LIBRARY
Te Tumu Herenga
THE UNIVERSITY OF AUCKLAND

RESEARCHSPACE@AUCKLAND

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand). This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.
http://researchspace.auckland.ac.nz/feedback

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form

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

ABSTRACT………………………………………………………………….. ii

ACKNOWLEDGEMENTS…..…………………………………………………… iii

TABLE OF CONTENTS……………………………………………………… iv

LIST OF FIGURES…………………………………………………………... ix

LIST OF TABLES……………………………………………………………. xii

CHAPTER 1: INTRODUCTION…………………………………………….. *1.1*

1.1 REFERENCES………………………………………………………… 1.8

CHAPTER 2: CURIOUS PARALLELS AND CURIOUS CONNECTIONS - PHYLOGENETIC THINKING IN BIOLOGY AND HISTORICAL LINGUISTICS………………………………….. *2.1*

2.0 ABSTRACT…………………………………………………………... 2.1

2.1 CURIOUS PARALLELS IN THE DOCUMENTS OF EVOLUTIONARY HISTORY………………………………………………………….. 2.1

2.2 IN THE BEGINNING – TWO ANCIENT GREEK OBSESSIONS…………… 2.4

2.3 BEFORE THERE WERE TREES………………………………………… 2.6

2.4 TANGLED TREES – EVOLUTION AND THE COMPARATIVE METHOD…. 2.10

2.5 AND THEN THERE WERE ALGORITHMS……………………………… 2.16

2.6 THE NEW SYNTHESIS OF BIOLOGY AND LINGUISTICS………………... 2.19

2.7 FUTURE CHALLENGES………………………………………………..... 2.21

2.8 REFERENCES………………………………………………………… 2.27

CHAPTER 3: LANGUAGE TREE DIVERGENCE TIMES SUPPORT THE ANATOLIAN THEORY OF INDO-EUROPEAN ORIGIN… *3.1*

3.0 ABSTRACT…………………………………………………………… 3.1

3.1 INTRODUCTION……………………………………………………… 3.2

3.2 MATERIALS AND METHODS…………………………………………… 3.3

3.2.1 DATA CODING…………………………………………………… 3.3

3.2.2 TREE CONSTRUCTION…………………………………………... 3.4

3.2.3 DIVERGENCE TIME ESTIMATES………………………………… 3.4

3.3 RESULTS AND DISCUSSION…………………………………………….. 3.5

3.4 REFERENCES…………………………………………………………... 3.9

CHAPTER 4: ARE ACCURATE DATES AN INTRACTABLE PROBLEM FOR HISTORICAL LINGUISTICS? *– TESTING HYPOTHESES ABOUT THE AGE OF INDO-EUROPEAN*…... ***4.1***

