# **Multi-trait analysis of rare-variant association summary statistics of autism, schizophrenia, and bipolar disorder**

**Pengjun Guo**

**RDFZ CHAOYANG BRANCH SCHOOL**

**Abstract**

Mental illness has seriously affected people's life. Autism (ASD), Schizophrenia (SCZ) and bipolar disorder (BP) pose a significant challenge to patients’ mental well-being and remarkably reduce patient’s life expectancy. Recently large-scale exome sequencing studies of these three mental disorders reveal a strong association between the effect sizes and rare variants, especially rare protein-truncating variants (PTVs; for example, nonsense, frameshift and essential splice-site mutations). In this study, we use Fisher’s exact test to compute p-value of rare PTVs in the case-control cohort of each disorder respectively, and carry out meta-analysis to detect the association between mental disorders and rare PTVs, compared to the results of meta-analysis of three disorders using a combination of genetic mutations including rare PTVs, rare missense variants and *de novo* mutations. We illustrate our findings in Manhattan plots and find even though rare PTVs play a great genetic role on mental disorder, analyzing PTVs exclusively does not account for much of mental disorders compared with analysis of a combination of diverse genetic mutations.

Keywords: autism, schizophrenia, bipolar disorder, exome sequencing, rare variant analysis, meta-analysis

**Table of Contents**

Abstract **2**

Introduction **4**

Methods and results **5**

Data Collections 5  
Gene-based analysis approach for each mental disorder 6  
Combining three mental disorders data in meta-analysis 7  
Generating Plots 7

