Journal of Biomedical Informatics xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis

Dairong Wang [a], Henry P. Parkman [b], Michael R. Jacobs [c], Anurag K. Mishra [d], Evgeny Krynetskiy [c,d], Zoran Obradovic [a,*]

[a] Data Analytics and Biomedical Informatics Center, Temple University, Philadelphia, PA, USA
[b] Gastroenterology Section, Temple University School of Medicine, Philadelphia, PA, USA
[c] Temple University School of Pharmacy, Philadelphia, PA, USA
[d] Jayne Haines Center for Pharmacogenomics and Drug Safety, Temple University School of Pharmacy, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 17 February 2011
Accepted 21 November 2011
Available online xxxx

Keywords:
Microarray
GWAS
Least angle regression
Gastroparesis
Domperidone
Side effects

ABSTRACT

Background: Domperidone treatment for gastroparesis is associated with variable efficacy as well as the potential for side effects. DNA microarray single nucleotide polymorphism (SNP) analysis may help to elucidate the role of genetic variability on the therapeutic effectiveness and toxicity of domperidone.
Aim: The aim of this study was to identify SNPs that are associated with clinical efficacy and side effects of domperidone treatment for gastroparesis from DNA microarray experiments. This will help develop a strategy for rational selection of patients for domperidone therapy.
Methods: DNA samples extracted from the saliva of 46 patients treated with domperidone were analyzed using Affymetrix 6.0 SNP microarrays. Then least angle regression (LARS) was used to select SNPs that are related to domperidone efficacy and side effects. Decision tree based prediction models were constructed with the most correlated features selected by LARS.
Results: Using the most stable SNP selected by LARS a prediction model for side effects of domperidone achieved (95 ± 0)% true negative rate (TN) and (78 ± 11)% true positive rate (TP) in nested leave-one-out tests. For domperidone efficacy, the prediction based on five most stable SNPs achieved (85 ± 7)% TP and (61 ± 4)% TN. Five identified SNPs are related to ubiquitin mediated proteolysis, epithelial cell signaling, leukocyte, cell adhesion, and tight junction signaling pathways. Genetic polymorphisms in three genes that are related to cancer and hedgehog signaling were found to significantly correlate with efficacy of domperidone.
Conclusion: LARS was found to be a useful tool for statistical analysis of domperidone-related DNA microarray data generated from a small number of patients.

1. Introduction

Gastroparesis is a symptomatic disorder of delayed gastric emptying with no mechanical obstruction. Patients can experience early satiety, nausea, vomiting, and/or abdominal pain. Common approaches in the management of gastroparesis include dietary interventions, antiemetic agents, and prokinetic agents. Domperidone, a dopamine-2 receptor antagonist, is a potentially useful medication for treatment of gastroparesis. Unfortunately, not every patient responds to treatment with domperidone and side effects can develop during treatment. Oral domperidone, although not approved by the FDA for use in the United States, may be obtained through a FDA investigational new drug application. In our previous study, we demonstrated that genetic variability in genes involved in domperidone absorption, metabolism, and drug target interaction, influenced the therapeutic effectiveness and toxicity of domperidone [1]. This prior study assessed candidate genes that seemed appropriate to investigate with the known action and metabolism of domperidone.

DNA microarray single nucleotide polymorphism (SNP) analysis holds great promise to elucidate the role of genetic variability on the clinical efficacy of drug therapy [2,3]. This technology collects millions of data points per genome, providing an unprecedented volume of genetic information from a single experiment. However, assessment of data quality, statistical evaluation, and interpretation tools required for DNA microarray experiments far exceed the capacity of approaches accepted in traditional molecular studies [4].

Abbreviations: LARS, least angle regression analysis; SNP, single nucleotide polymorphism.

* Corresponding author. Address: Data Analytics and Biomedical Informatics Center, 303 Wachman Hall (038-24), 1805 North Broad Street, Philadelphia, PA 19122-5082, USA. Fax: +1 215 204 5082.
E-mail address: zoran@ist.temple.edu (Z. Obradovic).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.11.013

In the present study, our goal was to use a genome wide search for putative genes that may be associated with clinical efficacy and/or side effects with domperidone treatment of patients with gastroparesis. Since the number of patients is small compared to the large amount of SNPs, we applied a least angle regression method to select related SNPs. Here we report the results of LARS regression analysis for identifying association between genotypes of 46 patients generated by DNA microarray experiments using Affymetrix 6.0 SNP arrays, and efficacy/side effects of domperidone. The selection process was repeated multiple times and the most stable SNPs were used for developing decision tree based models for domperidone efficacy and side effects prediction.