4.0 ABSTRACT…………………………………………………………... 4.1

4.1 LIMITATIONS OF THE COMPARATIVE METHOD AND GLOTTOCHRONOLOGY …………………………………………….... 4.1

4.2 A BIOLOGICAL SOLUTION TO A LINGUISTIC PROBLEM……………… 4.7

4.2.1 FROM WORD LISTS TO BINARY MATRICES – OVERCOMING INFORMATION LOSS…………………………………………….. 4.8

4.2.2 LIKELIHOOD MODELS AND BAYESIAN INFERENCE – OVERCOMING INACCURATE TREE-BUILDING METHODS………. 4.10

4.2.2.1 Likelihood Models of Evolution……………………….. 4.10

4.2.2.2 Bayesian Inference of Phylogeny…………………….... 4.14

4.2.3 NETWORK METHODS – OVERCOMING BORROWING…………… 4.17

4.2.4 RATE SMOOTHING AND ESTIMATING DATES – OVERCOMING RATE VARIATION THROUGH TIME…………………………….... 4.17

4.3 THE ORIGIN OF INDO-EUROPEAN – ILLUMINATION OR MORE MOTHS TO THE FLAME?……………………………………………………… 4.19

4.3.1 TWO THEORIES…………………………………………………. 4.20

4.3.2 DATA AND CODING…………………………………………….. 4.22

4.3.3 PHYLOGENETIC INFERENCE…………………………………….. 4.22

4.3.4 DIVERGENCE TIME ESTIMATION……………………………….. 4.26

4.3.5 TESTING ROBUSTNESS………………………………………….. 4.29

4.3.5.1 Bayesian Priors…..…………………………………….. 4.29

4.3.5.2 Cognacy Judgements………………………………….... 4.29

4.3.5.3 Calibrations and Constraint Trees…………………….. 4.30

4.3.5.4 Missing Data…………………………………………… 4.31

4.3.5.5 Root of Indo-European……………………………… 4.32

4.4 DISCUSSION…………………………………………………… 4.32

4.5 RESPONSE TO OUR CRITICS……………………………………… 4.33

4.5.1 THE POTENTIAL PITFALLS OF LINGUISTIC PALAEONTOLOGY…... 4.33

4.5.2 MODEL MISSPECIFICATION AND THE INDEPENDENCE OF CHARACTERS…………………………………………………… 4.36

4.5.2.1 Models are Lies That Lead Us to the Truth…………… 4.36

4.5.2.2 Can we Model Language Evolution?.......…………….. 4.37

4.5.2.3 The Independence Assumption………………………… 4.38

4.5.3 CONFIDENCE IN LEXICAL DATA……………………………… 4.39

4.6 CONCLUSION………………………………………………………... 4.44

4.7 REFERENCES………………………………………………………... 4.46

CHAPTER 5: FROM WORDS TO DATES -** WATER INTO WINE, MATHEMAGIC OR PHYLOGENETIC INFERENCE?**............ 5.1

5.0 ABSTRACT………………………………………………………... 5.1

5.1 WORDS INTO DATES OR WATER INTO WINE? ……………………….. 5.1

5.2 A NEW SET OF ANCIENT DATA…………………….……………... 5.3

5.3 STOCHASTIC MODELS AND BAYESIAN INFERENCE OF PHYLOGENY… 5.4

5.4 TWO DIFFERENT APPROACHES TO LIKELIHOOD INFERENCE AND MODELLING……………………………………………………… 5.5

5.4.1 TIME-REVERSIBLE MODEL AND METHOD 1…………………….. 5.6

5.4.2 STOCHASTIC-DOLLO MODEL AND METHOD 2…………………... 5.8

5.4.3 OTHER INFERENCE ISSUES……………………………………… 5.10

5.5 RESULTS……………………………………………………………….. 5.11

5.6 CONTROLLED MIRACLES – SYNTHETIC DATA VALIDATION………… 5.14

5.7 DISCUSSION……………………………………………………… 5.17

5.8 CONCLUSION………………………………………………………... 5.21

5.9 REFERENCES………………………………………………………... 5.21

CHAPTER 6: MAYAN LANGUAGE ORIGINS AND DIVERSIFICATION EXAMINED THROUGH PHYLOGENETIC ANALYSIS OF LEXICAL DATA………………………………………………… 6.1

6.0 ABSTRACT………………………………………………………………… 6.1
6.1 INTRODUCTION………………………………………………………… 6.2
6.2 MATERIALS AND METHODS…………………………….……………... 6.6
6.3 RESULTS……………………………………………………………….. 6.8
6.4 DISCUSSION…………………………………………………………. 6.14
6.5 CONCLUSION…………………………………………………………... 6.21
6.6 REFERENCES…………………………………………………………..... 6.21

CHAPTER 7: THE PERILS OF DATING EVE** – IMPROVED ANALYSES OF HUMAN MTDNA SEQUENCES HIGHLIGHT TIME DEPENDENCY OF MOLECULAR DATE ESTIMATES AND A TWO-PHASE POPULATION EXPANSION**……....... ***7.1***

