Network Modeling of Lateral Inheritance in Genome and Language Evolution

Inaugural-Dissertation

For the attainment of the title of Doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Shijulal Nelson-Sathi
from Thiruvananthapuram, India

Düsseldorf, June 2013

From the Institute of Molecular Evolution
at the Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. William F. Martin
Co-supervisor: Prof. Dr. Tal Dagan

Date of the oral examination: 05.07.2013

Statement of authorship

I hereby certify that this dissertation is the result of my own work. No other person's work has been used without due acknowledgement. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Shijulal Nelson-Sathi

To my Amma

"The formation of different languages and of distinct species and the proofs that both have been developed through a gradual process, are curiously parallel"

- Charles Darwin, "The Descent Of Man, 1871, Chapter 3, pp 79"

Summary

Recent advances in genomics and linguistics have generated vast data that provide a useful benchmark to study micro- and macro- evolutionary processes. Several evolutionary process such as recombination, hybridization, genome fusions and lateral gene transfer/horizontal gene transfer (LGT or HGT) in genome evolution are fundamentally non-treelike in nature. Analogies for all major evolutionary processes in genome evolution are also recognized in language evolution. Consequently, networks, in addition to bifurcating trees, become an essential tool for modeling conflicting signals and evolutionary complexity in genomic and linguistic research. Studying genome and language evolution using phylogenetic networks traces both vertical as well as lateral component during their evolution. Because similar evolutionary processes shaped both genome and language evolution into contemporary forms, it is also possible to use methods that are developed to study genome evolution to study language evolution.

In the course of this thesis the frequency and impact of lateral transfers during the evolution of genomes (Haloarchaea) and languages (Indo-European and Polynesian) were investigated. Phylogenomic networks were reconstructed using genomes or languages as nodes and their evolutionary relationships as edges. The evolution of ten halorarchaeal genomes using a phylogenomic network approach with respect to 1,143 eubacterial reference genomes identified extensive inter domain LGT during haloarchaeal genome evolution. The results exemplify the role of LGT in transforming a strictly anaerobic, chemolithoautotropic methanogen into a heterotrophic, oxygen-respiring and bacteriorhodopsin-photosythetic organism. In the second and third studies presented here, the evolutionary history of 84 Indo-European and 33 Polynesian languages were examined. In both cases reconstructed phylogenomic networks identified a higher frequency of lexical borrowings than previously thought.

Modeling genome and language evolution using phylogenomic networks opens up new insights and provides more precise quantitative inferences about both vertical and lateral components during evolution.

Zusammenfassung

Jüngste Fortschritte in der Genomik und den Sprachwissenschaften haben eine Flut an Daten generiert welche eine gute Grundlage bieten, um kleinere und größere evolutionäre Prozesse zu studieren. Viele dieser Prozesse, wie Rekombination, Hybridisierung, Genomfusionen und horizontaler Gentransfer, lassen sich aufgrund ihrer Natur nicht durch bifurkierende Bäume darstellen. Für alle wichtigen evolutionären Prozesse innerhalb der Genomik wurden auch analoge Prozesse in der Evolution der Sprachen gefunden. Dementsprechend sind Netzwerke, nach bifurkierenden Bäumen, essentielle Werkzeuge, um widersprüchliche Signale und evolutionäre Komplexität in den beiden Wissenschaften darzustellen. Bei der Untersuchung der Evolution von Genomen und der Geschichte von Sprachen sichert der Gebrauch von phylogenetischen Netzwerken, dass nicht nur die vertikale sondern auch die horizontale Komponente der evolutionären Prozesse erfasst wird. Aufgrund ihrer Ähnlichkeiten ist es möglich, Methoden, die zur Untersuchung der Genomevolution entwickelt wurden, auch zu verwenden, um die Geschichte von Sprachen zu untersuchen.

Im Verlauf dieser Arbeit wurde die Häufigkeit und der Einfluss lateralen Transfers auf die Evolution mehrerer bakterieller Genome (Halobakterien) und mehrerer Sprachen (indogermanische und polynesische Sprachen) untersucht. Für diesen Zweck wurden phylogenetische Netzwerke erstellt, deren Knoten Genome bzw. Sprachen repräsentieren, und deren Kanten die evolutionäre Beziehung zwischen Genomen bzw. Sprachen darstellen. Zuerst wurde die Evolution von zehn halobakteriellen Genomen studiert. Durch einen Vergleich mit 1.143 eubakteriellen Genomen wurde eine hohe Häufigkeit an lateralen Gentransferprozessen zwischen den beiden Domänen gefunden. Anhand der Ergebnisse wird beispielhaft die Rolle lateralen Gentransfers erklärt, welcher ein strikt anaerobes, chemolithoautotrophes und methanogenes Archaeum, in einen heterotrophen, Sauerstoff atmenden und bacteriorhodopsin-photosynthetischen Organismus transformiert. In zwei weiteren Studien, welche hier präsentiert werden, wurde der evolutionäre Hintergrund von 84 Indoeuropäischen und 33

Polynesischen Sprachen untersucht. In beiden Fällen wiesen phylogenetische Netzwerke auf einen viel höheren Grad an lexikalischer Entlehnung hin, als bisher angenommen wurde.

Die Darstellung von Genomevolution und Sprachgeschichte mit Hilfe von phylogenetischen Netzwerken gewährt neue und quantitative Einblicke in die Bedeutung der vertikalen sowie der horizontalen Komponente evolutionärer Prozesse.

Publications included in this thesis

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. ***Proc Natl Acad Sci U S A*** 109:20537-20542.

Nelson-Sathi S, List JM, Geisler H, Fangerau H, Gray RD, Martin W, Dagan T (2011) Networks uncover hidden lexical borrowing in Indo-European language evolution. ***Proc Roy Soc Lond B*** 278:1794-1803.

Nelson-Sathi S, List JM, Greenhill S, Geisler H, Cohen O, Pupko T, Landan G, Martin WF, Dagan T, Gray RD (2013) Polynesian language networks reveal complex history of contacts during the Pacific settlement (*submitted)*.

Additional publications

Sousa FL , Thiergart T, Landan G, Nelson-Sathi S, Pereira IAC, Allen, JF, Lane N, Martin WF (2013) Early bioenergetic evolution. ***Phil Trans Roy Soc Lond B*** 368:20130088.

List J-M, Nelson-Sathi S, Martin WF, Geissler H (2013) Using phylogenetic networks to model Chinese dialect history *(Submitted)*.

Shijulal Nelson-Sathi, Ovidiu Popa, Johann-Mattis List, Hans Geisler, William F. Martin, and Tal Dagan (2013) in ***Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization***, eds Fangerau H, Geisler H, Halling T, Martin W (Steiner, Stuttgart) (*Book Chapter, In Press).*

Table of Contents

1. Introduction

"… it might be that some very ancient language had altered little, and had given rise to few new languages, whilst others (owing to the spreading and subsequent isolation and states of civilisation of the several races, descended from a common race) had altered much, and had given rise to many new languages and dialects. …"

(Charles Darwin (1859) The origin of species, chapter 13)

Genomes and languages have much in common, *genes* or *words* both evolve via vertical inheritance and lateral transfers. Traditionally, both processes were described in terms of family trees with ancestral species/languages diverging into its descendants. Immediately after Darwin (1859) published family trees to describe evolutionary relationships among biological species, German linguist August Schleicher (1863) introduced trees in linguistics to explain language evolution. Soon after the work published by Darwin (1859) and Schleicher (1863), tree models rapidly became a common tool to study evolution in both fields. However, lateral components in both entities (genomes and languages) play an important role during their evolution, its impacts get largely discarded. Over the past 500 years, different metaphors and models have been developed to describe the natural systems and genealogical relationships (Ragan 2009), but realistic models that can explain evolution of genomes and languages in addition to vertical inheritance are still lacking.

1.1 Similarities between genome and language evolution

The parallels between biological and linguistic evolution were evident both to Charles Darwin, who briefly addressed the topic of language evolution in *The Origin of species* (Darwin 1859), and to the linguist August Schleicher, who in an open letter to Ernst Haeckel discussed the similarities between language classification and species evolution (Schleicher 1863). As genomes contain all the necessary biological hereditary information of species, a language system contains all linguistic requirements for replicating communication within a speech community. Both entities are constituted by discrete evolving elements that allow building highly elaborate functional complexes. Discrete heritable units such as nucleotides, amino acids and genes in biological evolution are similar to that of words, phonemes and syntax in language evolution, also similar evolutionary forces shaped both genomes and languages into contemporary form (Pagel 2009). Evolutionary relationship between genomes and languages shows that both systems undergone similar evolutionary shifts to attain increasing level of complexity (Ji 1989). Both genomes and languages evolved by evolutionary strategies that affected their properties and complexity, both were constantly subject to change and affected by lateral gene transfers in genomes and lexical borrowings in languages.

1.2 Lateral component of genome and language evolution

In biology, statistical methods were developed in late 1960's to infer phylogenetic trees from sets of homologous molecular sequences (Fitch and Margoliash 1967, Dayhoff 1969) and phylogenetic trees reconstructed using universal small-subunit ribosomal genes were soon assumed to represent a vertical bifurcating tree of life (Woese et al. 1990). As further genome sequences became available and phylogenetic inference grew more-sophisticated, single-gene topologies failed to tell a fully consistent phylogenetic story. In prokaryotes, patterns of topological incongruence

among different trees, or between a gene tree and a reference tree were interpreted as a primary evidence for lateral gene transfer (Jain et al. 1999, Beiko et al. 2005, Dagan et al. 2008, Soria-Carrasco and Castresana 2008, Dagan and Martin 2009, Haggerty et al. 2009, Ragan 2009). Genome evolution includes both vertical as well as lateral components and it can thus take the form of networks (Kunin et al. 2005, Dagan et al. 2008, Soria-Carrasco and Castresana 2008, Dagan and Martin 2009, Haggerty et al. 2009, Ragan 2009). Recent studies showed that substantial amount of lateral gene transfer events occurred between different taxonomic groups during bacterial evolution (Koonin et al. 2001, Dagan and Martin 2007).

Lateral gene transfer processes in genome evolution have strong resemblance with lexical borrowings in language evolution. Biologists and linguists were well aware of the fact that evolutionary relationships are not always necessarily *vertical* (genealogical). However tree models were long used as a metaphor for modeling genome and language evolution. Gradually tree models were rejected by several scholars arguing against the use of simple tree model for describing the complicated evolution of genomes and languages. In 1872, the German linguist Johannes Schmidt (1843-1901) proposed a "*Wave theory*" to better explain the patterns of language evolution, followed by many alternative ways such as "*chain*" or even "*animated pictures*" (Schuchardt 1870). Inadequacy of explaining vertical and lateral components of language evolution, none of these models gained acceptance among all linguistic scholars. An early explicit network approach can be found in a study by Bofante (1931) where more complex relations between languages are considered. Nevertheless, many independent models were developed in both fields to account for the lateral component in evolution (Bryant et al. 2005, Nakhleh et al. 2005, Huson and Bryant 2006), while these models can show the conflicting signals in the data, none of them is capable of giving an estimate regarding quantitative measurement of lateral transfers during evolution.

1.2.1 Lateral gene transfers (LGTs) in microbial genome evolution

Microbes evolve not only via vertical inheritance but also by acquiring genetic material from their environment via a process called lateral or horizontal gene transfer (LGT or HGT), during which a recipient genome acquires genetic material from a donor genome. The importance of LGTs for microbial genome evolution was not recognized until 1950's. Freeman first demonstrated LGT in 1951 by transferring a viral gene on to a non-virulent *Corynebacterium diphtheriae* strain so that it turned into a virulent stain. In 1959 Ochiai et al. described the existence of long distance inter-bacterial gene transfer and in mid 1980's Syvanen predicted the existence, biological significance and role of LGT's in biological evolution. LGT plays a crucial role in microbial genome evolution (Doolittle 1999, Ochman et al. 2000) giving the bacterial genomes a rather dynamic structure compared to the previously assumed static one (Martin 1999).

i. Basic mechanism of lateral gene transfers (LGTs)

The occurrence of lateral gene transfers (LGTs) that blurs the boundaries between species has been generally accepted for many years (Popa and Dagan 2011). There are several mechanisms discovered so far which mediate lateral gene transfer. They include, *transformation* - the uptake of DNA from the environment (Chen et al. 2005), *conjugation* – the transfer of genetic material via plasmid (Norman et al. 2005) and *transduction* – the transfer of DNA by phages (Thomas and Nielson 2005). In addition, genetic material can be transferred using *gene transfer agents* (GTA, Lang and Beatty 2007) or *nanotubes* that are membrane tubular protrusions connecting between cells (Dubey and Yehuda 2011). Microbial genomes are in a constant state of flux i.e., any segment of genetic material in a large bacterial population might have the chance to be laterally transferred. The current mechanistic understanding of the processes that facilitate LGT events has come from the study of model organisms and the environmental factors that promote or limit LGT events in nature are not well known. Even though only a minor

proportion of the transferred DNA between species is likely to be maintained in the new host over generations, there are many factors limiting evolutionarily successful LGT including mechanismistic barriers to their establishment, expression and function (Popa and Dagan 2011). In addition temporal and spatial factors limit the spread of the transferred genetic material in bacterial populations (Thomas and Neilson 2005).

ii. Impact of lateral gene transfers (LGTs) in genome evolution

The proportion of protein families affected by LGT during microbial evolution as inferred from gene phylogenies is estimated to be between 60% (Kunin et al. 2005, Dagan et al. 2008) and 90% (Mirkin et al. 2003). Since the amount of lateral transfers in microbial genomes is much more important than mutation in evolving new functions (Lawrence and Ochman 1998), the underlying process of microbial evolution would be fundamentally at odds with the concept of a single bifurcating tree, because lateral transfers are not a tree like process (Martin 1999, Doolittle 2004, Gogarten and Townsend 2005, Pal et al. 2005, Dagan and Martin 2006). A simple tree model that uses genealogical relationships is not capable of adequately describing the evolution of prokaryotes (Bapteste et. al. 2009) because the current phylogeny itself may be defined in a large part by LGT (Doolittle 1999). A few examples of abundant lateral gene transfer during early evolution include transfers from organellar to nuclear genomes in eukaryotes (Doolittle 1998, Martin 2003) and transfer of plasmids between bacterial species (Naik et al. 1994). Archaea - Eubacteria inter domain lateral gene transfers were frequent during early evolution and they played an important role for the evolution of the archaeal domain (Nelson et al. 1999, Deppenmeir et al. 2001, Allers and Mevarech 2005). Regarding bacterial pathogenicity, LGTs are considered as the primary mechanism for spreading antibiotic resistance genes in microbes (Nielsen 1998, Koonin et al. 2001).

iii. Trends and barriers of lateral gene transfers (LGTs)

Although inter-domain LGTs were frequent during early evolution (Nelson 1999, Mongodin 2005), most recent LGTs seem to occur between closely related species (≥95% similar nucleotide composition) suggesting an existence of donor-recipient *similarity barrier* (Popa and Dagan 2011). A *functional barrier* exists that suggesting most of the transferred genes perform preferred metabolic functions. The majority of the LGTs occur within the same habitat suggesting an *ecological barrier* and the frequency of LGTs between species negatively correlate with the physical distance suggesting a *spatial barrier* (Popa and Dagan 2011). In the context of long distance gene transfers, the transduction mechanism is considered to have the longest range (Majewski 2001).

1.2.2 Borrowings in language evolution

Lexical borrowing is the transfer of a *word* from a donor language to a recipient language as a result of a certain kind of contact between the speakers of the two languages (Trask 2000). Lexical borrowing can be reciprocal or unidirectional and occurs at variable rates during evolution. Lateral interactions during language evolution can range from the exchange of a few words to deep interference. Factors affecting the rate of lexical borrowing during language evolution include socio-cultural situation, the intensity of contact between the speakers of the respective languages, the dignity of specific language varieties within a given speech community, the genetic or typological closeness of the languages that facilitates the inclusion of foreign words, the amount of bi- or multilingual speakers in the respective linguistic communities, or combinations thereof (Thomason and kaufman 1998, Aikhenvald 2006). For example, English has been heavily influenced throughout its history by different languages such as Celtic, Norse and Norman French (Fox 1995).

i. Basic mechanism of lexical borrowing

Generally *words* from a donor language enter in to a recipient language as a technical term in contrast to the exposure to a foreign culture. Mostly donor languages may be players in dominant field of activities such as arts, religion, business, science and philosophy. Once a borrowed word looses its foreign cultural associations, it passes into general use in the languages. Borrowing processes can be a *direct transfer* or a *semantic transfer*. In *direct transfer,* form and meaning of words are transferred as a whole from the donor to recipient language. e.g., word *flor* was directly transferred from Old French to English *flower*. In case of *semantic transfer*, a word is reproduced in the recipient languages by expanding the meaning of a given word to match the *form-meaning* unity in the donor language (Weinreich 1953). For example German *maus* has two meanings: "animal" and a "computer device", the second meaning is a semantic borrowing from English to German. Basic vocabulary of languages is supposed to be more resistant to borrowing than its whole lexicon.

ii. Impact of lexical borrowings in language evolution

Similar to LGT in genome evolution, lexical borrowing is a non-tree-like evolutionary event that cannot be reconstructed using phylogenetic trees that are common in evolutionary biology (Soukhanov 1992, Orel 2003). In language evolution, lexical borrowing resulting from contacts, linguists were well aware of the existence of non-geological component in language evolution. For example at least 60% of cognates (words having same etymological origin) in Indo-European languages have been affected by at least one borrowing event during evolution (Nelson-Sathi et al. 2011). A recent study shows that English has borrowed eight percentage of its basic vocabulary from Old Norse and Old French (Embleton 2000). Icelandic, on the other hand, has preserved most of its original words (Bergsland 1962), maybe because of its geographical isolation.

iii. Trends and barriers of lexical borrowings

Since in most cases *word* borrowings happen as a result of interaction between two speech communities, there also exist some trends and barriers in borrowing process of language evolution. Since the sound systems of languages may differ crucially, not all words that might be borrowed are equally easy to pronounce for the speakers of different languages. In cases where the sound system of possible donor and recipient languages is similar, direct borrowing will happen result in a *similarity barrier*. Usually word borrowing occurs when a recipient language lacks certain words for some concepts that is present in the donor language, borrowing heavily depends on the meaning of the items being borrowed. For example, words representing basic concepts that are very essential for daily life are less likely to be transferred, resulting in a *functional barrier* (Hock and Joseph 2009). Since borrowing occurs as a result of interaction between two speech communities, it is obvious that borrowing events will be less frequent between geographically distant ones, resulting in a *spatial barrier*. The spatial barrier is closely connected with what one might call a *socio-cultural* or *socio-political barrier* for lexical borrowing: Due to social, cultural, or political reasons a given language variety may either be promoted or marginalized by the ones who speak it, resulting in a high or low borrowing rates (Tadmor 2009).

1.3 Networks to study lateral transfers in genome and language evolution

Traditionally, shared traits among genomes and languages were used to thought to include close relationships in the family tree, hence trees became the leading metaphor to describe their evolutionary relationships. Nevertheless, biologists and linguists have long been aware of the problems that lateral transfers poses to the tree model. Given the specific need to model both vertical and lateral processes, biologists and linguists naturally turn to networks as a format to represent evolving entities. Network

representation of relations is not new and it has been documented even before Darwin's species tree was popularized (Ragan 2009).

A network is a mathematical model of pairwise relations among entities. It is described as a collection of pairwise relations between entities where the entities are called *nodes* (or *vertices*) and the relations between them *edges* (Newman 2010). A network of N vertices can be fully defined by matrix, $A = [a_{ij}]_{N*N}$, with $a_{ij} = a_{ji} \neq 0$ if a connection exists between node i and j, and $a_{ij} = a_{ji} = 0$ otherwise. In a binary network, the information is limited to whether the vertices are connected or not. In a weighted Network edges can also have a certain weight that signifies the strength of the connection. Network based approaches are common in almost all fields of science including social science, cell biology, ecology and statistical physics. Networks provide complete overview of the whole system as interacting entities and its properties and connectivity patterns can tell us about to the topology, dynamics and development of the modelled system (Strogatz 2001, Alon 2007, Newman 2010).

Networks are generally used in phylogenetic research for reconstruction of evolutionary processes that are non-tree like in nature including hybridization, recombination, genome fusion and lateral gene transfer (Dagan 2011). A phylogenomic network represents completely sequenced genomes or lexicon of languages as nodes and their relationship as edges (Dagan et al. 2008, Dagan 2011, Nelson-Sathi et al. 2011). The network relationships can be reconstructed by means of genetic information from shared gene content (Halary et al. 2010), shared similarity (Dagan and Martin 2007, Lima-Mendez et al. 2008) or from phylogenetic trees (Beiko et al. 2005, Dagan and Martin 2008, Popa et al. 2011).

In contrast to phylogenetic trees, phylogenomic networks have many advantages when studying genome and language evolution. An in-depth analysis of network structure and properties enables the application of networks to study evolution in a much more quantitative way (Dagan 2011). Since they consider the lateral component, phylogenomic networks show more a dynamical picture of evolution rather than a static picture of relationship between taxa.

2. Aim of the thesis

In light of the forgoing, the aims of this thesis were to quantify the frequency and impact of lateral transfers during genome and language evolution using phylogenomic networks and publically available data (microbial genomes and lexicon of languages).

i. In the case of genomes, the goal was to provide a detailed investigation towards the amount of ancient eubacteria-archaea inter-domain LGTs and its impact on physiologically transforming an anerobic chemolithoautotroph (methanogen) into aerobic heterotroph (haloarchaea).

ii. In the case of language evolution, the goal was to use phylogenomic network approach, investigate the rate and frequency of hidden *lexical* borrowings during the evolution of Indo-European and Polynesian languages.

Thematic contents of the thesis

This thesis deals with phylogenomic network approaches to model genome and language evolution and it is mainly divided into two complementary sections comprising a total of three publications. The first part deals with the phylogenomic approach to infer the lateral gene transfers between archaeal and eubacterial domains and its impact on haloarchaeal evolution (Nelson-Sathi et al. 2012). The second part deals with the application of phylogenomic networks to infer the impact of lexical borrowings during Indo-European and Polynesian language evolution (Nelson-Sathi et al. 2011, Nelson-Sathi et al., submitted).

3. Publications

3.1

Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea

Shijulal Nelson-Sathi[1], Tal Dagan[2], Giddy Landan[1,2], Arnold Janssen[3], Mike Steel[4], James McInerney[5], Uwe Deppenmeier[6], William F. Martin[1*]

[1] Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Germany
[2] Institute of Genomic Microbiology, Heinrich-Heine University Düsseldorf, Germany
[3] Mathematisches Institut, Heinrich-Heine University Düsseldorf, Germany
[4] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand
[5] Department of Biology, The National University of Ireland, Maynooth, Kildare, Ireland
[6] Institute of Microbiology and Biotechnology, University of Bonn, Bonn, Germany
* Corresponding author: bill@hhu.de

Keywords: Lateral gene transfer, horizontal gene transfer, networks, Haloarchaea, respiration, methanogens.

The presented manuscript was published in the Journal of "*Proceedings of the National Academy of Sciences (PNAS)"*, 2012 Dec.

Impact Factor – 9.6

Contribution of Shijulal Nelson-Sathi – First author

Experimental design: 25%
Execution and analysis of experiments: 80%
Manuscript writing: 20%

Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea

Shijulal Nelson-Sathi[a], Tal Dagan[b], Giddy Landan[a,b], Arnold Janssen[c], Mike Steel[d], James O. McInerney[e], Uwe Deppenmeier[f], and William F. Martin[a,1]

[a]Institute of Molecular Evolution, [b]Institute of Genomic Microbiology, [c]Mathematisches Institut, Heinrich Heine University, 40225 Düsseldorf, Germany; [d]Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand; [e]Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland; and [f]Institute of Microbiology and Biotechnology, University of Bonn, 53115 Bonn, Germany

Edited* by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved October 25, 2012 (received for review May 29, 2012)

Archaebacterial halophiles (Haloarchaea) are oxygen-respiring heterotrophs that derive from methanogens—strictly anaerobic, hydrogen-dependent autotrophs. Haloarchaeal genomes are known to have acquired, via lateral gene transfer (LGT), several genes from eubacteria, but it is yet unknown how many genes the Haloarchaea acquired in total and, more importantly, whether independent haloarchaeal lineages acquired their genes in parallel, or as a single acquisition at the origin of the group. Here we have studied 10 haloarchaeal and 1,143 reference genomes and have identified 1,089 haloarchaeal gene families that were acquired by a methanogenic recipient from eubacteria. The data suggest that these genes were acquired in the haloarchaeal common ancestor, not in parallel in independent haloarchaeal lineages, nor in the common ancestor of haloarchaeans and methanosarcinales. The 1,089 acquisitions include genes for catabolic carbon metabolism, membrane transporters, menaquinone biosynthesis, and complexes I–IV of the eubacterial respiratory chain that functions in the haloarchaeal membrane consisting of diphytanyl isoprene ether lipids. LGT on a massive scale transformed a strictly anaerobic, chemolithoautotrophic methanogen into the heterotrophic, oxygen-respiring, and bacteriorhodopsin-photosynthetic haloarchaeal common ancestor.

Halophilic archaebacteria (Haloarchaea) require concentrated salt solutions for survival and can inhabit saturated brine environments such as salt lakes, the Dead Sea, and salterns (1). In rRNA and phylogenomic analyses of informational genes, Haloarchaea always branch well within the methanogens (2–4). Haloarchaea can thus be seen as deriving from methanogen ancestors, but the physiology of methanogens and halophiles could hardly be more different. Methanogens are strict anaerobes, most species are lithoautotrophs that use electrons from H_2 to reduce CO_2 to methane (obligate hydrogenotrophic methanogens), thereby generating a chemiosmotic ion gradient for ATP synthesis in their energy metabolism, although some species can generate methane from reduced C_1 compounds, or acetate in the case of aceticlastic forms (5–7). Their carbon metabolism involves the Wood–Ljungdahl (acetyl-CoA) pathway of CO_2 fixation (5–7). In contrast, Haloarchaea are obligate heterotrophs that typically use O_2 as the terminal acceptor of their electron transport chain, although many can also use alternative electron acceptors such as nitrate in addition to light harnessing via a bacteriorhodopsin-based proton pumping system (8). The evolutionary nature of that radical physiological transformation from anaerobic chemolithoautotroph to aerobic heterotroph is of interest.

Many individual reports document that lateral gene transfer (LGT) from eubacteria was involved in the origin of at least some components of haloarchaeal metabolism. These include the operon for gas vesicle formation, which allows Haloarchaea to remain in surface waters (9), the newly identified methylaspartate cycle of acetyl-CoA oxidation (10), various components of the haloarchaeal aerobic respiratory chain (11–18), and proteins involved in the assembly of FeS clusters (19). The sequencing of the first haloarchaeal genome over a decade ago identified some eubacterial genes that possibly could have been acquired by lateral gene transfer (11, 20), and whereas substantial data that would illuminate the origin of haloarchaeal physiology have accumulated since then, those data have not been subjected to comparative evolutionary analysis. Investigating the role of the environment in haloarchaeal genome evolution, Rhodes et al. (21) recently showed that Haloarchaea are indeed far more likely to acquire genes from other halophiles, but they did not address the issues at the focus of our present investigation, namely: How many eubacterial acquisitions are present in haloarchaeal genomes? How was the physiological transformation of methanogens to Haloarchaea affected by LGT? Do those acquisitions trace to the haloarchaeal common ancestor as a single acquisition or not?

To discern whether the eubacterial genes in haloarchaeal genomes are the result of multiple independent transfers in individual lineages or the result of a single ancient mass acquisition, here we have analyzed 10 sequenced haloarchaeal genomes—*Haloarcula marismortui* (22), *Halobacterium salinarum* (23), *Halobacterium* sp. (20), *Halomicrobium mukohataei* (24), *Haloquadratum walsbyi* (25), *Halorhabdus utahensis* (26), *Halorubrum lacusprofundi* (27), *Natrialba magadii* (28), *Natronomonas pharaonis* (29), and *Haloterrigena turkmmenica* (30)—in the context of 65 other archaebacterial and >1,000 eubacterial reference genomes.

Results and Discussion

We first clustered the 172,531 proteins encoded in the chromosomes of 75 archaebacterial genomes into families using the standard Markov cluster (MCL) procedure (31) yielding 16,061 protein families. Comparison with 1,078 completely sequenced eubacterial genomes delivered 1,479 protein families that are present in at least two Haloarchaea and contain archaebacterial and eubacterial homologs (Fig. 1*A*). Gene trees for the protein families were reconstructed using maximum likelihood inference (*Methods*).

Of 1,479 trees, 1,089 (73%) uncovered Haloarchaea as monophyletic and rooting within (or branching next to) eubacterial rather than archaebacterial homologs (Fig. 1*B*). For 414 of these trees, no homologs at all were detected in nonhalophilic

Author contributions: T.D., U.D., and W.F.M. designed research; S.N.-S. and T.D. performed research; S.N.-S., G.L., A.J., M.S., and J.O.M. analyzed data; and S.N.-S., G.L., and W.F.M. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: bill@hhu.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1209119109/-/DCSupplemental.

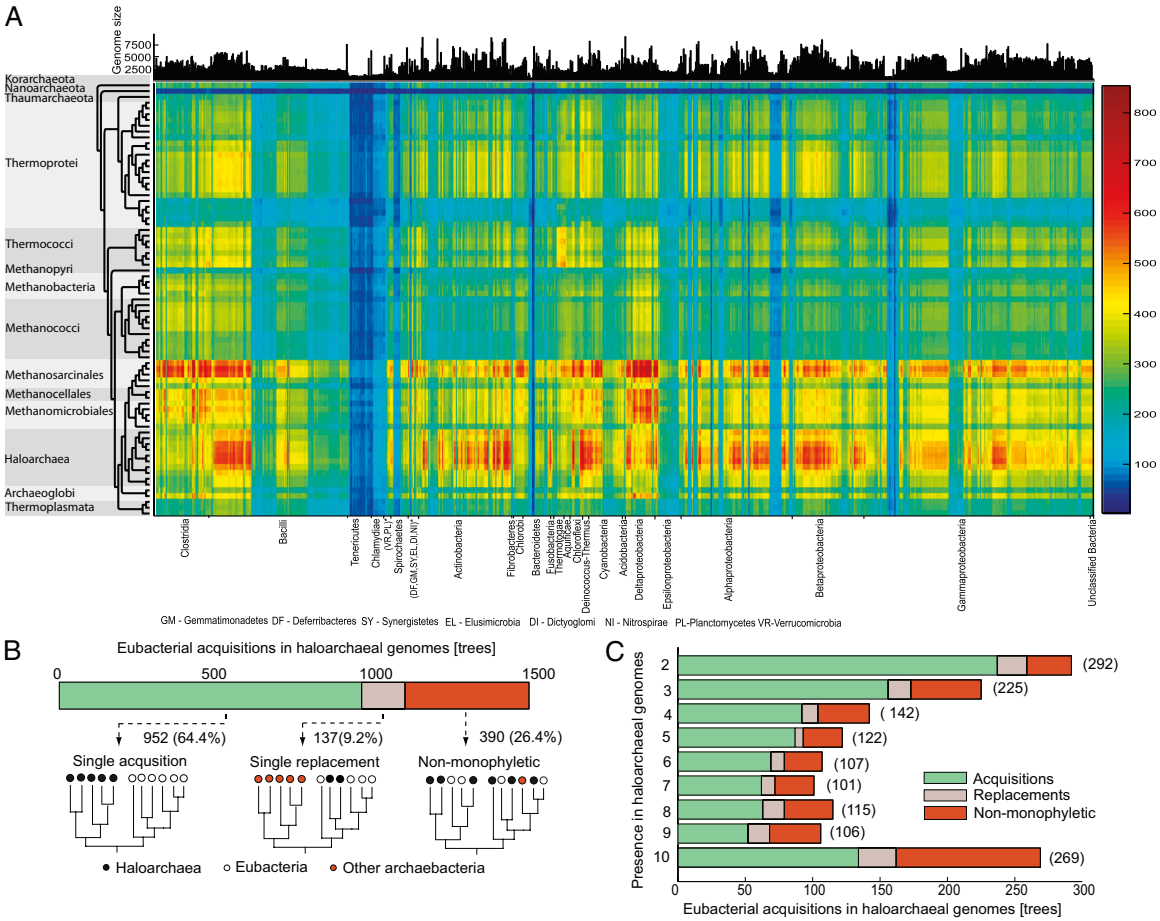


Fig. 1. (*A*) Number of shared genes between 1,078 bacterial genomes and 75 archaebacterial genomes. (*B*) Types of phylogenetic trees obtained with respect to the relationship of Haloarchaea, nonhalophilic archaea, and eubacterial genes. (*C*) Types of phylogenetic trees detailed by the number of haloarchaeal taxa.

archaebacteria. An additional 538 families had only very distant homologs (E values $>10^{-10}$ or amino acid identity <30%) in some nonhalophilic archaebacteria, together we designate these 952 cases as "acquisitions." An additional 137 genes yielded trees in which Haloarchaea branch within eubacteria to the exclusion of readily detectable archaebacterial homologs, we designate these genes as "replacements"; acquisitions and replacements we designate collectively as "imports" (Fig. 1*B*). The 390 cases of Haloarchaea nonmonophyly included 76 trees in which one haloarchaeon branched deviantly and 105 trees in which the Haloarchaea were split into two groups of two or more species. Because LGT is common in prokaryotes (32, 33), among haloarchaeans in particular (21), these 181 gene trees could well depict secondary transfers into or from the Haloarchaea.

