Genome coverage, literally speaking

The challenge of annotating 200 genomes with 4 million publications

Paul Janssen, Leon Goldovsky, Victor Kunin, Nikos Darzentas & Christos A. Ouzounis

In late 2004, 200 complete genomes had been sequenced and made available to the research community. At the time of writing this viewpoint, that number had further risen to 221 and will have undoubtedly increased again before publication. These genomes, which represent a wide range of species from archaea to human, are a highly valuable knowledge resource for the scientific community. However, the sequencing of a full genome is just the first step in research; it must be followed by the functional characterization of genes and proteins. In this context, it is interesting to see how well represented these sequenced species are in terms of publications. We have thus obtained the number of abstracts published per species and normalized that count by the number of genes in that species to obtain a comparable measure for the number of publications per gene for all completed and published genomes. This simple measure highlights the current knowledge gap between various organisms and could further serve as a guideline for selecting genomes for sequencing projects, high-throughput functional genomics and database annotation efforts.

The 200 complete genome sequences published by December 2004 included 118 genera, 166 species and 34 additional strains for 21 species. This rate translates to a doubling time of available genome sequences of less than two years (Janssen *et al*, 2003a). And it remains steady: in 2003, an average of one complete genome was released per week; 47 genomes were made available in the first 44 weeks of 2004. This trend will accelerate further, as more than 1,000 genome projects are currently underway (Bernal *et al*, 2001).

For the 221 genomes currently available, the total number of predicted proteins is

822,114, according to the COGENT database (Janssen *et al*, 2003b). One of the great challenges for computational and experimental genomics is the functional characterization of the genes and proteins encoded in these genomes (Eisenberg *et al*, 2000), a process that should be considered as continuous (Ouzounis & Karp, 2002). To achieve this goal, it is important to rely on existing knowledge and draw from previous studies conducted on these organisms. We have analysed how well each of the currently available genomes is characterized in terms of the number of publications pertinent to the corresponding species. To achieve a reliable measure for the available knowledge per genome, we obtained the number of abstracts per species—but not strain—from Medline and divided this number by the number of predicted protein-encoding genes. The corresponding ratio, which we term the Species Knowledge Index (SKI), thus reflects our current understanding of each of these species. More detailed information on the literature coverage of sequenced organisms, tracked by COGENT, is available through our GenMed server (<http://cgg.ebi.ac.uk/cgi-bin/genmed/genmed.pl>).

This analysis encompasses 3,806,293 Medline abstracts corresponding to 200 genomes. On average, there are 5 abstracts per gene for the first 200 genomes in the COGENT database, ranging from 1 abstract per 1000 genes for poorly characterized species, to 55.1 abstracts per gene for *Escherichia coli* (Fig 1A). This arithmetic mean value is grossly distorted by a few outliers (Fig 1B; colour-coded), namely *E. coli*, human (48.5), *Staphylococcus aureus* strains (ca. 16–17), mouse (15.6), *Helicobacter pylori* strains (ca. 13) and *Saccharomyces cerevisiae* (10.6). If these outliers are

excluded, then the average SKI value drops to 0.9 (580,016 abstracts for 651,183 genes). Yet, this value is still dominated either by important pathogens, such as *Chlamydia trachomatis* (9.5) and *Haemophilus influenzae* (8.9), or by model organisms such as *Pseudomonas aeruginosa* (5.7) and *Bacillus subtilis* (4.8).

...the SKI measure reflects our current status of knowledge for each species or strain at the molecular level...

Other model species follow at lesser ranks, for instance *Drosophila melanogaster* (1.3), *Schizosaccharomyces pombe* (1.1) and *Caenorhabditis elegans* (0.5). The most studied archaeal species are *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii*, both with SKI values of 0.3. At the very end of the scale are species that have been recently characterized or are of environmental interest (Janssen *et al*, 2003a), such as *Gloeobacter violaceus* and *Oceanobacillus iheyensis* which have only a few abstracts, corresponding to an average of 1 per 1,000 genes. It is worth noting that there is a tendency for the SKI value to decrease in value with respect to the order in which a genome was completed (not shown).

These results demonstrate a number of limitations to our approach. First, the time of characterization of the corresponding species is not taken into account; obviously, recently characterized species have lower SKI values. A rate value that takes into account the number of abstracts per gene per year would produce a more accurate estimate for the increase in

Fig 1 | Relationship between the number of genes and number of abstracts for the first 200 genomes made available in COGENT (v. 207, 29 Oct 2004, Janssen *et al*, 2003b; abstracts collected from Medline on 5 Nov 2004). (A) The x-axis represents the genome rank, sorted by SKI value corresponding to the order of appearance (release and publication). The y-axis corresponds to the number of genes (blue), abstracts (grey) and the SKI value (red), on a logarithmic scale. (B) The x-axis represents the number of genes, the y-axis the number of abstracts and the radius of circles the SKI value (scale shown). The five principal outliers mentioned in the text are coloured as follows: *Escherichia coli* strains (green), human–mouse pair (purple), *Staphylococcus aureus* strains (red), *Helicobacter pylori* strains (brown) and *Saccharomyces cerevisiae* (grey). In both representations, it is evident that there is a range of two orders of magnitude for gene numbers and a range of six orders of magnitude for abstract numbers, which results in a range of four orders of magnitude for the SKI value, between 0.001 and <100.

scientific interest for each organism. Second, naming conventions are not strictly followed in the literature, thus making it difficult to accurately retrieve the number of abstracts for *Homo sapiens*, for

example. Not all abstracts that contain the word ‘human’ necessarily relate to molecular biology and genes, but they might contain relevant information; as a filter, we used the keyword ‘protein’ for human and

mouse to focus on molecular information. In addition, species that are used in biotechnology, such as *E. coli*, are unquestionably over-represented. Similarly, there is no sufficient resolution for strain names; however, it is safe to assume that closely related strains have similar properties and the body of existing knowledge might therefore be easily transferable across strains in most cases. Third, Medline indices primarily cover medical literature and related biological sciences. It is therefore conceivable that journals that publish studies on organisms of environmental or industrial interest are not included.

