 1 we summarize all the relevant interactions present in the GWAS in the graph. The ultimate goal is to be able to accurately understand all the shown couplings in large-scale case-control studies while also comprehending the biological processes that lead to disease development. Thus, while statistical understanding is important, developing methods that can point in the direction of the appropriate biological processes taking place is the next ultimate goal.

Overview of Bayesian data analysis: Statistical conclusions about an unknown parameter θ (or unobserved data y_{unobs}) in the Bayesian approach to parameter estimation are described utilizing probability statements which are conditional on the observed data y: $p(\theta|y)$ and $p(y_{unobs}|y)$. Additionally, implicit conditioning is performed on the values of any covariates (Gelman *et al*., 2004).

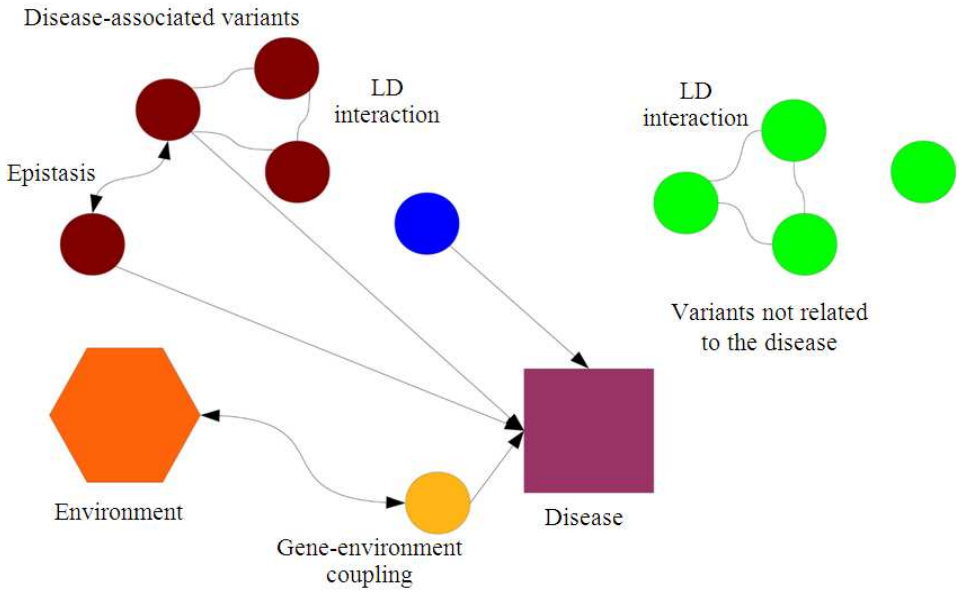

Fig. 1: Schematic graph representation of all the relevant interactions in the genome and paths to disease formation. SNPs are shown as circles with color indicating their disease connection: 'green' SNPs are not associated with the phenotype of interest, 'blue' are marginally associated, 'brown' are influencing disease formation either through epistasis or they are in Linkage Disequilibrium (LD) with such variants and 'orange' ones can lead to disease formation through gene-environment interactions. LD between different variants is depicted as lines without arrows, while gene-gene and gene-environment couplings are represented by lines ending with arrows at both ends. In the paper we review which of the interactions in the graph can be efficiently discovered using the novel Bayesian approaches.

The concept of conditioning on the observed data is what separates Bayesian statistics from other inference approaches which estimate unknown parameter over the distribution of the possible data values while conditioning on the true, yet unknown parameter value (Gelman *et al*., 2004; Rice, 2006).

