

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

Matching curated genome databases: a non trivial task

BMC Genomics 2008, **9**:501 doi:10.1186/1471-2164-9-501

Stephane Descorps-Declere (stephane.descorps-declere@igmors.u-psud.fr)
Matthieu Barba (matthieu.barba@igmors.u-psud.fr)
Bernard Labedan (bernard.labedan@igmors.u-psud.fr)

ISSN 1471-2164

Article type Research article

Submission date 12 June 2008

Acceptance date 24 October 2008

Publication date 24 October 2008

Article URL <http://www.biomedcentral.com/1471-2164/9/501>

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

<http://www.biomedcentral.com/info/authors/>

Matching curated genome databases: a non trivial task

*Stéphane Descorps-Declère**, *Matthieu Barba** and *Bernard Labedan[§]*

Institut de Génétique et Microbiologie, Université Paris Sud XI, CNRS UMR 8621,
Bât. 400, 91405 Orsay Cedex, France

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

SDD: stephane.descorps-declere@igmors.u-psud.fr

MB: matthieu.barba@igmors.u-psud.fr

BL: bernard.labedan@igmors.u-psud.fr

Abstract

Background

Curated databases of completely sequenced genomes have been designed independently at the NCBI (RefSeq) and EBI (Genome Reviews) to cope with non-standard annotation found in the version of the sequenced genome that has been published by databanks GenBank/EMBL/DDBJ. These curation attempts were expected to review the annotations and to improve their pertinence when using them to annotate newly released genome sequences by homology to previously annotated genomes. However, we observed that such an uncoordinated effort has two unwanted consequences. First, it is not trivial to map the protein identifiers of the same sequence in both databases. Secondly, the two reannotated versions of the same genome differ at the level of their structural annotation.

Results

Here, we propose CorBank, a program devised to provide cross-referencing protein identifiers no matter what the level of identity is found between their matching sequences. Approximately 98% of the 1,983,258 amino acid sequences are matching, allowing instantaneous retrieval of their respective cross-references. CorBank further allows detecting any differences between the independently curated versions of the same genome. We found that the RefSeq and Genome Reviews versions are perfectly matching for only 50 of the 641 complete genomes we have analyzed. In all other cases there are differences occurring at the level of the coding sequence (CDS), and/or in the total number of CDS in the respective version of the same genome.

CorBank is freely accessible at <http://www.corbank.u-psud.fr>. The CorBank site contains also updated publication of the exhaustive results obtained by comparing RefSeq and Genome Reviews versions of each genome. Accordingly, this web site

allows easy search of cross-references between RefSeq, Genome Reviews, and UniProt, for either a single CDS or a whole replicon.

Conclusions

CorBank is very efficient in rapid detection of the numerous differences existing between RefSeq and Genome Reviews versions of the same curated genome.

Although such differences are acceptable as reflecting different views, we suggest that curators of both genome databases could help reducing further divergence by agreeing on a minimal dialogue and attempting to publish the point of view of the other database whenever it is technically possible.

Background

Public genomic databanks are inexorably inundated by newly sequenced genomes.

The number of complete sequence of prokaryotic genomes that are published per year has increased more than tenfold in the last seven years with a present rate close to four newly published prokaryotic genomes per week. One of the main challenges encountered by genome databanks is that complete genomic sequences are submitted with a heterogeneous and (too) often crude gene annotation [1-4]. To cope with these major problems and to improve the representation of genomic information, NCBI and EBI are proposing curated versions, the Reference Sequence (RefSeq) [5] and Genome Reviews [6], respectively. Each database team is working independently but they share the same main goal of delivering an up-to-date, standardized and comprehensive view of the completely sequenced genomes that are present in the International Nucleotide Sequence Database (INSD) repository (GenBank/EMBL/DDBJ),

To facilitate the use of these standardized genomic data in comparative genomics studies, both RefSeq and Genome Reviews include manually curated information.

Noticeably, RefSeq and Genome Reviews provide cross-references to public databases to facilitate database searches. Interestingly, many of these cross-references (/db_xref) are specific to the curated database: for instance, RefSeq has /db_xref to Entrez [7] and often to CDD [8], whereas Genome Reviews has /db_xref to Gene Ontology [9], InterPro [10], and UniProt [11], and occasionally to HOGENOM [12], and PDB [13].

Thus, it would be advantageous to work with both curated databases since they look more complementary than concurrent. However, there is no immediate way to match the respective sequence identifiers listed by either RefSeq or Genome Reviews for the same gene of the same reannotated genome, although the knowledgebase UniProt [11] began to add links to both genome databases as this paper was in preparation.

