Genome Biology

RESEARCH HIGHLIGHT

# Filling the gaps in the genomic landscape

David Williams[1], J Peter Gogarten*[1] and Pascal Lapierre[2]

**Abstract**

A new initiative provides comparative genomicists with a more complete picture of genome diversity. Here we discuss the improved sampling strategy.

The relentless progress in sequencing technology continues to open up new opportunities for biologists. Since surveys of the first complete genetic code of single organisms only 15 years ago, findings from comparative genomics are now commonplace thanks to the more than 1,000 sequenced genomes available. Among the most striking discoveries is the high level of variation in gene content between closely related microbial strains. However, a representative sample of genomes from cultured bacteria and archaea has, until very recently, been out of reach. Many previous sequencing efforts have focused on useful, dangerous or unusual microorganisms, providing a patchy sampling of the known phylogenetic diversity. A new initiative, the 'Genomic Encyclopedia of Bacteria and Archaea' (GEBA) reported recently by Wu *et al.* [1], aims to fill in the gaps to provide a more complete picture of genomic diversity. The initial stage of the project aims to complete 159 genomes across the Bacteria and the Archaea selected according to their position in a phylogenetic tree of small subunit (SSU) rRNA, with in-depth sampling of the Actinobacteria. By analyzing 56 of the newly sequenced genomes the authors demonstrate improvements in the rate of novel protein discovery and extend the diversity and distribution of known protein families - a clear indication of the success of the new sampling strategy. On this basis we can expect further revelations in the near future. Here we discuss the advantages of the new sampling strategy and its limitations in the light of the apparent non-tree-like histories of whole genomes inferred from recent comparative genomic studies.

*Correspondence: gogarten@uconn.edu
[1]Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125, USA
Full list of author information is available at the end of the article



## Genome content and diversity

During the past decade, our understanding of evolution at the genomic level has been shaken to its core by many reports showing that genomes from closely related species can vary greatly in terms of gene content. The rapid alteration of gene content in genomes was first demonstrated by Welch *et al.* [2], with the comparison of three strains of *Escherichia coli*. They found that only about 39% of the non-overlapping set of genes were present in all three strains, leaving the majority of the genes to have either been gained through gene transfers and internal duplication, or lost along the evolutionary path of the different strains. The extreme plasticity of genome composition is illustrated by the comparison of genomes from three *Frankia* strains, a class of nitrogen-fixing soil bacteria whose members form symbiotic relationships with actinorhizal plants [3]. It was found that the biggest of the three genomes almost doubles the number of ORFs found in the smallest one, a feature that can be associated with the range of plants each can infect and their geographic locations.

A measure of protein diversity among related species can be derived by looking at the pan-genome of the whole group - that is, the pool of genes present in the group collectively, including those that are not present in all individuals. Tettelin *et al.* [4] sequentially sampled genes from eight Group-B *Streptococcus* (GBS) genomes and concluded that on the basis of the number of unique genes found in those eight genomes, one should expect to find an average of 30 new genes for every additional GBS genome sequenced. This was an outstanding finding because it implied an infinite number of proteins present in the pan-genome of GBS. When the concept was extended to the Bacteria more widely by analyzing 573 bacterial species using a gene frequency sampling approach [5], the number of expected unique genes per genome increased to an average of about 200 with no sign of leveling off. Results from the comparisons of the 56 genomes in the GEBA project confirm the existence of a surprising number of previously unknown gene families. Wu *et al.* [1] found that these 56 genomes provided a discovery rate of more than 1,000 novel protein families per genome. By sequencing bacterial genomes from under-represented phyla, they revealed that currently recognized protein diversity is likely to represent only a small fraction of the diversity existing in nature.

## Gene trees and genome networks

So how can this meta-genomic structure be modeled realistically when it comes to prokaryotic phylogenetics? Informed by the strong ancestral lineages seen in higher organisms, a tree-like model of evolution was originally extended to include microbial life by modeling sequence evolution in SSU rRNA [6]. This approach provided an early indication of the staggering diversity to be found among microorganisms and led to the classification of life into three domains. However, subsequent analyses of other gene families revealed clear incongruities between gene trees consisting of similar organisms [7]. In the light of evidence from more recent analyses, the once clear lines of the tree model for the history of species and their genomes have become somewhat blurred [8]. Although Wu *et al*. [1] used a SSU rRNA tree-guided sampling approach, in their initial assessment of the first 56 GEBA genomes they found 1,768 out of 16,797 protein families with no significant sequence similarity to known proteins. Furthermore, when comparing the 53 new bacterial genomes with 53 randomly sampled previously sequenced bacterial genomes, 2.8 to 4.4 times more phylogenetic diversity was observed for a concatenated alignment of 31 broadly conserved protein-coding genes.

