Positive Natural Selection in the Human Genome

Min Hu

Darwin College

University of Cambridge

August 2012

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This thesis describes my work undertaken at The Wellcome Trust Sanger Institute, in fulfillment of the requirements for the degree of Doctor of Philosophy, at Darwin College, University of Cambridge. This dissertation is the result of my own work and contains nothing that is the outcome of work done in collaboration, except where specifically indicated in the text. The work described here has not been submitted for a degree, diploma, or any other qualification at any other university or institution. I confirm that this thesis does not exceed the word limit set by the Biology Degree Committee.

Min Hu

Cambridge, August 2012

Acknowledgements

I can hardly express in words how thankful I am to people who made my past four years full of growth, joy and cheer, and who made this thesis possible. In 2008, excited but a little unsure, I entered the world of cutting-edge science in genomics and genetics, and for the first time, I started my life in a country with a different language, unfamiliar culture and perhaps many more opportunities. Big thanks to Matt Hurles and Alex Bateman, who kindly provided me with opportunities to rotate in their labs when I knew very little about the subjects, and taught me the basics of being a good scientist. Great thanks to Don Conrad and Ni Huang in Matt's group, who patiently taught me all the important and trivial things I needed to know to start coding and using the powerful Sanger computing farm.

The most grateful thanks go to my supervisor, Chris Tyler-Smith. He has always been extremely generous in devoting his time to teach and guide me in my research and scientific writing. There had been countless idea exchanges in our weekly meetings during the last three years, and lots of encouragements when I was facing challenges in my projects. Big thanks to Yali Xue, who held my hands throughout my very first project in the group and took care of me in every aspect; Qasim Ayub, who conducted excellent experiments for me in the lab; and Yuan Chen, who helped me a lot in retrieving data from the databases. Great thanks to previous colleagues Daniel MacArthur and Bryndis Yngvadottir, who shared with me lessons they learned from their PhD studies, and provided me with lots of support. Also thanks go to other members in team 19, who had made my life joyful with banters, parties and BBQs, and who had always been caring and helpful in difficult times.

Great thanks to Toomas Kivisild, my external supervisor, and to Jeff Barrett and Alex Bateman in my thesis committee, who provided me with valuable suggestions and ideas, as well as kind support and praise. Also thanks to Annabel Smith and Christina Hedberg-Delouka, who provided support on each critical step in my PhD. Great thanks to Wellcome Trust for the excellent PhD programme and generous scholarship.

I was lucky enough to get involved in the 1000 Genomes Project, collaborating with excellent scientists from allover the world. I enjoyed all the meetings, conferences and phone calls, and learnt a lot from this perhaps one of the world's biggest research projects in life sciences. Big thanks to all the participating scientists in the 1000 Genomes Project, who had been extremely helpful in providing data and exchanging ideas.

Finally, I would like to thank all my friends and my family, without whom I would not have been where I am today. Thanks to my fellow PhD students at Sanger, who formed a fun and supportive community around me. Thanks to all the friends I made during my time in Cambridge, who had made my life abroad more enjoyable than I could have ever hoped for. Special thanks to my parents, who had always been supportive to every decision I have made in my life.

Publications

Publications arising during the course of the work described in this thesis by the time of submission:

Hu M, Ayub Q, Guerra-Assunção JA, Long Q, Ning Z, Huang N, Romero IG, Mamanova L, Akan P, Liu X, Coffey AJ, Turner DJ, Swerdlow H, Burton J, Quail MA, Conrad DF, Enright AJ, Tyler-Smith C and Xue Y (2012). Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. *Human genetics*, 131;5;665-74.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, **Hu M**, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB and Tyler-Smith C (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* (New York, N.Y.), 335;6070;823-8.

1000 Genomes Project Consortium (including **Hu M** and Tyler-Smith C) (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467;7319;1061-73.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, **Hu M**, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW and Hurles ME (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464;7289;704-12.

