
A Kernel based Study of the Association between Copy Number Variants and Disease-related Traits

by
Nastaran Maus Esfahani

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

under the supervision of
Professor Paul Kennedy
and
Professor Daniel Catchpoole

University of Technology Sydney
Faculty of Engineering and Information Technology
October 2022

© Copyright by Nastaran Maus Esfahani, 2022

Certificate of Original Authorship

I, Nastaran Maus Esfahani, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

SIGNATURE: _____
[Nastaran Maus Esfahani]

DATE: 16th July, 2023

PLACE: Sydney, Australia

Acknowledgements

To my mother Nasrin, my father Mahdi and my only brother Saeed, who gave me their endless emotional support during my Ph.D. journey.

To my supervisor Prof Paul Kennedy who taught me the problem-solving skill that I benefit from it not only in my study but also in my personal life.

To my co-supervisor A/Prof Daniel Catchpoole, for his positive attitude and encouragement whenever things got too difficult for me to handle.

To my friends who were always there for me no matter how far away they were from me.

To myself for not giving up on my Ph.D. journey for what happened to me at the beginning stages, which should not happen to any woman anywhere in the world, and starting from square one to achieve one of my life goals.

Contents

List of Figures	viii
List of Tables	xiii
List of Publications	xv
List of Abbreviations and Symbols	xvii
Abstract	xix
1 Introduction	1
1.1 Background	1
1.2 Research questions	3
1.3 Contributions to knowledge	4
1.4 Thesis structure	7
2 Literature review	9
2.1 Genetic Variations	9
2.1.1 DNA sequence variations	9
2.1.2 Copy Number Variations	11
2.1.3 CNVs and Diseases	14
2.2 Genetic Association Studies	15
2.2.1 Collapsing Methods	15
2.3 Challenges of Studying CNVs	19
2.4 Methods for Studying CNVs	20
2.4.1 CNV Collapsing Random Effects Test	20
2.4.2 Kernel-based Association Tests	21
2.5 Research Gaps	27

3	MCKAT, a multi-dimensional copy number variant kernel association test	29
3.1	Introduction	29
3.2	Model Development	30
3.2.1	Single-pair CNV Kernel	31
3.2.2	Whole Chromosome CNV Kernel	31
3.2.3	Kernel-based Association Test	32
3.3	Model Evaluation and Simulation Results	33
3.4	Real Data Application Results	39
3.4.1	Autism and Rhabdomyosarcoma Data	39
3.4.2	Real Data Results	40
3.5	Discussion	54
3.6	Conclusion	55
4	SMCKAT, a sequential multi-dimensional copy number variant kernel association test	57
4.1	Introduction	57
4.2	Model Development	58
4.2.1	Pair CNV Group Kernel	58
4.2.2	Whole Genome CNV Group Kernel	61
4.2.3	Kernel-based Association Test	63
4.3	Model Evaluation using Simulated Data	63
4.3.1	Simulation Results	65
4.4	Real Data Application Results	71
4.4.1	CNV Analysis on Rhabdomyosarcoma Data Set	71
4.4.2	CNV Analysis on Cytogenetic Bands in RMS	73
4.4.3	CNV Analysis on Autism Data Set	75
4.5	Discussion	75
4.6	Conclusion	76
5	CNV-gene intersection effect on testing the association between CNVs and disease-related traits	79
5.1	Introduction	79
5.2	Effects of CNVs on Gene Expressions	80
5.3	Simulation studies	81

5.4	Real data application results	91
5.5	Discussion	94
5.6	Conclusion	95
6	Conclusion	97
6.1	Identifying the Association between Copy number Variants and Disease related Traits	98
6.2	Work Limitations and Future Works	99
6.3	Conclusion	100
7	Appendix	101
	Bibliography	107

