Evolution by Gene Loss?

A genome-wide survey of human SNPs that introduce premature termination codons

Bryndís Yngvadóttir

Queens' College
University of Cambridge
September 2008

This dissertation is submitted for the degree of Doctor of Philosophy

UNIVERSITY OF CAMBRIDGE

Declaration

This thesis describes my work undertaken in the laboratory of Dr Chris Tyler-Smith, at The Wellcome Trust Sanger Institute, in fulfilment of the requirements for the degree of Doctor of Philosophy, at Queens' College, University of Cambridge. 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. 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 page limit specified by the Biology Degree Committee.

Bryndís Yngvadóttir
Cambridge, September 2008

Abstract

Nonsense-SNPs introduce premature termination codons into genes, and can result in the absence of a gene product or a truncated and potentially harmful protein, so are often considered disadvantageous and associated with disease susceptibility. As such, the disrupted allele might be expected to be rare and, in healthy people, observed only in a heterozygous state. However, some, like those in the caspase-12 and actinin-3 genes, are found at high frequencies with many homozygotes and seem to have been advantageous in recent human evolution.

The goal of this project was to perform a genome-wide survey of nonsense SNPs in the human genome and evaluate the selective forces acting on them. Most available nonsense-SNPs (n=805) and a set of synonymous control SNPs (n=731) were genotyped in 1,151 individuals from 56 geographically distinct worldwide populations.

I identified 169 genes containing nonsense-SNPs that were polymorphic in the samples, of which 99 were found in a homozygous state, showing that both copies of these genes can be truncated in healthy subjects without any obvious consequences. This study illustrates how much the human gene content varies between individuals: on average by 24 genes (out of about 20,000) by nonsense-SNPs alone. Gene Ontology analysis revealed that there was significant overrepresentation of genes involved in olfactory reception and the nervous system.

As might be expected, these SNPs as a class were found to be slightly disadvantageous over evolutionary timescales, but a few nevertheless showed signs of being advantageous, indicated by unusually high levels of population differentiation or a departure from neutrality in tests based on resequencing the region surrounding the SNP in multiple individuals. In addition to caspase-12, a *SEMA4C* nonsense-SNP was confined to the Americas where it reached high frequency, while a *MAGEE2* nonsense-SNP was present at high frequency only in East Asia and showed evidence of positive selection. Several examples of beneficial

gene loss could thus be found, and have contributed in a small but significant way to human evolution.

Publications

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

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C (2009). A genomewide survey of the prevalence and evolutionary forces acting on human nonsense-SNPs. *American Journal of Human Genetics*, **84**(2):1-11.

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, **Yngvadottir B**, Nica AC, Woodwark C, Chen Y, Ayub Q, Mehdi SQ, Li P, Tyler-Smith, C (submitted). Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation.

Yngvadottir B, Carvalho-Silva DR. (2008) *Reconstructing Human History Using Autosomal, Y-Chromosomal and Mitochondrial Markers*. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020819

Yngvadottir B. (2007) Insights into modern disease from our distant evolutionary past. *European Journal of Human Genetics,* **15**(5):603-6.

Xue Y, Daly A, **Yngvadottir B**, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J, Huckle E, Burton J, Leonard S, Rogers J, Tyler-Smith C. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *American Journal of Human Genetics*, **78**(4):659-70.

Gutala R, Carvalho-Silva DR, Jin L, **Yngvadottir B**, Avadhanula V, Nanne K, Singh L, Chakraborty R, and Tyler-Smith C. 2006. A shared Y-chromosomal heritage between Muslims and Hindus in India. *Human genetics,* **120**(4):543-551.

Acknowledgements

Firstly, I would like to thank my supervisor, Chris Tyler-Smith, whose office door was always open and who believed in me and helped me in every way throughout my PhD. I feel truly privileged to work with him. Many thanks go to Yali Xue who (in addition to giving me moral support and all the time in the world) has taught me so many things in the lab and helped me to understand the data and analysis in every detail possible.