2. Materials and methods

2.1. Patients

This study was approved by the Temple University Institutional Review Board. All patients who participated in the study provided written informed consent. Patients with gastroparesis treated with domperidone under an IRB approved FDA investigational new drug application were approached for participation. In this protocol, patients were typically started on domperidone 10 mg orally four times a day given 30 min before meals and at bedtime. If there was no response or a suboptimal response after 6 weeks of treatment and the patient was not experiencing any toxicities, the dose was increased to 20 mg orally four times a day. Some patients had doses increased to 30 mg orally four time a day, again depending on response and toxicity. The inclusion criteria were: (1) patients 18 years of age and older of any age, sex, ethnic group, or economic status; (2) currently receiving or having received domperidone treatment at Temple University Hospital Section of Gastroenterology as part of their regular clinical care; and (3) able to provide informed consent to participate in the study.

Patients were interviewed and information regarding age, gender, weight, height, domperidone dose, and frequency of administration were collected. Therapeutic response was enquired using the modified Clinical Patient Grading Assessment Scale (CPGAS) [5]. Patients reported gastroparesis symptoms with the following gradations: +3 = completely better; +2 = considerably better; +1 = somewhat better; 0 = no change; −1 = somewhat worse; −2 = considerably worse; −3 = very considerably worse. Patients were questioned as to if they experienced any side effects with domperidone treatment. Patients were specifically questioned about any CNS, cardiovascular, gastrointestinal, musculoskeletal, and endocrine side effects.

2.2. Sample collection and genome-wide genotyping

Patients provided saliva samples using the Oragene DNA Self-Collection Kit (DNA Genotek, Ottawa, Ontario). The saliva samples were collected and saved for subsequent analysis. DNA was extracted from saliva specimens, purified, and preserved as described earlier [1]. Genome-wide SNP genotyping was performed by DNA microarray for 48 patients. Specifically, each DNA sample was genotyped using Affymetrix Genome-wide Human 6.0 SNP microarray, which assays 906,600 single nucleotide polymorphisms. Genotyping experiments were performed at the Genomics Facility of the Fox Chase Cancer Center (Philadelphia, PA). The collected data were processed by genotype console software from Affymetrix to convert the optical results into discrete values for SNPs. The samples from two patients did not pass quality control and were excluded from further analysis. Among the remaining 46 samples, quality of data was validated by comparing the genotypes at rs1045642, rs2229109, and rs9282564 in *ABCB1* gene; genotypes were 100% consistent with the results generated by TaqMan technology [1]. Genome-wide SNP data for 46 patients is available at www.dabi.temple.edu/~zoran/data/domperidone_supplements.

2.3. Data analysis

The patient-reported response on the CPGAS was used to categorize patients as responders (scores of 2 or 3) and non-responders (scores of 1 or less). For the least angle regression related data analysis these two groups were represented as a binary variable with values 1 and −1, respectively. Similarly, a binary 1/−1 variable was used to represent patients according to the presence or absence of side effects irrespective of side effect severity. To reduce data dimensionality single nucleotide polymorphisms (SNPs) were coded a single three-valued variable (AA = 0, AB = 1, BB = 2, null = AB = 1). Initially, there were 906,600 SNPs. Data was screened to ensure the presence of polymorphisms. Constant invariant SNPs were excluded, leaving 853,943 SNPs. In the dataset (46 samples of 906,600 SNPs each) we found 237,065 missing values (null) constituting 0.6% of the total data. Only two patients had more than 0.5% of missing SNPs and even for these two patients the fraction of missing SNPs was less than 3% (see Fig. 1). In addition, more than 83% of SNPs were present in all patients and only 0.25% of SNPs were missing in more than four patients (see Table 1). The missing values were imputed as AB as to be equally distant from codes for AA and BB. The final preprocessing step consisted of normalizing data to a mean of 0 and standard deviation of 1.

