Correspondence

Beyond 100 genomes

Paul Janssen*†, Benjamin Audit‡, Ildefonso Cases‡, Nikos Darzentas‡, Leon Goldovsky‡, Victor Kunin‡, Nuria Lopez-Bigas‡, José Manuel Peregrin-Alvarez‡, José B Pereira-Leal‡, Sophia Tsoka‡ and Christos A Ouzounis‡

Addresses: *Centre d' Ingénierie des Protéines (CIP), Université de Liège, 4000 Liège, Belgium. ‡Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK. †Current address: Laboratory of Microbiology, Belgian Nuclear Research Centre, SCK/CEN, Boeretang 200, B-2400-MOL, Belgium.

Correspondence: Christos A Ouzounis. E-mail: ouzounis@ebi.ac.uk

Published: 28 April 2003

Genome ***Biology*** 2003, **4**:402

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/5/402

Since the publication of the first entire genome sequence seven years ago [1], a multitude of other genomes have been - or are in the process of being - sequenced [2]. By the end of 2002, we witnessed the landmark submission of the 100th complete genome sequence in the databases [3]. There are now 106 complete genomes in the public domain, thanks to advances in sequencing technology and sustained funding. An overview, and in particular the rank ordering, of these genomes reveals certain interesting trends and provides valuable insights into possible future developments.

First, the contribution of genome sequencing projects in terms of actual protein sequence entries has been staggering. There are 433,238 protein sequences derived exclusively from entire genomes [4] (Figure 1), out of a total of a million protein sequences known to date. In contrast, there are only 101,602 entries in Swiss-Prot (release 40), underlining the significant effort that is required for high-quality annotation [5]. The growth of protein sequence data coming from entire genomes is expected to reach over 1 million entries in two years' time (Figure 1). Given that approximately 40% of genes in any organism cannot be assigned a specific functional role [6], this suggests that in just a few years hundreds of thousands of sequences will be uncharacterized. While the large-scale characterization of protein function obtained from high-throughput experimental techniques [7] will

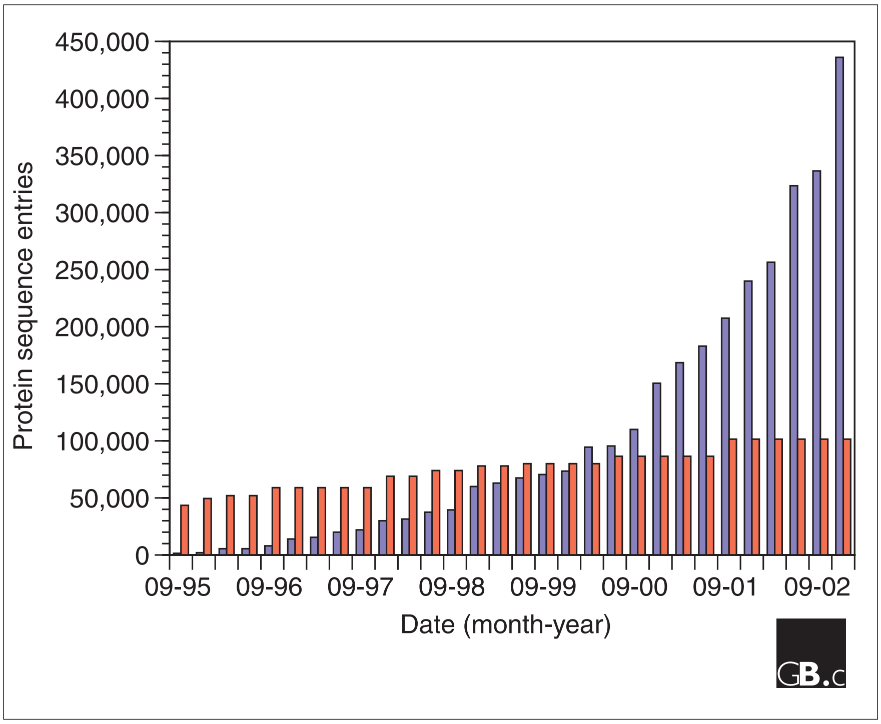

Figure 1
Cumulative number of protein sequence entries (*y*-axis) in completed genomes (CoGenT, in blue) and Swiss-Prot (in red) as a function of time (*x*-axis).

alleviate some of the above problems, it is clear that to capitalize on the information explosion in genome biology, more research should also be devoted to the development of intelligent automated genome-annotation systems that are able to predict functional properties of protein sequences [8].