Single Ancestral Acquisition. Are the 1,089 eubacterial imports in haloarchaeal genomes the result of a single ancestral acquisition or multiple parallel acquisitions? Monophyly alone does not completely decide the issue, because it is possible that a bacterial gene could be acquired recently in one haloarchaeal lineage and then passed around to other Haloarchaea by LGT. Such a process could, in principle, also generate monophyly for imported genes in a phylogenetic tree. However, in that case, individual gene trees for imported genes would be very different from one another as opposed to the case of single acquisition, where trees for imports should be the same due to vertical inheritance from the haloarchaeal common ancestor. Moreover, trees for ancestrally acquired eubacterial imports should not only be similar to each other, they should also be similar to trees for endogenous haloarchaeal genes that are shared only with other archaebacteria, which we call recipient genes. There are 364 haloarchaeal recipient genes that are present as single copies in all 10 Haloarchaea sampled and 109 haloarchaeal imports that are present as single copies in all 10 Haloarchaea (Fig. 2*A*), providing comparable tree sets. To avoid oversampling, the *H. salinarum* and the *Halobacterium* sp. genomes were condensed to one genome, because they share almost exactly the same genes and would have skewed the test by enhancing the congruence of the two sets.

Comparing the distributions of phylogenetic splits observed in the 364 recipient trees and the 109 imported trees containing all 10 (condensed to 9) Haloarchaea shows that the two sets exhibit a very similar phylogenetic signal (Fig. 2*B*). The six most common splits in the two sets of trees are identical and comprise 51% and 46% of the splits in the two sets, respectively. Moreover, these six splits exactly define the haloarchaeal phylogeny

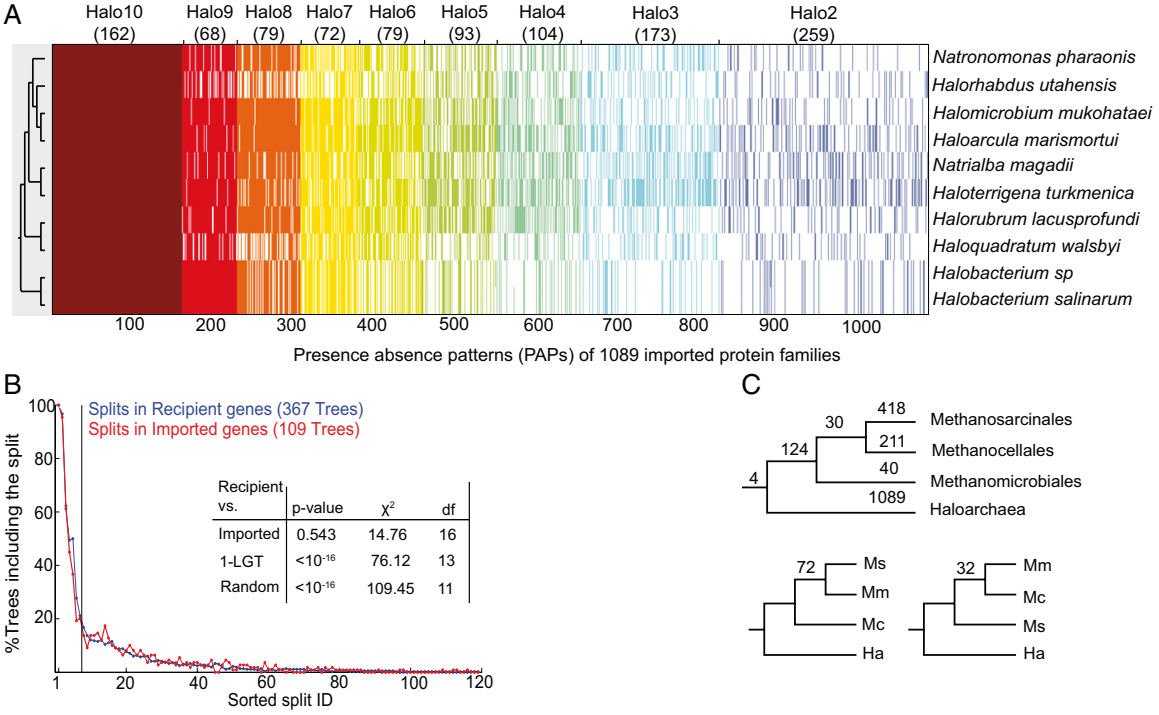


Fig. 2. Eubacterial genes in Haloarchaea. (*A*) Distribution of eubacterial imports present in at least two Haloarchaea. (*B*) Histogram of phylogenetic splits in imported and recipient trees. (*Inset*) Statistical test supporting single acquisition of imported eubacterial genes into the haloarchaeal common ancestor (*Methods*). df, degrees of freedom. Note that incompatible split frequency correlates with topological distance to the reference tree ($P = 7{\cdot}10^{-13}$ for recipient genes, $r = 0.76$; $P = 7{\cdot}10^{-19}$ for the imports), as expected for phylogenetic errors but not for competing biological signals (*SI Text*). (*C*) Numbers of eubacterial acquisitions and replacements in the ancestors of haloarchaea (Ha), methanosarcinales (Ms), methanomicrobiales (Mm), and methanocellales (Mc) shown for the reference topology in Fig. 1*A* and for the alternative topologies with respect to the Ms/Mm/Mc branching order (for unlabeled branches, the number of imports is identical with that shown for the reference topology; numbers of acquisitions and replacements are given in *SI Text*). Note that the uncultured methanogenic archaeon RC-I is not classified with methanocellales in GenBank taxonomy, but it branched with *Methanocella paludicola* SANAE in our reference topology, for which reason it was treated as an Mc member here. The frequency distributions of eubacterial imports across genomes and functional categories for haloarchaea is given in Table S1. The numbers of acquisitions and replacements, respectively, for the numbers of imports shown in *C* are 4: (4, 0); 124: (101, 23); 30: (22, 8); 418: (373, 45); 211: (141, 70); 40: (30, 10); 1,089: (952, 137); 72: (58, 14); and 32: (17, 15). For the methanogens in *C*, all species names and corresponding frequency distributions for functional categories are given in Table S3.

generated by 56 universally distributed archaebacterial genes (Fig. 1*A*, *Left*). To test the statistical significance of this evidence in favor of single acquisition, we used a goodness-of-fit test to compare the distributions of topologies in the two tree sets. The null hypothesis is that the two samples of trees are drawn from the same distribution, whereas the alternative hypothesis is that the two samples differ in their distribution (*Methods*). The test's *P* value was 0.543, meaning that the null hypothesis could not be rejected (Fig. 2*B*, *Inset*). To complement this result, we examined two additional sets of trees. One set consisted of 109 random trees, and the second consisted of the observed 109 single copy imported gene trees subject to one LGT rearrangement (one random prune and graft operation) each. The latter case of one LGT rearrangement represents the slightest possible LGT-induced deviation from the null hypothesis of single acquisition in the haloarchaeal common ancestor followed by vertical evolution. Both sets were tested against the recipient trees and found to be significantly different (*P* values $<<10^{-16}$; Fig. 2*B*, *Inset*), strongly rejecting the one LGT rearrangement per gene case.

When we include the 53 multiple copy genes that are present in all 10 genomes, the one LGT rearrangement per gene is also excluded, although the significance (*P* values $<10^{-8}$, see *SI Text*) drops. That drop is expected, however, because horizontal transfer, not duplication, drives the expansion of gene families in prokaryotes (34), hence the inclusion of multicopy genes preferentially includes genes for which LGT is more prevalent. We note that the goodness-of-fit test does not exclude one LGT for each gene, up to 34% of the 109 single-copy recipient trees can accept a single random prune and graft operation without the test rejecting the one LGT rearrangement case for the recipient set as a whole. However, for the 162 genes that are present in all 10 genomes, the possibility that the majority of imported genes are monophyletic because of import into one of the haloarchaeal lineages and subsequent passing around of the same gene can be excluded.

For the imports present in eight or fewer haloarchaeal genomes, excluding the (perhaps unlikely) possibility that monophyly is not due to acquisition in the haloarchaeal common ancestor but to lineage-specific acquisition and subsequent spread, is more difficult, mainly for reasons of sample size. The goodness-of-fit test based on split distributions cannot be used because few comparisons yield identical leaf sets for import vs. recipient trees. For the ≤8-species cases, we therefore developed a less direct test, comparing the sets of recipient and import trees via their phylogenetic compatibility with the recipient trees for the 10 haloarchaeal species (*Methods*). Here, too, the null hypothesis of common

ancestry for import and recipient genes could not be rejected in any of the ≤8-species cases, although the acquire-and-spread scenario was also not rejected for the 4-, 5-, and 6-species single and multiple copy cases (268 imports total; *SI Text*). Given that (*i*) the conventional interpretation of monophyly is presence in the common ancestor, that (*ii*) the 151 eight and seven species cases reject the acquire-and-spread scenario (*SI Text*) as an alternative explanation of monophyly, and that (*iii*) the data that most directly address the acquire-and-spread scenario—the 162 eubacterial imports present in all 10 genomes—most strongly reject it (Fig. 2*B*), the simplest interpretation of monophyly for the 1,089 imports is that their origin traces to a single acquisition in the haloarchaeal common ancestor followed mainly by vertical descent and widespread differential loss, with some subsequent LGT among haloarchaea (21, 32, 33), notably for multicopy genes (34), not being excluded.

Methanogens Are Affine for Eubacterial Genes. As seen in Fig. 1*A*, not only the 10 Haloarchaea, but also the five Methanosarcinales (Ms), the two Methanocellales (Mc), and the five Methanomicrobiales (Mm) sampled share many genes with eubacteria, raising the question of when these imports entered these methanogen lineages. Repeating our phylogenetic analyses for these groups (Fig. 2*C*) reveals that merely four eubacterial imports (three predicted membrane proteins and a glycosyl transferase) can be traced to their common ancestor, and that these are present in at most 6 of the 22 descendant genomes. Whereas 124 imports can be traced to the Ms/Mc/Mm common ancestor, these imports are also sparsely distributed, with only two (COG1032, an FeS-oxidoreductase and COG1387, histidinol phosphatase) being present in all 12 descendant methanogens. This contrasts to the 1,089 haloarchaeal imports that are specific to the haloarchaeal lineage, 162 of which (15%) have been retained in all 10 haloarchaeans sampled. The Ms, Mc, and Mm lineages have—like the haloarchaea—independently acquired hundreds of eubacterial genes, but the crucial observation is that they have remained strict anaerobes, and they have furthermore remained obligatory methanogenic (5–7). In stark contrast, the halophiles became aerobic heterotrophs and lost methanogenesis altogether. Collectively, the data point to a very different nature of the gene acquisition process in the halophiles and methanogens sampled here.

Donor Lineages. The acquisition of >1,000 genes is reminiscent of massive gene acquisitions surrounding the origin of mitochondria (35, 36) or plastids (37, 38). From what donor were these genes acquired? Because bacterial chromosomes undergo gene influx and gene export over time, it is unlikely that any one contemporary bacterial lineage would emerge as the donor of all eubacterial genes in haloarchaeal chromosomes (36, 39). All of the higher level taxa sampled appear as the sole sister group to the haloarchaeal gene or appeared in a sister group of mixed phylogenetic composition, as one might expect due to frequent LGT among bacteria (Figs. S1 and S2*A*). The most frequent apparent donor lineage was the actinobacteria with 131 occurrences as the sole taxon in the sister group to Haloarchaea and 169 occurrences in the mixed sister group cases, followed by α-proteobacteria (88 sole plus 97 mixed), γ-proteobacteria (51 sole plus 111 mixed), and δ-proteobacteria (53 sole plus 100 mixed).

Function of Imported Genes. Trees generated from 56 recipient genes present as a single copy in all archaebacteria place the Haloarchaea branching from within the methanogens, but not specifically as sisters to the Methanosarcinales (Fig. 1). Rather, the Haloarchaea appear to have emerged from simpler and more primitive methanogens, ones that lack both cytochromes and methanophenazine (5). Methanogens that lack cytochromes and methanophenazine are capable only of H_2-dependent methanogenesis, and have a single coupling site in their energy metabolism (5, 40). Haloarchaea have a respiratory chain with several coupling sites (1). Methanogens are strict autotrophs and strict anaerobes (5), whereas Haloarchaea are heterotrophs and can use O_2 as their terminal acceptor. Thus, the essential metabolic functional units for transforming a methanogen into the haloarchaeal common ancestor are (*i*) membrane transporters for reduced carbon compounds; (*ii*) a heterotrophic carbon metabolism that directs the oxidation of organic substrates to support carbon and energy metabolism; (*iii*) a respiratory chain for terminal oxidation and chemiosmotic ion pumping; and (*iv*) genes for the synthesis of any additional cofactors required, for example menaquinone, the quinone universally present in all halophiles (41). Those four essential functional units are very clearly represented within the eubacterial imports in haloarchaeal genomes.

Among the 1,089 haloarchaeal imports from eubacteria almost half (482, 44%) of the imports are related to metabolism, with amino acid transport and metabolism (114) and energy conversion (95) being the most abundant classes, followed by inorganic ion transport and metabolism (86) (Table S1; Fig. S2 *B* and *C*). Whereas methanogens without cytochromes grow on gases, which traverse membranes freely without transporters, Haloarchaea abound in eubacterial transporters: 157 of the acquired families are annotated as permease, importer, or transporter. Although the true substrate spectrum of these transporters is yet unknown, 49 trace to amino acid or carbohydrate metabolism (Tables S1 and S2), and they operate in a membrane consisting of typical archaebacterial lipids (1).

Methanogens cannot use exogenous carbohydrates for growth (5, 42); their sugar synthetic pathways are anabolic, whereas carbon metabolism in Haloarchaea runs in the catabolic direction. For a methanogen to become heterotrophic, it needs to acquire the enzymes underpinning the heterotrophic lifestyle from a heterotrophic donor (43). Among the eubacterial genes imported into Haloarchaea are pyruvate kinase, glucose-6-phosphate isomerase, phosphoglyceromutase, 6-phosphogluconate dehydrogenase, the eubacterial type fructose 1,6-bisphosphatase, as well as genes for 2-keto-3-deoxy-6-phosphogluconate aldolase of the Entner–Doudoroff pathway. Eubacterial enzymes of pyruvate breakdown were also found, including two copies of pyruvate:ferredoxin oxidoreductase, and genes for pyruvate dehydrogenase complex E1 and E2 subunits.

Earlier studies showed that five haloarchaeal respiratory chain components are eubacterial acquisitions in two Haloarchaea (15). Fig. 3 shows that most of the 11 subunits of NADH dehydrogenase (complex I) are present in all 10 Haloarchaea. Complexes I–III require quinones. Haloarchaea possess the naphthoquinone menaquinone (41) and several of the imported genes are involved in menaquinone biosynthesis, including menA. Finally, among the imported genes, 26 are annotated as transcriptional regulators and 8 are annotated as chaperones, including members of the DnaJ family.

Conclusion

Were these 1,000 genes accrued in the haloarchaeal ancestor one by one or in a single mass acquisition? The former possibility is unlikely, because in the absence of corresponding interaction partners to form functional complexes, individual protein subunits of catabolic carbon metabolism, the respiratory chain, or cofactor biosynthesis lack selectable function, which would allow them to become fixed in a methanogenic recipient. This argues in favor of mass transfer of genes for the entire pathways and complexes over a short period of evolutionary time. The origin of Haloarchaea was thus an evolutionary leap that transformed a methanogenic host into an oxygen-respiring heterotroph—the founder haloarchaeon. A possible context of that cellular association is anaerobic syntrophy (44, 45), that is, a H_2-producing heterotrophic bacterial donor in association with a H_2-

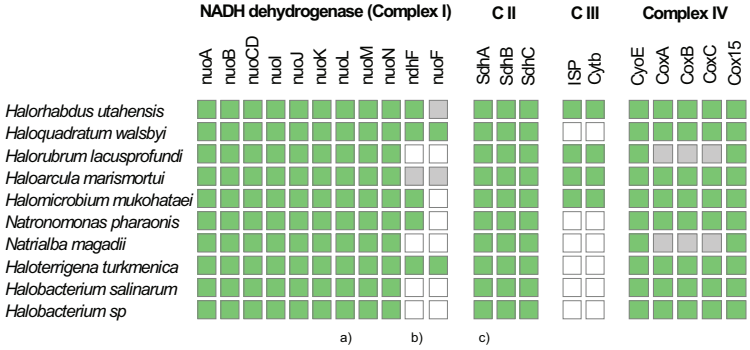


Fig. 3. Eubacterial respiratory chain components in Haloarchaea. Green boxes indicate presence of the gene in the corresponding Haloarchaea genome and that the gene is more similar to eubacterial than to archaebacterial homologs in the corresponding phylogenetic trees. Gray boxes indicate that homologs can be detected in the corresponding genome by BLAST searches, but that the clustering procedure did not included them within the 16,061 archaeal clusters. White boxes indicate that no homolog was detected. (*A*) Haloarchaeal nuoL sequences are monophyletic but an additional paralogous copy is present in *Halorhabdus*. (*B*) *Salinibacter* has acquired a copy of ndhF from Haloarchaea, which are otherwise monophyletic. (*C*) Haloarchaeal sdhA sequences are monophyletic but additional paralogous copies of eubacterial origin are present in several genomes (see also Table S4).

dependent methanogenic recipient. Anaerobic syntrophy is common in nature and has been suggested as the selective force at the origin of eukaryotes (43, 46). If similar processes underlie the origin of haloarchaea and eukaryotes, why did Haloarchaea remain prokaryotic, whereas eukaryotes became complex? The main physiological difference between Haloarchaea and eukaryotes concerns the location of the bioenergetic membrane. In Haloarchaea it is the archaebacterial plasma membrane (1). In eukaryotes it is the mitochondrial inner membrane—the key to eukaryote genome complexity (47). Mitochondria afforded ancient eukaryotes many orders of magnitude more energy per gene than their prokaryotic ancestors. That boost surmounted the energetic constraints imposed by reliance upon the cytoplasmic membrane as the source of chemiosmotic potential, thus allowing eukaryotic genomes and proteomes to expand freely, resulting in eukaryotic cell complexity (47). The Haloarchaea have long figured into issues of early microbial evolution (48). From the standpoint of genome chimaerism, they now appear to have undergone the very same physiological transformation as the eukaryotes, and the kind of gene transfer involved—from symbionts to the host chromosomes—is still ongoing in eukaryotic cells today (49). Haloarchaea remained prokaryotic because they failed to preserve a genome-containing bioenergetic organelle.

Methods

Data. Completely sequenced genomes of 1,153 microbial species were downloaded from the National Center for Bioinformatics Information (NCBI) website (www.ncbi.nlm.nih.gov). This includes 75 archaebacterial genomes (version April 2010) and 1,078 eubacterial genomes (version September 2010). Taxonomic classification of the species was downloaded from the NCBI Taxonomy database (www.ncbi.nlm.nih.gov/Taxonomy/).

Clusters of Homologous Proteins. Clusters of homologous proteins were reconstructed from a total of 172,531 proteins encoded within the archaeal chromosomes. An all-against-all genomes BLAST (50) yielded 147,071 reciprocal best BLAST hits (rBBH) (51) using E value $<10^{-10}$ and ≥30% amino acid identity as a threshold. Protein pairs were globally aligned using the Needleman–Wunsch algorithm with *needle* program (EMBOSS package) (52). A total of 137,022 protein pairs having global amino acids identities ≥30% were clustered into protein families using the MCL algorithm (31) with default parameters. This yielded a total of 16,061 archaeal protein families of ≥2 proteins. The remaining 35,509 proteins were classified as singletons. Eubacterial homologs to archaeal proteins were found using an rBBH analysis as described above, which yielded 8,451 archaeal protein families having one or more eubacterial homologs. The functional classification of protein families was based on the eukaryotic orthologous groups database (KOG) database (53). Protein families that overlapped with KOG clusters were annotated to the same function as the matching KOG. The remaining protein families were manually classified by sequence similarity to known KOGs using the KOGnitor tool (http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html). The haloarchaeal respiratory chain component genes were identified from the Kyoto Encyclopedia of Genes and Genomes database (http://www.genome.jp/kegg/).

Phylogenetic Trees. Protein families were aligned using MAFFT (multiple alignment using fast Fourier transform) (54), and trees were reconstructed using Phyml (55) with the best fitting model in individual trees as inferred by ProtTest3 (56) using the AIC measure. An archaebacterial reference tree was reconstructed from a weighted concatenated alignment of 56 archaebacterial single copy universal genes using Phyml with the IG+I+G model, which was the most frequent best fitting model, rooted using *Nanoarchaeota* and *Koarchaeota* as an outgroup. Trees of recipient genes were reconstructed from sequences of all 10 Haloarchaea and one nonhaloarchaeal sequence using the same procedure. For polarizing the direction of gene transfers, the root of Jain et al. (33) was used.

Reconstruction of Lateral Gene Transfer Events. Eubacterial acquisitions within halophilic archaeal genomes were identified by presence absence pattern (PAP) analysis and BLAST protein sequence similarity searches. Of the total 8,451 bacterial-like protein families in archaebacteria 1,479 had ≥2 Haloarchaea species. Of these, 952 do not possess other nonhaloarchaeal homologs in the same families and correspond to unique acquisitions within Haloarchaea from eubacterial species. Archaebacterial xenologous genes that were replaced by a eubacterial acquisition are expected to be more similar to their eubacterial ancestors than to their orthologs in other archaebacterial species (57). Putative replaced halophilic proteins were identified by comparing the E value of their BBHs within eubacterial and archaebacterial genomes. Proteins having a eubacterial BBH of lower E value than that of the archaebacterial BBH were classified as putative acquisitions from eubacteria, corresponding to 527 protein families. All 1,479 protein families were aligned with their eubacterial homologs including the three best eubacterial hits per archaebacterial protein (but excluding redundant eubacterial sequences), and phylogenies were reconstructed as described above. The trees were classified into groups by the branching topology of Haloarchaea and eubacteria using an in-house PERL script. A group is considered as monophyletic for Haloarchaea if there exists a bipartition (branch) in the tree that splits between Haloarchaea and the rest. Single eubacterial sequences branching with the haloarchaeal clade, and vice versa were tested manually. In each tree, the branch connecting the monophyletic Haloarchaea clade to the eubacteria serves to split the eubacteria clade into two groups, the nearest neighbor of Haloarchaea was assigned as described in Thiergart et al. (36).

Comparison of Tree Sets. Two sets of trees were compared using a χ^2 goodness-of-fit test (58), operating on a $2\times m$ contingency table. The m cells were defined in an adaptive procedure as follows. The two samples were pooled together into a single set of size n, and the n trees converted into splits. Each split was ranked according to its frequency in the pooled split sets. Each tree was labeled by its lowest ranking split, and the pooled tree set was sorted by this label. Cells were defined as a collection of split ranks by sequential addition of split ranks from the sorted list, and creation of a new cell when the current cell included at least $\sqrt{n}$ trees, resulting in $m \leq \sqrt{n}$ cells. In the last step, trees from the two sets were added to a $2\times m$ contingency table based on their least ranked split. We have studied the adaptive cell procedure and goodness-of-fit testing in a series of permutation analyses, and the resulting χ^2 test proved to be an unbiased α-level test (*SI Text*, Table S5, and Figs. S3 and S4).

Phylogenetic Compatibility with a Reference Set. Two sets of trees were compared by their compatibility with a reference set of trees. Each n taxon tree was decomposed into its $(n-3)$ splits, and each split was scored by the

fraction of splits in the reference set that are phylogenetically compatible with it. The *(n-3)* split compatibility scores were averaged to produce a tree compatibility score. The distributions of the tree compatibility scores for the two sets of trees was compared using the Kolmogorov–Smirnov test (58) (*SI Text*).

ACKNOWLEDGMENTS. We thank Martin Embley and Dan Graur for critical comments on an earlier version of the manuscript. We thank the central computing resources of the University of Düsseldorf for technical support. G.L. acknowledges financial support from the University of Düsseldorf rectorate, W.F.M. and T.D. are funded by the European Research Council and the German Ministry of Science and Education, J.O.M. is funded by the Science Foundation of Ireland, M.S. is funded by the Allan Wilson Centre and the Alexander von Humboldt Foundation, and A.J. and U.D. are funded by the German Research Foundation.

1. Oren A (2006) Life at high salt concentrations. *The Prokaryotes* 3:263–282.
2. Makarova KS, Yutin N, Bell SD, Koonin EV (2010) Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat Rev Microbiol* 8(10):731–741.
3. Kelly S, Wickstead B, Gull K (2011) Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc Biol Sci* 278(1708):1009–1018.
4. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6(3):245–252.
5. Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R (2008) Methanogenic archaea: Ecologically relevant differences in energy conservation. *Nat Rev Microbiol* 6 (8):579–591.
6. Thauer R (1998) Biochemistry of methanogenesis: A tribute to Marjory Stephensen. *Microbiol Aust* 144:2377–2406.
7. Ferry JG (2010) How to make a living by exhaling methane. *Annu Rev Microbiol* 64: 453–473.
8. Oren A (2006) The order Halobacteriales. *The Prokaryotes: A Handbook on the Biology of Bacteria* (Springer, New York), 113–164.
9. Li N, Cannon MC (1998) Gas vesicle genes identified in *Bacillus megaterium* and functional expression in *Escherichia coli*. *J Bacteriol* 180(9):2450–2458.
10. Khomyakova M, Bükmez O, Thomas LK, Erb TJ, Berg IA (2011) A methylaspartate cycle in haloarchaea. *Science* 331(6015):334–337.
11. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S (2001) Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res* 11(10):1641–1650.
12. Lemos RS (2002) Quinol:fumarate oxidoreductases and succinate:quinone oxidoreductases: Phylogenetic relationships, metal centres and membrane. *Biochim Biophys Acta* 1553:1–13.
13. Baymann F, Schoepp-Cothenet B, Lebrun E, van Lis R, Nitschke W (2012) Phylogeny of Rieske/cytb complexes with a special focus on the Haloarchaeal enzymes. *Genome Biol Evol* 4(8):720–729.
14. van Ooyen J, Soppa J (2007) Three 2-oxoacid dehydrogenase operons in *Haloferax volcanii*: Expression, deletion mutants and evolution. *Microbiology* 153(Pt 10): 3303–3313.
15. Boucher Y, et al. (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* 37:283–328.
16. Ichiki H, et al. (2001) Purification, characterization, and genetic analysis of Cu-containing dissimilatory nitrite reductase from a denitrifying halophilic archaeon, *Haloarcula marismortui*. *J Bacteriol* 183(14):4149–4156.
17. Pfeifer F, Griffig J, Oesterhelt D (1993) The fdx gene encoding the [2Fe—2S] ferredoxin of *Halobacterium salinarium* (*H. halobium*). *Mol Gen Genet* 239(1–2):66–71.
18. Bickel-Sandkotter S, Gartner W, Dane M (1996) Conversion of energy in halobacteria: ATP synthesis and phototaxis. *Arch Microbiol* 166(1):1–11.
19. Boyd JM, Drevland RM, Downs DM, Graham DE (2009) Archaeal ApbC/Nbp35 homologs function as iron-sulfur cluster carrier proteins. *J Bacteriol* 191(5):1490–1497.
20. Ng WV, et al. (2000) Genome sequence of *Halobacterium* species NRC-1. *Proc Natl Acad Sci USA* 97(22):12176–12181.
21. Rhodes ME, Spear JR, Oren A, House CH (2011) Differences in lateral gene transfer in hypersaline versus thermal environments. *BMC Evol Biol* 11:199.
22. Baliga NS, et al. (2004) Genome sequence of *Haloarcula marismortui*: A halophilic archaeon from the Dead Sea. *Genome Res* 14(11):2221–2234.
23. Pfeiffer F, et al. (2008) Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. *Genomics* 91(4):335–346.
24. Tindall BJ, et al. (2009) Complete genome sequence of Halomicrobium mukohataei type strain (arg-2). *Stand Genomic Sci* 1(3):270–277.
25. Bolhuis H, et al. (2006) The genome of the square archaeon *Haloquadratum walsbyi* : Life at the limits of water activity. *BMC Genomics* 7:169.
26. Anderson I, et al. (2009) Complete genome sequence of Halorhabdus utahensis type strain (AX-2). *Stand Genomic Sci* 1(3):218–225.
27. Franzmann P, Stackebrandt E (1988) CSA. *Halobacterium lacusprofundi* sp. nov., a halophilic bacterium isolated from Deep Lake, Antarctica. *Syst Appl Microbiol* 11: 20–27.
28. Kamekura M, Dyall-Smith ML, Upasani V, Ventosa A, Kates M (1997) Diversity of alkaliphilic halobacteria: Proposals for transfer of *Natronobacterium vacuolatum, Natronobacterium magadii*, and *Natronobacterium pharaonis* to *Halorubrum, Natrialba*, and *Natronomonas* gen. nov., respectively, as *Halorubrum vacuolatum* comb. nov., *Natrialba magadii* comb. nov., and *Natronomonas pharaonis* comb. nov., respectively. *Int J Syst Bacteriol* 47(3):853–857.
29. Falb M, et al. (2005) Living with two extremes: Conclusions from the genome sequence of *Natronomonas pharaonis*. *Genome Res* 15(10):1336–1343.
30. Saunders E, et al. (2010) Complete genome sequence of *Haloterrigena turkmenica* type strain (4k). *Stand Genomic Sci* 2(1):107–116.
31. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.
32. Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284 (5423):2124–2129.
33. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci USA* 96(7):3801–3806.
34. Treangen TJ, Rocha EPC (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7(1):e1001284.
35. Abhishek A, Bavishi A, Bavishi A, Choudhary M (2011) Bacterial genome chimaerism and the origin of mitochondria. *Can J Microbiol* 57(1):49–61.
36. Thiergart T, Landan G, Schenk M, Dagan T, Martin WF (2012) An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol* 4(4):466–485.
37. Lane CE, Archibald JM (2008) The eukaryotic tree of life: Endosymbiosis takes its TOL. *Trends Ecol Evol* 23(5):268–275.
38. Deusch O, et al. (2008) Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol* 25(4):748–761.
39. Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 105 (29):10039–10044.
40. Kaster A-K, Moll J, Parey K, Thauer RK (2011) Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. *Proc Natl Acad Sci USA* 108(7):2981–2986.
41. Collins MD, Jones D (1981) Distribution of isoprenoid quinone structural types in bacteria and their taxonomic implication. *Microbiol Rev* 45(2):316–354.
42. Siebers B, Schönheit P (2005) Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr Opin Microbiol* 8(6):695–705.
43. Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392(6671):37–41.
44. Schink B (1997) Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* 61(2):262–280.
45. Stams AJM, Plugge CM (2009) Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nat Rev Microbiol* 7(8):568–577.
46. Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440(7084):623–630.
47. Lane N, Martin W (2010) The energetics of genome complexity. *Nature* 467(7318): 929–934.
48. Lake JA, et al. (1985) Eubacteria, halobacteria, and the origin of photosynthesis: The photocytes. *Proc Natl Acad Sci USA* 82(11):3716–3720.
49. Wang D, Lloyd AH, Timmis JN (2012) Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. *Proc Natl Acad Sci USA* 109(7): 2444–2448.
50. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
51. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278(5338):631–637.
52. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.
53. Tatusov RL, et al. (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
54. Katoh K, Misawa K, Kuma K-I, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30 (14):3059–3066.
55. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.
56. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164–1165.
57. Deppenmeier U, et al. (2002) The genome of *Methanosarcina mazei*: Evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* 4(4): 453–461.
58. Zar JH (2010) *Biostatistical Analysis* (Prentice Hall, Upper Saddle River, NJ), 5th Ed.