Our data analysis used a Matlab R2010a implementation of the least angle regression (LARS) for selection of variables one at a time in piecewise linear forward steps [6]. Least angle regression (LARS) is a regression algorithm based on L1 norm optimization suitable for high-dimensional data containing many irrelevant and many correlated variables. Suppose we expect a response variable to be determined by a linear combination of a subset of covariates. Then the LARS algorithm provides a means of producing an estimate of which small subset of variables to include in the model as it automatically reduces the remaining regression coefficients to zero. The estimated parameters are increased in a direction equiangular to each one's correlations with the residual. This is achieved by taking into account the direction between the already selected variables and following equiangular paths in a process that is less greedy than the traditional forward feature selection.

Following previous studies [7], in control experiments the values of the response variables (efficacy and side effect) were

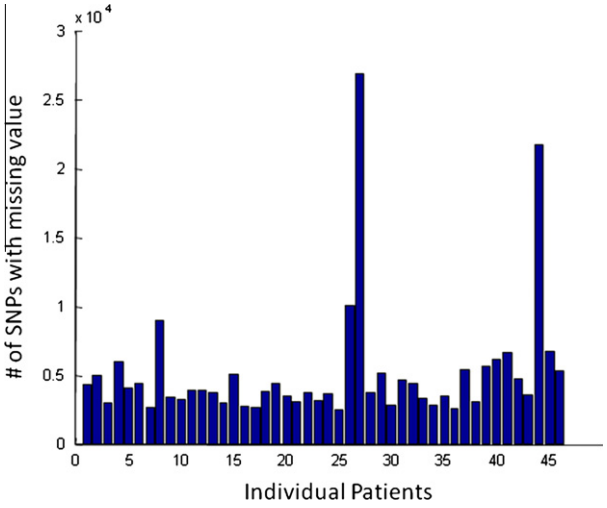

Fig. 1. The number of SNPs with missing values for each of 46 patients.

Please cite this article in press as: Wang D et al. DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis. J Biomed Inform (2011), doi:10.1016/j.jbi.2011.11.013

Table 1
The number of SNPs with up to five missing values in 46 patients.

Number of missing value for a specific SNP	0	1	2	3	4	5
Number of such SNPs	754,613	63,885	15,528	6611	3511	2247

shuffled (sorted in a random order), and LARS was used to check if certain SNPs would be identified as associated with such a random function that has the same marginal distribution as two functions of our interest. We repeated the random shuffle based experiments 1024 times for the domperidone side effects study, and 1024 times for the efficacy study (Tables 4, 5, 9, and 10). In each of these experiments the outcome value is randomly permuted while preserving the marginal distribution. The result obtained in randomly shuffled data were only trivia predictors providing evidence that accuracies obtained on observed data that were much better than random are good estimates of true generalization properties of identified models.

2.4. Gene analysis

The unique SNP identifier, chromosome location, gene name and other information for each probe set on a microarray was acquired with the NetAffx software available at Affymetrix web page (www.Affymetrix.com). SNPs that were associated with effectiveness or side effects through LARS were queried through the Affymetrix web site (https://www.affymetrix.com/user/login.jsp?-toURL=/analysis/netaffx/xmlquery.affx?netaffx=mapping) to map them to their associated gene.

2.5. Signaling pathways analysis

The Pathway Express component of Onto-Tools software was used for signaling pathway analysis of related genes (http://vortex.cs.wayne.edu/ontoexpress/). The list obtained from gene analysis of LARS result was uploaded to Onto-Tools [8] to generate a summary of related signaling pathways. This pathway analysis tool uses the genes of interest to identify associations with signaling pathways determined by previous biomedical studies.

2.6. Nested leave-one-out SNP analysis

To find the most related SNPs in terms of predicting either drug efficacy or drug side effects, we designed nested leave-one-out experiments [9,10] for decision tree learning with the four most stable SNPs picked by LARS (the data analysis procedure is outlined in Fig. 3). In each of 46 sets of experiments, we used a different patient as the test sample. On the remaining 45 samples, we repeated 45 experiments each time using LARS on different 44 samples to select 8, 16, or 24 SNPs related to the target variable and testing on the remaining 45th. Then, among all SNPs picked in these 45 experiments, we chose the top four most frequently picked SNPs to build decision tree models with one, two, three or four variables. The accuracy of these models was tested on the sample left out of any feature selection and learning process. So, for each of two prediction problems (drug efficacy or drug side effects prediction) in total we trained 2070 predictive models. The SNPs picked over 50% of times (1035 times) through the whole process were considered as the most related to the class variable (domperidone efficacy/side effects). These SNPs are shown in Tables 6 and 11.

