Learning the Genetic Causes for Schizophrenia through Copy-Number Variations

Xin Gao, Shan Liu

1. Introduction

Schizophrenia is a mental disorder that causes lifelong disability. The most common symptoms are auditory hallucinations, paranoia, and delusions. At least 1% of Americans have this illness. [1] Factors that cause schizophrenia are both genetic and non-genetic. Research has shown that gene plays a significant role in schizophrenia through studies on twins. However, the genetic aberrations associated with schizophrenia have been hard to pin down. A large number of genome scan projects have reported linkage in three chromosomal regions, but none of these regions has produced strong support for linkage in the majority of the projects. [2] Unlike some diseases that can be caused by a single defective gene such as Huntington's disease, schizophrenia is much more complex and elusive. [1, 3]

Instead of a single disease-causing gene, research has shown promises in DNA copy-number variation (CNV). The old belief was that a person has two copies of each gene in a genome (one from the father and one from the mother). However, recent discoveries revealed that large segments of DNA can vary in copy-number (i.e. some people are missing a gene—deletion, and some have three or more copies—duplication). Scientists believe DNA copy-number may be a common underlying factor in genetic disease. [3-5]

2. Objective

To evaluate DNA copy number variations in schizophrenia patients to look for insights into genetic risk factors for this disease.

3. Preprocessing of Data

We worked on a database which consists of 2,552 cases of schizophrenia patients and 3,532 controls (6,084 in total), each with copy-number intensity measured by Affymetrix technology using 600,470 probes for each person on 22 chromosomes. Each probe is identified with a unique index number that is matched to a unique chromosome location. The data contains all the CNV information for each person identified by these probes (locations where the CN is other than 2).

Due to the potential inaccuracy of copy-number measurement, we first ignored the reported variations when the length of the region is smaller than 4 probes. Then we combined some adjacent variations into one larger variation region if those adjacent variations are really close to each other, such that the normal region in between can be treated as the inaccuracy of measurement. We labeled normal copy-number (2) as 0, duplication (3 or 4) as 1, and deletion (1 or 0) as -1.

In Matlab, we designed a feature matrix X as a 6,084 by 600,445 sparse matrix with each column containing the copy-number variation label (0, 1 or -1) at each probe position from 1 to 600,445. We also made a vector Y of 6,084 dimensions with 1 indicating for case and 0 for control.

4. Result

4.1 Naïve Bayes

First, we implemented the Na ïve Bayes classifier on a modified data set. This problem is similar to a spam filter in text classification. We modified the input data in the following way. The input x_{ij}'s are discrete valued (if a patient i has CNV in the j^{th} position probe, we set $x_{ij} = 1$; otherwise $x_{ij}= 0$. i =1, …, m=6,084; j =1,…, n=600,445). The y_i's are binary indicator for case or control. We assumed the CNVs are conditionally independent given case or control. This assumption resolved the scarcity of training data problem, since the dimension of our feature is much greater than the number of data (n >> m). We implemented Na ïve Bayes using Laplace smoothing and 10-fold cross-validation. The estimated generalization error (averaged over 10) is 0.0222. The averaged accuracy is 97.78%. We identified the top 100 probe locations that may be particularly indicative of a patient having schizophrenia in chromosome 11, 3, 14, and 20. These locations are listed by their ranking order in Table 1.

Table 1. Top Chromosome Locations Indicating Schizophrenia (Na ïve Bayes)

Probe Number	**Chromosome Number**	**Location**
394,784 to 394,812	11	55,127,597 to 55,209,499
116,693 to 116,699	3	89,485,137 to 89,499,861
464,654 to 464,693	14	18,298,712 to 18,372,086
484,458 to 484,476	14	105,421,439 to 105,775,821
574,565 to 574,568	20	26,235,048 to 26,241,985

4.2 Support Vector Machine

Second, we ran LIBLINEAR SVM on the processed data as described above, and used 10-fold cross validation. The averaged accuracy on cross validation is 99.68%, which is very high and slightly better than the Na ïve Bayes' method. By plotting the magnitude of model parameter w^T, we have Figure 1 below. The locations that correspond to the largest positive magnitudes of w are on chromosome 11, 13 and 3. The locations that correspond to the largest negative magnitudes are around positions of chromosome 19. The positive magnitudes show that copy-number duplication (3 or 4) around the region tends to cause schizophrenia, and negative magnitudes show that copy-number deletion (1 or 0) around the region tends to have similar effects.