Conclusions **10**

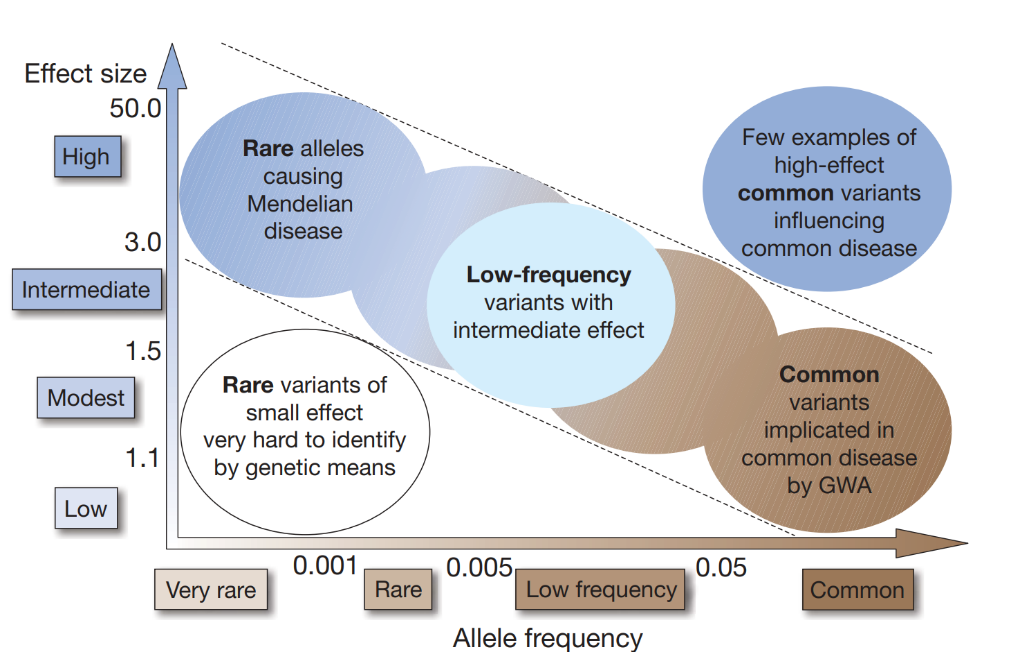
References **11**

Acknowledgement **13**

**Introduction**

In the recent century, mental illness has seriously distressed the life of a great portion of the population[1][5]. The years lived with disability, of the mental disease in 2017 was nearly 122,746,000 years[10], meaning a great gap in life expectancy. The three diseases concerned in this research are widely found across the world: autism spectrum disorder (ASD) which causes social disorder and mental retardation [2]; Schizophrenia which leads to hallucinations and delusions [3]; and bipolar disorder which results in a great difficulty for the patients to be in charge of their own emotions [4]. These three diseases severely devastated the life quality for the patients and also disrupted the life their family. Besides, their incidence rates are strikingly high: bipolar disorder occurs among 1-2% of the population[4], for schizophrenia, it affects 0.5% of the human beings[3], and 1/54 of the kids suffers from autism[2], and therefore these three diseases keep raising concerns of the scientist.



In the previous studies, researchers have unearthed the association of the symptom and the genetic features in the patients[1][5-8]. Their results show there are connections between several significant genes and the typical symptoms of the three diseases. Later research also showed important connections between the rare mutations and effect sizes. In general, the degree of effect size increases when the mutation is rare. Due to the significant impact on the occurrence of disease which the rare mutations have caused, it has been clearly shown there is a great importance of analyzing the rare gene mutation in both the case and control group.

**Figure 1** Schematic plot for the relationship between variant frequency and effect size. Image referenced from Manoli et al

Recent studies provide large-scale exome sequencing data of these three mental disorders and identify some risk genes of individual disorders[5-7]. To boost statistical power over single-disorder analysis using the exome sequencing data from all three mental disorders which have high genetic correlations can increase the sample size, making analyzing the decisive exome factor possible[7][8].

Despite these considerable advantages, we still possibly cannot get the ideal result without the appropriate type of genetic variants data source. Associated with previous evidence mentioned in the report, the rare PTVs could be important factors affecting the occurrence of some mental diseases.

Analyzing rare PTVs in the combined large-scale exome sequencing data increase the power to seek the risk genes affecting the occurrences and severity of the diseases possibly.

**Methods and results**

**Data collections**

We used rare-variant association summary statistics of three mental disorders, Autism, Bipolar, and Schizophrenia, publicly shared by Broad Institute. By downloading and importing the .csv files, we imported the data frame into the R.

Besides, we used tidyverse (including ggplot2, purrr, tibble, dplyr, tidyr, stringr, readr, forcats), qqman, skimr, metap, data.table for data processing.

In order to make the results clear and readable, we downloaded gene annotation (Homo\_sapiens.GRCh38.104.gff3) from Ensembl official website. We got the specific information about genes, and joined the data and the gene information by using *left\_join()* from the *tidyverse* pack. Because the generation of Manhattan plot needs the chromosome on which the gene is located, and the specific gene name, we extracted this information from the data.

The result shows the Hugo gene name and other useful information are combined with PTV data and ENFI sequences.

**Gene-based analysis approach for each mental disorder**

To find the remarkable mutations in the data, it’s important to calculate and quantify the statistical features(including P-values and odds ratios) of each gene. Facing extensive sample sizes and small statistical units matching with each gene, especially when the data is divided into two different groups (control and case), Fisher’s exact test is suitable. Named after Ronald Fisher, this test is especially suitable for categorical data, i.e., data that is classified into two different groups. It can calculate how significantly a statistical test deviates from the null hypothesis in an exact way.

Diagram

Description automatically generated

**Figure 2** Flow chart for data analysis to find significant genes.

As an example, a data set is divided to 2\*2 different cases, and the P-value will be calculated for each gene like this in the following equation (eq. 1):

|  |  |  |
| --- | --- | --- |
|  | Number of Cases | Number of Controls |
| Carrying at least one PTVs | a | b |
| Not carrying any PTVs | c | d |

|  |  |  |
| --- | --- | --- |
|  |  | (eq. 1) |

In practice, in the analysis of each mental disorder, we counted the number of cases/controls carrying at lease one rare PTVs and the number of cases/controls not carrying any rare PTVs for each gene. Then we inputted the counts to the function of the Fisher’s exact test, which is implemented in R packages and got the returning p-values and odds ratios for each gene. After calculating the analysis of the PTVs for each of the three disease, we joined the three datasets together using full\_join() from dplyr for the next step calculation of meta-analysis values. Figure 2 pictures this procedure. Finally we visualized the results in Manhattan plot for each of the analysis using the functionality offered in *qqman* package.