At the heart of all the Bayesian approaches for detection of gene-gene interactions lies the concept of Bayesian inference and, specifically, Bayesian model selection. The goal is to determine the posterior distribution of all parameters in the problem (disease association, epistatic interactions and block structures), given the common variants data for the case-control study while incorporating prior beliefs about parameter values. The conditional probability of all parameters given the observed data is proportional to the product of the likelihood function of the data and prior distribution on the parameters (Rice, 2006):

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})} \quad (1)$$

For all large data sets encountered in GWAS P(data) cannot be explicitly calculated (Zhang and Liu, 2007) and, therefore, P(parameters|data) can be known only up to the proportionality constant as shown in Eq. 1. However, advanced computational techniques (iterative sampling methods) can be used to determine posterior distribution of parameters (Liu, 2008; Rice, 2006). The main task is to make appropriate choices of statistical models to describe P (data|parameters) and also to choose appropriate prior distributions on the values of parameters: P(parameters).

Overview of Bayesian variable partition: Instead of testing each SNP set in a stepwise manner (Marchini *et al*., 2005; Liu *et al*., 2011), Bayesian approaches fit a single statistical model to all of the data simultaneously (Zhang and Liu, 2007; Zhang *et al*., 2011a; 2011b) allowing for increased robustness when compared to hypothesis testing methods (McCarthy *et al*., 2008; Zhang, 2012). Another advantage of Bayesian approach to the problem is the ability to quantify all the uncertainties and information and to incorporate previous knowledge about each specific SNP marker into the statistical model through the priors (Zhang and Liu, 2007; Rice, 2006).

In the Bayesian model selection framework, we are interested in figuring out which of the set of models {M} is the most likely one given the observed data (X). In analogous way to Eq. 1, we can find the posterior probability for a particular model M_i given data, by replacing parameters with M_i:

$$P(M_i \mid X) \propto P(X \mid M_i)P(M_i) \quad (2)$$

Thus, through comparison of $P(M_i|X)$ and $P(M_j|X)$ it can be determined using Eq. 2 whether model M_i or M_j is more likely (Rice, 2006). Now let's consider how this conceptual framework is applied in practice to the extraction of multilocus interactions in GWAS.

Epistasis analysis in genome-wide data sets: While statistical methods like BGTA (Zheng *et al*., 2006), MARS (Cook *et al*., 2004) and CPM (Nelson *et al*., 2001) are capable of detecting epistatic associations, the Bayesian Epistasis Association Mapping (BEAM) algorithm (Zhang and Liu, 2007) was the first practical approach capable of handling genome-wide case-control data sets. BEAM algorithm gives for each SNP marker posterior probabilities for disease association and epistatic interaction with other markers given the case-control genotype SNP data. The core of the Bayesian marker partition model used can be briefly summarized as follows.

BEAM can detect both interacting and non-interacting disease loci among a large number of variants. It is an application of Bayesian model selection procedure. Particularly, all the markers are split into three non-overlapping groups: (1) markers not associated with the disease, (2) marginally disease-associated variants and (3) those with interaction associated disease effect. Thus, using the priors on the marker memberships and Markov Chain Monte Carlo (MCMC) methods, posterior probabilities for group memberships are determined. Specifically, by interrogating each SNP marker conditionally on the current status of others via MCMC method the algorithm produces posterior probabilities (Zhang and Liu, 2007). Particularly, the genotype counts are modeled by the multinomial distribution with the frequency parameters described by the Dirichlet prior. In order to determine the posterior probability of each marker's group membership (represented by I) the Metropolis-Hastings (MH) algorithm (Liu, 2008) is used to sample from P(I|D,H) as given in Eq. 3:

$$P(I \mid D,H) \propto P(D1 \mid I)P(D2 \mid I)P(D0,H \mid I)P(I) \quad (3)$$

Where:

D = The patient data set (with disease)
H = The control data set (healthy) and then D0
D1 and D2 = Correspondingly partitions of the patient data set into three categories described above

The assumption is that case genotypes at the disease associated markers will have different distributions when compared to control genotypes. Furthermore, the likelihood model assumes independence between markers in control group.

While BEAM algorithm was one of the first few to be able to handle GWAS data, it suffered from an assumption that SNPs dependence structure could be described by the Markov chain (Zhang *et al*., 2011a; Zhang and Liu, 2007). In fact, SNP markers are highly correlated within haplotype blocks which are separated by the recombination events (IHMC, 2005; Reich *et al*., 2001). Therefore, despite its successful approach, BEAM model is not able to capture the block-like human genome structure.

