Genome Update: annotation quality in sequenced microbial genomes

Genomes of the month

There are five new microbial genomes described in this month’s Genome Update, four from bacterial species and one from a fungus. The bacterial genomes are those of Desulfovibrio vulgaris, Listeria monocytogenes, Mycobacterium avium and a Mediaevalis strain of Versinia pestis. The fungal genome is that of Phanerochaete chrysosporium, the fungus responsible for ‘white rot’ in decaying trees.

D. vulgaris is a sulfate-reducing bacterium found almost everywhere in nature, and is responsible for biocorrosion of metal infrastructures (e.g. oil drilling and pumping machinery); it can be also used for bioremediation of toxic metal ions such as cadmium and uranium. D. vulgaris subsp. vulgaris strain Hildenborough is the third genome of the δ-Proteobacteria to be published. Its genome is GC-rich and encodes about 3400 genes (see Table 1). The main chromosome is 3.5 Mbp long, and the genome also includes a 0.2 Mbp plasmid (Heidelberg et al., 2004).

Listeria monocytogenes is a common environmental bacterium which can cause food poisoning. Three different strains of Listeria monocytogenes associated with food-borne infections have been sequenced: the genome of strain F2365 (serotype 4b, cheese isolate) was fully sequenced, while the genomes of strains F6854 (serotype 4b, frankfurter isolate) and H7858 (serotype 4b, meat isolate) were sequenced to give about 8× coverage, although for these two chromosomes the gaps were not closed (Nelson et al., 2004). This report is a nice example of the power of comparative genomics – by comparing the relatively small number of genes unique to each genome (e.g. one genome contained only 51 unique genes), it is possible to model strain-specific and serotype-specific differences. One of the conclusions of the authors of this report is that ‘L. monocytogenes strains prevalent in human and animal illness have surprisingly high genomic stability, and rely on a relatively small number of unique regions for antigenic diversity and epidemiologically relevant attributes’ (Nelson et al., 2004).

The sequences of two other bacterial genomes listed in Table 1 have been deposited in the EMBL/GenBank libraries, but have not been published yet. The sequenced genome of Mycobacterium avium subsp. paratuberculosis strain k10 was searched for short sequence repeats (SSRs), which were used to design probes that were tested for discrimination amongst 33 different strains of the same species (Amponsah et al., 2004). The genome sequence of Y. pestis bv. Mediaevalis str. 91001 (see Table 1) has also been deposited recently in the EMBL/GenBank libraries, and seems similar in size and many characteristics to the other two Y. pestis strains sequenced to date.

Finally, the 30 Mbp genome of the white rot-causing basidiomycete P. chrysosporium RP78 has been published recently (Martinez et al., 2004); it is organized into ten chromosomes and encodes about 11 800 genes. This organism secretes enzymes into its environment which allow it to efficiently degrade lignin. The number of rRNA operons in the genome of this organism is not given, although the authors state, with admirable truthfulness and clarity, that ‘Typical of shotgun sequencing of eukaryotes, extended repeats, telomeres and rRNA clusters were excluded from the assembly. Nevertheless, substantial numbers of noncoding repetitive sequences and putative mobile elements were assembled’ (Martinez et al., 2004). This is the first genome to be sequenced from a member of the fungal phylum Basidiomycota.

Method of the month – comparison of genome annotation

For the past several months, we have been systematically going through the columns of genomic data shown in Table 1. Thus, for example, last month’s Genome Update included a discussion of comparison of tRNAs and codon usage in various sequenced genomes. This month we come to the last column, which is the number of genes annotated in a given genome. For sequenced bacterial genomes, the range is from a mere 480 genes in Mycoplasma genitalium (Fraser et al., 1995) to 8317 genes in Bradyrhizobium japonicum (Kaneko et al., 2002); the upper limit is already known to be a bit larger than this, as the genome of the mycobacterium ‘Sorangium cellulosum’ is about 12 Mbp long, and is likely to contain more than 10 000 genes (Pradella et al., 2002). Thus, the number of genes in bacterial genomes varies by more than 20-fold. Furthermore, only a small
fraction of the 480 genes in *Mycoplasma genitalium* are well-conserved in other bacteria. (We find less than 10%, although the number depends, of course, on what threshold one chooses; with an e-value cut-off of about $1 \times 10^{-10}$, we find only about 40 genes that are conserved throughout the sequenced bacterial genomes).

We had originally intended to make a plot of the number of genes in bacterial genomes, broken down into phyla, but since the coding density for most bacteria is quite high and roughly the same, this plot looks essentially the same as the one shown a few months ago when the length of genomes was discussed [see Fig. 1 in Ussery & Hallin (2004)]. Instead, this month we will briefly discuss how we can estimate the quality of genome annotation.

