From Species To Languages

A phylogenetic approach to human prehistory

QUENTIN DOUGLAS ATKINSON

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology,
Department of Psychology,
The University of Auckland, 2006.

ABSTRACT

Languages, like species, evolve. Just like biologists, historical linguists infer relationships between the lineages they study by analysing heritable features. For linguists, these features can be words, grammar and phonemes. This linguistic evidence of descent with modification plays an important role in our understanding of human prehistory. However, conventional methods in historical linguistics do not employ an explicit optimality criterion to evaluate evolutionary language trees. These methods cannot quantify uncertainty in the inferences nor provide an absolute chronology of divergence events. Previous attempts to estimate divergence times from lexical data using glottochronological methods have been heavily criticized, particularly for the assumption of constant rates of lexical replacement. Computational phylogenetic methods from biology can overcome these problems and allow divergence times to be estimated without the assumption of constant rates. Here these methods are applied to lexical data to test hypotheses about human prehistory. First, divergence time estimates for the age of the Indo-European language family are used to test between two competing theories of Indo-European origin – the Kurgan hypothesis and the Anatolian farming hypothesis. The resulting age estimates are consistent with the age range implied by the Anatolian farming theory. Validation exercises using different models, data sets and coding procedures, as well as the analysis of synthetic data, indicate these results are highly robust. Second, the same methodology was applied to Mayan lexical data to infer historical relationships and divergence times within the Mayan language family. The results highlight interesting uncertainties in Mayan language relationships and suggest that the family may be older than previously thought. Finally, returning to biology, similar tree-building and model validation techniques are used to draw inferences about human origins and dispersal from human mitochondrial DNA sequence data. These analyses support a human origin 150,000-250,000 years ago and reveal time dependency in rates of mitochondrial DNA evolution. Population size estimates generated using a coalescent approach suggest a two-phase human population expansion from Africa. Potential correlations between human genetic and linguistic diversity are highlighted. I conclude that there is much to be gained by linguists and biologists using the same methods and speaking the same language.

ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr Russell Gray - I could not have asked for a better supervisor. Thank you for your inspiration, motivation, support and guidance, and for sharing your skills as a researcher. It has been a wonderful four years working with you.

To my office buddy, Simon Greenhill, thank you for the good company and for all your help and advice. You are a scholar and a gentleman…and a viable alternative to Google. I am also very grateful to Professor Lyle Campbell at the University of Utah for his expert help and generous hospitality. To Elisabeth Norcliffe, thank you for sharing your knowledge of Mayan languages and the Ecuadorian hat-making industry. To Dr Geoff Nicholls, thank you for writing such a great program and for our lengthy sessions at the whiteboard. Thanks to David Welch for all your time and effort implementing a GUI that I could drive (and for that free trip to Switzerland). Thanks also to Dr Alexei Drummond for helping me tame the BEAST and take over the world and to Marcel van de Steeg for his valiant attempts to keep the recalcitrant "cluster" running. Thank you also to Professor Mike Corballis for providing sage advice when needed and to Professor Mark Pagel for problem solving when needed. I would also like to thank Scott Allan, Bob Blust, Lounès Chikhi, Penny Gray, Roger Green, Jeff Hamm, Niki Harré, John Huelsenbeck, Mark Liberman, David Penny, Allan Rodrigo, Fredrik Ronquist, Michael Sanderson and Stephen Shennan for useful advice and/or comments on manuscripts. To Mum and Dad, I really appreciate all your support over the years. And I owe an especially big thank you to Emma for so much "souper" advice and encouragement and for keeping me sane and happy.

Finally, I would like to acknowledge funding from the Royal Society of New Zealand, the University of Auckland, and also the Foundation for Research Science and Technology, who kindly provided a Bright Futures Scholarship.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One

INTRODUCTION

> **gene** n. the basic unit capable of transmitting characteristics from one generation to the next[1]. From Greek *genos* and *genea,* from Indo-European **gen*-, to give birth, beget; with derivatives referring to aspects and results of procreation and to familial and tribal groups; derivatives include kin, king, jaunty, genius, pregnant, gingerly, and nature[2].
>
> **word** n. a meaningful sound or combination of sounds that is a unit of language or its representation in a text[1]. From Old English *word*, from Germanic **wurdam*, from Indo-European *wer*-, to speak; derivatives include verb, rhetoric, rhyme and irony[2].

Languages, like species, are the product of descent with modification. Just as species can split into two populations which begin to evolve along independent evolutionary trajectories, so too can languages divide and begin to diverge. Like genes, hereditary units of language, such as words, syntax and grammar, are passed on from generation to generation with almost perfect fidelity. As a result of this process, the genealogy of the world's languages is preserved in the hereditary units of the languages themselves. Hence, just as the ancestry of biological species or organisms can be determined by comparing genes, so too can the history of languages, and the people who speak them, be revealed by comparing linguistic characters.

The parallels of process between linguistic and biological evolution mean that both disciplines employ similar methods for historical inference. Indeed, as is discussed in Chapter 2, the fields of biology and linguistics have been curiously connected almost

[1] Encarta® World English Dictionary © 1999 Microsoft Corporation. All rights reserved. Developed for Microsoft by Bloomsbury Publishing Plc.
[2] The American Heritage® Dictionary of the English Language: Fourth Edition. 2000.

since their inception. Today, an essential tool in evolutionary biology and historical linguistics is the evolutionary tree, or "phylogeny", depicting historical relationships between species or languages. In both disciplines, characters are mapped onto a tree and their fit is evaluated according to a set of implicit or explicit optimality criterion in order to draw inferences about the evolutionary history of the languages or species of interest. The traditional approach in both fields has been to perform this process manually. Biologists would compare homologous morphological or genetic traits between species and construct a tree that best explained the variation in the observed traits; the "best" tree generally being the tree that implied the least amount of evolutionary changes. Similarly, linguists compare homologous words (known as "cognates"), sounds and grammatical elements between languages and use these to infer the optimal language tree. Unfortunately, this optimization process is complex and computationally intensive, especially for large numbers of species/languages and large data sets – there are over 8×10^{21} possible rooted evolutionary trees relating just 20 species/languages. This can make determining the most likely evolutionary scenario extremely slow, or effectively impossible for large datasets[1]. In addition, without an explicit optimality criterion or model of evolution, determining the "optimal" tree may be somewhat subjective.

Over the last few decades, however, evolutionary biology has been revolutionized by a dramatic increase in the amount of genetic information available for analysis and the concurrent development of powerful computational tools for phylogenetic inference. For example, Bayesian inference of phylogeny and likelihood modelling allow trees with many taxa to be quickly and objectively evaluated using large genetic data sets and explicit models of evolution (Huesenbeck *et al*., 2001). As well as inferring phylogenetic relationships between species, given some assumptions about rates of evolution through time, it is also possible to estimate dates for lineage divergence events (Zuckerkandl and Pauling, 1962). These methods provide biologists with a rigorous empirical framework for hypothesis testing that has proved enormously successful for drawing inferences about evolutionary processes and elucidating the past.

[1] Foulds and Graham (1982) demonstrated that finding the best tree was NP-hard under parsimony. I.e. the computation time rises faster than exponentially as taxa are added.

The parallels between biological and linguistic evolution mean that these computational methods from biology can also be applied to the study of languages. Instead of modelling gene evolution, it is possible to model the evolution of words, sounds or other language features. This is particularly exciting because many of the problems of phylogenetic inference and divergence time estimation faced by historical linguists have biological equivalents that biologists have managed to overcome. For example, if linguistic evidence is to be constructively linked with archaeological and genetic evidence, it is necessary to establish an independent language chronology that can be correlated with genetic divergence times and the archaeological record (Kirch and Green, 1987). Unfortunately, previous attempts to extract chronological information from language data have been based on glottochronology - a method that is now largely discredited due to a number of problems, including the flawed assumption of constant rates of lexical replacement through time (Berglsand and Vogt, 1962; Blust, 2000; Campbell, 2004). However, biologists face similar problems in estimating species divergence times on the basis of genetic data and have been able to develop methods that can address these problems. Sanderson's rate-smoothing algorithm (Sanderson, 2002a, 2002b), for example, allows divergence times to be estimated without the assumption of strictly clock-like rates of evolution.

The aim of this thesis is to explore the species-languages analogy through the application of different computational phylogenetic methods from biology to linguistic data and, in so doing, to shed light on some important questions about human prehistory. Whilst a number of different phylogenetic tools are used here to answer a range of questions about human history and language evolution, a key focus of this thesis will be the estimation of language divergence times by comparing lexical (vocabulary) data between languages. The basic structure of the thesis is outlined below, with a brief discussion of the content of each chapter.

Chapter 2 provides an historical overview of the "curious parallels and curious connections" between the study of evolutionary biology and historical linguistics. Intriguingly, as well as analogies between the processes of biological and linguistic evolution, the fields themselves appear to have co-evolved. Based on a recent paper in *Systematic Biology* (Atkinson and Gray, 2005), this chapter traces the history of the

two disciplines from the ancient Greek philosophers to the present, identifying examples of the borrowing of ideas and methods between biologists and linguists, as well as instances of parallel and convergent evolution. This leads to a discussion of the current status of phylogenetic methods in biology and linguistics and a brief review of recent attempts to apply phylogenetic methods from biology to language data. The chapter concludes with a discussion of the potential contribution that recently developed methods from biology can make to historical linguistics. This highlights a number of exciting potential areas for future work involving not only improving phylogeny and divergence time estimates, but also reconstructing ancestral character states on a phylogeny, automating judgements of word similarity and relatedness, and testing hypotheses about deeper language relationships and borrowing.

The next section of the thesis, comprising Chapters 3 to 5, is based on a succession of publications applying model-based phylogenetic methods to resolve what has been dubbed the most well-studied and challenging problem in historical linguistics (Diamond and Bellwood, 2003) - the origin of the Indo-European language family. In addition to the obvious appeal of tackling historical linguistics' *grand prix*, there are a number of practical reasons why Indo-European constitutes a suitable starting point for the current research project. First, after over 200 years of Indo-European scholarship there is plenty of material available on almost all of the languages in the family, including detailed knowledge of the relationships between languages within the 10 major sub-groups of the family. Second, a long written record in many Indo-European languages provides a detailed account of the history of a number of the lineages, in some cases stretching back as far as three or four thousand years. Third, the most ancient texts are also valuable as fossilized forms of now extinct languages. The plethora of well-documented ancient and modern Indo-European sources increases the amount of deep historical signal in the data, making linguistic analysis of Indo-European especially powerful as a source of historical information. Finally, an extensive recorded history provides the opportunity of testing the methodology by comparing inferred relationships and divergence times to those implied by the historical record.

Chapter 3, based on a paper published in *Nature* in 2003 (Gray and Atkinson, 2003), uses likelihood models and Bayesian inference to test between two competing theories for the origin of the Indo-European language family – the Kurgan hypothesis and the more controversial Anatolian farming hypothesis. Each theory implies a very different age for the most recent common ancestor of the Indo-European languages – approximately 6,000 years and 8,500 years respectively. By using Bayesian inference of phylogeny and rate-smoothing algorithms it was possible to construct a confidence interval for the age at the base of the Indo-European language tree without the assumption of strictly clock-like rates of lexical evolution. The inferred date estimates clearly supported the time-depth predictions of the Anatolian farming hypothesis. Crucially, rather than providing a single point estimate for the age of the family, the uncertainty in the inferred time scale was also estimated. In addition, by repeating the analyses under a number of conditions the robustness of results to different methods of encoding the data, different age constraints and a range of other prior assumptions was established.

The findings reported here in Chapter 3 generated considerable international media interest[2]. In response, two book chapters were written presenting a more detailed explanation of the rationale and methodology employed, as well as addressing

[2] Boston Globe – "A new word on birth of western languages", 27 November, 2003.
Christchurch Press – "NZ research into language stuns world", 1 December, 2003.
Guardian (UK) – "Scientists trace evolution of Indo-European languages to Hittites", 27 November, 2003.
Hindustan Times (India) – "Mother of all Indo-European languages born in Turkey?", 26 November, 2003.
LA Times – "Language family traced to Turkey", 29 November, 2003, A23.
Le Monde – "Une étude relance le débat sur l'origine des langues indo-euroéennes", 27 November, 2003. (French daily newspaper)
Nature - News and Views – "Trees of life and of language", Nature, 426, 391-392.
Nature Science Update – "Language tree rooted in Turkey", 27 November, 2003.
New York Times – "A biological dig for the roots of language", 16 March, 2004.
New Zealand Herald – "NZ study cracks origin of English", 1 December, 2003, A01.
Onze Taal (Netherlands) – "L'origine dell'indoeuropeo", 27 November, 2003.
Pravda (Russia) – "Indians and Europeans Built Tower of Babel. What About Others?", 17 December, 2003.
Reuters – "Anatolian roots seen for Indo-European language tree", 26 November, 2003.
Science News – "Early date for the birth of Indo-European languages", Science, 302, 1490-1491.
ScienceNOW online – "An earlier birth for Indo-Eurpean languages?", 27 November, 2003.
Sydney Morning Herald – "English language traced to Turkish farmers", 1 December, 2003.
Telegraph (India) – "Bengali roots traced to Turkey", 28 November, 2003.
Telegraph (UK) – "Roots of English Traced to Turkey 8,000 years ago", 27 November, 2003.
Time Magazine – "New look at old words", 9 February, 2004, 52-52.
Washington Post – "Mother tongue may be older than many think", 27 November, 2003, A09.

concerns that had been raised about the initial paper (Atkinson and Gray, 2006; Atkinson and Gray, in press). Chapter 4 is an amalgamation of these two book chapters. It begins with an explanation of the nature of the linguistic data and how computational phylogenetic methods from biology, such as likelihood models of evolution and Bayesian inference of phylogeny, can be usefully applied to model the evolution of language lexicons. The question of Indo-European origins is then examined in more detail than in Chapter 3 and an expanded set of results is presented. The fit between these results and existing archaeological and linguistic evidence is examined, including a critical discussion of the much lauded "argument from the wheel", which has been claimed to conclusively support a younger Indo-European origin. The final section of the chapter comprises a response to critics.

Chapter 5 (based on Atkinson *et al*. [2005]) seeks to test the validity of the results presented in Chapters 3 and 4 by making use of an alternative model of lexical evolution and a second dataset of ancient Indo-European languages. As well as providing an alternative, independently compiled dataset with which to validate our previous findings, the ancient languages are phylogenetically "closer" to the point of Indo-European origin. This means they can be expected to contain more phylogenetic information about the oldest Indo-European relationships and to show less influence from modern borrowing. This dataset of ancient languages was analysed alongside the original dataset of chiefly contemporary languages using the time-reversible likelihood model from Chapters 3 and 4 as well as a very different, Dollo likelihood model (both models are described in Chapter 5). Divergence time estimates were shown to be robust across both models and datasets, producing remarkably consistent support for the age range implied by the Anatolian farming theory of Indo-European origin. The ability of both methods to reconstruct phylogeny and divergence times accurately was also tested using synthetic data. Examining the performance of two very different methodologies using synthetic data allowed the underlying process of lexical evolution to be investigated in more detail. The methods performed well under a range of scenarios, including scenarios of widespread and localized borrowing.

In Chapter 6, we venture beyond the confines of Indo-European and investigate another potential expansion with agriculture - the origin of the Mayan language family of Mesoamerica. The origin and history of the Maya has evoked the curiosity

of western scholars ever since John Lloyd Stephens' (1843) account of his journeys to the ancient ruins at Copan, Palenque and Tikal. The ancient hieroglyphic inscriptions these ruins bear, as well as more recent manuscripts known as the *Codices,* make Mayan one of the few New World language families with a relatively well-attested pre-Colombian written record[3]. Consequently, Mayan is an excellent candidate for language-based historical inference in Mesoamerica. Unfortunately, the current understanding of Mayan history derives from unreliable glottochronology-based estimates for the age of the family (e.g. Kaufman, 1976). Chapter 6 seeks to provide a more reliable Mayan chronology and to investigate the proposed Mayan expansion with agriculture (Kaufman, 1976; Diamond and Bellwood, 2003) by analysing a Mayan lexical database using the two methodologies described in Chapter 5. The results of this analysis suggest that the Mayan languages may be older than previously thought. Phylogenetic network methods were also employed to identify interesting patterns of reticulation in the data and a Highland Mayan homeland was inferred by mapping geographic characters onto the language trees.

Perhaps as a result of the failure of glottochronology, linguists are particularly sceptical of divergence time estimates based on lexical data. Consequently, Chapters 3 to 6 devote considerable attention to quantifying uncertainty in date estimates and to demonstrating the robustness of the results. Generally speaking, biologists have been less critical of divergence time estimates derived from genetic data. However, as Graur and Martin (2004) clearly demonstrate, biologists also have good reason to be sceptical – a large body of previous work is based on inappropriate models, unfounded rate calibrations and inaccurate estimates of uncertainty. Whilst Chapters 3 to 6 involved a biological approach to linguistics, in Chapter 7 the tables are turned as the approach to hypothesis testing and model validation developed in the preceding chapters using language data is applied to a biological dataset of human mitochondrial DNA (mtDNA) sequences.

Previous analyses of human mtDNA data have been instrumental in establishing a recent African human origin (Cann *et al*., 1987; Hasegawa and Horai, 1991; Vigilant *et al*., 1991; Stoneking *et al*., 1992; Penny *et al*., 1995; Watson *et al*., 1997; Ingman *et*

[3] Whilst other pre-Columbian writing systems existed (e.g. related systems in Zapotec and Epi-Olmec cultures), they are not as well attested as in Mayan.

al., 2000). However, the precision and accuracy of these findings has been limited by the the available data and methods of analysis. In addition, with a few exceptions (e.g., Templeton, 1992; Penny *et al*., 1995; Ruvolo, 1996) there has been little attention given to accurately quantifying uncertainty associated with tree topology and divergence time estimates. Furthermore, Ho *et al*. (2005) observed time-dependency in estimated rates of primate mtDNA evolution, which they argue could cause a systematic bias in divergence time estimates. Chapter 7 presents a thorough re-evaluation of the mtDNA evidence using newly available complete mtDNA sequence data and recently developed Bayesian phylogenetic methods to shed light on the origin and diversification of modern humans from Africa. As well as estimating divergence times, human population size is modelled through time using a coalescent approach. As in Chapters 3 to 6, particular attention is paid to quantifying uncertainty and testing the robustness of the results to variations in the assumptions of the method and choice of calibration points. The implications of these results for the currently accepted understanding of human origins are discussed. In keeping with the underlying theme of the thesis, this includes a discussion of evidence for an association between human genetic and linguistic variation.

To begin, however, we step back and look at why languages can sometimes be viewed very much like species, why words can sometimes be viewed like genes, and why historical linguists have so much in common with evolutionary biologists.

REFERENCES

Atkinson, Q. D. and R. D. Gray. 2005 Curious Parallels and Curious Connections – Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology* 54(4):513-526.

Atkinson, Q. D. and R. D. Gray. 2006. Are accurate dates an intractable problem for historical linguistics? Pages 269-296 in *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory.* (eds.) C. Lipo, M. O'Brien, S. Shennan & M. Collard. Aldine, Chicago.

Atkinson, Q. D. and R. D. Gray. In press. How old is the Indo-European language family? Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge.

Atkinson, Q. D. Nicholls, G., Welch, D. and R. D. Gray. 2005. From Words to Dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2):193-219.

Bergsland, K., and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3:115–153.

Blust, R. 2000. Why lexicostatistics doesn't work: the 'Universal Constant' hypothesis and the Austronesian languages. Pages 311-332 in *Time Depth in Historical Linguistics,* (eds.) C. Renfrew, A. McMahon, and L. Trask. McDonald Institute for Archaeological Research, Cambridge.

Cann, R. L., Stoneking, M. and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36.

Campbell, L. 2004. *Historical linguistics: An introduction, 2nd edition*. Edinburgh University Press, Edinburgh.

Diamond, J. and P. Bellwood. 2003. Farmers and their languages: the first expansions. *Science* 300:597.

Foulds, L. R., and R. L. Graham. 1982. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3:43-49.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *TRENDS in Genetics* 20:80-86.

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435-439.

Hasegawa, M. and S. Horai. 1991. Time of the Deepest Root for Polymorphism in Human Mitochondrial DNA. *Journal of Molecular Evolution* 32:37-42.

Ho, S. Y. W., Phillips, M. J., Cooper, A., and A. J. Drummond. 2005. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. *Molecular Biology and Evolution* 22(7):1561-1568.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310-2314.

Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*. 408:708-713.

Kaufman, T. S. 1976. Archaeological and linguistic correlations in Mayaland and associated areas of Meso-America. *Archaeology and Linguistics* 8(1):101-18.

Kirch P. V. and R. C. Green. 1987. History, phylogeny, and evolution in Polynesia. *Current Anthropology* 28:431–456.

Penny, D., Steel, M., Waddell, P. J. and M. D. Hendy. 1995. Improved Analyses of Human mtDNA Sequences Support a Recent African Origin for Homo sapiens. *Molecular Biology and Evolution* 12(5):863-882.

Ruvolo, M. 1996. A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. *Molecular Phylogenetics and Evolution* 5:202–219.

Sanderson, M. 2002a. *R8s, Analysis of Rates of Evolution, version 1.50*. http://ginger.ucdavis.edu/r8s/

Sanderson, M. 2002b. Estimating absolute rates of evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.

Stephens, J. L. 1843. *Incidents of Travel in the Yucatan*. Smithsonian Institute Press, Washington D.C.

Stoneking, M. Sherry, S. T., Redd, A. J., and L. Vigilant. 1992. New approaches to dating suggest a recent age for the human mtDNA ancestor. *Phil. Trans. R. Soc. Lond. B* 337:167-175.

Templeton, A. R., 1993. The "Eve" Hypothesis: A genetic critique and reanalysis. *American Anthropologist* 95(1):51-72.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. and A. C. Wilson. 1991. African Populations and the Evolution of Human Mitochondrial DNA. *Science* 253:1503-1507.

Watson, E., Forster, P., Richards, M., and H.-J. Bandelt. 1997. Mitochondrial Footprints of Human Expansions in Africa. *American Journal of Human Genetics* 61:691-704.

Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189-225 in *Horizons in biochemistry*, (eds.) M. Kasha and B Pullman. Academic Press, New York.

Chapter Two

CURIOUS PARALLELS AND CURIOUS CONNECTIONS
PHYLOGENETIC THINKING IN BIOLOGY AND HISTORICAL LINGUISTICS[1]

ABSTRACT

In The Descent of Man *(1871), Darwin observed "curious parallels" between the processes of biological and linguistic evolution. These parallels mean that evolutionary biologists and historical linguists seek answers to similar questions and face similar problems. As a result, the theory and methodology of the two disciplines has evolved in remarkably similar ways. In addition to Darwin's curious parallels of process, there are a number of equally curious parallels and connections between the development of methods in biology and historical linguistics. Here we briefly review the parallels between biological and linguistic evolution and contrast the historical development of phylogenetic methods in the two disciplines. We then look at a number of recent studies that have applied phylogenetic methods to language data and outline some current problems shared by the two fields.*

2.1 CURIOUS PARALLELS IN THE DOCUMENTS OF EVOLUTIONARY HISTORY

In *The Descent of Man,* Darwin (1871) noted that the process of evolution is not limited to just the biological realm:

> "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. … We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation." (pp. 89-90)

[1] Based on "Atkinson, Q. D. and R. D. Gray. 2005 Curious Parallels and Curious Connections – Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology* 54(4):513-526."

Many of the fundamental features of biological and linguistic evolution are demonstrably analogous (see Table 2.1 and Croft, 2000). Just as DNA sequences contain discrete heritable units, so too do languages in their grammatical and phonological structures and their vocabularies (lexicons). These may differ from language to language, and can be inherited as the languages are learned by subsequent generations. With knowledge of the processes of linguistic change it is possible to identify homologous linguistic characters that, like homologous biological structures, indicate inheritance from a common ancestor, or *proto-language*. For example, homologous words, or *cognates*, meaning "water" exist in English (*water*), German (*wasser*), Swedish (*vatten*) and Gothic (*wato*), reflecting descent from proto-Germanic (**water*; inferred *proto forms* are denoted with a '*'). Cognates are words of similar meaning with systematic sound correspondences indicating they were related due to common ancestry. Processes of mutation and random drift can operate on linguistic characters, just as they do on genes. An example of lexical mutation, or *innovation* as it is known in linguistics, is the word *boy* which arose at some point after English split off from the other West-Germanic languages (Campbell, 2004). A phonological example, is the unconditional sound change of *t* to *k* in Hawaiian. So the ancestral Polynesian word **tapu* "forbidden" changed to *kapu* and **tolu* "three" changed to *kolu,* and so on (Crowley, 1992). As well as these "point mutations", words, like gene sequences, can show insertions (e.g. Old Swedish **bökr* "books" to *böker;* Campbell, 2004), deletions (e.g. Proto-Oceanic **tupa* "derris root" to Selau *tua;* Blust, 2003) and reversals, or *metathesis* (e.g. Old English *brid* to Modern English *bird*; Campbell, 2004). Linguistic changes, like changes in biological form, are also sometimes structurally and/or functionally linked. Terms for "five", for instance, tend to be correlated with terms for "hand" for obvious reasons. As in biology, mutation and drift create variation that may be subject to selection. For example, Pawley and Syder (1983) found evidence that Darwinian selection pressures have acted on English syntax. Specifically, they identified differences between vernacular and literary English grammar that they argue were "adaptive to the particular conditions imposed by the mode of language use" (p. 577). The fundamental process of language formation involves cladogenesis, where a single lineage splits to form two new languages. Often, as in biology, this is due to geographic separation or migration events. Horizontal gene transfer and hybridisation also have a linguistic equivalent in borrowing between languages. For example, the

English word *mountain* is borrowed from French, *montagne*. Borrowing between languages can produce reticulated evolution in a similar fashion to horizontal gene transfer in plants or bacteria. Extreme cases of contact between languages can produce a form of language "hybrid", as in the case of some *creoles*. For example, Sranan, a creole spoken in Surinam, has elements of English, Dutch, Portuguese and a number of African and Indian languages, although it is essentially English-based. Many creoles involve less mixing, and are instead cut down versions of the language upon which they are based. One current view is that this is partly the result of a linguistic "founder effect", where the complexity of the original (usually plantation) language declines due to the small initial population and is revived with the arrival in greater numbers of speakers from other linguistic backgrounds. Lastly, like species, languages can become extinct and can even be 'fossilized' in the form of ancient texts. For example, we have archaic manuscripts of ancient languages like Hittite, Homeric Greek, Sanskrit and Mayan. A more detailed discussion of the parallels between biological and linguistic evolution, which builds on Hull's (1988) general account of evolutionary processes, can be found in Croft (2000).

TABLE 2.1: Conceptual parallels between biological and linguistic evolution.

Biological Evolution	**Language Evolution**
Discrete heritable units – e.g. genetic code, morphology, behaviour	Discrete heritable units – e.g. lexicon, syntax, and phonology
Homology	Cognates
Mutation – e.g. Base-pair substitutions	Innovation – e.g. Sound changes
Drift	Drift
Natural selection	Social selection
Cladogenesis – e.g. allopatric speciation (geographic separation) and sympatric speciation (ecological/reproductive separation)	Lineage splits – e.g. geographical separation and social separation
Anagenesis	Change without split
Horizontal gene transfer – e.g. hybridisation	Borrowing
Plant Hybrids – e.g. wheat, strawberry	Language Creoles – e.g. Surinamese
Correlated genotypes/phenotypes – e.g. allometry, pleiotropy.	Correlated cultural terms – e.g. 'five' and 'hand'.
Geographic clines	Dialects/Dialect chains
Fossils	Ancient Texts
Extinction	Language death

In 1965, Zuckerkandl and Pauling characterized molecules as "documents of evolutionary history":

> "Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history." (p. 357)

Languages are also documents of evolutionary history. By comparing features between languages linguists can make inferences about their historical relationships and thus gain insight into human prehistory. In attempting to make historical inferences linguists and biologists must ask similar questions: - What is the most reliable type of data? What is the probability of different types of change? Does a tree accurately reflect the evolutionary history or is some other representation more appropriate? Where phylogenies are of interest, which trees are best and how confident can we be in a particular result? It is perhaps not surprising then that biologists and linguists have developed similar phylogenetic methods to answer these questions (see, for example, Platnick and Cameron, 1977 and O'Hara, 1996). In fact, the theory and methodology of evolutionary biology and historical linguistics have evolved along related paths throughout their history. In the following discussion, we will examine the curious parallels and connections between the history of phylogenetic thinking in Western biology and linguistics. We note that this paper is not intended to be a comprehensive description of the history of either field – a large amount of literature already exists on the development of historical linguistics (e.g., Pedersen, 1931; Robins, 1997; Campbell, 2000) and evolutionary biology (e.g., Mayr, 1982; Gould, 2002). Instead, we present a brief comparative history highlighting some of the intriguing interdisciplinary relationships between the two fields.

2.2 IN THE BEGINNING - TWO ANCIENT GREEK OBSESSIONS

The Ancient Greek philosophers were obsessed with archetypal form, especially the definition and classification of plants and animals according to these archetypes. Plato's, perhaps offhand, definition of man as a "featherless biped" was famously rebuffed when the cynic Diogenes presented him with a plucked chicken. Plato's student, Aristotle, was more methodical and was the first to use a hierarchical system of animal classification based on discrete characters (Thompson, 1913). Although he did not outline his classification system formally, Aristotle is generally considered the founder of the comparative method in biology and many of his groupings still hold today (Thompson, 1913; Mayr, 1982). He identified the birds, amphibians, fish and mammals (with the exception of whales which he placed in their own group) as distinct groups and part of a broader group of animals "with blood". He also identified "bloodless" cephalopods, higher crustaceans, insects and the "lower animals". The

Greek preoccupation with classification facilitated a hierarchical conception of the natural world, which was a precursor to modern phylogenetic classification. However, these groupings were presented as immutable archetypes - they had always existed and always would exist. Aristotle's classification was thus ahistorical. Although the Greeks identified biological 'kinds', they had no concept of species existing through time. The prevailing view in Aristotle's time, a view that persisted up until the 17th century, was that organisms could arise through spontaneous generation from non-living matter. As a result, questions about species lineages and historical relationships never arose.

In contrast, when the Ancient Greek philosophers turned to language, they were obsessed, not with hierarchical relationships, but with explaining the process of change in linguistic structures. The writings of Homer (c. 730BC) were of fundamental importance in Greek schooling. By the time of Socrates (469-399BC) it was evident that the Greek language had changed considerably since Homer. The study of language was largely oriented towards keeping tabs on these changes in Greek (Campbell, 2000). Socrates sought to relate contemporary words to *prota onomata* or "first words", the primordial words of some ancient Greek tongue (Percival, 1987). Socrates believed that change was a process of decay and claimed that contemporary words were the product of semantic shift and phonetic degeneration from the *prota onomata*. In Plato's *Republic*, Socrates argued that:

> "…the primeval words have already been buried by people who wanted to embellish them by adding and removing letters to make them sound better, and disfiguring them totally, either for aesthetic considerations or as a result of the passage of time." 414C

As well as laying the foundations for modern biology, Aristotle advanced the Greek understanding of grammatical categories and developed ideas about the nature of linguistic change. He identified four key types of linguistic change that will be familiar to any biologist today – insertion, deletion, transposition and substitution (*Categories* 15a13 and *On Coming-to-be and Passing-away* 314b27, cited in Householder, 1981).

Unfortunately, the Greek inquiry into language change was very Greco-centric – the Greeks were chiefly concerned with Greek (Percival, 1987). Ideas about how Greek

etymology and grammar may have been related to other languages were thus not put forward. Change in the Greek language was merely a process of decay, as the language shifted from some ancestral ideal. Thus, in biology the Ancient Greeks had a hierarchical classification system but no notion of change, whilst in linguistics they understood something of the process of change (albeit involving decay from an ideal) but were not interested in language relationships.

It is worth noting that linguistic traditions in the East, including China, Mesopotamia and India, stretch back earlier than the Greek tradition, and were in many respects more advanced at this time (Robins, 1997). Eastern linguistics had a greater focus on grammar and phonetics than etymology (Pedersen, 1931). For example, the work of the Indian linguist, Panini, circa 5th century BC, comprised a detailed grammar of Sanskrit. However, the Eastern scholars, like the Greeks, were chiefly interested in describing the particulars of their own language and its change from an archaic form. This, combined with very little contact between eastern and Western linguists, meant that Eastern linguistics had little impact on the early development of the Western linguistic tradition (Robins, 1997).

2.3 BEFORE THERE WERE TREES

During the early Christian era and the Middle Ages, the still unchallenged Ancient Greek conception of constantly regenerating 'natural kinds', combined with creationist accounts of the origin of life, made positing historical species relationships both illogical and heretical. In linguistics, creationism promoted some historical thinking – for example, how to get peoples/languages aligned with the descendants of Noah. Unfortunately, the Old Testament account of the creation of all languages following the destruction of the Tower of Babel undermined any attempt to accurately infer the historical relationships between languages. St Augustine (354-430AD, *City of God,* 413–426/427AD) was the first in a long tradition of intellectuals who attempted, sometimes quite creatively, to integrate science and scripture by claiming that all languages had descended from Hebrew (Percival, 1987) – something akin to claiming that all species are descended from the woolly mammoth. It was not until the 17th century and the "Age of Reason" that scholars in both disciplines began to look critically at the accounts offered by scripture.

In linguistics, the effects of this revolution were realized more quickly than in biology. One interesting reason for this was the invention and increasing use of the printing press. The printing press greatly increased the accessibility and quantity of raw material describing foreign languages (Pedersen, 1931). One might even draw an analogy between this boom and the current proliferation of sequence data in genetics. Where previously linguists may have only been exposed to their own and neighbouring languages, Latin and perhaps some Greek, by the 17th century most scholars had access to Greek texts and many of the languages of Europe and the Near East, as well as the newly discovered languages of the Americas (Pedersen, 1931). This proliferation of material increased interest in language comparison. In short, linguistics became oriented towards classification.

One of the first to challenge the idea of a Hebraic root to the languages of Europe was J. J. Scaliger (1540-1609). Scaliger was able to identify Greek, Germanic, Romance and Slavic language groups by comparing the word for *God* between a number of European languages (Pedersen, 1931). He understood homologous characters as reflecting descent from parent to daughter languages and recognized their importance in reconstructing language relationships. Scaliger failed to find (or chose to ignore) any relationships between the main groups and so his explanations were still essentially ahistorical (Pederson, 1931). However, his work, combined with the existing knowledge of linguistic structure and processes of change, provided the raw ingredients for the comparative method in linguistics.

During this time, the biological comparative method also continued to build on the work of Aristotle. Leonardo da Vinci's (1452-1519) detailed anatomical studies compared humans to other species and recognized structural homologies (although most scholars maintained a purely anthropocentric interest in anatomy until the 18th century). Carolus Linnaeus (1707-1778), "the father of modern taxonomy", introduced a hierarchical classification system using precise species descriptions and Georges Cuvier (1769-1832) led the shift from anthropocentric anatomy to comparing anatomical structures between species (Mayr, 1982). Another significant milestone was the work of English naturalist John Ray (1628-1705). He was one of the first to suggest that 'species' existed through time and were not simply the result of spontaneously generating organisms of various 'kinds'. This discovery, which was

initially rejected as heretical, made historical explanations of species diversity possible for the first time.

An important figure in the development of both linguistic and biological theory before the 19th century was the philosopher Gottfried W. von Leibniz (1646-1716). Leibniz was parodied as Dr Pangloss in Voltaire's *Candide* for his belief that the world must be the best of all possible worlds because God had created it. More recently, the "Panglossian paradigm" was revisited in the famous Gould and Lewontin (1979) critique of adaptationism in biology. However, despite his theological idealism Leibniz advocated a dynamic conceptualization of the natural world that was at odds with a theistic account of biological and linguistic diversity. In biology, the Greek conception of immutable archetypes, which fitted so nicely with scripture, was beginning to be challenged. Leibniz (1712) argued that nature is constantly changing, and what is more, this change occurs gradually:

> "Everything goes by degrees in nature, and nothing leaps, and this rule controlling changes is part of my law of continuity." (pp.376)

Leibniz was influential in shifting interest during this time toward processes of change, although it was not until Jean-Baptiste Lamarck (1744-1829) that ideas of change began to be debated seriously and even then creationist accounts were still favoured (Mayr, 1982). Just as historical linguists had found it difficult to integrate their historical hypotheses with accounts offered by scripture, evolutionary accounts of species diversity were hampered by the biblical chronology, which was thought to imply an age for the earth of no more than 6,000 years.

Leibniz (1710) also applied his ideas of gradualism and uniformitarianism to linguistics. He argued that languages, as natural phenomena, must change in a gradual and continuous manner. Leibniz rejected doctrinaire arguments for a Hebraic root to all languages as well as Scaliger's proposition of a large number of unrelated language groups (Pedersen, 1931). Instead, he tried to construct a genealogy of the languages of Europe, Asia and Egypt, arguing that all these languages had descended from some common ancestor (Pedersen, 1931). To this end, he advocated the creation of grammars and dictionaries for all of the languages of the world (Robins, 1997). Although, Leibniz's genealogical conclusions were full of errors, his ideas of

gradualism and uniformitarianism remain fundamental (if somewhat controversial) in linguistics, as in biology.

Notions of gradual, continuous change were also expressed implicitly on the other side of the Atlantic by none other than Thomas Jefferson. In *Notes on the State of Virginia* (written 1781-82), he suggests the possibility of using linguistic data not only to infer historical relationships, but also to infer divergence times:

> "A separation into dialects may be the work of a few ages only, but for two dialects to recede from one another till they have lost all vestiges of their common origin, must require an immense course of time; perhaps not less than many people give to the age of the earth. A greater number of those radical changes of language having taken place among the red men of America, proves them of greater antiquity than those of Asia." (p. 227)

Most modern histories of historical linguistics begin with Jefferson's contemporary, Sir William Jones (1746-1794). In 1786 the British Orientalist and judge identified similarities between Sanskrit, Greek, Celtic, Gothic and Latin that led him to conclude that these languages had "sprung from some common source, which perhaps no longer exists" (pp. 34-35). Jones is generally given credit for the rapid subsequent acceptance of the Indo-European language family and the proliferation of broad comparative studies of Indo-European grammar, phonology and lexicon (Robins, 1997). In fact, Jones' methods were not at all novel, he was not the first to suggest a link between Sanskrit and some European languages and his conclusions were full of errors (Campbell, 2000). Most notably, he mistakenly identified Pahlavi (Persian, an Indo-European language) as Semetic and argued for a genealogical connection between the Hindus, Egyptians, Phoenicians, Chinese, Japanese, and Peruvians (Campbell, 2000)! Nonetheless, Jones' announcement did mark the beginning of a distinctly historical orientation in linguistics that would last throughout the following century. Even today, the nature, location and timing of the "common source" or *common ancestor* of Indo-European is still hotly contested.

By the end of the 18th century, both biologists and linguists recognized varying degrees of relatedness and used the concept of homology. Linguists understood that

diversity could be explained via descent with modification, they had linked homology with common ancestry and they were beginning to concern themselves with genealogical language relationships. Biologists had a highly refined system of classification and were beginning to question the immutability of species.

2.4 TANGLED TREES – EVOLUTION AND THE COMPARATIVE METHOD

Despite the efforts of evolutionists such as Lamarck, the creationist account of the biological world was not seriously challenged until Darwin's (1809-1882) *Origin of Species by means of Natural Selection* (1859). Darwin's theory of evolution by natural selection provided an alternative mechanism that could explain the diversity and complexity of nature without requiring divine influence. Like the linguists, biologists became interested in common ancestry, descent with modification and family trees. Figure 2.1 shows one of Darwin's early sketches from his notebook depicting an evolutionary tree. Although Darwin was not the first to use an evolutionary tree (Lamarck, for example, included a rudimentary tree in his 1809 *Philosophie Zoologique*), *The Origin of Species* elegantly linked affinity between species with proximity of descent, making tree diagrams historically meaningful and useful as explanations of the natural world (Mayr, 1982).

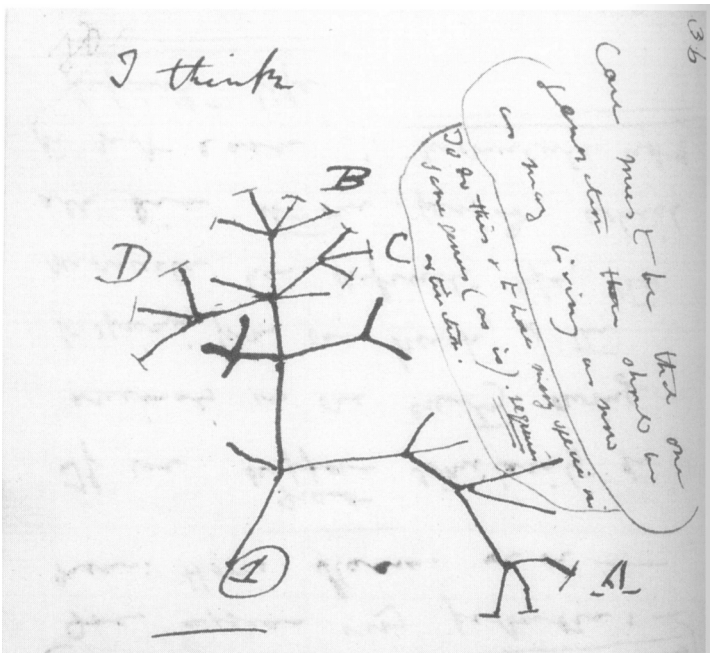


FIGURE 2.1: An early sketch of an evolutionary tree from one of Charles Darwin's notebooks (1837). (Darwin collection, by permission of the Syndics of Cambridge University Library)

In the spirit of the Enlightenment, Darwin did not restrict himself to speculation about biological evolution. In the *Origin of Species* (1859) he muses on the topic of linguistic evolution:

> "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one. Yet it might be that some ancient languages had altered very little and had given rise to few new languages, whilst others had altered much owing to the spreading, isolation, and state of civilisation of the several co-descended races, and had thus given rise to many new dialects and languages. The various degrees of difference between the languages of the same stock, would have to be expressed by groups subordinate to groups; but the proper or even the only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue." (Chap. 13, p.422)

It is tempting to credit Darwin with the subsequent proliferation of language trees in linguistics. Certainly, it seems likely that some borrowing of ideas may have occurred between biology and linguistics at this time. During the first half of the 19th century a string of influential linguists made reference to botany, comparative anatomy and physiology, including Franz Bopp, Jacob Grimm, Rasmus Rask, and Friedrich Schlegel (Koerner, 1983). In 1863, just four years after the *Origin of Species* was first published, the linguist August Schleicher (1821-68) published a paper depicting an Indo-European language tree entitled, *Die Darwinshce Theorie und die Sprachwissenschaft* (see Figure 2.2). The English translation was published in 1869 under the title *Darwinism Tested by the Science of Language*. Schleicher wrote the paper as an open letter to a friend, the biologist and committed Darwinian, Ernst Haeckel (1834-1919). Haeckel introduced Schleicher to the *Origin of Species* in 1863 and the linguist was evidently well informed in biology and the discourse surrounding Darwinian theory (Maher, 1983). However, in his paper Schleicher (1863) pointed out that a "family tree" approach had been part of linguistics since well before Darwin:

> "First, as regards Darwin's assertion that species change in course of time, a process repeated time and again which results in one form arising from another, this same process has long been generally assumed for linguistic organisms….We set up family trees of languages known to us in precisely the same way as Darwin has attempted to do for plant and animal species." (p. 7)

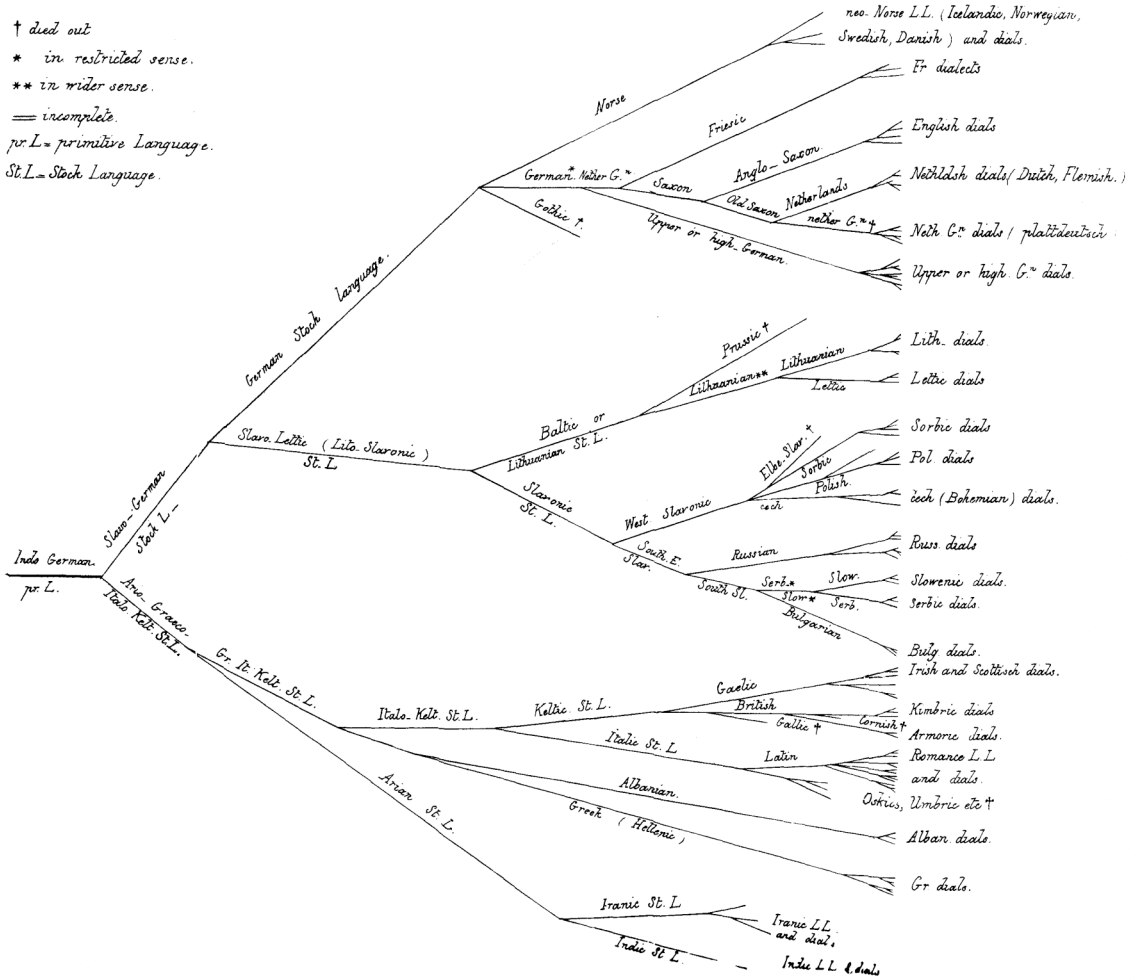


FIGURE 2.2: August Schleicher's (1863) Indo-European *Stammbaum*, or language family tree, from *Die Darwinshce Theorie und die Sprachwissenschaft* (*The Darwinian Theory and the Science of Language*). (Schleicher, A. 1983. Darwinism Tested by the Science of Language. (A. Bikkers, Trans.). John Benjamins Publishing Co., Amsterdam. (original work published in 1863), by permission of John Benjamins Publishing Co.)

Actually, Schleicher had used language trees, or *stammbaum*, in two 1853 publications, some six years before *The Origin of Species* was first published (Koerner, 1983). Koerner credits the idea of the genealogical language tree to Friedrich Schlegel (1772-1829) who introduced a "stammbaum approach" in an 1808

publication on comparative grammar (Schlegel, 1808). Schlegel, however, drew ideas from biology and comparative anatomy, where tree diagrams had been used to represent classification systems, and this may have contributed to his choosing a family tree approach to represent language relationships. The matter is further complicated by the contemporaneous publication of tree diagrams, or *stemma*, by philologists studying manuscript evolution. The first published manuscript phylogeny (see Figure 2.3) was drawn by Carl Johan Schlyter (1747-1805) in 1827 (Collin and Schlyter, 1827) and , as O'Hara (1996) points out, the idea of establishing the most authentic version of a text by reconstructing its ancestry may have been part of the monastic tradition for much longer. Although there could well have been some interdisciplinary cross-fertilization, Darwinian ideas of descent with modification were less revolutionary in linguistics than they were in biology. Phylogenetic understanding and methodology in linguistics had already developed rapidly before Darwin and this continued throughout the 19th century.

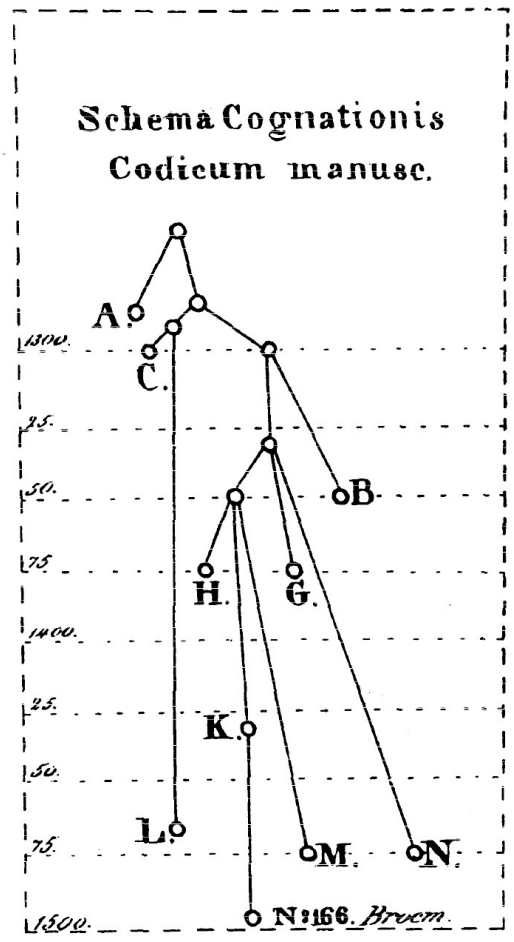


FIGURE 2.3: Carl Schlyter's 1827 manuscript phylogeny, or *stemma*, showing the relationships between copies of the Västgöta Law (Collin & Schlyter, 1827).

In 1822, Jacob Grimm (1785-1863), perhaps more widely known as half of the *brothers Grimm* of fairytale fame, introduced *Grimm's Law,* a series of sound changes in Germanic. Grimm formulated rules for these sound changes seen in cognate words of related languages (for example a change of original **p* to *f* in Germanic languages, as illustrated, for example, by Latin *pater* 'father' and English *father*, where Latin, not being a Germanic language, did not change the **p*, but English, a Germanic language, did change it to *f*). Later, the *Neogrammarians* identified a host of similar patterns and argued that all sound change was governed by regular sound laws. A passionate debate ensued about whether sound change was absolutely regular. Although initially controversial, this form of linguistic uniformitarianism was generally accepted by the end of the 19th century. This debate in linguistics occurred contemporaneously with debates over the relative merits of uniformitarianism in geology and biology. One of the most lively of these debates was between the geologist Charles Lyell (1797-1875), the physicist Lord Kelvin (1824-1907) and Charles Darwin over the age of the earth. In a letter written in 1837 to his sister, Darwin notes a linguistic argument put forward for the age of the earth by John F. W. Herschel (1792-1871) (Darwin, 1985). Herschel expresses his ideas in an 1836 letter to Charles Lyell (published in Cannon, 1961):

> "when we see what amount of change 2000 years has been able to produce in the languages of Greece and Italy or 1000 in those of Germany, France and Spain we naturally begin to ask how long a period must have lapsed since the Chinese, the Hebrew, the Delaware and the Malesass [Malagasy] had a point in common with the German and Italian and each other.- Time! Time! Time!- we must not impugn the Scripture Chronology, but we *must* interpret it in accordance with whatever shall appear on fair enquiry to be the truth for there cannot be two truths."

The Neogrammarians' principle that "sound laws suffer no exceptions" eventually won out with the publication in 1878 of the "neogrammarian manifesto" by Karl Brugmann (1849-1919) and Hermann Osthoff (1847-1909). The comparative method in linguistics developed gradually from the end of the 17th and was perfected with the Neogrammarians. It is a method designed to compare related languages and, on the basis of shared materials, to postulate, or "reconstruct", the sounds, words, and

structures of the parent language from which the related languages descend. It is the most commonly used method for inferring language relationships.

In 1884, the comparative method was further advanced when Brugmann made the important distinction between "innovations" and "retentions" (Hoenigswald, 1990). Innovations are shared characters that were not present in the ancestral form, whilst retentions are shared characters inherited from a common ancestor. Brugmann realized that shared innovations are much more informative for phylogenetic classification than shared retentions. The criterion of shared innovations is now central to working out the family-tree classification of related languages using the comparative method. This distinction was made in biology some 70 years later, in 1950, when Willi Hennig (1913-1976) differentiated symplesiomorphies (shared retentions) from synapomorphies (shared innovations). Despite the methodological similarities between the comparative method in linguistics and biological cladistics, up until very recently historical linguists did not typically use computer algorithms to search for the best language tree(s). This is surprising given that the task of finding optimal trees for even a moderate number of languages is one of considerable computational complexity (Swofford *et al.*, 1996). Some attempts have been made, however, to formalize the criterion implicit in the method (e.g. Gleason, 1959; Hoenigswald, 1960). Thomason and Kaufman (1988) identified six steps used in the comparative method to demonstrate genetic relationships between languages: (1) determining phonological (sound) correspondences in words of the same or related meaning; (2) establishing phonological systems; (3) establishing grammatical correspondences; (4) reconstructing grammatical systems; (5) identifying subgroup of languages; and (6) producing a model of diversification.

Central to the comparative method is the determination of systematic sound correspondences among related languages in order to reconstruct ancestral sounds and hence the most likely series of phonological innovations that lead to the attested sounds in the related languages. Exclusively shared innovations are used to infer historical relationships and construct a language family tree. To return to Grimm's Law as an example, the Germanic language family is characterized by a sound change of initial **p* in the ancestral Proto-Indo-European to *f* in Proto-Germanic. So, the *p,* preserved in its unchanged form in Latin *pater*, Greek *pater* and Sankskrit pitár, was

replaced by an *f* in Proto-Germanic, giving rise to English *father* and Gothic *fadar* and so on in the other Germanic languages (Campbell, 2004). This pattern occurs repeatedly between cognate words in the Germanic languages and their cognates in non-Germanic Indo-European languages (e.g. English *foot* and Greek *podos*). The comparative method also involves the reconstruction of ancestral words in reconstructed (hypothetical) proto-languages (analogous to ancestral character states in biology). For example, based on the types of patterns highlighted above, linguists can reconstruct with relative certainty the Proto-Indo-European word **ped-* for 'foot' (Campbell, 2004). Ancestral reconstructions can sometimes be checked against historically attested ancient languages. For instance, reconstructions based on the comparison of Romance languages (descended from Latin) can often be checked against attested forms in Latin documents.

Some 19th century scholars opposed the Stammbaum ('family tree') model of language evolution. The most often cited (although he was not the first) is Schleicher's student, Johannes Schmidt (1843-1901), who proposed a wave theory of language change (Schmidt, 1872). Schmidt studied dialects and thus became aware of the extent of borrowing between neighbouring populations. The wave theory proposed that language change spread in waves emanating from some epicentre. This process could happen repeatedly and from varying epicentres. As a result, in any given language different words could have different histories. The wave theory was therefore often seen as challenging the neogrammarian conception of language change, though today the two are seen to complement one another, with both needed to get the full history of language (i.e. determine what is inherited and what is diffused). Schmidt's model bears an obvious resemblance to the demic diffusion models used more recently in biology, for example, by Menozzi, Piazza and Cavalli-Sforza (1978).