Incorporating block-type genome structure: A new Bayesian model that infers LD-blocks and chooses SNP markers in the blocks that are disease associated, therefore successfully incorporating diplotype blocks in the human genome into the Bayesian approach proposed by Zhang and Liu (2007) is known as BEAM2 algorithm developed by Zhang *et al*. (2011a). The statistical Bayesian model for the LD-block structure is summarized in Zhang *et al*. (2011a) and Kozyryev and Zhang (2012). The main assumption is that diplotypes of individuals come from a multinomial distribution with frequency parameters described by the Dirichlet prior and that genotype combinations of SNPs in different LD blocks are mutually independent, which is a good approximation to reality (Zhang *et al*., 2011a). Therefore, the compact expression for the marginal probability of the data for a specific block is given by Eq. 4:

$$P(D_{[s,b)} \mid [s,b) = \text{block}) = \left(\prod_{i=1}^{3^{b-s}} \frac{\Gamma(n_i + a_i)}{\Gamma(a_i)}\right) \frac{\Gamma\left(\sum a_i\right)}{\Gamma\left(\sum (n_i + a_i)\right)} \quad (4)$$

where, a block of SNPs considered is (s,…,b-1); Γ is the gamma function, $\vec{a}$ is the vector of Dirichlet parameters and n_i refers to the number of counts for a specific diplotype. For joint inference of the diplotype blocks and disease association status we use the joint statistical model for the observed genotype data in cases and controls, the marker membership and block partition variable as in Eq. 5:

$$P(D,H,B,I) = P(D,H \mid B,I)P(B)P(I) \quad (5)$$

Finally, in order to determine the posteriors P(B|D,H) and P(I|D,H) the model uses a combination of MH algorithm and Gibbs sampler (Liu, 2008; Zhang *et al*., 2011a).

Detailed interaction partition structure determination: While successful in inferring epistatic interactions in GWAS, both BEAM and BEAM2 had a disadvantage of using saturated models which limited the ability of the algorithms to accurately determine the structure of the epistatic interactions among different disease related markers. However, recent studies showed that such interaction details arising due to encoding of the complicated regulatory mechanisms might play an important role in the disease formation (Zhang *et al*., 2011b; Yang *et al*., 2009; WTCCC, 2007). In order to be able to carefully explore the etiopathogenesis and genetic mechanisms of diseases, Zhang *et al*. (2011b) proposed the Recursive Bayesian Partition (RBP) algorithm. The RBP approach attempts to search for conditional independence and independence groups among interacting markers. RBP first recursively infers all the marginally independent interaction groups (no interaction between groups) and then infers the conditional independence within each group using chain-dependence model. RBP therefore successfully recursively determines dependence structure among interacting variants in the GWAS setting. Figure 2 shows an example of the possible outcomes of the RBP algorithm applied to GWAS data when determining the epistatic interactions independence structure.

Bayesian graph models and networks: Here we describe BEAM3 algorithm developed by Zhang (2012) and how it improves on BEAM and BEAM2 models and what genetic problems and questions it can help to address. Through the use of Bayesian graphical method, BEAM3 detects flexible interaction structures instead of using saturated models (like BEAM and BEAM2 do), therefore, highly reducing the multi-SNP model complexity. Moreover, because only the disease association graphs are constructed, BEAM3 provides for higher computational efficiently in the GWAS settings (Zhang, 2012).

