The Evolutionary Role of Human-Specific Genomic Events

Yuval Itan

SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY COLLEGE LONDON (UCL)

September 2009

Centre for Mathematics & Physics in the Life Sciences and Experimental Biology (CoMPLEX)
Department of Genetics, Evolution and Environment
UCL

Supervisor: Prof. Mark G. Thomas
Second Supervisor: Dr. Kevin Bryson

Declaration of ownership.

I, Yuval Itan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract.

In the short evolutionary time since the human-chimpanzee divergence, approximately 6.6 million years ago, humans have acquired a range of traits that are unique among primates. These include tripling brain size, enhanced cognitive abilities, complex culture, descended larynx structure that enables spoken language, longevity, specific diseases, inferior olfaction, and (in some human populations) adult lactase persistence. These traits were likely to have evolved through various genomic mechanisms, among them gene duplications and gene-culture co-evolution. Several studies have estimated the dates for some of these human lineage genomic events. However, no study to date has performed a genomewide estimate of the dates of all human gene duplications. Moreover, as many of these traits were likely to have evolved via gene-culture coevolutionary mechanisms, investigating the evolution of one of these human-specific traits – lactase persistence – provides a model example for in-depth future investigations of specific human phenotypes.

In this study I have investigated an important class of human-specific genomic events – gene duplications (otherwise known as human inparalogues). I have developed a new bioinformatics approach for detecting human lineage-specific inparalogues and the duplication dates for those genes. I show that human-specific inparalogues are non-randomly distributed among biological function classes, and their duplication event dates are non-randomly distributed on a timeline between the date of the human-chimpanzee split and the present. I have also investigated the evolution of the human-specific polymorphic trait – lactase persistence. I have performed a worldwide correlation analysis comparing frequency data on all currently known lactase persistence-associated alleles and the distribution of the lactase persistence phenotype in different human populations. I have also performed a gene-culture co-evolution analysis, employing spatially explicit simulation and Approximate Bayesian Computation to condition simulations on genetic and archaeological data, in order to make inferences on the evolution of lactase persistence and dairying in Europe.

Table of Contents.

List of Figures.

List of Tables.

Abbreviations.

ABC	Approximate Bayesian Computation
AMH	Anatomically Modern Human
ANND	Average Nearest Neighbour Distance
BG	Blood Glucose
BH	Breath Hydrogen
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST Like Alignment Tool
BP	Before Present
CD	Cultural Diffusion
cDNA	Coding Deoxyribonucleic Acid
CDS	Coding Sequence
CNV	Copy Number Variation
COG	Clusters of Orthologues Groups
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DD	Demic Diffusion
DNA	Deoxyribonucleic Acid
F_d	Dairying Farmers
F_{nd}	Non-dairying Farmers
GB	Genetic Background
GC	Guanine/Cytosine
GO	Gene Ontology
HG	Hunter Gatherers
INPARANOID	In-paralogue and Orthologue Identification
KYA	Thousand (i.e. Kilo) Years Ago
LBK	Linearbandkeramik
LP	Lactase Persistence
MAFFT	Multiple sequence Alignment employing Fast Fourier Transform
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MRCA	Most Recent Common Ancestor
mtDNA	Mitochondrial Deoxyribonucleic Acid

MYA	Million (i.e. Mega) Years Ago
MYH	Myosin Heavy Chain
OR	Olfactory Receptors
PAM	Partitioning Around Medoids
RNA	Ribonucleic Acid

Acknowledgements.

I would like to thank the following for funding my studies over the last 4 years: UCL Graduate School, ORS, B'nai B'rith, Annals of Human Genetics, and the Anglo-Jewish Association. I owe special thanks to the three Israeli individuals that prefer to remain anonymous who were so kind to financially support me during the first year of my studies. I thank the following for their professional and scientific help: Richard Emes, the very patient Ensembl Helpdesk (especially Bert Overduin), and the Perl Monks (may their tormented souls find some well deserved peace). I thank Ziheng Yang for his valuable time and advice. Many thanks to all the members of CoMPLEX, especially Andrew Pomiankowski (POM), Rachel Wolfson, and Hugh McCready who passed away.

Thanks to all my UCL friends that made work and Friday pub lunches even more enjoyable: Pascale Gerbault, Anke Liebert, Bryony Jones, Sarah Browning, Adam Powell, Laura Horsfall, Lauren Johnson, Rosemary Ekong, Ayele Tarekegn, Catherine Ingram, Krishna Veeramah, Chris Plaster. I thank Dallas Swallow for her time and patience for all my random questions. I owe special thanks to Neil Bradman for his kind interest and support, as well as his wife Gwen for her warm hospitality.

I thank my family, including my parents Hagay and Leah Itan, all my 4 sisters, my grandparents and my cousins Yael Rosenblum, Daniela Bar-El, and her husband Oshri. Thanks to my Israeli friends Dror Fried, Adar Paz, Gur Pines, Neta Eckstein, Michal Wolff, Eyal Lewinson, Yuli Barkan, Benny Vazana, Michal Eldar, Raz Liebreich, Yono Cohen, Doron Musel, Eyal Kalie, Lewis Gelfand and many more. Thanks to my Euro friends Marie-Luise (a.k.a. Ise) Mechias, Inge Broekman the Dutchie one, Anne Steinbrück the Deutsche sister, David Burton, and Michael Bailey (an Australian exile in European lands).

I would like to thank all off-work people/places/other that created an interesting variety: UCL Basketball team, London Shootfighters club, Notting Hill Arts Club, Clapham Book Club and the brilliant Victoria Howes. I thank budget airlines and my Beetle that enabled my travelling while making the European climate cosier, my electric guitar (and amp) for a neighbour-friendly therapy, and the Beatles, Pink Floyd and Nirvana for accompanying me in the long dark hours of the programming. I thank my favourite Blackheath neighbours: the polydactyl (i.e. six toed) ginger female cat for helping me understand who really owns the territory, the nocturnal churchyard foxes that made it feel like in the bush, the huge pigeon that unlawfully targeted the Beetle (I forgive him for everything), and Faye and Jared the friendly humans downstairs. I thank John Guillan for being, probably, the coolest landlord ever.

Many thanks to my dedicated supervisors Mark Thomas and Kevin Bryson that trusted me to explore my ideas while providing invaluable advice and directions when needed. They have made possible these four projects that I hope you'll enjoy reading. And this section will be incomplete without huge thanks to Zippy, Danny, and Orit(i) Pinchevsky, my mother, father, and sister in law, respectively. I will finish with the most important thanks to my beautiful and loving wife Yael Pinchevsky, my best friend and the best fashion designer in town.

1. Introduction.

1.1. Rationale of the Study.

Charles Darwin's theory of natural selection (Darwin, 1872) indirectly implied that the emergence of modern human has been a product of slow evolutionary process for which all organisms are, and have been, subjected – descent from earlier organisms, rather than an organism which is above and unrelated to other species.

After Darwin's revolutionary breakthrough, the discoveries of ancient hominid fossils in Africa during the 1920's have driven forward the scientific field of palaeoanthropology – the study of ancient human fossils – the first scientific field that dealt exclusively with human evolution (Figure 1.1). Palaeoanthropology had confirmed for the first time that human has evolved on a time scale of millions of years. The discovery of structure of the DNA ((Watson and Crick, 1953), popularly credited mostly to James Watson and Francis Crick, but involved to a large extent the work of Rosalind Franklin, Maurice Wilkins, and Raymond Gosling) has led to the establishment the central dogma of molecular biology: DNA→RNA→Protein (Figure 1.1).

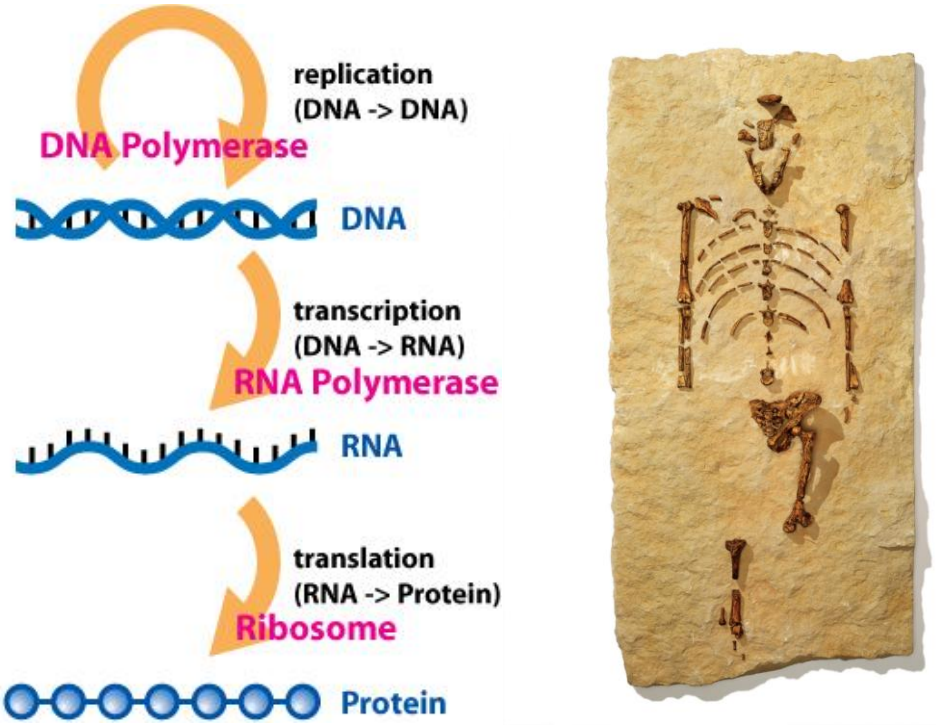

Figure 1.1. The two main methodologies used in human evolution studies. Molecular biology on the left and palaeoanthropology on the right. The left part of the Figure is a basic illustration of the dogma of molecular biology which is the basis for molecular genetics, and on the right is "Lucy" – the famous *Australopithecus afarensis* specimen dated from about 3.2 million years ago. Left image credit: Daniel Horspool. Right image credit: The Houston Museum of Natural Sciences.

Human evolution is a broad subject that has been contributed to by various disciplines. Palaeoanthropology provides evidence for human evolution through major morphological changes (notably bipedalism and large cranium) that differentiate modern human from chimpanzee, human's closest living relative, and all other primates (see section 1.3.1). Molecular evolution is the scientific field studying evolution at the DNA, RNA, and protein scales (see sections 1.3.2 and 1.6), and can be applied for research of the various genomic processes that have contributed to the modern human phenotype. Recent advances in molecular biology techniques allowed efficient and cost effective sequencing of the genomes of different species, among them human and chimpanzee. With the availability of these genomes, large scale interspecies comparisons and analyses provided new insights into human genomic evolution. Recent human population genetic variation studies give a different angle of human evolution – genetic differences among individuals and various human groups. Anthropological observation of humans and primates provide insights about the evolution of human behaviour, while archaeology studies the material evidence and hypotheses about the evolution of modern human culture.

When considering the vast subject of the evolution of the human phenotype, one may ask "what makes human special". Being human-centric rather than comparative, this question is likely to shift the study from the broader scope that I preferred my studies to follow. I consider the following question as more apt to ask: "what makes human different from chimpanzee". Indeed this is a vast (and arguably too general) question to ask, but with the recent availability of human and chimpanzee genomic data (Hubbard et al., 2009, Smedley et al., 2009, Lander et al., 2001, Mikkelsen et al., 2005), the integration of anthropology and archaeology with these data (Mace, 1993, Burger et al., 2007, Pinhasi et al., 2005), and high performance computing for bioinformatics analyses and computational simulations (Remm et al., 2001, Katoh et al., 2002, Kent, 2002, Itan et al., 2009) – it is now possible to start to address one aspect of this question: what are the genomic events that have lead to the human phenotype. This question can be tackled from different angles. I will do this through four studies: (1) detecting all gene duplication candidates in the human lineage; (2) estimating the dates and the functionalities of the duplicated genes found in (1) and thus correlating human genomic events with palaeoanthropological data; (3) worldwide Correlating lactase persistence (a trait unique to human) genotype and phenotype; and (4) modelling the origins and evolution of lactase persistence in Europe.

Studies (1) and (2) provide a large scale understanding of one genetic event class – gene duplication – that is likely to have played a strong role in shaping the human phenotype. Studies (3) and (4) are case studies of the evolution of one human-specific trait that provided people with a very strong selective advantage (Bersaglieri et al., 2004, Ingram et al., 2009a, Itan et al., 2009). The interdisciplinary nature of the full study requires the integration of different data types across life sciences, computer science and mathematics, and social sciences. Altogether, this whole work is under one umbrella: human-specific genomic events. I hope that this work will provide novel and important advances to the field of human evolution, especially the "human-specific", and will be a good framework for future studies in this field.

1.2. The Human Specific Phenotype: Human-Chimpanzee Differences.

To better understand the aspects of the human specific phenotype, the most obvious approach is to look for traits that exist in human but not in chimpanzee. In this section I will review some of the significant traits that are unique to humans among apes. The evolutionary perspective of the human phenotype will be reviewed in section 1.3.

The **human brain** weighs 1,300 – 1,400 grams on average, while the chimpanzee's brain weighs 420 grams on average (Nieuwenhuys et al., 2007). A common method to estimate the cognitive ability of a species (especially when comparing evolutionary close species) is to calculate the percentage of body weight made up by the brain. In humans the brain consists of about 2% of the body weight, while in chimpanzee the ratio is about 0.8% on average – about 2.5 times less than human: the human brain is about 3 times bigger (Carroll, 2003). The increased size of the human brain is mostly explained by increase in the size of the cerebral cortex, the largest brain structure and the location of most higher cognitive functions (Kornack and Rakic, 1998). The human cerebral cortex also shows functional asymmetries which are much more significant than in chimpanzee – most humans are right handed and have the language function located in the brain's left hemisphere, while chimpanzees show much weaker asymmetry in handedness (Hopkins and Cantalupo, 2004).

The **larynx** (also called the voice box) is the organ located inside the mammals' respiratory tract, which has a function of protecting the lower respiratory tract from passage of food and foreign particles. The larynx contains the vocal chords which produce vocal sounds. Humans are the only primate that has a descended larynx – humans are incapable of raising the larynx high enough so it will connect with the nasal passage. This human-specific (among apes) characteristic of the larynx is, interestingly, shared with some aquatic mammals, which has lead to the controversial *aquatic ape hypothesis*, which maintains that some unique human characteristics (such as descended larynx, hairlessness, and bipedalism) had evolved through a period of humans inhabiting aquatic environments (Morgan, 1999). The descended larynx has extended the length of human vocal tract, and so it is suggested that it was a crucial element in the development of **speech** and **language** – a major difference between the human and the

chimpanzee phenotypes (although descended larynx evolved in other mammals and vertebrates, such as red-deer stags and birds) (Fitch and Reby, 2001).

Obligate bipedalism is defined as locomotion on two legs that is the organism's only alternative. Obligate bipedalism is unique for humans among primates, and had evolved in various mammals, reptiles, and birds. Human bipedalism enabled carrying food for long distances, the potential of handling tools, as well as the ability to run for long distances. As human bipedalism evolved early in human history, it is likely that is was a key element in later major developments of the human phenotype (Hunt, 1994). The different perspectives regarding the evolution and function of human bipedalism will be discussed in section 1.3.

A significant difference between human and chimpanzee is human's **longevity** – the average life expectancy of humans in places with good health conditions is about 80 years (from CIA – The World Factbook, https://www.cia.gov/library/publications/the-world-factbook/), while the life expectancy for chimpanzee in captivity is about 50 years, and less in the wild (Jones et al., 1996). One key element for human longevity is the human growth rate and maturation process, which is slower in human than in chimpanzee. A human infant is helpless and totally dependent on the mother for a minimum of 2-3 years, while in chimpanzee, total dependency is only for a few months. The slow development of human infants is strongly related to their brain development – reaching one quarter of its final size at birth and half of the final size after one year, while a chimpanzee is born with its brain already half the final size (Campbell, 1999). Several genes and genetic pathways that may be involved in human aging and longevity have been identified (Browner et al., 2004). More implications of the long period of immaturity and longevity in human will be discussed in section 1.3.3.

There is a range of **diseases and disorders** that are unique to human. Among them is autism – a brain development disorder that is likely to have a complex genetic basis of interaction between multiple genes, the environment, and epigenetic factors (Amaral et al., 2008, Abrahams and Geschwind, 2008); Alzheimer's disease – the most common form of dementia, a degenerative terminal disease with causes that are only partly understood, associated with amyloid plaques (dead cells and protein deposits) and neurofibrillary tangles (overactive enzymes resulting in neuron cells death) in the brain

(Tiraboschi et al., 2004); and the acquired immunodeficiency syndrome (AIDS) – a modern incurable disease of the immune system caused by susceptibility to the human immunodeficiency virus (HIV). The HIV virus is thought to have been originated from the primates' SIV virus, which is non-pathogenic (Sepkowitz, 2001). Another significant disease unique to humans is smallpox – a potentially lethal infectious disease that is thought to have originated about 10,000 years ago, caused by two virus variants. Smallpox is thought to have caused 300-500 million human deaths during the 20th century (Barquet and Domingo, 1997). Smallpox is unique in being the only human infectious disease that had been completely eradicated (in 1979) after successful vaccination campaigns (Barquet and Domingo, 1997).

Culture is defined as an "integrated pattern of human knowledge, belief, and behaviour that is both a result of and integral to the human capacity for learning and transmitting knowledge to succeeding generations. Culture thus consists of language, ideas, beliefs, customs, taboos, codes, institutions, tools, techniques, works of art, rituals, ceremonies, and symbols. It has played a crucial role in human evolution, allowing human beings to adapt the environment to their own purposes rather than depend solely on natural selection to achieve adaptive success" (Britannica Concise Encyclopaedia, 2006). Culture is not strictly a phenotype, and it is argued that culture is not unique to human since it was characterised in different chimpanzee communities (Whiten et al., 1999). This is an open discussion which is beyond the scope of this work. However, three features have been suggested for a distinction of modern human culture from chimpanzee's (Tomasello, 1999): (1) Creating and using of conventional symbols, including written language and mathematical symbols and notations; (2) Creating and using complex tools and instrumental technologies; and (3) creating and participating in complex social organization and institutions. Because of these reasons, I consider "human culture" to be a significant human specific phenotype. See section 1.3.3 and (Powell et al., 2009) for the evolution of modern human behaviour and culture.

In this section I have briefly reviewed some significant elements of the human-specific phenotype. This is by no means a comprehensive list, but rather an attempt to give a broad perspective of the human-chimpanzee phenotypic differences.

1.3. The Human Phenotype Evolution.

Human evolution, the process that has shaped the modern human phenotype, is a very broad subject that was traditionally tackled by the fields of palaeoanthropology (fossil record) and archaeology (evidence for early human culture). With the advances of molecular biology techniques, human genomics is now being incorporated into the human evolution studies. In this section I will review significant human evolutionary events from human-chimpanzee divergence until present, separately bringing examples for each methodology of research. Climate and ecology played crucial roles in the evolution of early and modern human. However, these vast subjects will be only briefly discussed in this section since they are not a major aspect of my study, and they will be described in the different chapters whenever relevant (particularly in chapter 5).

1.3.1. Palaeoanthropology Perspective on Human Phenotype Evolution – Fossil Record and Morphology.

Hominids (the anglicised form of "Hominidae") is the genera of human and all extinct species since the human-chimpanzee divergence about 6.5 million years ago (mya) (Jobling et al., 2004). Figure 1.2 is an estimate of the hominid evolution timeline and phylogenetic relationships. In this section I will present the morphological evolution of the major genera and species leading to modern human. Note that due to sparse data and the nature of reconstruction techniques and inference in the field of palaeoanthropology, major disagreements are common among scientist in the field, so it is likely that each element presented here would be controversial among some researches in the field. However, I will attempt to present those among which there seems to be general agreement.

The earliest known hominid (that is, relatively, non controversial) is *Orrorin tugenensis*, "Millennium man", from Kenya, dated about 6mya. This species fossil includes a fragmentary thigh bone – indicating some degree of bipedalism – and thick enamelled molars that relate *Orrorin tugenensis* to the human lineage rather than to the chimpanzees (Senut et al., 2001).

Most fossils dating after about 4.2mya and until the appearance of *Homo* are of the genus *Australopithecus* (Jobling et al., 2004). The genus is assumed to have been bipedal, with evidence including the Laetoli volcanic ash footprints (Leakey and Hay, 1979) and "Lucy" – a well preserved partial skeleton – both belonging to the *A. afarensis* species (Figures 1.1 and 1.2). The most significant discovery about Lucy was her valgus knee – a strong indication for bipedalism. *Australopithecus* brain / body mass proportion was similar to chimpanzee. *Australopithecus* was 1-1.5m tall. It is suggested that a loss of body hair gradually took place in parallel with more modern species of *Australopithecus* as they became fully bipedal between 2 and 3mya (Wheeler, 1984), which leads to the theory that loss of hair contributed to the evolution of dark skin (Jablonski, 2008). The ongoing question of which *Australopithecus* species – if any – is the direct ancestor of *Homo* is controversial. For many years *Homo habilis* was considered to be the link between the *Australopithecus* and the genera *Homo*, based on evidence of a partial skull and jaw fossils from Olduvai Gorge, Tanzania dated 2.5mya (Jobling et al., 2004, Leakey et al., 1964). However, *habilis* does not show all characteristics of *Homo*: it has larger teeth and different body size and shape (the "body size and shape" is described differently among the different researchers studying *Homo habilis*), and thus it is now generally agreed that *habilis* was an extinct branch of the *Australopithecus* genus (Jobling et al., 2004). The first *Homo* species that is generally agreed to be distinct from *Australopithecus* is *Homo ergaster*, and its first fossils are dated from about 2mya (Wood and Collard, 1999). A theory that was widely accepted claimed that *Homo ergaster* and *Homo erectus* were two separate species, where the former lived in Africa and the latter outside Africa. However, the difficulty in making significant morphological distinction between the two species and the finding of a fossil in Africa dated 1mya and having all the erectus characteristics – has lead to the current prevalent theory that ergaster and erectus were one widespread species, and so I will now term both as *Homo erectus* (Asfaw et al., 2002, Jobling et al., 2004). The best preserved and complete early hominid skeleton is the "Nariokotome Boy" from Lake Turkana, Kenya, dated about 1.6mya (Walker and Leakey, 1993). The fossil shows some modern human characteristics of body and brain size. Mature *Homo erectus* male was estimated to have reached 1.8m tall and weighing 70kg, while its brain size was estimated to be 909cc, significantly smaller than mature modern human average brain size (1,450-1,500cc) and about 60% of the modern human brain / body mass proportion, but yet within the range of modern human brain size (830-2,300cc) (Clegg and Aiello,

1999). It has been suggested that this branch of *Homo erectus* survived until 27 thousand years ago (kya) in Java, which would have made them contemporaries of modern humans, while the *Homo floresiensis* species is thought to have survived in Flores until 12,000 years ago, making it the latest lasting non-human hominid (Swisher et al., 1994, Jobling et al., 2004, Morwood et al., 2005).

The definition of the different species (notably *Homo mauritanicus* and *Homo heidelbergensis*) in the genus dated from about 1mya until about 200kya (*Homo erectus* to *Homo sapiens*) is disputed, and so these species are generally termed as archaic *sapiens* (Jobling et al., 2004). Archaic *sapiens* had a less robust bone and muscle structure, and had larger brains, around 1,200cc. See section 1.4.1 for the potential genetic trigger for this significant brain expansion.

Homo neanderthalensis (Neanderthal) is a distinct branch of archaic *sapiens* that inhabited Europe and western Asia between 250 and 28kya, having a robust bone structure and a large brain, around 1400cc (Jobling et al., 2004). Whether Neanderthals interbred with modern human and contributed to the modern human gene pool is a matter of great controversy (Tattersall and Schwartz, 1999, Serre et al., 2004). However, Neanderthal ancient mitochondrial DNA (mtDNA) studies show that Neanderthals did not contribute to modern human mtDNA diversity (Serre et al., 2004).

The origins of anatomically modern human (AMH) is, yet again, a matter of great controversy among palaeoanthropologists. A recent study had made a system for distinction of AMH from archaic *sapiens*, where the main distinct AMH features are the globular shape of the skull and the degree of retraction of the face (Lieberman et al., 2002). The earliest AMH fossils found are the Omo remains, dated about 198kya (Fleagle et al., 2008) and fossils from Herto (Ethiopia), dated about 154-160kya, with the AMH features of a 1450cc brain and a globular skull, and the archaic feature of protruding brows (White et al., 2003). These early AMH fossils show post mortem modifications including cut marks – an indication for mortuary practices. However, these specimen are described today as *Homo sapiens hidaltu* - a sub-species predating *Homo sapiens* (White et al., 2003). The earliest known fully modern Homo sapiens fossils are from Omo-Kibish (Ethiopia, discovered by Richard Leakey in 1967), dated about 130kya, with a controversial recent study dating these fossils to be 196kya

(McDougall et al., 2005). Mitochondrial DNA studies estimate that "Mitochondrial Eve" (the most recent common matrilineal ancestor of AMH) is dated 171±50kya (Ingman et al., 2000, Gonder et al., 2007).

Both fossil and genetic evidence support the Out of Africa theory – where AMH evolved exclusively in Africa between 200-100kya and then a branch left Africa about 55-70kya and gradually replaced native *Homo erectus* and Neanderthal populations (Liu et al., 2006). The competing theory is the Multiregional Hypothesis which maintains that the evolution of AMH had been continuous and worldwide, and within only one human species (Wolpoff et al., 1988).

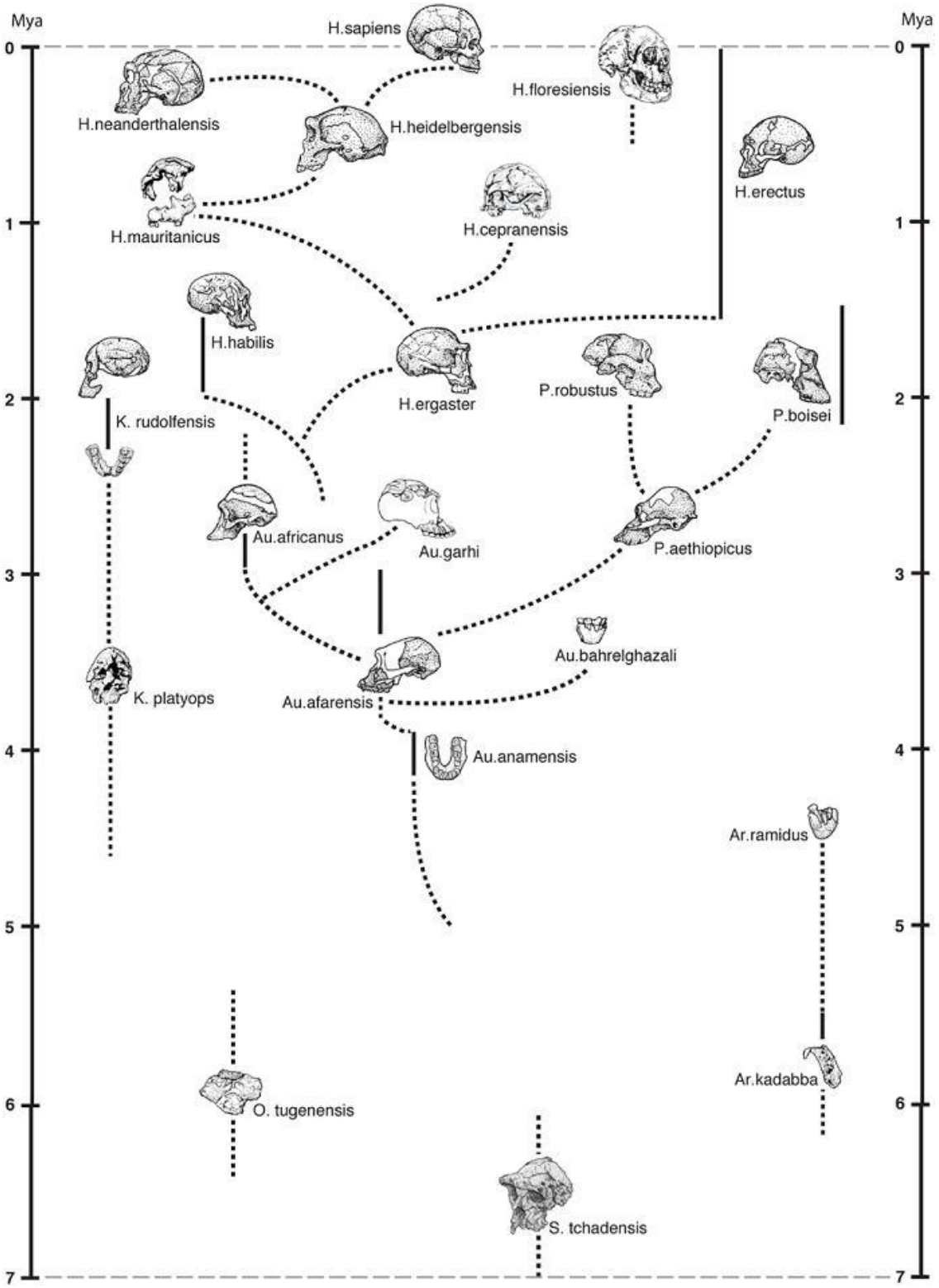

Figure 1.2. Phylogenetic relationships within the family Hominidae. The timeline is on the vertical axis. Solid lines show stratigraphic ranges - assigning time ranges for fossils. This diagram shows that typically several different hominid species have coexisted at any one point in time, and it is the exception that Homo sapiens is the lone hominid in the world today. Image credit: Ian Tattersall, American Museum of Natural History.

1.3.2. Genomic Perspective Human Phenotype Evolution – The Various Types of Genomic Events.

In this section I will review the different classes of large scale genomic events that are likely to have contributed to the evolution of human phenotype in the relatively short evolutionary time of about 6.5 million years (Steiper and Young, 2006). For notable examples of genomic events that contributed to the early and modern human phenotype see section 1.4.

Gene duplication is mainly caused by unequal crossing over or retroposition (Koonin, 2005) (Figure 1.3). **Unequal crossing over** occurs during the meiosis in regions of repetitive DNA, when the two homologous chromosomal regions are not precisely aligned (mismatched). While regular crossing over results in identical lengths of DNA exchanged between the two chromosomes, in the case of unequal cross over one of the chromosomes receives extra DNA sequence while the other chromosome loses it. The result is a **segmental duplication** of a region (which may contain a gene or set of genes) being duplicated in one chromosome (Koonin, 2005, Cheung et al., 2003). **Retroposition** is a process whereby repetitive DNA fragments are inserted into the chromosome by reverse transcription from mRNA molecules. Retroposition accounts for about 1,000 duplicated genes in the human genome (Emerson et al., 2004). A seminal work regarding the fate of the duplicated genes suggests that one copy maintains the original functionality of the gene, while the other copy "escapes" the constraint of purifying selection, and thus becomes "free" to accumulate genetic mutations that might give rise to novel functionalities (neo-functionalisation) or loss of function (non-functionalisation) (Ohno, 1970). Later experiments on duplicated gene expression levels have shown that extant gene pairs might partition between them the functions of the single ancestral gene (Prince and Pickett, 2002). The Sub-functionalisation model (also called the duplication-degeneration-complementation model) proposes that the two gene copies acquire complementary loss of function, and together they produce the full functionality of the ancestral gene (Force et al., 1999). A study that tested expression levels of gene copies in various human tissues provides an example for acquisition of new function through gene duplication (by retroposition) in human lineage. The study has found several cases where one gene copy was expressed in several different tissues while the other copy was expressed exclusively in testis

tissues (Marques et al., 2005). For more examples of human gene duplication studies and for my own research of human gene duplications see chapters 2 and 3. **Pseudogenes** are DNA sequences with features that resemble conventional genes, but that do not code for viable proteins, mostly due to stop codons and frameshifts (Figure 1.3). **Processed pseudogenes** emerge via retrotransposition – a portion of mRNA that is reverse transcribed back into the genome, inserting a new sequence lacking regulatory elements and thus being non functional – "dead on arrival" (Graur et al., 1989). **Non-processed pseudogenes** evolve after a gene duplication event, where one copy retains the original function and the other becomes dysfunctional (Wang et al., 2006). **Unitary pseudogenes** are elements of rapid evolution – where the only copy of a functional gene becomes dysfunctional, the genotype is fixed in the population (mostly via genetic drift or a population bottleneck), and the loss of function can give rise to new functionalities – the "less is more" hypothesis (Olson, 1999). Recent studies show that some genes that were traditionally annotated as pseudogenes are actually functional (coding to proteins) using alternative molecular mechanisms (Zheng et al., 2007, Zheng and Gerstein, 2006). A notable example for pseudogenization in the human lineage is the loss of olfactory receptor genes (Gilad et al., 2003), which will be discussed in section 1.4.1.

Gene fusion is a **chromosomal rearrangement** event where two separate genes form a hybrid gene, following a recombination event. When gene fusion happens in non-coding regions it may affect the regulation and expression of the gene, while gene fusion in coding regions may lead to new functionalities of the hybrid gene (Durrens et al., 2008). A major human lineage chromosomal rearrangement event by gene fusion is the fusion of the chimpanzee's chromosomes 12 and 13 into one chromosome in human, which is termed human chromosome 2 for annotation convenience (Shimada et al., 2005). It is proposed that the fusion of the UPS32 and TBC1D3 genes in the hominoid lineage has strongly contributed to the hominoid speciation (Paulding et al., 2003). In **gene fission** a gene splits into several parts by either recombinatorial or single-base mutation events, which can result in changes in regulation, production of a less complex protein due to domain deletion, or pseudogenization (Durrens et al., 2008).

A regulatory region is a DNA sequence, mostly upstream of the coding sequence of a gene, where transcription factors and other regulatory proteins can bind preferentially and thus regulate the expression levels of the gene (Stepanova et al., 2005). A

genomewide study has investigated transcription factor sites that are conserved among chimpanzee and mouse while absent in human. The study has shown that the human lineage loss was not random, but rather correlated to the biological function of the associated genes, which have an over-representation of sensory perception functions. This study suggest that these genes may highlight potential pathways underlying human-chimpanzee divergence (Donaldson and Gottgens, 2006).

Retroviral insertion is executed by retroviruses, which are unique among RNA viruses in their ability to integrate DNA copies of their genomes into the genome of the infected cell. On occasion, integration takes place in a human germline cell, giving rise to a human endogenous retrovirus (HERV), which can be inherited by the offspring of the infected host, and may eventually become fixed in the gene pool of the host population (Johnson and Coffin, 1999). The pathological effects of HERVs include susceptibility to cancer and autoimmune diseases (Lower, 1999, Dunn et al., 2003), while it was suggested that HERVs may have beneficial roles in protection against exogenous retroviral infection and in the formation of the placenta (Sverdlov, 2000, Villarreal, 1997). A study of polymorphic HERVs among different human populations shows that HERVs can be applied as good population genetics and forensics markers (Herrera et al., 2006).

"**Epigenetics** refers to heritable changes in gene function that do not change the DNA sequence but, rather, provide an "extra" layer of transcriptional control that regulates how genes are expressed" (Rodenhiser and Mann, 2006, Egger et al., 2004). Although not strictly genomic events, epigenetic effects have a direct influence on human genomics, and thus it is feasible to include them in the genetic category. Epigenetics effects of gene expression regulation are performed via the mechanisms of DNA methylation and histone modifications (Feinberg and Tycko, 2004). The loss of normal DNA methylation patterns may result in various human diseases that relate to X chromosome inactivation (Avner and Heard, 2001), genomic imprinting (Verona et al., 2003), and cancer (Feinberg and Tycko, 2004). I could not find any human evolution study for detecting human-lineage epigenetic elements and their functionalities when compared to chimpanzee. This subject is beyond the scope of this study, and is suggested as a potentially important future study.

Of the genomic classes presented in this section, human gene duplications are the most relevant to my research (especially regarding chapters 2 and 3) since they are genomewide events for which their date can be estimated. See chapters 2 and 3 for further details. Pseudogenization can also be potentially dated, but this is beyond the scope of this study, as will be discussed in chapters 2 and 3.

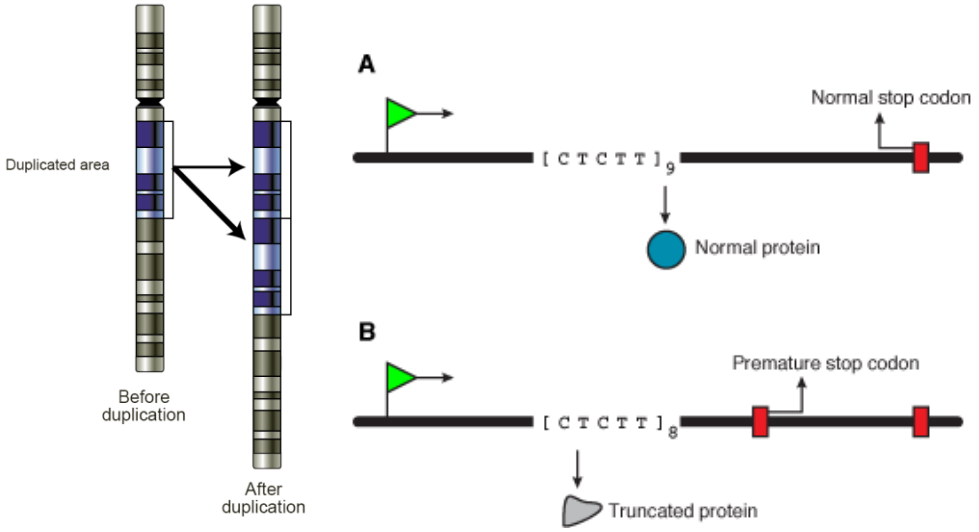

Figure 1.3. Illustrating gene duplications and pseudogenization. The left image illustrates a segmental duplication of a region that includes a gene, and the right image illustrates a premature stop codon that results in a dysfunctional truncated protein. The left image is taken from NHGRI, a public domain, and the right image is taken from (Craig, 2003).

1.3.3. Cultural Perspective Human Phenotype Evolution – From Early Human to Farming and Modernity.

The cultural evolution of humans from the time of human-chimpanzee divergence until several thousand years ago is mostly investigated by archaeology, the science that aims to understand pre-historical human culture, mostly through recovery of human material remains such as artefacts, architecture, and more (Aldenderfer and Maschner, 1996). This section will briefly review the archaeological evidence for human lineage culture, from early humans tool use, through ancient art, and the transition from modern humans hunting-gathering to a farming society.

Modern chimpanzees use a variety of tools for gathering food: sticks for extracting termites from mounds and stones for breaking open nuts (Whiten et al., 1999, Jobling et al., 2004), most of these tools would not be preserved in archaeological records, and

would not be identified as distinct from natural objects (Mercader et al., 2002). Since no tools were found for the *Orrorin* and most *Australopithecus* genera timeline, it could be assumed by parsimony that *Orrorin* and early Australopithecus have had a culture equivalent to the chimpanzee genus (Jobling et al., 2004).

Archaeological records begin 2.5mya with the tools of the Oldowan culture at the Olduvai Gorge, Tanzania. The makers of these tools are likely to be *A. habilis* – a relatively modern *Australopithecus*, before the transition to *Homo*. The tools include hammerstones, flakes, and cores, and it is assumed that they were used for scavenging large animal carcasses and breaking open bones for the highly nutritional bone marrow. This could give the tool users advantage over other scavengers, such as hyenas, that could not break these bones open (Napier, 1962).

A major shift occurred about 1.65mya; symmetrical teardrop shaped handaxes started to be made by *Homo erectus* in West Turkana, Kenya (Scarre, 2009). The technology and the culture of manufacturing these tools is termed Acheulean (after the French site St. Acheul). This technology was so successful that it dominates the Old World archaeological record until about 150kya. The tools were potentially used for tree hacking, and cutting carcasses and hides. It is likely that these stone tools were combined with other materials to create more sophisticated tools such as spears (Thieme, 1997). Over-sophistication of some of the tools (beyond needed functionality) suggests that the tools also served for social interactions (O'Brien, 1981) and as early means of artistic expression (Mania and Mania, 1988). It is hypothesized (controversially) that the Acheulean tool users possessed the ability for early language, because the parts of the brain that are correlated to the fine control required for the tool construction are also correlated to speech (Isaac, 1976).

The control of fire was a cornerstone in human history. It introduced cooked proteins and carbohydrates into the human diet, and allowed the extension of activity into night time, while providing protection from predators (Price, 2005). The earliest evidence for hominin (*Homo erectus*) use of fire is red clay sherds dated about 1.42mya, from various sites in East Africa (James, 1989). It is hypothesized that the change of diet as a consequence of fire control allowed humans to absorb more calories and as a result triggered brain expansion (Wrangham and Conklin-Brittain, 2003). The earliest

indication for fire being used as an engineering tool – treatment of stone tools – is from about 164kya, Pinnacle Point, South Africa. This use of fire required an elevated cognitive skill, and is associated with widespread evidence for symbolic behaviour from the same time range (Brown et al., 2009).

The use of art is often associated with modern human culture. Examples of abstract art in various South African sites are dated back to 75kya (Henshilwood et al., 2002). The oldest example of figurative art, a proxy for advanced symbolic communication, is of the Upper Palaeolithic Aurignacian culture in Schelklingen, Germany, dated about 40kya (Conard, 2009).

Hunting-gathering (HG) was the subsistence method for humans since 2mya and until about 10kya, when farming was introduced. A HG society obtain most food (about 80%) by gathering, and the rest by hunting (Barnard, 2004). The social and cultural structure of the HG is often being inferred by modern anthropological studies of such indigenous societies. It is thought that HG had a non-hierarchical society and mostly nomadic, and thus not tending to store food or support a full-time leaders class or artisans (Gowdy, 1997). The HG lifestyle required a wide territory for each individual (in comparison to farmers), and so the carrying capacity – the maximum number of individuals per area – of such societies was low. HG carrying capacity estimates vary and depends on several factors (see chapter 5), and a rough average approximation would be 0.1 individuals per km^2 (Bellwood, 2005). The need of the HG mothers to carry and care for the children for several years prevented them from fully participating in food collection for long periods, and more importantly – the minimum spacing between child births was about 4 years (Ethenberg, 2008). Domestic dog is likely to be the first animal to have been domesticated by humans, most likely by HG. Genetic and fossil record date the emergence of domestic dogs back to about 15kya (while other studies give dog domestication the range of 9-30kya) (Savolainen et al., 2002).

The earliest evidence of agriculture, the precursor for today's modern human culture, is dated about 10-11kya (though the date is disputed) in the Near East where people pioneered domestication and farming of wild cereal (Bellwood, 2005). The cognitive skill of the pioneering farmers was likely to have been similar to this of humans that lived 40kya (where complex art artefacts were constructed). That leads to the question

of what triggered, or rather, what prevented farming from starting for about 30 thousand years. Several answers and scenarios are being offered in numerous studies and the one that seems the most widely accepted today (although controversial) goes as follows: the last ice age ended about 14kya, followed by climatic stability and growth of the HG populations and extinction of large vertebrates (Jobling et al., 2004). The geographic locations where agriculture had started are correlated to the availability of wild grass species with a potential of domestication in mild climates: wheat and barley in the Near East, rice in the Far East, and so on (Diamond, 1998). Archaeozoology record show that the domestication of goat, cow, pig, and cattle was likely to have co-evolved with agriculture between 12-10kya (Ucko, 2007). The origins of dairy farming in Europe is the subject of chapter 5 of this work.

The agricultural subsistence has resulted in a significant change in lifestyle. Although farming reduced life expectancy in its earlier years, at later stages it allowed higher population density due to increased yield of food and the option of one carer for several infants, which allowed shorter intervals between births, while constant food supply was likely to have resulted in fewer miscarriages (Diamond, 2002). With the increase of farmers population size and the establishment of larger and permanent settlements, the dominant social unit became the household (rather than the whole group in HG), and new non-portable technologies could be developed. Private property gave rise to social hierarchical systems and bureaucracy, while surplus in food supply had resulted in the development of modern forms of trade. Written language is thought to have originated by economic administration (Jobling et al., 2004).

There are two main hypotheses that explain the process of the transition of the majority of human society from HG into an agricultural society: (1) Cultural Diffusion (CD) maintains that farming had spread with the spread of knowledge of technology (Zvelebil and Zvelebil, 1988), while (2) Demic Diffusion (DD) claims that farming had spread by means of physical migration of populations (Cavalli-Sforza et al., 1994). This subject will be thoroughly discussed in chapter 5 of this work.

1.4. Notable Genomic Events Contributing to Human Phenotype.

As discussed in section 1.3.2, there are various classes of genomic events that are likely to have a major role in the evolution of the modern human phenotype. In this section I will describe a few examples of some of these notable genetic events, first in early humans and then in modern humans (*Homo sapiens*).

1.4.1. Notable Genomic Events in Early Humans.

Most primates, including the extinct Australopithecus genus, have strong masticatory muscles, which require a massive and thick braincase. The gene that encodes in these primates masticatory muscles is the **myosin heavy chain** (MYH). In contrast, the *Homo* genus (modern human and earlier *Homo* species) have significantly smaller masticatory muscles and thinner braincases (White et al., 2000), while in modern human the MYH gene is inactivated (Stedman et al., 2004). A molecular evolution study has shown that MYH inactivation took place in the human lineage approximately 2.4mya (just before the transition from *Australopithecus* to *Homo*) as a result of a frame shift mutation. The loss of this protein is associated with the reduction of human muscle fibres, and of the entire masticatory muscles. The timing of the mutation predates modern human anatomy, and represents the first proteomic difference between human and chimpanzee that can be correlated to anatomic imprint in the fossil record. It is hypothesized that this mutation was a trigger for the thinner braincase, and thus for the brain expansion in the human lineage (Stedman et al., 2004).

Human is the only mammal that lacks the common mammalian **sialic acid** N-glycolylneuraminic acid (Neu5Gc). Neu5Gc is developmentally regulated, tissue specific, and has various biological roles in mammals (Angata and Varki, 2002). The human deficiency in Neu5Gc is a consequence of an Alu-mediated inactivating mutation of the gene encoding the enzyme CMP-N-acetylneuraminic acid (CMP-Neu5Ac) hydroxylase (CMAH), dated about 2.8mya (the *Australopithecus* genus). It is thought that in chimpanzee the CMAH is involved in down regulation of brain Neu5Gc (Kawano et al., 1995). It is suggested that the inactivation of CMAH in human had released human ancestors from this constraint, and thus had an evolutionary role in human brain and cognition development (Chou et al., 2002). Further evidence is the

CMAH inactivation in Neanderthals, which had a common ancestor with *Homo* sapiens dated 500-600kya (Hayakawa et al., 2001).

Olfactory receptors (OR) are the largest mammalian gene super-family, with consist more than 1,000 genes. 60% of these genes are pseudogenes in human (Glusman et al., 2001). A study has shown that human lineage had accumulated mutation in the OR super-family at a 4-fold faster rate than in chimpanzee, gorilla, orangutan, and rhesus macaque (Gilad et al., 2003, Glusman et al., 2001). The deterioration of OR genes in modern human suggests that human relied on their sense of smell less than chimpanzee and other primates, which may have contributed to the evolution of different behavioural patterns in human. Non-human primates use the sense of smell for sexual behaviour and social interaction, and thus humans were required to develop different strategies with the significant loss of olfactory capacities (Glusman et al., 2001).

1.4.2. Notable Genomic Events in Modern Humans.

Language is a trait unique to human, and is likely to have been a prerequisite for modern human culture (Wall and Przeworski, 2000). The ability to develop the modern human articulate speech capacities depends on fine control of the larynx and the mouth, traits that lack in chimpanzee and all other non-human primates (Lieberman et al., 2002). The gene **FOX2P** was identified to be correlated to the modern human ability to develop language, in a study that found that the gene is mutated in human individuals which suffer from severe speech and language disorders (Lai et al., 2001). FOX2P is extremely conserved among mammals, and the human FOX2P has two amino acids that are different from chimpanzee, where at least one of the differences is thought to have a functional consequence (Enard et al., 2002). It is suggested that two functional copies of the FOX2P gene are required for acquisition of normal spoken language (Fisher et al., 1998). The fixation of the FOX2P gene in humans is estimated to be 200kya (Enard et al., 2002), at the time of the emergence of anatomically modern human (Homo sapiens), which is compatible with the model maintaining that the expansion of modern humans was driven by the appearance of spoken language (Klein, 1989).

Another trait that is unique to some modern humans is **lactase persistence** (LP), a dominant Mendelian trait that determines the ability of adult human to digest lactose,

the main sugar in milk. For newborn mammals, milk is the only source of nutrition. Lactase is the enzyme responsible for cleaving lactose from disaccharides to digestible monosaccharide, and is coded by the LCT gene. Following weaning in mammals, there is a downregulation of the LCT gene which result in the inability to digest lactose throughout adult human life – lactose non persistence (Ingram et al., 2009a). About 40% of modern human populations, including Europeans and some African and Asian groups (mostly ones that have a history of pastoralism subsistence) are lactase persistent. There are currently 4 known alleles that are associated with lactase persistence. In chapter 4, I investigate the worldwide correlation between the LP associated alleles and LP phenotype. LP gives a strong selective advantage to individuals that have a constant supply of milk, which has lead to the hypothesis that LP originated in a pastoralist population. In chapter 5, I investigate the European origins and gene-culture coevolutionary dynamics of this evolutionary very recent human specific trait. The main finding of this study is that European LP is likely to have originated about 7,500 years before present in the region between central Europe and the northern Balkans, in a gene-culture coevolutionary process on the wave front of the Neolithic expansion. The background and various aspects of LP will be thoroughly described in chapters 4 and 5.

1.5. Integrating Early and Modern Human Genomic Studies.

A main motivation of my study is to investigate the evolution human phenotype from different perspectives – genomewide (human chimpanzee comparison), worldwide (lactase persistence in different human populations), and gene-culture coevolution (the evolution of lactase persistence in Europe). Chapters 2 and 3 are investigating the duplications in human lineage, dating these duplications, applying functions to the duplicated genes, and correlating an aspect of human genomics to fossil record. The scale of the times in these studies is tens of thousands of years, since they deal with a timeline of about 6.5 million years – from human-chimpanzee divergence until present. Since modern human emerged only about 200kya, these chapters will naturally have more focus on early human genomics, while chapters 4 and 5 that focus on lactase persistence are presenting a case study of modern human-specific genomics.

Interestingly, chapter 3 show a disproportionately large number of duplicated genes in the modern human lineage, which are likely to have contributed to the modern human phenotype. There are also several statistically significant clusters of gene duplications around a few dates in human history, which may suggest that these “bursts” of gene duplications have contributed to a strong evolutionary drive in human history. Focusing on the duplication times together with their function may give a clearer picture of the genomic transition from early into modern human. Altogether, the combination of large scale and fine scale early and modern human genomic studies should give a clearer picture about the evolution of human. This will be further discussed in chapter 6.

2. Detecting Human-Chimpanzee Lineage Inparalogues.

2.1. Introduction.

Gene duplications are likely to represent an important class of the evolutionary events that have shaped the unique human phenotype in the short evolutionary time since the Human-Chimpanzee divergence approximately 6.6 million years ago (Steiper and Young, 2006). Furthermore, together with pseudogenization, gene duplications are evolutionary events for which time of occurrence can be estimated (Yang and Yoder, 2003, Brawand et al., 2008).

With the availability of both human and chimpanzee genomes in high re-sequencing coverage assemblies (Lander et al., 2001, Mikkelsen et al., 2005), and the high annotation quality of most known human genes (Hubbard et al., 2009), it should now be possible to identify all human lineage specific gene duplication events (i.e. human inparalogues) using bioinformatics approaches. A few pioneering studies have attempted to do that (Tatusov et al., 1997, Remm et al., 2001). However, due to problems that arise from the different natures of the Human and Chimpanzee's genomes assemblies and level of annotation, these methods have been based on some problematic assumptions and oversimplifications in the algorithm and the datasets used, leading to inaccuracies in detecting human inparalogues.

This chapter describes an attempt to collect a reliable and representative set of human inparalogues, overcoming the conceptual errors that are prevalent in past studies, using methods that I have developed for tackling these trivial and non-trivial obstacles. This chapter is focusing on the methodology and algorithm developed for finding human-lineage gene duplications, rather than on the characterization of these duplications and estimation of duplication dates, which will be explored in detail in chapter 3.

2.1.1. Definitions of Evolutionary Terms Employed.

The evolutionary relations between genes in the same species (paralogues) and among different species often lead to confusing or misleading definitions due to terminology inconsistencies in different studies (Koonin, 2005), therefore I will define in this section the evolutionary terms that are relevant to this study of detecting human inparalogues. See Figure 2.1 for a graphical description of the different evolutionary relations.

Homology, the most general definition, designates a relationship of common descent between any DNA sequence entities, without further specification of the evolutionary scenario that gave rise to observed homology. Accordingly, the entities related by homology, in particular genes, are called homologues (Koonin, 2005). Because the term 'homologues' can refer to orthologues, inparalogues, or outparalogues, it is improper to use it for describing any specific evolutionary relations between genes. All genes in Figure 2.1 are homologous.

Orthologues are homologues produced by species divergence – they represent genes derived from a common ancestral copy in the ancestral species. Orthologues tend to have similar function (Jenuth, 2000). Orthologues can provide useful information regarding functionality, conservation, evolutionary constraint / selection, and evolution of similar genes among different species (Remm et al., 2001). For example – genes XA, XB, and XC in Figure 2.1 are orthologues, since A, B, and C are different species with one common ancestor, while X is the ancestral gene in that common ancestor.

Paralogues are homologues produced by gene duplication and represent genes derived from a common ancestral copy that duplicated within an organism followed by divergence (Jenuth, 2000). Paralogues can have different functions that emerge over time, for example – the two paralogous human genes *AMY1* and *AMY2*, where the former is coding for salivary amylase and the latter is coding for pancreatic amylase (Samuelson et al., 1988). Until recently, the term paralogues was the only one used for describing gene duplications within one species. However, since no distinction was made between duplications that occurred before speciation and duplications that occurred after speciation – two subgroups were needed to be defined – outparalogues

and inparalogues (Remm et al., 2001). Inclusion in one or other of these groups is defined by the speciation event being considered.

"Outparalogues: paralogous genes that evolved via ancient duplication(s) preceding the given speciation event" (Koonin, 2005). In other words – outparalogues are gene duplications that occurred **before** the speciation event, and as such they do not represent any "lineage-specific" gene duplication event unless one copy is lost in one species. For example – genes XB, YB, and ZB1 in Figure 2.1 are outparalogues since X, Y, and Z are different genes in the same species (B) that were also separate genes in the common ancestor – gene duplications before speciation.

"**Inparalogues**: paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event" (Koonin, 2005). In other words – inparalogues are gene duplications that occurred **after** a specific speciation event, and as such they are suitable for "lineage-specific" gene duplications studies. For example – genes ZA1, ZA2, and ZA3 in Figure 2.1 are inparalogues since they are a result of two separate duplication events of the gene Z in the lineage of species A (i.e. after the A-B speciation event).