Table 2
Prediction of side effects of domperidone based on most stable SNPs picked by LARS.

SNP ids	rs17291650	rs17291650 rs9977558	rs17291650 rs9977558 rs9632703	rs17291650 rs9977558 rs9632703 rs7637788
Initially pick 8 SNPs				
Freq	27/46	31/46	33/46	29/46
TP	0.11	0.78	0.78	0.89
TN	0.89	0.92	0.95	0.95
(TP + TN)/2	0.5	0.85	0.865	0.92
SNP ids	rs17291650	rs17291650 rs9632703	rs17291650 rs9632703 rs9977558	rs17291650 rs9632703 rs9977558 rs17702569
Initially pick 16 SNPs				
Freq	32/46	29/46	31/46	24/46
TP	0.22	0.56	0.56	0.67
TN	0.89	0.92	0.92	0.95
(TP + TN)/2	0.56	0.74	0.74	0.81
SNP ids	rs17291650	rs17291650 rs9632703	rs17291650 rs9632703 rs9977558	rs17291650 rs9632703 rs9977558 rs17702569
Initially pick 24 SNPs				
Freq	29/46	32/46	28/46	27/46
TP	0.11	0.44	0.67	0.78
TN	0.92	0.86	0.92	0.95
(TP + TN)/2	0.52	0.65	0.80	0.87

2.7. Decision tree based classification

To create a model that predicts side effects and efficacy of domperidone based on one, two, three or four SNPs identified as the most stable by LARS, we applied inductive decision trees [11]. The idea behind this approach is that the most stable features selected by LARS are more relevant to the target (domperidone efficacy and side effects) instead of noises caused by individual samples. Correlated or irrelevant features can degrade dramatically the accuracy of decision tree models and this problem is greatly reduced by relaying on features selected by LARS. The rule to select the number of SNPs is based on the accuracy achieved on training data by a decision tree model. SNPs were included incrementally according to LARS ranking until accuracy was not improved by using more SNPs, which suggests that no more SNPs are needed for prediction. We used the SNPs selected as explained in the previous section to develop an easy way to interpret decision tree prediction model.

Specifically we used the classregtree function in MatLab R2010a to develop the decision tree model and the eval function in Matlab R2010a for the model validation [12].

3. Results

3.1. Patient data

The majority of the patients were Caucasian (43/46), female (38/46). The average of age for the patients was 43 years old. In this cohort of patients, for domperidone drug efficacy study, 30 of the 46 patients were classified as responders to domperidone and 16 patients were classified as non-responders. For the study about the side effects of domperidone, nine patients out of 46 patients (19.6%) reported side effects [1]. Since the data was unbalanced, following a common practice [13], the results reported in Tables 1, 2, 5 and 6 measured accuracy as the average of true positive

Please cite this article in press as: Wang D et al. DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis. J Biomed Inform (2011), doi:10.1016/j.jbi.2011.11.013

Table 3
Accuracy of domperidone side effect prediction when using up to nine most stable SNPs as determined by LARS. Using more than four SNPs did not improve prediction accuracy.

# of SNPs	1	2	3	4	5	6	7	8	9
TP	0.11	0.44	0.67	0.78	0.78	0.78	0.78	0.67	0.67
TN	0.92	0.86	0.92	0.95	0.89	0.89	0.89	0.86	0.86
(TP + TN)/2	0.52	0.65	0.79	0.86	0.83	0.83	0.83	0.77	0.77

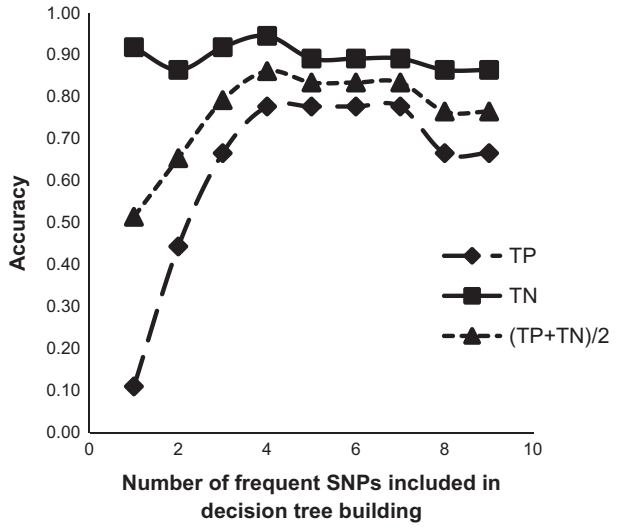

Fig. 2. Accuracy of domperidone side effects prediction depending on the number of SNPs included in decision tree building. TP, true positive rate; TN, true negative rate, (TP + TN)/2, accuracy.