3.2

Networks uncover hidden lexical borrowing in Indo-European language evolution

Shijulal Nelson-Sathi[1], Johann-Mattis List[2], Hans Geisler[2], Heiner Fangerau[3], Russell D Gray[4], William Martin[1], Tal Dagan[1]

[1] Institute of Botany III and [2] Faculty of Philosophy, Heinrich-Heine University Düsseldorf, Germany
[3] Institute for the History, Philosophy and Ethics of Medicine, Ulm University, Germany
[4] Department of Psychology, University of Auckland, Auckland 1142, New Zealand

Corresponding author: Tal Dagan, Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany, Tel: +49 211 811 2736 Fax: +49 211 811 3554, e-mail: tal.dagan@uni-duesseldorf.de

The presented manuscript was published in the Journal of "*Proceedings of the Royal Society B: Biological Sciences*"(*ProcB*), 2011 June.

Impact Factor – 5.4

Contribution of Shijulal Nelson-Sathi – First author

Experimental design: 30%
Execution and analysis of experiments: 80%
Manuscript writing: 35%

PROCEEDINGS OF THE ROYAL SOCIETY B

FirstCite® e-publishing

Proc. R. Soc. B
doi:10.1098/rspb.2010.1917
Published online

Networks uncover hidden lexical borrowing in Indo-European language evolution

Shijulal Nelson-Sathi[1], Johann-Mattis List[2], Hans Geisler[2], Heiner Fangerau[3], Russell D. Gray[4], William Martin[1] and Tal Dagan[1,*]

[1]*Institute of Botany III, and* [2]*Faculty of Philosophy, Heinrich-Heine University Düsseldorf, Germany*
[3]*Institute of the History, Philosophy and Ethics of Medicine, Ulm University, Germany*
[4]*Department of Psychology, University of Auckland, Auckland 1142, New Zealand*

Language evolution is traditionally described in terms of family trees with ancestral languages splitting into descendent languages. However, it has long been recognized that language evolution also entails horizontal components, most commonly through lexical borrowing. For example, the English language was heavily influenced by Old Norse and Old French; eight per cent of its basic vocabulary is borrowed. Borrowing is a distinctly non-tree-like process—akin to horizontal gene transfer in genome evolution—that cannot be recovered by phylogenetic trees. Here, we infer the frequency of hidden borrowing among 2346 cognates (etymologically related words) of basic vocabulary distributed across 84 Indo-European languages. The dataset includes 124 (5%) known borrowings. Applying the uniformitarian principle to inventory dynamics in past and present basic vocabularies, we find that 1373 (61%) of the cognates have been affected by borrowing during their history. Our approach correctly identified 117 (94%) known borrowings. Reconstructed phylogenetic networks that capture both vertical and horizontal components of evolutionary history reveal that, on average, eight per cent of the words of basic vocabulary in each Indo-European language were involved in borrowing during evolution. Basic vocabulary is often assumed to be relatively resistant to borrowing. Our results indicate that the impact of borrowing is far more widespread than previously thought.

Keywords: community structure; lateral transfer; phylogenetics

1. INTRODUCTION

Genome evolution and language evolution have a lot in common. Both processes entail evolving elements—genes or words—that are inherited from ancestors to their descendants. The parallels between biological and linguistic evolution were evident both to Charles Darwin, who briefly addressed the topic of language evolution in *The origin of species* [1], and to the linguist August Schleicher, who in an open letter to Ernst Haeckel discussed the similarities between language classification and species evolution [2]. Computational methods that are currently used to reconstruct genome phylogenies can also be used to reconstruct evolutionary trees of languages [3,4]. However, approaches to language phylogeny that are based on bifurcating trees recover vertical inheritance only [3,5–7], neglecting the horizontal component of language evolution (borrowing). Horizontal interactions during language evolution can range from the exchange of just a few words to deep interference [8]. In previous investigations, which focused only on the component of language evolution that is described by a bifurcating tree [3,5–7], the extent of borrowing might therefore have been overlooked.

Lexical borrowing is the transfer of a word from a donor language to a recipient language as a result of a certain kind of contact between the speakers of the two languages [9]. This is one of the most common types of interaction between languages. Lexical borrowing can be reciprocal or unidirectional, and occurs at variable rates during evolution. Factors affecting the rate of lexical borrowing during evolution include the intensity of contact between the speakers of the respective languages, the genetic or typological closeness of the languages (which facilitates the inclusion of foreign words), the amount of bi- or multi-lingual speakers in the respective linguistic communities, or a combination thereof [10,11]. For example, English has been heavily influenced throughout its history by different languages such as Old Norse and Old French [12], it has been estimated that 8 per cent of its basic vocabulary is borrowed from those languages [13]. Icelandic, on the other hand, has preserved most of its original words [14].

A key part of inferences in historical linguistics is the identification of cognate sets. These are sets of words from different languages that are etymologically related. The words in a cognate set are derived from a single common ancestral form that was present in an ancestral language. Cognate judgement is an arduous enterprise since it includes the complete evolutionary reconstruction of all words in the sampled languages for a certain concept. Historical linguists usually make use of an in-depth analysis of structural resemblances between the

* Author for correspondence (tal.dagan@uni-duesseldorf.de).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2010.1917 or via http://rspb.royalsocietypublishing.org.

Received 6 September 2010
Accepted 3 November 2010

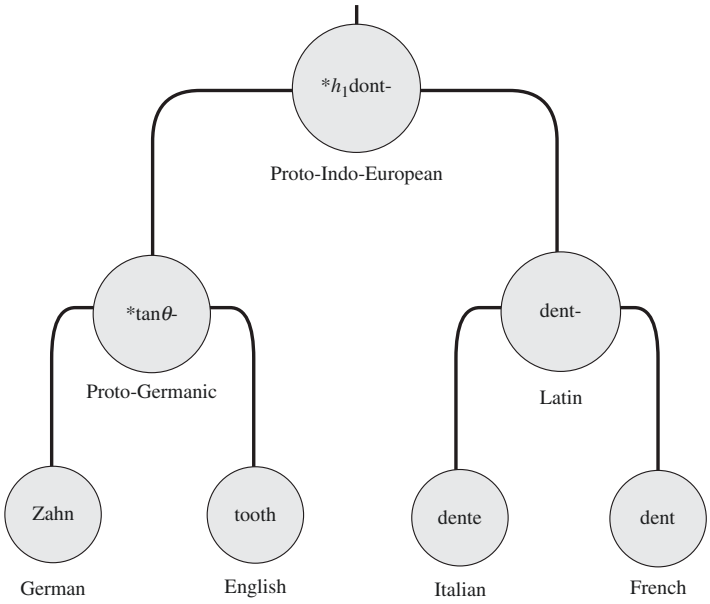


Figure 1. Etymological reconstruction of the concept tooth. The English and German word forms have descended from the Proto-Germanic ancestor [52]. The Italian and French words are descendants of Latin, and the Proto-Germanic and Latin forms stem from Proto-Indo-European [43,53].

word forms, looking for sound correspondences in specific environments. The identification of a cognate is thus much more than just a hunt for resemblant forms or 'lookalikes'. Only a set of words that have regular sound correspondences provide good evidence for genealogical relatedness and thus only these words can be grouped into a single cognate set (COG). For example, the concept 'tooth' has a cognate set that unites English *tooth*, German *Zahn*, Italian *dente* and French *dent* as etymologically related (figure 1). However, similar word forms can arise not only by inheritance, but also by lexical borrowing. Unfortunately, the further we go back in time, the more difficult it becomes to distinguish inheritance from transfer, and reconstructed COGs may include hidden borrowing events that are erroneously coded as vertical inheritance.

Lexical borrowing is a non-tree-like evolutionary event that cannot be reconstructed using phylogenetic trees that are common in evolutionary biology [15,16]. Linguists have long been aware of the problems that borrowing introduces. At about the same time that Darwin suggested the tree metaphor for the evolution of species in 1859 [1], August Schleicher introduced the family tree to linguistics [17]. Few years later, his model was rejected by several scholars arguing against the use of a simple tree model to describe the evolution of languages, which they noted to be reticulated by nature [18,19]. Other non-tree-like models were proposed by linguists to study language evolution—including waves [18,20] and networks [21]—but they lacked either quantitative parameters, historical dimensions or both. At the other extreme, quantitative estimates for language divergence lacked an explicit model to explain language relatedness [22,23]. Apart from some sporadic attempts to visualize language evolution of specific words by a combination of a bifurcating family tree with the non-tree-like component superimposed on it [24], linguists have, for lack of better alternatives, largely stuck to the tree model, while emphasizing its inadequacies.

Phylogenetic methods that were developed to take into account horizontal transfer of genes during microbial evolution offer an alternative model for the horizontal aspects of language evolution. Recent years have witnessed several applications of reticulated trees and split networks to language evolution [25–28], yet none of these have either specifically uncovered borrowing events or delivered an estimate for the borrowing frequency during language evolution. Here, we apply phylogenetic networks to recover the frequency of hidden borrowings during the evolution of Indo-European languages using the criterion of word inventory dynamics over time, proposing a general model for language evolution that includes both vertical and horizontal components of word transfer during evolution.

2. METHODS

(a) *Data*

Here, we used two publicly available cognate datasets: Dyen [29] and Tower of Babel (ToB) [30]. For the analysis, all COGs in both datasets are converted into a binary presence/absence pattern (PAP). A PAP within the Dyen dataset includes 84 digits; if a cognate set includes one or more words from language i, then digit x_i in its corresponding pattern is '1'; otherwise, it is '0'. The same conversion method is used for the ToB dataset where the PAPs include 73 digits.

(b) *Shared COGs network*

The number of shared COGs between each language pair is calculated as the number of cognate sets in which both languages are present. A division of the network into modules

is based on maximizing a modularity function defined as the number of edges within a community minus the expected number of edges [31]. Initially, an optimal division into two components is found by maximizing this function over all possible divisions by using spectral optimization, which is based on the leading eigenvector of the matching modularity matrix. To further subdivide the network into more than two modules, additional subdivisions are made, each time comparing the contribution of the new subdivision with the general modularity score of the entire network. This process is carried out until there are no additional subdivisions that will increase the modularity of the network as a whole [31].

(c) *Reference trees*

Language trees were inferred by a Bayesian approach using MrBayes [32] as detailed by Gray & Atkinson [3]. In addition, neighbour-joining (NJ) trees [33] were reconstructed from Hamming distances using SplitsTree [34]. A reference tree with English internal to the Germanic clade was produced manually from the Bayesian tree. A randomized reference tree for the Dyen dataset was produced by randomizing the language names in the Bayesian reference tree. Trees are available in Newick format at http://www.molevol.de/resources.

(d) *Borrowing models and the minimal lateral network*

In the loss-only (LO) model, all COGs are assumed to have originated at the root of the reference tree. The loss events for each COG are estimated by using a binary recursive PERL algorithm that scans the reference tree and infers the minimum number of losses [35]. When a COG is absent in a whole clade, a single loss event is inferred in the common ancestor of that clade. In the single-origin (SO) model, each cognate is assumed to have originated at its first occurrence on the reference tree. A binary recursive algorithm scans the reference tree from root to tips to identify the first ancestral node that is the common ancestor of all cognate 'present' cases.

In the BOR1 model, each cognate is allowed to have two word origins, where one is a borrowing. A preliminary origin is inferred as in the SO model, followed by researching for a cognate origin in each of the two clades branching from the preliminary origin node. If the hypothetical taxonomic unit that was inferred as the preliminary origin has no cognate 'absent' descendants, the cognate is inferred to have an SO. Once the nodes of the two origins are set, losses are inferred as in the LO model.

We tested additional models allowing four, eight and 16 origins, where one is an origin, and the rest are borrowings. These are implemented in the same way as in the BOR1 model, except that the origin search is iterated. For example, a search for origins under the BOR3 model entails (i) a search for a preliminary origin (as in the SO model), (ii) a search for the next origin in descendants (as in the BOR1 model) and, (iii) for each next origin, another search. If an origin has no cognate-absent descendants, the number of origins inferred is smaller than the maximum allowed. Ancestral vocabulary size at a certain internal node is inferred as the total COG origins that were inferred to occur at that node. The distributions of ancestral and modern vocabulary sizes were compared by using the Wilcoxon non-parametric test [36].

The minimal lateral network (MLN) [37] is calculated for each dataset by the allowance model that was statistically accepted by the test described above. The MLN comprises the reference tree, with additional information of the vocabulary size in all internal nodes. Lateral cognate sharing among internal and external nodes is summarized in a 167×167 matrix that includes all tree nodes, where $a_{ij} = a_{ji} =$ number of laterally shared COGs between nodes i and j. The MLN is then depicted by an in-house script using MATLAB.

3. RESULTS AND DISCUSSION

(a) *Community structure in the network of shared cognate sets*

For the study of evolution by borrowing, we analysed two independent, publicly available collections of cognate sets from Indo-European languages. Both datasets comprise words from individual languages or dialects corresponding to concepts that are included in Swadesh lists [38]. Basic concepts are expressed by simple words rather than compounds or phrases and contain names for body parts, pronouns, common verbs and numerals, but exclude technological words and words related to specific ecologies or habitats. Words expressing basic concepts are supposed to exist in all languages and thus may serve as a *tertium comparationis* for language comparison [39]. Moreover, basic concepts are rarely replaced by other words, either through external (lexical borrowing) or internal factors (semantic shift) [13,16].

The Dyen dataset [29] includes word forms for 84 languages (including Greek, Armenian, Celtic, Romance, Germanic, Slavic, Albanian and Indo-Iranian languages) corresponding to 200 basic vocabulary concepts [39] sorted into 2346 COGs [3]. While obvious borrowings were excluded in the original Dyen dataset [29], we used an edited version where 124 marked borrowings are coded into their respective COGs [25]. Detailed reinspection of Romance cognates revealed an additional six hidden borrowings [40] (electronic supplementary material, table S1).

The second dataset is based on etymological dictionaries and Swadesh lists published by the ToB project [30]. It is based on word forms for 110 basic vocabulary items for a total of 98 languages from which we extracted 73 contemporary ones, including languages from the Celtic, Romance, Germanic, Slavic, Albanian and Indo-Iranian branches of Indo-European, sorted into 722 COGs. Detectable borrowings were excluded in the original database; however, a recent detailed screening revealed five undetected borrowings within Romance languages [40].

A network analysis of the distribution of cognate word forms across Indo-European languages should provide new insights into the frequency and distribution of borrowing in Indo-European language history. Networks are mathematical structures used to model pairwise relations between entities. The entities are called vertices and they are linked by edges that represent the connections or interactions between the vertices. A network of N vertices can be fully defined by the matrix $A = [a_{ij}]_{N^* N}$, with $a_{ij} = a_{ji} \neq 0$ if a link exists between nodes i and j, and $a_{ij} = a_{ji} = 0$ otherwise. In the study of Indo-European languages, each language is represented by a vertex, i, whereas the elements of the matrix, A,

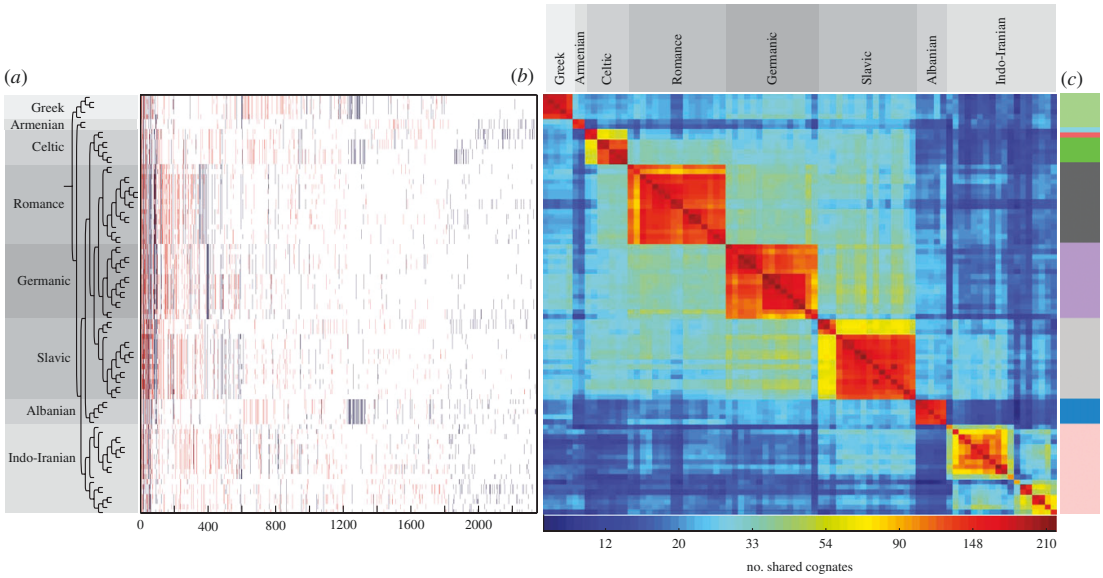


Figure 2. Modules in the shared COGs network. (*a*) A graphic representation of cognate PAPs. Languages are sorted by their order on the reference phylogenetic tree [3]. COGs are sorted by their size in ascending order. A presence case of a certain COG in a certain language is coloured in blue if the COG pattern is congruent with the tree branching patterns and red otherwise. (*b*) A matrix representation of the shared COGs network in Indo-European languages. Cells in the matrix are edges in the network. Edges are colour-coded by the frequency of shared cognate according to the colour bar at the bottom. The languages in the matrix are sorted by order of appearance in the phylogenetic tree on the left. (*c*) Modules within the shared COGs network. Languages included in the same module are coloured in the same colour.

correspond to the number of shared cognate sets between language pairs, a_{ji}. Cognate sharing can result either from vertical inheritance or from borrowing.

For network reconstruction, cognate sets were converted into a binary format of PAPs for each COG in each language [3]. For the 2346 COGs in the Dyen dataset [29], 1169 different PAPs were observed, of which 942 (80%) are unique and 227 are recurring (figure 2*a*). Closely related languages typically share the most frequent PAPs. For example, Panjabi and Lahnda, two Indian languages, share 78 cognates that are unique to both languages. The ToB dataset includes 532 different PAPs, none of which are unique (electronic supplementary material, figure S1). The frequency of shared COGs among languages in the main branches uncovers components of both inheritance and borrowing.

The binary PAPs of the Dyen COGs are readily assorted into an 84×84 matrix representation of the cognate-sharing network that consists of vertices (languages) connected by edges (shared cognates), the edge weights are the number of shared cognates per vertex pair. There are 3486 edges in the network, all vertices of which are connected, thereby forming a 'clique' in network terms (figure 2*b*). Some groups of languages are more strongly interconnected among themselves than with others in the cognate-sharing network, thereby forming communities.

We examined the community structure in the network by division into modules [31,41]. Modules correspond to 'natural' groups within a network, that is, groups of vertices that are more highly connected to each other than they are to other vertex sets. With only two exceptions, the nine modules calculated within the cognate-sharing network correspond exactly to the main branches of Indo-European languages. One exception concerns the Armenian dialects Adapazar (*Armenian List* in Dyen dataset [42]) and eastern modern Armenian (*Armenian Mod* in Dyen dataset [42]), which are grouped with the Greek languages into one module. This is because Armenian shares significantly ($p \ll 0.01$, using the Wilcoxon test) more cognates with the Greek languages (30 ± 2, $n = 5$) than with the other languages (22 ± 3, $n = 79$). This module has been independently recognized by linguists [43]. The other exception is the split of both Irish dialects from Celtic (figure 2*c*). The same network-based analysis of the ToB dataset yields only four modules: (i) Slavic and Albanian; (ii) Armenian, Greek, Celtic, Germanic and Romance; (iii) Indo; and (iv) Iranian (electronic supplementary material, figure S2).

Language communities that do not correspond to monophyletic clades in the tree are the result of patchy COG distributions that could not be reconciled with the phylogenetic tree. For example, Romani, which branches with Indo-Iranian languages, shares 25 COGs with Modern Greek, such as the COGs for 'flower' (Modern Greek: λουλούδι (*louloudi*); Romani: *lulugi*) and 'because' (Modern Greek: επειδή (*epeide*); Romani: *epidhi*). Since the Romani dialect in the Dyen dataset [29] is a variety spoken in Greece [42], these are probably borrowed from Greek to Romani.

(b) *Borrowing frequency during Indo-European language evolution*

In the Dyen dataset, there are 1391 (59%) patchily distributed PAPs that are incongruent with the tree

branching pattern (figure 2*a*). In principle, such patchy COG distributions could arise solely through independent parallel evolution, through vertical inheritance from the common ancestor of all languages and differential loss of lexica during language evolution, or via lexical borrowing among languages. The first possibility seems sufficiently unlikely as to exclude *a priori*. There is no clear estimation for the frequency of parallel evolution during language evolution, but we can assume that it is rather rare and cannot, therefore, be used to explain the distribution pattern of all patchy COGs. If we invoke the second scenario to explain all COGs of patchy distribution, then the result is a common ancestral language that includes each and every COG existing in contemporary languages. In order to entertain such a claim, one would have to assume that the proto-language employed many different, but redundant, words for the same basic concepts, far more than every known contemporary language. This runs contrary to uniformitarianism, a key principle in historical sciences such as geology, biology and linguistics, which states that processes in the past should not be assumed to differ fundamentally from those observed today [44,45]. Hence, if ancient and modern languages were of similar nature, then the number of words that were used to express fundamental concepts (basic vocabulary size) in ancestral languages should be similar to that used in contemporary languages. This principle can be used to infer the minimum amount of lexical borrowing in Indo-European languages that is required in order to bring the distribution of basic vocabulary size in ancestral languages into agreement with that of contemporary languages.

This network method to address non-tree-like patterns of shared characters requires the use of a reference tree [37]. Here, we use a phylogenetic tree reconstructed by a Bayesian approach [3]. First, we designate an evolutionary scenario that uses vertical inheritance and LO (model), according to which current COG distribution is governed solely by loss. Each ancestral language contains all cognates present in its descendants, and vocabulary size hence becomes progressively larger back through time (figure 3*a*). Note that a loss event applies only to the sample of basic vocabulary and does not mean a loss from the language as a whole. With the Dyen dataset [29] and the reference tree, the common Indo-European ancestor would have had a vocabulary size of 2346 for basic words, expressing 200 basic concepts. This estimate is 11 times larger than the average basic vocabulary size in our sample ($p = 1.05 \times 10^{-24}$, using the Wilcoxon test). Such large vocabulary sizes are indeed unrealistic, but so is the assumption that new words do not arise during language evolution. In the SO model, we allow new words to arise over time, placing the word origin at the most parsimonious place that is the common ancestor of all COG-present cases (figure 3*b*). This model results in smaller ancestral vocabularies of up to 317 COGs, but these are still significantly larger than the contemporary vocabularies ($p = 1.65 \times 10^{-19}$, using the Wilcoxon test). The SO model entails an average of three losses per COG (electronic supplementary material, table S2).

Thus, we either have to embrace the untenable assumption that ancestral vocabulary sizes were fundamentally different in the past than they are today

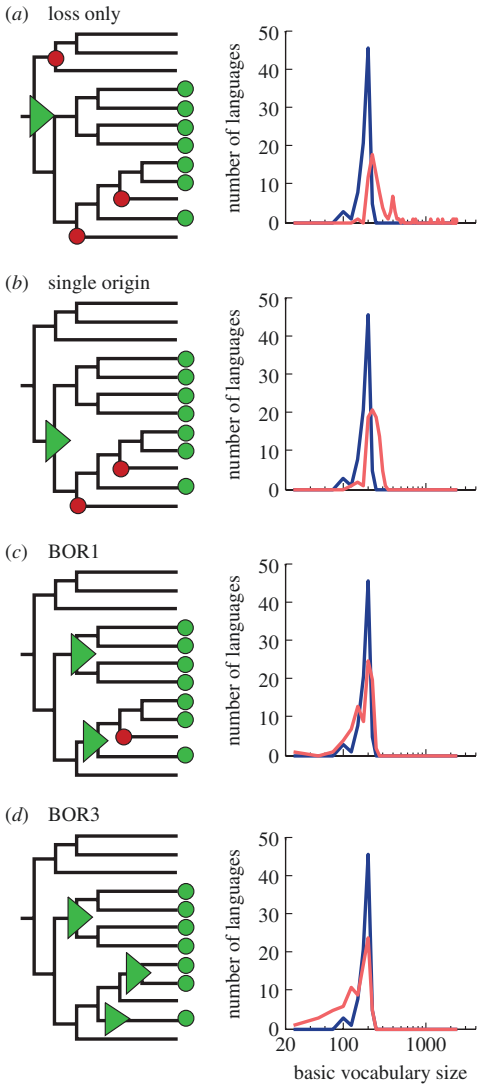


Figure 3. Inference of borrowing frequency by ancestral vocabulary size. (*a–d*) Schematic (left) and dynamics of ancestral and contemporary vocabulary size (right) under the different borrowing models. The fraction of interquartile range ($(\text{Median}_{\text{ancestral}} - \text{Median}_{\text{contemporary}})/\text{IQR}_{\text{contemporary}}$) in the different models is as follows. Loss only: 2.92; origin only: 1.93; BOR1: 0.12; BOR3: −0.86. Green triangles, origin; red circles, loss; green circles, word presence; blue line, contemporary languages; red line, ancestral languages.

or, preferably, we have to allow some amount of borrowing during evolution. We start by allowing only one borrowing event per COG, the BOR1 model. This model allows each COG to have two origins in the reference tree, one of which is by borrowing from any source (figure 3*c*). The result of this model is reduced ancestral vocabularies during the early evolution of languages, and an overall ancestral vocabulary size distribution that is not significantly different from that of contemporary languages ($p = 0.61$, using the Wilcoxon test). Of the total Dyen COGs, 918 (39%) are monophyletic, hence

their distribution is readily explained by an SO, while the remaining 1373 (61%) are patchy enough to infer two origins (one borrowing event). This frequency translates to an average rate of 0.6 borrowing events per COG during Indo-European language evolution.

If we allow up to three borrowings per COG (the BOR3 model; figure 3*d*), inferred ancestral vocabulary shrinks towards sizes that are again significantly different from modern ones, but this time are smaller than those of contemporary languages ($p = 4.43 \times 10^{-5}$, using the Wilcoxon test); that is, too much borrowing and not enough vertical descent are incurred from the standpoint of ancestral vocabulary sizes. Furthermore, under the BOR3 model, the average number of inferred word losses per COG is less than 1. But loss of COGs within basic vocabulary occurs quite frequently in language evolution [7], hence the BOR3 model is also unrealistic in that sense. Additional models allowing up to 15 borrowings per COG result in even smaller ancestral vocabulary sizes (electronic supplementary material, figure S3). Hence, ancestral basic vocabulary sizes demand borrowings to keep them realistically small, but too much borrowing makes them unrealistically small.

Testing the present evolutionary models with the help of a reference tree that is inferred from the same data might bias the inference of origin and loss events. However, using the Bayesian approach to reconstruct the tree yields the majority signal in the data. If the majority of COGs evolve mainly by vertical inheritance, then the tree is expected to be a reliable representation of the language phylogeny [46]. High frequency of borrowing events may mask the vertical signal and lead to less reliable reconstruction. To test the robustness of our borrowing frequency estimates, we repeated our analysis using various reference trees. Use of an alternative phylogenetic tree reconstructed by NJ [33] results in the same BOR1 model ($p = 0.7$, using the Wilcoxon test; electronic supplementary material, figure S3). In both reference trees, English is basal to the Germanic clade. However, this position is debated among linguists, and traditional classifications put English inside that clade [12,47]. To test the influence of the English position within the tree on our borrowing assessment, we tested all models using a reference tree with English in an internal position. Using that reference tree also yielded the BOR1 model ($p = 0.78$, using the Wilcoxon test), with all other models rejected ($\alpha = 0.05$). Using a random phylogenetic tree eliminates all patterns of vertically inherited COGs and accordingly results in the BOR15 model ($p = 0.16$, using the Wilcoxon test; electronic supplementary material, figure S4).

Performing the same tests on the ToB dataset yielded higher borrowing frequencies, with BOR3 being the only statistically accepted model ($p = 0.59$, using the Wilcoxon test; electronic supplementary material, figure S5). Inference by this model results in 155 COGs of SO, 181 COGs of two origins, 307 COGs of three origins and 79 COGs of four origins. Hence, in 567 (79%) of the 722 COGs, we detected one or more borrowing event. The average rate of borrowing events per COG during language evolution in the ToB dataset is 1.4 (electronic supplementary material, table S2). The higher borrowing rate inferred for the ToB dataset in comparison to the Dyen dataset might have to do with differences in their reconstruction. The cognate judgements in ToB are based on a deeper etymological reconstruction in comparison to the Dyen dataset. This results in more words that are distributed over fewer cognate sets, which leads to patchy COG distribution patterns that are frequently incongruent with the phylogenetic tree.

The sample of languages is crucial for the distinction between COG origin by birth or borrowing because what may seem to be a word birth within a given sample of languages in our data could in fact be a borrowing event from a non-sampled language. How severe is the effect of external borrowing on our results? If we assume the extreme case, for example, that all COGs in the dataset originated by borrowing from external languages, then we have to add one borrowing event to the average rate for each COG. In that case, the average borrowing rate would increase from 0.6 to 1.6 events per COG using the Dyen dataset. However, this extreme scenario is unlikely because it entails the assumption that the Indo-European groups sampled here lacked the wherewithal to invent even one new COG. Nonetheless, external borrowing has almost certainly had an effect on these data. Although we currently lack a dataset that would allow us to quantify the rate of external borrowing, if we assume that it is similar to the internal borrowing rate within our sample, the overall borrowing rate would be double our current estimate. Again we stress that the borrowing frequency inferred from the present sample of languages using our method delivers a minimum value (a conservative lower bound).

Another aspect of the data sample used in our analysis is the collection of cognates. Here, we study the dynamics of vocabulary size during evolution through the proxy of basic vocabulary (i.e. the Swadesh list). However, origin and loss of words in the COGs sample can occur by semantic shift where the word is present in the language but absent from the sample. It is possible that different meaning collections evolve under regimens different from the ones described here. Application of similar methods to study vocabulary size dynamics over time using different cognate datasets will help to clarify this issue.

Notwithstanding certain amounts of cognate misjudgements and parallel evolution [48] resulting in tree-incompatible COG distributions, our inference uncovers abundant, and hitherto unrecognized, borrowing during the evolution of the Indo-European languages.

Scholars usually agree that nouns are more easily borrowed than verbs [49]. When classified according to the English gloss, the Dyen dataset includes 887 (53%) cognate sets corresponding to nouns within basic vocabulary and 766 (46%) cognate sets corresponding to verbs. A total of 503 (53%) nominal cognate sets and 450 (47%) verbal cognate sets were identified as including hidden borrowing events. A comparison of these frequencies shows that there is no significant difference in borrowing frequencies between nouns and verbs ($p = 0.4$, using the *G*-test).