Second, in addition to the well-defined collection of 106 completed and published genomes, there are another 544 ongoing projects, covering a large number of taxa. Yet, the known taxa of Bacteria and Archaea are far better represented among the completed genome projects compared to the Eukarya (Figure 2). Using comparative genomics we have already obtained a glimpse of the bewildering biological diversity of the prokaryotic world [9]. Very soon, a similar trend might emerge for the Eukarya: 208 out of the 544 ongoing genome projects are dedicated to eukaryotic species. However, many eukaryotic taxa are still not represented (Figure 2). A better sampling of phylogenetic diversity might be required, to fully explore the genomes of eukaryotic cells.

Third, over time, both the range of sequenced genome sizes and the selection of species on the basis of their social impact has expanded [10] (Figure 3). Sequenced genome sizes range from 0.5 to 300 Megabases (Mb), with the exception of the human and mouse genomes, which span 2,900 and 2,500 Mb respectively (and together constitute almost 90% of the data in the 106 available DNA sequences). Although species of medical and academic interest were initially the main targets of genome projects, there has been a recent trend to sequence genomes from species with impact on agriculture, environmental sciences or industrial processes. In addition, a growing number of genomes are being sequenced in order to provide a better perspective for the structure and function

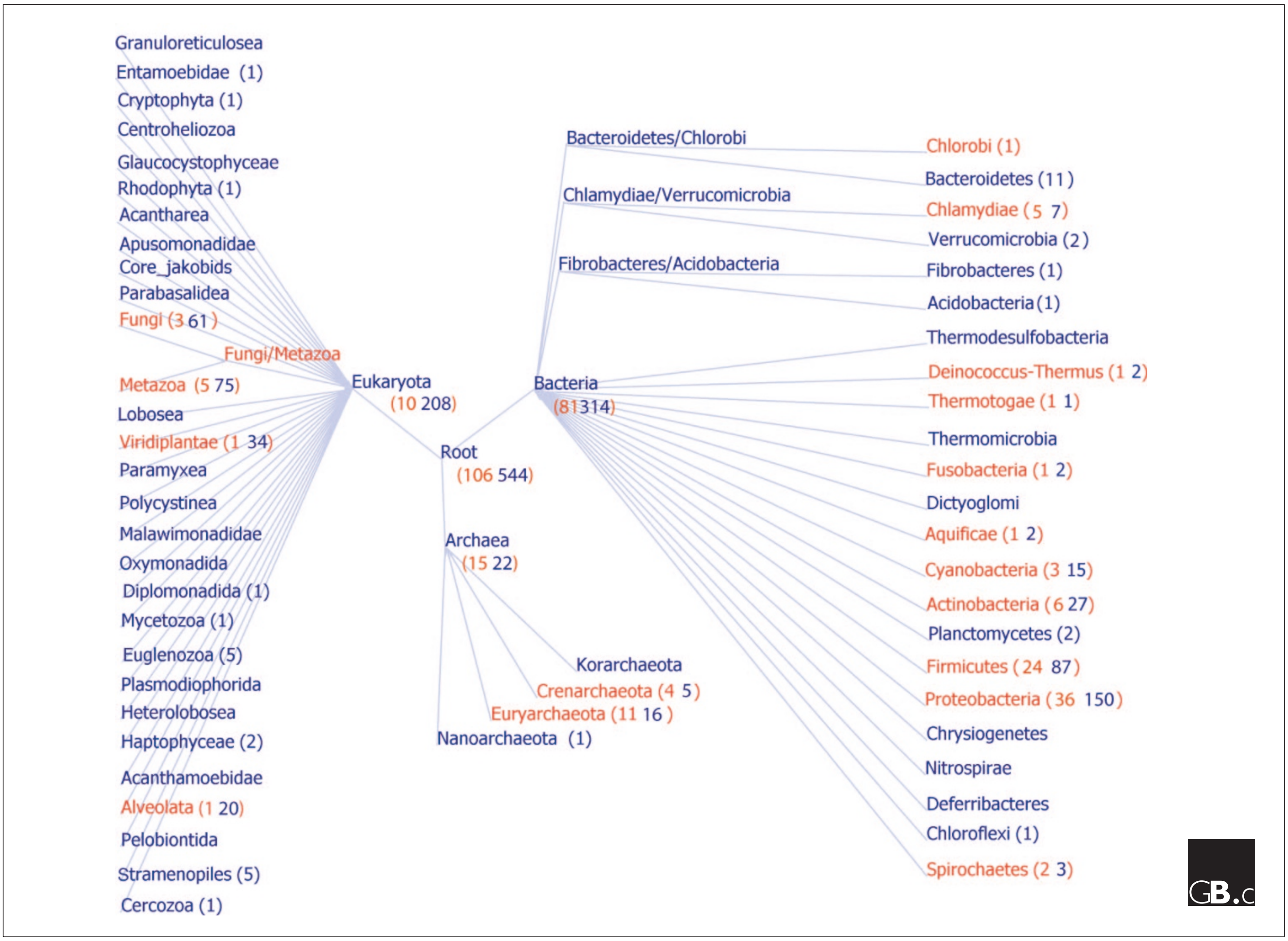

Figure 2
Phylogenetic distribution of genome sequencing projects. Archaea and Bacteria are shown to the phylum level and Eukarya to their first taxonomic branching, with the exception of Metazoa and Fungi. The numbers in parentheses represent the number of completed, published (red) and ongoing (blue) genome projects. The tree is based on the taxonomy database from the National Center for Biotechnology Information (NCBI). Information about ongoing genome projects has been obtained from the Genomes OnLine Database (GOLD) [14], as of 22 January 2003.