Certain communities might not successfully advocate the sequencing of their favourite species, despite its relative importance in terms of available abstracts in Medline...

The accuracy of the SKI measure for organisms could further improve if the dynamics of their nomenclature were considered. For instance, *Ralstonia solanacearum* had 159 hits in Medline, but this organism was first classified as *Burkholderia solanacearum* (50 hits) and thereafter as *Pseudomonas solanacearum* (167 hits). The keyword ‘solanacearum’ produced 320 hits, close to the sum for all three species (376 hits). The slight discrepancy results from the use of more than one name in abstracts, for instance “...in *Ralstonia solanacearum* (formerly *Pseudomonas solanacearum*), a phytopathogenic bacterium...”. In this particular case, the term ‘solanacearum’ would thus result in a more accurate count of scientific papers for the organism *R. solanacearum*, assuming that no other species name contains this term. However, it would be unrealistic to consider the dynamics in nomenclature for every single species. Nevertheless, it may be possible to obtain improved SKI values in the future, assuming that the literature keeps pace with taxonomic modifications. In this respect, it may be worthwhile to use the yearly average of SKI values over a suitably long period and compare those values only with each other. In essence, global and average SKI values are not static but will change over time, thus reflecting renewed or dwindling scientific interest in a particular organism.

Despite the above limitations—namely the time of species characterization, taxonomic conventions and potential biases—the SKI measure reflects our current status of knowledge for each species or strain at the molecular level and essentially delimits the number of published studies according to the number of genes. To achieve an improvement, better descriptions of strains and species in the literature are necessary. We have recently shown that current definitions of species and genera in particular are not entirely satisfactory, according to a measure of genome sequence similarity called genome conservation (Kunin *et al*, 2005). It is thus imperative that sufficient literature coverage, which reflects the active experimental interest of research communities, is available for a given organism before a genome sequencing project is initiated. Certain communities might not successfully advocate the sequencing of their favourite species, despite its relative importance in terms of available abstracts in Medline, whereas closely related species with less literature support may be sequenced.

To exemplify this point, we have analysed the group of *Streptomyces* species that have been or are in the process of being sequenced, according to COGENT and the Genomes OnLine Database (GOLD; Bernal *et al*, 2001) respectively. The *Streptomyces* genus is a complex and important group of Actinobacteria, with many unresolved branches and a variety of phenotypic attributes (Anderson & Wellington, 2001). Two species whose genomes have been sequenced, *S. coelicolor* and *S. avermitilis*, are represented in Medline by 1014 and 119 abstracts, respectively, corresponding to SKI values of 0.13 and 0.014. According to GOLD, five other *Streptomyces* species are currently being sequenced: *S. ambofaciens* (116 abstracts), *S. diversa* (no abstracts), *S. noursei* (81 abstracts),

S. peucetius (87 abstracts) and *S. scabies* (60 abstracts). It is therefore surprising to find that *S. aureofaciens* (365 abstracts), *S. antibioticus* (277 abstracts) and *S. griseus* (1223 abstracts) are not listed in GOLD as ongoing projects, although all three strains are representative species of key *Streptomyces* groups (Anderson & Wellington, 2001).

Obviously, the criterion of published abstracts alone is not sufficient to prioritize genome sequencing targets, yet it provides a rational measure of current knowledge and interest by taking into account the number of published studies. It is conceivable that a more elaborate listing of all species and/or strains could be obtained and ranked by the corresponding number of published abstracts. These corpora could serve as focal points for the experimental communities and would facilitate the identification of 'neglected' organisms that might be considered for genome sequencing in the future. Other uses could obviously benefit computational analysis, including database annotation and text mining.

We believe that the SKI measure demonstrates the significant variation in the number of publications for each genome and the huge challenge of using this literature to accurately annotate these genome sequences. Simply put, we would need to achieve more than 26,000 publications for *Mycoplasma genitalium*—which has 479 genes and is currently covered by about 400 abstracts—to reach the current SKI value for *E. coli*. We have a long way to go.

REFERENCES

- Anderson AS, Wellington EM (2001) The taxonomy of *Streptomyces* and related genera. *Int J Syst Evol Microbiol* **51**: 797–814
 Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29**: 126–127

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. *Nature* **405**: 823–826

Janssen P *et al* (2003a) Beyond 100 genomes. *Genome Biol* **4**: 402

Janssen P, Enright AJ, Audit B, Cases I, Goldovsky L, Harte N, Kunin V, Ouzounis CA (2003b) Complete GENome Tracking (COGENT): a flexible data environment for computational genomics. *Bioinformatics* **19**: 1451–1452

Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA (2005) Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Res* **33**: 616–621

Ouzounis CA, Karp PD (2002) The past, present and future of genome-wide re-annotation. *Genome Biol* **3**: COMMENT2001

Paul Janssen (top left) is at the Belgian Nuclear Research Centre in Mol, Belgium. Leon Goldovsky (top right), Victor Kunin (middle left), Nikos Darzentas (middle right) and Christos A. Ouzounis (bottom) are at the European Bioinformatics Institute, Hinxton, UK.
 E-mail: ouzounis@ebi.ac.uk
 doi:10.1038/sj.embor.7400412