The importance of recognizing and reporting sequence database contamination for proteomics

Olivier Pible, Erica M. Hartmann, Gilles Imbert, Jean Armengaud*

CEA, DSV, IBEB, Lab Biochim System Perturb, F-30207 Bagnols-sur-Cèze, France

ARTICLE INFO

Article history:
Available online 16 April 2014

Keywords:
Database
Proteomics
Metaproteomics
Contamination
Blast analysis
Curation

ABSTRACT

Advances in genome sequencing have made proteomic experiments more successful than ever. However, not all entries in a sequence database are of equal quality. Genome sequences are contaminated more frequently than is admitted. Contamination impacts homology-based proteomic, proteogenomic, and metaproteomic results. We highlight two examples in the National Center for Biotechnology Information non-redundant database (NCBInt) that are likely contaminated: the bacterium Enterococcus gallinarum EGD-AAK12 and the insect Ceratitis capitata. We hope to incite users of this and other databases to critically evaluate submitted sequences and to contribute to the overall quality of the database by signaling potential errors when possible.

© 2014 The Authors. Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

1. Opinion paper

Genome sequencing has exploded, and the wealth of information generated from genome sequences completed and annotated over the last decade has allowed the reconstruction of individual genomes for non-cultivable organisms from metagenomic data, the determination of community diversity and structure from environmental samples [1] and the exponential increase in protein identifications in proteomic and metaproteomic studies [2,3]. The rapid progress in both genomics and proteomics has been lauded and the technical challenges raised, but the time has come to talk about the proverbial elephant: our sequences are contaminated. The objective of this paper is to highlight the need to understand how genomic data are obtained, compiled and archived, and the importance of curated data repositories for proteomics of non-model organisms and metaproteomic surveys.

The presence of contaminant reads, contigs originating from DNA extraneous to the organism of interest, comes as no surprise, especially to those who perform the sequencing. The Department of Energy Joint Genome Institute, a popular resource behind many genomes, issues the following warning to users of its Microbial Single Cell Program: “Despite our best efforts, it is likely that there are some contigs in your single cell genome(s) that are from contaminant organisms. Common contaminants that are known to be in the reagents we purchase are Delftia, Pseudomonas, and Ralstonia. Other contaminants that we commonly see are Propionibacterium and Lactobacillus. In addition, there may be contaminants from your particular sample in the form of free DNA that made it into the well along with your single cell. Although we do an automated screen of your data for the known common contaminants and this information is provided to you in the JGI Single-cell Assembly QC report, this data is not removed because this could result in the removal of legitimate,
highly conserved genes from your genome” [4]. Contamination has many sources beyond consumables and reagents at the sequencing facility, even when the best laboratory practices are used. Contamination can occur at any point during sample preparation and DNA extraction [5], and even the supposedly pure cultures and individual organisms used to generate the source DNA can prove to be co-cultures or to have symbiotic partners whose DNA finds its way into sequencing reactions [6]. For example, the discovery and clinical detection of a novel parvovirus-like hybrid virus was later demonstrated to have come from infected diatoms used to produce the silica in DNA extraction spin columns [5], and quantitative PCR revealed that a putatively pure culture of the current-producing bacterium Geobacter sulfurreducens DL1 was in fact a co-culture of two G. sulfurreducens strains, despite the use of standard microbiological techniques to maintain pure cultures [6].

