environmental microbiology

Environmental Microbiology (2015) **17**(1), 4–9 doi:10.1111/1462-2920.12622

Genomics update

A Proteome Quality Index

Jan Zaucha,[1,2]* Jonathan Stahlhacke,[1] Matt E. Oates,[1] Natalie Thurlby,[1,2] Owen J. L. Rackham,[3] Hai Fang,[1] Ben Smithers[1] and Julian Gough[1]

[1]*Department of Computer Science and* [2]*Bristol Centre for Complexity Sciences, University of Bristol, Bristol, UK.*

[3]*Medical Research Council Clinical Sciences Centre, Faculty of Medicine, Hammersmith Hospital, Imperial College London, London, UK.*

Summary

We present the Proteome Quality Index (PQI; http://pqi-list.org), a much-needed resource for users of bacterial and eukaryotic proteomes. Completely sequenced genomes for which there is an available set of protein sequences (the proteome) are given a one-to five-star rating supported by 11 different metrics of quality. The database indexes over 3000 proteomes at the time of writing and is provided via a website for browsing, filtering and downloading. Previous to this work, there was no systematic way to account for the large variability in quality of the thousands of proteomes, and this is likely to have profoundly influenced the outcome of many published studies, in particular large-scale comparative analyses. The lack of a measure of proteome quality is likely due to the difficulty in producing one, a problem that we have approached by integrating multiple metrics. The continued development and improvement of the index will require the contribution of additional metrics by us and by others; the PQI provides a useful point of reference for the scientific community, but it is only the first step towards a 'standard' for the field.

Introduction

There is a strong need in the scientific community for ways to quantify the quality of protein sequence datasets deduced from the sequenced genomes. This need arises because there is an enormous variability in the quality and consistency of proteomes, both in terms of the individual sequences of each protein and in terms of the completeness of the protein collection and how representative it is of the proteins in the complete genome (Chothia and Gough, 2009). In other fields, such as nucleic acid sequencing, 3D protein structure determination or collection of gene expression data, there have been community-wide agreements settled among journals, data repositories [e.g. the International Nucleotide Sequence Database Collaboration (Nakamura *et al.*, 2013), Protein Data Bank (Rose *et al.*, 2013) or Gene Expression Omnibus (Barrett *et al.*, 2013)], funding bodies and scientists. These agreements have been crucial to the advancement of the field. To the detriment of the field of protein sequence analysis, in particular comparative genomics, there are currently no clear standards to guide the publishing and depositing in databases of complete proteomes. This is largely because of a lack of metrics by which the quality of a 'complete' proteome can be systematically assessed.

The purpose of introducing a Proteome Quality Index (PQI) is twofold. First, by providing systematic metrics, a protein quality index will help journals and databases to impose standards on the producers of the proteomes, and in turn will provide signposts to help guide the work of data production. Second, for users of complete proteomes, a protein quality index will provide a way to select data with an appropriate trade-off between genome coverage and completeness of the proteome for each particular study. It will also aid in interpreting results (e.g. by highlighting the potential for a result to be caused by a dataset artefact).

The causes of the growing number of poor-quality proteomes are several (including the higher reward for publishing first over publishing good quality). However, we should not neglect to note that *de novo* assembly of a genome is still a very challenging task and that generating a complete proteome is often one of the last steps in a long pipeline. It should also be noted that the production of a high-quality proteome dataset may not be even an objective of a project, and the authors are often not claiming that their proteome is of high quality. At present, this information is not evident to scientists accessing the data,

*For correspondence. E-mail Jan.Zaucha@bristol.ac.uk; Tel. +44(0)117 394 1212; Fax +44(0)117 95 45208. All authors contributed equally to the work.