In detail, Zhang (2012) allowed for higher-order couplings via saturated interactions within cliques (non-overlapping partition of SNPs) and pairwise interactions between them. It can be shown (Zhang, 2012) that the joint probability of all SNPs X, parameters, including disease graph and association status (G, I) and disease status indicator (Y) is given by:

$$P(X,Y,G,I) \propto \frac{P_A(X_1 \mid Y,G)}{P_0(X_1)} P(G \mid I) P(I) \quad (6)$$

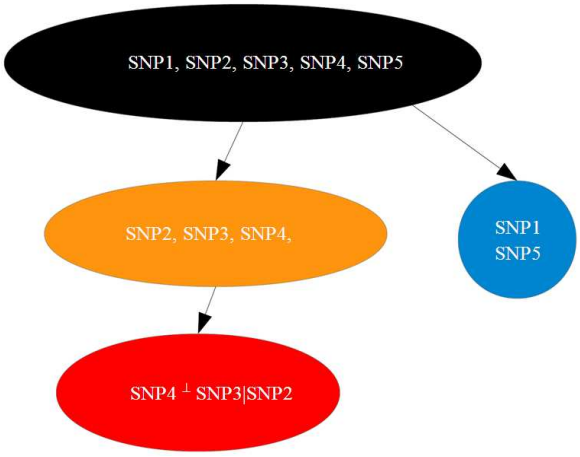

Fig. 2: A diagram of the procedure for the inference of a detailed dependence structure among disease related variants or mutations when using recursive Bayesian partition (RBP) as done in Zhang *et al*. (2011b). In this simple example, five SNPs (numbered SNP1 though SNP5) were assumed to be associated with the phenotype in question. The independence groups within the set of those SNPs are singled out using circles/ovals and different colors. There is a strong conditional independence in the group of 'red' SNPs {2, 3, 4} while 'blue' SNP1 and SNP5 are independent of the other three disease-associated variants

where, G = (C,Δ) is an undirected disease graph constructed on disease associated SNPs (X_1) and including partition of SNPs into cliques (C) and interaction between cliques (Δ); probability function of X_1 set under the phenotype association hypothesis is described by P_A. Therefore, as can be seen from Eq. 6, only a few disease-associated SNPs are modeled (in set X_1) and hence a significant portion of computational time is saved due to avoiding explicit modeling of complicated dependence structures of all SNPs which could be millions (Zhang, 2012; Zhang and Liu, 2011; Jiang *et al*., 2011). Additionally, through the choice of a proper baseline probability function $P_0(X_1)$, the model automatically accounts for the complex LD effects among dense SNPs employing graphs. Thus, a significant number of repetitive false interactions are avoided reducing computational burden (Zhang, 2012).

In a different direction, Bayesian methodology has been also applied to data-mining and machine learning approaches to improve detection of gene-gene interactions in GWAS. Chen *et al*. (2011a) proposed to use a Bayesian classification tree model for identification of multilocus interactions in the large-scale data sets.

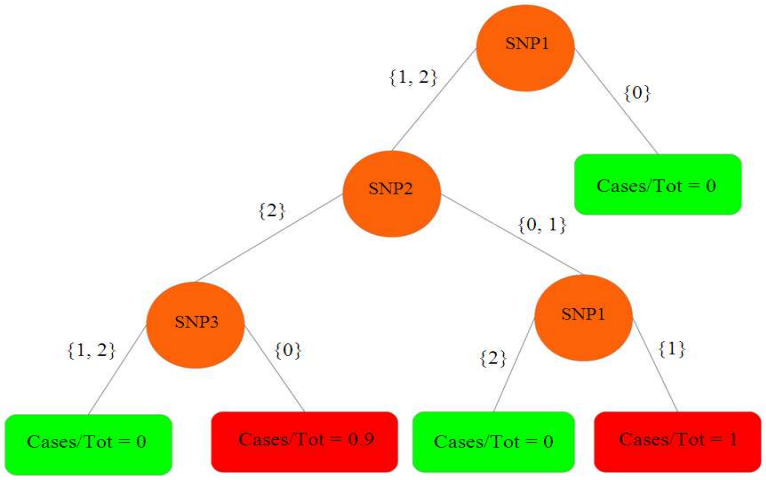

Fig. 3: A simple example of the classification tree structure representation of the disease associated multilocus interactions as used in Chen *et al*. (2011a) for Bayesian classification tree search method. Predictor variables (SNPs here) are shown as orange circles while edges are marked with the genotype values. The terminal nodes in the graph represent the partition of the feature space with each subject eventually being assigned to only one such node. Each terminal node is marked with the proportion of the case individuals assigned there, with green-colored nodes containing mainly controls and red-colored ones containing mainly cases. As an example, from this simple tree model we can conclude that subjects with SNP1 value of {1} and SNP2 value of {0,1} are very likely to have a disease in question, therefore it is possible that interaction between this common variants leads to the formation of the disease. The goal of the method is the determination of the posterior distribution for the binary tree given the observed data in case-control study