2.5 AND THEN THERE WERE ALGORITHMS…

During the first half of the 20th century another major methodological revolution occurred in biology. The work of Ronald Fischer (1930) and Theodosius Dobzhansky (1937) and later Sir Julian Huxley, Ernst Mayr and George G. Simpson, allowed Darwinian evolutionary theory to be explained in terms of Mendelian characters of

inheritance and population genetics, thus giving rise to "the modern synthesis". In 1953, Watson and Crick announced the structure of DNA and by the beginning of the 1960's the first protein sequences had been published. This produced much more data than could be analysed by inspection and biologists began working on numerical, algorithmic methods of inferring phylogenies. Felsenstein (2004) provides an excellent overview of the development of numerical phylogenetic methods, beginning with the distance methods of Sokal and Sneath in the late 1950s (Michener and Sokal, 1957; Sneath 1957).

Five years earlier, however, Morris Swadesh introduced similar distance methods into linguistics. Swadesh's (1952, 1955) approach, known as lexicostatistics, used the percentage of shared cognates between languages to produce a pair-wise distance matrix. These distance matrices were then analysed using clustering algorithms to infer tree topologies. Swadesh (1952) also introduced *glottochronology*, based on the idea of a *glottoclock*, or constant rate of lexical replacement. He realized that, under the assumption of constant rates, one could also infer divergence times from the distance data. A decade later, Zuckerkandl and Pauling (1962) introduced the idea of the molecular clock to infer species divergence times in biology. This is, however, almost certainly an example of convergent evolution, not borrowing - Zuckerkandl and Pauling's molecular clock proposal evolved from earlier work in biology linking the distance between species to variation found in haemoglobin (e.g. Reichert and Brown, 1909). Scholars of the time did, however, recognize parallels between modern evolutionary biology and historical linguistics. Stevick (1963) presents "…an extended adaptation to linguistic interests of a biologist's statement about classification" (p. 162). Below is an extract in which Stevick paraphrases several pages of Dobzhansky's (1941) *Genetics and the origin of species* "by substituting linguistic for biological terms and examples":

> "The conclusion that is forced on us is that the discontinuous variation encountered in natural speech, except that based on single feature differences, is maintained by means of preventing the intercommunication of representatives of the now discrete language groups. This conclusion is evidently applicable to discrete groups of any rank whatever, beginning with languages and up to and including branches and families. The development of isolating mechanisms is therefore a *conditio sine qua non* for emergence of

> discrete groups of forms in linguistic development…This conclusion is certainly not vitiated by the well-known fact that the isolation between groups may be complete or only partial. An occasional exchange of language materials, not attaining to the frequency of random interchange, results in the production of some intergrades, without, however, entirely swamping the differences between groups." (p. 163)

Whilst biologists have embraced computational phylogenetic methods, the same cannot be said for historical linguists. Despite some initial enthusiasm, the numerical approaches introduced by Swadesh were heavily criticised and are now largely discredited (Bergsland and Vogt, 1962; Blust, 2000; Campbell, 2004). Criticisms of lexicostatistics and glottochronology tend to fall into four main categories that will be familiar to evolutionary biologists. First, the conversion of lexical character data to distance scores between languages results in the loss of information, reducing the power of the method to reconstruct evolutionary history accurately (Steel, Hendy and Penny, 1988). Using distance data also makes it difficult to deal with polymorphisms (i.e. multiple terms in a language for a given meaning). Second, the clustering methods employed produced inaccurate trees, grouping together languages that evolve slowly rather than languages that share a recent common ancestor (Blust, 2000). Third, language contact and borrowing of lexical items between languages make purely tree-based methods inappropriate (Hjelmslev, 1958; Bateman *et al.*, 1990). And fourth, the assumption of constant rates of lexical replacement through time and across all meaning categories does not hold for linguistic data, making date estimates unreliable (Bergsland and Vogt, 1962). With the possibility of such errors, and no way of quantifying uncertainty, lexicostatistics fell out of favour and methods in biology and linguistics drifted apart.

Two notable exceptions to this trend are the mathematicians David Sankoff and Joseph Kruskal. Sankoff is well-known in biology for his dynamic programming algorithm for counting character state changes on a phylogeny (Sankoff, 1975) and his work on rate variation and invariant sites (e.g. Sankoff, 1990). However, his early work was in lexicostatistical methods and rates of lexical evolution (Sankoff, 1969, 1970, 1973). Sankoff (1973) introduced the gamma-distribution to linguistics as a means of modelling rate variation between words shortly after Uzzell and Corbin

(1971) used a gamma distribution to model rate variation in molecular evolution. More recently, Sankoff and Kruskal have outlined how similar algorithms can be used to solve computational problems in different fields, including biology and linguistics (Sankoff and Kruskal, 1983). Kruskal was also involved with a number of lexicostatistical studies of Indo-European languages in the 1970s that elaborated on Swadesh's methods (Kruskal, Dyen and Black, 1971, 1973; Dyen, Kruskal and Black, 1992).

During the last fifty years, computational phylogenetic methods and statistical inference have revolutionized evolutionary biology. A burgeoning of sequence data has produced enormous databases that can only be investigated using computational techniques. Conversely, the field of linguistics, haunted perhaps by the "ghost of glottochronology past", has remained curiously averse to computational phylogenetic methods.

2.6 THE NEW SYNTHESIS OF BIOLOGY AND LINGUISTICS

> "The poets made all the words and therefore language is the archives of history."
> - The Poet (1844), Ralph Waldo Emerson (1906)

In a landmark paper published in 1988, Cavalli-Sforza *et al*. reported a figure directly comparing a human genetic and linguistic tree. Although extremely controversial (O'Grady *et al*., 1989; Bateman *et al*., 1990; Penny, Watson and Steel, 1993), the Cavalli-Sforza *et al*. paper highlighted the similarities between processes of historical inference in biology and linguistics, as well as the potential importance of linguistic data for inferences about human population history. In the wake of this paper, there has been a proliferation of studies attempting to test hypotheses about prehistoric language expansions (see Barbujani and Pilastro, 1993; Cavalli-Sforza *et al.*, 1994; Semino *et al.*, 2000; Bellwood and Renfrew, 2002; Chikhi *et al.*, 2002; Diamond and Bellwood, 2003; Hurles *et al.*, 2003), and something of a resurgence of interest in computational phylogenetic methods in historical linguistics. This "new synthesis" of biology and linguistics (McMahon and McMahon, 2003) has provided solutions to many of the problems that plagued lexicostatistics and glottochronology. For example, character-based tree-building techniques retain individual character state

information, thus avoiding the problem of information loss associated with distance-based methods. Ringe *et al*. (1997, 2002) used compatibility methods to infer an Indo-European language tree from discrete grammatical, phonological and lexical characters. Gray and Jordan (2000) conducted a parsimony analysis of over 5000 discrete lexical characters to find an optimal tree for 77 Austronesian languages. They then used this tree to test competing scenarios for the settlement of the Pacific. Holden (2002) applied similar methods to test migration scenarios in the Bantu language family and Rexová, Frynta, and Zrzavy (2003) constructed an Indo-European language tree, also using parsimony methods.

In biological phylogenetics over the last fifteen years there has been a gradual move away from parsimony analysis to likelihood models and Bayesian inference of phylogeny (see Swofford *et al.*, 1996; Huelsenbeck *et al.*, 2001). Explicitly modelling the process of evolution makes the assumptions of the method clear, makes it easy to implement more complex and realistic models of sequence evolution and allows different models to be compared easily (Page and Holmes, 1998; Pagel, 2000). Moreover, statistical modelling techniques make it easier to quantify uncertainty in results and to test between competing hypotheses (Swofford *et al.*, 1996). The process of character evolution can also be modelled in linguistics. Pagel (2000) used an explicit likelihood model of lexical replacement to make inferences about different rates of word evolution. More recently, Pagel and Meade (in press) compared rates of change between meanings in Indo-European and Bantu languages. They not only found a relationship between rates of meaning evolution between language families, but also, quite remarkably, that rates of word replacement in these languages are correlated with rates of word use in English today (Pagel and Meade, in press). Gray and Atkinson (2003; see Chapter 3) combined a likelihood model of lexical evolution with Bayesian inference of phylogeny (Huelsenbeck and Ronquist, 2001) to construct a distribution of the most probable trees for the Indo-European language family. We then used a penalised likelihood rate-smoothing algorithm (Sanderson, 2002a, 2002b) to infer the age of the Indo-European language family. Sanderson developed the rate smoothing approach to allow biologists to infer divergence times without having to assume a constant molecular clock. By applying this algorithm to linguistic data we were able to overcome one of the fundamental problems of glottochronolgy (the "glottoclock" is not constant), and thus test between two competing hypotheses for

the age of the Indo-European language family. By analysing linguistic and genetic data in a common analytical framework much more precise inferences about human history should be possible (see Gray and Jordan [2000] and Hurles *et al*. [2003] for attempts to synthesise linguistic, genetic and archaeological inferences about Pacific settlement). Interestingly, quantitative phylogenetic methods have also been used in the study of manuscript evolution to produce a phylogeny of the Canterbury Tales (Barbrook *et al*., 1998).

2.7 FUTURE CHALLENGES

Today, researchers using computational methods in evolutionary biology and historical linguistics aim to answer similar questions and hence face similar challenges. One emerging challenge in computational historical linguistics lies in developing algorithms to determine the probability that lexical characters are cognate (Covington, 1996; Heeringa *et al*. 2000; Kondrak, 2001). The sounds comprising each word must be compared across large sets of data to determine cognacy. Accurate comparisons between words must allow for insertions, deletions and metathesis (reversals) and incorporate complex models of phonological change. These comparisons are fundamentally similar to reconstructions of character change on a phylogeny. Hence, when historical linguists make cognacy judgements using the comparative method they quite rightly consider prior knowledge of the relationships between these languages. Unfortunately, since there is no explicit optimality criterion used to make the cognacy judgements, it is difficult to evaluate the evidence supporting a relationship objectively. Identifying cognates between languages has obvious parallels with the problem of sequence alignment in biology. Biologists must also deal with insertions, deletions and reversals, and are beginning to consider phylogeny (Felsenstein, 2004). As a result, biologists have proposed methods that simultaneously perform alignment and reconstruction of phylogeny (e.g. *Clustal W;* Thompson, Higgins and Gibson, 1994). These methods require further development and should be applicable to historical linguistics.

Model fitting and comparison is another challenge jointly faced by phylogenetic methods in biology and linguistics. Burnham and Anderson (1998) describe model choice as a balance between under- and over-fitting parameters. A model that is too

simple may produce biased results if it fails to capture important parts of the evolutionary process. Conversely, adding extra parameters may improve the apparent fit of a model to data but at the cost of increasing sampling error and computational complexity as there are more parameters to estimate. Traditional lexicostatistics and glottochronological methods implicitly assumed a very simple model of constant rates of change between different meanings and over time. Divergence dates between languages were estimated using the formula:

$$t = (\log c) / (2\log r)$$

where *t* is time, *c* is the percentage of shared cognates, and *r* is the retention rate per thousand years. The retention rate was assumed to be roughly constant at around 81% per thousand years for the Swadesh 200 word list (a list of basic vocabulary terms). However, as mentioned above, this approach produced some obviously incorrect results. In the next chapter, we attempt to overcome these problems by inferring an Indo-European language phylogeny from discrete rather than distance data, by using a gamma distribution to model rate variation between cognate sets, and by using rate smoothing to allow for rate variation through time. The likelihood model used in Chapter 3 assumes that gains and losses of cognates were equally likely. This is not particularly realistic. Assuming that the effects of borrowing can be excluded, it is much more probable that cognates would evolve only once but could be lost multiple times. Atkinson *et al.* (2005; see Chapter 5), have implemented a "Dollo" likelihood model of cognate evolution as a move in the direction of greater realism. Interestingly, the tree topologies and divergence times inferred for Indo-European languages using this model are congruent with those reported in Gray and Atkinson (2003; Chapter 3).

The models of lexical evolution discussed so far all make the standard "rates across sites" assumption. "Rates across sites" models (see Yang, 1993, 1994) essentially modify the independent and identically distributed (IID) rates assumption to allow for rate variation between sites. This is usually achieved by treating the rate of evolution at each site as a variable drawn from some distribution (usually a gamma distribution) and/or by allowing for a proportion of invariant sites. In other words, these models assume that, whilst historical, social, and cultural contingencies can undoubtedly influence the process of linguistic change, fundamental factors such as similarities in the way humans acquire language, and the need to communicate in an expressive and

intelligible way, mean that there are sufficient commonalities in the way different words will evolve to justify the "rates across sites" assumption as a useful starting point. In the words of Ringe *et al*. (2002, p. 61) –

> "Languages replicate themselves (and thus 'survive' from generation to generation) through a process of native-language acquisition by children. Importantly for historical linguistics, that process is tightly constrained".

Warnow *et al.* (in press) reject the "rates across sites" assumption. Instead, they advocate a "no common mechanism" model (see Steel and Penny, 2000) of language evolution in which rates of change are unrelated between meanings and across branches. Such a model does not allow branch lengths or divergence times to be inferred.

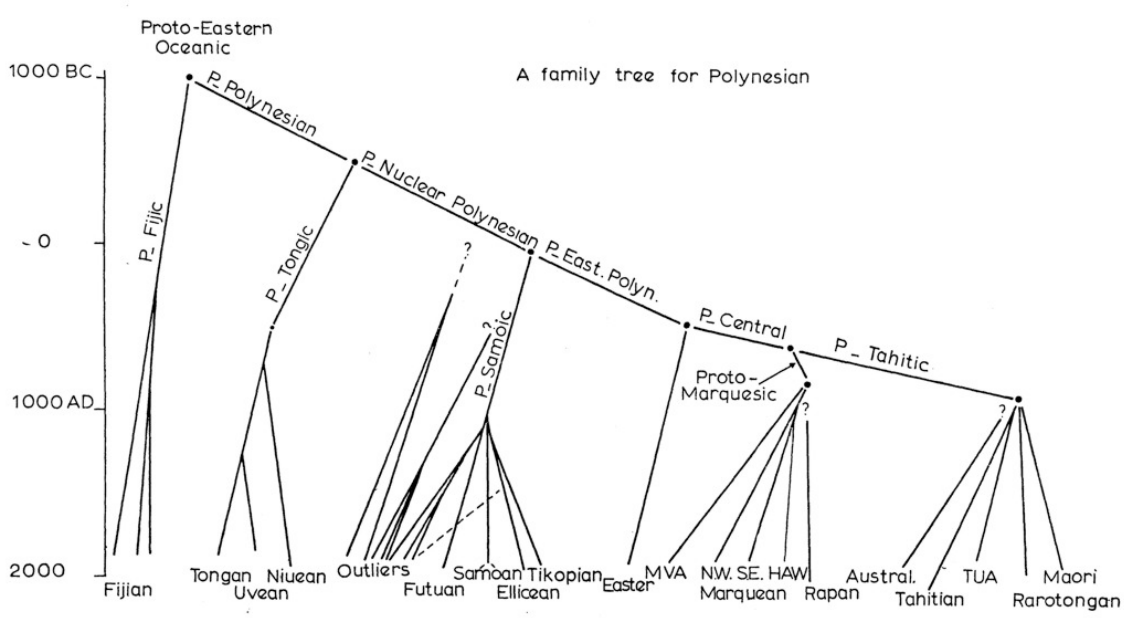


FIGURE 2.4: A phylogenetic tree for Polynesian languages from Green (1966). The tree is unusual for a language tree in that it depicts divergence times and suggests an increase in the rate of linguistic change in Eastern Polynesia (note the flattening of the branch across the base of Eastern Polynesia). (courtesy of the Journal of the Polynesian Society).

In studies of biological evolution, investigations of factors such as generation time, body size, temperature, metabolic rate and population size that may affect the rate of molecular evolution are a major area of current inquiry (e.g. Bromham and Penny, 2003). A similar list of factors has been proposed to affect the rate of linguistic evolution. Nettle (1999), for example, found that in a plausible computer simulation smaller speech communities will have higher rates of change and a greater probability of borrowing. In an explicit analogy with biological evolution, Green (1966; Kirch

and Green, 1987) suggested that the successive "founder events" that occurred in the settlement of the Pacific led to accelerated rates of linguistic change (see Figure 2.4). Other researchers have suggested that factors like contact between languages and the population mobility might also affect language evolution rates (e.g. Blust, 2000; Pawley, 2002). One benefit that could follow from the use of explicit likelihood models in studies of language evolution is the possibility of rigorous comparative tests of hypotheses about factors that affect the rate of linguistic evolution.

Another important challenge in both computational biology and historical linguistics lies in developing methods to investigate reticulate evolution. Tree models of evolution dominate historical linguistics, just like evolutionary biology. A persistent criticism of this approach is that human history is far from tree-like (Terrell, 1988; Moore, 1994; Terrell, Kelly and Rainbird, 2001). Not only might patterns of genetic, linguistic, and cultural diversity reflect different histories (Bateman *et al.*, 1990), each of these histories might be strikingly reticulate. Casual consideration of the history of the English language would lead one to believe that language evolution is anything but tree-like. English is a veritable fruit salad of a language with chunks of vocabulary from the Celts, Romans, Angles, Saxons, Jutes, Vikings, Normans, and slices of Latin, French, Greek, and Italian tossed with some more recent garnishes from Arabic, Persian, Turkish and Hindi. There is even the odd Polynesian borrowing like "tattoo". Ninety-nine percent of words in the Oxford English Dictionary are in fact borrowings from other languages (McWhorter, 2001), and over fifty percent of the total English lexicon comes from Romance languages post the Norman conquest. This figure, however, falls to around 6% for basic vocabulary such as the Swadesh 200 word list (Embleton, 1986). In the case of the Indo-European language family, a number of extinct languages are attested in ancient texts and for many sub-groups linguists have been able to reconstruct a detailed account of the sound changes that occurred since Proto-Indo-European. This makes it possible to identify many, but not all, of the borrowed terms, which do not fit into the regular pattern of sound change. However, most language families are not so well understood. In such cases, the existence of dialect chains, borrowing of cultural terms and contact between languages can therefore pose major problems for attempts to infer language trees. Similarly, biologists studying plants and prokaryotic evolution must deal with hybridization and horizontal gene transfer. Hence, an important challenge for both

biologists and historical linguistics is the development of methods to investigate reticulate evolution. Minett and Wang (2003) developed a method to detect borrowing between languages through identifying incompatible characters on a phylogeny. They propose that in some circumstances even the direction of borrowing may be determined. In biology, a number of new methods have recently been proposed for visualising conflicting signals in the data (e.g. split decomposition and *NeighborNet* analysis, see Huson, 1998; and Bryant and Moulton, 2002). Bryant, Filimon and Gray (2005) have outlined how these methods can be used to investigate conflicting signal caused by lexical borrowing. One test of these methods is their ability to recover evidence of known reticulation. Sranan is a creole language developed by African slaves in Surinam. The English established Surinam on the northern coast of South America in 1651 as a slave colony. However, Dutch has been the official language since 1667 and hence Sranan's lexicon contains words derived from both English and Dutch (McWhorter, 2001). Figure 2.5 shows the split decomposition and *NeighborNet* graphs reported by Bryant *et al.* (2005) for some Germanic languages, including Sranan. Both analyses recovered the conflicting signal generated by Sranan's hybrid history. A further challenge in both biology and linguistics is to explicitly model reticulate evolution. As is described in Chapter 5, Atkinson *et al.* (2005) have taken a step along this road, using an evolutionary model that incorporates word borrowing between random languages to synthesize data and test the robustness of their results to borrowing.

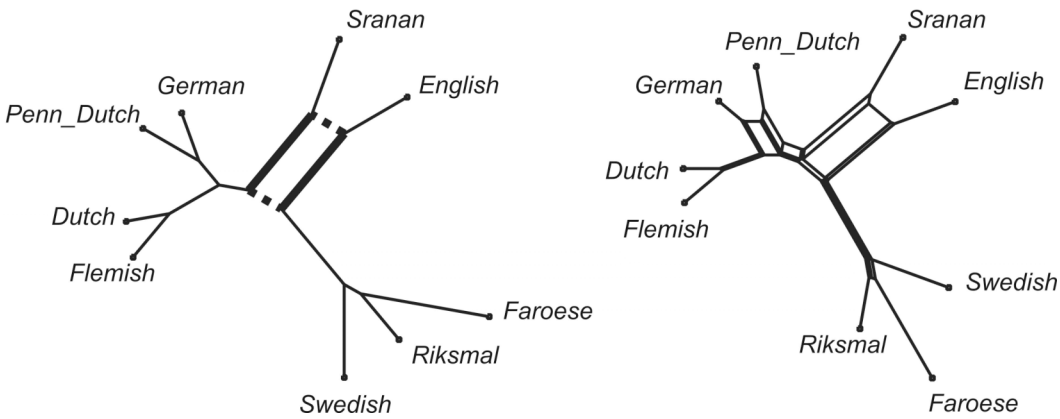


FIGURE 2.5: The networks produced by split decomposition (left) and NeighborNet analyses (right) for a selection of Germanic languages (from Bryant *et al*, 2005). Both networks display the conflicting signal introduced by the creole Sranan. In the split decomposition graph (left), the split grouping Sranan and English is shown in bold, while the conflicting split grouping Sranan with German, Penn. Duct, Dutch, and Flemish is shown as a dotted line. (From Mace, R, Holden, C and Shennan, S (eds), The Evolution of Cultural Diversity: A Phylogenetic Approach, Chapter 5 (© UCL Press, 2005))

Finally, in historical linguistics, as in biology, there is the question of whether it is possible to infer distant genetic relationships reliably. Campbell (2004) identifies a number of controversial attempts to establish linguistic super-families, including Nostratic (comprising Indo-European, Uralic, Altaic, Dravidian, Kartvelian and Afroasiatic), Amerind (comprising all of the languages of the Americas except Eskimo-Aleut and Na-Dene) and even Proto-World, the global mother tongue (Shevoroshkin, 1990). The problem with making inferences about these very deep relationships is that languages change much more quickly than gene sequences. Most linguists believe that after about 8,000-10,000 years, the comparative method breaks down as it becomes impossible to differentiate between homology and chance resemblances or borrowings (see Nichols, 1992). For instance, although Maori *mata* "eye" and Modern Greek *mati* "eye" appear cognate, the resemblance is actually due to chance. Linguists are thus highly sceptical of arguments for ancient language relationships, especially when cognacy judgements are made with less than the normal standard of rigour. Joseph Greenberg (1987) and Merritt Ruhlen (1994) proposed a 12,000-year-old "Amerind" family of Native American languages based on a technique called mass-comparison. They examine large numbers of words between languages in search of words with a similar form and related meanings. For example, Ruhlen (1994, p. 168) offered as evidence for Amerind, words ostensibly related to a hypothetical Proto-Amerind term **t'ana* "child, sibling". As Campbell (2003) points out, the semantic variation Ruhlen allowed (meanings including small, woman, cousin, son-in-law, old man, friend and some 15 other terms) coupled with relatively loose phonetic matches (Ruhlen treats *tsuh-ki* and *u-tse-kwa* as related to **t'ana*) make chance resemblance highly likely. Campbell (2003) goes on to cite examples of words from English (*son*), German (*tante* "aunt") and Maori (*tiena* "younger sibling") that would be misidentified as related by Ruhlen's criteria. On a different time scale, in biology the much less controversial, but still highly contentious, tree of life continues to provoke debate (Gribaldo and Commarano, 1998). One challenge is to push the depth at which it is feasible to reconstruct a phylogeny back further and to develop criteria for accepting or rejecting genetic relationships. Pagel (2000) has shown that some words evolve slowly enough to make it possible to, at least in principle, resolve 20,000-year-old language relationships. The practical challenge of discriminating these deep homologies from more recent borrowings and chance similarities still remains however. A possible solution is to

analyse the kind of abstract grammatical characters that are claimed to have slower rates of evolutionary change (Nichols, 1992). Deep relationships may then be able to be resolved by combining different forms of linguistic data in a single analysis, each with a different model. Biologists have developed methods to combine nucleotide data with protein, restriction site, gene order and morphological data (e.g. *MrBayes,* Huelsenbeck and Ronquist, 2001). Linguists may be able to achieve the same benefit by treating lexical, grammatical and phonological data simultaneously. Finally, by analysing synthetic data on a phylogeny under a given model, we can measure the ability of a given method to reconstruct deep relationships from different data types. Alternatively, Mossel and Steel (2005) have proposed analytical methods for assessing the extent to which deep phylogenetic relationships can be inferred from a number of biological data types. Again, this has obvious applications to linguistics.

These common challenges are a reflection not only of the curious parallels of process that exist between biological and linguistic evolution, but they also reflect over two millennia of co-evolution between research in biology and historical linguistics. In the light of such parallels it seems likely that biology and linguistics will remain curiously, and let's hope productively, connected.

2.8 REFERENCES

Atkinson, Q. D., Nicholls, G., Welch, D. and R. D. Gray. 2005. From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2):193–219.

Barbujani, G., and A. Pilastro. 1993. Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proceedings of the National Academy of Science USA* 90:4670– 4673

Barbrook, A. C., C. J. Howe, N. Blake and P. Robinson. 1998. The phylogeny of the Canterbury Tales. *Nature* 394:839.

Bateman, R., I. Goddard, R. O'Grady, V. Funk, R. Mooi, W. Kress, and P. Cannell. 1990. Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* 31:1-24.

Bellwood, P. and C. Renfrew. (eds.) 2002. *Examining the farming/language hypothesis.* MacDonald Institute for Archaeological Research, Cambridge.

Bergsland, K., and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3:115–153.

Blust, R. 2000. Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. Pages 311-332 in *Time depth in historical linguistics.* (C. Renfrew, A. McMahon, and L. Trask, eds.). The McDonald Institute for Archaeological Research, Cambridge.

Blust, R. 2003. Vowelless words in Selau. Pages 143-152 in *Issues in Austronesian historical phonology.* (J. Lynch, ed.). Pacific Linguistics, The Australian National University, Canberra.

Bromham, L., and D. Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* 4:216-224

Brugmann, K. and H. Osthoff. 1878. Preface to Morphologische Untersuchungen auf dem Gebiet der indogermanischen Sprachen. Vol. 1. Winfred Lehmann, trans. In Winfred Lehmann (ed). *A Reader in Nineteenth-Century Historical Indo-European Linguistics*.
Indiana University Press, Bloomington. 1963.

Bryant, D., and V. Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. Workshop in Algorithms for Bioinformatics, Proceedings 2002:375–391.

Bryant, D., F. Filimon, and R. D. Gray. 2005. Untangling our past: Pacific settlement, phylogenetic trees and Austronesian languages. Pages 69-85 in *The evolution of cultural diversity: Phylogenetic approaches.* (R. Mace, C. Holden, and S. Shennan, eds.). UCL Press, London.

Burnham, K. P. and D. R. Anderson. 1998. *Model selection and inference: A practical information-theoretic approach.* Springer, New York.

Campbell, L. 2000. The history of linguistics. Pages 81-104 in *The handbook of linguistics.* (M. Aronoff and J. Rees-Miller, eds.). Blackwell Publishing, Oxford.

Campbell, L. 2003. How to show languages are related: methods for distant genetic relationship. Pages 262-282 in *The handbook of historical linguistics.* (B. D. Joseph and R. D. Janda, eds.). Blackwell Publishing, Malden (MA).

Campbell, L. 2004. *Historical linguistics: An introduction. 2nd edition.* Edinburgh University Press, Edinburgh.

Cannon, W. 1961. The impact of uniformitarianism: Two letters from John Herschel to Charles Lyell, 1836-1837. *Proc. Amer. Phil. Soc.* 105:301-14.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi, and J. Mountain. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. USA*. 85:6002-6006.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The history and geography of human genes.* Princeton University Press, Princeton, N.J.

Chikhi, L., R. A. Nichols, G. Barbujani, and M. A. Beaumont. 2002. Y genetic data support the neolithic demic diffusion model. *Proc. Natl. Acad. Sci.*, 99:11008–11013.

Collin H. S. and C. J. Schlyter. 1827. *Corpus Iuris Sueo-Goto-rum Antiqui. Z.* Haeggstrom, Stockholm, 1.

Covington, M. 1996. An algorithm to align words for historical comparison. *Computational Linguistics* 22:481-496.

Croft, W. 2000. *Explaining language change: An evolutionary approach.* Pearson Education Ltd., Harlow

Crowley, T. 1992. *An introduction to historical linguistics. 3rd Ed.* Oxford University Press, Oxford.

Darwin, C. 1859. *The origin of species by means of natural selection.* Oxford University Press, Oxford.

Darwin, C. 1871. *The descent of man.* Murray, London.

Darwin, C. 1985. S. Smith and F. Burkhardt. *The correspondence of Charles Darwin*. Cambridge University Press, Cambridge.

Diamond, J., and P. Bellwood. 2003. Farmers and their languages: The first expansions. *Science* 300:597.

Dobzhansky, T. 1937. *Genetics and the origin of species.* Columbia University Press, New York.

Dobzhansky, T. 1941. *Genetics and the origin of species (2nd ed.)*. Columbia University Press, New York.

Dyen, I., J. B. Kruskal, and P. Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5):1-132. American Philosophical Society, Philadelphia.

Embleton, S. 1986. *Statistics in historical linguistics*. Brockmeyer, Bochum.

Emerson, R. 1906. *Essays, 1st and 2nd series*. Dent, London.

Felsenstein, J. 2004. *Inferring phylogenies.* Sinauer Associates Inc., Massachusetts.

Fisher, R. A. 1930. *The genetical theory of natural selection*, Oxford University Press, Oxford.

Gleason, H. A. 1959. Counting and calculating for historical reconstruction. *Anthropological Linguistics* 1:22-32.

Gould, S. J. and R. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B. Biol. Sci.* 205: 581-98.

Gould, S. 2002. *The structure of evolutionary theory*. Belknap Press, Cambridge.

Gray, R. D., and Q. D. Atkinson. 2003. Language tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 405:435-439.

Gray, R., and F. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405:1052–1055.

Green, R. C. 1966. Linguistic subgrouping within Polynesia: the implications for prehistoric settlement. *Journal of the Polynesian Society* 75:6-38.

Greenberg, J. H. 1987. *Language in the Americas*. Stanford University Press, Stanford, (CA).

Gribaldo, S., and P. Commarano. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *Journal of Molecular Evolution* 47:508-516.

Grimm, J. 1822. *Deutsche grammatik. Part I: Zweite ausgabe*. Dieterich, Gottingen.

Heeringa, W., J. Nerbonne, and P. Kleiweg. 2002. Validating dialect comparison methods. Pages 445-452 in *Classification, automation, and new media*. (W. Gaul, and G. Ritter, eds.). Proceedings of the 24th Annual Conference of the Gesellschaft fur Klassifikation e. V., University of Passau, March 15-17, 2000, Springer, Berlin, Heidelberg and New York.

Hennig, W. 1950. *Grundzuge einer theorie der phylogenetischen systematik*. Deutscher Zentralverlag, Berlin.

Hjelmslev, L. 1958. *Essai d'une critique de la methode dite glottochronologique.* Proceedings of the Thirty-second International Congress of Americanists, Copenhagen, 1956. Munksgaard, Copenhagen.

Hoenigswald, H. 1960. *Language change and linguistic reconstruction.* The University of Chicago Press, Chicago.

Hoenigswald, H. M. 1990. Language families and subgroupings, tree model and wave theory, and reconstruction of protolanguages. Pages 441-454 in *Research guide on language change* (E. C. Polome, ed.). Trends in Linguistics, Studies and Monographs, 48. Berlin and New York: Mouton de Gruyter.

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: A maximum-parsimony analysis. *Proc. R. Soc. Lond. B*. 269:793–799.

Householder, F. 1981. *The syntax of Appolonius dyscolus*. John Benjamins B. V., Amsterdam.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310.

Hull, D. 1988. *Science as process: an evolutionary account of the social and conceptual development of science.* University of Chicago Press, Chicago.

Hurles, M. E., L. Matisoo-Smith, R. D. Gray, and D. Penny. 2003. Untangling oceanic settlement: the edge of the knowable. *Trends in Ecology and Evolution* 18:531-540.

Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* 1.4:68–73.

Jefferson, T. 1781-1782. reprint of *Notes on the State of Virginia*. (M. D. Peterson, ed.). 1984. Library of America, Literary classics of the United States, New York.

Jones, Sir W. 1786. *Third anniversary discourse: 'On the Hindus'*, reprinted in The collected works of Sir William Jones. V III, 1807. Stockdale, London.

Kirch, P. V. and R. C. Green. History, Phylogeny and Evolution in Polynesia. *Current Antrhopology* 28(4):431-456.

Koerner, K. (ed.) 1983. Preface. In K. Koerner (ed). *August Schleicher: Die Sprachen Europaas in systematischer Übersicht.* John Benjamins Publishing Co., Amsterdam.

Kondrak, G. 2001. Identifying cognates by phonetic and semantic similarity. Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001). 103-110. Pittsburgh. June, 2001.

Kruskal, J., I. Dyen, and P. Black. 1971. The vocabulary method of reconstructing language trees: Innovations and large-scale applications. Pages 361-380 in *Mathematics in the archeological and historical sciences*. (F. R. Hodson, D. G. Kendall, and P. Tautu, eds.). Edinburgh University Press, Edinburgh.

Kruskal, J. B., I. Dyen, and P. Black. 1973. Some results from the vocabulary method of reconstructing language trees. Pages 30-55 in *Lexicostatistics in genetic linguistics.* (I. Dyen, ed.). The Hague, Mouton.

Lamarck, J. B. 1809. *Philosophie zoologique* trans. by Hugh Elliot as Zoological Philosophy: An Exposition with Regard to the Natural History of Animals with introductory essays by David L. Hull and Richard W. Burkhardt Jr., Chicago, 1984.

Leibniz, G. W. 1710. *Miscellanea berolinensia in Berlin memoirs*. Berlin Academy, Berlin.

Leibniz, G. W. 1712. Monadology. In G. W. Leibniz: *Philosophical papers.* 2nd Ed. 1969. (L. E. Loemker, ed.). Reidel, Dordrecht.

Maher, J. 1983. Introduction. In *Linguistics and evolutionary theory.* (K. Koerner, ed.). John Benjamins Publishing Co., Amsterdam.

Mayr, E. 1982. *The growth of biological thought.* Harvard University Press, Cambridge, Massachusetts.

McMahon, A., and R. McMahon. 2003. Finding families: Qualitative methods in language classification. *Transactions of the Philological Society* 101:7-55.

McWhorter, J. H. 2001. *The power of Babel.* Arrow Books, London.

Menozzi, P., A. Piazza, and L. L. Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201:786-792.

Michener, C. and R. Sokal. 1957. A quantitative approach to a problem in classification. *Evolution* 11:130-162.

Minett, J. and W. Wang. 2003. On detecting borrowing: Distance-based and character-based approaches. *Diachronica* 20:289-330.

Moore, J. H. 1994. Putting anthropology back together again: the ethnogenetic critique of cladistic theory. *American Anthropologist* 96:925-948.

Mossel E. and M. Steel. 2005. How much can evolved characters tell us about the tree that generated them? Pages 384-412 in *Mathematics of Evolution and Biology* (O. Gascuel ed.), Oxford University Press.

Nettle, D. 1999. Is the rate of linguistic change constant? *Lingua* 108:119-136.

Nichols, J. 1992. *Linguistic diversity in space and time*. University of Chicago Press, Chicago.

O'Grady, R., I. Goddard, R. M. Bateman, W. A. DiMicheal, V. A. Funk, W. J. Kress, R. Mooi, P. F. Cannell. 1989. Genes and tongues. *Science* 243:1651.

O'Hara, R. 1996. *Trees of history in systematics and philology*. Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano, 27: 81–88.

Page, R. D. M., and E. C. Holmes. 1998. *Molecular evolution: phylogenetic approach*. University Press, Cambridge.

Pagel, M. 2000. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. *Pages 189–207 in Time depth in historical linguistics.* (C. Renfrew, A. McMahon, and L. Trask, eds.). McDonald Institute for Archaeological Research, Cambridge.

Pagel, M., and A. Meade. In press. Estimating rates of meaning evolution on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages.* (J. Clackson, P. Forster and C. Renfrew, eds.). McDonald Institute for Archaeological Research, Cambridge.

Pawley, A. 2002. The Austronesian dispersal: languages, technologies and people. Pages 251-274 in *Examining the farming/language hypothesis.* (Bellwood, P. and C. Renfrew, eds.) MacDonald Institute for Archaeological Research, Cambridge.

Pawley, A., and F. Syder. 1983. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7:551-579.

Pedersen, H. 1931. *The discovery of language – Linguistic science in the nineteenth century.* Indiana University Press, Bloomington.

Penny, D. Watson, E. and M. Steel. 1993. Trees from languages and genes are very similar. *Systematic Biology* 42:382-384.

Percival, K. 1987. Biological analogy in the study of language before the advent of comparative grammar. Pages 3-38 in *Biological metaphor and cladistic classification* (H. Hoenigswald and L. Wiener, eds.). University of Pennsylvania Press, Philadelphia.

Platnick, N. and D. Cameron. 1977. Cladistic methods in textual, linguistic and phylogenetic analysis. *Systematic Zoology* 26:380-385.

Reichert, E. and A. Brown. 1909. The differentiation and specificity of corresponding proteins and other vital substances in relation to biological classification
and organic evolution. Carnegie Institute Washington. Pub. No. 116.

Rexová, K., D. Frynta, and J. Zrzavy. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19:120–127.

Richards, M., V. Macaulay, E. Hickey, E. Vega, B. Sykes, V. Guida, C. Rengo, D. Sellitto, F. Cruciani, T. Kivisild, R. Villems, M. Thomas, S. Rychkov, O. Rychkov, Y. Rychkov, M. Golge, D. Dimitrov, E. Hill, D. Bradley, V. Romano, F. Cali, G. Vona, A. Demaine, S. Papiha, C. Triantaphyllidis, G. Stefanescu, J. Hatina, M. Belledi, A. Di Rienzo, A. Novelletto, A. Oppenheim, S. Norby, N. Al-Zaheri, S. Santachiara-Benerecetti, R. Scozari, A. Torroni, HJ. Bandelt. 2000. Tracing

European founder lineages in the near eastern mtDNA pool. *American Journal of Human Genetics* 67:1251–1276.

Ringe, D., T. Warnow, A. Taylor, A. Michailov, and L. Levison. 1997. Computational cladistics and the position of Tocharian. Pages 391–414 of Monograph 26 in *The bronze age and early iron age peoples of eastern central Asia.* (V. Mair, ed.). A special volume of the Journal of Indoeuropean Studies.

Ringe, D., T. Warnow, and A. Taylor. 2002. Indo-European and computational cladistics. *Trans. Phil. Soc.* 100:59–129.

Robins, R. 1997. *A short history of linguistics*. Longmans, London.

Ruhlen, M. 1994. *The origin of language: Tracing the origin of the mother tongue*. John Wiley and Sons, New York.

Sanderson, M. 2002a. Estimating absolute rates of evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.

Sanderson, M. 2002b. *R8s, Analysis of Rates of Evolution, version 1.50.* http://ginger.ucdavis.edu/r8s/

Sankoff, D. 1969. *Historical linguistics as a stochastic process*. PhD thesis, McGill University.

Sankoff, D. 1970. On the rate of replacement of word-meaning relationships. *Language* 46: 564-569.

Sankoff, D. 1973. Mathematical developments in lexicostatistical theory. Pages 93-112 In Current Trends in Linguistics 11: Diachronic, areal and typological linguistics (ed. T. A. Sebeok). Mouton, The Hague.

Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics* 28:35-42.

Sankoff, D. 1990. Designer invariants for large phylogenies. *Molecular Biology and Evolution* 7: 255-269.

Sankoff, D. and J. Kruskal. 1983. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Addison-Wesley Publishing Co., Reading, Massachusetts.

Schlegel, F. 1808. Uber die sprache und weisheit der Indier: Ein beitrag zur begrundung der alterthumskunde. In Amsterdam Classics in Linguistics (S. Timpanaro, ed.). John Benjamins Publishing Co., Amsterdam.

Schleicher, A. 1863. *Die Darwinshce theorie und die sprachwissenschaft*. Hermann Bohlau, Weimar.

Schmidt, J. 1872. *Die verwantschaftsverhaltnisse der Indogermanische sprachen*. Herman Bohlau, Weimar.

Shevoroshkin, V. 1990. The mother tongue: How linguists have reconstructed the ancestor of all living languages. *The Sciences*, May/June, p. 20-27.

Semino, O., G. Passarino, P. Oefner, A. Lin, S. Arbuzova, L. Beckman, G. De Benedictis, P. Francalacci, S. Limborska, A. Kouvatsi, M. Marcikiæ, D. Primorac, S. Santachiara-Benerecetti, L.L. Cavalli-Sforza, P. Underhill. 2000. The genetic legacy of palaeolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. *Science* 290:1155–1159.

Sneath, P. 1957. The application of computers to taxonomy. Journal of General *Microbiology* 17:201-226.

Steel, M., M. Hendy, and D. Penny. 1988. Loss of information in genetic distances. *Nature* 333, 494-495.

Steel, M. and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* 17:839-850.

Stevick, R. 1963. The biological model and historical linguistics. *Language* 39, 159-169.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96:453–463.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21:121–137.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-514 in *Molecular systematics, second ed*. (D. M. Hillis, C. Moritz, and B. K. Marble, eds.). Sinauer, Sunderland, Mass.

Terrell, J. 1988. History as a family tree, history as an entangled bank: Constructing images and interpretations of prehistory in the South Pacific. *Antiquity* 62:642-657.

Terrell, J., K. M. Kelly and P. Rainbird. 2001. Foregone conclusions? *Current Anthropology* 42:97-124.

Thomason, S. and T. Kaufman. 1988. *Language contact creolization, and genetic linguistics*. University of California Press, Berkeley (CA).

Thompson, D. 1913. *On Aristotle as a biologist*. Oxford: Clarendon Press. Essay reprinted in D'Arcy Wentworth Thompson, Science and the classics (London: Oxford University Press, H. Milford, 1940).

Thompson, J, D. Higgins and T. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22**:** 4673-4680.

Uzzell, T., and K. Corbin. 1971. Fitting discrete probability distribution to evolutionary events. *Science* 172:1089-1096.

Warnow, T., S. N. Evans, D. Ringe, and L. Nakhleh. In press. Stochastic models of language evolution and an application to the Indo-European family of languages.In *Phylogenetic methods and the prehistory of languages.* (J. Clackson, P. Forster and C. Renfrew. ed.). McDonald Institute for Archaeological Research, Cambridge.

Watson, J. D. and F. H. C. Crick. 1953. Molecular structure of nucleic acids. *Nature* 171: 737-738.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10:1396-1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39:306-314.

Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189-225 in *Horizons in biochemistry* (M. Kasha and B Pullman eds.). Academic Press, New York.

Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. *Journal of Theoretical Biology* 8:357-366.

Chapter Three

LANGUAGE TREE DIVERGENCE TIMES SUPPORT THE ANATOLIAN THEORY OF INDO-EUROPEAN ORIGIN[1]

ABSTRACT

Languages, like genes, provide vital clues about human history (Gray and Jordan, 2000; Pagel, 2000). The origin of the Indo-European language family is 'the most intensively studied, yet still most recalcitrant, problem of historical linguistics' (Diamond and Bellwood, 2003). Numerous genetic studies of Indo-European origins have also produced inconclusive results (Richards, et al., 2000; Semino et al., 2000; Chikhi et al., 2002). Here we analyse linguistic data using computational methods derived from evolutionary biology. We test between two theories of Indo-European origin – the 'Kurgan expansion' and 'Anatolian farming' hypotheses. The former centres on possible archaeological evidence for an expansion into Europe and the near-East by Kurgan horsemen beginning in the sixth millennium BP (Gimbutas, 1973; Mallory, 1989). The latter claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000 to 9,500BP (Renfrew, 1987, 2000). In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800BP and 9,800BP. The results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the Bayesian analysis.

[1] Based on "Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435-439."

3.1 INTRODUCTION

Historical linguists traditionally use the 'comparative method' to construct language family trees from discrete lexical, morphological and phonological data. Unfortunately, whilst the comparative method can provide a *relative* chronology, it cannot provide *absolute* date estimates. A derivative of lexicostatistics, glottochronology, is an alternative, distance-based approach to language tree construction that enables absolute dates to be estimated (Swadesh, 1952, 1955). Glottochronology uses the percentage of shared 'cognates' between languages to calculate divergence times by assuming a constant rate of lexical replacement or 'glottoclock'. Cognates are words inferred to have a common historical origin because of systematic sound correspondences and clear similarities in form and meaning. Despite some initial enthusiasm, the method has been heavily criticised and is now largely discredited (Bergsland and Vogt, 1962; Blust, 2000; Campbell, 2004). Criticisms of glottochronology, and distance-based methods in general, tend to fall into four main categories: first, by summarizing cognate data into percentage scores, much of the information in the discrete character data is lost, greatly reducing the power of the method to reconstruct evolutionary history accurately (Steel, Hendy and Penny, 1988); second, the clustering methods employed tend to produce inaccurate trees when lineages evolve at different rates, grouping together languages that evolve slowly rather than languages that share a recent common ancestor (Swofford *et al*., 1996; Blust, 2000); third, substantial borrowing of lexical items between languages makes tree-based methods inappropriate; and fourth, the assumption of a strict *glottoclock* rarely holds, making date estimates unreliable (Bergsland and Vogt, 1962). For these reasons historical linguists have generally abandoned efforts to estimate absolute ages on the basis of lexical diversity.

Recent advances in computational phylogenetic methods, however, provide possible solutions to the four main problems faced by glottochronology. First, the problem of information loss that comes from converting discrete characters into distances can be overcome by analysing the discrete characters themselves to find the optimal tree(s). Second, the accuracy of tree topology and branch length estimation can be improved by using explicit likelihood models of evolution. Maximum likelihood methods generally outperform distance and parsimony approaches in situations where there are

unequal rates of change (Kuhner and Felsenstein, 1994). Moreover, uncertainty in the estimation of tree topology, branch lengths and parameters of the evolutionary model can be estimated using Bayesian Markov chain Monte Carlo (MCMC; Metropolis *et al*., 1953) sampling methods in which the frequency distribution of the sample approximates the posterior probability distribution of the trees (Huelsenbeck *et al*., 2001). All subsequent analyses can then incorporate this uncertainty. Third, lexical items that are obvious borrowings can be removed from the analysis, and computational methods such as split decomposition (Huson, 1998) and NeighbourNet (Bryant and Moulton, 2002), which do not force the data to fit a tree model, can be used to check for non-treelike signals in the data. Finally, the assumption of a strict clock can be relaxed by using rate smoothing algorithms to model rate variation across the tree. The penalized-likelihood (Sanderson, 2002a, 2002b) method allows for rate variation between lineages whilst incorporating a 'roughness penalty' that penalizes changes in rate from branch to branch. This smoothes inferred rate variation across the tree so that the age of any node can be estimated even under conditions of rate heterogeneity.

3.2 MATERIALS AND METHODS

3.2.1 DATA AND CODING

Data were sourced from Dyen, Kruskal and Black's (1992, 1997) comparative Indo-European database. The database records word forms and cognacy judgments in 95 languages across the 200 items in the Swadesh word list. This list consists of items of basic vocabulary such as pronouns, numerals and body-parts that are known to be relatively resistant to borrowing. For example, while English is a Germanic language it has borrowed around 50% of its total lexicon from French and Latin. However, only about 6% of English entries in the Swadesh 200 word list are clear Romance language borrowings (Embleton, 1986). Where borrowings were obvious Dyen *et al*. (1992) did not score them as cognate, and thus they were excluded from our analysis. 11 of the speech varieties that were not coded by Dyen *et al.* (1992) were also excluded. To facilitate reconstruction of some of the oldest language relationships, we added three extinct Indo-European languages, thought to fit near the base of the tree (Hittite, Tocharian A and Tocharian B). Word form and cognacy judgements for all three

languages were made on the basis of multiple sources to ensure reliability (Hoffner, 1967; Tischler, 1973, 1997; Guterbock and Hoffner, 1986; Gamkrelidze and Ivanov, 1995; Adams, 1999). Presence or absence of words from each cognate set was coded as '1' or '0' respectively to produce a binary matrix of 2449 cognates in 87 languages.

3.2.2 TREE CONSTRUCTION

Language trees were constructed using a 'restriction site' likelihood model of evolution that allows for unequal character-state frequencies and gamma distributed character specific rate heterogeneity *(MrBayes version 2.01*; Huelsenbeck and Ronquist, 2001*)*. We used default 'flat' priors for the rate matrix, branch lengths, gamma shape parameter and site-specific rates. The results were found to be robust to changes in these priors.

The program was run ten times using four concurrent Markov chains. Each run generated 1,300,000 trees from a random starting phylogeny. On the basis of an autocorrelation analysis only every 10,000th tree was sampled to ensure that consecutive samples were independent. A 'burn-in' period of 300,000 trees for each run was used to avoid sampling trees before the run had reached convergence. Log-likelihood plots and an examination of the post burn-in tree topologies demonstrated that the runs had indeed reached convergence by this time. For each analysis a total of 1,000 trees were sampled and rooted with Hittite. The branch between Hittite and the rest of the tree was split at the root such that half its length was assigned to the Hittite branch and half to the remainder of the tree - divergence time estimates were found to be robust to threefold alterations of this allocation.

3.2.3 DIVERGENCE TIME ESTIMATES

11 nodes corresponding to the points of initial divergence in all of the major language sub-families were given minimum and/or maximum ages based on known historical information (see Table 4.5). The ages of all terminal nodes on the tree, representing languages spoken today, were set to zero by default. Hittite and the Tocharic languages were constrained in accordance with estimated ages of the source texts. Relatively broad date ranges were chosen in order to avoid making disputable, *a*

priori assumptions about Indo-European history. A likelihood ratio test with the extinct languages removed revealed that rates were significantly non-clocklike (χ^2=787.3, df=82, p<.001). Divergence time estimates were thus made using the semi-parametric, *penalized likelihood* model of rate variation implemented in *R8s* (*version 1.50*; Sanderson, 2002a). The cross-validation procedure was applied to the majority-rule consensus tree (Figure 3.1a) to determine the optimal value of the rate smoothing parameter. Step-by-step removal of each of the 14 age constraints on the consensus tree revealed that divergence time estimates were robust to calibration errors. For 13 nodes, the reconstructed age was within 390 years of the original constraint range. Only the reconstructed age for Hittite showed an appreciable variation from the constraint range. This may be attributable to the effect of missing data associated with extinct languages and is investigated further below.

3.3 RESULTS AND DISCUSSION

Examining subsets of languages using split decomposition revealed a strong tree-like signal in the data, and a preliminary parsimony analysis produced a consistency index of 0.48 and retention index of 0.76, well above what would be expected from biological data sets of a similar size (Sanderson and Donoghue, 1989). The consensus tree from an initial analysis is shown in Figure 3.1a. The topology of the tree is consistent with the traditional Indo-European language groups (Gamkrelidze and Ivanov, 1995). All of these groups are monophyletic and supported by high posterior probability values. Recent parsimony and compatibility analyses have also supported these groupings, as well as a Romano-Germano-Celtic supergroup, the early divergence of Greek and Armenian lineages (Rexova, Frynta and Zrzavy, 2003), and the basal position of Tocharian (Ringe, Warnow and Taylor, 2002). The consensus tree also reflects traditional uncertainties in the relationships between the major Indo-European language groups. For instance, historical linguists have not resolved the position of the Albanian group and our results clearly reflect this uncertainty (the posterior probability of the Albanian/Indo-Iranian group is only 0.36).

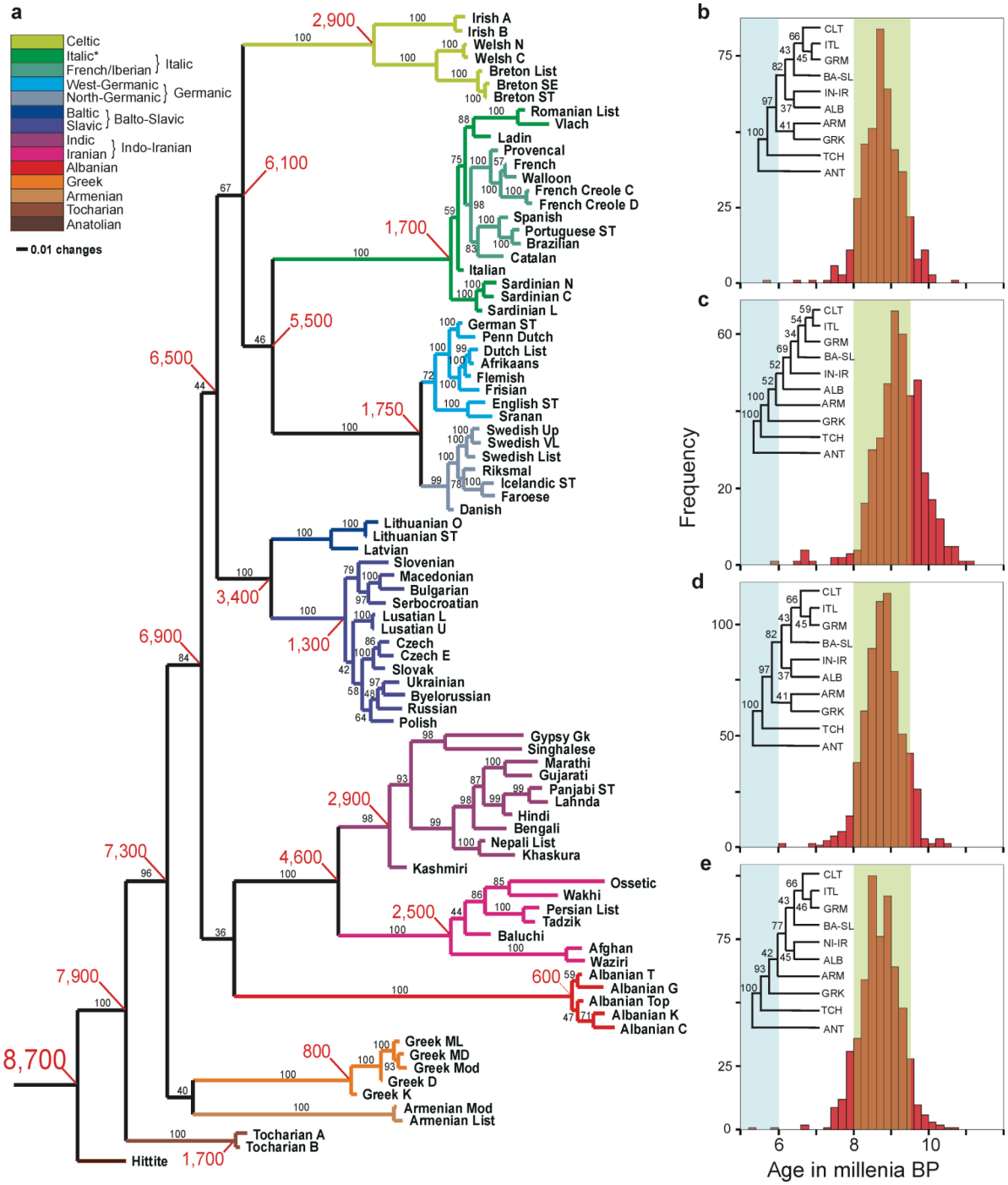


FIGURE 3.1: **a** Majority-rule consensus tree based on the MCMC sample of 1,000 trees. The major language groupings are colour coded. Branch-lengths are proportional to the inferred maximum-likelihood estimates of evolutionary change per cognate. Values above each branch (in black) express the Bayesian posterior probabilities as a percentage. Values in red show the inferred ages of nodes in years BP. *Italic also includes the French/Iberian sub-group. Panels **b-e** show the distribution of divergence time estimates at the root of the Indo-European phylogeny for: **b**, initial assumption set using all cognate information and most stringent constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), ((NorthGermanic, WestGermanic), Italic, Celtic)))]; **c**, conservative cognate coding with doubtful cognates excluded; **d**, all cognate sets with minimum topological constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (NorthGermanic, WestGermanic), Italic, Celtic))]; **e**, missing data coding with minimum topological constraints and all cognate sets. Shaded bars represent the implied age ranges under the two competing theories of Indo-European origin – blue for the Kurgan hypothesis and green for the Anatolian farming hypothesis. The relationship between the major language groups in the consensus tree for each analysis is also shown, along with posterior probability values.