7.0 ABSTRACT………………………………………………………………… 7.1
7.1 INTRODUCTION …………………………………………………………... 7.2
7.1.1 FIRST DATES………………………………………………………. 7.2
7.1.2 OVERCOMING DATING PROBLEMS……………………………… 7.4
7.1.2.1 Data…………………………………………………… 7.4
7.1.2.2 Tree Building and Quantifying Phylogenetic Uncertainty……………………………………………. 7.5
7.1.2.3 Estimating Rates and Dates……………………………. 7.6
7.1.3 OBJECTIVES OF THE CURRENT STUDY………………………….. 7.8
7.2 MATERIALS AND METHODS…………………………………………… 7.8
7.3 RESULTS AND DISCUSSION…………………………………………….. 7.11
7.3.1 DATING EVE AND TIME DEPENDENCY OF RATES……………..…. 7.11
7.3.1.1 Calibration Error………………………………………. 7.13
7.3.1.2 Purifying Selection…………………………………….. 7.13
7.3.1.3 Model Misspecification and Mutational Saturation........ 7.15
7.3.1.4 Multiple Constraints and Relaxing the Clock Assumption.. 7.17
7.3.1.5 Implications for the Age of Mitochondrial Eve……….. 7.18
7.3.2 HUMAN DISPERSAL OUT OF AFRICA…………………………… 7.19
7.3.3 MODELLING THE HUMAN POPULATION EXPANSION……………. 7.20
7.3.3.1 Some Caveats………………………………………….. 7.23

7.3.4 TOWARDS GENETIC AND LINGUISTIC CONSILIENCE……………. 7.24

7.4 CONCLUSION………………………………………………………… 7.28

7.5 REFERENCES………………………………………………………… 7.29

CHAPTER 8: CONCLUSION……………………………….….……..... 8.1

8.1 REFERENCES………………………………………………………… 8.5

LIST OF FIGURES

FIGURE 2.1 Evolutionary tree from one of Darwin’s (1837) notebooks……… **2.10**

FIGURE 2.2 Schleicher’s (1863) Indo-European language tree……………….. **2.12**

FIGURE 2.3 Schlyter’s (1827) manuscript phylogeny………………………… **2.13**

FIGURE 2.4 Phylogenetic tree for Polynesian languages from Green (1966)… **2.23**

FIGURE 2.5 Phylogenetic networks produced by split decomposition and NeighborNet analyses for a selection of Germanic languages…... **2.25**

FIGURE 3.1 Indo-European majority-rule consensus tree and divergence time distributions……... **3.6**

FIGURE 4.1 Indo-European language tree constructed using the comparative method (after Campbell, 1998)…………………………………………… **4.3**

FIGURE 4.2 Curve showing exponential decay of the percentage of shared cognates with time………………………………………………………. **4.4**

FIGURE 4.3 Illustration of the effect of rate variation on distance based tree building methods (after Blust, 2000)……………………………………… **4.6**

FIGURE 4.4 The gamma distribution, used to model rate variation between sites……………………………………………………………………. **4.12**

FIGURE 4.5 Calculation and comparison of likelihood for two language phylogenies…………………………………………………… **4.13**

FIGURE 4.6 Majority-rule consensus tree from Bayesian MCMC sample distribution for the Germanic sample data set………………………………... **4.16**

FIGURE 4.7 Estimating divergence times using the sample Germanic data set…………………………………………………………………... **4.19**

FIGURE 4.8 Majority-rule consensus tree from the initial Bayesian MCMC sample of 1,000 trees for the Indo-European data set…………………... **4.24**

FIGURE 4.9 Consensus network from the Bayesian MCMC sample of trees for the Indo-European data set…………………………….................... **4.25**

FIGURE 4.10 Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of Indo-European trees for the initial assumption set………... **4.28**

FIGURE 4.11 Frequency distribution of age estimates for the North and West Germanic subgroups across filtered Bayesian MCMC sample of Indo-European trees………………………………………………………… **4.28**

FIGURE 4.12 Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of Indo-European trees for analysis with doubtful cognates excluded……………………………………………….. **4.30**

FIGURE 4.13 Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of Indo-European trees using minimum set of topological constraints…………………………………………………….. **4.31**