especially when the proteome is part of a larger repository including many proteomes, all of varying quality. The NCBI's RefSeq microbial genome database has recently instituted certain quality controls for the submitted genomes, including the number of frameshifts and the presence of complete rRNAs and essential conserved proteins (Tatusova *et al*., 2014). The consequences of ignoring proteome quality in multi-organism studies can be severe and easily lead to incorrect conclusions. In a recent study, we used existing proteomes to resolve the tree of sequenced life (sTOL) (Fang *et al*., 2013). The study demonstrated some discrepancies between the classification of species from molecular characters (complete proteome repertoire) and from taxonomy (the *status quo* in the literature). In attempting to understand the differences, we observed that many assembled proteomes are lacking essential housekeeping genes, contain errors due to inadequate assembly or have been built too closely on another genome. A similar observation of proteome incompleteness has also been reported in a recent analysis of all complete *Escherichia coli* genomes (Cook and Ussery, 2013). As a result, it should be anticipated that the results of other published research, particularly in the field of comparative genomics, could have been affected by the same problems and might need to be revisited.

The principal reason that there is not already a PQI is that there is no single metric that could objectively evaluate the quality of a proteome. At the moment, the best judge would be an experienced scientist who can consider all relevant factors on a project-by-project basis. But with over 24 000 genome projects currently pending completion (Pagani *et al*., 2012), automated computational metrics are the only feasible option, particularly for evaluating the proteomes of less studied species. Coming up with a measure for proteome quality will require a joint effort from the scientific community. Here we propose a concrete starting point: the PQI database (http://pqi-list.org), which is largely based on our SUPERFAMILY database (de Lima Morais *et al*., 2011). The PQI provides a minimum starting point from which the future – much needed – measure(s) can eventually emerge. Although we anticipate much criticism of the technical details of the PQI, we would like to point out that this is just the first quality measure and the one that satisfies the most basic requirements for utility. We also encourage and invite active criticism from the community because it will drive the addition of more and improved metrics to the measure, hopefully from multiple providers.

Methods

Automatic scoring pipeline implemented in PQI

All proteomes coming from completely sequenced genomes are automatically loaded into the PQI resource for assessment. At the time of this writing, the resource contains proteomes for 1707 species (comprising 1156 bacteria, 122 archaea and 429 eukaryotes; including all available strains for each species). There is a form provided to submit proteomes to be added to PQI.

List of metrics implemented

Below is the list of automated methods currently implemented in the PQI resource. Note that metrics that compare a proteome to its local phylogenetic clade are labelled with a clade flag (clade based). For these methods, the 'metric' is the modified Z-score (which approximates difference from median in standard deviation units) as compared with the local phylogenetic clade. For non-clade based methods, the 'metric' is the same value as the 'raw score'.

X content. The score is the percentage of amino acids in all proteins for this proteome that are undefined (i.e. represented by an 'X' in the sequence). The first residue of the protein is excluded from the statistics because there is a high bias for it to be uncertain, even in the highest quality proteomes, due to uncertain translation start sites.

PubMed publication count. The raw score is total number of publications related to the genome as listed for that entry in the PubMed database (NCBI Resource Coordinators, 2014).

Number of domain superfamilies (clade based). The SUPERFAMILY database (de Lima Morais *et al*., 2011) provides protein domain assignments at the structural classification of proteins (SCOP) version 1.75 (Andreeva *et al*., 2008) superfamily level using hidden Markov models (HMMs) (Gough *et al*., 2001). This measure compares the number of proteins assigned by SUPERFAMILY to domain superfamilies with the average one for that clade.

Number of domain families (clade based). The raw score is the number of distinct SCOP protein domain families that are annotated to the proteome using a hybrid HMM/pairwise similarity method from the SUPERFAMILY resource (Gough, 2006) compared with the average one for that clade.

Per cent of sequences covered (clade based). The raw score is the percentage of the proteome sequence (in amino acid residues) that is covered by SCOP domain superfamily assignments.

Core Eukaryotic Gene domain architecture inclusion. This method checks for domain-architecture similarity to the Core Eukaryotic Gene (CEG) library used by

the CEGMA tool (Parra *et al*., 2007), originally based on the Eukaryotic Orthologous Groups (KOG) database (Koonin *et al*., 2004). The SUPERFAMILY HMM library is scored against all instances of the KOG entries found in the CEG set to obtain domain assignments. The raw score is the proportion of the total CEG set domain architectures found in the proteome's unique annotations.