(TP) rate and true negative rate (TN). More precisely, accuracy was computed as (TP + TN)/2.

3.2. Identification of SNPs related to side effects

Our predictor (Table 2) achieved (87 ± 6)% accuracy (calculated by averaging true positive rate and true negative rate) on side effect prediction. In contrast, the control experiments on data obtained by shuffling the response variable achieved less than 55% accuracy (Tables 4 and 5). We observe that the frequency for the SNPs to be selected as the most stable features may change based on how many SNPs/features were picked initially from LARS (row freq score reported in the tables).

21 SNPs were picked by the "nested leave-one-out" LARS method for domperidone side effects. Of these SNPs those picked over 50% of the time are listed in Table 6. The corresponding genes are MRPL39/ JAM2, SEMA3E, ATF1, UBE2E2/UBE2E1. Using signaling pathway analysis we found that they are related to ubiquitin mediated proteolysis, epithelial cell signaling, leukocyte, cell adhesion, and tight junction signaling pathways (Table 7). JAM2 was found to be associated with four signaling pathways while the remaining genes were found in single pathways.

3.3. Identification of SNPs related to clinical efficacy

The prediction results of clinical efficacy were less accurate. Using a single SNP (for FBXL20) our predictor achieved (73 ± 4)% accuracy as shown at Table 8. This was still much better than the accuracy of 57% obtained in control experiments on shuffled response (Table 10). The frequency for the SNPs to be selected as the most stable features may change based on how many SNPs/features were picked initially from LARS (row freq score reported in the tables). 45 SNPs were picked by the nested leave one out LARS method for the domperidone efficacy study, since the prediction

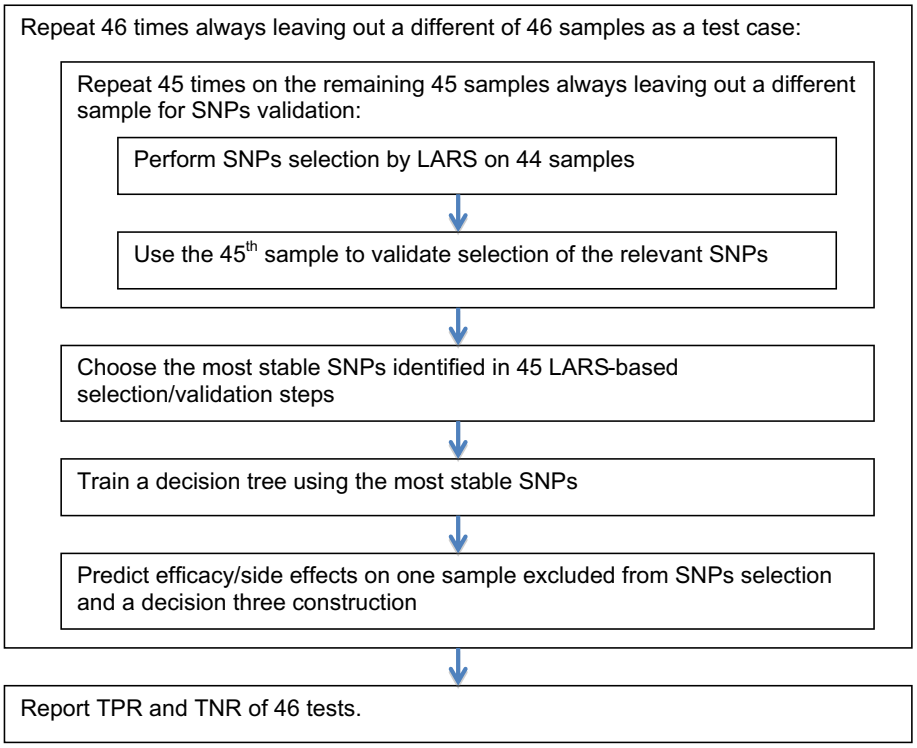

Fig. 3. Data analysis process for nested leave-one-out SNP analysis.

