UNIVERSITI TEKNOLOGI MARA

GENOMICS OF THE ORANG ASLI AND MALAYS: FUNCTIONAL MODELLING OF THE PATHOGENIC SNPS FOR CANCERS AND RISK PREDICTION FOR EATING BEHAVIOURS, NUTRIENT DEFICIENCIES AND METABOLIC DISORDERS

NURUL AIN BINTI KHORUDDIN

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Science)

Faculty of Applied Sciences

March 2023

ABSTRACT

Single-nucleotide polymorphisms (SNPs) are the most common genetic variations associated with various human diseases, including cancers and nutritional disorders. Genome-wide association studies (GWAS) had discovered numerous SNPs related to increased risk of cancers such as breast cancer, colorectal cancer, and leukaemia as well as nutritional disorders. However, GWAS are frequently carried out in certain populations, for which the Orang Asli and Malays were excluded. By mining the whole-genome sequence of the 98 Orang Asli and 96 Malays, genome variations were identified and two bioinformatic pipelines were established to identify and evaluate the impact of pathogenic SNPs that may increase the risk of cancers, eating behaviours, nutrient deficiency, and metabolic disorders among them. A database of genotype-predicted phenotypes was built in this study. For the cancer risk prediction pipeline, five different in silico tools, SIFT, PROVEAN, PolyPhen-2, Condel, and PANTHER, were utilised to predict and analyse the functional effect of the SNPs. Out of the 80 cancer-related nsSNPs from the GWAS dataset, 52 nsSNPs were found among the Orang Asli and Malays. Three nsSNPs, rs1126809 (TYR), rs10936600 (LRRC34), and rs757978 (FARP2) were identified as the most damaging pathogenic variants associated with basal cell carcinoma or squamous cell carcinoma, multiple myeloma and chronic lymphocytic leukaemia, respectively. These mutations modify the protein interface and change the allosteric sites of the respective proteins. As the TYR, LRRC34, and FARP2 genes are involved in so many biological processes, including cell proliferation, differentiation, growth, and survival, any loss of protein function might lead to cancer formation. Thus, rs1126809, rs10936600, and rs757978 are the significant pathogenic variants that are likely to increase the risks of cancers among the Orang Asli and Malays. For the nutrient-related variant prediction pipeline, three bioinformatics tools (VCFtools, ANNOVAR and VEP) were used to identify and annotate the SNPs. The genetic risks of the eating behaviours, nutrient deficiencies, and metabolic disorders of both cohorts, Orang Asli and Malays, were profiled. The Orang Asli and the Malays genomes have an average of 70 SNPs associated with eating behaviours, 81 SNPs associated with nutrient deficiencies and 80 SNPs associated with metabolic disorders. The genetic markers identified in this study provided the basis for phenotype-genotype studies to be conducted within the Malaysian populations so that an association between genetic markers with cancers, eating behaviours, nutrient deficiency, and metabolic disorders can be established. It is believed that this bioinformatics approach would complement the healthcare providers to offer appropriate preventive or corrective measures for individuals at risk. Thus, the developed pipeline and the database generated from this study are fundamental in implementing precision medicine for cancers and nutrient disorders. The data may be used to strategise educational programmes or interventions to increase awareness and promote a healthy lifestyle among the OA and Malays. However, the functions and effects of the identified variants still require wet lab experiments for further investigations.

ACKNOWLEDGEMENT

First and foremost, I am thankful to The Almighty Allah S.W.T. for blessing me with the courage, knowledge and strength throughout my research work to complete my PhD journey.

I am extremely grateful to my supervisors; Prof. Dato' Dr. Mohd Zaki Salleh and Prof Dr. Hajah Farida Zuraina Mohd Yusof for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and experience have inspired me not only in my academic research work but also life in general. Additionally, I would like to express my deepest gratitude to Prof. Dr. Teh Lay Kek for her treasured support which was influential in shaping my study methods and critiquing my results.

I would like to express my sincere thanks to Dr. Nur Fakhruzzaman Noorizhab, Dr. Lim Wai Feng and Asso. Prof. Dr. Richard Johari James for their mentorship and technical support on my study. I enjoyed working with fellow postgraduate students in particular Nourul Emmilia Mohd Fazli and Norsuhaila Rosmimi Rosli. Thank you for the insightful discussion, input and support. The suggestions, thoughts, and constructive comments had undoubtedly helped to provide different perspectives to my research. I would also like to thank all staff in iPROMISE. It is their kind help and support that have made my study and life in iPROMISE wonderful.

I am tremendously thankful to my late father and my mother for their love, prayers, strength, and sacrifices in educating and preparing me for this journey. Special gratitude to my beloved husband, Ahmad Muzammil Abdul Aziz for his love, understanding and sacrifices that allows me to complete my PhD. Also my gratitude to my siblings for the tremendous support, encouragement and hope they had given to me over these years. Without all of them, it would be impossible for me to finish my study. Thank you all for the strength you gave me. I love you all!

Last but not least, I would also like to acknowledge the Ministry of Higher Education, Malaysia through the Long-Term Research Grant (LRGS) scheme and "Geran Insentif Penyeliaan" (GIP) for funding this study.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1 Research Background

Precision medicine aims to achieve better healthcare for every patient by personalising prevention and treatment. It can also improve patients quality of life while reducing unnecessary diagnostic testing and therapies (Dzau, 2016). On the other hand, precision health promises the delivery of health at the individual level, notably in disease prevention and risk assessment. Rapid genomic discoveries of variants were made possible through genome-wide association studies and decreasing costs of sequencing (Mardis, 2017) and genotyping. These had helped to reposition precision medicine from an academic exercise closer to clinical reality (Collin & Vamus, 2015; Mirnezami et al., 2012). To date, at least fourteen countries have practised precision medicine through national genomic-medicine initiatives funded by their government (Stark et al., 2019). Some of these countries, such as Saudi Arabia, France, Turkey, Australia, and the United Kingdom, focus on variants of certain diseases including non-communicable diseases, cancers, and rare diseases (Stark et al., 2019). However, Malaysia is still in the early phase of adopting precision medicine, as good sequencing facilities, skilful and knowledgeable geneticists, scientists, and bioinformaticians are lacking (Jamal, 2017).

A rapid improvement in the big data analytics field has played an essential role in the evolution of healthcare practices and research. Access to large omics data, such as proteomics, nutrigenomics, genomics, transcriptomics, metagenomics, and metabolomics data, have revolutionised biology, leading to a better understanding of biological mechanisms (Hasin et al., 2017). In fact, traditional observational epidemiology or biology alone is no longer sufficient to fully interpret complex heterogeneous disorders, which indirectly limits the prevention and treatment of certain diseases (Alyass et al., 2015). Besides, big data analytics have been applied recently towards helping the process of diagnosis and treatment of disease as well as disease exploration (Chen et al., 2017). Consequently, precision medicine has rapidly evolved with the recent development of large-scale biological databases, robust methods for