A novel copy number variants kernel association test with application to autism spectrum disorders studies

Xiang Zhan[*], Santhosh Girirajan[†], Ni Zhao[*], Michael C. Wu[*], and Debashis Ghosh[‡]

Abstract — **Copy number variants (CNVs) have been implicated in a variety of neurodevelopmental disorders, including autism spectrum disorders, intellectual disability and schizophrenia. Recent advances in high-throughput genomic technologies have enabled rapid discovery of many genetic variants including CNVs. As a result, there is increasing interest in studying the role of CNVs in the etiology of many complex diseases. Despite the availability of an unprecedented wealth of CNV data, methods for testing association between CNVs and disease-related traits are still under-developed due to the low prevalence and complicated multi-scale features of CNVs. We propose a novel CNV kernel association test (CKAT) in this paper. To address the low prevalence, CNVs are first grouped into CNV regions (CNVR). Then, taking into account the multi-scale features of CNVs, we first design a single-CNV kernel which summarizes the similarity between two CNVs, and next aggregate the single-CNV kernel to a CNVR kernel which summarizes the similarity between two CNVRs. Finally, association between CNVR and disease-related traits is assessed by comparing the kernel-based similarity with the similarity in the trait using a score test for variance components in a random effect model. We illustrate the proposed CKAT using simulations and show that CKAT is more powerful than existing methods, while always being able to control the type I error.**

Keyword: *Autism spectrum disorders; Copy number variants; Kernel association test.*

1 Introduction

Copy number variants (CNVs) are deletions and duplications of DNA segments, which have been implicated in a variety of neurodevelopmental disorders, including autism spectrum disorders, intellectual disability and schizophrenia [Girirajan *et al.*, 2011, Sebat *et al.*, 2007]. Understanding the relationship between CNVs and these diseases can contribute important new insights into the underlying genetics etiology and may further lead to effective means in prevention and treatments. A useful means to study the complex relationship between CNVs and human health conditions is through association studies [McCarroll and Altshuler, 2007]. A powerful mode of genetic association analysis is collapsing methods, which study the association between a group of genetic variants and traits. Such methods have been widely used in SNPs association analysis [Wu *et al.*, 2010] and rare variants association analysis [Wu *et al.*, 2011]. However, these collapsing methods cannot be directly applied to CNV association analysis due to its unique features (low prevalence, multi-scale features, phenotypic heterogeneity, etc.). New methods are necessary.

In this paper, to utilize both type and size information in a CNV, we propose the CNV kernel association test (CKAT). We first design a single-CNV kernel which accounts for the multi-scale features of a CNV. Intuitively speaking, the kernel is used as a similarity measure between two CNVs. To overcome low prevalence of CNVs, we pool CNVs together to form CNV regions (CNVRs) and carefully aggregate the single-CNV kernel to a CNVR kernel which describes the similarity between two CNVRs. Compared with a single CNV, more samples are likely to have CNVs detected in a region which can makes the CNVR kernel more informative. Finally, association between CNVR and the trait is tested by comparing the similarity in CNVRs (captured by the CNVR kernel) to that in the trait. In particular, the trait we considered in this paper is disease status. If the CNVR similarity between two patients (or two healthy controls) is consistently higher than the CNVR similarity between one patient and one healthy control, then it may suggest existence of association between the CNVR and the disease risk. Statistically speaking, the similarity comparison is evaluated in a logistic random effect model and the p-value for the association test is also analytically calculated via a variance component score test in the

[*] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.

[†] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA.

[‡] Department of Biostatistics and Informatics, University of Colorado, Aurora, CO 80045, USA.

logistic regression framework. Using extensive simulation studies, we demonstrate that the proposed CKAT always has correct type I error rate and high power in a wide range of settings.