Anticipating genome content is a difficult problem. Wu *et al*. [1] demonstrated that the use of a SSU phylogenetic tree as a sampling guide provides a substantial improvement in new information per genome sequenced. Their analysis of 31 broadly conserved protein-coding gene families confirmed the utility of phylogenetic sampling in obtaining a richer sample of protein diversity. While such a tree provides some measure of average protein diversity [9], this average signal does not necessarily represent the history of the genomes. By combining genes with different histories into a single supermatrix, the conflicting phylogenetic signals are likely to lead to artifacts in a tree-only reconstruction. The resulting tree may be dominated by signals due to highways of gene sharing [8] between certain lineages and may not be representative of the history of the organism, its genome, or of a single major cellular component [10].

But how do we reconcile a tree-like relationship between whole organisms with the varied evolutionary history of individual genes found in genomes? One solution offered is to use a combination of genes that we know are more resilient to gene transfers and have a higher likelihood of reflecting the true evolutionary history of the organisms. Examples include genes coding for highly integrated cellular components such as the ribosome or ATP synthases, for which a tree-like history is more likely. Inferred histories of the remaining gene families can be added to provide a more accurate reconstruction of the network-like evolutionary history of genomes [10].

Wu *et al*. [1] have demonstrated that selecting genomes to sequence on a phylogenetic basis is a far more profitable use of resources in terms of diversity exploration than the previous, less coordinated approach. The GEBA initiative will thus provide the data necessary to answer important questions in microbiology sooner than would otherwise be possible. As the authors anticipate, the final piece of the puzzle will be effective means to sequence genomes from organisms lacking representatives in pure culture. When this is achieved we will be able to approach a complete picture of genomic diversity.

### Abbreviations

GBS, Group-B *Streptococcus*; GEBA, Genomic Encyclopedia of Bacteria and Archaea; ORF, open reading frame; SSU, small subunit.

### Acknowledgements

Work in the authors' lab is supported through the NSF Assembling the Tree of Life (DEB 0830024) and NASA exobiology (NNG05GN41G and NNX07AK15G) programs.

### Author details

[1]Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125, USA
[2]University of Connecticut Biotechnology Center, 91 North Eagleville Road, Storrs, CT 06269-3149, USA

Published: 16 February 2010

### References

1. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova N, Kunin V, Goodwin L, Wu M, Tindall B, Hooper S, Pati A, Lykidis A, Spring S, Anderson I, Dhaeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, *et al.*: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.** *Nature* 2009, **462:**1056-1060.
2. Welch R, Burland V, Plunkett III G, Redford P, Roesch P, Rasko D, Buckles E, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew G, Rose D, Zhou S, Schwartz D, Perna N, Mobley H, Donnenberg M, Blattner F: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*.** *Proc Natl Acad Sci USA* 2002, **99:**17020-17024.
3. Normand P, Lapierre P, Tisa L, Gogarten J, Alloisio N, Bagnarol E, Bassi C, Berry A, Bickhart D, Choisne N, Couloux A, Cournoyer B, Cruveiller S, Daubin V, Demange N, Francino M, Goltsman E, Huang Y, Kopp O, Labarre L, Lapidus A, Lavire C, Marechal J, Martinez M, Mastronunzio J, Mullin B, Niemann J, Pujic P, Rawnsley T, Rouy Z, *et al.*: **Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography.** *Genome Res* 2007, **17:**7-15.
4. Tettelin H, Masignani V, Cieslewicz M, Donati C, Medini D, Ward N, Angiuoli S, Crabtree J, Jones A, Durkin A, DeBoy R, Davidsen T, Mora M, Scarselli M, Margarit Y Ros I, Peterson J, Hauser C, Sundaram J, Nelson W, Madupu R, Brinkac L, Dodson R, Rosovitz M, Sullivan S, Daugherty S, Haft D, Selengut J, Gwinn M, Zhou L, *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome".** *Proc Natl Acad Sci USA* 2005, **102:**13950-13955.
5. Lapierre P, Gogarten J: **Estimating the size of the bacterial pan-genome.** *Trends Genet* 2009, **25:**107-110.
6. Woese C, Fox G: **Phylogenetic structure of the prokaryotic domain: The primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74:**5088-5090.
7. Hilario E, Gogarten J: **Horizontal transfer of ATPase genes - The tree of life becomes a net of life.** *BioSystems* 1993, **31:**111-119.
8. Beiko RG, Harlow TJ, Ragan MA: **Highways of gene sharing in prokaryotes.** *Proc Natl Acad Sci USA* 2005, **102:**14332-14337.
9. Pardi F, Goldman N: **Resource-aware taxon selection for maximizing phylogenetic diversity.** *Syst Biol* 2007, **56:**431-444.
10. Swithers K, Gogarten J, Fournier G: **Trees in the web of life.** *J Biol* 2009, **8:**54.

doi:10.1186/gb-2010-11-2-103
**Cite this article as**: Williams D, *et al.*: Filling the gaps in the genomic landscape. *Genome Biology* 2010, **11:**103.