Toward Automatic Reconstruction of a Highly Resolved Tree of Life

Francesca D. Ciccarelli,[1,2,3]* Tobias Doerks,[1]* Christian von Mering,[1] Christopher J. Creevey,[1] Berend Snel,[4] Peer Bork[1,5]†

We have developed an automatable procedure for reconstructing the tree of life with branch lengths comparable across all three domains. The tree has its basis in a concatenation of 31 orthologs occurring in 191 species with sequenced genomes. It revealed interdomain discrepancies in taxonomic classification. Systematic detection and subsequent exclusion of products of horizontal gene transfer increased phylogenetic resolution, allowing us to confirm accepted relationships and resolve disputed and preliminary classifications. For example, we place the phylum Acidobacteria as a sister group of δ-Proteobacteria, support a Gram-positive origin of Bacteria, and suggest a thermophilic last universal common ancestor.

Reconstructing the phylogenetic relationships among all living organisms is one of the fundamental challenges in biology. Numerous attempts to derive a tree of life using various methods have been published [for a review, see (1)], and its principal existence has been questioned recently (2, 3). Moreover, even under the assumption of a tree of life, numerous groupings and taxonomic entities still remain heavily debated, and the advent of molecular and genomic data has increased the variety of classifications rather than reducing the problem (1).

Theoretical and practical limits to reconstructing a tree of life have been put forward, such as the insufficient amount of discriminating characters available, even in information-rich genomic data sets (4), and the computing resources required to cope with large numbers of species (1). Furthermore, there are factors that hamper accurate reconstruction of phylogenetic trees regardless of the methods used, such as sampling biases of species included (5) and dilution of phylogenetic signal by horizontal gene transfer (HGT) (6), the

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany. [2]Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy. [3]Institute of Molecular Oncology Foundation, Fondazione Italiana per la Ricerca sul Cancro, Via Adamello 16, 20139 Milan, Italy. [4]Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Toernooiveld 1, 6525 ED, Nijmegen, Netherlands. [5]Max Delbrück Centrum, 13092 Berlin-Buch, Germany.

*These authors contributed equally to this work.
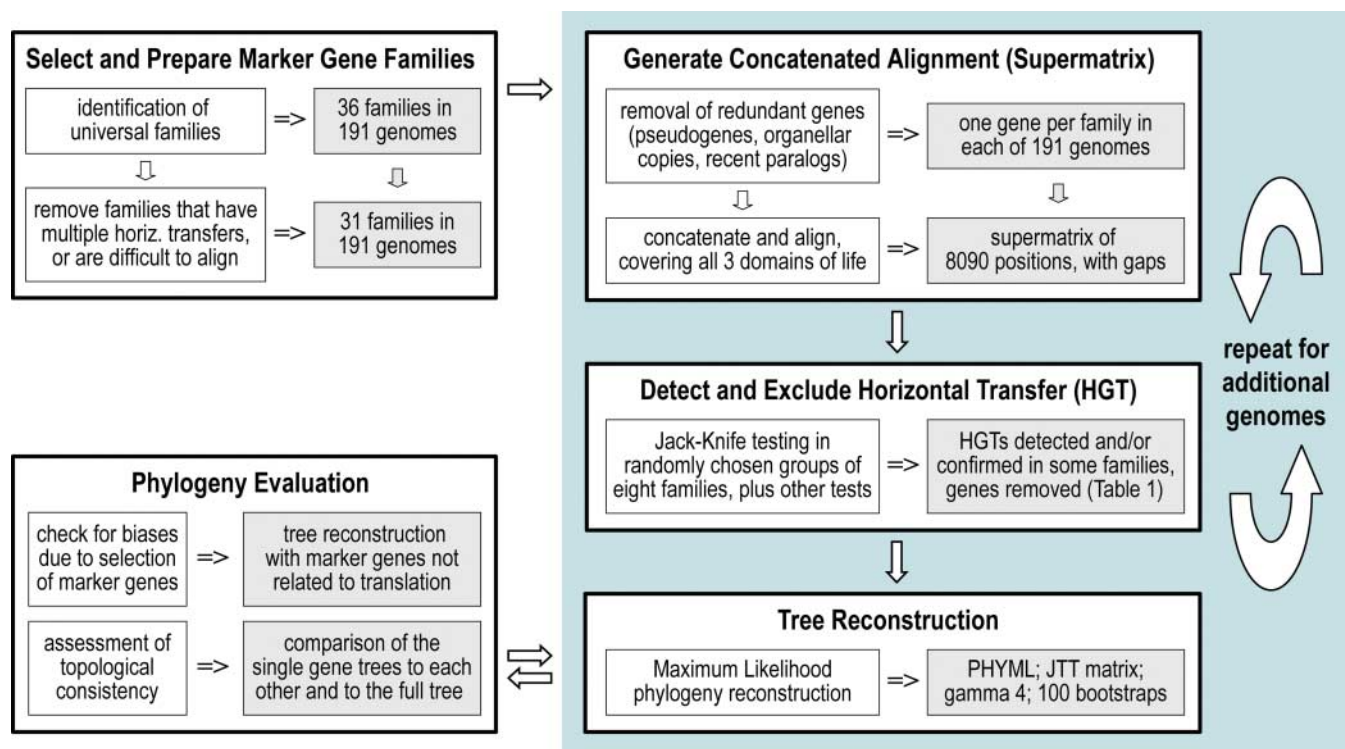†To whom correspondence should be addressed. E-mail: peer.bork@embl.de