Abstract

The detection of positive natural selection in the human lineage is of great interest for the understanding of modern human phenotypes and adaptations to different environmental conditions. Although extensive genome-wide scans for signatures of positive selection have been performed using genotype data, these have significant limitations, illustrated by the low overlap among different studies. Thanks to the Next-Generation Sequencing technology, near-complete sequence data for both the whole genome and targeted regions are now available, allowing a nearly unbiased genome-wide scan for positive selection as well as the possibility of localizing the specific variants selected.

The theme of this PhD thesis is to detect and localize positive selection targets in the human genome using sequencing data. This includes three projects:

(1) Localizing selection targets in candidate regions identified by LD-based tests on genotype data, by applying frequency-spectrum based tests (Tajima's *D*, Fay and Wu's *H*, and a Composite Likelihood Ratio test) to targeted resequencing data. Two regions were resequenced at high coverage and putative selection targets were identified.

(2) A genome-wide scan of selective sweeps using frequency-spectrum based tests on 1000 Genomes Project low coverage Pilot data. Candidate positively selected regions and genes were identified and some interesting examples and their plausible selected functions are discussed.

(3) A genome-wide search for regions with very recent ancestry among all humans. Regions with shared recent coalescence times indicate positive selection affecting all modern humans, which has an older age than the recent positive selection identified by neutrality tests. We calculated the Time to the Most Recent Common Ancestor (TMRCA) of low diversity/divergence regions in the human genome, with the aim of identifying regions with very recent common ancestor, which may have been positively selected during early modern human evolution.

These three projects altogether demonstrated the value and impact of low-coverage or high-coverage, targeted or whole-genome sequencing data on providing new insights into positive natural selection in the modern human history, and built up the first steps of the exciting new sequencing era for the exploration of human evolution.

Abbreviations

aCGH	array-comparative genomic hybridization
ASW	African ancestry in Southwest USA
CEU	Utah residents with European ancestry
CGI	Complete Genomics Inc.
CHB	Chinese Han in Beijing
CLR	composite likelihood ratio
cM	centimorgan
CMS	composite of multiple signals
CNV	copy number variant
CRT	cyclic reversible termination
DAF	derived allele frequency
DBP	diastolic blood pressure
DNA	deoxyribonucleic acid
EHH	extended haplotype homozygosity
ENCODE	encyclopedia of DNA elements
eQTL	expression quantitative trait loci
FDR	false discovery rate
FoSTeS	fork stalling and template switching
Gb	gigabases

GIH	Gujarati Indian in Houston
GWAS	genome wide association studies
HLA	human leukocyte antigen
IQR	interquartile range
JPT	Japanese in Tokyo
kb	kilobases
KYA	thousand years ago
LD	linkage disequilibrium
LSA	later stone age
LWK	Luhya in Webuye, Kenya
MAF	minor allele frequency
Mb	megabases
MHC	major histocompatibility complex
miRNA	micro RNA
MKK	Maasai in Kinyawa, Kenya
MP	middle Paleolithic
MRCA	most recent common ancestor
mtDNA	mitochondrial DNA
MXL	Mexican ancestry in Los Angeles
MYA	million years ago
NAHR	non-allelic homologous recombination

ncRNA	non-coding RNA
NCS	non-coding sequences
NGS	next generation sequencing
NHEJ	non-homologous end joining
NHGRI	National Human Genome Research Institute
OoA	out of Africa
PCR	polymerase chain reaction
piRNA	piwi-interacting RNA
PUR	Puerto Rican in Puerto Rico
PWM	position weight matrix
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rRNA	ribosomal RNA
SBS	sequencing by synthesis
siRNA	small Interfering RNA
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA
SNV	single nucleotide variant
SV	structural variant
TF	transcription factor
TIRF	total internal reflection fluorescence

TMRCA	time to the most recent common ancestor
tRNA	transfer RNA
TSI	Toscans in Italy
UP	upper Paleolithic
VNTR	variable number tandem repeat
XP-EHH	cross-population extended haplotype homozygosity
YRI	Yoruba in Ibadan

Table of contents