One major advantage of the Bayesian MCMC approach is that any inferences are not contingent upon a specific tree topology. Trees are sampled in proportion to their posterior probability, providing a direct measure of uncertainty in the tree topology and branch-length estimates. By estimating divergence times across the MCMC sample distribution of trees we can explicitly account for variability in the age estimates due to phylogenetic uncertainty, and hence calculate a confidence interval for the age of any node. As described above, divergence times were estimated by constraining the age of 14 points on each tree in accordance with historically attested events (see Table 4.5). We then used penalised-likelihood rate-smoothing to calculate divergence times without the assumption of rate constancy (Sanderson, 2002a, 2002b). Another advantage of the Bayesian framework is that prior knowledge about language relationships can be incorporated into the analysis. To ensure that the sample was consistent with well-established linguistic relationships, we filtered the 10,000-tree sample using a constraint tree (see caption, Figure 3.1b). We used the resulting distribution of 3,500 basal divergence time estimates to create a confidence interval for the age of the Indo-European language family (Figure 3.1b).

A key part of any Bayesian phylogenetic analysis is an assessment of the robustness of the inferences. One important potential cause of error is cognacy judgements. In the initial analysis we included all cognate sets in the Dyen *et al.* (1992, 1997) database in an effort to maximise phylogenetic signal. To assess the impact of different levels of stringency in the cognacy judgements we repeated the analysis with all cognate sets identified by Dyen *et al*. as 'doubtful' removed. 'Doubtful cognates' (for instance possible chance similarities) could falsely increase similarities between languages and thus lead to an underestimate of the divergence times. Unrecognised borrowing between closely related languages would have a similar effect. Conversely, borrowing between distantly related languages will falsely inflate branch-lengths at the base of the tree and thus increase divergence time estimates. With the doubtful cognates removed, the conservative coding lead to a similar estimate of Indo-European language relationships to that produced using the original coding. The relationships within each of the 10 major groups were unchanged. Only the placement of the weakly supported basal branches differed (see Figure 3.1c). More significantly, the divergence time estimates increased, suggesting that the effects of chance similarities and unrecognised borrowings between closely related languages may have

outweighed those of borrowings between distantly related languages. In other words, our initial analysis is likely to have underestimated the age of Indo-European.

The constraint tree used to filter the MCMC sample of trees also contained assumptions about Indo-European history that may have biased the results. We therefore repeated the analyses using a more relaxed set of constraints (see caption, Figure 3.1d). This produced a divergence time distribution and consensus tree almost identical to the original sample distribution (see Figure 3.1d).

Another potential bias lay in the initial coding procedure that made no allowance for missing cognate information. The languages at the base of the tree (Hittite and Tocharian A and B) may appear to lack cognates found in other languages because our knowledge of these extinct languages is limited to reconstructions from ancient texts. This uneven sampling may have increased basal branch-lengths and thus inflated divergence time estimates. We tested this possibility by recoding apparently absent cognates as uncertainties (absent or present) and rerunning the analyses. Whilst divergence time estimates decreased slightly, the effect was only small (see Figure 3.1e).

Finally, although there is considerable support for Hittite (an extinct Anatolian language) as the most appropriate root for Indo-European (Gamkrelidze and Ivanov, 1995; Rexova, *et al*., 2003), rooting the tree with Hittite could be claimed to bias the analysis in favour of the Anatolian hypothesis. We thus reran the analysis using the consensus tree in Figure 3.1a rooted with Balto-Slavic, Greek and Indo-Iranian as outgroups. This *increased* the estimated divergence time from 8,700BP to 9,600, 9,400 and 10,100BP respectively.

The pattern and timing of expansion suggested by the four analyses in Figure 3.1 is consistent with the Anatolian farming theory of Indo-European origin. Radiocarbon analysis of the earliest Neolithic sites across Europe suggest that agriculture arrived in Greece at some time during the ninth millennium BP and had reached as far as Scotland by 5,500BP (Gkiasta *et al*., 2003). Figure 3.1a shows the Hittite lineage diverging from Proto-Indo-European around 8,700BP, perhaps reflecting the initial migration out of Anatolia. Tocharian, and the Greco-Armenian lineages are shown as

distinct by 7,000BP, with all other major groups formed by 5,000BP. This scenario is consistent with recent genetic studies supporting a Neolithic, Near Eastern contribution to the European gene-pool (Richards *et al*., 2000; Chikhi *et al*., 2002). The consensus tree also shows evidence of a rapid period of divergence giving rise to the Italic, Celtic, Balto-Slavic and perhaps Indo-Iranian families, that is intriguingly close to the time suggested for a possible Kurgan expansion. Thus, as Cavalli-Sforza, Mennozi and Piazza (1994) observed, these hypotheses need not be mutually exclusive.

Phylogenetic methods have revolutionised evolutionary biology over the last 20 years and are now starting to take hold in other areas of historical inference (Barbrook *et al*., 1998; Gray and Jordan, 2000; Ringe *et al*., 2002; Holden, 2002; McMahon and McMahon, 2003; Rexova *et al*., 2003). The model-based Bayesian framework employed in this paper offers several advantages over previous applications of computational methods to language phylogenies. This approach allowed us to: - identify sections in the language tree that were poorly supported; explicitly incorporate this uncertainty in tree typology and branch length estimates in our analysis; test the possible effects of borrowing, chance similarities, and Bayesian priors on our analysis; and estimate divergence times without the assumption of a strict glottoclock. The challenge of making accurate inferences about human history is an extremely demanding one, requiring the integration of archaeological, genetic, cultural and linguistic data. The combination of computational phylogenetic methods and lexical data to test archaeological hypotheses is a step forward in this challenging and fascinating task.

3.4 REFERENCES

Adams, D. Q. *A Dictionary of Tocharian B* (Leiden Studies in Indo-European 10). Amsterdam: Rodopi. Available via online database at S. Starostin and A. Lubotsky (Eds.) Database Query to A dictionary of Tocharian B. http://iiasnt.leidenuniv.nl/ied/index2.html (1999).

Barbrook, A. C., Howe, C. J., Blake, N. and Robinson, P. The phylogeny of *The Canterbury Tales*. *Nature* **394**, 839 (1998).

Bergsland, K. and Vogt, H. On the validity of glottochronology. *Current Anthropology* **3**, 115-153 (1962).

Blust, R. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. and Trask, L.) 311-332 (The McDonald Institute for Archaeological Research, Cambridge, 2000).

Bryant, D., F. Filimon, and R. D. Gray. In *The evolution of cultural diversity: Phylogenetic approaches* (eds R. Mace, C. Holden, and S. Shennan) 65-89 (UCL Press, London, 2005).

Campbell, L. *Historical linguistics: An Introduction, 2nd edition*. (Edinburgh University Press, Edinburgh, 2004).

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. *The history and geography of human genes.* (Princeton University Press, Princeton, 1994).

Chikhi, L. Nichols, R. A., Barbujani, G., and Beaumont, M. A. Y genetic data support the Neolithic Demic Diffusion Model. *Proc. Natl. Acad. Sci. USA* **99**, 11008-11013 (2002).

Diamond, J. and Bellwood, P. Farmers and Their Languages: The First Expansions. *Science* **300**, 597 (2003).

Dyen, I., J. B. Kruskal, and P. Black. *An Indoeuropean Classification: A Lexicostatistical Experiment.* American Philosophical Society, Transactions 82(5). Philadelphia (1992).

Dyen, I., Kruskal, J. B. and Black, P. FILE IE-DATA1. World Wide Web. Available online: http://www.ntu.edu.au/education/langs/ielex/IE-DATA1 (1997).

Embleton, S. *Statistics in Historical Linguistics.* (Brockmeyer, Bochum, 1992).

Gamkrelidze, T. V. and Ivanov, V. V. *Indo-European and the Indo-Europeans. (Trends in Linguistics 80)*. (Mouton de Gruyter, Berlin, 1995).

Gimbutas, M. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500-2500 B.C. *Journal of Indo-European Studies* **1**, 163-214 (1973).

Gkiasta, M. Russell, T. Shennan, S. and Steele, J. Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* **77**, 45-62 (2003).

Gray, R.D., and Jordan, F.M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052-1055 (2000).

Guterbock, H. G. and H. A. Hoffner. *The Hittite dictionary of the Oriental Institute of the University of Chicago.* (The Institute, Chicago, 1986).

Hoffner, H. A. *An English-Hittite Dictionary*. (American Oriental Society, New Haven, 1967).

Holden, C.J. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. Roy. Soc. London* **269**, 793-799 (2002).

Huelsenbeck, J. P and Ronquist. F. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754-755 (2001).

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314 (2001).

Huson, D.H. SplitsTree: Analyzing and visualizing evolutionary data, *Bioinformatics* **14**, 68-73 (1998).

Kuhner, M. K., and J. Felsenstein. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution* 11:459–468 (1994).

Mallory, J. P. *In search of the Indo-Europeans: Languages, Archaeology and Myth.* (Thames and Hudson, London, 1989).

McMahon, A. and McMahon, R. Finding families: Quantitative methods in language classification. *Transactions of the Philological Society* **101**, 7-55 (2003).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M.N., Teller, A.H., and Teller. E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087-1091 (1953).

Pagel, M. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. and Trask, L.) 189-207 (The McDonald Institute for Archaeological Research, Cambridge, 2000).

Renfrew, C. *Archaeology and Language: The Puzzle of Indo-European Origins*. (Cape, London, 1987).

Renfrew, C. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. and Trask, L.) 413-439 (The McDonald Institute for Archaeological Research, Cambridge, 2000).

Rexova, K., Frynta, D. and Zrzavy, J. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* **19**, 120-127 (2003).

Richards, M. *et al*. Tracing European founder lineage in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251-1276 (2000).

Ringe, D., Warnow, T. and Taylor, A. IndoEuropean and computational cladistics. *Trans. Philol. Soc.* **100**, 59-129 (2002).

Sanderson, M. *R8s, Analysis of rates of evolution, version 1.50*. (University of California, Davis, 2002a).

Sanderson, M. Estimating absolute rates of evolution and divergence times: A penalised likelihood approach. *Molecular Biology and Evolution* 19, 101–109 (2002b.).

Sanderson, M. J. and Donoghue, M. J. Patterns of variation in levels of homoplasy. *Evolution*, **43**, 1781-1795 (1989).

Semoni, O. *et al*. The genetic legacy of Paleolithic Homo sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155-1159 (2000).

Steel, M.A., Hendy, M.D. and Penny, D. Loss of information in genetic distances. *Nature* **333**, 494-495 (1988).

Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 453-463 (1952).

Swadesh, M. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 121–137 (1955).

Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. in *Molecular Systematics* (eds Hillis, D., Moritz, C. and Mable, B. K.) 407-514 (Sinauer Associates, Inc. Sunderland, Massachusetts, 1996).

Tischler, J. *Glottochronologie und Lexicostatistik*. (Innsbrucker Verlag, Innsbruck, 1973).

Tischler, J. *Hethitisch-Deutsches Worterverzeichnis*. (Probedruck, Dresden, 1997).

Chapter Four

ARE ACCURATE DATES AN INTRACTABLE PROBLEM FOR HISTORICAL LINGUISTICS?
TESTING HYPOTHESES ABOUT THE AGE OF INDO-EUROPEAN[1]

ABSTRACT

Ancient population movements and cultural transformations have left us with a fascinating legacy of archaeological, genetic and linguistic evidence. Synthesizing these historical archives allows us to reconstruct the past with more detail and greater certainty than would be possible on the basis of any discipline on its own. One important aspect of historical inference is the dating of ancient events. Unfortunately, date estimates based on historical linguistics are often viewed with scepticism, limiting our ability to test hypotheses with linguistic data. Instead pre-historians prefer dating methods available in archaeology and genetics. However, as was demonstrated in the previous chapter, the problems associated with traditional linguistic date estimation techniques can be overcome by applying methods from evolutionary biology. Here, we explain our rationale and methodology in more detail, elaborate on the results reported in Gray and Atkinson (2003) and critically discuss the implications of these result for Indo-European origins.

[1] Based on "Atkinson, Q. D. and R. D. Gray. 2006. Are accurate dates an intractable problem for historical linguistics? Pages 269-296 in *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory*. (eds.) C. Lipo, M. O'Brien, S. Shennan & M. Collard. Aldine, Chicago." and "Atkinson, Q. D. and R. D. Gray. In press. How old is the Indo-European language family? Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge."

4.1 LIMITATIONS OF THE COMPARATIVE METHOD AND GLOTTOCHRONOLOGY

Traditionally, the most popular means of extracting diachronic information from linguistic data has been the "comparative method." This method groups languages through knowledge of current and historically attested language syntax, word form, and phonology. By examining systematic sound correspondences among languages, linguists can reconstruct a likely series of innovations. Exclusively shared innovations are used to infer historical relationships and construct a language-family tree. As was described in Chapter 2, for example, the Germanic language family can be characterized by a sound change from an initial **p* in the ancestral Proto-Indo-European to *f* in Proto-Germanic (Campbell, 2004). Hence *pater* in Greek and Latin has changed to *father* in English.

The comparative method provides two useful sources of historical information. First, the inferred family tree reveals major language groupings and the relative chronology of divergence events. Figure 4.1 shows the Indo-European language tree as constructed using the comparative method (Campbell, 1998). We can see that the Celtic languages, for example, form a monophyletic group and that the initial Celtic split occurred between the Insular and Continental varieties. A second source of information lies in an approach known as "linguistic palaeontology." The meaning of inherited words can be used to draw inferences about the environment, culture, and daily life of a protolanguage's speakers. For instance, as is discussed in Chapter 6, we can be reasonably confident that Proto-Mayan culture possessed agriculture because the vocabulary of Proto-Mayan, reconstructed from thirty-one of its descendant languages, exhibits a large number of agricultural terms, including those for "maize," "corncob," and "to harvest" (Campbell, 1997).

Although language trees and linguistic palaeontology are important lines of investigation, neither can produce estimates of absolute time depths. In the case of language family trees, although linguists can infer relative chronology with some confidence, estimates of absolute time depth are at best intuitive guesses based on the perceived similarity between languages. There is no objective criterion for calculating time depth and no measure of the statistical error associated with an estimate.

Conversely, date estimates based on linguistic palaeontology can be obtained only by identifying a reconstructed protolanguage with a particular culture evident in the archaeological record.

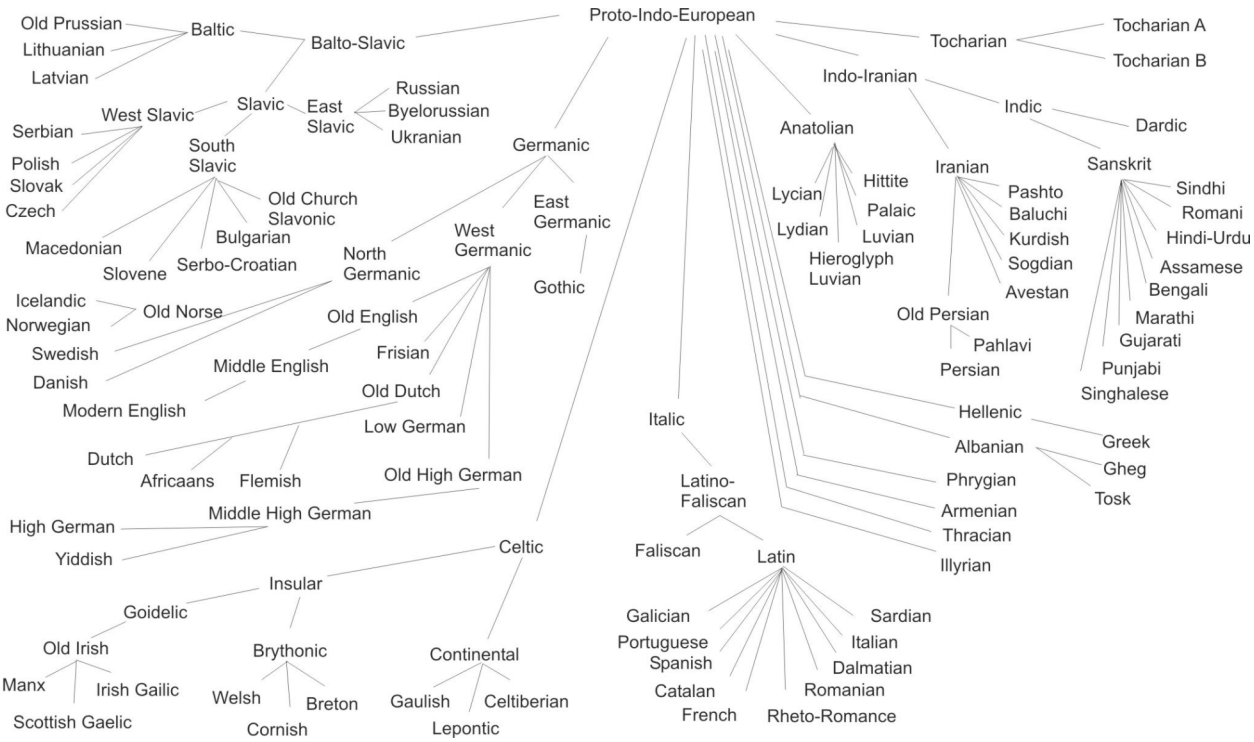


FIGURE 4.1: Indo-European language tree constructed using the comparative method (after Campbell, 1998). Although it is possible to infer the relative chronology of divergence events from this topology, absolute date estimates are at best educated guesses.

TABLE 4.1: A sample dataset of five Swadesh List terms across six Germanic languages (and Greek). The orthography has been simplified in order to emphasize cognacy relationships. Cognacy is indicated by the numbers in superscript.

English	And[1]	Big[1]	Fire[1]	Meat[1]	Rub[1]	Water[1]
German	Und[1]	Gross[2]	Feuer[1]	Fleisch[2]	Reiben[1]	Wasser[1]
Dutch	En[1]	Groot[2]	Vuur[1]	Vleesch[2]	Wrijven[1]	Water[1]
Swedish	Och[2]	Stor[3]	Eld[2]	Kott[3]	Gnida[2]	Vatten[1]
Icelandic	Og[2]	Stor[3]	Eldr[2]	Hold[3]	Nua[3]	Vatn[1]
Danish	Og[2]	Stor[3]	Ild[2]	Kod[4]	Gnide[2]	Vand[1]
Greek	Ke[3]	Meghalos[4]	Fotia[3]	Kreas[5]	Trivo[4]	Nero[2]

An alternative approach is Morris Swadesh's (1952, 1955) lexicostatistics and its derivative "glottochronology". These methods use lexical data to determine language relationships and to estimate absolute divergence times. Lexicostatistical methods infer language trees on the basis of the percentage of shared cognates between languages – the more similar the languages, the more closely they are related. Words are judged to be cognate if they can be shown to be related via a pattern of systematic sound correspondences and have similar meanings. Table 4.1 shows a sample of Swadesh word-list items across a selection of Germanic languages as well as Greek. This information can be used to construct evolutionary language trees. Glottochronology is an extension of this approach to estimate divergence times under the assumption of a

"glottoclock", or constant rate of language change. The following formula can be used to relate language similarity to time along an exponential decay curve: -

$$t = \frac{\log C}{2 \log r}$$

where t is time depth in millennia, C is the percentage of cognates shared and r is the "universal" constant or rate of retention (the expected proportion of cognates remaining after 1000 years of separation; Swadesh, 1955). Usually, analyses are restricted to the Swadesh word list, a collection of 100-200 basic meanings that are thought to be relatively culturally universal, stable and resistant to borrowing. These include kinship terms (e.g. mother, father), terms for body parts (e.g. hand, mouth, hair), numerals and basic verbs (e.g. to drink, to sleep, to burn). For the Swadesh 200 word list, a value of 81% is often used for r. Thus, if two languages share terms for 150 (75%) of the 200 words in the list, then according to the above equation we would infer that they separated about 680 years ago (see Figure 4.2).

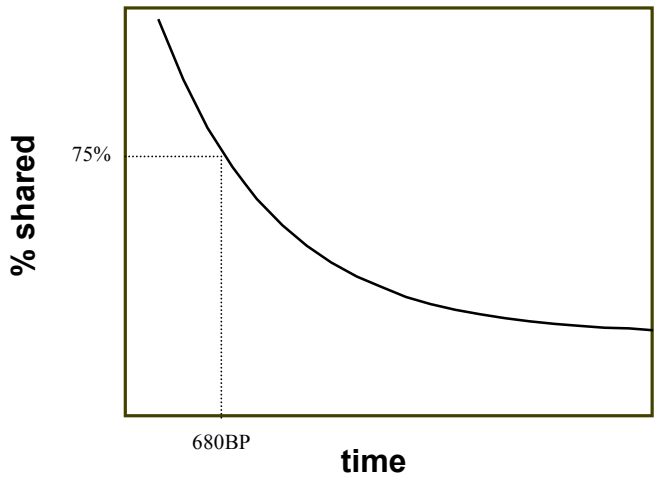


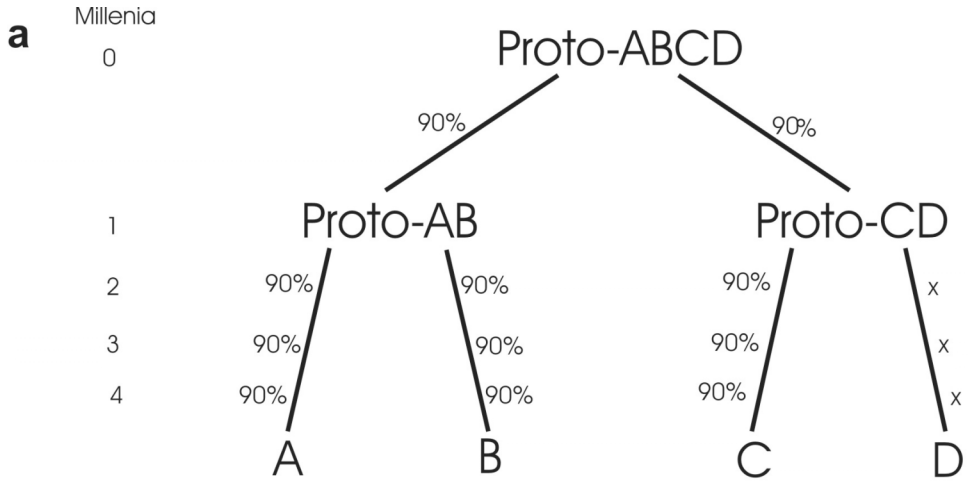
FIGURE 4.2: Curve showing exponential decay of the percentage of shared cognates with time. E.g., if two languages share cognates for 75% of the terms in the Swadesh list then glottochronology would predict a divergence time of 680BP.

Although glottochronology was initially received with enthusiasm, the method has since fallen out of favor as a result of problems identified with both glottochronology and distance-based methods of historical inference in general. These problems can be grouped into four main categories:

1. *Information loss* – The presence or absence of each cognate in each language usually represents a specific evolutionary event and constitutes evidence for grouping a particular subset of languages. By summarizing individual cognate

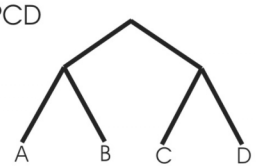
data into percentage-similarity scores between languages, much of the information in the data is lost, greatly reducing the power of the method to reconstruct topology and branch lengths accurately (Steel, Hendy and Penny, 1988).

2. *Inaccurate tree-building techniques* – Distance-based tree-building techniques, such as the "unweighted pair group method with arithmetic mean" (UPGMA) can produce inaccurate trees under some conditions. Where rates of change are unequal, these methods tend to group languages that have evolved more slowly rather than languages that share a recent common ancestor (Blust, 2000). Figure 4.3 illustrates an example of this sort of error. The correct relationship among four languages is shown on the tree. If rates of retention are equal (e.g., 90 percent per millennium) across the tree, then distance methods reconstruct the correct relationships among the languages. Language A will be grouped most closely with language B, and language C will be grouped most closely with language D. However, if language D evolves more slowly than languages A, B, and C after it has split from language D (e.g., retention rate of 80 percent per millennium), then the UPGMA distance methods will group language C with languages A and B rather than the correct reconstruction, with language D.
3. *Borrowing* – Borrowing of words between languages can produce erroneous trees and confound divergence-time estimates. Where borrowing is common, the percentage of shared cognates may be a poor indication of relatedness, producing spurious relationships between languages. Borrowing may also confound divergence-time estimates, given that apparent lexical change depends on the degree of contact between languages as well as on the time since separation.



b **Equal Rates (x=90%)**

Millenia	x	**A:B**	A:C	A:D	B:C	B:D	**C:D**	PAB:PCD
1		**-**	-	-	-	-	-	.81
2	90	**.81**	.66	.66	.66	.66	**.81**	-
3	90	**.66**	.53	.53	.53	.53	**.66**	-
4	90	**.53**	.43	.43	.43	.43	**.53**	-
5	90	**.43**	.35	.35	.35	.35	**.43**	-



c **Rate variation (x=80%)**

Millenia	x	**A:B**	**A:C**	A:D	**B:C**	B:D	C:D	PAB:PCD
1		**-**	**-**	-	**-**	-	-	.81
2	80	**.81**	**.66**	.58	**.66**	.58	.72	-
3	80	**.66**	**.53**	.42	**.53**	.42	.52	-
4	80	**.53**	**.43**	.30	**.43**	.30	.37	-

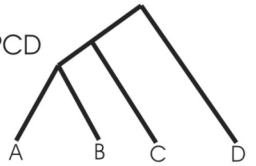


FIGURE 4.3: Illustration of the effect of rate variation on distance based tree building methods (after Blust, 2000). **a** The true topological relationship between four hypothetical languages A, B, C and D. The vertical axis corresponds to time and numbers along each branch are rates of retention per millennia (90%, or x% in the case of the leaf leading to D). **b** The distance calculations for constant rates (x=90%). Here the method retrieves the true tree. **c** The same calculation under conditions of rate heterogeneity (x=80%). In this case the method will reconstruct the wrong ancestral relationships.

4. *Rate variation* – Glottochronology is based on the assumption of a constant rate of lexical change, but languages do not always evolve at a constant rate. Rather, they vary as a result of sociolinguistic, cultural, and environmental factors. Rate variation can occur between languages, between individual meanings within a language, or through time (Pagel, 2000). Bergsland and Vogt (1962) compared a number of contemporary languages with their archaic forms and found significant rate variation between languages. For example, Modern Icelandic and Norwegian were compared with their common ancestral language, Old Norse, spoken between 1,200 and 800 years ago. The retention rate for Norwegian since the time of divergence from Old Norse was found to be 81%, consistent with a 1,000-year-old divergence event. However, Modern

> Icelandic produced a retention rate of 99% for the same interval, wrongly suggesting that it had diverged from Old Norse less than 200 years ago.

These problems have led many linguists to completely abandon any attempt to derive dates from lexical data. For example, Clackson (2000, p. 451) claims that the data and methods "do not allow the question 'When was Proto-Indo-European spoken?' to be answered in any really meaningful or helpful way."

4.2 A BIOLOGICAL SOLUTION TO A LINGUISTIC PROBLEM

Fortunately, none of these problems are unique to linguistics. As we saw in Chapter 2, linguists use information about current and historically attested languages to infer their history in much the same way as evolutionary biologists use DNA sequence, morphological and sometimes behavioural data to construct evolutionary trees of biological species. Questions of relatedness and divergence dates are of interest to biologists just as they are to linguists. As a result biologists must also deal with the problems outlined above: nucleotide sequence information is lost when data is analysed as distance matrices (Steel, Hendy and Penny, 1988); distance-based tree-building methods may not accurately reconstruct phylogeny (Kuhner and Felsenstein, 1994); different genes (and nucleotides) evolve at different rates and these rates can vary through time (Excoffier and Yang, 1999); and finally, evolution is not always tree-like due to phenomena such as hybridisation and horizontal gene transfer (Faguy and Doolittle, 2000).

Despite these obstacles, computational methods have revolutionised evolutionary biology. Rather than giving up and declaring that time-depth estimates are intractable, biologists have developed techniques to overcome each problem. Here, we describe how these biological methods can be adapted and applied to lexical data.

4.2.1 FROM WORD LISTS TO BINARY MATRICES – OVERCOMING INFORMATION LOSS

Distance-based tree-building methods result in information loss in biology just as they do in linguistics. Discrete information about the presence or absence of genes, nucleotides, morphological features, or behavior is condensed into distance scores between species. Steel *et al*. (1988) showed that information is lost when sequence data are converted to distances because the resulting distance matrix does not allow the original sequence to be recovered. Thus biologists prefer to use character-based phylogenetic methods such as parsimony and maximum-likelihood. Rather than using genetic distance as a proxy for evolutionary change, parsimony and maximum-likelihood methods retain individual character information and reconstruct the evolution of each character across a phylogeny. Applied to linguistics, character-based methods ensure that information about the presence or absence of individual word forms or grammatical and phonological features (the characters) can be retained in the analysis.

Lexical data are particularly well-suited to phylogentic analysis because of the large number of well-studied characters available. These can be divided into meaningful evolutionary units known as *cognate sets* (as described above, words are judged to be cognate if they can be shown to be related via a pattern of systematic sound correspondences and have similar meanings). Cognate words from different languages can be grouped into cognate sets that reflect patterns of inheritance. Due to the possibility of unintuitive or misleading similarities between words from different languages, expert knowledge of the sound changes involved is required in order to make cognacy judgements accurately. For example, knowledge of regular sound correspondences between the languages is required to ascertain that the English word *when* is cognate with Greek *pote* of the same meaning. Conversely, English *have* is not cognate with Latin *habere* despite similar word form and meaning.

We can represent the cognate information in Table 4.1 most simply as binary characters in a matrix, where the presence or absence of a particular cognate set in a particular language is denoted by a 1 or 0 respectively. Table 4.2 shows a binary representation of the cognate information from Table 4.1. Each row represents a

language and each column a character (in this case a cognate set). The sequence of 1's and 0's for each language can be viewed as analogous to the gene sequence of a species. Character-based methods use this information directly, without having to convert the data into distance scores. Hence, *unlike lexicostatistics and glottochronology, the number of cognates shared between languages are not counted, nor are pair-wise distances between languages calculated.* Instead, the distribution of cognates is mapped onto an evolutionary language tree (see Figure 4.5) and likely character state changes are inferred across the whole tree. Alternative coding methods are also possible, such as representing the data as a set of meaning categories each with multiple character states. It has been argued that semantic categories are the fundamental "objects" of linguistic change (Evans, Ringe and Warnow, in press) and that binary coding of the presence or absence of cognate sets is thus inappropriate. However, cognate sets constitute discrete, relatively unambiguous heritable evolutionary units with a birth and death (see Chapter 5; Atkinson *et al*., 2005; Nicholls and Gray, in press) and there is no reason to suppose they are any more or less fundamental to language evolution than semantic categories. Further, coding the data as semantic categories makes it difficult to deal with polymorphisms (i.e. when a language has more than one word for a given meaning – e.g. for the meaning "sea" German has both *See* and *Meer*). It also significantly increases the number of parameters required to model the process of evolution. Pagel (2000) points out that, if each word requires a different set of rate parameters, then for just 200 words in 40 languages there are 1278 parameters to estimate. A binary coding of cognate presence/absence information is thus used here.

TABLE 4.2: Germanic (and Greek) cognates from table 4.1 expressed in a binary matrix showing cognate presence (1) or absence (0).

Meaning	**and**			**big**				**fire**			**meat**					**rub**				**water**	
Cognate set	1	2	3	1	2	3	4	1	2	3	1	2	3	4	5	1	2	3	4	1	2
English	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0
German	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
Dutch	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
Swedish	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Icelandic	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0
Danish	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
Greek	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1

4.2.2 LIKELIHOOD MODELS AND BAYESIAN INFERENCE - OVERCOMING INACCURATE TREE-BUILDING METHODS

Recently developed phylogenetic methods in biology employ likelihood models of evolution and Bayesian inference of phylogeny. These methods allow us to overcome the problems identified with the distance-based tree-building methods used in lexicostatistics.

4.2.2.1 Likelihood Models of Evolution

Likelihood evolutionary modelling has become the method of choice in phylogenetics (Swofford *et al*., 1996). Likelihood models have a number of advantages over other approaches. First, we can work with explicit models of evolution and test between competing models. The assumptions of the method are thus overt and easily verifiable. Second, we can increase the complexity of the model as required. For example, as explained below, we were able to test for the influence of rate variation between cognate sets and, as a result, incorporate this into the analysis using a gamma distribution. And third, model parameters can be estimated from the data itself, thus avoiding restrictive *a priori* assumptions about the evolutionary processes involved (Pagel, 1997). By improving the accuracy of tree topology and branch-length estimation, maximum-likelihood methods generally outperform distance and parsimony approaches in situations where there are unequal rates of change (Kuhner and Felsenstein, 1994).

Likelihood methods integrate three related components: – the observed data, a model of character evolution, and an evolutionary tree or a set of trees. This approach is based on the premise that we should favor the explanation that makes our observed data most likely given the assumptions of our model - i.e. we should favour the tree with the highest likelihood score. In biology the observed data are usually a set of gene sequences. For languages, the observed data can take the form of a binary matrix coding cognate presence or absence, as with the example shown in Table 4.2.

Likelihood methods combine the data with an explicit statistical model of character evolution to reconstruct character-state changes across a tree. Likelihood models of

evolution are usually expressed as a rate matrix representing the relative rates of all possible character state changes. More complicated models can be implemented to account for phenomena such as site-specific rate variation and unequal character state frequencies. Table 4.3 shows the "general time-reversible" rate matrix used by biologists to model nucleotide substitution. Each cell corresponds to a probability of character-state change per unit time. A gamma shape parameter can also be added to allow for rate variation between sites. We can model lexical evolution by applying the same approach to linguistic data to allow rate variation between words and through time (Pagel, 2000). Because rate variation can be incorporated into the tree-building process, likelihood methods are not as susceptible to problems associated with rate heterogeneity.

TABLE 4.3: The general time-reversible rate matrix used to model nucleotide evolution (Swofford *et al*., 1996). The model parameters are u (the mean substitution rate), a, b, $c…f$ (the relative rate parameters which allow all of the possible transformations to occur at different rates), and π_A, π_C, π_G and π_T (which represent the relative frequencies of the different bases A, C, G and T).

	A	C	G	T
A	$-u(a\pi_C+b\pi_G+c\pi_T)$	$ua\pi_C$	$ub\pi_G$	$uc\pi_T$
C	$ua\pi_A$	$-u(a\pi_A+d\pi_G+e\pi_T)$	$ud\pi_G$	$ue\pi_T$
G	$ub\pi_A$	$ud\pi_C$	$-u(b\pi_A+d\pi_C+f\pi_T)$	$uf\pi_T$
T	$uc\pi_A$	$ue\pi_C$	$uf\pi_G$	$-u(c\pi_A+e\pi_C+f\pi_G)$

TABLE 4.4 - Simple likelihood time-reversible rate matrix adapted for modelling lexical replacement in language evolution. This is a time-reversible model that allows for unequal equilibrium frequencies of 1's and 0's (cognate presence and absence). The model parameters are u (the mean substitution rate), and π_0 and π_1 (which represent the relative frequencies of 1's and 0's).

	0	1
0	$-u\pi_1$	$u\pi_1$
1	$u\pi_0$	$-u\pi_0$

Here, we are interested in the processes of cognate gain and loss, respectively represented by 0 to 1 changes and 1 to 0 changes on the tree (see Figure 4.5c). We can model this process effectively with a relatively simple two-state time-reversible model of lexical evolution. The rate matrix for such a model is shown in Table 4.4. Each cell represents the relative rate of gaining or losing a cognate. A gamma shape parameter can also be added to allow for rate variation between words. The gamma distribution provides a range of rate categories for the model to choose from when assigning rates to each cognate set. The distribution of these rates is determined by the

gamma shape parameter (α). Figure 4.4 shows the gamma distribution for three possible values of α. For small values of α (e.g. α=0.5), most cognate sets evolve slowly, but a few can evolve at higher rates. As α increases, the distribution becomes more peaked and symmetrical around a rate of 1 (e.g. α=50) – i.e. rates become more equal (Swofford *et al*., 1996). A suitable value for α can be estimated from the data.

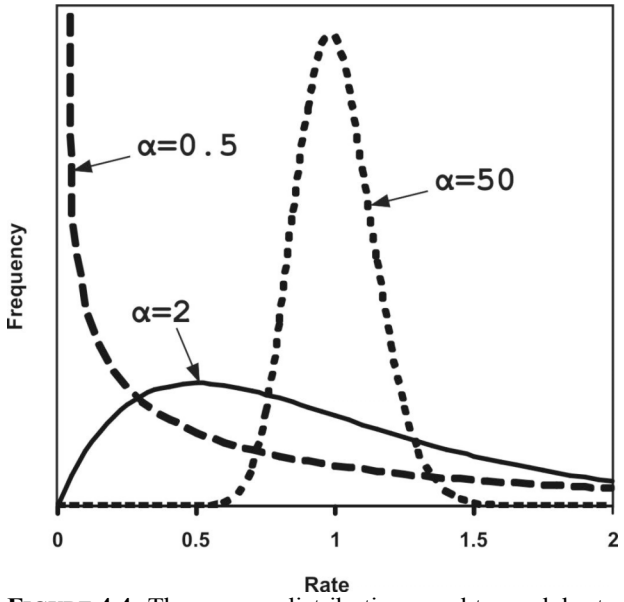
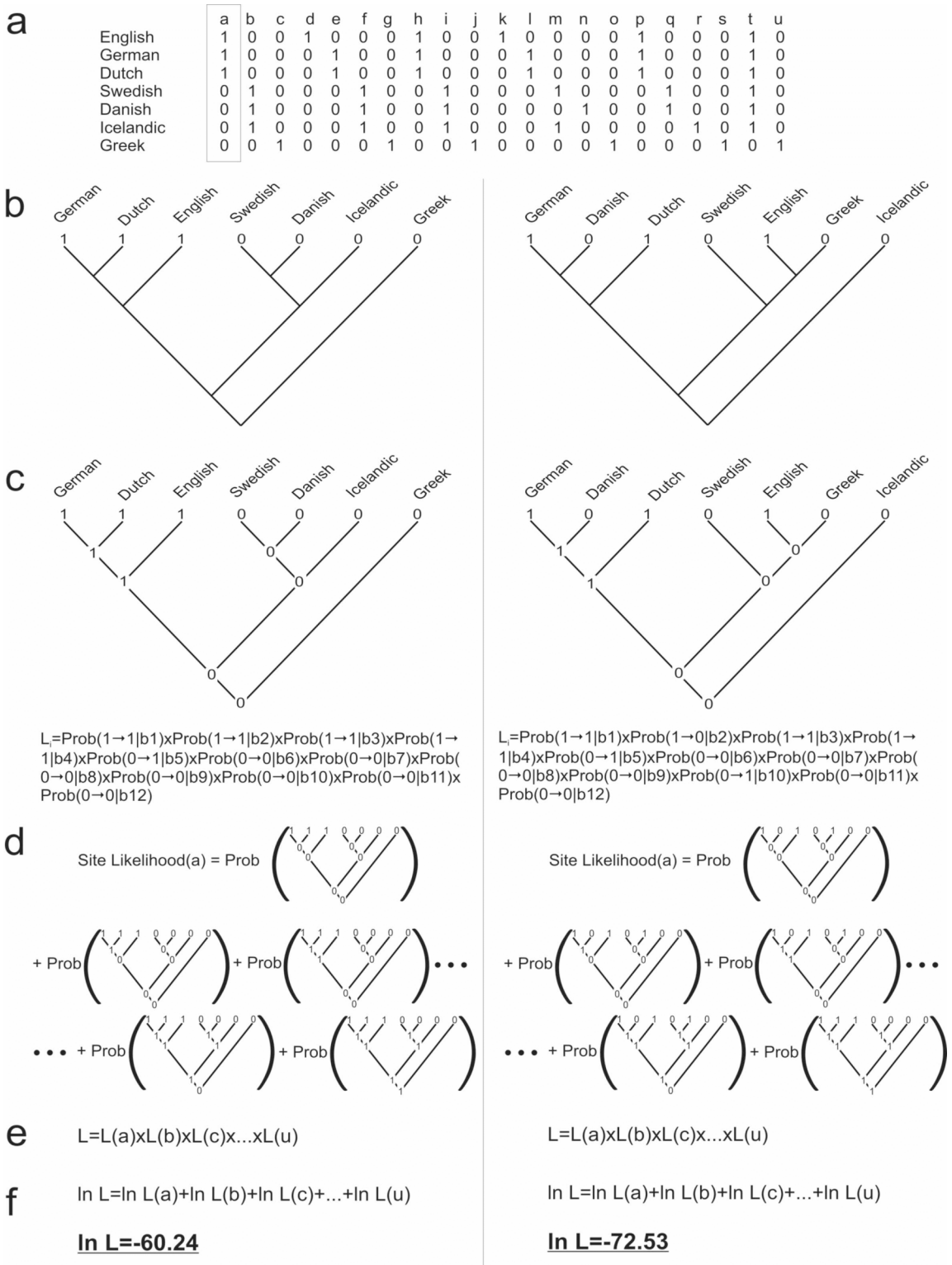


FIGURE 4.4: The gamma distribution, used to model rate variation between sites. Three possible values for α are shown. For small values of α (e.g. α=0.5), most cognate sets evolve slowly, but a few can evolve at higher rates. As α increases, the distribution becomes more peaked and symmetrical around a rate of 1 – i.e. rates become more equal (e.g. α=50).

For a given model of evolution, trees are evaluated according to their likelihood scores, which represent the probability of a specific tree giving rise to the observed data under the model. The greater the likelihood of producing the observed data, the more favorable the tree. The maximum-likelihood tree is that tree or trees making the data most likely. The basic procedure for calculating the likelihood score is outlined in Figure 4.5 using the data from Table 4.2. Two trees are evaluated so that their likelihood scores can be compared. First, presence/absence information for a particular cognate set is selected (Figure 4.5a) and mapped onto each tree (Figure 4.5b). Next, ancestral character states are hypothesized, and the likelihood of the resulting scenario is calculated (Figure 4.5c). The likelihood is contingent on the substitution probabilities per unit time as defined in the model. This process is repeated for all possible combinations of ancestral character states. The likelihood of each tree, for this single cognate set, is the sum of the probabilities of all the possible ancestral-state combinations (Figure 4.5d).



	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
English	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0
German	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
Dutch	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
Swedish	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Danish	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
Icelandic	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0
Greek	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1

FIGURE 4.5: Calculation and comparison of likelihood for two language phylogenies. **a** Presence/absence information for a particular cognate set is selected from the data in table 4.2. **b** This information is mapped onto each tree. **c** Ancestral character states are hypothesized and the probability or likelihood of the resulting scenario is calculated. **d** The likelihood of each tree, for this single cognate set, is the sum of the probabilities of all the possible ancestral state combinations. **e-f** The overall log-likelihood score is calculated by summing the logs of the likelihood scores of all cognate sets. The tree with the least negative log-likelihood is the most favourable. Comparing the final log-likelihood scores of each tree we see that, consistent with what linguists already know of Germanic language relationships, the tree on the left is more favourable.

The overall likelihood of each tree for all of the data in Table 4.2 can be calculated by taking the product of all 21 individual cognate-set likelihoods (Figure 4.5e). This value is often very small (especially for larger data sets), so the log of the likelihood is usually used instead. The less negative the log-likelihood is, the more likely the tree is. The overall log-likelihood score can thus be calculated by summing the logs of the likelihood scores of all cognate sets (Figure 4.5f). In comparing the final log-likelihood scores of each tree, we see that, consistent with what linguists already know about Germanic-language relationships, the tree on the left is more favorable. It should be noted that this is a highly simplified example based on a restricted data set. More information in the form of more characters allows for more powerful inferences; hence real linguistic and biological data sets typically include thousands of characters.

4.2.2.2 Bayesian Inference of Phylogeny

It is not usually computationally feasible to evaluate the likelihood of all possible language trees in order to find the maximum-likelihood tree. The number of possible tree topologies that must be evaluated increases exponentially with the number of taxa. For six taxa there are 945 possible rooted trees, for 10 taxa there are over 34 million trees, and for 20 taxa there are over 8×10^{21} trees. Current processor speeds do not allow us to reconstruct character-state changes, estimate model parameters, and calculate likelihood scores across all possible trees for anything more than a handful of taxa. Further, the vast number of possibilities combined with finite data means that inferring a single tree will be misleading - there are usually many trees with likelihood scores which are close to the likelihood of the maximum-likelihood tree. Even when the model is an accurate description of cognate evolution there is a very high probability that the true tree will not coincide with the maximum-likelihood tree. There is thus uncertainty in the inferred phylogeny. Bayesian inference is an alternative approach to phylogenetic analysis that allows us to draw inferences from a large amount of data using powerful probabilistic models without searching for the 'optimal tree' (Huelsenbeck *et al*., 2001). In this approach trees are sampled according to their posterior probabilities. The posterior probability of a tree (the probability of the tree given the priors, data and the model) is related by Bayes' theorem to its likelihood score (the probability of the data given the tree) and its prior

probability (a reflection of any prior knowledge about tree topology that is to be included in the analysis). Unfortunately, we cannot evaluate this function analytically. However, we can use a Markov chain Monte Carlo (MCMC; Metropolis *et al*., 1953) algorithm to generate a sample of trees in which the frequency distribution of the sample is an approximation of the posterior probability distribution of the trees (Huelsenbeck *et al*. 2001). To do this, we used *MrBayes,* a Bayesian phylogenetic inference programme (Huelsenbeck and Ronquist, 2001).

MrBayes uses MCMC algorithms to search through the realm of possible trees. From a random starting tree, changes are proposed to the tree topology, branch-lengths and model parameters according to a specified prior distribution of the parameters. The changes are either accepted or rejected based on the likelihood of the resulting evolutionary reconstruction – i.e. reconstructions that give higher likelihood scores tend to be favoured. In this way the chain quickly goes from sampling random trees to sampling those trees which best explain the data. After an initial 'burn-in' period, trees begin to be sampled in proportion to their likelihood given the data. This produces a distribution of trees. A useful way to summarise this distribution is with a consensus tree or consensus network (Holland and Moulton, 2003) depicting uncertainty in the reconstructed relationships. These graphs are, however, just useful pictorial summaries of the analysis. The fundamental output of the analysis is the distribution of trees.

Figure 4.6 shows a Bayesian consensus tree generated from the sample Germanic data set. As well as tree topology and branch length information, the consensus tree shows the degree of support for each subclade within the phylogeny. Clade support is usually expressed in the form of the posterior probability of a clade – approximated by the percentage of time that the clade appears in the sample distribution. A value of 100, for example, indicates that the clade occurs in all sampled trees. Lower values indicate an increasing degree of statistical uncertainty. Although clade-support values are very low in Figure 4.6 because of the small artificial example data set, the topology is broadly consistent with what linguists know of Germanic history.

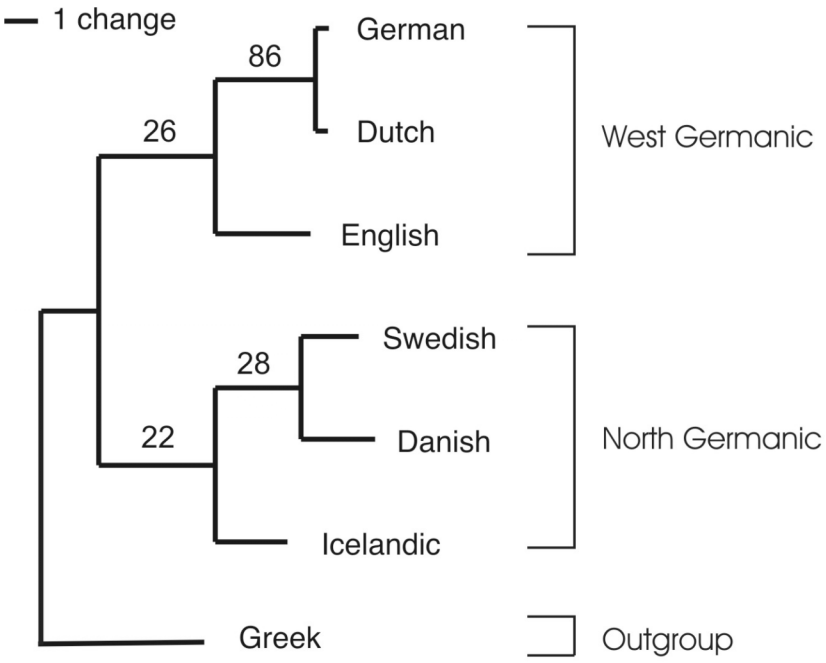


FIGURE 4.6: Bayesian consensus tree for the Germanic sample data set. Branch lengths are proportional to the inferred amount of change. Posterior probability values for clades are shown above the branch immediately ancestral to each clade.

The Bayesian sample distribution allows us to approximate phylogenetic uncertainty (uncertainty in tree topology and branch lengths), given the data, and to incorporate this into subsequent analyses. This is not possible using either the comparative method or traditional glottochronology, and yet the effect of phylogenetic uncertainty is a crucial consideration if results are to be used to test historical hypotheses. For example, rather than restricting divergence-time analyses to a single optimal tree, such as in Figure 4.6, we can analyze the entire Bayesian sample distribution and determine the error in date estimates resulting from phylogenetic uncertainty (Huelsenbeck *et al*. 2001). This would include the 26 percent of trees with a West Germanic clade as well as the 74 percent without the West Germanic clade. The resulting distribution of divergence times would thus not be contingent on the existence of a West Germanic clade.

Alternatively, we may wish to incorporate prior knowledge into the analysis by imposing constraints on the tree topology. In the case of Germanic, we happen to know on the basis of other linguistic and historical evidence that the true tree must in fact include a West Germanic clade. If we wanted to include this information in our

analysis, we could repeat the above process searching only within the subset of trees that include a West Germanic clade.

4.2.3 NETWORK METHODS - OVERCOMING BORROWING

Borrowing between languages is analogous to horizontal gene transfer in biology. Biologists use computational methods such as split decomposition (Huson, 1998), which do not force the data to fit a tree model, to check for non-treelike signals in the data. When patterns of reticulation can be identified, we can account for the effects. This procedure has also been applied to historical linguistics (Bryant, Filimon and Gray, 2005). In linguistics, the influence of borrowing can be minimized by restricting analyses to basic vocabulary such as the Swadesh word list. For example, although English is a Germanic language, it has borrowed around 50 percent of its total lexicon from French and Latin. However, only about 6 percent of English entries in the Swadesh 200 word list are clear Romance language borrowings (Embleton, 1986). These known borrowings can be removed prior to analysis. For example, the English word *mountain* would not be coded as cognate with French *montagne,* since it was obviously borrowed from French into English after the Norman invasion. Any remaining reticulation can be detected using methods that can identify conflicting signal, such as split decomposition (Huson, 1998) and *NeighbourNet* (Bryant and Moulton, 2002). In addition, because maximum-likelihood methods incorporate an explicit model of evolution, it is much easier to test assumptions about evolutionary processes such as borrowing. For example, it is possible to include borrowing between languages as part of the evolutionary model itself. We can then simulate data sets with varying degrees of borrowing and compare them to real data. This approach will be employed in Chapter 5.

4.2.4 RATE SMOOTHING AND ESTIMATING DATES - OVERCOMING RATE VARIATION THROUGH TIME

There are at least two types of rate variation in lexical evolution. First, rate variation can occur between cognates. For example, even in the Swadesh word list, the Indo-European word for *five* is highly conserved (1 cognate set) whilst the word for *dirty* is highly variable (27 cognate sets). This is akin to site-specific rate variation in biology.

Biologists can account for this type of rate variation by allowing a distribution of rates. As mentioned above, we used a model of cognate evolution that allowed for gamma distributed rate variation between cognates.

Second, rates of lexical evolution can vary through time and between lineages. Clearly this will cause problems if we are trying to estimate absolute divergence times on the inferred phylogenies since inferred branch-lengths are not directly proportional to time. Again, biologists have developed a number of methods for dealing with this type of rate variation. One such approach is the penalised likelihood method of rate smoothing implemented in *r8s* (*version 1.7;* Sanderson, 2002a), which allows for rate variation between lineages while incorporating a "roughness penalty" that costs the model more if rates vary excessively from branch to branch. Sanderson (2002b) has shown that the penalised likelihood optimization procedure performs significantly better under conditions of rate heterogeneity than procedures that assume a constant rate of evolution.

We can apply the same methods to linguistic data. Again, using the sample Germanic data set as an example, we begin by constructing an evolutionary tree with branch lengths proportional to the inferred amount of change, such as the consensus tree from the sample Germanic data set. Known divergence times based on historically attested dates are then used to calibrate rates of change across the tree (Figure 4.7). For example, we know from historical information that the Anglo-Saxons began to settle in Britain in A.D. 449. This would suggest that the English lineage split from the other West Germanic languages at some point during the fifth century A.D. We can constrain the age of this node on the tree accordingly. Similarly, we can date the break up of the North Germanic languages to around the end of the first millennium A.D. and constrain the age of this node on the tree. The penalised likelihood model is then used to smooth rates of evolution across the tree and to calculate divergence times. We can thus reconstruct the age of the node of the tree in Figure 4.6, representing the break-up of West Germanic and North Germanic. This procedure can then be repeated on all of the trees in the MCMC Bayesian sample distribution. The result is a distribution of divergence times rather than a single estimate. This distribution can be used to create a confidence interval for the age at any node.

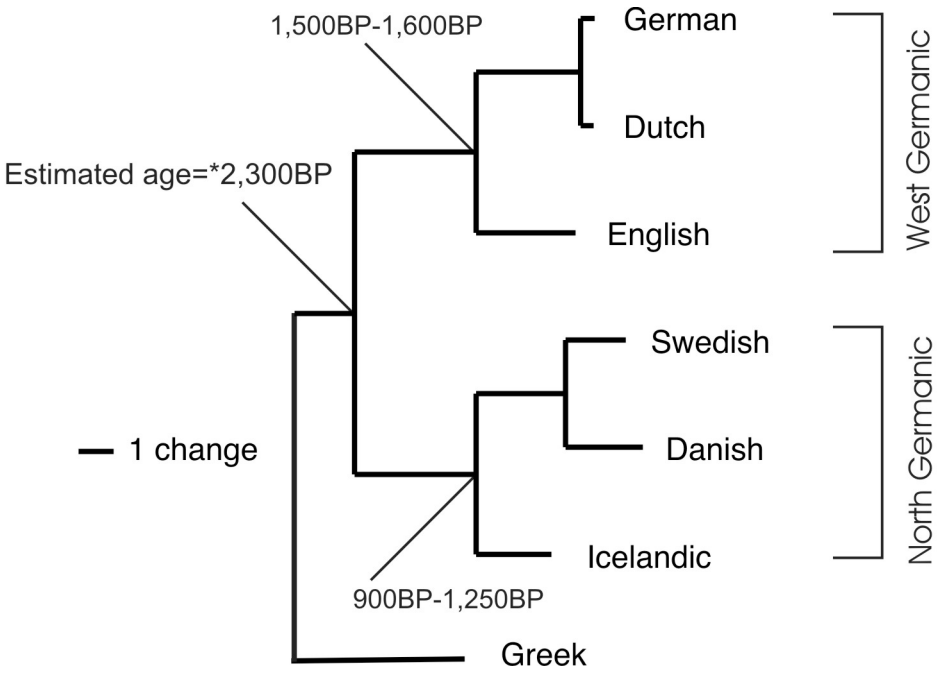


FIGURE 4.7: Estimating divergence times using the sample Germanic data set. The age of the nodes corresponding to the North and West Germanic divergence are constrained in accordance with historically attested dates. Rate smoothing is then used to estimate the age of the split between the North and West Germanic lineages.