Fig. 1. Overview of the procedure. The white boxes represent the major steps for building the pan-domain phylogeny presented here. Steps in gray represent automatable parts of the procedure that need to be carried out for including further species. For the 31 clusters of orthologous groups (COGs) used in the analysis, we manually derived 1:1 orthologs by removing mitochondrial and chloroplast paralogs from corresponding multiple alignments. We built domain-specific alignments by using corresponding proteins encoded by the 31 orthologs and aligned the resulting profiles. With this procedure, we maximized the number of positions of the global alignment and reduced the number of misaligned residues. For a detailed description of the methods, see (8).
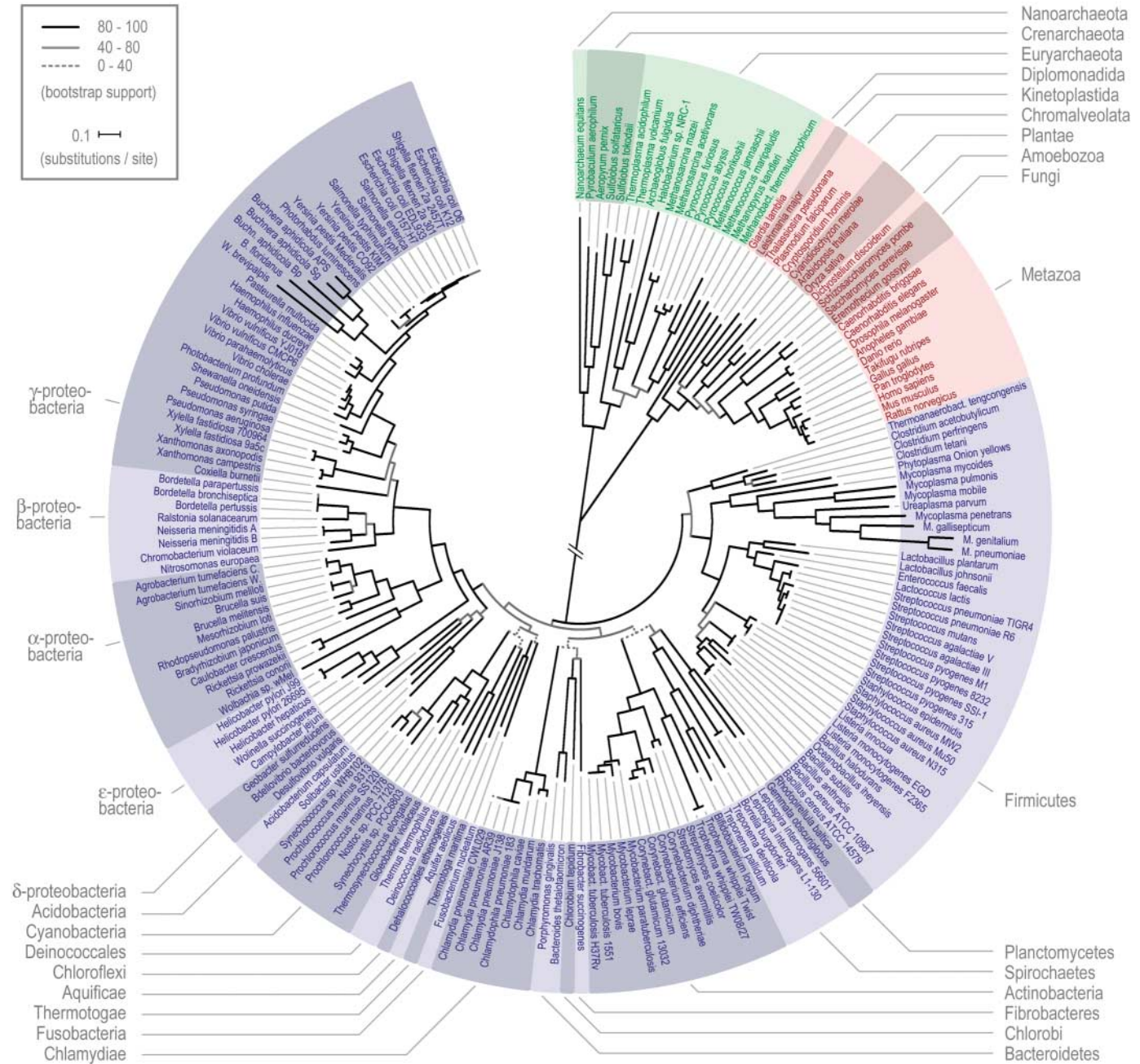
Fig. 2. Global phylogeny of fully sequenced organisms. The phylogenetic tree has its basis in a cleaned and concatenated alignment of 31 universal protein families and covers 191 species whose genomes have been fully sequenced (*14*). Green section, Archaea; red, Eukaryota; blue, Bacteria. Labels and color shadings indicate various frequently used subdivisions. The branch separating Eukaryota and Archaea from Bacteria in this unrooted tree has been shortened for display purposes.