C**opy number variation** (**CNV**) is a DNA segment (that may or may not include genes) which has a different number of copies among two or more chromosomes sampled from a population. The size of the segment can be up to several megabases. CNVs arise due to sequence duplications or deletions (Cook and Scherer, 2008). This work will not deal with CNVs. However, it is important to understand the main conceptual difference between CNV and the other homology terms that were explained above, as CNV deals with gene duplications and so may cause confusion. CNV, as with any genetic variation term, is based on more than one individual. For this reason it cannot be used (in its strict sense) as a "representative" of the species, though CNV does have a potential use as a tool to measure if the human-lineage duplications detected are representatives of the majority of modern human population, or if they are duplications that represent only specific group or individuals. This genetic variation work is beyond the scope of this thesis.

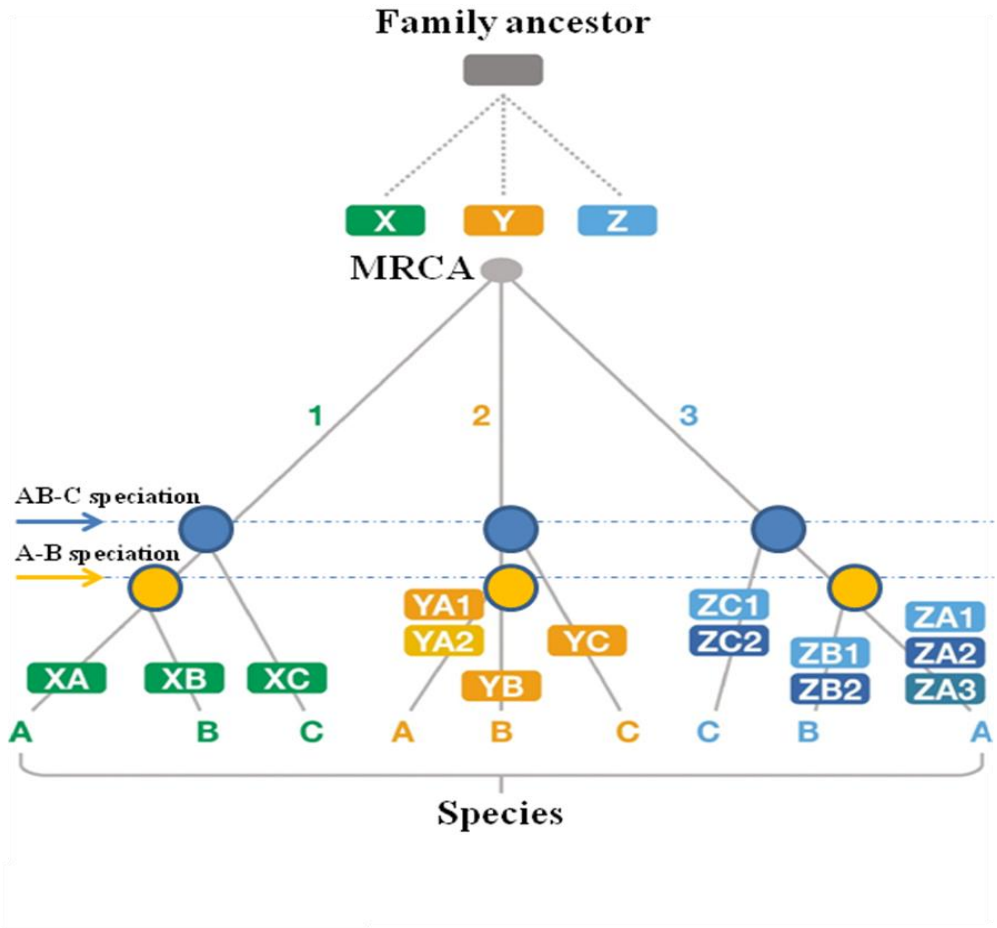

Figure 2.1. A hypothetical phylogenetic tree illustrating orthologous and paralogous relationships between three ancestral genes (X, Y, and Z) and their descendants in three species (A, B, and C). The divergence of the three genes was prior to the species most recent common ancestor (MRCA). The hypothetical timeline from the family ancestor until present is from the top to the bottom of the Figure. The blue circles and blue arrow represent the first divergence event (between species AB and C), and the yellow circles and yellow arrow represent the second divergence event (between species A and B). Adapted from Koonin (2005) (Figure 2, page 313).

2.1.2. Review of Orthologues and Paralogues Detection Methods.

The first step in identifying inparalogues in a specific species for a particular species pair is identifying the corresponding orthologues in the reference species, to make a distinction between 'out-' and 'in-' paralogues – duplications that happened before MRCA speciation and duplications that happened after MRCA speciation, respectively. However, until recently most studies focused on paralogues without the distinction between inparalogues and outparalogues (Remm et al., 2001), and some pair the paralogous studies with segmental duplications – continuous portions of DNA that map

to two or more locations on one genome, and tend to form core ‘duplicons’ in the human genome (Bailey and Eichler, 2006, Jiang et al., 2007).

The “traditional” process of identifying lineage-specific duplications is laboratory based, applying molecular genetics techniques. An example of such methodologies is a recent extensive study of comparing primates segmental duplications (She et al., 2006) combining bioinformatics analysis using BLAST-based detection schemes (Bailey et al., 2001, Bailey et al., 2002) with fluorescent in situ hybridization (FISH) analyses (Nath and Johnson, 2000) for detecting lineage-specific segmental patterns.

With the accumulation and availability of whole genome data from several species, due to the development of cheaper and more efficient sequencing techniques, the only practical way of analyzing the homology relationship between sets of genes is by applying (or combining) bioinformatics methods, which mostly follow one of two main approaches: best reciprocal hit and phylogenetic reconstruction (Koonin, 2005, Altenhoff and Dessimoz, 2009).

The more commonly used approach is based on best reciprocal hit using sequence database search algorithms such as BLAST or BLAT (Altschul et al., 1990, Kent, 2002) or reciprocal smaller distance using a substitution rate matrix such as JC69, F84, or HKY85 (Jukes and Cantor, 1969, Felsenstein, 1989, Hasegawa et al., 1985). Both approaches are much more computationally efficient than phylogeny methods (and arguably some are at least as accurate, as will be explained in the section describing the InParanoid algorithm), and thus can be applied for complete genomes orthology and paralogy detection.

The second, and less commonly used approach – phylogenetic reconstruction – is a natural way of detecting orthology and paralogy, as the specific type of homology is being directly inferred from the topology of the tree. A simple example for homology type inferred from tree topology would be three sequences – two human sequences (which will be called H1 and H2) and one chimpanzee sequence (called C1), where we want to know if H1 and H2 are inparalogues, and given that C1 is an orthologue for at least one of the two human sequences. The possible topologies, described using the Newick tree format (Felsenstein, 2003), of the full tree space are: (1) ((H1,H2),C1) , (2)

(H2,(H1,C1)), and (3) (H1,(H2,C1)). See Figure 2.2 for the full tree space of these 3 sequences. Trees (2) and (3) represent scenarios where the H1-H2 duplication occurred before the human-chimpanzee speciation event, and thus they are not inparalogues. Tree (1) is the only scenario where H1 and H2 are inparalogues, since the duplication event took place after human- chimpanzee speciation.

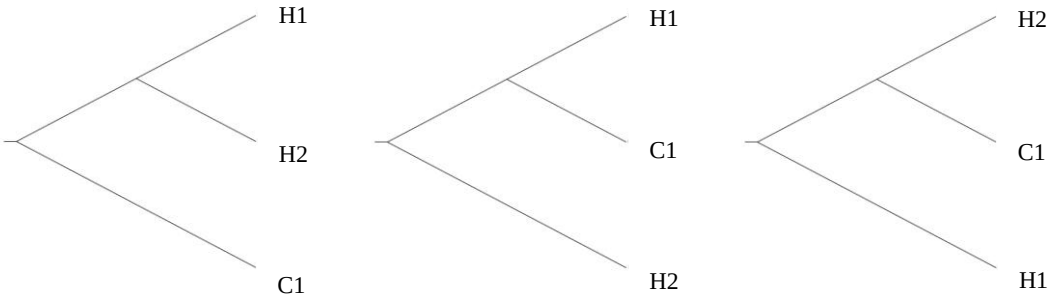

Figure 2.2. The tree space of two human sequences and one chimpanzee sequence. The left tree is the only one that represents two human inparalogues and their chimpanzee orthologue. In the central and right trees the human sequences are outparalogues.

However, this example deals with only 3 sequences. The full combinatorial space of a genome-wide comparison is immense – the number of possible rooted bifurcating trees is for n sequences is $\frac{(2n-5)!}{2^{n-3}(n-3)!}$ and for unrooted trees it is $\frac{(2n-3)!}{2^{n-2}(n-2)!}$ (Cristianini and Hahn, 2006). Before inferring each topology - multiple sequence alignment, an error-prone and time consuming procedure (Remm et al., 2001) – must be performed. The number of possible rooted and unrooted trees for n=5 is 15 and 105, respectively, while for n=10 the number of possible rooted and unrooted trees jumps to 2,027,025 and 34,459,425, respectively (calculated by me using the Python programming language, http://www.python.org/). As a result, the computational time required for such analyses on a genome-wide scale (with tens of thousands of sequences for each species) makes this method very computationally intense even when using tree space searching optimization methods that reduce the tree space (Koonin, 2005) or by treating some of the sequences as having a non-random relations to each other.

Below, I will present two well-established methods as representatives of the modern computational methods of genome-wide orthology and paralogy detection. The methods are both non-phylogenetic, and so present practical options for whole-genome paralogues detection.

COG - Clusters of Orthologues Groups (Tatusov et al., 1997), was the first platform created to identify large scale clusters/groups of orthologues and paralogues, as opposed to previous methods that identified smaller and separate sets of orthologues and paralogues. The main assumption of COG is that any set of at least three proteins from relatively distant genomes that are more similar to each other than they are to any other proteins from the same genomes are most likely to be orthologues. The prediction holds even if sequence similarity between some of the compared proteins are relatively low, thus COG can also group genes that are fast evolving (Koonin, 2005). The COG algorithm consists of the following steps (Koonin, 2005, Tatusov et al., 1997): (1) An all-against-all BLASTP (Altschul et al., 1990) comparison of protein sequences from multiple genomes. (2) Detection and clustering of orthologues and paralogues, following the assumption that if a gene from one of the genomes has its two best BLAST hits (BeTs) in two other genomes (i.e. the two genes most similar to a specific gene are from distant genomes rather than from the gene's own genome) then it is likely that they are orthologues. (3) Identification of triangles of genome-specific best hits, treating paralogues detected at step 2 as single entities. (4) Forming COG's from triangles with a common side. See Figure 2.3 for examples of orthologues and paralogous identified using the COG algorithm.

Although COG identifies orthologues and inparalogues, it tends to have high false positive rates when large protein families include both in- and outparalogues or when multidomain proteins that are included in the analysis artificially bridge unrelated COGs (Koonin, 2005, Altenhoff and Dessimoz, 2009) since multidomain proteins don't fully represent their corresponding genes' full DNA sequence. Moreover, the minimum number of species for a COG is three, and so a COG represents sequences with conserved functions across different and distant lineages. It becomes a problem when there is a need to find orthologous groups (and detecting inparalogues in these groups) between closely related species, such as human and chimpanzee – two species that are too closely related for their paralogues being detected in COG. Another obvious problem may arise when sequences are available from only two species.

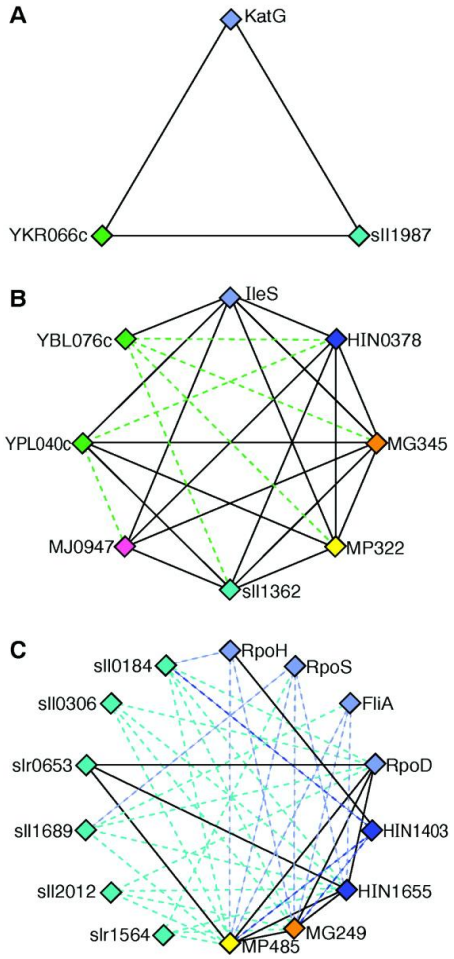

Figure 2.3. Three examples of orthologues and paralogues obtained by COG (Clusters of Orthologous Groups). Different colours representing different species, solid lines show symmetrical BeTs (and thus orthologous relation), broken lines show asymmetrical BeTs, and proteins of the same colour are paralogues. (A) The minimal COG of 3 orthologues. (B) A COG with two yeast paralogous - YBL076c and YPL040c. (C) A complex COG where 3 species have multiple paralogous (for example, Sll0184 and Slr1564) and 2 species have no paralogous but do have orthologues identified in other species (for example, MP485 has no paralogues, and is orthologous with MG249, RpoD, and Slr0653). Figure taken from Tatusov et al. (1997).

The problems of COG include disentangling inparalogues from outparalogues and dealing with closely related genomes, and this has led to the development of **InParanoid** - In-paralogue and Orthologue Identification (Remm et al., 2001, O'Brien et al., 2005). The algorithm identifies orthologues and inparalogues between any given pair of genomes (two species only), while the programme MultiParanoid allows finding orthology and paralogy among multiple species, making it conceptually more similar to COG (Berglund et al., 2008). Given the proteomes (in this case – exactly one protein

from each coding gene) of two given species B and C with a most recent common ancestor A (Figure 2.4a), the InParanoid algorithm works as follows (Remm et al., 2001, O'Brien et al., 2005): (1) Find all sequence pairwise similarities between B-C, C-B, B-B, and C-C using BLASTP (Altschul et al., 1990). (2) Mark two-way best hits as potential orthologues (these are inter-species seed-orthologues). (3) Add potential inparalogues for each seed-orthologues pair, by assuming that two inparalogues (which are, by definition, from the same species) are closer to each other than the distance between the seed-orthologues, otherwise the gene duplication is assumed to be before the divergence of B and C, and thus the two sequences are considered to be outparalogues (Figure 2.4b). (4) Calculate relative distance scores for the potential inparalogues (Figure 2.4b), using the equation: $C3 = \frac{Blast[C2:C3] - Blast[C2:B2]}{Blast[C2:C2] - Blast[C2:B2]}$, where Blast[X:Y] is the averaged BLASTP score between X and Y in bits. (5) Resolve overlapping groups of orthologues and inparalogues.

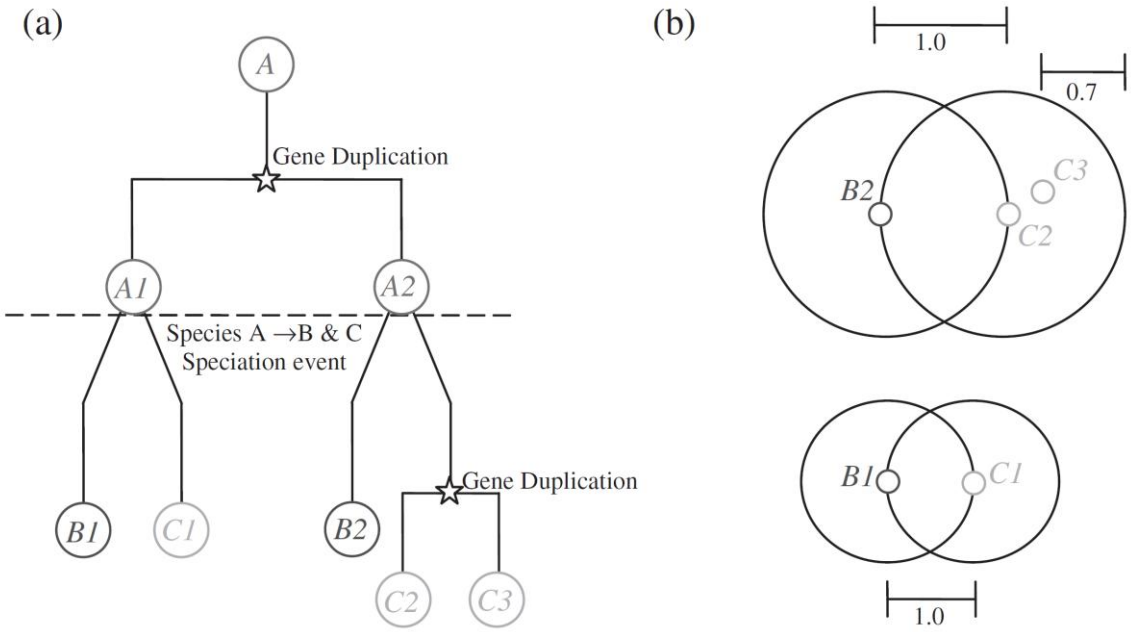

Figure 2.4. The InParanoid algorithm. (a) Showing a protein in the ancestral species 'A' that underwent a gene duplication event. After speciation event into species 'B' and 'C', gene C2 was duplicated into the inparalogues C2 and C3. (b) Showing the clustering method. The best reciprocal hit proteins B2 and C2 are regarded as inter-species seed-orthologues, around which paralogues are clustered. The distance between the seed-orthologues is set to 1.0, and accordingly the distance between an orthologue and its inparalogue is always between 0 and 1.0 (as is the case with inparalogues C2 and C3), while a distance between an orthologue and an outparalogue (that is rejected) is greater than 1.0 (as is the case with inparalogues C2 and C3). The image was taken from O'Brien et al. (2005) Figure 1.

I chose InParanoid as the main algorithm to work with because the focus of the algorithm is detecting inparalogues, rather than paralogues (or only orthologous) in general. Also, a benchmark test of several of the most popular orthologue detection methods using a Human, Mouse, and C. elegans protein expression and sequence data (Hulsen et al., 2006) has shown that InParanoid had the best performance in Human-Mouse orthologue detection. The benchmark tested used the Pearson correlation between conservation of function (determined by known protein expression levels) and the orthologue prediction performed with the following six methods: (1) COG (Tatusov et al., 1997), (2) best bidirectional hit (essentially a simple version of InParanoid implemented by the authors of the benchmark study), (3) InParanoid (Remm et al., 2001), (4) OrthoMCL (Li et al., 2003) – a markov-clustering (Enright et al., 2002) based algorithm, (5) Z / Hundred – estimating statistical significance of dynamic alignment scores through the use of a Monte-Carlo process (Comet et al., 1999), and (6) PhyloGenetic Tree – based on time consuming multiple alignments (van Noort et al., 2003). Interestingly, InParanoid outperformed even the phylogeny based method (van Noort et al., 2003) and was shown to perform exceptionally well in detecting orthologues among relatively closely related species (Human-Mouse, as opposed to the much more distantly related Human-C. elegans and Mouse-C. elegans).

However, the InParanoid algorithm introduces problems when attempting to implement it for detecting inparalogues in projected genomes (genomes of non-model organisms that their genes are experimentally unknown, and thus these genes are being identified by projection – transferring their nearest species experimentally known genes to the corresponding location in the non-model organism genome (Hubbard et al., 2009)). This problem is a critical issue in detecting Human-Chimpanzee inparalogues where the chimpanzee's genome is projected. This will be discussed in the following section.

2.1.3. Problems with Inparalogues Detection using InParanoid.

InParanoid is the only comprehensive method that focuses on detecting inparalogues (rather than paralogues in general), which makes it a potentially ideal tool for detecting human inparalogues. However, some critical problems were encountered when I attempted using the InParanoid Human-Chimpanzee orthologues/inparalogues database (O'Brien et al., 2005), and, alternatively, attempting to locally use InParanoid with the

Human and Chimpanzee proteomes (Hubbard et al., 2009). I will now elaborate on the problems that were encountered, and which this project has attempted to tackle. Unless otherwise stated, all automating procedures in this section were performed by Perl scripts written by me.

2.1.3.1. Human Haplotype Data.

As a part of the effort to map human genomic variants that may be associated with common diseases susceptibility – two projects were conducted to identify two haplotypes of the major histocompatibility complex (MHC) on Human chromosome 6 (COX and QBL), to which susceptibility to more than 100 diseases has been mapped (Stewart et al., 2004, Traherne et al., 2006).

The Ensembl annotated human genome database (Hubbard et al., 2009) includes 246 COX haplotype alleles and 234 QBL alleles, and altogether 741 proteins (Hubbard et al., 2009, Smedley et al., 2009) (Table 2.1.) The InParanoid database of orthologues and inparalogues (http://inparanoid.sbc.su.se/cgi-bin/index.cgi) was attained by using the InParanoid algorithm with the full known proteome of each species from which it attempts to identify inparalogues - it includes the longest protein sequence from each human coding gene. However, the InParanoid algorithm does not filter for haplotype data, resulting in using chromosome 6 COX and QBL protein sequences. The result is many variants of the same genes collected from different genomes, which leads to false detection of inparalogues (i.e. false positives), as the haplotype genes and proteins are likely to be very similar to each other, and so they will be identified as inparalogues even though they are actually variants of the same gene among different individuals rather than being genes that are representing duplications in the human lineage.

The use of haplotype data in InParanoid's inparalogues detection has an effect similar to artificially adding hundreds of almost identical copies of hundreds of genes to the human genome database. This leads to erroneous clustering and an overprediction of human inparalogues.

2.1.3.2. Proteome Data.

The input data for InParanoid consists of the proteomes of the two species, which orthologues and inparalogues are to be sought. The longest peptide sequence from each coding gene is used, creating a non-redundant representation of the organism's genome. As a result of using proteome data, only the gene's coding sequence (CDS) is represented. A gene consists mostly of introns and regulatory regions, while the CDS is a small part of the whole gene (12.11% on average, calculated by me using the BioMart (Smedley et al., 2009) information for all human coding gene and CDS lengths). This poses a problem since the majority of the gene's sequence information is lacking.

On the conceptual level, there is a problem in using peptide sequences for detecting physical DNA duplications, as the peptides are the product of codons that contain redundancies (i.e. a few different codons that code for the same amino acid) and so the peptide sequence will miss DNA silent mutations. In general, a peptide sequence does not perfectly represent the DNA sequence that it was derived from.

For these reasons, it is problematic to use protein sequences for detecting lineage-specific gene duplications. However, the proteome can be very useful as a first pass filter for detecting inparalogue candidates as will be demonstrated in Section 2.2.

2.1.3.3. Ambiguous Data.

As a part of the Ensembl (Hubbard et al., 2009) gene annotation process of each species, transcripts are aligned to the whole sequenced genome to identify the chromosomal location of each gene. Due to low sequence coverage or low transcript quality, there are cases where a transcript cannot be mapped to specific chromosomal regions, and consequently the gene's chromosomal location is identified as 'random' (when a specific chromosome is identified), 'Un' (when the chromosome is unknown), or 'NT' (essentially like 'Un', with the original contig's name specified as a chromosome). I will refer to the three classes of ambiguous annotation data as 'ambiguous'. The numbers of these ambiguous genes vary among the different annotated species.

The Ensembl database (Hubbard et al., 2009) includes 221 ambiguous human genes and 1,268 ambiguous chimpanzee genes (Table 2.1), which are used in the current InParanoid human-chimpanzee database (O'Brien et al., 2005). Although the sequence quality for some of these genes may be adequate, the fact that they are 'ambiguous' (as described above) makes it not impossible to detect if they overlap with other genes (see next section for gene overlaps), and if they are identified as gene duplications it is difficult to know if they are tandem duplications (on a similar chromosome) or duplications among different chromosomes. Altogether, the fact that these genes cannot be traced into a specific location suggests a problem in the quality of the genes annotation, and so using these genes makes the dataset used much less reliable.

2.1.3.4. Gene Conversion.

Following gene duplication, adjacent paralogous are prone to reciprocal unequal crossovers by virtue of the high degree of homology between them. As a consequence of these unequal crossovers, the 'acceptor' sequence is replaced, wholly or partly, by a sequence that is copied from the 'donor', whereas the sequence of the donor remains unaltered. This process is termed **gene-conversion** (Chen et al., 2007). As a consequence of a full gene conversion, the two copies of the gene have very high degree of similarity. In the case of a gene duplication occurring before the most recent speciation event followed by gene conversion, any currently available inparalogues detection method is likely to identify the two copies as inparalogues, when in fact they are outparalogues. There is no currently available bioinformatics filter gene conversion regions, and so I expect that all currently available inparalogues detection methods will include false positive inparalogues.

2.1.3.5. Non-Model Organisms.

The most critical problem that I have encountered when locally using InParanoid to detect human-chimpanzee inparalogues was the use of the chimpanzee proteome.

Chimpanzee is a non-model organism whose genome has been sequenced and annotated by Ensembl. The majority of annotated genomes available from Ensembl and BioMart are of non-model organisms (such as the orangutan, macaque, horse, cat, platypus, and

more) as well (Hubbard et al., 2009, Mikkelsen et al., 2005). One major implication of annotating non-model organisms' genomes is that they have a very low proportion of experimentally known genes. The non-model organism's unknown genes are being annotated by projection – aligning its transcripts to the known genes from the evolutionary nearest genome(s). See Figure 2.5 for the phylogenetic relations between mammals and vertebrates, which determine the genomes from which non-model organisms or unknown genes are being projected, and Table 2.1 for the numbers of known and projected genes among several model and non-model organisms. Note that for lower coverage genomes or where genes cannot be annotated in model organisms, Ensembl is applying another annotation category termed "novel genes" – a process that is essentially following the same process as projection, but unlike projection, it allows the projection sequence to change the original assembly (Hubbard et al., 2009). For convenience I will term both "projected" and "novel" genes as "projected".

The chimpanzee's unknown (i.e. in-silico predicted or projected) genes are wholly projected from the human known genes, while (for example) the majority of the horse's gene annotations are projected from several model-organisms genomes, including human and mouse.

The majority of chimpanzee's annotated genes are being projected from known human genes, and currently there is no algorithm for identifying chimpanzee-specific genes. Furthermore, comparing a human genome/proteome with the chimpanzee's genome/proteome that is projected from human is essentially as if human genes are being compared with "less annotated" genes. .

The lack of chimpanzee-specific genes/peptides is demonstrated by performing an InParanoid run where the human and chimpanzee's proteomes are used as input, after applying various filtering as will be elaborated in section 2.2. The output of the InParanoid run was the full set of human and chimpanzee orthologous group (that some of them contain inparalogues). I detected cases of human and chimpanzee orthologous groups where one species has inparalogues while the other species has no inparalogues (in other words – human- or chimpanzee- specific inparalogue groups). This gives a measurement of how much the two genomes' annotation is balanced, with the null hypothesis being that human and chimpanzee have a similar number of species-specific

gene duplications. The numbers of human- and chimpanzee-specific gene duplications show a massive bias towards human duplications – 192 human-specific inparalogues groups, and only 33 in chimpanzee (Table 2.2), a difference of almost 6 times. While at face value this could have been a very exciting discovery, indicating that human lineage has had a significantly accelerated gene duplication rate compared to chimpanzee (or alternatively, continuing this line of thought, that chimpanzee had a significant deceleration), running InParanoid with the human proteome as in input against several other species demonstrated that this is not the case. Rather, there seems to be a bias stemmed in the nature of the specific species' genome – depending whether it is a genome of a model or a non-model organism, which reflects on the annotation of the genes – being known or projected. When running the human proteome against organisms in which the majority of the genes are known (e.g. mouse and cow, see Tables 2.1 and 2.2) the tendency of human "having" more lineage specific gene duplications was reversed. I detected differences of 1.27 and 1.57 times more lineage specific inparalogues-containing groups in cow and mouse, respectively, than in human. The number of human peptide sequences used is only 1.07 times larger than the chimpanzee's, while the number of mouse peptides is 1.1 times larger than human and the cow's is 1.1 times smaller – so the differences in number of peptide sequences among the different species are not likely to account for the bias witnessed in the human-chimpanzee test. Performing similar InParanoid runs and species-specific duplications analyses of the human proteome against other non-model primates revealed similar patterns that were witnessed with chimpanzee: a 3.64 times more human-specific inparalogues groups than orangutan-specific, and 1.84 times more human-specific inparalogues groups than macaque-specific. Importantly, Ensembl added human genes to the orangutan and macaque's database where the Ensembl projection failed to identify acceptable gene models for these species, and this may explain the smaller numerical bias that these species have when compared to chimpanzee bias.

Performing the analyses described above for the human proteome against horse's, a non-primate non-model organism, revealed high similarities in the number of species-specific inparalogues groups among the two species – 196 human-specific vs. 204 horse-specific (Table 2.2). The horse's Ensembl genome annotation is projected from all known mammalian genes, and – at a lower priority – from non-mammalian vertebrates. Also, the horse genome assembly coverage was (as for Ensembl version 52)

x6.79 (where the assembly coverage unit represents the average number of times that each unit of the genome was sequenced), a relatively very good quality for whole genome sequencing (Ensembl unofficially defines “low coverage” as about x2.5 or less). The similar human- and horse-specific groups, together with the horse annotation process and its high coverage show that when all (available) high vertebrate genomes are taken into consideration then there is no numeric bias. This does not mean that the horse’s genome annotation can be used for inparalogues prediction, but rather that on average the different lineages seem to have about the same gene duplication rates.

In summary, the use of a non-model organism’s proteome as one of the species when performing inparalogues prediction using InParanoid, or any other inparalogues prediction algorithm, produces an underestimation of the non-model organism’s inparalogues. This has presented a critical problem in using the chimpanzee’s proteome for detecting the human lineage gene duplications, and required developing new methodologies for doing that. Section 2.2 will describe the algorithm that was developed to detect lineage specific duplications in cases resembling the human-chimpanzee relations – one genome is of a model organism while the other one is non-model, while resolving the problems of human haplotype data, proteome data, gene conversion, and the use of non-model organisms, which were presented in sections 2.1.3.1 - 2.1.3.5.

Table 2.1. Gene categories in model and non-model organisms' genomes and proteomes. The different categories are explained in section 2.1.3.5.

Organism	Number known protein-coding genes	Number projected and novel protein-coding genes	Number ambiguous genes	Number overlapping genes	Number haplotype genes
Human (*Homo sapiens*)	21388	28	221	2125	741
Chimpanzee (*Pan troglodytes*)	2647	17182	1268	1226	-
Orangutan (*Pongo pygmaeus abelii*)	3813	16255	1245	1007	88
Macaque (*Macaca mulatta*)	874	21031	1123	1854	-
Mouse (*Mus musculus*)	23019	98	273	1976	-
Cow (*Bos taurus*)	20471	583	2745	874	-
Horse (*Equus caballus*)	723	19599	153	1024	-

Table 2.2. The number orthologous clusters having species-specific inparalogues, detected by InParanoid. Non-model organisms are identified by '*'. For hypothetical species j and k, a cluster was detected for having species-specific inparalogues by counting the number of inparalogues for j and k, then if the number of j inparalogues is greater than 0 and the number k inparalogues is equal to 0 then the cluster is considered as having j-specific inparalogues (and vice versa for k-specific inparalogues).

Organisms tested	N estimated human-specific duplications	N estimated species-specific duplications
Human-Chimpanzee*	192	33
Human-Mouse	207	326
Human-Orangutan*	171	47
Human-Macaque*	208	111
Human-Cow	220	279
Human-Horse*	196	204

2.2. The Human Inparalogues Detection Algorithm.

As was demonstrated in the previous sections of this chapter, there are various reasons why past efforts to detect human inparalogues may produce unreliable results with projected or poorly annotated genomes. I will now describe the full process that I have developed for finding human inparalogues that overcomes many of these problems.

Importantly, this algorithm can be applied to identify inparalogues among any two species, where one is a model organism in which a proteome is available and the other is a non-model organism. The process requires the availability of both their genomes assemblies, and the availability of another outgroups model organism's proteome.

The programming language used for writing the various scripts for the algorithm was Perl (http://www.perl.org/) which is used extensively in Bioinformatics applications mainly because its implementation of regular expressions (identification of patterns in text) which makes the language ideal for handling genetic and proteomic sequences. Other applications that were used will be described at the relevant sections.
The algorithm is first filtering the input data of human and mouse for InParanoid by removing ambiguous data and resolving gene overlaps, then detecting human inparalogue candidates using InParanoid. The inparalogue candidates are used to identify potential chimpanzee orthologues and human inparalogues on chimpanzee and human genomes, respectively. The full duplication lengths of these candidates are being identified, and phylogenetic trees are inferred, while removing topologies that suggest human outparalogues and filtering molecular clock violations. The final step of the algorithm is filtering for gene conversions and acquiring the human inparalogue genes in the duplicated regions. The full algorithm is described in Figure 2.6.

2.2.1. Choosing an Outgroup and filtering data.

The first part of the algorithm is identifying potential inparalogues applying the InParanoid software, using the human proteome and the proteome of the model organism that is nearest to human, in this case – the mouse. See Figure 2.5 (Benton and Donoghue, 2007).

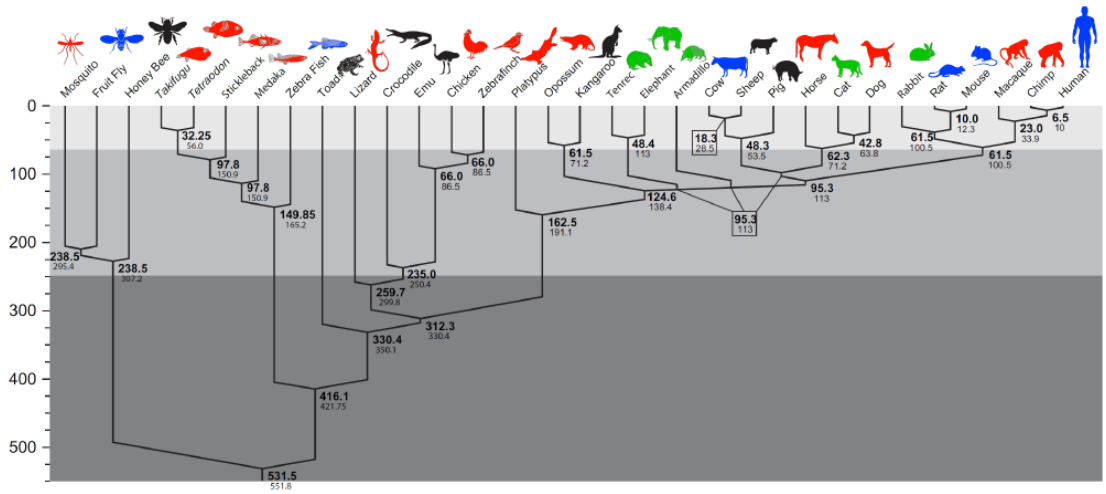

Figure 2.5. A phylogenetic diagram ranging from the insect to the primates clades. Numbers represent divergence/speciation times. The different colours represent the Ensembl version 55 genome annotation type: blue represents model organisms, red represents non-model organisms with high sequence coverage (more than x4), green represents non-model organisms with low sequence coverage (equal or less than x4), and black represents organisms that their annotation is currently not available in Ensembl. Adapted from (Benton and Donoghue, 2007) Figure 8, page 43.

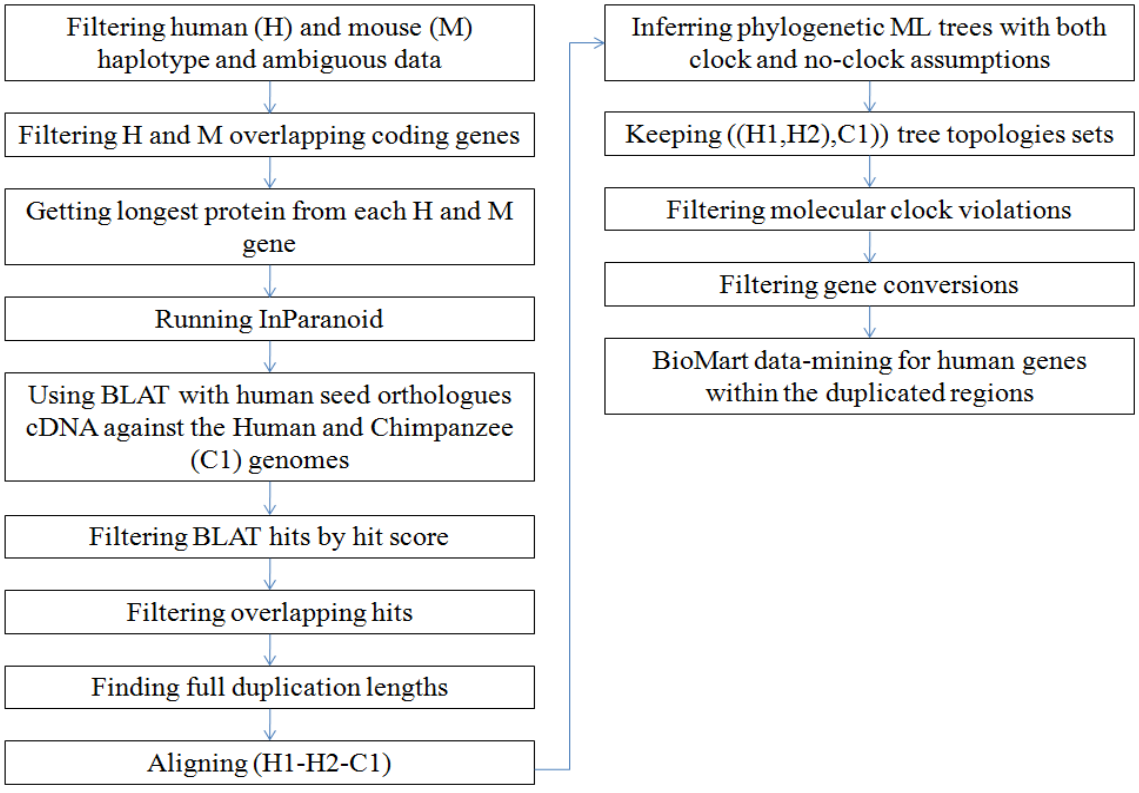

Figure 2.6. The filtering and analyses stages in the human-lineage gene duplication detecting algorithm. Each stage provides the input for the next stage, while the initial input is the full human and mouse proteomes. H1 and H2 represent the human orthologue and human inparalogue-candidate sequences, respectively. The algorithm is fully described in section 2.2.

Using the biological data mining website BioMart (Smedley et al., 2009), the following Ensembl (Hubbard et al., 2009) annotation features for all 21,388 human and 23,019 mouse protein coding gene were obtained: (1) Chromosome number/symbol. (2) Start location. (3) End location.

Genes were removed if: (1) The chromosome's symbol is either ambiguous or haplotype data. (2) Two following entries are overlapping, following the logical rule: IF (chr_i==chr_{i-1} AND $start_i$<end_{i-1}) THEN $gene_i$ filtered. This rule means that if two following genes are located on the same chromosome and the start location of one gene is located within the other gene then it is overlapping and thus removed. This step was repeated until there were zero overlaps.

From each human and mouse non-ambiguous and non-overlapping known coding genes, the longest protein sequence was acquired. Altogether, the final set acquired peptides representing the human and mouse proteomes consisted of 18,522 human peptide sequences and 21,043 mouse peptide sequences. The filtering process has removed 2,866 human peptides and 1,976 mouse peptides.

2.2.2. Human-Mouse InParanoid Run.

As demonstrated in section 2.1.3.5, InParanoid (Remm et al., 2001, O'Brien et al., 2005) did not provide reliable human inparalogues results when used with the known human and the projected chimpanzee's proteomes. However, InParanoid provides a very robust and accurate platform for detecting inparalogues among two model organisms such as human and mouse (van Noort et al., 2003). For these reasons, after the initial filtering process described above, InParanoid was ideal for detecting human-mouse inparalogues with the filtered human and mouse proteomes as input.

Running InParanoid with the 18,522 human peptide sequences and 21,043 mouse peptide sequences obtained in section 2.2.1 resulted in 16,227 clusters of human-mouse seed orthologues, among them 305 contain one or more human inparalogue. It is important to note that each of the human inparalogues detected in this stage are in regard to the human-mouse lineage, making the majority of them to be human-

chimpanzee lineage inparalogues, as human-mouse divergence occurred about 60mya while human-chimpanzee divergence occurred about 6.5mya.

2.2.3. Human-Chimpanzee BLAT Run.

BLAT (BLAST Like Alignment Tool) is a software identifying DNA or peptide sequences in a database, such as a full genome (Kent, 2002).

To detect gene duplications, the DNA sequence was required to be used against the human and chimpanzee genomes (the chimpanzee's assembly has a high coverage of x6 coverage, making the sequence reliable to use).

For each of the 305 clusters of human-mouse containing one or more human inparalogue (as described in section 2.2.2), the cDNA sequence of the human seed-orthologue peptide was acquired using BioMart, a biological data mining web interface (Smedley et al., 2009). Then BLAT (Kent, 2002) was used to identify the chimpanzee orthologues and human inparalogues. BLAT's characteristics are tailored to identify DNA sequence duplications on genomes with a high degree of similarity. This makes BLAT suited for species with a small evolutionary distance such as human and chimpanzee, and consequently suited for finding human inparalogues which are assumed to have a smaller distance from their human orthologue than the distance between the human and chimpanzee orthologues. The BLAT run of the 305 human cDNA sequences against the human and chimpanzee genomes on the UCSC web server was automated by using the Perl script that is available at the following website: http://genomewiki.ucsc.edu/index.php/Image:BlatBot_pl.txt.

The chimpanzee orthologues and human inparalogue candidates were detected from all BLAT hits by applying the following criteria: (1) Highest bit scores (which the BLAT algorithm uses to determine the best match). For human inparalogues detection, a minimum threshold of half of that of the best hit was applied. (2) Sequence length similarity of at least 50%, since local alignment may capture various regions of the BLATed cDNA sequence scattered on huge regions of the chromosome, which may result in a (say) 300 base pairs cDNA sequence being match to a 1 million base pairs hit.

The human inparalogue-candidates were then filtered for overlaps, following the same process described in section 2.2.1.

2.2.4. Finding the Full Extent of Human Duplicated Regions.

Although using cDNA sequences provides more evidence for DNA duplications than using peptide sequence only, each sequence will represent only a portion of the full actual segmental sequence duplication that may extend upstream and downstream from the orthologues and candidate human inparalogue sequence detected by the human cDNA.

To find the full extent of each duplication, the Ensembl Perl API interface (http://www.ensembl.org/info/data/api.html) was applied, and was automated with a Perl script written by myself. The dataset was divided into triplets of (1) Human orthologue. (2) Human inparalogue-candidate. (3) Chimpanzee orthologue. Upstream from the start of each of the 3 sequences, sliding windows of 100 base pair slices were obtained and compared to each other. In case there was a similarity greater than 90% (a heuristic value, greater than the similarity between two random sequences and lower than the expected 95%-100% human inparalogues / human-chimpanzee orthologues comparisons (Britten, 2002, Mikkelsen et al., 2005)) another 100 base pair slice upstream of the previous slice was obtained and the same similarity check was made. The window continued its upstream slide until similarity went below 90%. The same process was performed downstream of each human and chimpanzee sequence's end. Importantly, as genome are represented by only one strand, whenever a sequence was on the opposite strand the complementary sequence was inferred and the upstream-downstream directions were reversed.

By checking for no overlap between the extended tandem duplications, the full human duplications and their full length chimpanzee orthologue DNA sequences were obtained.

2.2.5. Alignment, Phylogenetic Trees and Molecular Clock Testing.

At this stage, all human-human-chimpanzee orthologue sets were still inparalogue candidates, as they rely were identified as potential human-chimpanzee inparalogues only by comparison of human and mouse proteomes. The human-mouse divergence was ~61.5mya while the human-chimpanzee split was ~6.5mya, so the majority of duplications identified at this stage are expected to be outparalogues with respect to human-chimpanzee comparison.

As described in section 2.1.2, phylogenetic tree inference is a very robust way to estimate homology types. The InParanoid clustering and various filtering described above had reduced the potential tree space from one that is completely impractical (see section 2.1.2 for the number of possible bifurcating trees – millions for 10 sequences, and so an inconceivably large number for sequences of two full genomes) into a scale of only hundreds of human-human-chimpanzee triplets. This has made possible the use of phylogenetic tree inference for the human inparalogues and their chimp orthologue triplets, for differentiating human inparalogues from outparalogue.

The first step in any phylogenetic inference is performing multiple sequence alignment. The software I chose for that was MAFFT - Multiple sequence Alignment employing Fast Fourier Transform (Katoh et al., 2002). The advantage of the method is its flexibility and reliability tradeoff – automatically optimizing the alignment according to the different DNA sequence lengths used, which is very important due to the large variety of sequences used in this case – from a few hundred to hundreds of thousands of base pairs. A benchmark test (Katoh et al., 2005) has shown high performance of MAFFT when compared to other well established methods, including MUSCLE (Edgar, 2004), T-Coffee (Notredame et al., 2000), and ClustalW (Thompson et al., 1994). Moreover, due to memory constraints MUSCLE is incapable of aligning sequences larger than a few thousand base pairs, T-Coffee's very slow computation time makes it impractical to use for this study, and ClustalW is slower and less accurate than MAFFT. Manual alignment testing that I performed with MAFFT (inspecting by eye) also confirmed the high performance of the software.

I wrote a Perl script automating the DNAML and DNAMLK maximum likelihood phylogeny inference programs, which are a part of the Phylip package (Felsenstein, 1989), with the aligned sequences as input. DNAML doesn't assume a molecular clock and DNAMLK does. Both tree topology and maximum likelihood score were obtained from each set of sequences. The first filtering process kept only the trees with the topology of ((H1,H2),C1) – representing two human inparalogues and their chimpanzee orthologue (see the left tree in Figure 2.2). Then a likelihood ratio test of the molecular clock was applied to make sure that the molecular clock cannot be rejected (Felsenstein, 1981). The test was as follows: $2\Delta ML = 2(ML_1 - ML_0)$, where ML_1 is the DNAMLK (clock) maximum likelihood score and ML_0 is the DNAML (no clock) log maximum likelihood score. In cases where $2\Delta ML > 10.83$ (p-value<0.001 for a χ^2 distribution with 1 degree of freedom: d.f.=*s*-2 where *s* is the number of sequences. Note that the p-value is conservatively low to avoid false positives, while the change from p-value=0.05 to p-value=0.001 has resulted in rejecting only two extra sets) the molecular clock was considered to be violated.

2.2.6. Gene Conversion.

As explained in section 2.1.3.4, gene conversion may cause outparalogues to be detected as inparalogues because gene conversions can cause closely related sequences to become more similar. With respect to inparalogue detection and duplication data estimation (the subject of chapter 3), this has the effect of resetting the inferred data to zero or near zero. Because of that, and because the genomic regions upstream and downstream of the gene conversion are likely to continue diverging at the expected evolutionary rate, detecting gene conversion by means of genetic distance or phylogenetic inference is very difficult.

However, an important characteristic of gene conversions is that most gene converted sequences have a high content of G and C nucleotides. Various studies give the range of 60%-90% (Galtier, 2003, Galtier et al., 2001, Chen et al., 2007, Marais, 2003, Spencer et al., 2006). With the fact that two gene converted sequences always have a very short genetic distance from each other, it was possible to take both factors into account to provide, to a first order of approximation, gene conversions.

For each human inparalogue-candidate pair on the same chromosome, the genetic distance between was calculated using the DNADIST program with the F84 substitution matrix (Felsenstein, 1989). Where two sequences have a relatively high similarities (up to 10% difference, such as is the case between human and chimpanzee orthologues and human inparalogues) the F84 genetic distance log scale (Felsenstein, 1989) is roughly linearly correlated to the percentage of difference between the nucleotides of the two sequences – genetic distance of 0 represents 0% difference between the sequences, and genetic distance of 0.1 is roughly 8% difference. After calculating the genetic distances, GC content of the inparalogue-candidates was calculated. Pairs where the genetic distance was smaller than 0.02 on a scale of 0 to 0.1 and their GC content was greater than 60% were considered to be gene conversion candidates. Figure 2.7 shows that all inparalogue-candidates with a high GC content also have a very short genetic distance from each other and cluster into one well defined group, which is very like to consist of some gene conversion sequences (importantly, since the scope of this work allowed only a preliminary basic attempt to approximate gene conversions, this cluster does not represent statistical significance, but rather a visual representation of the sequences that fall into the criteria of a small genetic distance and high GC content). All inparalogue-candidates that were detected as gene conversion were removed. It is important to note that this candidate gene converted set will consist of false positives – genes that were assigned gene converted candidate status but which are genuine inparalogues. However, since only 9 such sequences were removed, and since a reliable final dataset of true inparalogues is of a greater importance, this should not be considered as a problem.

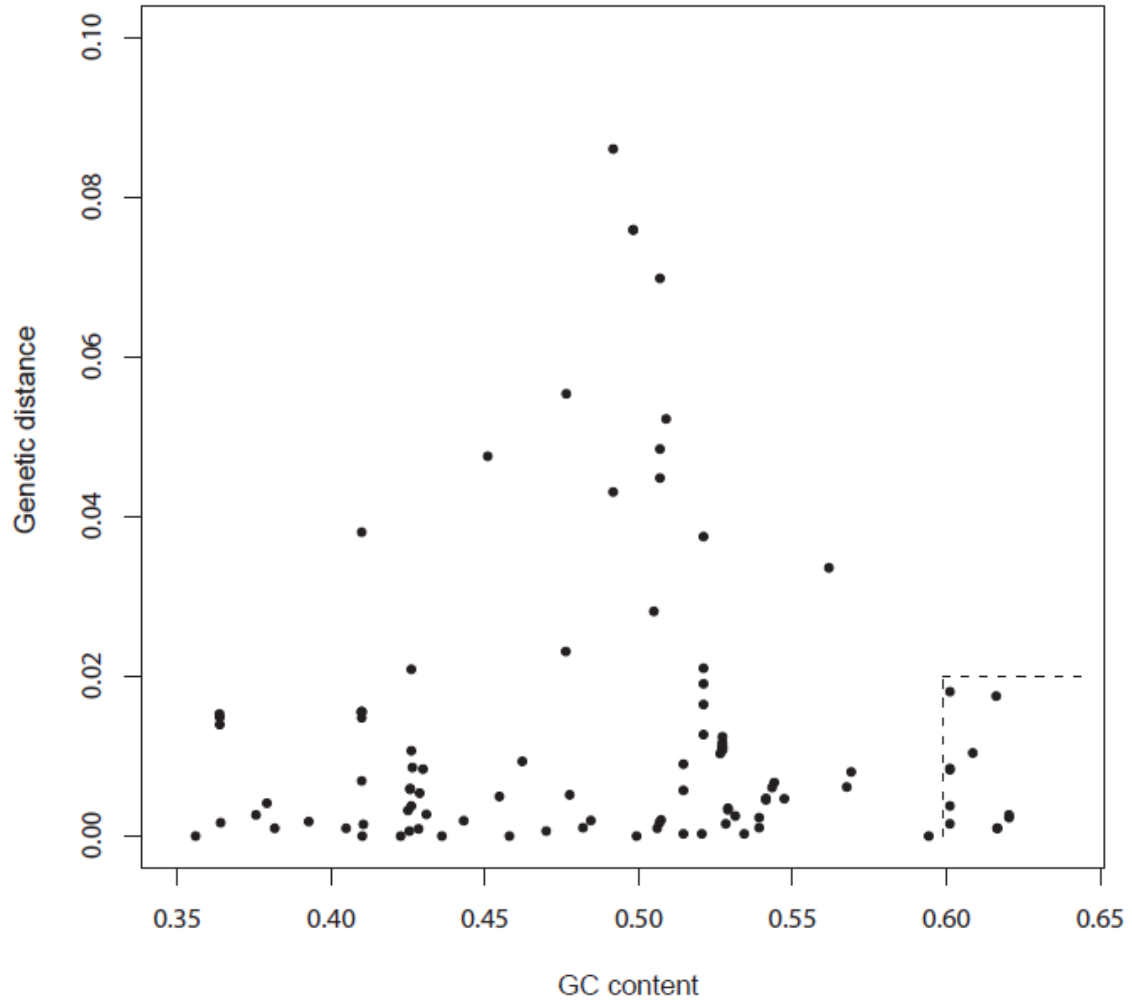

Figure 2.7. The results of testing same-chromosome human gene duplications for gene conversions. For each duplication event genetic distance was calculated between the human orthologue and its inparalogue candidate using the F84 substitution matrix in the Phylip package (Felsenstein, 1989). GC content was calculated by counting the G and C bases in each inparalogue candidate, then dividing by the full sequence length. The dashed lines area shows that all duplications having high GC content are also having a short genetic distance from their orthologues, making them likely to be gene conversions.

2.3. The Final Candidate Human Inparalogues Set.

After applying the full process described in section 2.2, 138 human inparalogues were identified, 104 of them are duplication that occurred on the same chromosome, while 34 are duplications among different chromosomes.

This chapter describes the algorithm that I have developed for finding all human inparalogues, and its application for detecting the candidate human inparalogues and their chimpanzee orthologues dataset. Chapter 3 will explore various characteristics of

the human inparalogue candidates that were detected in this chapter, focusing on estimating the duplication times and functionalities of these genes.

2.4. Discussion.

This chapter described the problems that interfere with acquiring a reliable set of human inparalogues when using the currently available homology detection methods. These problems include the human haplotype data, proteome data, gene conversion, and the use of non-model organisms. Then the chapter described an algorithm that was consequently developed to find a good quality set of human inparalogues.

The algorithm that I have developed and the filtering processes applied are relevant for any model-non model organism inparalogues detection. For an example, in a proposed future project the algorithm can be applied to find cow (an organism with a majority of genes known) – dog (an organism with a majority of genes projected) inparalogues, using a rat (a model organism with a majority of known genes) as an outgroup (see Figure 2.4 for the phylogeny among the 3 species – cow and dog diverged about 62.3mya, while cow-dog diverged from rat about 95.3mya). In this example, since cow and dog are more distantly related than human-chimpanzee, it would be suggested to use BLAST instead of BLAT, as it is more sensitive for more distant homologies (see section 2.2.3). It should be noted that the more distantly related the species are, the more likely it is for conserved genes to be detected as orthologues.

Future applications for this algorithm may be, for example, identifying the full inparalogues datasets for all model/non model organism pairs.

The combination of the GC content and the genetic distance test results (Figure 2.7) could be further analysed in future studies. For examples, it is evident that there are two prominent clusters (determined visually by me and not though statistical testing) of inparalogues having a very short genetic distance (suggesting recent gene duplications) having GC content between 42.5%-44% and between 50%-55%, respectively. Investigating the molecular and evolutionary implications of these could prove informative.

Some improvements envisaged for future versions of the algorithm are a more robust process of detecting gene conversion, and creating a fully automated pipeline of the process described in this chapter. Such an application would take as an input two evolutionary neighbouring species and their outgroup species proteomes, and providing, as an output, the full set of inparalogues after performing the filtering, clustering, and tree inference procedures.

CNV data can also be incorporated for testing the robustness of the inparalogues results. For example – each human-lineage duplication could be tested against the equivalent human CNV gene data and check whether the duplication is polymorphic, or rather if it is a duplication that is fixed in that species.