List of Figures

Figure	Page
2.1 Single nucleotide polymorphism. A, T, G and C stand for adenine, thymine, guanine and cytosine respectively.	10
2.2 Characteristics of copy number variants: type, chromosomal position and dosage.	12
2.3 Molecular mechanisms of SNP phenotypes. The paired black lines represent chromosomal regions. Squared brackets ([]) represent the CNV region, both black and white squares show a gene, the dotted lines represent deletion or amplification.	13
2.4 An overview of the CCRET with the dosage model as an example from (Tzeng et al. 2015). C_{1-4} : cases, N_{1-4} : controls, CNVR: copy number variation region, red rectangle: deletion, blue rectangle: duplication, green rectangle: gene. DS: dosage, Len: length, GI: gene intersection. In part (I), CNVRs are created, in part (II) CNV information for each subject is stored in a matrix, and in part (III) the association between CNV characteristics and disease related traits is tested.	22
2.5 Diagram of copy number profile curves and common area under the curve by Brucker et al. (2020). (a) Example of CNV data describing individuals' CNV profile in chromosome 1. (b) Copy number (CN) profile curves of two individuals with overlapping deletions of dosage 0. (c) CN profile curves of two individuals with overlapping with overlapping duplications of dosage 3 and 4. (d) The cAUC between two individuals who have overlapping deletions of dosage 1 and overlapping duplications of dosage 3, so that the cAUC between the individuals is the sum of the two areas.	27

3.1	P-value based QQ-plots of MCKAT, CKAT and CONCUR under first (a) and second (b) simulation scenarios.	36
3.2	Empirical power of MCKAT and CKAT under first simulation scenario, rare CNV data.	37
3.3	Empirical power of MCKAT and CKAT under second simulation scenario, frequent CNV data.	38
3.4	Empirical power of MCKAT and CONCUR under first simulation scenario, rare CNV data.	38
3.5	Empirical power of MCKAT and CONCUR under second simulation scenario, frequent CNV data.	39
3.6	Manhattan plot showing the $-\log(\text{pvalue})$ of testing association between CNVs on the chromosome cytogenetic bands and RMS sub types. Those with $-\log(\text{pvalue})$ above the threshold line, are significantly associated with the RMS subtype	44
3.7	Chromosomal ideograms showing statistically significant cytogenetic bands that CNVs on them are associated with the RMS subtype for chromosomes 2, 8, 11 and 13.	45
3.8	Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 1, 3 and 4.	48
3.9	Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 5, 6 and 7.	49
3.10	Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 9, 10 and 12.	50
3.11	Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 14, 15 and 16.	51
3.12	Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 17, 18 and 19.	52

3.13	Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 20, 21 and 22.	53
4.1	SMCKAT workflow diagram. Firstly, preparing CNV groups for each CNV profiles and aligning relevant CNV groups of each subject. Secondly, measuring the similarity between CNV groups by the pair CNV group kernel. Thirdly, extracting CNV group series for each subject and measuring the similarity between all CNV profiles by the whole genome CNV group kernel. Finally, testing the association between CNV characteristics and sequential order with disease-related traits. .	59
4.2	Generating CNV profile R_i where CNVs are sorted with respect to their chromosomal position. A, B,..., and F are arbitrary CNVs at m^{th} , m^{th+1} , ..., and m^{th+n} positions and G_i is a group of CNVs of size n . .	59
4.3	Aligning CNVs within two CNV groups of size n , G_i and G_j , to generate n CNV pairs.	60
4.4	Sliding window of size n across CNV profile to extract CNV groups of size n	62
4.5	Aligning G_z^i to either of G_{z-1}^j , G_z^j or G_{z+1}^j of the highest similarity. .	62
4.6	P-value based QQ-plots of SMCKAT and MCKAT under the first simulation scenario, the rare CNVs application.	66
4.7	P-value based QQ-plots of SMCKAT and CONCUR under the first simulation scenario, the rare CNVs application.	66
4.8	P-value based QQ-plots of SMCKAT and CKAT under first simulation scenario, the rare CNVs application.	67
4.9	P-value based QQ-plots of SMCKAT and MCKAT under the second simulation scenario, the frequent CNVs application.	68
4.10	P-value based QQ-plots of SMCKAT and CONCUR under the second simulation scenario, the frequent CNVs application.	69
4.11	P-value based QQ-plots of SMCKAT and CKAT under the second simulation scenario, the frequent CNVs application.	69
4.12	Empirical power of SMCKAT, MCKAT, CONCUR and CKAT under the first simulation scenario, rare CNV data.	70
4.13	Empirical power of SMCKAT, MCKAT, CONCUR and CKAT under the second simulation scenario, frequent CNV data.	71