4.3 THE ORIGIN OF INDO-EUROPEAN – ILLUMINATION OR MORE MOTHS TO THE FLAME?

In 1786 Sir William Jones noted similarities among Sanskrit, Greek, Celtic, Persian, Gothic, and Latin that led him to conclude that these languages had "sprung from some common source" (Jones 1807, pp. 34-5). The duly christened Indo-European language family today includes most of the indigenous languages of Europe and many from the near East, all of which are descendants of an ancient ancestral tongue, Proto-Indo-European. Despite over 200 years of scrutiny, scholars have been unable to locate the origin of Indo-European definitively in time or place. Indeed, the origin of the Indo-European language family has been described as "the most intensively studied, yet still most recalcitrant, problem of historical linguistics" (Diamond and Bellwood, 2003, p. 597). Theories have been put forward advocating ages ranging from 4000 to 23,000 years BP (Otte, 1997), with hypothesised homelands including central Europe (Devoto, 1962), the Balkans (Diakonov, 1984), and even India (Kumar, 1999). Mallory (1989) acknowledges 14 distinct homeland hypotheses since 1960 alone. He rather colourfully remarks that:

> "the quest for the origins of the Indo-Europeans has all the fascination of an electric light in the open air on a summer night: it tends to attract every species

> of scholar or would-be savant who can take pen to hand" (Mallory, 1989, p. 143).

Unfortunately, archaeological, genetic and linguistic research on Indo-European origins has so far proved inconclusive. Whilst numerous theories of Indo-European origin have been proposed, they have proven difficult to test. In what follows, we apply the techniques described above to test between competing hypotheses about the age of the Indo-European language family.

4.3.1 TWO THEORIES

There are currently two main theories about the origin of Indo-European. The first theory, put forward by Marija Gimbutas (1973a, 1973b) on the basis of linguistic and archaeological evidence, links Proto-Indo-European (the hypothesized ancestral Indo-European tongue) with the Kurgan culture of southern Russia and the Ukraine. The Kurgans were a group of semi-nomadic, pastoralist, warrior-horsemen who expanded from their homeland in the Russian steppes during the 5th and 6th millennia BP, conquering Danubian Europe, central Asia and India, and later the Balkans and Anatolia. This expansion is thought to roughly match the accepted ancestral range of Indo-European (Trask, 1996). As well as the apparent geographical congruence between Kurgan and Indo-European territories, there is linguistic evidence for an association between the two cultures. Words for supposed Kurgan technological innovations are notably consistent across widely divergent Indo-European sub-families. These include terms for "wheel" (**rotho-*, **k^{w}(e)k^{w}l-o-*), "axle" (**aks-lo-*), "yoke" (**jug-o-*), "horse" (**ekwo-*) and "to go, transport in a vehicle" *(*wegh-;* Mallory, 1989; Campbell, 2004). It is argued that these words and associated technologies must have been present in the Proto-Indo-European culture and that they were likely to have been Kurgan in origin. Hence, the argument goes, the Indo-European language family is no older than 5,000-6,000 BP. Mallory (1989) argues for a similar time and place of Indo-European origin – a region around the Black Sea about 5,000-6,000 BP (although he is more cautious and refrains from identifying Proto-Indo-European with a specific culture such as the Kurgans).

The second theory, proposed by archaeologist Colin Renfrew (1987, 2000), holds that Indo-European languages spread, not with marauding Russian horsemen, but with the

expansion of agriculture from Anatolia between 8,000 and 9,500 years ago. Radiocarbon analysis of the earliest Neolithic sites across Europe provides a fairly detailed chronology of agricultural dispersal. This archaeological evidence indicates that agriculture spread from Anatolia, arriving in Greece at some time during the ninth millennium BP and reaching as far as the British Isles by 5,500BP (Gkiasta *et al*., 2003). Renfrew maintains that the linguistic argument for the Kurgan theory is based on only limited evidence for a few enigmatic Proto-Indo-European word forms. He points out that parallel semantic shifts or widespread borrowing can produce similar word forms across different languages without requiring that an ancestral term was present in the proto-language. Renfrew also challenges the idea that Kurgan social structure and technology was sufficiently advanced to allow them to conquer whole continents in a time when even small cities did not exist. Far more credible, he argues, is that Proto-Indo-European spread more passively with the spread of agriculture - a scenario that is also thought to have occurred across the Pacific (Bellwood, 1991; Bellwood, 1994), Southeast Asia (Glover and Higham, 1996) and sub-Saharan Africa (Holden, 2002). On the basis of linguistic evidence, Diakonov (1984) also argues for an early Indo-European spread with agriculture but places the homeland in the Balkans – a position that may be reconcilable with Renfrew's theory.

One way of potentially resolving the debate is to look outside the archaeological record for independent evidence that allows us to test between the two main theories. Genetic studies offer one potential source of evidence. Unfortunately, due to problems associated with admixture, slow rates of genetic change and the relatively recent time-scales involved, genetic analyses have been unable to resolve the debate (Cavalli-Sforza, Menozzi and Piazza, 1994; Rosser *et al*., 2000). Early genetic studies based on protein polymorphisms (e.g., Menozzi, Piazza and Cavalli-Sforza, 1978) supported a Neolithic population spread originating from the Near East. More recent studies using mitochondrial DNA and non-recombining Y-chromosome markers (e.g., Richards *et al*., 2000; Semino *et al*., 2000; Chikhi *et al*., 2002) also support a Near Eastern origin, although there is debate over the relative contribution of Neolithic farmers and Palaeolithic hunter-gatherers to the European gene pool.

Languages change much faster than genes and so contain more historical information at shallower time depths. Thus there is an opportunity for a linguistic contribution to

the question of Indo-European origin. Here we take advantage of the fact that the Kurgan hypothesis and the Anatolian-farming hypothesis both imply very different age ranges–6,000 BP and 8,500 BP, respectively—and test the hypotheses by constructing a confidence interval for the age at the base of the Indo-European language tree.

4.3.2 DATA AND CODING

Linguistic data were derived from Dyen, Kruskal and Black's (1992) comparative Indo-European database, which records word forms and cognancy judgments in 95 languages across the 200 semantic categories of the Swadesh word list. Some of the languages in the database are represented by multiple speech varieties, 11 of which were not included in the analyses because they had been identified by Dyen *et al*. (1992) as less desirable data sources. Three extinct languages (Hittite, Tocharian A and Tocharian B) were added to the database in an attempt to improve the resolution of basal relationships in the inferred phylogeny. Multiple references were used to corroborate cognacy judgements (Hoffner, 1967; Tischler, 1973, 1997; Guterbock and Hoffner, 1986; Gamkrelidze and Ivanov, 1995; Adams, 1999). For each meaning in the database, languages were grouped into cognate sets. Similar word forms known to be the result of borrowing were not coded as cognates. From the modified database we created a binary matrix representing the presence/absence of 2,449 cognate sets in 87 languages. We found only limited evidence of reticulation in the data (see Bryant *et al*. [2005] for a more detailed examination of reticulation in Indo-European). Examining subsets of languages using split decomposition revealed a strong treelike signal in the data, and a preliminary parsimony analysis produced a consistency index of 0.48 and a retention index of 0.76, well above what would be expected from biological datasets of a similar size (Sanderson and Donoghue, 1989).

4.3.3 PHYLOGENETIC INFERENCE

A likelihood model of evolution and Bayesian inference of phylogeny were used to reconstruct Indo-European language relationships. We adapted the "restriction-site" time-reversible model of binary character evolution implemented in *MrBayes* (*version 3*; Huelsenbeck and Ronquist, 2001). This allowed for unequal character-state

frequencies and gamma-distributed character-specific rate heterogeneity. Model parameters, including the rate matrix, branch lengths, and gamma distribution shape parameter were estimated from the data. For each of the reported analyses, ten million post-burn-in trees were generated from ten four-chain MCMC runs, each started from a random tree. To ensure that consecutive samples were independent, only every 10,000th tree was sampled from this distribution, producing an effective sample size of 1,000. Trees were rooted with Hittite in accordance with linguistic evidence (Gamkrelidze and Ivanov, 1995; Rexová *et al.*, 2003).

Figure 4.8 shows a majority-rule consensus tree for the sample distribution of trees. The posterior probability values above each internal branch give an indication of the uncertainty associated with each clade on the consensus tree (the percentage of trees in the Bayesian distribution that contain the clade). For example, the value 67 above the branch leading to the Italo-Celto-Germanic clade indicates that that clade was present in 67% of the trees in the sample distribution. The tree topology is consistent with the traditional Indo-European language groups (Campbell, 2004), correctly grouping the Celtic, Germanic (including North and West divisions), Romance (including Iberian-French division), Balto-Slavic (including Baltic and Slavic divisions), Indo-Iranian (including Indic and Iranian divisions), Albanian, Greek, Armenian, and Tocharic groups. All these groups are monophyletic and are supported by high posterior probability values. Recent parsimony and compatibility analyses have also supported these groupings as well as a Romano-Germano-Celtic supergroup, the early divergence of Greek and Armenian lineages (Rexová *et al.*, 2003), and the basal position of Tocharian (Ringe *et al.*, 2002).

Although the consensus tree summarizes much of the information in the Bayesian sample distribution, it should not be interpreted as the "true" tree. There is uncertainty in the branch-length estimates and topology of any reconstructed phylogeny. For instance, historical linguists have not resolved the position of the Albanian group, and this uncertainty is clearly reflected in the tree—the posterior probability of the Albanian/Indo-Iranian group is only 0.36. One major advantage of the Bayesian approach is that results need not be contingent on a single phylogeny. One way of representing conflicting support for various groupings is the consensus network (Holland and Moulton, 2003). Figure 4.9 shows a consensus network from the same

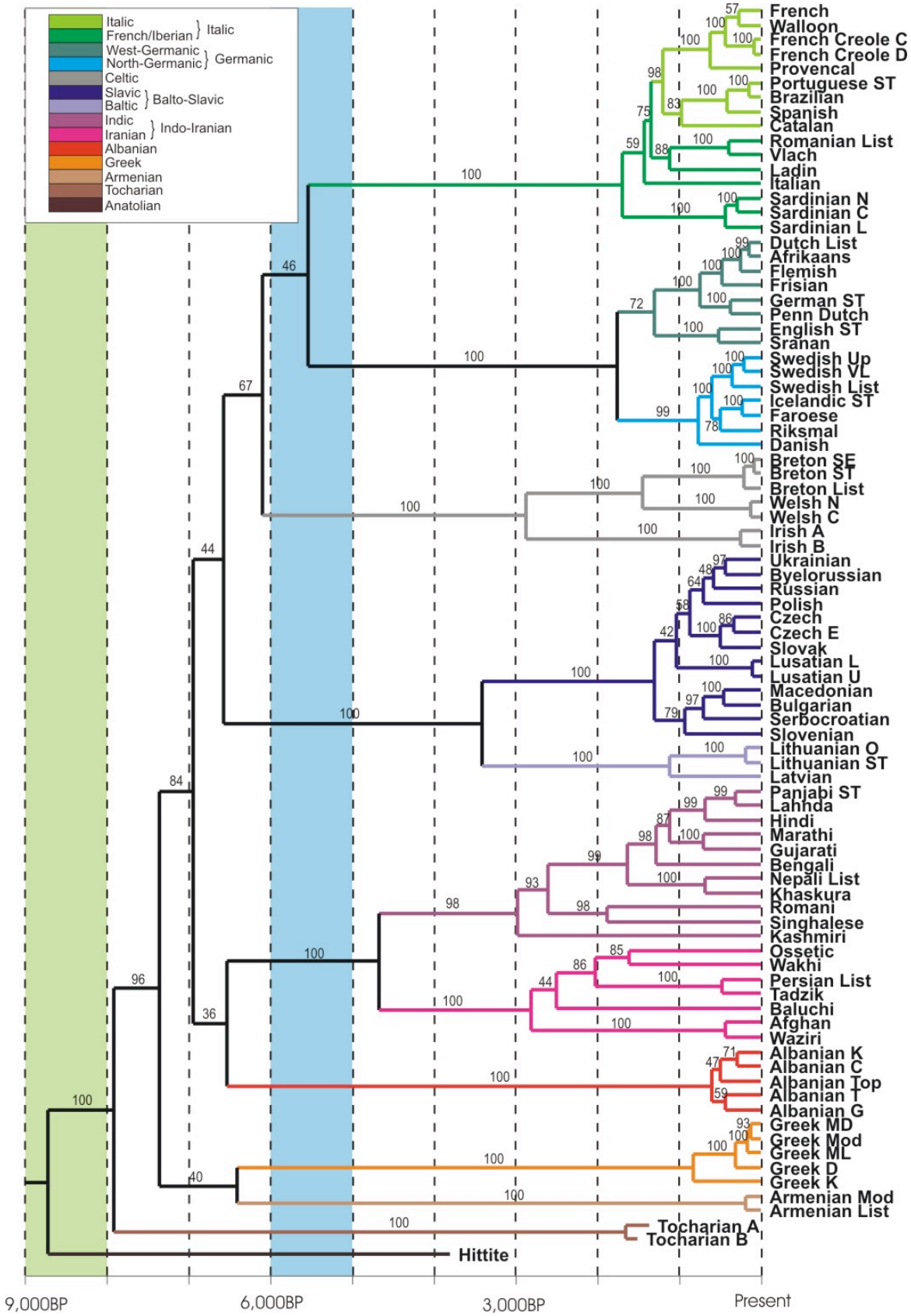


FIGURE 4.8: Majority-rule consensus tree from the initial Bayesian MCMC sample of 1,000 trees. Values above each branch indicate uncertainty (posterior probability) in the tree as a percentage. Branch-lengths are proportional to time. Coloured bars represent the age range proposed by the two main theories – the Anatolian theory (green bar) and the Kurgan theory (blue bar). The basal age (8,700BP) supports the Anatolian theory.

Bayesian sample distribution of 1,000 trees. The values next to splits represent posterior probabilities for that split. For example, the value 41 next to the parallel lines separating Italic and Celtic from the rest of the Indo-European sub-families indicates that that split was present in 41% of the trees in the sample distribution. Similarly, the split grouping Italic and Germanic languages was present in 46% of the sample distribution. By estimating divergence times across the Bayesian sample of trees we can account for variation in the age estimates that results from this phylogenetic uncertainty.

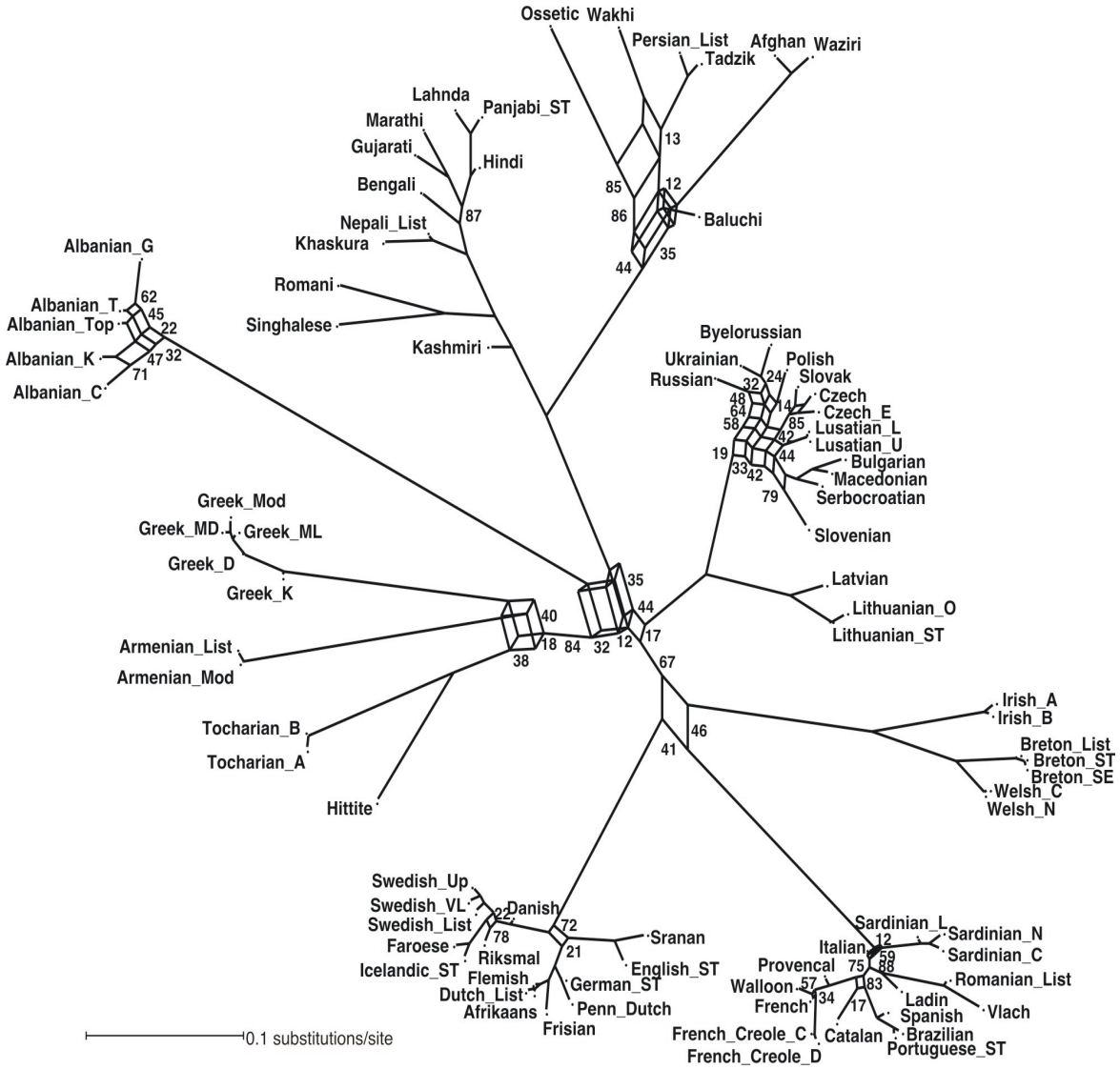


FIGURE 4.9: Consensus network (Holland and Moulton, 2003) from the Bayesian MCMC sample of trees. Values express the posterior probability of each split (values above 90% are not indicated). A threshold of 10% was used to draw this splits graph – i.e. only those splits occuring in at least 10% of the observed trees are shown in the graph. Branch lengths represent the median number of inferred changes per site across the sample distribution.

4.3.4 DIVERGENCE TIME ESTIMATION

The age of 14 Indo-European divergence points were constrained in accordance with historical evidence (see Table 4.5). These known node ages were combined with branch-length information to estimate rates of evolution across each tree. Using the penalised likelihood method of rate smoothing described above, rates were allowed to vary across the tree whilst incorporating a 'roughness penalty' that incurs a cost if rates vary excessively from branch to branch. This allows us to derive age estimates for each node on the tree without assuming strictly clock-like rates of change. Using this procedure, the branch-lengths in Figure 4.8 were estimated proportional to time. As can be seen in the figure, the estimated age at the base of the tree is 8,700BP - within the range predicted by the Anatolian farming theory of Indo-European origin.

A single divergence time, with no estimate of the error associated with the calculation, is of limited value. To test between historical hypotheses we need some measure of the error associated with the date estimates. Specifically, uncertainty in the phylogeny gives rise to a corresponding uncertainty in age estimates. In order to account for phylogenetic uncertainty we estimated the age at the base of the trees in the post-burn-in Bayesian MCMC sample to produce a probability distribution for the age of Indo-European. One advantage of the Bayesian framework is that prior knowledge about language relationships can be incorporated into the analysis. In order to eliminate trees that conflict with known Indo-European language groupings, the original 1,000 tree sample was filtered using a constraint tree representing these known language groupings [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), ((NorthGermanic, WestGermanic), Italic, Celtic)))]. This constraint tree was consistent with the majority-rule consensus tree generated from the entire Bayesian sample distribution. The filtered distribution of divergence time estimates was then used to create a confidence interval for the age of the Indo-European language family. This distribution could then be compared with the age ranges implied by the two main theories of Indo-European origin (see Figure 4.10). The results are clearly consistent with the Anatolian hypothesis.

Table 4.5: AGE CONSTRAINTS USED TO CALIBRATE THE INDO-EUROPEAN DIVERGENCE TIME CALCULATIONS, BASED ON KNOWN HISTORICAL INFORMATION.

Divergence	Age constraint	Source	Historical Information
Iberian-French	450AD-800AD	1,3	Death of last writers knowing classical Latin and repetition of Latin liturgical formulas without comprehension in 6th to 8th centuries AD. Strasburg Oaths, 842AD.
Italic-Romanian	150AD-300AD	1,2,3	Last Roman troops withdrawn to south of Danube, 270AD. Dacia conquered by Rome, 112AD.
North/West Germanic	50AD-250AD	1,2,3,6	Germanic tribes united against Rome, 1AD. Gothic migration to Eastern Europe, 180AD. Earliest attested North Germanic inscriptions date from 3rd century AD.
Welsh/Breton	400AD-550AD	2,5	Migrants from Britain colonize Brittany in 5th century AD.
Irish/Welsh	before 300AD	5,6	Archaic Irish inscriptions date back to the 5th century AD – divergence must have occurred well before this time.
Indic	before 200BC	2,3	Singhalese records dating from as early as 2nd century BC indicate that Indic languages had begun to diverge by this time.
Iranian	before 500BC	2,3,6	By 500BC Old Persian was distinct from the Eastern Iranian dialects.
Indo-Iranian	before 1,000BC	2,3,6	Rgveda, an identifiably Indic epic, is thought to date from between 1450-1000BC. The Avesta, a similar Iranian epic, has been recorded in oral tradition since before 800BC.
Slavic	before 700AD	2	Old Church Slavonic and East Slavic texts date to beginning of 9th century and indicate significant divergence by this time. The split must have occurred after the Balto-Slavic divergence.
Balto-Slavic	1,400BC-100AD	2,3	Distinct Slavic culture and language known to pre-date 100AD on the basis of Tacitus's "Germany". Archaeological evidence suggests the split may have occurred as early as 1,400BC.
Greek split	before 1,500BC	2,4,6	Earliest form of an ancient Greek dialect is Mycenaean, attested in Linear B texts dating from 15th century BC.
Tocharic	140BC-350AD	2,3	Tocharian languages are thought to have diverged shortly after the fall of Bactria (135BC) and no later than 100 years before the first known inscriptions of Tocharian B.
Tocharian A & B	500AD-750AD	2,3	Earliest texts from later half of 1st millennia AD. No texts after 750AD by which time Tocharians are thought to have been assimilated with Turkish invaders.
Hittite	1,800BC-1,300BC	2,3	Oldest Hittite text of King Anittas from the 18th century BC. Latest texts from the 14th-13th centuries BC.

Sources:1 - Embleton, 1991.
2 – Gamkrelidze and Ivanov, 1995.
3 - *Indo-European Chronology*, 2002. Available online at: -
http://www.indoeuro.bizland.com
4 – Champion *et al.*, 1984.
5 – Ringe et al., 1998.

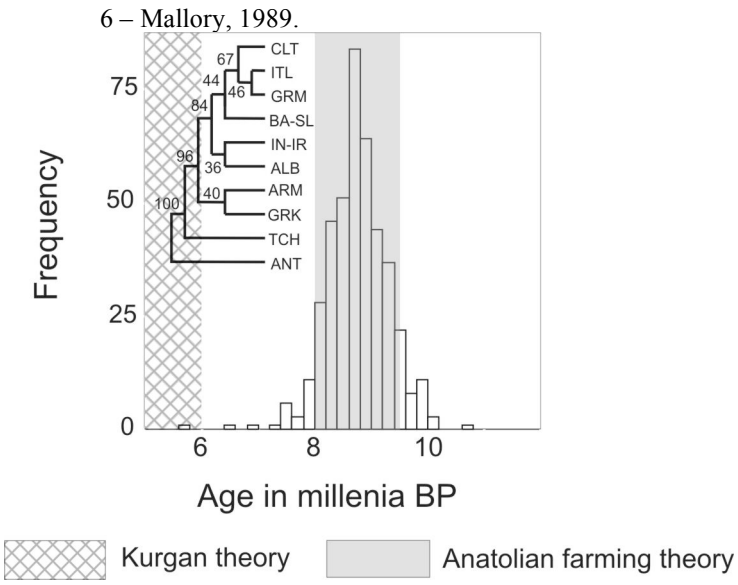


FIGURE 4.10: Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees for the initial assumption set (n=435). The majority-rule consensus tree for the entire (unfiltered) sample is shown in the upper left.

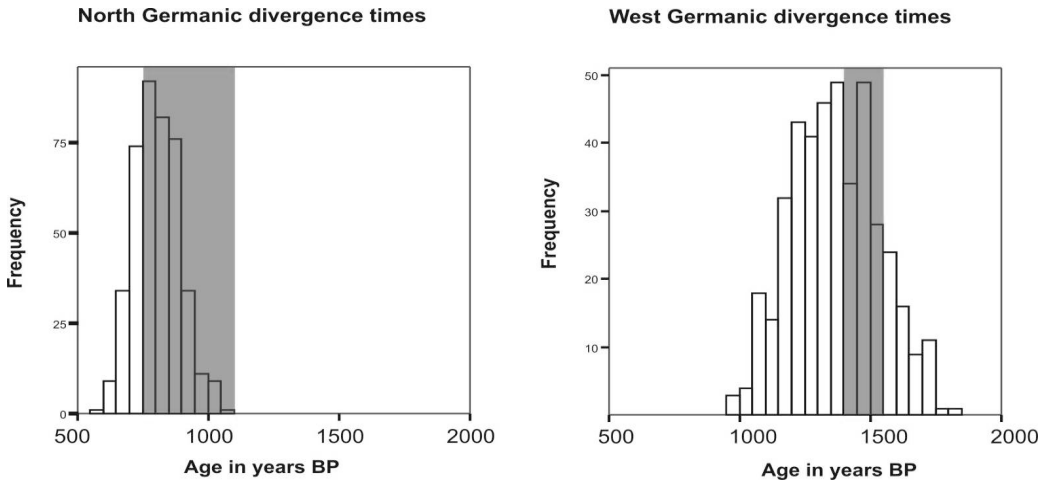


FIGURE 4.11: Frequency distribution of age estimates for the North and West Germanic subgroups across filtered Bayesian MCMC sample of trees (n=433). The grey bands indicate the historically attested time of divergence.

Not all historically attested language splits were used in our analysis. One means of validating our methodology is to produce divergence time distributions for nodes that were not constrained in the analysis and compare this to the historically attested time of divergence. For example, Figure 4.11 shows the inferred divergence time distributions for the North and West Germanic subgroups. The grey band in these figures indicates the likely age of each subgroup based on the historical record. The age estimates for the North Germanic clade correspond with written evidence for the

break up of these languages between 900AD and 1250AD. Similarly, estimated ages of the West Germanic clade are consistent with historical evidence dating the Anglo-Saxon migration to the British Isles about 1500 years ago.

4.3.5 TESTING ROBUSTNESS

A key part of any Bayesian phylogenetic analysis is an assessment of the robustness of the inferences. To do this we tested the effect of altering a number of different parameters and assumptions of the method.

4.3.5.1 Bayesian 'Priors'

Initialising each Bayesian MCMC chain required the specification of a starting tree and prior parameters ('priors') for the analysis. The MCMC sample distribution was the product of ten separate runs from different random starting trees. Divergence time and topology results for each of the separate runs were consistent. Other test analyses were run using a range of priors for parameters controlling the rate matrix, branch-lengths, gamma distribution and character state frequencies. The inferred tree phylogeny and branch-lengths did not noticeably change when priors were altered.

4.3.5.2 Cognacy Judgements

The Dyen *et al*. (1992, 1997) database contained information about the certainty of cognacy judgements. Words were coded as 'cognate' or 'doubtful cognates'. In the initial analysis we included all cognate information in an effort to maximise any phylogenetic signal. However, we wanted to test the robustness of our results to changes in the stringency of cognacy decisions. For this reason, the analysis was repeated with doubtful cognates excluded. This produced a similar age range to the initial analysis, indicating that our results were robust to errors in cognacy judgements (see Figure 4.12).

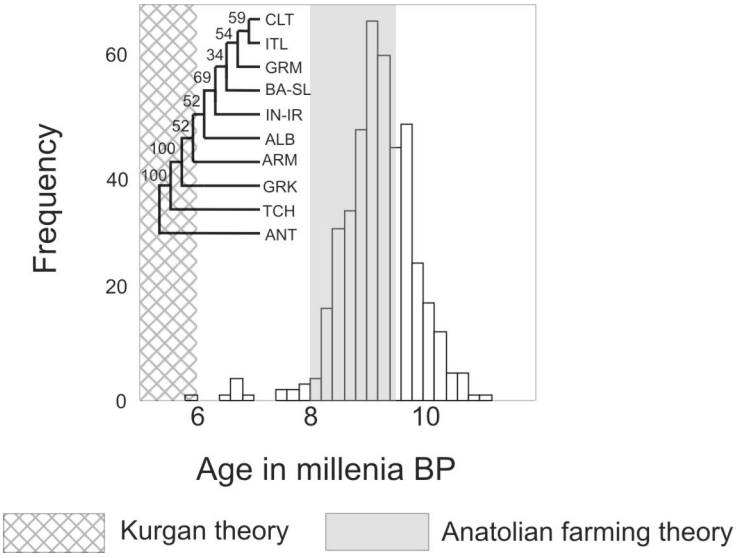


FIGURE 4.12: Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees for analysis with doubtful cognates excluded (n=433). The majority-rule consensus tree for the entire sample is shown in the upper left.

4.3.5.3 Calibrations and Constraint Trees

We wanted to test the robustness of our results to variations in age constraints. The step-by-step removal of each of the 14 age constraints on the consensus tree revealed that divergence time estimates were robust to calibration errors across the tree. For 13 nodes, the reconstructed age was within 390 years of the original constraint range. Only the reconstructed age for Hittite showed an appreciable variation from the constraint range. This may be attributable to the effect of missing data associated with extinct languages. Reconstructed ages at the base of the tree ranged from 10,400BP with the removal of the Hittite age constraint, to 8,500BP with the removal of the Iranian group age constraint. The results are highly robust calibration errors because of the large number of age constraints we used to calibrate rates of lexical evolution across the tree.

We also wanted to be sure that the constraint tree used to filter the Bayesian distribution of trees had not systematically biased our results. Figure 4.13 shows the divergence time distribution for the initial data set after filtering using a minimum set of topological constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (NorthGermanic, WestGermanic), Italic, Celtic))]. Again, the divergence time distribution was consistent with the Anatolian farming theory.

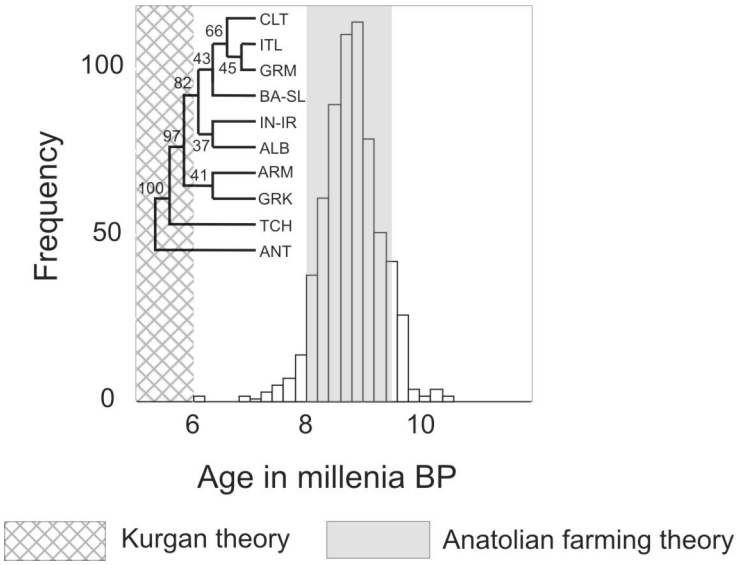


FIGURE 4.13: Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees using minimum set of topological constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (NorthGermanic, WestGermanic), Italic, Celtic))] (n=670). The majority-rule consensus tree for the entire sample is shown in the upper left.

4.3.5.4 Missing Data

Another possible bias was the effect of missing data. Some of the languages in the Dyen *et al*. (1992, 1997) database may have contained fewer cognates because information about these languages was missing. For example, the three extinct languages (Hittite, Tocharian A and Tocharian B) are derived from a limited range of source texts and it is possible that some cognates were missed because the terms were not referred to in the source text. This may have biased divergence time estimates by falsely increasing basal branch-lengths. Nicholls and Gray (in press) point out that we should expect fewer cognates to be present in the languages at the base of a tree anyway - the fact that Hittite has 94 cognates whilst most languages have around 200, does not necessarily imply that data is missing. Nonetheless, we tested for the effect of missing data by including information about whether or not the word for a particular term was missing from the database. If we could not rule out the possibility that a cognate was absent from a language because it had not been found or recorded, then that cognate was coded as missing (represented by a '?' in the matrix). Encoding missing cognate information in this way means that we can account for uncertainty in the data itself using the likelihood model – the unknown states become parameters to

be estimated. Analysing this recoded data also produced an age range consistent with the Anatolian theory (see Figure 4.14).

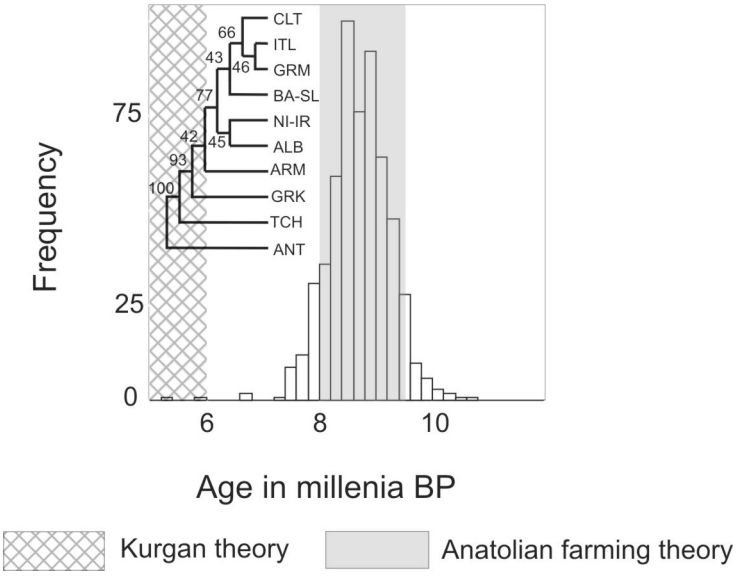


FIGURE 4.14: Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees with information about missing cognates included (n=620). The majority-rule consensus tree for the entire sample is shown in the upper left.

4.3.5.5 Root of Indo-European

Finally, we tested the effect of the rooting point for the trees. In the previous analyses, trees were rooted with Hittite. Although this is consistent with independent linguistic analyses (Gamkrelidze and Ivanov, 1995; Rexova, Frynta and Zrzavy, 2003), other potential root points are possible. It could be claimed that a Hittite root biases age estimates in favour of the Anatolian hypothesis. We thus reran the rate smoothing analysis rooting the consensus tree in Figure 4.8 with Balto-Slavic, Greek, Tocharian and Indo-Iranian groups. In all four cases the estimated divergence time increased to between 9,500BP and 10,700BP.

4.4 DISCUSSION

The time depth estimates reported here are consistent with the times predicted by a spread of language with the expansion of agriculture from Anatolia. The branching pattern and dates of internal nodes are broadly consistent with archaeological evidence indicating that between the tenth and sixth millennia BP a culture based on cereal cultivation and animal husbandry spread from Anatolia into Greece and the

Balkans and then out across Europe and the near-East (Gkiasta *et al.*, 2003; Renfrew, 1987). Hittite appears to have diverged from the main Proto-Indo-European stock around 8,700 years ago, perhaps reflecting the initial migration out of Anatolia. Indeed, this date exactly matches estimates for the age of Europe's first agricultural settlements in southern Greece (Renfrew, 1987). Following the initial split, the language tree shows the formation of separate Tocharian, Greek, and then Armenian lineages, all before 6,000BP, with all of the remaining language families formed by 4,000BP. We note that the received linguistic orthodoxy (Indo-European is only 6,000 years old) does approximately fit the divergence dates we obtained for most of the branches of the tree. Only the basal branches leading to Hittite, Tocharian, Greek and Armenian are well beyond this age. Interestingly, the date range hypothesised for the Kurgan expansion does correspond to a rapid period of divergence on the consensus tree. According to the divergence time estimates shown in Figure 4.8, many of the major Indo-European sub-families - Indo-Iranian, Balto-Slavic, Germanic, Italic and Celtic - diverged between 6,000 and 7,000 years ago – intriguingly close to the hypothesised time of the Kurgan expansion. Thus it seems possible that there were two distinct phases in the spread of Indo-European: an initial phase, involving the movement of Indo-European with agriculture, out of Anatolia into Greece and the Balkans some 8,500 years ago; and a second phase (perhaps the Kurgan expansion) which saw the subsequent spread of Indo-European languages across the rest of Europe and east, into Persia and central Asia.

4.5 RESPONSE TO OUR CRITICS

4.5.1 THE POTENTIAL PITFALLS OF LINGUISTIC PALAEONTOLOGY

A number of linguists have claimed that linguistic palaeontology offers a compelling reason why the arguments we have presented must be wrong: Proto-Indo-Europeans are claimed to have had a word for "wheel" (*$*k^{w}(e)k^{w}l$-o-*) but wheels did not exist in Europe 9,000 years ago. The case is based on a widespread distribution of apparently related words for wheel in Indo-European languages. Trask (2003a) repeats this well-known argument in a commentary on our work that appeared on the Web:

> "we can reconstruct for PIE [Proto-Indo-European] a number of words pertaining to wheeled vehicles in general and horse-drawn chariots in

> particular. The last speakers of PIE must therefore have been familiar with these things. But the archaeologists tell us that these things were not invented before about 6000 BP. So, how could a language that was last spoken around 10,000 years ago have words for things that were not invented until 4,000 years later?"

This argument is nowhere near as compelling as Trask and other commentators have claimed. There are at least two alternative explanations for the distribution of terms associated with wheel and wheeled transport: independent semantic innovations from a common root and/or widespread borrowing of technological terms. To describe the case for the former, we can do no better than quote Trask (1996:355-356) himself:

> "There is a PIE word **ekwo-* "horse," as well as **wegh-* "convey, go in a vehicle," **kwekwlo-"*wheel," **aks-* "axle," and **nobh-* "hub of a wheel." This has led some scholars to conclude that the PIE-speakers not only rode horses but had wagons and chariots as well. This is debatable, however, since everyone places PIE at least 6,000 years in the past, while hard evidence for wheeled vehicles is perhaps no earlier than 5,000 years ago. Watkins (1969) considers that these terms pertaining to wheeled vehicles were chiefly metaphorical extensions of older IE words with different senses (*nobh-, for example, meant "navel"). The word *kwekwlo-"wheel" itself is derived from the root *kwel-"turn, revolve." Nevertheless, the vision of fierce IE warriors, riding horses and driving chariots, sweeping down on their neighbours brandishing bloody swords, has proven to be an enduring one, and scholars have found it difficult to dislodge from the popular consciousness the idea of the PIE-speakers as warlike conquerors in chariots."

In other words, independent semantic innovations from a common root are a likely mechanism by which we can account for the supposed Proto-Indo-European reconstructions associated with wheeled transport. Linguists can reconstruct word forms with much greater certainty than their meanings. For example, upon the development of wheeled transport, words derived from the Proto-Indo-European term **kwel-* (meaning "to turn, rotate") may have been independently co-opted to describe the wheel. On the basis of the reconstructed ages shown in Figure 4.8, as few as three such semantic innovations around the sixth millennium BP could have accounted for the attested distribution of terms related to $*k^{w}(e)k^{w}l$-*o-* "wheel" (one shift just before

the break up of the Italic-Celtic-Germanic-Balto-Slavic-Indo-Iranian lineage, one shift in the Greek-Armenian lineage, and one shift [or borrowing] in the Tocharian lineage).

The second possible explanation for the distribution of terms pertaining to wheeled vehicles is widespread borrowing. Good ideas spread. Terms associated with a new technology are often borrowed along with the technology. The spread of wheeled transport across Europe and the near East 5,000–6,000 years ago seems a likely candidate for borrowing of this sort. Linguists are able to identify many borrowings (particularly more recent ones) on the basis of the presence or absence of certain systematic sound correspondences. However, our date estimates suggest that most of the major Indo-European groups were just beginning to diverge when the wheel was introduced. We would thus expect the currently attested forms of any borrowed terms to look as if they were inherited from Proto-Indo-European – they may thus be impossible to reliably identify.

This does not mean that there is no issue here. Ideally, we should aim to synthesise all lines of evidence relating to the age of Indo-European. Ringe (unpublished manuscript) presents a careful summary of the terms related to wheeled vehicles in Indo-European. He argues that words for 'thill' (a pole that connects a yoke or harness to a vehicle) and 'yoke' can confidently be reconstructed for Proto-Indo-European. He notes that reflexes of *$*k^{w}(e)k^{w}l$-o-* 'wheel' have not been found in Anatolian languages but exist in Tocharian A and B and other Indo-European languages, and hence can be reconstructed for the common ancestor of all non-Anatolian Indo-European languages. Ringe claims that the specific forms of these words make parallel semantic changes or borrowing extremely implausible. It would be extremely useful to attempt to quantify just how unlikely such alternative scenarios are. Until all the assumptions of these arguments are formalised, and the probability of alternative scenarios quantified, it will remain difficult to synthesise all the different lines of evidence on the age of Indo-European.

4.5.2 MODEL MISSPECIFICATION AND THE INDEPENDENCE OF CHARACTERS

The model of lexical evolution employed here has been criticized for two reasons. First, Warnow *et al*. (in press) argue that it is not possible to estimate divergence times on the basis of lexical data because the process of lexical evolution is so unconstrained as to make any attempt to estimate branch-lengths futile. Instead, they advocate a no common mechanism model to infer language relationships (but not divergence times). Second, Evans *et al*. (in press) claim the evolutionary model of binary character evolution employed here is inappropriate because it assumes independence between characters when our characters are clearly not independent. Here we address each of these concerns.

4.5.2.1 Models are lies that lead us to the truth

When biologists model evolution, they lie: they lie about the independence of character state changes across sites; they lie about the homogeneity of substitution mechanisms; and they lie about the importance of selection pressure on substitution rates. But these are lies that lead us to the truth. Biological research is based on a strategy of model-building and statistical inference that has proved highly successful (e.g. Hillis, 1992; Hillis, Moritz and Marble, 1996; Pagel, 1999). The goal for biologists is not to construct a model so complex that it captures every nuance and vagary of the evolutionary process, but rather to find the simplest model available that can reliably estimate the parameters of interest.

Model choice is thus a balance between over- and under-fitting parameters (Burnham and Anderson, 1998). Adding extra parameters can improve the apparent fit of a model to data, however, sampling error is also increased because there are more unknown parameters to estimate (Swofford *et al*. 1996). Depending on the question we are trying to answer, this added uncertainty can prevent us from estimating the model parameters from the data to within a useful margin of error. In many cases, adding just a few extra parameters can create a computationally intractable problem. Conversely, a model that is too simple can produce biased results if it fails to capture

an important part of the process (Burnham and Anderson, 1998). There is thus a compromise between biased estimates and variance inflation.

The strategy that has proved successful in biology is to start with a simple model that captures some of the fundamental processes involved and increase the complexity as necessary. For example, nucleotide substitution models range from a simple equal rates model (Jukes and Cantor, 1969), to more complex models that allow for differences in transition/transversion rates, unequal character state frequencies, site specific rates, and auto-correlation between sites (Swofford *et al*. 1996). Although even the most complicated models are simplifications of the process of evolution, often the simplest substitution model captures enough of what is going on to allow biologists to extract a meaningful signal from the data. Levins (1966) gives three reasons why we should use a simple model. First, violations of the assumptions of the model are expected to cancel each other out. Second, small errors in the model should result in small errors in the conclusions. And third, by comparing multiple models with reality we can determine which aspects of the process are important.

4.5.2.2 Can we model language evolution?

Our simple, time-reversible, binary model of lexical evolution assumes that the rate of appearance and disappearance of cognates is randomly distributed about some mean value. This rate can vary between cognate sets according to a gamma distribution and, with the addition of rate smoothing, rates of change can vary through time in a constrained way. Whilst historical, social, and cultural contingencies can undoubtedly influence the process of linguistic change, we explicitly reject Warnow *et al*.'s (in press) counsel of despair, that language evolution is so idiosyncratic and unconstrained that inferring divergence dates is impossible. Language evolution is subject to real-world constraints, such as human language acquisition, expressiveness, intelligibility, and generation time. As we pointed out in Chapter 2, Ringe, Warnow and Taylor (2002, p. 61) argue this point themselves:

> "Languages replicate themselves (and thus 'survive' from generation to generation) through a process of native-language acquisition by children. Importantly for historical linguistics, that process is tightly constrained".

These constraints create underlying commonalities in the evolutionary process that we can, and should, be trying to model.

4.5.2.3 The Independence Assumption

Evans *et al*. (in press) argue that our model is "patently inappropriate" because it assumes that all characters are independent. In biology, this is known as the I.I.D. (identically and independently distributed) assumption. Evans *et al*. correctly point out that the I.I.D. assumption is violated when individual meanings in the Swadesh word list are broken up into characters representing multiple cognate sets. Specifically, if a particular cognate set is present in a language, it will be less likely that other cognate sets for the same meaning will also be present. However, we do not think that this lack of independence biases our time depth estimates.

We note that the assumption of independence does not hold for nucleotide or amino acid sequence data either. For example, compensating substitutions in ribosomal RNA sequences result in correlation between paired sites in stem regions (Felsenstein, 2004). However, biologists still get reasonably accurate estimates of phylogeny despite violations of this assumption. In fact, Evans *et al.* (in press) were not able to demonstrate that coding the data as binary characters, rather than as multi-state characters, will produce biased results. Pagel and Meade (in press) demonstrated that, on the contrary, binary and multi-state coded data produce trees that differ in length by a constant of proportionality. In other words, the binary and multi-state trees are just scaled versions of one another. Since we estimate rates of evolution for each tree using the branch lengths of that tree, scaling the branch lengths does not affect our results. Pagel and Meade (in press) also approximated the effect of violations of the independence assumption on the MCMC analysis by 'heating' the likelihood scores. They inferred that violations of independence may produce higher posterior probability values but would have little effect on the consensus tree topology. This means that we may have underestimated the error due to phylogenetic uncertainty but our estimates will not be biased towards any particular date.

Finally, treating cognate sets as the fundamental unit of lexical evolution does not, as Evans *et al*. (in press) argue, constitute an "extreme violation" of the independence

assumption. Almost all of the languages in the Dyen *et al*. (1992, 1997) database contain polymorphisms, meaning that for a given language there exist multiple words of the same meaning. The polymorphisms in our data are a reflection of the nature of lexical evolution. Specifically, they demonstrate a lack of strict dependence between cognate sets within meaning categories – i.e. a word with a given meaning can arise in a language that already has a word of that meaning. Models of lexical evolution that do not allow polymorphisms (e.g. Ringe *et al*., 2002) could also be labeled as inappropriate because they assume that for a word to arise in a language any existing words with that meaning must be concomitantly lost from the language. This is not always the case. Ringe *et al* (2002) note themselves that although the words "small" and "little" have very similar meanings, they have persisted together in English for over a thousand years. Our binary coding procedure allows us to represent such polymorphisms with ease. The presence of polymorphisms means that dependencies between cognate sets are not as strong as Evans *et al*. (in press) claim. A further factor that weakens the dependencies between the cognate sets arises from the 'thinning" process that occurs in lexical evolution. The observed cognate sets do not represent the full compliment of actual cognates that arose in Indo-European. Some cognates that existed in the past will not have persisted into present day languages and any unique "cognates" were not included in the analysis. This "thinning" of the cognates also acts to reduce dependencies between characters in the analysis and thus further weakens any effect of violations of independence. The issue of independence is investigated further in Chapter 5 using synthetic data. There it is shown that violations of the independence assumption do not significantly affect date estimates.

4.5.3 CONFIDENCE IN LEXICAL DATA

From a phylogenetic viewpoint the lexicon is a tremendously attractive source of data because of the large number of possible characters it affords. However, we are aware that many historical linguists are sceptical of inferences based purely on lexical data. Garrett (in press) argues that borrowing of lexical terms, or advergence, within the major Indo-European subgroups could have distorted our results. He identifies a number of cases where an ancestral term has been replaced by a different term in all or some of the daughter languages, presumably via borrowing:

> "Thus Latin *ignis* 'fire' has been replaced by reflexes of Latin *focus* 'hearth' throughout Romance, and archaic Sanskrit *hanti* 'kills' has been re-placed by reflexes of a younger Sanskrit form *marayati* throughout Indo-Aryan."

Garret argues correctly that, where a word has been borrowed across a subgroup after the initial divergence of the group, our method will infer that the word evolved in the branch leading up to that subgroup (see Latin *focus* example, figure 4.15a). This will falsely inflate the branch lengths below the subgroups and deflate branch lengths within each group. Since we estimate rates of evolution on the basis of within-group branch lengths, it is argued that we will underestimate rates of change and hence overestimate divergence lower in the tree. However, this argument requires that two special assumptions hold. First, any borrowing must occur across a whole subgroup and only across a whole subgroup. When terms are not borrowed across the whole group there is no systematic bias to infer changes in the branch leading to the group. Depending on the distribution of borrowed terms, advergence can even produce the opposite effect, falsely inflating branch lengths within subgroups and hence causing us to underestimate divergence times. It seems unlikely that all or even most borrowed terms were borrowed across an entire subgroup. Garrett (in press) highlights 16 instances of borrowing within Indo-European subgroups[1]. These were presumably selected because they were thought to reflect the sort of advergence pattern that would bias our results. Of these, at least 6 are unlikely to favour inferred language change at the base of a subgroup[2]. Figure 4.15b shows the example of the Romance term for "head".

Second, even if we accept the first assumption, we must assume that the proposed process of advergence is unique to contemporary languages. As Garrett (in press) puts it, this requires "the unscientific assumption that linguistic change in the period for which we have no direct evidence was radically different from change we can study directly". Rather than arguing that borrowing was rare at one stage and then suddenly became common across all of the major lineages at about the same time, it seems

[1] The proposed borrowings were: in Romance - ear, fire, liver, count, eat, head and narrow; in Germanic - leaf, sharp and think; and in Indic – kill, night, play, suck, flower and liver. We note that this list was not intended by Garrett to be a comprehensive account of all possible borrowings.

[2] Borrowings that are unlikely to favour inferred language change at the base of a subgroup or that would favour inferred language change within a subgroup are: in Romance - ear, head, narrow; in Germanic – leaf; and in Indic – flower and liver.

more plausible to suggest that borrowing has always occurred. If the same process of advergence in related languages has always occurred then the effect of shifting implied changes to more ancestral branches will be propagated down the tree such that there should be no net effect on divergence time calculation. For example, borrowing within Italic may shift inferred changes from the more modern branches to the branch leading to Italic, but borrowing between Proto-Italic and its contemporaries will also shift inferred changes from this branch to ancestral branches. This means that whilst we may incorrectly reconstruct some Proto-Indo-European roots, our divergence time calculation will not be affected. We maintain that although advergence has undoubtedly occurred throughout the history of Indo-European, and that this may have affected our trees, this effect is likely to be random and there is no reason to think it will have significantly biased our results. By analysing synthetic data with simulated borrowing, in Chapter 5 we show that date estimates are highly robust to even high levels of borrowing.

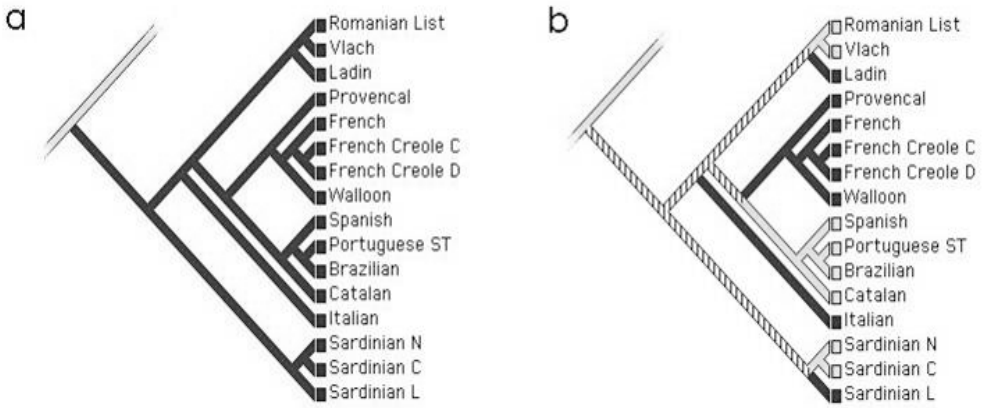


FIGURE 4.15: a *Parsimony character trace for reflexes of Latin* focus *(originally 'hearth' but borrowed as 'fire') on the Romance consensus tree. Black indicates presence of the character, grey indicates absence and dashed indicates uncertainty. This shows borrowing across the whole Romance subgroup – evolutionary change is inferred at the base of the subgroup with no change within the subgroup, falsely inflating divergence time estimates.* ***b*** *As with panel (a) but for reflexes of Latin* testa *(originally 'cup, jar, shell' but borrowed as 'head'). Here the borrowing is not across the whole Romance subgroup – evolutionary change is inferred within the subgroup.*

Ringe *et al*. (2002) argue that non-lexical characters such as grammatical and phonological features are less likely to be borrowed (although they also note that parallel changes in phonological and morphological characters are possible). To avoid potential problems due to lexical borrowing they coded 16 phonological and 22 morphological characters as strict constraints in their analyses (they did not throw out the remaining 333 lexical characters). While we agree that phonological and

morphological characters would be very useful, we believe there are good reasons to trust the inferences based on the lexical data in our case. The Dyen *et al*. (1992, 1997) data has had much of the known borrowing filtered from it. Further, the relationships we infer between Indo-European languages are remarkably similar to those inferred by linguists using the comparative method. Our results are not only consistent with accepted language relationships, but also reflect acknowledged uncertainties, such as the position of Albanian. Our time depth estimates for internal nodes of the Indo-European tree are also congruent with known historical events (see Sections 4.3.4 and 4.3.5.3 and Gray and Atkinson, 2003). Significantly, if we constrain our trees to fit the Ringe *et al*. (2002) typology we get very similar date estimates to our initial consensus tree topology. In short, there is nothing to indicate that either our tree typologies or date estimates have been seriously distorted by the use of just lexical data.

Determined critics might still claim that any remaining undetected lexical borrowing that exists in the Dyen *et al*. data has led us to make erroneous time depth inferences at the root of the Indo-European tree. The Swadesh 100 word list is expected to be more resistant to change and less prone to borrowing than the 200-word list (Embleton, 1991, McMahon and McMahon, 2003). If undetected borrowing has biased our tree topology and divergence time estimates then the 100-word list might be expected to produce different estimates. To assess this possibility we repeated the analysis using only the Swadesh 100 word list items. Figure 4.16 shows the results of this analysis. Predictably, with a smaller data set variance in the age estimates increased. However, the resulting age range was still consistent with the Anatolian theory of Indo European origin. Interestingly, the majority rule consensus tree (shown in Figure 4.17) is slightly different to that obtained from the full data set. It contains a Balto-Slavic-Indo-Iranian group and an Italo-Celtic group. Ringe *et al.*'s (2002) compatibility analysis also found these clades. The low posterior probability values for these groups mean that we should not over-interpret the certainty of these deeper relationships, but clearly the possibility that undetected lexical borrowing is obscuring some of the deeper relationships would repay further attention. We emphasise that this possible borrowing does not appear to affect our time depth estimates for the root of the tree however.