We can also observe that there are several vertical lines of blue dots in some regions, which may indicate that many probes around the region are redundant in predicting schizophrenia.

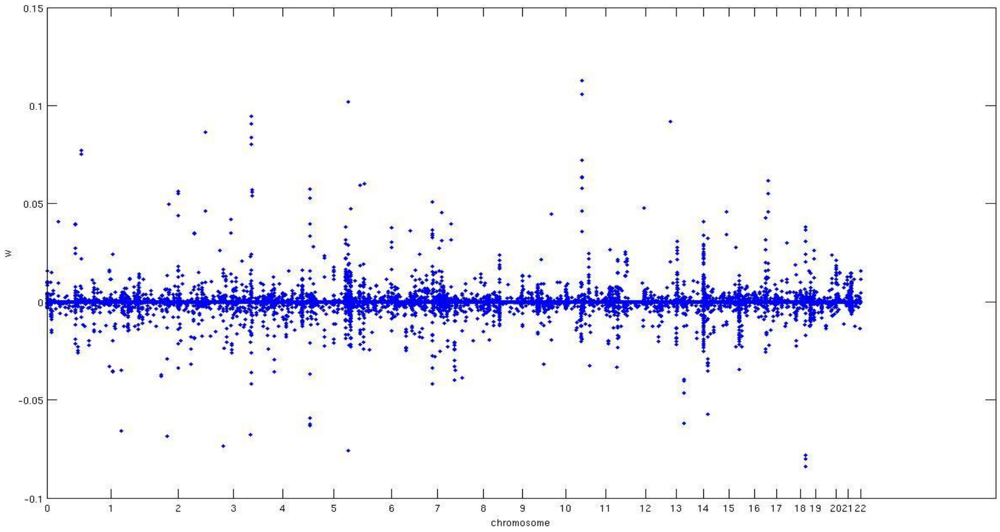


Figure 1. Magnitude of w^T vs. Chromosome Locations in the SVM Algorithm

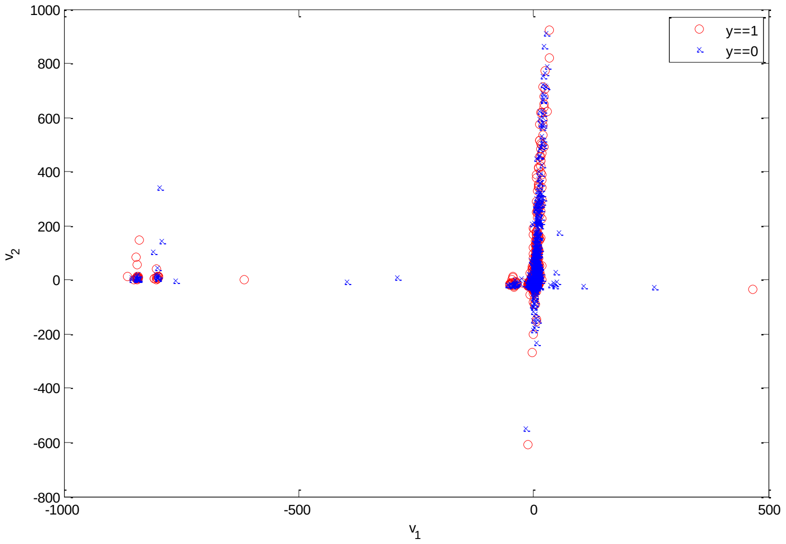
We identified the top 100 probe locations that may be particularly indicative of a patient having schizophrenia in chromosome 11, 13, 3, 4, 1, and 6. These locations are listed by their ranking order in Table 2.

Table 2. Top Chromosome Locations Indicating Schizophrenia (SVM)

Probe Number	**Chromosome Number**	**Location**
394,783 to 394,812	11	55,127,597 to 55,209,499
459,855 to 459,858	13	97,328,242 to 97,330,758
116,693 to 116,699	3	89,485,137 to 89,499,861
151,060 to 151,219	4	69,064,675 to 70,236,428
21,243 to 25,391	1	103,907,158 to 147,427,061
234,277 to 234,280	6	81,341,226 to 81,346,109
193,818 to 193,831	5	97,073,409 to 97,098,575
89,712 to 89,715	2	212,895,279 to 212,899,634
440,378 to 440,381	12	127,796,466 to 127,798,966
8,261 to 8,265	1	40,794,563 to 40,799,336
501,754 to 532,512	15	95,616,714 to 51,519,463
372,408 to 372,411	10	90,934,196 to 90,935,788
135,758 to 135,762	3	190,847,118 to 190,849,457
268,243 to 268,248	7	52,701,022 to 52,711,270