extent of which is still extremely controversial (*2*, *3*, *7*). In addition to these difficulties, different data sets have been used with a variety of methods and parameter settings, making it almost impossible to quantitatively compare the proposed results. Hence, there exists the challenge and requirement for a reproducible and updatable pipeline to reconstruct the tree of life by means of a commonly available data set, such as completely sequenced genomes. Here, we demonstrate the feasibility of the tree construction and present a phylogeny based on an

alignment of sufficient length and resolution to accurately calculate comparable branch lengths across all three domains of life. We have created for this purpose a supermatrix of 31 concatenated, universally occurring genes with indisputable orthology in 191 species with completely annotated genomes (Fig. 1 and table S1). Although initial identification and analysis of these genes required considerable manual effort (*8*), the inclusion of additional species with completely annotated genomes has pipeline character (Fig. 1). Because the 31 universal genes are all

involved in translation, we applied the same tree-building procedure to independent sets of domain-specific nontranslational genes (*8*).

For the tree reconstruction, we mostly used standard approaches (Fig. 1) with the exception of a procedure for detection and selective exclusion of HGTs, which turned out to be essential for obtaining a highly resolved tree. We started with 36 genes universally present in all 191 species for which orthologs could be unambiguously identified (*8*) and eliminated five of them from the analysis (mostly tRNA synthetases) because they

have undergone multiple horizontal transfers and/or were difficult to align (Fig. 1 and table S1). Although the 31 remaining genes are unlikely to be subjected to lateral transfers because they mainly encode for ribosomal proteins (9), we systematically tested them for any HGT event not yet identified. We randomly allocated the 31 gene products into four groups, and for each group we derived the corresponding subsets of trees where each protein was in turn missing from the alignment (resampling with displacement: jackknife test). We subsequently checked for topological incongruence within each subset of trees and further tested candidate HGTs by two other independent measures (8). If at least one of these two measures could confirm the jackknife indication, the gene was considered horizontally transferred and removed from the corresponding alignment (Fig. 1 and table S2).