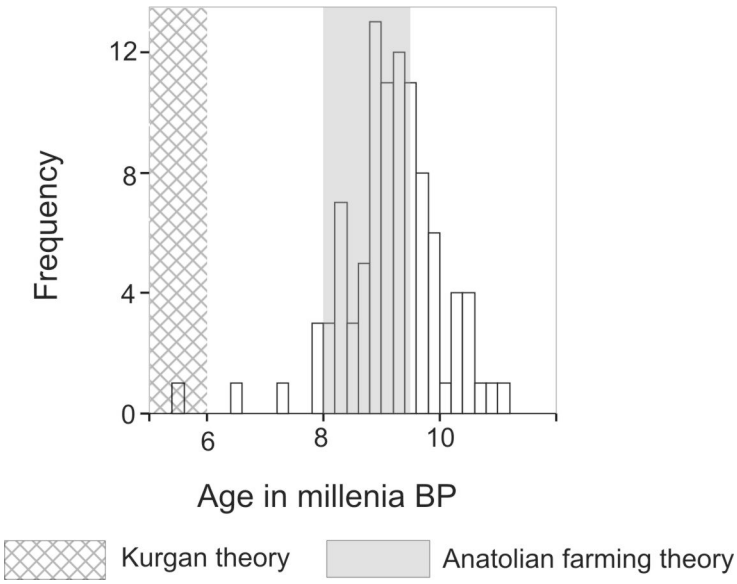


FIGURE 4.16: Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees using Swadesh 100 word list items only (n=97).

It is interesting to note that whilst our methodology produced consistent results using the Swadesh 100 and 200 word list, Tischler's (1973) glottochronological analysis was affected by the choice of word list. Tischler generated Indo-European divergence times using pair-wise distance comparisons between languages under the assumption of constant rates of lexical replacement. Using the Swadesh 200 word list, he calculated that the core Indo-European languages (Greek, Italic, Balto-Slavic, Germanic and Indo-Iranian) diverged around 5500BP whilst Hittite diverged from the common stock around 8400BP. This is in striking agreement with the timing depicted in Figure 8. However, the same calculation using the Swadesh 100 word list, produced a Hittite divergence time of almost 11000BP. Other inferred divergence times were also older. Tischler favoured the 200 word list results because they tended to be more consistent and were based on a larger sample size. However, the disparate 100 word list ages led Tischler to conclude that the divergence times for Hittite (and a number of other peripheral Indo-European languages, including Albanian, Armenian and Old Irish) were in fact anomalous and he instead favoured an age for Indo-European of between 5000 and 6000 years, reflecting the break-up of the core languages. He explained the apparent earlier divergence of Hittite, Albanian, Old Irish and Armenian as an artifact of borrowing with non-Indo-European languages or increased rates of change.

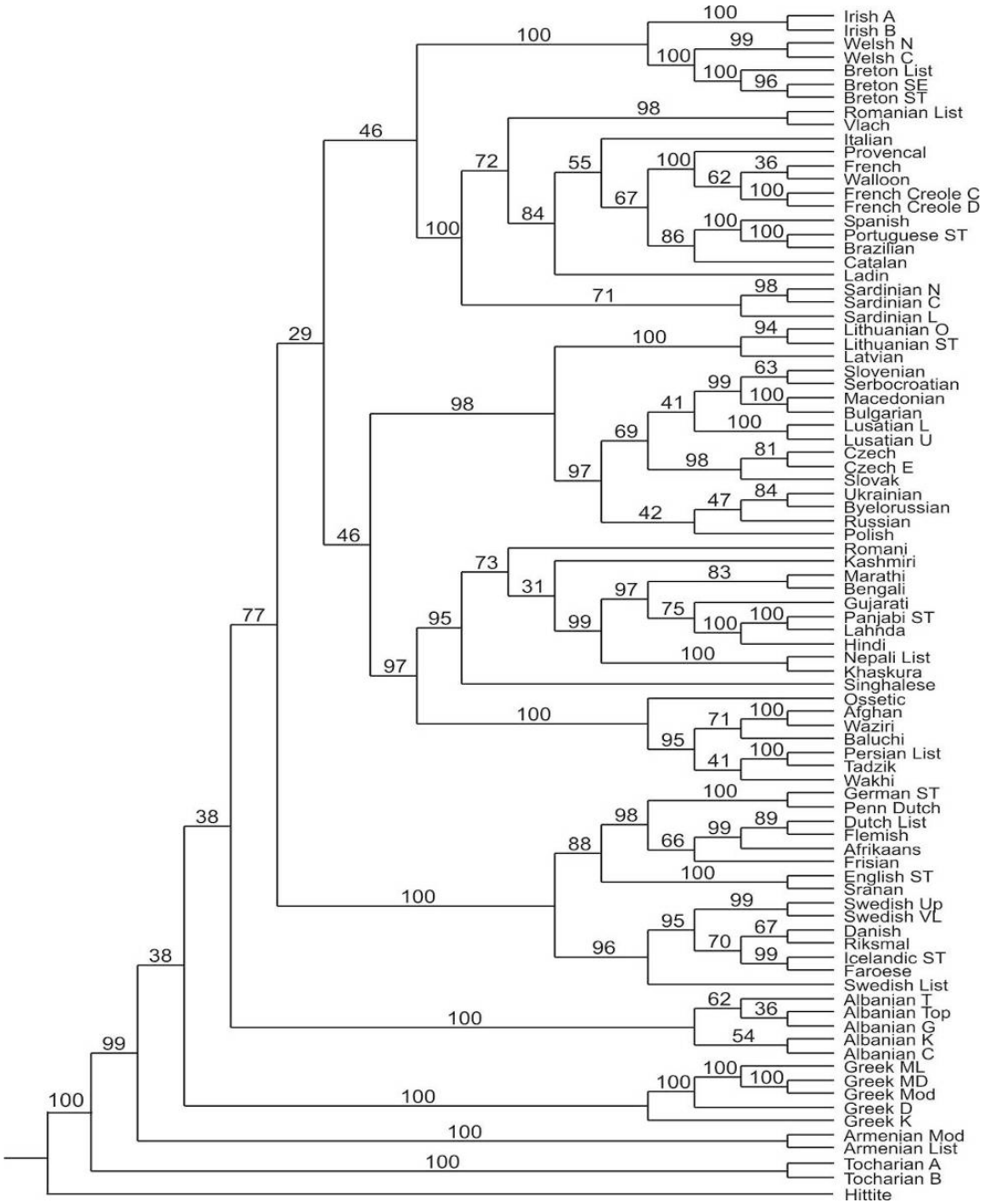


FIGURE 4.17: *Majority-rule consensus tree (unfiltered) for Swadesh 100 word-list items only. Values above each branch express uncertainty (posterior probability) in the tree as a percentage.*

4.6 CONCLUSION

The analyses we have presented here are far from the last word on the vexed issue of Indo-European origins. We expect that "every species of scholar and would be savant who can take pen to hand" will still be drawn to the question of Indo-European origins. However, in contrast to some of the more pessimistic claims of our critics, we do not think that estimating the age of the Indo-European language family is an

intractable problem. Some of these critics have argued that it is hard enough to get the tree typology correct, let alone branch lengths or divergence times. From this point of view all efforts to estimate dates should be abandoned until we can get the tree exactly right. We think that would be a big mistake. It would prematurely close off legitimate scientific inquiry. The probability of getting the one "perfect phylogeny" from the 6.66×10^{152} possible unrooted trees for 87 languages is rather small. Fortunately we do not need to get the tree exactly correct in order to make accurate date estimates. Using the Bayesian phylogenetic approach we can calculate divergence dates over a distribution of most probable trees, integrating out uncertainty in the phylogeny. We acknowledge that estimating language divergence dates is difficult, but maintain it is possible if the following conditions are satisfied:

a. a data set of sufficient size and quality can be assembled to enable the tree and its associated branch lengths to be estimated with sufficient accuracy,
b. most of the borrowing is removed from the data,
c. an appropriate statistical model of character evolution is used (it should contain sufficient parameters to give accurate estimates but not be over-parameterised),
d. multiple nodes on the tree are calibrated with reliable age ranges,
e. uncertainty in the estimation of tree topology and branch lengths are incorporated into the analysis.
f. variation in the rate of linguistic evolution is accommodated in the analysis.

The analyses of Indo-European divergence dates we have outlined above go a long way to meeting these requirements. The Dyen *et al*. (1997) data set we used in our analyses contains over two thousand carefully coded cognate sets (condition a). Dyen *et al*. excluded known borrowings from these sets (condition b). The two state, time-reversible model of cognate gains and losses with gamma distributed rate heterogeneity produced accurate trees (i.e. congruent with the results of the comparative method and known historical relationships) (condition c). When the branch lengths were combined with the large number of well-calibrated nodes (condition d), the estimated divergence dates were also in line with known historical events. The Bayesian MCMC approach allowed us to incorporate phylogenetic uncertainty into our analyses (condition e), and to investigate the consequences of variations in the priors, tree rooting, and stringency in cognate judgements. Finally,

rate smoothing allowed us to estimate divergence dates without the assumption of a strict glottoclock (condition f). We challenge our critics to find any paper on molecular divergence dates that uses as many calibration points, investigates the impact of so many different assumptions, or goes to the same lengths to validate its results.

In the words of W. S. Holt, history is "a damn dim candle over a damn dark abyss". Although we see reason for careful scholarship when attempting to estimate language divergence dates, we see no justification for pessimism here. Far from dancing around the question of Indo-European origins like moths around a flame, with the light of computational phylogenetic methods we can illuminate the past.

4.7 REFERENCES

Adams, D. Q. 1999. *A Dictionary of Tocharian B* (Leiden Studies in Indo-European 10). Amsterdam: Rodopi. Available via online database at S. Starostin and A. Lubotsky (Eds.) Database Query to A dictionary of Tocharian B. http://iiasnt.leidenuniv.nl/ied/index2.html

Atkinson, Q. D., Nicholls, G., Welch, D. and R. D. Gray. 2005. From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103(2):193–219.

Bellwood, P. 1991. The Austronesian dispersal and the origin of languages. *Scientific America,* 265:88-93.

Bellwood, P. 1994. An archaeologist's view of language macrofamily relationships. *Oceanic Linguistics,* 33:391-406.

Bergsland, K., and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology,* 3:115–53.

Blust, R. 2000. Why lexicostatistics doesn't work: the 'Universal Constant' hypothesis and the Austronesian languages. Pages 311-332 in *Time Depth in Historical Linguistics,* (eds.) C. Renfrew, A. McMahon, and L. Trask. McDonald Institute for Archaeological Research, Cambridge.

Bryant, D., F. Filimon, and R. D. Gray. 2005. Untangling our past: Pacific settlement, phylogenetic trees and Austronesian languages. Pages 69–85 in *The evolution of cultural diversity: Phylogenetic approaches* (eds.) R. Mace, C. Holden, and S. Shennan. UCL Press, London.

Bryant, D., and V. Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. *Workshop in Algorithms for Bioinformatics, Proceedings* 2002:375–391.

Burnham, K. P. and D. R. Anderson. 1998. *Model Selection and Inference – A practical Information-Theoretic Approach.* Springer, New York.

Campbell, L. 1997. *American Indian Languages*. Oxford University Press, Oxford.

Campbell, L. 1998. *Historical Linguistics: An Introduction*. Edinburgh University Press, Edinburgh.

Campbell, L. 2004. *Historical linguistics: An Introduction, 2nd edition*. Edinburgh University Press, Edinburgh.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press, Princeton (NJ).

Champion, T., Gamble, C., Shennan, S. and A. Whittle. 1984. *Prehistoric Europe*. Academic Press, New York.

Chikhi, L., R. A. Nichols, G. Barbujani, and M. A. Beaumont. 2002. Y Genetic Data Support the Neolithic Demic Diffusion Model. *Proceedings of the National Academy of Sciences*, 99:11008–11013.

Clackson, J. 2000. Time depth in Indo-European. Pages 441-454 in *Time Depth in Historical Linguistics,* (eds.) C. Renfrew, A. McMahon, and L. Trask. McDonald Institute for Archaeological Research, Cambridge.

Devoto, G. 1962. *Origini Indeuropeo.* Instituto Italiano di Preistoria Italiana, Florence.

Diakonov, I. M. 1984. On the original home of the speakers of Indo-European. *Soviet Anthropology and Archaeology* 23, 5–87.

Diamond, J. and P. Bellwood. 2003. Farmers and their languages: the first expansions. *Science* 300, 597.

Dyen, I., J. B. Kruskal, and P. Black. 1992. *An Indoeuropean Classification: A Lexicostatistical Experiment.* American Philosophical Society, Transactions 82(5). Philadelphia.

Dyen, I., J. B. Kruskal, and P. Black. 1997. FILE IE-DATA1. Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.

Embleton, S. 1986. *Statistics in Historical Linguistics.* Brockmeyer, Bochum.

Embleton, S. M. 1991. Mathematical methods of genetic classification. Pages 365–388 in *Sprung from Some Common Source*, eds S. L. Lamb and E. D. Mitchell. Stanford University Press, Stanford (CA).

Excoffier, L., and Z. Yang. 1999. Substitution Rate Variation among Sites in Mitochondrial Hypervariable Region I of Humans and Chimpanzees. *Molecular Biology and Evolution* 16:1357–68.

Evans, S. N., Ringe, D. and Warnow, T. in press. Inference of divergence times as a statistical inverse problem. In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge.

Faguy, D.M. and W.F. Doolittle. 2000. Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. *Trends in Genetics* 16, 196-7.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer, Sunderland (MA).

Gamkrelidze, T. V., and V. V. Ivanov. 1995. *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture*. Mouton de Gruyter, Berlin.

Garrett, A. (in press) Convergence in the formation of Indo-European subgroups: Phylogeny and chronology. In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge.

Gimbutas, M. 1973a. Old Europe c. 7000–3500 B.C., the earliest European cultures before the infiltration of the Indo-European peoples. *Journal of Indo-European Studies* 1, 1-20.

Gimbutas, M. 1973b. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500–2500 B.C. *Journal of Indo-European Studies* 1, 163–214.

Gkiasta, M., T. Russell, S. Shennan, and J. Steele. 2003. Neolithic transition in Europe: The radiocarbon record revisited. *Antiquity* 77, 45–62.

Glover, I., and C. Higham. 1996. New evidence for rice cultivation in S., S.E., and E. Asia. Pages 413-442 in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. (ed.) D. Harris. Blackwell, Cambridge.

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435-9.

Guterbock, H. G. and H. A. Hoffner. 1986. *The Hittite dictionary of the Oriental Institute of the University of Chicago.* The Institute, Chicago.

Hillis, D. M. 1992. Experimental phylogenetics generation of a known phylogeny. *Science* 255, 589-92.

Hillis, D. M., Moritz, C. and B. K. Marble. 1996. *Molecular Systematics*, second ed. Sinauer, Sunderland (MA).

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: A maximum-parsimony analysis. *Royal Society of London, Proceedings Series B* 269, 793–9.

Holland, B. and Moulton, V. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. Pages 165-176 in *Algorithms in bioinformatics, WABI 2003*, (eds.) G. Benson and R. Page. Springer-Verlag, Berlin.

Hoffner, H. A. 1967. *An English-Hittite Dictionary*. American Oriental Society, New Haven.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–5.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310-2314.

Huson, D. H. 1998. SplitsTree: Analyzing and Visualizing Evolutionary Data. *Bioinformatics* 14:68–73.

Jones, Sir W. 1807. Third Anniversary Discourse, 'On the Hindus.' Pages 24-46 in *The Collected Works of Sir William Jones, vol. 3*. Stockdale and Walker, London.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 in *Mammalian Protein Metabolism*, Vol. 3, ed. M.N. Munro. Academic Press, New York.

Kuhner, M. K., and J. Felsenstein. 1994. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution* 11:459–468.

Kumar, V. K. 1999. *Discovery of Dravidian as the Common Source of Indo-European*. Retrieved Sept. 27th 2002 from http://www.datanumeric.com/dravidian/

Levins, R. 1966. The strategy of model building in population biology, *American Scientist* 54, 421-31

Mallory, J. P. 1989. *In Search of the Indo-Europeans: Languages, Archaeology and Myth.* Thames and Hudson, London.

McMahon, A. and R. McMahon. 2003. Finding Families: Quantitative methods in language classification. *Transactions of the Philological Society* 101, 7-55.

Menozzi, P., A. Piazza, and L. L. Cavalli-Sforza. 1978. Synthetic Maps of Human Gene Frequencies in Europeans. *Science* 201:786–792.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087–91.

Nicholls, G., and R. Gray. in press. Quantifying uncertainty in a stochastic dollo model of vocabulary evolution. In *Phylogenetic methods and the prehistory of languages.* (eds.) J. Clackson, P. Forster and C. Renfrew. McDonald Institute for Archaeological Research, Cambridge.

Otte, M. 1997. The diffusion of modern languages in prehistoric Eurasia. Pages 74-81 in *Archaeology and Language.* (eds.) R. Blench and M. Spriggs. Routledge, London.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26, 331–48.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401, 877-84.

Pagel, M. 2000. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. Pages 189-207 in *Time Depth in Historical Linguistics*. (eds.) C. Renfrew, A. McMahon and L. Trask. The McDonald Institute for Archaeological Research, Cambridge.

Pagel, M., and A. Meade. In press. Estimating rates of meaning evolution on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages.* (eds.) J. Clackson, P. Forster and C. Renfrew. McDonald Institute for Archaeological Research, Cambridge.

Renfrew, C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins.* Cape, London.

Renfrew, C. 2000. 10,000 or 5000 Years Ago? Questions of Time Depth. Pages 413-439 in *Time Depth in Historical Linguistics*. (eds.) C. Renfrew, A. McMahon, and L. Trask. McDonald Institute for Archaeological Research, Cambridge.

Rexová, K., D. Frynta, and J. Zrzavy. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19, 120–127.

Richards, M., *et al*. (2000). Tracing European Founder Lineages in the Near Eastern mtDNA Pool. *American Journal of Human Genetics* 67:1251–1276.

Ringe, D. n.d. Proto-Indo-European wheeled vehicle terminology. *Unpublished manuscript.*

Ringe, D., T. Warnow, and A. Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100, 59–129.

Ringe, D., T. Warnow, and A. Taylor, A. Michailov, and L. Levinson. 1998. Computational Cladistics and the Position of Tocharian. *Journal of Indo-European Studies* 26:391–414.

Rosser, Z. H. *et al.* 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by Language. *American Journal of Human Genetics* 67, 1526–43.

Sanderson, M. 2002a. *R8s, Analysis of Rates of Evolution, version 1.50*. http://ginger.ucdavis.edu/r8s/

Sanderson, M. 2002b. Estimating absolute rates of evolution and divergence times: A penalised likelihood approach. *Molecular Biology and Evolution* 19, 101–109.

Sanderson, M. J., and M. J. Donoghue. 1989. Patterns of Variation in Levels of Homoplasy. *Evolution* 43:1781–1795.

Schleicher, A. A. 1863. *Die Darwinische Theorie und die Sprachwissenschaft*. Bohlau, Weimar, Germany.

Semino, O., *et al*. 2000. The Genetic Legacy of Palaeolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. *Science* 290:1155–1159.

Steel, M., M. Hendy, and D. Penny. 1988. Loss of information in genetic distances. *Nature* 333, 494-495.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96, 453–63.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 121–137.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic Inference. Pages 407-514 in *Molecular Systematics*, second ed. (eds.) D. M. Hillis, C. Moritz, and B. K. Marble. Sinauer, Sunderland (MA).

Tischler, J. 1973. *Glottochronologie und Lexicostatistik*. Innsbrucker Verlag, Innsbruck.

Tischler, J. 1997. *Hethitisch-Deutsches Worterverzeichnis*. Probedruck, Dresden.

Trask, R. L. 1996. *Historical Linguistics.* Arnold, New York.

Trask, R. L. 2003a. Re: Language Tree Rooted in Turkey. http://groups.yahoo.com/group/evolutionary-psychology/message/28240

Trask, R. L. 2003b. Re: Language Tree Rooted in Turkey. http://groups.yahoo.com/group/evolutionary-psychology/message/28291

Warnow, T., S. N. Evans, D. Ringe, and L. Nakhleh. In press. Stochastic models of language evolution and an application to the Indo-European family of languages.In *Phylogenetic methods and the prehistory of languages.* (eds.) J. Clackson, P. Forster and C. Renfrew. McDonald Institute for Archaeological Research, Cambridge.

Watkins, C. 1969. *Indogermanische Grammatik III/1. Geschichte der Indogermanischen Verbalflexion.* Carl Winter Verlag, Heidelberg.

Chapter Five

FROM WORDS TO DATES - WATER INTO WINE, MATHEMAGIC OR PHYLOGENTIC INFERENCE?[1]

ABSTRACT

The application of quantitative phylogenetic methods to Dyen, Kruskal and Black's (1992, 1997) Indo-European lexical data presented in Chapters 3 and 4 produced controversial divergence time estimates. Here we test the robustness of these results using an alternative data set of ancient Indo-European languages. We employ two very different stochastic models of word evolution - Gray and Atkinson's (2003) time-reversible model, used in the previous chapter, and a stochastic-Dollo model of word evolution introduced by Nicholls and Gray (in press). Results of this analysis support the findings reported in Chapters 3 and 4. We also tested the ability of both methods to reconstruct phylogeny and divergence times accurately from synthetic data. The methods performed well under a range of scenarios, including widespread and localized borrowing.

5.1 WORDS INTO DATES OR WATER INTO WINE?

Morris Swadesh (1952, 1955) formalized the idea of inferring language divergence times from word lists when he developed lexicostatistics and glottochronology. Although initially received with some enthusiasm, as was discussed in Chapter 4, glottochronology was heavily criticised and most linguists today view it as discredited (Campbell, 2004). The well-known criticisms of glottochronology outlined in Chapter 4 were:

1. loss of information reduces the power of the method,
2. inaccurate tree-building techniques produce erroneous trees,

[1] Atkinson, Q. D. Nicholls, G., Welch, D. and R. D. Gray. 2005. From Words to Dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2):193-219.

3. borrowing between languages confounds divergence time estimates,
4. the assumption of constant rates through time does not hold.

We argued that whilst the criticisms of glottochronology reflect legitimate concerns, it is possible to estimate dates using a different class of methods. As we showed in Chapters 3 and 4, current statistical phylogenetic methods used widely in biology allow us to overcome the problems associated with glottochronology. First, character-based stochastic models of evolution retain phylogenetic information from the source data and allow us to reconstruct phylogeny accurately, even under conditions of rate heterogeneity (Huelsenbeck *et al*., 2001). Character-based methods can also account for polymorphisms (multiple words for the same meaning) and uncertainty in vocabulary assignments and cognacy judgements. Further, by using an explicit model of evolution, the assumptions of the model are clear and the effects of changing these assumptions can be tested easily. Second, Bayesian phylogenetic inference allows us to quantify random error in tree topology and branch-lengths, a crucial factor if results are to be used to test between competing hypotheses. Third, the degree of reticulate evolution and borrowing between languages can be assessed using visualization tools, like *SplitsTree* (Huson, 1998) and *NeighborNet* (Bryant and Moulton, 2002), that do not assume a tree-like model of evolution (Bryant, Filimon and Gray, 2005). A statistical framework also makes it possible to systematically test the robustness of results to some violations in the assumptions of the method, including the nature and magnitude of borrowing. For example, in Chapter 4 we presented evidence that divergence time estimates were robust to borrowing by comparing a number of different coding procedures and analysing subsets of the lexicon thought to be more resistant to borrowing. Finally, violations of the assumption of rate constancy can be investigated in a similar fashion. Where there is a concern about its effect, rate smoothing algorithms allow divergence times to be estimated without the assumption of a strict glottoclock. In Chapter 3 and 4, this approach was used to test between two competing hypotheses for the age of Indo-European.

Many features of vocabulary evolution are not represented in the classes of models used here. As a consequence uncertainty arises from two sources: random error, which would be present if the models were perfect descriptions of vocabulary evolution, and which arises from stochastic fluctuations predictable in distribution; and systematic error, the estimation bias caused by model misspecification. Random

error will be well-estimated in the Bayesian framework employed here. However, the problem of quantifying the uncertainty due to model misspecification is much harder, since the number of "real" language evolution models one might reasonably entertain is enormous. Progress can be made only by focusing on those effects thought likely to be important and estimating the size (and direction) of the biases they cause (see Section 5.6)

5.2 A NEW SET OF ANCIENT DATA

The analyses presented in Chapters 3 and 4 were based on a lexical dataset comprising 200 Swadesh list meanings in 87 languages derived from Dyen, Kruskal and Black's (1992) Indo-European lexical database. Divergence time estimates from Chapter 3 (Gray and Atkinson, 2003) suggested a root age of Indo-European of between 7,800 and 9,800BP, consistent with the Anatolian theory of Indo-European origin (Renfrew, 1987). Crucially, this age range was outside the 5,000 to 6,000BP age range implied by the alternative, Kurgan, theory of Indo-European origin (Gimbutas, 1973a,b). These results were supported by further analyses presented in Chapter 4 (Atkinson and Gray, 2006, in press) and by Nicholls and Gray (in press). Here we apply a number of new techniques and tools of analysis to an alternative Indo-European dataset compiled by Ringe, Warnow and Taylor (2002). This data includes 330 meaning categories in 20 extinct and 4 extant Indo-European languages.

Repeating the analyses from Chapters 3 and 4 on a second Indo-European dataset has a number of benefits. The data include different languages and some different meaning categories to those in the Dyen *et al*. (1992) database. In addition, cognacy judgements were made by Don Ringe, an independent expert in the field, who was not involved with the Dyen *et al*. (1992) study. These facts alone make this analysis a particularly interesting test of the methodology. Different date estimates for each set of data would cast doubt on the reliability of the method and/or the amount of temporal information in the data. Conversely, consistent dates across both datasets would support the idea that there is a strong temporal signal in lexical data and that the methods employed are robust to variations in the cognacy judgement criteria, meaning categories used and even the languages analysed.

Of course, agreement across data sets might simply indicate a bias in the inferential methodology. If this were the case we would expect the methodology to show similar biases on some synthetic data. In Section 5.6, we test the methodology on a range of synthetic data simulated under models that were designed to include features thought to be problematic for the methodology. By deliberately setting up model misspecification in this way we can test for its effect.

Another advantage of using the Ringe *et al*. (2002) data is that it comprises mainly ancient languages. First, all other things being equal, this should improve the resolution of some of the deeper Indo-European branching structure. Second, as discussed in Chapter 4, Garrett (in press) has found evidence for modern advergence processes in a number of Indo-European sub-groups. He points out that the analyses presented in Gray and Atkinson (2003) used mainly contemporary languages and hence may be biased by the effect of certain types of unidentified modern borrowing. We discussed in Chapter 4 why this is unlikely. However, here, by analysing a dataset of ancient languages, many of which are two or three thousand years old, we can all but eliminate the effect of modern borrowing. Moreover, our synthetic studies show that if local borrowing is present in ancient and modern languages this need not produce a bias. Where model distortion is mild, and uniform over time, parameters distorted to fit calibration points near the leaves predict ages for unattested branching events.

5.3 STOCHASTIC MODELS AND BAYESIAN INFERENCE OF PHYLOGENY

As described in Chapter 4, stochastic models of evolution and Bayesian inference of phylogeny allow us to overcome the problems identified with the distance-based tree-building methods used in lexicostatistics and glottochronology. Bayesian and related likelihood-based inference can outperform distance and parsimony tree-building methods in situations where models are reliable and there are unequal rates of change (Kuhner and Felsenstein 1994). The basic procedure for calculating the likelihood score for the model of Gray and Atkinson (2003) is described in Chapter 4 - a more detailed explanation in a biological context can be found in Swofford *et al*. (1996).

Bayesian inference of phylogeny and Markov chain Monte Carlo (MCMC) algorithms (Metropolis *et al*., 1953) were used to generate a sample distribution of trees that reflects the component of phylogenetic uncertainty in our analysis due to *random* error – the potential for *systematic* error due to model misspecification is investigated in Section 5.6. This inference is based on the data, an observation model and a set of prior beliefs (or *priors*) about all unknown parameters of the model, including the tree topology, branch-lengths, and rate matrix. We favour priors that are uninformative with respect to the hypotheses being investigated. However, an important component of any Bayesian MCMC analysis is to ensure that results are robust to a range of sensible priors.

Crucially, the MCMC posterior sample of trees is much more informative about the phylogenetic signal in the data than methods that return any single "optimal" phylogeny (*cf.* Ringe *et al*., 2002). The sample distribution allows us to approximate phylogenetic uncertainty (uncertainty in tree topology and branch lengths), given the data, and to incorporate this into our results. This is impossible using either the comparative method or traditional glottochronology, and yet the effect of phylogenetic uncertainty is a crucial consideration if results are to be used to test historical hypotheses. We can calculate divergence times across the entire Bayesian sample distribution and determine the error in date estimates resulting from phylogenetic uncertainty. This also means that we can make inferences about the age of Indo-European without having to commit to a particular topology.

5.4 TWO DIFFERENT APPROACHES TO LIKELIHOOD INFERENCE AND MODELING

Most scientists would agree that the best way to validate a result is to repeat the analysis, preferably on an independent data set. A close second, however, may be to reanalyse the data using a different methodology and model. Here then, as well as analysing an alternative dataset, we employ two very different models of language evolution, based on a very different set of core assumptions. Each model and methodology is described briefly below.

5.4.1 TIME-REVERSIBLE MODEL AND METHOD 1

Chapters 3 and 4 use a model of binary character evolution implemented in the program *MrBayes* (Huelsenbeck and Ronquist, 2001) to generate a sample distribution of trees with branch lengths proportional to the inferred amount of evolutionary change. We can model the process of cognate gain (0 to 1) and loss (1 to 0) using three parameters - μ, the mean substitution rate, and π_0 and π_1, which represent the relative frequencies of 1's and 0's. The substitution rate is estimated in the MCMC analysis and the equilibrium frequency of 1's and 0's can be estimated from their frequency in the data. Missing data is treated as another unknown binary parameter to be estimated.

As there are just two states, 0 and 1, this model is time-reversible – we cannot tell the direction in which the cognate evolved from its history in a single language. This model allows a single cognate to appear in and disappear from a single language more than once over the course of time, allowing the model to mimic the effect of word-borrowing. As the direction of time is not determined, we cannot determine the root of the tree from the data – we need to provide an outgroup as a root. For all the method 1 analyses reported here, trees were rooted with Hittite, consistent with independent linguistic analyses (Gamkrelidze and Ivanov, 1995; Rexova, Frynta and Zrzavy, 2003). The analyses in Chapters 3 and 4 showed that changing the root point did not affect age estimates significantly. A gamma shape parameter (α) was used to allow for rate variation between words. As with the overall rate parameter, α was estimated from the data. An α value of 5 was observed, indicating moderate rate variation.

Rate variation between lineages and through time was accounted for by relaxing the assumption of a strict glottoclock. The 87 languages in the modified Dyen *et al*. (1997) data set allowed for 11 internal clade constraints. Terminal nodes representing contemporary languages were set to 0 years whilst 3 extinct languages (Hittite and Tocharian A & B) were constrained in accordance with estimated ages of the source texts (see Table 4.5, summarized in Table 5.1). For the 24 languages in the Ringe *et al*. (2002) data, 12 internal node constraints were available, whilst 20 extinct

languages were constrained in accordance with estimated ages of the source texts (see Table 5.2). Again, Sanderson's (2002a) penalised likelihood algorithm, as implemented in *r8s* (Sanderson, 2002b), was then used to smooth rates of evolution across each tree and to calculate divergence times. Interestingly, high smoothing factors were found to fit the data best, suggesting that the process of evolution is in fact relatively tightly constrained. The procedure was repeated on all of the trees in the MCMC Bayesian sample distribution. The distribution of divergence times at the root can be used to create a confidence interval for the age of Indo-European.

TABLE 5.1: Age constraints, for the Dyen *et al*. (1997) data set, used to calibrate the divergence time calculations on the basis of known historical information. Summary of Table 4.5 in Chapter 4. Terminal node constraints representing ancient languages are shown in italics.

Calibration	**Age constraint**
Iberian-French	450AD-800AD
Italic-Romanian	150AD-300AD
North/West Germanic	50AD-250AD
Welsh/Breton	400AD-550AD
Irish/Welsh	before 300AD
Indic	before 200BC
Iranian	before 500BC
Indo-Iranian	before 1,000BC
Slavic	before 700AD
Balto-Slavic	1,400BC-100AD
Greek split	before 1,500BC
Tocharic	140BC-350AD
Tocharian A & B	*500AD-750AD*
Hittite	*1,800BC-1,300BC*

There are two potential criticisms of this model. First, the same rate parameter is used to estimate cognate gains and losses. Whilst cognates can be lost relatively easily, the innovation events that produce them are rare - it is very unlikely that two languages would ever independently gain the same cognate. Thus, it may be argued that trying to fit a single rate parameter to a model of cognate gain and loss is inappropriate (Evans, Ringe and Warnow, in press). Indeed, this may be problematic if the rate of gain and loss are widely different. However, processes of borrowing and dialect chains at divergence mean that models which allow cognates to be gained more than once may still be reasonable. In fact, as we will see below, this feature of the model may allow it to accommodate moderate reticulation in the data. A second potential criticism is that method 1 uses a "finite sites" model from biology. This means

character state changes are modelled through time across a fixed number of characters, or 'sites'. However, in reality, the number of cognates or sites is not finite and depends on the number of languages we are looking at and how long they have been evolving. In what follows we describe an alternative model introduced by Nicholls and Gray (in press) that does not assume finite sites.

TABLE 5.2: Age constraints for the Ringe *et al*. (2002) data set, used to calibrate the divergence time calculations on the basis of known historical information. Terminal node constraints representing ancient languages are shown in italics.

Calibration	**Age constraint**
Italic	before 800BC
Germanic	750BC-250BC
North-West Germanic	50AD-250AD
West Germanic	400AD-500AD
Celtic	650BC-300AD
Indic	before 200BC
Iranian	before 500BC
Indo-Iranian	before 1,000BC
Baltic	600AD-700AD
Balto-Slavic	1,400BC-400BC
Greek split	before 1,500BC
Tocharic	140BC-350AD
Vedic	*1,500BC-800BC*
Old Persian	*600BC-300BC*
Avestan	*600BC-400BC*
Old Prussian	*1,250AD-1,600AD*
Old Church Slavonic	*900AD-1,100AD*
Old High German	*850AD-1,050AD*
Old English	*900AD-1,100AD*
Old Norse	*1,150AD-1,350AD*
Gothic	*300AD-400AD*
Armenian	*400AD-800AD*
Greek	*500BC-300BC*
Latin	*200BC-100AD*
Oscan	*400BC-50BC*
Umbrian	*300BC-50BC*
Old Irish	*600AD-900AD*
Tocharian A & B	*500AD-750AD*
Lycian	*500BC-200BC*
Luvian	*1,700BC-1,200BC*
Hittite	*1,700BC-1,200BC*

5.4.2 STOCHASTIC-DOLLO MODEL AND METHOD 2

Dollo's Law states that traits can evolve only once (Farris, 1977). In this context, we treat cognates as traits and assume that the same cognate cannot be independently created in different languages. This assumption is equivalent to asserting that the

cognate data is homoplasy free (*cf*. Ringe *et al*., 2002). Based on this assumption, we outline a stochastic model of language change appropriate to the cognate data used here and described in Chapter 4.

The model allows language change to occur in three different ways: words (and corresponding cognate sets) are created, words are lost, and words reproduce (when languages split, forming two child copies of a parent language). We assume that words are created in any given language at rate λ. When a word is created, it falls into a new cognate class, so word creation and cognate class creation are synonymous. If there are k languages extant at time t, new cognates are created at total rate $k\lambda$. Each word is lost from a given language independently at rate μ. If at time t, there are k languages and language i contains l_i words, word death occurs at a total rate of $\mu(l_1+l_2+...+l_k)$.

Each language splits at rate θ. When a language splits, two child copies of the language are made and the parent language dies. At the time of splitting, the child languages are indistinguishable from the parent language and thereafter evolve in exactly the same way as the parent language did. If there are k languages at time t, language splitting occurs at total rate $k\theta$.

We assume that the times between all events causing language change are exponentially distributed and that all rates – the cognate birth rate, λ, the cognate loss rate, μ, and the language splitting rate, θ – are constant across time and space. We assume also that all languages and cognates evolve independently.

The data described in section 2 is collected in such a way that cognates that are present in no languages or only one language at the time of collection are not recorded. Thus the observed cognate birth rate λ^* is different from the actual cognate birth rate λ since words must be born and survive into at least two languages in order to be observed. This data thinning process may result in the birth times of cognates in the data being unevenly distributed over the tree. This effect is accounted for in the likelihood calculation for a given tree, the details of which are given in Nicholls and Gray (in press).

There are two obvious features of the data that this model fails to capture. The first is missing data. We do not account for missing data and recode any missing cognates as absent. It is necessary to check for biases caused by this approximation. We repeat analyses omitting languages with a significant amount of missing data (13% is used as the cut-off below). The effect of doing this on the relatively complete data sets treated in this paper is negligible.

A more important issue is that of borrowing between languages. If one language gains a new word by borrowing it from another, the Dollo assumption is violated. While it is relatively simple to include borrowing when building a model of language change, we are currently unable to analyse such a model. In order to quantify the magnitude of this misspecification, in Section 5.6 we present a series of analyses of data synthesised under models with borrowing but analysed under the Stochastic-Dollo model.

Inference for the Stochastic-Dollo model is again made within a Bayesian framework and the data is analysed using a MCMC algorithm implemented in *Matlab* by Geoff Nicholls and David Welch. The relevant software, called *TraitLab*, can be downloaded from (aitken.math.auckland.ac.nz/~nicholls/TraitLab/).

5.4.3 OTHER INFERENCE ISSUES

Both of the above models assume that the characters in the binary data are independent. Evans *et al*. (in press) argue that the results of any analysis using these models are invalid due to violations of the assumption of independence. As is discussed in Chapter 4, Evans *et al.* (in press) make the point that the independence assumption is violated when individual meanings in the Swadesh word list are broken up into characters representing multiple cognate sets. Specifically, if a particular cognate set is present in a language, it will be less likely that other cognate sets for the same meaning will also be present. Conversely the Swadesh meaning categories must be occupied at all times in all languages, whereas both models allow meaning categories to be empty. In Section 5.6, we attempt to detect bias caused by this type of misspecification using synthetic data. The problem of empty meaning categories

does not seem to be important. We surmise that this is because ancestral meaning categories need not be filled by cognates present in the data. They may be occupied by ancient cognates with no instances in the data.

5.5 RESULTS

Figure 5.1 shows the results of a series of analyses of both data sets using both of the models of evolution described above. Table 5.3 summarizes these results including the data, priors and other conditions used for each analysis from Figure 5.1.

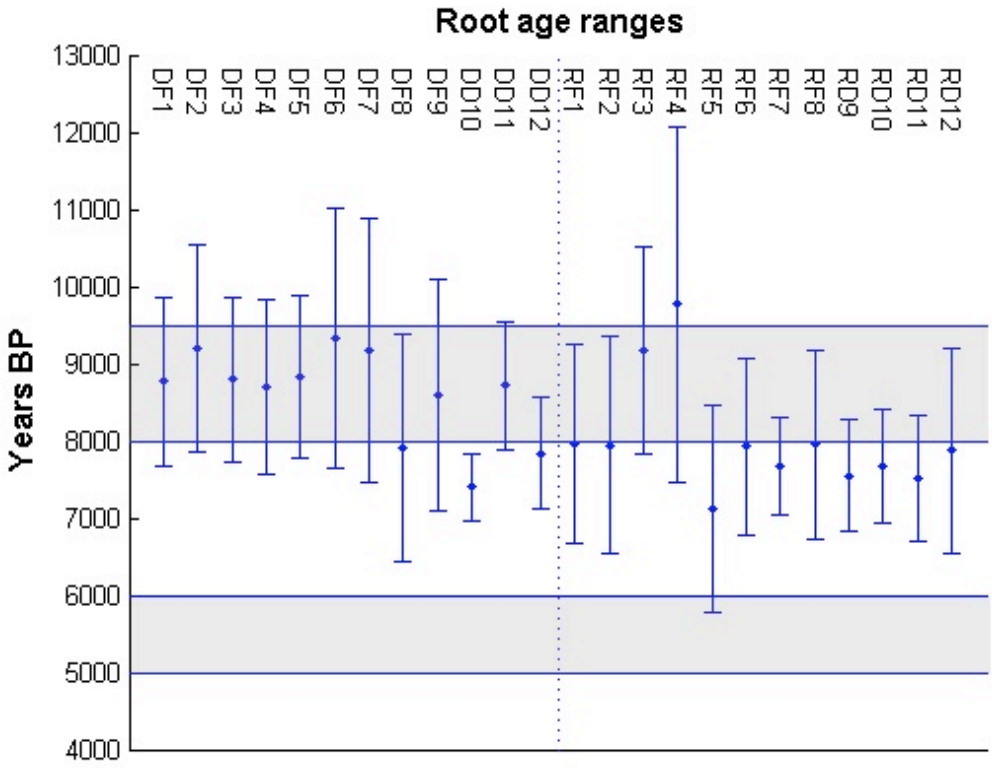


FIGURE 5.1: Divergence time estimates for the Dyen *et al*. (1997) dataset (to the left of the dotted line, with labels beginning with 'D') and Ringe *et al*. (2002) dataset (to the right of the dotted line, with labels beginning with 'R') using the time-reversible (labelled 'DF' and 'RF') and stochastic-Dollo (labelled 'DD' and 'RD') models. Each analysis is summarized with a plotted mean and error bars representing a 95% confidence interval. The horizontal bands indicate the age range implied under the two competing theories of Indo-European origin – the Kurgan hypothesis (5,000-6,000BP) and the Anatolian hypothesis (8,000-9,500BP).

TABLE 5.3: Summary of analyses from Figure 5.1, including the mean and standard deviation for the age at the root of Indo-European.

#	Data	Model	μ	S.D	Comments
DF1	Dyen *et al.* (1997)	Time-rev	8764	544	All cognacy information, uniform branch length priors, Gamma distributed rates across sites, PL rate smoothing.
DF2	Dyen *et al.* (1997)	Time-rev	9201	670	As for DF1 with conservative cognacy judgements only.
DF3	Dyen *et al.* (1997)	Time-rev	8794	533	As for DF1 with loose constraints.
DF4	Dyen *et al.* (1997)	Time-rev	8699	561	As for DF1 with missing data coded as '?'
DF5	Dyen *et al.* (1997)	Time-rev	8829	520	As for DF3 with strict clock assumption.
DF6	Dyen *et al.* (1997)	Time-rev	9336	843	As for DF2 with exponential branch length priors.
DF7	Dyen *et al.* (1997)	Time-rev	9176	855	As for DF2 with Swadesh 100 word list terms only.
DF8	Dyen *et al.* (1997)	Time-rev	7900	735	As for DF2 with selection of 20 languages.
DF9	Dyen *et al.* (1997)	Time-rev	8590	748	As for DF8 with selection of a different 20 languages.
DD10	Dyen *et al.* (1997)	Dollo	7400	218	Conservative cognacy judgements, uniform MRCA-time prior.
DD11	Dyen *et al.* (1997)	Dollo	8714	410	As for DD10 with Swadesh 100 word list terms only.
DD12	Dyen *et al.* (1997)	Dollo	7834	360	As for DD10 with selection of 31 languages.
RF1	Ringe *et al.* (2002)	Time-rev	7964	645	Uniform branch length priors, gamma distributed rates across sites, PL rate smoothing
RF2	Ringe *et al.* (2002)	Time-rev	7942	702	As for RF1 with exponential branch length priors.
RF3	Ringe *et al.* (2002)	Time-rev	9166	671	As for RF1 with fine grained cognacy judgements for some characters.
RF4	Ringe *et al.* (2002)	Time-rev	9770	1150	As for RF1 with Swadesh 200 word list items only.
RF5	Ringe *et al.* (2002)	Time-rev	7106	671	As for RF1 with Swadesh 100 word list items only.
RF6	Ringe *et al.* (2002)	Time-rev	7932	572	As for RF1 with topology constraint in accordance with Ringe *et al.* (2002).
RF7	Ringe *et al.* (2002)	Time-rev	7665	313	As for RF1 with equal rates across sites.
RF8	Ringe *et al.* (2002)	Time-rev	7956	612	As for RF1 with strict clock.
RD9	Ringe *et al.* (2002)	Dollo	7552	366	Uniform MRCA-time prior, 15 languages with over 10% sampling.
RD10	Ringe *et al.* (2002)	Dollo	7671	368	As for RD9 with fine grained cognacy judgements for some characters.
RD11	Ringe *et al.* (2002)	Dollo	7513	406	As for RD9 with Swadesh 200 word list items only.
RD12	Ringe *et al.* (2002)	Dollo	7869	659	As for RD9 with Swadesh 100 word list items only, 17 languages with over 10% sampling.

In Chapters 3 and 4 divergence time estimates for the root of the Indo-European tree were found to be robust to a wide range of plausible rooting points, Bayesian priors,

cognacy judgement criteria, age constraints and the effect of missing information in the data. Key results from these analyses are summarized in Figure 5.1 (DF1-6). These results are consistent with a number of subsequent analyses, using subsets of 20 languages (DF8&9) and even when the data is limited to the highly conserved and borrowing-resistant Swadesh 100 word list (DF7). Using the stochastic-Dollo model, we found evidence for similar ages using the Swadesh 100 word list items (DD11) and slightly younger divergence times using the whole data set (DD10) and a subset of 31 languages (DD12). These ages inferred under the stochastic-Dollo model are thus also broadly consistent with the results from Chapters 3 and 4.

Results from the analysis of the Ringe *et al*. (2002) data (RF1) are consistent with the Dyen *et al*. (1997) data. Varying branch-length priors (RF2), rates across sites (RF7), and rates through time (RF8), and the cognacy judgement criteria (RF3, RD10) had little affect on divergence times estimated from the Ringe *et al*. data. Analysing progressively more refined data sets (Swadesh 200 word list items only – RF4, RD11; and Swadesh 100 word list items only – RF5, RD12), also had little affect on the mean age estimates, although, predictably, as the amount of information decreased, variance in the estimated dates increased. The analysis RF5, of Swadesh 100 list terms, appears to go against this trend, showing less variation than the analysis of Swadesh 200 list terms in RF4. This is misleading, however, as the weak signal in the data meant that most of the trees produced in this analysis had to be filtered out because they were inconsistent with the clade constraints for known Indo-European language groups. As a result, variation was artificially decreased in this analysis, a case where Method 1 fails to produce reliable error bars. In analyses of larger data sets almost all of the trees in the Bayesian sample were consistent with recognized Indo-European groups. This anomaly did not occur using the Trait Lab analysis of the Swadesh 100 dataset because the program allows topology to be constrained prior to the analysis. Finally, in RF6, the topology of the Indo-European tree was constrained to that obtained by Ringe *et al*.'s (2002) own analysis of lexical, phonological and morphological characters. Again, the estimated age at the root of the tree is unaffected. This is a consequence of the fact that data can be informative of total tree length even when uninformative of topology

5.6 CONTROLLED MIRACLES - SYNTHETIC DATA VALIDATION

No statistical model of language change will capture all aspects of a process as complicated as the evolution of a language lexicon. In this sense all models are lies. However, as was argued in Chapter 4, models can be used as "lies that lead us towards the truth". In other words, they can allow us to answer questions of interest with sufficient precision and accuracy so as to provide a meaningful result. The crucial caveat is that we must test the performance of our model for a given task. To determine the reliability and accuracy of our results, we need to be able to quantify the systematic uncertainty in parameter estimates.

One way to do this is to generate synthetic data on a known phylogeny under plausible alternative models of evolution, including models which violate the assumptions of our inference model. Validating a methodology on synthetic data is useful for a number of reasons. First, we know the "true" phylogeny - the topology and branch lengths on which the data was generated. This means we can test the ability of our method to reconstruct a phylogeny from a given set of data accurately. Second, it is generally much easier to generate data on a phylogeny than it is to estimate the phylogeny that has produced a given set of data. As a result, we can test the performance of our methods on data generated under more sophisticated models. Third, we know the "true" data, (i.e. the real data). By comparing the real data with synthetic data generated under different models, we may be able to identify certain assumptions of our model that are especially important.

Figure 5.2 shows the mean root age and 95% confidence interval for a series of synthetic data analyses carried out using TraitLab. Results to the left of the vertical line are for the time-reversible model of method 1, whilst those to the right are for method 2, fitting the stochastic-Dollo model. These results and the evolutionary models used to generate each set of data are summarized in Table 5.4.

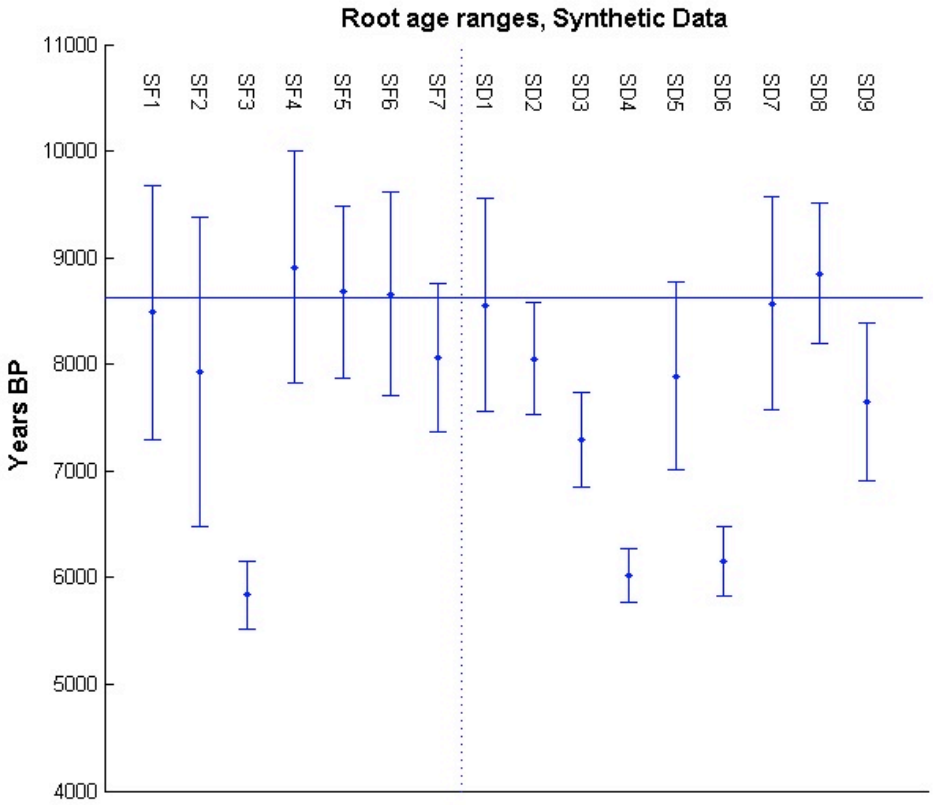


FIGURE 5.2: Mean root age and 95% confidence interval for a series of synthetic data analyses carried out using TraitLab. Data was generated under a number of models of evolution on a tree with a root age of 8680 years BP (indicated by the horizontal line) chosen from the posterior distribution in RD16. Results to the left of the vertical line are for method 1, fitting the time-reversible model, whilst those to the right are for method 2, fitting the stochastic-Dollo model using TraitLab.

TABLE 5.4: Summary of results shown in Figure 5.2 for synthetic data analyses, including the mean and standard deviation for the estimated age at the root of the tree on which data was synthesized. Data was synthesized on a tree chosen from the posterior distribution in RD16. The true age was 8680 units.

Analysis	μ	S.D.	Synthetic data model
SF1	8488	595	Time-Reversible
SF2	7925	725	Stochastic-Dollo with 20% global borrowing
SF3	5838	160	Stochastic-Dollo with 100% global borrowing
SF4	8911	545	Stochastic-Dollo with 20% local borrowing, 4000 years
SF5	8677	404	Dependent model
SF6	8660	479	Dependent model with 20% local borrowing, 1000 years
SF7	8061	350	Dependent model with 20% local borrowing, 1500 years
SD1	8558	500	Stochastic-Dollo model
SD2	8054	262	Stochastic-Dollo Model with 10% global borrowing
SD3	7291	220	Stochastic-Dollo Model with 20% global borrowing
SD4	6021	123	Stochastic-Dollo Model with 100% global borrowing
SD5	7891	443	Stochastic-Dollo Model with 100% local borrowing, 500 years
SD6	6157	161	Stochastic-Dollo Model with 20% local borrowing 4000yrs
SD7	8567	499	Dependent model
SD8	8853	331	Dependent model with 20% local borrowing, 1000 years
SD9	7642	370	Dependent model with 20% local borrowing, 1500 years

Both the methods were able to reconstruct the age of data synthesized under their respective models (SF1, SD1). More interestingly, TraitLab allows data to be simulated under a number of other evolutionary scenarios. Data was synthesized on a number of different trees; however, the results presented here all use the same tree, chosen from the posterior distribution in RD16. The true age was 8680 years.

First, we investigated the effect of borrowing on divergence time estimates. We generated a series of synthetic data sets using models of evolution that allow for horizontal as well as vertical transmission of cognates between random pairs of languages. As with the standard models, cognates evolve through time along each lineage according to the stochastic-Dollo model of word birth/death. In addition, however, for any given time interval, cognates can be borrowed from one lineage to another randomly selected existing lineage with a certain probability. By varying this probability we can simulate the effect of different rates of borrowing. Even relatively high rates of borrowing, at 20% of the cognate death rate (SF2 and SD3), had only a small effect on divergence time estimates, causing a slight underestimation of ages in both methods of analysis. As the rate of borrowing increased the extent of underestimation also increased. For a borrowing rate of 100%, date estimates were reduced by 30% under method 1 and 2 (SF3 and SD4).

Second, it may be unrealistic to assume that words can be borrowed from any language to any other language with equal probability. For example, Western-European languages may be far more likely to borrow from other Western-European languages than from Iranian languages. For this reason, we also synthesized data with borrowing limited to local areas of the tree. This was achieved by restricting the borrowing process to only those languages that had diverged within a certain threshold cut-off time. If, for example, a 4,000 year threshold was used, at any given time interval on the tree, lineages that had been separated for over 4,000 years could not borrow words between them. This has the effect of eliminating borrowing between the most distantly related languages, such as between the Germanic and Iranian languages. Using a 4,000 year threshold and 20% borrowing rate caused method 2 (SD6), but not method 1 (SF4), to underestimate divergence times. More localized borrowing tended to result in more accurate date estimates. Interestingly, using a much lower borrowing threshold, such as 500 years, and relatively high rates

of borrowing allows us to approximate the effects of dialect chain divergence, where languages remain in contact for a period after beginning to separate. Again, this had little effect on divergence time estimates (SD5).

Finally, data were generated under a dependent model of cognate evolution to test the effect of violations of the independence assumption. The dependent model of evolution models multiple subsets of cognates representing meaning categories. Under this model, cognates can evolve independently between meaning categories but are subject to constraints within each meaning category. Each language must always have at least one cognate in each meaning category. In addition, by varying parameters controlling the expected vocabulary size relative to the number of meaning categories, we can alter the expected number of words within each meaning category in each language. The resulting model is dependent in that each language must have at least one word for each meaning category. The converse feature of the model proposed by Evans *et. al* (in press) is not explicitly modeled: if a word is present, other words of the same meaning are born and die independently. However, the probability distribution for the number of words in each category does decline from a maximum at one word for the parameters we chose when we generate synthetic data. This dependent model may more accurately reflect the true process of language evolution. Both methods perform very well at recovering the true age (SF5, SD7). This remains the case even if we also introduce borrowing (SF6, SF7, SD8, SD9).

5.7 DISCUSSION

The divergence time estimates derived from the two independent lexical data sets using two very different models of word evolution are strikingly consistent. The findings of Gray and Atkinson (2003), reported here in Chapters 3 and 4, seem to be robust to the choice of languages sampled, to the meaning categories analysed, to who makes vocabulary assignments and cognacy judgements, and even to the age of the languages sampled. The fact that ancient and modern data sets have produced similar date estimates strongly supports the notion that the process of language evolution itself is sufficiently constrained and robust to human socio-cultural change as to make date estimates based on lexical comparison a feasible possibility. This is also evidence

against Garret's (in press) suggestion that contemporary borrowing could bias the age estimates.