To better detect gene duplications, InParanoid could be adapted to use BLASTN instead of BLASTP. This would result in the input being a non-redundant filtered genome, rather than the proteome data of the species. However, testing this option has revealed that current conventional computing power is insufficient for such a task. A test run of 3,000 human and chimpanzee sequences (1500 from each species), where the upper threshold for one sequence length was 300,000 base pairs required about 4GB RAM. A full genome InParanoid run with about 20,000 gene from each species and no sequence length threshold (which may include sequences of a million base pairs or more) would require about 50 RAM (a very rough approximation, assuming that the full non-redundant genomes will be more than 10-15 times larger than the 3,000 human and chimpanzee dataset that was tested). Adapting InParanoid for distributed computing may enable such a task at a feasible time (i.e. in a number of weeks or less) and computer memory. Another possibility would be to make the BLAST algorithm more memory efficient.

In the same spirit of this chapter, identifying other “all human-lineage” genomic events of other classes may be performed, such as: pseudogenization (Wang et al., 2006), regulatory regions changes (Montgomery, 2009), retroviral insertions and sequence deletions (Costantini and Bernardi, 2009), genomic rearrangement (Zhang et al., 2009), and various (not strictly genetic) epigenetic effects (Lee and Mahadevan, 2009). For all such searches, the particulars of differences in annotation quality would need to be accounted for.

3. Estimating Dates of Human Lineage-Specific Gene Duplications.

3.1. Introduction.

Gene duplication is a class of large scale genomic events that is likely to have contributed to the shaping of the human phenotype in the short evolutionary time since the divergence of human and chimpanzee – approximately 6.6 million years before present (Steiper and Young, 2006). A duplication of a gene can result in several outcomes: pseudogenization of one copy, different expression levels, or (the less common option) one copy retains the original function while the other copy (or copies) develops new functionalities (Ohno, 1970, Prince and Pickett, 2002). Chapter 2 describes the method that I have developed for detecting human lineage gene duplications (inparalogues).

The major evolutionary events that have lead to the modern human phenotype are traditionally being studied by palaeoanthropology – fossil record of the various human genera from the human-chimpanzee divergence and until present. The most distinct morphological characteristics of modern human are bipedalism and a brain three times larger than chimpanzee's. Carbon 14 dating of human fossil record provides a timeline of these significant morphological changes. See sections 1.2 and 1.3.1 for human-chimpanzee phenotypic differences and for human fossil record, respectively.

Molecular evolution techniques can be applied to estimate the dates of gene duplication events, and specifically – the dates of human inparalogues that were identified in chapter 3. Under the Null hypothesis the dates of human gene duplications are expected to be randomly distributed along the timeline from human-chimpanzee divergence until present. However, gene duplications are large-scale genomic events that are likely to have had a significant impact on the human phenotype in a short evolutionary time. Therefore, I hypothesize that the human lineage timeline contains clusters of duplication events. Moreover, it is possible that these duplication events are correlated to some of the significant human morphological changes that are documented in fossil record. If

such clusters are identified, the functionalities of these duplicated genes and the correlation of their duplication date with fossil record could provide for the first time a genomewide correlation between fossil record and human genomics.

In this chapter I will estimate the dates for all human inparalogue candidates, using maximum likelihood and Bayesian techniques from the software PAML – Phylogenetic Analysis by Maximum Likelihood (Yang, 2007, Yang, 1997), automating the duplications dating process for all human inparalogues. I will then identify the function and possible gene enrichment (a statistically significant overrepresentation of a specific function) for all duplicated genes using the Gene Ontology (GO) (Ashburner et al., 2000) and DAVID (Database for Annotation, Visualization, and Integrated Discovery) interfaces (Dennis et al., 2003), and finally – will attempt to identify clusters of gene duplication times by Quality Threshold (QT) partitional clustering algorithm (Heyer et al., 1999). I will then discuss the correlation between human fossil record, genomics, and evolution of function.

3.1.1. Primate Evolution and Human-Chimpanzee Divergence.

Chimpanzee (together with bonobo) is human's nearest living organism. For this reason it is ideal to use chimpanzee orthologues as outgroups for finding human inparalogues (see chapter 2) and for rooting the molecular clock estimating the dates of the inparalogues duplications. This section will briefly review the primate evolution timeline, which leads to the divergence of the human and chimpanzee lineages. See Figure 3.1 for primate phylogeny and estimates divergence times.

Although fossil record is considered to be a more robust evidence for species divergence than molecular clock dating, there is a scarcity of ancient primate fossils and so the earliest fossils for a genus are unlikely to be available. For estimating a time range for species divergence, fossils provide a good estimate for the "minimum" age of branching, but they are poorer for estimating a "maximum age" (Benton and Donoghue, 2007). For this reason molecular clock estimates for species divergence often predates the fossil record estimate.

The first fossil evidence for primates is dated about 65mya, at just about the time of the Cretaceous-Tertiary mass extinction event. It is suggested that the first primates, the Plesiadapiforms, were small tree dwelling insect eating mammals (Van Valen and Sloan, 1965).

There is more conclusive primate fossil evidence the Eocene era, between ~55-35mya, where the major clades of Prosimians (which include lemurs, lorises, etc.) and Simians (which include old and new world monkeys) started to show their distinct characteristics. Eocene primates were widespread in the Old World and North America, with their population declining at the mass extinction caused by global cooling at the end of the Eocene (Fleagle, 1998).

Molecular clock dating estimates the divergence time of Catarrhini (Old World monkeys and apes) and Platyrrhini (New World monkeys) to be approximately 42.9mya. Earliest Old World monkey fossil is dated between 32 and 37mya (Benefit and McCrossin, 1997). New World monkeys are thought to have diverged from Catarrhini by migrating from Africa to South America across the Atlantic Ocean in a natural raft of floating mangrove vegetation (Sellers, 2000). Some prominent differences between Old World and New World primates are the flat nose and side-facing nostrils of the Platyrrhini, most Platyrrhini males lacking trichromatic vision (e.g. being colour blind), and unlike Old World monkeys – most Platyrrhini have monogamous pair bonds with paternal care of infant (Garber et al., 2008, Sellers, 2000, Jacobs et al., 1996).

There is a relative wealth of African Hominidae (great apes) fossils from the Miocene period, 23-25mya, suggesting that the numbers and diversity of apes was greater than today (Begun et al., 1997). Molecular clock estimates the divergence between great apes (orangutan, gorilla, chimpanzee, bonobo, and human) and Old World monkeys (rhesus macaque, baboon, langur and more) to be about 30.5mya. Old World monkeys differ from apes by having a smaller body and mostly having tails (Sellers, 2000). Apes show high cognitive abilities when compares to the other monkeys, with abilities including the use of tools, complex problem solving, and arguably the ability to acquire a basic form of language and culture (Whiten et al., 1999).

The mechanism and date of the divergence between human and chimpanzee is a matter of a long ongoing debate. One hypothesis maintains that human and chimpanzee underwent allopatric speciation – one group separated from the ancestral species group and a geographical barrier (possibly the Rift Valley) has separated the two groups for long enough to prevent gene flow and eventually breeding between the two groups became impossible. This mode of human-chimpanzee speciation is supported by a recent genetic study (Webster, 2009). Another theory claims that the mechanism was sympatric speciation – groups separating as a result of sexual preference or specialisation in a specific niche. It is argued that such speciation cannot be captured by conventional genetic studies, but rather should be investigated through modelling and computer simulations (Fitzpatrick et al., 2008). A genetic study has demonstrated that after human and chimpanzee first diverged about 10mya, the two groups had inhabited again the same geographic space less than 6.3mya and interbred before their final speciation (Patterson et al., 2006).

The estimates for the human-chimpanzee divergence time range between 4mya (Hobolth et al., 2007) and 10mya (Benton and Donoghue, 2007). In this chapter I will use the human-chimpanzee divergence estimate of 6.6mya, which was obtained from a primate divergence times study that performed Bayesian analyses of genomic data from 13 primates and 6 mammalian outgroups, while considering the context of divergence time estimates from past studies (Steiper and Young, 2006).

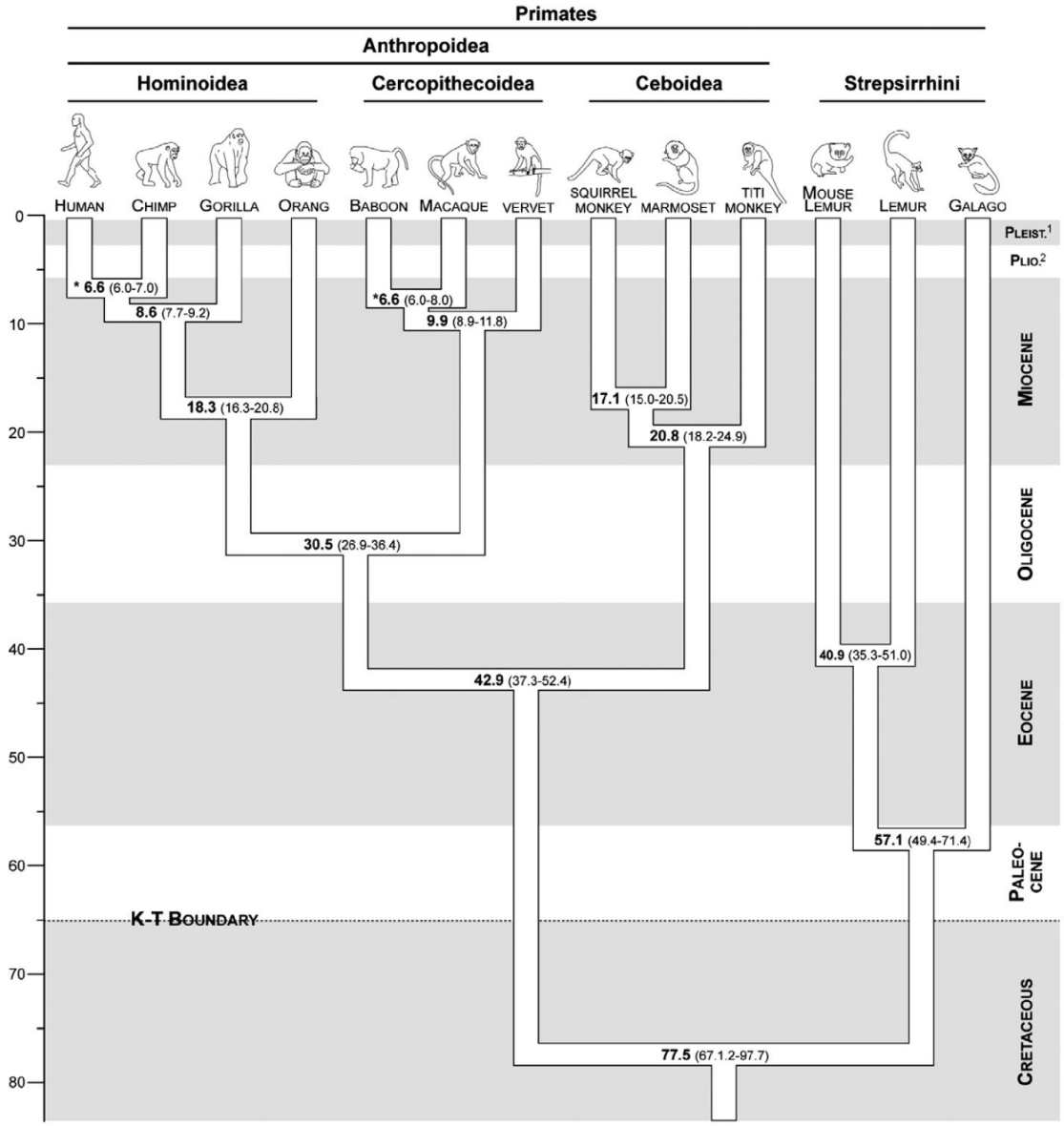

Figure 3.1. Primates phylogeny obtained by molecular clock estimates. The dates on the nodes are in a millions of years scale, with a lower and upper bound divergence time estimate. The K-T (Cretaceous-Tertiary) Boundary is the large scale mass extinction event that had occurred during a short time about 65.5mya. Taken from Steiper and Young (2006), Figure 1.

3.1.2. Hypothesis and Rationale – Clusters of Duplication Events in Human Lineage.

Molecular evolution is the field describing evolution at the genomic and proteomic level. The basic principles of molecular evolution maintain that the main force in evolution is mutations at the DNA level, where harmful (deleterious) mutations are removed and favourable (beneficial) mutations accumulate more than neutral mutations. Selection and drift are the main factors determining whether a mutation will remain or will be removed from the genome and the population gene pool (Kimura, 1968, King

and Jukes, 1969, Yang, 2006a). We will assume that the same basic molecular evolution principles apply to gene duplications as they apply to single mutations.

The core element that I will attempt to determine in this chapter is the dates for all human-lineage duplication events – from human-chimpanzee divergence 6.6mya and until present. If one assumes conditions of no selection and no genetic drift, then according to the molecular evolution principles, the **Null Hypothesis** would be: the dates for human-lineage gene duplications events are randomly distributed along the human-chimpanzee timeline, and don't tend to cluster or to be absent from one time frame or another. However, in various previous sections (including sections 1.3.2, 1.4.1, and 1.4.2) I have described various genomic events in the human lineage that had a direct correlation to the evolution of early and modern human phenotypes – genomic events that have triggered brain expansion, language, and more. Since these genomic events are not randomly distributed along the human timeline, I would expect human-lineage gene duplications to have the same non-random behaviour. For this reason, my **Research Hypothesis** for this study is as follows: human specific gene duplications are dateable events that are likely to have a major role in shaping the unique human phenotype. The dates of these genomic events are clustered around key periods along the timeline from human-chimpanzee divergence until present.

In this study I will not only attempt to identify clusters of human-lineage gene duplication events, but also attempt to detect enrichment for genes with particular functions in these gene clusters (if they are found) – over-representation of specific function around a specific time indicate that these genomic events had a strong drive for this phenotypic trait. I will further elaborate about the correlation between genomics, function, and fossil record in section 3.1.6.

3.1.3. Molecular Clocks and Estimating Duplication Times.

Estimating the date for a gene (or any genetic sequence) duplication event requires the use of a **molecular clock**. The molecular clock basic hypothesis asserts that DNA and protein sequences evolve at a constant rate over time among different organisms (Yang, 2006b). Therefore, the molecular clock hypothesis maintains that the number of nucleotides or amino acid differences between two sequences is proportional to the time

of divergence due to the constant mutation rate over time (a phenomenon termed genetic equidistance), as was asserted by a study of cytochrome C residue differences between mammals, birds, and fish (Margoliash, 1963). Later studies had proposed the **Neutral Theory** of molecular evolution – suggesting that a large fraction of mutations are "neutral" and thus do not affect natural selection, and so these mutations can either be permanently fixated in the whole population, or disappear as a result of genetic drift (Kimura, 1968). This model was demonstrated as over-simplistic in cases where distantly related species divergence was estimated, as evolutionary rates and molecular clock models tend to depend on at least five major factors (Ayala, 1999): (1) generation time of a species; (2) population size; (3) species-specific differences; (4) evolution of function; and (5) changes in selective pressure. To deal with this problem, more sophisticated, realistic, and parameterised "relaxed" models of molecular clocks were developed; in the **global-clock** model the evolutionary rate is around a constant average value determined by a point calibration (see next paragraph), while in the **local-clock model** the evolutionary rate can vary among the different branches of the tree (Yang, 2006b, Yoder and Yang, 2000).

To estimate the evolutionary mutation rate, the molecular clock is calibrated with a known divergence time. For example, assume a phylogenetic tree of two human inparalogues (H1 and H2) and their chimpanzee orthologue C1: ((H1,H2),C1). The divergence time between human and chimpanzee is estimated to be 6.6mya (Steiper and Young, 2006), and so the divergence time between any human and chimpanzee orthologue is also 6.6 million years. The mutation rate is estimated using the substitution rate matrix of choice with a particular calibration point (see sections 2.1, 2.2, and 3.2 for more information about the different substitution matrices). A very simplistic example for demonstrating the estimation of divergence time of two inparalogues using a molecular clock would be as follows: assume a phylogenetic tree of ((X1,X2),Y1@1.0) where X1 and X2 are inparalogues of species X and Y1 is their orthologue from species Y, that diverged from species X 1.0 million years ago (i.e. the calibration point). Assume that using the JC69 substitution matrix which gives the same weight for transition and transversion mutations (Jukes and Cantor, 1969), an average of 10 residue differences is found between the two inparalogues and their orthologue, which is in average one difference per 200,000 years (e.g. the mutation rate, assuming similar rates among orthologues and paralogues). Now assume that there are 3 residue

differences between X1 and X2. Multiplying the mutation rate by the number of differences, the divergence time of X1 and X2 is estimated 300,000 years ago. Note that substitution rate matrices, substitution rate, and calibration points are major elements of dating estimates, but modern models of molecular clocks are much more complicated and parameter rich (Yang, 2007, Huelsenbeck et al., 2001). Reviewing each of these parameters is beyond the scope of this work, though I will explain each of the other relevant parameter that I estimate in section 3.2 (rather than ones that I will keep fixed to their default values, which are determined by the program's authors based on empirical evidence).

There are two major methods for applying the molecular clock – **maximum likelihood** (ML) (Yang, 2007) and **Bayesian** (Huelsenbeck et al., 2001, Yang and Rannala, 2006). ML is a statistical methodology for estimating the parameter value in a model and testing hypotheses concerning the parameters (Yang, 2006a). The output of ML for a parameter is an estimate of a single value which has the highest likelihood (probability) to fit the model. Bayesian methods are based on a prior range for a parameter as input, and a posterior range (rather than a single estimate) of the parameter as an output, where the posterior can be represented in different statistical distribution, such as Gaussian or Binomial (Yang, 2006a, Huelsenbeck et al., 2001). Bayesian statistics were not commonly applied until recently due to the different calculations required by the different methods. While ML is calculated analytically with a few differential equations, Bayesian computation requires iterative and stochastic calculations, where the most common method applied in Bayesian is **Markov Chain Monte Carlo** (MCMC) – an algorithm for sampling from probability distributions of a parameter. In this study I will apply both methods to estimate the dates of human gene duplications – using ML for estimating the evolutionary rate, which will be used as a prior for Bayesian estimation of the dates. I consider Bayesian as more appropriate to make the human inparalogues duplication date estimates because this method allows using soft bounds for the molecular clock calibration values – allowing a range for the human-chimpanzee divergence time rather than a single value (Yang and Rannala, 2006). The ML method can arguably be described as Bayesian since it infers probability.

3.1.4. Studies Dating Divergence Events.

There are only 2 studies known to me that have focused on dating all gene duplications in the human lineage (Gu et al., 2002, Cotton and Page, 2005). However, the studies analysed all human paralogues without differentiating inparalogues from outparalogues, while the timescale examined was of 3,500 million years rather than the 6.6 million years from human-chimpanzee divergence until present. The studies counted gene duplications in time frames of 50 million years and did not attempt to identify clusters or to estimate the function of the duplicated genes. Furthermore, the identification process of human paralogues lacked the filtering processes that I have demonstrated as essential in chapter 2.

Another study has attempted to identify the duplication time of human gene family blocks. It did not use molecular clock for dating, but rather estimated phylogenetic tree topologies for the different duplication events (without using an outgroup) and then compared the trees and branch lengths with the primate-rodent divergence time, which is a crude timescale (Friedman and Hughes, 2003). In this section I will review a few examples of dating species divergence and duplication events using the methodologies that I will apply in this study. Note that the Bayesian dating techniques are more recent and thus there are less examples of Bayesian dating.

A mitochondrial genome study of extinct and extant bear species estimated divergence time of the different species (Krause et al., 2008). The study demonstrated a correlation between climatic changes and speciation and evolution patterns. The dating technique used for the study was mcmctree – a Bayesian MCMC method (part of the PAML package) for estimating divergence times (Yang and Rannala, 2006, Yang, 2007). See section 3.2 for further information about the method. Another study applying mcmctree for divergence time found that the extinct American mastodon diverged from the Elephantidae genera between 24-28mya, African elephants diverged from the mammoth-Asian African lineage about 7.6mya, while mammoth diverged from Asian elephant about 6.7mya (Rohland et al., 2007). A study estimating the divergence time of fish species (Finn and Kristoffersen, 2007) has applied MrBayes (Huelsenbeck and Ronquist, 2001), a Bayesian method equivalent to mcmctree.

A study of human chromosome 1 PRAME (Preferentially expressed antigen of melanoma) genes has shown two large segmental duplications that have occurred approximately 3mya in a cluster that had arisen due to translocation between 85 and 95mya, while both duplication events were shown to have evidence for a strong selective advantage (Birtle et al., 2005). The study applied the baseml program (Goldman and Yang, 1994) for ML analysis of coding and non-coding nucleotide sequences (while the codeml program analyses coding sequence only. Both programs are part of the PAML package).

3.1.5. The Novelty of the Study – Correlating Genomics with Fossil Record.

Previous studies for dating human gene duplications have referred to "human lineage" as the timeline from the emergence of life until present (Gu et al., 2002, Cotton and Page, 2005). Other studies have timed the gene duplications of single specific chromosomes (Birtle et al., 2005), or of gene family blocks without applying a molecular clock (Friedman and Hughes, 2003). There is yet no study that estimates the dates of all human inparalogues – the genes that duplicated from human-chimpanzee divergence 6.6mya until present. Moreover, I believe that any attempt perform such a study using currently available human inparalogue databases would produce results that are not reliable – I have demonstrated in chapter 2 the various problems of previous human inparalogue detection studies (Remm et al., 2001, O'Brien et al., 2005, Tatusov et al., 1997, Hubbard et al., 2009, Stewart et al., 2004). I believe that in chapter 2 I have produced for the first time a reliable set of human inparalogue candidates, and so I made it possible for the first time to produce a reliable study dating human inparalogue duplication events, as I will attempt to perform in this chapter.

A few studies have attempted to correlate dates of human lineage genetic events with fossil record and with significant changes in human morphology and phenotype, such as brain expansion (Stedman et al., 2004) and language capacities (Enard et al., 2002). However, there was yet no attempt to correlate an entire class of genomic events in the human lineage to fossil record. With the full dataset of human inparalogues that I have detected in chapter 2, I will attempt to automate a process of estimating their duplication times. Once this is performed, a correlation between human genomics and human fossil record could be done for the first time, where the availability of many duplication

events (or alternatively, clusters) around or before key periods in human fossil record would show that bursts of large scale genomic events in human history have had a strong role in shaping the modern human phenotype, while if this work detects clusters at times that are less significant in fossil record then it may suggest that fossil record evidence is lacking or that the correlation between genomics and fossil record is following a more complex dynamics that needs to be further studies. Another possibility could be that the duplication events are randomly distributed along the human timeline, and that would suggest that the null hypothesis is correct. I consider each of these potential results as important and novel. This is the first time that such an attempt is performed, and I hope that it will contribute to the interdisciplinary field of human evolution.

3.2. Materials and Methods.

In my attempt to date human lineage gene duplications and correlate them with fossil record I have followed 3 main stages: (1) estimating gene duplication times; (2) clustering gene duplication times; and (3) investigating function and gene enrichment. Each of the stages contains non-trivial and subjective elements, such as prior parameters fine-tuning (stage 1), defining what is a cluster (stage 2), and attributing a biological function to a duplicated segment (stage 3). In this section I will follow the full process that I have employed, using as input the dataset of human inparalogues that was detected in chapter 2.

3.2.1. Human Inparalogues Input.

In chapter 2 I have described the process that I have developed for detecting all human inparalogue candidates. The result was 138 duplication events, of them 34 duplications among different chromosomes, while 104 duplications are on the same chromosome. See Table 3.1 for a summary of all duplications. Although the molecular mechanisms that caused the same and different chromosomal duplications are likely to be different (segmental duplication on same chromosome duplications and retrotransposition on different chromosomes duplications) and the patterns of duplication dates may differ from these two classes, the relatively small number of different chromosomes duplications makes it impossible to perform reliable clustering, so I will analyse the two

classes together as one set. However, I will briefly explore the differences between the same and different chromosome duplications density distributions in section 3.3.

3.2.2. Estimating Gene Duplication Times.

For estimating the human inparalogues duplication dates I used baseml and mcmctree, both are part of the PAML package that contains several phylogenetic analysis tools (Yang, 2007). For a detailed manual of using the different PAML programs see (Yang, 2009).

The program mcmctree implements MCMC methods for estimating sequences divergence times on a given rooted tree using a calibration point or range (Yang and Rannala, 2006, Rannala and Yang, 2007). The two main advantages of this MCMC dating method over ML dating methods (such as baseml or codeml) are: (1) mcmctree allows using soft boundaries prior for the tree root calibration, which reflects the uncertainty of the estimated time range for human-chimpanzee divergence time (Steiper and Young, 2006), and (2) mcmctree calculates a posterior distribution for divergence time, which has a probability of being beyond the soft bound range for human-chimpanzee divergence time. In cases where the duplication time is estimates beyond the upper bound the duplication event is detected as an outparalogue – a false positive that was not detected in chapter 2. Dating such an outparalogue with a ML method would give an estimate of the upper bound, and it would be impossible to ascertain whether it is an inparalogue dated very near to human-chimpanzee divergence time, or rather if it is an outparalogue for which the duplication time is unknown.

The first requirement for mcmctree is to provide the scale (α) and shape (β) for the gamma distribution prior values of the substitution rate. These values were estimated by first evaluating average substitution rate (s) using baseml, a ML likelihood method for parameters estimates, using the following tree: ((H1, H2), C1@0.066), where H1 and H2 are the human inparalogues, C1 is their chimpanzee orthologue, and @0.066 is the point estimate for the human-chimpanzee 6.6mya divergence time calibration point (one unit is 100 million years). I used the F84 substitution model (Hasegawa et al., 1985) which computes genetic distance by considering experimental substitution rates for transitions and transversions. Then the scale and shape where calculated as following:

$\alpha=(s/sd)^2$ and $\beta=s/sd^2$, where *sd* is the standard deviation for *s*, a heuristic value fixed to *sd*=*s*/2.

The main challenge in using mcmctree for 138 sets was the requirement for parameters' fine tuning. The MCMC algorithm uses 4 parameters for the step lengths used in the proposals in the MCMC algorithm. These proposals (1) change the divergence times, (2) change the rates, (3) perform the mixing step (page 225 in Yang and Rannala 2006), and (4) change parameters in the substitution model (such as π_i, one F84 substitution rate parameter). The optimal acceptance proportion for one of these 4 parameters is 0.3, while the acceptance interval is between 0.2 and 0.4 (Yang, 2009). The estimate of these parameters is performed manually by the user, where if the result for the proportion value for a parameter falls below the minimum then decreasing the parameter's value will increase its value for the next run and vice versa – increasing its value in case of a to high value will lower its value in the next run. It is impossible (using mcmctree only) to predict the effect of changing the parameters' values since they dependant on the nature of the sequences used, and are rarely linear (as I've witnessed from various manual testing that I have performed, data not shown). Although after some intuitive trial and error it is very doable to fine tune each parameter to give an acceptance proportion within the interval, it is much more difficult to automate it for 138 sequences due to the reasons explained above. To solve this problem, I have used the binary search algorithm that is designed to locate an element in a sorted list, eliminating half of the list at each iteration, giving a computation complexity of $2(log_2K)$ for a list with no upper bound and (log_2K) for a list with an upper bound, where *K* is the number of elements – an efficient computation complexity (for example, a sorted list with an upper bound of 65,000 elements would require a maximum of 65,000 attempt to locate a random number, while a binary search will find the element in a maximum of 16 attempt). The acceptance proportion for a specific set of genes is monotonically decreased / increased with the parameters values being increased / decreased, respectively, as explain above. Since the acceptance proportion values represent a range rather than a discrete list, I have divided the range into units of 0.01, so I accepted parameters values that gave a high accuracy acceptance value between 0.29 and 0.31. For more information about binary search see (Cormen et al., 2009).

After estimating and fine tuning the parameters, mcmctree was executed using the following tree: ((H1, H2), C1 'B(0.06,0.07)') , where B(0.06,0.07) is the soft bound range for human chimpanzee divergence time that is estimated to be between 6.0 and 7.0mya (Steiper and Young, 2006). I have automated the procedure for the 138 human lineage gene duplications with a Perl script that I have written, and have obtained the estimated duplication dates (see Table 3.1).

Table 3.1. The chromosomal location, biological function, and estimated duplication time of all human inparalogue candidates. Sorted by duplication time – from present to human-chimpanzee divergence.

Chromosome	Start position	End position	strand	Function	Duplication time (mya)
X	52798483	52806246	+	transcription and translation regulation	0.02
11	57738793	57739770	+	sensory perception	0.05
X	70900227	70903498	+	unknown	0.07
8	7776354	7777500	+	immune system	0.08
10	135330645	135331913	+	transcription and translation regulation	0.08
10	135333955	135335223	+	transcription and translation regulation	0.08
10	135343873	135345141	+	transcription and translation regulation	0.08
10	135347172	135348440	+	transcription and translation regulation	0.08
1	610959	611897	-	sensory perception	0.09
5	180726894	180727832	+	sensory perception	0.09
2	89680942	89681746	+	immune system	0.1
10	135337264	135338532	+	transcription regulation	0.1
16	28298620	28322440	-	transcription and translation regulation	0.11
15	81002607	81005939	-	transcription and translation regulation	0.12
2	240633242	240634186	-	inter/intra cellular signalling	0.13
10	47867676	47872093	+	immune system	0.14
10	135340563	135341831	+	transcription and translation regulation	0.14
2	95654725	95655906	+	transcription and translation regulation	0.15
11	55351271	55352206	+	sensory perception	0.15
1	159817852	159828656	+	non-coding duplication	0.18
5	69381269	69408154	+	transcription and translation regulation	0.22
10	81361534	81363821	+	cellular regulation	0.25
17	31648372	31649759	-	immune system	0.31
8	12212843	12220196	-	immune system	0.33
5	69426242	69460017	-	inter/intra cellular signalling	0.42
10	18081268	18127693	+	membrane protein	0.51
4	75699863	75707136	+	cellular signalling	0.52
X	52993880	52994521	-	membrane protein	0.55
10	81594145	81600342	+	unknown	0.56
17	41728277	41770847	+	membrane protein	0.62
8	7716940	7718770	-	immune system	0.63
1	246718513	246719460	+	sensory perception	0.64
10	18138461	18239400	+	cellular transport	0.74
X	153115156	153132153	-	inter/intra cellular signalling	0.77
2	130995528	131001938	+	non-coding duplication	0.8
X	47875014	47876855	+	metabolic and catabolic processes	0.85
10	46578843	46593924	-	inter/intra cellular signalling	0.86
X	153152274	153169958	-	inter/intra cellular signalling	0.91
2	106395969	106451176	-	cellular transport	0.94
9	40690320	40696387	+	membrane protein	1.02
9	41490708	41496772	+	membrane protein	1.02
10	47414578	47468859	+	immune system	1.02
9	41311260	41317335	-	membrane protein	1.03
9	65243336	65249401	-	membrane protein	1.04
X	72009011	72012253	-	transcription and translation regulation	1.07
9	39875004	39881057	+	membrane protein	1.09
9	39345728	39351806	+	membrane protein	1.1
7	143600471	143614992	-	inter/intra cellular signalling	1.11
22	20068430	20073067	+	metabolic and catabolic processes	1.14
8	7742979	7758467	+	immune system	1.22
10	81136040	81142247	+	unknown	1.22
4	69010133	69045117	+	membrane protein	1.25
22	20230346	20234983	-	metabolic and catabolic processes	1.29
10	47221150	47232193	+	cellular regulation	1.31

17	60281179	60323837	-	membrane protein	1.35
8	7720109	7723885	+	immune system	1.53
X	154264958	154266073	+	nucleus activity	1.7
X	154340340	154341455	-	nucleus activity	1.71
12	102948687	102949451	+	non-coding duplication	1.74
22	19892427	19909843	-	metabolic and catabolic processes	1.8
10	135288593	135290236	-	nucleus activity	1.82
10	48873417	48877823	-	unknown	1.85
8	7871325	7872908	-	metabolic and catabolic processes	1.96
8	7259893	7261759	-	immune system	2.04
19	60974983	60976357	+	inter/intra cellular signalling	2.04
X	148851465	148852750	-	unknown	2.11
8	106086	107024	-	sensory perception	2.16
1	246150968	246151840	+	sensory perception	2.18
15	80838110	80838688	-	non-coding duplication	2.38
8	7177319	7178911	+	metabolic and catabolic processes	2.38
9	99000406	99001614	-	transcription and translation regulation	2.49
2	89897079	89897555	+	immune system	2.63
6	170790619	170791556	+	unknown	2.65
16	68765683	68777924	+	extracellular binding	2.66
10	88978236	88984445	+	transcription and translation regulation	2.73
10	89110454	89116661	+	unknown	2.75
8	7182047	7183639	+	metabolic and catabolic processes	2.77
11	76649	77586	-	unknown	2.79
10	47210318	47214775	+	cellular regulation	2.82
2	113911916	113969790	+	non-coding duplication	2.84
7	72272617	72287760	+	inter/intra cellular signalling	2.94
15	100233690	100234627	+	non-coding duplication	2.98
2	109916647	109962880	+	inter/intra cellular signalling	3.02
6	50919	51856	-	unknown	3.05
7	74210536	74225683	-	immune system	3.05
19	107279	108216	-	non-coding duplication	3.13
7	6805380	6832353	-	cellular regulation	3.14
16	73000405	73012669	-	extracellular binding	3.15
8	7866593	7868185	-	non-coding duplication	3.19
2	87852800	87897619	-	cellular transport	3.25
2	87022878	87078395	+	cellular transport	3.29
2	110715607	110760256	-	inter/intra cellular signalling	3.51
16	4381	8789	-	inter/intra cellular signalling	3.53
2	89180465	89180942	-	immune system	3.56
2	89325386	89326322	-	immune system	3.57
2	112852098	112896715	-	inter/intra cellular signalling	3.57
2	89046894	89047372	-	immune system	3.63
9	4807	9213	-	non-coding duplication	3.67
1	4559	8963	-	non-coding duplication	3.7
2	114068374	114072787	+	non-coding duplication	3.7
2	89776410	89776887	+	immune system	3.82
2	89849045	89849560	+	immune system	3.84
2	89100627	89101296	-	unknown	3.86
15	100329588	100333992	+	inter/intra cellular signalling	3.86
2	89276695	89277432	-	non-coding duplication	3.87
2	89662063	89662799	+	immune system	3.9
2	89856037	89856778	+	non-coding duplication	3.91
19	40554102	40555142	+	inter/intra cellular signalling	4.2
9	70046768	70104323	+	metabolic and catabolic processes	4.39
4	8978698	8979891	+	metabolic and catabolic processes	4.44
2	97367012	97367743	-	non-coding duplication	4.49
9	69672249	69729888	-	metabolic and catabolic processes	4.5
9	68494783	68552307	-	cellular signalling	4.57
1	144636184	144651472	-	inter/intra cellular signalling	4.65
1	145933157	145948450	-	inter/intra cellular signalling	4.67
19	1828199	1832565	-	metabolic and catabolic processes	4.73
4	8969207	8970799	+	non-coding duplication	4.86
4	8935989	8937581	+	metabolic and catabolic processes	4.89
4	8945482	8947074	+	metabolic and catabolic processes	4.89
4	8954972	8956564	+	metabolic and catabolic processes	4.89
4	8964462	8966054	+	metabolic and catabolic processes	4.89
4	8973953	8975545	+	metabolic and catabolic processes	4.89
9	106903	108119	-	transcription regulation	4.9
4	8940735	8942327	+	metabolic and catabolic processes	4.9
4	8959717	8961309	+	metabolic and catabolic processes	4.9
4	8950227	8951819	+	metabolic and catabolic processes	4.95
2	89090592	89091069	-	immune system	5.18
11	57963329	57964044	-	sensory perception	5.28
2	130613598	130619039	-	non-coding duplication	5.39
2	89836021	89836467	+	immune system	5.5
8	12280552	12282144	-	immune system	5.51

10	52109175	52115379	+	non-coding duplication	5.55
7	143378450	143379233	+	sensory perception	5.69
11	57668647	57669436	+	sensory perception	5.7
2	89126601	89127078	-	non-coding duplication	5.73
2	89830253	89830730	+	immune system	5.73
2	89012310	89013010	-	non-coding duplication	5.83
11	6847688	6848401	+	sensory perception	5.85

3.2.3. Detection Duplication Dates Clusters.

After estimating the duplication dates I have attempted to check whether the duplications have a tendency to cluster, and if they are clustered then to detect these clusters. All the procedures in this section were performed using the R language (http://www.r-project.org/).

For estimating the degree of clustering in the set of 138 human duplication dates I have first calculated the average nearest neighbour distance (ANND) as follows:

$$ANND = \frac{\sum_{i=1}^{138} D_i}{138}$$

Where i is a gene duplications elements and D_i is the distance (a unit is one million years) between i and its closest date. For example, in a uniform distribution of 138 elements at a time interval of 6.6 million years we would expect ANND=6.6/137=0.048. The ANND value calculated for the set of human inparalogues was 0.021 – much smaller than the uniform distribution. However, since we expect that the duplication dates distribution is random, I simulated 100,000 random sets of 138 elements ranging between 0 and 6.6mya and calculated the ANND value for each of these sets (Figure 3.2). In 95.3% of the simulations, the observed ANND was smaller than the simulated ANND, giving a value of p=0.047 which shows statistical significance for the dates being clustered clustering (which may be due to bias in the dating methodology).

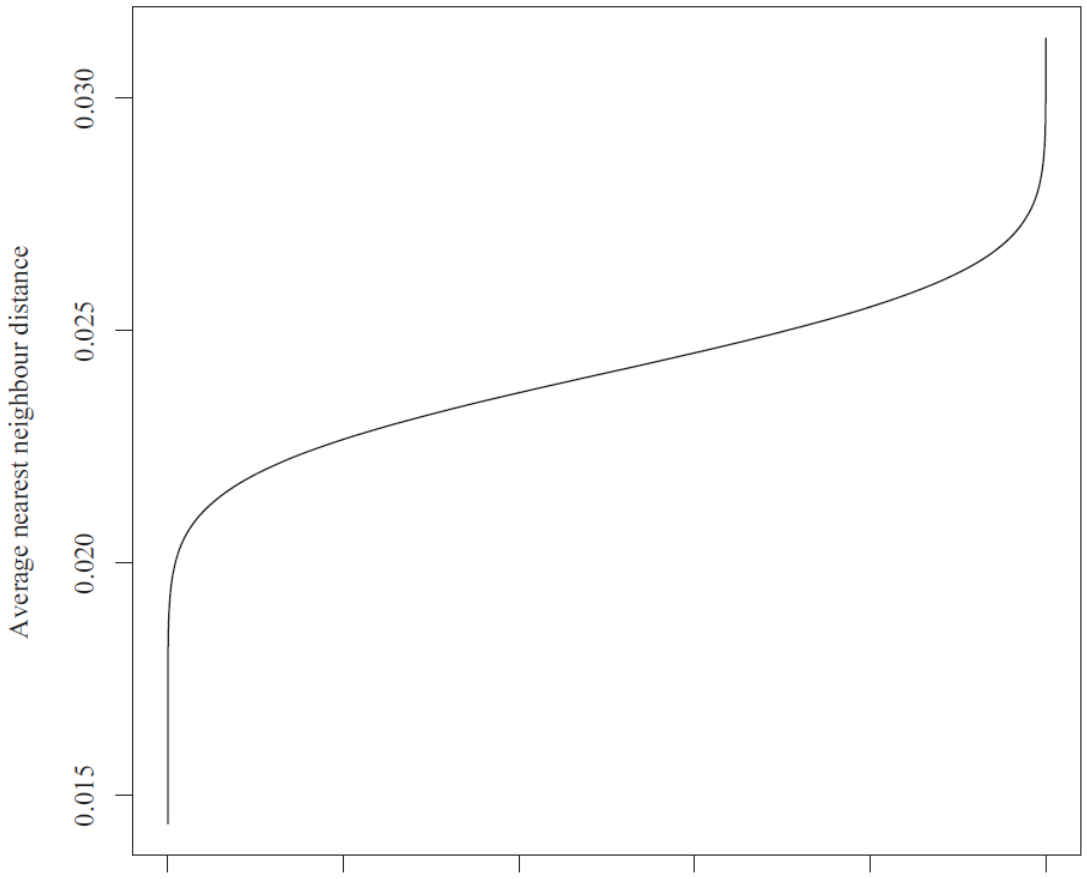

Figure 3.2. Simulated random ANND. The horizontal axis is for the simulated sets sorted by ANND in ascending order.

The next problem was to define a cluster, and such a definition inevitably includes several subjective elements. The main clustering paradigm chosen was partitional clustering – which gives a finite one level set of clusters and cluster centres for the dataset. There are two main ways to define a partitional cluster: (1) by a fixed predetermined number of clusters, and (2) by determining a minimum radius for a cluster, assuming that the number of clusters is unknown. Approach (1) clusters all the elements in the dataset and approach (2) clusters only the elements that are within the radius, and thus there may be elements that are not being clustered. I believe that only relevant dates should be clustered, and so I chose to use and focus on approach (2) (although I accept that clustering with approach (1) could prove informative, the emphasis of this study is not to compare clustering methods). For clustering in the chosen approach (2) I used Quality Threshold (QT) flexible partitional clustering algorithm (Heyer et al., 1999), which is a part of the flexclust R package (http://cran.r-project.org/web/packages/flexclust/). The algorithm is creating a candidate cluster for

each point, where the point is surrounded by other points up to the maximum radius, saves the cluster with most points as the first true cluster, and then recourses with the non-clustered reduced set of points. The "centroid" of a QT cluster is the mean value of all the cluster's points. I first needed to determine a fixed maximum radius value and then minimum number of elements for each cluster. Both values are heuristic, and I used the radius value as the observed ANND*2, and a minimum of 5 elements per cluster (which corresponds to the expected number of elements around one central point).

3.2.4. Assigning Biological Function to Duplications.

The set of 138 duplications obtained in chapter 2 are sequences that are likely to contain coding genes, as they were identified by using BLAT with human coding sequences against the human genome, and the best hits are likely to have functional similarities to the original sequences. Moreover, since I have identified the full length of each duplication, some of the duplications may contain more than one coding gene. However, it is also possible that the full sequence does not contain any gene or that it is a pseudogene.

For estimating the functions of the duplicated genes, I used BioMart, a biological and genomic data mining online tool (Smedley et al., 2009), with an input of the chromosome, start position, end position, and strand of each duplication (see Table 3.1). For each entry I identified the Gene Ontology (GO) (Ashburner et al., 2000) biological and molecular (a lower level) function, as well as obtaining Uniprot (Bairoch et al., 2005, Jain et al., 2009) function data. Whenever available I used the higher biological function. I then manually merged the categories into 10 parent categories: (1) cellular regulation, (2) cellular transport and membrane proteins, (3) extracellular binding, (4) immune system, (5) inter/intra cellular signalling, (6) metabolic and catabolic processes, (7) nucleus activity, (8) sensory perception, (9) transcription and translation regulation, and (10) unknown.

3.2.5. Detecting Gene-Enrichment in Human Inparalogues.

After describing the distribution of the different functions over the human lineage, I checked whether there are functions that are over-represented. For example: assume that in the observed set of duplications, 10% of the genes have an immune system function. If human immune system genes consist of (say) 1% of all human genes, the observed immune system function is 10 times more than expected. Gene enrichment is the term for describing this significant over-representation of specific function in a dataset of genes when compared to a background set of genes. I used the online resource DAVID (Database for Annotation, Visualization and Integrated Discovery) for detecting clusters of gene enrichment (or lack of) in the dataset of human inparalogues using fuzzy heuristic clustering (weighting the degree of belonging of each element to each cluster). The program accepts a set from one specific annotation resource. I used Entrez (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi) annotated genes for DAVID as it was the resource with which the largest number of functions could be identified. The strength of DAVID clustering is that it considers functional information from various experimental resources but has an algorithm to avoid redundancies, and so it provides a broad coverage with a robust evidence for gene enrichment. I have used DAVID with high classification stringency, which is likely to discard false positive clusters. The significance of each biological cluster is measured by a group enrichment score, which is the geometric mean (in log scale) of the annotation cluster member's p-values. Thus, the top ranked annotation groups (with values greater than 1) are most likely to have consistently lower p-values for their annotation members.

3.3. Results.

3.3.1. Distribution of Human Lineage Gene Duplications and Function.

The set of human inparalogues with their locations, duplication date estimates, and estimated function is summarised in Table 3.1. I detected 138 gene duplication events in the human lineage – from human-chimpanzee divergence until present. There are duplication events in each chromosome except in chromosomes 3, 13, 14, 18, 20, and 21.

A density distribution of all human inparalogues is illustrated in Figure 3.3, which demonstrates that there is a large density of duplications between present and 1mya, while there is a secondary high density are of duplications between 3 and 4mya.

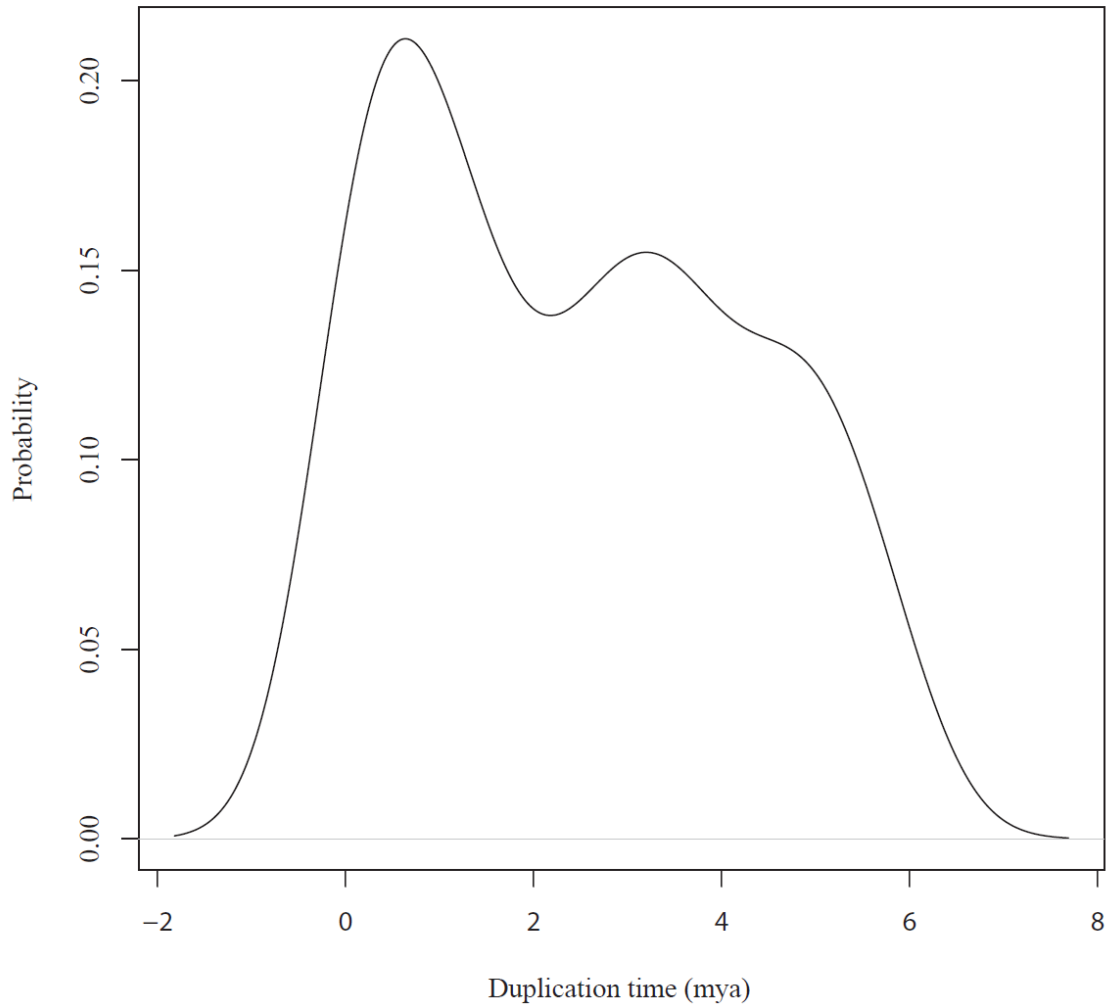

Figure 3.3. The kernel density plot for all human gene duplications. Due to the algorithm used the x-axis shows values beyond the minimum and maximum values for the human duplications dates, and therefore the plot should be used only as a visual representation for duplications density rather than for their precise values.

The number of duplications and the different functions of these duplications in time windows of 500,000 years are shown at the histogram plot of Figure 3.4. The plot reconfirms the density function of Figure 3.3: the average (expected) number of duplications for each time frame is 10.62 (138 duplications divided by 13 time windows), while there are 25 gene duplication events between 500,000 years ago until present, while there are 16 duplications between 3.5 and 4mya. Interestingly, although

human-chimpanzee divergence time is estimated as 6.6mya, the oldest duplication date estimate is 5.85mya.

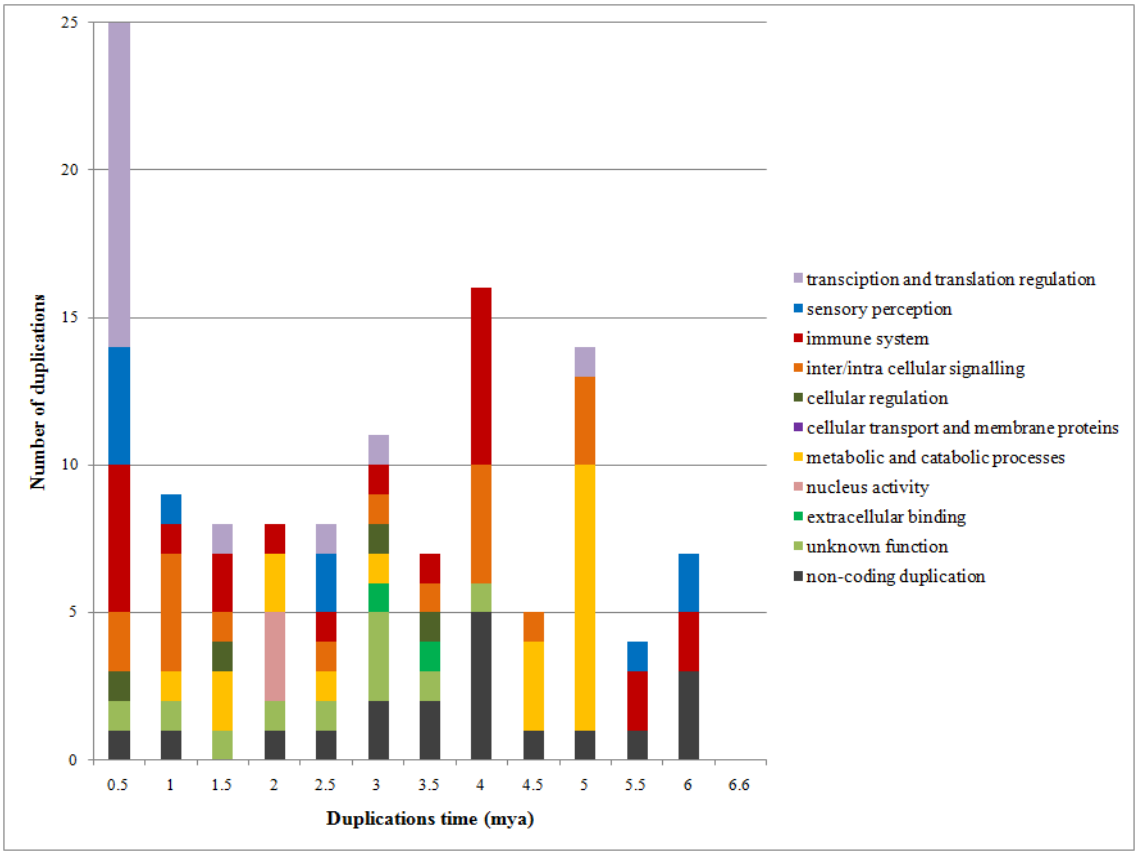

Figure 3.4. Distribution of human lineage gene duplications and functions. Each time window represents 500,000 years, where the histogram with the duplication time value of 0.5 is representing all gene duplication between 0 and 0.5mya, the histogram with the duplication time value of 1 is representing all gene duplication between 0.5 and 1mya, and so on. The colour scheme is similar in figures 3.7 and 3.9.

There is a different distribution of duplication dates between the classes of duplications among different chromosomes and duplication on similar chromosomes (Figure 3.5): the different chromosome duplications tend to accumulate between 3.5 and 4mya (with a secondary smaller peak of very recent duplications), while same chromosome duplications accumulate on a very recent time window – approximately between present and 1mya.

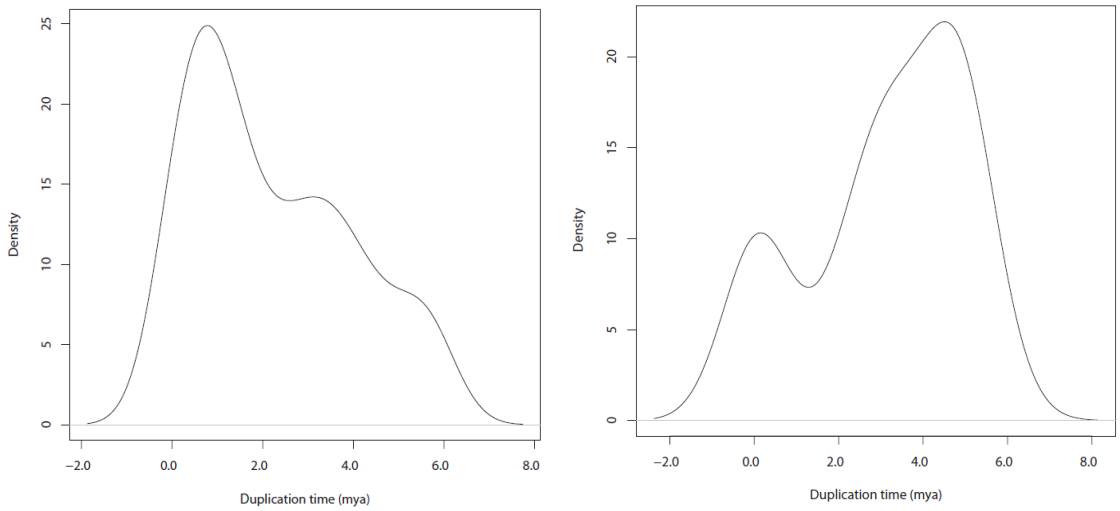

Figure 3.5. The kernel density plot for human gene duplications on same and different chromosomes. The left plot is for same chromosome duplications and the right plot is for different chromosome duplications and. Due to the algorithm used the x-axis shows values beyond the minimum and maximum values for the human duplications dates, and therefore the plots should be used only as a visual representation for duplications density rather than for their precise values.

The distribution of the different duplication functions over time varies with the different time windows (Figure 3.6). Except for the cellular regulation function, all functions have multiple density peaks, while some functions occurred throughout the full range of human lineage timeline and others occurred within a limited time range. The distribution of all human gene duplication functions is shown in Figure 3.7.