Mean sequence length (clade based). The raw score is the mean number of amino acids in all proteins from the given proteome.

Mean hit length (clade based). The raw score is the mean number of amino acids in the superfamily assignments of the proteome.

Number of domain architectures (clade based). A 'domain architecture' is an assignment of a protein to a sequential order of SCOP protein domain superfamilies and gaps by the SUPERFAMILY resource. The raw score is the number of the unique domain architectures of the proteome.

Per cent of sequences with an assignment (clade based). The raw score is the percentage of proteins in the proteome that have a SCOP superfamily assignment according to SUPERFAMILY.

Per cent of sequences in UniProt. The raw score is the percentage of sequences in the proteome that appear in the UniProt database with 100% sequence identity (The UniProt Consortium, 2014).

Defining the local phylogenetic clade for clade-based metrics

Metrics that compare characteristics of proteomes to others (e.g. average sequence length) can indicate outliers and hence suggest a possible systematic error in the creation of the proteome set (e.g. fragmentary assembly). Because there can be a great variation of these characteristics across the tree of life, the comparison should be done locally among similar organisms. This requires a procedure for the selection of a local phylogenetic clade.

The species tree and branch lengths are taken from the sTOL (Fang *et al*., 2013). An organism's local clade is defined as all common descendants of the most recent ancestor satisfying the following requirements: (i) the clade includes at least 10 distinct species, and (ii) the branch length to the parent node is at least 0.01 (ensuring enough variation to compare against in the case of many closely related species). Here, the branch length serves as a weighting scheme ensuring that the representation of each species is normalized with respect to its phylogenetic placement, as described by Gerstein and colleagues (1994); see Appendix S1 for details. Note that the phylogenetic clade is obtained for each proteome independently. Thus, even closely related species whose clades partly overlap may still be compared against a slightly different background distribution. It is rarely necessary to go to the root of a kingdom to satisfy the above criteria, but e.g. for unique representatives of new phyla, it is possible that the clade used for comparison may be very broad.

Five-star rating

All scores are mapped to a human-readable one- to five-star rating. For this purpose, we developed a single universal function to map all metrics, independent of distribution, to a star rating – see Appendix S2 for details.

In practice the five-star rating assigns a high score to proteomes that do not stand out significantly from the median within their phylogenetic clade (clade-based methods) and show no alarming features (as shown by the other metrics).

Results

PQI web resource

At the time of this writing, the PQI web resource contains 3220 annotated proteomes from all major providers including NCBI (Geer *et al*., 2010), Ensembl (Flicek *et al*., 2014) and many others. It contains 11 automatically generated metrics (Table 1). In addition to the value of the metric, a simple five-star rating is provided for each, and combined to give the overall 'PQI score' for each proteome. Also shown alongside the PQI score is a separate 'user score', designed for developer and user voting. This voting score is especially useful for highlighting proteomes where the automatically generated PQI score should be treated with caution. Admittedly, there are exceptional organisms where the PQI score is not expected to perform as well as other cases, and this provides a mechanism by which they can be flagged.

The website has the facility for users to upload their own metrics and publish them in the PQI. All uploaded metrics from external providers will be optionally displayed alongside the core PQI metrics; it is our expectation that suitable metrics will be migrated to the core set and become part of the overall PQI score after consultation with the providers. As stated above, this first release of the PQI is a starting point for the development of a reliable standard by the community, not the final word.

Table 1. An overview of the automatic metrics implemented in PQI.

Method	Description	Number of proteomes	Number with ≤ 2★	Number with ≥ 4★
0	Overall PQI score	3270	40	521
1	X content	3270	80	2971
2	PubMed publication count	2752	861	11
3	Number of superfamilies[a]	3270	282	1240
4	Number of families[a]	3270	371	1204
5	Per cent of sequence covered[a]	3270	563	1136
6	CEG domain architecture inclusion	498	129	76
7	Mean sequence length[a]	3270	596	1152
8	Mean hit length[a]	3270	448	1243
9	Number of domain architectures[a]	3270	417	1183
10	Percent of sequences with a SUPERFAMILY assignment[a]	3270	584	1164
11	Percent of sequences in UniProt	3270	216	2992

a. Clade-based methods.