FIGURE 4.14 Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of Indo-European trees with information about missing cognates included………………………………………………… **4.32**

FIGURE 4.15 Parsimony character traces for reflexes of Latin *focus* and *testa* in Romance languages... **4.41**

FIGURE 4.16 Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of Indo-European trees using Swadesh 100 word list items only……………………………………………………… **4.43**

FIGURE 4.17 Majority-rule consensus tree (unfiltered) from the Bayesian sample for the Indo-European Swadesh 100 word-list items only…………... **4.44**

FIGURE 5.1 Mean root age and 95% confidence interval for the Dyen et al. (1997) and Ringe et al. (2002) Indo-European dataset using the time-reversible and stochastic-Dollo models……………………………….... **5.11**

FIGURE 5.2 Mean root age and 95% confidence interval for a series of synthetic data analyses carried out using TraitLab……………………………… **5.15**

FIGURE 5.3 Majority-rule consensus tree from the initial Bayesian MCMC sample of 1,000 trees based on the Ringe *et al.*(2002) Indo-European data.. **5.19**

FIGURE 5.4 Consensus network from the initial Bayesian MCMC sample of 1,000 trees based on the Ringe *et al.*(2002) Indo-European data……… **5.20**

FIGURE 6.1 Phylogenetic network of the Mayan language data produced using *SplitsTree4* for: **a** complete 32-language Mayan data set; and **b** with Chuj and Tojolabal removed………………………………………….. **6.10**

FIGURE 6.2 Consensus network of the Bayesian sample distribution of Mayan trees analysed using the time-reversible model……………………….. **6.11**

FIGURE 6.3 Consensus trees from the Bayesian sample distribution of Mayan trees under the time-reversible model **a** and stochastic-Dollo model **b** of character evolution…………………………………………………. **6.12**

FIGURE 6.4 Confidence intervals for the age at the root of the Mayan language tree and for the age of the split separating Eastern Mayan/Q'anjobalan from Yucatecan/Cholan/Tzeltalan across a range of analyses……….. **6.14**

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**………………… **7.22**

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…………………………………………………………… **7.25**

LIST OF TABLES

TABLE 2.1 Conceptual parallels between biological and linguistic evolution. **2.3**

TABLE 4.1 Sample dataset of five Swadesh List terms across six Germanic languages (and Greek)………………………………………… **4.3**

TABLE 4.2 Germanic (and Greek) cognates from table 4.1 expressed in a binary matrix showing cognate presence or absence…………………… **4.9**

TABLE 4.3 The general time-reversible rate matrix used to model nucleotide evolution…………………………………………………………. **4.11**

TABLE 4.4 Simple likelihood time-reversible rate matrix adapted for modelling lexical replacement in language evolution………………………. **4.11**

TABLE 4.5 Age constraints used to calibrate the Indo-European divergence time calculations, based on known historical information……………. **4.27**

TABLE 5.1 Age constraints for the Dyen et al. (1997) Indo-European data set, used to calibrate the divergence time calculations, based on known historical information…………………………………………………… **5.7**

TABLE 5.2 Age constraints for the Ringe et al. (2002) data set, used to calibrate the divergence time calculations, based on known historical information…………………………………………………… **5.8**

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…………….. **5.11**

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………………… **5.15**

TABLE 6.1 Age constraints used to calibrate the Mayan divergence time calculations, based on known historical information………………………….. **6.8**

TABLE 7.1 Summary of key studies attempting to estimate the age of mitochondrial Eve………………………………………………………………. **7.3**

TABLE 7.2 Human mitochondrial DNA sequence sources, year of publication, region sampled, number of sequences used from this data set in the current study and Genbank accession numbers……………………………….. **7.9**

TABLE 7.3 95% HPD intervals for estimated substitution rate (per site per Myrs) and age of mitochondrial Eve derived from the coding region using single constraints with a strict clock and GTR + Γ + I substitution model ………………………………………………………………. **7.12**

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. **7.12**

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………………………………… **7.13**

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………………….. **7.16**

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 with a relaxed clock assumption……………………………………………………… **7.17**