Specifically, this kind of machine-learning approach produces tree-structure models where each nonterminal node determines the splitting rule based upon the predictor variables like SNP genotypes and edges between nodes correspond to different possible values for the variable in the top parent node. In summary, a path along such a tree till the terminal node represents a specific combination of predictor variables along the path, in such sense, accommodating for the multilocus interactions (Cordell, 2009; Chen *et al*., 2011a). For example, Fig. 3 shows an example of such a tree model.

There are various ways for searching through tree space in such recursive partitioning approaches including greedy algorithms (Hastie *et al*., 2009), random forests approach (Breiman, 2001; Cordell, 2009) and MCMC (Chipman *et al*., 1998; Denison *et al*., 1998). Bayesian variable partition and Bayesian classification trees are, in fact, conceptually very similar in that prior is assigned to all the tree models with the purpose of controlling the tree size (Chen *et al*., 2011a). One main advantage of this approach is in a possible enhancement of finding probability for epistatic interactions with weak marginal effects due to ensuring the variable splitting through the prior specification (Chen *et al*., 2011a). Moreover, due to the adaptivity of the MCMC algorithm, such Bayesian tree models detect higher-order interactions by performing thorough searches near trees with the interacting variables determined in previous iterations (Chen *et al*., 2011a). However, one of the major drawbacks of such approaches is that they do not test for interactions directly, but instead allow for them, while testing for associations in the data (Cordell, 2009).

RESULTS

Even though practical Bayesian approaches for GWAS multilocus interactions analysis have emerged relatively recently, such methods have already helped to make important advances in determination of disease etiology. Table 1 succinctly summaries all the Bayesian methods described above as well as their success in determination of the previously known disease loci and, more importantly, in the discovery of new multilocus interactions responsible for complex diseases. For example, Zhang *et al*. (2012) discovered 319 high-order interactions across the genome that can potentially explain the missing genetic component of the Rheumatoid Arthritis (RA) susceptibility. Moreover, their findings indicate that nervous system, in addition to autoimmune one, potentially performs a crucial role in RA development. This is an example of the statistical study in which disease underlying biological processes can be extracted from determined statistical associations.

Table 1: A comparison of novel Bayesian approaches for GWAS epistasis analysis. As can be seen from the table, studies applying Bayesian methodology not only confirmed previously detected disease loci in large-scale data sets, but moreover have already identified potential missing heritability in the form of multilocus interactions

Statistical method	Brief description	Genome-wide data set	Results/detected loci
BEAM (Zhang and Liu, 2007)	Epistasis detection	AMD GWA data set[a]	More powerful than previous approaches
BEAM2 (Zhang *et al*., 2011a)	Epistasis/LD-block detection	WTCCC T1D[b]	Many previous loci+new two-way associations
RBP (Zhang *et al*., 2011b)	Detailed independence structure of epistasis	dbMHC[c] T1D data set	Confirmed previously known saturated interactions
BEAM3 (Zhang, 2012)	Bayesian graph model for epistasis/LD	WTCCC IBD[d] data set	All previous IBD loci+2 new+2 interchr.[f] interactions
Bayesian Classification Tree (Chen *et al*., 2011a)	Classification tree model/ recursive partitioning	Crohn's disease data	Possible epistasis identified
Haplotype Block Differences (Kozyryev and Zhang, 2012)	Separate LD-block determination for cases and controls	WTCCC T1D and RA[e] chr6 data sets	Detected differences around previously known loci + near new positions
BEAM+BEAM2 (Zhang *et al*., 2012)	High-order epistatic interactions study	WTCCC RA data set	319 high-order interactions found