Our approach [confirmed by single tree analysis (8)] detected a total of 7 HGT candidates [i.e., orthologous gene displacements (10)] among 31 orthologs from 191 species, with some species being involved in more than one HGT event (table S2). Three out of the four aminoacyl-tRNA synthetases (aa-RSs) used in this analysis have undergone HGT, including Valyl-RS (COG0525), which had been reported before (11), thus confirming the mobility of these enzymes (12). Clostridia is the only class that acquired ribosomal proteins by lateral transfer, likely in a single ancient event, because the displaced orthologs are present in all sequenced Clostridia species (table S2). To our knowledge, only one other horizontal transfer of ribosomal proteins has been reported so far (13). The identification of 7 HGTs in the 31 translation-related genes compares with the 30 (10 per domain) lateral transfers detected in domain-specific trees from 24 nontranslational genes (8).

Species-specific exclusion of HGTs and concatenation of all gene product alignments resulted in a supermatrix of 8090 positions for 191 species. This supermatrix was subsequently used to reconstruct the tree of life shown in Fig. 2 (14).

The global tree topology was supported by two independent measurements: First, by using domain-specific subtrees from nontranslational genes we could confirm the monophyly of all major divisions and reproduce most of their branching orders (8), albeit with weaker statistical support. This is due to lower sequence coverage and/or conservation as well as a higher number of excluded characters because of the higher incidence of HGTs (8). Secondly, three independent tests carried out on individual gene trees revealed that, although they are not identical, they share similarities with both the obtained tree of life and with each other (8). Although it may be possible to reject the null hypothesis of each of these tests without much difficulty, their combined evidence suggests that the gene trees have a cohesive phylogenetic signal.

Within the tree of life, as many as 65% of the branches are supported by a bootstrap proportion

Table 1. Noteworthy selected features of the tree of life phylogeny that are novel, debated, or difficult to reproduce according to current literature. An extended version of the table is available as table S6. In the case of Firmicutes as the earliest branching bacterial phylum, it is noteworthy that the remaining 33% of the BP show at least a subclade of the Firmicutes at the earliest division.

Domain	Topological feature	BP (%)
	Coelomata hypothesis	100
Eukaryota	Amoebozoa related to Opistokonta	41
	Deep branching of Diplomonadida	100
	Relationships within phyla	
	Separation between β- and γ-Proteobacteria	100
	Disruption of Chroococcales monophyly	100
	Disruption of Actinomycetales monophyly	100
	Acidobacteria-Proteobacteria clade	98
	Cluster of *F. succinogenes* next to the Chlorobium-Bacterioidales (Sphingobacteria hypothesis)	62
	Cluster of *F. nucleatum* with hyperthermophylic Bacteria	36
Eubacteria	*Relationships between phyla*	
	Grouping of Chlamydiae, Spirochetes, Actinobacteria, and Bacteriodales-Chlorobi	67
	Grouping of Cyanobacteria, hyperthermophylic, and Deinococcales-Chloroflexi	51
	Relationships between super-phyla	
	Grouping of Proteobacteria with Cyanobacteria, hyperthermophylic, and Deinococcales-Chloroflexi	74
	Deep branching of Firmicutes	66
	Relationship within phyla	
	A. fulgidus with halobacterium and methanosarcina	99
Archaeabacteria	*Relationship between phyla*	
	Nanoarchaea as a sister branch of Crenarchaea	100

(BP) of 100%, and 81.7% have more than 80% BP support, enabling us to propose resolutions to debated classifications at both the root and the tips of the tree (Table 1). Although in Prokaryota statistical support for deeper branches is generally weaker than that for the recent ones, it is noteworthy that, within Bacteria, the Firmicutes appear to comprise the earliest branching phylum, in agreement with a proposed Gram-positive ancestor for all Bacteria (15) (Fig. 2 and Table 1). In our tree, Firmicutes are placed at the earliest division of Eubacteria with 66% BP support, and 33% of remaining BP show at least a subclade of Firmicutes at the earliest division. This placement and the fact that the 15 slowest evolving taxa of the Bacteria are all Gram positive (8) support the theory of a Gram-positive origin of Bacteria. Furthermore, the thermophilic Firmicute *Thermoanaerobacter tengcongensis* is the taxon with the shortest overall phylogenetic distance to the root of Bacteria (Fig. 2) and as such is most likely to have retained ancestral states (16). Together with the fact that slowest evolving, ancestral Archaea (table S7) are also (hyper)thermophilic (8), this lends support to the hypothesis that the last universal common ancestor was living at high temperatures.