Please cite this article in press as: Wang D et al. DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis. J Biomed Inform (2011), doi:10.1016/j.jbi.2011.11.013

Table 4
A specific example of control on shuffled domperidone side effects data.

SNP ids	rs6990279	rs6990279 rs4130579	rs6990279 rs4130579 rs8013234	rs6990279 rs4130579 rs8013234 rs17397533
Freq	32/46	41/46	32/46	26/46
TP	0.33	0.33	0.44	0.22
TN	0.92	1	0.89	0.92
(TP + TN)/2	0.625	0.665	0.665	0.57

Table 5
Average accuracy of domperidone side effect prediction on randomly shuffled data (repeated 1024 times).

# of SNP	1	2	3	4	5
TP	0.15	0.26	0.26	0.29	0.35
TN	0.94	0.79	0.69	0.68	0.68
(TP + TN)/2	0.55	0.53	0.48	0.49	0.51

accuracy dropped significantly after selecting the first SNP. Therefore, in Table 11 we list only the two most stable SNPs. The corresponding genes (FBXL20, C7orf25/GLI3) are related to cancer, hedgehog signaling (Table 12). The gene associated with this pathway is GLI3.

4. Discussion

This study used LARS along with nested-leave-out cross validation to identify SNPs related to efficacy and side effects of domperidone. LARS is a regression algorithm for high-dimensional data. It provides a means of producing an estimate of those variables to include, as well as their coefficients for regression analysis. The estimated parameters are increased in a direction equiangular to each one's correlations with the residual. It is a very effective method for feature selection when the number of dimensions is significantly greater than the number of samples. The method has been used successfully for microarray analysis before [14]. We performed feature selection 2070 times (46×45) through nested leave-one-out experiments and the features selected most frequently (over 50% of the time in our experiments) are deemed as the relevant feature for the study.

We have previously used similar stability based feature selection successfully [15]. SNPs identified most often as related to domperidone efficacy (Table 11) have been shown to be linked to cancer and hedgehog cellular signaling. Of note, for the prediction of drug efficacy, the accuracy drops when we add more SNPs to the dataset after the first SNP. Along with it, the co-appearance of SNPs drops significantly too (the freq value in Table 8). The data suggest there are negative synergisms among the most frequently picked SNPs for domperidone efficacy study by LARS.

A different group of SNPs and associated genes have been linked with domperidone side effects (Table 6). Genetic polymorphism in *ATF1* was found to significantly correlate with side effects of domperidone. The product of this gene is activating transcription factor 1 (ATF1) that belongs to the leucine zipper family, and regulates cAMP and Ca^{2+}-responsive genes. Importantly, ATF1 is reported to regulate expression of *MUC2* gene encoding synthesis of the most abundant secretory mucin in the intestine [16]. Mucin forms a physical barrier for absorption of small molecules and bacterial toxins by the intestinal epithelium, and therefore might modulate the uptake of domperidone. Our data suggest that ubiquitin-mediated proteolysis (UBE2E2, UBE2E1) might also be associated with the side effects of domperidone.

Among other genes, *JAM2* may have functions related to efficacy/side effects of domperidone. *JAM2* (junctional adhesion molecule 2) encodes a protein that is localized in the tight junctions between high endothelial cells. Because this protein forms a physical barrier to prevent solutes and water from passing through the paracellular space [17], genetic variations in its expression or functioning may contribute to domperidone gut absorption. Also we noticed that the accuracy prediction for side effects of domperidone increased sequentially when we added more SNPs into the prediction model (Table 2), and the co-appearance of SNPs does not drop significantly (the freq value for Table 2), which is different from the results for the efficacy study (Table 8). The data suggests that there might be positive synergism among different SNPs picked for developing side effects of domperidone. A possible follow-up molecular medicine experiment may help validate these findings.

It has been reported that the method of least angle regression is sensitive to noise [18]. Therefore, we selected only the top four most stable SNPs as the most trustworthy features for prediction model building. Including more SNPs did not improve the accuracy of prediction as confirmed by results reported at Table 3 and Fig. 2.