[a]Age-related macular degeneration genome-wide association data set with 116,204 SNPs for 96 cases and 50 controls. [b]Type 1 Diabetes (T1D) data generated by the Welcome Trust Case-Control Consortium; [c]This data contained resequenced haplotypes of exons for DRB1 and DQB1 genes in the MHC region (Zhang *et al*., 2011). [d]Inflammatory bowel disease (IBD) data set from the Welcome Trust Case-Control Consortium with 2,005 IBD patients and 3,004 combined controls (Zhang, 2012). [e]Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D) data generated by the Welcome Trust Case-Control Consortium; [f]Interchromosome

For sure, many more studies will follow in the near future that apply Bayesian methods either to existing GWAS data or to new large scale studies that will be produced soon (Hayden, 2012).

DISCUSSION

Certain issues need to be considered when using Bayesian approaches described above. For example, a combination of genotyping errors, disease heterogeneities and population substructures could have adverse effect on the statistical results of the methods (Zhang and Liu, 2007). Currently, the major problem with GWAS approaches is that the determined disease associated genetic regions explain only a small part of the disease heritability (Donnelly, 2008; WTCCC, 2007). However, it is possible that with the improved statistical methods outlined above the situation will soon change after the detailed understanding of the interactions involved emerges. Additionally, the main criticism of the GWAS based on the SNPs analysis, is that it is hard to understand the causal biology taking place in the disease formation (Hall, 2010); however, with the development of the recent Bayesian models that provide the detailed structure of the multilocus interactions (Zhang *et al*., 2011b; 2012) detailed etiopathogenesis of many diseases may soon be elucidated.

Improvements to the Bayesian approaches mentioned in this article can include incorporation of environmental factors and population structures as covariates in the statistical model (Zhang *et al*., 2011b; Lobach *et al*., 2010). Another possible improvement is to impute untyped SNPs and missing genotypes from the reference panel (Zhang *et al*., 2011b; Zhang, 2011; Marchini *et al*., 2007). Moreover, utilization of sophisticated methods for incorporation of prior biological knowledge (like pathway topology) can increase the probability of making discoveries in association studies (Chen *et al*., 2011b).

While the main focus of this review article was on statistical methods to determine disease-related interactions among genetic variants, it is important to keep in mind the relationship between determined mathematical coupling and its biochemical underpinnings. Particularly, a common view is that disease development at large is prompted by biomolecular or protein-protein interactions at the molecular level (Cordell, 2009; Gibson, 2012; Jiang *et al*., 2011). While studying specifics of multilocus interactions has potential to convey the details of biological and biochemical disease pathways, the biological interpretation of the determined single- and multi-variant effects is the current burning issue in genetics (Cordell, 2009). The crust of the problem is the necessity to infer biological interaction from a statistical one and the straightforwardness of this process is highly debated by geneticists and epidemiologists (Cordell, 2009; Greenland, 2009). It has been even suggested that functional epistasis might not be detectable in the current GWAS as statistical interactions (Greenland, 2009; Vanderweele, 2009). Soon we will be able to solve this debate using actual results from the ongoing studies. One possible solution is to merge together data-driven and hypothesis-motivated approaches (Jiang *et al*., 2011; Hall, 2010).

CONCLUSION

In conclusion, Bayesian approaches are filling an important previously empty niche in bioinformatics and

genomics research and the future of this scientific area looks extremely exciting and, for sure, will promptly bring a multitude of important surprises.

ACKNOWLEDGEMENT

Zhang was supported by the start-up funding and Sesseel Award from Yale University.