At the base of the Proteobacteria, the monophyletic Acidobacteria appear as a sister group to the δ-Proteobacteria (Fig. 2). The 64% BP sup-

port for this relationship indicates that the Acidobacteria may be a sixth divergent class within Proteobacteria. The Proteobacterial-Acidobacterial monophyly is supported with a BP of 98%, further raising the question whether Acidobacteria should indeed be an independent phylum (17).

Toward the tips of the tree, within Cyanobacteria *Synechococcus* (sp.WH8102) groups with *Prochloroccales* and *Nostoc* groups with *Synechocystis*, a result that has been supported by some ribosomal RNA (rRNA) studies (18) and challenges the classical order Chroococcales (19).

Within Archaea, the position of Nanoarchaeota remains debated [e.g., (20)]. We find (with 100% BP support) that they are a sister group of Crenarchaeota, without an indication of reported HGTs from Crenarchaeota (20) in all core genes studied.

Within Eukaryota, our tree gives clear support for the classical Coelomata hypothesis that groups Arthropoda with Deuterostomia (chordates) in a monophyletic clade. This is in contrast to the "new animal phylogeny" that groups nematodes and arthropods into the monophyletic Ecdysozoa (21, 22). The ecdysozoan clade has been supported by small subunit rRNA and single-gene phylogenies [(23) and references therein] but has been rejected by a number of recent studies on the basis of genomics features and whole-genome phylogenies [(24) and references therein]. Current sampling biases and accelerated evolution of se-
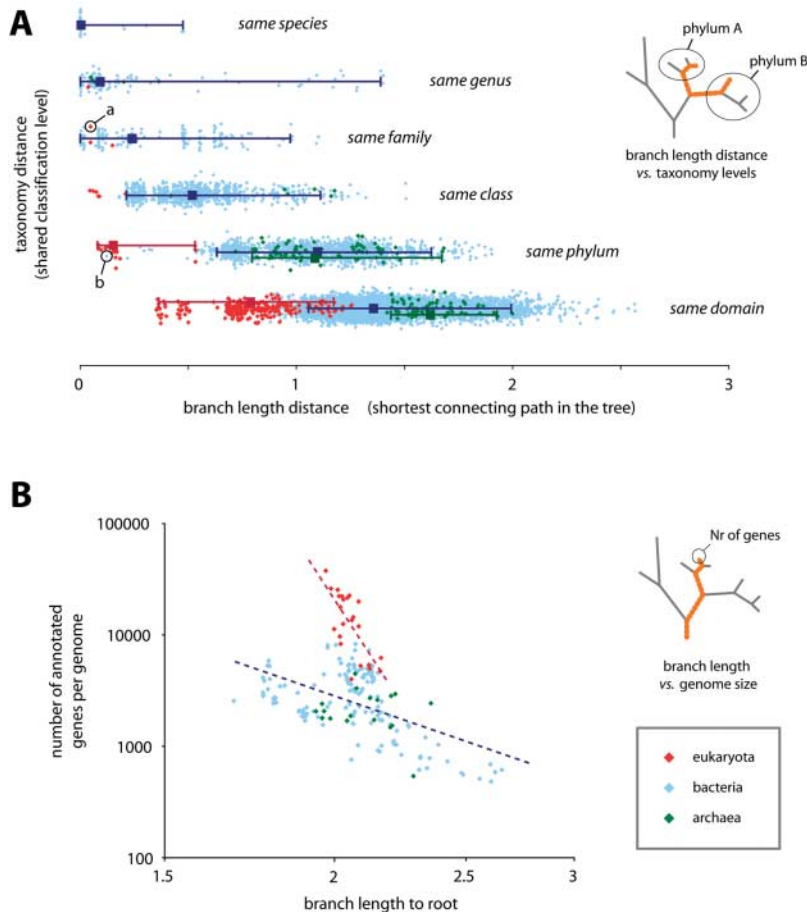
Fig. 3. Global analysis of branch length information. (**A**) Average sequence divergence within taxonomic classification units. Each data point denotes a pairwise comparison of two taxa, relating their intertaxa branch-length distance (i.e., sequence divergence) with their level of relatedness according to the National Center for Biotechnology Information taxonomy ("taxonomy distance"). Horizontal bars denote 95% intervals and medians of the data. Some minor taxonomy hierarchy levels have been omitted. Marked items: (Point a) *Homo sapiens* versus *Pan troglodytes*. The sequence divergence between human and chimp is low; they most likely would have been assigned the same genus if they had been prokaryotes [see also (*30*) for a proposed revision]. (Point b) *Synechococcus* (sp. WH8102) versus *Prochlorococcus marinus 9313*. The two species are annotated as distinct orders, but they appear quite closely related, challenging the classical order of Chroococcales. (**B**) Evolutionary speed and genome size. For each taxon, cumulative branch lengths from the tip to the root is plotted against genome size (measured here as number of genes). Dashed lines are linear regressions.

