THE FUNCTIONAL IMPACT OF COPY NUMBER VARIATION IN THE HUMAN GENOME

This dissertation is submitted for
the degree of Doctor of Philosophy,
by

NI HUANG

Wellcome Trust Sanger Institute
Darwin College, University of Cambridge

December 2011

PREFACE

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except specifically indicated in the text and acknowledgements. No part of this dissertation has been submitted for a degree or diploma or other qualification at the University of Cambridge or any other university. This dissertation does not exceed the 60,000 words excluding bibliography and appendices.

ACKNOWLEDGEMENT

This thesis would not have been possible without the constant guidance, help, encouragement and support of my wonderful supervisor, Matthew Hurles, to whom I am forever in debt. It is hard to overstate how much I have benefited from his knowledge, insight, patience and enthusiasm. Without him, I would have been lost.

I wish to thank members of my thesis committee: my secondary supervisor, Gos Micklem, Inês Barroso and Helen Firth for the helpful advices. I wish to thank Sadaf Farooqi and Elena Bochukova at Addenbrooke's Hospital, for their collaboration and the inspiring discussions. I am grateful to Richard Redon, whose enthusiastic supervision during my first rotation drew me to the study of copy number variation, and Chris Tyler-Smith and Yali Xue, whose teaching and caring made my second rotation an unforgettable experience.

Special thanks to people at the graduate program at Sanger Institute, Alex Bateman, Christina Hedberg-Delouka and Annabel Smith for the warm support and caring that helped me through my most difficult period during the writing of this thesis. Lots of thanks to all members of team 29 and 19 past and present at the Sanger Institute for all the help, encouragement and friendship.

I also thank the Wellcome Trust for the generous financial support.

Finally, I wish to thank my parents, whose love and understanding supported me throughout this four-year endeavor abroad, as always.

SUMMARY

The functional impact of copy number variation in the human genome

Ni Huang

Copy number variation (CNV) is a class of genetic variation where large segments of the genome vary in copy number among different individuals. It has become clear in the past decade that CNV affects a significant proportion of the human genome and can play an important role in human disease. With array-based copy number detection and the current generation of sequencing technologies, our ability to discover genetic variants is running far ahead of our ability to interpret their functional impact. One approach to close this gap is to explore statistical association between genetic variants and phenotypes. In contrast to the successes of genome-wide association studies for common disease using common single nucleotide polymorphism (SNP) as markers, the majority of disease CNVs discovered so far have low population frequencies and are mainly involved in rare developmental disorders. Another strategy to improve interpretation of genomic variants is to establish a predictive understanding of their functional impact. Large heterozygous deletions are of particular interest, since *i*) loss-of-function (LOF) of coding sequences encompassed by large deletions can be relatively unambiguously ascribed and *ii*) haploinsufficiency (HI), wherein only one functional copy of a gene is not sufficient to maintain normal phenotype, is a major cause of dominant diseases.

This thesis explored both approaches. Initially, I developed an informatics pipeline for robust discovery of CNVs from large numbers of samples genotyped using the Affymetrix whole-genome SNP array 6.0, to support both the association-based and prediction-based study. For the disease association strategy, I studied the role of

both common and rare CNVs in severe early-onset obesity using a case-control design, from which a rare 220kb heterozygous deletion at 16p11.2 that encompasses *SH2B1* was found causal for the phenotype and an 8kb common deletion upstream of *NEGR1* was found to be significantly associated with the disease, particularly in females. Using the prediction-based approach, I characterized the properties of HI genes by comparing with genes observed to be deleted in apparently healthy individuals and I developed a prediction model to distinguish HI and haplosufficient (HS) genes using the most informative properties identified from these comparisons. An HI-based pathogenicity score was devised to distinguish pathogenic genic CNVs from benign genic CNVs. Finally, I proposed a probabilistic diagnostic framework to incorporate population variation, and integrate other sources of evidence, to enable an improved, and quantitative, identification of causal variants.

PUBLICATIONS

Publications arising from work associated with this thesis:

1. E. G. Bochukova*, **N. Huang***, J. Keogh, E. Henning, C. Purmann, K. Blaszczyk, S. Saeed, J. Hamilton-Shield, J. Clayton-Smith, S. O'Rahilly, M. E. Hurles, and I. S. Farooqi. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463:666–70, 2010.

2. **N. Huang**, I. Lee, E. M. Marcotte, and M. E. Hurles. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet*, 6:e1001154, 2010.

3. N. J. Prescott, K. M. Dominy, M. Kubo, C. M. Lewis, S. A. Fisher, R. Redon, **N. Huang**, B. E. Stranger, K. Blaszczyk, B. Hudspith, G. Parkes, N. Hosono, K. Yamazaki, C. M. Onnie, A. Forbes, E. T. Dermitzakis, Y. Nakamura, J. C. Mansfield, J. Sanderson, M. E. Hurles, R. G. Roberts, and C. G. Mathew. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum Mol Genet*, 19:1828–39, 2010.

4. S. Nik-Zainal, R. Strick, M. Storer, **N. Huang**, R. Rad, L. Willatt, T. Fitzgerald, V. Martin, R. Sandford, N. P. Carter, A. R. Janecke, S. P. Renner, P. G. Oppelt, P. Oppelt, C. Schulze, S. Brucker, M. Hurles, M. W. Beckmann, P. L. Strissel, and C. Shaw-Smith. High incidence of recurrent copy number variants in patients with isolated and syndromic Müllerian aplasia. *J Med Genet*, 48:197–204, 2011.

5. D. G. MacArthur, S. Balasubramanian, A. Frankish, **N. Huang**, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, 1000 Genome Project Consortium, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335:823–8, 2012.

*Join first authors

TABLE OF CONTENTS