(c) *Minimal lateral networks of Indo-European languages*

COG distributions that do not map exactly onto the phylogenetic tree, with borrowing constrained by ancestral

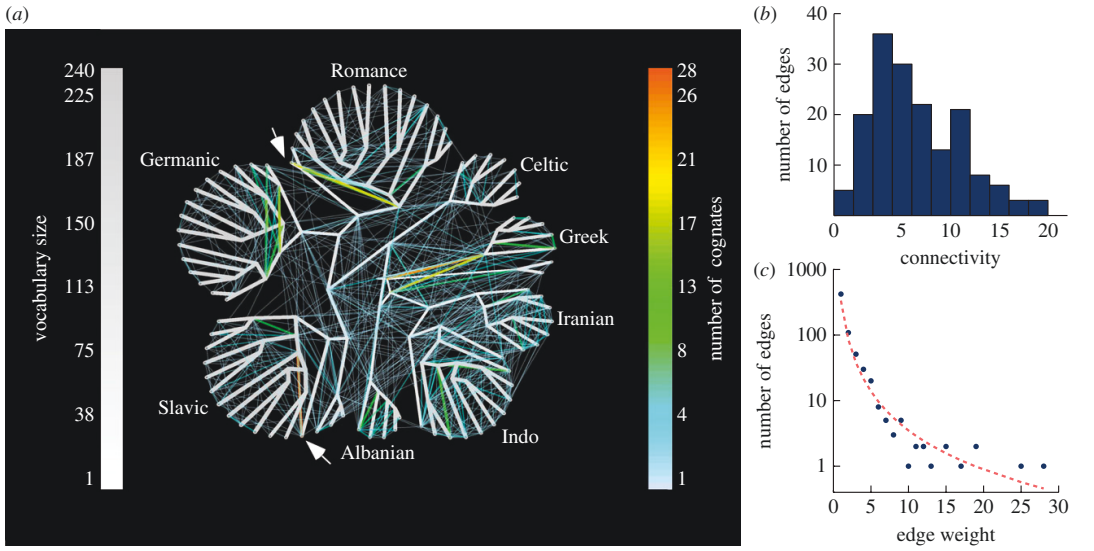


Figure 4. The MLN of Indo-European languages. (*a*) An MLN for 84 contemporary languages reconstructed under the BOR1 model. Vertical edges are indicated in grey, with both the width and the shading of the edge shown proportional to the number of inferred vertically inherited COGs along the edge (see the scale). The lateral network is indicated by edges that do not map onto the vertical component, with the number of cognates per edge indicated in colour (see the scale). Lateral edges that link ancestral nodes represent laterally shared COGs among the descendent languages of the connected nodes, whose distribution pattern could not be explained by origin and LO under the ancestral vocabulary size constraint. The two heaviest edges of Slovene (Slavic) and Romanian (Romance) are marked by an arrow. (*b*) Distribution of connectivity, the number of one-edge-distanced neighbours for each vertex, in the network. (*c*) Frequency distribution of edge weight in the lateral component of the network.

vocabulary size only, constitute the MLN [37]. The MLN reconstructed from the Dyen dataset consists of 167 vertices, of which 84 are contemporary and 83 are ancestral languages (internal nodes in the reference tree). The vertices are interconnected either by the branches of the reference tree, representing vertical inheritance, or by lateral edges, representing horizontal transfer (figure 4*a*).

The internal and external vertices in the MLN for the broad sample of COGs are linked by 666 lateral edges. The connectivity (number of edges per vertex) within the MLN ranges between 0 and 21 edges per language, with a median of 7 (figure 4*b*). The most highly connected node is Ossetic (21 edges), an east Iranian language, which is connected with Indo-Iranian, Greek and Slavic languages. Lateral edges connected to external nodes correspond to comparatively recent borrowing events. On average $8 \pm 7\%$ COGs per language are involved in recent borrowing (electronic supplementary material, table S3). This result suggests that English, at 8 per cent borrowing rate [13], is not exceptional; it is merely the most studied language. The clustering coefficient of the MLN is 0.22, and the mean shortest path is 3.128 edges. Combined with the high level of clustering, this means that the MLN forms a small-world network.

The edge weight distribution within the MLN is characterized by a majority of small edge weights. Of the total edges, 422 (63%) are of a single laterally shared COG, while edges of multiple COGs are rare (figure 4*c*). The two heaviest lateral edges include an edge between Slovene and the remaining Slavic languages (28 COGs), and an edge between Romanian and the remaining Romance languages (19 COGs). These lateral edges uncover a certain kind of language change that results from the same evolutionary process. Both Slovene and Romanian, being heavily influenced by neighbouring languages, underwent a process of linguistic revival starting from the early 19th century, in which the original

Table 1. Reconstructed borrowing events. The origin node that includes the reinserted borrowing is shaded in light grey.

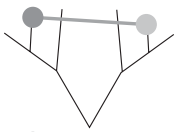
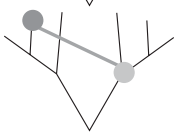
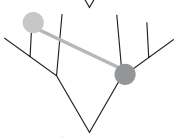
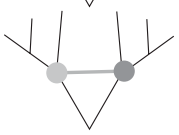
edge type	origin node	number of reinserted borrowings
external–external		1
external–internal		18
		58
internal–internal		40

Table 2. Lateral edge (LE) frequencies between and within groups in the MLN.

		normalized borrowing		median LE weight[b]		H_0:$LE_{int} \leq LE_{ext}$ frequency[c,d]
group	n[a]	int	ext	int	ext	p-value
Greek	9	1.22	0.25	2	1	<0.05
Armenian	3	0	0.17	0	1	n.a.
Celtic	13	1.61	0.29	2	1	$\ll 0.05$
Romance	31	2.45	0.36	1	1	$\ll 0.05$
Germanic	29	2.37	0.44	1	1	$\ll 0.05$
Slavic	31	2.35	0.64	1	1	$\ll 0.05$
Albanian	9	1.55	0.18	4	1	$\ll 0.05$
Indic	21	3.33	0.68	2	1	$\ll 0.05$
Iranian	14	2.35	0.75	2	1	$\ll 0.05$

[a]Number of languages within group.
[b]Range of median number of COGs per lateral edge.
[c]One-side Kolmogorov–Smirnov test for lateral edge distribution.
[d]For internal edges (int), number of internal edges per number of nodes within the group; for external edges (ext), number of external edges per number of nodes outside the group.

traits that had been lost during long periods of contact were artificially reintroduced into the languages by the speakers in order to bring them back to a stage of earlier 'purity' [50,51]. Before the 19th century, Slovene comprised several dialects spoken in the Alpine provinces of the Austrian Empire, which were dominated by German and Italian. Romanian, on the other hand, was heavily influenced by neighbouring Slavic and Greek varieties, with which it formed the so-called Balkan *Sprachbund*. Along with the nationalist movements in Europe starting from the end of the 18th century, both languages were successively 'purified' by replacing the loanwords of non-Slavic or non-Romance origin with 'native' words from Slavic or Romance languages, respectively [50,51]. This process is somewhat different from the process of borrowing as it was defined in the beginning of this paper. It nonetheless illustrates additional horizontal complexities in the processes of language evolution that are readily detected in the MLN.

The comparison between the edges reconstructed using the two reference trees that differ in their English position supplies a few interesting observations regarding the applicability of our approach to detect borrowing events. While both reference trees yielded the same borrowing model (i.e. the same overall borrowing rates), there are 23 lateral edges connecting to English in the basal position and only 15 lateral edges connecting to English in the internal position. A closer inspection of the COGs in which the lateral edges connecting to English were detected revealed that seven of the eight COGs detected as borrowings in the basal position could not be verified as borrowings by traditional historical linguistics. Thus, using different reference trees with the same COG distribution patterns does not much affect the resulting borrowing model, but it may increase the accuracy of concrete predictions made by this approach (see electronic supplementary material, table S4 for detailed etymological reconstruction of the COGs). Consequently, the borrowing inference accuracy in our approach is expected to increase with the accuracy of the reference tree.

The MLN inferred from the ToB dataset shows similar network characteristics, with the ancestors of Indian and Iranian clades found also as highly connected nodes and a majority (676; 76%) of single laterally shared COGs (electronic supplementary material, figure S6).

Of the total 666 edges in the MLN reconstructed for the Dyen dataset, 148 (22%) edges connect between two external nodes—that is, between two contemporary languages. The 301 (45%) edges that connect between an internal node and an external node represent COGs that are shared between a group and an outlier. The 217 (33%) edges that connect between two internal nodes represent COGs that are common to two different groups, yet their distribution pattern could not be explained by vertical inheritance alone under the vocabulary size criterion. As a control to see whether our method is inferring spurious borrowing, we examined the edges within cognates that included the 124 reinserted borrowing events. In seven cognates, the algorithm detected no borrowings, while in all other 117 (94%) cognates a borrowing event was inferred. In 59 (48%), the reinserted borrowing language was inferred as an external node. In the remaining 58 (47%), reinserted borrowing languages were inferred within descendants of an internal node (table 1).

The data can address the issue of whether words are exchanged more frequently within than between main branches of Indo-European. We can compare the probability of a certain language to be laterally connected with languages that are either from the same main branch or from different main branches of the Indo-European languages. With the exception of the Armenian branch, the probability for a lateral edge within the branch (internal edge) is considerably higher than between branches (external edge). Furthermore, lateral edge weights are significantly larger in internal lateral edges than in external lateral edges (table 2). Hence, lexical borrowing in Indo-European languages is much more frequent among languages within the same branch in comparison to languages from different branches. This provides new evidence for the existence of certain cultural barriers to lexical borrowing during language evolution [10].

The study was supported by the German Federal Ministry of Education and Research (S.N.S., J.M.L., H.G., T.D. and W.M.) and the European Research Council (W.M.). We are

thankful to Frank Kressing, Matthis Krischel, Thorsten Halling and Sven Sommerfeld for helpful discussions, and to Dan Graur for his help in refining the manuscript. We thank Liat Shavit-Grievink for her help in phylogenetic reconstruction.

REFERENCES

1 Darwin, C. 1859 *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London, UK: John Murray. See http://www.nla.gov.au/apps/cdview/nla.gen-vn4591931.
2 Schleicher, A. 1863 *Die Darwinsche Theorie und die Sprachwissenschaft offenes Sendschreiben an Herrn Dr. Ernst Häckel*, 3rd edn (1873). Weimar, Germany: Böhlau.
3 Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)
4 Pagel, M. 2009 Human language as a culturally transmitted replicator. *Nat. Rev. Genet.* **10**, 405–415.
5 Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. 2005 Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075. (doi:10.1126/science.1114615)
6 Lansing, J. S. *et al.* 2007 Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl Acad. Sci. USA* **104**, 16 022–16 026. (doi:10.1073/pnas.0704451104)
7 Pagel, M. & Atkinson, Q. D. 2007 A Meade frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)
8 Thomason, S. G. 2001 *Language contact: an introduction*. Edinburgh, UK: Edinburgh University Press.
9 Trask, R. L. 2000 *The dictionary of historical and comparative linguistics*. Edinburgh, UK: Edinburgh University Press.
10 Thomason, S. & Kaufman, T. 1988 *Language contact, creolization, and genetic linguistics*. Berkeley, CA: University of California Press.
11 Aikhenvald, A. Y. 2006 Grammars in contact: a cross-linguistic perspective. In *Grammars in contact: a cross-linguistic typology* (eds A. Y. Aikhenvald & R. M. Dixon), pp. 1–66. Oxford, UK: Oxford University Press.
12 Fox, A. 1995 *Linguistic reconstruction: an introduction to theory and method*. Oxford, UK: Oxford University Press.
13 Embleton, S. 2000 Lexicostatistics/glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In *Time depth in historical linguistics* (eds C. Renfrew, A. McMahon & L. Trask), pp. 143–165. Cambridge, UK: The McDonald Institute for Archaeological Research.
14 Bergsland, K. & Vogt, H. 1962 On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153. (doi:10.1086/200264)
15 Boyd, R., Borgerhoff, M. M., Durham, W. H. & Richerson, P. J. 1997 Are cultural phylogenies possible? In *Human by nature, between biology and the social sciences* (eds P. Weingart, P. J. Richerson, S. D. Mitchell & S. Maasen), pp. 355–386. Mahwah, NJ: Erlbaum.
16 Atkinson, Q. D. & Gray, R. D. 2006 How old is the Indo-European language family? Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 91–109. Cambridge, UK: McDonald Institute for Archaeological Research.
17 Schleicher, A. 1853 Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*, **September**, 786–787.
18 Schmidt, J. 1872 *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Weimar, Germany: Hermann Böhlau.
19 Schuchardt, H. 1922 Über die Klassifikation der romanischen Mundarten. In *Hugo Schuchardt-Brevier. Ein Vademekum der allgemeinen Sprachwissenshaft. Als Festgabe zum 80. Geburtstag des Meisters zusammengestellt und eingeleitet von Leo Spitzer* (ed. L. Spitzer), pp. 144–166. Halle, Germany: Max Niemeyer.
20 Hirt, H. 1905 *Die Indogermanen. Ihre Verbreitung, ihre Urheimat und ihre Kultur*, vol. 1. Strassburg, France: Trübner.
21 Bonfante, G. I. 1931 I dialetti indoeuropei. *Annali del R. Istituto Orientale di Napoli* **4**, 69–185.
22 Dyen, I., James, A. T. & Cole, J. W. L. 1967 Language divergence and estimated word retention rate. *Language* **43**, 150–171. (doi:10.2307/411390)
23 Ringe, D. A. 1992 On calculating the factor of chance in language comparison. *Trans. Am. Phil. Soc.* **82**, 1–110. (doi:10.2307/1006563)
24 Southworth, F. C. 1964 Family-tree diagrams. *Language* **40**, 557–565. (doi:10.2307/411938)
25 Bryant, D., Filimon, F. & Gray, R. D. 2005 Untangling our past: languages, trees, splits and networks. In *The evolution of cultural diversity: phylogenetic approaches* (eds R. Mace, C. Holden & S. Shennan), pp. 67–84. London, UK: UCL Press.
26 Nakhleh, L., Ringe, D. & Warnow, T. 2005 Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* **81**, 382–420. (doi:10.1353/lan.2005.0078)
27 McMahon, A., Heggarty, P., McMahon, R. & Slaska, N. 2005 Swadesh sublists and the benefits of borrowing: an Andean case study. *Trans. Phil. Soc.* **103**, 147–170. (doi:10.1111/j.1467-968X.2005.00148.x)
28 Ben Hamed, M. & Wang, F. 2006 Stuck in the forest: trees, networks and Chinese dialects. *Diachronica* **23**, 29–60.
29 Dyen, I., Kruskal, J. B. & Black, P. 1997 *Comparative Indo-European database: file IEdata1*. See http://www.wordgumbo.com/ie/cmp/iedata.txt.
30 Starostin, G. 2008 *Tower of Babel: an etymological database project*. See http://starling.rinet.ru.
31 Newman, M. E. J. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S003614450342480)
32 Ronquist, F. & Huelsenbeck, J. P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)
33 Saitou, N. & Nei, M. 1987 The neighbor-joining method. A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
34 Huson, D. H. & Bryant, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267. (doi:10.1093/molbev/msj030)
35 Dagan, T. & Martin, W. 2007 Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl Acad. Sci. USA* **104**, 870–875. (doi:10.1073/pnas.0606318104)
36 Zar, J. H. 1999 *Biostatistical analysis*, 4th edn. Englewood Cliffs, NJ: Pearson Prentice-Hall.
37 Dagan, T., Artzy-Randrup, Y. & Martin, W. 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl Acad. Sci. USA* **105**, 10 039–10 044. (doi:10.1073/pnas.0800679105)

38 Swadesh, M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137. (doi:10.1086/464321)

39 Swadesh, M. 1952 Lexicostatistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proc. Am. Phil. Soc.* **96**, 452–463.

40 Geisler, H. & List, J.-M. 2011 Beautiful trees on unstable ground: notes on the data problem in lexicostatistics. In *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik, Akten der Arbeitstagung der Indogermanischen Gesellschaft Würzburg, 24–26 September 2009* (ed. H. Hettrich). Wiesbaden, Germany: Reichert.

41 Girvan, M. & Newman, M. E. J. 2002 Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **12**, 7821–7826.

42 Dyen, I., Kruskal, J. B. & Black, P. 1992 An Indoeuropean classification: a lexicostatistical experiment. *Trans. Am. Phil. Soc.* **82**, 3–132.

43 Mallory, J. P. & Adams, D. Q. 2006 *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford, UK: Oxford University Press.

44 Wells, R. S. 1973 Uniformitarianism in linguistics. In *Dictionary of the history of ideas* (ed. P. Wiener), pp. 423–431. New York, NJ: Scribner.

45 Christy, C. 1983 *Uniformitarianism in linguistics*. Amsterdam, The Netherlands: John Benjamins.

46 Greenhill, S. J., Currie, T. E. & Gray, R. D. 2009 Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306. (doi:10.1098/rspb.2008.1944)

47 Lewis, P. M. 2009 *Ethnologue: languages of the world*, 16th edn. Dallas, TX: SIL International. See http://www.ethnologue.com.

48 Garrett, A. 2006 Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 139–151. Cambridge, UK: McDonald Institute for Archaeological Research.

49 Hock, H. H. & Joseph, B. D. 2009 Language history, language change and language relationship. In *An introduction to historical and comparative linguistics*, 2nd edn. Berlin, Germany: Mouton de Gruyter.

50 Auty, R. 1963 The formation of the Slovene literary language against the background of the Slavonic national revival. *SEER* **41**, 391–402.

51 Mallinson, G. 1988 The Romance languages in. In *The Romance languages* (eds M. Harris & V. Nigel), pp. 391–419. London, UK: Croom Helm.

52 Orel, V. 2003 *A handbook of Germanic etymology*. Leiden, The Netherlands: Brill.

53 Soukhanov, A. H. 1992 *The American heritage dictionary of the English language*. Boston, MA: Mifflin.

3.3.

Polynesian language networks reveal complex history of contacts during the Pacific settlement

Shijulal Nelson-Sathi[1], Johann-Mattis List[2], Simon Greenhill[6], Hans Geisler[4], Ofir Cohen[3], Tal Pupko[8], Giddy Landan[5], William F. Martin[1], Tal Dagan[5*], Russell Gray[7]*

[1]Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Germany
[2]Research Center Deutscher Sprachatlas, Philipps-University Marburg, Germany
[3]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel
[4]Institute of Romance Languages and Literature, Heinrich-Heine University Düsseldorf, Germany
[5]Institute of Genomic Microbiology, Heinrich-Heine University Düsseldorf, Germany
[6]School of Culture, History and Language, ANU College of Asia and Pacific, Australian National University, Canberra, Australia.
[7]Department of Psychology, University of Auckland, Auckland 1142, New Zealand
[8]Current address: Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

Keywords: Polynesian language networks, borrowing, dialect networks, horizontal inheritance, cultural evolution, Pacific settlement.

*Corresponding authors

The presented manuscript was submitted in the Journal of "*Proceedings of the National Academy of Sciences (PNAS), 2013,May*

Contribution of Shijulal Nelson-Sathi – First author

Experimental design: 60%
Execution and analysis of experiments: 80%
Manuscript writing: 50%

Polynesian language networks reveal complex history of contacts during the Pacific settlement

Shijulal Nelson-Sathi[1], Johann-Mattis List[2],Simon Greenhill[6], Hans Geisler[4], Ofir Cohen[3,8], Tal Pupko[3], Giddy Landan[5], William F. Martin[1], Tal Dagan[5]*, Russell Gray[7]*

[1]Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Germany [2]Research Center Deutscher Sprachatlas, Philipps-University Marburg, Germany [3]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel [4]Institute of Romance Languagesand Literature, Heinrich-Heine University Düsseldorf, Germany [5]Institute of Genomic Microbiology, Heinrich-Heine University Düsseldorf, Germany [6]School of Culture, History and Language, ANU College of Asia and Pacific, Australian National University, Canberra, Australia. [7]Department of Psychology, University of Auckland, Auckland 1142, New Zealand and Philosophy Program, Research School of the Social Sciences, Australian National University 0200 Canberra, ACT, Australia. [8]Current address: Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Trees are the standard way of representing historical relationships in both biology and linguistics. However, trees only depict the vertical component of inheritance. In both fields there are also horizontal components (lateral gene transfer in genome evolution and borrowing in language history). Phylogenetic networks are increasingly being used in studies of microbial evolution to infer both the vertical and the horizontal components of evolutionary change. Here, we apply network methods to study the borrowing dynamics among 33 Polynesian languages. The data reveal unexpectedly high levels of borrowing, between languages from distantly separated Polynesian islands. The average level of borrowing per cognate in the basic vocabulary is 51%. This frequency of borrowing is substantially higher than that estimated for Indo-European languages using the same approach. Furthermore, and again in contrast to Indo-European languages, the borrowing rate in basic vocabulary is not substantially lower than that found in the whole lexicon (72%). This borrowing in basic vocabulary indicates a process of gradual breakup of dialect networks, suggesting in turn that the history of Polynesian settlement was far from a series of one-way voyages. Instead it involved extended contact across the vast distances of the Pacific Ocean and the subsequent slow breakup of at least three large dialect networks.

borrowing | cultural evolution | dialect networks | horizontal inheritance | Language evolution

INTRODUCTION

Genome evolution and language change have a lot in common (1). The domains of both processes are elements that are constantly subject to change and both involve the splitting of lineages. The cumulative effects of such changes produce new languages or new species. Given the "curious parallels" in both processes it is not surprising that biologists and linguists developed similar methods to reconstruct these genealogical relations, and the concept of the family tree, that was independently adopted by linguists and biologists (2,3), has became the paradigm to model how species evolve and how languages develop. However, while the concept of a "Great Tree of Life" has only recently come under fire among biologists (4), linguists began quite early to question the adequacy of trees to depict the complexity of language evolution. Words and other aspects of language are not always vertically inherited from ancestral languages, but can also be borrowed from other languages. In 1872, the German linguist Johannes Schmidt proposed an alternative model of language evolution known as the Wave Theory (5), according to which innovations spread over speech communities like waves, making it impossible to depict the complex processes of language change with family trees. Although many scholars followed Schmidt's example and emphasized the inadequacy of the family tree in linguistics, none of the many alternative models that were proposed, be it waves (5,6), set diagrams (7), or even animated pictures (8), have gained general acceptance among language scholars. Accordingly, most historical linguists continue to construct family trees, while at the same time acknowledging their inadequacy (9).

The debate about the extent to which language change is accurately represented by a tree is a microcosm of the more general debate about the extent to which human cultural evolution is treelike. As far back as 1948 the influential anthropologist Kroeber explicitly contrasted Darwin's idea of a 'tree of life' with that of a 'tree of cultures'. Kroeber (10) argued cultural evolution is highly reticulate, with frequent borrowing and diffusion of traits between cultures. Despite the recent growth of cultural phylogenetics (11-13), considerable doubt remains about just how treelike different aspects of culture are (14). Borgerhoff Mulder et al. (15) conclude their review of cultural phylogenetics with the cautionary statement that our, 'Current understanding of the relative importance of horizontal and vertical transmission is shaky, to say the least' (p. 62).

Network methods are a rapidly developing area in evolutionary biology (16). These biological methods have been used to construct language networks using either the split decomposition or the NeighbourNet algorithm (14,17-20). However, while the resulting splits graphs give a graphical representation of how tree-like the data are, they do not yield estimates of borrowing frequencies, nor do they identify specific borrowing events. Thus,

Significance

As a culture evolves, so does its language. Words can be vertically inherited during language evolution, generating treelike patterns, or laterally spread across languages — borrowed — through cultural contact. In the evolution of Polynesian languages, vast distances between Polynesian islands pose seemingly steep natural barriers to language contact. Hence the history of Polynesian languages should be tree-like. We studied homolgous vocabulary in Polynesian languages with network methods to detect both vertical and lateral components of Polynesian language evolution. We found surprisingly high levels of borrowing. Many detected borrowings reflect myriad connections among expertly seafaring cultures during South Pacific settlement.

Reserved for Publication Footnotes

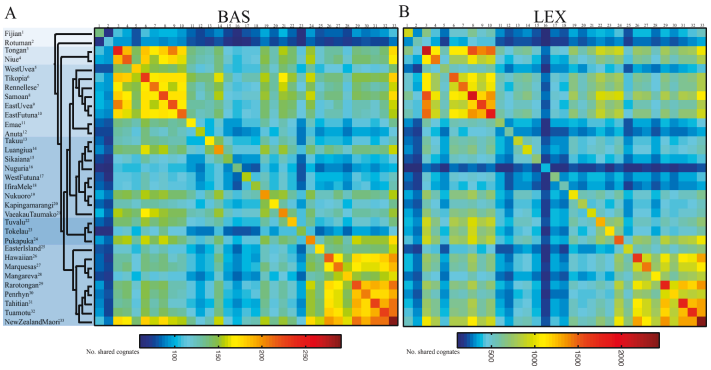


Fig. 1. Polynesian shared cognate network. A) A graphical representation of the basic vocabulary shared cognate network. B) Full lexicon shared cognate network. Cells in the matrix are edges in the network. Edges are color-coded by the frequency of shared cognate according to the color bar in the bottom. Languages are sorted by their order on the Bayesian reference tree.

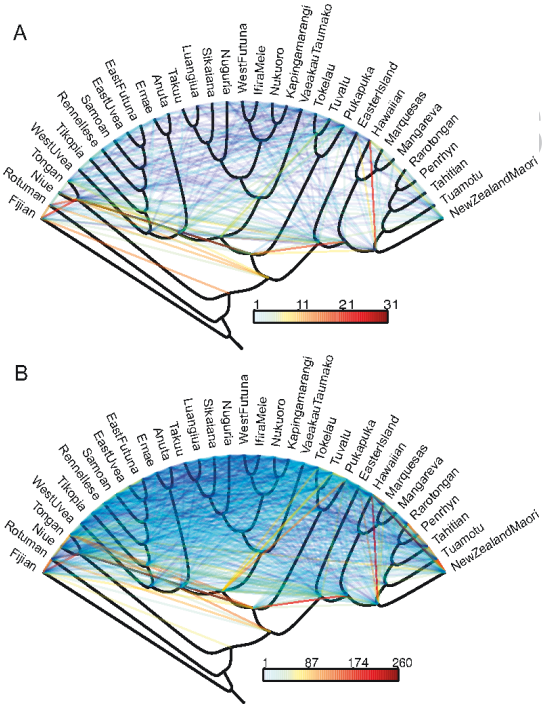


Fig. 2. A complete Polynesian language borrowing network showing both vertical and lateral edges reconstructed using MLN approach. Branches in the MLN represent vertical inheritance and are determined by the underlying reference tree. Lateral edges in the MLN represent inferred borrowing events. They connect external (contemporary) and internal (ancestral) tree nodes (languages) whenever a gain-loss scenario involving more than one gain was inferred for at least one of the cognate sets. Heavy edges in Polynesian borrowing networks clearly showing some of the dialect networks postulated by Pawley (2009).

even though these splits networks are based on quantitative data, they are best thought of as a tool for visualizing data rather making detailed inferences about complex histories.

A more recent Minimal Lateral Network (MLN) approach can directly address borrowing events in linguistic datasets (21) Originally developed to study lateral gene transfer in microbial genome evolution (22), this method estimates the amount of horizontal evolution under the uniformitarian assumption (23,24) that ancestral genome (or vocabulary) sizes were not significantly different from those seen today. Given a patchy distribution of genes (or cognates) that is incompatible with a reference tree, the method infers the minimum number of origins required to explain the distribution (unlimited losses are allowed). If there were only vertical inheritance, then only a single origin would be required. However, postulating only a single origin for many genetic and linguistic datasets would require accepting that the ancestral genome (or vocabulary) size was substantially larger than that attested today. The MLN approach estimates the minimal number of additional origins (borrowings in the linguistic case) required to be consistent with modern genome or vocabulary sizes.

The method had remarkable success in identifying known borrowings in a dataset of Indo-European basic vocabulary items (25) (the so-called *Swadesh list*). Nelson-Sathi et al. (21) found that it recovered around 94% of the known borrowings. However, items on the Swadesh list are generally thought to be resistant to borrowing and stable over time. Here we apply a modified version of the MLN method to both the basic vocabulary and a sample of the total lexicon from 33 Polynesian languages. The Polynesian islands have often been seen as an ideal "laboratory for the study of cultural adaptation" (26). The vast ocean distances between these "far flung islands" has prompted some commentators to claim that their language evolution should be highly treelike. Even critics of the family tree model, such as Dixon (27) argue that, "For some groups of languages — for instance, Semitic and Polynesian — the family tree model is entirely applicable". However, while most linguists agree that all Polynesian languages form a subgroup of the Austronesian language family, the internal subgrouping of the Polynesian languages has been the subject of long-standing debates (28-31), and archaeologists have documented substantial post-settlement contact (32- 34). Here we use the MLN approach to quantify the degree of borrowing amongst Polynesian languages, and to infer the detailed dynamics of any complex, non-treelike language histories.

RESULTS

Conflicting Signals in Polynesian Language Evolution

A total of 4,800 cognate sets from 33 Polynesian languages in the POLLEX database (ver. June 2012) were classified into two datasets – one of basic vocabulary and one of the full lexicon. Each dataset was analyzed separately. For the analysis, all cognate sets in the dataset were converted into a binary matrix of presence/absence patterns (PAPs), consisting of 33 cells in a row. If a given cognate set had one or more reflexes in a language *i*, the i^{th} cell in the row was set to 1, otherwise it was set to 0.

The basic vocabulary subset includes a total of 410 PAPs; 398 are unique and 12 are recurrent (Supp. Fig. 1). The dataset

Table 1. Borrowing dynamics inference using Minimal Lateral Netowrk (MLN) and Stochastic Mapping (SM) approaches.

	MLN inference							SM inference	
Reference Tree	Accepted Model	P-value	No. Origins 1	2	>2	Avg. gain	Avg. loss	Avg. gain	Avg. loss
BASIC									
Bayesian	BOR_7	0.33	100	64	265	1.93	4.5	4.74	5.78
Bayesian-Modified	BOR_7	0.23	73	97	259	1.95	4.39	4.58	5.01
	BOR_{15}	0.30	73	97	259	1.95	4.5		
MP	BOR_7	0.76	17	40	372	2.9	4.14	4.0	3.8
NJ	BOR_7	0.54	20	43	366	2.7	4.0	4.47	4.5
	BOR_{15}	0.23	20	43	366	3.5	3.28		
FULL									
Bayesian	BOR_7	0.94	360	649	3,344	2.4	5.2	5.6	6.19
Bayesian-Modified	BOR_7	0.85	257	795	3,310	2.3	5.5	5.35	5.5
MP	BOR_3	0.29	135	920	3,307	1.98	5.35	5.4	5.1
	BOR_7	0.10	135	571	3,656	3.02	4.15		
NJ	BOR_3	0.18	104	512	3,746	2.0	4.4	5.31	5.0
	BOR_7	0.41	104	512	3,746	2.87	3.38		

MLN -Minimum Lateral Network, SM-Stochastic Mapping, MP- Maximum Parsimony, NJ –Neighbor Joining. Bayesian-Modified reference tree is a manually edited Bayesian tree

Table 2. dMLN and wMLN network statistics.

		Degree				Edge weight				Connectivity			
		Mean	Median	Min	Max	Mean	Median	Min	Max	Ext-Ext	Ext-Int	Int-Int	Total Edges
dMLN	BASIC	10.6	10	0	29	2.4	1	1	31	120	178	48	346
	FULL	30.6	31	0	55	10.4	3	1	260	359	484	154	997
wMLN	BASIC	8.2	7	0	30	3.0	2	1	31	104	130	35	269
	FULL	21.5	19	0	56	14.8	4	1	300	294	324	83	701

Ext-*External,* Int- *Internal node*

includes seven cognate sets reflected in all languages: Proto-Polynesian **fale* "house", **koe* "thou", **quha* "rain", **qate* "liver", **kai* "eat", Proto-Malayo-Polynesian **tuqud* "to stand", and Proto-Central-Malayo-Polynesian **waiʀ* "water". Five further cognates, Proto-Polynesian **foqou* "new", **laŋi* "sky", Proto-Malayo-Polynesian **telu* "three", **mata* "eye", and Proto-Austronesian **lima* "five", are only missing in Tokelauan. The full dataset includes 4,370 cognates that were converted into 3,590 PAPs, of which 3325 are unique and 265 are recurring (Supp. Fig. 1). Only six of these cognates, Proto-Polynesian **fohe* "paddle", **futi* "to pluck", **hifo* "downwards", **nofo* "to put down", and **waka* "canoe", had reflexes in all 33 Polynesian languages.

