Whole-genome Sequencing-based Association Studies of Cardiovascular Biomarkers

Jie Huang

This dissertation is submitted to the University of Cambridge
Faculty of Biology for the degree of Doctor of Philosophy

Darwin College

University of Cambridge

February 2015

PREFACE

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

The dissertation does not exceed the page limit of 300 specified by the Biology Degree Committee

ABSTRACT

Background: Genome-wide association studies (GWAS) have significantly advanced the genetic study of complex human traits. With the advent of whole-genome sequencing (WGS) technologies and the increased capacity to identify rare variants, GWAS that use WGS data are expected to provide further opportunities for the discovery of variants that have larger and even causal effects. The UK10K project is one of the largest studies that use WGS to investigate the contribution of low frequency and rare genetic variants to medical traits.

Research aims: My research aims to address the utility of WGS-based imputation and associations for identifying the genetic determinants of a select quantitative traits that are associated with cardiovascular risks. Under the UK10K project framework, I study a suite of circulating biomarkers that have been reported for association with CVD. Specifically, I seek to evaluate the following three broad aspects: 1. what are the characteristics of phasing and imputation with WGS data? 2. what novel analytic methods could be applied to a large scale WGS based association study on a rich of phenotypes? 3. can I identify novel and potentially stronger effect genetic variants that are associated with the chosen CVD traits?

Methods: My study leverages existing WGS data from the UK10K project (N = ~4,000) and further uses it as a reference to impute more samples (N > 10,000) that have genome-wide SNP array data. In doing so, I first evaluate the quality of the WGS data and its utility for imputation, by comparing it to WGS data from the 1000 Genomes Project. Then, I examine the associations between genotypes and phenotypes for 13 quantitative traits, first in samples having WGS and then in samples having imputed data. The 13 CVD related biomarkers include four lipid traits (high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), total cholesterol (TC), triglycerides (TG)), one inflammatory biomarker (C-reactive protein (CRP)), and eight haematological traits (hemoglobin (HGB), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), packed cell volume (PCV), platelet counts (PLT), red blood cell counts (RBC), white blood cell counts (WBC)).

To my dear parents: YuanYu Huang & Youquan Deng

To my wife: Weilin Chen

To my daughter Valerie and my son Jimmy

ACKNOWLEDGEMENTS

From the three years of my PhD study there are a lot of people I want to mention and thank for giving me help and advice. The first and most important person is Dr. Nicole Soranzo, my PhD thesis supervisor who gave me an opportunity to come to Cambridge and Sanger Institute to pursue my PhD study and work on a most important and frontier project on WGS. Her dedication to science and attentive guidance to trainees inspires me to become a good researcher and scientific leader myself.

I would like to thank the other faculty members who served on my thesis committee and provided strategic guidance for my PhD project: Dr. Richard Durbin, Dr. Carl Anderson, Dr. Eleftheria Zeggini from Sanger Institute, and Dr. Adam Butterworth from University of Cambridge.

I would like to thank our wonderful teammates (team 151): Lu Chen, Klaudia Walter, Louella Vasquez, Valentina Iotchkova, Massimiliano Cocca, Matthias Geihs, Yasin Memari, So-Youn Shin. Over the past three years, we sit close together and worked even closely with each other. Whenever there is a question, I feel that I could simply turn around my chair and get insightful feedback and help instantly.

I would also like to thank the larger community of the UK10K project and outsider collaborators who contributed data for our large meta-analysis, particularly to Drs. Nic Timpson and Josine Min from Bristol University.