Moreover, the independent efforts of NCBI and EBI curators in improving the structural annotation of a few CDS, lead to increasingly different genomic versions of the same organism. Three different instances are expected when comparing the structural annotations made independently by RefSeq and Genome Reviews curators: (i) the amino acid sequences are exactly identical, (ii) both CDS share an overlapping identical segment but differ in length, (iii) a few CDS are found exclusively in one genome database. This last instance corresponds often to the redefinition of a putative CDS as being a pseudogene on the basis of structural features.

We aimed to obtain immediate and exhaustive cross-references of each protein-coding gene when dealing with such possible divergences that reflect different points of view between RefSeq and Genome Reviews. Accordingly, we designed CorBank, a software (see [14]) that detects not only perfect identities but also any differences between RefSeq and Genome Reviews databases.

Results

Complete sequences of each replicon of each prokaryotic organism endowed with the same Taxonomy ID in both RefSeq and Genome Reviews were downloaded from each database and mapped by their common INSD identification numbers. Then, as schematized on Fig. 1, we compared both database versions of the same genomic data to identify the cross-references for each gene and to measure their level of matching. Accordingly, the different scripts that make up the CorBank program [14] were applied to these mapped data in two successive steps in order first to find exact matches and then to identify the nature and location of any difference in imperfect matches.

Matching gene sequences in independently curated genome databases

To be as fast as possible, we did not compare the sequence partners by using efficient but slow programs such as BLASTClust [15]. Rather, we used the Perl language to build hash tables where each amino acid sequence is a key that indexes its encoding CDS. Matching is straightforward when the same key is found for the two versions of the same gene sequence - one in RefSeq and the other in Genome Reviews (Fig. 1, yellow part). In rare instances, more than two identical sequences were found for the two versions of the same genome. This occurred for example with strictly identical insertion sequences present at different locations on the analyzed genome. Moreover, we could not dismiss the hypothesis that in very very rare cases pairs of completely conserved paralogues could form bidirectional best matches that may be erroneously interpreted. To handle these problems, we further used the respective gene positions to identify the pertinent couples of corresponding sequences (Fig. 1, yellow part). Using this approach based on hash tables, we found that 98% of copies of the 1,983,258 genes described in both databases are matching, allowing instantaneous retrieval of their respective cross-references (see, for instance, Fig. 2 Table C).

However, the view was more contrasted when comparing complete genome annotation instead of looking at each individual gene. Table 1 shows that only 50 of the 641 complete genomes we have analyzed are perfectly matching at the level of their structural annotation. The other ones differ in terms of their respective total number of sequences and/or distribution of perfect matching sequences (Table 1). The copies in both curated databases of 260 genomes differ by their total numbers of genes and by a significant proportion (up to 12.5%, see below *Xanthomonas oryzae pv. oryzae KACC10331* in Table 3) of inexact matching of individual genes. The two versions of 321 species differ by their respective total numbers of genes but their corresponding CDS are matching exactly. For instance, *Bordetella petrii DSM 12804* has 5004 CDS that are matching exactly but RefSeq contains 23 CDS that are absent from Genome Reviews, whereas Genome Reviews display four additional CDS and 24 pseudo-CDS (amino acid sequence without a protein_id) that are not present in the RefSeq file. Finally, only 10 genomes have the same total numbers of genes but up to 7.4% of their corresponding genes display inexact matching. For instance, *Xanthomonas campestris pv. campestris str. 8004* displays 4273 CDS in both genomic databases but the respective amino acid sequence of the product of 310 of them differ between RefSeq and Genome Reviews. Complete data are available in Additional file 1 and on the CorBank site [14]).

Defining peculiarities of gene sequences that are partially identical between independently curated genome databases

We further studied these imperfectly matching sequences by measuring their similarity using an alignment-free approach (for a review and references inside, see [16]). Indeed, such an approach is fast and well-adapted to comparison of varying versions of the same sequence that share a significant common part. As detailed in Methods, we calculated the Euclidean distance that separates the distributions of

words of length L ($=10$) for each copy of the same gene in RefSeq and Genome Reviews, respectively (Fig. 1, blue part). This allows finding the cross-references between the respective imperfectly matching copies of the same gene (see, for instance, Fig. 2 Table D). A large variety of differences explaining these imperfect matches have been found using the CorBank program (Fig. 1, pink part). All of these differences - including the very rare ones - have been categorized as summarized in the Additional file 1 and on the page <http://www.corbank.u-psud.fr/help.html>. CorBank is able to filter any differences in any sequence locations (see, for instance, Fig. 2 Table D).