In the basic vocabulary dataset, only 24 cognates (5.5%) were congruent with the topology of the reference tree. The remaining 94.5% of the PAPs did not tidily fit the tree branching pattern. An example of patchily distributed cognate **tafuqa* (platform, foundation, base) is shown in Supplementary Fig. 2. Numerals dominate these congruent cognates, such as Proto-Polynesian **tasi* "one", **rua* "two", and **lima* "five." Ten of the 24 congruent cognates seem to be highly resistant to borrowing as they appear in the upper half of the Leipzig-Jakarta rank (35) (Supp Table S1), that ranks concepts according to borrowing resistance based on a detailed statistical survey of 41 languages including many different language families (36). In the full lexicon subset, the distribution of 98% of all cognates conflicts with the reference tree topology. The presence/absence distribution of only 92 cognates (2%) can be explained by strict vertical inheritance alone, i.e., without independent losses. (Supp. Table S2).

To study patterns of shared cognates, the binary PAPs of each dataset were summarized into a network of shared cognates where the vertices are languages that are connected by edges representing the frequency of shared cognates between them. The matrix representation of the network is defined as $A=[a_{ij}]$ with size 33×33, where a_{ij} is calculated as the number of cognates shared by languages i and j. The resulting networks comprise 33 vertices and 528 edges connecting all vertices (Fig. 1). The edge weight distribution in the network reveals that many closely related languages are strongly interconnected, confirming their genealogical relatedness. However, there exist several distantly related languages that show strong interconnections in the networks of shared cognates. For example, in the full lexicon subset New Zealand Maori shares 44 cognates uniquely with Rarotongan, and 38 cognates uniquely with Hawaiian. In the basic dataset, Luangiua shares 140 common cognates with Tikopia, much more than with any language in its clade (Fig. 1).

Borrowing Frequency during Polynesian Language Evolution

The MLN approach reconstructs scenarios of language history dynamics that comprise of vertical inheritance complemented by varying amounts of borrowings. We tested six evolutionary models, ranging from no borrowing and up to 31 bor-

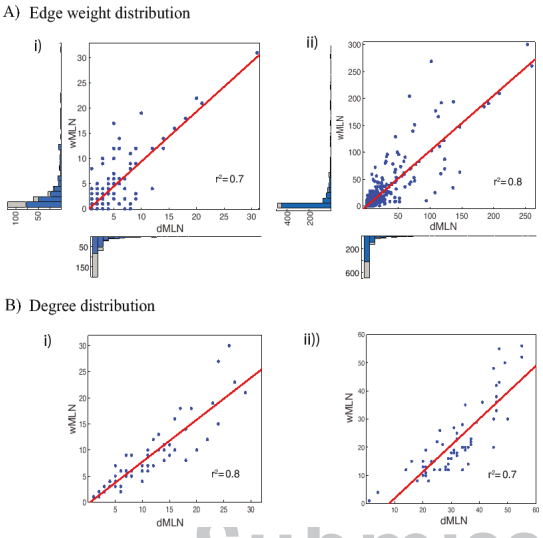


Fig. 3. Comparison of network properties between the dMLN and wMLN. A) Edge weight distribution of basic and full datasets B) Degree distribution of basic and full lexicon.

rowing events (BOR_{31}). Comparing the ancestral vocabulary size distributions resulting from these models with the vocabulary size distribution attested in contemporary languages, revealed that the BOR_7 model (up to 7 borrowings per cognate) yields the best fit between ancestral and contemporary vocabulary size distributions in both the basic and full datasets (P_{BASIC}=0.33, $P_{LEXICON}$=0.94, using the Wilcoxon test). All other models were rejected (using α=0.05). These estimated borrowing levels are substantially higher than those estimated for Indo-European, where the BOR_1 model was the best fit (21).

The presence/absence distribution of only 24 cognates (5.5%) in the basic dataset could be explained by a single origin (monophyly). The distribution of the remaining 94.5% cognates could only be explained by adding one or more borrowing events. In the full subset, only 92 cognates (2%) could be explained by a single origin. The total gain events translate into an average frequency of 1.9 and 2.4 borrowing events per cognate in the basic and full datasets respectively. The most frequently borrowed words in the basic dataset include reflexes of Proto-Polynesian **fatitili* "thunder", **soko* "to exchange", **fafa* "mouth", **pokoqulu* "skull, head", **mamawa* "yawn", **mai* "come", **luaki* "vomit", and **koi* "sharp". There are 87 words in the full dataset with five or more borrowings inferred and the most frequently borrowed word is **qaoa* "*Banyan tree*".

Since the MLN approach uses a reference tree for the inference of origin-loss scenarios, alternative tree topologies may result in different estimates of borrowing events. To test the robustness of our borrowing inference, we repeated the analysis using three alternative reference trees: a manually modified version of Bayesian tree with a modified branching positions of Pukapuka and Tokelau, a Neighbor-Joining tree (37), and a Wagner maximum parsimony (MP) tree (38). The modified Bayesian tree yielded similar mean borrowing frequency per cognate in both the basic and full datasets (Table 1). Using the MP and NJ trees for the basic vocabulary dataset resulted in slightly higher average borrowings per cognate while the frequency of losses remains similar (Table 1). The use of an MP reference tree to reconstruct borrowing events in the full data yielded slightly lower mean borrowings/cognate while the NJ tree inference was similar to that of the Bayesian reference trees (Table 1). Applying a stochastic

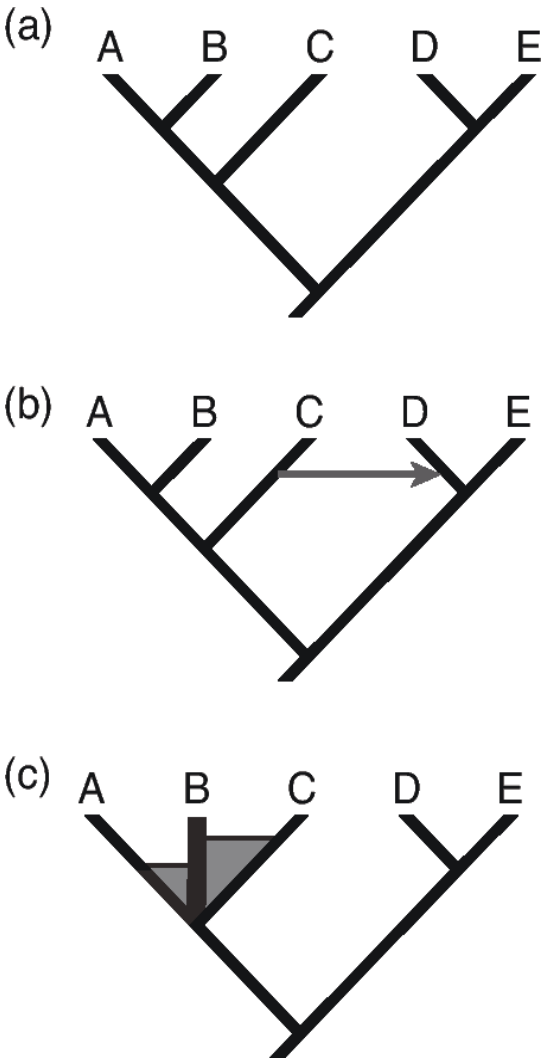


Fig. 4. Three models of language differentiation. (a) the standard family tree model, (b) a tree model with borrowing, and (c) dialect network breakup in which innovations partially diffuse across the network as it slowly breaks up, producing substantial conflicting non-treelike signal.

mapping (SM) approach (39) to both datasets yielded slightly higher averages of borrowings and losses per cognate using all reference trees (Table 1). The higher expectations of borrowing and losses using the SM approach as compared with MLN may be expected. It may be partially explained by the observation that in SM approach inferred events are driven by both "parsimonious" principles (i.e., along a branch that starts with zero and ends with one, there must be a gain event) and also driven by the continuous time Markov chain assumption (i.e., along a branch even if it starts and ends with the same character the expectation of events is >0).

Comparing our results to the borrowing frequency in basic vocabulary calculated using the same approach for 84 Indo-European languages (21) (0.6 borrowings per cognate) reveals that the borrowing frequency in Pacific languages is approximately three times higher.

Minimal Lateral Network (MLN) of Polynesian Language Evolution

In contrast to a phylogenetic tree that displays only vertical connections between a set of taxa, a phylogenetic network displays both vertical and lateral connections. A minimal lateral network (MLN) of Polynesian languages represents both vertical and lateral components of their evolutionary history. The reconstructed MLN consists of 65 vertices, representing 33 contemporary and 32 ancestral languages (i.e. internal and external nodes in the reference tree). To estimate the most parsimonious number of edges in the network we used two different algorithms optimizing either the phylogenetic distance between the connected nodes (dMLN) or the total edge weights connecting the nodes (wMLN). The dMLN minimizes the total evolutionary distance between the laterally connected languages, assuming that borrowing events are more likely to occur between closely related languages. The reasoning behind the wMLN is that borrowing events are more likely to occur between languages were many lateral connections were inferred (Fig 2). The two algorithms we used to prune superfluous edged in the MLN yielded overall similar networks (Table 2). The distribution of node degree and edge weight is similar between the dMLN and wMLN (Fig 3). Most of the dMLN and wMLN edges are identical and their weight distributions are significantly linearly correlated ($R^2_{BASIC}=0.7$, $R^2_{FULL}=0.8$, (40)). The degree and edge weight distribution of dMLN and wMLN shows similar trends in the Polynesian word borrowing frequencies (Fig 3). In what follows we present in detail the dMLN network (Fig. 2).

The internal and external vertices in the dMLN for the Polynesian basic vocabulary subset are linked by 346 lateral edges. The mean node degree is 7 edges per node with the most highly connected nodes being Pukapuka (29 edges), Tongan (27 edges) and the common ancestor of Tongan and Niuean – proto-Tongic (26 edges) (Table 2). Lateral edges between external nodes correspond to recent borrowing events. For all 33 languages at least one recent borrowing event was inferred, among these, Tongan (15 edges), New Zealand Maori (15 edges) and Pukapuka (14 edges) show a high frequency of contemporary connections. All those languages occur in specific geographical positions, Pukapuka and Tongan being in the geographical center of the languages in our dataset. Some languages such as Ifira-Mele (1 edge) and Takuu (1 edge) are rarely connected within the network even though they are not that geographically isolated.

The edge weight distribution within the dMLN is characterized by a majority of small edge weights and a minority of heavy edges. Most of the edges in the network are substantiated by only one inferred borrowing event (Table 2). The heaviest edges in the network include, an edge connecting the common ancestor of Tongan and Niuean with the common ancestor of the Western Polynesian languages (31 cognates), an edge between Fijian and the common ancestor of Tongan and Niuean (21 cognates), and one edge between Hawaiian and the common ancestor of New Zealand Maori, Rarotongan, Penrhyn, Tahitian, and Tuamotu (21 cognates). These heavy edges point to extended language contacts among these islands during their settlement in the Pacific (Table 2)

The internal and external vertices in the dMLN of the Polynesian total full lexicon subset cognates are linked by 997 lateral edges. Similar to the basic vocabulary dataset, the connectivity within each node range from few to multiple edges per language. The most highly connected node is 'Tongan'. Of the total 997 edges, only 278 (28%) are of a single laterally shared cognate, while edges of multiple cognates are frequent. The heaviest edges include an edge connecting the common ancestor of Tongan and Niuean and the ancestor of Western Polynesian languages (260 cognates), and another edge connecting between New Zealand Maori and Hawaiian (253 cognates). In full lexicon dMLN, most the of edges connect between contemporary languages, suggesting a high frequency of recent borrowings events (Table 2). As a validation of the MLN approach, we compared the "known" Pukapuka and Rotuman borrowings inferred by linguists (see the POLLEX database http://pollex.org.nz) with those inferred in are analyses. The algorithm detected 20 of the 25 Pukapuka borrowings, and 10 out of the 10 Rotuman loans.

DISCUSSION

Despite the vast oceanic distances between the Polynesian islands, and claims that their evolution should thus be highly treelike (27), our analyses revealed exceptionally high levels of conflicting signal in the distribution of the cognate sets. In total, 95% of the basic vocabulary and 98% of the full lexicon do not fit the expected tree-like language family subgrouping. Our analyses of the MLNs show that these patchy distributions cannot be explained merely by frequent cognate loss. Instead, we find remarkably high levels of borrowing. The best fitting model for both the basic vocabulary and the total lexicon was the BOR_7 model (up to seven borrowings per cognate). This is substantially more than basic vocabulary in Indo-European languages, where the BOR_1 model was the best fit (21).

One way of expressing these results is that the average level of borrowing in a language per cognate in the basic vocabulary is 51%, and 72% for the whole lexicon (Supp. Table S3). If the borrowing only occurred between two languages, this would equate to an average of approximately 25% of the basic vocabulary and 36% of the total lexicon in language being borrowed. Not only are these figures substantially higher than those in Indo-European, they are far higher than the average of 5% borrowings in basic vocabulary reported by Bowern et al. (41) in a survey of 122 hunter-gatherer languages from northwest Amazonia, northern Australia, and California and the Great Basin. These languages are spoken in small-scale societies, and in regions commonly thought to have high levels of borrowing. Rates of borrowing in basic vocabulary are generally much lower than for the total lexicon. The English lexicon, for example, contains around 60% loans from Romance languages that were borrowed after the Norman conquest, however, only 6% of the basic vocabulary is borrowed (42). Perhaps the most surprising thing about our present results for Polynesian is not that there are high levels of borrowing (36), but rather that borrowing for the basic vocabulary is so frequent.

Why might there be so much apparent borrowing in Polynesian basic vocabulary? There are three possible explanations. The first explanation is that the data we used in these analyses were patchy. Were this the case, then multiple origins for any given cognate might not indicate borrowing, but poor data sampling instead. However, there are good reasons not to believe this is the primary explanation of the data. There has been a long history of data collection for most Polynesian languages with comparative wordlists available since the 1700s (43-45). The POLLEX database we used dates to 1965 and is one of the longest-standing and largest comparative lexical databases in existence (45). Thus, there is no pressing cause to suspect a severe sampling bias in these data.

A second alternative explanation for the apparent high levels of borrowing could be semantic shift. Some of the most borrowed words are those that are known to have undergone frequent semantic shift. One example is **fatitili* (thunder). In Polynesia there are multiple reconstructions for words relating to thunder including Proto-Oceanic **kuruk* (thunder, to thunder), Proto-Central Polynesian **kurukuru* and Proto-Polynesian **mana* (thunder, supernatural force). Often these forms have shifted into other semantic slots, such as **mana*, which has been recruited as a supernatural term as impressive natural events often are (46). These competing reconstructions of near-synonyms would

switch in and out, mimicking the effect of borrowing. However, these repeated moves in semantic space are unlikely to be the main explanation for the high apparent levels of borrowing, as the reconstructions in POLLEX often cross semantic boundaries.

The third, and most likely explanation for the high levels of apparent borrowing in basic vocabulary is that that the lexical evolution in these languages departs, not just from the standard family tree model (Fig. 4a), but also from the standard borrowing model (Fig 4b). In the standard borrowing model speech communities split completely, the languages then diverge producing nested sets of innovations, and subsequently some items are borrowed between the discrete languages. As the languages have been out of contact for a considerable time items related to trade or useful parts of technology are much more likely to be borrowed than basic vocabulary. In contrast, the breakup of dialect networks is more subtle, and might be analogous to recombination during microbial speciation (47). Innovations arise in speech communities and partially diffuse along the dialect network. The slow breakup of these networks produces contradictory or incompatible innovations, as some speech communities will have partially split off from the network before the innovation has spread completely. This pattern of language differentiation will therefore produce what (48) has labeled as linkages i.e. "innovation linked" rather than "innovation defined" subgroups. As these speech communities are still in considerable contact, innovations in basic vocabulary diffuse at much the same rate as innovations in other parts of the lexicon.

The MLNs provide clear evidence for some of the Polynesian dialect networks postulated by Pawley (49). The gradual breakup of ancient dialect networks will produce connections to internal vertices of the network. The heaviest edge in the network connects the common ancestor of Tongan and Niuean with the common ancestor of the Western Polynesian languages (31 cognates, Fig. 2a). This is consistent with the messy breakup of a Proto Polynesian dialect network. The second strongest edge connects Fijian and the common ancestor of Tongan and Niuean (21 cognates) – consistent with Geraghty's suggestion of an earlier Central Pacific dialect network (50-51). The other major internal connection in the network links Hawaiian and the common ancestor of New Zealand Maori, Rarotongan, Penrhyn, Tahitian, and Tuamotu (21 cognates) and links between Eastern Polynesia (excluding EasterIsland) and Western Polynesia (18 cognates) fits with Pawley's dialect chain break up of nuclear Polynesian group (49). The inclusion of Hawaiian in this linkage conflicts with its traditional placement in the recently debated Tahitic subgroup (52), but is compatible with suggestions that Hawaii was initially settled from the Marquesas but had long-standing population contacts with Tahitic languages such as Tahitian, Penrhyn, Rarotongan (53).

What social and technological processes might have produced the strikingly high levels of borrowing across the vast distances of the Pacific Ocean that we have detected here? Evidence from studies of simulated voyaging supports the claim that the settlement of the far-flung islands of the Pacific was not the consequence of chance one-way voyages, but rather required complex sailing and navigational skills (54). These skills enabled Polynesians to remain in contact long after the initial settlement of an island. According to Irwin (54), Polynesians on many islands were forced to stay in close contact with other neighboring islands by the lack of social and ecological resources. Early European explorers were amazed both by the distances Polynesian regularly voyaged (55), and their detailed knowledge of numerous islands thousands of kilometers away (56). Archaeological evidence suggests that there was indeed substantial ongoing contact between remote islands (32-34), which could have served as a vehicle for the borrowing observed. This contact is amply reflected in the language networks. For example, the spread of the Tongan empire between 1,200-1,500 AD (28) to the islands of East Uvea, Rotuma, Futuna, Samoa and Niue is reflected in the high number of edges connected to Tongan in both the basic vocabulary and total lexicon networks (Fig. 2)

The approach we developed here shows considerable promise for future investigations of dialect networks language and borrowing. The vocabularies of Rotuman and Pukapuka are known to contain numerous borrowings. Biggs's classic 1965 paper on direct and indirect inheritance in Rotuman used the presence of phonological irregularities to document examples of Polynesian borrowing into Rotuman. A similar approach was used by Clark (57) to diagnose Eastern Polynesian borrowings into Pukapuka. Both the basic vocabulary and total lexicon networks recover this borrowing signal. The MLN approach successfully inferred all ten of the borrowings listed in POLLEX. The approach also did well in recovering 20 out of the 25 Pukapuka borrowings documented in POLLEX. However, it would be wrong to suggest that MLNs are the complete answer to the problem of detecting horizontal inheritance in language diversification. The present approach does not enable the directionality of borrowing to be inferred, and nor does it always correctly identify the correct internal edge of the network from which the borrowing occurred. Instead, MLNs are best viewed as a very useful supplement to the standard linguistic tool kit – perhaps a necessary one – when borrowing is a factor. They enable linguists to escape from the constraints of the pure family tree model without falling into the vague obscurantism of wave models or "tangled banks" (58). In combination with careful analyses of sound change, MLNs will enable increasingly more precise quantitative inferences about both borrowing and dialect networks in the diversification of language families across the globe.

METHODS

Data

The lexical data were downloaded from the June 2012 version of the POLLEX-Online database (45). One cognate in the basic dataset and eight cognates in the full dataset had only one reflex in all 33 languages (i.e. were unique). These singletons were removed from the analysis.

Reference trees

An initial language tree was inferred by applying a Bayesian approach to the binary-coded basic vocabulary data using *BEAST* 1.7 (59). Following Gray et al (60) a covarion model was used to allow for rate heterogeneity. A strict clock was enforced with a root age of 3,000 before past (BP). Since the internal classification of Polynesian languages remains controversial, several reference trees were constructed for the analysis. First, the tree topology was constrained to match the subgroupings proposed by Pawley (60-63). Second, we reconstructed an additional reference tree where the inferred position of Pukapuka was manually reassigned to a deeper ancestral node in the tree, and Rotuman was made a sister taxon to Fijian. A distance-based tree was reconstructed from the basic dataset using neighbor-joining (NJ) approach (37) using SplitsTree program (64). A maximum parsimony tree was reconstructed with the Wagner method using Mix software (65). We used *Fijian* to root the trees and topology comparisons of all reference trees used in this study are provided in the supplementary material (Supp. Fig. 2).

Borrowing estimation and minimal lateral network reconstruction

For the inference of borrowing rates we assume that the ancestral vocabulary size distribution was not fundamentally different from the vocabulary size distribution attested in the contemporary languages (21). The inference procedure is based on the testing of different evolutionary models that allow varying amounts of borrowings in order to explain the gain-loss dynamics of all cognates in a given dataset. The ancestral vocabulary size distribution was calculated using seven different evolutionary models. Two of these models (LossOnly and SingleOrigin) allow only the vertical transfer of characters. While in the LossOnly model only the loss of cognate sets is allowed, the SingleOrigin model allows one gain per cognate set. The remaining five BOR_n models (where n={1,3,7,15,31}), allow for n+1 gains (a single origin and n borrowing events), and an unlimited amount of losses per cognate set (21). The goodness of fit of each model was tested by comparing the ancestral and the contemporary vocabulary size distributions using the Wilcoxon non-parametric test (66). We additionally used a probabilistic stochastic mapping (SM) based approach to estimate the expected number of cognate gain and loss events. The expectation of cognate gains and losses along the reference tree branches were calculated using the GLOOME server (67).

The minimal lateral network (MLN) was reconstructed for both datasets using the gain-loss inference produced by the evolutionary model that

explained the data best. Branches in the MLN represent vertical inheritance and are determined by the underlying reference tree. Lateral edges in the MLN represent inferred borrowing events. They connect external (contemporary) and internal (ancestral) tree nodes (languages) whenever a gain-loss scenario involving more than one gain was inferred for at least one of the cognate sets. To represent all possible borrowing events in a cognate set of *n* gains, a total of $(n^2\text{-}n)/2$ edges connecting all nodes are required. While this connectivity approach surely covers the full realm of possibilities, it may lead to an overestimation of borrowing frequency. To solve this shortcoming of the MLN, we determine a more parsimonious frequency of *n-1* edges using a *minimal spanning tree* (MST) search algorithm (68). In other words, we treat the edges inferred for each cognate set having *n* nodes as a small network, looking for a sub-network of *n-1* edges connecting all nodes and having the smallest possible sum of edge weights. For the reduction of superfluous edges we used two different methods. In the distance-based approach (dMLN) edge weights were calculated as the evolutionary distance between the connected nodes by the total branches connecting the nodes in the tree. In the weighted frequency approach (wMLN), edge weights are calculated from the total MLN by the number of laterally shared COGs between the connected languages. From the edges connecting the nodes per COG, the set of *n-1* edges that maximizes the total edge weight in the COG are selected.

ACKNOWLEDGMENTS.

The study was supported by German Federal Ministry of Education and Research (TD, WM) and the European Research Council (grant No. 232975 to WFM; grant No. 281357 to TD). We thank Andrew Pawley and Ross Clark for useful discussion.

1. Atkinson Q, Gray R (2005) Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Syst Biol* 54:513–526.
2. Darwin C (1859) in *On the Origin of Species*, 6th Ed. Available at *http://www.gutenberg.org/etext/2009.*
3. Schleicher A (1853) Die ersten Spaltungen des indogermanischen Urvolkes [Early divergence events of the Proto-Indo-European Volk]. *Allgemeine Monatsschrift für Wissenschaft und Literatur* pp. 786–787. German.
4. Doolittle WF, Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA* 104:2043–2049.
5. Schmidt J (1872) in *Die Verwantschaftsverhältnisse der indogermanischen Sprachen* [Family relations among Indo-Germanic languages]. eds Hermann Böhlau. German.
6. Hirt H (1905) in *Die Indogermanen. Ihre Verbreitung, ihre Urheimat und ihre Kultur* [The Indo-Europeans. Their Dispersion, their Homeland, and their Culture] (Strassburg: Trübner), German.
7. Bloomfield L (1933) in *Language* (London: George Allen & Unwin Ltd).
8. Schuchardt H (1900) *Über die Klassifikation der romanischen Mundarten* [On the classification of Romance dialects]. [Probe-Vorlesung, gehalten zu Leipzig am 30 April 1870], Graz. German.
9. Pawley A (1996) in *Oceanic Culture history: Essays in Honour of Roger Green*, On the Polynesian subgroup as a problem for Irwin's continuous settlement hypothesis. eds Davidson J et al., pp. 387–410.
10. Kroeber AL (1948) in *Anthropology : race, language, culture, psychology, prehistory*. (New York: Harcourt, Brace).
11. Mace R, Holden C (2005) A phylogenetic approach to cultural evolution. *Trends Ecol Evol (Amst)* 20:116–121.
12. O'Brien, MJ, Lyman RL, Collard M, Holden CJ, Gray RD, Shennan SJ (2008) in *Cultural Transmission and Archaeology: Issues and Case Studies.* eds O'Brien MJ, Society for American Archaeology (Washington, USA), pp. 39-58.
13. Nunn CL (2011) in *The comparative approach in evolutionary anthropology and biology* (Chicago University Press).
14. Gray RD, Bryant D, Greenhill SJ (2010) On the shape and fabric of human history. *Phil Trans Roy Soc B* 365:3923–3933.
15. Borgerhoff Mulder M, Nunn CL, Towner MC (2006) Cultural macroevolution and the transmission of traits. *Evol Anthropol* 15:52–64.
16. Huson DH, Rupp R, Scornavacca C (2010) in *Phylogenetic networks: concepts, algorithms and applications* (Cambridge University Press).
17. Hurles ME, Matisoo-Smith E, Gray RD, Penny D (2003) Untangling Oceanic settlement: the edge of the knowable. *Trends Ecol Evol (Amst)* 18:531–540.
18. Bryant D, Filimon F & Gray RD (2005) in *The evolution of cultural diversity: phylogenetic approaches* eds Mace R, Holden CJ, Shennan S (London, UK: UCL Press), pp. 67–84.
19. Ben Hamed M, Wang F (2006) Stuck in the forest: trees, networks and Chinese dialects. *Diachronica* 23:29–60.
20. Bowern C (2010) Historical linguistics in Australia: trees, networks and their implications. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:3845–3854.
21. Nelson-Sathi S et al. (2011) Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proc R Soc B* 1713:1794-803.
22. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104:870–875.
23. Wells RS (1973) in Dictionary of the history of ideas, Uniformitarianism in Linguistics, eds Wiener PP (Scribner), pp 423-431.
24. Christy C (1983) in *Studies in the history of linguistics,* Uniformitarianism in linguistics (John Benjamins, Amsterdam and Philadelphia).
25. Swadesh M, Morris (1955) Towards greater accuracy in lexicostatistic dating. *Int J Am Ling* 21:121–137.
26. Sahlins MD (1958) in *Social stratification in Polynesia*. (Seattle, University of Washington Press).
27 Dixon R (1997) in *The rise and fall of languages.* (Cambridge University Press).
28. Geraghty PA (2004) in *Language contact and change in the Austronesian world*, eds Dutton and D.T. Tryon. Berlin, Mouton de Gruyter, Trends in Linguistic Studies and Monographs, pp 77: 223–349.
29. Greenhill SJ, Gray RD (2012) Basic vocabulary and Bayesian phylolinguistics: Issues of understanding and representation. *Diachronica* 29:523–537.
30. Wilson WH (1985) Evidence for an outlier source for the Proto Eastern Polynesian pronomial system. *Oceanic Linguistics*, 24:85-133.
31. Wilson WH (2012) Whence the East Polynesians?: Further Linguistic Evidence for a Northern Outlier Source. *Oceanic Linguistics* 51:289–359
32. Walter R, Sheppard PJ (1996) The Ngati Tiare adze cache: further evidence of prehistoric contact between West Polynesia and the southern Cook Islands: *Archaeology in Oceania*, 31:33–39.
33. Weisler MI, Kirch PV (1996) Interisland and interarchipelago transfer of stone tools in prehistoric Polynesia. *Proc Natl Acad Sci U S A* 93:1381–1385.
34. Weisler MI (1998) Hard Evidence for Prehistoric Interaction in Polynesia. *Curr Anthropol* 39:521–532.
35. Haspelmath M, Tadmor U (2009) in *Loanwords in the World's Languages: A Comparative Handbook* (Berlin: De Gruyter Mouton).
36. Tadmor U, Haspelmath, M, Taylor B (2010) Borrowability and the notion of basic vocabulary. *Diachronica* 2:226-246.
37. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406-425.
38. Eck RV, Dayhoff MO (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences *Science* 152:363-6.
39. Cohen O, Rubinstein ND, Stern A, Gophna U, Pupko T (2008) A likelihood framework to analyse phyletic patterns. *Phil Trans Roy Soc B* 363:3903–3911.
40. Chatterjee S, Hadi AS (1986) Influential observations, high leverage points, and outliers in linear regression. *Statistical Science* 1:379-393.
41. Bowern C et al. (2011) Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages? *PLoS ONE* 6:e25195.
42. Embleton SM (1986) in *Statistics in Historical Linguistics* (Bochum: Studienverlag Brockmeyer).
43. Pratt G (1911). Pratt's Grammar & Dictionary of the Samoan Language. (Apia, Western Samoa, Malua Printing Press).
44. Forster JR (1778) in *Observations Made During a Voyage Round the World on Physical Geography*, Natural History and Ethnic Philosophy, London, p.231.
45. Greenhill SJ, Clark R (2011) POLLEX-Online: The Polynesian Lexicon Project Online. *Oceanic Linguistics* 50:551–559.
46. Blust R (2007) Proto-Oceanic* mana revisited. *Oceanic Linguistics* 46:404-423.
47. Shapiro BJ et al. (2012) Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* 336:48–51.
48. Ross, Malcolm D (1988) in *Proto Oceanic and the Austronesian languages of western Melanesia*. (Canberra: Pacific Linguistics).
49. Pawley A (2009) Polynesian paradoxes: subgroups, wave models and the dialect geography of Proto Polynesian. *Paper presented at 11-ICAL*, Aussois.
50. Geraghty, P, (1983) Proto-Oceanic horticultural practices. M.A. thesis, Department of Anthropology, University of Auckland.
51. Geraghty, PA (2004) in *Borrowing: a pacific perspective* (eds J. Tent & P. Geraghty), Canberra: Pacific Linguistics. pp. 65–98.
52. Walworth, M (2012) Eastern Polynesian: The linguistic evidence revisited. Working Papers in Linguistics, Department of Linguistics, University of Hawai'i at Manoa, 43(5), 1-13.
53. Kirch P, Green R (2001) Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology (Cambridge Univ. Press, Cambridge, UK).
54. Irwin GJ (1998) The colonisation of the Pacific Plate: chronological, navigational and social issues *Polynesian Soc* 107:111–143.
55. Hale H (1846) United States exploring expedition during the years 1838, 1839, 1840, 1841, 1842, under the command of Charles Wilkes U.S.N: Ethnography and Philology. Philadelphia, Lea and Blanchard.
56. Salmond A (2003) *The Trial of the Cannibal Dog: Captain Cook in the South Seas*. (Penguin, New Zealand).
57. Clark R (1980) East Polynesian borrowings in Pukapukan. *J Polynesian Soc* 89:259–265.
58. 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
59. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29:1969–1973.
60. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.
61. Pawley A (1966) Polynesian languages: a subgrouping based on shared innovations in morphology *Polynesian Soc* 75:39-64.
62. Pawley A (1967) The relationships of Polynesian Outlier languages. *Polynesian Soc* 75:259-296
63. Marck JC (2000) *Topics in Polynesian language and culture history*. (Canberra: Pacific linguistics).
64. Huson DH (2005) Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol* 23:254–267.
65. Felsenstein J (2004) PHYLIP (Phylogeny Inference Package), Seattle (WA): Department of Genome Sciences, University of Washington.
66. Zar JH (1999) *Biostatistical Analysis*. (Pearson Prentice-Hall, Upper Saddle River, NJ), 4th Ed.

67. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T (2010) GLOOME: gain loss mapping engine. *Bioinformatics* 26:2914–2915.
68. Kruskal JB (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc Am Math Soc* 7.1: 48–50

4. Summary of the results

Lateral transfer is an ongoing process of natural variation. Phylogenomic networks – in contrast to phylogenetic trees - provide a better way to model genome and language evolution. Since networks can accommodate both vertical as well as lateral components of evolution, they give a dynamic picture of the larger process. Since similar evolutionary processes shaped both genomes and languages into contemporary forms, it is possible to apply methods that are developed to study genome evolution to study language evolution.

Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea

Halophilic archaeabacteria (Haloarchaea) are oxygen-respiring heterotrophs, known to involved in LGTs from eubacteria. Here the evolution of 10 haloarchaeal genomes with respect to 1,143 reference genomes were studied, and it was found that massive number of lateral transfers from eubacteria to the ancestor of Halobacteria transformed a anaerobic, chemolithoautotropic methanogen into heterotrophic oxygen-respiring haloarchaeal organism. About 1089 haloarchaeal gene families were identified that were acquired by a methanogenic recipient from eubacteria. Analyses showed that these genes were acquired by the common ancestor of haloarchaea and those transferred families include genes for catabolic carbon metabolism, membrane transporters, meanquinone biosynthesis and complexes I-IV of the eubacterial respiratory chain.

Networks uncover hidden lexical borrowing in Indo-European language evolution

Similar to ribosomal genes in genomes, basic vocabulary of languages is often assumed to be relatively resistant to borrowing. Here a phylogenomic network approach was used to recover the frequency of hidden lexical

borrowings during the evolution of Indo-European languages using the criterion of word inventory dynamics over time, proposing a general model for language evolution that captures both vertical and horizontal evolutionary components. The reconstructed Minimal Lateral Networks (MLN) that capture both the vertical and the lateral evolutionary history of 84 Indo-European languages revealed that, on average, eight percent of the words of basic vocabulary in each language were involved in borrowing during evolution. This indicates that the impact of borrowing is far more widespread than previously thought.

Polynesian language networks reveal complex history of contacts during the Pacific settlement

In the evolution of Polynesian languages, vast distances between Polynesian islands pose seemingly steep natural barriers to language contact, hence the evolutionary history of those languages expected to be tree-like. However network methods to detect both vertical and lateral components among 33 Polynesian languages reveal unexpected high levels of borrowing. At least 51% of basic vocabulary and 72% of whole lexicon of Polynesian languages experienced at least one borrowing during evolution. The estimated Polynesian lexical borrowing frequency is substantially higher than that estimated for any other language family.

The frequency of LGT and lexical borrowing during genome and language evolution clearly shows it is inappropriate to ignore the lateral component in both fields. As a result, family tree can no longer be taken as the basic model of genome or language evolution. Instead, phylogenomic networks can offer a fuller understanding of evolution in both fields.

5. Discussion

A clear mechanistic understanding of how genomes and languages evolve over time still remains as a challenging problem in evolutionary biology and historical linguistics. In the last two decades, many new approaches to phylogenetic reconstruction have been proposed to model the genealogical processes that might lead to the diversification of entities (genomes and languages). But given the frequency estimates of lateral transfers in genome and language evolution, it is very unlikely that a simple bifurcating tree that does not take in to account lateral relationships among genomes or languages is sufficient to model their evolution in a realistic manner. In biology, recent transitions from single gene analysis to whole genome comparisons of entire microbial populations enhanced our understanding of evolution, while questioning the early assumptions of a tree like microbial evolution process. Current whole genome analyses do not support a bifurcating tree of life, instead they favour more realistic pictures involving phylogenetic networks (Doolittle 1999, Kunin et al. 2005), which can better represent the true relationships (vertical and lateral) among genomes that are characterized by high rate of LGTs. Similarly, language change is not only based on the modification of inherited items but also driven by direct or indirect exchange of units. Bifurcating trees can only provide a reduced version of their evolution often may also be misleading. Network approaches are a straightforward way to solve this problem and they can combine both vertical and lateral component, providing a more realistic picture of evolution. Thus networks approaches are best viewed as a useful supplement to the existing evolutionary toolkit to model both genome and language evolution.

6. References

Aikhenvald AY (2006) in *Grammars in Contact: A Cross-linguistic Typology,* eds Aikhenvald AY, Dixon RM (*Oxford University Press,* Oxford*)* pp 1-66.

Allers T, Mevarech M (2005) Archaeal genetics — the third way. *Nat Rev Genet* 6:58–73.

Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet* 8:450-461.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe FJ, Dupré J, Dagan T, Boucher Y, Martin W (2009) Prokaryotic evolution and the tree of life are two different things. *Biol Direct* 4:34.

Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 102:14332–14337.

Bergsland K, Vogt H (1962) On the Validity of Glottochronology. *Curr Anthropol* 3:115-153.

Bonfante G (1931) in *I Dialetti Indoeuropei* (Istituto Orientale di Napoli, Annali del R*)* 4, 69–185.

Bryant D, Filimon F, and Gray RD (2005) in *The Evolution of Cultural Diversity: A Phylogenetic Approach*, eds Mace R (UCL Press) 67-84.

Chen I, Christie PJ, Dubnau D (2005) The ins and outs of DNA transfer in bacteria. *Science* 310:1456–1460.

Dagan T (2011) Phylogenomic networks. *Trends Microbiol* 19:483–491.

Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 105:10039–10044.

Dagan T, Martin W (2006) The tree of one percent. *Genome Biol* 7**:**118

Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104:870–875.

Dagan T, Martin W (2009) Getting a better picture of microbial evolution en route to a network of genomes. *Philos Trans R Soc Lond B Biol Sci* 364:2187–2196.

Darwin C (1859) in *On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life* (John Murray, London).

Dayhoff MO (1969) Computer Analysis of Protein Evolution. *Scientific American* 221:86–95.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R et al. (2002) The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* 4:453–461.

Doolittle WF (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14:307–311.

Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284:2124–2129.

Doolittle WF (1999) Lateral genomics. *Trends Cell Biol* 9:M5–M8.

Doolittle WF (2004) in *Microbial Phylogeny and Evolution*: *Concepts and Controversies*, eds Sapp J (Oxford University Press, New York) pp 119-133.

Dubey GP, Ben-Yehuda S (2011) Intercellular Nanotubes Mediate Bacterial Communication. *Cell* 144:590–600.

Embleton S (2000) in *Time depth in historical linguistics*, eds Renfrew C, McMahon A, Trask L (The McDonald Institute for Archaeological Research, Cambridge) pp 143-165.

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. *Science* 155:279–284.

Fox A (1995) in *Linguistic Reconstruction: An Introduction to Theory and Method* (Oxford University Press, Oxford).

Freeman VJ (1951) Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. *J Bacteriol* 61:675–688.

Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 3:679–687.

Haggerty LS, Martin FJ, Fitzpatrick DA, McInerney JO (2009) Gene and genome trees conflict at many levels. *Philos Trans R Soc Lond B Biol Sci* 364:2209–2219.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci USA* 107:127–132.

Hamed MB, Wang F (2006) Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23:29–60.

Hock HH , Joseph BD (1995) in *An Introduction to Historical and Comparative Linguistics* (Mouton de Gruyter, Berlin and New York).

Huson DH , Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.

Huson DH, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol* 3:23–35.

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806.

Ji XL (1989) in *Encyclopedia of China: Language and Character* (Encyclopedia of China Publishing House, Beijing, China).

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 55:709–742.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: Reconstructing the microbial phylogenetic network. *Genome Res* 15:954–959.

Lang AS, Beatty JT (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol* 15:54–62.

Lawrence JG, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci USA* 95:9413–9417.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* 25:762–777.

Majewski J (2001) Sexual isolation in bacteria. *FEMS Microbiol Lett* 199:161-169

Martin W (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* 21:99–104.

Martin W (2003) Gene transfer from organelles to the nucleus: Frequent and in big chunks. *Proc Natl Acad Sci USA* 100:8612–8614.

McMahon A and McMahon R (2005) in *Language Classification by Numbers* (Oxford University Press, Oxford).

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3:2.

Mongodin , Nelson KE, Daugherty S, Deboy RT, Wister J, Khouri H, Weidman J, Walsh DA, Papke RT, Sanchez Perez G et al. (2005) The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc Natl Acad Sci USA* 102:18147–18152.

Naik GA, Bhat LN, Chopade BA, Lynch JM (1994) Transfer of Broad-Host-Range Antibiotic-Resistance Plasmids in Soil Microcosms. *Curr Microbiol* 28:209–215.

Nakhleh L, Ringe DA, Warnow T (2005) Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language* 81:382–420.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 399:323–329.

Nelson-Sathi S, List JM, Geisler H, Fangerau H, Gray RD, Martin W, Dagan T (2011) Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proc Roy Soc Lond* 1713:1794-803.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci USA* 109:20537-42.

Newman MEJ (2010) in *Networks: An Introduction* (Oxford University Press, Oxford).

Nielsen KM (1998) Barriers to horizontal gene transfer by natural transformation in soil bacteria. *APMIS Suppl* 106:77–84.

Norman A, Hansen LH, Sørensen SJ (2009) Conjugative plasmids: vessels of the communal gene pool. *Philos Trans R Soc Lond B Biol Sci* 364:2275–2289.

Ochiai K, Yamanaka T, Kimura K, Sawada, O (1959). Inheritance of drug resistance (and its transfer) between Shigella strains and Between Shigella and E. coli strains. Hihon Iji Shimpor (in Japanese)1861: 34.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.

Orel V (2003) in *A handbook of Germanic etymology* (Brill, Leiden).

Pagel M (2009) Human language as a culturally transmitted replicator. *Nat Rev Genet* 10:405-415.

Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37:1372–1375.

Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* 14:615–623.

Popa O, Hazkani-Covo E, Landan G, Martin W and Dagan T (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* 21: 599–609.

Ragan MA (2009) Trees and networks before and after Darwin. *Biol Direct* 4:43–discussion 43.

Schleicher A (1863) in *Die Darwinsche Theorie und die Sprachwissenschaft Offenes Sendschreiben an Herrn Ernst Häckel.* (Böhlau, Weimar).

Schuchardt H (1900) *Über die Klassifikation der Romanischen Mundarten*. [Probe-Vorlesung, gehalten zu Leipzig am 30 April 1870], Graz.

Soria-Carrasco V, Castresana J (2008) Estimation of phylogenetic inconsistencies in the three domains of life. *Mol Biol Evol* 25:2319–2329.

Soukhanov AH (1992) in *The American heritage dictionary of the English language* (Mifflin, Boston).

Syvanen M (1985) Cross-species gene transfer; implications for a new theory of evolution. *J Theor Biol* 112:333–343.

Strogatz SH(2001) Exploring complex networks. *Nature*. 410:268-276.

Tadmor U (2009) in *Loanwords in the world's languages, A comparative Handbook,* eds Haspelmath M and Tadmor U (de Gruyter Berlin and New York), 55–75.

Thomas CM, Nielsen KM (2005) Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat Rev Microbiol* 3:711–721.

Thomason S, Kaufman T (1988) in *Language Contact, Creolization, and Genetic Linguistics* (University of California Press, Berkeley).

Trask RL (2000) in *The Dictionary of Historical and Comparative Linguistics* (Edinburgh Univ. Press, Edinburgh).

Weinreich U (1953) in *Languages in contact*. With a preface by André Martinet. (Mouton, The Hague and Paris), 8th ed.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87:4576–4579.

7. Appendix

7.1 Abbreviations

DNA	Deoxyribonucleic Acid
GTA	Gene Transfer Agent
HGT	Horizontal Gene Transfer
LGT	Lateral Gene Transfer
MLN	Minimal Lateral Network

7.2 Supplementary material

This section lists supplementary material of three publications described in the thesis.

Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea

Shijulal Nelson-Sathi[1], Tal Dagan[2], Giddy Landan[1,2], Arnold Janssen[3], Mike Steel[4], James McInerney[5], Uwe Deppenmeier[6], William F. Martin[1*]

SUPPORTING ONLINE INFORMATION

Because of larger size of the table, Supp. Table S3 and Supp. Table S4 are not included in the thesis, they are accessible at
http://www.pnas.org/content/early/2012/11/21/1209119109/suppl/DCSupplemental

Supporting Information

Nelson-Sathi et al. 10.1073/pnas.1209119109

SI Text

Statistical Methods. The task at hand is to compare two collections of trees, 367 trees reconstructed from recipient genes and 109 trees reconstructed from imported genes. The trees in each set differ from one another, either due to noisy data or due to estimation errors and biases, but our null hypothesis is that genes in both sets evolved along the same phylogeny from a single origin and therefore should display the same phylogenetic signal. In the alternative scenarios, the trees are not related by the same underlying phylogeny, either because of multiple origins or due to lateral gene transfer (LGT) between lineages. To gain a perspective on how those alternate scenarios will look, we generated two additional synthetic datasets: 109 random trees sampled uniformly from the entire tree space and 109 one-LGT trees, constructed by a minimal perturbation of the imported dataset where a random subtree was pruned and then regrafted at a random branch of the remaining trunk. This simulates a single lateral transfer event from the grafting branch to the pruned clade.

The phylogenetic signal contained within each tree can be summarized in several ways (1). We have examined three basic units of phylogenetic information: phylogenetic partitions (splits), taxa quartets assertions, and triple taxa assertions. Splits and quartets were applied to both the rooted and unrooted versions of the trees, for a total of five phylogenetic signal units.

To test the hypotheses: H_0: Trees in the two sets are drawn from the same underlying tree distribution, vs. H_1: The two sets of trees differ in their underlying phylogenetic signal, we have developed three methodologies: goodness of fit between tree distributions, Euclidean distance between frequencies of phylogenetic assertions, and comparison of distances to a common consensus tree.

Goodness of fit between tree distributions. The two sets of trees were recorded into a $2\times m$ contingency table, where the m categories were defined in an adaptive procedure based on one of the five phylogenetic units. First, the two samples were pooled together into a single set of size n, and the n trees converted into tuples of phylogenetic assertions, or states. Each state was ranked according to its frequency in the pooled state sets. Next, each tree was labeled by the rank of its lowest ranking state, and the pooled tree set was sorted by this label. Bins were defined as a collection of states by sequential addition of states from the sorted list, and creation of a new bin when the current bin included at least $\sqrt{n}$ trees, resulting in $m \leq \sqrt{n}$ bins (the choice of $\sqrt{n}$ is a common practice to ensure a balance between the number of bins and the average sample size for each bin). In the last step, trees from the two sets were added to a $2\times m$ contingency table (with the two rows corresponding to the two sets) based on their label, i.e., their least ranked state. The resulting contingency table was used to derive a standard goodness-of-fit statistic (2). The significance of the goodness-of-fit statistic was tested in a permutation test and the P value estimated from a Monte Carlo simulation with 10^5 permutations. One advantage of the goodness-of-fit statistic is that asymptotically it is χ^2 distributed with $m-1$ degrees of freedom, and the P value can be approximated using the χ^2_{m-1} cumulative distribution function (Table S5*A*).

Euclidean distance between frequencies of phylogenetic assertions. Each of the two sets of trees was converted to a set of phylogenetic assertions, using one of the five phylogenetic units. The two distributions of phylogenetic states were represented as frequency vectors, and the similarity between the two sets was measured by the Euclidean distance between the two frequency vectors. The significance of the Euclidean distance statistic was tested in a permutation test and the P value was estimated from a Monte Carlo simulation with 10^5 permutations (Table S5*B*).

Comparison of distances to a common consensus tree. First, a greedy consensus tree (3) was computed from the pooled set of trees. Next, the distance from the pooled consensus to each tree in the two tree sets was calculated based on one of the five phylogenetic units (1). The distributions of the tree distances for the two sets of trees were compared using the Kolmogorov–Smirnov test (2). (Table S5*C*).

Phylogenetic compatibility with a reference set. The comparison of sets of trees by the foregoing methodologies is applicable only when all trees include the same set of taxa. To extend the analysis to trees that include only a subset of taxa, we examined such trees in terms of their phylogenetic compatibility with a reference set comprised of all recipient trees that do include the full set of taxa. Recipient and imported trees that include only a subset of taxa were grouped based on the number of taxa n, and each group was analyzed separately. Each n taxon tree was decomposed into its $(n-3)$ splits, and each split was scored by the fraction of splits in the reference set that are phylogenetically compatible with it. The split compatibility scores for all splits of all trees in the group forms the split compatibility distributions of the group. Additionally, the $(n-3)$ split compatibility scores of a specific tree were averaged to produce a tree compatibility score. The distributions of compatibility scores for the recipient and imported groups of trees were compared using the Kolmogorov–Smirnov test (2). (Table S5*D*).

Multiple copy genes. The foregoing tests can be applied only when gene families are present as (at most) single copies (SC) in the several genomes. To apply the tests to trees where multiple copies (MC) of a gene are present in some genomes, we converted the MC trees into SC-like trees by removal of some of the additional copies, using several removal strategies:

i) Condensing of tips: When all copies of a gene in a specific genome form a monophyletic clade in the tree, they can be condensed into a single leaf without affecting the phylogenetic relationships between the several taxa. Only a few MC trees could be converted into SC trees using this strategy.

ii) Retaining exactly one copy per genome: In this approach, we created two sets of SC-like trees, one containing the copies that best fit a reference tree and the second containing the copies with the worst fit to the reference tree. A MC tree was first reduced to a collection of SC-like subtrees by taking all possible combinations of a single copy form each of the several genomes. Next we scored each of the subtrees by its compatibility with the reference tree and retained the two extreme scoring trees as members of the best/worst sets. When several trees were tied with minimal/maximal score, we randomly selected one of the tied trees. We restricted this approach to cases where there are less than 1,024 possible subtrees, only a few cases of very high copy number MC trees were omitted due to this restriction.

iii) Retaining only those genomes where the gene is present in a single copy. This approach can be applied to all MC trees, but some of the resulting SC-like subtrees have less than four taxa and are therefore uninformative.

The goodness-of-fit tests are shown in Table S5*E*, and the tree compatibility tests in Table S5*F*.

Power of the goodness-of-fit test. The goodness-of-fit test based on unrooted splits is powerful enough to reject the recipient vs. one-

LGT comparison. In the one-LGT dataset, every gene is affected by one LGT, raising the question how will the test fair if only some of the genes are affected by LGT. To address this question, we repeated the analysis using random mixtures of the one-LGT and imported datasets (Fig. S3). The goodness-of-fit test based on unrooted splits is powerful enough to reject a mixture of 34% LGT/66% imports at the 5% level.

Common conflicting splits. In Fig. 2*B* (modified version reproduced as Fig. S4), we observed that the six most common splits are compatible and that the tree they define is identical to the haloarchaeal phylogeny generated by 56 universally distributed archaebacterial genes. Moreover, these six splits comprise 51% and 46% of the splits in the recipient and imported sets, respectively. However, other splits are also present in a sizeable proportion of the trees. For example, splits ranking as the 7th to 20th most common are present in about 10% of the trees. The question arises whether these splits indicate an alternative biological signal or whether they are the result of random phylogenetic reconstruction error. If the next 12 or so splits are attributable to random phylogeny errors (as opposed to a biological signal), then the most frequent splits should correspond to alternative topologies that are very close to the reference tree (only one branch being "wrong," for example). If, on the other hand, it is a biological signal, there should be no correlation between split frequency and topological distance to (compatibility with) the reference tree. In Fig. S4, which is a modified version of Fig. 2*B*, we plotted the compatibility of splits with the reference tree (which is also the tree for the first six splits), alongside the split frequencies in the recipient and imported trees.

Clearly, the most frequent splits that are incompatible with the reference tree are also those that are most compatible with it. The correlation is very high (Spearman rank correlation $r = 0.75$; $P = 7{\cdot}10^{-13}$ for the recipients, $r = 0.76$; $P = 7{\cdot}10^{-19}$ for the imports). This strongly indicates that there is no alternative biological signal in this data, but that the second-best splits are behaving exactly as one would expect for the case that phylogeny methods are doing the best they can, but are slightly imperfect.

1. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
2. Zar JH (2010) *Biostatistical Analysis* (Prentice Hall, Upper Saddle River, NJ), 5th Ed.
3. Bryant D (2003) A classification of consensus methods for phylogenies. *BioConsensus*, eds Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS (American Mathematical Society, Providence, RI), pp 163–183.

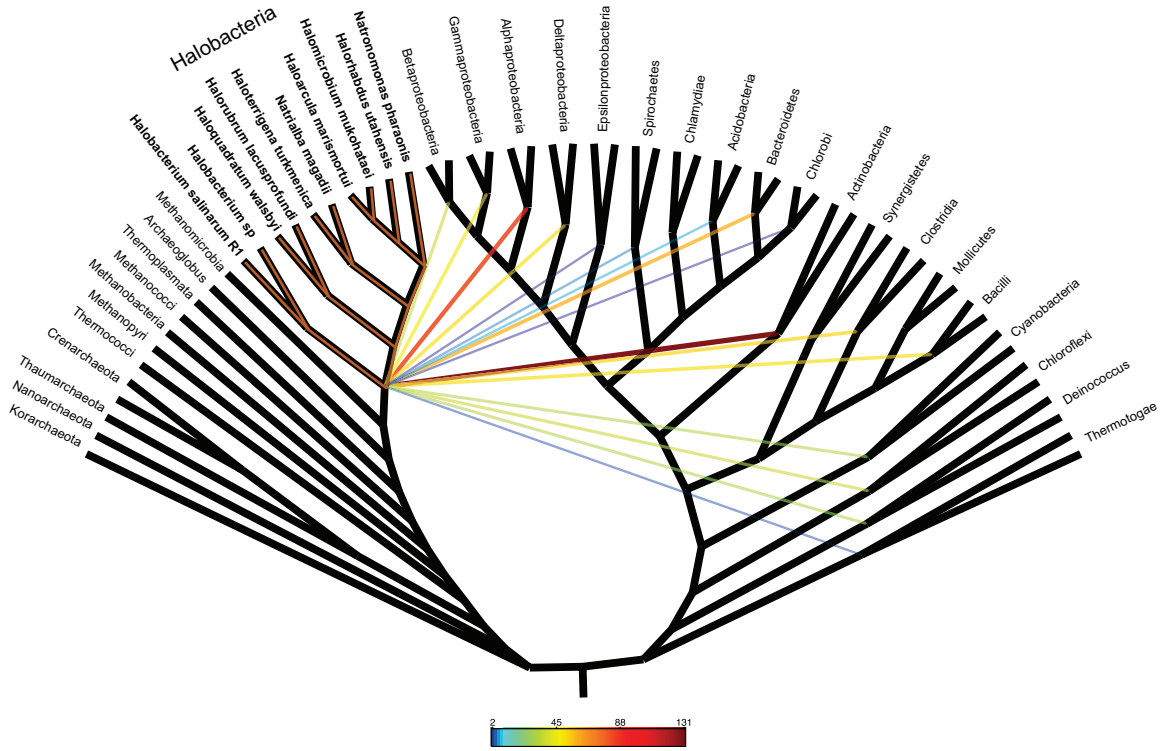


Fig. S1. Acquisition network showing sole donor lineages in what is best understood as a single acquisition from a chimeric donor genome.

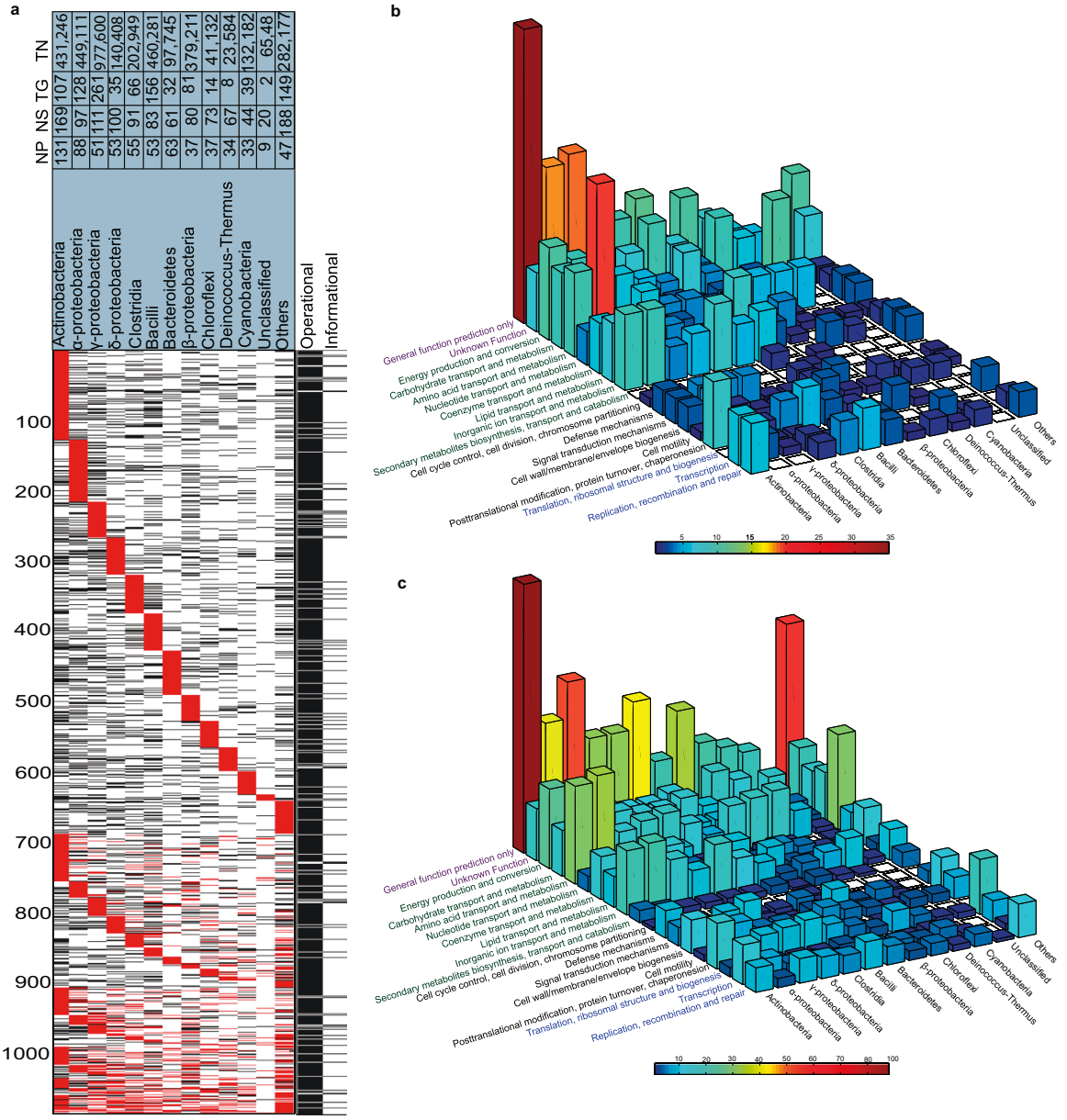


Fig. S2. Phylogenetic affinities and functional classes of eubacterial genes imported into Haloarchaea. (*A*) Presence of eubacterial groups in the sister clade to the haloarchaeal imports (red) and presence in the tree but not in the sister clade (black). The assignment to informational and operational classes for each import is indicated on the right hand side of *A*. Numbers in *A*, *Top* are as follows: NP, number of trees in which the taxon was the only taxon present in the sister clade to the Haloarchaea (the top 691 entries); NS, number of times that the taxon was present in the sister clade to the Haloarchaea (either the sole taxon present or in addition to other taxa); TG, number of genomes sampled for the taxon; TN, total number of genes sampled for the taxon. (*B*) Number of trees in which the taxon was the only taxon present in the sister clade to the Haloarchaea plotted against functional categories. (*C*) Number of trees in which the taxon was present in a mixed sister clade plotted against functional categories.

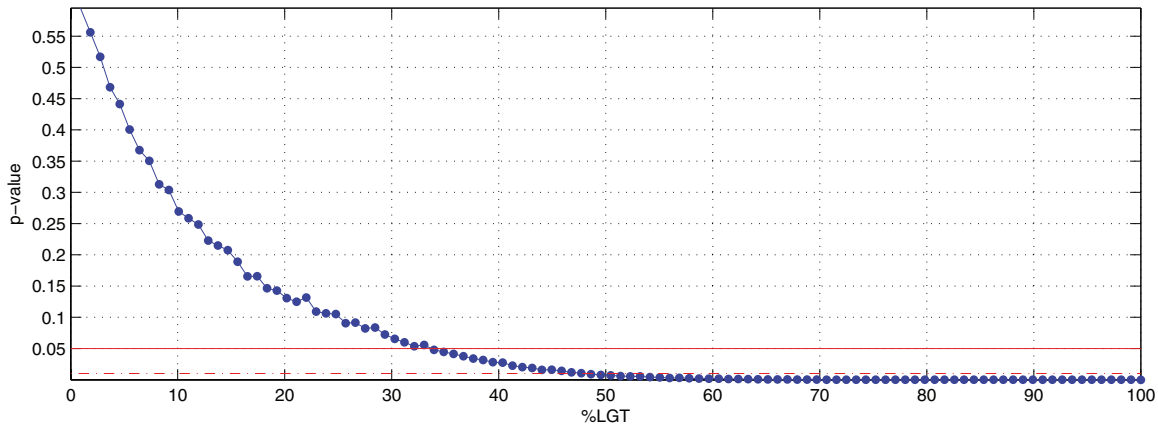


Fig. S3. Power of the goodness-of-fit test.

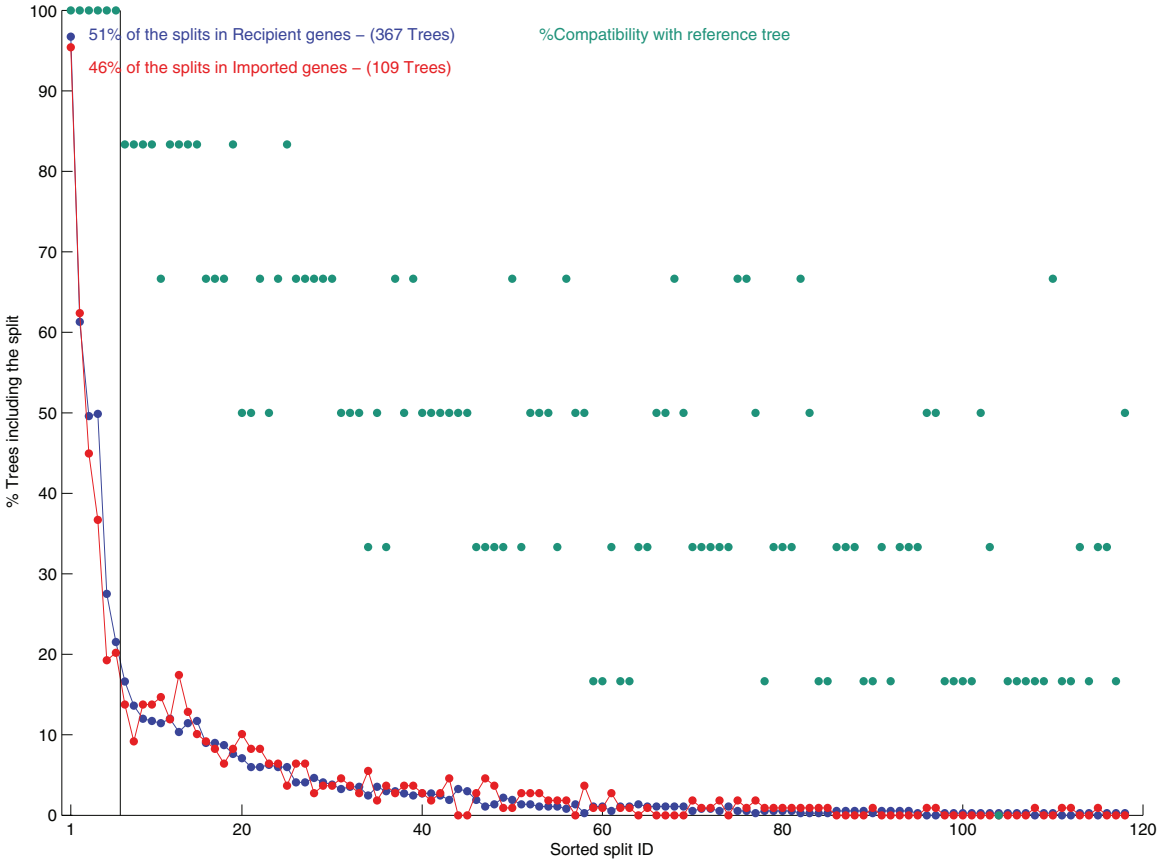


Fig. S4. Split frequencies and their compatibility with the reference tree.