Finally, I would like to thank my former supervisor, Dr. Chris O'Donnell at Framingham Heart Study, for offering me research experience in cardiovascular genetics and continuous support for my PhD study and career growth.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1000GP	1000 Genomes Project
ADH	Autosomal dominant hypercholesterolemia
ALSPAC	Avon Longitudinal Study of Parents and Children
Apo-A1	apolipoprotein A-I
Apo-B	apolipoprotein B
Apo-E	apolipoprotein E
AMD	age-related macular degeneration
BF	Bayes' factor
BGI	Beijing Genomics Institute
BP	blood pressure
CAD	coronary artery disease
CBR	Cambridge BioResource
CHD	coronary heart disease
CNV	copy number variation
CKD	chronic kidney disease
CRP	C-reactive protein
CVD	cardiovascular disease
DALYs	disability-adjusted life years
DHS	DNaseI hypersensitive sites
EAF	effect allele frequency
EMR	electronic medical records
ERFC	Emerging Risk Factors Collaboration
FHS	Framingham Heart Study
FVG	Friuli Venezia Giulia
GWAS	genome-wide association studies
HDL	high-density lipoprotein
HGB	hemoglobin
HELIC	HELlenic Isolated Cohorts study
HMM	hidden markov model
HWE	hardy-weinberg equilibrium
IBD	identify by descent

IBS	identify by state
InDel	insertion/deletion polymorphism
INGI	Italian Network of Genetic Isolates
LD	linkage disequilibrium
LDL	low-density lipoprotein
LMT	lipid modification therapies
LoF	loss of function
LOLIPOP	London Life Sciences Population study
LURIC	Ludwigshafen Risk and Cardiovascular Health
MAF	minor allele frequency
HGB	haemoglobin
MCH	mean corpuscular hemoglobin
MCHC	mean corpuscular hemoglobin concentration
MCV	mean cell volume
MDS	multidimensional scaling
MI	myocardial infarction
MR	mendelian randomisation
OR	odds ratio
PCA	principle component analysis
PCV	packed cell volume
PLT	platelet count
PROCARDIS	Precocious Coronary Artery Disease Study
QC	quality control
RBC	red blood cell
RCT	reverse cholesterol transport
SKAT	sequence kernel association test
SKAT-O	sequence kernel association test - optimized
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
TC	total cholesterol
TFBS	transcription factor binding sites
TG	triglycerides
TSS	transcription start site

TwinsUK	UK Adult Twin Registry
UK10K	10,000 UK genome sequencing project
UTR	untranslated regions
VB	Val Borbera
WBC	white blood cell
WGS	Whole Genome Sequencing
WTCCC	Wellcome Trust Case Control Consortium
WTSI	Wellcome Trust Sanger Institute

PUBLICATIONS ARISING FROM THIS DISSERTATION

- ** Co-first author*
- *For papers with more than 10 authors, my name is listed together with the first 3 and the last 3 authors. When there are more than 3 co-starred first-authors, all of them are listed.*

1. Gormley P*, Downes K*, **Huang J***, Kettunen J, Aki S, …, Palotie A, Ripatti S, Soranzo N. A polygenic panel of platelet-associated SNPs is associated with risk of incident ischaemic stroke. (*submitted*)
2. Walter K*, Min M*, **Huang J***, Lucy Crooks*, …, Timpson NJ, Durbin R , Soranzo N. The UK10K project: rare variants in health and disease. *(under revision)*
3. **Huang J**, Howie B, Memari M, …, Timpson NJ, Marchini J, Soranzo N, UK10K Project. A reference panel of 3,781 genomes from the UK10K Project increases imputation performance over the 1000 Genomes Project. *Nature Communications*. (*accepted*)
4. Taylor P, Porcu E, Chew S, … **Huang J**, …, Soranzo N, Timpson NJ, Wilson S, the UK10K Consortium. Whole genome sequence based analysis of thyroid function. *Nature Communications*. 2015 Mar 6;6:5681
5. Timpson NJ, Walter K, Min JL, …, **Huang J**, ..., Humphries SE, Zeggini E, Soranzo N; UK1OK consortium members. A novel low-frequency variant near APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Communications*. 2014 Sep 16;5:4871
6. O'Connell J, Gurdasani D, Delaneau O, …, **Huang J**, …, Soranzo N, Sandhu MS, Marchini J. A general approach for haplotype phasing across the full spectrum of relatedness. *PLOS Genetics* 2014 Apr 17;10(4):e1004234

PUBLICATIONS ARISING ELSEWHERE (from 2012-01 to 2015-01)

- * *Co-first author*
- *For papers with more than 10 authors, my name is listed together with the first 3 and the last 3 authors. When there are more than 3 co-starred first-authors, all of them are listed.*