Table 6
SNPs most frequently selected by the nested leave one out LARS prediction of domperidone side effects.

rs id	Freq	Chromosome	Gene relation	Gene symbol	Gene function
rs9977558	1880/2070	21	Upstream	MRPL39	Mitochondrial ribosomal protein L39
			Upstream	JAM2	Junctional adhesion molecule 2
rs9632703	1815/2070	7	Intron	SEMA3E	Sema domain, immunoglobulin domain (Ig), Short basic domain, secreted, (semaphorin)3E
rs17291650	1797/2070	12	CDS	ATF1	Activating transcription factor 1
rs7637788	1180/2070	3	Downstream	UBE2E2	Ubiquitin-conjugating enzyme E2E2
			Upstream	UBE2E1	Ubiquitin-conjugating enzyme E2E1

Table 7
Signaling pathways associated with the most frequent SNPs associated with drug side effects.

Signaling pathway	Number of genes	Genes found
Ubiquitin mediated proteolysis	138	UBE2E1, UBE2E2
Epithelial cell signaling in Helicobacter pytori infection	68	JAM2
Leukocyte transendothelial migration	119	JAM2
Axon guidance	129	SEMA3E
Cell adhesion molecules	134	JAM2
Tight junction	135	JAM2

Please cite this article in press as: Wang D et al. DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis. J Biomed Inform (2011), doi:10.1016/j.jbi.2011.11.013

Table 8
Prediction of efficacy of domperidone based on most stable SNPs selected by LARS.

SNPids	rs4511574	rs4511574 rs952890	rs4511574 rs952890 rs3891683	rs4511574 rs952890 rs3891683 rs10831474
Pick 8 SNP initially				
Freq	46/46	22/46	8/46	4/46
TP	0.93	0.63	0.4	0.43
TN	0.63	0.38	0.25	0.31
(TP + TN)/2	0.78	0.505	0.325	0.37
SNPids	rs4511574	rs4511574 rs952890	rs4511574 rs952890 rs4648633	rs4511574 rs952890 rs6705961 rs2098058
Pick 16 SNP initially				
Freq	41/46	19/46	9/46	3/46
TP	0.83	0.67	0.53	0.56
TN	0.56	0.56	0.31	0.31
(TP + TN)/2	0.70	0.61	0.42	0.44
SNPids	rs4511574	rs4511574 rs4648633	rs4511574 rs4648633 rs952890	rs4511574 rs4648633 rs952890 rs6705961
Pick 24 SNP initially				
Freq	41/46	11/46	8/46	4/46
TP	0.8	0.73	0.67	0.67
TN	0.63	0.56	0.56	0.38
(TP + TN)/2	0.71	0.65	0.61	0.52

Table 9
A specific example of control prediction for domperidone efficacy on shuffled data.

SNPids	rs1491719	rs1491719 rs844294	rs1491719 rs844294 rs1106039	rs1491719 rs844294 rs1106039 rs686275
Freq	30/46	24/46	10/46	11/46
TP	0.77	0.67	0.73	0.77
TN	0.25	0.44	0.44	0.5
(TP + TN)/2	0.51	0.555	0.585	0.635

Table 10
Average accuracy of domperidone efficacy prediction on randomly shuffled data (repeated 1024 times).

# of SNP	1	2	3	4	5
TP	0.79	0.70	0.66	0.68	0.64
TN	0.34	0.39	0.43	0.44	0.39
(TP+TN)/2	0.57	0.55	0.55	0.56	0.51

Nested leave-one-out experiments provided evidence that the predictors generated by decision tree models based on stable features picked by LARS achieved higher accuracy than random controls (Tables 2–5 and 8–10) [19]. The results suggest that the identified SNPs might provide useful prognostic information for drug response and toxicity in similar patients. A more comprehensive clinical study is needed to validate the prediction model.

Because of a small number of samples (46 genomes interrogated), and a high number of independent tests (more than 800,000 SNPs assayed), a standard analysis using the multiple testing adjustments will yield no statistically significant predictors. Therefore, we consider this study as exploratory data analysis without pre-specified hypotheses. We expect our studies to generate testable hypotheses rather than be able to statistically evaluate

Table 11
SNPs most frequently selected by the nested leave one out LARS for prediction of domperidone efficacy.

rs id	Freq.	Chromosome	Gene relation	Gene symbol	Gene function
rs4511574	2070/2070	17	Intron	FBXL20	F-box and leucine-rich repeat protein 20
rs952890	1562/2070	7	Downstream	C7orf25	Chromosome 7 open reading frame 25
			Upstream	GLI3	GLI family zinc finger 3

Table 12
Signaling pathways associated with the most frequent SNPs associated with domperidone efficacy.