Table S1. Functional categories and distribution of eubacterial imports in Haloarchaea

Table S1

Table S2. Total of 1,089 imports and their functional annotations

Table S2

Table S3. Functional categories and distribution of eubacterial imports in Methanosarcinales (Ms), Methanomicrobialesm (Mm), and Methanocellales (Mc)

Table S3

Table S4. Gene names, functional annotations, and gene distribution among Haloarchaea for components of the respiratory chain

Table S4

Table S5. Statistical tests

Table S5

Table S1

Function: COG category	No.	Presence in 10 Haloarchaea: Tr.	2	3	4	5	6	7	8	9	10
Information storage and processing:	**84 (8%)**										
[J] Translation, ribosome struct. and biogenesis	8	.	.	3	2	.	.	.	.	1	2
[K] Transcription	32	.	5	5	2	4	5	1	3	2	6
[L] Replication, recombination and repair	44	.	14	5	2	1	1	2	2	6	11
Cellular processes and signaling:	**146 (14%)**										
[D] Cell cycle, cell div., chromosome partitioning	3	.	1	.	1	.	6	.	.	.	.
[V] Defense mechanisms	15	8	5	1	2	2	2	.	.	.	3
[T] Signal transduction mechanisms	40	.	13	10	6	3	4	1	.	.	3
[M] Cell wall/membrane/envelope biogenesis	33	.	10	6	5	1	2	2	2	4	1
[N] Cell motility	4	.	2	2	.	.	.	.	.	.	.
[O] Posttransl. mod., prot. turnover, chaperones	51	5	11	7	4	5	3	2	1	3	15
Metabolism:	**482 (44%)**										
[C] Energy production and conversion	95	.	9	18	10	5	5	7	12	6	23
[G] Carbohydrate transport and metabolism	57	15	20	8	7	6	1	2	1	7	4
[E] Amino acid transport and metabolism	114	34	32	21	7	8	9	11	11	3	12
[F] Nucleotide transport and metabolism	32	4	7	3	4	2	3	1	.	2	10
[H] Coenzyme transport and metabolism	54	2	7	6	5	3	.	7	7	8	11
[I] Lipid transport and metabolism	30	1	4	3	2	3	4	1	7	2	4
[P] Inorganic ion transport and metabolism	86	63	30	9	7	11	5	4	10	5	5
[Q] Secondary metabolite biosynth. and transport	14	1	1	4	1	3	2	1	.	.	2
Poorly characterized:	**377 (34%)**										
[R] General function prediction only	276	34	59	45	30	29	27	21	20	15	30
[S] Unknown function	101	.	26	18	7	7	6	9	3	4	20

Tr., number of clusters in that category annotated as transporters, importers, or translocators.

Table S2 Available at
http://www.pnas.org/content/early/2012/11/21/1209119109/suppl/DCSupplemental

Table S3

Function:	Presence in 12 Ms–Mm–Mc												Presence in 5 Ms						Presence in 5 Mm						Presence in 2 Mc		
COG category	No.	Tr.	2	3	4	5	6	7	8	9	10	11	12	No.	Tr.	2	3	4	5	No.	Tr.	2	3	4	5	No.	Tr.	2
Information storage and processing:	**8 (6%)**													**33 (8%)**						**2 (5%)**						**18 (9%)**		
[J] Translation, ribosome struct. and biogenesis	·	·	·	·	·	·	·	·	·	·	·	·	·	3	·	1	1	1	·	·	·	·	·	·	·	2	·	2
[K] Transcription	4	·	·	2	2	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	7	·	7
[L] Replication, recombination and repair	4	·	·	2	1	·	1	·	·	·	·	·	·	30	·	19	9	1	·	2	·	2	·	·	·	9	·	9
Cellular processes and signaling:	**14 (11%)**													**47 (11%)**						**5 (12.5%)**						**37 (17%)**		
[V] Defense mechanisms	3	2	2	1	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	2	·	2	·	·	·	7	·	7
[T] Signal transduction mechanisms	·	·	·	·	·	·	·	·	·	·	·	·	·	20	·	12	7	1	·	1	·	·	1	·	·	8	·	8
[M] Cell wall/membrane/envelope biogenesis	7	·	2	2	1	2	·	·	·	·	·	·	·	15	·	4	8	3	·	·	·	·	·	·	·	9	·	9
[N] Cell motility	·	·	·	·	·	·	·	·	·	·	·	·	·	2	·	2	·	·	·	·	·	·	·	·	·	·	·	·
[O] Posttransl. mod., prot. turnover, chaperones	3	·	2	·	·	·	1	·	·	·	·	·	·	8	·	2	3	3	·	2	·	2	·	·	·	13	·	13
[U] Intracell., traffick, sec., vescl., transport	1	1	1	·	·	·	·	·	·	·	·	·	·	2	·	·	1	1	·	·	·	·	·	·	·	·	·	·
Metabolism:	**46 (37%)**													**108 (26%)**						**12 (30%)**						**64 (30%)**		
[C] Energy production and conversion	13	·	1	4	2	4	·	1	·	·	·	·	1	22	·	14	6	2	·	4	·	4	·	·	·	8	·	8
[G] Carbohydrate transport and metabolism	7	·	2	·	2	3	·	·	·	·	·	·	·	11	·	5	4	2	·	4	·	4	·	·	·	13	·	13
[E] Amino acid transport and metabolism	6	4	·	·	3	·	3	·	·	·	·	·	·	29	6	11	10	8	·	1	1	1	·	·	·	9	·	9
[F] Nucleotide transport and metabolism	3	·	2	1	1	·	·	·	·	·	·	·	·	3	·	2	1	·	·	1	·	1	·	·	·	3	·	3
[H] Coenzyme transport and metabolism	3	·	1	1	1	·	·	·	·	·	·	·	·	16	·	10	4	2	·	·	·	·	·	·	·	11	·	11
[I] Lipid transport and metabolism	2	·	·	1	1	·	·	·	·	·	·	·	·	4	·	3	1	·	·	1	·	·	1	·	·	2	·	2
[P] Inorganic ion transport and metabolism	8	6	·	4	1	·	2	1	·	·	·	·	·	21	15	11	6	4	·	1	·	1	·	·	·	17	10	17
[Q] Secondary metabolite biosynth. and transport	4	·	1	1	1	·	1	·	·	·	·	·	·	2	1	1	·	·	·	·	·	·	·	·	·	1	·	1
Poorly characterized:	**56 (45%)**													**230 (55%)**						**145 (47%)**						**92 (44%)**		
[R] General function prediction only	20	·	4	3	5	7	1	·	·	·	·	·	·	53	·	18	22	12	1	2	·	2	·	·	·	20	·	20
[S] Unknown function	36	·	10	10	3	2	4	4	1	·	·	1	1	177	·	87	69	18	3	19	·	17	2	·	·	72	·	72

Tr., number of clusters in that category annotated as transporters, importers, or translocators.

Genomes sampled:

Methanosarcinales: *Methanosarcina mazei* Gö, *Methanosarcina acetivorans*, *Methanosarcina barkeri* fusaro, *Methanococcoides burtonii* DSM6242, *Methanosaeta thermophila* PT.

Methanomicrobiales: *Methanoculleus marisnigri* JR1, *Methanospirillum hungatei* JF-1, *Candidatus Methanosphaerula palustris* E1 9c, *Candidatus Methanoregula boonei* 6A8, *Methanocorpusculum labreanum* Z.

Methanocellales: *Methanocella paludicola* SANAE and Uncultured methanogenic archaeon RC-I, which is not classified as methanocellales in GenBank taxonomy, but it always branched with Methanocella paludicola SANAE in our analyses, for which reason it was treated as an Mc member here.

Table S4

Available at *http://www.pnas.org/content/early/2012/11/21/1209119109/suppl/DCSupplemental*

Table S5

A. Goodness-of-fit	**Recipient vs.**		
	Imported	**One LGT**	**Random**
9 Taxa 367 Recipient vs. 109 Other	χ^2 (MC) *p-value*	χ^2 (MC) *p-value*	χ^2 (MC) *p-value*
Unrooted splits bins	0.543 (0.552)	$<10^{-16}$ ($<10^{-5}$)	$<10^{-16}$ ($<10^{-5}$)
Unrooted quartets bins	0.167 (0.167)	$2.66 \cdot 10^{-9}$ ($<10^{-5}$)	$<10^{-16}$ ($<10^{-5}$)
Rooted splits bins	0.859 (0.866)	$3.40 \cdot 10^{-5}$ ($3 \cdot 10^{-5}$)	$<10^{-16}$ ($<10^{-5}$)
Rooted quartets bins	0.933 (0.936)	$1.77 \cdot 10^{-10}$ ($<10^{-5}$)	$<10^{-16}$ ($<10^{-5}$)
Rooted triplets bins	0.507 (0.510)	$1.42 \cdot 10^{-7}$ ($<10^{-5}$)	$<10^{-16}$ ($<10^{-5}$)

MC: 10^5 Monte Carlo permutations.

B. Euclidean distance	**Recipient vs.**		
	Imported	**One LGT**	**Random**
9 Taxa 367 Recipient vs. 109 Other	MC *p-value*	MC *p-value*	MC *p-value*
Unrooted splits frequencies	0.257	$<10^{-5}$	$<10^{-5}$

Unrooted quartets frequencies	0.164	$<10^{-5}$	$<10^{-5}$
Rooted splits frequencies	0.062	$<10^{-5}$	$<10^{-5}$
Rooted quartets frequencies	0.041	$<10^{-5}$	$<10^{-5}$
Rooted triplets frequencies	0.030	0.002	$<10^{-5}$

MC: 10^5 Monte Carlo permutations.

C. Distances to consensus tree	**Recipient vs.**		
	Imported	**One LGT**	**Random**
9 Taxa 367 Recipient vs. 109 Other	KS *p-value*	KS *p-value*	KS *p-value*
Unrooted splits distances	0.190	$2.69 \cdot 10^{-11}$	$5.37 \cdot 10^{-48}$
Unrooted quartets distances	0.021	$2.61 \cdot 10^{-9}$	$2.58 \cdot 10^{-47}$
Rooted splits distances	0.530	$9.83 \cdot 10^{-7}$	$3.17 \cdot 10^{-45}$
Rooted quartets distances	0.100	$1.54 \cdot 10^{-9}$	$8.58 \cdot 10^{-44}$
Rooted triplets distances	0.376	$5.12 \cdot 10^{-4}$	$3.68 \cdot 10^{-17}$

KS: Kolmogorov-Smirnov two sample test.

D. Compatibility with a reference set

D1. Tree compatibility

Number of OTUs	Number of trees		Recipient vs. Imported	Recipient vs. One LGT	Recipient vs. Random
	Recipient	Other	KS *p-value*	KS *p-value*	KS *p-value*
4	16	80	0.768	0.045	0.004
5	7	55	0.255	0.040	$7.9 \cdot 10^{-04}$
6	16	60	0.081	0.745	0.002
7	23	48	0.673	0.005	$1.5 \cdot 10^{-08}$
8	47	57	0.094	$2.5 \cdot 10^{-05}$	$4.9 \cdot 10^{-19}$

D2. Split compatibility

Number of OTUs	Number of splits		Recipient vs. Imported	Recipient vs. One LGT	Recipient vs. Random
	Recipient	Other	KS *p-value*	KS *p-value*	KS *p-value*
4	16	80	0.768	0.045	0.004
5	14	110	0.139	0.010	$1.8 \cdot 10^{-04}$
6	48	180	0.064	0.509	$1.9 \cdot 10^{-05}$
7	92	192	0.141	$2.9 \cdot 10^{-06}$	$8.3 \cdot 10^{-20}$
8	235	285	0.768	0.045	0.004

KS: Kolmogorov-Smirnov two sample test. Significant at 5% FDR[4]

E. Goodness-of-fit tests including MC trees, based on unrooted splits bins	Number of trees		**Recipient vs.**		
			Imported	**One LGT**	**Random**
	Recipient	Other	χ^2_{m-1} χ^2 *p-value*	χ^2 *p-value*	χ^2 *p-value*
Single copy (from A)	367	109	0.543	$<10^{-16}$	$<10^{-16}$
Condensed tip duplicates	371	114	0.588	$2.47 \cdot 10^{-11}$	$<10^{-16}$
Best copy of multiple copies	432	162	0.184	$3.92 \cdot 10^{-13}$	$<10^{-16}$
Worst copy of multiple copies	432	162	0.487	$1.14 \cdot 10^{-09}$	$<10^{-16}$

F. Tree compatibility tests including MC trees

F1. Condensed tip duplicates

Number of OTUs	Number of trees		**Recipient vs.**		
			Imported	**One LGT**	**Random**
	Recipient	Other	KS *p-value*	KS *p-value*	KS *p-value*
4	16	88	0.753	0.053	0.002
5	9	58	0.101	0.010	$3.5 \cdot 10^{-05}$
6	17	64	0.076	0.869	$6.4 \cdot 10^{-04}$
7	23	48	0.673	0.005	$1.5 \cdot 10^{-08}$
8	49	57	0.127	$2.5 \cdot 10^{-05}$	$4.8 \cdot 10^{-18}$

F2. Best copy of multiple copies

Number of OTUs	Number of trees		**Recipient vs.**		
			Imported	**One LGT**	**Random**
	Recipient	Other	KS *p-value*	KS *p-value*	KS *p-value*
4	27	111	0.724	0.007	$2.5 \cdot 10^{-04}$
5	16	95	0.069	$7.5 \cdot 10^{-04}$	$1.9 \cdot 10^{-06}$
6	22	93	0.182	0.417	$8.6 \cdot 10^{-04}$
7	28	81	0.536	0.004	$2.8 \cdot 10^{-13}$
8	69	85	0.034	$2.5 \cdot 10^{-05}$	$2.8 \cdot 10^{-28}$

F3. Worst copy of multiple copies

Number of OTUs	Number of trees		**Recipient vs.**		
			Imported	**One LGT**	**Random**
	Recipient	Other	KS *p-value*	KS *p-value*	KS *p-value*
4	27	111	0.435	0.038	0.038
5	16	95	0.043	$5.1 \cdot 10^{-04}$	$4.3 \cdot 10^{-05}$
6	22	93	0.213	0.281	0.018
7	28	81	0.053	$4.7 \cdot 10^{-04}$	$1.6 \cdot 10^{-08}$
8	69	85	0.015	$5.4 \cdot 10^{-06}$	$4.0 \cdot 10^{-24}$

F4. Taxa with multiple copy removed			**Recipient vs.**		
Number of OTUs	Number of trees		**Imported**	**One LGT**	**Random**
	Recipient	Other	KS *p-value*	KS *p-value*	KS *p-value*
4	23	123	0.383	0.013	$1.4 \cdot 10^{-04}$
5	20	88	0.014	0.001	$8.8 \cdot 10^{-08}$
6	32	91	0.996	0.015	$1.5 \cdot 10^{-05}$
7	49	66	0.808	$4.0 \cdot 10^{-05}$	$4.8 \cdot 10^{-18}$
8	65	77	0.261	$2.8 \cdot 10^{-05}$	$1.3 \cdot 10^{-25}$

KS: Kolmogorov-Smirnov two sample test. Significant at 5% FDR[4]

Networks uncover hidden lexical borrowing in Indo-European language evolution

Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, Tal Dagan

SUPPORTING ONLINE MATERIAL

Figure S1. Cognate presence absence pattern (PAPs) in ToB dataset. Languages are sorted by their order on the reference phylogenetic tree. COGs are sorted by their size in ascending order. A presence case of certain COG in a certain language is colored in blue if the COG pattern is congruent with the tree branching patterns and red otherwise.

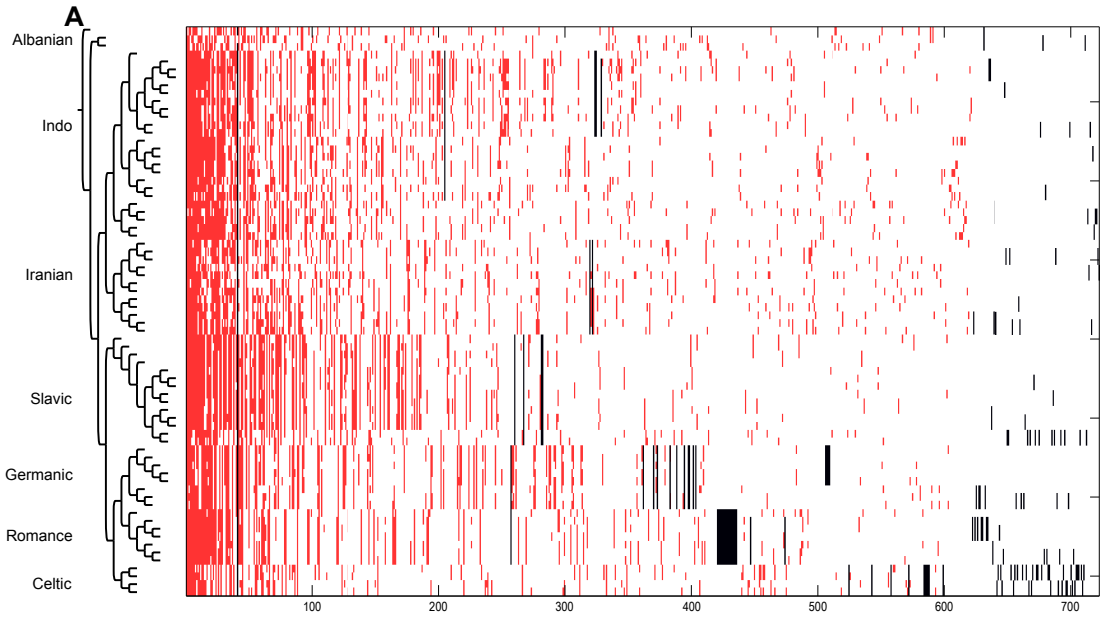


Figure S2: Modules in the shared COGs network (ToB).
(A) A matrix representation of the shared COGs network in Indo-European languages. Cells in the matrix are edges in the network. Edges are color-coded by the frequency of shared cognate according to the colorbar at the bottom. The languages in the matrix are sorted by order of appearance in the phylogenetic tree on the left. (B) Modules within the shared COGs network. Languages included in the same module are colored in the same color.

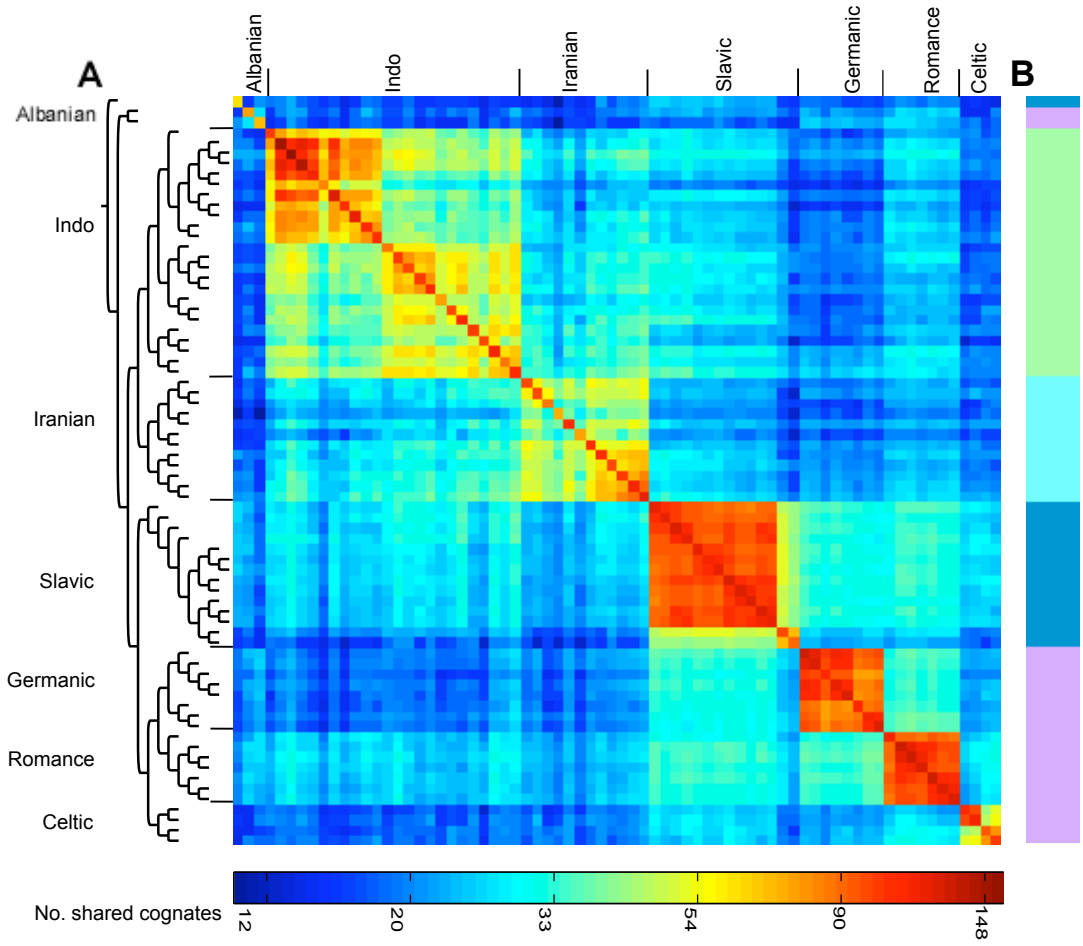


Figure S3: Acceptance rate of the different borrowing models using the different reference trees.
Analysis of the Dyen dataset using (a) Bayesian and (b) Neighbor-Joining reference tree (outliers in the *Loss Only* model are excluded).

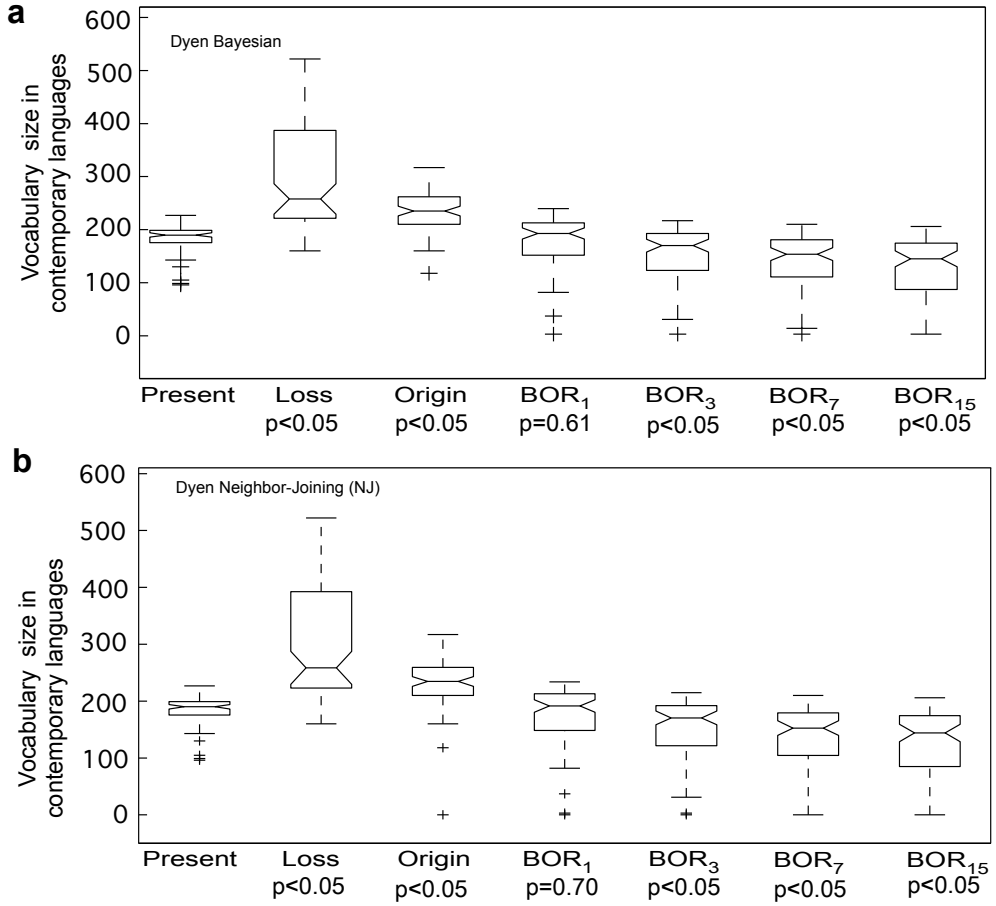


Figure S4: The affect of a randomized reference tree using Dyen dataset (A) A matrix representation of the shared COGs network in Indo-European languages. Cells in the matrix are edges in the network. Edges are color-coded by the frequency of shared cognate according to the colorbar at the bottom. The languages in the matrix randomized. (B) Vocabulary size under the different models using the randomized reference tree.

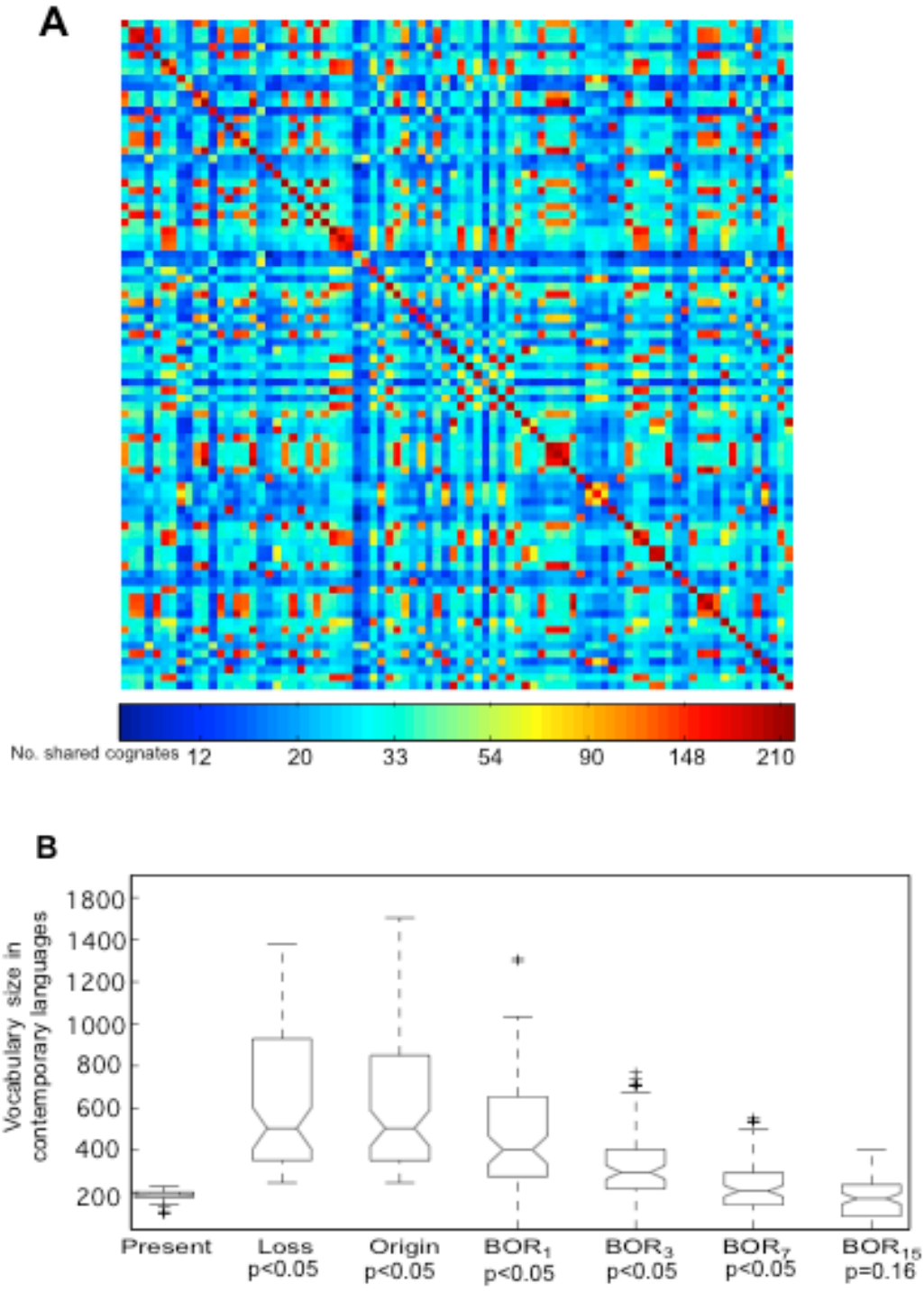


Figure S5: Acceptance rate of the different borrowing models using the Tower of Babel dataset the different reference trees.
Analysis of ToB dataset using (a) Bayesian and (b) Neighbor-Joining reference tree (outliers in the *Loss Only* model are excluded).

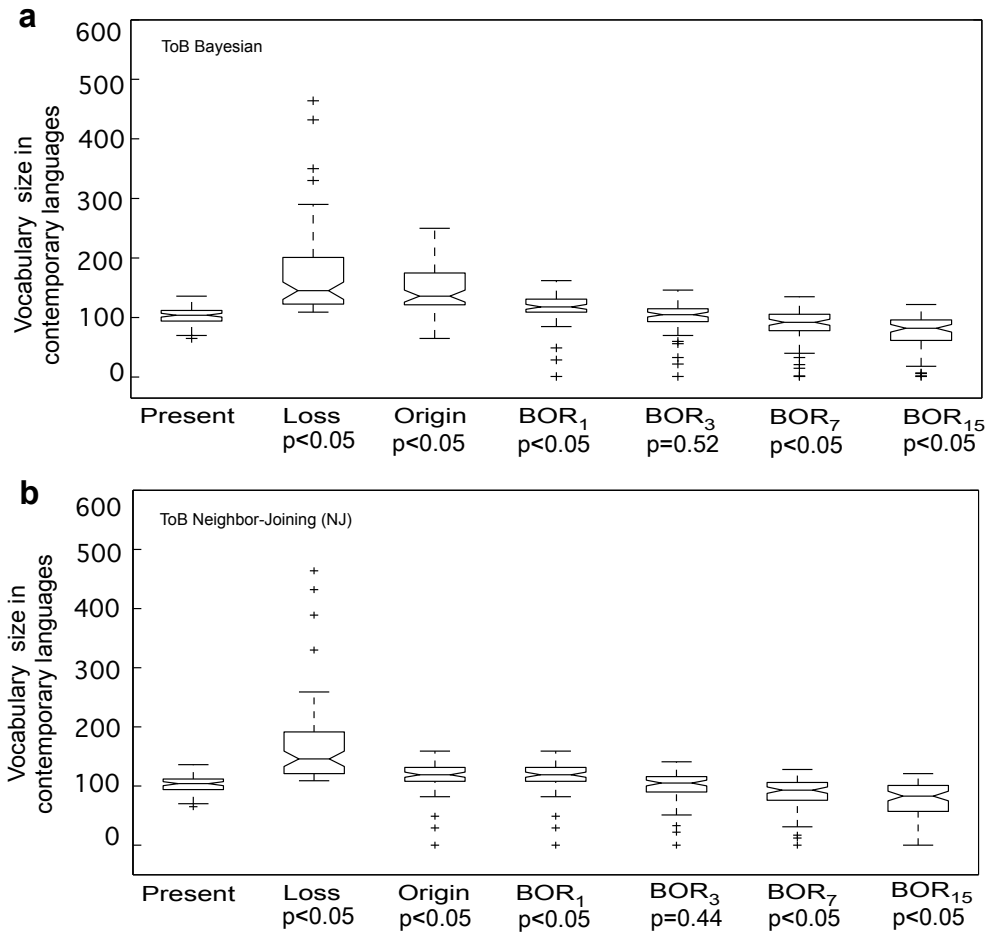


Figure S6: The minimal lateral network (MLN) of Indo-European languages using ToB. (A) An MLN for 73 contemporary languages reconstructed under the *BOR3* model. Vertical edges are indicated in gray, with both the width and the shading of the edge shown proportional to the number of inferred vertically inherited COGs along the edge (see the scale). The lateral network is indicated by edges that do not map onto the vertical component, with number of cognates per edge indicated in color (see the scale). Lateral edges that link ancestral nodes represent laterally shared COGs among the descendant languages of the connected nodes, whose distribution pattern could not be explained by origin and loss only under the ancestral vocabulary size constraint (B) Distribution of connectivity, the number of one-edge-distanced neighbors for each vertex, in the network. (C) Frequency distribution of edge weight in the lateral component of the network.