quenced representatives of certain metazoan lineages (e.g., arthropods and nematodes) (Fig. 2) may factor in these results. This highlights the need for the sequencing of slow-evolving species (*16*), which may resolve such controversies in the tree.

Despite a highly resolved and robust tree, we cannot exclude a few uncertainties in tree topology due to biased species sampling or long branch attraction (LBA) (*25*). For example, the grouping of *Thermotoga* and *Aquifex* in our and other trees might be partially caused by their common thermophilic life-styles (*26*), whereas LBA might account for the placement of diplomonadida (*Giardia lamblia*) as the most basal eukaryal taxon (Table 1).

The use of a common protein set across all three domains of life also ensures that the observed branch lengths are comparable across the entire tree. This enables, for example, an ob-

jective, quantitative analysis of the consistency of traditional taxonomic groupings (Fig. 3). As expected, the hierarchy of taxonomic groups correlates with phylogenetic diversity measured between and within them (e.g., species belonging to the same family have a shorter branch length distance than species belonging only to the same phylum). Within each taxonomic level, branch lengths distances vary considerably (*27*), apparently owing to factors that influence substitution rates, such as differences in lifestyle or population size. However, even when taking this effect into account, we observe a strong discrepancy between taxonomic divisions within Eukaryota and Prokaryota (Fig. 3A). Organisms that have been assigned to separate phyla in Eukaryota would clearly belong to the same phylum in the prokaryotic classification. Historically, eukaryotes have obviously been

given more taxonomic resolution than prokaryotes, a testament to their greater morphological diversity.

Another universal trend is that smaller genomes evolve faster [i.e., have longer branch lengths (Fig. 3B)]. This has been noted before for pathogenic or endosymbiontic organisms with reduced genomes, and it is easily explained because they have only limited capabilities to remove mutations by means of recombination or DNA repair (*28*, *29*). However, we observe this trend also for genomes of larger sizes, including free-living prokaryotes and eukaryotes. Intriguingly, there is not a single organism sequenced that is fast-evolving and has a large genome (Fig. 3B). This suggests that the coupled processes of genome reduction and evolutionary acceleration may be irreversible: Genomes apparently do not grow again after a prolonged phase of genome reduction.

The pan-domain phylogeny that resulted from the procedure presented here will increase in resolution with more species being sequenced. This updatable reference phylogeny of completely sequenced species allows accurate comparisons of branch lengths across domains. The resulting tree of life will be an invaluable tool in many areas of biological research, ranging from classical taxonomy, via studies on the rate of evolution, to environmental genomics where DNA fragments of unknown phylogenetic origin need to be assigned.