2 Methodology

2.1 Kernel-based association analysis

Notationally, let y_i be the disease status with $y_i = 1$ denoting the disease group and $y_i = 0$ denoting the control group, where $i = 1, \ldots, n$ are subjects. Let $R_i = (X_1^i, \ldots X_{p_i}^i)$ be the CNVs within the CNVR from subject i. The following logistic regression model is used to relate the disease risk to CNVs

$$logit[Pr(y_1 = 1)] = \beta_0 + f(R_i), \tag{2.1}$$

where $f(\cdot)$ is a centered unknown function in the space spanned by the CNVR kernel $k_R(\cdot, \cdot)$. Based on (2.1), the hypothesis of no association between disease and CNVs can be tested as $H_0 : f(\cdot) = 0$. To test $H_0 : f(\cdot) = 0$, one way is to treat the CNV effect vector $F = (f(R_1), \ldots, f(R_n))'$ as a random effect vector which is distributed as $N(0, \tau K)$, where $\tau \geq 0$ and K is the $n \times n$ CNVR kernel matrix. It has been shown that testing $H_0 : f(\cdot) = 0$ is equivalent to testing $H_0 : \tau = 0$ in the logistic random effect model, and moreover, τ is a variance component parameter in the logistic random effect model, which can be tested using a likelihood-based score test [Wu *et al.*, 2010, 2011]. The remaining task is to design appropriate CNV kernels for association analysis.

2.2 Single CNV kernel

Let $X = (X^{(1)}, X^{(2)})$ denote a CNV, where $X^{(1)}$ is length/size of the CNV which equals to end position minus start position, and $X^{(2)}$ is the type information of the CNV, taking values 1 (deletion CNV) and 3 (duplication CNV). Considering two arbitrary CNVs X_1 and X_2, we define the kernel function between two CNVs as

$$k(X_1, X_2) = \exp\left\{ -\frac{\left(X_1^{(1)} - X_2^{(1)}\right)^2}{\rho} \right\} \times \left[\frac{I(X_1^{(2)} = X_2^{(2)}) + 1}{2} \right] \tag{2.2}$$

As mentioned before, $k(X_1, X_2)$ is used to describe the similarity between X_1 and X_2. As defined in (2.2), both size and type of CNV contribute to the kernel similarity measure. The size of a CNV, $X^{(1)}$, can be in the order of thousands of base pairs. Even the size difference $(X_1^{(1)} - X_2^{(1)})$ can take a wide range of values. Compared with the second term, the first term can be really small. Hence the shape parameter $\rho > 0$ can balance the contribution of size and type in describing CNV similarities. The selection of ρ depends on one's belief on the data. If one thinks that CNV size is more important in determining the disease, then a larger ρ should be used otherwise the first term would be dominated by the second term. On the other hand, if one thinks CNV type are more likely to be disease-related, then smaller ρ's are preferred. In practice, without such background knowledge, we chose the ρ such that contributions of size and type are comparable.

2.3 CNV region kernel

Kernel-based association analysis are often conducted in the variant-set level rather than single variant level [Wu *et al.*, 2010, 2011]. Hence, kernel-based CNV association analysis should focus on CNVRs with multiple CNVs instead of a single CNV [Tzeng *et al.*, 2015]. Therefore, we propose a CNVR kernel which describes the sample pairwise similarity between all CNVs in a CNVR. Suppose the CNVR is pre-fixed, and let $R_i = (X_1^i, \ldots X_{p_i}^i)$ be the CNV profiles of sample i in that region, where $X_1^i, \ldots X_{p_i}^i$ are CNVs sorted according to their positions and p_i is the number of CNVs in sample i in the region. Similarly, we have a corresponding CNVs series $R_j = (X_1^j, \ldots X_{p_j}^j)$ for another sample j. Then the CNVR kernel function between sample i and j in this particular region is defined as

$$k_R(R_i, R_j) = \max_{l=0,1,\ldots,p_i-p_j} \sum_{t=1}^{p_j} k(X_{t+l}^i, X_t^j); \quad if \;\; p_i \geq p_j > 0, \tag{2.3}$$

where $k(\cdot,\cdot)$ is the single-CNV kernel defined in (2.2). The maximum operation in the definition of $k_R(\cdot,\cdot)$ searches for the best CNV-to-CNV correspondence in the CNV profiles of sample i and j in the CNVR.

3 Numerical studies

Without loss of generality, we assumed the CNVR to be the interval [0,1] throughout this simulation. We compared CKAT to the widely used Fisher's exact test [Agresti and Kateri, 2011]. A total of 600 subjects were simulated with 300 cases and 300 controls. For subject i, we randomly generated m_i CNVs, where m_i took values 0, 1, 2, 3 with probabilities 0.6, 0.2, 0.1, 0.1 respectively. We simulated $2m_i$ endpoints and sorted them from smallest to largest. The first two endpoints formed the position information of the first CNV, the next two formed the second CNV, and so on. Finally, we randomly simulated a Bernoulli variable with success probability 0.5 as the type of each CNV.