Many people assume that, since gene-finding is relatively easy in bacterial and archaeal genomes, the genes reported in EMBL or GenBank files have THE correct annotation. However, in writing these Genome Update articles over the past several months, it has become clear that the genomes are not all annotated to the same standards. For example, about 10% of the bacterial genomes (e.g. 15 genomes out of 150 published) do not have the rRNA gene sequences annotated in their GenBank files. It is clear that there can be occasional large differences in the quality of annotation of the genes as well. For example, consider the *Leptospira interrogans* Copenhageni strain Fiocruz L1-130 genome (Nascimento et al., 2004). As mentioned in last month’s Genome Update, this is nearly identical in size to another *Leptospira interrogans* genome (Ren et al., 2003), which has nearly 1000 extra genes (3728 vs 4727 genes, for two bacterial genomes of the same species, both about 4.7 Mbp in length). It seems to be a general rule of thumb that there is very roughly one gene every 1000 bp in many bacterial genomes. Using this criterion, one might expect there to be about 4700 genes encoded by the *Leptospira interrogans* genomes – so perhaps the earlier estimate is closer to what is expected. But what if a genome has undergone some sort of decay, or perhaps had a large insertion of non-coding regions? Is there a practical way for estimating how many genes there should be in a given bacterial genome? A few years ago, we utilized three different statistical measures to estimate the expected number of genes (Skovgaard et al., 2001), and found that most of the 28 bacterial genomes examined at that time were over-annotated by about 20%, compared to our estimates (which, admittedly, could be conservative). The problem is separating the ‘ORFs from the ELFs’ (Lawrence, 2003; Ochman, 2002) – i.e. trying to accurately determine the small open reading frames (ORFs) which truly encode proteins, versus random ORFs which occur by chance and might not reflect genes which are ever expressed. It is difficult to prove that a given sequence is not a gene – just that it is not expressed under a given set of experimental conditions. However, the problem of over-annotation is unfortunately not one that can be easily

### Table 1. Summary of the published genomes discussed in this Update

Note that the accession number for each chromosome is the same for GenBank, EMBL and the DNA DataBase of Japan (DDBJ).

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size (bp)</th>
<th>AT content (%)</th>
<th>rRNA operons</th>
<th>tRNAs</th>
<th>CDS</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Desulfovibrio vulgaris</em> subsp. <em>vulgaris</em> Hildenborough</td>
<td>3 570 858</td>
<td>36.9</td>
<td>5</td>
<td>68</td>
<td>3 395</td>
<td>AE017285</td>
</tr>
<tr>
<td><em>Listeria monocytogenes</em> F2365 (serotype 4b, cheese isolate)</td>
<td>2 905 310</td>
<td>62.0</td>
<td>6</td>
<td>67</td>
<td>2 847</td>
<td>AE017262</td>
</tr>
<tr>
<td><em>Mycobacterium avium</em> subsp. <em>paratuberculosis</em> k10</td>
<td>4 829 781</td>
<td>30.7</td>
<td>1</td>
<td>46</td>
<td>4 350</td>
<td>AE016958</td>
</tr>
<tr>
<td><em>Yersinia pestis</em> bv. Mediaevalis str. 91001</td>
<td>4 595 065</td>
<td>52.3</td>
<td>7</td>
<td>72</td>
<td>3 895</td>
<td>AE017042</td>
</tr>
<tr>
<td><em>Phanerochaete chrysosporium</em> RP78</td>
<td>~ 29 900 000</td>
<td>48.5</td>
<td>?</td>
<td>199</td>
<td>11 777</td>
<td>AADS00000000</td>
</tr>
</tbody>
</table>

Fig. 1. Over-annotation of archaeal and bacterial genomes. Box-and-whiskers plots show data from the superkingdoms *Bacteria* and *Archaea* and phyla within these groups; more information, including the values for each genome, can be found in the supplemental web pages associated with this article. Note that alpha, beta, gamma and epsilon refer to classes within the *Proteobacteria*. 

![Graph showing over-annotation of genomes]
ignored, if one wants to compare bacterial proteomes. For example, is it REALLY likely to be true that one Leptospira interrogans genome encodes an extra 1000 proteins compared to another?

Fig. 1 shows the fraction of genes estimated to be over-annotated for 159 sequenced genomes, sorted by superkingdom and phylum. Nearly all genomes are ‘over-annotated’ by about 20%. The good news, in a way, is that most of the genomes seem to have roughly the same ratio of genes annotated, compared to our estimates. Note that Crenarchaeota genomes seem to be more over-annotated; however, this is based on only four genomes, two of which are by far the most over-annotated, with about twice as many genes predicted as might be expected (e.g. one gene every 500 bp, rather than one gene every 1000 bp as for most other genomes). These were both annotated by the same group, which actually reported all the ORFs over a certain length. It should also be noted that a gene-finder has not been run through the sequences.

For the sequenced bacterial genomes, one of the most over-annotated genomes is that of the above-mentioned Leptospira interrogans strain (Copenhageni strain Fiocruz L1-130). We estimate its genome to be over-annotated by about 60%, whilst the genome of Leptospira interrogans strain S6601 is about 30% over-annotated, closer to the values for the other spirochaete genomes. It could well be that our estimates for the ‘true number of proteins’ are too low. However, regardless of this fact, at least this provides us with some measure to allow us to see which genomes differ significantly in their annotation criteria. Full results, including a table of all published archaeal and bacterial genomes, sorted by their ‘over-annotation’ value, as well as protein length distribution plots for each genome, can be found on our supplemental web pages. Only about ten genomes are significantly different from the average – perhaps these genomes should be treated with caution when doing proteome comparisons based solely on the EMBL or GenBank files.

Next month, the method of genome comparison discussed will be the Artemis Comparison Tool (ACT) (http://www.sanger.ac.uk/Software/ACT/).

**Supplemental web pages**

Web pages containing supplemental material related to this article can be accessed from the following url: http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp006/

**Acknowledgements**

This work was supported by a grant from the Danish Center for Scientific Computing.

**David W. Ussery and Peter F. Hallin**

Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

Correspondence: David W. Ussery (dave@cbs.dtu.dk)


Lawrence, J. (2003). When ELFs are ORFs, but don’t act like them. Trends Genet 19, 131–132.


DOI 10.1099/mic.0.27338-0

http://mic.sgmjournals.org