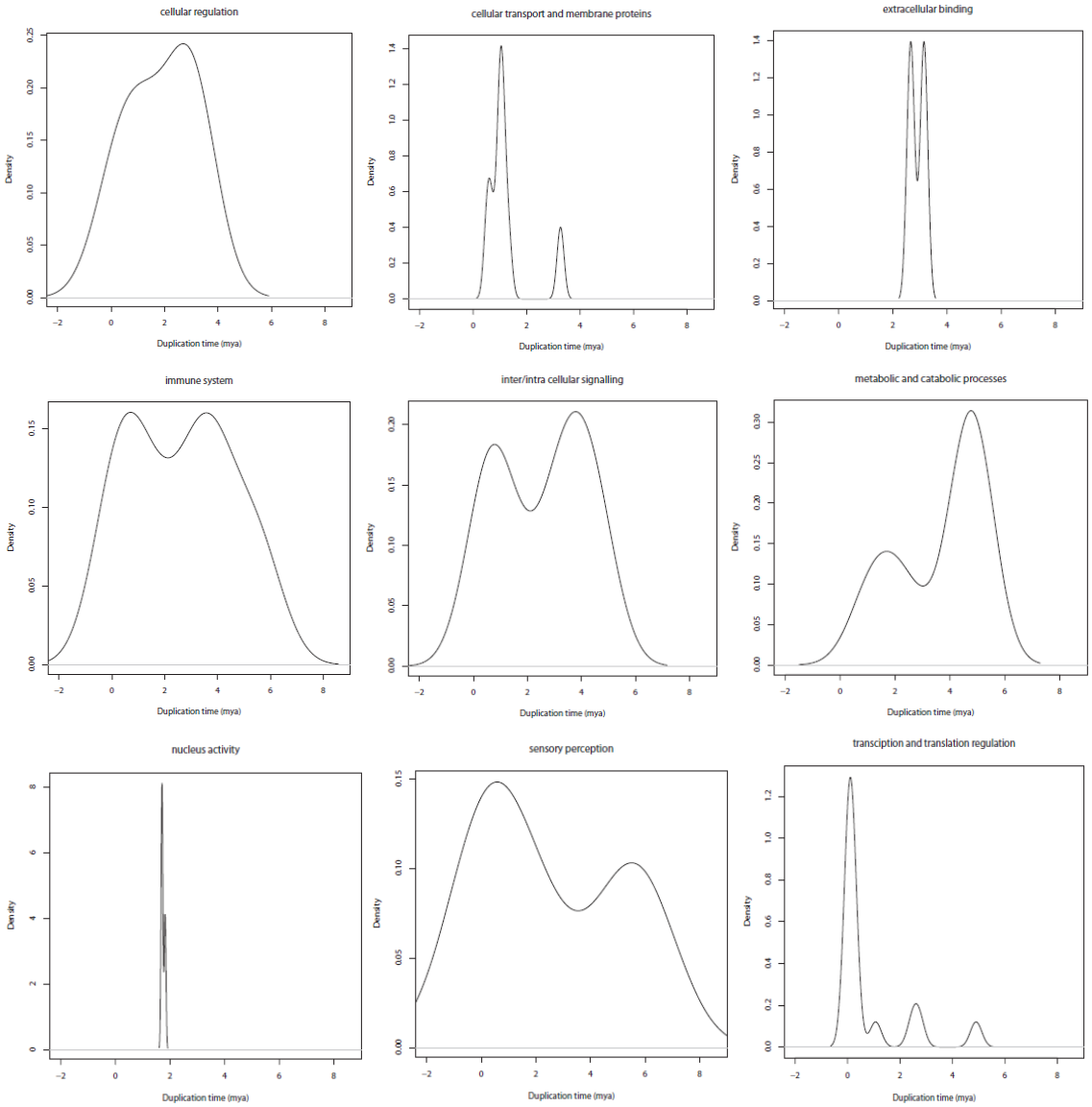

Figure 3.6. The kernel density plots for the biological functions of human lineage gene duplications. Due to the algorithm used the x-axis shows values beyond the minimum and maximum values for the human duplications dates of different functions, and therefore the plots should be used only as a visual representation for density rather than for their precise values.

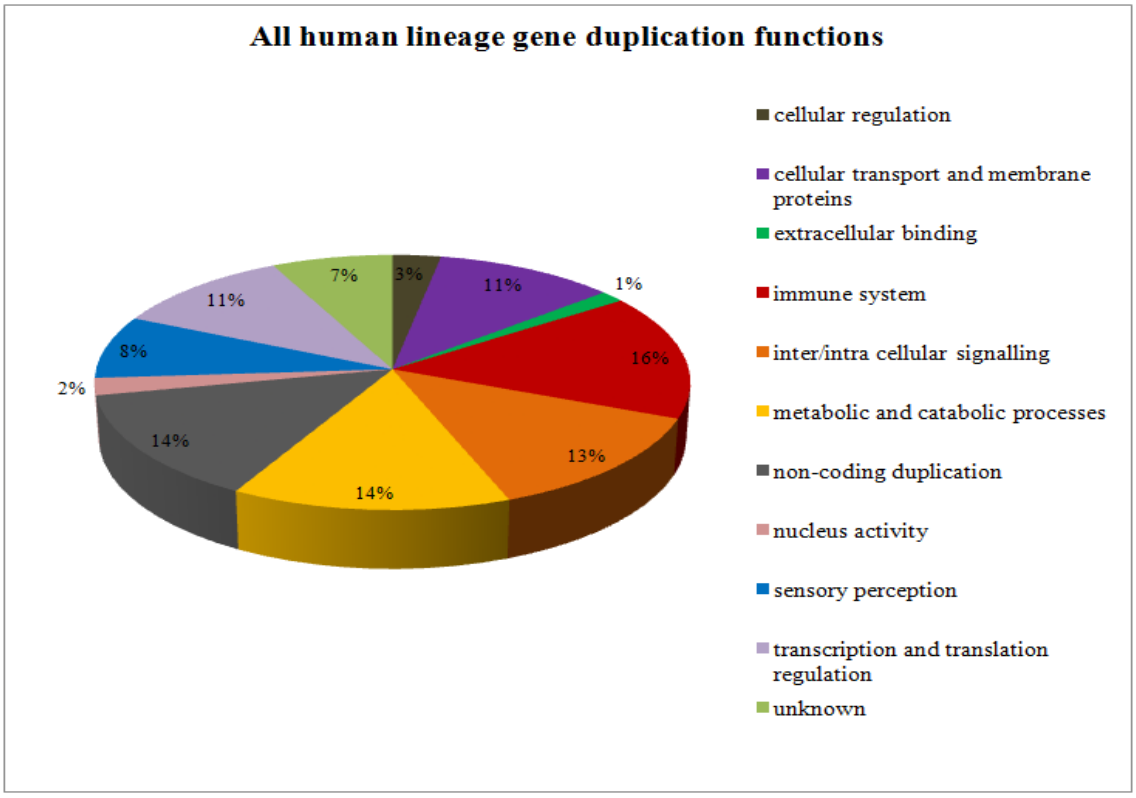

Figure 3.7. Distribution of all human lineage gene duplication functions.

3.3.2. Clusters of Human Inparalogues Duplication dates.

As mentioned in section 3.2.3, I consider the QT clustering as a more robust form of clustering in the context of human duplication dates, and therefore I will focus on the QT analysis that identified 5 clusters. See Table 3.2 for a summary of the clusters and their centroids, with Pam clustering ((Kaufman and Rousseeuw, 2005) used as a control. Pam is an improved version of the commonly used K-means algorithm (MacQueen, 1967) that clusters n observations into k clusters, where each observation belongs to the cluster with the nearest mean. The "medoid" of a Pam cluster is the median value for all the points in the cluster.

Table 3.2. Five clusters of human inparalogues dates. Using QT as the main method and Pam as the control method.

QT clusters			Pam clusters		
N duplications	Centroid (mya)	Diameter (million years)	N duplications	Medoid (mya)	Diameter (million years)
17	0.105	0.13	29	0.130	0.54
8	1.049	0.08	33	1.090	1.23
8	3.548	0.06	27	2.750	1.23
5	3.858	0.08	19	3.700	0.95
6	4.890	0.04	30	4.900	1.46

Considering the cluster centroid times from the most recent to the most ancient, the largest cluster is also the most recent one, centred 105kya and ranging from 118kya and 92kya – within the anatomically modern human time period (see section 1.3.1 for human lineage palaeoanthropological times and Figure 3.9 for the five clusters). The second cluster is around 1.049mya – the Homo erectus genus, the third and fourth clusters are around 3.548 and 3.858mya, respectively, and both are from the Australopithecus genus, while the oldest cluster is around 4.89mya – a disputed era in fossil record, which was a transition between Orrorin (the most ancient species genus in human lineage) and Australopithecus. See section 3.4 for discussion of the clusters times.

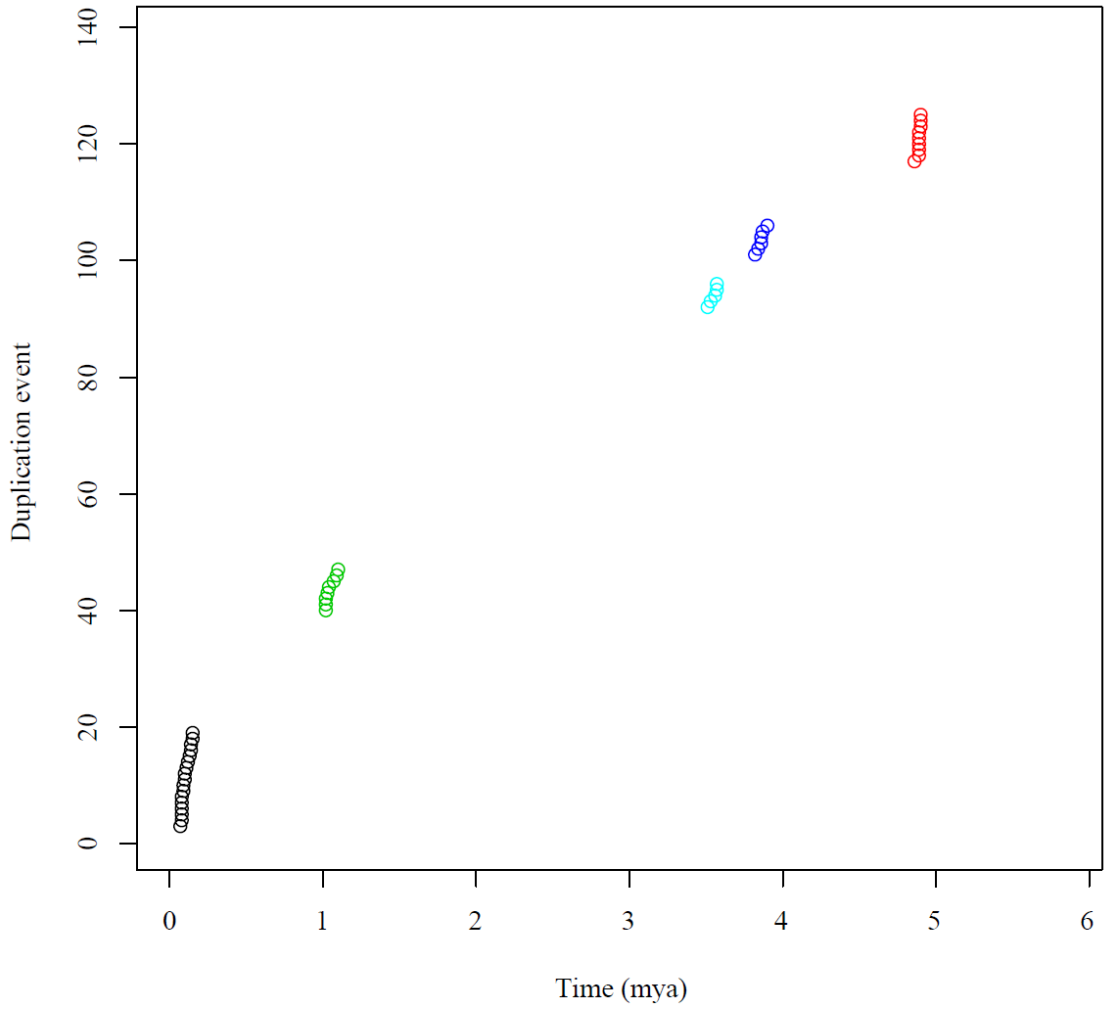

Figure 3.8. The five clusters of human gene duplication dates obtained by the QT clustering method. The colours only serve to visually differentiate between the clusters.

The different clusters contain different dominant functions, as could be expected from Figure 3.5. There is a limited number of functions for each cluster – while there is a total of 10 functional classes, the number of different functional classes in each cluster ranges between 2 and 5. The most ancient cluster contains mostly metabolic and catabolic functions whereas this function does not appear in any other cluster, while the other clusters are dominated by regulatory, immune system, cellular signalling and transport functions. See section 3.4 for a discussion of functions in different clusters and Figure 3.9 for the distribution of functions in the different clusters.

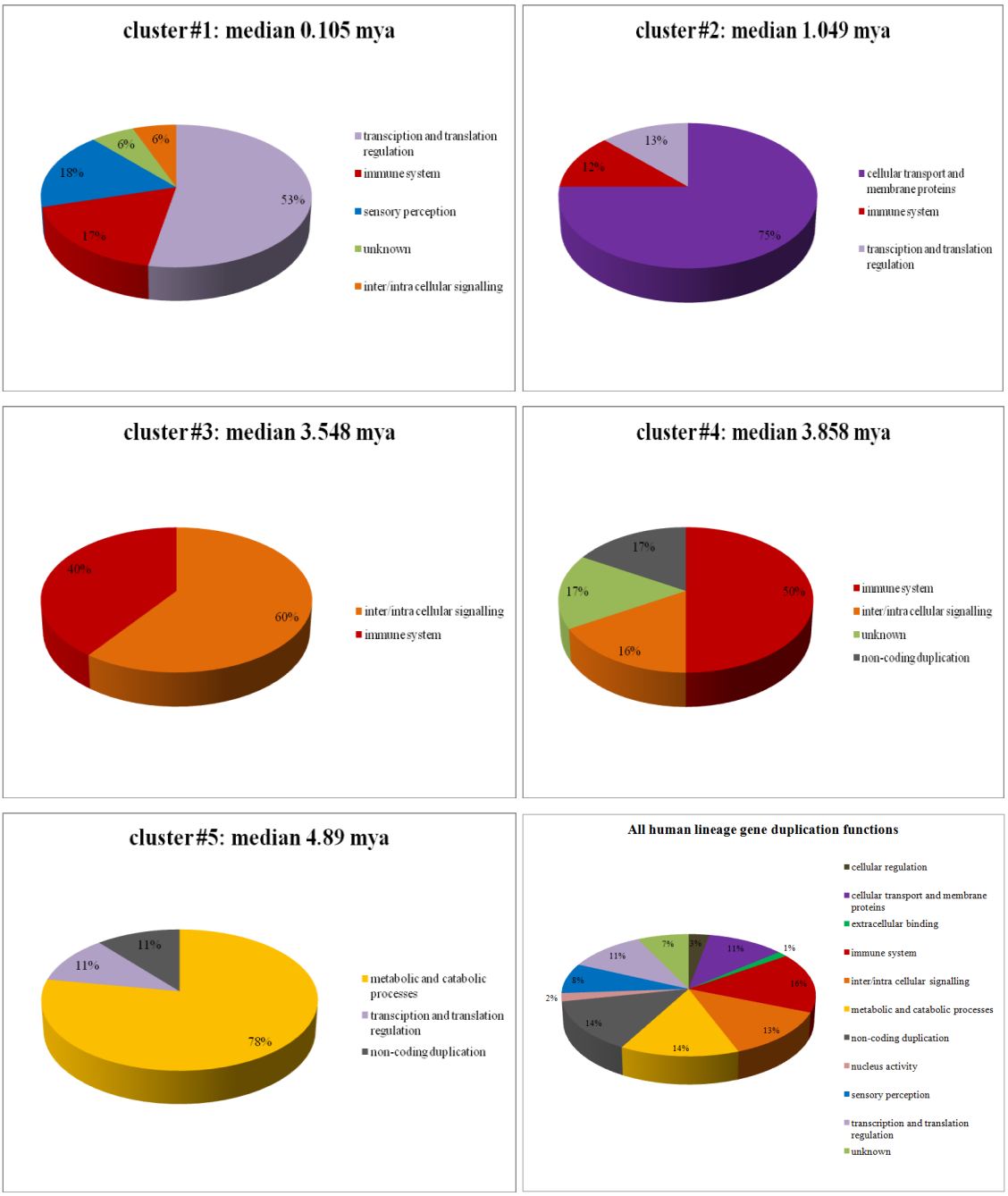

Figure 3.9. The distribution of function in the five clusters of human gene duplication dates obtained by the QT clustering method. The bottom right cluster is similar to Figure 3.7. There is one specific colour designated for each function.

3.3.3. Gene Enrichment in Human Inparalogues.

Describing the different functions of human inparalogues can give insights regarding acquisition of different functions over different time windows. However, it may be that some of the functions that seem to "dominate" a specific time window or cluster is simply because their number is greater than the other functions (i.e. they are actually represented in the same proportion of their actual distribution in the human genome, and

thus their representation is not significantly greater than expected). I have described in section 3.2.5 the DAVID gene enrichment tool that I have used for identifying functional cluster in the human inparalogues set.

Table 3.3. The human inparalogues gene enriched functional clusters identified by DAVID.

Biological function	Group enrichment score
Immune system	5.39
Sensory perception	1.63
Metabolic and catabolic processes	1.37

Table 3.3 shows the three gene-enriched biological groups that were identified by DAVID with high classification stringency. The group with the highest score (i.e. having the highest significance) is immune system, while sensory perception and metabolic and catabolic processes also show highly significant representation in the human inparalogues dataset.

3.4. Discussion.

In this work I have attempted for the first time to estimate the dates of all gene duplications in the human lineage. I have then attributed biological functions to these gene duplications, identified clusters of duplication times, described the accumulation of different functions over time, and then performed a gene enrichment test for over-representation of biological functions in my dataset.

One of the main motivations in this study was to correlate human gene duplications with fossil record. I found that the clusters tend to be in two major time windows: a recent one between 1mya and present, and an ancient one between 3.5 and 5mya. These two time windows are significant in human history: the last one million years were the transition from early *Homo erectus* to anatomically modern human (AMH), while the time between 5 and 3.5mya was the transition between *Orrorin* (the first know species in the human genera) and *Australopithecus* where fossil record shows a significant change towards bipedalism. Interestingly, the largest cluster identified (see Table 3.3) is also the most recent one – centred at about 100,000 years ago – after the “out of Africa” event and before the gradual replacement of Neanderthals in Europe. Another

significant finding is the complete lack of gene duplications at the earliest stages of human evolution – there are no duplications between 5.85 and 6.6mya. This might due to a bias in the dataset as I have obtained the set of human inparalogues with a rigorous process of verifying true inparalogues (see chapter 2) and thus rejecting some true ancient inparalogues that are near the human-chimpanzee divergence boundary.

I attempted to check whether specific functions tended to duplicated in specific time windows (Figures 3.6 and 3.9). Some of the functions (cellular regulation, extracellular binding, and nucleus activity) are a very small portion of the full duplication set – 1%-3%; therefore their distribution description may be biased. Moreover, it is important to consider the original density distribution of human inparalogues (Figure 3.3) and assume that this should be the expected distribution of the different functions over time. I discovered that some of the functions tend to accumulate around specific time windows, for example – there was a burst of transcription and translation regulation gene duplications very recently, while the most ancient duplications cluster was dominated (78%) by metabolic and catabolic processes genes. An in-depth association between specific duplicated functions at a specific time window with the fossil record morphological change that occurred during and after the duplications is proposed as an extensive future study. There are abundant possibilities to investigate duplication dates in the context of the various available human genomic classes databases (Lander et al., 2001).

The gene enrichment test identified three biological functions that are over-represented in the human inparalogues set: immune system, sensory perception, and metabolic and catabolic processes. The immune system function, that has the greatest enrichment score, is distributed all over the human lineage timeline from human-chimpanzee divergence until present, suggesting consistent immune system gene duplications and evolution of the immune system had a significant role throughout all human lineage. Sensory perception gene duplications mostly occurred at both very recent and very old time periods, with a stronger representation in AMH time. Moreover, the only duplication times cluster where sensory perception duplicated genes are represented is the most recent one, centred at about 100,000 years ago. I believe that this demonstrates the genetic contribution of human recent cognitive development and that changes in the genes that contributed to the human senses have also played a role during the early

stages of human evolution. This is a preliminary suggestion, and I think that it has the potential to develop into a vast interdisciplinary study that combines human cognition, fossil record, and human genomics. Another over-represented function was metabolic and catabolic processes. Very interestingly, this function appears exclusively in the oldest duplication dates cluster, centred on 4.89mya. This could suggest an adaptation to different diet at the earlier stages of human history, which is reasonable since the climate in East Africa started to become drier about 5mya, where jungles where gradually replaced by savannas (Behrensmeyer et al., 1997), and it is likely that changes in the digestive system allowing consuming food from the new environment would have given a selective advantage. A future study could combine nutrition, fossil record, climatology, and human genomics to further investigate this ancient burst of function and its affect on human evolution.

There are many other possible future studies that could use the results that I have presented in this study. The set of human inparalogues could be tested for its molecular properties: the characteristics of the duplication sizes, the distance between tandem duplications, and the molecular mechanisms that were likely to have caused the duplications. The subject of genome obesity could also be investigated in the context of human gene duplications. This study could be extended to detect all gene duplications in the primates' lineage. Since it extends over approximately 77.5 million years (see Figure 3.1) it will be possible to make the distinction between the date estimates that were obtained on same chromosome duplications and ones that were obtained from different chromosomes duplications. It was impossible to perform this interesting task in depth in this study because the small number of different chromosome duplications did not allow significant and robust clustering, while the density plot of the dates distribution in these two duplication classes (Figure 3.5) visually shows differences between them, where same chromosome duplications tend to accumulate at recent times while different chromosome duplications accumulate at much older times. A correlation between the two different duplication classes and different functions could also prove informative.

Since the molecular clock is inferring dates in a Poisson distribution, it is possible the genuine ancient human lineage duplications were detected as outparalogues. It is also important to note that since the human inparalogues were detected by first using the

human and mouse (about 10 times further evolutionary from human than chimp) proteomes, we should expect a high rate of genes with high conservation.

The methodology that I have developed in this chapter could be applied to investigate other species duplication times, clustering, and function, and correlate them with their fossil record. For example, there is an abundant fossil evidence of the elephant lineage, including the African and Asian elephants, the extinct mammoth and their common ancestor – the mastodon (Lister and Sher, 2001).

Concluding the major findings of this work, I found that among all human inparalogues a disproportionately large number of them were duplicated very recently – around 100kya. I demonstrated that gene duplications of some biological functions tend to have accumulated at narrow time windows rather than being evenly distributed along the whole human lineage. I identified that there are three biological functions that are over-represented in the human inparalogues sets, and I hypothesise that these functions have had an important role in shaping the modern human phenotype.

4. A Worldwide Correlation of Lactase Persistence Phenotype and Genotypes.

This chapter is based on the following article that was submitted to the BMC Evolutionary Biology journal on the 28.07.2009: Itan Y, Jones BL, Ingram CJE, Swallow MS, Thomas MG (2009) A Worldwide Correlation of Lactase Persistence Phenotype and Genotypes BMC Evol Biol.
At the time of writing the manuscript is under the status of "Editorial Assessment".

The content of this chapter will resemble in many parts the original article, with some changes: I will integrate the original article's supplementary information Tables into the main body of this chapter and further elaborate on some relevant subjects that were only briefly mentioned in the original article, such as the *GenoPheno* and *Natural Neighbour* methods. The core work of this study (including data analysis and article writing) was mostly performed by me. Bryony Jones and Catherine Ingram contributed to collating the data, while Dallas Swallow contributed her lactase persistence expertise.

4.1. Introduction.

An estimated 65% of human adults (and most adult mammals) downregulate the production of intestinal lactase after weaning. Lactase is necessary for the digestion of lactose, the main carbohydrate in milk (Ingram et al., 2009a), and without it, milk consumption can lead to bloating, flatulence, cramps and nausea (Simoons, 1969, Heyman, 2006, Swallow et al., 2001, Castiglia, 1994). Continued production of lactase throughout adult life (lactase persistence, LP) is a genetically determined trait and is found at moderate to high frequencies in Europeans and some African, Middle Eastern and Southern Asian populations, but is rare or absent elsewhere (see Table 4.1 and Figure 4.1).

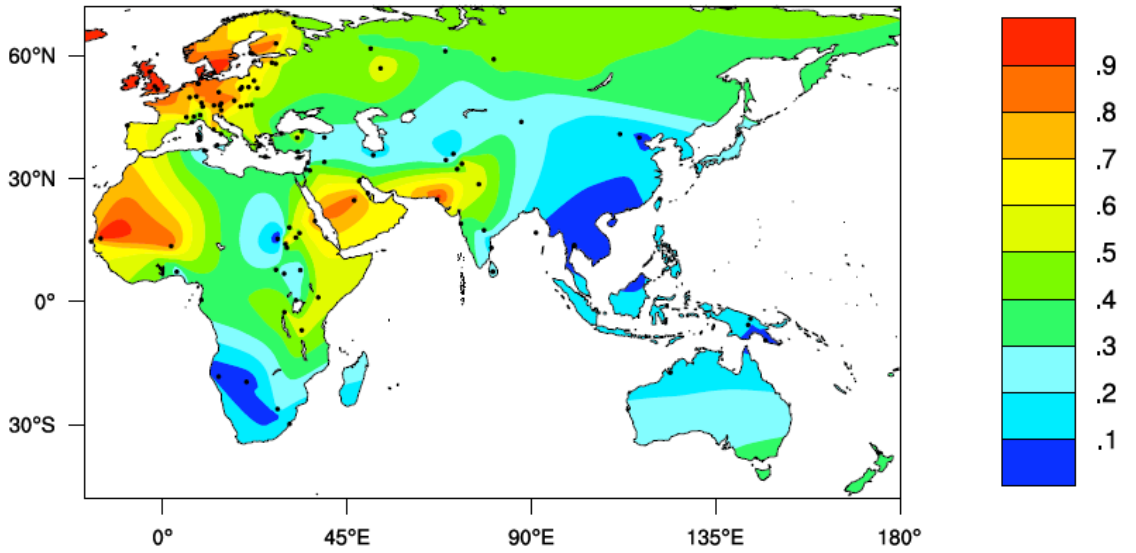

Figure 4.1. Old World LP phenotype frequencies based on all phenotype frequencies. Dots represent collection locations. Colour key shows the frequencies of the LP phenotype.

The most frequently used non-invasive methods for identifying the presence of intestinal lactase are based upon detecting digestion products of lactose produced by the subject (Blood Glucose, BG) or gut bacteria (Breath Hydrogen, BH). For both methods a lactose load is administered to the subject following an overnight fast. In individuals producing lactase this leads to a detectable increase in blood glucose. In individuals who are not producing lactase, the undigested lactose will pass into the colon where it is fermented by various gut bacteria, producing fatty acids and various gases, particularly hydrogen. Hydrogen is highly soluble in the blood and so can be detected in the breath using a portable hydrogen analyser. Both the BG and the BH tests have asymmetric type I (false positive) and type II (false negative) error rates. Thus any study seeking association between a particular polymorphism and LP should take these error rates into account. In addition it should be noted that while in most cases the presence / absence of intestinal lactase in an adult is likely to be genetically determined, the absence of lactase can also be caused by gut trauma such as gastroenteritis (Newcomer et al., 1975, Peuhkuri, 2000). Other non-invasive methods for detecting the presence / absence of lactase include assaying for urine galactose and detecting metabolites of Carbon-14-labelled lactose. These methods are rarely used today. The most reliable method is intestinal biopsy, which provides a direct determination of intestinal lactase activity. However, this procedure is very rarely used for diagnosing healthy individuals because of its invasive nature (Mulcare, 2006b).

With the recent discovery of nucleotide changes associated with LP comes the prospect of direct genetic tests for the trait (Enattah et al., 2007, Mulcare et al., 2004, Rasinpera et al., 2004, Swallow, 2004, Enattah et al., 2002). However, it has become clear that there are multiple, independently derived LP-associated alleles with different geographical distributions (Mulcare et al., 2004, Ingram et al., 2007, Ingram et al., 2009a, Tishkoff et al., 2007, Swallow, 2006). LP is particularly common in Europeans and certain African and Middle Eastern groups. As a consequence these are the regions where most genetic studies have been focused and all currently known LP alleles have been identified (Mulcare, 2006b, Tishkoff et al., 2007, Ingram et al., 2007). The first allelic variant that was shown to be strongly associated with increased lactase activity is a C>T change 13,910 bases upstream of the LCT gene in the 13th intron of the MCM6 gene (Enattah et al., 2002). Functional studies have indicated that this change may affect lactase gene promoter activity (Lewinsky et al., 2005) but, as with all LP-associated variants, there remains the possibility that linkage to an as yet unknown causative nucleotide change may explain observed associations. Haplotype length conservation (Bersaglieri et al., 2004), linked microsatellite variation (Coelho et al., 2005) and ancient DNA analysis from early European farmers (Burger et al., 2007) later confirmed that this allele has a recent evolutionary origin and had been the subject of strong positive natural selection. Furthermore, as I will present in chapter 5, using a simulation model of the origins and evolution of lactase persistence and dairying in Europe, I have inferred that natural selection started to act on an initially small number of lactase persistent dairyers around 7,500 BP in a region between Central Europe and the northern Balkans, possibly in association with the Linearbandkeramik culture.

However, the presence of this allele could not explain the frequency of LP in most African populations (Mulcare et al., 2004). Further studies identified three additional variants that are strongly associated with LP in some African and Middle Eastern populations and/or have evidence of function, all are upstream of the *LCT* gene in the 13th intron of the *MCM6* gene: -13,907*G, -13,915*G and -14,010*C (Ingram et al., 2007, Tishkoff et al., 2007, Enattah et al., 2008, Ingram et al., 2009b). Where data was sufficient, some of these alleles also showed genetic signatures of a recent origin and strong positive natural selection (Tishkoff et al., 2007, Enattah et al., 2008).

Although at least four strong candidate causative alleles have been identified, only a small number of populations have been studied, and those are confined to Europe,

Africa and the Middle East. It is therefore unlikely that all LP-associated or LP-causing alleles are currently known. As a consequence, genetic tests based on current knowledge would underestimate the frequency of LP in most world populations. As part of the first study to seek a genetic explanation for the distribution of LP in Africa (Mulcare et al., 2004), a statistical procedure (*GenoPheno*) was developed to test if the frequency of an LP-associated allele could explain reported LP frequency in ethnically matched populations. Crucially, this statistical procedure was designed to account for sampling errors and the asymmetric type I and type II error rates associated with different phenotype tests (BH and BG).

In this study I have sought to extend this approach to the whole of the Old World. However, while there is a rich literature on the frequencies of LP in different geographic regions (Ingram et al., 2009a) and a growing body of publications reporting the frequencies of candidate LP-causing alleles, in most cases the genetic and phenotypic data are not from the same individuals and often not of the same or closely neighbouring groups. To overcome this problem I performed surface interpolation of various data categories (genetic, phenotypic, sample numbers, phenotype tests used and their associated error rates) and applied the statistical procedures described on a fine grid covering the Old World landmass. This has allowed identification of regions where reported LP-associated allele frequencies are insufficient to explain the presence of LP. These regions should be good candidates for future genotype/phenotype studies.

4.2. Material and Methods.

4.2.1. Data.

My global LP phenotype dataset consists of 112 locations (Ingram et al., 2009a) (see Table 4.1). These data were carefully selected from a large literature on LP frequencies so as to remove data collected from (1) children, (2) patients selected for likely lactose intolerance, (3) family members, and (4) people with twentieth/twenty-first century immigrant status. Genotype data was obtained for 118 locations where the frequency of the -13,910 C>T allele had been estimated (Bersaglieri et al., 2004, Enattah et al., 2007, Ingram et al., 2007, Ingram, 2008, Mulcare, 2006b, Mulcare et al., 2004, Almon et al., 2007), and from 45 locations where the frequency of all 4 currently known LP

associated allelic variants had been estimated ((Enattah et al., 2008, Ingram, 2008, Myles et al., 2005, Tishkoff et al., 2007) and the unpublished work of Ingram et al., 2009. See Table 4.2)). Where there was more than one dataset for a particular location (for either genotype or phenotype data), a weighted average frequency was calculated. The type I and type II error rates used were 8.621% and 6.849%, respectively, for BG and 6.818% and 4.167%, respectively, for BH (Mulcare et al., 2004). Predicted LP frequencies, from the LP genotype frequencies, were calculated by assuming *Hardy-Weinberg* equilibrium and dominance (see Table 4.2).

The geographic space explored for all analyses was from longitude -19 to 180, and from latitude -48 to 72.

4.2.2. Surface Interpolation.

To estimate the distribution of LP and LP-associated allele frequencies in continuous space, from irregularly spaced data, surface interpolation was performed using the *Natural Neighbour* algorithm (Sibson, 1981, Watson, 1992), as implemented in the *PyNGL* module of the Python programming language (Berglund et al., 2008, Watson, 1994, Berndt and Berndt, 1994). This algorithm first divides a 2-dimensional space into polygons according to the locations of the observed data points, then estimates the value at locations for which data is absent by weighting each of the neighbouring locations by their relative overlap. The equation used for *Natural Neighbour* is:

$$F(x,y) = \sum_{i=1}^{n} w_i f(x_i, y_i)$$

where $F(x,y)$ is the estimated frequency at location (x,y) where data is lacking, n is the set of (x,y) location's bordering (i.e. natural neighbours) data points with known frequency data, w_i is the weight for a known data point, and $f(x_i,y_i)$ is the known (observed) frequency for location i.

Other simpler alternative methods for surface interpolation include the *Nearest Neighbour* (Knuth, 1973) and the *Inverse Distance Weighting* (Shepard, 1968) algorithms, where the former uses only one known value when estimating each unknown value (making it over-simplistic) and the latter weights all known values for each unknown value, making it less suitable for a global scale surface interpolation.

4.2.3. Quantitative Difference Correlation Analysis.

I also performed an analysis to quantify the difference between phenotype frequency and predicted phenotype frequency based on the frequency of LP-associated alleles. As in section 4.2.1, I assumed Hardy-Weinberg equilibrium and performed surface interpolation using the data provided in Tables 4.1 and 4.2. I then subtracted the surface representing expected frequencies from that representing observed LP frequencies. Maps were plotted using *PyNGL* (http://www.pyngl.ucar.edu/) (Berndt and Berndt, 1994).

4.2.4. *GenoPheno* Correlation Analysis.

To identify regions where LP-associated allele frequencies are insufficient to explain observed LP incidence I applied the Monte Carlo based statistical test *GenoPheno* (Mulcare et al., 2004). The *GenoPheno* algorithm is defined as follows (Mulcare et al., 2004): (1) A value for p (the frequency of the -13,910C>T allele in the genotyped group) was drawn from a Beta(T+1, C+1) distribution, where T is the number of -13,910C>T alleles and C is the number of -13,910*C alleles found in the genotyped group. This beta distribution describes the posterior distribution for p, given the genotype data, assuming a Uniform(0,1) prior. (2) The predicted frequency of true lactase persistence in the population, L_{true}, was calculated as $p^2+2p(1-p)$ (i.e., the expected frequency of TT+CT genotypes under Hardy-Weinberg equilibrium). (3) Values for f_n (the frequency of false negatives according to the phenotyping method used) and f_p (the frequency of false positives according to the phenotyping method used) were drawn from Beta distributions of the error rates and sampling size. These beta distributions describe the posterior distribution for f_n and f_p, given the combined false error rate data reported above and assuming a Uniform(0,1) prior. (4) The predicted frequency of apparent lactose digesters accounting for phenotyping error, L_{app} (the frequency of apparent lactase persistence in the phenotyped group), was calculated as $L_{true}(1-fp) + (1-L_{true})f_n$. (5) A simulated value for n_L, the number of lactose digesters observed in the phenotyped group was drawn from a Binomial(n,L_{app}) distribution, where n is the number sampled in the phenotyped group. (6) Steps 1–5 were repeated 10,000 times (N=10,000) to build up a Monte Carlo sampling distribution for n_L under the null hypothesis that the C/T genotype and phenotyping error alone account for the

apparent frequency of lactose digesters. (7) Let S_g be the sum of simulated n_L values greater than or equal to the observed n_L value, and let S_l be the sum of simulated n_L values less than or equal to the observed n_L value. A two-tailed *P* value for the observed n_L under the null hypothesis was found as $2\min(S_g,S_l)/N$. In this case the null hypothesis is that the C/T genotype and phenotyping error alone account for the apparent frequency of lactose digesters (Mulcare et al., 2004).

GenoPheno was applied to each cell in a 198 (west-east) by 119 (south-north) grid of covering the Old World. For each cell it was necessary to provide information on LP-associated allele frequencies and LP incidence (see above) as well as on sample numbers used for each data type and type I and type II error rates for the LP phenotype tests used. These parameters were estimated by surface interpolating values from genetic and phenotypic studies to provide 6 surface interpolated 'layers' of information.

4.3. Results.

4.3.1. Interpolated LP Phenotype Frequencies.

Figure 4.1 shows an interpolated map of the frequencies of LP based on phenotype tests (also see Table 4.1, (Ingram et al., 2009a)). Although this map should provide a reasonable representation of frequencies in Europe and western Asia, it should be noted that (1) data is sparse at eastern and northern Asia, Indonesia, Melanesia, Australia and Polynesia, and (2) in Africa and the Middle East it is often the case that populations living in close proximity to each other have dramatically different LP frequencies, depending to an extent on traditional subsistence strategies (Ingram et al., 2009a).

4.3.2. Interpolated Predicted LP Phenotype Frequencies.

Figure 4.2 shows an interpolated map of the frequencies of LP predicted by all 4 currently known LP associated allelic variants, based on genotyping tests (see Table 4.2, (Bersaglieri et al., 2004, Enattah et al., 2007, Ingram et al., 2007, Ingram, 2008, Mulcare, 2006b, Mulcare et al., 2004, Almon et al., 2007)). As with the phenotype data, the genotype data is sparse in eastern and northern Asia, Indonesia, Melanesia, Australia and Polynesia.

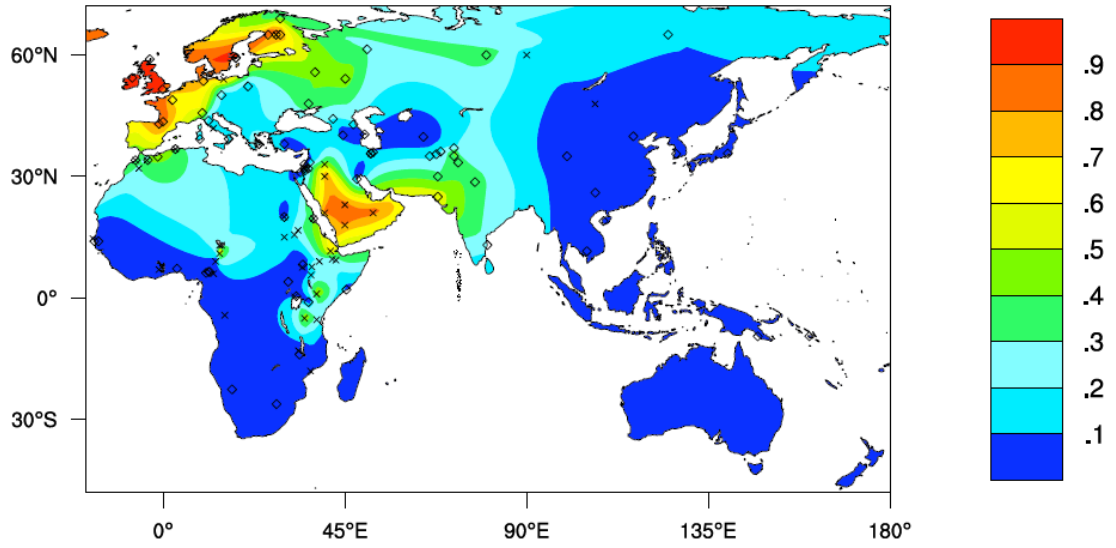

Figure 4.2. Predicted Old World LP phenotype frequencies based on all genotype frequencies. The prediction is assuming Hardy-Weinberg equilibrium. Crosses represent collection locations where all 4 currently known LP-correlated alleles were genotyped, and diamonds represent collection locations where the only data on the -13,910 C>T allele is available. Colour key shows the predicted LP phenotype frequencies.

Figure 4.3 shows an interpolated map of the frequencies of LP predicted by the -13,910 C>T allele data only (see Table 4.2, (Bersaglieri et al., 2004, Enattah et al., 2007, Ingram et al., 2007, Ingram, 2008, Mulcare, 2006b, Mulcare et al., 2004, Almon et al., 2007)).

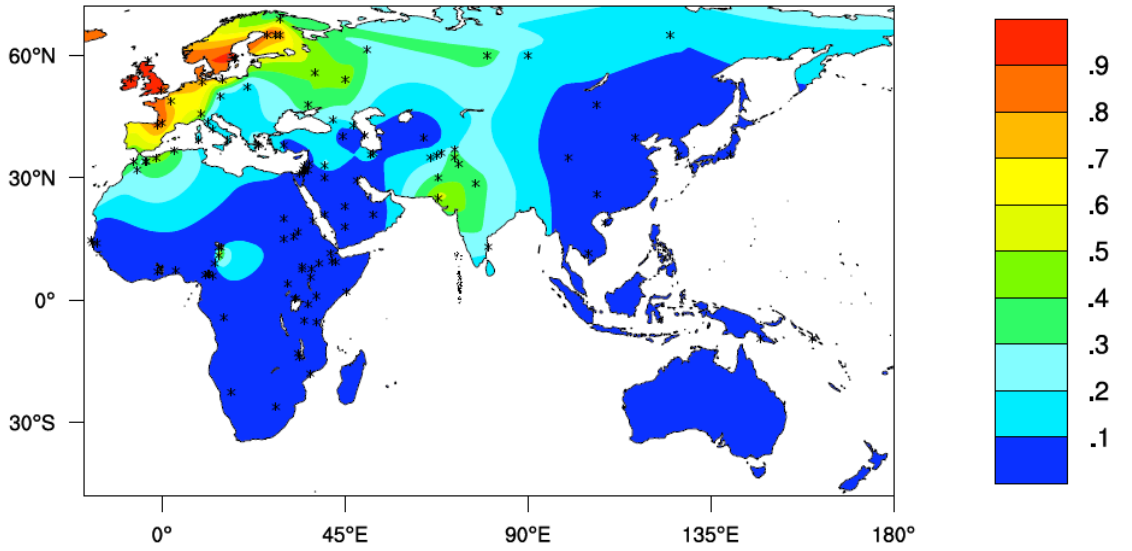

Figure 4.3. Predicted Old World LP phenotype frequencies based on frequency data for the -13,910 C>T allele only. The prediction is assuming Hardy-Weinberg equilibrium. Stars represent collection locations. Colour key shows the predicted LP phenotype frequencies.

Figure 4.4 shows an interpolated map of the frequencies of LP predicted by the 3 currently known LP associated allelic variants, excluding the -13,910 C>T allele (see Table 4.2, (Enattah et al., 2008, Ingram, 2008, Myles et al., 2005, Tishkoff et al., 2007)). This map should provide a reasonable representation of frequencies the 3 LP associated allelic variants in eastern Africa and the Middle East, while data is sparse at the rest of the world.

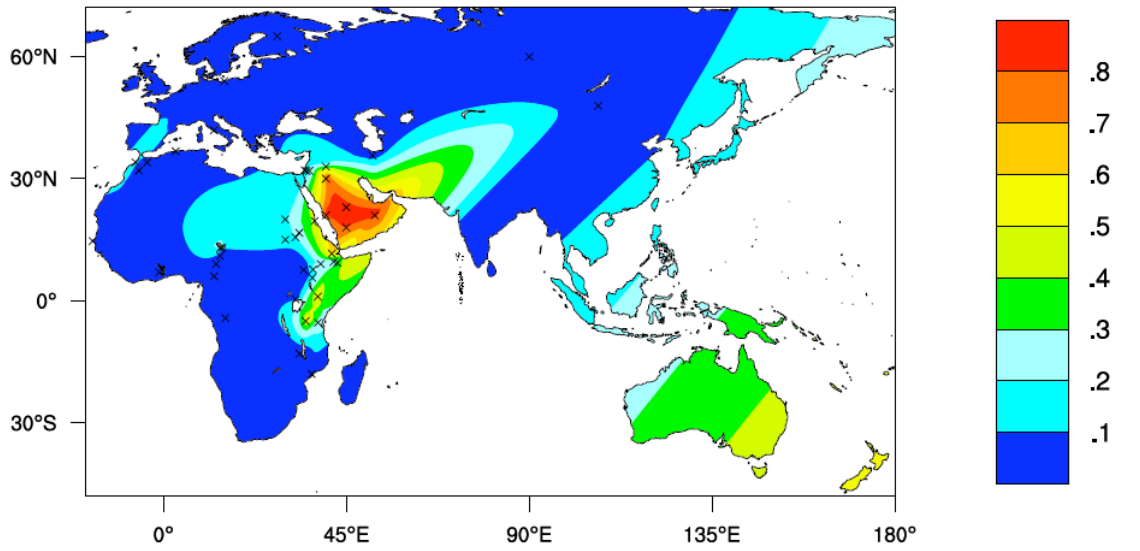

Figure 4.4. Predicted Old World LP phenotype frequencies based on frequency data for the 3 currently known LP associated allelic variants, excluding the -13,910 C>T allele. The prediction is assuming Hardy-Weinberg equilibrium. Crosses represent collection locations. Colour key shows the predicted LP phenotype frequencies.

4.3.3. LP Genotype-Phenotype Correlations.

Figure 4.5 shows the quantitative difference between observed phenotype frequency and predicted phenotype frequency based on the frequency of 4 LP-associated alleles. This map was obtained by subtracting the surface shown in Figure 4.2 from that shown in Figure 4.1. It represents the extent to which current knowledge of the frequencies various LP-associated alleles explains the distribution of the LP trait. In many cases sample numbers used to obtain molecular and phenotype data were small. Additionally, phenotype testing error rates are appreciable. It is therefore possible that, for some regions, where the discrepancies between predicted and observed LP frequencies are high, such differences can be explained by sampling and testing errors alone.

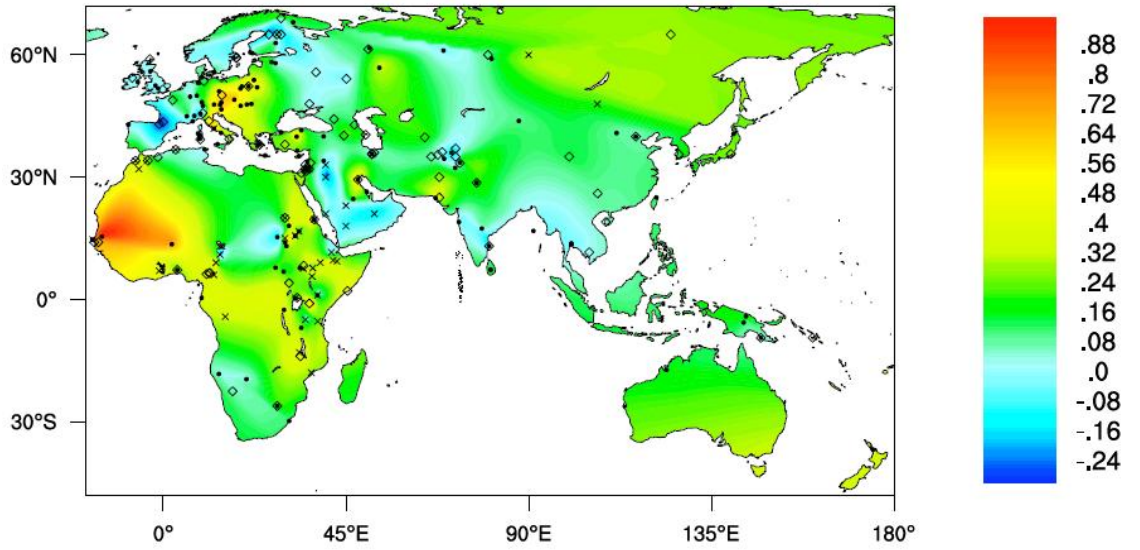

Figure 4.5. Old World LP genotype-phenotype correlation, obtained by calculating the quantitative difference between observed phenotype frequency and predicted phenotype frequency based on the frequency of 4 LP-associated alleles. Positive and negative values represent cases of LP-correlated genotype under- and over-predicting the LP phenotype, respectively. Dots represent LP phenotype collection locations, crosses represent data collection locations for all currently known 4 LP-correlated alleles, and diamonds represent -13,910 C>T only data collection locations. Colour key shows the values of the predicted LP phenotype frequencies (Figure 4.2) subtracted from the observed LP phenotype frequencies (Figure 4.1).

To account for the sampling and testing errors, I have applied the Monte Carlo based statistical test *GenoPheno* (Mulcare et al., 2004) to the surfaces presented in Figures 4.1 and 4.2. Performing this test also requires data on sample numbers and error rates, for which I generated interpolated surfaces by applying the same reasoning as I have to LP frequencies. By applying the *GenoPheno* test to 23562 locations on a on a 198 by 119 cell grid I obtained the surface presented on Figure 4.6. These p-values approximate the probability of the observed genotype and phenotype data under the null hypothesis that the LP-associated alleles and phenotyping errors alone account for the observed LP frequency.

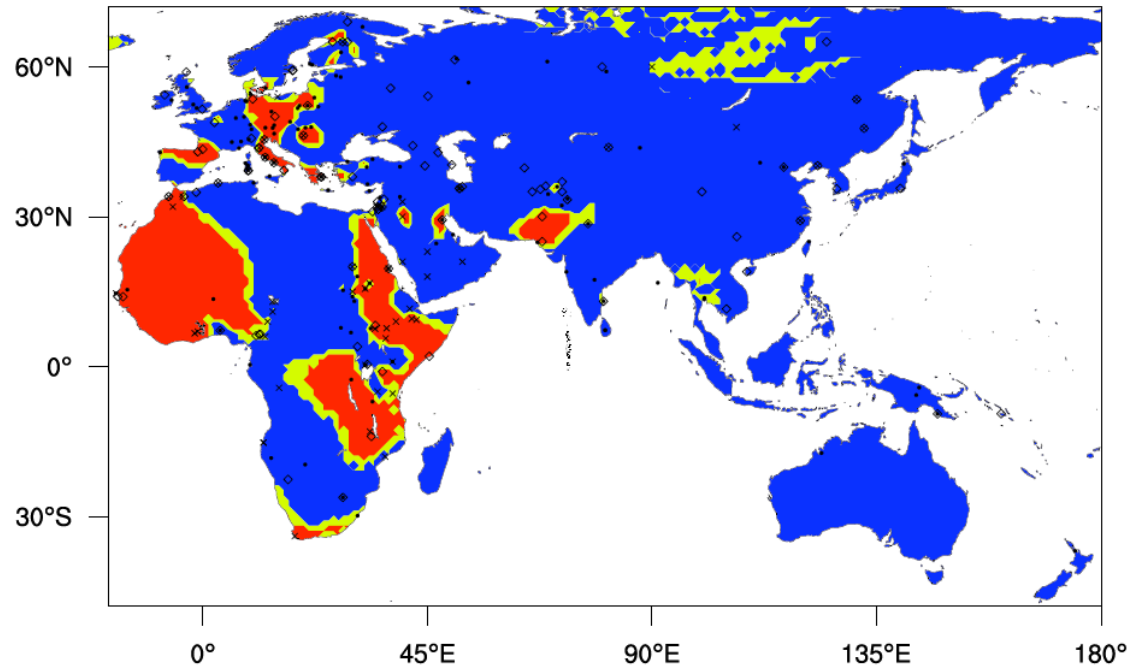

Figure 4.6. Old World LP genotype-phenotype correlation, obtained by the *GenoPheno* Monte Carlo test. Dots represent LP phenotype collection locations, crosses represent data collection locations for all currently known 4 LP-correlated alleles, and diamonds represent -13,910 C>T only data collection locations. Colour key shows the *p* value obtained by the *GenoPheno* test. All values of $p<0.01$, indicating a very significant lack of correlation, are shown in red colour, yellow colour represents a statistical significance of $p<0.05$, while blue colour is for non significance of $p\geq0.05$.

4.4. Discussion.

In this study I have identified regions where the current data on LP-associated allele frequencies is insufficient to explain the estimated LP phenotype frequencies, by surface interpolating LP genotype and phenotype data. The analyses also indicate regions where genotypic or phenotypic data is sparse or non-existent. Data collection from these regions is likely to be of value in developing a fuller understanding of the distribution and evolution of LP. I suggest that regions where LP-associated genotypes are under-predicting LP are good candidates for further genetic studies.

While on a broad scale most regions of the Old World have been sampled for the -13,910*T allele, data on frequencies of the other three LP-associated alleles is localised mainly to Africa and the Middle East. It is likely that further studies will identify appreciable frequencies of the -13,907*G, -13,915*G or -14,010*C alleles, or reveal new LP-associated alleles, in other regions.

The analysis indicated a few regions (the Horn of Africa, Arabia, and the Basque region) where the LP-associated allele frequency appears to over-predict LP phenotype frequency. If we assume that all four LP-associated alleles considered here are causative of the trait, or very tightly linked to causative variants, then it is likely that over-prediction is a result of population sampling problems. For example, the pastoralist Bedouin in Saudi Arabia have high frequencies of LP, while non-Bedouin Arabs from the same region typically have lower frequencies (Hijazi et al., 1983). Similar issues may explain over-prediction in the Horn of Africa (Eritrea, Djibouti, Ethiopia and Somalia), where ethnic diversity is particularly high and the phenotypic and genotypic data are derived in many cases from different ethnic groups with different subsistence strategies (Blench, 2006, Tishkoff et al., 2009). To an extent these problems of matching population groups from the same geographic regions applies to the whole analysis. However, it is notable from Figure 4.5 that where a lack of correspondence between LP phenotype and predicted phenotype frequencies occurs, it is usually when genotype over-predicts phenotype, while under-prediction is rare.

By applying *GenoPheno* statistical procedure to interpolated layers of phenotype and genotype associated data (Figure 4.6), I have identified west and parts of southeast Africa, eastern and southern Europe, and parts of western, central, and southern Asia as potential targets for further genetic studies. A paucity of frequency data for the -13,907*G, -13,915*G and -14,010*C alleles in most of these regions may partly explain this under-prediction (Figure 4.5). The population sampling problems described above may explain the under-prediction I infer in eastern Europe and parts of southern Asia, as in each of these regions, the locations where phenotype and genotype data were obtained are mostly well separated. This population data-matching problem is, however, unlikely to explain the lack of correspondence between LP and allele frequency-based predicted LP frequencies in the region around Pakistan and Afghanistan, as well as in West Africa and Italy. Further genetic studies in these regions should prove informative. I also suggest that the information that I present here could potentially be in use for international health and food aid organisations, to aid with understanding the region population's estimated genotype and phenotype of lactase persistence.

In this study I have demonstrated that lactase persistence genotype data is currently insufficient to explain lactase persistence phenotype frequency in western and southern Africa and several other Old World regions. The identification of additional LP-associated or LP-causative alleles, especially in these regions, will help not only in developing a better understanding of the evolution of LP but also in elucidating the physiological mechanisms that underlie the trait. The interpolation and mapping approach that I have applied in this study may also be of value in studying the underlying genetic basis and evolution of other phenotypic variation that impacts on human health, such as the distribution of functional variation in drug metabolising enzymes (Xie et al., 2001).