5.1	Scenario 1, no intersections between CNVs and genes. Each row is a CNV profile of a subject. CNVR: copy number variation region, blue rectangle: amplification and red rectangle: deletion.	82
5.2	Scenario 2, genes have intersection only with CNVs of amplification type. Each row is a CNV profile of a subject. CNVR: copy number variation region, red rectangle: deletion.	82
5.3	Scenario 3, genes have intersection only with CNVs of deletion type. Each row is a CNV profile of a subject. CNVR: copy number variation region, blue rectangle: amplification.	83
5.4	Scenario 4, genes have intersection with CNVs of both amplification and deletion types. Each row is a CNV profile of a subject. CNVR: copy number variation region, blue rectangle: amplification and red rectangle: deletion.	83
5.5	P-value based QQ-plots of MCKAT and SMCKAT under the first simulation scenario, no CNV-gene intersections.	86
5.6	P-value based QQ-plots of MCKAT and SMCKAT under the second simulation scenario, only CNV-gene amplification intersections. . . .	87
5.7	P-value based QQ-plots of MCKAT and SMCKAT under the third simulation scenario, only CNV-gene deletion intersections.	88
5.8	P-value based QQ-plots of MCKAT and SMCKAT under the fourth simulation scenario, CNV-gene both deletion and amplification intersections.	88
5.9	Empirical power of MCKAT under CNV-gene intersections and no CNV-gene intersections simulated scenarios.	90
5.10	Empirical power of SMCKAT under the CNV-gene intersections and the no CNV-gene intersections simulated scenarios.	90

List of Tables

Table	Page
2.1 Examples of disorders conveyed by CNVs	15
3.1 P-values of testing the association between RMS subtype and CNVs in each chromosome. (*) denotes significant association between RMS subtype and CNVs by MCKAT, CKAT and CONCUR, (#) denotes the total number of CNVs on that chromosome.	41
3.2 P-values of the testing association between ASD status and CNVs in each chromosome by MCKAT, CKAT and CONCUR. (*) denotes significant association between ASD and CNVs, (#) denotes the number of total CNVs on that chromosome.	43
3.3 P-values of the testing association between RMS subtype and CNVs in each cytogenetic bands of chromosome 8 by MCKAT. (*) denotes significant association between RMS subtype and CNVs, (#) denotes the number of total CNVs on the band.	46
3.4 The cytogenetic bands across the whole genome identified as significantly associated with the RMS subtype by MCKAT. (#) denotes the number of CNVs on the band.	47
4.1 P-values of the chromosomes that their CNV sequential orders are identified significantly associated with the RMS sub types for the different CNV group sizes.	72
4.2 P-values of the testing association between RMS subtype and CNVs in the chromosome 8 cytogenetic bands by SMCKAT, MCKAT and CKAT. (*) denotes significant association between RMS subtype and CNVs, (#) denotes the number of total CNVs on the band.	74

4.3	P-values of testing the association between CNV sequential order and ASD status trying different CNV group sizes.	75
5.1	Genes with significant frequency of somatic mutation across RMS patients	84
5.2	Genes reported by Shern et al. (2014) as embryonal and alveolar RMS cancer sub types classifier genes.	91
5.3	P-values of testing the association between RMS subtype and CNVs, both CNV characteristics and CNV-gene intersection, in each chromosome. (*) denotes significant association identified by MCKAT. . . .	93
5.4	P-values of testing the association between RMS subtype and CNVs, both CNV characteristics and CNV-gene intersections, in chromosomes 2, 11, 8 and 13 with group size of 5. (*) denotes significant association identified by SMCKAT.	93

List of Publications

Listed below are the publications and other outputs associated with the research presented in this thesis.