**Combining three mental disorders data in meta-analysis**

We meta-analyzed the p values of above separated Fisher exact test for each disorder using weighted Z-scores. The function *sumz* from *metap* package is used to combine p values from multiple sources. It requires the original p-values and weights of them. The equation for calculating Z-scores is shown below, wherein the *k* means the number of the studies and *w* means weights.

|  |  |  |
| --- | --- | --- |
|  |  | (eq. 2) |

To calculate the weights of each disease, we used the square root of the sample size (including both the case and control groups). For the data we did meta-analysis, the sample size of each disease and their square root is shown in the table below,. When calculating the statistics with only the PTVs for each disease, the weights have to be re-evaluated using the selected case and control group.

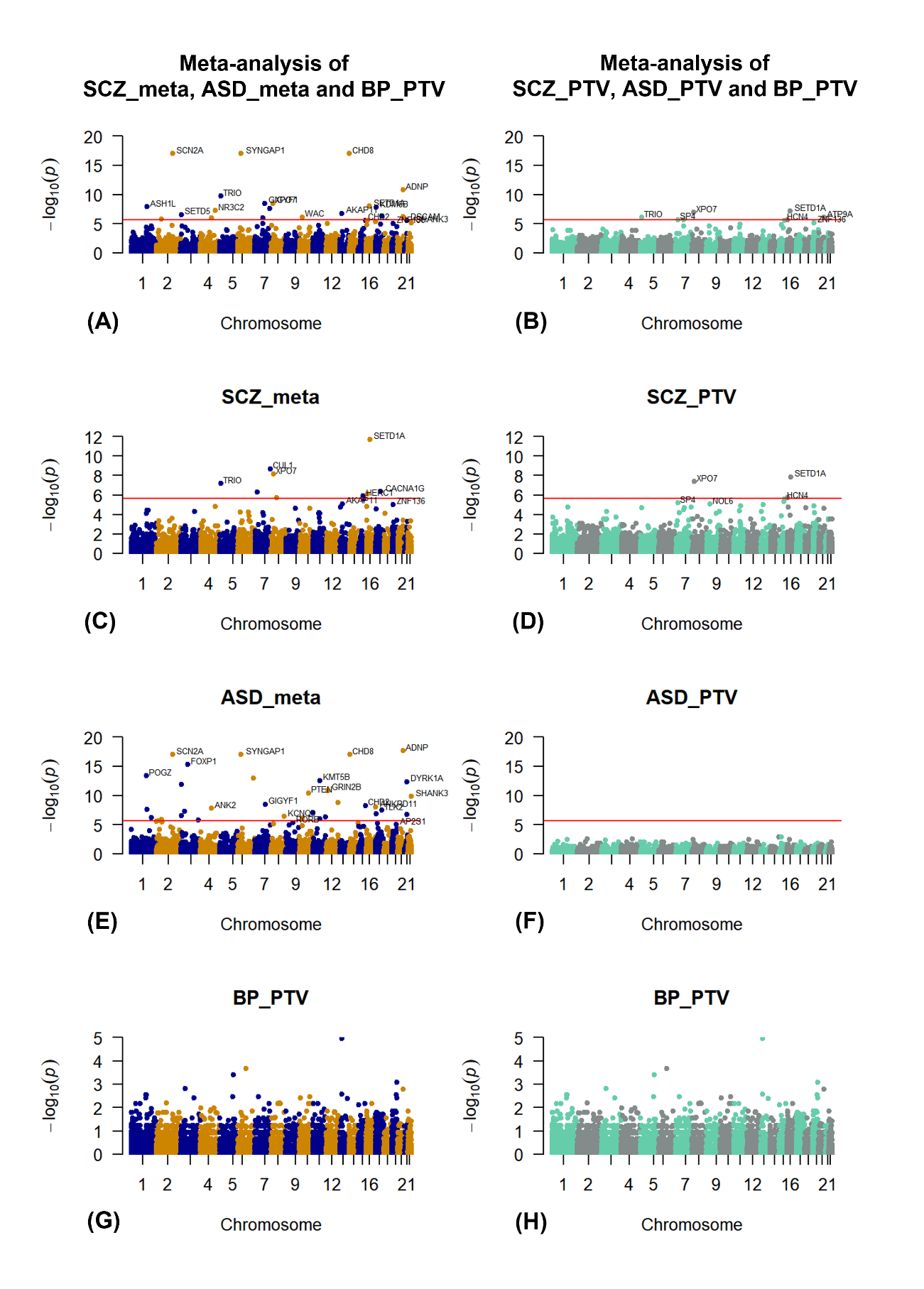
|  |  |  |  |
| --- | --- | --- | --- |
|  | SCZ | ASD | BP |
| Sample Size | 121579 | 29758 | 28344 |
| Weight | 348.67 | 160.49 | 168.36 |