Table 4.1. The lactase persistence phenotype frequencies. Columns show location (continent, country, longitude and latitude), population group, number of individuals tested, frequency of lactase persistent individuals, LP test method, and the primary source reference. The Americas were excluded from the Table due to paucity of data. Other reasons for data exclusion were: recent immigrant populations, children (under 12 years old), or biased individuals selection criteria (such as individuals reported being lactase non persistent or related individuals). Wherever only country name was available, location was determined by the capital city or the estimated central point of the country.

COUNTRY	POPULATION	LONGITUDE	LATITUDE	N	FREQUENCY OF DIGESTORS	TESTING METHOD (BG / BH / UG / BIOPSY)	REFERENCE
BOTSWANA	Shua	25.00	-25.00	22	0.09	BG	Nurse, G. T., & Jenkins, T. (1974) Br. Med. J. 2, 728.
EGYPT	Cairo and Giza	31.25	30.05	67	0.33	BH	Hussein, L., *et al.* (1982) Hum.Hered. 32, 94.
EGYPT	Nile Delta	32.00	31.50	291	0.27	BH	Hussein, L., *et al.* (1982) Hum.Hered. 32, 94.
EGYPT	Suez Canal Zone	32.50	31.00	16	0.31	BH	Hussein, L., *et al.* (1982) Hum.Hered. 32, 94.
EGYPT	Upper Egypt, North	28.00	31.00	111	0.15	BH	Hussein, L., *et al.* (1982) Hum.Hered. 32, 94.
EGYPT	Upper Egypt, South	28.00	28.00	85	0.40	BH	Hussein, L., *et al.* (1982) Hum.Hered. 32, 94.
ETHIOPIA	Somali	41.86	9.59	90	0.24	BH	Ingram *et al.* (2009, submitted) J Mol Evol.
GABON	Bantu	9.45	0.38	20	0.40	BH	Gendrel, D., *et al.* (1989) J.Pediatr.Gastroenterol.Nutr. 8, 545.
KENYA	Borana	38.00	1.00	7	0.71	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Burji	38.00	1.00	6	0.50	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	El Molo	38.00	1.00	6	0.67	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Gabra	38.00	1.00	8	1.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Kikuyu	38.00	1.00	2	0.50	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Konso	38.00	1.00	4	0.50	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Maasai	38.00	1.00	26	0.88	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Marakwet	38.00	1.00	5	0.60	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Nandi	38.00	1.00	2	0.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Ogiek	38.00	1.00	11	0.55	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Pokot	38.00	1.00	10	0.60	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Rendille	38.00	1.00	7	0.71	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Sabaot	38.00	1.00	4	0.75	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Samburu	38.00	1.00	9	0.89	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Sengwer	38.00	1.00	12	0.17	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Somali	38.00	1.00	1	1.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Tugen	38.00	1.00	11	0.73	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Turkana	38.00	1.00	8	0.50	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Wata	38.00	1.00	1	0.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
KENYA	Yaaku	38.00	1.00	11	0.73	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
NAMIBIA	!Kung	20.50	-19.60	40	0.03	BG	Jenkins, T., *et al.* (1974) Br.Med.J. 2, 23.
NAMIBIA	Herero	13.70	-18.30	39	0.03	BG	Currie B, *et al.* (1978). S. Afr. J. Sci. 74:227.
NIGER	Tuareg	2.12	13.52	118	0.87	BH	Flatz, G., *et al.* (1986) Am.J.Hum.Genet 38, 515.
NIGERIA	Hausa/Fulani	3.47	7.23	15	0.40	BG	Olatunbosun, D. A., *et al.* (1971) Am.J.Dig.Dis. 16, 909.
NIGERIA	Ibo	3.47	7.23	11	0.18	BG	Olatunbosun, D. A., *et al.* (1971) Am.J.Dig.Dis. 16, 909.
NIGERIA	Yoruba	3.47	7.23	48	0.17	BG	Olatunbosun, D. A., *et al.* (1971) Am.J.Dig.Dis. 16, 909.
RWANDA	Hutu-Hutu	29.74	-2.60	36	0.58	UG	Cox, J. A., *et al.* (1974) Am.J.Dig.Dis. 19, 714.
RWANDA	Hutu-Tutsi	29.74	-2.60	11	0.45	UG	Cox, J. A., *et al.* (1974) Am.J.Dig.Dis. 19, 714.
RWANDA	Shi	29.74	-2.60	28	0.04	UG	Cox, J. A., *et al.* (1974) Am.J.Dig.Dis. 19, 714.
RWANDA	Tussi-Tutsi	29.74	-2.60	27	0.93	UG	Cox, J. A., *et al.* (1974) Am.J.Dig.Dis. 19, 714.

SENEGAL	Diolas	-17.43	14.67	40	0.73	BG	Arnold, J., *et al.* (1980) C R.Seances Soc.Biol Fil. 174, 983.
SENEGAL	Peuhls	-15.12	15.40	29	1.00	BG	Arnold, J., *et al.* (1980) C R.Seances Soc.Biol Fil. 174, 983.
SENEGAL	Sereres	-17.43	14.67	38	0.71	BG	Arnold, J., *et al.* (1980) C R.Seances Soc.Biol Fil. 174, 983.
SENEGAL	Toucouleurs	-17.43	14.67	40	0.90	BG	Arnold, J., *et al.* (1980) C R.Seances Soc.Biol Fil. 174, 983.
SENEGAL	Wolof	-17.43	14.67	53	0.51	BG	Arnold, J., *et al.* (1980) C R.Seances Soc.Biol Fil. 174, 983.
SOUTH AFRICA	Shangaan	28.08	-26.20	7	0.14	BH	Segal, I., *et al.* (1983) Am.J.Clin.Nutr. 38, 901.
SOUTH AFRICA	Sotho	28.08	-26.20	23	0.35	BH	Segal, I., *et al.* (1983) Am.J.Clin.Nutr. 38, 901.
SOUTH AFRICA	Swazi	28.08	-26.20	12	0.25	BH	Segal, I., *et al.* (1983) Am.J.Clin.Nutr. 38, 901.
SOUTH AFRICA	Tswana	28.08	-26.20	24	0.17	BH	Segal, I., *et al.* (1983) Am.J.Clin.Nutr. 38, 901.
SOUTH AFRICA	Xhosa	28.08	-26.20	17	0.18	BH	Segal, I., *et al.* (1983) Am.J.Clin.Nutr. 38, 901.
SOUTH AFRICA	Zulu	31.02	-29.85	47	0.11	BG	O'Keefe, S. J.& Adam, J. (1983) S.Afr.Med.J. 63, 778.
SOUTH AFRICA	Zulu	28.08	-26.20	32	0.19	BH	Segal, I., *et al.* (1983) Am.J.Clin.Nutr. 38, 901.
SUDAN	Ama	30.00	14.00	2	0.50	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Amarar	37.22	19.62	82	0.87	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Artega	37.22	19.62	22	0.82	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Bedja	30.95	18.05	9	0.89	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Beja Banuamir	30.00	14.00	6	1.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Beja Hadandawa	30.00	14.00	11	0.82	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Beni Amir	37.22	19.62	40	0.88	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Bisharin	37.22	19.62	22	0.86	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Dinka	33.63	7.67	208	0.25	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Dinka	30.00	14.00	7	0.86	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Dongolawi	30.95	18.05	16	0.19	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Gomoeia	30.95	18.05	31	0.68	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Habbani	30.35	13.08	19	0.47	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Haddendoa	37.22	19.62	137	0.80	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Jaali	32.53	15.59	113	0.53	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Jaali	33.43	16.69	94	0.48	BH	Ingram, C. J., *et al.* (2007) Hum Genet 120, 779.
SUDAN	Kahli	30.95	18.05	21	0.62	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Koalib	30.00	14.00	1	1.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Liguri/Logorik	30.00	14.00	1	0.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Masalit	30.00	14.00	1	1.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Misseri	30.35	13.08	20	0.40	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Nilotic	27.67	7.77	18	0.33	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Nuba	29.68	6.80	58	0.21	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Nubians	30.95	18.05	21	0.33	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Nuer	33.63	7.67	23	0.22	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Nuer	30.00	14.00	2	1.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
SUDAN	Shaygi	30.95	18.05	42	0.38	BH	Bayoumi, R. A., *et al.* (1981) Hum.Genet 57, 279.
SUDAN	Shilluk	33.63	7.67	8	0.38	BH	Bayoumi, R. A., *et al.* (1982) Am.J.Phys.Anthropol. 58, 173.
SUDAN	Shilook	30.00	14.00	4	0.75	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Akie	34.00	-7.00	11	0.55	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Burunge	34.00	-7.00	16	0.38	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Datog	34.00	-7.00	1	0.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Dorobo	34.00	-7.00	6	0.67	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Fiome	34.00	-7.00	7	0.14	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Hadza	34.00	-7.00	15	0.60	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Iraqw	34.00	-7.00	19	0.95	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Maasai	34.00	-7.00	15	0.67	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Mbugu	34.00	-7.00	23	0.43	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Mbugwe	34.00	-7.00	8	0.50	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Pare	34.00	-7.00	8	0.75	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Rangi	34.00	-7.00	26	0.65	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Samba'a	34.00	-7.00	2	0.00	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.
TANZANIA	Sandawe	34.00	-7.00	23	0.35	BG	Tishkoff, S. A., *et al.* (2007) Nat Genet 39, 31.

TUNISIA	Tunisian	10.18	36.80	43	0.16	BH	Filali, A., *et al.* (1987) Gastroenterol.Clin.Biol 11, 554.
UGANDA	Baganda	32.57	0.32	12	0.00	BG	Cook, G. C., & Dahlquist, A . (1968) Gastroenterology 55, 328.
UGANDA	Batutsi	32.57	0.32	5	1.00	BG	Cook, G. C., *et al.* (1968) Gastroenterology 55, 328.
UGANDA	Nilotic	32.57	0.32	9	0.56	BG	Cook, G. C., *et al.* (1966) Lancet 1, 725.
UGANDA	Ugandan Bantu	32.57	0.32	17	0.06	BG	Cook, G. C., *et al.* (1966) Lancet 1, 725.
ZAMBIA	Bantu of Zambia	28.11	15.28	26	0.00	BG	Cook, G. C., *et al.* (1973) Gastroenterology 64, 405.
CHINA	Kazakh	87.58	43.80	195	0.24	BH	Yongfa, W., *et al*. (1984) Hum.Genet. 67, 103.
CHINA	Mongols	111.65	40.81	198	0.12	BH	Yongfa, W., *et al*. (1984) Hum.Genet. 67, 103.
CHINA	Northern Han	116.39	39.93	248	0.08	BH	Yongfa, W., *et al*. (1984) Hum.Genet. 67, 103.
INDIA	Indians	72.83	18.98	100	0.36	BG	Desai, H. G., *et al.* (1970) Indian J.Med.Sci. 24, 729.
INDIA	Indians	78.47	17.38	18	0.39	BG	Reddy, V., Pershad, J. (1972) Am.J.Clin.Nutr. 25, 114.
INDIA	Indians	80.28	13.08	38	0.00	BIOPSY	Swaminathan *et al*. (1970) Clin Chim Acta 30, 707.
INDIA	Northern Indians	77.20	28.60	70	0.73	BG	Gupta, P. S., *et al.* (1971) J.Trop.Med.Hyg. 74, 225.
INDIA	Northern Indians	77.20	28.60	66	0.36	BG	Tandon, R. K., *et al*. (1981) Am.J.Clin.Nutr. 34, 943.
JAPAN	Japanese	140.47	40.59	40	0.28	BG	Yoshida, Y., *et al.* (1975) Gastroenterol.Jpn. 10, 29.
MYANMAR	Burmese	91.17	16.78	50	0.08	BG	Aung-Than-Batu *et al.* (1972), Union Burma J Life Sci, 5, 133-135
PAKISTAN	Baloochi	67.05	24.87	4	1.00	BG	Rab, S. M., *et al.* (1976) Br. Med. J. 1, 436.
PAKISTAN	Baluchistani	71.92	32.27	32	0.38	BH	Ahmad, M., Flatz,G. (1984) Hum. Hered. 34, 69.
PAKISTAN	Kashmiri	71.92	32.27	27	0.30	BH	Ahmad, M., Flatz,G. (1984) Hum. Hered. 34, 69.
PAKISTAN	Mohajir	67.05	24.87	15	0.80	BG	Rab, S. M., *et al.* (1976) Br. Med. J. 1, 436.
PAKISTAN	Pathan	67.05	24.87	15	1.00	BG	Rab, S. M., *et al.* (1976) Br. Med. J. 1, 436.
PAKISTAN	Punjabi	71.92	32.27	322	0.41	BH	Ahmad, M., Flatz,G. (1984) Hum. Hered. 34, 69.
PAKISTAN	Punjabi	67.05	24.87	9	1.00	BG	Rab, S. M., *et al.* (1976) Br. Med. J. 1, 436.
PAKISTAN	Sindhi	71.92	32.27	33	0.42	BH	Ahmad, M., Flatz,G. (1984) Hum. Hered. 34, 69.
PAKISTAN	Sindhi	67.05	24.87	12	1.00	BG	Rab, S. M., *et al.* (1976) Br. Med. J. 1, 436.
RUSSIA	Khanty (Northern)	69.02	61.04	115	0.29	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	Komi-Izhems	50.81	61.67	56	0.38	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	Mansi	69.02	61.04	81	0.28	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	Nenets (West Siberia)	80.86	59.06	9	0.22	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	Udmurtians	53.23	56.85	30	0.60	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	West-Siberian	80.86	59.06	47	0.51	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
SRI LANKA	Sri Lankan	80.64	7.30	135	0.29	BG	Thomas, S., *et al.* (1990) J Trop Pediatr 36, 80.
SRI LANKA	Sri Lankan	80.64	7.30	135	0.29	BG	Thomas, S., *et al.* (1990) J Trop Pediatr 36, 80.
SRI LANKA	Sri Lankans ("Ceylonese")	80.60	7.26	200	0.28	BG	Senewiratne, B., *et al.* (1977) Gastroenterology 72, 1257.
TAIWAN	Chinese	121.45	25.02	50	0.12	BG	Sung, J., *et al.* (1972) Asian Journal of Medicine 8, 149.
THAILAND	Thai	100.52	13.75	140	0.03	BG	Keusch, G. T., *et al.* (1969) Am.J.Clin.Nutr. 22, 638.
THAILAND	Thai	100.49	13.45	40	0.00	BG	Troncale *et al.* (1967) Br. Med. J. 4, 578.
AUSTRALIA	Aboriginal	123.97	-17.30	45	0.16	BH	Brand, J. C., *et al.* (1983) Am.J.Clin.Nutr. 37, 449.
NEW ZEALAND	Maori	174.77	-36.87	28	0.36	BH	Abbott W.G., Tasman-Jones C. (1985) N Z Med J. 10;98(776), 228.
PAPUA NEW GUINEA	Central (inc. Port Moresby)	147.19	-9.46	14	0.07	BG	Cook, G. C. (1979) Ann.Hum.Biol 6, 55.
PAPUA NEW GUINEA	E and W Sepik provinces	143.52	-4.18	35	0.23	BH	Arnhold R.G. *et al.* (1981) Ann Hum Biol 5, 481
PAPUA NEW GUINEA	Gulf & Western	147.19	-9.46	13	0.08	BG	Cook, G. C. (1979) Ann.Hum.Biol 6, 55.
PAPUA NEW GUINEA	Highlands	147.19	-9.46	13	0.00	BG	Cook, G. C. (1979) Ann.Hum.Biol 6, 55.
PAPUA NEW GUINEA	Huli, Mendi, and Dunai	142.95	-5.70	30	0.10	BG	Jenkins, T., *et al.* (1981) Ann.Hum.Biol 8, 447.
PAPUA NEW GUINEA	Milne Bay	147.19	-9.46	2	0.00	BG	Cook, G. C. (1979) Ann.Hum.Biol 6, 55.
PAPUA NEW GUINEA	Morobe & Northern	147.19	-9.46	5	0.00	BG	Cook, G. C. (1979) Ann.Hum.Biol 6, 55.
PAPUA NEW GUINEA	N. Solomons & E. New Britain	147.19	-9.46	3	0.00	BG	Cook, G. C. (1979) Ann.Hum.Biol 6, 55.
AUSTRIA	Austrian	14.00	47.75	118	0.75	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.
AUSTRIA	Austrian	14.00	47.75	57	0.79	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.
AUSTRIA	Austrian	14.00	47.75	88	0.80	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.
AUSTRIA	Austrian	14.00	47.75	32	0.81	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.
AUSTRIA	Karnten Austrian	14.31	46.62	46	0.80	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.
AUSTRIA	Oberosterreich Austrian	14.30	48.30	45	0.84	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.
AUSTRIA	Tirol Austrian	9.77	47.50	124	0.83	BH	Rosenkranz, W., *et al.* (1982) Hum.Genet 62, 158.

CYPRUS	Greek Cypriots	33.37	35.17	50	0.34	BG	Kanaghinis, T., *et al.* (1974) Am.J.Dig.Dis. 19, 1021.
DENMARK	Danes	12.58	55.73	91	0.96	BG	Busk, H. E., *et al.* (1975) Ugeskr Laeger 137, 2062-4.
ESTONIA	Estonian	26.70	58.23	112	0.75	BG	Lember, M., *et al.* (1991) Eur J Gastroenterol Hepatol 3, 479.
ESTONIA	Setus	27.64	57.96	100	0.51	UG	Lember, M., *et al.* (1991) Eur J Gastroenterol Hepatol 3, 479.
FINLAND	Finnish-speaking Finns	21.43	60.60	91	0.92	BG	Sahi, T. (1974) Scand.J.Gastroenterol. 9, 303.
FINLAND	Finns	27.68	62.90	638	0.83	BIOPSY	Jussila, J. (1969) Ann.Clin.Res. 1, 199.
FINLAND	Rural Finn	21.93	60.41	159	0.83	BG	Jussila, J., *et al.* (1970) Scand.J.Gastroenterol. 5, 49.
FINLAND	Swedish-speaking Finns	21.43	60.60	156	0.83	BG	Sahi, T. (1974) Scand.J.Gastroenterol. 9, 303.
FORMER CZECHOSLOVAKIA	Czech	17.50	49.00	17	0.82	BG	Leichter, J. (1972) Am.J.Dig.Dis. 17, 73.
FRANCE	French	5.82	44.93	102	0.76	BH	Cloarec, D., *et al.* (1991) Gastroenterol.Clin.Biol 15, 588.
FRANCE	French	6.63	49.75	85	0.71	BH	Cuddenec, Y., *et al.* (1982) Gastroenterol.Clin.Biol 6, 776.
FRANCE	Maghrebins (Northern African Muslims)	7.25	43.70	55	0.22	BG	O'Morain, C., *et al.* (1978) Acta Gastroenterol Belg 41, 56-63.
FRANCE	Northern French	6.63	49.75	76	0.78	BH	Cuddenec, Y., *et al.* (1982) Gastroenterol.Clin.Biol 6, 776.
FRANCE	Southern French	6.63	49.75	40	0.43	BH	Cuddenec, Y., *et al.* (1982) Gastroenterol.Clin.Biol 6, 776.
FRANCE	Southern French	7.25	43.70	55	0.58	BG	O'Morain, C., *et al.* (1978) Acta Gastroenterol Belg 41, 56-63.
GERMANY	Baden-Wurttemberg Germans	9.50	48.40	136	0.76	BH	Flatz, G., *et al.* (1982) Hum.Genet 62, 152.
GERMANY	Bayern Germans	12.53	47.80	221	0.86	BH	Flatz, G., *et al.* (1982) Hum.Genet 62, 152.
GERMANY	Eastern Germans	13.75	51.05	246	0.78	BH	Flatz, G., *et al.* (1982) Hum.Genet 62, 152.
GERMANY	Germans	8.52	53.18	60	0.87	BIOPSY	Howell, J. N., *et al.* (1980) Hepatogastroenterology 27, 208.
GERMANY	Northwest Germans	8.80	53.08	341	0.91	BH	Flatz, G., *et al.* (1982) Hum.Genet 62, 152.
GERMANY	Rheinland and Pfalz Germans	8.27	50.00	182	0.86	BH	Flatz, G., *et al.* (1982) Hum.Genet 62, 152.
GERMANY	Schleswig-Holstein Germans	9.55	54.52	100	0.94	BH	Flatz, G., *et al.* (1982) Hum.Genet 62, 152.
GREECE	Continental Greeks	23.73	37.98	600	0.55	BG	Kanaghinis, T., *et al.* (1974) Am.J.Dig.Dis. 19, 1021.
GREECE	Cretan Greek	25.13	35.33	50	0.44	BG	Kanaghinis, T., *et al.* (1974) Am.J.Dig.Dis. 19, 1021.
GREECE	Greek	23.73	37.98	16	0.63	BG	Spanidou, E. P., & Petrakis, NL (1972) Lancet 2, 872.
GREECE	Greeks	23.73	37.98	200	0.25	BH	Ladas, S., *et al.* (1982) Gut 23, 968.
GREECE	Greeks	23.73	37.98	250	0.77	BG	Zografos *et al.* (1973), The Lancet, 301,367.
HUNGARY	Eastern Hungarian	19.08	47.50	70	0.71	BH	Czeizel, A., *et al.* (1983) Hum.Genet 64, 398.
HUNGARY	Hungarian	19.08	47.50	262	0.59	BH	Czeizel, A., *et al.* (1983) Hum.Genet 64, 398.
HUNGARY	Matyo	20.58	47.82	172	0.63	BH	Czeizel, A., *et al.* (1983) Hum.Genet 64, 398.
HUNGARY	Northeastern Hungarian	19.08	47.50	103	0.58	BH	Czeizel, A., *et al.* (1983) Hum.Genet 64, 398.
HUNGARY	Romai	21.72	47.95	113	0.44	BH	Czeizel, A., *et al.* (1983) Hum.Genet 64, 398.
HUNGARY	Western Hungarian	19.08	47.50	100	0.72	BH	Czeizel, A., *et al.* (1983) Hum.Genet 64, 398.
IRELAND	Native Irish	-6.25	53.33	50	0.96	BG	Fielding, J. F., *et al.* (1981) Ir.J.Med.Sci. 150, 276.
ITALY	Italians	9.20	45.47	42	0.38	BH	Bozzani, A., *et al.* (1986) Dig.Dis.Sci. 31, 1313.
ITALY	Italians	9.20	45.47	89	0.48	BG	Cavalli-Sforza *et al.* (1987) Am J Clin Nutr 45, 748
ITALY	Italians	12.48	41.90	65	0.82	BG	Cavalli-Sforza *et al.* (1987) Am J Clin Nutr 45, 748
ITALY	Italians	14.25	40.83	51	0.59	BG	Cavalli-Sforza *et al.* (1987) Am J Clin Nutr 45, 748
ITALY	Italians	14.25	40.83	44	0.23	BIOPSY	Rossi *et al.*, (1997), Gastroenterology. 112(5), 1506.
ITALY	Italians	9.20	45.47	20	0.25	BH	Zuccato, E., *et al.* (1983) Eur J Clin Invest 13, 261.
ITALY	Napolitans	14.25	40.83	99	0.46	BH	Rinaldi, E., *et al.* (1984) Lancet 1, 355-7.
ITALY	Neapolitan	14.25	40.83	9	0.00	BG	De Ritis, F., *et al.* (1970) Enzymol.Biol Clin.(Basel) 11, 263.
ITALY	Northern Italians	7.67	45.05	208	0.49	BH	Burgio, G. R., *et al.* (1984) Am.J.Clin.Nutr. 39, 100.
ITALY	Sardinians	8.56	40.73	50	0.14	BH	Meloni, G. F., *et al.* (2001)Am.J.Clin.Nutr. 73, 582.
ITALY	Sardinians	9.00	39.40	47	0.15	BH	Meloni, T., *et al.*, (1998) Ital J Gastroenterol Hepatol 30, 490.
ITALY	Sardinians	9.00	40.10	53	0.11	BH	Meloni, T., *et al.*, (1998) Ital J Gastroenterol Hepatol 30, 490.
ITALY	Sardinians	9.00	40.30	38	0.18	BH	Meloni, T., *et al.*, (1998) Ital J Gastroenterol Hepatol 30, 490.
ITALY	Sicilians	13.37	38.12	100	0.29	BH	Burgio, G. R., *et al.* (1984) Am.J.Clin.Nutr. 39, 100.
POLAND	Eastern Polish	23.13	52.03	35	0.63	BH	Socha, J., *et al.* (1984) Ann.Hum.Biol 11, 311.
POLAND	Northeastern Polish	22.35	53.83	34	0.59	BH	Socha, J., *et al.* (1984) Ann.Hum.Biol 11, 311.
POLAND	Polish	21.00	52.25	21	0.71	BG	Leichter, J. (1972) Am.J.Dig.Dis. 17, 73.
POLAND	Polish	19.00	51.73	29	0.62	BH	Socha, J., *et al.* (1984) Ann.Hum.Biol 11, 311.
POLAND	Polish	19.37	52.23	92	0.63	BH	Socha, J., *et al.* (1984) Ann.Hum.Biol 11, 311.
POLAND	Polish	19.37	52.23	85	0.64	BH	Socha, J., *et al.* (1984) Ann.Hum.Biol 11, 311.

RUSSIA	Kildin Saami	32.00	68.00	50	0.52	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	Komi-Permiaks	32.00	68.00	112	0.50	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
RUSSIA	Udmurtians	32.00	68.00	75	0.41	BG	Kozlov, A. I. (1998) Int J Circumpolar Health 57, 18.
SPAIN	Galician	-8.55	42.88	338	0.66	BH	Leis, R., *et al*. (1997) J.Pediatr.Gastroenterol.Nutr. 25, 296.
UK	British	-3.20	55.95	150	0.95	BG	Ferguson, A., *et al.* (1984) Gut 25, 163.
UK	British natives	-1.25	51.75	75	0.95	BIOPSY	Ho, M. W., *et al.* (1982) Am.J.Hum.Genet 34, 650.
UK	White British	-1.92	52.47	67	0.97	BIOPSY	Iqbal, T. H., *et al*. (1993) Br. Med. J. 306, 1303.
AFGHANISTAN	Hazara	69.18	34.52	10	0.20	BG	Rahimi, A. G., *et al*. (1976) Hum.Genet. 34, 57.
AFGHANISTAN	Mixed urban	69.18	34.52	34	0.24	BG	Rahimi, A. G., *et al*. (1976) Hum.Genet. 34, 57.
AFGHANISTAN	Pasha-I	71.00	36.00	60	0.13	BG	Rahimi, A. G., *et al*. (1976) Hum.Genet. 34, 57.
AFGHANISTAN	Pashtun	69.18	34.52	71	0.21	BG	Rahimi, A. G., *et al*. (1976) Hum.Genet. 34, 57.
AFGHANISTAN	Tajik	69.18	34.52	79	0.18	BG	Rahimi, A. G., *et al*. (1976) Hum.Genet. 34, 57.
AFGHANISTAN	Uzbek	69.18	34.52	16	0.00	BG	Rahimi, A. G., *et al*. (1976) Hum.Genet. 34, 57.
IRAN	Iranian	51.42	35.67	21	0.14	BG	Sadre, M., *et al*. (1979) Am.J.Clin.Nutr. 32, 1948.
ISRAEL	Arabs	34.95	32.23	67	0.19	BG	Gilat, T *et al.* (1971) Digestive Diseases 16, 203
JORDAN	Jordanian Arabs	35.93	31.95	148	0.25	BH	Hijazi, S. S., *et al.* (1983) Trop.Geogr.Med. 35, 157.
JORDAN	Mediterranean origin Jordanian Arabs	35.93	31.95	56	0.23	BG	Snook, C. R., *et al*. (1976) Trop.Geogr.Med. 28, 333.
JORDAN	Urban/agricultural Jordanian Arabs	35.93	31.95	162	0.76	BH	Hijazi, S. S., *et al.* (1983) Trop.Geogr.Med. 35, 157.
KUWAIT	Arab Kuwaiti	47.98	29.37	70	0.53	BH	Sanae, H. A., *et al.* (2003) Med. Princ. Pract. 12, 160.
KUWAIT	Asian Kuwaiti	47.98	29.37	79	0.42	BH	Sanae, H. A., *et al.* (2003) Med. Princ. Pract. 12, 160.
LEBANON	Lebanese	35.51	33.87	74	0.22	BG	Nasrallah, S. M. (1979) Am.J.Clin.Nutr. 32, 1994.
PAKISTAN	Punjabi	73.07	33.60	53	0.55	BG	Abbas H., Ahmad M. (1983) Hum. Genet. 64:277.
SAUDI ARABIA	Arabs	50.11	26.43	109	0.43	BH	Dissanayake, A.S. *et al.*, (1990) Annals of Saudi Medicine, 10, 598.
SAUDI ARABIA	Bedouin	50.11	26.43	21	0.81	BH	Dissanayake, A.S. *et al.*, (1990) Annals of Saudi Medicine, 10, 598.
SAUDI ARABIA	Beduin and Urban Saudi	46.77	24.64	14	0.86	BG	Cook & Al Torki (1975) Br. Med. J. 3,135.
SAUDI ARABIA	Yemenites	50.11	26.43	17	0.53	BH	Dissanayake, A.S. *et al.*, (1990) Annals of Saudi Medicine, 10, 598.
TURKEY	Central Anatolia	39.50	34.00	104	0.29	BH	Flatz, G., *et al.* (1986) Am.J.Hum.Genet 38, 515.
TURKEY	Eastern Anatolia	39.50	40.00	122	0.26	BH	Flatz, G., *et al.* (1986) Am.J.Hum.Genet 38, 515.
TURKEY	North Coast of Turkey	34.00	41.50	64	0.31	BH	Flatz, G., *et al.* (1986) Am.J.Hum.Genet 38, 515.
TURKEY	South Coast of Turkey	33.00	36.50	54	0.28	BH	Flatz, G., *et al.* (1986) Am.J.Hum.Genet 38, 515.
TURKEY	Turks	32.86	39.93	30	0.63	BG	Tuncbilek, *et al*. (1973), The Lancet July 21, 151
TURKEY	Western Anatolia and European Turkey	28.96	41.02	126	0.30	BH	Flatz, G., *et al.* (1986) Am.J.Hum.Genet 38, 515.

Table 4.2. The lactase persistence associated allele frequencies. Columns show location (continent, country, longitude and latitude), population group, number of individuals tested, frequency of -13910*T, -13,907*G, -13,915*G and -14,010*C LP-associated alleles, and the primary literature and own data source. Data taken from SNP typing tests (where only -13,910*T is shown) or from resequencing. The Americas were excluded from the Table due to paucity of data. The predicted lactase persistence frequency was calculated by assuming Hardy-Weinberg equilibrium and dominance using the sum of the all available LP-associated alleles at a specific location. Wherever only country name was available, location was determined by the capital city or the estimated central point of the country. The "sum" column is the result of adding together the 4 LP-associated alleles. It should be noted that the collection location for the Indian and North Indian genotype data was Singapore. As an exception, I placed these data in the location of the ancestral population because of lack of genetic data from India.

COUNTRY	POPULATION	LONG	LAT	N	-14010 G>C	-13915 T>G	-13907 C>G	-13910 C>T	SUM	PREDICTED LP FREQUENCY	REFERENCE
Algeria	Berber Mzab	3.05	36.76	66	0.00	0.00	0.00	0.17	0.17	0.31	Myles *et al.* (2005) Hum Genet. 117, 34.
Algeria	Mozabite	3.05	36.76	60	-	-	-	0.22	0.22	0.39	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Cameroon	Cameroonian	12.50	6.00	130	0.00	0.00	0.00	0.00	0.00	0.00	Jones *et al.* (2009, unpublished)
Cameroon	Cameroonian	14.50	13.00	108	0.00	0.08	0.00	0.02	0.10	0.19	Jones *et al.* (2009, unpublished)
Cameroon	Fulani	14.00	11.00	110	0.00	0.00	0.00	0.39	0.39	0.63	Ingram *et al.* (2009, submitted) J Mol Evol.
Cameroon	Fulani	11.55	6.47	98	-	-	-	0.11	0.11	0.21	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Cameroon	Hausa	11.55	6.47	36	-	-	-	0.14	0.14	0.26	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Cameroon	Mambila	13.00	9.00	74	0.00	0.00	0.00	0.00	0.00	0.00	Ingram *et al.* (2009, submitted) J Mol Evol.
Cameroon	Mambila	11.28	6.45	244	-	-	-	0.00	0.00	0.01	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Cameroon	Nso	10.67	6.20	252	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Cameroon	Shuwa Arab	14.00	13.00	30	0.00	0.13	0.00	0.00	0.13	0.25	Ingram *et al.* (2009, submitted) J Mol Evol.
Cameroon	Yamba	11.55	6.47	42	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Congo	Congolese	15.28	-4.26	90	0.00	0.00	0.00	0.00	0.00	0.00	Jones *et al.* (2009, unpublished)
Ethiopia	Afar	41.44	11.56	74	0.00	0.12	0.30	0.01	0.43	0.68	Ingram *et al.* (2009, submitted) J Mol Evol.
Ethiopia	Amharic	38.70	9.03	38	0.00	0.13	0.05	0.00	0.19	0.34	Ingram *et al.* (2009, submitted) J Mol Evol.
Ethiopia	Ethiopian	34.50	7.58	120	0.00	0.00	0.00	0.00	0.00	0.00	Jones *et al.* (2009, unpublished)
Ethiopia	Ethiopian	36.65	5.65	132	0.00	0.05	0.02	0.00	0.07	0.14	Jones *et al.* (2009, unpublished)
Ethiopia	Ethiopian	36.83	7.67	146	0.01	0.08	0.07	0.00	0.16	0.29	Jones *et al.* (2009, unpublished)
Ethiopia	Ethiopian	38.70	9.03	130	0.00	0.02	0.06	0.00	0.10	0.19	Jones *et al.* (2009, unpublished)
Ethiopia	Ethiopian	41.44	11.56	148	0.01	0.19	0.25	0.01	0.46	0.71	Jones *et al.* (2009, unpublished)
Ethiopia	Phenotyped Somali	41.87	9.58	218	0.01	0.05	0.06	0.02	0.13	0.24	Ingram *et al.* (2009, submitted) J Mol Evol.
Ethiopia	Somali	42.80	9.35	74	0.03	0.04	0.10	0.00	0.16	0.30	Ingram *et al.* (2009, submitted) J Mol Evol.
Ethiopian	Nuer	34.58	8.25	238	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Ghana	Ghanaian	-1.00	7.00	114	0.00	0.00	0.00	0.00	0.00	0.00	Jones *et al.* (2009, unpublished)
Kenya	Borana	38.00	1.00	16	0.13	0.19	0.13	0.00	0.44	0.68	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Burji	38.00	1.00	16	0.06	0.00	0.00	0.00	0.06	0.12	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	El Molo	38.00	1.00	18	0.11	0.00	0.00	0.00	0.11	0.21	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Gabra	38.00	1.00	18	0.00	0.28	0.11	0.00	0.39	0.63	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Kikuyu	38.00	1.00	4	0.75	0.00	0.00	0.00	0.75	0.94	Tishkoff *et al.* (2007) Nat Genet. 39, 31

Kenya	Konso	38.00	1.00	12	0.08	0.08	0.00	0.00	0.17	0.30	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Maasai	38.00	1.00	64	0.58	0.00	0.03	0.00	0.61	0.85	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Marakwet	38.00	1.00	14	0.36	0.07	0.00	0.00	0.43	0.67	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Nandi	38.00	1.00	8	0.25	0.00	0.00	0.00	0.25	0.44	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Ogiek	38.00	1.00	22	0.36	0.00	0.00	0.00	0.36	0.60	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Pokot	38.00	1.00	28	0.29	0.04	0.00	0.00	0.32	0.54	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Rendille	38.00	1.00	16	0.13	0.13	0.06	0.00	0.31	0.53	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Sabaot	38.00	1.00	12	0.17	0.00	0.00	0.00	0.17	0.31	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Samburu	38.00	1.00	18	0.28	0.06	0.06	0.00	0.40	0.64	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Sengwer	38.00	1.00	32	0.06	0.00	0.00	0.00	0.06	0.12	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Somali	38.00	1.00	2	0.00	0.50	0.00	0.00	0.50	0.75	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Tugen	38.00	1.00	32	0.19	0.00	0.00	0.00	0.19	0.34	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Turkana	38.00	1.00	26	0.21	0.00	0.00	0.00	0.21	0.37	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Wata	38.00	1.00	2	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Kenya	Yaaku	38.00	1.00	28	0.54	0.00	0.04	0.00	0.58	0.82	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Malawi	Bantu	33.78	-13.98	310	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Malawi	Malawian	33.50	-13.00	100	0.00	0.00	0.00	0.00	0.00	0.00	Jones *et al.* (2009, unpublished)
Morocco	Arabs	-6.84	34.03	180	-	-	-	0.18	0.18	0.33	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Morocco	Berber	-3.77	34.05	154	-	-	-	0.14	0.14	0.25	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Morocco	Moroccan	-6.84	34.03	24	0.00	0.08	0.00	0.21	0.29	0.50	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Morocco	Saharawi	-6.84	34.03	114	-	-	-	0.26	0.26	0.45	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Morocco (High-Atlas)	Amizmiz	-4.00	34.00	78	0.00	0.00	0.00	0.14	0.14	0.26	Myles *et al.* (2005) Hum Genet. 117, 34.
Morocco (Mid-Atlas)	Berber Moyen-Atlas	-6.00	32.00	66	0.00	0.00	0.00	0.16	0.16	0.29	Myles *et al.* (2005) Hum Genet. 117, 34.
Mozambique	Mozambicans	36.50	-18.00	102	0.00	0.00	0.00	0.00	0.00	0.00	Jones *et al.* (2009, unpublished)
N.E. Kenya	Bantu	36.00	-1.00	24	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Namibia	San	17.08	-22.57	14	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Namibia	San	17.08	-22.57	30	-	-	-	0.00	0.00	0.00	Mulcare (2006) London: University of London PhD.
Nigeria	Yoruba	3.47	7.23	50	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Senegal	Manjak	-16.00	14.00	186	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Senegal	Wolof	-17.38	14.67	118	0.00	0.00	0.00	0.00	0.00	0.00	Ingram *et al.* (2009, submitted) J Mol Evol.
Senegal	Wolof	-17.00	14.00	20	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Somalia	Somali	45.37	2.07	158	-	-	-	0.03	0.03	0.06	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
South Africa	Bantu	28.08	-26.20	16	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
South Africa	Bantu	28.08	-26.20	50	-	-	-	0.00	0.00	0.00	Mulcare (2006) London: University of London PhD.
Sudan	Ama	30.00	20.00	4	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Beja (Banuamir)	30.00	20.00	12	0.00	0.17	0.25	0.00	0.42	0.66	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Beja (Hadandawa)	30.00	20.00	22	0.00	0.09	0.18	0.00	0.27	0.47	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Beni Amer	37.22	19.62	38	-	-	-	0.05	0.05	0.10	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) London: University of London PhD.
Sudan	Beni Amer	37.22	19.62	162	0.00	0.25	0.01	0.01	0.26	0.45	Ingram *et al.* (2009, submitted) J Mol Evol.
Sudan	Dinka	30.00	20.00	18	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Dunglawi	30.00	20.00	12	0.00	0.00	0.08	0.00	0.08	0.16	Ingram *et al.* (2009, submitted) J Mol Evol.
Sudan	Fulani	30.00	20.00	88	-	-	-	0.48	0.48	0.73	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Sudan	Gaali	32.53	15.59	20	0.00	0.00	0.05	0.00	0.05	0.10	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Sudan	Jaali	33.43	16.69	172	0.00	0.13	0.01	0.01	0.15	0.27	Ingram *et al.* (2009, submitted) J Mol Evol.
Sudan	Koalib	30.00	20.00	2	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Liguri/Logorik	30.00	20.00	2	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Mahas	32.53	15.59	30	0.00	0.17	0.00	0.00	0.17	0.31	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.

Sudan	Masalit	30.00	20.00	2	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Nuer	30.00	20.00	10	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Shaigi	30.00	20.00	18	0.00	0.06	0.00	0.00	0.06	0.11	Ingram *et al.* (2009, submitted) J Mol Evol.
Sudan	Shilook	30.00	20.00	16	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Sudan	Sudanese	30.00	15.00	60	0.00	0.07	0.00	0.02	0.09	0.17	Jones *et al.* (2009, unpublished)
Sudanese	Dinka	31.00	4.00	68	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Sudanese	Ga'ali	30.00	20.00	60	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Sudanese	Nuer	31.00	4.00	26	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Sudanese	Shaigi	30.00	20.00	22	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Tanzania	Akie	35.00	-5.00	28	0.25	0.00	0.00	0.00	0.25	0.44	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Burunge	35.00	-5.00	36	0.38	0.00	0.00	0.00	0.38	0.62	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Datog	35.00	-5.00	8	0.63	0.00	0.00	0.00	0.63	0.86	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Dorobo	35.00	-5.00	20	0.40	0.00	0.00	0.00	0.40	0.64	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Fiome	35.00	-5.00	24	0.55	0.00	0.00	0.00	0.55	0.80	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Hadza	35.00	-5.00	36	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Iraqw	35.00	-5.00	78	0.58	0.00	0.00	0.00	0.58	0.82	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Maasai	35.00	-5.00	38	0.45	0.00	0.00	0.00	0.45	0.69	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Mbugu	35.00	-5.00	60	0.31	0.00	0.00	0.00	0.31	0.52	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Mbugwe	35.00	-5.00	26	0.27	0.04	0.00	0.00	0.31	0.52	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Pare	35.00	-5.00	20	0.10	0.00	0.00	0.00	0.10	0.19	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Rangi	35.00	-5.00	70	0.27	0.00	0.00	0.00	0.27	0.47	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Samba'a	35.00	-5.00	6	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Sandawe	35.00	-5.00	62	0.13	0.00	0.00	0.00	0.13	0.25	Tishkoff *et al.* (2007) Nat Genet. 39, 31
Tanzania	Tanzanian	38.05	-5.38	92	0.14	0.00	0.00	0.00	0.14	0.26	Jones *et al.* (2009, unpublished)
Uganda	Bantu	32.98	0.43	44	-	-	-	0.00	0.00	0.00	Mulcare *et al.* (2004) Am J Hum Genet. 74, 1102.
Uganda	Ugandan	32.57	0.32	76	0.03	0.00	0.00	0.00	0.03	0.06	Jones *et al.* (2009, unpublished)
Afghanistan	Pashtu (Pushtu)	72.00	35.00	16	-	-	-	0.13	0.13	0.23	Mulcare (2006) London: University of London PhD.
Afghanistan	Tadjik	68.71	36.13	98	-	-	-	0.10	0.10	0.19	Mulcare (2006) London: University of London PhD.
Afghanistan	Uzbek	67.64	35.50	76	-	-	-	0.08	0.08	0.15	Mulcare (2006) London: University of London PhD.
Algeria	Algerian	-1.32	34.88	21	-	-	-	0.33	0.33	0.56	Mulcare (2006) London: University of London PhD.
Armenia	Armenian	44.51	40.18	88	-	-	-	0.01	0.01	0.02	Mulcare (2006) London: University of London PhD.
Azerbaijan	Azerbaijani	49.88	40.40	44	-	-	-	0.02	0.02	0.04	Mulcare (2006) London: University of London PhD.
Cambodia	Cambodian	104.92	11.55	22	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Dai	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Daur	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Han	109.00	19.00	90	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Han	100.00	35.00	200	-	-	-	0.00	0.00	0.00	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
China	Hezhen	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Lahu	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Miaozu	107.00	26.00	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Mongola	116.39	39.93	20	-	-	-	0.10	0.10	0.19	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Naxi	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Oroqen	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.

China	She	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Tu	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Tujia	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Uygur	116.39	39.93	20	-	-	-	0.05	0.05	0.10	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Xibo	116.39	39.93	18	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
China	Yizu	116.39	39.93	20	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
India	Indian	80.28	13.08	68	-	-	-	0.13	0.13	0.25	Mulcare (2006) London: University of London PhD.
Iran	Iranians	51.42	35.67	42	0.00	0.00	0.00	0.10	0.10	0.19	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Mongolia	Mongolian	106.92	47.92	102	0.00	0.00	0.00	0.04	0.04	0.08	Jones *et al.* (2009, unpublished)
North India	Indian	77.20	28.60	128	-	-	-	0.19	0.19	0.34	Mulcare (2006) London: University of London PhD.
Russia	Erzas	45.11	54.11	60	-	-	-	0.27	0.27	0.47	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Russia	Mokshas	45.11	54.11	60	-	-	-	0.28	0.28	0.48	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Russia	Ob-Ugric	80.00	60.00	40	-	-	-	0.03	0.03	0.06	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Russia	Russian	90.00	60.00	76	0.00	0.00	0.00	0.07	0.07	0.14	Jones *et al.* (2009, unpublished)
Russia	Udmurts	80.00	60.00	60	-	-	-	0.33	0.33	0.55	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Russia (Komi republic)	Komi	50.49	61.40	20	-	-	-	0.15	0.15	0.28	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Siberia	Yakut	125.00	65.00	50	-	-	-	0.06	0.06	0.12	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
South Korea	South Korean	127.00	35.57	46	-	-	-	0.00	0.00	0.00	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Syria, Iraq, Lebanon, West Bank	Arabs	40.00	33.00	40	0.00	0.11	0.00	0.13	0.24	0.41	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Papua New Guinea	Papuan	147.19	-9.46	34	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Solomon Islands	Melanesian (NAN)	159.95	-9.43	44	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Czechoslovakia	Roma	14.47	50.08	162	-	-	-	0.10	0.10	0.19	Mulcare (2006) London: University of London PhD.
Finland	Finns	28.00	65.00	1876	0.00	0.00	0.00	0.58	0.58	0.82	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Finland	Saami	29.00	69.00	60	-	-	-	0.17	0.17	0.31	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Finland and Sweden	Scandinavians	18.05	59.33	360	-	-	-	0.82	0.82	0.97	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Finns	eastern	29.00	65.00	77	-	-	-	0.55	0.55	0.80	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Finns	western	26.00	65.00	308	-	-	-	0.62	0.62	0.86	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
France	Basques	0.00	43.50	170	-	-	-	0.66	0.66	0.88	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
France	French	2.33	48.87	58	-	-	-	0.43	0.43	0.68	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
France	French	2.33	48.87	34	-	-	-	0.34	0.34	0.56	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
France	French Basque	-1.00	43.00	48	-	-	-	0.67	0.67	0.89	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Germany	German	10.00	53.55	60	-	-	-	0.56	0.56	0.80	Mulcare (2006) London: University of London PhD.
Greeks	Greece	23.73	37.98	82	-	-	-	0.13	0.13	0.25	Mulcare (2006) London: University of London PhD.
Italy	North Italian	9.72	45.68	28	-	-	-	0.36	0.36	0.59	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Italy	S. European	12.48	41.90	66	0.00	0.00	0.00	0.09	0.09	0.17	Ingram *et al.* (2009, submitted) J Mol Evol.
Italy	Sardinian	9.12	39.22	56	-	-	-	0.07	0.07	0.14	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Italy	South Italians	16.25	39.30	200	-	-	-	0.05	0.05	0.10	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Italy	Tuscan	11.25	43.77	16	-	-	-	0.06	0.06	0.12	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Mixed	N. European	15.00	54.00	110	0.00	0.00	0.00	0.62	0.62	0.85	Ingram *et al.* (2009, submitted) J Mol Evol.
Poland	Ashkenazi	21.00	52.25	96	-	-	-	0.08	0.08	0.16	Mulcare (2006) London: University of London PhD.

Russia	Russian	37.62	55.75	50	-	-	-	0.24	0.24	0.42	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Russian (Caucasus)	Adygei	42.06	44.22	34	-	-	-	0.12	0.12	0.22	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Sweden	Swedish	17.98	59.23	784	-	-	-	0.74	0.74	0.93	Almon *et al. (*2007) Scand J Gastroenterol. 42, 165.
UK	English, London	-0.12	51.50	64	-	-	-	0.73	0.73	0.93	Mulcare (2006) London: University of London PhD.
UK	Northern Ireland	-7.63	54.37	65	-	-	-	0.95	0.95	1.00	Mulcare (2006) London: University of London PhD.
UK	Orcadian (Orkney Islands)	-3.30	58.95	32	-	-	-	0.69	0.69	0.90	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Russia	Druss	47.12	42.83	34	-	-	-	0.12	0.12	0.23	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Russia	mixed	47.12	42.83	46	-	-	-	0.13	0.13	0.24	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Russia	Nog	47.12	42.83	40	-	-	-	0.07	0.07	0.14	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Iran	Iranian	52.00	36.00	90	-	-	-	0.04	0.04	0.09	Mulcare (2006) London: University of London PhD.
Iran	Qashqai	51.42	35.67	20	-	-	-	0.05	0.05	0.10	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Israel	Bedouin	34.77	32.07	38	0.00	0.13	0.00	0.03	0.16	0.29	Ingram *et al.* (2009, submitted) J Mol Evol.
Israel	Bedouin	34.00	31.00	98	-	-	-	0.03	0.03	0.06	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Israel	Druze	35.00	33.00	96	-	-	-	0.02	0.02	0.04	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Israel	Druze	34.77	32.07	28	0.00	0.11	0.00	0.04	0.14	0.27	Ingram *et al.* (2009, submitted) J Mol Evol.
Israel	Non-Bedouin Arabs	34.77	32.07	160	0.00	0.05	0.00	0.00	0.05	0.10	Ingram *et al.* (2009, submitted) J Mol Evol.
Israel	Palestinian Arabs	35.13	31.47	102	-	-	-	0.04	0.04	0.08	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Israel/PAA	Palestinian Arabs	35.20	31.90	36	0.00	0.00	0.00	0.00	0.00	0.00	Ingram *et al.* (2009, submitted) J Mol Evol.
Japan	Japanese	139.75	35.69	62	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Jordan	Jordanian	35.93	31.95	112	0.00	0.05	0.00	0.05	0.11	0.20	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Jordan	Jordanian Bedouin	35.93	31.95	52	-	-	-	0.00	0.00	0.00	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) London: University of London PhD.
Jordan	Jordanian Bedouin	35.93	31.95	46	0.00	0.35	0.00	0.00	0.35	0.57	Ingram *et al.* (2009, submitted) J Mol Evol.
Kuwait	Kuwaiti	47.98	29.37	28	-	-	-	0.00	0.00	0.00	Mulcare (2006) London: University of London PhD.
Pakistan	Balochi	73.04	33.43	50	-	-	-	0.36	0.36	0.59	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Pakistan	Balti	68.00	30.00	46	-	-	-	0.00	0.00	0.00	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Baluch	68.00	30.00	38	-	-	-	0.34	0.34	0.56	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Brahui	73.04	33.43	50	-	-	-	0.34	0.34	0.56	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Pakistan	Brahui	68.00	30.00	60	-	-	-	0.27	0.27	0.47	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Burusho	72.00	37.00	50	-	-	-	0.10	0.10	0.19	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Pakistan	Burusho	68.00	30.00	60	-	-	-	0.02	0.02	0.04	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Hazara	66.00	35.00	50	-	-	-	0.08	0.08	0.15	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Pakistan	Hazara	68.00	30.00	28	-	-	-	0.04	0.04	0.08	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Kalash	73.04	33.43	50	-	-	-	0.00	0.00	0.00	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Pakistan	Kalash	68.00	30.00	60	-	-	-	0.00	0.00	0.00	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Kashmiri	68.00	30.00	40	-	-	-	0.12	0.12	0.23	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Makrani Baluch	68.00	30.00	58	-	-	-	0.17	0.17	0.31	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Mohannes	68.00	30.00	58	-	-	-	0.28	0.28	0.48	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Parsi	68.00	30.00	58	-	-	-	0.14	0.14	0.26	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Pathan	72.00	35.00	50	-	-	-	0.30	0.30	0.51	Bersaglieri *et al.* (2004) Am J Hum Genet. 74, 1111.
Pakistan	Pathan	68.00	30.00	56	-	-	-	0.30	0.30	0.51	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Pakistan	Sindhi	68.00	25.00	50	-	-	-	0.32	0.32	0.54	Bersaglieri *et al.* (2004) Am J Hum Genet. 74,

											1111.
Pakistan	Sindi	68.00	30.00	56	-	-	-	0.41	0.41	0.65	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Saudi Arabia	Bedouin	45.00	23.00	94	0.00	0.48	0.00	0.00	0.48	0.73	Ingram *et al.* (2009, submitted) J Mol Evol.
Saudi Arabia	Central	45.00	23.00	180	0.00	0.61	0.00	0.00	0.61	0.84	Imtiaz *et al.* (2007) J Med Genet. 44, e89.
Saudi Arabia	Eastern	52.00	21.00	164	0.00	0.62	0.00	0.00	0.62	0.85	Imtiaz *et al.* (2007) J Med Genet. 44, e89.
Saudi Arabia	Northern	40.00	30.00	164	0.00	0.52	0.00	0.01	0.53	0.78	Imtiaz *et al.* (2007) J Med Genet. 44, e89.
Saudi Arabia	Southern	45.00	18.00	184	0.00	0.58	0.00	0.00	0.58	0.82	Imtiaz *et al.* (2007) J Med Genet. 44, e89.
Saudi Arabia	Western	40.00	21.00	172	0.00	0.65	0.00	0.01	0.65	0.88	Imtiaz *et al.* (2007) J Med Genet. 44, e89.
Saudi Arabia	Arabs	45.00	23.00	248	0.00	0.57	0.01	0.00	0.58	0.83	Enattah *et al.* (2008) Am J Hum Genet. 82, 57.
Syria	Assyrians	36.30	33.50	80	-	-	-	0.04	0.04	0.07	Mulcare (2006) London: University of London PhD.
Turkey	Anatolian Turks	30.00	38.00	98	-	-	-	0.03	0.03	0.06	Mulcare (2006) London: University of London PhD.
Ukraine	Ukraine	36.00	48.00	92	-	-	-	0.22	0.22	0.39	Mulcare (2006) London: University of London PhD.
Uzbekistan	Uzbekistani	64.43	39.77	36	-	-	-	0.00	0.00	0.00	Mulcare (2006) London: University of London PhD.
West Bank	Palestinian Arabs	35.00	32.00	34	-	-	-	0.03	0.03	0.06	Mulcare (2006) London: University of London PhD.

5. Simulating the Origins and Evolution of Lactase Persistence in Europe.

This chapter is based on the following published article: Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The Origins of Lactase Persistence in Europe. PLoS Comput Biol 5(8): e1000491. doi:10.1371/journal.pcbi.1000491

The content of this chapter will resemble in many parts the original article, with some changes: I will integrate relevant parts of the original article's supplementary information into the main body of this chapter and further elaborate on some relevant subjects that were only briefly mentioned (or not at all) in the original article, such as the spread of farming and animal domestication. It is important to note that the original article (and consequently this chapter) is a result of a collaborative study. Mark Beaumont developed and supervised the use of the Approximate Bayesian Computation (ABC) method that is applied in this study, Adam Powell analysed the results with the ABC method, and Joachim Burger contributed his archaeology and ancient DNA expertise.