Of course, one problem could be with the methodology. It is therefore impressive that two different methods using very different models give the same result. We were able to address the principal concerns that have been raised about model misspecification using synthetic data. The models were found to be robust to key criticisms of borrowing and independence. Significantly, if the age implied by the Kurgan hypothesis were the true age of Indo-European, model misspecification would have to cause us to over-estimate the age of the common ancestor by a factor of one and one half in order for us to find support for the 8,000BP to 9,500BP age range implied by the Anatolian theory. None of the types of model misspecification tested here produced appreciable overestimation. Analyses of data synthesized under models incorporating various degrees of global and local borrowing produced progressively greater underestimation of age estimates as the degree and extent of the borrowing was increased. Interestingly, whilst the time-reversible model performed relatively well in reconstructing the root age from data synthesized under a stochastic-Dollo model with borrowing (e.g. SF2 & SF4), it could not reconstruct reasonable branch-lengths from data synthesized under the strict Dollo model, with no borrowing. The time reversible nature of the model appears to allow it to effectively accommodate for homoplasy due to borrowing, however, when applied to homoplasy-free synthetic data the model has difficulty. When data was generated under a "dependent" model of evolution both models were able to estimate the root age relatively accurately. This suggests that whilst these models assume that the evolution of characters is independent, they are in fact robust to violations of this assumption. Certainly, there is no evidence to suggest that violations of the independence assumption might cause us to grossly over-estimate ages.

A key aspect of our approach is that we can make date estimates without having to postulate just one phylogeny. Instead, inferences about divergence times are made on the basis of the Bayesian sample distribution of trees, allowing us to quantify the phylogenetic uncertainty implicit in our date estimates. We can, however, use a consensus tree or consensus network (Holland and Moulton, 2003) as a visualization tool.

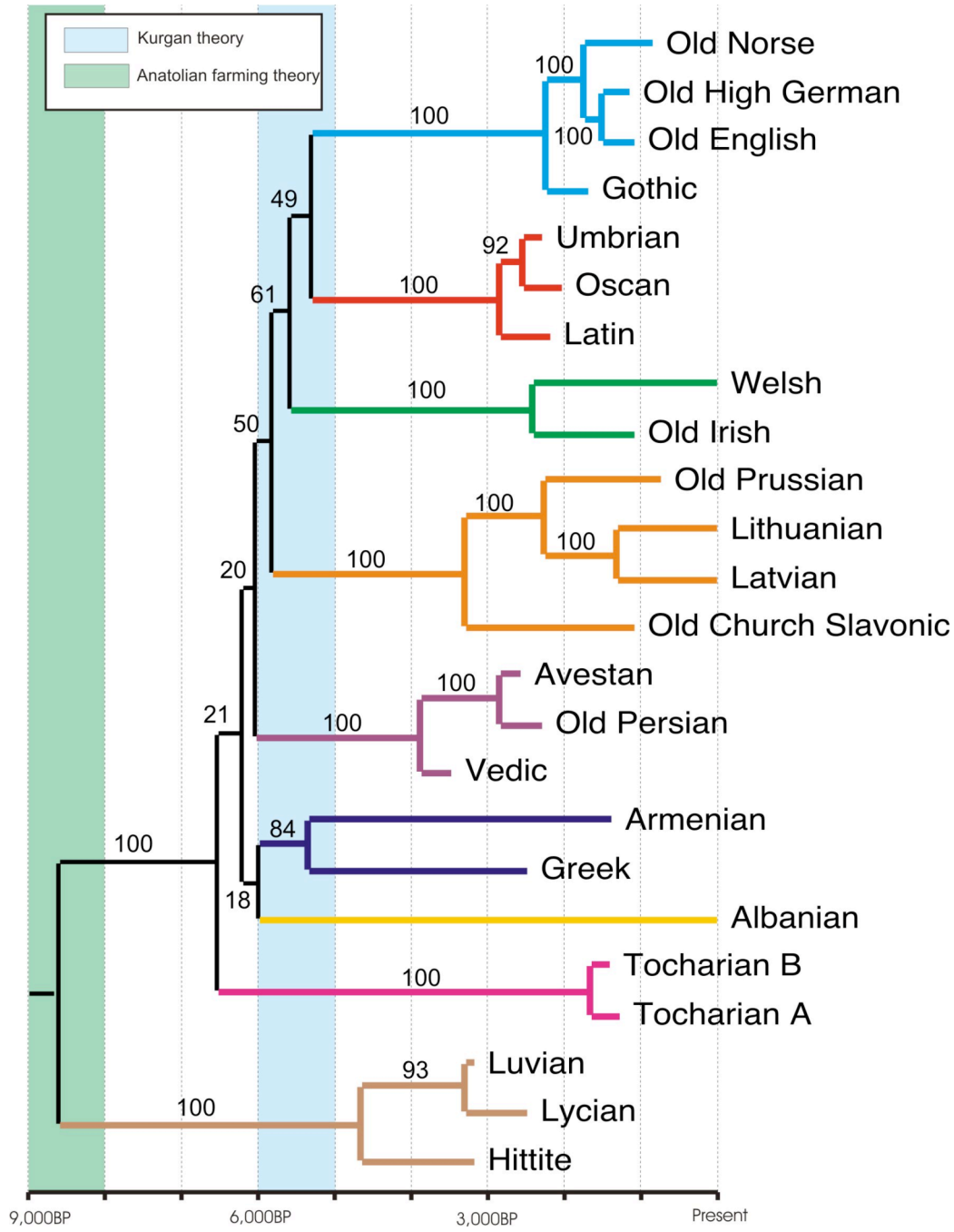


FIGURE 5.3: Majority-rule consensus tree from the initial Bayesian MCMC sample of 1,000 trees based on the Ringe *et al.*(2002) data. Values above each branch indicate uncertainty (posterior probability) in the tree as a percentage. Branch-lengths are proportional to time. Branches are coloured to distinguish recognized sub-groups. Shaded bars represent the age range proposed by the two main theories – the Anatolian theory (green bar) and the Kurgan theory (blue bar). The basal age (8,680BP) supports the Anatolian theory. While consensus trees are a useful visual aid, they cannot display the strength of evidence for conflicting clades. The consensus network in Figure 5.4 is a better representation of the results, although perhaps more difficult to interpret.

Figure 5.3 shows a consensus tree constructed using the distribution of trees from the standard RF1 analysis of the Ringe *et al.* (2002) data. Branch-lengths are proportional to time. As well as tree topology and branch length information, the consensus tree shows the degree of support for each sub-clade within the phylogeny, expressed in the

form of the "posterior probability" of a clade — the percentage of time that the clade appears in the Bayesian MCMC sample distribution. A value of 100, for example, indicates that the clade occurs in all sampled trees. Lower values indicate an increasing degree of statistical uncertainty. It is evident from Figure 5.3 that the two theories of Indo-European origin may not, in fact, be mutually exclusive – a possibility identified by Cavalli-Sforza, Menozzi and Piazza (1994). Whilst the basal age (8,680BP) supports the Anatolian theory of Indo-European origin, there is a period of rapid divergence during the hypothesized time of the Kurgan expansion, between 5,000BP and 6,000BP. We saw a similar pattern in the analysis of the Dyen *et al.* (1997) data in Chapters 3 and 4.

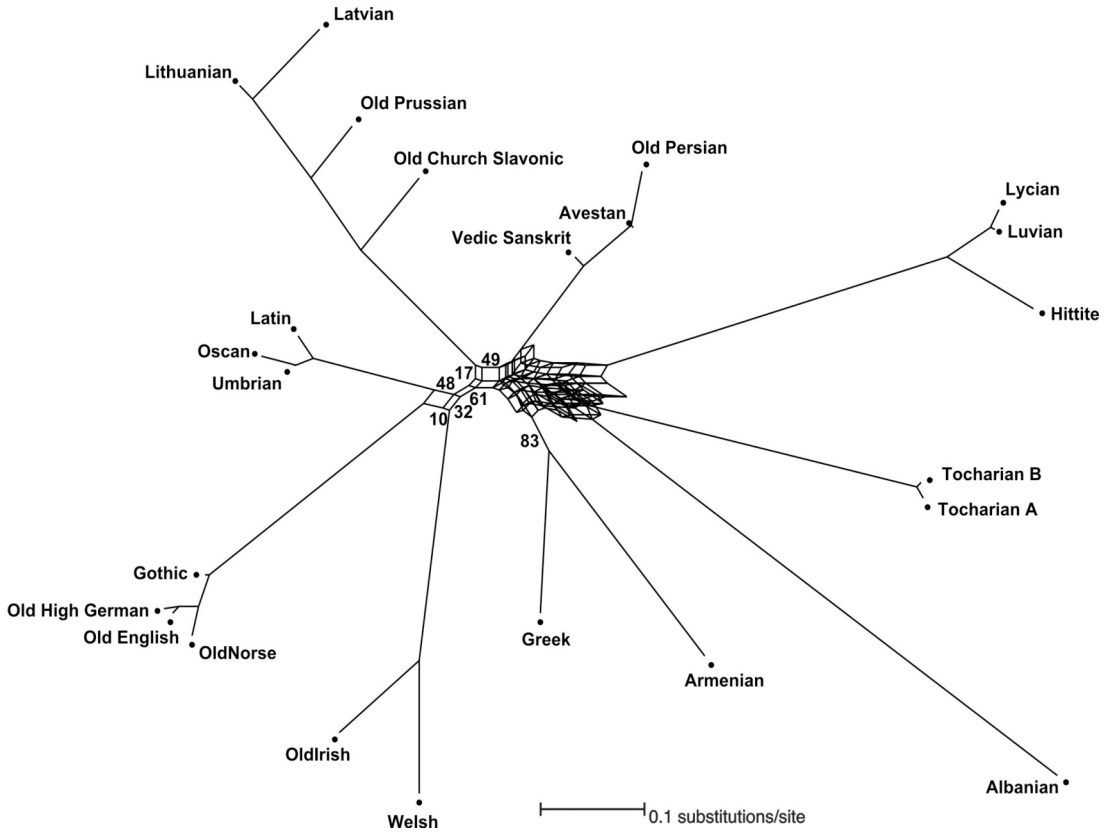


FIGURE 5.4: Consensus network from the initial Bayesian MCMC sample of 1,000 trees based on the Ringe *et al.*(2002) data, constructed using SplitsTree (Huson, 1998). Values express percentage support for some of the splits. A threshold of 10% was used to draw this splits graph – i.e. only those splits occurring in at least 10% of the observed trees are shown in the graph. Branch lengths represent the median number of reconstructed substitutions per site across the sample distribution.

One problem with using consensus trees, is that they cannot display the strength of evidence for conflicting clades. For example, we may be able to show that the Germanic languages group with the Celtic languages 49% of the time, but we cannot

simultaneously show that 32% of the time the Germanic languages group with the Italic languages, even though this may be very interesting. One way of summarizing a distribution of trees without losing this information is to display conflicting clades or 'splits' simultaneously. We can do this using consensus networks (Holland and Moulton, 2003). Figure 5.4 shows the RF1 distribution of trees summarized as a consensus network displaying all those clades with greater than 10% support. Each edge or 'split' separating one set of languages from another corresponds to a clade. This clearly shows the lack of resolution at the base of the tree – the box like structures in the centre of the figure indicate incompatible clades in the sample distribution of trees. This picture is consistent with acknowledged uncertainties, such as the position of Albanian and the relationship between Italic and Celtic. It is this uncertainty in the branching structure that we can integrate out by estimating divergence times across the sample distribution of trees.

5.8 CONCLUSION

The well-known criticisms of glottochronology have led many researchers to reject the possibility of estimating dates from lexical data. Any approach that attempts to turn words into dates is dismissed as attempting the impossible - trying to turn water into wine. Here we have shown that estimating divergence time confidence intervals from lexical data is far from impossible or miraculous. New statistical tools from evolutionary biology enable us to estimate phylogeny and divergence times without falling victim to the pitfalls of glottochronology. The availability of these methods means that it is no longer valid to dismiss all attempts at divergence date estimates simply because Swadesh's approach was problematic. If used sensibly these new methods offer a powerful set of tools with the potential to resolve some of the long-standing debates in historical linguistics.

5.9 REFERENCES

Atkinson, Q. D. and R. D. Gray. 2006. Are accurate dates an intractable problem for historical linguistics? Pages 269-296 In *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory.* (eds.) C. Lipo, M. O'Brien, S. Shennan & M. Collard. Aldine, Chicago.

Atkinson, Q. D. and Gray, R. D. in press. How old is the Indo-European language family? Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge.

Bryant, D., F. Filimon, and R. D. Gray. 2005. Untangling our past: Pacific settlement, phylogenetic trees and Austronesian languages. Pages 69–85 in *The evolution of cultural diversity: Phylogenetic approaches* (eds.) R. Mace, C. Holden, and S. Shennan. UCL Press, London.

Bryant, D., and V. Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. *Workshop in Algorithms for Bioinformatics, Proceedings* 2002:375–391.

Campbell, L. 2004. *Historical linguistics: An introduction.* 2nd edition. Edinburgh University Press, Edinburgh.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press, Princeton, N.J.

Dyen, I., J. B. Kruskal, and P. Black. 1992. An Indoeuropean Classification: A Lexicostatistical Experiment. *American Philosophical Society, Transactions* 82(5). Philadelphia.

Dyen, I., J. B. Kruskal, and P. Black. 1997. FILE IE-DATA1. Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.

Evans, S. N., Ringe, D. and Warnow, T. in press. Inference of divergence times as a statistical inverse problem. In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge.

Farris, J. S. 1977. Phylogentic analysis under Dollo's Law. *Systematic Zoology*, 26:77-88.

Gamkrelidze, T. V., and V. V. Ivanov. 1995. *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture.* Mouton de Gruyter, Berlin.

Garrett, A. in press. Convergence in the formation of Indo-European subgroups: Phylogeny and chronology. In *Phylogenetic methods and the prehistory of languages* (eds.) J. Clackson, P. Forster and C. Renfrew. MacDonald Institute for Archaeological Research, Cambridge.

Gimbutas, M. 1973a. Old Europe c. 7000–3500 B.C., the earliest European cultures before the infiltration of the Indo-European peoples. *Journal of Indo-European Studies* 1, 1-20.

Gimbutas, M. 1973b. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500–2500 B.C. *Journal of Indo-European Studies* 1, 163–214.

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426, 435-9.

Holland, B., and V. Moulton. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. Pages 165–176 in *Algorithms in bioinformatics, WABI 2003.* (eds.) G. Benson and R. Page. Springer-Verlag, Berlin, Germany.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian Inference of Phylogeny. *Bioinformatics* 17:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294:2310–2314.

Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* 1.4:68–73.

Kuhner, M. K., and J. Felsenstein. 1994. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution* 11:459–468.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21:1087–1091.

Nicholls, G., and R. Gray. in press. Quantifying uncertainty in a stochastic Dollo model of vocabulary evolution. In *Phylogenetic methods and the prehistory of languages.* (eds.) J. Clackson, P. Forster and C. Renfrew. McDonald Institute for Archaeological Research, Cambridge.

Renfrew, C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins.* London: Cape.

Rexová, K., D. Frynta, and J. Zrzavy. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19, 120–127.

Ringe, D., T. Warnow, and A. Taylor. 2002. Indo-European and Computational Cladistics. *Philological Society, Transactions* 100:59–129.

Sanderson, M. 2002a. Estimating absolute rates of evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* 19, 101–109.

Sanderson, M. 2002b. R8s, Analysis of Rates of Evolution, version 1.50. http://ginger.ucdavis.edu/r8s/

Steel, M., M. Hendy, and D. Penny. 1988. Loss of information in genetic distances. *Nature* 333, 494-495.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *American Philosophical Society, Proceedings* 96:453–463.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. International *Journal of American Linguistics* 21:121–137.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic Inference. Pages 407-514 in *Molecular Systematics*, second ed., (eds.) D. M. Hillis, C. Moritz, and B. K. Marble. Sinauer, Sunderland (MA).

Chapter Six

MAYAN LANGUAGE ORIGINS AND DIVERSIFICATION EXAMINED THROUGH PHYLOGENETIC ANALYSIS OF LEXICAL DATA

ABSTRACT

Linguistic evidence plays an important role in our understanding of Mayan prehistory. However, traditional linguistic methods have a number of limitations. Language family trees based on the "comparative method" yield a relative chronology of lineage divergence events, but cannot quantify uncertainty or provide absolute divergence dates. An alternative method, glottochronology, attempts to provide absolute dates but these estimates are generally thought to be unreliable, due largely to the assumption of a universal and constant rate of lexical replacement. The paucity of alternative historical sources has meant that, despite these problems, theories of Mayan origin and diversity have utilized glottochronology-based divergence time estimates. Here, we apply stochastic models of language evolution and computational phylogenetic methods to Mayan lexical data to infer the relationships within the Mayan language family and to estimate confidence intervals for lineage divergence times without the assumption of rate constancy. Our analyses highlight a number of areas of uncertainty in Mayan language relationships. Divergence time estimates also showed uncertainty but suggest the family may be older than previously thought. We evaluate how these results relate to existing archaeological and linguistic evidence for the origin and diversification of the Mayan language family. We conclude with suggestions for future research including ways to improve the precision of date estimates.

6.1 INTRODUCTION

The challenge of reconstructing human prehistory is a demanding one, requiring the triangulation of archaeological, genetic and linguistic evidence (Kirch and Green, 2001; Cavalli-Sforza, Menozzi and Piazza, 1994). By correlating evidence for demographic changes in the archaeological record with specific linguistic or human genetic groupings we can trace complex population histories through time. Languages, which evolve much faster than genes, are particularly important for elucidating Neolithic population migrations, which fall beyond recorded history but, due to slow rates of genetic mutation, may be too recent for genetic studies to provide conclusive answers. The importance of an interdisciplinary approach to historical inference is exemplified in debates about the history of the Maya of southern Mexico, Guatemala, Belize, and Honduras. The ruins and hieroglyphic inscriptions of the ancient Maya, dating mainly from the "Classic period" between 250AD and 900AD, bear witness to a complex agriculturalist society with a well-attested writing system and accurate celestial calendar. However, despite the help of dated hieroglyphic inscriptions, there remains controversy over when, how and from where the Mayan language groups came to occupy their current distribution. Much of our current understanding of Mayan origins is based on linguistic evidence.

Historical linguistics contributes to our understanding of Mayan prehistory via three lines of evidence. The standard approach to historical linguistics is the "comparative method", in which historical relationships are determined on the basis of regular patterns of sound change inferred by comparing corresponding sounds in *cognate* words across languages and by matching grammatical elements across languages. Cognates are words of similar meaning that can be shown to be related due to common ancestry - e.g. English *water* and German *wasser*. Although our understanding of Mayan language relationships based on the comparative method is improving, there are still areas of uncertainty (Campbell and Kaufman, 1985). The division of the family into six or seven principal subgroups (see Figure 6.1a) is generally accepted, but the deeper relationships among these groups, which are especially interesting for inferences about settlement patterns and Mayan origins, remain controversial (Campbell, 1984; Campbell and Kaufman, 1985). For example,

Huastecan, spoken in the Gulf Coast Veracruz region, and Yucatecan, spoken on the lowlands of the Yucatan peninsular, are sometimes claimed to have diverged first, one after the other, but there is no consensus on this (*cf.* Kaufman, 1976; Campbell, 1977; Fox, 1978; Campbell and Kaufman, 1985). Even the relationship between the hieroglyphic Classic Maya language/s and contemporary languages, which has important implications for the decipherment of the hieroglyphs, is unresolved.

The timing of divergence events is of crucial importance for correlating linguistic and archaeological evidence. Whilst language relationships inferred using the comparative method imply a relative chronology of divergence events, the approach cannot provide absolute date estimates. A second line of linguistic evidence, which does attempt to provide absolute dates is glottochronology (Swadesh, 1952, 1955). Glottochronology uses the percentage of cognates shared between pairs of languages to calculate divergence times under the assumption of a constant rate of lexical replacement, or "glottoclock". Usually, comparisons are restricted to the "Swadesh word-list" of 100 or 200 items of basic vocabulary thought to be culturally universal and relatively resistant to borrowing. Based on Swadesh's (1952,1955), original findings a universal retention rate (the percentage of cognates retained in a lexicon after 1,000 years evolution) of 81% is generally used for the 200 word list. Initial glottochronology-based estimates put the age of Proto-Mayan, the hypothesized common ancestor of all known Mayan languages, at 3,600BP (Swadesh, 1967). This estimate was subsequently revised to 4,200BP by Kaufman (1976) using more extensive and more complete lexical data, which were more thoroughly analyzed. There are, however, a number of widely recognized problems with glottochronology. First, the conversion of lexical character data to distance scores between languages results in a loss of information, reducing the power of the method to reconstruct evolutionary history accurately (Steel, Hendy and Penny, 1988). Second, the clustering methods employed produce inaccurate trees, grouping together languages that evolve slowly rather than languages that share a recent common ancestor (Blust, 2000). Third, the assumption of constant rates of lexical replacement through time and across all meaning categories does not hold for linguistic data, making date estimates unreliable (Bergsland and Vogt, 1962). Previous studies of Mayan languages stretched this assumption of rate constancy even further by using a rate

derived principally from the Indo-European language family (Swadesh, 1960; 1967; Kaufman; 1976) even though there is no reason to assume that rates for Mayan are the same as for the much larger and more widespread Indo-European family[1]. Fourth, it is argued that language contact and borrowing of lexical items between languages makes purely tree-based methods inappropriate (Hjelmslev, 1958; Bateman *et al.*, 1990). Finally, glottochronology does not provide quantification of the uncertainty associated with the inferred relationships or divergence times. Whilst attempts have been made to provide a means of quantifying uncertainty by allowing the proposed 81% universal retention rate to vary by +/- 2% (Swadesh, 1952), this calculation is rarely used and, regardless, does not allow for phylogenetic uncertainty in the reconstructed tree topology and branch lengths. Despite these problems, glottochronology has been used as the primary source of Mayan language chronology (Kaufman, 1974, 1976). The 4,200BP age of the family broadly agrees with the time-scale of agricultural expansion in Mesoamerica and is used by some to argue that the Mayan language family spread with the spread of intensive maize cultivation at this time (Josserand, 1975; Kaufman, 1976; Diamond and Bellwood, 2003; Bellwood, 2005).

A third line of evidence is *linguistic palaeontology,* a technique that seeks to draw inferences about the ancestral culture of a language group by comparing terms for different cultural items across the languages of the family. Common terms across a family imply that the items the terms refer to were present in the proto-language. Proto-Mayan terms associated with maize agriculture, as well as terms for other cultivars such as beans and squash, imply that Proto-Mayans had agriculture (Kaufman, 1976; Campbell and Kaufman, 1985). Linguistic palaeontology has also been used to argue for a highland Mayan homeland (Kaufman, 1976). Reconstructed Proto-Mayan has terms for both exclusively highland and lowland plants and animals. As highlanders today are aware of lowland species but lowlanders tend to be ignorant of the highland zones, it is argued that the Proto-Mayans must have originated from the Guatemalan highlands. However, recent evidence suggests that Mesoamerican

[1] Swadesh compared ancient and modern forms of thirteen languages to derive his universal retention rate. Only two of these thirteen languages (Coptic and Mandarin) were non-Indo-European. Later work on Kannada, Japanese, Arabic, Georgian, Armenian and Sardinian called into question Swadesh's claim of constant rates (Bergsland and Vogt, 1962; Campbell, 2004).

maize crops were derived from a lowland, not a highland, wild species (Piperno and Pearsall, 1993; Bennetzen *et al*., 2001) suggesting a lowland origin may be more likely. One possibility is that Proto-Mayan originated in the lowlands, perhaps in the Gulf Coast Veracruz region, and spread eastward to their present location (Willey, 1982).

Recently, studies have applied phylogenetic methods from biology to linguistic data to test hypotheses about the history of Indo-European (Rexova *et al*., 2003; Gray and Atkinson, 2003; Atkinson *et al.*, 2005), Bantu (Holden, 2002, 2003), Austronesian (Gray and Jordan, 2000), and Papuan (Dunn *et al*., 2005) language families. These methods can complement conventional approaches to historical linguistics and overcome the problems associated with date estimation using glottochronology. First, parsimony and maximum-likelihood approaches analyse character state data, thus avoiding the problem of information loss associated with distance-based methods like lexicostatistics/glottochronology. Second, Bayesian inference of phylogeny and Markov Chain Monte Carlo (MCMC; Metropolis *et al.*, 1953) sampling algorithms can evaluate large amounts of data accurately and quickly and allow us to quantify phylogenetic uncertainty due to random error. This provides an objective measure of the level of support for specific subgroups. Third, biologists have developed sophisticated methods of estimating divergence times on a phylogeny that do not assume strictly clock-like rates of evolution. For example, Sanderson's (2002a) rate smoothing algorithm allows rates to vary across a tree within a set of known age constraints, whilst incorporating a smoothing factor that penalises excessive rate variation. Hence, unlike glottochronology, this approach estimates divergence times without the assumption of a glottoclock, and by using multiple calibration points rather than a constant *a priori* rate, accuracy can be improved despite the added uncertainty of allowing rates to vary (Sanderson, 2002a). In addition, by combining this approach with Bayesian inference of phylogeny, we can quantify the uncertainty in date estimates and calculate confidence intervals for divergence times. Finally, phylogenetic trees cannot represent non-tree-like signal in the data due to borrowing or the presence of dialect-chains. To investigate the extent to which a tree-like model of evolution is appropriate and to identify interesting patterns of non-tree-like signal in the data, we can use network visualization techniques that do not assume a tree-like

model, such as the phylogenetic network methods available in *SplitsTree4* (Huson and Bryant, 2006; Bryant, Filimon and Gray, 2005).

Speaking about Indo-European, Renfrew (1987) laments that "the conventional view has been built up in a system of logic where the historical linguists generally accept some of the premises offered by the prehistoric archaeologists, and the archaeologists accept those of the linguists" (pp.263-264). To avoid this potential circularity, we need to examine independent lines of evidence from both fields critically and seek interdisciplinary agreement. Here, we apply phylogenetic methods from biology to Mayan lexical data as an example of four ways in which these methods can contribute to the linguistic evidence. First, we use network-based methods to identify non-tree-like signal in the data and uncover evidence of borrowing and/or dialect chain divergence. Second, we use two different stochastic models of language evolution to produce a sample distribution of trees and quantify the strength of support for Mayan language family relationships, including deeper relationships and the position of Classic Mayan. Third, we quantify the relative support of a highland versus lowland Maya homeland by mapping geography onto the trees in our sample distribution and inferring the most probable ancestral state. And fourth, we re-evaluate the Mayan linguistic chronology supporting a spread with agriculture by calculating confidence intervals for lineage divergence times across the sample distribution of trees.

6.2 MATERIALS AND METHODS

We compiled a lexical database comprising roughly 1200 cognates for Swadesh 100-word-list terms in 30 extant and 5 archaic Mayan languages. Word form and cognacy judgements were made on the basis of multiple sources (Deinhart, 1989; Kaufman, 2003; Boot, 2002; Hernandez, 1929; de Coto, 1983) and checked to ensure the accuracy and reliability of the data. Hieroglyphic Classic Maya terms occurring only after 700AD were ignored to a) limit the date range assigned to the Classic Maya node, and b) minimize the influence of later Yucatecan terms in the hieroglyphs (see Section 6.4). Colonial sources were only used for Colonial Kaqchikel and Colonial Yucatec. The database is available online at http://language.psy.auckland.ac.nz/.

For analysis, data were coded in a binary matrix representing the presence (1) or absence (0) of cognate sets in each language, with missing data also coded (?). Ch'olti' was excluded from the analyses due to poor sampling (86% missing data). Modern Yucatec and Kaqchikel were excluded in favour of their colonial sources. To investigate evidence of reticulation in the data we used splits graphs produced in *SplitsTree4* (Huson and Bryant, 2006). Two related languages, Tojolabal and Chuj, showed evidence of extensive borrowing and were removed from subsequent phylogenetic analyses.

Tree topology and branch lengths were estimated from the data matrix under a likelihood framework using Bayesian inference of phylogeny and MCMC sampling algorithms. We analysed the data using two very different models of lexical evolution, outlined in detail in Atkinson *et al.* (2005; see Chapter 5). The first is a time-reversible model of binary character evolution, based on the restriction-site model from biology, as implemented in *MrBayes* (*version 3.1.1*; Huelsenbeck and Ronquist, 2001). This model allows for unequal character state frequencies and gamma distributed rate heterogeneity between cognates. We used *MrBayes*' default priors for the reported analyses, however, results were found to be robust to a range of priors (some of which are shown in Figure 6.4). Branch-lengths were calculated in proportion to the amount of inferred change. A likelihood ratio test for clock-like rates of evolution allowed us to reject a clock under the time-reversible model (χ^2=106.91, df=28, p<0.001). To estimate divergence times without assuming clock-like evolution, we applied *r8s* (*version 1.7*; Sanderson, 2002b) to the Bayesian sample distribution of trees. Unknown divergence times were estimated using rates of evolution calibrated on the basis of archaeological and ethno-historical evidence for language splits (see Table 6.1). The second model, implemented in TraitLab (downloadable from http://aitken.math.auckland.ac.nz/~nicholls/TraitLab/) and described in Chapter 5, assumes a stochastic-Dollo (Farris, 1977) process in which each cognate set evolves only once. As this software cannot yet accommodate missing data, Classic Maya (>50% missing data) was excluded from the stochastic-Dollo analyses. Using TraitLab, dates were estimated directly as branch lengths are calculated proportional to time, with rates calibrated using the same historically attested divergence points as for the time-reversible model. Both analyses were run

for 1.3 million generations sampling every 1,000 trees. A 'burn-in' period of 300,000 trees for each run was used to avoid sampling trees before the run had reached convergence. Log-likelihood plots and an examination of the post burn-in tree topologies demonstrated that the runs had indeed reached convergence by this time.

TABLE 6.1: Age constraints used to calibrate the Mayan divergence time calculations.

Constraint	Age Range	Evidence	Reference
Classic Maya	300A.D.-700A.D.	Earliest dated monument, Stela 29 at Tikal, 292A.D., although most stela are later. Items identified by Boot (2002) as being younger than 700A.D. were excluded.	Boot, 2002. A Preliminary Classic Maya-English/English-Classic Maya Vocabulary of Hieroglyphic Readings.
Colonial Yucatec	1570A.D. - 1590A.D.	Source document written around 1580.	Bocabulario de Maya Than de Viena. 1993. Bocabulario de Maya Than, Facsimile and edited version by René Acuña.
Colonial Kaqchikel	1630A.D. - 1647A.D.	Sources compiled in the years before publication.	De Coto (1647/1983) Thesasavrvs Verborv - Vocabvlario de la Lengua Cakchiquel v(el) Guatemalteca.
Sipakapense /Sakapulteko split	Pre 1551A.D.	Titulo of Sacapulas (1551) recognizes the two languages as being distinct	Campbell, L. 1978. Quichean Prehistory: Linguistic contribution. Carmack, R. M. 1973. Quichean Civilization.
Poqomam/Poqomchi split	1250A.D. - 1450A.D.	Ethno-historical evidence of Rabinal Quiche invasion "driving a wedge" between what is now Poqomam and Poqomchi.	Campbell, L. 1978. Quichean Prehistory: Linguistic contribution.
Kaqchikel/Tz'utujil	Pre 1493A.D.	Ethno-history indicates the populations had evidently been separated before this time.	Carmack, R. M. 1973. Quichean Civilization.

To infer the ancestral Maya homeland the geographic location of each language was recorded as either Chiapas Highland, Maya Lowland or, uniquely for Huastecan, the Gulf Coast Veracruz region. These characters were mapped onto the sample distribution of trees. We used the MCMC sample distribution of trees produced under the stochastic-Dollo model so that all trees were rooted independent of external linguistic evidence. We then used *MacClade* (*v. 4.06*; Maddison and Maddison, 1992) to infer geographic character states at ancestral nodes on the tree. Character state changes across the tree imply migration events. For each tree in the sample

distribution, we calculated the most parsimonious assignment of ancestral states – i.e. the reconstruction favouring the minimum number of migration events.

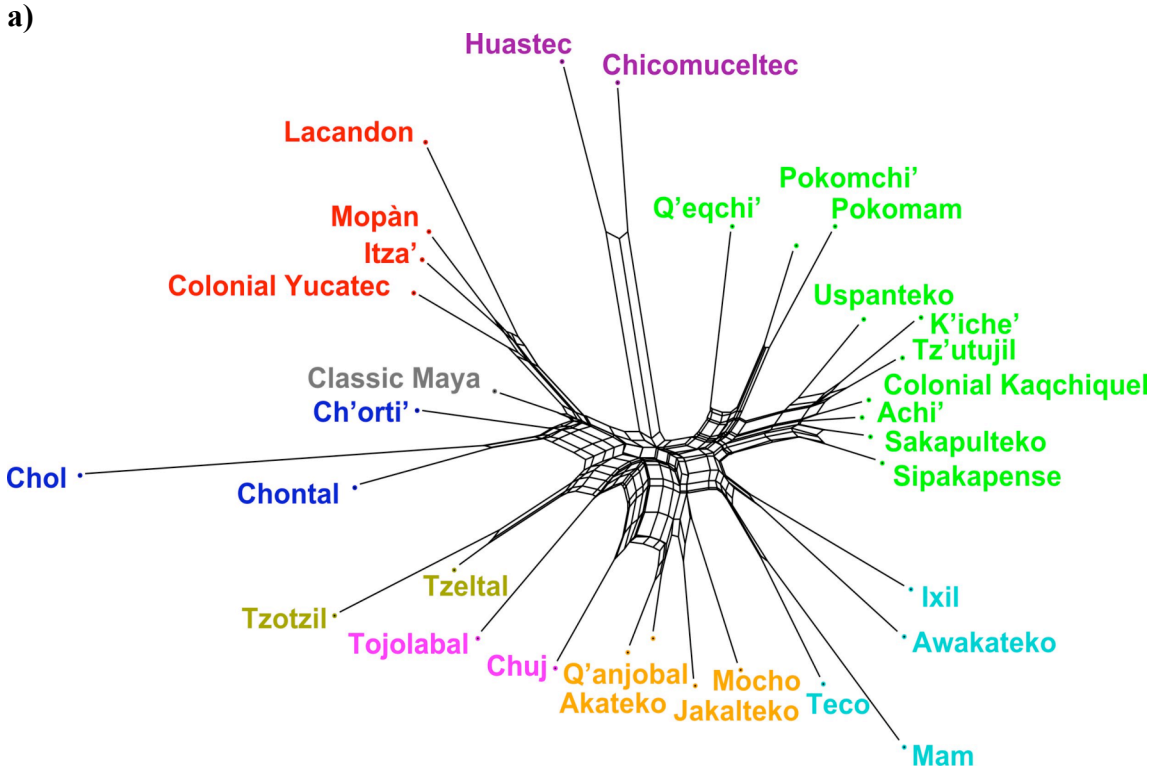
6.3 RESULTS

Figure 6.1a shows a phylogenetic network of the Mayan language data constructed using *SplitsTree4* (Huson and Bryant, 2006). Each set of parallel lines represents a split in the data separating one set of languages from all of the other languages. Splits are drawn proportional to the mean amount of inferred change. Incompatible language groupings implied by these splits, representing reticulate evolution, appear as box-like structures. All of the recognized subgroups (colour coded) cluster together, with the exception of Chuj and Tojolabal. The network shows Tojolabal grouping with Tzeltalan (olive) and Chuj grouping with Q'anjobalan (orange), but with an exceptionally high degree of reticulation, suggesting extensive borrowing or involvement in a dialect chain. Interestingly, the Tojolabal/Chuj grouping is the subject of some disagreement. Whilst some linguists put these languages together in a subgroup most closely related to the Q'anjobalan languages (e.g. Campbell, 1997; Kaufman, 1976), others prefer to place Tojolabal with Tzeltalan (Robertson, 1977). This disagreement is reflected in genuine conflicting signal in the lexical data. Removing Chuj and Tojolabal from the network (Figure 6.1b) produces a considerably more tree-like, although still somewhat reticulate, picture. These languages were thus removed from the tree-based phylogenetic analyses that follow.

A key advantage of using a Bayesian framework is that we can estimate phylogenetic uncertainty due to random error implicit in our model of evolution. Hence, rather than returning a single, optimal tree, the method returns a distribution of trees, sampled in proportion to their likelihood, given the data, the model and a set of prior assumptions. This uncertainty can be expressed as a consensus network (Holland and Moulton, 2003) of the sample distribution of trees. Figure 6.2 shows a consensus network generated under the binary, time-reversible model of cognate gains and losses. Again, all of the recognized Mayan subgroups are grouped together. The highland K'ichean (green) and Mamean (turquoise) languages are grouped together with 95% support, forming an "Eastern Mayan" grouping, consistent with reconstructions based on the comparative method (Campbell and Kaufman, 1985).

There is also evidence for an early split separating the lowland Yucatecan (red) and Cholan (blue) subgroups plus Tzeltalan (olive) from the highland Eastern Mayan and Q'anjobalan (orange) languages (69%) - there is less support for the highland Q'anjobalan languages grouping with the lowland languages and Tzeltalan (22%). Finally, there is uncertainty about the relationship between Classic Mayan, the Yucatecan and the Cholan languages.

a)



b)

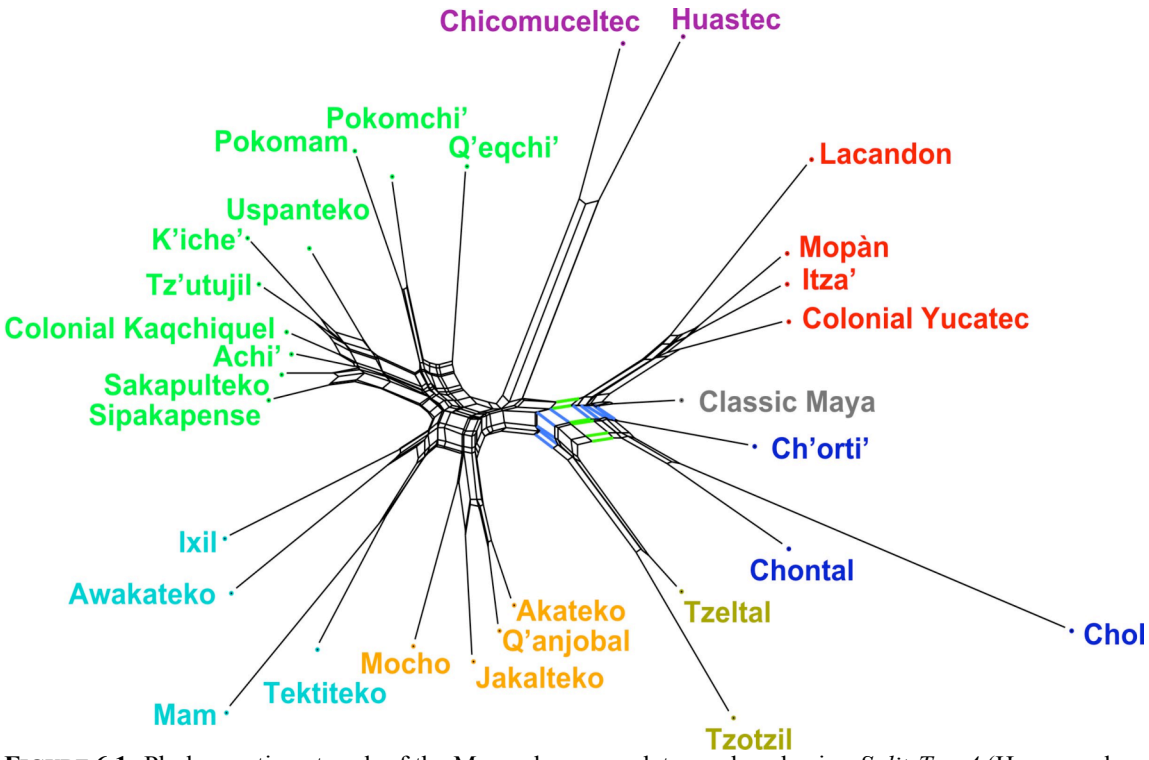


FIGURE 6.1: Phylogenetic network of the Mayan language data produced using *SplitsTree4* (Huson and Bryant, 2006) for: **a** complete 32-language data set; and **b** with Chuj and Tojolabal removed. Two interesting conflicting splits are highlighted in panel b. The green split represents characters supporting a Cholan-Yucatecan group, whilst the blue split represents characters supporting a Cholan-Tzeltalan group. Language names are colour coded according to recognized subgroups (Campbell, 1997). Green = K'ichean, purple = Huastecan, red = Yucatecan, blue = Cholan, olive = Tzeltalan, orange = Q'anjobalan, turquoise = Mamean, pink = Chujean.

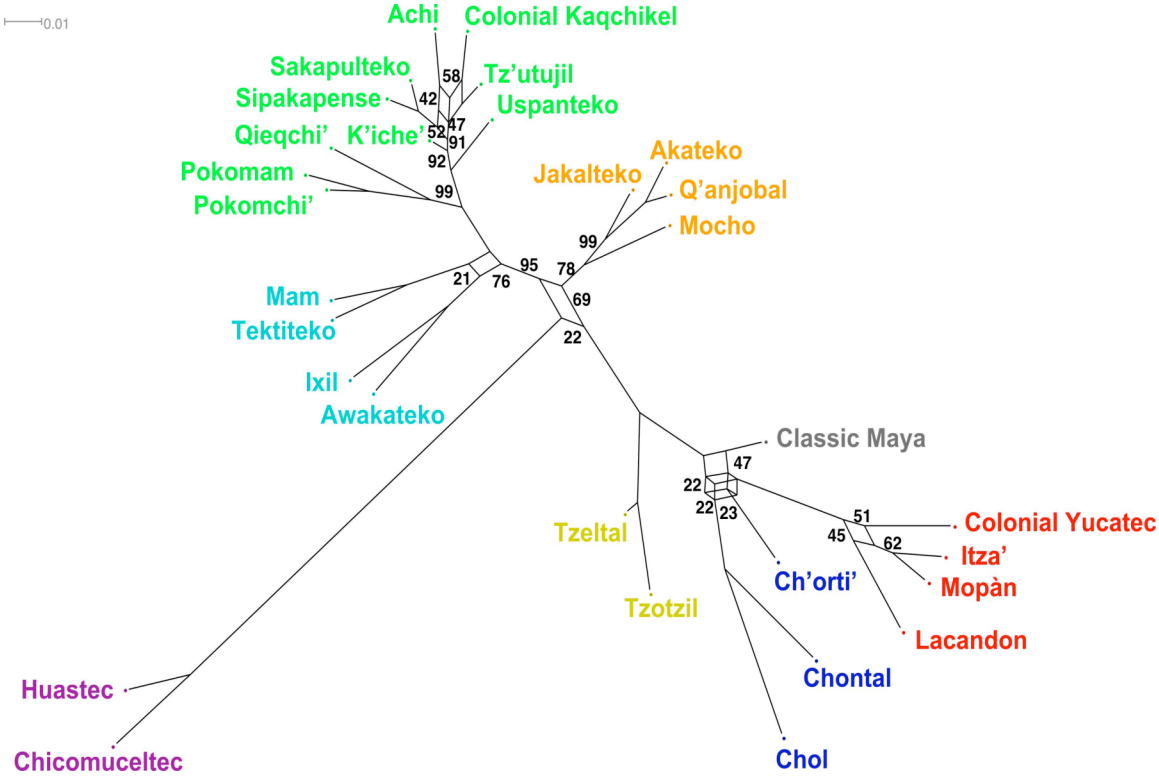
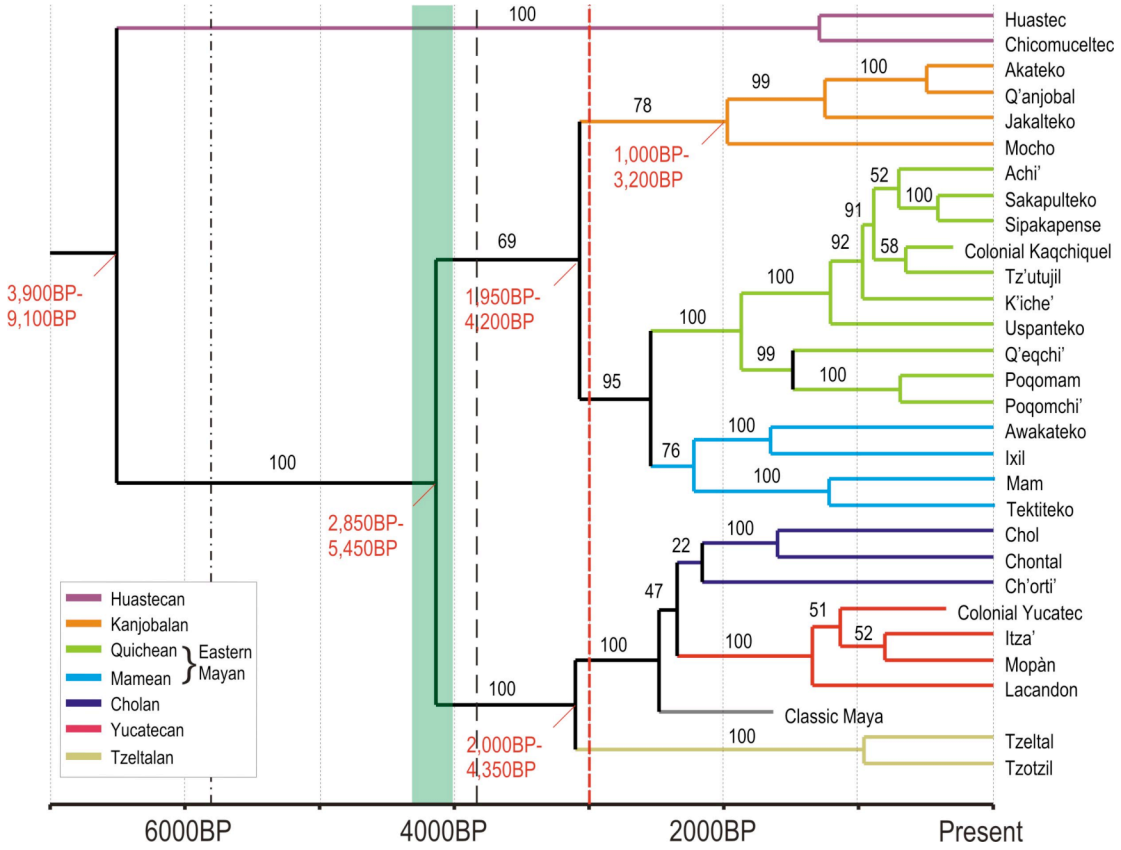


FIGURE 6.2: Consensus network of the Bayesian sample distribution of trees under the Time-reversible model. Bayesian posterior probability values are shown for splits with less than 100% support. All splits that occur in over 20% of the sample distribution are included. Branch lengths are proportional to the inferred amount of change. Language names are colour coded according to recognized subgroups (Campbell, 1997).

Figure 6.3a shows the sample distribution summarized as a consensus tree for the time-reversible model. Since a tree cannot represent incompatible groupings, only the most well-supported clades are shown, along with Bayesian posterior probability values. Branch lengths are drawn proportional to time, with 95% confidence intervals for key divergences shown in red. Whilst a Bayesian approach allows us to quantify random error it cannot account for systematic error – i.e. error in our estimates due to the effects of model misspecification. We thus repeated the analyses using a second, very different, stochastic-Dollo model of cognate birth and death in which each cognate can evolve only once (see Section 6.2), the results of which are shown in Figure 6.3b. One advantage of this model is that we do not have to assign an outgroup because the root of the tree can be inferred from the data. The inferred language relationships were very similar to those found under the time-reversible model. Huastec was reconstructed as the outgroup in 100% of the sample distribution of trees under the stochastic-Dollo model.

a) time reversible model



b) stochastic-Dollo model

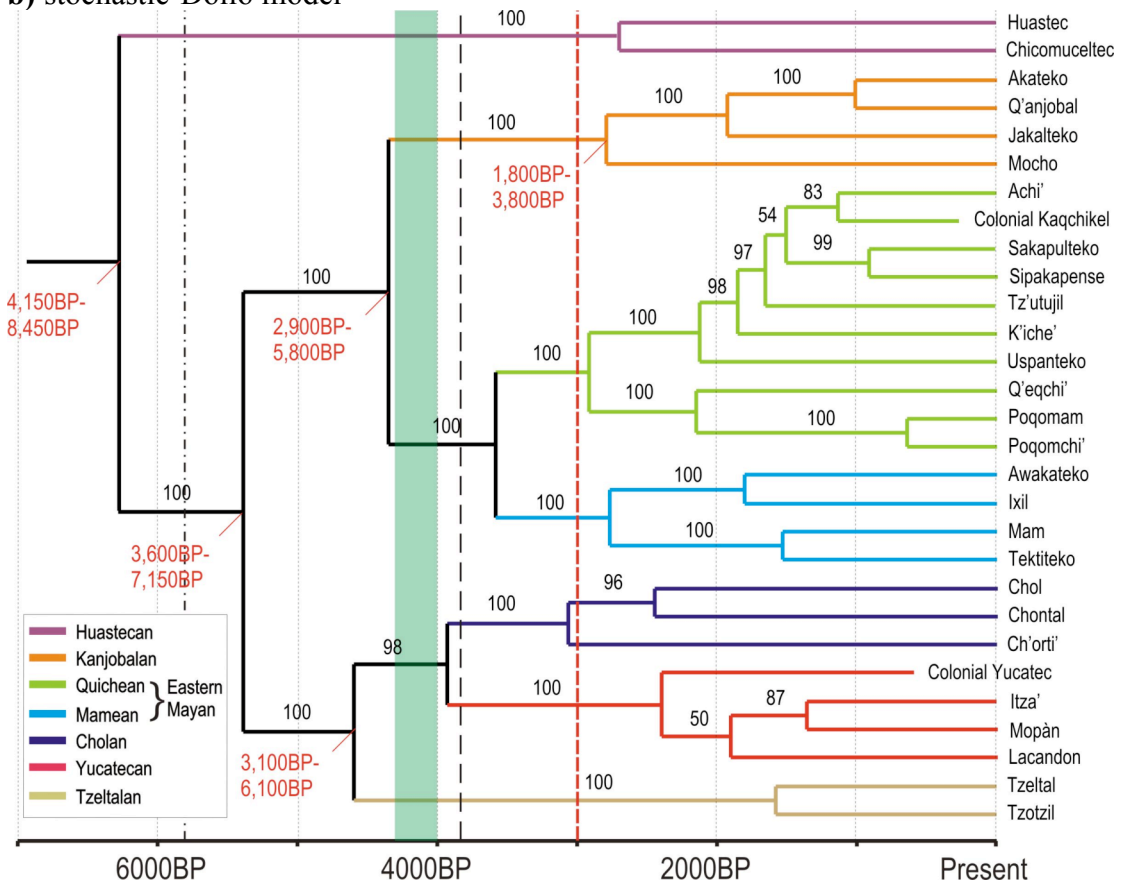


FIGURE 6.3: Consensus trees from the Bayesian sample distribution of trees under the time-reversible model **a** and stochastic-Dollo model **b** of character evolution. Bayesian posterior probability values indicating support for the clades are shown above the branch immediately ancestral to each clade. Branch lengths are drawn proportional to time. 95% confidence intervals for key divergences are shown in red. The dot-dashed line indicates the time of first appearance of maize pollen in the archaeological record of the Maya area (Pohl *et al*., 1996). Green bar indicates date range associated with marked increase in agriculture in the lowland area. Black dashed line indicates appearance of Barra complex, the earliest distinctly "Maya" pottery from the Pacific Coast of Guatemala (McKillop, 2004). Red dashed line indicates first appearance of pottery (called Swasey) in the Maya lowlands (McKillop, 2004).

As a further validation exercise, we repeated the above analyses under a range of model assumptions. Tree topology was highly robust to assumptions of the method, however, divergence times showed some variation. For each of these validation analyses, Figure 6.4 shows confidence intervals for the age at the root (left-hand graph) and at the primary split separating Eastern Mayan/Q'anjobalan from Yucatecan/Cholan/Tzeltalan (right-hand graph). The divergence time estimates of the time-reversible model with a yule branch length prior (6.4a) were largely unaffected by the assumption of strictly clock-like evolution (6.4b). A flat prior on branch lengths produced slightly older date estimates (6.4c). One way to test for the effect of unidentified borrowing between closely related languages is to analyse a subset of languages sampled from across the tree. This reduces the number of closely related languages in the analysis and should reduce the effect of borrowing between closely related languages. An analysis of a subset of 16 languages representing all of the major groups again produced slightly older divergence time estimates (6.4d), but did not differ significantly from the 30-language analyses. Analyses using the stochastic-Dollo model with a yule branch length prior (6.4e) and flat branch length prior (6.4f) produced similar results. Assuming Indo-European rates of evolution (6.4g; rates from Atkinson *et al*, 2005) produced similar mean age estimates but with substantially less variance. Finally, a reduced, 16-language analysis under the stochasitic-Dollo model (6.4h) again produced similar, although slightly younger, divergence time estimates.

The ancestral Mayan homeland was inferred across the sample distribution of trees generated under the stochastic Dollo model. All of the trees produced an unequivocal reconstruction supporting a homeland in the Chiapas Highland region. All of the trees also support a highland location for the split separating Eastern Mayan/Q'anjobalan from Yucatecan/Cholan/Tzeltalan.

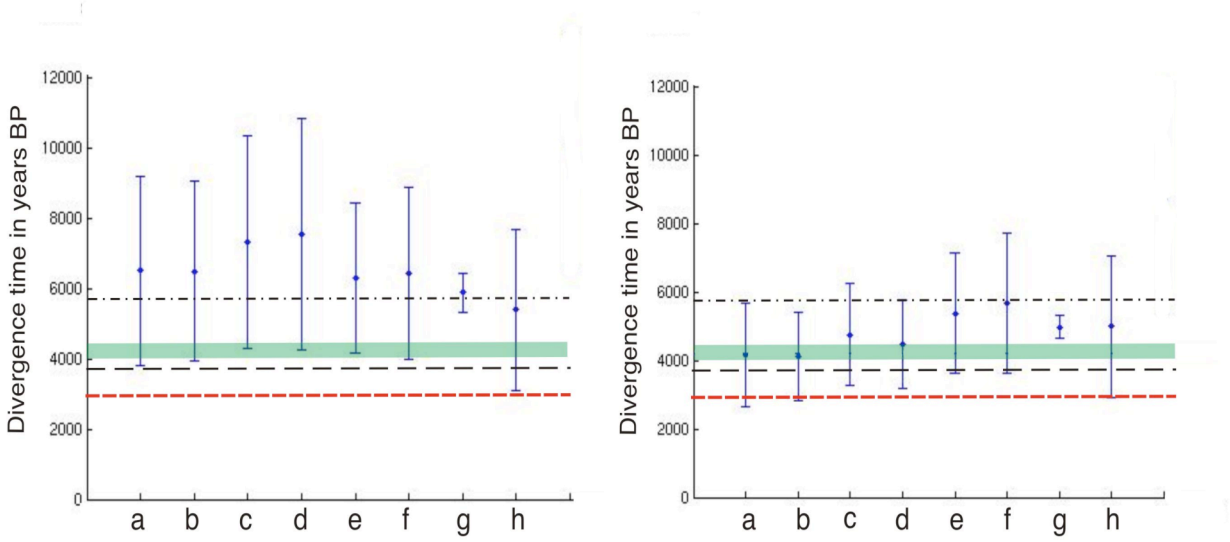


FIGURE 6.4: Confidence intervals for the age at the root of the Mayan language tree (left-hand graph) and for the age of the split separating Eastern Mayan/Q'anjobalan from Yucatecan/Cholan/Tzeltalan (right-hand graph): **a** time-reversible model yule branch length-prior, **b** assuming a strict clock, **c** flat branch-length priors, **d** subset of 16 languages; **e** stochastic-Dollo model yule branch length prior, **f** flat branch length prior, **g** Indo-European rates, **h** subset of 16 languages. The dot-dashed line indicates the time of first appearance of maize pollen in the archaeological record of the Maya area (Pohl *et al*., 1996). Green bar indicates date range associated with marked increase in agriculture in the lowland area. Dashed line indicates appearance of Barra complex, the earliest distinctly "Maya" pottery from the Pacific Coast of Guatemala (McKillop, 2004). Red dashed line indicates first appearance of pottery (called Swasey) in the Maya lowlands (McKillop, 2004).