When contaminant sequences are not removed, they confound taxonomic identification in metagenomic samples using techniques such as fragment recruitment [7] and they translate to improperly classified protein sequences in databases such as the widely used non-redundant (nr) protein database maintained by the National Center for Biotechnology Information (NCBI). Genome sequence quality control measures exist, but their use is not universal. Standard parameters for sequence quality control include GC content, the expected genome length, the number of reads, and the N50 [8]. Additional algorithms, such as Kontaminant, can look for overlap between newly generated sequences and specified reference contaminants [9]. Thus, a myriad of criteria and automated programs exist to assist in the proof-reading of genome sequences, but their use is neither mandatory nor consistent. The burden of quality control for sequences in public repositories is generally placed upon those submitting the sequences, but information regarding what measures have been taken, if any, is not transmitted to database end-users. The NCBI Reference Sequence (RefSeq) collection is a manually curated subset of the NCBI nr database; however much of the additional standards are focused on genome annotation (see inset), and again a range of tools exist [10]. Furthermore, the status of microbial genomes in the RefSeq database is typically “provisional,” meaning, “The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff” [10]. The Universal Protein Resource Knowledgebase (UniProtKB) is similarly based on nucleotide submissions to databases like the NCBI’s GenBank [11]. Like RefSeq, entries in the SwissProt collection of UniProt, as opposed to TrEMBL, which is not curated, are further subject to extensive manual curation, again generally focused on annotation.

Results of the 2010 NCBI workshop on genome annotation [12]

1. MINIMAL GENOME ANNOTATION SHOULD HAVE
   a. rRNAs (5S, 16S, 23S) and corresponding genes with locus_tags,
   b. tRNAs and corresponding genes with locus_tags,
   c. protein-coding genes with locus_tags (see below) and corresponding CDS.

2. ANNOTATION SHOULD FOLLOW INSDC SUBMISSION GUIDELINES.

Annotation standards should follow INSDC submission guidelines (GenBank/EMBL/DDBJ) part of which were documented as part of the workshop.

a. prior to genome submission a submitted Bioproject record with a registered locus_tag prefix is required according to accepted guidelines http://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf,
b. the genome submission should be valid according to feature table documentation http://insdc.org/documents/feature_table.html.

3. METHODOLOGIES AND SOPS (STANDARD OPERATING PROCEDURES).

Genomes should be linked to the SOPs used to create the annotation and with the evidence used to create annotation.

4. EXCEPTIONS.

Exceptions (unusual annotations, annotations not within expected ranges) should be documented and strongly supporting (experimental) evidence should be provided.

5. PSEUDGENES.A proposal for pseudogenes was submitted to INSDC and is being examined.

6. ANNOTATION ABOVE THE MINIMUM

Additional (enriched) annotation should follow the above procedures (small RNAs, promoter features, operons, CRISPR sequences), INSDC guidelines, and be documented in linked SOPs.

7. CATALOG OF REPUTABLE SOFTWARE AND ANNOTATION SOURCES AND DATABASES.

A non-exhaustive catalog of reliable software, sources, and databases for the production of microbial genome annotation is both a useful community resource that aids in producing high quality genome annotation. A catalog will be reported here.

8. VALIDATION CHECKS AND ANNOTATION MEASURES.

Validation checks should be done prior to the submission. NCBI has already provided numerous tools to validate and ensure correctness of annotation. Additional checks will be put in place to ensure the minimal standards are met.

With all of the various quality control standards and methods for meeting them, it is difficult for end-users to evaluate the trustworthiness of database entries, so by default most sequences are given the benefit of the doubt. In the absence of proof of error, most sequences are assumed to be correctly assembled and annotated and attributed to the correct taxonomic group. However, during the course of an analysis, it sometimes becomes obvious that something is amiss. Notable recent cases include the mis-annotation of a keratin-derived sequence as a plant protein [13] and the mistaken inclusion of bacterial sequences in the genomic sequences of two Caenorhabditis species [14]. A string of such incidents inspired us to perform a cursory survey of the NCBI nr database, revealing a glimpse at the extent to which contamination is present. We present the following BLAST alignments as evidence of sequence contamination.