5.1. Introduction.

Lactase persistence (LP) is an autosomal dominant trait enabling the continued production of the enzyme lactase throughout adult life. Lactase non-persistence is the ancestral condition for humans, and indeed for all mammals (Swallow, 2003). Production of lactase in the gut is essential for the digestion of the milk sugar lactose. LP is common in northern and western Europeans as well as in many African, Middle Eastern and southern Asian pastoralist groups, but is rare or absent elsewhere in the world (Ingram et al., 2007, Swallow, 2003, Mulcare et al., 2004, Tishkoff et al., 2007). In Europeans LP is strongly associated with a single C to T transition in the *MCM6* gene (*-13,910*T*), located 13.91kb upstream from the lactase gene (Enattah et al., 2002). Furthermore, *in vitro* studies have indicated that the *-13,910*T* allele can directly affect *LCT* gene promoter activity (Lewinsky et al., 2005). The *-13,910*T* allele ranges frequency from 6%-36% in eastern and southern Europe, 56%-67% in Central and western Europe, to 73%-95% in the British Isles and Scandinavia (Mulcare, 2006a,

Bersaglieri et al., 2004) while LP ranges in frequency from 15%-54% in eastern and southern Europe, 62%-86% in Central and western Europe, to 89%-96% in the British Isles and Scandinavia (Ingram et al., 2009a). This makes the *-13,910*T* allele a good candidate for predicting LP in Europe. However, genotype/phenotype frequency comparisons have shown that the *-13,910*T* allele cannot account for LP frequencies in most African (Mulcare et al., 2004) and Middle Eastern populations (Enattah et al., 2008). Instead, different LP-associated alleles occurring in the same genomic region have been reported, indicating convergent evolution (Tishkoff et al., 2007, Ingram et al., 2007, Enattah et al., 2008, Enattah et al., 2007). In chapter 4 I explore all four known LP-associated alleles and their worldwide distribution and correlation with the LP phenotype.

Using long-range haplotype conservation (Bersaglieri et al., 2004) and variation in closely linked microsatellites (Coelho et al., 2005) as proxies for allelic age, the *-13,910*T* variant has been estimated to be between 2,188 and 20,650 years old and between 7,450 and 12,300 years old, respectively. These recent age estimates, when considered in conjunction with modern allele frequencies, indicate that *-13,910*T* has been subjected to very strong natural selection (s = 0.014 - 0.19; (Bersaglieri et al., 2004)). It is interesting to note that similar estimates for the strength of selection have been obtained for one of the major African LP variants (Tishkoff et al., 2007).

It is unlikely that lactase persistence would provide a selective advantage without a supply of fresh milk and this has lead to a gene-culture co-evolutionary model where lactase persistence is only favoured in cultures practicing dairying (Kretchmer, 1972, Simoons, 1970, McCracken, 1971b, Aoki, 1986), and dairying is more favoured in lactase persistent populations (Bayless et al., 1971, Nei and Saitou, 1986, Simoons, 1970, McCracken, 1971a). The reasons why LP, in conjunction with dairying, should confer such a strong selective advantage remain open to speculation. Flatz and Rotthauwe (1973) proposed the *calcium assimilation hypothesis*, whereby a lactase persistence allele is favoured in high-latitude regions because reduced levels of sunlight do not allow sufficient synthesis of vitamin-D in the skin. Vitamin D is required for calcium absorption and milk provides a good dietary source of both nutrients. Additional factors are likely to include the ability to consume a calorie and protein-rich food source, the relative constancy in the supply of milk (in contrast to the boom-and-

bust of seasonal crops), and the value of fresh milk as a source of uncontaminated fluids. It is likely that the relative advantages conferred by these various factors differ in Europe and Africa.

Estimates of the age of the *-13,910*T* correspond well with estimates of the onset of dairying in Europe. Slaughtering age profiles in sheep, goats and cattle suggest dairying was present in south-eastern Europe at the onset of the Neolithic (Vigne and Helmer, 2007, Bartosiewicz, 2007), while residual milk proteins preserved in ceramic vessels provide evidence for dairying in present day Romania and Hungary 7,900-7,450 years BP (Craig et al., 2005). Furthermore, residual analyses of fats indicate dairying at the onset of the Neolithic in England, some 6,100 years BP (Copley et al., 2003, Copley et al., 2005), and after to 8,500 BP in the western parts of present day Turkey (Evershed et al., 2008). Allelic age estimates are also consistent with the results of a recent ancient DNA study (Burger et al., 2007) which showed that the *-13,910*T* allele was rare or absent among early farmers from Central and Eastern Europe. These observations lend support to the view that *-13,910*T*, and thus LP, rose rapidly in frequency only after the onset of dairying, as opposed to the 'reverse-cause' hypothesis (Nei and Saitou, 1986, Bayless et al., 1971, Simoons, 1970, McCracken, 1971a), whereby dairying developed in response to the evolution of LP.

Archaeological studies estimate that farming originated in the Near East about 10-11kya as a result of a mild climate and the availability of wild crops that were potential for farming (Bellwood, 2005). The change into a farming lifestyle from hunting-gathering had brought to a substantial change in lifestyle. Although in early stages of farming the life expectancy of farming was lower than this of hunter gatherers, it introduced the option of one carer for several infant, which enabled women to have shorter intervals between child births, and as a result the population density was increased (Diamond, 2002). The Neolithic transition in Europe from hunting gathering to farming started approximately 9,000 years BP, and has been attempted to explain by two major mechanisms: the Demic Diffusion (DD) (Cavalli-Sforza et al., 1994) and Cultural Diffusion (Zvelebil and Zvelebil, 1988) models. According to the DD model, farming had spread in Europe and replaced hunting gathering by means of physical migration, while the CD model asserts that farming had spread by means of the spread of idea and technology. Assuming the DD model, we would expect the modern European gene pool

to consist mostly of Anatolian/Near Eastern ancestry, while if we assume the CD model then the modern European genetic ancestry would be expected to consist mostly of earlier hunter gatherers. A simulation method testing both hypotheses has suggested that it is likely that the mechanism was a complex combination between the DD and CD models (Currat and Excoffier, 2005). The spread in Europe had a south-eastern – north-western cline, with faster migration along coastlines (Clark, 1965). Figure 5.1 shows the spread of farming in Europe that I obtained by applying the *Natural Neighbour* algorithm (Watson, 1994) surface interpolating the calibrated c-14 dates of the arrival of farming to 761 locations around Europe and West Asia (Pinhasi et al., 2005).

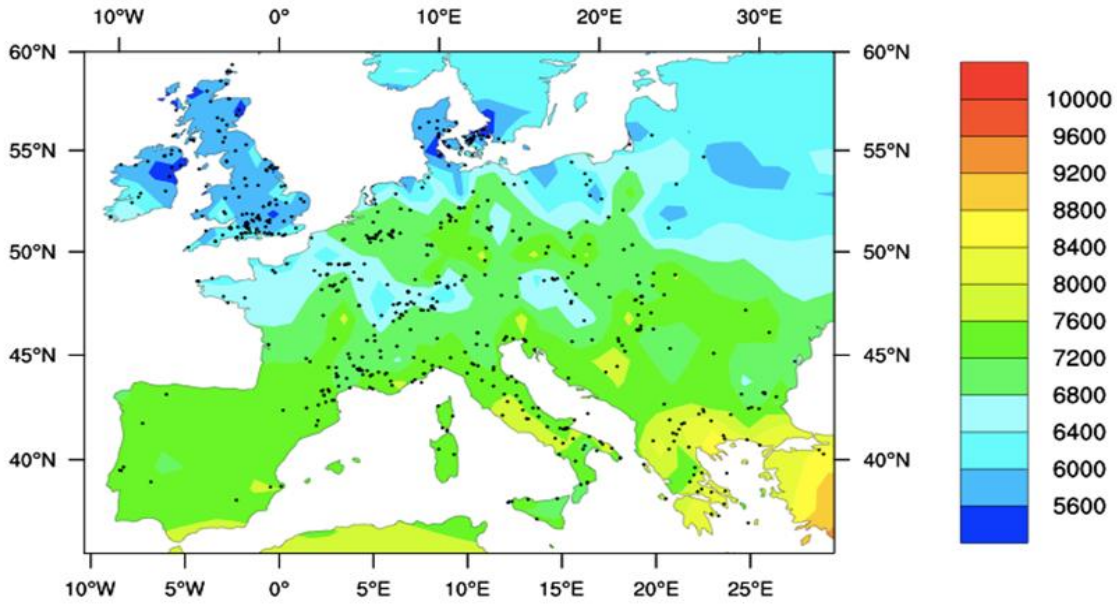

Figure 5.1. The dates of farming to different parts of Europe and West Asia. The contour map was calculated by using the *Natural Neighbour* surface interpolation method. The colour bar represents years before present, dots represent the archaeological sites where data were collected. The map was plotted using the *PyNGL* module (http://www.pyngl.ucar.edu/).

Archaeozoology record show that the domestication of goat, cow, pig, and cattle was likely to have co-evolved with agriculture between 12-10kya (Ucko, 2007). A study has demonstrated a substantial geographic coincidence between high diversity in cattle milk genes, locations of the European Neolithic cattle farming sites, and present day lactase persistence in European, suggesting a gene-culture coevolution between cattle and Neolithic Europeans (Beja-Pereira et al., 2003).

Important questions remain regarding the location of the earliest *-13,910*T*-carrying dairying groups and the demographic and gene-culture co-evolutionary processes that

shaped the modern distribution of LP in Europe. The present-day distribution of the *-13,910*T* allele might be taken to indicate an origin in Northwest Europe. However, the earliest archaeozoological and residual lipid and protein evidence for dairying is found in the Near East, in Southeast Europe and in Mediterranean Europe (Vigne and Helmer, 2007, Evershed et al., 2008, Vigne, 2006). While these observations can seem contradictory, forward computer simulations have shown that the centre of distribution of an allele can be far removed from its location of origin when a population expands along a wave front (Edmonds et al., 2004, Klopfstein et al., 2006).

Assuming that the *-13,910*T*-allele was only subjected to strong natural selection in dairying groups, it is likely that *-13,910*T*-carrying dairyers underwent demographic expansion to a greater extent than non-dairying groups. While gene flow between dairying and non-dairying groups would ultimately lead to genetic homogeneity, under conditions of limited gene flow between cultural groups, it is plausible that the earliest LP peoples would have made a higher contribution to the European gene pool than their non-LP neighbours. In this study I used demic forward computer simulations to examine potential scenarios for the spread of LP in Europe. I simulated three interacting cultural groups (hunter gatherers, non-dairying farmers and dairying farmers) and tracked the spread of an allele that is selected only in one group (dairying farmers). I also tracked the expected proportion of genetic ancestry from the geographic region where LP/dairying coevolution began. I parameterized intrademic gene flow between cultural groups, interdemic gene flow, sporadic longer-distance migration, the cultural diffusion of subsistence practices and selection favouring lactase persistent dairyers. I compared the predicted frequency of a LP allele and arrival dates of farmers – from simulation outcomes – to known frequencies of the *-13,910*T* allele (Mulcare et al., 2004, Bersaglieri et al., 2004) and carbon-14 based estimates of the arrival dates of farmers (Pinhasi et al., 2005) at different locations throughout Europe. Approximate Bayesian computation (ABC) was employed – a set of methods that allow the estimation of parameters under models too complex for a full-likelihood approach (Beaumont et al., 2002). By comparing summary statistics on the observed data with those computed on the simulated datasets, ABC enables estimation of the key demographic and evolutionary parameters including the region where LP-dairying coevolution in began in Europe.

5.2. Material and Methods.

5.2.1. The Simulation Model.

The simulation approach is motivated by a previous demic computer simulation study (Barbujani et al., 1995) and has features in common with more recent applications of this approach (Ray et al., 2003, Currat and Excoffier, 2005, Excoffier, 2004). Geographic space is modelled as a series of rectangular demes arranged to approximate the European landmass (2375 land demes and 1511 sea demes). Each deme has attributes of elevation, area (which varies due to the curvature of Earth and is calculated accordingly for each individual deme), and a climate (Mediterranean, Temperate, or Cold/Desert – see Figures 5.2 and 5.3).

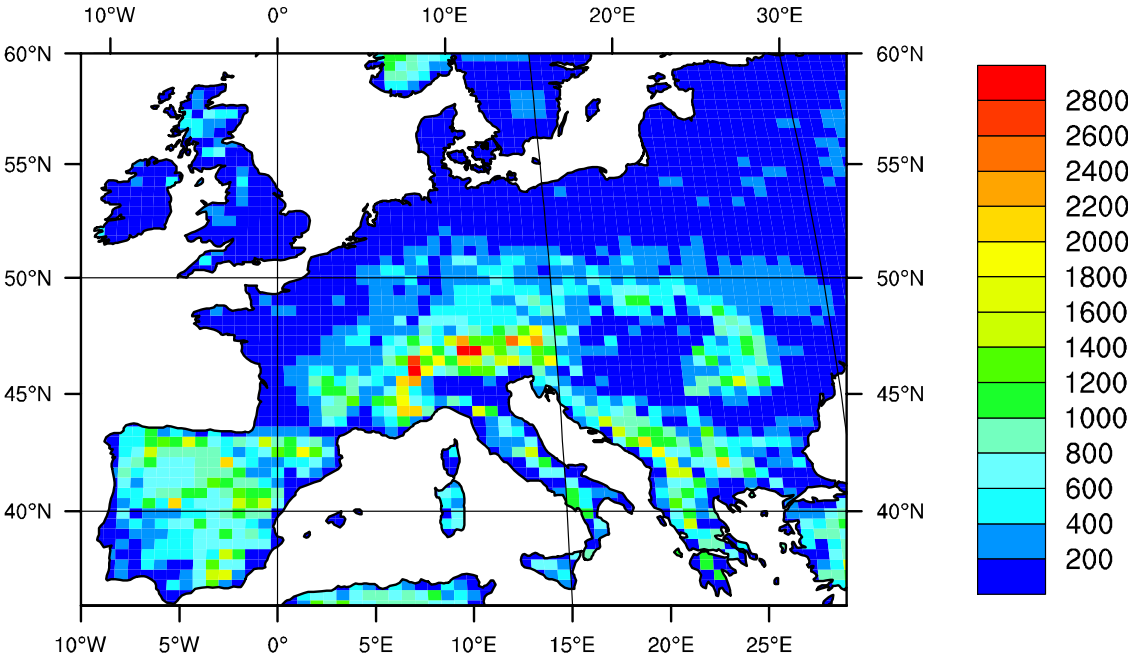

Figure 5.2. The average elevation at each simulated deme. The colour bar represents average elevation in meters above sea level.

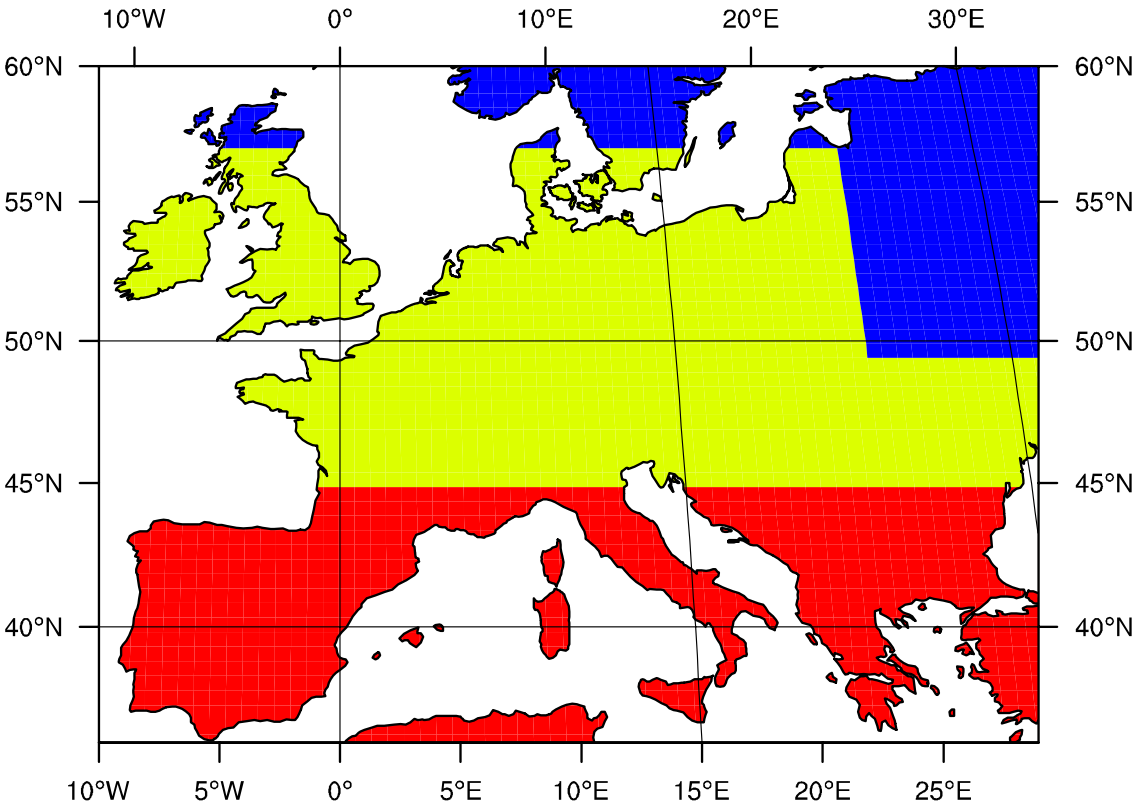

Figure 5.3. The climate at each simulated deme. Red colour represents Mediterranean climate, yellow represent temperate climate, and blue is for cold/desert climate.

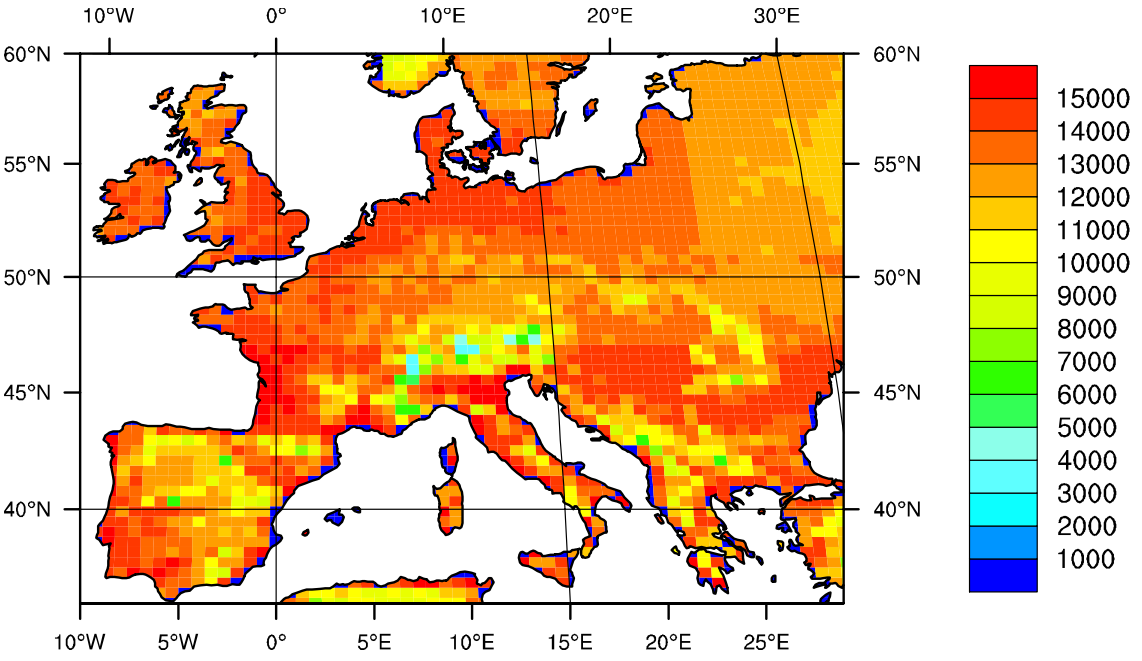

Figure 5.4. The carrying capacity at each simulated deme. Values dependent on the deme's average elevation and climate (Figures 5.2 and 5.3, respectively).

A maximum total population size is specified for each land deme taking into account its area, and assuming that lower elevation and mild Mediterranean climate results in a

greater potential population size, while harsher conditions, such as high elevations and cold/desert climates, result in a smaller potential population size (Colledge et al., 2004). The ratio for the relative contribution coefficients of climate and elevation factors to the population size is fixed at 1:4 in this study; meaning that elevation has a more dramatic effect than climate on population size. The sum of the carrying capacities of the three cultural groups (the deme's maximum population size, Figure 5.4) is calculated by:

$$K_{deme} = (0.2cl + 0.8el)D_{\max}A_{deme} \quad (1)$$

where *cl* and *el* are the climatic and relative elevation factors, respectively; *cl* having values of 1 for Mediterranean, 2/3 for Temperate, and 1/3 for cold/desert climates (Colledge et al., 2004) (see Figure 5.3), and *el* being calculated as:

$$el = 1 - \frac{deme\_elevation}{max\_elevation} \quad (1.1)$$

So *el* ranges between 0 at the highest elevation and 1 at sea level (see Figure 5.2). D_{max} is the maximum population density and is fixed at 5 individuals per km^2 (i.e. in a sea level Mediterranean climate deme (Hassan, 1981)), and A_{deme} is the area of the deme in km^2.

Each deme contains three distinct cultural groups: non-dairying farmers (F_{nd}), dairying farmers (F_d), and hunter-gatherers (HG). The ratios of ceiling population size for F_{nd}, F_d, and HG (as a proportion of the total maximum population size for the deme, K_{deme}) are 50:50:1 respectively (Bellwood, 2005, Hassan, 1981). Each cultural group in each deme is assigned a frequency for an allele that is subjected to genetic drift (modelled by intergenerational binomial sampling) and an allele at an unlinked locus that is not (as explained below). Initially the frequency of both 'alleles' is set at zero. The former represents a LP allele and is subject to selection of intensity *s*, only in the F_d group. The latter, here termed the GB (genetic background) 'allele', is used to track the general genetic ancestry component from the region where the LP allele is first found among dairying farmers. It will be used to infer the *expected* proportion of genes that originate from this region. The two alleles are assumed to be unlinked and are modelled

separately. I treat *s* as an unknown but bounded parameter, and choose random values ranging from 0 to 0.2 in simulations (Bersaglieri et al., 2004).

The LP and GB 'allele' frequency dynamics are determined in each generation by five processes: (1) intrademic bidirectional geneflow between cultural groups; (2) bidirectional geneflow between demes (interdemic) within the same cultural groups; (3) sporadic unidirectional migration within the same cultural groups; (4) cultural diffusion (CD); and (5) selection operating on LP allele-carrying individuals within the F_d group. Hardy-Weinberg equilibrium within each cultural group within each deme is assumed. Population size increase for each cultural group in each deme is modelled by logistic growth, limited by the carrying capacity of each group within each deme. The growth rate is fixed to r = 1.3 per generation, a value estimated from data of world population growth rate over the last 10,000 years, excluding the post-Industrial Revolution population boom (US Census Bureau: www.census.gov). In addition, the F_d group is allowed to increase in size as a function of the selective advantage of the LP allele, *s*, by considering the number of LP individuals and the selective advantage to being a LP dairyer (see equation 5).

I define *intrademic bidirectional geneflow* as the exchange of individuals between different cultural groups within a deme (see Figure 5.5). A proportion of individuals in each cultural group, P_c, are deemed 'available to change group'. The actual number of individuals that exchange genes between cultural groups *i* and *j*, $B_{i\leftrightarrow j}$, is determined as follows:

$$B_{i\leftrightarrow j} = \frac{N_j}{N_i + N_j} P_c N_i \qquad (2)$$

Where N_i *and* N_j are the total number of individuals belonging to each cultural group. I treat P_c as an unknown but bounded parameter, and choose random values ranging from 0 to 0.2 in simulations (Spielmann and Eder, 1994, Mace, 1993).

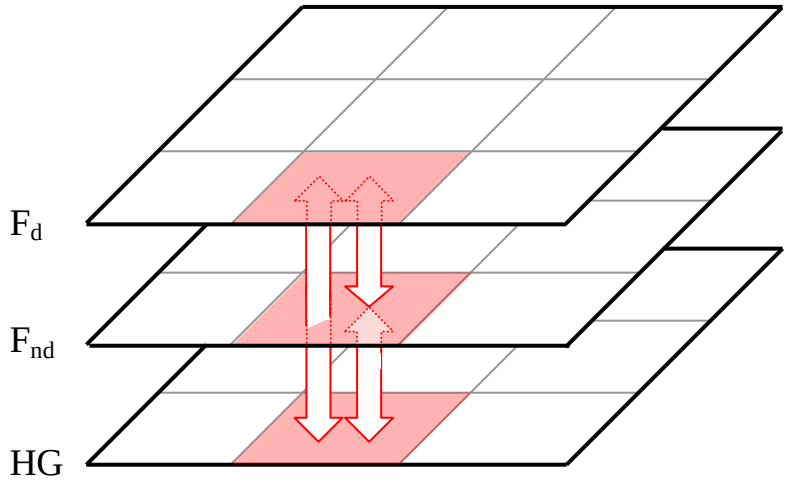

Figure 5.5. Intrademic bidirectional geneflow between all cultural groups within a deme. The number of individuals exchanged between two cultural groups is determined by equation (2).

I define *interdemic bidirectional geneflow* as the exchange of individuals between the same cultural groups in neighbouring demes (see Figure 5.6). A proportion of individuals in each cultural group, P_d, are deemed 'available to change deme'. The actual number exchanged is determined using the same formula as for *intrademic bidirectional geneflow* (equation 2), except I substitute P_d for P_c, and N_i *and* N_j are the total number of individuals belonging to each cultural group in each neighbouring deme. In each generation, each cultural group in each deme undergoes *bidirectional geneflow* with one neighbouring deme, randomly chosen from the 8 possible.

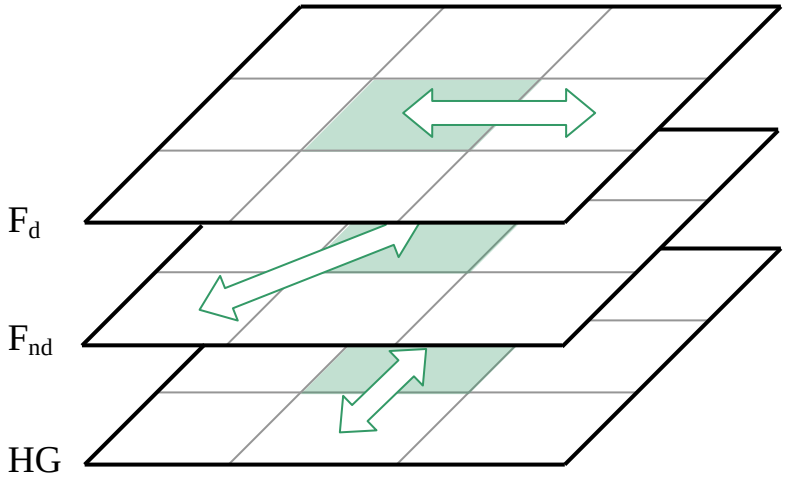

Figure 5.6. Interdemic bidirectional geneflow between similar cultural groups in different demes. The number of individuals exchanged between two groups is determined by equation (2).

I define *sporadic unidirectional migration* as the movement of some individuals in a particular cultural group and deme to the same cultural group in a different deme (see Figure 5.7). A proportion of individuals in each cultural group, P_s, are deemed 'available to migrate'. The actual number of individuals that migrate, N_{mig}, is dependent on the 'pressure' to leave the current deme and the availability of unoccupied carrying capacity in the destination deme ('attractiveness'), and is determined as follows:

$$N_{mig} = \frac{1}{2}\left(\frac{K_{dest} - N_{dest}}{K_{deme_{dest}}} + \frac{N_{curr}}{K_{curr}}\right) P_s N_{curr} \qquad (3)$$

Where $K_{deme_{dest}}$ is the value of K_{deme} (see equation 1) in the destination deme, K_{curr} and K_{dest} are the carrying capacities for a specific cultural group, and N_{curr} and N_{dest} are the number of people in the same cultural group, in the current home and destination demes respectively. I treat P_s as an unknown but bounded parameter, and choose random values ranging from 0 to 0.2 in simulations. The destination deme is chosen by a Gaussian random-walk process, which takes into account the mobility of the cultural group and the topography of the home deme. The Gaussian distribution is centred on the home deme; and its standard deviation is the product of the mobility of the cultural group, M_i, and the relative mobility factor of the home deme, M_{curr}. I treat M_i as a separate unknown but bounded parameter for each of the three cultural groups, and choose random values ranging from 0 to 3 (demes) in simulations. M_{curr} is determined for each deme by its elevation, allowing greater mobility at lower elevations (Weale et al., 2001, Thomas et al., 2008), with fixed values of 0.5 (demes) at mountainous terrain (above 1,100 meters), 1.0 at lowlands (below 1,100 meters), and 1.5 at coastal demes. The *sporadic unidirectional migration* function allows movement overseas, but whenever a sea deme is identified as a non-realistic destination deme the nearest neighbouring coastal deme is chosen instead. This feature, together with the attractiveness of low elevation land and the higher M_{curr} value for coastal demes, creates the realistic tendency of a faster spread of farming along coastlines, consistent with archaeological data (Clark, 1965).

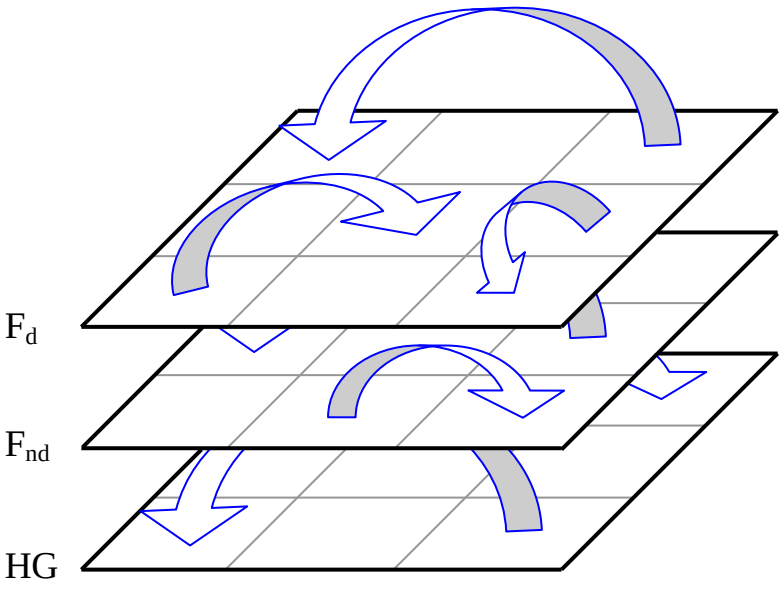

Figure 5.7. Sporadic unidirectional migration. Illustrating only one potential scenario, as migrants potentially leave and migrate to every populated deme. The migrants' destination deme is chosen by a Gaussian random walk process, centred on the home deme and with a standard deviation of the product of the cultural group mobility, M_i, and the relative mobility factor of the home deme, M_{curr}. See equation (3).

I define *Cultural Diffusion* (CD) as the spread of culture and technology by learning through exposure rather than by migration (see Figure 5.8). In the simulations a proportion of individuals in each cultural group, P_{dif}, are deemed 'available to convert' from one cultural group to another. The number of individual that convert from cultural group *i* to cultural group *j*, $N_{i \to j}$, is determined by this parameter and the proportion of the carrying capacity (*K*) of the home deme (deme *0*) and in the 8 neighbouring demes (demes *1* to *8*) that is taken up by cultural group *j*, as follows:

$$N_{i_0 \to j} = N_{i_0} P_{dif} \left(b \frac{N_{j_0}}{K_{j_0}} + (1-b) \frac{1}{8} \sum_{n=1}^{8} \frac{N_{j_n}}{K_n} \right) \qquad (4)$$

where *b* is the relative influences of the home deme and the 8 neighbouring demes (fixed to 0.75). I treat P_{dif} as an unknown but bounded parameter, and choose a random value ranging from 0 to 0.2 in each simulation. That value is then applied to 'conversions' between all 3 cultural groups.

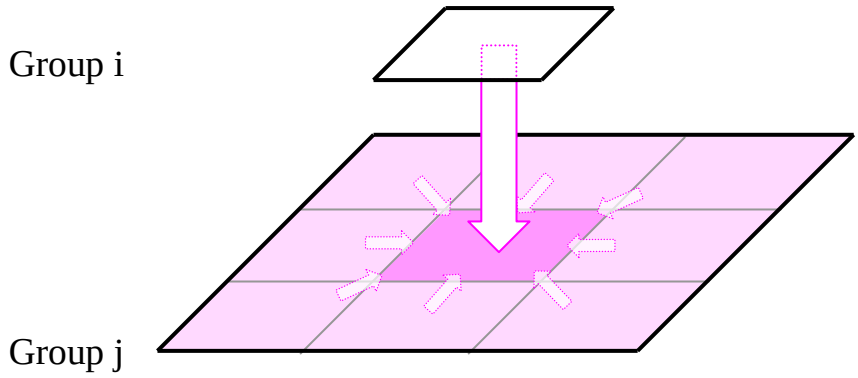

Figure 5.8. Cultural diffusion. The number of individuals in cultural group i converting to cultural group j is determined by the proportion of the carrying capacity taken by group j in the home deme of group i and the eight surrounding demes. See equation (4).

The geographic location where LP / dairying gene-culture coevolution starts is chosen at random from all land demes. This LP mutation is initialized at a frequency of 0.1 in F_d when their population size reaches a critical size in the chosen start deme, set to a minimum of 20 individuals per deme in simulations. While we would expect any *de novo* mutation to always have an initial frequency of 1/2N, we also expect that it will have a high probability of extinction unless selection is very strong (Haldane, 1927). Indeed, in preliminary simulations this was observed (data not shown). Thus, for computational efficiency I condition on the LP mutation having already reached a frequency of 0.1 in F_d in the deme of origin. However, such a starting frequency means that little more than four LP alleles are initialized in simulations. Selection acting on the LP allele, *p*, increases its frequency in F_d only, as follows (Maynard Smith, 1998):

$$p' = \frac{p^2(1+s) + pq(1+s)}{1 + s(p^2 + 2pq)} \tag{5}$$

where *s* is the selection coefficient for *p*, and *p'* is the new LP allele frequency. In addition, selection acting on the LP allele increases the number, *N*, in F_d as follows:

$$N' = N(1 + s(p^2 + 2pq)) \tag{6}$$

where *N'* is the new number of F_d in a particular deme.

All simulations were run for 360 generations which, assuming a generation time of 25 years (Thomas et al., 2006, Tremblay and Vezina, 2000), corresponds to the 9,000-year history of farming in Europe. I performed 200,000 simulations in total.

The genetic contribution of the population living in the region of origin of LP / dairying gene-culture coevolution to the overall European population is tracked over generations by calculating the GB 'allele' frequency over all demes in all 3 cultural groups. In the generation when the LP allele is initialized, all cultural groups in the origin deme and 8 neighbouring demes are assigned the unlinked GB 'allele' at a frequency of 1. The GB 'allele' is subjected to the same intra- and inter-deme geneflow and migration processes as described above, but is not subject to drift, as modelled by binomial sampling, or to selection. At the end of each simulation this GB allele is taken to represent the general genetic contribution of the population living in the region of origin of LP to the modern European population. The ancestry component of Europeans, at any generation, that originates from people living in the region of origin of the LP allele (F_{GB}) is calculated as follows:

$$F_{GB} = \frac{\sum_{i}^{n} \sum_{j \in \{F_{nd}, F_{d}, HG\}} p_{GB_{ij}} N_{i_j}}{\sum_{i}^{n} N_i} \qquad (7)$$

where n is the number of land demes, N_i is the total number of people in deme *i*, and p_{GBij} and N_{ij} are the frequency of the GB 'allele' and the population size in deme *i* / cultural group *j*, respectively.

5.2.2. Parameters Estimation.

To estimate parameters of interest an ABC approach was applied, following (Beaumont et al., 2002). By comparing summary statistics computed on each simulated dataset to those from the observed data, only those simulations with summary statistics sufficiently close to the target (i.e. the observed summary statistics) are accepted, remainder are rejected. Then a weighted local-linear regression was performed on these retained parameter sets, with weight determined by the "distance" between the simulation summary statistics and the target (all details below). This generates

approximate marginal posterior probability distributions for each parameter of interest, from which the modal point estimates are derived. The chosen summary statistics, **U**, are the frequencies of the *-13,910*T* allele at 12 different sample locations around Europe, the Near East and western Asia (Mulcare, 2006a, Bersaglieri et al., 2004). In addition, the times to arrival of farming at 11 of the same locations (the Anatolia location is excluded as the simulation model is initialized with this as the origin of the spread of farming into Europe) are included as summary statistics. These are not summary statistics *sensu stricto* but are parameters in the model for which independent estimates are obtained. However, the simulations, being stochastic, generate a distribution of arrival times, and should be conditioned on those that are consistent with the known archaeological evidence (Figure 5.1). The most straightforward way to do this is to place a point prior on the arrival dates, and then condition on these using the ABC machinery, as if they are summary statistics. The point priors for the arrival dates of farming at 11 of the 12 sampling locations considered (Anatolia was set to 9,000 years as the simulations begin 360 generations ago in 'an Anatolia' populated by farmers) were calculated as follows: (1) The average nearest-neighbour distance (ANND) between each sampling location was calculated (557.13km). (2) A 2-D Gaussian sampling region was constructed around each of the 11 sampling locations, of standard deviation = ANND / 1.96 (this ensures that 95% of each Gaussian sampling region will be within the ANND). (3) A weighted average of all dates within 3 standard deviations of the sampling location was calculated using all calibrated carbon-14 earliest farming arrival dates from Pinhasi *et al*. (Pinhasi et al., 2005), and weighting using the distance from the sampling location and the standard probability density function for a Gaussian distribution. Assuming a generation time of 25 years (Thomas et al., 2006, Tremblay and Vezina, 2000) these observed dates are converted to generations from the start of the simulation, which was set at 9,000 years BP or 360 generations ago (see Table 5.1). Two Spearman's rank-order correlation coefficients are also included, calculated separately for the 12 T-allele frequencies and the 11 times to arrival of farming, giving a total of 25 summary statistics. When calculating these statistics for the simulated data: LP frequencies are taken in the final generation of the simulation at the 12 corresponding geographic locations; and the time to arrival of farming is defined as the simulation generation at which either F_d or F_{nd} reach 1% of their carrying capacity within each of the 11 corresponding location demes. All time to arrival of farming statistics are scaled to the interval [0,1] by dividing by the total

number of simulated generations (360).

Table 5.1. -13,910*T allele frequencies, inferred farming start dates and geographic coordinates of 12 locations data used in ABC analysis. Inferred arrival of farming dates were based on: [1] a weighted average of all calibrated carbon-14 earliest farming arrival dates from Pinhasi et al. [31] within 853 km of each sampling location, weighted using the distance from the sampling location and the standard probability density function for a Gaussian distribution of s.d. 285 km; and [2] by assuming a constant rate of spread of farming (estimated at 0.9 km/year (Pinhasi et al., 2005)) and calculating the great circle distance from Anatolia to each sampling location. All inferred generations after the start of farming were calculated by assuming a generation time of 25 years (Thomas et al., 2006, Tremblay and Vezina, 2000).

Location	−13,910 *T allele frequency	N individuals used to assess −13,910*T allele frequency	Reference for 13,910 *T allele frequency	Great circle distance from central Anatolia (km)	Inferred farming arrival date in years BP [1] (generations after start of simulation)	Inferred farming arrival date in years BP [2] (generations after start of simulation)	Latitude	Longitude
Turkey	0.031	49	[7]	0	9000 (0)	9000 (0)	38.00	30.00
Greece	0.134	41	[7]	550	7932 (43)	8389 (24)	37.98	23.73
Tuscany	0.063	16	[8]	1699	7274 (69)	7112 (76)	43.77	11.25
Sardinia	0.071	56	[8]	1829	7371 (65)	6968 (81)	39.00	9.00
North Italy	0.357	28	[8]	1880	6992 (80)	6911 (84)	45.68	9.72
Scandinavia	0.815	360	[8]	2523	5833 (127)	6197 (112)	59.33	18.05
Germany	0.556	60	[7]	2309	6396 (104)	6434 (103)	53.55	10.00
France	0.431	58	[8]	2523	6552 (98)	6197 (112)	48.87	2.33
French Basque	0.667	48	[8]	2666	7078 (77)	6037 (119)	43.00	−1.00
Southern UK	0.734	64	[7]	2785	5954 (122)	5905 (124)	51.50	−0.12
Orkney	0.688	32	[8]	3325	5778 (129)	5306 (148)	58.95	−3.30
Ireland	0.954	65	[7]	3349	5807 (128)	5260 (150)	54.37	−7.63

Parameters of interest, ϕ, are: the east-west and north-south coordinates of the location where the LP-allele first undergoes selection among F_d; the generation at which this selection starts; the selective advantage of LP within the F_d group, s; the proportion available for interdemic bidirectional geneflow, P_d; the proportion available for intrademic bidirectional geneflow among cultural groups, P_c; the rate of cultural diffusion, P_{dif}; the proportion of people available for sporadic migration, P_s; the mobility of each of the three cultural groups, M_i; and the contribution of people living in the deme where LP-dairying gene-culture coevolution began and its 8 surrounding demes, F_{GB}, to the modern European gene-pool. The uniform prior distributions for each parameter are given in Tables 5.2 and 5.3.

Table 5.2. Posterior estimates of demographic and evolutionary parameters (mean, mode and 95% credibility interval). Posterior distributions were by estimated by ABC employing regression adjustment and weighting of simulations accepted at the 0.5% tolerance level (Beaumont et al., 2002).

Parameter	Parameter symbol	Prior range	Units	Posterior 95% CI	Mode	Mean
Interdemic BD GF	P_d	0 to 0.2	Proportion	0.00716 - 0.171	0.0440	0.0620
Intrademic BD GF	P_c	0 to 0.2	Proportion	0.00206 - 0.0867	0.0153	0.0339
Cultural Diffusion	P_{dif}	0 to 0.2	Proportion	0.00113 - 0.0847	0.0136	0.0321
Selective Advantage	s	0 to 0.2	Proportion	0.0518 - 0.159	0.0953	0.0957
Proportion available for Sporadic migration	P_s	0 to 0.2	Proportion	0.0575 - 0.251	0.129	0.132
Sporadic migration mobility HG	M_{HG}	0 to 3	Demes	0.333 - 2.17	1.16	1.20
Sporadic migration mobility F_{nd}	M_{Fnd}	0 to 3	Demes	0.311 - 1.18	0.733	0.713
Sporadic migration mobility F_d	M_{Fd}	0 to 3	Demes	2.15 - 3.93	3.12	3.08
Time of Origin of Gene-Culture coevolution	Not a parameter *sensu stricto*	[0 to 9000]	Years	6256 - 8683	7441	7553
Genetic contribution to modern European genepool	Not a parameter *sensu stricto*	[0 to 100]	Percent	2.83 - 27.4	7.47	11.1

Table 5.3. Parameters of simulation model. 'Flat' indicates that a uniform prior was used.

Symbol	Fixed/ variable (F/V)	Value	Description
D_{max}	F	5	Maximum population density per per km^2.
cl	F	cold=1/3, temperate= 2/3, med=1	Climatic factor modifying carrying capacity.
el	F	0 to 1, depending on elevation values as a proportion of max elevation.	Elevation factor modifying capacity.
a	F	0.2	Coefficient for relative contribution of climatic factor (a) and elevation factor (1-a) to deme carrying capacity.
ratios	F	1:50:50 for HG, Fd, Fnd, respectively.	Ratios between the carrying capacities of the cultural groups, summing to deme's carrying capacity.
r	F	1.3	Logistic population growth rate.
gen	F	360	Number of generations in one simulation run. One generation = 25 years.
M_{curr}	F	1.5, 1, 0.5 for coastal, lowland, and mountains, respectively.	Topography factor modifying sporadic migration distance. Mountains defined as elevation>1100m
b	F	0.75	Cultural diffusion coefficient for relative contribution of local population density and 1-b for surrounding demes' population density.
s	V	0 to 0.2 (flat)	Selective advantage. Affects gene frequencies and population growth.
P_c	V	0 to 0.2 (flat)	Proportion of people available to move to another cultural group within a deme (bidirectional).
P_d	V	0 to 0.2 (flat)	Proportion of people available to move to the same cultural group in a neighbouring deme (bidirectional).
P_s	V	0 to 0.2 (flat)	Proportion of people available for sporadic migration.
M_{Fnd}	V	0 to 3 (flat)	Sporadic migration mobility of non-dairying farmers (s.d. of the Gaussian random walk distribution given by the product of this value, M_i, and the relative mobility factor of the home deme, M_{curr}).
M_{Fd}	V	0 to 3 (flat)	Sporadic migration mobility of dairying farmers.
M_{HG}	V	0 to 3 (flat)	Sporadic migration mobility of hunter-gatherers.
P_{dif}	V	0 to 0.2 (flat)	Maximum proportion of people available for converting into another cultural group.
location	V	Any land deme	Start location coordinates for LP-dairying coevolution.

The full ABC algorithm is as follows: (1) choose the summary statistics **U** as outlined above and calculate their values, **u**, for the observed data (these are given in Table 5.1), (2) choose a tolerance level δ (a proportion of the best fitting simulations, P_δ, is predefined to accept and from this calculate an implicit tolerance level δ), (3) sample a parameter set ϕ_i from the pre-determined prior distribution of ϕ, (4) simulate forward under the model using parameter set ϕ_i, (5) in the final generation of the simulation the

summary statistics, $\mathbf{u_i}$, is calculated for this simulated data, (6) If $||\mathbf{u_i} - \mathbf{u}|| \leq \delta$ (where $||.||$ is the Euclidean norm between the two vectors) the parameter set ϕ_i is accepted, (7) steps 3 to 6 are repeated until a sufficient number of retained parameter sets is obtained, (8) A local-linear standard multiple regression is then performed to adjust the ϕ_i, with each ϕ_i weighted according to the size of $||\mathbf{u_i} - \mathbf{u}||$ using the Epanechnikov kernel function $K_\delta(t)$ (see (Beaumont et al., 2002) for details), (9) The resulting fitted parameter sets ϕ_i^* form a random sample from the approximate joint posterior distribution $P(\phi|\mathbf{U}=\mathbf{u})$. All retained parameters – except for the two coordinate values and the generation at which the co-evolutionary process starts – were log transformed prior to the regression step, and subsequently back-transformed to produce the fitted parameter sets ϕ_i^*, as suggested by Beaumont *et al*. (Beaumont et al., 2002).

The simulation and ABC analysis procedures were written in the Python Programming Language (URL: http://www.python.org/) employing the numarray and Numpy array handling libraries. Maps were generated using the Python library *PyNGL* (http://www.pyngl.ucar.edu/). Post-ABC analysis data was processed and visualised using the statistical package 'R' (URL: http://www.R-project.org/).

5.3. Results.

Simulation time. Unlike the simulation models used in related studies (Barbujani et al., 1995, Ray et al., 2003, Excoffier, 2004, Currat and Excoffier, 2005) the one presented is stochastic and more parameter-heavy. In addition, it was written in Python using the object orientated paradigm which, while utilizing some highly efficient array-handling libraries such as numarray and Numpy, is considerably slower than purely procedural simulations written in a lower-level programming language such as C++. A single simulation takes about 170 seconds on a 3.0GHz Athlon™ 64 processor.

Demographic parameter estimation. The regression adjustment and weighting step of ABC were applied to simulations accepted at the 0.5% tolerance level (Beaumont et al., 2002). As can be seen in Figure 5.9, for some parameters, such as the sporadic migration mobility of hunter-gatherers, little information could be obtained using the observed data (also see Table 5.2). This is unsurprising since we would expect the value for this parameter to make little difference to either the arrival time of farming or the

distribution of a LP allele. However, the analyses did appear informative for some key parameters. (1) The 95% credibility interval (CI) for selective advantage of the LP allele among dairying farmers, *s*, is considerably narrower (0.0518 - 0.159; mode = 0.0953) than its prior (0 - 0.2); (2) The 95% CI for the proportion of individuals available for intrademic bidirectional geneflow between cultural groups, P_c, (0.00206 - 0.0867; mode = 0.0153) falls in the lower end of its prior range (0 - 0.2); and (3) The sporadic migration mobility of dairying farmers, M_{Fd}, is significantly higher than that for non-dairying farmers; 99.998% of 100,000 random draws from the former are greater than those from the latter. I note that for some parameters the estimated 95% credible intervals lie outside the upper prior bound. This is a consequence of using regression adjustment in a model with rectangular priors (Beaumont et al., 2002). Points in which the parameter value is close to the boundary, but with summary statistics that are distant from those observed, may have their parameter values projected outside the boundary by the regression method.

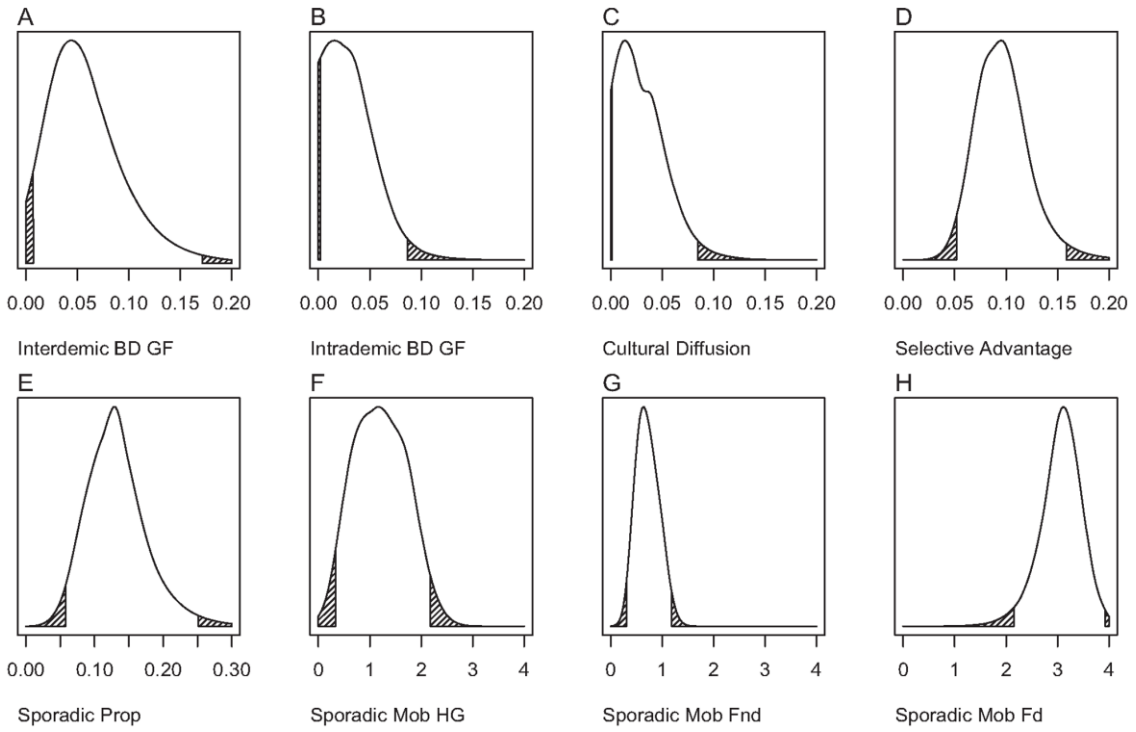

Figure 5.9. Approximate marginal posterior density estimates of demographic and evolutionary parameters. ABC was performed using regression adjustment and weighting, following acceptance at the 0.5% tolerance level (Beaumont et al., 2002). The upper and lower 2.5% of each distribution are shaded. For some parameters the estimated 95% credible intervals lie outside the upper prior bound. This is a consequence of the regression adjustment stage of ABC when using rectangular priors (Beaumont et al., 2002). Points in which the parameter value is close to the boundary, but with summary statistics that are distant from those observed, can have their parameter values projected outside the boundary. Parameters estimated are (A) Interdemic bidirectional geneflow, (B) Intrademic bidirectional geneflow, (C) the rate of cultural diffusion of subsistence practices, (D) the selective advantage of a LP allele among dairying farmers, (E) the proportion of individuals in a deme available for sporadic long-distance migration, and the average mobility – in number of demes moved – of (F) hunter-gatherers, (G) non-dairying farmers, and (H) dairying farmers.

To investigate relationships among demographic and evolutionary parameters Spearman's R^2 and p-values were calculated for all possible pairwise joint posterior parameter distribution (see Table 5.4), following acceptance at the 0.5% level and regression adjustment (Beaumont et al., 2002). Figure 5.10 shows those with $R^2 > 0.024$. The following parameter pairs, in order of decreasing R^2, showed non-independence by this criteria: (A) proportion available for sporadic migration and the sporadic mobility of dairying farmers, (B) proportion available for sporadic migration and the sporadic mobility of non-dairying farmers, (C) selective advantage and sporadic mobility of non-dairying farmers, and (D) sporadic mobility of dairying farmers and sporadic mobility of hunter-gatherers. That the first two joint distributions show

negative correlation is unsurprising since changes in the proportion available for sporadic migration, or in the sporadic migration mobility of dairying and non-dairying farmers, will have similar effects on the timing of arrival of farming at different locations.

Table 5.4. Correlations among demographic and evolutionary parameters. Spearman's R^2 (above diagonal) and p-values (below diagonal) are given for all pairwise joint posterior parameter distribution. Posterior distributions were estimated by ABC employing regression adjustment and weighting of simulations accepted at the 0.5% tolerance level (Beaumont et al., 2002). Parameter joint distributions are shown in Figure 2 (main article) for combination returning a Spearman's R^2 value > 0.024.

	Sporadic Prop	Selective Advantage	Sporadic Mob F_d	Sporadic Mob F_{nd}	Interdemic BD GF	Intrademic BD GF	Cultural Diffusion	Sporadic Mob HG
Sporadic Prop		0.00680	0.458	0.118	1.69E-04	0.00229	0.00241	0.00710
Selective Advantage	0.00901		00727	0.0746	0.00222	0.0175	0.0111	5.96E-04
Sporadic Mob F_d	1.40E-135	0.00691		5.94E-04	0.00829	0.0137	0.0124	0.0255
Sporadic Mob F_{nd}	4.01E-29	1.27E-18	0.441		0.0208	1.90E-05	8.89E-05	0.00418
Interdemic BD GF	0.681	0.136	0.00390	4.52E-06		0.0239	0.00451	0.00197
Intrademic BD GF	0.130	2.59E-05	2.08E-04	0.890	8.53E-07		0.00580	4.18E-05
Cultural Diffusion	0.121	8.21E-04	4.14E-04	0.766	0.0334	0.0159		2.32E-04
Sporadic Mob HG	0.00760	0.440	3.70E-07	0.0406	0.160	0.838	0.630	

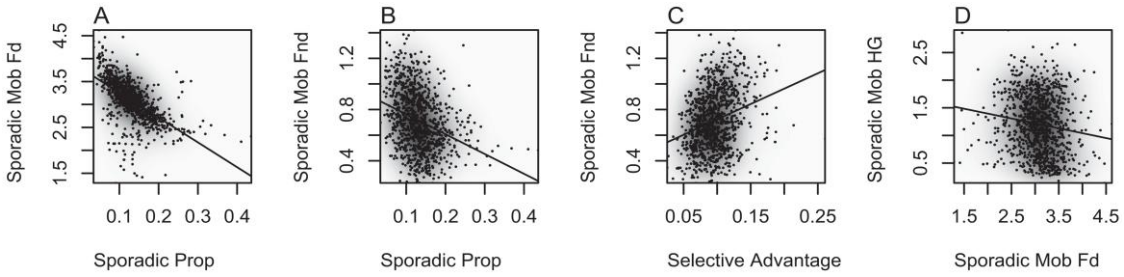

Figure 5.10. Pairwise joint approximate posterior density estimates of demographic and evolutionary parameters showing high degrees of correlation (Spearman's R^2 > 0.024). Points represent regression adjusted parameter values from simulations accepted at the 0.5% tolerance level. Shading was added using 2D kernel density estimation. Parameter combinations shown are the proportion of individuals in a deme available for sporadic long-distance migration versus the average mobility – in number of demes moved – of (A) dairying farmers, and (B) non-dairying farmers, (C) the selective advantage of a LP allele among dairying farmers versus the average mobility of non-dairying farmers, and (D) the average mobility of dairying farmers versus the average mobility of hunter-gatherers.