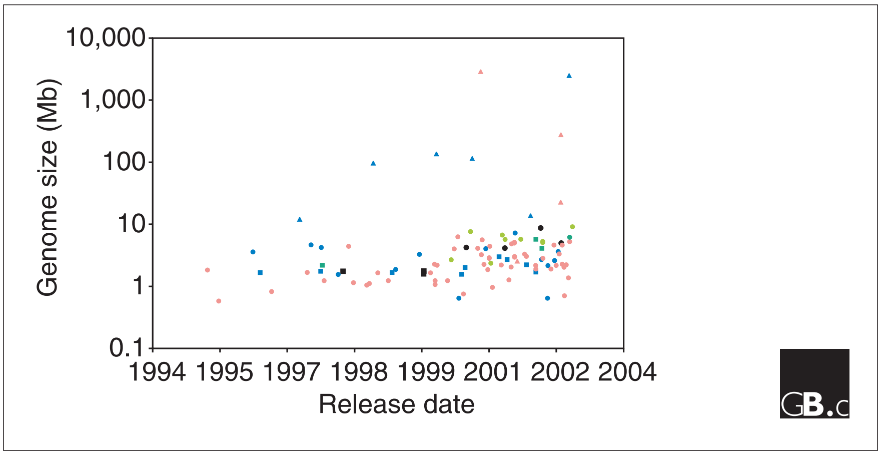

Figure 3
Representation of completed genome sequences over time (*x*-axis) and size (*y*-axis, in Mb, logarithmic scale) labeled according to their social impact. Genomes from Archaea (squares), Bacteria (circles) and Eukarya (triangles) are colored according to their academic (blue), medical (pink), agricultural (light green), ecological (dark green) and industrial (black) relevance.

of evolutionarily related genes and genomes through comparative analysis.

Thus, 10 years after the computational analysis of the first eukaryotic chromosome [11] and seven years after an exhaustive analysis of the first complete genome [1,12], genomic science has become a stand-alone discipline, and genome sequencing and computational analysis have become mutually dependent, intertwined in a fascinating interplay. Not so long ago it would have been unthinkable that from a set of DNA fragments, it would be possible to assemble a single genome, find the genes, translate them into proteins, identify their potential functional roles and ultimately integrate all this structural and functional information into complex biochemical networks [13]. Although there are still significant challenges, these technologies, along with scientific advances, have now come of age and are expected to have a growing impact on various aspects of human welfare.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269:**496-512.
2. Nelson KE, Paulsen IT, Heidelberg JF, Fraser CM: **Status of genome projects for nonpathogenic bacteria and archaea.** *Nat Biotechnol* 2000, **18:**1049-1054.
3. Akman L, Yamashita A, Watanabe HO, K., Shiba T, Hattori M, Aksoy S: **Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*.** *Nat Genet* 2002, **32:**402-407.
4. **Complete Genome Tracking Database** [http://maine.ebi.ac.uk:8000/services/cogent]
5. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-8.
6. Iliopoulos I, Tsoka S, Andrade MA, Janssen P, Audit B, Tramontano A, Valencia A, Leroy C, Sander C, Ouzounis CA: **Genome sequences and great expectations.** *Genome Biol* 2000, **2:**interactions0001.1-0001.3.
7. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM: **A biochemical genomics approach for identifying genes by the activity of their products.** *Science* 1999, **286:**1153-1155.
8. Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C: **Automated genome sequence analysis and annotation.** *Bioinformatics* 1999, **15:**391-412.
9. Torsvik V, Ovreas L, Thingstad TF: **Prokaryotic diversity - magnitude, dynamics, and controlling factors.** *Science* 2002, **296:**1064-1066.
10. Doolittle RF: **Biodiversity: microbial genomes multiply.** *Nature* 2002, **416:**697-700.
11. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: **What's in a genome?** *Nature* 1992, **358:**287.
12. Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C: **Challenging times for bioinformatics.** *Nature* 1995, **376:**647-648.
13. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405:**823-826.
14. Bernal A, Ear U, Kyrpides N: **Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.** *Nucleic Acids Res* 2001, **29:**126-127.

comment
reviews
reports
deposited research
refereed research
interactions
information