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

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.

SIGNATURE: Production Note: Signature removed prior to publication.

[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

Figure **Page**

List of Tables

Table **Page**

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 evaluatd 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. SMKAT 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.