Pathway name	Number of genes	Genes found
Basal cell carcinoma	55	GLI3
Hedgehog signaling pathway	57	GLI3
Pathways in cancer	330	GLI3

the existing hypotheses. The outcomes of our experiments should be tested prospectively once it has become possible to formulate specific hypotheses [20]. A direct method to address the clinical significance of our findings is the functional validation of identified genetic variants *in vitro* and *in vivo* [21], an ongoing project in our labs.

In summary, LARS has been successfully used to select SNPs related to drug efficacy and side effects of domperidone from microarray data. Based on the data, we have generated simple predictors which achieved high accuracy with the most frequent SNPs by nested leave-one-out cross validation. It is intriguing for biomedical scientists to design additional molecular medicine experiments to unveil the mechanisms for linkage between these SNPs and domperidone efficacy and side effects.

Acknowledgments

We are grateful to Drs. J.R. Testa and J.-M. Pei (Genomics Facility, Fox Chase Cancer Center, Philadelphia, PA) for their help in performing SNP microarray analysis. This work was supported by Jayne Haines Center for Pharmacogenomics and Drug Safety, Temple University School of Pharmacy, Seed Grant from Temple University (to HP, MJ, and EK), and in part by a grant from the Pennsylvania Department of Health (to ZO). The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

References

[1] Parkman HP, Jacobs MR, Mishra A, Hurdle JA, Sachdeva P, Gaughan JP, et al. Dig Dis Sci 2011;56:115–24.
[2] Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–6.
[3] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucl Acids Res 2001;29:308–11.
[4] Benjamini Yoav HY. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc 1995;57:289–300.
[5] Maranki JL, Lytes V, Meilahn JE, Harbison S, Friedenberg FK, Fisher RS, et al. Predictive factors for clinical improvement with Enterra gastric electric stimulation treatment for refractory gastroparesis. Dig Dis Sci 2008;53:2072–8.
[6] Efron B, Hastie T, Johanstone I, Tibshirani R. Least angle regression. Ann Stat 2004;32:407–99.

[7] Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 2005;102:2454–9.
[8] Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S. Recent additions and improvements to the Onto-Tools. Nucl Acids Res 2005;33:W762–5.
[9] McLachlan GJ, Do KA, Ambroise C. Analyzing microarray gene expression data. Hoboken, New Jersey: Wiley-Interscience; 2004.
[10] Yang AJ, Song XY. Bayesian variable selection for disease classification using gene expression data. Bioinformatics 2010;26:215–22.
[11] Yang T. Computational verb decision trees. Int J Comput Cogn 2006;4:34–46.
[12] Breiman L, Friedman R, Olshen R, Stone C. Classification and regression trees. Boca Raton, FL: CRC Press; 1984.
[13] Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.
[14] Segal MR, Dahlquist KD, Conklin BR. Regression approaches for microarray data analysis. J Comput Biol 2003;10:961–80.
[15] Radivojac P, Obradovic Z, Dunker AK, Vucetic S. Feature selection filters based on permutation test. Proceedings of the 15th European conference on machine learning; 2004. p. 334–46.
[16] Hokari R, Lee H, Crawley SC, Yang SC, Gum Jr JR, Miura S, et al. MUC2 intestinal mucin via CREB/ATF1. Am J Physiol Gastrointest Liver Physiol 2005;289:G949–59.
[17] Bazzoni G. The JAM family of junctional adhesion molecules. Curr Opin Cell Biol 2003;15:525–30.
[18] Weisberg S. Discussion on "least angle regression". Ann Stat 2004;32:490–4.
[19] Perner P. Improving the accuracy of decision tree induction by feature preselection. Appl Artif Intell 2001;15:747–60.
[20] Sullivan PF. Spurious genetic associations. Biol Psychiatry 2007;61:1121–6.
[21] Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. N Engl J Med 2011;364:1144–53.

Please cite this article in press as: Wang D et al. DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis. J Biomed Inform (2011), doi:10.1016/j.jbi.2011.11.013