**Generating plot**

Using *mutate( )* from *tidyverse*, we generated a column of combined p-value for further exploration.

Manhattan plot is a type of [scatter plot](https://www.r-graph-gallery.com/scatterplot/) that specializes in showing statistics of genome wide association study. It generates points representing a mutation/gene in the selection area corresponding to its chromosome and we can choose to highlight the significant mutations/genes (need to be input from a list) with special color, and also label them with the name, making the critical mutations/genes possible to be noticed.

We used *QQman* package from the R library. It’s based on *ggplot2* and requires only a few inputs including the p-value of genes, the chromosome on which the gene is located, the location bp, and its name. Besides, we should generate a list of concerned genes which is below a very small p-value, based on our meta-analysis, and in this work we chose 0.00001 and highlight them in each plot to show the association between the meta results and the outcome of each single-disorder.

Therefore, we generated images of the three diseases, and a composite image of the data processed through the *sumz* described above in the right panel in Figure 3 for the results based on rare PTVs. To be able to detect the genetic differences between PTVs and a combination of diverse genetic mutations (including PTVs, missenses and de novo mutations), we did the same meta-analysis for the three diseases, but with the p values published in their studies which were calculated based on PTVs, missenses and de novo mutations. The result is plotted in the right part in Figure 3.

**Figure 3** Manhattan Plot for the P-value. Left panel (blue and orange plots) shows the p-value with full considerations of all disease cases for **(A)** the meta-analysis of all three diseases **(B)** the SCZ meta-analysis **(C)** the ASD meta-analysis and **(D)** the bipolar PTV; and the right panel (aquamarine and azure colored plot) shows with only the PTV considered in the P-value for **(E)** the meta-analysis of all three diseases (F) the SCZ PTV (G) the ASD PTV and (H) the bipolar PTV. We remarked important genes whose P-value is less than 0.00001. The red line: P value = 2.14 \* 10-6

In the PTV only analysis (right panel in Figure 3), it’s obvious that schizophrenia plays a leading role in the final result of this analysis, not only because of its huge sample size, but also because PTV is the leading factor in the occurrence of schizophrenia itself. In contrast, the results of autism are quite insignificant. It is very likely that the gene mutations of the *de novo* cases are the most important factor in autism. For bipolar disorder, the result shows no significant risk genes are identified, which is closely related to the small sample size and lack of statistical power.

**Conclusions**

In this work, we used meta-analysis to search for links between mental disorder and gene variants. We also did the analysis using only PTV data for the three diseases and found PTV only data can not fully explain the significant genes. I learned important analysis methods such as Fisher’s exact test and Z-score, and for the next step, we will study carefully which significant genes can be explained by PTV only data and find the profound associations between the diseases and the gene mutations.