Geographic and temporal origin of LP-dairying co-evolution: Following acceptance at the 0.5% level and regression adjustment it is estimated that the most probable location where an LP allele first underwent selection among dairying farmers lies in a region between the central Balkans and central Europe (see Figure 5.11). It should be noted that, as simulated, it was not attempted to identify the location where the LP *-13,910*T* allele first arose. Instead it was assumed that it started to rise to appreciable frequencies only after selection began among dairying farmers, initially at the particular location estimated. The timing of the start of this gene-culture coevolution process was therefore strongly influenced by the arrival time of dairying farmers at the location where selection began in simulations. Since simulations that give a good fit to the timing of the arrival of farming were selected at different locations (Pinhasi et al., 2005), a narrow range of dates for when selection began was estimated (95% CI 6,256 to 8,683 years BP; mode = 7,441 years BP; see Figure 5.12A).

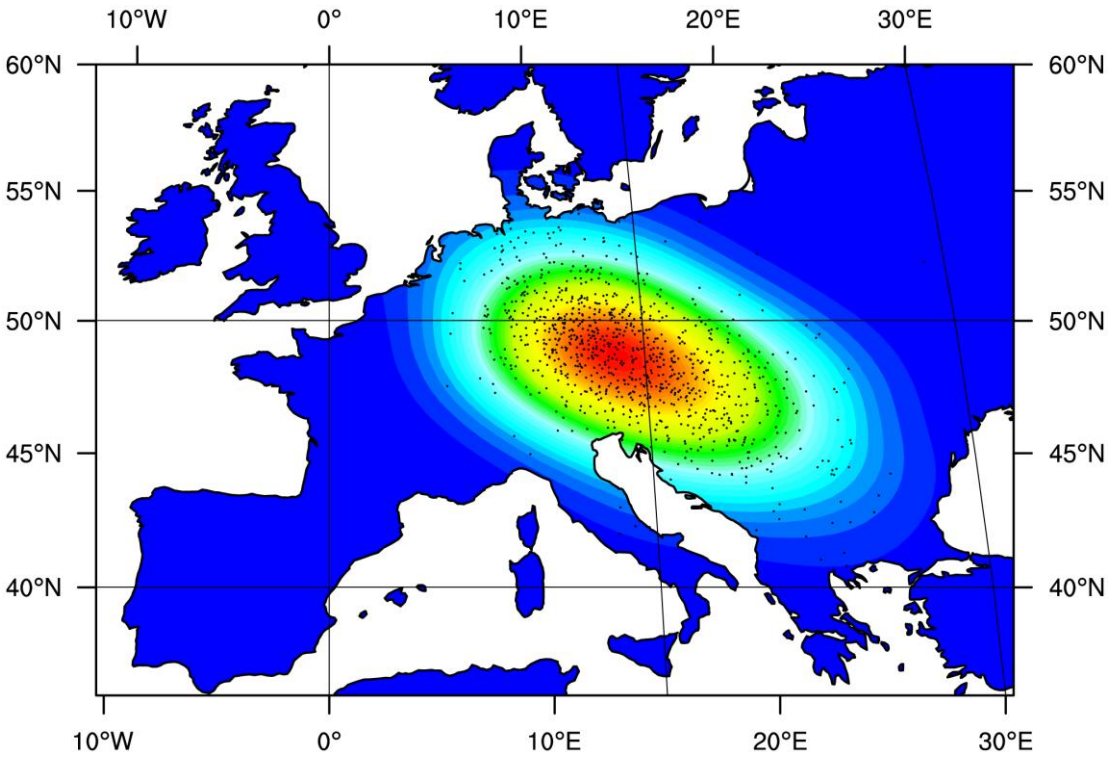

Figure 5.11. Approximate posterior density of region of origin for LP / dairying co-evolution. Points represent regression-adjusted latitude and longitude coordinates from simulations accepted at the 0.5% tolerance level. Shading was added using 2D kernel density estimation.

Genetic contribution of the earliest LP dairying farmers to the modern European gene pool: Although not strictly a parameter of the model presented, the ABC approach had been applied to estimate the genetic contribution of people living in the deme where LP-

dairying gene-culture coevolution began, and its 8 surrounding demes, to the modern European gene-pool (95% CI 2.83 to 27.4%; mode = 7.47%; see Figure 5.12B).

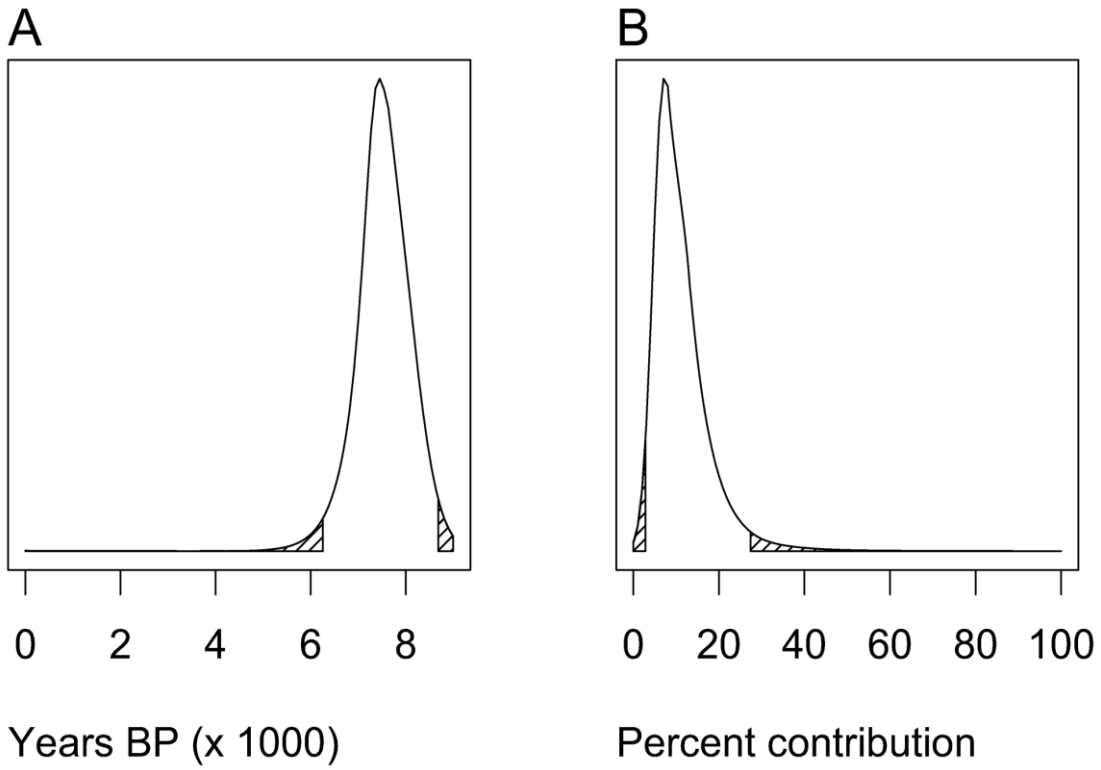

Figure 5.12. Estimates of the date of origin for LP / dairying coevolution and the contribution of people living in the deme of origin for LP / dairying co-evolution, and its eight surrounding demes, to the modern European gene pool. Although not parameters of the model *sensu stricto*, estimates were calculated as with all model parameters by using ABC with regression adjustment and weighting, following acceptance at the 0.5% tolerance level (Beaumont et al., 2002). The date of origin for LP / dairying coevolution (A) is given in thousands of years before present, and the contribution of people living in the deme of origin for LP / dairying co-evolution, and its 8 surrounding demes, to the modern European gene pool (B) is given as a percentage. The upper and lower 2.5% of each distribution are shaded.

The genetic contribution will, to a large extent, be determined by the start location of LP-dairying gene-culture co-evolution. For example, if this process started in Anatolia or the Greek peninsula then we would expect the people living in that region to make a greater contribution to overall European ancestry than if it started in Northwest Europe. With respect to LP a more pertinent question is: Does the advent of LP-dairying coevolution increase the genetic contribution of people living in a particular region to the modern European gene pool? To investigate this, two extra sets of 5,000 simulations were performed, each by picking parameter values at random from the marginal

posterior distributions obtained above. Each set of 5,000 simulations was run with identical sets of parameter value combinations except that in one set the level of selection acting on the LP allele was fixed to zero. Then the distributions of genetic contribution (of people living in and around the LP-dairying start deme to the modern European genepool) were compared with and without selection acting. It was surprising to find that the two distributions are nearly identical (see Figure 5.13).

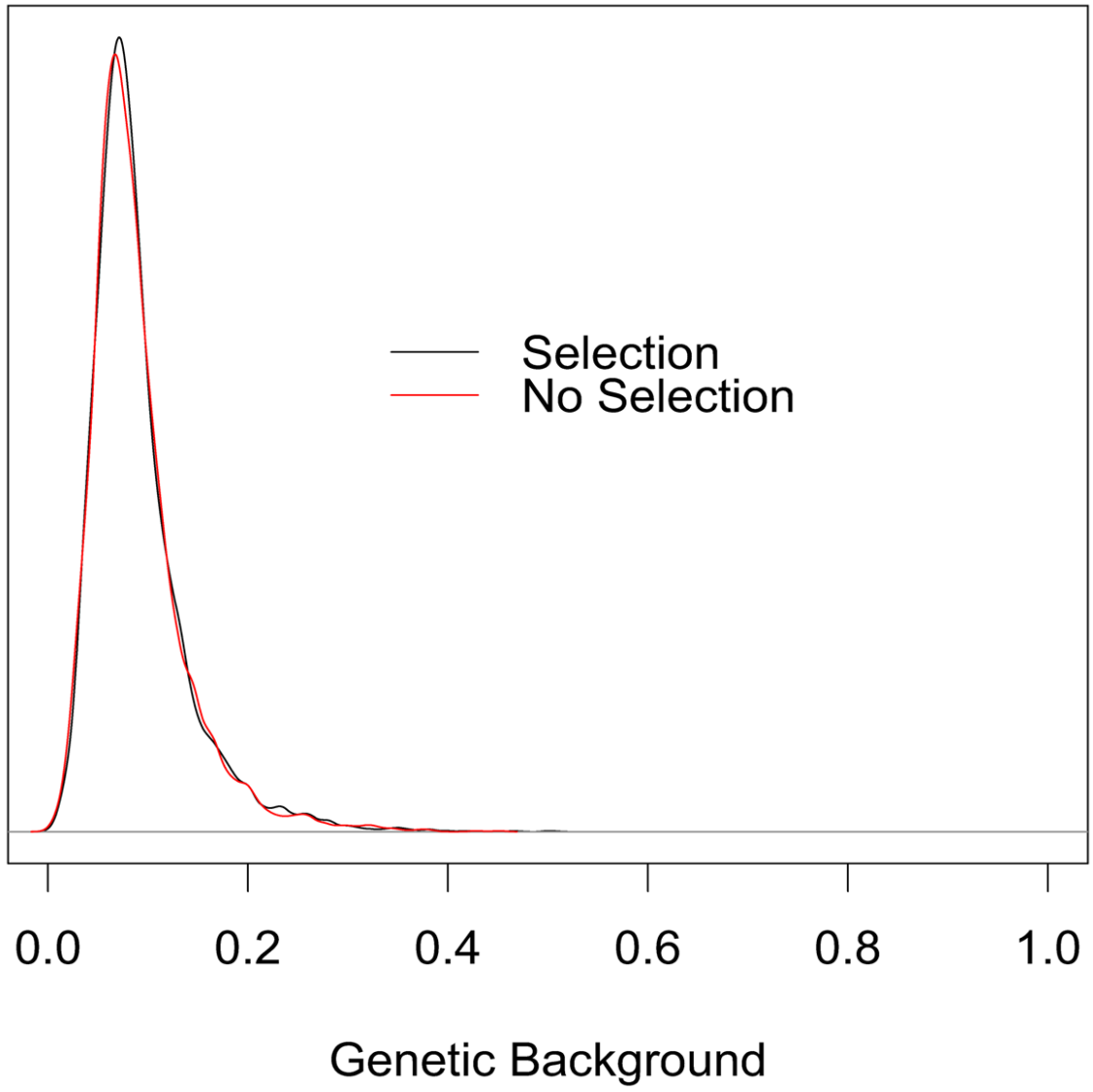

Figure 5.13. Contribution of people living in the deme of origin for LP / dairying co-evolution, and its 8 surrounding demes, to the modern European gene pool with and without selection on LP. Value distributions were taken from 5,000 simulations assuming selection (black line), and 5,000 simulations assuming no selection (red line). Simulation parameter values were sampled at random from the marginal posterior density estimates presented in Figure 5.9 and were identical for each set of 5,000 simulations, except that in the 'no selection' set the selection acting on the LP allele in dairyers parameter was set to zero.

Performance of model in explaining observed data: To explore the power of the model to explain the two data sets considered (13,910*T allele frequency at 12 European locations and farming arrival date at 11 European locations), the following for each data type and at each location considered were plotted: (1) the observed value, (2) the distribution of values from simulations accepted at the 0.5% tolerance level, and (3) the distribution of values from all simulations in which the 13,910*T allele arose and did not go extinct (see Figures 5.14 and 5.15). Although it will necessarily be the case that the 0.5% closest points will be nearer to the observed summary statistics than those simulated from the prior, it is still possible that an observed value will be an outlier from the distribution of simulated points, possibly indicating poor fit of the model. However, as can be seen from Figure 5.15, simulations accepted at the 0.5% tolerance level generate narrow ranges of expectations for the farming arrival date, in very good accordance with the observed (target) values. This can be taken to indicate that with the ABC-estimated parameter values, the model explains the farming arrival dates very well. When considering the 13,910*T allele frequency at the 12 European locations for which data was available (Figure 5.14) it is notable that the observed (target) values are within the 95% equal tail probability interval of expectations generated from simulations accepted at the 0.5% tolerance level. However, a number of the target values are somewhat offset from the expectation modes. In particular, it is notable that for northern European locations the observed frequency is lower than the mode of the expected values and the opposite is the case for southern European locations.

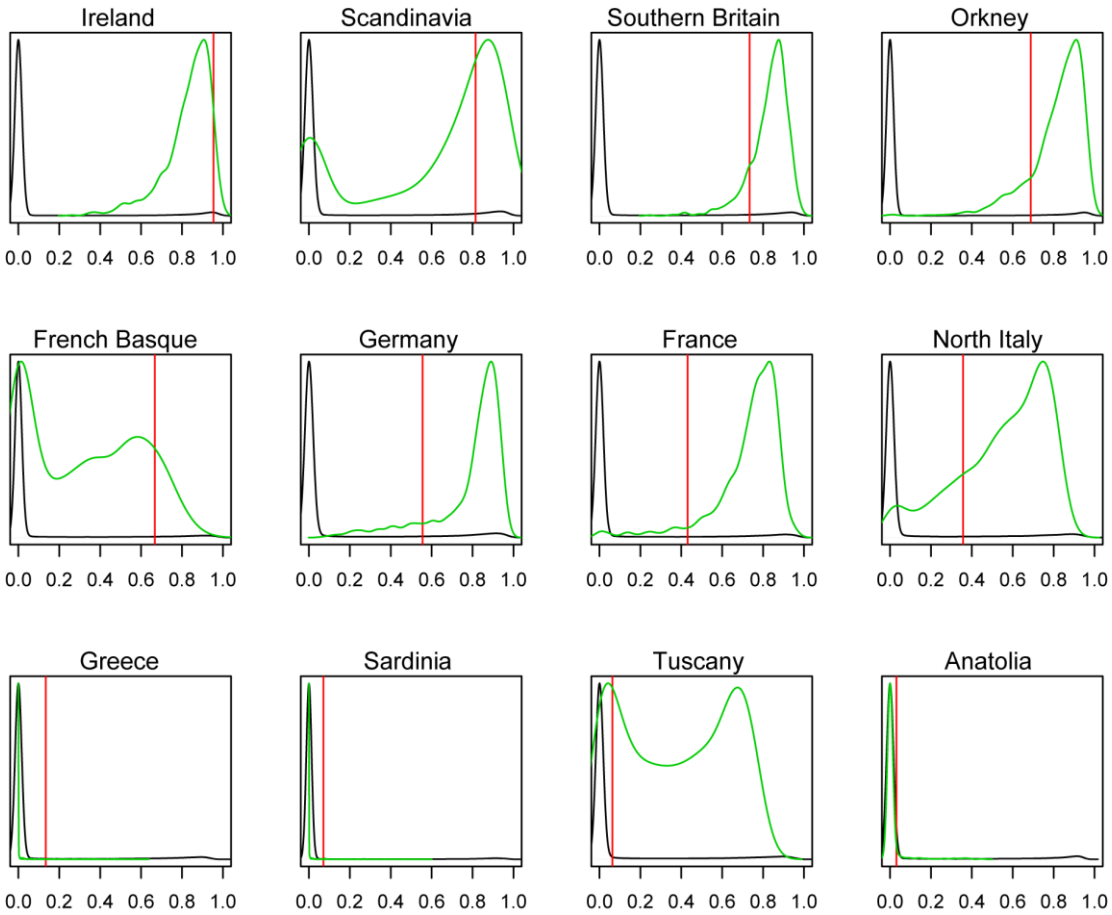

Figure 5.14. Performance of model in explaining observed data on −13,910*T allele frequency at 12 locations throughout Europe. The observed point values are indicated by vertical red lines. The distributions of expected values from all simulations in which the 13,910*T allele arose and did not go extinct are indicated by black lines. The distributions of expected values from all simulations accepted at the 0.5% tolerance level in ABC analysis are indicated by green lines.

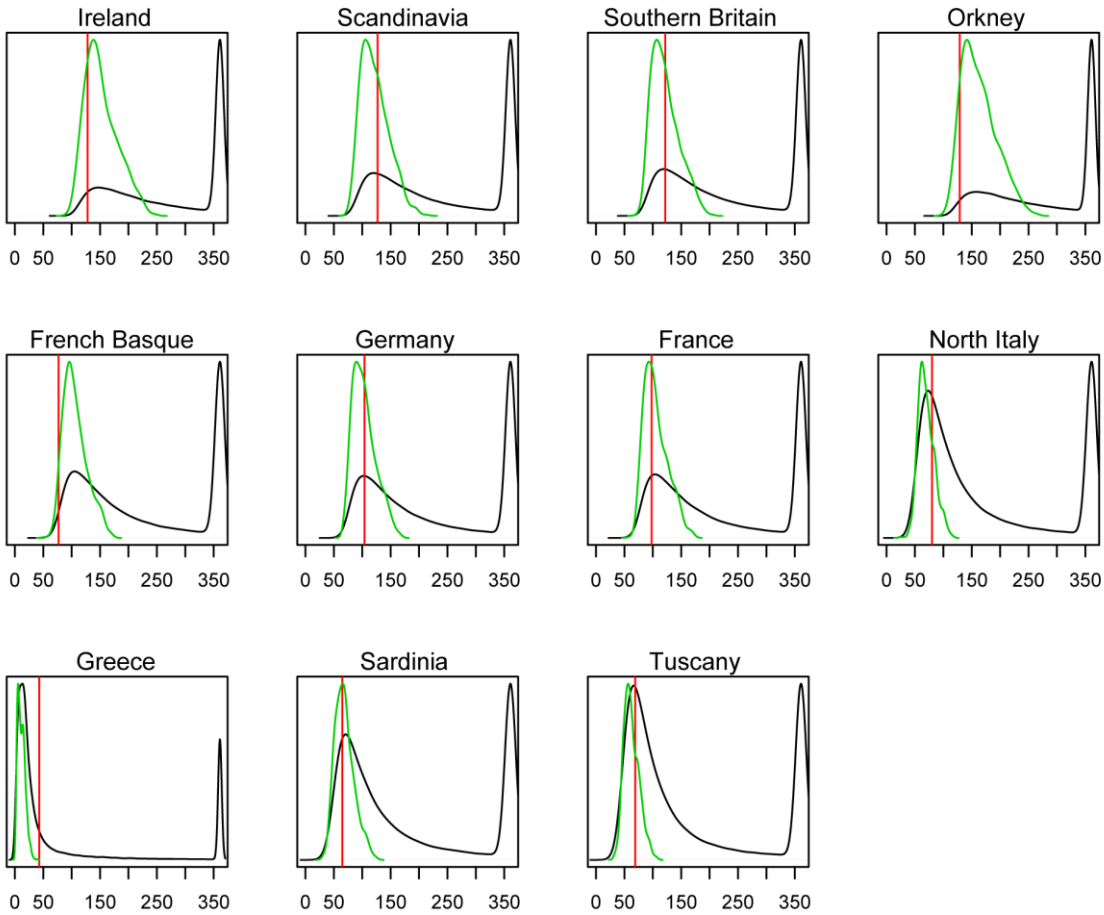

Figure 5.15. Performance of model in explaining observed data on the estimated time of arrival of farming at 11 locations throughout Europe. The observed point values are indicated by vertical red lines. The distributions of expected values from all simulations in which the 13,910*T allele arose and did not go extinct are indicated by black lines. The distributions of expected values from all simulations accepted at the 0.5% tolerance level in ABC analysis are indicated by green lines.

5.4. Discussion.

The simulation model I have employed here is relatively complex compared to related human demographic / evolutionary models reported (Barbujani et al., 1995, Ray et al., 2003, Excoffier, 2004, Currat and Excoffier, 2005). The inclusion of a selected allele and three distinct but interbreeding cultural groups is necessary for the type of questions addressed in this study. But the inclusion of four parameters related to sporadic migration activity, namely the proportion of individuals available for sporadic long-distance migration and the sporadic mobility of each of the 3 cultural groups (modeled separately as a Gaussian random walk process) both allows to tackle the problem of migration overseas and adds, in my view, an extra level of realism to the model. However, as with any simulation model of population history, many simplifying

assumptions have to be made and the extent to which these assumptions may lead to erroneous conclusions remains unknown. For example, I have not considered the 'reverse-cause' hypothesis (Nei and Saitou, 1986, Bayless et al., 1971, Simoons, 1970, McCracken, 1971a) – which proposes that dairying first arose in populations that were already LP – because both ancient DNA evidence (Burger et al., 2007) and data from lipid residues on pots (Evershed et al., 2008) are inconsistent with this view. However, this does not mean that once LP-dairying gene-culture coevolution was established, conversion to the culture of dairying was more likely in high LP frequency populations. Such a process is captured in the model to an extent, in that 'cultural' conversion is determined by the frequency of the receiving cultural group (see equation 4), and LP is unlikely to rise to high frequencies anywhere without the presence of dairying. Nonetheless, a more explicit treatment of this process may lead to different conclusions. Some parameters, such as those relating to the effects of climate zone / elevation, and the logistic growth rate, are fixed based on realistic assumptions (Bellwood, 2005, Colledge et al., 2004, Hassan, 1981). For those parameters that are allowed to vary within a range I note that an important shortcoming is that in any single simulation their value is constant over the 360-generation duration of the run. This may be a particular issue for selection acting on an LP allele in F_d (see below). Since 'good' simulations are identified by using their fit to only two data sets (arrival time of farming and LP allele frequency, both at a range of geographic locations) it is unsurprising that the analysis is relatively uninformative for some parameters. However, inclusion of these parameters does serve to reflect uncertainty in their values.

Estimates of the arrival dates for farming the 11 locations considered here were calculated as local weighted averages of calibrated carbon-14 dates (Pinhasi et al., 2005) from a Gaussian sampling region (also see Figure 5.1). The standard deviation of this region was set at the average nearest neighbour distance to ensure that most of the carbon-14 data was used. However, the geographic density of carbon-14 dates is highly uneven across Europe and so the number of such dates that are informative for farming arrival time at any of the 11 locations will vary. Also, there appears to be a considerable amount of noise in the dates for the first farmers. For example, the earliest carbon-14 date for farming in Ireland predates those for Great Britain, the Low Countries and Denmark. To test if these concerns had a major effect on the results, the simulation date was reanalysed by setting the target farming arrival dates as those inferred by assuming

a constant rate of spread of farming (estimated at 0.9 km/year (Pinhasi et al., 2005)) and calculating the great circle distance from Anatolia to each sampling location. The results of this reanalysis were very similar to those presented above (see Figure 5.16).

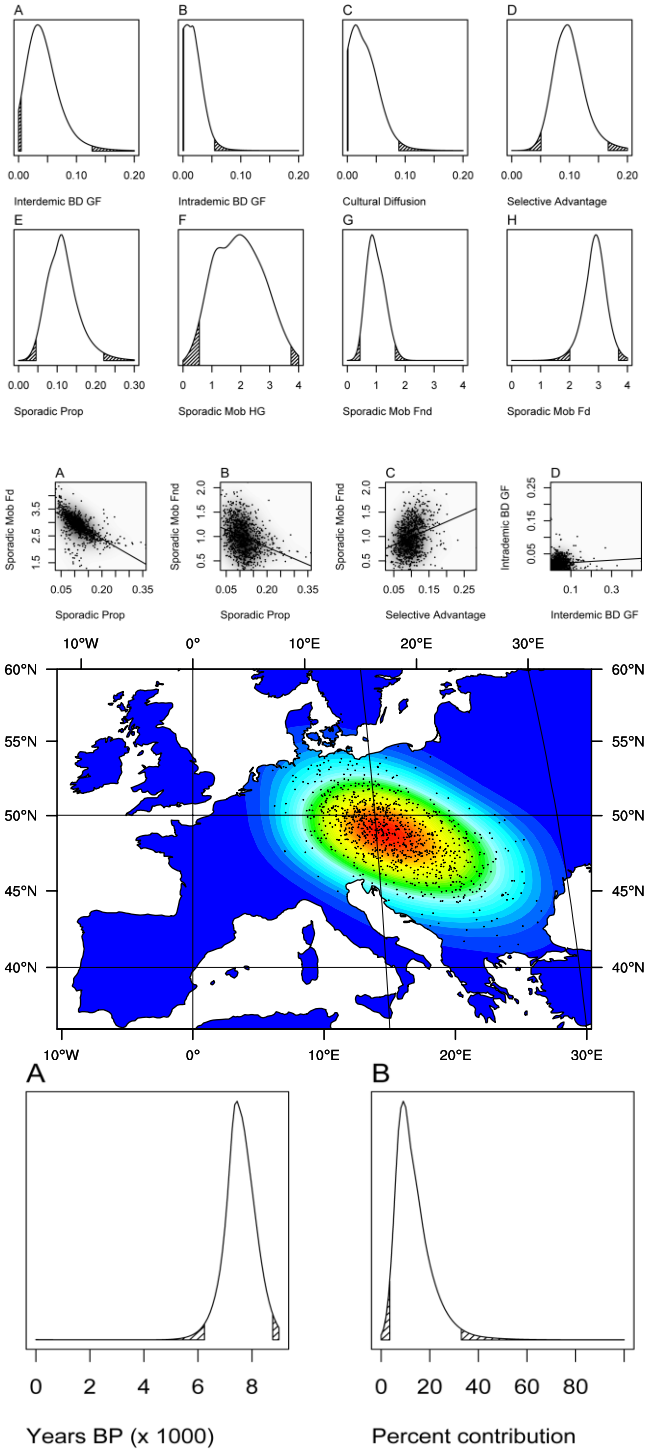

Figure 5.16. Reanalyses Images. Equivalent to Figures 5.9-5.12 from top to bottom ,respectively, reanalysed by setting the target farming arrival dates as those inferred by assuming a constant rate of spread of farming (estimated at 0.9 km/year (Pinhasi et al., 2005)) and calculating the great circle distance from Anatolia to each sampling location.

I am well aware that the spread of the Neolithic over Europe was not as constant as the model assumes. After the arrival of the Neolithic in the Balkans, there is a pause of approximately 800 years before it starts to spread to Central Europe, and there is another pause of 1,000 years before it spreads further into the northern German lowlands and other parts of the northern Europe. Clearly, the carbon-14 dates used to estimate the farming arrival times will not fully reflect the complex history of neolithisation in all parts of the continent.

The list of parameters for which the marginal posterior distributions are notably narrower than their corresponding prior ranges (selective advantage, intrademic gene flow, the sporadic migration distance of F_d and F_{nd}, and the geographic origin location of LP / dairying co-evolution) – which I interpret as those parameters for which the analysis is informative – is an unsurprising one since we would expect these parameters to have the greatest influence on the spread of an LP allele and farming in Europe. Likewise, it is unsurprising that the proportion available for sporadic migration and the sporadic mobility of (a) dairying farmers, and (b) non-dairying farmers are both strongly negatively correlated (Figure 5.10A and 5.10B) since we would expect these parameters to be confounded in influencing the arrival time of farmers at different locations.

The estimated selective advantage conferred by a LP allele (mode = 0.0953; 95% CI = 0.0518 - 0.159) is in good agreement with previous estimates for Europeans (0.014 - 0.15 (Bersaglieri et al., 2004)). However, it should be noted that (1) this estimate is for selection only in dairying farmers, who make up just under half of the population that is simulated, and (2) it is assumed that selection is constant over time. It is possible that selection favouring LP has in fact been episodic and possibly spatially structured in different climate zones (Flatz and Rotthauwe, 1973, Beja-Pereira et al., 2003, Simoons, 1980, Simoons, 1978, Bloom and Sherman, 2005, Simoons, 2001). Episodic selection would be difficult to model without additional information on when those episodes were likely to have occurred. But I reason that constant selection strength is a more parsimonious assumption in the absence of evidence to the contrary. If, as modelled here, dairying farmers made up less than half of the European post-Neolithic population then we would expect the real continent-wide selection values for LP to average less

than half of what estimated here. Such a range of selection values are, however, still consistent with previous estimates based on haplotype decay (Bersaglieri et al., 2004).

Perhaps the most interesting result presented here is the estimation of the geographic and temporal origins of LP-dairying co-evolution. The highest posterior probabilities were found for a region between the central Balkans and central Europe (see Figure 5.11). At first sight such a location of origin may seem counter intuitive since it is far-removed from Northwest Europe, where the *-13,910*T* allele is found at highest frequency. However, previous simulations have shown that the geographic centroid of allele can be offset from its location of origin, particularly when it occurs on the wave front of a demographic expansion (Klopfstein et al., 2006, Edmonds et al., 2004). The lactase-dairying coevolution origin region inferred here is consistent with a number of archaeologically attested patterns concerning the emergence and spread of dairying. Recent carbon isotope ratios from lipids extracted from archaeological sherds show the presence of milk fats in present-day western Turkey and connect these findings to an increased importance of cattle herding (Röhrs and Herre, 1961, Boessneck and Driesch, 1979, Buttenhuis, 1995, Benecke, 1998, Evershed et al., 2008). In general, the spread of the Neolithic lifestyle from the Aegean to Central Europe goes hand in hand with the decline of the importance of sheep and goat and the rise in frequency of cattle bones in archaeological assemblages. While the Balkans at the beginning of the Neolithic still shows a variety of subsistence strategies (Bartosiewicz, 2005), the middle Neolithic in SE-Europe and the earliest Neolithic in Central Europe after 7,500 BP show a clear preponderance of cattle. Benecke (Benecke, 1994b) gives the following averaged rates for the respective domestic species: cattle 55.2%, sheep and goat 32.6%, pig 12%. The proportion of cattle in Central Europe increases during the following centuries to an average of 73% and then stays (with a few exceptions) stable for most prehistoric periods of Middle and northern Europe. Thereby, cattle herding is in most cases connected with kill-of profiles indicative for dairying (Arbogast, 1994, Balasse and Tresset, 2002, Tresset, 1996, Tresset, 1997, Benecke, 1994a, Benecke, 1994b, Bartosiewicz, 2007). Milk consumption and dairying have been proposed to be as early as the Pre-Pottery Neolithic B of the Near East and may even be a reason for domestication (Cribb, 1987). Without doubt, it was a common cultural practice during all phases and regions of the European Neolithic, especially for goat and cattle. However, a fully developed dairying-based farming economy emerges first during the late Neolithic in Southeast Europe and the Middle Neolithic Cultures following the

Linearbandkeramik (LBK) in Central Europe, and is connected mainly to cattle and partly also to goat (for the Rössen culture see (Benecke, 1994b, Benecke, 1994a)). In the Mediterranean, milking of cattle occurs episodically (Vigne, 2006) and sheep and goat remain the dominant domestics, as they were earlier in Anatolia and the Aegean. It is very likely that the goat and sheep, and to a lesser extent cattle, based economies of the Mediterranean used processed milk in the form of yoghurt, cheese and other milk-derived products instead of fresh milk. The nutritional and agricultural differences between southern Europe, the Mediterranean and central and northern Europe, as well as historic reports, point to this. For instance, the Romans used goat and sheep milk for the production of cheese, and cattle as a draught animal. In contrast the Germanic peoples and other inhabitants of central and northern Europe practised cattle dairying and drank fresh milk in significant amounts. Strabo reports in his Geography (Strabo, 1969): "Their [sc. "the men of Britain"] habits are in part like those of the Celti, but in part more simple and barbaric - so much so that, on account of their inexperience, some of them, although well supplied with milk, make no cheese; and they have no experience in gardening or other agricultural pursuits."

Overall, by considering the results from the simulations and archaeological, archaeozoological, and archaeometric findings, it seems very plausible to connect the geographic origin of the spread of LP to the increasing emergence of a cattle-based dairying economy during the 6th millennium BC. The geographic region of origin of the LBK – in modern day Northwest Hungary and Southwest Slovakia (Pavúk, 2005, Bánffy, 2004) – certainly correlates well with the results (see Figure 5.17). The date of origin of LP-dairying coevolution estimated here (mode = 7,441 years BP; 95% CI = 6,256 to 8,683 years BP; see Figure 5.12A and Table 5.2) also fits well with dates for the early LBK in Central Europe (~7,500 years BP) and its proposed main predecessor, the Starčevo culture of the northern Balkan Peninsula and south of Lake Balaton (8,100 to 7,500 years BP; (Baldia, 2003)). However, as explained above, the date estimate is conditioned by farming arrival dates in the estimated LP-dairying coevolution origin region. As a result, the date and location estimates are not independently derived. Nonetheless, a role for LP-dairying coevolution in the later rapid spread of LBK culture – from its origins in the Carpathian Basin – into central and Northwest Europe would be consistent with the significantly higher sporadic migration distances inferred for of F_d when compared to F_{nd}. This is also consistent with the rapid dissemination of the LBK

culture over a territory of 2,000 km width and approximately one million square kilometres within less than 500 years (Lüning, 2005).

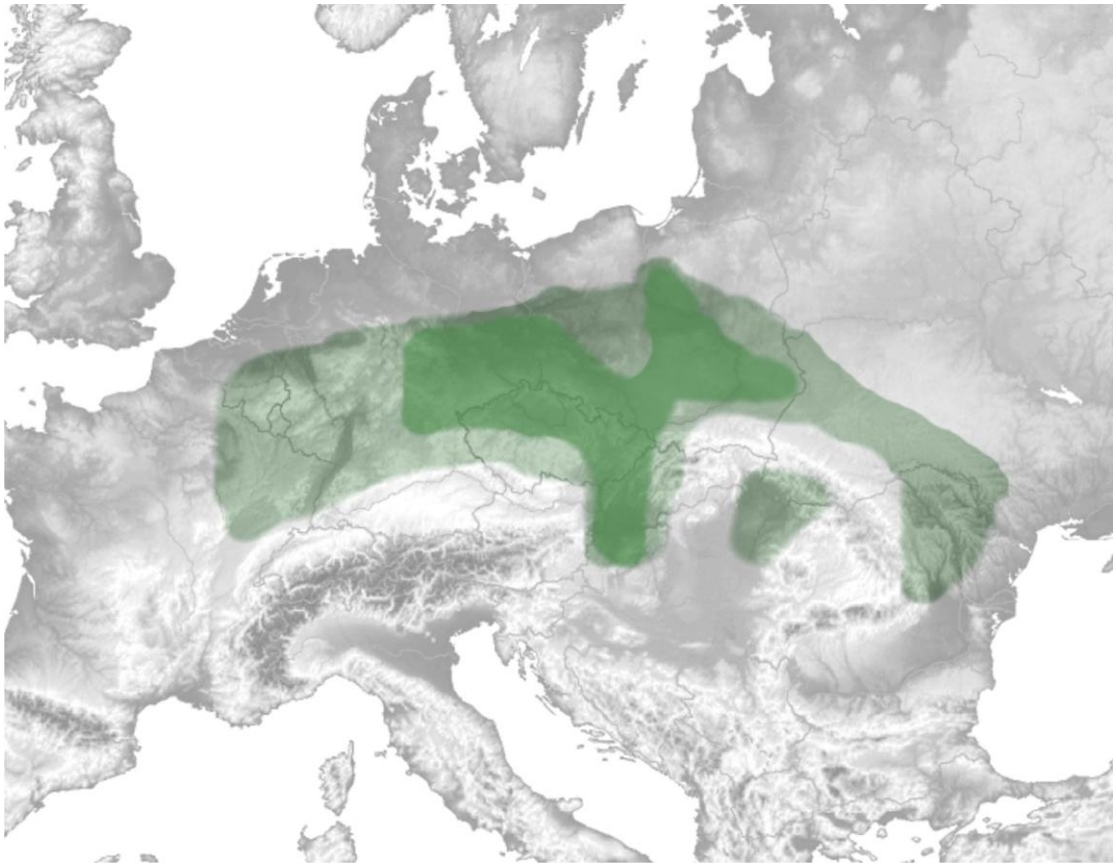

Figure 5.17. Main regions of the spread of the Linearbandkeramik culture from its origins in modern day northwest Hungary and southwest Slovakia. Early phase is in dark green and late phase is in light green.

Contrary to my expectations, I did not find that the presence of a positively selected LP allele in early dairying groups increases the unlinked genetic contribution of people living in the region where LP-dairying coevolution started to the modern European gene pool, when using demographic parameter values estimated here. The main reason for this is likely to be the relatively high inferred rates of intra- and interdemic gene flow between dairying and non-dairying farmers and between neighbouring demes, respectively, leading to a rapid erosion of any demographic 'hitchhiking' of unlinked genomic regions. Additionally, the simulation tracked only the genetic contribution of people living in and around the deme of LP / dairying coevolution from the inception of this process. Since it takes some time for the LP allele to rise to appreciable frequencies, any demographic 'hitchhiking' effect may become important only after the allele centroid has moved some distance away from its origin deme.

Another notable result was obtained when comparing the range of expected 13,910*T allele frequencies at different European locations – from simulations accepted at the 0.5% tolerance level – to those observed. While all observed values were within the 95% equal tail probability interval of the simulated values, many were somewhat offset from the modes. This could indicate that the simulation model does not fully explain the distribution of the 13,910*T allele in Europe. One possible explanation for this is that migration activity – as modeled here by interdemic gene flow and sporadic unidirectional migration – has increased subsequent to the expansion of farming into the northwestern reaches of Europe. In this scenario the farming expansion phase, occurring 9,000 to 5,500 years BP, would be mainly responsible for generating the 13,910*T allele frequency cline in Europe but higher migration activity following this period would then have a homogenizing effect in LP allele frequencies. Intriguingly, a general pattern can be seen (Figure 5.14) whereby observed frequencies are lower than expected in northern Europe and higher than expected in southern Europe. Such a pattern is the opposite of what we would expect if selection for LP was higher in northern latitudes through a greater requirement for dietary vitamin D and calcium because low-sunlight conditions reduce UV-mediated vitamin D production in the skin (Flatz and Rotthauwe, 1973). This frequently cited mechanism (Simoons, 1980, Simoons, 1978, Simoons, 2001, Weiss, 2004, Hollox et al., 2001, Akey et al., 2004, Ingram et al., 2009a) was not included in the model and thus would seem to have negative explanatory power. Thus the simulations indicate that geographically and temporally homogeneous selection in combination with well-attested underlying demographic processes are sufficient to explain, indeed, to over-explain, the LP / latitude correlation in Europe. However, it should be noted that since a parameterised latitudinal effect on selection was not explicitly included in the model, there may be scenarios where such an effect could also explain patterns of LP in Europe.

As inferred here, the spread of a LP allele in Europe was shaped not only by selection but also by underlying demographic processes; in this case the spread of farmers from the Balkans into the rest of Europe. I propose that this combination of factors could also explain the apparent homogeneity of LP-associated mutations in Europe. In Africa there are at least four known LP-associated alleles, including three that are likely to be of African origin (Tishkoff et al., 2007, Ingram et al., 2007) as well as *-13,910*T*, which is likely to be of European origin (Mulcare et al., 2004, Coelho et al., 2005). The greater

apparent diversity of LP-associated mutations in Africa may reflect a greater genetic diversity in general, leading to the availability of more mutations upon which selection can act following the advent of dairying. However, I suggest that this diversity is the result of an 'imposition' of dairying culture on a pre-existing farming people, rather than the spread of dairying being tied to the spread of dairyers. Such a model would require the availability of a number of, albeit low-frequency, LP-causing mutations; either through a high mutation rate or a large number of potential LP-causing sites. It is therefore possible that, in the absence of the spread of dairying being linked to a major demographic expansion, high LP-allele diversity will also be found in the Indian subcontinent.

The model used does not accommodate all data (both genetic and archaeological) that is potentially informative on the coevolution of LP and dairying in Europe. Future improvements can be made by adding more 'realism' to the model and by increasing the number of data types that are used in the ABC analysis, leading to more integrative inference. The former should include both adding more fixed parameter information (such as the effects of past vegetation, climate variation and other geographic features on migration parameters and carrying capacities (Özdogan and Basgelen, 1999, Cavalli-Sforza et al., 1994, Özdogan, 2007)) and estimating currently fixed parameters such as the ratio of dairying to non-dairying farmers. The latter could be achieved by writing the simulation model so that it generates expectations for other data types. For example, including the movement of domestic cattle could be used to generate expectations on patterns of ancient and modern cattle genetic diversity, for which considerable data is available (Troy et al., 2001, Bollongino et al., 2006, Edwards et al., 2007, Achilli et al., 2008, Achilli et al., 2009). For an extra level of realism I also suggest applying a more accurate model for human population grown, based on the logistic equation that increases its carrying capacity according to advance in technology (Marchetti et al., 1996). The proportion of vitamin D and lactose consumption in different cultures could also be applied as an extra level of realism in future simulations. Finally, it should be possible to extend the approach that was used here to study the evolution of LP and dairying in other parts of the world.

I infer that the coevolution of European LP and dairying originated in a region between central Europe and the northern Balkans around 6,256 to 8,683 years BP. I propose the

following scenario: after the arrival of the Neolithic in south-eastern Europe and the increasing importance of cattle herding and dairying, natural selection started to act on a few LP individuals of the early Neolithic cultures of the northern Balkans. After the initial slow increase of LP frequency in those populations and the onset of the Central European LBK culture around 7,500 BP, LP frequencies rose more rapidly in a gene-culture co-evolutionary process and on the wave front of a demographic expansion, leading to the establishment of highly developed cattle- (and partly also goat-) based dairying economies during the Middle Neolithic of central Europe around 6,500 BP. A latitudinal effect on selection for LP, through an increased requirement for dietary vitamin D (Flatz and Rotthauwe, 1973), is unnecessary to explain the high frequencies found in northern Europe.

6. Discussion.

When planning and performing the different parts of my PhD studies that are presented in the previous chapters I had one major goal: to understand the role of various genomic events in human evolution. I was interested in large-scale understanding of processes and in a top-down approach, and this reflected in the nature of the four different studies that I have performed: detecting all human lineage gene duplications (chapter 2), estimating all human lineage gene duplications' date and function (chapter 3), worldwide interpolating and correlation of lactase persistence genotypes and phenotype (chapter 4), and simulating the origins and demography of lactase persistence in Europe (chapter 5). I am well aware that due to the "large-scale" interests and the time constraint of 4 years, I will have overlooked or chosen not to tackle several related issues. However, I attempted to suggest these as subjects for future studies in the relevant chapters. That is not to say that I did not attempt to consider the "small details" that build the studied mechanisms – I believe that all chapters demonstrate attempts to make a comprehensive understanding of the various evolutionary processes, with a careful consideration of the trade-off (especially in simulations) between realism (i.e. parameter heavy) and computation time / results analyses options. In all four studies that I have performed I have dealt with large data sets and enjoyed the challenges of estimating missing data by different methods.

In chapter 2 I attempted to identify all human inparalogues, and developed a systematic method to tackle several major problems that were prevalent in the previous methods that have attempted to achieve the same aim in the past. The main result of this chapter is an algorithm that I believe is the most robust one that is available today for detecting species inparalogues in cases where one of the genomes used is of a non-model organism. Consequently I consider that set of human inparalogues that I have detected with this method as the most robust and most comprehensive that is available today. When I planned the human inparalogues detecting project I expected that it would be a relatively straightforward process, as both human and chimpanzee proteomes had just become available, and there was a well-established method (InParanoid) that could automatically identify inparalogues given two species' proteome sets. The results that I obtained on first time using InParanoid with the human and chimpanzee proteomes seemed very exciting and even sensational – the human genome underwent 6 times

more duplications than the chimpanzee's genome. However, further examination of the result taught me an important lesson – sensational results are likely (but not always) to be a result of various biases – in chapter 2 I present the full range of problems that went undetected in all previous studies, which brings me to my second major insight – well established methods and studies may contain errors and should be individually tested to better understand the method and data that being used.

In chapter 3 I estimated the dates of the human inparalogues duplication events, their function, and attempted to check whether the dates are clustered or if they are, as expected, randomly distributed (the null hypothesis). I found that the dates of the duplication events are clustered, and I believe that the main issue in such a clustering is how one defines a cluster. For example – clusters can be identified by pre-determining the number of clusters or by setting a maximum radius for a cluster (and in this case there is the question of what is a sensible radius – one that show some statistical degree for clustering, or rather one that empirically represent a meaningful human evolution time unit). I have clustered the dates in two different approaches, and presented the distribution of the different functions within these clusters. I found that there was a burst of gene duplications in anatomically modern human. The most recent time window (between 500,000 years ago and present) is also the one that contains the largest number of human gene duplication – 27 gene duplications where, for comparison, the expected number would be 10.62 and the second largest burst of duplication (between 3.5 and 4 million years ago) contains 16 duplications. Interestingly, the most ancient time window (between 6 and 6.6 million years ago) does not contain any gene duplication. Although these results may suggest a bias, I believe that they are reliable since I have taken strict precautions to avoid the counting of dates around both human-chimpanzee divergence time and present date by phylogenetic estimates, molecular clock validation, and removing genes that are suspected to have undergone gene convertion from my dataset. Moreover, I found that different time windows are enriched for genes in different functional classes, and three biological classes are over-represented in the human inparalogues set. For example – the metabolic and catabolic processes function class in enriched for gene duplications, and occurred exclusively in the oldest cluster, within an average estimate of 4.89 million years ago. The biological class with the highest gene enrichment score is the immune system and the second is sensory perception, with the former duplicated throughout the timeline of the human lineage, and the latter appearing

in two "bursts" – a recent and an ancient one. I believe that in this study I present for the first time a large scale correlation between human genome and the palaeoanthropological record, and I hope that these results could provide useful for various future studies that could find more detailed correlation between these functions, their time of duplication, and human fossil record.

In chapter 4 I have applied a population level approach to the association of genotypes and phenotypes at worldwide scale. The idea for this study arose as a result of some collaborative analysis I contributed to Ingram et al. (2009) where I collated human LP phenotype frequencies from all available literature, filtered for reliable frequency estimates, and performed surface interpolation mapping of that data. In the present study I have correlated all known LP-associated alleles with the LP phenotype from populations of the same regions. In this study I have dealt with two major challenges: (1) since there are only data from 120 genotype and 112 reliable phenotype collection locations, I performed surface interpolation for estimating the missing data, and (2) to correlate LP genotypes and phenotype I used a method that was designed for estimating correlation using observed data of genotype, phenotype, sample size, and method error rate. This required an automation process and further interpolation to estimate the missing data. This study can be very useful as a tool for researchers to determine areas for further LP genotype studies, since in places that have high frequencies of lactase persistence but that also have low frequencies of LP-associated alleles, we would expect to discover new LP-associated genotypes. For example, this study suggests that West Africa is a region that seems to be a strong candidate for such further LP genotype studies. I believe that this study could prove very useful for other global genotype-phenotype association studies, such as human drug metabolising enzymes that might be of strong interest to both academy and the pharmaceutical industry.

In chapter 5 I present an integrative simulation modelling-based inference study of the evolution of lactase persistence in Europe. This study utilizes genetic data (the frequencies of the European LP associated -13,910*T allele in different European populations and selective advantage modelling), archaeological data (arrival of farming to different parts in Europe), geographic information (topography and Earth's curvature considerations), anthropology (for estimating the dynamics between different cultural groups) and other information sources. My role in this study was writing the program's

code, developing the model's mathematical equations together with Mark Thomas, collating empirical data for lactase persistence and the arrival of farming, running the simulations and collating the results, and plotting maps. Although I was involved in the ABC analyses and the LBK/archaeological discussion, these subjects were mainly dealt by other collaborators: Adam Powell and Mark Beaumont performed the ABC analyses, while Joachim Burger contributed his LBK/archaeological knowledge. I found that LP and dairying gene-culture coevolution has begun in the Central Europe / Northern Balkans region approximately 7,500 years ago in association with the LBK culture. Moreover, I demonstrated that the calcium assimilation hypothesis (which maintains that milk gave a selective advantage to individuals in northern latitudes because it contains vitamin D which is lacking in places with low sun exposure) is not necessary for explain the current distribution of LP in Europe. I believe that this study presents a good example of interdisciplinary research and could be a platform for other evolutionary or parameter rich studies. I also think that various results of this research could be subjects for future studies, such as the strong correlation between the proportion of dairying and non-dairying farmers' availability to migrate and their respective migration rate. Future studies could also consider using the extra -13,910*T data points in Europe that I have presented in chapter 4, and possibly fix certain parameters that have shown a narrow distribution (such as non-dairying farmers migration rate) to allow introduction of new parameters without significantly increasing the complexity and computation time of this simulation model. It would be interesting to show how lactase persistence evolved in other parts of the world, probably via convergent evolution and different demographic dynamic in different pastoralist population in Africa and Asia.

I believe that the different research approaches that I have presented in these four studies could potentially be combined. For example – one could focus on the oldest human inparalogues cluster dated 4.89mya and where the dominant function is metabolic and catabolic processes, and simulate a scenario of the evolution of nutrition in early hominids as a result of the beginning of transition of the African climate into a drier one, and as a result the change of the terrain from jungles into savannas and consequently the change in food resources and the selective advantage that mutations allowing digestion of the new foods were likely to have had given. Another option is to collate human CNV data in the same way that I collated the human LP phenotype data,

and integrate it in the algorithm of human inparalogues detection, possibly in a project that aims to find inparalogues that are unique to different human populations and their estimated duplication date and correlated biological function.

In summary, I have presented in this work four studies of genomic events that have contributed to the human phenotype, following different approaches and methodologies. I hope that they have made a significant contribution to knowledge on human evolution.

References.

ABRAHAMS, B. S. & GESCHWIND, D. H. (2008) Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet,* 9**,** 341-55.

ACHILLI, A., BONFIGLIO, S., OLIVIERI, A., MALUSA, A., PALA, M., KASHANI, B. H., PEREGO, U. A., AJMONE-MARSAN, P., LIOTTA, L., SEMINO, O., BANDELT, H. J., FERRETTI, L. & TORRONI, A. (2009) The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PLoS One,* 4**,** e5753.

ACHILLI, A., OLIVIERI, A., PELLECCHIA, M., UBOLDI, C., COLLI, L., AL-ZAHERY, N., ACCETTURO, M., PALA, M., KASHANI, B. H., PEREGO, U. A., BATTAGLIA, V., FORNARINO, S., KALAMATI, J., HOUSHMAND, M., NEGRINI, R., SEMINO, O., RICHARDS, M., MACAULAY, V., FERRETTI, L., BANDELT, H. J., AJMONE-MARSAN, P. & TORRONI, A. (2008) Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr Biol,* 18**,** R157-8.

AKEY, J. M., EBERLE, M. A., RIEDER, M. J., CARLSON, C. S., SHRIVER, M. D., NICKERSON, D. A. & KRUGLYAK, L. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol,* 2**,** e286.

ALDENDERFER, M. & MASCHNER, H. D. G. (1996) *Anthropology, Space, and Geographic Information Systems*, Oxford University Press, USA.

ALMON, R., ENGFELDT, P., TYSK, C., SJOSTROM, M. & NILSSON, T. K. (2007) Prevalence and trends in adult-type hypolactasia in different age cohorts in Central Sweden diagnosed by genotyping for the adult-type hypolactasia-linked LCT -13910C > T mutation. *Scand J Gastroenterol,* 42**,** 165-70.

ALTENHOFF, A. M. & DESSIMOZ, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol,* 5**,** e1000262.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

AMARAL, D. G., SCHUMANN, C. M. & NORDAHL, C. W. (2008) Neuroanatomy of autism. *Trends Neurosci,* 31**,** 137-45.

ANGATA, T. & VARKI, A. (2002) Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chem Rev,* 102**,** 439-69.

AOKI, K. (1986) A stochastic model of gene-culture coevolution suggested by the "culture historical hypothesis" for the evolution of adult lactose absorption in humans. *Proc Natl Acad Sci U S A,* 83**,** 2929-33.

ARBOGAST, R. M. (1994) Premiers élevages néolithqiues du Nord-Est de la France. Liège, Ètudes et Rech.Arch.Univ.Liège.

ASFAW, B., GILBERT, W. H., BEYENE, Y., HART, W. K., RENNE, P. R., WOLDEGABRIEL, G., VRBA, E. S. & WHITE, T. D. (2002) Remains of Homo erectus from Bouri, Middle Awash, Ethiopia. *Nature,* 416**,** 317-20.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

AVNER, P. & HEARD, E. (2001) X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet,* 2**,** 59-67.

AYALA, F. J. (1999) Molecular clock mirages. *Bioessays,* 21**,** 71-5.

BAILEY, J. A. & EICHLER, E. E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet,* 7**,** 552-64.