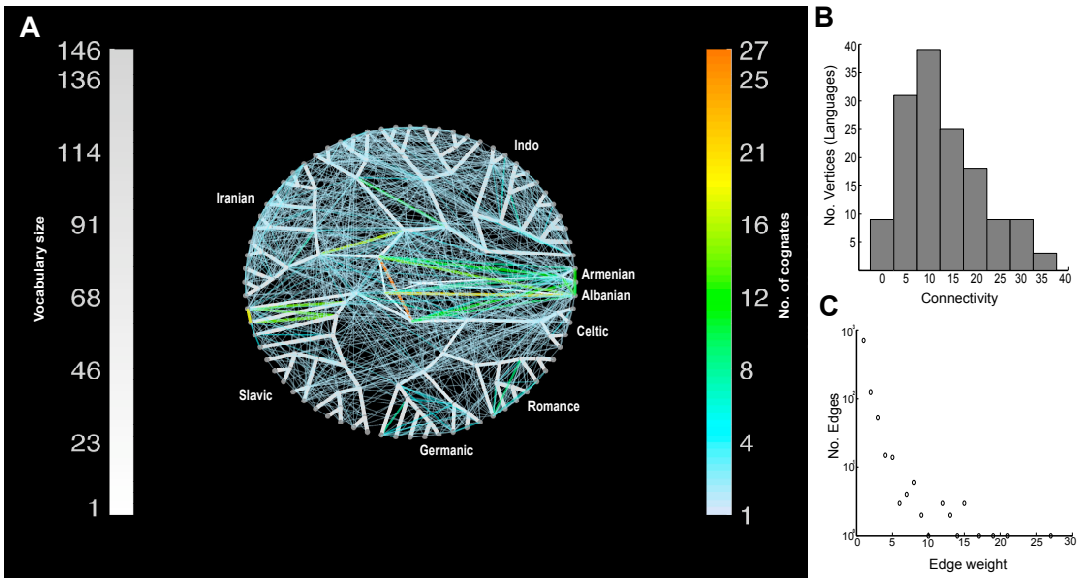


Table S1: Detected borrowings in the Dyen dataset.

Item	**Donor Language**	**Source Lang.**	**Recipient Languages**				
			Rom.	**Italian**	**Provencal**	**French**	**Spanish**
KILL	French	*tuer*			tua		
ROAD	Greek	*drómos*	drum				
SKIN	Latin	*cutis*					cutis
WALK	Old Franc.	*marka*			marcha	marcher	
WOMAN	Greek	*familia*	femeie				

Table S2: Borrowing and cog-word loss statistics under different borrowing allowances

Dyen	Borrowing allowances					
	Loss Only	SO	BOR_1	BOR_3	BOR_7	BOR_{15}
Average per cognate borrowing rate	0	0	0.6	0.9	1.2	1.4
Percent of cognates accepting LGTmax	0	0	58	4	0	0
Average losses per cognate	8.5	2.7	1.3	0.7	0.4	0.2
Percent of cognates with no losses	0.2	41	67	80	87	91
Average origin/loss ratio	-	1/3	1/1	2/1	5/1	9/1

ToB	Borrowing allowances					
	Loss Only	SO	BOR_1	BOR_3	BOR_7	BOR_{15}
Average per cognate borrowing rate	0	0	0.7	1.4	2.0	2.5
Percent of cognates accepting LGTmax	0	0	78	10	0.8	0
Average losses per cognate	9.6	5.5	3.2	2.0	1.2	0.7
Percent of cognates with no losses	0.1	21	46	60	70	78
Average origin/loss ratio	-	1/5	1/2	1/1	2/1	5/1

Table S3. COGs connected with a lateral edge in external nodes (contemporary languages).

		COGs connected with a lateral edge	
Language	**Total COGs**	**No.**	**Proportion**
Greek_ML	192	9	5
Greek_MD	203	10	5
Greek_Mod	194	5	3
Greek_D	198	23	12
Greek_K	168	39	23
Armenian_Mod	150	11	7
Armenian_List	143	7	5
Irish_A	175	11	6
Irish_B	172	7	4
Welsh_N	195	7	4
Welsh_C	190	4	2
Breton_List	193	8	4
Breton_SE	191	3	2
Breton_ST	198	4	2
Romanian_List	178	29	16
Vlach	147	13	9
Italian	206	5	2
Ladin	193	9	5
Provencal	208	7	3
French	194	1	1
Walloon	188	4	2
French_Creole_C	203	5	2
French_Creole_D	189	0	0
Spanish	198	9	5
Portuguese_ST	210	6	3
Brazilian	199	2	1
Catalan	190	35	18
Sardinian_N	178	12	7
Sardinian_L	190	6	3
Sardinian_C	183	10	5
German_ST	202	5	2
Penn_Dutch	178	4	2
Dutch_List	210	3	1
Afrikaans	213	6	3
Flemish	210	8	4
Frisian	180	10	6
Swedish_Up	218	7	3
Swedish_VL	208	3	1
Swedish_List	218	9	4
Danish	202	4	2
Riksmal	197	6	3
Icelandic_ST	200	5	3
Faroese	211	13	6

English_ST	181	23	13
Takitaki	158	22	14
Lithuanian_O	189	8	4
Lithuanian_ST	197	9	5
Latvian	155	25	16
Slovenian	178	48	27
Lusatian_L	191	2	1
Lusatian_U	192	2	1
Czech	210	6	3
Slovak	217	9	4
Czech_E	193	9	5
Polish	194	13	7
Ukrainian	209	17	8
Byelorussian	185	6	3
Russian	191	8	4
Macedonian	194	8	4
Bulgarian	170	12	7
Serbocroatian	187	9	5
Albanian_T	190	9	5
Albanian_Top	187	9	5
Albanian_G	175	12	7
Albanian_K	173	28	16
Albanian_C	166	48	29
Gypsy_Gk	105	40	38
Kashmiri	154	38	25
Marathi	162	11	7
Gujarati	174	16	9
Panjabi_ST	180	8	4
Lahnda	178	10	6
Hindi	201	18	9
Bengali	174	30	17
Nepali_List	227	29	13
Khaskura	190	17	9
Singhalese	99	41	41
Ossetic	96	43	45
Afghan	185	25	14
Waziri	174	15	9
Persian_List	176	12	7
Tadzik	190	20	11
Baluchi	143	27	19
Wakhi	129	36	28

Table S4: Reconstruction differences due to English position on the tree.

(A) The following COGs were inferred as borrowing using the basal-English tree and not inferred as borrowing using the internal-English tree (Etymologies based on Orel 2003).

Item	COG	Word Form	Etymology
HAIR	HAIR_3	hair	PGM *xēran 'hair'
GOOD	GOOD_2	good	PGM *gōđaz 'good'
EARTH	EARTH_SOIL_3	earth	PGM *erþō 'earth'
HERE	HERE_2	here	PGM *xēr 'here'
ROTTEN	ROTTEN_LOG_11	rotten	ON rotinn 'rotten'
SWELL	TO_SWELL_1	swell	PGM *swellanan 'swell'
WET	WET_4	wet	PGM *wētaz 'wet'
WIDE	WIDE_6	wide	PGM *wīđaz 'wide'

(B) The revised tree with English in internal position results in: four miscoded COGs, seven correctly detected borrowings (coded as cognates in Dyen), and four possible cases of parallel evolution. Here follows a detailed descriptions regarding the accuracy of the predictions using the internal-English tree:

Category	Cognate and Word Form	Description
Miscoding	BAD_11 English 'bad'	Miscoded in Dyen. This is no borrowing between English and Indian or Iranian languages as suggested by the method, but simply a resemblance in form, which is usually considered as coincidence by scholars of Indo-European and etymologists.
	DULL_KNIFE_5 English 'dull'	Apparently a miscoding in Dyen 1997: The form has an obscure etymology and no conclusions can be drawn. It is usually not connected to Breton 'dall'. So we have a coincidental resemblance and a miscoding by Dyen here, no borrowing event.
	WOODS_15 English 'woods'	Miscoded in Dyen as cognate with German 'Wald' and Flemish 'woud', so this is no borrowing, but a coding error, since the words are not etymologically related at all!
	DUST_14 English 'dust'	Etymology is unclear. Probably a miscoding in Dyen. Anyway, this is probably not a borrowing event.

Borrowing	BARK_7 English 'bark'	Borrowing from Old Norse (Scandinavian Languages): English 'bark' from Old Norse borkr 'bark'
	TO_COUNT_11 English 'count'	Borrowing from Old French 'conter' 'add up'. Correctly identified as borrowing from Romance or French.
	FRUIT_2 English 'fruit'	Borrowing from Old French 'fruit', correctly coded as borrowing.
	TO_PUSH_1 English 'to push'	Borrowing from Old French 'poulser', correctly coded as borrowing.
	SKIN_OF_PERSON_9 English 'skin'	Borrowing from Old Norse 'skinn', correctly coded as borrowing.
	SKY_11 English 'sky'	Borrowing from Old Norse 'sky' 'cloud'. Correctly coded as borrowing.
	WING_1 English 'wing'	Borrowing from Old Norse 'vœngr' 'wing of a bird'. Correctly coded as a borrowing.
Parallel evolution	LEAF_9 English 'leaf'	Goes back to Proto-Germanic *laubaz, so this is no borrowing, yet it may point to parallel evolution in Scandinavian and English (in German the corresponding word is "Laub" which is not given in the database).
	SMALL_10 English 'small'	From Proto-Germanic *smalaz, apparently no loan, but parallel evolution in the languages showing this COG.
	TO_THROW_18 English 'throw'	Apparently no borrowing but parallel evolution in Frisian and English (German corresponding word is 'drehen', which has a different meaning and is therefore not reflected in the dataset).
	WITH_ACCOMPANYING_11 English 'with'	From Proto-Germanic 'withro'. Probably a case of parallel evolution in Nordic languages and English, or a case of borrowing which is very problematic to prove, since the German corresponding word is

		'wider', i.e. 'against', so we have a semantic shift from 'against' to 'with' here.

Polynesian language networks reveal complex history of contacts during the Pacific settlement

Nelson-Sathi S, List JM, Greenhill S, Geisler H, Cohen O, Pupko T, Landan G, Martin WF, Dagan T, Gray RD

ONLINE SUPPORTING INFORMATION

Figure S1: Cognate presence absence pattern (PAPs) which are recurring highlighted by red others by black A) Basic subset B) Lexicon subset

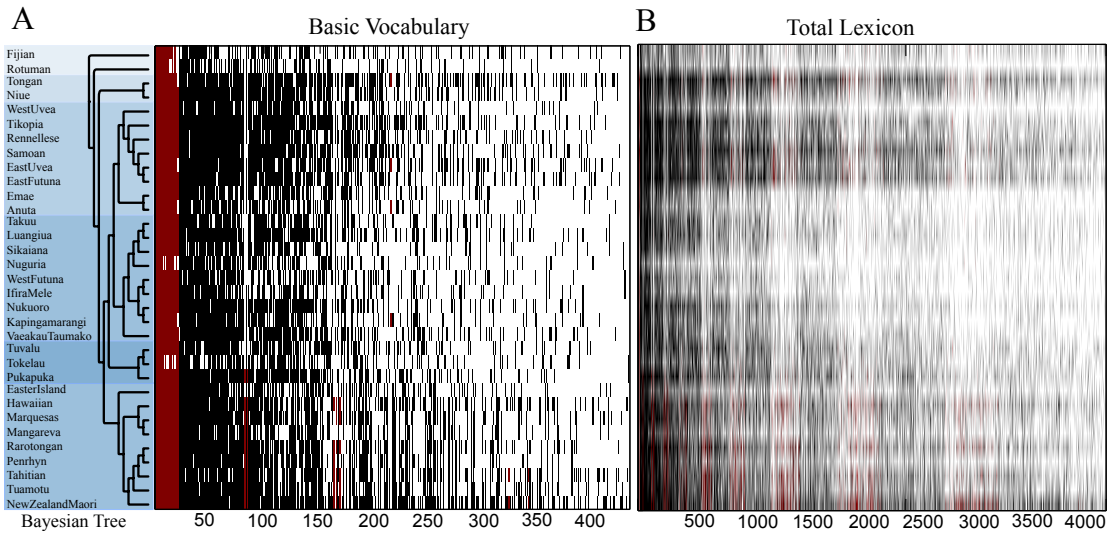


Figure S2: A patchily distributed cognate set **tafuqa* (platform, foundation, base) . This distribution can be explained by a single gain during the development of Proto-Eastern Polynesian and two subsequent borrowings into Pukapuka and Niue. Green filled circles represent origins and edge represent possible borrowings.

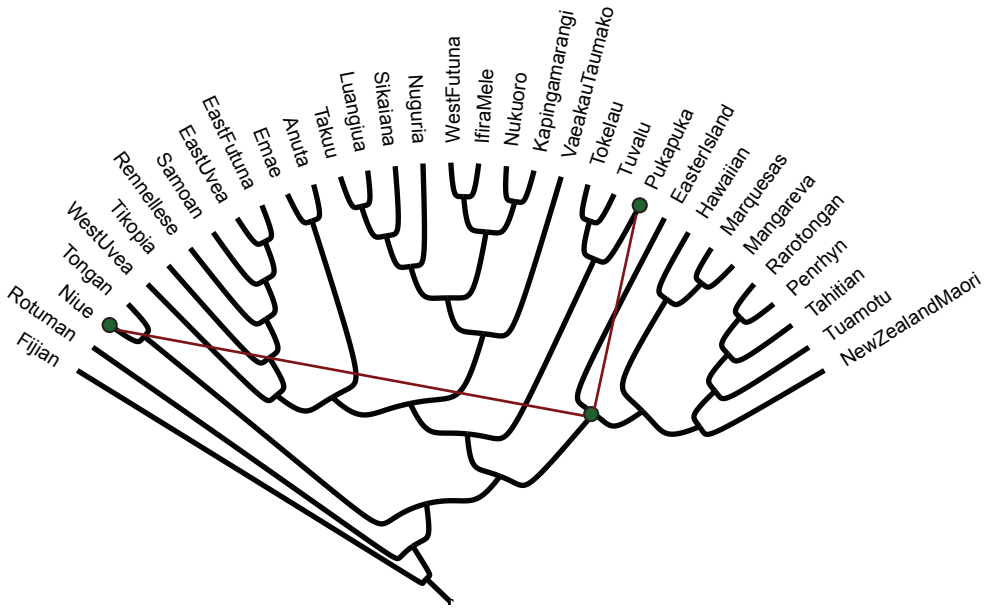


Figure S3: Comparison different reference ree topologies A) Bayesian and Bayesian-modified B) Bayesian and Maximum Parsimony (MP) C) Bayesian and Neighbor Joining (NJ) tree.

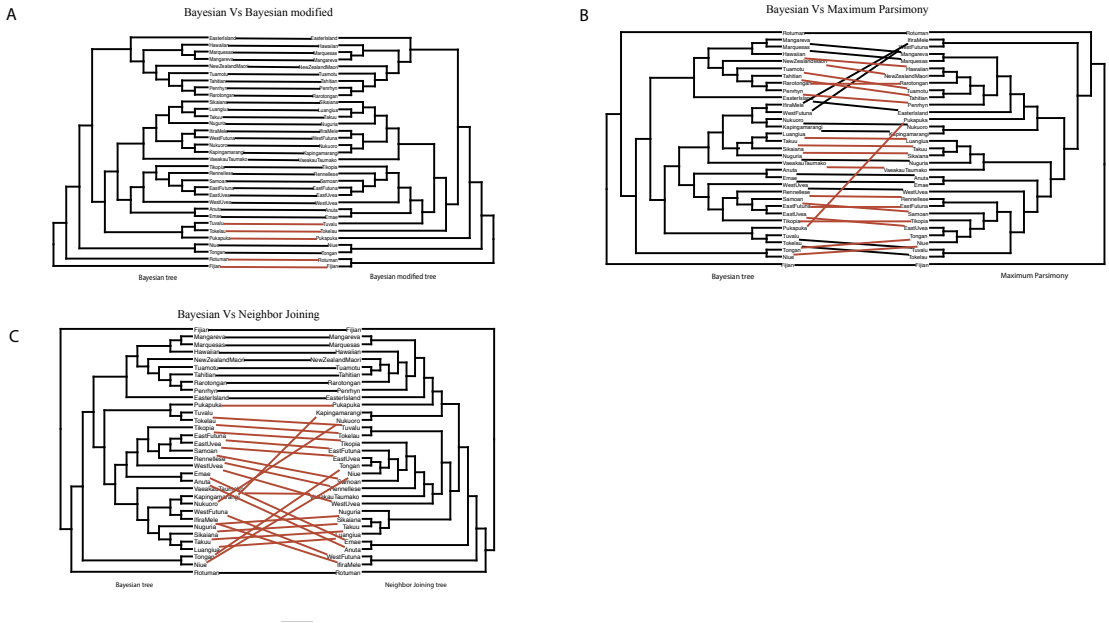


Table S1: Highly borrowing resistant Polynesian basic vocabulary cognates according to Leipzig-Jakarta rank

Rank	ID	Protoform	Label	Meaning
1	19	AN	AFI	!fire
6	97	PN	QALELO	!tongue
14	239	OC	AU	!I
27	541	AN	FATU.2A	!stone
x	680	AN	FITU	!Seven
54	735	AN	FOQOU	!new
61	1232	PN	KATA.1	!to laugh
x	1274	TA	KAUII.*	!left
9	1399	PN	KO-TOU	!you
x	1471	CK	KOFU.1C	!to cook
35	1724	OC	IA.1	!he/she
x	1807	MP	LAGI.2	!sky
x	2003	AN	LIMA.A	!Five
x	2093	AN	RUA	!Two
83	2439	AN	MATA.1A	!eye
48	2453	OC	MATAGI	!wind
21	3240	MP	POO.1	!night
x	3559	CO	SAAVARI.*	!to spit
33	4031	NP	TASI.1	!One
x	4148	MQ	TEKO.2	!white
x	4304	CE	TOKE.1B	!worm
x	4338	AN	TOLU	!Three
36	4429	AN	TUKI	!to pound
x	4708	XO	WELEWELE.*	!spider

Rank-Leipzig-Jakar Stability Rank

Table S2: List of Polynesian lexicon cognates that can be explained by strict vertical inheritance

ID	Label	Protoform	Description
62	QAKI.2	TO	Preposition marking instrument
130	ANAUXO		First person singular personal pronoun
238	AWATEA.*	CE	Late morning to early afternoon, midday
255	EE-IA TA		These (near speaker)
303	FAAGAI.B	CE	Adopt, foster
400	FAKA-TASA	PN	Together
404	FAKA-TEMU	SF	Touch lightly
573	FIHA	AN	How many?
595	FINE.1	TO	Armpit

623	FOA.C	TA		Headache
665	FOTU.3B	EC		Genital orifice
706	HUERO	CE		Seed
737	FUNA.2	CE		One of Buck's middle period nights of the moon (Grn)
750	FUTAA	KN		
766	GAHAHA	TO		Rattle, rustle
854	GOIO.1A	CE		Common noddy (Anous stolidus) (Clk): *go(o)io
905	HOO	TO		Pant
965	KAI-LUU	TO		Eat scraps
986	KAAKAAIA.*	CK		White Tern (Gygis sp.)
1093	KAAPITI.2	CO		Mat woven from coconut fronds
1128	KATIRU.*	MQ		(Cucumis melo)
1300	KO.2	SO		Progressive aspect marker
1381	KOO-ROA	CK		Index finger
1404	KOO-MATA	MQ		?
1428	KOO-NUI	CK		Thumb
1457	KOO-PURE	TA		Spotted
1470	KOTAKE	MQ		A bird
1506	KUAGO	XW		A fish
1507	KUEI	XW		??
1535	KULU-KULU.2		PN	A kind of yam
1586	INA-INA	TA		Singe a pig; expose to radiant heat
1588	I-NAAKUANEI		TA	Just now, earlier today
1589	I-NANAFI	CE		Yesterday
1648	RAQA-KAU	PN		Wood, tree
1664	LALO-KOKA	XW		Placename
1757	REMU.2	TA		Moss
1802	RIRO	EP		Be taken, become something else
1841	ROGO-MA-TAANE		CK	
1856	RORE	TA		Stilts
2133	MANO.B	MQ		Four thousand
2138	MANU	MP		Living creature (excluding humans, fish)
2173	MASAGA.3	EC		Turtle sp
2282	MAA-UTOLU	PN		First person exclusive plural pronoun (independent): *(ki)maa-utolu
2309	MESO	TA		A bird
2325	MIO	MQ		Extinguished
2368	MOKO-ROA-I-ATA		CK	Milky Way
2520	NII-KAU	CE		Coconut frond
2557	NUKA.1	TO		
2576	OFI.B	TU		To fit in, to pass through a narrow place
2586	OIRE	LO		Village
2622	QOTI.A	AN		Completed, finished
2644	PAA-FATA	TA		Shelf or platform
2661	PAKA.1B	EP		Scab
2694	PAKOKO.B	MQ		Male flower of the breadfruit
2715	PAA-RAKU	TA		Rake
2802	PATA.4	TA		Fillip, flick with fingers
2819	PAA-TIKI.*	CE		A fish, Flounder, Flatfish
2820	PAA-TITI	TA		Drive in (as a nail, stake)
2831	PATU.3	TA		Build in stone

2931	POFORE.*	MQ	Skinned
3000	POU	OC	Post n
3017	PUA-LIKI	TO	A tree
3108	SA	PN	Non-specific article
3140	SAAKERE	CK	: *(f, s)a(q)kere
3145	SAKI-NA.	TO	Catch by the foot
3189	SAPE.B	EP	Crooked, wrong
3229	SEA-GA	MQ	Victim, corpse
3300	SOA	CP	Friend
3342	SOPE.2	EP	Buttocks, posteriors, rear end
3402	TAEAKE	TA	Relative of same generation as ego
3435	TAFU FJ	To	light a fire, burn v.t
3455	TAGATOO	TO	Loose
3566	TALO.1	AN	Taro (Colocasia esculenta)
3572	TAA-MAA	TA	Clean (vt)
3596	TANU	OC	Bury
3624	TAA-PEE	CE	Cause to ripen or decay
3661	TAA-TOU	NP	First person inclusive plural pronoun (independent): *(ki)taa-tou
3689	TAULOKO	SO	A plant (Solanum sp.)
3712	TAA-UTOLU	PN	First person inclusive plural pronoun (independent): *(ki)taa-utolu
3714	TAU-TURU	TA	Support, help
3916	TOOREA	CE	Bird sp. (Pluvialis dominica)
3925	TOLOA.C	CO	A shorebird (Numenius sp.)
4052	TUNU	AN	Cook on open fire; roast, grill
4055	TUUPAAPAKU	EP	Corpse, cadaver
4059	TUPU.A	MP	Grow: *t(u,i)pu
4116	UKI	TA	Generation, age, epoch
4120	QURA	AN	Crayfish
4178	UTE.1	MQ	Paper Mulberry
4208	WAI-RAGA	TA	Place where something is kept
4239	WARUPN	Scrape v	
4256	WEKA.3	XW	
4292	VILU	XW	Former times

Table S3. Average borrowings per languages in basic and total lexicon subset

a) Basic subset

Language	SO ori	OTU ori	HTU ori	Vocab size	HTU bor %	OTU bor %	Total bor %
Anuta	92	8	54	154	35.1	5.2	40.3
EasterIsland	93	20	61	174	35.1	11.5	46.6
EastFutuna	92	9	111	212	52.4	4.2	56.6
EastUvea	92	12	118	222	53.2	5.4	58.6
Emae	92	16	63	171	36.8	9.4	46.2
Fijian	54	83	0	137	0	60.6	60.6
Hawaiian	96	38	82	216	38	17.6	55.6
IfiraMele	93	8	52	153	34	5.2	39.2

Kapingam-arangi	92	8	67	167	40.1	4.8	44.9
Luangiua	93	16	91	200	45.5	8	53.5
Mangareva	96	9	76	181	42	5	47
Marquesas	96	18	94	208	45.2	8.7	53.9
NewZeala-ndMaori	95	54	141	290	48.6	18.6	67.2
Niue	93	27	96	216	44.4	12.5	56.9
Nuguria	92	10	35	137	25.5	7.3	32.8
Nukuoro	92	12	94	198	47.5	6.1	53.6
Penrhyn	96	3	106	205	51.7	1.5	53.2
Pukapuka	92	39	70	201	34.8	19.4	54.2
Rarotongan	96	9	117	222	52.7	4.1	56.8
Rennellese	92	18	99	209	47.4	8.6	56
Rotuman	71	41	0	112	0	36.6	36.6
Samoan	92	15	117	224	52.2	6.7	58.9
Sikaiana	93	4	52	149	34.9	2.7	37.6
Tahitian	96	21	112	229	48.9	9.2	58.1
Takuu	93	5	69	167	41.3	3	44.3
Tikopia	92	18	123	233	52.8	7.7	60.5
Tokelau	92	16	42	150	28	10.7	38.7
Tongan	93	58	96	247	38.9	23.5	62.4
Tuamotu	96	17	114	227	50.2	7.5	57.7
Tuvalu	92	27	66	185	35.7	14.6	50.3
VaeakauTa-umako	92	19	91	202	45	9.4	54.4
WestFutuna	93	9	52	154	33.8	5.8	39.6
WestUvea	92	12	76	180	42.2	6.7	48.9
Total	%				39.8	11.1	**50.9**

b) Lexicon subset

Language	SO ori	OTU ori	HTU ori	Vocab size	HTU bor %	OTU bor %	Total bor %
Anuta	280	68	250	598	41.8	11.4	53.2
EasterIsland	289	321	314	924	34	34.7	68.7
EastFutuna	287	199	1392	1878	74.1	10.6	84.7
EastUvea	287	102	1211	1600	75.7	6.4	82.1
Emae	280	90	388	758	51.2	11.9	63.1
Fijian	161	752	0	913	0	82.4	82.4
Hawaiian	307	707	654	1668	39.2	42.4	81.6
IfiraMele	280	41	254	575	44.2	7.1	51.3
Kapingam-arangi	281	60	464	805	57.6	7.5	65.1
Luangiua	285	102	568	955	59.5	10.7	70.2
Mangareva	312	195	571	1078	53	18.1	71.1
Marquesas	312	363	754	1429	52.8	25.4	78.2

NewZeala-ndMaori	309	1008	1114	2431	45.8	41.5	87.3
Niue	290	312	817	1419	57.6	22	79.6
Nuguria	282	44	60	386	15.5	11.4	26.9
Nukuoro	281	119	646	1046	61.8	11.4	73.2
Penrhyn	324	94	769	1187	64.8	7.9	72.7
Pukapuka	280	601	615	1496	41.1	40.2	81.3
Rarotongan	324	322	1117	1763	63.4	18.3	81.7
Rennellese	281	229	971	1481	65.6	15.5	81.1
Rotuman	207	353	0	560	0	63	63
Samoan	282	371	1402	2055	68.2	18.1	86.3
Sikaiana	284	66	371	721	51.5	9.2	60.7
Tahitian	323	372	1074	1769	60.7	21	81.7
Takuu	285	68	524	877	59.7	7.8	67.5
Tikopia	280	225	1007	1512	66.6	14.9	81.5
Tokelau	281	287	581	1149	50.6	25	75.6
Tongan	290	1019	817	2126	38.4	47.9	86.3
Tuamotu	320	346	1050	1716	61.2	20.2	81.4
Tuvalu	281	314	692	1287	53.8	24.4	78.2
VaeakauTa-umako	280	139	557	976	57.1	14.2	71.3
WestFutuna	280	70	325	675	48.1	10.4	58.5
WestUvea	280	56	374	710	52.7	7.9	60.6
Total %					50.5	21.8	**72.3**

SO ori – no. of Single Origins, OTU ori – no. of OTU origins, HTU ori – no. of HTU origins, Vocab size – Vocabulary Size, HTU bor % - Percentage of HTU
borrowings cognates , OTU % - Percentage of OTU borrowings cognates, Total bor % - percentage of total borrowing cognates.

7.3 Conferences and Workshops

Talks

Nelson-Sathi S, Martin W, Dagan T: A network approach to study vertical inheritance and lateral transfer during the evolution of Indo-European languages. "Evolution and Classification in Biology, Linguistics and the History of Science. An Interdisciplinary Workshop", Schloss Mickeln, Düsseldorf 2009/06/11-12.

Nelson-Sathi S: Networks uncover hidden lexical borrowing in Indo-European language evolution, *Bridging Disciplines — Evolution and Classification in Biology, Linguistics and the History of Sciences*. Schloss Reisenburg, Ulm University, Germany. 2011/06/24-26.

Poster Presentations

Nelson-Sathi S, J Mattis List, Hans Geisler, Heiner Fangerau, Russell D Gray, William Martin, Tal Dagan. Networks uncover hidden lexical borrowing in Indo-European language evolution. SMBE 2010 - Annual Meeting of the Society for Molecular Biology and Evolution, Lyon, France - July 4-8, 2010.

Nelson-Sathi S, Dagan T, Martin W. Phylogenomic networks of archaebacteria reveal frequent lateral gene transfer and eubacterial acquisitions, SMBE 2011 – Annual Meeting of the Society for Molecular Biology and Evolution, Kyoto University, Japan –July 26-30, 2011.

Nelson-Sathi S, Greenhill Simon, Gray Russell, Dagan Tal, Martin W. Polynesian borrowings networks, SMBE 2012 – Annual Meeting of the Society for Molecular Biology and Evolution, Dublin Convention Centre, Dublin, Ireland - June 23-26, 2012.

Workshops

The Future of Phylogenetic Networks, Lorentz Center, Oort, Netherlands - 15 Oct 2012 through 19 Oct 2012.

Presenting Science 1, Interdisciplinary Graduate and Research Academy Düsseldorf (iGRAD) Workshop, Heinrich Heine University, 6-7 Dec 2012.

Preparing for Conflicts, Interdisciplinary Graduate and Research Academy Düsseldorf (iGRAD) Workshop, Heinrich Heine University, 28-29 Jan 2013.

Fundamentals of Project Management for Doctoral Researchers, Interdisciplinary Graduate and Research Academy Düsseldorf (iGRAD) Workshop, Heinrich Heine University, 4-5 Mar 2013.

7.4 Acknowledgements

This PhD thesis is a result of help and support from many people. First and foremost I want to thank my supervisor Prof. Dr. William F. Martin (Bill). It has been an honor to be his PhD student. I appreciate all his contributions of time, ideas and funding to make my PhD experience productive and stimulating. He was really encouraging and I enjoyed working with him at Institute of Molecular Evolution.

I want to thank my guide Prof. Dr. Tal Dagan for her support throughout my PhD. They joy and enthusiasm she has for her research was contagious and motivational for me. She is my primary resource for getting science questions answered and was instrumental in helping me with each step of work.

I thank German Federal Ministry of Education and Research and European Research Council for supporting my research.

I want to express my gratitude to Dr. Giddy Landen, Dr. Filipa Sousa for their discussions, support and encouragement during my research.

I thankful to all my friends and colleagues, Dr. Sven Gould, Dr. Christian Esser, Dr. Vereena Zimorski, Thorsten Thiergart, Kathrin Hoffmann, Mayo Röttger, Gary Kusdian, Christian Wöhle, Nabor Lozada Chávez, Thorsten Kloesges, Ovidiu Popa, Sriram Garg, David Bogumil, Doris Matthée, Ariane Baab and all members of molevol team for their help, co-operation and for making the working environment really enjoyable.

I thank Prof. Dr. Hans Geisler, Prof. Dr. Heiner Fangerau, Dr. Mattis List, Frank Kressing and all members of EvoClass interdisciplinary programme for their support and help to study language evolution.

I thank Dr. Stephan Raub and members of HPC computing facility of Heinrich Heine University for computational resource support.

Many thanks to my parents Nelson Nadar and Sathi Rosamma, my brother Dr. Binulal Nelson, sister-in-law Sreeja Narayanan and my love Sreelekshmi Sreekumaran for their continuous support and inspiration.

Finally I thank all my former supervisors, teachers and friends who help me to pursue my degree.