Proteome scores

About a third of all proteomes were awarded at least four stars, giving no indication of serious problems. The majority of proteomes (about two thirds) have three stars. Only 22 proteomes fail to attain two stars. These proteomes are not all necessarily incorrect, but they stand out in most metrics and should be examined carefully before being used in any analysis. Most individual metrics award more than half of the proteomes at least four stars, and about 10% of proteomes are found to be outliers with less than two stars for that specific metric. The only metrics for which most proteomes fail to achieve a high score are the PubMed publication count and CEG domain-architecture inclusion.

Figure 1 shows an example of a proteome with a PQI rating of more than four stars. *Neosartorya fischeri* has been sequenced by. Wortman and colleagues(2006) in the hope of gaining insight into a pathogen of the same family, *Aspergillus fumigatus*, which is the primary cause of invasive aspergillosis. This proteome has no alarming characteristics as seen in the PQI metrics.

Figure 2 shows another example of a proteome, but with a PQI rating of less than two stars. The proteome of *Lactobacillus fermentum* CECT 5716 seems to be the result of a quick sequencing effort (Jiménez *et al.*, 2010). The authors used the ultrafast 454 pyrosequencing technology, which was later shown to have a high systematic error rate and requires error correction post processing with the standard GS-FLX software (Gilles *et al.*, 2011). Compared with other species of the *Lactobacillus* genus (the proteome's local phylogenetic clade as defined by PQI), this proteome scores very poorly in all clade-based metrics, which suggests a lack of completeness. In the publication, the authors admit the genome was assembled on the scaffold from *L. fermentum* IFO 3956 (Morita *et al.*, 2008). Consequently, the proteome has a very low number of domain-architectures not seen in *L. fermentum* IFO 3956.

Discussion

We present the PQI, a much-needed resource for proteome quality metrics available for a comprehensive database of downloadable proteomes. Measuring proteome quality is difficult, and this is perhaps why there did not previously exist such a resource despite the clear

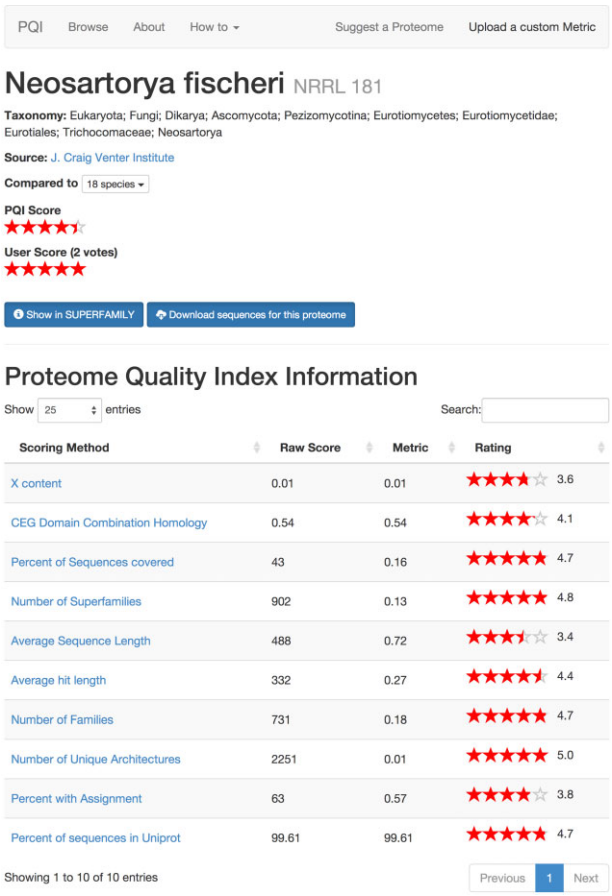

Fig. 1. The PQI score page for the proteome of *Neosartorya fischeri*.