1. Baumert J*, **Huang J***, McKnight B*, Sabater-Lleal M*, Steri M*, …, Strachan DP, Peters A, Smith NL. No evidence for genome-wide interactions on plasma fibrinogen by smoking, alcohol consumption and body mass index: results from meta-analyses of 80,607 subjects. *PLoS One*. December 31, 2014 DOI: 10.1371
2. Shin SY, Fauman EB, Petersen AK, …, **Huang J**, …, Kastenmüller G, Spector TD, Soranzo N. An atlas of genetic influences on human metabolism. *Nature Genetics* 2014 Jun;46(6):543-50
3. Han B, Luo H, Raelson J, **Huang J**, Li Y, Tremblay J, Hu B, Qi S, Wu J. TGFBI (βIG-H3) is a diabetes risk gene based on mouse and human genetic studies. *Hum Mol Genet*. 2014 Apr 11
4. **Huang J**, Huffman JE, Yamkauchi M, …, Lowenstein CJ, Strachan DP, O'Donnell CJ; CHARGE Consortium Hemostatic Factor Working Group. Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. *Arteriosclerosis,Thrombosis, and Vascular Biology*. 2014 Feb 27
5. Sabater-Lleal M*, **Huang J***, Chasman D*, Naitza S*, Dehghan A*, …, Strachan DP, Hamsten A, O'Donnell CJ. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*. 2013 Sep 17;128(12):1310-24.
6. **Huang J**, Liu Y, Welch R, Willer C, Hindorff LA, Li Y. WikiGWA: an open platform for collecting and using genome-wide association (GWA) results. *European Journal of Human Genetics.* 2013 Apr;21(4):471-3
7. O'Seaghdha CM, Wu H, Yang Q, …, **Huang J**, …, Bonny O, Fox CS, Bochud M. Meta-analysis of genome-wide association studies identifies six new loci for serum calcium concentrations. *PLoS Genetics*. 2013 Sep;9(9):e1003796
8. Kleber ME, Seppälä I, Pilz S, …, **Huang J**, …, Lehtimäki T, März W, Meinitzer A. Genome-wide association study identifies three genomic loci significantly associated with

serum levels of homoarginine – The AtheroRemo Consortium. *Circulation Cardiovascular Genetics.* 2013 Sep 18

9. McGrath LM, Cornelis MC, …, **Huang J**, …, Sullivan P, Perlis RH, Smoller JW. Genetic predictors of risk and resilience in psychiatric disorders: a cross-disorder genomewide association study of functional impairment in major depressive disorder, bipolar disorder, and schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2013 Sep 13
10. ALSGEN Consortium, Ahmeti KB, Ajroud-Driss S, …, **Huang J**, …, Veldink JH, Yang Y, Zheng JG. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. *Neurobiology of Aging*. 2013 Jan;34(1):357.e7-19
11. Chen M-H*, **Huang J***, Chen W-M, Larson MG, Fox CS, Vasan RS, Seshadri S, O'Donnell CJ, Yang Q. Using family-based imputation in genome-wide association studies with large complex pedigrees: the Framingham Heart Study. *PLoS ONE*. 2012;7(12):e51589. (*co-first author)
12. **Huang J**, Sabater-Lleal M, Asselbergs FW, …, Liu Y, O'Donnell CJ, Hamsten A. Genome-wide association study for circulating levels of plasminogen activator inhibitor-1 (PAI-1) provides novel insights into the regulation of PAI-1. *Blood.* 2012 Dec 6;120(24):4873-81
13. **Huang J**, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 data. *European Journal of Human Genetics.* 2012 Jul;20(7):801-5
14. Bis JC, DeCarli C, Smith AV, …, **Huang J**, …, Launer LJ, Ikram MA, Seshadri S; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. Common variants at 12q14 and 12q24 are associated with hippocampal volume. *Nature Genetics*. 2012 Apr 15;44(5):545-51.
15. Willour VL1, Seifuddin F, Mahon PB, …, **Huang J**, …, Gurling H, Purcell S, Smoller JW, Craddock N, DePaulo JR Jr, Schulze TG, McMahon FJ, Zandi PP, Potash JB. A genome-wide association study of attempted suicide. *Mol Psychiatry.* 2012 Apr;17(4):433-44.