BAILEY, J. A., GU, Z., CLARK, R. A., REINERT, K., SAMONTE, R. V., SCHWARTZ, S., ADAMS, M. D., MYERS, E. W., LI, P. W. & EICHLER, E. E. (2002) Recent segmental duplications in the human genome. *Science,* 297**,** 1003-7.

BAILEY, J. A., YAVOR, A. M., MASSA, H. F., TRASK, B. J. & EICHLER, E. E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res,* 11**,** 1005-17.

BAIROCH, A., APWEILER, R., WU, C. H., BARKER, W. C., BOECKMANN, B., FERRO, S., GASTEIGER, E., HUANG, H., LOPEZ, R., MAGRANE, M., MARTIN, M. J., NATALE, D. A., O'DONOVAN, C., REDASCHI, N. & YEH, L. S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res,* 33**,** D154-9.

BALASSE, M. & TRESSET, A. (2002) Early weaning of Neolithic domestic cattle (Bercy, France) revealed by intra-tooth variation in nitrogen isotope ratios. *Journal of Archaeological Science,* 29**,** 853-859.

BALDIA, M. O. (2003) Breaking unnatural barriers: Comparative archaeology, climate, and culture change in Central and Northern Europe (6000 - 2000 BC). Paper presented in the Session "Comparative archeology and paleoclimatology: sociocultural responses to a changing world" under the theme "Past human environments in modern contexts" at the Fifth World Archaeology Congress, June 23, 2003, Washington DC, USA. *Fifth World Archaeology Congress.* Washington DC, USA.

BÁNFFY, E. (2004) The 6th Millennium BC boundary in Western Transdanubia and its role in the Central European Neolithic transition (The Szentgyörgyvölgy-Pityerdomb Settlement). Budapest, Varia Arch. Hungarica.

BARBUJANI, G., SOKAL, R. R. & ODEN, N. L. (1995) Indo-European origins: a computer-simulation test of five hypotheses. *Am J Phys Anthropol,* 96**,** 109-32.

BARNARD, A. (2004) *Hunter-Gatherers in History, Archaeology and Anthropology,* Berg Publishers.

BARQUET, N. & DOMINGO, P. (1997) Smallpox: the triumph over the most terrible of the ministers of death. *Ann Intern Med,* 127**,** 635-42.

BARTOSIEWICZ, L. (2005) Animals, environment, and culture in the Neolithic of the Carpathian Basin and adjacent areas. IN BAILEY, D., WHITTLE, A. & CUMMINGS, V. (Eds.) *(un)settling the Neolithic.* Oxbow 2005.

BARTOSIEWICZ, L. (2007) Mammalian Bone. IN WHITTLE, A. (Ed.) *The early Neolithic on the Great Hungarian plain. Investigations of the Körös culture site of Ecsegfalva 23, County Békés.* Budapest.

BAYLESS, T. M., PAIGE, D. M. & FERRY, G. D. (1971) Lactose intolerance and milk drinking habits. *Gastroenterology,* 60**,** 605-8.

BEAUMONT, M. A., ZHANG, W. & BALDING, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics,* 162**,** 2025-35.

BEGUN, D. R., WARD, C. V. & ROSE, M. D. (1997) *Function, Phylogeny, and Fossils: Miocene Hominoid Evolution and Adaptations,* Springer.

BEHRENSMEYER, A. K., TODD, N. E., POTTS, R. & MCBRINN, G. E. (1997) Late pliocene faunal turnover in the turkana basin, kenya and ethiopia. *Science,* 278**,** 1589-94.

BEJA-PEREIRA, A., LUIKART, G., ENGLAND, P. R., BRADLEY, D. G., JANN, O. C., BERTORELLE, G., CHAMBERLAIN, A. T., NUNES, T. P., METODIEV,

S., FERRAND, N. & ERHARDT, G. (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nat Genet,* 35**,** 311-3.

BELLWOOD, P. S. (2005) *The first farmers : the origins of agricultural societies,* Malden, MA, Blackwell Pub.

BENECKE, N. (1994a) Archäozoologische Studien zur Entwicklung der Haustierhaltung in Mitteleuropa und Südskandinavien von den Anfängen bis zum ausgehenden Mittelalter. Berlin.

BENECKE, N. (1994b) *Der Mensch und seine Haustiere,* Stuttgart, Theiss.

BENECKE, N. (1998) Animal Remains From the Neolithic and Bronze Age Settlements at Kirklareli (Turkish Thrace). IN BUITENHUIS, H. (Ed.) *Archaeozoology of the Near East III.* Groningen ARC Publicaties.

BENEFIT, B. R. & MCCROSSIN, M. L. (1997) Earliest known Old World monkey skull. *Nature,* 388**,** 368-71.

BENTON, M. J. & DONOGHUE, P. C. (2007) Paleontological evidence to date the tree of life. *Mol Biol Evol,* 24**,** 26-53.

BERGLUND, A. C., SJOLUND, E., OSTLUND, G. & SONNHAMMER, E. L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res,* 36**,** D263-6.

BERNDT, R. M. & BERNDT, C. H. (1994) *The Speaking Land: Myth and Story in Aboriginal Australia,* Inner Traditions.

BERSAGLIERI, T., SABETI, P. C., PATTERSON, N., VANDERPLOEG, T., SCHAFFNER, S. F., DRAKE, J. A., RHODES, M., REICH, D. E. & HIRSCHHORN, J. N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet,* 74**,** 1111-20.

BIRTLE, Z., GOODSTADT, L. & PONTING, C. (2005) Duplication and positive selection among hominin-specific PRAME genes. *BMC Genomics,* 6**,** 120.

BLENCH, R. (2006) *Archaeology, Language, and the African Past,* AltaMira Press, U.S.

BLOOM, G. & SHERMAN, P. (2005) Dairying barriers affect the distribution of lactose malabsorption. *Evolution and Human Behavior,* 26**,** 301.e1–301.e33.

BOESSNECK, J. V. D. & DRIESCH, A. (1979) Die Tierknochenfunde aus der Neolithischen Siedlung auf dem Fikirtepe bei Kadiköy am Marmarameer. Munich, Munich, Inst. für Domestikationsforschung.

BOLLONGINO, R., EDWARDS, C. J., ALT, K. W., BURGER, J. & BRADLEY, D. G. (2006) Early history of European domestic cattle as revealed by ancient DNA. *Biol Lett,* 2**,** 155-9.

BRAWAND, D., WAHLI, W. & KAESSMANN, H. (2008) Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biol,* 6**,** e63.

BRITTEN, R. J. (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci U S A,* 99**,** 13633-5.

BROWN, K. S., MAREAN, C. W., HERRIES, A. I., JACOBS, Z., TRIBOLO, C., BRAUN, D., ROBERTS, D. L., MEYER, M. C. & BERNATCHEZ, J. (2009) Fire as an engineering tool of early modern humans. *Science,* 325**,** 859-62.

BROWNER, W. S., KAHN, A. J., ZIV, E., REINER, A. P., OSHIMA, J., CAWTHON, R. M., HSUEH, W. C. & CUMMINGS, S. R. (2004) The genetics of human longevity. *Am J Med,* 117**,** 851-60.

BURGER, J., KIRCHNER, M., BRAMANTI, B., HAAK, W. & THOMAS, M. G. (2007) Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc Natl Acad Sci U S A,* 104**,** 3736-41.

BUTTENHUIS, H. (1995) The Faunal Remains. IN ROODENBERG, J. (Ed.) *The Ilipinar Excavations I, Five Seasons of Fieldwork in NW Anatolia, 1987-91*. Istanbul, Nederlands Historisch-Archaeologisch Instituut.

CAMPBELL, B. (1999) *Human Evolution: An Introduction to Man's Adaptations*, AldineTransaction.

CARROLL, S. B. (2003) Genetics and the making of Homo sapiens. *Nature,* 422**,** 849-57.

CASTIGLIA, P. T. (1994) Lactose intolerance. *J Pediatr Health Care,* 8**,** 36-8.

CAVALLI-SFORZA, L. L., MENOZZI, P. & PIAZZA, A. (1994) *The History and Geography of Human Genes*, Princeton University Press.

CHEN, J. M., COOPER, D. N., CHUZHANOVA, N., FEREC, C. & PATRINOS, G. P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet,* 8**,** 762-75.

CHEUNG, J., ESTIVILL, X., KHAJA, R., MACDONALD, J. R., LAU, K., TSUI, L. C. & SCHERER, S. W. (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol,* 4**,** R25.

CHOU, H. H., HAYAKAWA, T., DIAZ, S., KRINGS, M., INDRIATI, E., LEAKEY, M., PAABO, S., SATTA, Y., TAKAHATA, N. & VARKI, A. (2002) Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc Natl Acad Sci U S A,* 99**,** 11736-41.

CLARK, J. G. D. (1965) Radiocarbon dating and the expansion of farming culture from the Near East over Europe. *Proc. Prehist. Soc.*, 31**,** 57– 73.

CLEGG, M. & AIELLO, L. C. (1999) A comparison of the nariokotome Homo erectus with juveniles from a modern human population. *Am J Phys Anthropol,* 110**,** 81-93.

COELHO, M., LUISELLI, D., BERTORELLE, G., LOPES, A. I., SEIXAS, S., DESTRO-BISOL, G. & ROCHA, J. (2005) Microsatellite variation and evolution of human lactase persistence. *Hum Genet,* 117**,** 329-39.

COLLEDGE, S., CONOLLY, J. & SHENNAN, S. (2004) Archaeobotanical evidence for the spread of farming in the Eastern Mediterranean. *Curr Anthropol,* 45**,** S35–S58.

COMET, J. P., AUDE, J. C., GLEMET, E., RISLER, J. L., HENAUT, A., SLONIMSKI, P. P. & CODANI, J. J. (1999) Significance of Z-value statistics of Smith-Waterman scores for protein alignments. *Comput Chem,* 23**,** 317-31.

CONARD, N. J. (2009) A female figurine from the basal Aurignacian of Hohle Fels Cave in southwestern Germany. *Nature,* 459**,** 248-52.

COOK, E. H., JR. & SCHERER, S. W. (2008) Copy-number variations associated with neuropsychiatric conditions. *Nature,* 455**,** 919-23.

COPLEY, M. S., BERSTAN, R., DUDD, S. N., DOCHERTY, G., MUKHERJEE, A. J., STRAKER, V., PAYNE, S. & EVERSHED, R. P. (2003) Direct chemical evidence for widespread dairying in prehistoric Britain. *Proc Natl Acad Sci U S A,* 100**,** 1524-9.

COPLEY, M. S., BERSTAN, R., MUKHERJEE, A. J., DUDD, S. N., STRAKER, V., PAYNE, S. & EVERSHED, R. P. (2005) Dairying in antiquity. III. Evidence from absorbed lipid residues dating to the British Neolithic. *Journal of Archaeological Science,* 32**,** 523-546.

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. & STEIN, C. (2009) *Introduction to Algorithms, Third Edition*.

COSTANTINI, M. & BERNARDI, G. (2009) Mapping insertions, deletions and SNPs on Venter's chromosomes. *PLoS One,* 4**,** e5972.

COTTON, J. A. & PAGE, R. D. (2005) Rates and patterns of gene duplication and loss in the human genome. *Proc Biol Sci,* 272**,** 277-83.

CRAIG, A. G. (2003) *Antigenic Variation,* Academic Press.

CRAIG, O. E., J., C., HERON, C. P., WILLIS, L. H., TAYLOR, G., WHITTLE, A. & COLLINS, M. J. (2005) Did the first farmers of central and eastern Europe produce dairy foods? *Antiquity,* 79**,** 882–894.

CRIBB, R. L. D. (1987) The Logic of the herd: A computer simulation of archaeological herd structure. *Journal Anthr.Arch.,* 6**,** 367.

CRISTIANINI, N. & HAHN, M. W. (2006) *Introduction to Computational Genomics: A Case Studies Approach*, Cambridge University Press.

CURRAT, M. & EXCOFFIER, L. (2005) The effect of the Neolithic expansion on European molecular diversity. *Proc Biol Sci,* 272**,** 679-88.

DARWIN, C. R. (1872) *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, John Murray.

DENNIS, G., JR., SHERMAN, B. T., HOSACK, D. A., YANG, J., GAO, W., LANE, H. C. & LEMPICKI, R. A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol,* 4**,** P3.

DIAMOND, J. (2002) Evolution, consequences and future of plant and animal domestication. *Nature,* 418**,** 700-7.

DIAMOND, J. M. (1998) *Guns, Germs and Steel: A short history of everybody for the last 13,000 years*, Vintage.

DONALDSON, I. J. & GOTTGENS, B. (2006) Evolution of candidate transcriptional regulatory motifs since the human-chimpanzee divergence. *Genome Biol,* 7**,** R52.

DUNN, C. A., MEDSTRAND, P. & MAGER, D. L. (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc Natl Acad Sci U S A,* 100**,** 12841-6.

DURRENS, P., NIKOLSKI, M. & SHERMAN, D. (2008) Fusion and fission of genes define a metric between fungal genomes. *PLoS Comput Biol,* 4**,** e1000200.

EDGAR, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EDMONDS, C. A., LILLIE, A. S. & CAVALLI-SFORZA, L. L. (2004) Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A,* 101**,** 975-9.

EDWARDS, C. J., BOLLONGINO, R., SCHEU, A., CHAMBERLAIN, A., TRESSET, A., VIGNE, J. D., BAIRD, J. F., LARSON, G., HO, S. Y., HEUPINK, T. H., SHAPIRO, B., FREEMAN, A. R., THOMAS, M. G., ARBOGAST, R. M., ARNDT, B., BARTOSIEWICZ, L., BENECKE, N., BUDJA, M., CHAIX, L., CHOYKE, A. M., COQUEUGNIOT, E., DOHLE, H. J., GOLDNER, H., HARTZ, S., HELMER, D., HERZIG, B., HONGO, H., MASHKOUR, M., OZDOGAN, M., PUCHER, E., ROTH, G., SCHADE-LINDIG, S., SCHMOLCKE, U., SCHULTING, R. J., STEPHAN, E., UERPMANN, H. P., VOROS, I., VOYTEK, B., BRADLEY, D. G. & BURGER, J. (2007) Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proc Biol Sci,* 274**,** 1377-1385.

EGGER, G., LIANG, G., APARICIO, A. & JONES, P. A. (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature,* 429**,** 457-63.

EMERSON, J. J., KAESSMANN, H., BETRAN, E. & LONG, M. (2004) Extensive gene traffic on the mammalian X chromosome. *Science,* 303**,** 537-40.

ENARD, W., PRZEWORSKI, M., FISHER, S. E., LAI, C. S., WIEBE, V., KITANO, T., MONACO, A. P. & PAABO, S. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature,* 418**,** 869-72.

ENATTAH, N. S., JENSEN, T. G., NIELSEN, M., LEWINSKI, R., KUOKKANEN, M., RASINPERA, H., EL-SHANTI, H., SEO, J. K., ALIFRANGIS, M., KHALIL, I. F., NATAH, A., ALI, A., NATAH, S., COMAS, D., MEHDI, S. Q., GROOP, L., VESTERGAARD, E. M., IMTIAZ, F., RASHED, M. S., MEYER, B., TROELSEN, J. & PELTONEN, L. (2008) Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am J Hum Genet,* 82**,** 57-72.

ENATTAH, N. S., SAHI, T., SAVILAHTI, E., TERWILLIGER, J. D., PELTONEN, L. & JARVELA, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet,* 30**,** 233-7.

ENATTAH, N. S., TRUDEAU, A., PIMENOFF, V., MAIURI, L., AURICCHIO, S., GRECO, L., ROSSI, M., LENTZE, M., SEO, J. K., RAHGOZAR, S., KHALIL,

I., ALIFRANGIS, M., NATAH, S., GROOP, L., SHAAT, N., KOZLOV, A., VERSCHUBSKAYA, G., COMAS, D., BULAYEVA, K., MEHDI, S. Q., TERWILLIGER, J. D., SAHI, T., SAVILAHTI, E., PEROLA, M., SAJANTILA, A., JARVELA, I. & PELTONEN, L. (2007) Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am J Hum Genet,* 81**,** 615-25.

ENRIGHT, A. J., VAN DONGEN, S. & OUZOUNIS, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res,* 30**,** 1575-84.

ETHENBERG, M. (2008) *Women in Prehistory*, University of Oklahoma Press.

EVERSHED, R. P., PAYNE, S., SHERRATT, A. G., COPLEY, M. S., COOLIDGE, J., UREM-KOTSU, D., KOTSAKIS, K., OZDOGAN, M., OZDOGAN, A. E., NIEUWENHUYSE, O., AKKERMANS, P. M., BAILEY, D., ANDEESCU, R. R., CAMPBELL, S., FARID, S., HODDER, I., YALMAN, N., OZBASARAN, M., BICAKCI, E., GARFINKEL, Y., LEVY, T. & BURTON, M. M. (2008) Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature,* 455**,** 528-31.

EXCOFFIER, L. (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol,* 13**,** 853-64.

FEINBERG, A. P. & TYCKO, B. (2004) The history of cancer epigenetics. *Nat Rev Cancer,* 4**,** 143-53.

FELSENSTEIN, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol,* 17**,** 368-76.

FELSENSTEIN, J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics***,** 164-166.

FELSENSTEIN, J. (2003) *Inferring Phylogenies*, Sinauer Associates.

FINN, R. N. & KRISTOFFERSEN, B. A. (2007) Vertebrate vitellogenin gene duplication in relation to the "3R hypothesis": correlation to the pelagic egg and the oceanic radiation of teleosts. *PLoS One,* 2**,** e169.

FISHER, S. E., VARGHA-KHADEM, F., WATKINS, K. E., MONACO, A. P. & PEMBREY, M. E. (1998) Localisation of a gene implicated in a severe speech and language disorder. *Nat Genet,* 18**,** 168-70.

FITCH, W. T. & REBY, D. (2001) The descended larynx is not uniquely human. *Proc Biol Sci,* 268**,** 1669-75.

FITZPATRICK, B. M., FORDYCE, J. A. & GAVRILETS, S. (2008) What, if anything, is sympatric speciation? *J Evol Biol,* 21**,** 1452-9.

FLATZ, G. & ROTTHAUWE, H. W. (1973) Lactose nutrition and natural selection. *Lancet,* 2**,** 76-7.

FLEAGLE, J. G. (1998) *Primate Adaptation and Evolution,* Academic Press.

FLEAGLE, J. G., ASSEFA, Z., BROWN, F. H. & SHEA, J. J. (2008) Paleoanthropology of the Kibish Formation, southern Ethiopia: Introduction. *J Hum Evol,* 55**,** 360-5.

FORCE, A., LYNCH, M., PICKETT, F. B., AMORES, A., YAN, Y. L. & POSTLETHWAIT, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics,* 151**,** 1531-45.

FRIEDMAN, R. & HUGHES, A. L. (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol,* 20**,** 154-61.

GALTIER, N. (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends Genet,* 19**,** 65-8.

GALTIER, N., PIGANEAU, G., MOUCHIROUD, D. & DURET, L. (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics,* 159**,** 907-11.

GARBER, P. A., ESTRADA, A., J.C., B.-M., E.W., H. & K.B., S. (2008) *South American Primates: Comparative Perspectives in the Study of Behavior, Ecology, and Conservation (Developments in Primatology: Progress and Prospects),* Springer.

GILAD, Y., MAN, O., PAABO, S. & LANCET, D. (2003) Human specific loss of olfactory receptor genes. *Proc Natl Acad Sci U S A,* 100**,** 3324-7.

GLUSMAN, G., YANAI, I., RUBIN, I. & LANCET, D. (2001) The complete human olfactory subgenome. *Genome Res,* 11**,** 685-702.

GOLDMAN, N. & YANG, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol,* 11**,** 725-36.

GONDER, M. K., MORTENSEN, H. M., REED, F. A., DE SOUSA, A. & TISHKOFF, S. A. (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol,* 24**,** 757-68.

GOWDY, J. (1997) *Limited Wants, Unlimited Means: A Reader On Hunter-Gatherer Economics And The Environment* Island Press.

GRAUR, D., SHUALI, Y. & LI, W. H. (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol,* 28**,** 279-85.

GU, X., WANG, Y. & GU, J. (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet,* 31**,** 205-9.

HALDANE, J. B. S. (1927) The mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Camb. Phil. Soc.,* 23.

HASEGAWA, M., KISHINO, H. & YANO, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution,* 22**,** 160-174.

HASSAN, F. A. (1981) *Demographic Archaeology,* New York, Academic Press.

HAYAKAWA, T., SATTA, Y., GAGNEUX, P., VARKI, A. & TAKAHATA, N. (2001) Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci U S A,* 98**,** 11399-404.

HENSHILWOOD, C. S., D'ERRICO, F., YATES, R., JACOBS, Z., TRIBOLO, C., DULLER, G. A., MERCIER, N., SEALY, J. C., VALLADAS, H., WATTS, I. & WINTLE, A. G. (2002) Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science,* 295**,** 1278-80.

HERRERA, R. J., LOWERY, R. K., ALFONSO, A., MCDONALD, J. F. & LUIS, J. R. (2006) Ancient retroviral insertions among human populations. *J Hum Genet,* 51**,** 353-62.

HEYER, L. J., KRUGLYAK, S. & YOOSEPH, S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res,* 9**,** 1106-15.

HEYMAN, M. B. (2006) Lactose intolerance in infants, children, and adolescents. *Pediatrics,* 118**,** 1279-86.

HIJAZI, S. S., ABULABAN, A., AMMARIN, Z. & FLATZ, G. (1983) Distribution of adult lactase phenotypes in Bedouins and in urban and agricultural populations of Jordan. *Tropical & Geographical Medicine,* 35**,** 157–161.

HOBOLTH, A., CHRISTENSEN, O. F., MAILUND, T. & SCHIERUP, M. H. (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet,* 3**,** e7.

HOLLOX, E. J., POULTER, M., ZVARIK, M., FERAK, V., KRAUSE, A., JENKINS, T., SAHA, N., KOZLOV, A. I. & SWALLOW, D. M. (2001) Lactase haplotype diversity in the Old World. *Am J Hum Genet,* 68**,** 160-172.

HOPKINS, W. D. & CANTALUPO, C. (2004) Handedness in chimpanzees (Pan troglodytes) is associated with asymmetries of the primary motor cortex but not with homologous language areas. *Behav Neurosci,* 118**,** 1176-83.

HUBBARD, T. J., AKEN, B. L., AYLING, S., BALLESTER, B., BEAL, K., BRAGIN, E., BRENT, S., CHEN, Y., CLAPHAM, P., CLARKE, L., COATES, G., FAIRLEY, S., FITZGERALD, S., FERNANDEZ-BANET, J., GORDON, L., GRAF, S., HAIDER, S., HAMMOND, M., HOLLAND, R., HOWE, K., JENKINSON, A., JOHNSON, N., KAHARI, A., KEEFE, D., KEENAN, S., KINSELLA, R., KOKOCINSKI, F., KULESHA, E., LAWSON, D., LONGDEN, I., MEGY, K., MEIDL, P., OVERDUIN, B., PARKER, A., PRITCHARD, B., RIOS, D., SCHUSTER, M., SLATER, G., SMEDLEY, D., SPOONER, W., SPUDICH, G., TREVANION, S., VILELLA, A., VOGEL, J., WHITE, S., WILDER, S., ZADISSA, A., BIRNEY, E., CUNNINGHAM, F., CURWEN, V., DURBIN, R., FERNANDEZ-SUAREZ, X. M., HERRERO, J., KASPRZYK, A., PROCTOR, G., SMITH, J., SEARLE, S. & FLICEK, P. (2009) Ensembl 2009. *Nucleic Acids Res,* 37**,** D690-7.

HUELSENBECK, J. P. & RONQUIST, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17**,** 754-5.

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294**,** 2310-4.

HULSEN, T., HUYNEN, M. A., DE VLIEG, J. & GROENEN, P. M. (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol,* 7**,** R31.

HUNT, K. D. (1994) The evolution of human bipedality: ecology and functional morphology. *Journal of human evolution,* 26**,** 183-202.

INGMAN, M., KAESSMANN, H., PAABO, S. & GYLLENSTEN, U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature,* 408**,** 708-13.

INGRAM, C. J. (2008) The evolutionary genetics of lactase persistence in Africa and the Middle East. London, University of London.

INGRAM, C. J., ELAMIN, M. F., MULCARE, C. A., WEALE, M. E., TAREKEGN, A., RAGA, T. O., BEKELE, E., ELAMIN, F. M., THOMAS, M. G., BRADMAN, N. & SWALLOW, D. M. (2007) A novel polymorphism

associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum Genet,* 120**,** 779-88.

INGRAM, C. J., MULCARE, C. A., ITAN, Y., THOMAS, M. G. & SWALLOW, D. M. (2009a) Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet,* 124**,** 579-91.

INGRAM, C. J., RAGA, T. O., TAREKEGN, A., BROWNING, S. L., ELAMIN, M. F., BEKELE, E., THOMAS, M. G., WEALE, M. E., BRADMAN, N. & SWALLOW, D. M. (2009b) Multiple Rare Variants as a Cause of a Common Phenotype: Several Different Lactase Persistence Associated Alleles in a Single Ethnic Group. *J Mol Evol*.

ISAAC, G. L. (1976) Stages of cultural elaboration in the Pleistocene: possible archaeological indicators of the development of language capabilities, in Origins and Evolution of Languages and Speech. *Annals of the New York Academy of Sciences,* 280**,** 276-288.

ITAN, Y., POWELL, A., BEAUMONT, M. A., BURGER, J. & THOMAS, M. G. (2009) The origins of lactase persistence in Europe. *PLoS Comput Biol,* 5**,** e1000491.

JABLONSKI, N. G. (2008) *Skin: a natural history*, University of California Press.

JACOBS, G. H., NEITZ, M., DEEGAN, J. F. & NEITZ, J. (1996) Trichromatic colour vision in New World monkeys. *Nature,* 382**,** 156-8.

JAIN, E., BAIROCH, A., DUVAUD, S., PHAN, I., REDASCHI, N., SUZEK, B. E., MARTIN, M. J., MCGARVEY, P. & GASTEIGER, E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics,* 10**,** 136.

JAMES, R. (1989) Hominid use of fire in the lower and middle Pleistocene. *Curr. Anthropol. ,* 30**,** 1-26.

JENUTH, J. P. (2000) The NCBI. Publicly available tools and resources on the Web. *Methods Mol Biol,* 132**,** 301-12.

JIANG, Z., TANG, H., VENTURA, M., CARDONE, M. F., MARQUES-BONET, T., SHE, X., PEVZNER, P. A. & EICHLER, E. E. (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet,* 39**,** 1361-8.

JOBLING, M., HURLES, M. & TYLER-SMITH, C. (2004) *Human Evolutionary Genetics: Origins, Peoples and Disease*, Garland Publishing.

JOHNSON, W. E. & COFFIN, J. M. (1999) Constructing primate phylogenies from ancient retrovirus sequences. *Proc Natl Acad Sci U S A,* 96**,** 10254-60.

JONES, C., JONES, C. A., KNOX JONES, J. & WILSON, D. E. (1996) Pan troglodytes. *Mammalian Species,* 529**,** 1-9.

JUKES, T. H. & CANTOR, C. R. (1969) Evolution of protein molecules. *Mammalian protein metabolism***,** 21-123.

KATOH, K., KUMA, K., TOH, H. & MIYATA, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res,* 33**,** 511-8.

KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30**,** 3059-66.

KAUFMAN, L. & ROUSSEEUW, P. J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis,* Wiley-Interscience.

KAWANO, T., KOYAMA, S., TAKEMATSU, H., KOZUTSUMI, Y., KAWASAKI, H., KAWASHIMA, S., KAWASAKI, T. & SUZUKI, A. (1995) Molecular cloning of cytidine monophospho-N-acetylneuraminic acid hydroxylase. Regulation of species- and tissue-specific expression of N-glycolylneuraminic acid. *J Biol Chem,* 270**,** 16458-63.

KENT, W. J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res,* 12**,** 656-64.

KIMURA, M. (1968) Evolutionary rate at the molecular level. *Nature,* 217**,** 624-6.

KING, J. L. & JUKES, T. H. (1969) Non-Darwinian evolution. *Science,* 164**,** 788-98.

KLEIN, R. G. (1989) *The Human Career: Human Biological and Cultural Origins,* University Of Chicago Press.

KLOPFSTEIN, S., CURRAT, M. & EXCOFFIER, L. (2006) The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol,* 23**,** 482-90.

KNUTH, D. E. (1973) *The Art of Computer Programming,* Addison Wesley.

KOONIN, E. V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet,* 39**,** 309-38.

KORNACK, D. R. & RAKIC, P. (1998) Changes in cell-cycle kinetics during the development and evolution of primate neocortex. *Proc Natl Acad Sci U S A,* 95**,** 1242-6.

KRAUSE, J., UNGER, T., NOCON, A., MALASPINAS, A. S., KOLOKOTRONIS, S. O., STILLER, M., SOIBELZON, L., SPRIGGS, H., DEAR, P. H., BRIGGS, A.

W., BRAY, S. C., O'BRIEN, S. J., RABEDER, G., MATHEUS, P., COOPER, A., SLATKIN, M., PAABO, S. & HOFREITER, M. (2008) Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evol Biol,* 8**,** 220.

KRETCHMER, N. (1972) Lactose and lactase. *Sci Am,* 227**,** 71-8.

LAI, C. S., FISHER, S. E., HURST, J. A., VARGHA-KHADEM, F. & MONACO, A. P. (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature,* 413**,** 519-23.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LEAKEY, L. S., TOBIAS, P. V. & NAPIER, J. R. (1964) A New Species of the Genus Homo from Olduvai Gorge. *Nature,* 202**,** 7-9.

LEAKEY, M. D. & HAY, R. L. (1979) Pliocene footprints in the Laetolil Beds at Laetoli, northern Tanzania. *Nature,* 278**,** 317-323.

LEE, B. M. & MAHADEVAN, L. C. (2009) Stability of histone modifications across mammalian genomes: Implications for 'epigenetic' marking. *J Cell Biochem*.

LEWINSKY, R. H., JENSEN, T. G., MOLLER, J., STENSBALLE, A., OLSEN, J. & TROELSEN, J. T. (2005) T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Hum Mol Genet,* 14**,** 3945-53.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

LIEBERMAN, D. E., MCBRATNEY, B. M. & KROVITZ, G. (2002) The evolution and development of cranial form in Homosapiens. *Proc Natl Acad Sci U S A,* 99**,** 1134-9.

LISTER, A. M. & SHER, A. V. (2001) The origin and evolution of the woolly mammoth. *Science,* 294**,** 1094-7.

LIU, H., PRUGNOLLE, F., MANICA, A. & BALLOUX, F. (2006) A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet,* 79**,** 230-7.

LOWER, R. (1999) The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends Microbiol,* 7**,** 350-6.

LÜNING, J. (2005) Bandkeramische Hofplätze und absolute Chronologie der Bandkeramik. IN LÜNING, J., FRIEDRICH, C. & ZIMMERMANN, A. (Eds.) *Die Bandkeramik im 21. Jahrhundert: Symposium in der Abtei Brauweiler bei Köln.*

MACE, R. (1993) Transitions Between Cultivation and Pastoralism in Sub-Saharan Africa. *Current Anthropology,* 34**,** 363-382.

MACQUEEN, J. (1967) Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob,* 1**,** 281-297.

MANIA, D. & MANIA, U. (1988) Deliberate engravings on bone artefacts of Homo Erectus. *Rock Art Research,* 5**,** 91-95.

MARAIS, G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends Genet,* 19**,** 330-8.

MARCHETTI, C., MEYER, P. S. & AUSUBEL, J. H. (1996) Human population dynamics revisited with the logistic model: how much can be modeled and predicted? *Technol Forecast Soc Change,* 52**,** 1-30.

MARGOLIASH, E. (1963) Primary Structure and Evolution of Cytochrome C. *Proc Natl Acad Sci U S A,* 50**,** 672-9.

MARQUES, A. C., DUPANLOUP, I., VINCKENBOSCH, N., REYMOND, A. & KAESSMANN, H. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol,* 3**,** e357.

MAYNARD SMITH, J. (1998) *Evolutionary genetics,* Oxford; New York, Oxford University Press.

MCCRACKEN, R. D. (1971a) Lactase Deficiency: An Example of Dietary Evolution. *Current Anthropology,* 12**,** 497-517.

MCCRACKEN, R. D. (1971b) Origins and implications of the distribution of adult lactase deficiency in human populations. *J Trop Pediatr Environ Child Health,* 17**,** 7-10.

MCDOUGALL, I., BROWN, F. H. & FLEAGLE, J. G. (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature,* 433**,** 733-6.

MERCADER, J., PANGER, M. & BOESCH, C. (2002) Excavation of a chimpanzee stone tool site in the African rainforest. *Science,* 296**,** 1452-5.

MIKKELSEN, T. S., HILLIER, L. W., EICHLER, E. E., ZODY, M. C., JAFFE, D. B., YANG, S.-P., ENARD, W., HELLMANN, I., LINDBLAD-TOH, K., ALTHEIDE, T. K., ARCHIDIACONO, N., BORK, P., BUTLER, J., CHANG, J. L., CHENG, Z., CHINWALLA, A. T., DEJONG, P., DELEHAUNTY, K. D., FRONICK, C. C., FULTON, L. L., GILAD, Y., GLUSMAN, G., GNERRE, S., GRAVES, T. A., HAYAKAWA, T., HAYDEN, K. E., HUANG, X., JI, H., KENT, W. J., KING, M.-C., KULBOKASIII, E. J., LEE, M. K., LIU, G., LOPEZ-OTIN, C., MAKOVA, K. D., MAN, O., MARDIS, E. R., MAUCELI, E., MINER, T. L., NASH, W. E., NELSON, J. O., PAABO, S., PATTERSON, N. J., POHL, C. S., POLLARD, K. S., PRUFER, K., PUENTE, X. S., REICH, D., ROCCHI, M., ROSENBLOOM, K., RUVOLO, M., RICHTER, D. J., SCHAFFNER, S. F., SMIT, A. F. A., SMITH, S. M., SUYAMA, M., TAYLOR, J., TORRENTS, D., TUZUN, E., VARKI, A., VELASCO, G., VENTURA, M., WALLIS, J. W., WENDL, M. C., K.WILSON, R., LANDER, E. S. &

WATERSTON, R. H. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* 437**,** 69-87.

MONTGOMERY, S. (2009) Current computational methods for prioritizing candidate regulatory polymorphisms. *Methods Mol Biol,* 569**,** 89-114.

MORGAN, E. (1999) *The Aquatic Ape Hypothesis: Most Credible Theory of Human Evolution,* Souvenir Press Ltd.

MORWOOD, M. J., BROWN, P., JATMIKO, SUTIKNA, T., SAPTOMO, E. W., WESTAWAY, K. E., DUE, R. A., ROBERTS, R. G., MAEDA, T., WASISTO, S. & DJUBIANTONO, T. (2005) Further evidence for small-bodied hominins from the Late Pleistocene of Flores, Indonesia. *Nature,* 437**,** 1012-7.

MULCARE, C. A. (2006a) The Evolution of the Lactase Persistence Phenotype. *Department of Biology.* London, University of London.

MULCARE, C. A. (2006b) The evolution of the lactase persistence phenotype. London, University of London.

MULCARE, C. A., WEALE, M. E., JONES, A. L., CONNELL, B., ZEITLYN, D., TAREKEGN, A., SWALLOW, D. M., BRADMAN, N. & THOMAS, M. G. (2004) The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet,* 74**,** 1102-10.

MYLES, S., BOUZEKRI, N., HAVERFIELD, E., CHERKAOUI, M., DUGOUJON, J. M. & WARD, R. (2005) Genetic evidence in support of a shared Eurasian-North African dairying origin. *Hum Genet,* 117**,** 34-42.

NAPIER, J. (1962) Fossil Hand Bones from Olduvai Gorge. *Nature,* 196**,** 409 - 411.

NATH, J. & JOHNSON, K. L. (2000) A review of fluorescence in situ hybridization (FISH): current status and future prospects. *Biotech Histochem,* 75**,** 54-78.

NEI, M. & SAITOU, N. (1986) Genetic relationship of human populations and ethnic differences in reaction to drugs and food. *Prog Clin Biol Res,* 214**,** 21-37.

NEWCOMER, A. D., MCGILL, D. B., THOMAS, P. J. & HOFMANN, A. F. (1975) Prospective comparison of indirect methods for detecting lactase deficiency. *N Engl J Med,* 293**,** 1232-6.

NIEUWENHUYS, R., VOOGD, J. & VAN HUIJZEN, C. (2007) *The Human Central Nervous System: A Synopsis and Atlas,* Springer.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol,* 302**,** 205-17.

O'BRIEN, E. (1981) The projectile capabilities of an Acheulian handaxe from Olorgesailie. *Current Anthropology,* 22**,** 76-79.

O'BRIEN, K. P., REMM, M. & SONNHAMMER, E. L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res,* 33**,** D476-80.

OHNO, S. (1970) *Evolution by gene duplication.* , Springer-Verlag.

OLSON, M. V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet,* 64**,** 18-23.

ÖZDOGAN, M. (2007) Von Zentralanatolien nach Europa. Die Ausbreitung der neolithischen Lebensweise. IN LANDESMUSEUM, B. (Ed.) *Vor 12000 Jahren in Anatolien- Die ältesten Monumente der Menschheit.* Stuttgart, Konrad Theiss Verlag.

ÖZDOGAN, M. & BASGELEN, N. (1999) Neolithic in Turkey. The cradle of civilization. Istanbul, Ancient Anatolians Civilizations Series 3.

PATTERSON, N., RICHTER, D. J., GNERRE, S., LANDER, E. S. & REICH, D. (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature,* 441**,** 1103-8.

PAULDING, C. A., RUVOLO, M. & HABER, D. A. (2003) The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc Natl Acad Sci U S A,* 100**,** 2507-11.

PAVÚK, J. (2005) Typologische Geschichte der Linearbandkeramik. IN LÜNING, J., FRIEDRICH, C. & ZIMMERMANN, A. (Eds.) *Die Bandkeramik im 21. Jahrhundert: Symposium in der Abtei Brauweiler bei Köln.*

PEUHKURI, K. (2000) Lactose, lactase, and bowel disorders. Helsinki, University of Helsinki.

PINHASI, R., FORT, J. & AMMERMAN, A. J. (2005) Tracing the origin and spread of agriculture in Europe. *PLoS Biol,* 3**,** e410.

POWELL, A., SHENNAN, S. & THOMAS, M. G. (2009) Late Pleistocene demography and the appearance of modern human behavior. *Science,* 324**,** 1298-301.

PRICE, D. (2005) Energy and human evolution *Population & Environment,* 16**,** 301-319.

PRINCE, V. E. & PICKETT, F. B. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet,* 3**,** 827-37.

RANNALA, B. & YANG, Z. (2007) Inferring speciation times under an episodic molecular clock. *Syst Biol,* 56**,** 453-66.

RASINPERA, H., SAVILAHTI, E., ENATTAH, N. S., KUOKKANEN, M., TOTTERMAN, N., LINDAHL, H., JARVELA, I. & KOLHO, K. L. (2004) A genetic test which can be used to diagnose adult-type hypolactasia in children. *Gut,* 53**,** 1571-6.

RAY, N., CURRAT, M. & EXCOFFIER, L. (2003) Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol,* 20**,** 76-86.

REMM, M., STORM, C. E. & SONNHAMMER, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol,* 314**,** 1041-52.

RODENHISER, D. & MANN, M. (2006) Epigenetics and human disease: translating basic biology into clinical applications. *CMAJ,* 174**,** 341-8.

ROHLAND, N., MALASPINAS, A. S., POLLACK, J. L., SLATKIN, M., MATHEUS, P. & HOFREITER, M. (2007) Proboscidean mitogenomics: chronology and mode of elephant evolution using mastodon as outgroup. *PLoS Biol,* 5**,** e207.

RÖHRS, M. & HERRE, W. (1961) Zur Frühentwicklung der Haustiere. Die Tierreste der Neolithischen Siedlung Fikirtepe am Kleinasiatischen Gestade des Bosporus. *Zeitschrift für Tierzüchtung und Züchtungsbiologie.*

SAMUELSON, L. C., WIEBAUER, K., GUMUCIO, D. L. & MEISLER, M. H. (1988) Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene. *Nucleic Acids Res,* 16**,** 8261-76.

SAVOLAINEN, P., ZHANG, Y. P., LUO, J., LUNDEBERG, J. & LEITNER, T. (2002) Genetic evidence for an East Asian origin of domestic dogs. *Science,* 298**,** 1610-3.

SCARRE, C. (2009) *The Human Past: World Prehistory & the Development of Human Societies,* Thames & Hudson.

SELLERS, W. (2000) Primate Evolution. University of Edinburgh.

SENUT, B., PICKFORD, M., GOMMERY, D., MEIN, P., CHEBOI, K. & COPPENS, Y. (2001) First hominid from the Miocene (Lukeino Formation, Kenya). *Comptes Rendus de l'Academie des Sciences,* 332.

SEPKOWITZ, K. A. (2001) AIDS--the first 20 years. *N Engl J Med,* 344**,** 1764-72.

SERRE, D., LANGANEY, A., CHECH, M., TESCHLER-NICOLA, M., PAUNOVIC, M., MENNECIER, P., HOFREITER, M., POSSNERT, G. & PAABO, S. (2004) No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol,* 2**,** E57.

SHE, X., LIU, G., VENTURA, M., ZHAO, S., MISCEO, D., ROBERTO, R., CARDONE, M. F., ROCCHI, M., GREEN, E. D., ARCHIDIACANO, N. & EICHLER, E. E. (2006) A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res,* 16**,** 576-83.

SHEPARD, D. (1968) A two-dimensional interpolation function for irregularly-spaced data. *ACM National Conference.*

SHIMADA, M. K., KIM, C. G., KITANO, T., FERRELL, R. E., KOHARA, Y. & SAITOU, N. (2005) Nucleotide sequence comparison of a chromosome rearrangement on human chromosome 12 and the corresponding ape chromosomes. *Cytogenet Genome Res,* 108**,** 83-90.

SIBSON, R. (1981) A brief description of natural neighbor interpolation. IN BARNETT, V. (Ed.) *Interpreting Multivariate Data (Probability & Mathematical Statistics).* John Wiley & Sons.

SIMOONS, F. (1980) Effects of culture: geographical and historical approaches. *Int J Obes,* 4**,** 387-94.

SIMOONS, F. J. (1969) Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. I. Review of the medical research. *Am J Dig Dis,* 14**,** 819-36.

SIMOONS, F. J. (1970) Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis. *Am J Dig Dis,* 15**,** 695-710.

SIMOONS, F. J. (1978) The geographic hypothesis and lactose malabsorption. A weighing of the evidence. *Am J Dig Dis,* 23**,** 963-80.

SIMOONS, F. J. (2001) Persistence of lactase activity among northern europeans: A weighing of the evidence for the calcium absorption hypothesis. *Ecology of Food and Nutrition,* 40**,** 397-469.

SMEDLEY, D., HAIDER, S., BALLESTER, B., HOLLAND, R., LONDON, D., THORISSON, G. & KASPRZYK, A. (2009) BioMart - biological queries made easy. *BMC Genomics,* 10**,** 22.

SPENCER, C. C., DELOUKAS, P., HUNT, S., MULLIKIN, J., MYERS, S., SILVERMAN, B., DONNELLY, P., BENTLEY, D. & MCVEAN, G. (2006) The influence of recombination on human genetic diversity. *PLoS Genet,* 2**,** e148.

SPIELMANN, K. A. & EDER, J. F. (1994) Hunters and Farmers : Then and Now. *Annu. Rev. Anthropol.,* 23**,** 303-323.

STEDMAN, H. H., KOZYAK, B. W., NELSON, A., THESIER, D. M., SU, L. T., LOW, D. W., BRIDGES, C. R., SHRAGER, J. B., MINUGH-PURVIS, N. & MITCHELL, M. A. (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature,* 428**,** 415-8.

STEIPER, M. E. & YOUNG, N. M. (2006) Primate molecular divergence dates. *Mol Phylogenet Evol,* 41**,** 384-94.

STEPANOVA, M., TIAZHELOVA, T., SKOBLOV, M. & BARANOVA, A. (2005) A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics,* 21**,** 1789-96.

STEWART, C. A., HORTON, R., ALLCOCK, R. J., ASHURST, J. L., ATRAZHEV, A. M., COGGILL, P., DUNHAM, I., FORBES, S., HALLS, K., HOWSON, J. M., HUMPHRAY, S. J., HUNT, S., MUNGALL, A. J., OSOEGAWA, K., PALMER, S., ROBERTS, A. N., ROGERS, J., SIMS, S., WANG, Y., WILMING, L. G., ELLIOTT, J. F., DE JONG, P. J., SAWCER, S., TODD, J. A., TROWSDALE, J. & BECK, S. (2004) Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res,* 14**,** 1176-87.

STRABO (1969) The Geography of Strabo in eight volumes with an English translation of H.L.Jones. London, Cambridge, Mass., Henry G. Bohn.

SVERDLOV, E. D. (2000) Retroviruses and primate evolution. *Bioessays,* 22**,** 161-71.

SWALLOW, D. M. (2003) Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet,* 37**,** 197-219.

SWALLOW, D. M. (2004) DNA test for hypolactasia premature. *Gut,* 55**,** 131-132.

SWALLOW, D. M. (2006) DNA test for hypolactasia premature. *Gut,* 55**,** 131; author reply 131-2.

SWALLOW, D. M., POULTER, M. & HOLLOX, E. J. (2001) Intolerance to lactose and other dietary sugars. *Drug Metab Dispos,* 29**,** 513-6.

SWISHER, C. C., 3RD, CURTIS, G. H., JACOB, T., GETTY, A. G., SUPRIJO, A. & WIDIASMORO (1994) Age of the earliest known hominids in Java, Indonesia. *Science,* 263**,** 1118-21.

TATTERSALL, I. & SCHWARTZ, J. H. (1999) Hominids and hybrids: the place of Neanderthals in human evolution. *Proc Natl Acad Sci U S A,* 96**,** 7117-9.

TATUSOV, R. L., KOONIN, E. V. & LIPMAN, D. J. (1997) A genomic perspective on protein families. *Science,* 278**,** 631-7.

THIEME, H. (1997) Lower Palaeolithic hunting spears from Germany. *Nature,* 385**,** 807-10.

THOMAS, M. G., BARNES, I., WEALE, M. E., JONES, A. L., FORSTER, P., BRADMAN, N. & PRAMSTALLER, P. P. (2008) New genetic evidence supports isolation and drift in the Ladin communities of the South Tyrolean Alps but not an ancient origin in the Middle East. *Eur J Hum Genet,* 16**,** 124-34.

THOMAS, M. G., STUMPF, M. P. & HARKE, H. (2006) Evidence for an apartheid-like social structure in early Anglo-Saxon England. *Proc Biol Sci,* 273**,** 2651-7.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22**,** 4673-80.

TIRABOSCHI, P., HANSEN, L. A., THAL, L. J. & COREY-BLOOM, J. (2004) The importance of neuritic plaques and tangles to the development and evolution of AD. *Neurology,* 62**,** 1984-9.

TISHKOFF, S. A., REED, F. A., FRIEDLAENDER, F. R., EHRET, C., RANCIARO, A., FROMENT, A., HIRBO, J. B., AWOMOYI, A. A., BODO, J. M., DOUMBO, O., IBRAHIM, M., JUMA, A. T., KOTZE, M. J., LEMA, G., MOORE, J. H., MORTENSEN, H., NYAMBO, T. B., OMAR, S. A., POWELL, K., PRETORIUS, G. S., SMITH, M. W., THERA, M. A., WAMBEBE, C., WEBER, J. L. & WILLIAMS, S. M. (2009) The genetic structure and history of Africans and African Americans. *Science,* 324**,** 1035-44.

TISHKOFF, S. A., REED, F. A., RANCIARO, A., VOIGHT, B. F., BABBITT, C. C., SILVERMAN, J. S., POWELL, K., MORTENSEN, H. M., HIRBO, J. B., OSMAN, M., IBRAHIM, M., OMAR, S. A., LEMA, G., NYAMBO, T. B., GHORI, J., BUMPSTEAD, S., PRITCHARD, J. K., WRAY, G. A. &

DELOUKAS, P. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet,* 39**,** 31-40.

TOMASELLO, M. (1999) The Human Adaptation for Culture. *Annual Review of Anthropology,* 28**,** 509-529.

TRAHERNE, J. A., HORTON, R., ROBERTS, A. N., MIRETTI, M. M., HURLES, M. E., STEWART, C. A., ASHURST, J. L., ATRAZHEV, A. M., COGGILL, P., PALMER, S., ALMEIDA, J., SIMS, S., WILMING, L. G., ROGERS, J., DE JONG, P. J., CARRINGTON, M., ELLIOTT, J. F., SAWCER, S., TODD, J. A., TROWSDALE, J. & BECK, S. (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet,* 2**,** e9.

TREMBLAY, M. & VEZINA, H. (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am J Hum Genet,* 66**,** 651-8.

TRESSET, A. (1996) Le rôle des relations homme/animal dans l'évolution économique et culturelle des sociétés des V-VI millénaires en Bassin Parisien. Paris, Université de Paris I, Panthéon- Sorbonne.

TRESSET, A. (1997) L'approvisionnement carné Cerny dans le contexte néolithique du Bassin Parisien. IN CONSTATIN, C., MORDANT, D. & SIMONIN, D. (Eds.) *La Culture de Cerny: Nouvelle économie, nouvelle société au Néolithique.* Nemours, Actes du colloque de Nemours.

TROY, C. S., MACHUGH, D. E., BAILEY, J. F., MAGEE, D. A., LOFTUS, R. T., CUNNINGHAM, P., CHAMBERLAIN, A. T., SYKES, B. C. & BRADLEY, D. G. (2001) Genetic evidence for Near-Eastern origins of European cattle. *Nature,* 410**,** 1088-91.

UCKO, P. (2007) *The Domestication and Exploitation of Plants and Animals*, Aldine Transaction.

VAN NOORT, V., SNEL, B. & HUYNEN, M. A. (2003) Predicting gene function by conserved co-expression. *Trends Genet,* 19**,** 238-42.

VAN VALEN, L. & SLOAN, R. E. (1965) The earliest primates. *Science,* 150**,** 743-5.

VERONA, R. I., MANN, M. R. & BARTOLOMEI, M. S. (2003) Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annu Rev Cell Dev Biol,* 19**,** 237-59.

VIGNE, J.-D. (2006) Maîtrise et usages de l'élevage et des animaux domestiques au Néolithique: quelques illustrations au Proche-Orient et en Europe. IN GUILAINE, J. (Ed.) *Populations néolithiques et environnements.* Paris, Errance éd.

VIGNE, J. D. & HELMER, D. (2007) Was milk a "secondary product" in the Old World Neolithisation process? Its role in the domestication of cattle, sheep and goats. *Anthropozoologica,* 42**,** 9.

VILLARREAL, L. P. (1997) On viruses, sex, and motherhood. *J Virol,* 71**,** 859-65.

WALKER, A. & LEAKEY, R. (1993) *The Nariokotome Homo Erectus Skeleton*, Springer.

WALL, J. D. & PRZEWORSKI, M. (2000) When did the human population size start increasing? *Genetics,* 155**,** 1865-74.

WANG, X., GRUS, W. E. & ZHANG, J. (2006) Gene losses during human origins. *PLoS Biol,* 4**,** e52.

WATSON, D. (1992) *Contouring: A Guide to the Analysis and Display of Spatial Data*, Pergamon.

WATSON, D. (1994) *nngridr: An implementation of natural neighbour implementation*, David Watson.

WATSON, J. D. & CRICK, F. H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature,* 171**,** 737-8.

WEALE, M. E., YEPISKOPOSYAN, L., JAGER, R. F., HOVHANNISYAN, N., KHUDOYAN, A., BURBAGE-HALL, O., BRADMAN, N. & THOMAS, M. G. (2001) Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. *Hum Genet,* 109**,** 659-74.

WEBSTER, M. T. (2009) Patterns of autosomal divergence between the human and chimpanzee genomes support an allopatric model of speciation. *Gene,* 443**,** 70-5.

WEISS, K. M. (2004) The unkindest cup. *Lancet,* 363**,** 1489-90.

WHEELER, P. E. (1984) The Evolution of Bipedality and Loss of Functional Body Hair in Hominoids. *Journal of Human Evolution,* 13**,** 91-98.

WHITE, T. D., ASFAW, B., DEGUSTA, D., GILBERT, H., RICHARDS, G. D., SUWA, G. & HOWELL, F. C. (2003) Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature,* 423**,** 742-7.

WHITE, T. D., SUWA, G., SIMPSON, S. & ASFAW, B. (2000) Jaws and teeth of Australopithecus afarensis from Maka, Middle Awash, Ethiopia. *Am J Phys Anthropol,* 111**,** 45-68.

WHITEN, A., GOODALL, J., MCGREW, W. C., NISHIDA, T., REYNOLDS, V., SUGIYAMA, Y., TUTIN, C. E., WRANGHAM, R. W. & BOESCH, C. (1999) Cultures in chimpanzees. *Nature,* 399**,** 682-5.

WOLPOFF, M. H., SPUHLER, J. N., SMITH, F. H., RADOVCIC, J., POPE, G., FRAYER, D. W., ECKHARDT, R. & CLARK, G. (1988) Modern human origins. *Science,* 241**,** 772-4.

WOOD, B. & COLLARD, M. (1999) The human genus. *Science,* 284**,** 65-71.

WRANGHAM, R. & CONKLIN-BRITTAIN, N. (2003) Cooking as a biological trait'. *Comp Biochem Physiol A Mol Integr Physiol,* 136**,** 35-46.

XIE, H. G., KIM, R. B., WOOD, A. J. & STEIN, C. M. (2001) Molecular basis of ethnic differences in drug disposition and response. *Annu Rev Pharmacol Toxicol,* 41**,** 815-50.

YANG, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci,* 13**,** 555-6.

YANG, Z. (2006a) *Computational Molecular Evolution*, Oxford University Press.

YANG, Z. (2006b) Molecular Clock and Estimation of Species Divergence Times. *Computational Molecular Evolution.* Oxford University Press.

YANG, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

YANG, Z. (2009) PAML: Phylogenetic Analysis by Maximum Likelihood

YANG, Z. & RANNALA, B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol,* 23**,** 212-26.

YANG, Z. & YODER, A. D. (2003) Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst Biol,* 52**,** 705-16.

YODER, A. D. & YANG, Z. (2000) Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol,* 17**,** 1081-90.

ZHANG, F., CARVALHO, C. M. & LUPSKI, J. R. (2009) Complex human chromosomal and genomic rearrangements. *Trends Genet,* 25**,** 298-307.

ZHENG, D., FRANKISH, A., BAERTSCH, R., KAPRANOV, P., REYMOND, A., CHOO, S. W., LU, Y., DENOEUD, F., ANTONARAKIS, S. E., SNYDER, M., RUAN, Y., WEI, C. L., GINGERAS, T. R., GUIGO, R., HARROW, J. & GERSTEIN, M. B. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res,* 17**,** 839-51.

ZHENG, D. & GERSTEIN, M. B. (2006) A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol,* 7 Suppl 1**,** S13 1-10.

ZVELEBIL, M. & ZVELEBIL, K. V. (1988) Agricultural transition and Indo-European dispersals. *Antiquity,* 62**,** 574–583.