The first taxon that we identified as suspicious was the bacterium Enterococcus gallinarum EGD-AAK12 (taxid: 1357296), which supposedly contains 12,518 genes, 12,300 of which are annotated as coding sequences (CDS). This 11 Mbp genome of this Firmicute of the order Lactobacillale
is in the “scaffolds or contigs” status, with 6194 contigs and a contig N50 at 3468 nt. Its proteome is available in the NCBI database but is not included in the RefSeq protein database. Because of the discrepancy between the genome size of this strain and those of other Enterococcus bacteria (generally approximately 3 Mbp), we examined this organism using BLAST queries. We created a proteome fasta file (ORGfasta) using an http request in the form http://www.ncbi.nlm.nih.gov/protein/?term=taxid1357296 [Organism:noexp] (for taxid: 1357296) and post-processed using a Python script to only retain non-redundant sequences. To identify which sequences could be related to a contamination of the proteome by a different organism, the ORGfasta proteome was processed as follows: (i) a first BLASTp search was performed using a small subset of ORGfasta against the NCBI database at an evalue threshold of 1E−20 to identify the contaminating genus (CONTgenus), (ii) a list of GIs associated to CONTgenus children taxa was compiled using the NCBI taxonomy gi_taxid_prot.dmp file, (iii) fasta entries corresponding to this GI list were retrieved from NCBI to build a BLAST database (CONTgenus BLASTdb), (iv) a BLASTp search of the ORGfasta proteome against the CONTgenus BLASTdb was performed, and (v) the output BLAST xml was parsed using a Python script to list ORGfasta sequences with hits at an evalue below 1E−20 for sequences of the exact same length on any CONTgenus sequence. A reference proteome (REFfasta) created using a strain of the same genus whose genome is complete, Enterococcus faecium DO, was submitted to the same BLAST search against the CONTgenus BLASTdb. An Enterococcus BLAST database excluding both the ORGfasta and REFfasta sequences was built, and the ORGfasta and REFfasta sequences were searched against this database, again using BLAST (Table 1). Of the 12,300 predicted E. gallinarum EGD-AAK12 proteins, 3198 were almost identical to a Klebsiella sequence (same length, BLAST evalue <1E−20), whereas for Enterococcus faecium DO, the closest “complete” reference organism, only 103 sequences met these criteria, giving an estimate for the number of false positives. Thus, at least one Klebsiella strain was likely present as a contaminant in the DNA used for sequencing. Only 244 E. gallinarum EGD-AAK12 sequences are almost identical to an Enterococcus sequence (excluding ORGfasta and E. faecium sequences), whereas E. faecium DO yielded 4122 hits. The high representation of species close to E. faecium in the Enterococcus database could explain the number of hits for E. faecium DO, but the low number of hits for E. gallinarum EGD-AAK12 raises questions. Because of the presence of Klebsiella, a Proteobacteria of the order Enterobacteriales, in the sample and because of the generation of short reads in the Illumina sequencing approach, assembly errors could lead to erroneous chimeric protein sequences in the output, reducing the number of correctly assembled coding sequences and thus reducing number of sequences shared with other enterococci.

The second dubious taxon that was selected for investigation was Ceratitis capitata (taxid: 7213), which is a eukaryote from the order Diptera, well-known for causing extensive damage to fruit crops. This organism was examined as for the first example, and the results are given in Table 2. Another well-known Diptera, namely Drosophila melanogaster (taxid: 7227, Bioproject 164), was used as the reference proteome. The C. capitata genome is in the “scaffolds or contigs” status, and its proteome is available both in the NCBI and RefSeq protein databases. We detected 789 C. capitata proteins that were almost identical to sequences from the Escherichia genus (same length, BLAST evalue <1E−20). In comparison, only nine such matches were obtained from the 21,402 non-redundant D. melanogaster sequences. As in the previous example, the number of BLAST hits on a Diptera order database excluding sequences from both proteomes is much higher for D. melanogaster than for C. capitata, which strengthens the hypothesis that taxid: 7213 is contaminated with Escherichia sequences. This contamination could

### Table 1 - Whole proteome BLAST analysis of Enterococcus gallinarum EGD-AAK12.a

| Genome sequence                          | Taxid       | Enterococcus genus (283,139)b | Klebsiella genus (contaminant) (181,226)b |
|-----------------------------------------|-------------|-------------------------------|------------------------------------------|
| Enterococcus gallinarum                 | 1357296 (12,300)b | 244                           | 3198                                     |
| EGD-AAK12 (contaminated)                |             |                               |                                          |
| Enterococcus faecium DO (not contaminated) | 333849 (3762)b | 4122                          | 103                                      |

*a* Hits (sequences of equal lengths, evalue threshold of 1E−20) against an Enterococcus genus database excluding both queried taxa and a Klebsiella genus database.