REFERENCES

Altshuler, D. and M. Daly, 2007. Guilt beyond a reasonable doubt. Nature Genet., 39: 813-815. DOI: 10.1038/ng0707-813

Branton, D., D.W. Deamer, A. Marziali, H. Bayley and S.A. Benner *et al*., 2008. The potential and challenges of nanopore sequencing. Nature Biotech. 26: 1146-1153. PMID: 18846088

Breiman, L., 2001. Random forests. Mach. Learn., 45: 5-32. DOI: 10.1023/A:1010933404324

Hayden, E.C., 2012. Open-data project aims to ease the way for genomic research. Nature. DOI: 10.1038/nature.2012.10507

Chen, M., J. Cho and H. Zhao, 2011a. Detecting epistatic SNPS associated with complex diseases via a bayesian classification tree search method. Ann. Hum. Genet., 75: 112-121. DOI: 10.1111/j.1469-1809.2010.00627.x

Chen, M., J. Cho and H. Zhao, 2011b. Incorporating biological pathways via a markov random field model in genome-wide association studies. PLoS Genet., DOI: 10.1371/journal.pgen.1001353

Chipman, H.A., E.I. George and R.E. McCulloch, 1998. Bayesian CART model search. J. Am. Stat. Assoc., 93: 935-948.

Cook, N.R., R.Y.L. Zee and P.M. Ridker, 2004. Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. Stat. Med., 23: 1439-1453. DOI: 10.1002/sim.1749

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

Cordell, H.J., 2009. Detecting gene–gene interactions that underlie human diseases. Nat. Genet. 10: 392-404. DOI: 10.1038/nrg2579

Denison, D.G.T., D.B.K. Mallick and A.F.M. Smith, 1998. A Bayesian CART algorithm. Biometrika, 85: 363-377. DOI: 10.1093/biomet/85.2.363

Donnelly, P., 2008. Commentary article progress and challenges in genome-wide association studies in humans. Nature, 456: 728-731. DOI: 10.1038/nature07631

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin, 2004. Bayesian Data Analysis. 2nd Edn., Chapman and Hall/CRC, Boca Raton, Fla., ISBN-10: 158488388X, pp: 668.

Gibson, G., 2012. Rare and common variants: Twenty arguments. Nature Rev., 13: 135-145. DOI: 10.1038/nrg3118

Greenland, S., 2009. Interactions in epidemiology: Relevance, identification and estimation. Epidemiology, 20: 14-17. DOI: 10.1097/EDE.0b013e318193e7b5

Hall, S.S., 2010. Revolution postponed. Sci. Am., 303: 60-67. DOI: 10.1038/scientificamerican1010-60

Hastie, T., R. Tibshirani and J.H. Friedman, 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edn., Springer, New York, ISBN-10: 0387848576, pp: 745.

Hirschhorn, J.N. and M.J. Daly, 2005. Genome-wide association studies for common diseases and complex traits. Nature Rev. Genet., 6: 95-108. DOI: 10.1038/nrg1521

Hoggart, C.J., T.G. Clark1, M.D. Iorio, J.C. Whittaker and D.J. Balding, 2008. Genome-wide significance for dense SNP and resequencing data. Genet. Epidemiol., 32: 179-185. DOI: 10.1002/gepi.20292

Jiang, X., R.E. Neapolitan, M.M. Barmada and S. Visweswaran, 2011. Learning genetic epistasis using Bayesian network scoring criteria. BMC Bioinform., 12: 89. DOI: 10.1186/1471-2105-12-89

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

Kozyryev, I. and J. Zhang, 2012. Bayesian determination of disease associated differences in haplotype blocks. Am. J. Bioinform., 1: 20-29. DOI: 10.3844/ajbsp.2012.20.29

Lobach, I., R. Fan and R.J. Carroll, 2010. Genotype-based association mapping of complex diseases: Gene-environment interactions with multiple genetic markers and measurement error in environmental exposures. Genet. Epidemiol., 32: 792-802. DOI: 10.1002/gepi.20523

Liu, J.S., 2008. Monte Carlo Strategies in Scientific Computing. 1st Edn., Springer, New York, ISBN-10: 0387952306, pp: 343.