Thanks go to Matt Hurles, Alex Bateman and Bill Amos for advice and support throughout the PhD. Thanks go also to Dan Turner who managed to teach me the ropes in the lab without getting impatient with my ignorance of lab-based techniques. I thank the sample donors, and Howard Cann for making the HGDP-CEPH data freely available. Thanks go to the many people passing through our Team 19 of human evolution in the past four years for their enthusiasm and expertise. In particular, I would like to mention Denise Carvalho-Silva, Tatiana Zerjal and Cara Woodwark. Thanks to all the great people in the Sanger Genotyping Platform and Large Scale Sequencing Pipeline groups, without whom I would have no data. I would especially like to thank Panos Deloukas and members of his team, Sarah Hunter, Pam Whittaker, Marcos Delgado and Rhian Gwilliam for the genotyping and QC processing. Special thanks go to Barbara Stranger and Manolis Dermitzakis for allowing me to make use of their gene expression data and then helping me understand what it was all about. I would like to thank Pardis Sabeti and Pat Varilly for advice on the LRH-tests and for giving me the source code for Sweep. Thanks go also to all the HelpDesk people at Ensembl and HapMap that I have bugged with endless queries throughout the years. Thanks to Areum Han who gave me the full data from SNP2NMD and to Ni Huang, Yuan Chen and Jim Stalker for various useful scripts. Great thanks to Joan Green and Andrew King for excellent "Journal Picks" pointers and for the endless renewal of my library books. I would also like to thank Christina Hedberg-Delouka deeply for all her support and kind words in the past years. Big thanks go to the Wellcome Trust for an excellent PhD program and for its generous fellowship that not only put a roof over my head but also allowed me to go to all those great conferences.

A special thanks go to Agnar Helgason for giving me a head start in my transition from the studies of social anthropology (old but not forgotten) to the exciting world of genetics—things are changing so fast in this line of work. Lots of love goes to my Sanger girls, Raffaella, Eleni and Antigone for "keeping it real". These past years have been tough, but because of you they have also been a joy ride of fabulous dinner parties, awesome red wine and some great vibes playing in the background. I'll miss us. Thanks go also to my Ice girls, Ellen, Þurý and Hulda, who have always encouraged me with their hugs and kisses, be they natural or electronic. I would especially like to thank my parents for their love and never-ending support, for standing behind me in the good times and the bad, and for making me less home-sick by making Cambridge their second home away from home. Thanks to Guðrún and Þorsteinn for being the greatest siblings a girl could hope for. And finally, my greatest thanks go to my love Bjarki, whom I could not have done this PhD without. Your support and understanding has been essential throughout the past eleven years and surprisingly so has your cooking for the past few weeks! Hopefully I can do the same for you in a year's time. This thesis is dedicated to you.

Table of Contents

Abbreviations

AMH	anatomically modern humans
ASO	allele-specific oligo
bp	base pairs
BP	biological process
CEU	CEPH Utah residents with ancestry from northern and western Europe
CHB	Han Chinese in Beijing
CNV	copy number variant
DAF	derived allele frequency
DAVID	Database for Annotation, Visualization and Integrated Discovery
EHH	extended haplotype homozygosity
GC	Gene Call
GO	gene ontology
HGDP-CEPH	CEPH Human Genome Diversity Cell Line Panel
HGMD	Human Gene Mutation Database
HLA	human leukocyte antigen
HWE	Hardy-Weinberg Equilibrium
iHS	Integrated Haplotype Score
JPT	Japanese in Tokyo
kb	kilobases
KYA	thousand years ago
LD	linkage disequilibrium
LNP	lactase nonpersistance
LP	lactase persistance
LRH	long-range haplotype
LSO	locus-specific oligo
LWK	Luhya in Webuye, Kenya
MAF	minor allele frequency
Mb	mega base
MF	molecular function
MYA	million years ago
NCBI	National Center for Biotechnology Information
NMD	nonsense-mediated mRNA decay
OR	olfactory receptor
ORF	open reading frame
PCR	polymerase chain reaction
PHASE	Phylogenetics And Sequence Evolution
PTC	premature termination codon
REHH	relative extended haplotype homozygosity
SNP	single nucleotide polymorphism
STS	sequence tag site
UCSC	University of California Santa Cruz
VNTR	Variable number of tandem repeat
WGA	whole-genome-amplification
WTCCC	Wellcome Trust Case Control Consortium
WTSI	Wellcome Trust Sanger Institute
YRI	Yoruba in Ibadan, Nigeria