**References**

[1] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753. https://doi.org/10.1038/nature08494

[2] Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J. Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., Brand, H., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell, 180(3), 568–584.e23. https://doi.org/10.1016/j.cell.2019.12.036

[3] Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., ... & Daly, M. J. (2020). Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. *MedRxiv*.

[4] Palmer, D. S., Howrigan, D. P., Chapman, S. B., Adolfsson, R., Bass, N., Blackwood, D., ... & Neale, B. M. (2021). Exome sequencing in bipolar disorder reveals shared risk gene AKAP11 with schizophrenia. *medRxiv*.

[5] Sullivan, P. F., Daly, M. J., & O'donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, *13*(8), 537-551.

[6] Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., ... & Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics*, *50*(2), 229-237.

[7] Wu, Y., Cao, H., Baranova, A., Huang, H., Li, S., Cai, L., ... & Wang, Q. (2020). Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Translational psychiatry*, *10*(1), 1-11.

|  |
| --- |
| [8]Luo, L., Shen, J., Zhang, H., Chhibber, A., Mehrotra, D. V., & Tang, Z. Z. (2020). Multi-trait analysis of rare-variant association summary statistics using MTAR. Nature communications, 11(1), 1-11. |

[9] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., ... & Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. Nature genetics, 47(11), 1236-1241.

[10] James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., ... & Briggs, A. M. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 392(10159), 1789-1858.

[11] Fisher, R. A. (1922). "On the interpretation of χ2 from contingency tables, and the calculation of P". Journal of the Royal Statistical Society. 85 (1): 87–94. doi:10.2307/2340521. JSTOR 2340521.

[12] Fisher, R.A. (1954). Statistical Methods for Research Workers. Oliver and Boyd. ISBN 0-05-002170-2.

[13] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

[14] Michael Dewey (2017). metap: meta-analysis of significance values. R package version 0.8.

[15] Becker BJ (1994). “Combining significance levels.” In Cooper H, Hedges LV (eds.), A handbook of research synthesis, 215–230. Russell Sage, New York.

[16] Rosenthal R (1978). “Combining results of independent studies.” Psychological Bulletin, 85, 185–193.

[17] Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RMJ (1949). The American soldier, vol 1: Adjustment during army life. Princeton University Press, Princeton.

[18] Zaykin DV (2011). “Optimally weighted Z–test is a powerful method for combining probabilities in meta–analysis.” Journal of Evolutionary Biology, 24, 1836–1841.

**Acknowledgments**

I have been in love with doing data analysis to find out connections between events and then making graphs and figures after data analysis. In this summer around June 2021, I had a great opportunity to study scientific research on doing data analysis with Dr. Jinjie Duan from Aarhus University using one of the most important datasets of European mental disorder sequencing data. I had some previous experience in computing programming, and then I quickly found R language to be interesting and powerful in doing data analysis and plotting.

Over the summer, I learned how to use R and Rstudio to read large amount of data. I found easy R packages and commands to look into the data frames, summarize data features and compare between datasets. Under the guidance from Dr. Duan, I was able to code my own program to properly clean up the large sequencing dataset and finally calculate Fisher’s exact test. Then I also learned how to plot Manhattan plot to show the significance of genes.

During the time when I was writing the paper, Dr. Duan gave me a lot of advise on how to write papers in a formal and scientific way, like how to make good figures and what I should describe in my report. Even if I am still just beginning on the study of research and writing papers, I feel I have learned a great deal. Thanks to Dr. Duan sincerely!

We have a plan of going forward with the R analysis and I will study the profound associations between the diseases and the genes. I would also want to thank my family and my school teachers for their help, especially when I have to focus my time on doing projects and writing reports.

Michael Michael

Michael

Michael

Michael

Michael

Michael

Michael

Michael