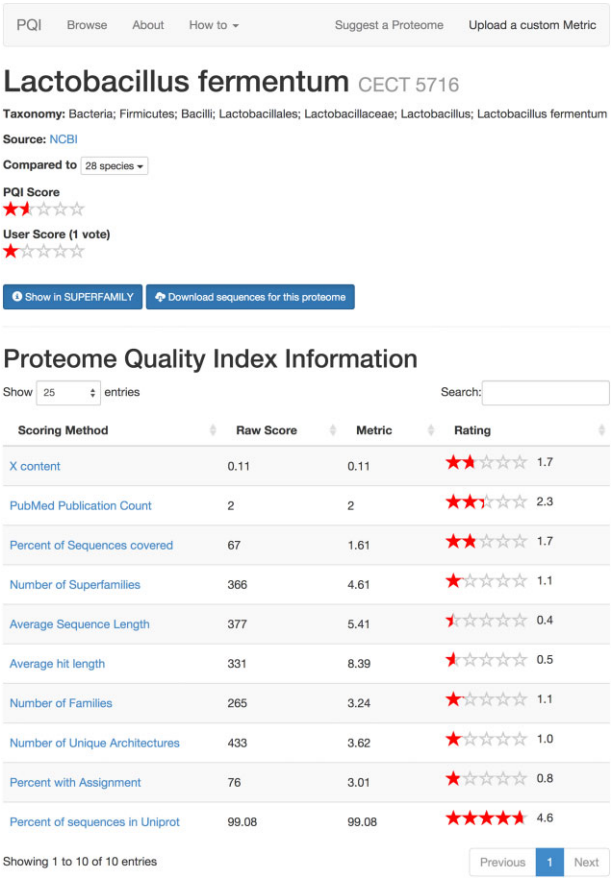

Fig. 2. The PQI score page for the proteome of *Lactobacillus fermentum* CECT 5716.

value and demand for it. Now that the PQI has been released, there is such a resource, which we believe is significantly better than nothing. We aim to improve the PQI over time, ideally with contributions from other groups. Because it is difficult to quantify proteome quality, multiple metrics need to be combined; we have provided 11 metrics as a starting point, but crucially we invite (and provide the facility for) anybody to upload and publish their own metrics on the website for all users to see.

We would like to stress that the primary objective of this tool is for scientists to be aware of the enormous variation in proteome quality and to provide a means for taking that into account in their research. We fully accept that the authors of some proteomes have made no claim as to their quality and have provided something nonetheless of great value. Sequencing groups should not be discouraged from sharing proteome data of all levels of quality.

The PQI has known limitations with the current metrics. By adding and improving metrics over the course of the continued development of the resource, it will become more reliable; however, because of the complex nature of organisms, it will probably never be possible to rate proteomes without at least some human interpretation. The most obvious limitation of the PQI is that organisms may be outliers because of a bias in what has been studied [e.g. human (Venter *et al.*, 2001)] because of some natural peculiarity in the genome [e.g. parasitic organisms such as *Guillardia theta* (Douglas and Penny, 1999), which have undergone drastic transcriptional reduction] or because they are an evolutionary outlier in the present sampling of organisms. Also, proteomes for which there is less equivalent comparative data are harder to assess so we include the facility for user scores and comments to flag such cases.

Despite accepted limitations of the quality metrics implemented in PQI, the development of proteome quality assurance is timely, important and should be pursued. Already, the PQI can be used as an aid to selection of proteome sets for large-scale comparative studies and provides a point of reference for editors, referees and data producers.

Finally, we extend an open invitation for others to engage in the production of metrics; in particular, we are aware of the need for metrics addressing: amino acid bias, sequencing depth, technology used and contamination detection (Kumar *et al.*, 2013; Pible *et al.*, 2014).

Acknowledgement

We are grateful to the Engineering and Physical Sciences Research Council (grant EP/I013717/1) for funding.