6.4 DISCUSSION

Phylogenetic network analysis generally supports the traditional Mayan subgroups but indicates considerable non-tree-like signal in the data. The observed pattern of reticulation is consistent with current debates in Mayan historical linguistics. The controversy over the correct position of Tojolabal and Chuj (Kaufman, 1976; Campbell, 1977; Robertson, 1977) appears to be the product of language contact resulting in extensive unrecognized borrowing. Interestingly, this pattern of borrowing is not unique to the lexicon. Phonological characters show an identical pattern of conflicting signal – e.g., the sound innovation of **b'* to *p'* groups Tojolobal with Chuj and the Eastern Mayan languages plus Q'anjobalan, whilst the innovation of **k* to *q* groups Tojolobal with Yucatecan, Cholan and Tzeltalan (Josserand, 1975, after Kaufman, 1969). Our analyses also show evidence of some reticulation in the highland Mayan languages, particularly K'ichean, perhaps indicating the presence of a highland dialect chain. In order to examine the effect of this non-tree-like signal on our analyses we can analyse synthetic data generated under a model that incorporates

borrowing between languages. Using this approach, Atkinson *et al*. (2005) found that tree topology and divergence time estimates were robust to borrowing under both the models used here (see Chapter 5). However, under higher rates of borrowing, the stochastic-Dollo model tended to underestimate divergence times. Whilst the Dollo assumption, which only allows cognates to be born once, may be intuitively more realistic than a time-reversible model, it may also make the stochastic-Dollo model less robust to the influence of borrowing, which can artificially scatter cognates across the tree (Atkinson *et al.*, 2005). For this reason, under conditions of widespread borrowing shown here, the time-reversible model may be preferable. However, regardless of the model, the observed patterns of reticulation need to be considered when interpreting the tree-based phylogenetic analyses discussed below.

The sample distribution of Mayan language phylogenies summarized in Figure 6.2 and Figure 6.3 are consistent with conventional lower order language subgroupings and also demonstrate some well-supported deeper relationships. Whilst trees generated under the time-reversible model were assigned a Huastecan outgroup *a priori*, the stochastic-Dollo analysis, which allows us to estimate the root of the tree from the lexical data alone, favours a Huastecan outgroup. This is consistent with the traditional view (Campbell and Kaufman, 1985). Both models find strong support for an Eastern Mayan group, comprising K'ichean and Mamean, and support for a major division separating Yucatecan, Cholan and Tzeltalan from Eastern Mayan plus Q'anjobalan. Given these deeper relationships, our analysis suggests that the Maya homeland is unlikely to be in the Maya lowland or coastal Veracruz area. Our reconstruction of geographic characters shows all trees support a Chiapas Highland origin. Additionally, the geographic origin of the next split, separating Eastern Mayan and Q'anjobalan from Yucatecan, Cholan and Tzeltalan, is reconstructed as highland in all trees. These results are consistent with the currently favoured highland origin theory.

The inferred relationships between the lowland languages raise a number of interesting questions. Cholan is reconstructed as most closely related to its lowland neighbour Yucatecan. However, previous comparative work based on phonological data has found support for a Cholan-Tzeltalan grouping (e.g. Kaufman, 1976;

Kaufman and Norman, 1984), placing Yucatecan on its own outside these two. There are two possible explanations for this. One is that the Cholan-Tzeltalan grouping, favoured by phonological evidence, is correct and our finding of a close relationship between Cholan and Yucatecan is merely the product of unrecognized lexical borrowing. This would make sense given the shared territory and widespread dominance of first Cholan, and then Yucatecan culture in the Classic and post-Classic periods, which may have promoted borrowing from the dominant culture (see Justeson *et al*., 1985). Indeed, the splits highlighted in Figure 6.1b show some conflicting lexical evidence grouping Cholan-Yucatecan (green split) and Cholan-Tzeltalan (blue split). Another possibility is that the Cholan-Yucatecan grouping, favoured by lexical evidence, is correct and the phonological characters shared between Cholan and Tzeltalan are either shared retentions or the product of areal diffusion or parallel development – similar phonological changes occurring independently in different lineages. This scenario can also explain the archaeological evidence for shared cultural traits across the Maya lowlands (e.g. the Swasey ceramic complex; Hammond, 2000) as the result of common ancestry rather than borrowed technology. Kaufman and Norman (1984) offer only a single sound correspondence in support of the Tzeltalan-Cholan grouping (**k* and **k'* to *ch* and *ch'* respectively). The alternative Yucatecan-Cholan grouping is also supported by phonological evidence - e.g. Yucatecan and Cholan *h* compared to Tzeltalan *j*. We performed a simple parsimony optimization of 25 Mayan phonological characters (data from Campbell, 1988) using *MacClade* (*v. 4.06*; Maddison and Maddison, 1992) and found the phonological data could not distinguish between a Cholan/Yucatecan or Cholan/Tzeltalan grouping – both arrangements required the same number of phonological changes. Given this uncertainty, further investigation of the relationships between the Cholan, Yucatecan and Tzeltalan languages seems prudent and may shed light on the early history of the Classic Maya.

The situation is further complicated when we consider the position of Classic Maya itself. It is generally accepted that there were at least two languages/dialects represented in the Classic Maya hieroglyphic inscriptions; an earlier, southern, Cholan, variety and a later, more northern, Yucatecan, variety. However, the boundary between the two in time and space is indistinct and the relationship of each

to specific contemporary languages within these subgroups remains controversial. We focused on the earlier Cholan variety of Classic Maya by restricting our sampling to only those terms attested prior to 700AD in an attempt to minimize the presence of later Yucatecan terms. Scholars have variously linked the earlier variety of Classic Maya to Proto-Cholan (Justeson and Fox 1989; Justeson and Campbell 1997), pre-Western-Cholan (the ancestor of Chol and Chontal; Josserand *et al*. 1985) and pre-Ch'olti'/Ch'orti'an (Houston, Robertson and Stuart, 2000). Our analyses place Classic Maya either at the base of a Cholan-Yucatecan group (47%), as deriving from the Ch'orti' lineage along with Yucatecan (23%), or as deriving from the Ch'orti' lineage with Yucatecan as a sister group to Cholan (17%). A substantial amount of missing data for Classic Maya (~50%) could have contributed to the lack of clarity here. The uncertainty in our reconstruction may also be due to limitations on the extent to which scholars can identify sites as representing the Cholan versus the Yucatecan variety of Classic Maya. Our Classic Maya data contains both uniquely Cholan and Yucatecan terms suggesting either a) extensive borrowing or perhaps a dialect chain, or b) that we are in fact dealing with more than one language. In both cases, the effect on our analysis would be to artificially place Classic Maya as ancestral to both Cholan and Yucatecan when in fact this may not be the case. It is thus interesting to note that we also see evidence for an association between Classic Maya and Ch'orti'an, supporting Houston, Robertson and Stuart's (2000) proposal that Ch'orti'/Ch'olti' speakers are the descendents of the early Classic Maya.

Mean divergence time estimates for the age at the root of the Mayan language tree under both models were approximately 6,000 to 6,500 years (Figure 6.3), considerably older than Kaufman's (1976) glottochronology-based estimate of 4,200 years. However, both models produced a 95% confidence interval for the root age of between approximately 4,000BP and 9,000BP, a range that does overlap with Kaufman's finding. Root age estimates were relatively robust between models, giving consistent results across the different analyses plotted in Figure 6.4. Mean internal age estimates showed some disagreement between the two models (see Figures 6.3 and 6.4), although error bars overlapped - e.g., the inferred age of Q'anjobalan was 1,000-3,200BP under the time-reversible model, but 1,800-3,800BP under the stochastic-Dollo model. We thus need to be cautious in interpreting these results. Where there is

disagreement, the chronology inferred under the time-reversible model may be more accurate. Classic Maya, and hence an important age constraint, had to be excluded from the stochastic-Dollo model analysis due to the amount of missing cognate information in the Classic Maya data. In addition, borrowing may have had a greater impact on the stochastic-Dollo model, which is more susceptible to the effects of non-tree-like signal in the data.

When we compare our results with the archaeological record a number of interesting patterns emerge. The original inhabitants of the current Maya area were pre-ceramic hunter-gatherers who arrived at least 10,000 years ago (Zeitlin and Zeitlin, 2000). Domesticated cultigens, principally maize and manioc, first appear in Mesoamerica around 6,350BP in the Oaxaca Valley region (Piperno and Flannery, 2001) and later in what is now Mayan territory in Belize (Pohl *et al*. 1996) and the Chiapas coast (Kennet and Voorhies, 1996) between 5,500BP and 5,000BP (Figure 6.3, dot-dashed line). However, large-scale deforestation associated with the expansion of agriculture in the region does not occur until after ca. 4,500BP (Figure 6.3, green bar), with some areas showing little or no evidence of maize utilization until the early-Preclassic[2], about 3,500-3,000BP (Pohl *et al*., 1996). Support for a theory of Mayan languages spreading with agriculture is currently based on the correspondence between this later agricultural fluorescence and Kaufman's (1976) 4,200BP glottochronology-based age estimate for the Mayan language family. Our findings indicate that the family as a whole may be older than this. The chronology shown in Figure 6.3a suggests that the divergence of Huastecan is more likely to have occurred some 6,000 years ago, perhaps with the initial spread of agriculture. Although previously dismissed on the basis of the glottochronological date estimates, this revised timescale is consistent with an apparent cultural continuity in the Maya area stretching from Preclassic times, back to the sixth millennia BP (Coe, 2005). Estimated divergence times under the time-reversible model for the split separating the highland Eastern Mayan and Q'anjobalan languages from the lowland Yucatecan and Cholan languages are centred around 4,200BP. This suggests that the increase in agricultural activity at this time may have been driven by the expansion of Mayan languages from an original highland homeland, giving rise to the first major division in the family after the split

[2] The Mesoamerican Preclassic period stretches from 4,000BP to 1,700 BP (McKillop, 2004).

with Huastecan. Finally, the Preclassic also marked the beginning of the adoption of pottery making and a gradual increase in cultural complexity culminating in the large cities and state level organization of the Classic period (Hammond, 2000; Shearer, 2000). Around 3,000BP the Swasey pottery complex emerges in the Maya lowlands (Figure 6.3, red dashed line; Hammond, 2000; McKillop, 2004). It is interesting to note that this time coincides with the inferred divergence of the lowland Cholan/Yucatecan plus Tzeltalan group under the time-reversible model (Figure 6.3a).

These correlations suggest important areas for future research but should not be over-interpreted. In order to overcome the concerns about circularity expressed in Renfrew's lament, quoted in the introduction, establishing congruence between archaeological, genetic and linguistic evidence requires an accurate assessment of uncertainty and a critical examination of the fit between independent lines of evidence. For example, whilst our date estimates for the split separating Eastern Mayan/Q'anjobalan from Cholan/Yucatecan/Tzeltalan fit nicely with agricultural intensification beginning 4,200 years ago, the 95% confidence interval for this split spans 3,000 years. In addition, the adoption of maize farming was gradual throughout the Maya lowlands, inconsistent with the notion of high status agriculturalists expanding into new territories and quickly replacing hunter-gatherer populations (Marcus, 2003; van der Merwe *et al*., 2000). A simple scenario of a dominant language/culture spreading with the expansion of agriculture may thus be unrealistic in the case of the Maya (Pohl *et al*., 1996). Integrating archaeological evidence for large-scale demographic/cultural shifts with evidence from historical linguistics will require a more fine-grained understanding of the chronology of Mayan language divergence.

There is, however, reason to believe that divergence time estimates can be substantially improved and uncertainty reduced. The analysis shown in Figure 6.4g, where we assumed that the rate of lexical replacement was the same as for Indo-European, stands out as having a much smaller confidence interval[3]. This shows that

[3] Whilst assuming Indo-European rates (Figure 6.4g) produced divergence times consistent with analyses based on internal age calibrations, this should not be interpreted as validation of a universal and constant rate of language change. Even if we assume constant rates across a tree, estimates can

with more information about the rates of evolution we can substantially reduce uncertainty to further clarify the timing of the Mayan expansion. This reduced uncertainty is achievable without assuming Indo-European rates, if we increase the number of calibration points based on historically attested Mayan language divergence events. For the current analyses only four age constraints could be identified, with additional rate information supplied by including extinct languages. The Poqomam-Poqomchi' internal age constraint was the only single constraint with both an upper and lower bound. This meant inferred rates were particularly sensitive to variation in the branch-lengths leading to the Poqomam-Poqomchi' clade. In our analysis of time depth in Indo-European (Gray and Atkinson, 2003; Atkinson *et al*., 2005; Atkinson and Gray, 2006, in press) we used 11 much deeper internal age constraints, 6 of which included an upper and lower bound. This greatly reduced the effect of phylogenetic uncertainty on estimated rates and produced much narrower divergence time confidence intervals. Increasing the sample size from the Swadesh 100 to the Swadesh 200 word list further reduced uncertainty. This suggests two important means of refining the age estimates presented here. First, more research is needed linking archaeological and ethno-historical evidence of colonial and late-pre-colonial population movements to language divergence events that can then be used as calibration points. Second, analysing an expanded dataset of Swadesh 200 word list terms in each of the languages should reduce uncertainty in branch-length estimates. Finally, relationships have been hypothesized between the Mayan languages and other Mesoamerican language families, including Mixe-Zoquean and Totonacan languages (Campbell and Kaufman, 1985). These relationships are notoriously difficult to validate due to problems with distinguishing between genuine cognate terms, chance resemblances and borrowed terms. Nonetheless, by accounting for such possible errors in cognacy judgements and including languages from these families as an outgroup, it may be possible to reduce uncertainty at the base of the tree and provide more precise Mayan divergence time estimates. Finally, Diamond and Bellwood (2003) and Bellwood (2005) have argued, also on the basis of glottochronology time

vary due to sample size and cognacy judgement criteria and can therefore be expected to vary between datasets, even for the same language family. For example, shorter inferred branch lengths meant that rate estimates based on the 16-language analysis shown in Figure 6.4h were on average 15% slower than the principal 30-language analysis – rate estimates are not transferable between the analyses even though divergence time estimates were virtually identical. For this reason it is important that dates are calibrated internally for each analysis rather than assumed to be universal.

estimates, that these language families had a common origin, diversifying as they spread with the spread of agriculture across Mesoamerica. An analysis combing languages from different families may be able to shed light on this hypothesis and Mesoamerican prehistory more generally.

6.5 CONCLUSION

The current study is a significant step forward in our understanding of Mayan prehistory. First, we highlighted interesting areas of reticulate evolution and uncertainty in the Mayan language family. In particular, our results indicate relationships between the lowland Mayan languages that warrant further investigation, including the position of Classic Maya. Second, we inferred a likely highland Mayan homeland by mapping geographic characters onto language trees. Third, we were able to infer language divergence times and quantify uncertainty in these estimates without the flawed assumptions of glottochronology. Our date estimates are older than traditional glottochronology-based estimates and coincide with a number of interesting patterns in the archaeological record that warrant further investigation. Our findings also have broader implications for our understanding of Mesoamerican prehistory, suggesting that the glottochronology-based dates used to argue for the spread of agriculture with the Uto-Aztecan, Oto-Manguean, and Mixe-Zoquean language families (Diamond and Bellwood, 2003; Bellwood, 2005) may also require revision. We are optimistic that further work in this area will allow uncertainty in language tree topology and divergence time estimates to be substantially reduced and can provide a solid framework for elucidating the prehistory of the Maya and Mesoamerica.

6.6 REFERENCES

Atkinson, Q. D. Nicholls, G., Welch, D. and Gray, R. D. 2005. From Words to Dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2):193-219.

Bateman, R., Goddard, I., O'Grady, R., Funk, V., Mooi, R., Kress, W. and Cannell P. 1990. Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* 31:1-24.

Bellwood, P. 2005. *First Farmers: The origins of agricultural societies.* Blackwell Publishing, Malden (MA).

Bennetzen, J., Buckler, E., Chandler, V., Doebley, J., Dorweiler, J., Gaut, B., Freeling, M., Hake, S., Kellogg, E., Poethig, R., Walbot, V., and Wessler, S. 2001. Genetic evidence and the origin of maize. *Latin American Antiquity,* 12: 84–86.

Bergsland, K., and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3:115–153.

Blust, R. 2000. Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. Pages 311-332 in *Time depth in historical linguistics*. (eds.) C. Renfrew, A. McMahon, and L. Trask. The McDonald Institute for Archaeological Research, Cambridge.

Bocabulario de Maya Than de Viena. 1993. Bocabulario de Maya Than, Facsimile and edited version by René Acuña. Instituto de Investigaciones Filológicas, Universidad Nacional Autónoma de México, México.

Boot, E. 2002. *A preliminary Classic Maya-English/English-Classic Maya Vocabulary of Hieroglyphic Readings.* Leiden University, NL.

Bryant, D., F. Filimon, and R. D. Gray. 2005. Untangling our past: Pacific settlement, phylogenetic trees and Austronesian languages. Pages 65-89 in *The evolution of cultural diversity: Phylogenetic approaches.* (eds.) R. Mace, C. Holden, and S. Shennan. UCL Press, London.

Campbell, L. 1977. *K'ichean Linguistic Prehistory*. Univ. Calif. Publ. Ling. 81. Univ. Calif. Press, Los Angeles.

Campbell, L. 1978. Quichean Prehistory: Linguistic contribution. Pages 25-54 in *Papers in Mayan Linguistics*. (ed.) N. C. England. (Miscellaneous Publications in Anthropology, 6, Studies in Mayan Linguistics, 2.) University of Missouri, Colombia, Missouri. 25-54.

Campbell, L. 1984. The implications of Mayan historical linguistics for glyphic research. Pages 1-16 in *Phoneticism in Mayan Hieroglyphic Writing*. (Eds.) John S. Justeson and L. Campbell. Institute for MesoAmerican Studies, State University of New York at Albany.

Campbell, L. 1988. *The linguistics of Southeast Chiapas, Mexico.* New World Archaeological Foundation, Brigham Young University, Provo (Utah).

Campbell, L. 1997. *American Indian Languages: The historical linguistics of Native America.* Oxford University Press, New York.

Campbell, L. 2004. *Historical linguistics: An introduction. 2nd edition.* Edinburgh University Press, Edinburgh.

Campbell, L. and Kaufman, T. 1985. Mayan Linguistics: Where are we now? *Annual Review of Anthropology* 14:187-98.

Carmack, R. M. 1973. *Quichean Civilization.* University of California Press, Los Angeles, CA.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press, Princeton (NJ).

Coe, M. D. 2005. *The Maya.* 7th Edition. Thames and Hudson, New York (NY).

De Coto, Thomas. 1647/1983. Thesasavrvs Verborv - Vocabvlario de la Lengua Cakchiquel v(el) Guatemalteca, Nueuamente Hecho y Recopilado Con Summo Estudio, Trauajo y Erudicion. Universidad Nacional Autonoma de Mexico, Mexico.

Diamond, J., and P. Bellwood. 2003. Farmers and their languages: The first expansions. *Science* 300:597.

Dienhart, J. M. 1989. *The Mayan Languages – A comparative vocabulary*. Odense University Press, Odense, Denmark.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., and Levinson, S. C. 2005. Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 309:2072-2075.

Farris, J. S. 1977. Phylogenetic analysis under Dollo's Law. *Systematic Zoology* 26:77-88.

Fox, J. A. 1978. Proto-Mayan accent, morpheme structure conditions, and velar innovations. PhD dissertation, University of Chicago

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435-9.

Gray, R., and F. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405:1052–1055.

Hammond, N. 2000. The Maya Lowlands: Pioneer Farmers to Merchant Princes. Pages 196-249 in *The Cambridge History of the Native Peoples of the Americas, v. II, Part 1* (eds.) R. E. W. Adams and M. J. Macleod. Cambridge University Press, Cambridge.

Hernandez, J. M. (ed) 1590/1929. *Diccionario de Motul*. Mérida.

Hjelmslev, L. 1958. *Essai d'une critique de la methode dite glottochronologique*. Proceedings of the Thirty-second International Congress of Americanists, Copenhagen, 1956. Munksgaard, Copenhagen.

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: A maximum-parsimony analysis. *Royal Society of London, Proceedings Series B* 269, 793–9.

Holden, C. J. 2003. Evolution of Bantu languages. Pages 121-124 in *Yearbook of Science and Technology*. McGraw-Hill, New York.

Holland, B., and V. Moulton. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. Pages 165–176 in *Algorithms in bioinformatics, WABI 2003.* (eds.) G. Benson and R. Page. Springer-Verlag, Berlin, Germany.

Houston, S., Robertson, J. and Stuart, D. 2000. The Language of Classic Maya Inscriptions. *Current Anthropology* 41(3):321-356.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian Inference of Phylogeny. *Bioinformatics* 17:754–755.

Huson D. H. and Bryant, D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23(2):254-267.

Josserand, J. K. 1975. Archaeological and linguistic correlations for Mayan prehistory. Actas del XLI Congreso Internacional de Americanistas, México, 2 al 7 de septiembre de 1974, vol. 1, pp. 501-510.

Josserand, K., Schele, L., and Hopkins, N. A. 1985. Auxiliary Verb + ti Constructions in the Classic Maya Inscriptions. Pages 87-102 in Fourth Palenque Round Table, 1980, Vol. VI, edited by E. P. Benson. Center for Pre-Columbian Art Research, San Francisco.

Justeson, J. S., and Campbell, L. 1997. The Linguistic Background of Maya Hieroglyphic Writing: Arguments against a "Highland Mayan" Role. Pages 41-67 in *The Language of Maya Hieroglyphs* (eds.) M. J. Macri and A. Ford. Pre-Columbian Research Institute, San Francisco.

Justeson, J. S., and J. A. Fox. 1989. Hieroglyphic evidence for the languages of the Lowland Maya. Unpublished MS in possession of author.

Justeson, J. S., Norman, W. N., Campbell, L., and Kaufman, T. 1985. *The Foreign Impact on Lowland Mayan Language and Script*. Publication 53. Middle American Research Institute, Tulane University, New Orleans.

Kaufman, T. S. 1969. *Some Recent Hypotheses on Mayan Diversification*. Working Paper No 26. Language-Behaviour Research Laboratory, Berkeley.

Kaufman, T. S. 1974. *Meso-American Indian Languages*. Encyclopaedia Britiannic, 15th Edition.

Kaufman, T. S. 1976. Archaeological and linguistic correlations in Mayaland and associated areas of Meso-America. *Archaeology and Linguistics* 8(1):101-18.

Kaufman, T. S. and Norman, W. M. 1984. An Outline of Proto-Cholan Phonology, Morphology and Vocabulary. Pages 77-166 in *Phoneticism in Mayan Hieroglyphic Writing.* (eds.) John S. Justeson and L. Campbell. Institute for MesoAmerican Studies, State University of New York at Albany.

Kaufman, T. S. 2003. *A preliminary Mayan Etymological Dictionary*. Unpublished.

Kennett, D. J., and Voorhies, B. 1996. Oxygen isotopic analysis of archaeological shells to detect seasonal use of wetlands on the southern Pacific coast of Mexico. Journal of Archaeological. *Science* 23: 689–704.

Kirch P. V. and Green R. C. 1987. History, phylogeny, and evolution in Polynesia. *Current Anthropology* 28:431–456.

Maddison, W. P. & Maddison, D.R. (1992). MacClade version 3. Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland (MA).

Marcus, J. 2003. Recent Advances in Maya Archaeology. *Journal of Archaeological Research* 11(2):71-148.

McKillop, H. 2004. *The Ancient Maya – new perspectives*. ABC Clio, Santa Barbara, CA.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087–91.

Piperno, D. R., and Flannery, K. V. 2001. The earliest archaeological maize (Zea mays L.) from highland Mexico: New accelerator mass spectrometry dates and their implications. *Proceedings of the National Academy of Sciences* 98(4): 2101–2103.

Piperno, D. M. and Pearsall, D. 1993. Phytoliths in the Reproductive Structures of Maize and Teosinte: Implications for the Study of Maize Evolution. *Journal of Archaeological Science* 20:337-362.

Pohl, M. D., Pope, K. O., Jones, J. G., Jacob, J. S., Pipemo, D. R., deFrance, S. D., Lentz, D. L., Gifford, J. A., Danforth, M. E., and Josserand, J. K. 1996. Early Agriculture in the Maya Lowlands. *Latin American Antiquity* 7(4):355-72.

Renfrew, C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins.* Cape, London.

Rexová, K., D. Frynta, and J. Zrzavy. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19:120–127.

Robertson, J. S. 1977. A proposed revision in Mayan subgrouping. *International Journal of American Linguistics* 43:105-20.

Sanderson, M. 2002a. Estimating absolute rates of evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.

Sanderson, M. 2002b. *R8s, Analysis of Rates of Evolution, version 1.50.* http://ginger.ucdavis.edu/r8s/

Shearer, R. J. 2000. The Maya Highland and the Adjacent Pacific Coast. Pages 448-499 in *The Cambridge History of the Native Peoples of the Americas, v. II, Part 1*, (eds.) R. E. W. Adams and M. J. Macleod. Cambridge University Press, Cambridge.

Steel, M., M. Hendy, and D. Penny. 1988. Loss of information in genetic distances. *Nature* 333, 494-495.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *American Philosophical Society, Proceedings* 96:453–463.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. International *Journal of American Linguistics* 21:121–137.

Swadesh, M. 1960. *Interrelaciones de las lenguas mayenses*. Anales Institutp Nacional de Anthropologia e Historia, Mexico.

Swadesh, M. 1967. Lexicostatistic Classification. Pages 79-115 in *Handbook of Middle American Indians, Volume 5 – Linguistics.* (ed.) N. A. McQuown. University of Texas Press, Austin.

van der Merwe, N. J., Tykot, R. H., Hammond, N., and Oakberg, K. 2000. Diet and animal husbandry of the Preclassic Maya at Cuello, Belize: Isotopic and zooarchaeological evidence. Pages 23-38 in *Biogeochemical Approaches to Paleodietary Analysis* (eds.) Ambrose, S. H., and Katzenberg, M. A. Kluwer Academic/Plenum, New York.

Willey, G. R. 1982. Maya Archaeology. *Science* 215:260-267.

Zetilin, R. N. and Zeitlin, J. F. 2000. The Paleoindian and Archaic Cultures of Mesoamerica. Pages 45-121 in *The Cambridge History of the Native Peoples of the Americas, v. II, Part 1,* (eds.) R. E. W. Adams and M. J. Macleod. Cambridge University Press, Cambridge.

Chapter Seven

THE PERILS OF DATING EVE
IMPROVED ANALYSES OF HUMAN MTDNA SEQUENCES HIGHLIGHT TIME DEPENDENCY OF MOLECULAR DATE ESTIMATES AND A TWO-PHASE POPULATION EXPANSION

ABSTRACT

Previous phylogenetic analyses of human mitochondrial DNA (mtDNA) have been instrumental in establishing the validity of a recent African human origin. However, date estimates for the age of "mitochondrial Eve", the most recent common ancestor of the human maternal lineage, still show considerable variation. This is not surprising, as previous studies have been limited by the available data, choice of tree-building methods and rate calibrations. Recently, however, there has been a proliferation of readily available gene sequence data, including complete human mtDNA sequences, and a corresponding increase in the sophistication of phylogenetic methods available to analyse sequence data. For example, it is now possible to account for variation in substitution rates between sites and through time, to model population demographics through time, and to quantify uncertainty in parameter estimates under a range of probabilistic models of sequence evolution. Here we attempt to overcome the limitations of previous mtDNA-based studies by applying these new methods to a data set of 252 complete human mtDNA sequences to infer a timescale for human origins and expansion. Our analyses support a human origin within the last 150,000-250,000 years. We show, however, that date estimates depend on the time depth of calibration points in a manner consistent with the J-shaped rate curve described by Ho et al. *(2005). The degree of change in rates through time is shown to depend partly on the choice of substitution model, with simpler models tending to exhibit greater time dependency of rates. Using a coalescent approach, we also model human population size through time and find support for a two-phase human population expansion. Finally, we explore the fit between the inferred mtDNA phylogeny and major language family groupings.*

7.1 INTRODUCTION

7.1.1 FIRST DATES

Phylogenetic analyses of global human genetic variation provide crucial information about the origin and evolution of modern humans. mtDNA is particularly well-suited for phylogenetic inference because it evolves much faster than most nuclear DNA (Howell *et al.*, 1996) and so is rich in the genetic variation required to extract phylogenetic signal from the data. In addition, mtDNA is haploid, inherited only maternally, and hence its evolution is not complicated by recombination during meiosis (Elson *et al*., 2001). Date estimates for the age of "mitochondrial Eve", the most recent common ancestor (MRCA) of the human mtDNA lineage, are critical to the currently accepted chronology of human expansion from Africa. Initial mtDNA analyses focused on testing between two competing hypotheses of modern human origins - the "multiregional hypothesis", implying a 1 million year old human expansion from Africa, and the "replacement hypothesis", which holds that modern humans originated in Africa less than 250,000 years ago (see Stringer and Andrews, 1988). Based on date estimates for mitochondrial Eve (Cann *et al*., 1987; Hasegawa and Horai, 1991; Vigilant *et al*., 1991; Stoneking *et al*., 1992; Penny *et al*., 1995; Watson *et al*., 1997; Ingman *et al*., 2000), as well as research involving other genetic markers (Dorit *et al.*, 1995; Ruvolo, 1996) and archaeological evidence (Klein, 1992), a recent African human origin as per the replacement hypothesis is now widely accepted. Interest has now shifted towards providing a more detailed description of the pattern and timing of divergence within the human mtDNA lineage (Forster *et al*., 1996; Saillard *et al*., 2000; Ingman and Gyllensten, 2003; Mishmar *et al*, 2003; Forster and Matsumura, 2005; Macaulay *et al*., 2005; Merriwether *et al*., 2005; Thangaraj *et al*., 2005). However, there is still considerable uncertainty about the age of mitochondrial Eve, and hence the time-scale of human evolution remains contentious.

TABLE 7.1: Summary of key attempts to estimate the age of mitochondrial Eve.

Paper	Mitochon-drial Eve	Data	Tree-building	Rate Calibration	Rate variation	Uncertainty
Cann *et al*. 1987	140-290kya	RFLP (n=147)	Parsimony	3 internal archaeological constraints – New Guinea, Australia	Strict clock + Constant rates across sites	Rate calibration error
Hasegawa and Horai, 1991	280 (191-386) kya[1]	D-loop sequence (n=871 partial sequences)	Maximum likelihood	4 million year human-chimp split	Strict clock + invariant sites	Stochastic variation in branch length estimates and rate estimate
Vigilant *et al*., 1991	208 (166-249) kya	D-loop sequence (n=189)	Parsimony	4-6 million year human-chimp split	Strict clock + Constant rates across sites	Rate calibration error
Stoneking *et al*., 1992	133 (63-356) kya[2]	D-loop sequence (n=239)	UPGMA and Neighbour-joining	1 internal archaeological constraint - 60kya PNG colonization	Strict clock + Constant rates across sites	Stochastic variation in branch length and rate estimates
Pesole *et al*., 1992	300-800 kya	D-loop sequence (n=14)	Markov Clock	5-7.4 million year human-chimp split	Strict clock + invariant sites	Rate calibration error, plus stochastic variation in branch length and rate estimates
Ruvolo *et al*., 1993	298 (129-536) kya[3]	mtDNA COII and ND4-5 sequences (synonymous substitutions used in final analysis; n=6)	Parsimony	6 million year human-chimp split	Strict clock + Constant rates across sites	Stochastic variation in branch length estimates
Horai *et al*., 1995	143 (107-179) kya	Complete mtDNA sequence (rates calculated using D-loop and synonymous substitutions; n=3)	UPGMA	13 million year orangutan-ape split produced 4.9My human-chimp split using synonymous substitution rates	Strict clock + Constant rates across sites	Stochastic variation in branch length estimates
Watson *et al*., 1997	111-148kya	D-loop sequence (n=407)	Median Network – distance-based date estimates	1 internal archaeological constraint – 11,300 years to Eskimo/Na-Dene founder lineages (from Forster *et al*., 1996)	Strict clock + Constant rates across sites	No formal assessment – age range associated with age of most ancient clades.
Ingman *et al*., 2000	172 (72-272) kya	Complete mtDNA sequence (excluding D-loop; n=53)	Neighbour-joining	5 million year human-chimp split	Strict clock + Constant rates across sites	Stochastic variation in branch length estimates
Mishmar *et al*., 2003	198 (160-236) kya	Complete mtDNA sequence (excluding D-loop; n=104)	Neighbour-joining	6.5 million year human-chimp split	Strict clock + Constant rates across sites	Stochastic variation in branch length estimates
Macaulay *et al*., 2005	205 (160-247) kya	Complete mtDNA sequence (excluding D-loop; n=31)	Median Network/Maximum-likelihood branch lengths	6.5 million year human-chimp split	Strict clock + Gamma distributed rates across sites	Stochastic variation in branch length estimates

[1] Hasegawa and Horai (1991) do mention that a 5Mya human-chimp constraint will produce 5/4 increase in date estimates.

[2] Also quote estimate of 137 (63-416) kya using slightly different method.

[3] Quote estimates based on a human-chimp constraint of 4, 6 and 10 Mya - 195 (+/-68), 298 (+/-105) and 506 (+/-178) respectively.

Table 7.1 summarizes the methodology, data and results of some key attempts to infer the age of mitochondrial Eve over the last two decades. Mean age estimates vary substantially, ranging from 133,000 years BP (Stoneking *et al*., 1992) to 298,000 years BP (Ruvolo *et al*., 1993) or, in the case of Pesole *et al*.'s (1992) study, at least 400,000 years BP. If we include uncertainty associated with these estimates, then the feasible date range extends from 63,000 years BP to 800,000 years BP. None of these studies stand out as a clear best estimate; all can be impugned on the basis of either problems with the data, their approach to tree-building, or their method of rate estimation. In what follows, we outline these problems and how they can be overcome using currently available complete mtDNA sequence data and recent advancements in phylogenetic methods.

7.1.2 OVERCOMING DATING PROBLEMS

7.1.2.1 Data

Early mtDNA-based date estimates were constrained by the suitability of the available data for phylogenetic analysis. RFLP (Restriction Fragment Length Polymorphism) data, used in pioneering research by Cann *et al*. (1987), contains only limited variation and can produce unreliable results due to difficulties with modeling RFLP evolution. Later work made use of sequence data from the mitochondrial D-loop (Hasegawa and Horai, 1991; Vigilant *et al*., 1991; Stoneking *et al*., 1992; Pesole *et al*., 1992; Penny *et al*., 1995; Watson *et al*., 1997), a relatively short (~1,000 base-pairs), quickly evolving segment of non-coding DNA. Whilst the D-loop contains considerably more phylogenetic information than RFLP data, this includes a large number of hyper-variable sites (Meyer *et al*., 1999). The fast, highly variable rates of D-loop evolution can create problems for nucleotide substitution models because of the high probability of multiple substitutions along a branch at a given site. Having to correct for a large number of multiple substitutions increases uncertainty and can lead to erroneous branch length and tree topology estimates. For example, Ruvolo (1996) points out that Vigilant *et al*.'s (1991) data are consistent with dates varying by at least a factor of two depending on the method used to correct for multiple substitutions.

More recent studies have tended to make use of complete mtDNA sequences, including the entire coding region in their analyses (e.g. Horai *et al*., 1995; Ingman *et al*., 2000; Mishmar *et al*., 2003; Macaulay *et al*., 2005). There are a number of reasons for this. First, although the mtDNA coding region evolves more slowly than the D-loop (Howell *et al*., 2003), it is much larger (~16,000 base-pairs) and hence contains a greater amount of phylogenetically informative sites. Second, the coding region evolves in a more predictable way that is easier to model. Currently, there are a large number of complete human mtDNA sequences readily available on Genbank (>1700 at time of publication). The analyses presented here were thus based on a representative sample of 252 such sequences plus 3 great ape sequences (see Section 7.2).

7.1.2.2 Tree Building and Quantifying Phylogenetic Uncertainty

It is well known that distance-based tree-building methods result in information loss (Steel, Hendy and Penny, 1988), which reduces the power of the method to infer phylogenies accurately and can give rise to erroneous trees. Character-based tree-building methods, such as parsimony and maximum-likelihood, do not suffer from the same problem of information loss. A likelihood approach also has the advantage of allowing us to explicitly model the process of sequence evolution using a range of models. Where the model is appropriate, likelihood methods generally outperform distance and parsimony tree-building methods (Kuhner and Felsenstein, 1994). However, evaluating trees under a likelihood framework can be computationally intensive. For this reason, previous mtDNA studies have tended to use distance-based methods (e.g. Stoneking *et al*., 1992; Pesole *et al*., 1993; Horai *et al*., 1995; Watson *et al*., 1997; Ingman *et al*., 2000; Mishmar *et al*., 2003), or parsimony (Cann *et al*., 1987; Vigilant *et al*., 1991; Ruvolo *et al*., 1993). Likelihood-based analyses have been restricted to the analysis of small data sets using a single, simple model (e.g. Hasegawa and Horai, 1991; Macaulay *et al*., 2005). Recently, however, a rapid increase in available sequence data, faster computers and the development of heuristic phylogenetic search methods, such as Bayesian inference of phylogeny and Markov chain Monte Carlos (MCMC; Metropolis *et al*., 1953) sampling algorithms, has made it possible to fit complex likelihood substitution models to large sequence data sets. Here we take advantage of these methods to perform a more rigorous analysis of

mtDNA lineage divergence times than has previously been possible, using a range of likelihood models to analyse a large set of complete mtDNA sequence data.

In addition to accurately inferring tree topology and divergence times, these methods allow historical population demographics to be inferred from the data. Previous work using mtDNA mismatch distributions has found conflicting evidence for major human population expansions within the last 100,000 years (e.g. Sherry *et al.*, 1994; Rogers, 1995; *cf.* Excoffier and Schneider, 1999). Mismatch distributions can be difficult to interpret, are distance-based and hence subject to information loss, and date estimates for any expansion are dependent on an *a priori* model of population growth. By assuming a coalescent prior for the tree under a likelihood framework, we can retain character information and explicitly model human population size through time, without assuming an *a priori* parametric growth model.

Another key advantage of using Bayesian inference of phylogeny and MCMC sampling is that we can quantify phylogenetic uncertainty in our parameter estimates, a crucial consideration if results are to be used to test historical hypotheses (see Templeton, 1993; Penny *et al*., 1995). By basing inferences on a single tree or distance matrix, previous estimates for the age of mitochondrial Eve have not fully accounted for phylogenetic uncertainty (*cf.* Penny *et al*. [1995], who focus explicitly on quantifying phylogenetic uncertainty but do not consider divergence times). Rather than yielding a single "optimal" tree, the Bayesian MCMC approach employed here produces a distribution of trees sampled in proportion to their likelihood given a model and the data. By estimating parameters across this sample of trees, we can quantify the degree of phylogenetic uncertainty in our results.

7.1.2.3 Estimating Rates and Dates

There is considerable uncertainty associated with estimating rates of mtDNA sequence evolution. First, rates are generally calibrated using palaeontological or archaeological evidence of known species or population divergence times. Most of the studies shown in Table 7.1 use a single age constraint based on palaeontological evidence for the timing of the divergence between the human and chimp lineages (Hasegawa and Horai, 1991; Vigilant *et al*., 1991; Pesole *et al*., 1992; Ruvolo *et al*.,

1993; Horai *et al*., 1995; Ingman *et al*., 2000; Mishmar *et al*., 2003; Macaulay *et al*., 2005). As can be seen in Table 7.1, over the last 15 years the accepted human-chimp divergence time has increased from 4Mya (Hasegawa and Horai, 1991) to over 6Mya (Mishmar *et al*., 2003; Macaulay *et al*., 2005). For studies based solely on the Human-Chimp age constraint, this shift translates into a one and a half fold increase in estimated divergence times. Second, stochastic variation in lineage coalescence times mean that the time to MRCA of a clade may be older or younger than the age of the founding population that it is taken to represent (Templeton, 1993). And third, even when a divergence time is known, due to the stochastic nature of sequence evolution there is uncertainty in the inferred rate of change. Unfortunately, recent mtDNA studies have tended to ignore calibration and rate uncertainty (Horai *et al*., 1995; Watson *et al*., 1997; Ingman *et al*., 2000; Mishmar *et al*., 2003; Macaulay *et al*., 2005). Here, we incorporate phylogenetic uncertainty and rate uncertainty implicit in our evolutionary model by sampling rates across the MCMC rate distribution. We incorporate calibration uncertainty into our analyses by using a probability distribution to constrain the age of calibration nodes rather than a single point estimate (see Section 7.2 for constraint ranges).

Another weakness of previous mtDNA-based date estimates is their reliance on a single calibration point - only Cann *et al*.'s (1987) study used multiple rate calibrations. By making use of multiple, carefully chosen age constraints we can reduce the impact of any single biased age constraint on our results. Small errors in chosen calibration points can be expected to average out across constraints to provide an ultimately more accurate rate estimate. Using multiple age constraints also allows us to investigate evidence for rate variation across the tree. Whilst previous studies have been based on the assumption of a molecular clock, or constant rate of sequence evolution through time, recent work by Ho *et al*. (2005) has shown that the assumption of constant rates though time may be misleading. Ho *et al*. (2005) demonstrate that rates of change in a number of genes, including the mtDNA D-loop region, show a systematic time dependency. Observed rates of sequence evolution were shown to decrease with increasing time depth according to a translated exponential function, or "J-shaped" rate curve. As a result, using a calibration point much older than the time-scale on which rates are being inferred may cause dates to be grossly over-estimated. This is particularly problematic for those mtDNA studies

using rate calibrations based on a human-chimp species divergence time to infer much younger internal date estimates associated with the human expansion. Analyses based on internal archaeological age constraints of roughly the same time depth as the dates being estimated should be less prone to this effect. Only three of the studies in Table 7.1 used internal archaeological age constraints (Cann *et al*., 1987; Stoneking *et al*., 1992; and Watson *et al*., 1997). According to Ho *et al*.'s (2005) findings, these studies should provide the best estimates of rates at time depths similar to the constraint points. Unfortunately, all of these studies were based on RFLP or D-loop data. Here, we use two internal constraints derived from archaeological evidence and one constraint based on palaeontological evidence for the human-chimp split to estimate the age of mitochondrial Eve and investigate evidence for Ho *et al*.'s (2005) J-shaped rate curve.

7.1.3 OBJECTIVES OF THE CURRENT STUDY

By combining the advantages of a Bayesian, likelihood framework with complete sequence data and multiple calibration points we aim to overcome the problems associated with previous mtDNA analyses and provide a reliable confidence interval for the age of mitochondrial Eve. We quantify phylogenetic uncertainty and uncertainty in rates and investigate the robustness of these results to the choice of model and age constraints. We examine evidence for the J-shaped rate curve and investigate the possibility of estimating dates given rate variation through time using a relaxed clock assumption. We then use a coalescent approach to investigate human population growth since mitochondrial Eve. Finally, we discuss the implications of our results for our understanding of human evolution. This includes an examination of evidence for an association between human genetic and linguistic evolution.

7.2 MATERIALS AND METHODS

A data set of 252 complete human mtDNA sequences was constructed for the present study from sequences available on GenBank (see Table 7.2). These comprised sequences from Africa (53), Eurasia (93), Australia and Papua New Guinea (46), the Pacific (6), the New World (9) and South East Asia (45), including sequences from two recent studies of Aboriginal Malay (Macaulay *et al*., 2005) and Andaman Island

(Thangaraj *et al*., 2005) populations. We also included one bonobo (Acc. No. NC001644), one chimp (Acc. No. D38113) and one gorilla sequence (Acc. No. X93347). Data were machine aligned in *ClustalX* (Thompson *et al*., 1997) and then manually checked and adjusted for errors. Codon positions for coding regions were determined using reference sequence J01415.1.

TABLE 7.2: Sequence sources, year of publication, region sampled, number of sequences used from this data set in the current study and Genbank accession numbers.

Author	Year	Region	n	Genbank acc. numbers
Ingman *et al.*	2000	Global	53	AF346963 - AF347015
Maca-Meyer *et al.*	2001	Global	33	AF381981- AF382013
Ingman and Gyllensten	2003	Global; plus Australia/PNG	52	AY289051 - AY289102
Kong *et al.*	2003	China	3	AY255145, AY255176, AY255180
Maca-Meyer *et al.*	2003	U haplogroup	3	AY275527, AY275529, AY275536,
Mishmar *et al*.	2003	Global	31	AY195748-49, AY195753, AY195755-57, AY195759-63, AY195766, AY195770-73, AY195776-80, AY195782-92
Palanichamy *et al.*	2004	India – N only	11	AY713976-AY714050
Achilli *et al.*	2005	U haplogroup	4	AY882379-AY882417
Friedlaender *et al.*	2005	New Britain, New Ireland	3	AY956412-AY956414
Macaulay *et al.*	2005	Malaysia	15	AY963572-AY963586
Thangaraj *et al.*	2005	Andaman and Nicobar Islands	15	AY950286-AY950300
Trejaut *et al*.	2005	Taiwan – B4 haplotype only	8	AJ842744-AJ842751
Starikovskaya *et al.*	2005	Siberia	20	AY519484-AY519497,AY570524-AY570526, AY615359-AY615361
Bandelt *et al.*	2005	China	1	AY972053

Tree topology, divergence time and population estimation was conducted using Bayesian inference and MCMC sampling, as implemented in *BEAST* (*version 1.3*, Drummond *et al*., 2002; Drummond and Rambaut, 2003). Analyses were performed on the mtDNA coding region under a General Time Reversible (GTR) substitution model allowing gamma distributed rates across sites and a proportion of invariant sites (henceforth GTR + Γ + I). Although the GTR + Γ + I model is routinely used to allow for rate variation across sites, Shapiro *et al*. (2006) have shown that incorporating codon-position-specific rate-variation in a model provides a better fit to coding region data. We thus repeated the analyses incorporating codon-position-specific rates in a GTR substitution model (GTR + CP). For comparison, we also analysed just the mtDNA D-loop region under the GTR + Γ + I model. MCMC analyses were run for 20,000,000 generations. Tree topology and parameter estimates

were sampled every 2,000 generations after an 8,000,000-generation burn-in. Examination of the post-burn-in sample distribution indicated that the chains had reached convergence by this time. Effective sample sizes for all the parameters of interest were over 100.

Rates of sequence evolution were calibrated using three independent age constraints, chosen to reflect a range of time-scales. Two internal age constraints are labeled in Figure 7.2. The age of Constraint 1, comprising 11 highland and coastal Papua New Guinea sequences and one Nasoi sequence from the Bismarck Archipelago was constrained to 50,000 years BP (+/- 5,000 years S.E.) based on archaeological evidence for modern human arrival in Papua New Guinea at least as early as 40,000 years BP (Groube *et al*., 1986). This value provided the lower bound on the age of the clade, as genetic coalescent times were expected to slightly pre-date the initial colonization event (Stoneking *et al*., 1992). These dates are also consistent with evidence for the settlement of Australia (then connected to Papua New Guinea as part of the super-continent, Sahul), at some time between 40,000 and 60,000 years BP (Bowler *et al*., 2003). The age of Constraint 2, comprising 11 sequences of Malayo-Polynesians (non-Taiwanese Austronesian speakers) from Papua New Guinea and Remote Oceania, was constrained to 4,000 years BP (+/- 500 years S.E.), based on archaeological evidence for the expansion of Neolithic technology, agriculture and the Lapita pottery complex (Pawley, 2002) from Taiwan. A final age constraint was obtained by adding the three ape sequences to the analysis and constraining the age of the human-chimpanzee divergence to 6.5M years BP (+/- 250,000 years S.E.). Andrews (1992) agues for a human-chimpanzee divergence of at least 5Mya, and two recent studies of human mtDNA variation used a 6.5M year human-chimp calibration (Mishmar *et al*., 2003; Macaulay *et al*., 2005), based on independent genetic evidence for a 6M year human-chimp split (Goodman *et al*., 1998), plus 500,000 years for lineage coalescence.

In order to investigate the time dependency of rate estimates, analyses were performed using each age constraint separately under the assumption of strictly clock-like evolution. Analyses were also performed with all three age constraints imposed simultaneously. For the multiple-constraint analyses, the clock assumption was relaxed by allowing rates to vary between lineages, with the rate on each branch being

drawn independently and identically from a logNormal distribution (see Drummond *et al*., 2006).

The model comparisons and branch-length optimization reported in Table 7.6 were performed in PAUP* (Swofford, 2003), with the exception of the GY94 model (Goldman and Yang, 1994) which was implemented using *codeml* in PAML (Yang, 1997). For each model we analysed a sub-set of 100 sequences, selected at random from the initial data set, with the chimp sequence as an outgroup. The tree topology was fixed to a consensus tree from an MCMC analysis of the trimmed dataset based on the GTR + Γ + I model.

A Bayesian coalescent approach was used to estimate human population size through time and produce a Bayesian skyline plot, as described in Drummond *et al*. (2005) and implemented in *BEAST* (*version 1.3*, Drummond *et al*., 2002; Drummond and Rambaut, 2003). Given a set of contemporary gene sequences and some model of sequence evolution, coalescent methods can estimate the size of a population back through time based on the shape of the inferred tree/s. Population size estimates were inferred based on the GTR+CP substitution model using the Papua New Guinea age constraint and a strict clock assumption.

7.3 RESULTS AND DISCUSSION

7.3.1 DATING EVE AND TIME DEPENDENCY OF RATES

Inferred rates of mtDNA evolution, and hence estimates for the age of mitochondrial Eve, were found to depend upon the choice of age constraint, in a manner consistent with the J-shaped rate curve proposed by Ho *et al*. (2005). Table 7.3 reports the mean and 95% highest posterior density (HPD) intervals for the rate of mtDNA coding region sequence evolution, and the implied age of mitochondrial Eve, inferred under the GTR + Γ + I substitution model using each of the three age constraints. Table 7.4 reports the same statistics inferred using the GTR + CP substitution model. Under both substitution models, rates based on the Malayo-Polynesian age constraint are roughly 5 times faster than rates based on the other two constraints. Rate estimates based on the Papua New Guinea age constraint are also faster than those based on the

human-chimp constraint, although the 95% HPD intervals overlap and the difference is greater under the GTR + CP model (~50%) than under the GTR + Γ + I model (~15%). The same pattern is exaggerated when analyses are restricted to the D-loop region (see Table 7.5). Rates of D-loop evolution based on the Malayo-Polynesian age constraint are over 7 times faster than rates derived from the Papua New Guinea calibration, which are in turn over 3 times faster than rates derived from the human-chimp calibration. Estimated rates of substitution for the D-loop are 5-20 times faster than for the coding region.

TABLE 7.3: 95% HPD intervals for estimated substitution rate and age of mitochondrial Eve derived from the coding region using single constraints with a strict clock and GTR + Γ + I substitution model

	Calibration (kyrs)			**Est. rate /site/Myrs**			**Mitochondrial Eve (kyrs)**		
	lower	**mean**	**upper**	**lower**	**mean**	**upper**	**lower**	**mean**	**upper**
Malayo-Polynesian	3	4	5	0.039	0.071	0.103	22.4	39.0	60.7
PNG	40	50	60	0.010	0.015	0.020	116.6	183.1	249.5
Human-Chimp	6000	6500	7000	0.012	0.013	0.015	170.2	207.2	247.7

TABLE 7.4: 95% HPD intervals for estimated substitution rate and age of mitochondrial Eve derived from the coding region using single constraints with a strict clock and GTR + CP substitution model

	Calibration (kyrs)			**Est. rate /site/Myrs**			**Mitochondrial Eve (kyrs)**		
	lower	**mean**	**upper**	**lower**	**mean**	**upper**	**lower**	**mean**	**upper**
Malayo-Polynesian	3	4	5	0.045	0.076	0.112	21.8	37.7	57.4
PNG	40	50	60	0.010	0.015	0.020	121.4	183.5	252.3
Human-Chimp	6000	6500	7000	0.009	0.010	0.011	221.4	269.1	319.0

TABLE 7.5: 95% HPD intervals for estimated substitution rate and age of mitochondrial Eve derived from the D-loop using single constraints and GTR + Γ + I substitution model.

	Calibration (kyrs)			**Est. rate /site/Myrs**			**Mitochondrial Eve (kyrs)**		
	lower	**mean**	**upper**	**lower**	**mean**	**upper**	**lower**	**mean**	**upper**
Malayo-Polynesian	3	4	5	0.836	1.513	2.268	8.4	16.1	24.8
PNG	40	50	60	0.138	0.216	0.299	67.4	110.2	154.1
Human-Chimp	6000	6500	7000	0.048	0.068	0.091	228.0	355.1	479.3

Ho *et al*. (2005) discuss a number of factors that may account for the observed time dependency in rate estimates. They consider the effect of calibration error, purifying selection, mutational hotspots and saturation, and sequencing error. Here we discuss the implications of each of these for our findings, with the exception of sequencing error, which Ho *et al*. (2005) demonstrated was unlikely to lead to substantial rate variation through time.

7.3.1.1 Calibration Error

It is possible that the observed trend in rates is merely a spurious result based on incorrect age constraints. The inferred rates are sensitive to calibration error, hence, if any of the three constraints were significantly biased, this may create the appearance of rate variation where none exists. In the case of the Malayo-Polynesian constraint, this seems unlikely, given that the calibration time would have to be increased by a factor of 5 to make it consistent with the Papua New Guinea constraint. Calibration error is a more plausible explanation for the difference in rate estimates based on the Papua New Guinea and human-chimp calibrations. Although mean coding region rate estimates are slowest based on the human-chimp constraint, the 95% HPD intervals currently overlap with rate estimates based on the Papua New Guinea constraint, meaning that, taken on their own, the rate information derived from these deeper calibrations is not sufficient to rule out a common rate. However, there are reasons to believe that the difference in rates between the Papua New Guinea and human-chimp constraints is not merely due to calibration error. First, the trend in mean rate estimates is in the same direction as that reported by Ho *et al*. (2005), who used a range of different data sets and constraints. Second, rates based on our analyses of the D-loop show an exaggerated version of the same trend. 95% HPD intervals do not overlap for any of the D-loop analyses and the degree of difference is difficult to explain away as calibration error - this would require that the constrained Papua New Guinea clade had a 160,000 year coalescence time, predating any evidence for human settlement outside Africa, or alternatively that humans diverged from chimps only 2 million years ago. Thus, it is unlikely that the observed time dependency in rates is solely the product of inaccuracies in calibration times.

7.3.1.2 Purifying Selection

Both Ho *et al*. (2005) and Penny (2005) argue that purifying selection could at least partly account for time dependency in rate estimates. It is well known that rates of human mtDNA evolution based on transmission data from pedigree studies are significantly faster than rates derived from phylogenetic divergence time calibrations (Howell *et al.*, 1996; Howell *et al*., 2003). One explanation for this is that, whilst

pedigree study rate estimates represent the mutation rate between generations, or instantaneous mutation rate, phylogenetically derived rates generally reflect the rate at which some fraction of these mutations become fixed in the population, known as the substitution rate. Purifying selection filters out deleterious or slightly deleterious mutations from a population, such that, over long time scales, the observed substitution rate is considerably slower than the mutation rate (Penny, 2005). Random genetic drift will have a similar effect since only a fraction of the mutations that arise will actually survive to be observed in the study population. Our very fast estimates for the rate of change in the coding region derived using the Malayo-Polynesian age constraint (mean rates of 0.071 and 0.076 substitutions/site/Myrs under the GTR + Γ + I and GTR + CP models respectively) closely match mtDNA coding region mutation rates inferred from pedigree studies (0.075 substitutions/site/Myrs; Howell *et al*., 2003). In other words, our results suggest that over a timescale of a few thousand years, the observed rate of change is essentially equivalent to the mutation rate. However, over the longer time periods used in the analyses based on the Papua New Guinea and human-chimp age constraints, we may be observing the much slower *substitution* rate. The J-shaped rate curve may thus reflect a transition from the instantaneous mutation rate to a long-term substitution rate fettered by purifying selection. To investigate this possibility, we performed a maximum-likelihood analysis of our data set using the Goldman and Yang codon model (GY94; Goldman and Yang, 1994), with the tree topology fixed to the topology shown in Figure 7.2. Based on the number of synonymous versus non-synonymous substitutions (dN/dS = ω) at each site, we can estimate the proportion of sites that are under neutral (ω = 1) or purifying (ω < 1) selection. A maximum-likelihood reconstruction using this model implied that over 90% of the sites are likely to be under purifying selection (mean ω = 0.032). Using a similar approach, Ho *et al*. (2005) compared the strength of purifying selection in intraspecific and interspecific branches of a primate mtDNA phylogeny. They found evidence for increased purifying selection at interspecific branches. This suggests that the observed decrease in rates with increasing time depth could plausibly be explained by the influence of purifying selection. One problem with this explanation, however, is that the time dependency in rates is much more pronounced in the non-coding D-loop region, which is less likely to be under selection pressure.