Liu, Y., H. Xu, S. Chen, X. Chen and Z. Zhang *et al*., 2011. Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. PLoS Genet., 7: e1001338-e1001338. PMID: 21437271

Marchini, J., P. Donnelly and L.R. Cardon, 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet., 37: 413-417. DOI: 10.1038/ng1537

Marchini, J., Howie, B., Myers, S., McVean, G. and P. Donnelly, 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet., 39: 906-913. DOI: 10.1038/ng2088

McCarthy, M.I., G.R. Abecasis, L.R. Cardon, D.B. Goldstein and J. Little *et al*., 2008. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat. Rev. Genet., 9: 356-369. DOI: 10.1038/nrg2344

Metzker, M.L., 2010. Sequencing technologies-the next generation. Nat. Rev. Genet., 11: 31-46. DOI: 10.1038/nrg2626

Moore, J.H., 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. Hered., 56: 73-82. DOI: 10.1159/000073735

Nelson, M.R., S.L.R. Kardia, R.E. Ferrell and C.F. Sing, 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res., 11: 458-470. DOI: 10.1101/gr.172901

Reich, D.E., M. Cargill, S. Bolk, J. Ireland and P.C. Sabeti *et al*., 2001. Linkage disequilibrium in the human genome. Nature, 411: 199-204. DOI: 10.1038/35075590

Rice, J.A., 2006. Mathematical Statistics and Data Analysis. 2nd Edn., Academic Internet Publishers, ISBN-10: 1428814051, pp: 85.

Risch, N. and K. Merikangas, 1996. The future of genetic studies of complex human diseases. Science 273: 1516-1517. DOI: 10.1126/science.273.5281.1516

Ritchie, M.D., L.W. Hahn, N. Roodi, L.R. Bailey and W.D. Dupont *et al*., 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet., 69: 138-147. PMID: 11404819

Schaffer, A., 2012. Nanopore sequencing. Technol. Rev.

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

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

Todd, J.A., N.M Walker, J.D. Cooper, D.J. Smyth1 and Kate Downes *et al*., 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet., 39: 857-864. DOI: 10.1038/ng2068

Vanderweele, T.J., 2009. Sufficient cause interactions and statistical interactions. Epidemiology, 20: 6-13. DOI: 10.1097/EDE.0b013e31818f69e7

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

Wiltshire, S., J.T. Bell1, C.J. Groves1, C. Dina and A.T. Hattersley *et al*., 2006. Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21-25 and 10q23-26 in Northern Europeans. Ann. Hum. Genet., 70: 726-737. DOI: 10.1111/j.1469-1809.2006.00289.x

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

Yang, Y., C. He and J. Ott, 2009. Testing association with interactions by partitioning chi-squares. Ann. Hum. Genet., 73: 109-117. DOI: 10.1111/j.1469-1809.2008.00480.x

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

Zhang, Y. and J.S. Liu, 2011. Fast and accurate approximation to significance tests in genome-wide association studies. J. Am. Stat. Assoc., 106: 846-857. DOI: 10.1198/jasa.2011.ap10657

Zhang, Y., J. Zhang and J.S. Liu, 2011a. 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

Zhang, J., Q. Zhang, D. Lewis and M.Q. Zhang, 2011b. 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., 2011. Bayesian epistasis association mapping via SNP imputation. Biostatistics, 12: 211-222. DOI: 10.1093/biostatistics/kxq063

Zhang, Y., 2012. A novel bayesian graphical model for genome-wide multi-SNP association mapping. Gen. Epidem., 36: 36-47. DOI: 10.1002/gepi.20661

Zhang, J., Z. Wu, C. Chao and M.Q. Zhang, 2012. High-order interactions in rheumatoid arthritis detected by bayesian method using genome-wide association studies data. Am. Med. J., 3: 56-66. DOI: 10.3844/amjsp.2012.56.66

Zheng, T., H. Wang and S.H. Lo, 2006. Backward Genotype-Trait Association (BGTA)-based dissection of complex traits in case-control designs. Hum. Hered., 62: 196-212. DOI: 10.1159/000096995