After the CNVs were simulated, we generated the group label y_i from the following logistic model

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{m_i} \left[\left\{ \beta_j^{Del} I[X_{ij}^{(2)} = 1] + \beta_j^{Dup} I[X_{ij}^{(2)} = 3] \right\} X_{ij}^{(1)} \right], \tag{3.1}$$

where $\pi_i = Pr(y_i = 1)$, $\beta_0 = -4$ implies a prevalence of roughly 0.018 for ASD, $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})$ is the jth CNV of the ith subject, and β_j^{Del}, β_j^{Dup} are the log of the odd ratio (OR) of CNV j for deletion and duplication respectively. β_j^{Del} and β_j^{Dup} shared the same absolute values but might have different signs. For simplicity, we called a CNV risk-associated (R) if the associated $\beta_j^{Del} > 0$ or $\beta_j^{Dup} > 0$, protective (P) if β-coefficient is smaller than 0, or neutral (N) on ASD if β-coefficient equals 0. A heterogeneous CNVR containing both deletions and duplications with even probability was considered in this simulation. The effects of deletions and duplications were (Del, Dup)= (R,R), (R,N), (R,P), (P,R). The results are reported in Figure 1 (type I error) and Figure 2 (power).

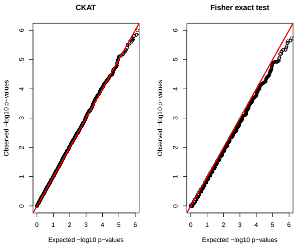

Figure 1: $-\log_{10}$ p-value based QQ plots of CKAT and Fisher's exact test. The x axis represents $-\log_{10}$ expected p-values and the y axis represents $-\log_{10}$ observed p-values.

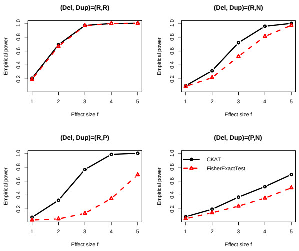

Figure 2: Empirical power of CKAT and Fisher's exact test. The black line is for CKAT and the red line is for Fisher's exact test.

4 Discussions and summary

We have proposed the CKAT to evaluate the association between CNVs and disease-related traits. The kernel implemented in CKAT is elaborately designed so that it can capture special features of CNVs, such as multi-dimensionality (type and size) and heterogeneity effects. The kernel (2.2) is defined in a rather ad hoc fashion; however, we do not pursue an optimal CNV kernel choice in this paper. After the kernel is designed, we then apply the kernel strategy in the literature [Wu *et al.*, 2010, 2011] to test the association between CNVR and disease-related outcomes. Simulation studies show that CKAT can always protect the type I error and have higher power than existing methods under a wide range of scenarios. Finally, CKAT is illustrated with a real data examining the association between CNV and autism. Many CNV regions are detected as significantly associated with ASD. Taking Chromosome 22 as an example, two regions are detected by CKAT. One has a well-established association with ASD in previous studies. The other contains a putative genes, ADORA2A, which might be functionally related to ASD. Further work is needed to understand the biological and genetic mechanisms of the region on ASD.

The proposed CKAT calculates the p-value of the association test analytically, which is computationally efficient and flexible for CNV association analysis, as demonstrated in our numerical studies. Compared with existing methods, it always has adequate power for detecting an existing association. Moreover, CKAT also has good performance when the nominal significance level of the test is extreme, which makes it a desirable tool in genome-wide association analysis where multiple testing burden is usually very high. Besides serving as a useful tool in CNV association analysis, the way of incorporating both CNV length information and CNV type information in the CNV kernels can be also extended to pooling information from different data types. Given the increasing availability of genome-wide information form different data sources, this mode of analysis can shed light on integrative genomics across multiple platforms in the foreseeable near future.

References

Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer Berlin Heidelberg.

Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., et al. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genetics*, **7**, e1002334.

McCarroll, S.A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, **39**, S37–S42.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., et al. (2007). Strong association of de novo copy number mutations with autism. *Science*, **316**(5823), 445–449.

Tzeng, J. Y., Magnusson, P. K., Sullivan, P. F., Szatkiewicz, J. P. and Swedish Schizophrenia Consortium. (2015). A New Method for Detecting Associations with Rare Copy-Number Variants. *PLoS Genetics*, **11**(10), e1005403.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, **86**, 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, **89**, 82–93.