*b* Number of non-redundant sequences.

### Table 2 - Whole proteome BLAST analysis of Ceratitis capitata.a

| Genome sequence                          | Taxid             | Diptera order (245,091)b | Escherichia genus (contaminant) (671,057)b |
|-----------------------------------------|-------------------|--------------------------|------------------------------------------|
| Ceratitis capitata (contaminated)        | 7213 (18,497)b    | 882                      | 789                                      |
| Drosophila melanogaster (not contaminated) | 7227, Bioproject 164 (21,402)b | 6079 | 9                                      |

*a* Hits (sequences of equal lengths, evalue threshold of 1E−20) against a Diptera order database excluding both queried taxa and an Escherichia genus database.

*b* Number of non-redundant sequences.
Conflicts of interest

The authors declare no conflict of interests.

Transparency document

The Transparency document associated with this article can be found in the online version.

Acknowledgements

This work was supported by the Commissariat à l’Energie Atomique et aux Energies Alternatives, and the Agence Nationale de la Recherche (ANR-12-BSV6-0012-01). EMH would like to express her sincere gratitude to the Fulbright commission for initial support.

References

[1] Carr R, Shen-Orr SS, Borenstein E. Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. PLoS Comput Biol 2015;9:e1003292.
[2] Armengaud J, Hartmann EM, Bland C. Proteogenomics for environmental microbiology. Proteomics 2013;13:2731–42.
[3] Armengaud J, Trapp J, Pible O, Geflard O, Chaumot A, Hartmann EM. Non-model organisms, a species endangered by proteogenomics. J Proteomics 2014.
[4] Clingenpeel S. Single cell data decontamination; 2012.
[5] Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J Virol 2013;87:11966–77.
[6] Shrestha PM, Nevin KP, Shrestha M, Lovley DR. When is a microbial culture pure? Persistent cryptic contaminant escapes detection even with deep genome sequencing. MBio 2015;4, e00591–12.
[7] Rappe MS. Stabilizing the foundation of the house that ‘omics builds: the evolving value of cultured isolates to marine microbiology. Curr Opin Microbiol 2013;16: 618–24.
[8] Batzoglou S, Jaffe DB, Stanley K, Butler J, Nimer S, Mauceli E, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res 2002;12:177–89.
[9] Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front Genet 2013;4:288.
[10] Fruitik K, Brown G, Tatusova T, Maglott D. In: McEntyre J, Ostell J, editors. The NCBi handbook. Bethesda, MD: National Center for Biotechnology Information (US); 2012.
[11] Apweiler R, Martin MJ, O’Donovan C, Magrane M, Alam-Faruque Y, Antunes R, et al. Ongoing and future developments at the universal protein resource. Nucleic Acids Res 2011;39:D214–9.
[12] NCBI genome annotation workshop. Bethesda, MD: National Center for Biotechnology Information (US); 2010.
[13] Nawrot R, Barylski J, Schulze WX. Incorrectly annotated keratin derived peptide sequences lead to misleading MS/MS data interpretation. J Proteomics 2013;91:270–3.
[14] Percudani R. A microbial metagenome (Leucobacter sp.) in Caenorhabditis whole genome sequences. Bioinform Biol Insights 2013;7:55–72.
[15] Behar A, Yuval B, Jurkevitch E. Gut bacterial communities in the Mediterranean fruit fly (Ceratitis capitata) and their impact on host longevity. J Insect Physiol 2008;54:1377–83.
[16] Armengaud J. A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr Opin Microbiol 2009;12:292–300.
[17] Armengaud J. Microbiome and proteomics, getting the best of both worlds! Environ Microbiol 2013;15:12–23.
[18] Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, et al. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res 2009;37:D471–8.