References and Notes
1. F. Delsuc *et al.*, *Nat. Rev. Genet.* **6**, 361 (2005).
2. W. F. Doolittle, *Science* **284**, 2124 (1999).
3. J. P. Gogarten *et al.*, *Mol. Biol. Evol.* **19**, 2226 (2002).
4. J. R. Brown *et al.*, *Nat. Genet.* **28**, 281 (2001).
5. D. M. Hillis *et al.*, *Syst. Biol.* **52**, 124 (2003).
6. H. Philippe, C. J. Douady, *Curr. Opin. Microbiol.* **6**, 498 (2003).
7. V. Daubin, N. A. Moran, H. Ochman, *Science* **301**, 829 (2003).
8. Materials and methods are available on *Science* Online.
9. R. Jain *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801 (1999).
10. E. V. Koonin *et al.*, *Trends Genet.* **12**, 334 (1996).
11. C. G. Kurland, S. G. E. Andersson, *Microbiol. Mol. Biol. Rev.* **64**, 786 (2000).
12. Y. I. Wolf *et al.*, *Genome Res.* **9**, 689 (1999).
13. C. Brochier *et al.*, *Trends Genet.* **16**, 529 (2000).
14. A detailed version of the tree in Fig. 2, including bootstrap support values and branch lengths, is available at www.bork.embl.de/tree_of_life/.
15. A. L. Koch, *Trends Microbiol.* **11**, 166 (2003).
16. F. Raible *et al.*, *Science* **310**, 1325 (2005).
17. P. Hugenholtz *et al.*, *J. Bacteriol.* **180**, 4765 (1998).
18. D. Honda *et al.*, *J. Mol. Evol.* **48**, 723 (1999).
19. K. W. Nägeli, *Allg. Schweiz. Ges. Gesammten Naturwiss.* **10**, 45 (1849).
20. C. Brochier *et al.*, *Genome Biol.* **6**, R42 (2005).
21. A. Adoutte *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4453 (2000).
22. K. M. Halanych, *Annu. Rev. Ecol. Evol. Syst.* **35**, 229 (2004).
23. H. Philippe *et al.*, *Mol. Biol. Evol.* **22**, 1246 (2005).
24. G. K. Philip *et al.*, *Mol. Biol. Evol.* **22**, 1175 (2005).
25. J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).
26. D. P. Kreil, C. A. Ouzounis, *Nucleic Acids Res.* **29**, 1608 (2001).
27. K. T. Konstantinidis, J. M. Tiedje, *J. Bacteriol.* **187**, 6258 (2005).
28. N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2873 (1996).
29. S. G. E. Andersson, C. G. Kurland, *Trends Microbiol.* **6**, 263 (1998).

30. D. E. Wildman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7181 (2003).
31. We thank S. Hooper for help in designing trees and are grateful to members of the Bork group as well as T. Gibson, M. Gouy, M. Huynen, A. Budd, and M. Wolf for stimulating discussions. This work was supported in part by Bundesministerium fuer Bildung und Forschung (Nationales

Germline Mutations in Genes Within the MAPK Pathway Cause Cardio-facio-cutaneous Syndrome

Pablo Rodriguez-Viciana,[1]* Osamu Tetsu,[1,2]* William E. Tidyman,[3] Anne L. Estep,[1] Brenda A. Conger,[6] Molly Santa Cruz,[6] Frank McCormick,[1,4] Katherine A. Rauen[1,5]†

Cardio-facio-cutaneous (CFC) syndrome is a sporadic developmental disorder involving characteristic craniofacial features, cardiac defects, ectodermal abnormalities, and developmental delay. We demonstrate that heterogeneous de novo missense mutations in three genes within the mitogen-activated protein kinase (MAPK) pathway cause CFC syndrome. The majority of cases (18 out of 23) are caused by mutations in *BRAF*, a gene frequently mutated in cancer. Of the 11 mutations identified, two result in amino acid substitutions that occur in tumors, but most are unique and suggest previously unknown mechanisms of B-Raf activation. Furthermore, three of five individuals without *BRAF* mutations had missense mutations in either *MEK1* or *MEK2*, downstream effectors of B-Raf. Our findings highlight the involvement of the MAPK pathway in human development and will provide a molecular diagnosis of CFC syndrome.

There is an emerging group of medical genetic syndromes that are due to activating mutations in genes associated with the Ras pathway, including Noonan syndrome (NS, *PTPN11*) (*1*) and Costello syndrome (CS, *HRAS*) (*2*, *3*). Cardio-facio-cutaneous syndrome [CFC; Online Mendelian Inheritance in Man (OMIM) 115150] has many features that overlap with NS and CS. CFC is a sporadic, complex developmental disorder involving characteristic craniofacial features, cardiac anomalies (most commonly atrial septal defect and pulmonic stenosis), hair and skin abnormalities, postnatal growth deficiency, hypotonia, and developmental delay (*4*). Because of the similarity between CFC and CS, we screened patients with CFC for mutations in *HRAS* (*3*). We found no mutations in this gene, supporting a distinct genetic etiology between CS and CFC syndromes. We therefore expanded our search and sequenced other Ras genes (see Materials and Methods in the supporting online material), as well as genes encoding downstream effectors of Ras: *BRAF*, *CRAF*, *MEK1*, and *MEK2*. Our cohort consisted of 23 unrelated individuals with the clinical diagnosis of CFC syndrome who did not have a mutation in *HRAS* or *PTPN11* (table S1).