Maus Esfahani, N., Catchpoole, D., Khan, J., & Kennedy, P. J. (2021). MCKAT: a multi-dimensional copy number variant kernel association test. BMC bioinformatics, 22(1), 1-16. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04494-w>

Maus Esfahani, N., Catchpoole, D., & Kennedy, P. J. (2021). SMCKAT, a Sequential Multi-Dimensional CNV Kernel-Based Association Test. Life, 11(12), 1302. <https://www.mdpi.com/2075-1729/11/12/1302>

List of Abbreviations and Symbols

Abbreviation	Description
AS-PCR	allele-specific PCR
ARMS	alveolar rhabdomyosarcoma
ASD	autism spectrum disorder
CAPS	cleaved amplified polymorphic sequence
CCRET	CNV collapsing random effects test
CKAT	CNV kernel association test
CNVRs	CNV regions
CAST	collapsing and summation test
cAUC	common area under the curve
CXN	copy number
CNP	copy number polymorphisms
CONCUR	copy number profile curve-based association test
CNV	copy number variation
dCAPs	derived CAPS
DR	difference from the Reference
ERMS	embryonal rhabdomyosarcoma
FWER	family-wise error rate
MCKAT	multi-dimensional CNV kernel-based association test
NIH	national Institute of Health
PCR	polymerase chain reaction
Q-Q	quantile-quantile
ROI	region of interest
RMS	rhabdomyosarcoma
SMCKAT	sequential multi-dimensional copy number variant kernel association test
SNP	single nucleotide polymorphism
VOUS	Variants of uncertain significance
WS	weighted-sum

Symbol	Description
X	a copy number variant
$X^{(1)}$	start chromosomal position
$X^{(2)}$	end chromosomal position
$X^{(3)}$	CNV type
$X^{(4)}$	CNV dosage
K	similarity matrix
K_{ij}	similarity between CNV profile i and j
K_s	single-pair CNV Kernel
R_i	list of a subject's CNV
K_w	whole Chromosome CNV Kernel
y_i	status of the phenotype
Z	covariant matrix
G	CNV group
K_{PG}	pair CNV Group Kernel
P	CNV group series
K_{WG}	Whole Genome CNV Group Kernel

Abstract

Copy number variants (CNVs) are the most common form of structural genetic variation, reflecting the gain or loss of DNA segments compared with a reference genome. Studies have shown that CNVs are linked to various disorders like autism, intellectual disability, and schizophrenia. Consequently, the interest in studying a possible association of CNVs to specific disease traits is growing. However, due to the specific multi-dimensional characteristics of the CNVs, methods for testing the association between CNVs and the disease-related traits are still few and underdeveloped. The research presented in this thesis addresses several aspects of research on the association between CNVs and disease related traits , and the broader concepts of the association between CNV sequential order with adverse phenotype, and the association of the CNV and other genetic variation interactions with disease related traits.

This work makes three contributions to knowledge, relating to the significance of CNVs on some chromosomal regions in association with disease related traits. Contribution 1 proposed a multi-dimensional CNV kernel based association test (MCKAT). MCKAT performs better than the state of the art methods and was evaluated on both simulated and real data. MCKAT can identify chromosomal regions at cytogenetic band level containing CNVs that are significantly associated with disease related traits. MCKAT considers all CNV characteristics in testing the association and can provide strong evidence, small p-values, to accept or reject the association hypothesis. MCKAT is applicable to both frequent and rare CNV data sets.

Contribution 2 is a sequential multi-dimensional CNV kernel based association test (SMCKAT). SMCKAT tests the association between the CNV sequential order and disease related traits. SMCKAT considers not only the CNV characteristics but the CNV sequential order. SMCKAT can identify the chromosomal regions that the CNV sequential order is significantly associated with disease related traits.

Based on our knowledge, SMCKAT is the first such method to test the association between the CNV sequential order and disease related traits.

Contribution 3 uses our proposed method to demonstrate that considering the CNV-gene intersection along with the CNV characteristics in testing the association between CNVs and disease related traits is informative and can provide more insights about the disease development. This is because CNVs can affect their intersected genes in different way like changing the gene expression or disturbing their function. Our proposed methods can be used not only in testing the association between the dual effect of CNVs and their intersected with disease related traits but any other genetic variations based on data availability.

Overall, this research confirms that having association tests, specific to CNVs and compatible with CNV characteristics, to identify the chromosomal regions which contain CNVs that are significantly associated with disease related traits are of biological significance. This work provides methods that can help biologists to identify CNV hot spots associated with disease related traits without doing extensive investigations at the individual level to find significant CNVs. The results of these methods may also provide them with a better understanding of how the interaction between CNVs and other genetic variations like genes can have association with a disease related traits.