4.3 Principal Component Analysis

The input feature (probe locations) has more than 600,000 dimensions. We examined how well the data can be represented using a reduced number of dimensions by applying principal component analysis. First we plotted the data mapped onto the two most significant principal component space. The result showed a large overlap between the cases and controls and they cannot be separated at all.



We then tried to compress the data using additional principal components vectors, and then run SVM on the data with reduced dimensions. The accuracy of the SVM for 100 principal components vectors is 54.6%. This result showed the case and control cannot be separated using a small number of principal components, which implied that the most significant variances between the data do not come from the differences between case and control. In another words, the positions of copy-number variation that most likely causing schizophrenia would only cause a tiny portion of all the variations between people. The copy-number values that cause the most significant variations between people may come from their racial and demographic differences.

5. Discussion

We evaluated DNA copy number variations in a schizophrenia patient database to look for insights into genetic risk factors for this disease using supervised machine learning methods, in particular Na ïve Bayes and Support Vector Machine. We identified two regions on chromosome 11 (55,127,597 to 55,209,499), and chromosome 3 (89,485,137 to 89,499,861) with CNV that are most likely to be associated with schizophrenia.

Our training errors from both Na ïve Bayes and SVM are low, which relieved some of our initial concerns on the assumption of CNV being conditionally independent given case or control. Given the large input feature set (6,084 by 600,445), we found most of the locations with CNV are irrelevant to the disease. The small numbers of DNA specimens, which may not be enough for our high-dimensional feature was resolved.

Besides trying to accurately predict whether a person has schizophrenia by applying a machine learning algorithm on his/her CNV information, it is more meaningful to understand which positions on the DNA actually cause the schizophrenia. Using Naïve Bayes and SVM, we identified some of the most significant chromosome locations in prediction. To check for internal validity of the study, we investigated the differences in the identified CNV locations between the two methods. The results showed concordance in the top regions being locations on chromosome 11 and 3. SVM is able to identify more diverse locations than Naïve Bayes, which is not surprising due to the extra information on duplication or deletion utilized in SVM, but lacked in Naïve Bayes. To check for external validity of the study, we compared our results with literature in the field. [6, 7] At least one genome-wide association study of schizophrenia indentified two locations on chromosome 11. [5] We could not check whether the exact locations matched our results due to different labeling method in the paper. The next step is to seek help from a genetics expert to interpret our findings.

One major limitation of our study is generalizability of the results. The excellent testing accuracy from our study may not achieve the same prediction power when applied to other data sets. There are many potential confounders such as characteristics of the case and control in this study and the testing environment. We could not guarantee generalizability until testing our prediction algorithm on a new data set.

6. Acknowledgement

We thank Dr. Douglas Levinson for letting us use the database, and for providing the methodologies for preprocessing the CNV data.

7. Reference

[1] National Institute of Mental Health website, <http://www.nimh.nih.gov/health/publications/schizophrenia/what-is-schizophrenia.shtml>

[2] Lewis, et. al. “Genome Scan Meta-Analysis of Schizophrenia and Bipolar Disorder, Part II: Schizophrenia.” *Am. J. Hum. Genet. 73: 34-48, 2003*

[3] “What is copy number variation?” website,< http://www.gene-quantification.de/cnv-faq.pdf>

[4] Wilson, G.M., Stephane Flibotte, Vikramjit Chopra, Brianna L. Melnyk, William G. Honer and Robert A. Holt, “DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling.” *Human Molecular Genetics*, 2006, Vol. 15, No. 5

[5] Conrad, et. al. “Origins and functional impact of copy number variation in the human genome.” *Nature 464, 704-712 (1 April 2010)*

[6] The International Schizophrenia Consortium, “Rare chromosomal deletions and duplications increase risk of schizophrenia.” *Nature 455, 237-241 (11 September 2008)*

[7] Kirov, et. al. “Support for the involvement of large copy number variants in the pathogenesis of schizophrenia.” *Human Molecular Genetics, 2009, Vol. 18, No. 8*