7.3.1.3 Model misspecification and Mutational Saturation

Some form of model misspecification may contribute to the observed decline in rates. Heyer *et al*. (2001) suggest that the difference between rates derived from pedigree and phylogenetic studies may be due to the increased effect of rapidly evolving sites and mutational hotspots on phylogenetic rate estimates. This may be particularly important for rate estimates based on the human-chimp age constraint. Mutational saturation along the long branch connecting human and chimp lineages may cause rates along this branch to be underestimated and hence the age of mitochondrial Eve to be over-estimated. We would expect this effect to be most noticeable in the D-loop, due to the faster rates of change and large number of hyper-variable sites, and indeed this is consistent with our findings - rate estimates for the D-loop (Table 7.5) show considerably more variation between age constraints than for the coding region (Table 7.3). Also, estimates for the age of mitochondrial Eve derived from the coding region differ between the GTR + Γ + I and GTR + CP models when the human-chimp calibration is used, but do not differ when the two internal calibrations are used. This suggests that the effect of model choice on rate and date estimates warrants further investigation.

In order to determine the effect of model choice on rate and date estimates, we optimized branch-length estimates on a fixed tree topology under a range of commonly used substitution models. For analyses based on the human-chimp split, the ratio of the human-chimp branch-length to the depth of the human clade determines the estimated age of mitochondrial Eve – a larger ratio implies a younger age for the human clade. Accordingly, rate estimates are also dependent on the length of the human-chimp branch. We can thus determine the relative effect of the different models on substitution rate and divergence time estimates by measuring this ratio. Table 7.6 shows the inferred ratio of the human-chimp branch-length to the depth of the human clade, as well as log-likelihood and Akaike Information Criterion scores (AIC; Akaike, 1974) for the 13 substitution models tested. The AIC is a measure of how well a model fits the data, given the number of parameters in the model – lower scores indicate a better fit. Sets of nested models in the table are separated by horizontal lines. Clearly, increasing model complexity within each nested set gives an

improved fit to the data and a generally larger ratio of the human-chimp branch-length to the depth of the human clade. The highest ratios were produced by the site substitution models incorporating gamma distributed rates across sties (Γ) and a proportion of invariant sites (I). This is consistent with the hypothesis that multiple substitutions at hyper-variable sites may be causing simpler substitution models that assume constant rates across sites to underestimate the human-chimp branch length. However, the best-fitting models are those that take into account codon position and the GY94 codon model. These models all produce a smaller branch-length ratio than the "Γ + I" models. One explanation for this is that the "Γ + I" models are over-estimating the human-chimp branch-length. An alternative explanation is that the codon position models are failing to identify multiple substitutions despite their better fit to the data. This has important implications for model choice in studies of molecular rates, suggesting that the "best fit" model may not necessarily give the "best" branch length estimates. This could be explored further by adding some form of site-specific rate variation to the codon models. Unfortunately, site-specific rate variation cannot be implemented in conjunction with codon-specific rates in the software used here.

TABLE 7.6: Log-likelihood scores, Akaike Information Criterion (AIC) scores and ratio of internal human divergence to human-chimp divergence for a range of commonly used substitution models – JC = Jukes-Cantor (1969), HKY85 = Hasegawa, Kishino and Yano (1985), GTR = General Time Reversible, GY94 = Goldman and Yang (1994).

Model	Parameters	log L	AIC	B.L. Ratio
JC	**0**	-27905.56	55811.14	31.3
HKY85	**4**	-25799.23	51606.47	33.2
GTR	**8**	-25767.03	51550.06	33.3
JC + Γ	**1**	-27651.65	55305.30	40.4
HKY85 + Γ	**5**	-25537.13	51084.25	49.4
GTR + Γ	**9**	-25498.38	51014.76	53.8
JC + Γ + I	**2**	-27639.09	55282.18	41.7
HKY85 + Γ + I	**6**	-25524.31	51060.63	53.3
GTR + Γ + I	**10**	-25489.82	50999.64	55.5
JC + CP	**2**	-27180.67	54365.33	33.7
HKY85 + CP	**6**	-24972.32	49956.64	37.2
GTR + CP	**10**	-24938.29	49896.59	37.1
GY94 Codon	**11**	-24290.87	48603.75	42.7

7.3.1.4 Multiple Constraints and Relaxing the Clock Assumption

The fact that the rate estimates reported in Tables 7.3, 7.4 and 7.5 are dependent on the choice of calibration point suggests that the assumption of constant rates of change through time does not hold. Given this evidence for rate variation it is inappropriate to fit a model that assumes strictly clock-like rates using multiple age constraints. Furthermore, even without the constraint information, a likelihood ratio test allows us to reject the molecular clock assumption ($\chi^2 = 441.4$, df = 253, $p < 0.001$). In order to estimate divergence times in the face of rate variation we can relax the clock assumption. Table 7.7 reports rate and date estimates for the coding region analysed under both substitution models using all three age constraints and a relaxed clock model (Drummond *et al.*, 2006). Unsurprisingly, for both models, the estimated age of mitochondrial Eve falls between the coding region rate estimates derived using the single Papua New Guinea and human-chimp constraints. The degree of rate variation across the tree is measured by the coefficient of variance of the universal rate. The mean coefficient of variance is higher under the GTR + CP model (0.187) than the GTR + Γ + I model (0.133), consistent with the increased observed time dependency of rates under the GTR + CP model.

TABLE 7.7: 95% HPD intervals for estimated substitution rate and age of mitochondrial Eve derived from the coding region using the human-chimp and two internal age constraints under a relaxed clock assumption. 95% HPD intervals are also given for the coefficient of variance of the universal rate.

Model	**Universal mean rate for relaxed clock (per/site/Myrs)**			**Coefficient of variance in rate for relaxed clock**			**Mitochondrial Eve (kyrs)**		
	lower	**mean**	**upper**	**lower**	**mean**	**upper**	**lower**	**mean**	**upper**
GTR + Γ + I	0.012	0.015	0.018	0.003	0.149	0.255	141.2	183.4	232.9
GTR + CP	0.010	0.013	0.015	0.059	0.187	0.308	162.1	215.7	247.0

Unfortunately, whilst the relaxed-clock model has the effect of smoothing rate variation across the tree, it does not model a systematic, time dependent, rate transition. Thus, although it is more appropriate than assuming a strict-clock and may be a useful proxy for time dependent rate variation, ideally, a more principled approach to estimating divergence times under a J-shaped rate curve needs to be developed. Ho *et al.*, (2005) fit a translated exponential function to the observed rate curve and use this to infer rates through time. Using this approach it is possible to apply a correction to existing age estimates. However, without a spectrum of

calibration points to fit the curve to, there can be considerable error associated with this method of correction and, again, there is no principled reason why the rate transition should follow a translated exponential curve. Part of the challenge of future work in this area will be identifying the causes of the observed time dependency of rates and incorporating these processes into our analytical models.

7.3.1.5 Implications for the Age of Mitochondrial Eve

Until now, systematic time dependency in rate estimates has not been considered in studies of mtDNA evolution. However, our results suggest that this phenomenon may be crucial to understanding the chronology of human evolution. We see at least a five-fold variation in divergence time estimates depending on the choice of calibration point. In addition, the strength of the effect is mediated by the choice of model and the data. Much of the apparent discrepancy in previous mtDNA-based divergence time estimates may thus be attributable to a combination of the limitations of the data for phylogenetic analysis, inappropriate or overly simplistic models, and the choice of age constraints. This is particularly problematic for studies based on D-loop data, simple substitution models, the human-chimp rate calibration, or some combination of these. The effect of these factors observed here is corroborated by comparing age estimates from previous work. For example, consistent with our findings, those studies in Table 7.1 that made use of the human-chimp age constraint produced on average older estimates for the age of mitochondrial Eve than the analyses based on internal archaeological constraints (Cann *et al*., 1987; Stoneking *et al*., 1992; Watson *et al*., 1997).

If rates do vary according to a J-shaped rate curve, then the true age of the human maternal MRCA should fall somewhere between our estimates based on the human-chimp constraint and the Papua New Guinea constraint. This would place the age of mitochondrial Eve somewhere between 183,000 (95% HDP from 117,000 to 250,000) and 207,000 (95% HDP from 170 to 248,000) years ago based on the GTR + Γ + I substitution model, or between 184,000 (95% HDP from 121,000 to 252,000) and 269,000 (95% HDP from 221,000 to 319,000) years ago based on the GTR + CP model.

The analyses reported in Table 7.7 go some way to approximating rate variation through time and give an estimate for the age of mitochondrial Eve of 188,000 (95% HPD from 146,000 to 233,000) years BP based on the GTR + Γ + I substitution model, or 215,000 (95% HDP from 162,000 to 247,000) years ago based on the GTR + CP model. These 95% HPD intervals cover a wide range of approximately 100,000 years – from a minimum of 146,000 years BP under the GTR + Γ + I model to a maximum of 247,000 years BP under the GTR + CP model. Although this range is comparable to previous estimates, unlike previous work, we included uncertainty in calibration times, rate estimates, tree topology and branch-length estimates. We thus believe these estimates constitute a significant improvement over previous estimates and represent realistic upper and lower bounds for the age of mitochondrial Eve.

7.3.2 HUMAN DISPERSAL OUT OF AFRICA

The question of when and how humans left Africa has also recently received considerable attention (Maca-Meyer *et al*., 2001; Ingman and Gyllensten, 2003; Forster and Matsumura, 2005; Macaulay *et al*., 2005; Merriwether *et al*., 2005; Thangaraj *et al*., 2005; Mellars, 2006). However, previous attempts to date the human expansion from Africa using mtDNA data have suffered from the same limitations as studies dating mitochondrial Eve. Maca-Meyer *et al*. (2001) used complete mtDNA sequence data to estimate an age for the oldest non-African mtDNA lineages of between 59,000 and 69,000 years BP, but their estimate was based on genetic distance data, assumed a rate of evolution derived form the human-chimp divergence, and did not account for uncertainty in the rate estimate. Forster *et al*. (2001) estimated an age of 54,000 years BP (+/- 8,000 years S.E.), however, this was based on RFLP data and did not account for rate uncertainty.

It is possible to infer the timing and pattern of the human dispersal from Africa using the analyses described above. Under the GTR + CP substitution model, we estimated the age of the two major non-African lineages, the M-haplotype and N-haplotype lineages, to be 71,000 years (95% HPD between 50,000 and 97,000) and 76,000 years (95% HPD between 52,000 and 107,000) respectively. Owing to the evidence for rate variation through time, these age estimates were based on the Papua New Guinea age constraint, which is temporally closest to the expected date of the African exodus. The

GTR + Γ + I substitution model produced similar age estimates. Whilst it is unclear whether these lineages began to diverge just before or just after the expansion from Africa, the close agreement between the N- and M- haplotype divergence times and the almost complete absence of either haplotype in Africa[4], suggests that the African exodus occurred close to the estimated time of lineage coalescence.

As well as the timing of the African exodus, we can also draw inferences about the likely route taken by the first successful human expansion into Eurasia. Two routes have been proposed – a northern route, up the Nile and across the Sinai Peninsular, and a southern route from East Africa across the Red Sea and along the Indian Ocean coast (Forster and Matsumura, 2005). Recent studies of complete mtDNA sequence data from aboriginal populations in Malaysia (Macaulay *et al*., 2005) and the Andaman Islands (Thangaraj *et al*., 2005) indicate that these populations on the Indian Ocean Coast contain some of the oldest mtDNA lineages outside Africa. This supports an expansion along the southern route. We included the sequences of Macaulay *et al*. (2005) and Thangaraj *et al*. (2005) in our analyses. Consistent with their findings, the aboriginal Malay and Andaman Island sequences represented some of the oldest mtDNA lineages outside Africa in our data. The tree in Figure 7.2 shows that these sequences were among the first non-African sequences to diverge. The fact that the aboriginal Malay sequences occur at the base of both the M- and N- haplotype sub-trees, suggests that these haplotypes are both the result of the same expansion event. It is also interesting to note that our date estimates for the African exodus are consistent with the proposed timing of a southern exodus between 60,000 and 65,000 years ago, when the sea level was at a local minimum, which would have made crossing the Red Sea temporarily much easier (Forster, 2004).

7.3.3 MODELING THE HUMAN POPULATION EXPANSION

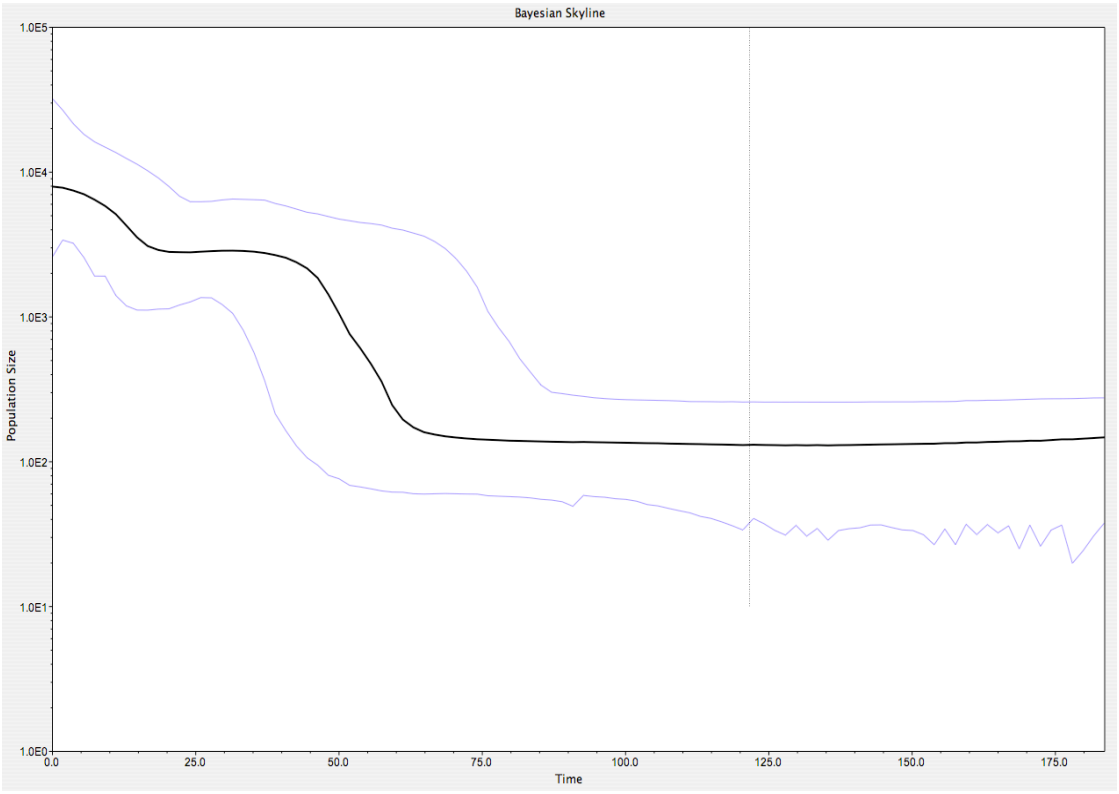
Figure 7.1a shows a Bayesian skyline plot of the inferred effective population size through time for an initial exploratory analysis of the coding region data using the GTR + CP substitution model and ten skyline plot time intervals. The effective population size is a variable commonly used in population genetics and considered to

[4] Whilst the M-haplotype is relatively common in Northern Africa, this is thought to be due to a reverse migration within the last 20,000 years (Forster, 2004).

be proportional to the actual population size. The black line represents the median inferred population size. The grey lines represent the upper and lower bounds of the 95% HPD region. Figure 7.1a indicates two distinct population growth phases in human history, separated by a long period of stasis. The two growth phases occur, first, between 30,000 and 80,000 years ago and second, within the last 25,000 years. There is considerable uncertainty associated with this reconstruction. This stems from uncertainty in the population size parameter estimates as well as uncertainty in the estimated rate of evolution. In an effort to try to reduce the effect of rate uncertainty and clarify the pattern of population expansion, the analysis was repeated with the rate fixed to the mean rate from the exploratory analysis. Factoring out rate uncertainty in this way means that the resulting error bars represent uncertainty implicit in our population demographic model. Figure 1b shows the result of this more refined analysis.

Clearly, much of the uncertainty in the timing of the two population expansions results from uncertainty in the rate of change. As Figure 1b shows, when rates are fixed, the 95% HPD region is substantially reduced and we see strong support for a two-phase population expansion. The first major expansion, when human population size is shown increasing by at least an order of magnitude, is dated to between 55,000 and 65,000 years BP. The absolute date range for this event is of course contingent upon the assumed mean rate. As we have shown above, there is uncertainty associated with the inferred rate estimates and thus the absolute timescale should be interpreted with caution. However, the timing of the expansion event relative to the tree divergence times is not contingent on the assumed rate, since the inferred timing of both will scale in proportion to the rate. We can thus be confident that the initial expansion occurred shortly after the emergence of the first non-African mtDNA lineages (or roughly 70,000 years ago based on the mean rate used here – see also Figure 7.2). This fits with a scenario of rapid population growth as humans expanded along the Indian Ocean coast and into Australia and Papua New Guinea. The initial major expansion appears to have been followed by a long period of very little population growth or possibly even decline. If we believe the inferred time-scale, this corresponds to a period of decreasing global temperatures and increasing glaciation that may have restricted human population growth, particularly in temperate and inland regions.

a)



b)

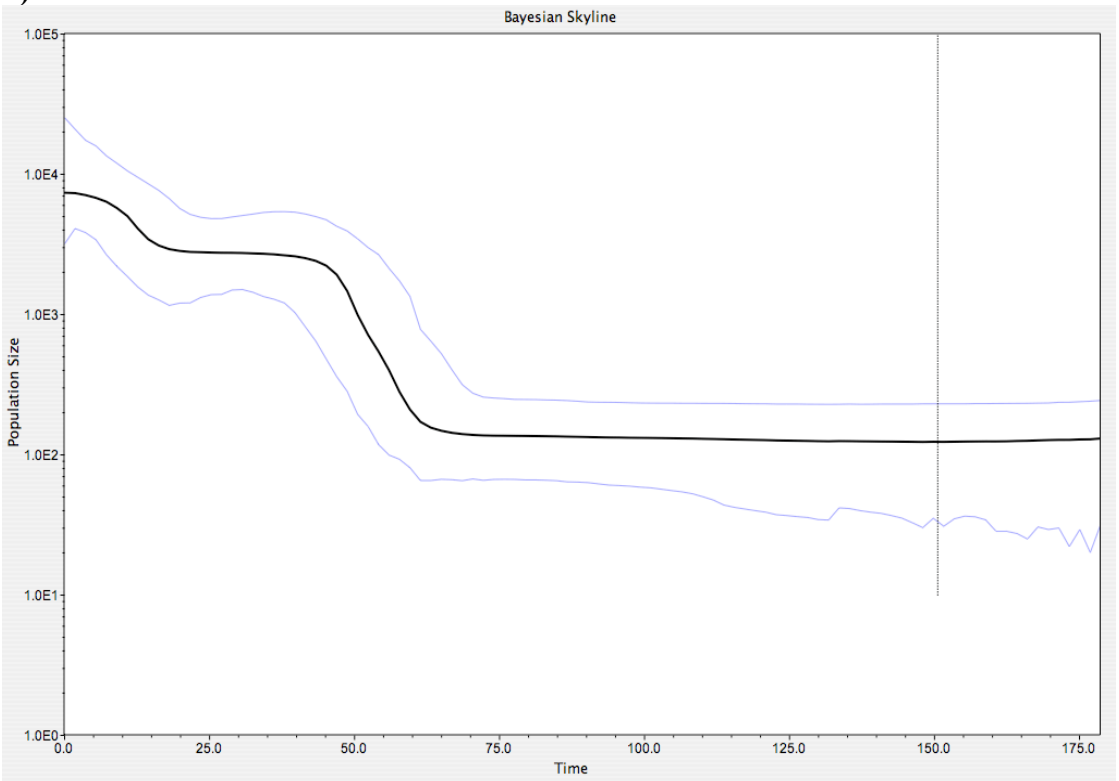


FIGURE 7.1: Bayesian skyline plot (Drummond *et al*., 2005) of effective population size **a** estimated from the coding region analysed under the GTR + CP substitution model using the Papua New Guinea age constraint, and **b** as for **a** but with rates fixed to the mean rate in **a**. The dotted line represents the 95% HPD lower bound for the age of mitochondrial Eve. The black line represents the median inferred population size. The grey lines represent the upper and lower bounds of the 95% HPD.

The second population growth phase is reconstructed beginning between 10,000 and 15,000 years BP. This is contemporaneous with the beginning of the Holocene, marking the end of the last ice age between 12,000 and 15,000 years BP, when rising temperatures would have made vast areas of Eurasia and the New World relatively hospitable for early human colonization (Forster, 2004). The second growth phase may thus be linked to this global increase in temperature. Population growth is also likely to have occurred with the spread of agricultural technology beginning around 11,000 years BP. Diamond and Bellwood (2003) argue that the expansion of agriculture was a major driving force in human evolution during the Holocene. They link the agricultural expansion with the spread of many of the world's language families. The demic diffusion model of farming dispersal (Ammerman and Cavalli-Sforza, 1973; Cavalli-Sforza, Menozzi and Piazza, 1994) holds that the expansion of agriculture and language were coupled with population dispersal. If this were the case, then we would expect some correspondence between linguistic and genetic variation. We examine this possibility further in Section 7.3.4, below.

7.3.3.1 Some Caveats

The coalescent model employed here to estimate population size assumes that taxa are randomly sampled from a single population (Drummond *et al.*, 2005). Clearly, this is not the case here. However, coalescent-based population estimates have been shown to be robust to violations of this assumption (Shapiro *et al.*, 2004). A more detailed investigation of these results will involve analysis of subsets of the mtDNA sequences, based on both geographic region and haplotype. This will allow us to determine the sensitivity of our results to sampling bias, as well as providing more fine-grained population demographic information. For example, Klein (1992) argues on the basis of archaeological evidence for a decline in population size in Africa between 20,000 and 40,000BP. It is possible to test this scenario by analyzing a sub-set of African sequences. A similar approach could be used to infer past population dynamics for Europe, the New World and the Pacific.

7.3.4 TOWARDS GENETIC AND LINGUISTIC CONSILIENCE

Templeton (1993) criticized early mtDNA research into human origins for being over-zealous in the interpretation of their findings, grossly underestimating, or choosing to ignore, uncertainty in their results. Penny *et al*. (1995, p. 867) stress the importance of constructing a "consilience of induction" (Whewell, 1967; p. 469), integrating evidence from multiple genes, as well as evidence from beyond biology. The above results need to be synthesized with evidence from other genes. Ruvolo's (1996) review of date estimates from mtDNA, microsatelite, protein polymorphism and Y-chromosome data is a paradigm example here. However, a review of the human genetic data is beyond the scope of this thesis. Instead, in keeping with our focus on the parallels between linguistic and biological evolution, we look for congruence between genetic and linguistic variation.

In 1988, Cavalli-Sforza *et al*. made a comparison between a human genetic and linguistic tree and, on the basis of tree similarity, argued for substantial linguistic and genetic co-evolution. This work highlighted the similarities between processes of historical inference in biology and linguistics, as well as the potential importance of linguistic data for inferences about human population history. Although the degree of similarity between the genetic population and linguistic trees was later challenged (O'Grady *et al*., 1989; Bateman *et al*., 1990), Penny, Watson and Steel (1993) were able to show that the two trees presented by Cavalli-Sforza *et al*. (1988) were more congruent than would be expected by chance. Unfortunately, linguists do not accept the language tree presented in the Cavalli-Sforza *et al*. paper, arguing that it is not possible to verify language relationships older than about 10,000 years (Nichols, 1992). In addition, Cavalli-Sforza *et al*. (1988) compared genetic *population* trees with language trees. One criticism of their approach was that their "populations" were defined either arbitrarily or on the basis of language affiliation, making it difficult to interpret the significance of any association with linguistic variation (O'Grady *et al*., 1989; Bateman *et al*., 1990).

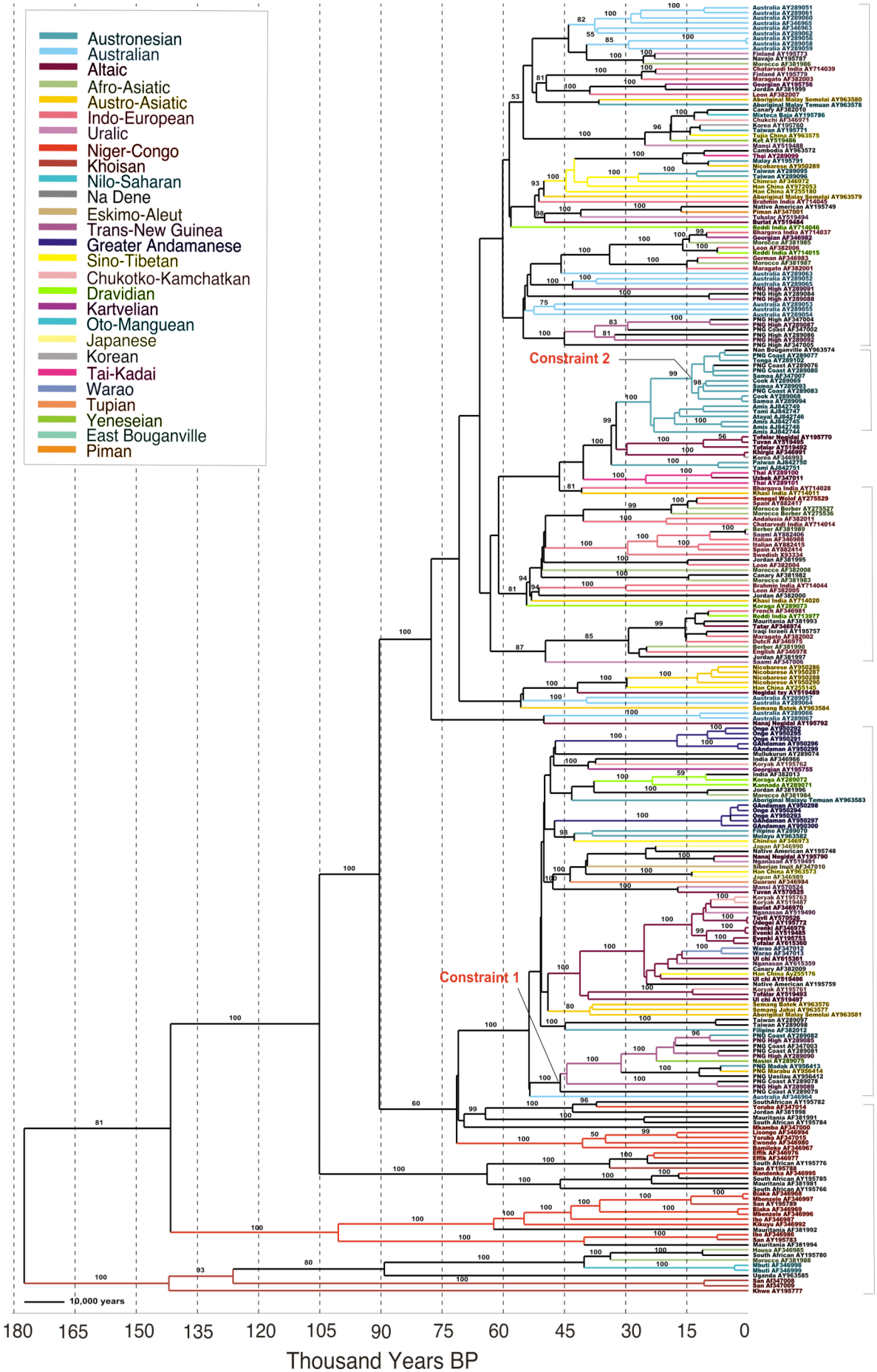

Austronesian
Australian
Altaic
Afro-Asiatic
Austro-Asiatic
Indo-European
Uralic
Niger-Congo
Khoisan
Nilo-Saharan
Na Dene
Eskimo-Aleut
Trans-New Guinea
Greater Andamanese
Sino-Tibetan
Chukotko-Kamchatkan
Dravidian
Kartvelian
Oto-Manguean
Japanese
Korean
Tai-Kadai
Warao
Tupian
Yeneseian
East Bouganville
Piman
Constraint 2
Constraint 1
8 Indo-European/1 Altaic
Austronesian
15 Indo-European/1 Altaic
0 Indo-European/13 Altaic
African
10,000 years
180
165
150
135
120
105
90
75
60
45
30
15
0
Thousand Years BP

FIGURE 7.2: Majority-rule consensus tree of a Bayesian MCMC sample distribution from the mtDNA coding region analysis using the Papua New Guinea age constraint and assuming a strict clock and GTR + CP substitution model. Branch lengths are drawn proportional to time - the scale is shown at the base of the figure. Clade posterior probability values above 50% are shown in black above the branch immediately ancestral to the clade. Branches are coloured based on the language family of the language spoken by the sequenced individual (key at top left). The clades representing the two internal human age constraints are labeled in red. Other clades showing interesting linguistic patterns are also labeled in the right margin.

One way of investigating genetic and linguistic co-evolution without requiring a global language tree or pre-defined populations is to map linguistic groups onto a gene tree. Figure 7.2 shows language families mapped onto a consensus tree from the mtDNA coding region analysis using the Papua New Guinea age constraint and GTR + CP substitution model. It is important to note that the tree in Figure 7.2 is a gene tree representing the ancestral relationships between individuals' mtDNA within the human population – i.e. not a population tree, representing the ancestral relationships between populations. Even if genes and languages have co-evolved, we would not expect linguistic characters to map perfectly (without "homoplasy") onto the gene tree because a single genetic population can be represented by multiple lineages in the mtDNA gene tree. Nonetheless, Figure 7.2 does highlight a number of interesting patterns that warrant further investigation. At the base of the tree is a clade of so-called Khoisan "click" languages (coloured brown in Figure 7.2) dating to over 100,000 years BP. This is consistent with very high Y-chromosome diversity among Khoisan speakers (Semino *et al.*, 2002) and has been used to argue that the Khoisan languages are an extremely ancient language group (Knight *et al*., 2003). Knight *et al,* (2003) propose that some of the features of these languages, such as the "click" consonants, could have been retained since the earliest human divergence. Alternatively, these languages, or some of the features they share, may have diffused much later into existing populations that happened to include ancient mitochondrial lineages. The issue is further complicated because it has not been clearly established that the Khoisan languages are in fact a monophyletic group (Campbell, 2004). By applying the phylogenetic techniques outlined in the previous chapters to Khoisan language data we may be able to shed light on the relationship between these languages and their fit with the genetic data. It should be possible to estimate the age of the proposed Khoisan language family and test whether these languages are in fact as old as the mtDNA lineages suggest. If the family is much younger, this would

support the notion that the languages, or the features they share, dispersed more recently across existing mtDNA lineages.

Another interesting feature is the striking association between the Austronesian language family and the B4a1a haplotype (highlighted in Figure 7.2). All of the B4ala sequences are from Austronesian speakers. Taiwanese sequences occur at the base of the B4a1a clade, consistent with linguistic and archaeological evidence for an Austronesian homeland in Taiwan (Pawley, 2002). This clade (excluding the Taiwanese sequences) was used for the Malayo-Polynesian age constraint of 4,000 years (+/- 500 years S.E.) based on evidence for an expansion of agriculture and the Lapita pottery complex at this time (Pawley, 2002; see Section 7.2). Clearly, there is a discrepancy between the estimated age of the Malayo-Polynesian clade based on the Papua New Guinea age constraint (~14,000 years BP in figure 7.2; 95% HPD of between 9,800 and 27,900 years BP) and the 4,000 year estimate based on archaeological data. This discrepancy is widely recognized. In a recent analysis of Austronesian mtDNA data, Trejaut *et al*. (2005) point out that their date estimate for the Malayo-Polynesian clade does not fit with the archaeological evidence but they do not investigate this further. If molecular rates of change shift from the instantaneous mutation rate over very short time-scales to a slower long-term substitution rate, then previously published mtDNA-based date estimates such as Trejaut *et al*.'s may simply be too early. As we observed in Section 7.3.1, the instantaneous mutation rate is precisely consistent with a 4,000-year age for Austronesian. This suggests that divergence time estimates for this and other recent Holocene expansions, including the colonization of the Americas (Forster *et al*., 1996; Saillard *et al*., 2000) and perhaps Europe (Maca-Meyer *et al*., 2001), may need to be re-examined.

Finally, we consider the distribution of two Eurasian language families across the tree – Indo-European, spoken throughout Europe and the near East, and Altaic, spoken in Russia and central Asia. Indo-European and Altaic language families have been grouped as part of a controversial linguistic "superfamily" known as Nostratic (Dolgopolsky, 1987; Kaiser and Shevoroshkin, 1988). Three major clades on the mtDNA phylogeny (highlighted in Figure 7.2), dated between 45,000 and 60,000 years BP, contain converse distributions of Indo-European and Altaic speakers – two contain a relatively large number of Indo-European sequences and almost no Altaic

speakers, whilst the other contains a large number of Altaic speakers and no Indo-European speakers. Although the sample sizes are relatively small, comparing the observed distribution of Altaic and Indo-European speakers on the tree with 100 random distributions indicated that these languages were clustered non-randomly on the tree (i.e. the number of implied changes of language affiliation was significantly less in the observed distribution than the random distributions; $p<0.001$). Whilst few linguists would accept that Altaic, Indo-European, or even the proposed Nostratic language families are as old as 45,000 to 60,000 years, the age of the observed genetic clades could significantly predate the age of any major Eurasian population divergence and expansion. This means the observed patterns could still be the result of language/genes co-evolution. One possible method for investigating this more formally is to use a structured coalescent approach to model populations through time (Notohara, 1990), treating language families as a proxy for human population structure. For example, Eurasian sequences can be subdivided into Indo-European and Altaic population groups and can be further subdivided within each language family. These language families may represent a better prior on long-term human population structure than the broad geographic groupings used previously (e.g. Cavalli-Sforza *et al*. 1988; Cavalli-Sforza, Menozzi and Piazza, 1994). This approach also has the potential to provide information on the age and population size of language family groups, as well as migration rates between families.

7.4 CONCLUSION

The perils of dating Eve and reconstructing human prehistory are many, but these problems are not insurmountable. Newly available data and methods allow us to overcome some of the limitations of previous work and improve our understanding of human origins and dispersal. The Bayesian framework employed here is particularly well-suited for quantifying uncertainty in results and testing methodological assumptions. Based on this approach, we have been able to address a number of important questions. First, using multiple calibration points we found evidence for time dependency in mtDNA rate estimates. We considered a number of possible explanations for this. Further analyses suggested purifying selection or mutational saturation as likely explanations. Second, by relaxing the clock assumption, we were able to estimate the age of mitochondrial Eve on the basis of multiple rate calibrations

without assuming strictly clock-like rates of molecular evolution. This supported a human origin between 150,000 and 250,000 years BP. Third, the inferred age and branching pattern of the oldest non-African mtDNA lineages add support to a scenario of human expansion from Africa along the Indian Ocean coast as far as Papua New Guinea and Australia. Fourth, we were able to explicitly model human population size through time. This revealed a previously undiscovered two-phase human population expansion. A period of rapid population growth is inferred occurring shortly after the human expansion from Africa. A second population expansion is also inferred, which may have been linked to the end of the last glaciation or the spread of agriculture. Finally, we identified a number of interesting associations between mtDNA and linguistic diversity that warrant further investigation.

Only by combining genetic, archaeological and linguistic evidence can we begin to build a complete picture of human origins and dispersal. The results presented here are an important part of this picture, but they are far from the last word. The exponentially increasing volume of genetic data and rapidly advancing methods of analysis, as well as the potential to combine genetic with linguistic evidence, will no doubt ensure an exciting future for reconstructing our past.

7.5 REFERENCES

Achilli, A., Rengo, C., Battaglia, V., Pala, M., Olivieri, A., Fornarino, S., Magri, C., Scozzari, R., Babudri, N., Santachiara-Benerecetti, A. S., Bandelt, H-J., Semino, O. and Torroni, A. 2005. Saami and berbers--an unexpected mitochondrial DNA link. *American Journal of Human Genetics.* 76(5):883-886.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control AC* 19:716–723.

Ammerman, A. J. and Cavalli-Sforza, L. L. 1973. A population model for the diffusion of early farming in Europe. Pages 343-358 in *The Explanation of Cultural Change: Models in Prehistory* (ed.) C. Renfrew. Duckworth, London.

Andrews, P. (1992) Evolution and environment in the Hominoidea. *Nature* 360:641-646.

Bandelt, H. J., Achilli, A., Kong, Q. P., Salas, A., Lutz-Bonengel, S., Sun, C., Zhang, Y. P., Torroni, A. and Yao, Y. G. 2005. Low 'penetrance' of phylogenetic

knowledge in mitochondrial disease studies. *Biochem. Biophys. Res. Commun.* 333(1), 122-130.

Bateman, R., I. Goddard, R. O'Grady, V. Funk, R. Mooi, W. Kress, and P. Cannell. 1990. Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* 31:1–24.

Bowler, J.M., Johnston, H., Olley, J.M., Prescott, J.R., Roberts, R.G., Shawcross, W., and Spooner, N.A. 2003. New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature* 421: 837–840.

Campbell, L. 2004. *Historical linguistics: An introduction. 2nd edition.* Edinburgh University Press, Edinburgh.

Cann, R. L., Stoneking, M. and Wilson, A. C. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza. A. 1994. *The History and Geography of Human Genes.* Princeton University Press, Princeton (NJ).

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P., and Mountain, J. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences USA* 85:6002–6006.

Diamond, J., and P. Bellwood. 2003. Farmers and their languages: The first expansions. *Science*. 300:597.

Dolgopolsky, A. B. 1987. The Indo-European homeland and lexical contacts of Proto-Indo-European with other languages. *Mediterranean Archaeological Review* 3:7-31.

Dorit, R. L., Akashi, H., and Gilbert, W. 1995. Absence of polymorphism at ZFY locus on the human Y chromosome. *Science* 268:1183-1186.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol* 4(5): e88.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon. W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.

Drummond, A. J., and Rambaut. A. 2003. *BEAST. Version 1.3.* Oxford University Press, Oxford. http://evolve.zoc.ox.ac.uk/ beast/

Drummond, A. J., Rambaut, A., Shapiro, B. and Pybus, O.G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 22, 1185-1192.

Elson, J. L., Andrews, R. M., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., Howell, N. 2001. Analysis of European mtDNAs for recombination. *American Journal of Human Genetics.* 68(1):145-153.

Excoffier, L., and Schneider, S. 1999. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proceedings of the National Academy of Sciences USA* 96(19):10597-10602.

Forster, P. 2004. Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Phil. Trans. R. Soc. Lond. B*. 359, 255-264.

Forster, P., Harding, R., Torroni, A., and Bandelt, H.-J. 1996 Origin and evolution of Native American mtDNA variation: A reappraisal. *American Journal of Human Genetics* 59: 935–945.

Forster, P., and Matsumura, S. 2005. Did Early Humans Go North or South? *Science* 308:965

Forster, P., Torroni, A., Renfrew, C., and Röhl, A. 2001. Phylogenetic Star Contraction Applied to Asian and Papuan mtDNA Evolution. *Molecular Biology and Evolution* 18(10):1864–1881.

Friedlaender, J., Schurr, T., Gentz, F., Koki, G., Friedlaender, F., Horvat, G., Babb, P., Cerchio, S., Kaestle, F., Schanfield, M., Deka, R., Yanagihara, R. and Merriwether, D.A. 2005. Expanding Southwest Pacific Mitochondrial Haplogroups P and Q. *Molecular Biology and Evolution* 22(6), 1506-1517.

Goldman, N., and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725–736.

Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J., Gunnell, G. and Groves, C. P. 1998. *Molecular Phylogenetics and Evolution* 9:585–598.

Gunnell,G., and Groves, C. P. 1998. Toward a Phylogenetic Classification of Primates Based on DNA Evidence Complemented by Fossil Evidence. *Molecular Phylogenetics and Evolution*. 9(3):585–598

Groube, L. M., Chappell, J., Muke, J., and Price, D. 1986. A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. *Nature.* 324:453-455.

Hasegawa, M. and Horai, S. 1991. Time of the Deepest Root for Polymorphism in Human Mitochondrial DNA. *Journal of Molecular Evolution* 32:37-42.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160–174.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., Labuda, D. 2001. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *American Journal of Human Genetics* 69:1113–1126.

Ho, S. Y. W., Phillips, M. J., Cooper, A., and Drummond, A. J. 2005. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. *Molecular Biology and Evolution* 22(7):1561-1568.

Horai, S., Hayasaka, K., Kondon, R., Tsugane, K., and Takahata, N. 1995. *Proceedings of the National Academy of Sciences USA* 92:532-536.

Howell, N., Kubacka, I., and Mackey, D. A. 1996. How rapidly does the human mitochondrial genome evolve? *American Journal of Human Genetics* 59:501-509.

Howell, N., Smejkal, C. B., Mackey, D. A., Chinnery, P. F., Turnbull, D. M., and Herrnstadt. C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *American Journal of Human Genetics* 72:659–670.

Ingman, M., and Gyllensten, U. 2003. Mitochondrial Genome Variation and Evolutionary History of Australian and New Guinean Aborigines. *Genome Research* 13:1600–1606.

Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*. 408:708-713.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. Pages 21-132 in *Mammalian Protein Metabolism, Vol. 3*, (ed.) M.N. Munro. Academic Press, New York.

Kaiser, M. and Shevoroshkin, V. 1988. Nostratic. *Annual Review of Anthropology* 17: 309-329.

Klein, R. G. 1992. The archaeology of modern human origins. *Evolutionary Anthropology* 1:5-14.

Kong, Q-P., Yao, Y-G., Sun, C., Bandelt, H-J., Zhu, C.-L. and Zhang, Y-P. 2003. Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *American Journal of Human Genetics.* 73 (3), 671-676.

Knight, A., Underhill, P. A. Mortensen, H. M., Zhivotovsky, L. A., Lin, A. A. Henn, B. M., Louis, D., Ruhlen, M., and Mountain, J. L. 2003. African Y Chromosome and mtDNA Divergence Provides Insight into the History of Click Languages. *Current Biology* 13:464–473.

Kuhner, M. K., and Felsenstein, J., 1994. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates, *Molecular Biology and Evolution* 11, 459–468.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N. K., Raja, J. M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H-J., Oppenheimer, S., Torroni, A., and Richards, M. 2005. Single, Rapid Coastal Settlement of Asia

Revealed by Analysis of Complete Mitochondrial Genomes. *Science.* 308:1034-1036.

Maca-Meyer,N., Gonzalez,A.M., Larruga,J.M., Flores,C. and Cabrera,V.M. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet.* 2(1):13.

Maca-Meyer, N., González, A. M., Pestano, J., Flores, C., Larruga, J. M., Cabrera, V. C. 2003. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genetics* 4:15.

Mellars,P. 2006. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* 439:931-5.

Merriwether, D. A., Hodgson, J. A., Friedlaender, F. R., Allaby, R., Cerchio, S., Koki, G., and Friedlaender, S. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proceedings of the National Academy of Sciences* 102(37): 13034-13039.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., 1953. Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics* 21, 1087–1091.

Meyer, S., Weiss, G., and von Haeseler, A. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* 152(3):1103-1110.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clarke, A. G., Hosseinib, S., Brandona, M., Easleyf, K., Cheng, E., Brown, M. D., Sukerniki, R. I., Olckersj, A., and Wallace, D. C. 2003. Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences USA* 100:171–176.

Nichols, J. 1992. *Linguistic diversity in space and time*. University of Chicago Press, Chicago.

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* 29: 59-75.

O'Grady, R., I. Goddard, R. M. Bateman, W. A. DiMicheal, V. A. Funk, W. J. Kress, R. Mooi, and P. F. Cannell. 1989. Genes and tongues. *Science* 243:1651.

Pawley, A. 2002. The Austronesian Dispersal: Languages, Technologies and People. Pages 251-273 in *Examining the farming/language dispersal hypothesis.* (eds.) P. Bellwood and C. Renfrew. McDonald Institute for Archaeological Research, Cambridge.

Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H-J., Kong, Q. P., Khan, F., Wang, C. Y., Chaudhuri, T. K., Palla, V. and Zhang, Y. P. 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing:

implications for the peopling of South Asia. *American Journal of Human Genetics* 75(6):966-978.

Penny, D. 2005. Relativity for molecular clocks. *Nature* 436:183184.

Penny, D., Steel, M., Waddell, P. J. and Hendy, M. D. 1995. Improved Analyses of Human mtDNA Sequences Support a Recent African Origin for Homo sapiens. *Molecular Biology and Evolution* 12(5):863-882.

Penny, D., Watson, E., and Steel, M. 1993. Trees from languages and genes are very similar. *Systematic Biology* 42:382–384.

Pesole, G., Sbisá, E., Preparata, G., and Saccone, C. 1992. The Evolution of the Mitochondrial D-Loop Region and the Origin of Modern Man. *Molecular and Biology Evolution* 9(4):587-598.

Rogers, A. 1995. Genetic evidence for a Pleistocene population explosion. *Evolution* 49(4):608-615.

Ruvolo, M., Zehr, S., von Dornum, M., Pan, D., Chang, B. and Lin, J. 1993. Mitochondrial COII Sequences and Modern Human Origins. *Molecular and Biology Evolution* 10(6):1115-1135.

Ruvolo, M. 1996. A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. *Molecular Phylogenetics and Evolution* 5:202 – 219.

Saillard, J., Forster, P., Lynnerup, N., Bandelt, H.-J. and Nørby, S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *American Journal of Human Genetics* 67:718–726.

Semino, O., Santachiara-Benerecetti, A. S., Falaschi, F., Cavalli-Sforza, L. L., and Underhill, P. A. 2002. Ethiopians and Khoisan Share the Deepest Clades of the Human Y-Chromosome Phylogeny. *Am. J. Hum. Genet*. 70:265–268.

Shapiro et al. 2004. Rise and Fall of the Beringian Steppe Bison. *Science* 306:1561-1565.

Shapiro, B., Rambaut, A., Drummond, A. J. 2006. Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Molecular Biology and Evolution* 23(1):7–9.

Sherry, S. T., Rogers, A. R., Harpending, H., Soodyall, H., Jenkins, T., and Stoneking, M. 1994. Mismatch Distributions of mtDNA Reveal Recent Human Population Expansions. *Human Biology*. 66(5):761-775.

Starikovskaya, E. B., Sukernik, R. I., Derbeneva, O. A., Volodko, N. V., Ruiz-Pesini, E., Torroni, A., Brown, M. D., Lott, M. T., Hosseini, S. H., Huoponen, K. and Wallace, D. C. Mitochondrial DNA diversity in indigenous populations of the

southern extent of Siberia, and the origins of Native American haplogroups. *Ann. Hum. Genet*. 69(1):67-89.

Steel, M., Hendy, M., and Penny, D. 1988. Loss of information in genetic distances. *Nature* 333, 494-495.

Stoneking, M., Sherry, S. T., Redd, A. J., and Vigilant, L. 1992. New approaches to dating suggest a recent age for the human mtDNA ancestor. *Phil. Trans. R. Soc. Lond. B*. 337:167-175.

Stringer, C. B. and Andrews, P. 1988. Genetic and Fossil Evidence for the Origin of Modern Humans. *Science* 239:1263-8.

Swofford, D. L. 2003. *PAUP*: phylogenetic analysis using parsimony (*and other methods)*. Sinauer Associates, Sunderland (MA).

Templeton, A. R., 1993. The "Eve" Hypothesis: A genetic critique and reanalysis. *American Anthropologist* 95(1):51-72.

Thangaraj, K., Chaubey, G., Kivisild, T., Reddy, A. G., Singh, V. K., Rasalkar, A. A., and Singh, L. 2005. Reconstructing the Origin of Andaman Islanders. *Science* 308:996.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 24:4876-4882.

Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J., Li, Z. Y., and Lin, M. 2005. Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. *PLoS Biology* 3(8):e247.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. and Wilson, A. C. 1991. African Populations and the Evolution of Human Mitochondrial DNA. *Science*. 253:1503-1507.

Watson, E., Forster, P., Richards, M., and Bandelt, H.-J. 1997. Mitochondrial Footprints of Human Expansions in Africa. *American Journal of Human Genetics*. 61:691-704.

Whewell, W. 1967. *The philosophy of the inductive sciences: founded upon their history. Vol. 2.* 2d ed. Cass, London. (Reprint of 1847 original.)

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13:555–556.

Chapter Eight

CONCLUSION

> "All perception of truth is the detection of an analogy" (p. 260)
> - Henry David Thoreau (1962)

Science frequently elucidates by analogy. Here, I have argued that the analogy between species and language evolution is more than just a curiosity. Parallels between the processes of biological and linguistic evolution mean that, by treating words and languages in much the same way as biologists treat genes and species, computational phylogenetic methods from biology can be productively applied to study language evolution.

The strategy of model-based phylogenetic inference employed by biologists is built around a powerful set of analytical tools. These tools can be adapted to the study of language evolution to provide an important compliment to more conventional methods in historical linguistics. By explicitly modeling the process of language change it is possible to objectively and efficiently evaluate language trees using large lexical data sets. Network-based phylogenetic methods also make it possible to identify evidence of non-tree-like language evolution indicating borrowing between languages. Combining models of lexical evolution with Bayesian inference of phylogeny and rate smoothing algorithms allows language trees and divergence times to be inferred without the flawed assumptions of lexicostatistics and glottochronology. By quantifying uncertainty in the resulting inferences, it is possible to use these methods to test between competing historical hypotheses.

Whilst analogies can be extremely helpful, they can also be misleading. As a result, testing the validity of the methods has been a fundamental focus of this thesis. The well-studied Indo-European language family was used to demonstrate the external

validity of our methodology, showing that inferred language trees and divergence times were consistent with established language relationships and a number of historically attested divergence dates. The statistical framework used here also makes it possible to systematically test the robustness of results to violations in the assumptions of the method. Results were shown to be robust to a range of model priors, data coding methods, and age constraints. Perhaps most impressively, Chapter 5 demonstrated that the method produced consistent results across two very different Indo-European datasets using two different models of language evolution. Chapter 5 also used synthetic data to address concerns that had been raised about model misspecification. The models were found to be robust to key criticisms including borrowing and character non-independence. The result of this process of model validation is a suite of analytical tools that can be confidently used to answer long-standing questions in historical linguistics.

Exploring the analogy between biological and linguistic evolution has provided some important insights into human prehistory. In Chapters 3, 4 and 5 lexical data was used to determine the age of the Indo-European language family. The resulting divergence time estimates were consistent with an expansion of Indo-European from Anatolia with the spread of agriculture, challenging the traditional interpretation of Indo-European origin. In Chapter 6, the same methods were applied in a new setting to re-evaluate Mayan language prehistory. This highlighted interesting uncertainties in Mayan language relationships and indicated that the family may be older than previously thought. Finally, in Chapter 7 the approach to tree-building and model validation that proved so useful for language analysis was applied to mitochondrial DNA sequence data to re-evaluate the history of the human maternal lineage. The results of this work have important implications for theories of human origins and dispersal, revealing time dependency in rates of mitochondrial DNA evolution and a previously unreported two-phase human population expansion.

As was discussed in Chapter 2, there are some exciting possibilities for future work combining evolutionary biology and historical linguistics. For example, the problem of comparing multiple words between languages to make cognacy judgments is analogous to the sequence alignment problem in biology. Hence, we may be able to determine the probability that lexical characters are cognate by employing algorithms

similar to those used by biologists for sequence alignment (Kondrak, 2001; Heeringa et al. 2000; Covington, 1996). Another interesting area involves investigating factors that affect rates of lexical replacement. In much the same way as biologists have been able to determine the effect of population size, temperature, topography and generation time on rates of species evolution (e.g. Bromham and Penny, 2003), it should be possible to investigate the effects of similar variables on rates of language evolution. For example, computer simulations (Nettle, 1999) suggest that smaller speech communities will have higher rates of change and a greater probability of borrowing. Explicitly modeling evolutionary change on a phylogeny also allows a probability distribution for ancestral character states to be inferred for any node on the tree (e.g. Lutzoni, Pagel and Reeb, 2003). Stochastic models of language change could thus be used to infer the likely vocabulary and phonology of proto-languages under a probabilistic framework. It may also be possible to extend our knowledge beyond the established proto-languages to infer much deeper language relationships by using slowly evolving words (Pagel, 2000) and non-lexical characters such as grammatical and phonological features of languages (Nichols, 1992). Analyzing phonological and morphological language data alongside lexical data also allows the results from any one data source to be validated against the other sources.

How the data itself is treated is another area for future development. Biologists have created Genbank (http://www.ncbi.nih.gov/Genbank/index.html), an online database for storing, organising and sharing large amounts of gene sequence data. This has proved extremely effective, not just for storing data, but as a collaborative research tool. Sequences are available to anyone who wishes to download them, either for their own work or to validate existing sequences and analyses. A similar, large-scale open-access data storage system could be developed for storing, organizing and distributing lexical, morphological and phonological language data.

Finally, there is more work to be done bringing together linguistic and genetic evidence for human population history. As we saw in Chapter 7, there are some interesting associations between patterns of linguistic and genetic variation that warrant further examination. This will involve improved phylogenetic analysis of genetic and linguistic data, as well as exploring ways of combining the two lines of evidence, perhaps using a structured coalescent approach.

Evolutionary biology and historical linguistics are indeed curiously connected. This observation is particularly significant for inferences about human history, where both fields can apply similar methods to answer similar questions using very different data. Here, I have been able to translate useful ideas and methods from one discipline to the other and there is no reason why this process of translation cannot continue. The ultimate goal for evolutionary biologists and historical linguists, however, should be to speak the same language (Feizkah, 2004).

REFERENCES

Bromham, L., and D. Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* 4:216-224

Covington, M. 1996. An algorithm to align words for historical comparison. *Computational Linguistics* 22:481-496.

Feizkah, E. 2004. New look at old words. *Time Magazine* 9 Feb., 52-52.

Heeringa, W., J. Nerbonne, and P. Kleiweg. 2002. Validating dialect comparison methods. Pages 445-452 in *Classification, automation, and new media.* (eds.) W. Gaul, and G. Ritter. Proceedings of the 24th Annual Conference of the Gesellschaft fur Klassifikation e. V., University of Passau, March 15-17, 2000, Springer, Berlin, Heidelberg and New York.

Kondrak, G. 2001. Identifying cognates by phonetic and semantic similarity. Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001). 103-110. Pittsburgh. June, 2001.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major Fungal Lineages Are Derived from Lichen Symbiotic Ancestors. *Nature* 411:937–940.

Nettle, D. 1999. Is the rate of linguistic change constant? *Lingua* 108:119-136.

Nichols, J. 1992. *Linguistic diversity in space and time*. University of Chicago Press, Chicago.

Pagel, M. 2000. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. Pages 189–207 in *Time depth in historical linguistics.* (C. Renfrew, A. McMahon, and L. Trask, eds.). McDonald Institute for Archaeological Research, Cambridge.

Thoreau, H. D. 1962. *The Journal of Henry D. Thoreau: In Fourteen Volumes Bound As Two : Vols. I-VII (1837-October, 1855) (1837-1855 Bound in 1 Volume)*, (eds.) B. Torrey and F. Allen, Dover, New York. Originally published in 1906.