References

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* **36:** D419–D425.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., *et al.* (2013) NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res* **41:** D991–D995.

Chothia, C., and Gough, J. (2009) Genomic and structural aspects of protein evolution. *Biochem J* **419:** 15–28.

Cook, H., and Ussery, D.W. (2013) Sigma factors in a thousand *E. coli* genomes. *Environ Microbiol* **15:** 3121–3129.

Douglas, S.E., and Penny, S.L. (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J Mol Evol* **48:** 236–244.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis*. New York, The United States of America: Cambridge University Press.

Fang, H., Oates, M.E., Pethica, R.B., Greenwood, J.M., Sardar, A.J., Rackham, O.J.L., *et al.* (2013) A daily-updated tree of (sequenced) life as a reference for genome research. *Sci Rep* **3**.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., *et al*. (2014) Ensembl 2014. *Nucleic Acids Res* **42:** D749–D755.

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., *et al*. (2010) The NCBI BioSystems database. *Nucleic Acids Res* **38:** D492–D496.

Gerstein, M., Sonnhammer, E.L., and Chothia, C. (1994) Volume changes in protein evolution. *J Mol Biol* **236:** 1067–1078.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.-F. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12:** 245.

Gough, J. (2006) Genomic scale sub-family assignment of protein domains. *Nucleic Acids Res* **34:** 3625–3633.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313:** 903–919.

Jiménez, E., Langa, S., Martín, V., Arroyo, R., Martín, R., Fernández, L., and Rodríguez, J.M. (2010) Complete genome sequence of *Lactobacillus fermentum* CECT 5716, a probiotic strain isolated from human milk. *J Bacteriol* **192:** 4800.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5:** R7.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* **4:** 237.

de Lima Morais, D.A., Fang, H., Rackham, O.J.L., Wilson, D., Pethica, R., Chothia, C., and Gough, J. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res* **39:** D427–D434.

Morita, H., Toh, H., Fukuda, S., Horikawa, H., Oshima, K., Suzuki, T., *et al*. (2008) Comparative genome analysis of *Lactobacillus reuteri* and *Lactobacillus fermentum* reveal a genomic island for reuterin and cobalamin production. *DNA Res* **15:** 151–161.

Nakamura, Y., Cochrane, G., and Karsch-Mizrachi, I. (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* **41:** D21–D24.

NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **42:** D7–D17.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M., Smirnova, T., Nosrat, B., *et al*. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40:** D571–D579.

Parra, G., Bradnam, K., and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23:** 1061–1067.

Pible, O., Hartmann, E.M., Imbert, G., and Armengaud, J. (2014) The importance of recognizing and reporting sequence database contamination for proteomics. *EuPA Open Proteom* **3:** 246–249.

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., *et al*. (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* **41:** D475–D482.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., and Tolstoy, I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* **42:** D553–D559.

The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42:** D191–D198.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., *et al*. (2001) The sequence of the human genome. *Science* **291:** 1304–1351.

Wortman, J.R., Fedorova, N., Crabtree, J., Joardar, V., Maiti, R., Haas, B.J., *et al*. (2006) Whole genome comparison of the *A. fumigatus* family. *Med Mycol* **44:** S3–S7.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Shows the distribution of PQI star ratings for the X content method. The red line represents the star rating; the purple line represents the corresponding X content for the proteome. For this metric, the majority of proteomes are awarded a perfect five-star rating because they have 0% of X content. A discontinuity in the star-rating curve occurs at the position of the first proteome that has non-zero X content.

Fig. S2. Shows the distribution of PQI star ratings for the number of superfamilies method. The red line represents the star rating; the purple line represents the corresponding modified Z-score for the proteome within its clade. For this input distribution the universal mapping function yields a continuous star-rating curve.

Score distributions for each scoring method can be viewed on the PQI website on each metric page (note that the graph displays the score distribution for the proteomes loaded on the page, thus to view the full distribution please select 'show all entries' under 'Per Genome Scores').

Appendix S1. The weighting scheme for clade-based metrics.

Appendix S2. Mapping scores to a five-star rating.