Using bidirectional sequencing of peripheral lymphocyte genomic DNA, we identified heterogeneous missense mutations in *BRAF* [GenBank accession (NM) 004333] in 18 out of 23 (78% of) individuals having CFC syndrome. Eleven distinct missense mutations clustered in two regions (Fig. 1A). Five individuals had a nucleotide (nt) switch, specifically nt770A→G transition in exon 6, with a predicted missense substitution of arginine for Gln^{257} (Q257R) (*5*) in the cysteine-rich domain (CRD) of the conserved region 1 (CR1) (Fig. 1B). The other cluster of mutations was in the protein kinase domain and involved exons 11, 12, 14, and 15. Five patients had heterogeneous missense mutations in exon 12 (table S2). Mutations identified at a lower frequency included missense mutations in the glycine loop encoded by exon 11 (*n* = 3), the catalytic domain encoded by exon 14 (*n* = 1), and the DFG motif in the activation segment (exon 15; *n* = 3). All parents and controls, totaling 40 phenotypically unaffected individuals, had none of these mutations, which supports the hypothesis that the occurrence of CFC is sporadic.

Although the causative mutations were heterogeneous, the distribution of mutations was specific and nonrandom. No frameshift, nonsense, or splice mutations were detected in the cohort of patients; thus, *BRAF* haploinsufficiency is not a likely causative mechanism of CFC. In 5 out of 23 (22% of) individuals with

CFC syndrome, no *BRAF* mutations were identified. Three of these individuals were found to have missense mutations in *MEK1* (NM 002755) and *MEK2* (NM 030662) that encode downstream effectors of B-Raf (Fig. 1C). Two individuals had heterogeneous mutations in *MEK1*: nt158T→C transition (F53S) and a nt389A→G transition (Y130C) in the protein kinase domain (fig. S1). CFC patient number 21 had a *MEK2* missense nt170T→G transversion, predicting a F57C substitution (Fig. 1D). Interestingly, F57 in MEK2 (MAPK kinase 2) is the equivalent position to F53 in the closely related MEK1, which suggests that substitutions of this residue may have similar functional consequences in the two family isoforms. The disease-causing mutations in the other two individuals remain to be identified.

The Raf/MEK/ERK cascade is the best understood of the MAPK pathways. (ERK, the extracellular signal–regulated kinase, is a type of MAPK.) In addition to B-Raf, the Raf family includes C-Raf-1 and the X-linked A-Raf. The expression pattern of each isoform is distinct (*6*), and genetic studies in mice have revealed nonredundant developmental functions (*7*, *8*). Somatic mutations in *BRAF* occur at high frequency in numerous human cancers (*9*). One mutation, ^{V600E}B-Raf, which confers increased kinase activity, accounts for more than 90% of these mutations. Its presence in benign nevi, as well as primary and metastatic melanoma, suggests that MAP kinase activation is important in melanocytic neoplasia but insufficient for tumorigenesis (*10*).

In contrast to the mutation spectrum seen in cancer, the *BRAF* missense mutations identified in individuals having CFC syndrome are more widely distributed (Fig. 1A). Of the 11 different missense amino acid substitutions, only five involve codons that are altered in cancer (table S2), yet only two individuals with CFC syndrome, both of whom have severe phenotypes, have the same substitution that has been reported in cancer (fig. S2). To explore the functional consequences of these mutations, we compared the kinase activity of the CFC B-Raf mutants to that of the wild-type protein (^{WT}B-Raf) and several cancer-derived mutants in transfected human embryonic kidney 293T cells (Fig. 2A; SOM). Four of the CFC B-Raf mutants had increased kinase activity compared with ^{WT}B-Raf, and this activity was as high as that of the ^{V600E}B-Raf mutant found in cancer. Two CFC B-Raf mutants had lower activity than ^{WT}B-Raf and appear to be kinase-impaired (*11*). Thus, the type of B-Raf mutations found in

[1]Comprehensive Cancer Center and Cancer Research Institute, [2]Department of Pathology, [3]Department of Anatomy, [4]Department of Microbiology and Immunology, [5]Department of Pediatrics, University of California, San Francisco, CA 94115, USA. [6]CFC International, Vestal, NY 13850, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: rauen@cc.ucsf.edu