We found that the differences between matching sequences that have unequal lengths were predominantly (98.7%) located at the N-terminal part. Indeed, it is often difficult to identify the start codon, especially when several methionines are found in this N-terminal region (see, for example, [17]).

Identifying the whole differences separating independently curated copies of a genome

Scanning paired versions of the same genome with CorBank allows computing the statistics of similarities and differences between genome databases. Figs. 2 and 3 detail the results obtained with the archaeon *Pyrococcus horikoshii*. The genomes of three *Pyrococcus* have been published ten years ago: *P. horikoshii* in 1998 [18], *P. abyssi* in 1999 [19] and *P. furiosus* in 2000 [20]. Since then, these genomes, sequenced and annotated by independent groups, have been curated several times. Fig. 2 shows that many differences have accumulated between the curated versions of the *P. horikoshii* genome in RefSeq and Genome Reviews (Fig. 2 Table B). First, the respective total numbers of genes are strikingly different. Among the 1955 sequences published in RefSeq and the 2076 ones listed in Genome Reviews, only 1789 are matching. Secondly, we have only 1667 of these matches that are exact (Fig. 2 Table

C), while 122 display various differences. Fig. 2 (Table D) details a few instances of these differences in length and location of the start and end of each gene. Thirdly, Fig. 3 shows that there are a significant number of sequences putatively encoded by the *P. horikoshii* genome that are found in uniquely one genome database: 166 genes in RefSeq (Fig. 3 Table E) and 272 in Genome Reviews (Fig. 3 Table F), respectively. However, Genome Reviews classifies as pseudo-CDS a list of 15 amino acid sequences which have no protein_id. Since these pseudo-CDS are found as standard coding sequences among the 166 sequences that are specific to RefSeq (Fig. 3 Table G), we ascertained this point. CorBank was further used to match these 15 pseudo-CDS using uniquely the position information that have been kept in both databases. As a result, Fig. 3 Table B was improved in Fig. 3 Table H after matching 13 of the 15 Genome Reviews pseudo-CDS as exact and two ones as inexact. Thus, it appears that RefSeq and Genome Reviews are producing increasingly divergent views of the same genome.

Table 2 shows the same trend for the two other *Pyrococcus* species, although the divergence is less marked. Such a discrepancy is strongly diminished when looking at the genome of the related *Thermococcus kodakarensis*, belonging to the same family (Thermococcaceae), which has been published more recently (in 2004 [21]).

However, this example does not reflect a general (statistical) trend between the amount of divergences and time elapsed since the completion of sequence that would be true for all analyzed genomes (see below Tables 3a and 3b and accompanying text).

Discussion

CorBank is fulfilling two complementary goals: (1) to deliver immediate cross-references between each copy of each gene published in both RefSeq and Genome

Reviews genome databases; (2) to identify any differences between both independently curated structural annotations. The first objective is achieved almost immediately: e.g. cross-referencing the two databases versions of a 3000 CDS genome is completed in less than 1 second on a basic home computer. Exhaustive comparison of the 641 prokaryotic species present in both databases at the end of July 2008 (Genome Reviews Release 94.0, 22nd July 2008 - RefSeq Release 30, July 11, 2008) has been completed in less than 60 min. Thus, the efficiency of CorBank is largely equivalent to that of the PICR tool that is described in a paper [22] that appeared as we were writing a first version of this manuscript. PICR, a web service allowing matching a large variety of protein sequence identifiers, is restricted to 100% identity matches and cannot discriminate the correct pair when recovering more than two identical sequences since it does not exploit information about genomic locations, contrarily to Corbank. Thus, this PICR tool and a previous one, MagicMatch [23], are not as efficient as CorBank to match exhaustively genome databases. This quality is especially true of our second goal that is achieved uniquely by CorBank. Its exhaustive comparison of the species currently present in both RefSeq and Genome Reviews shows dramatic differences in the structural annotations of a large portion of their copies of the same genomes (Tables 1 to 3, Figs. 2 and 3). Of the 641 compared genomes, 581 differ in their total numbers of CDS and 270 have from 1 to 781 coding sequences per genome that differ in length.

The large majority of the 50 perfectly matching genomes correspond to newly sequenced species where the manual curation has not been started. However, there is no direct correlation between the sequencing age and the level of divergence between the lastly curated versions of the same genome as shown on Tables 3a and 3b that list the top ten database-specific organisms in both RefSeq and Genome Reviews,

respectively. Actually, a Spearman test failed to show any correlation of the different parameters computed by CorBank with the time elapsed since the completion of sequence (not shown).

Surprisingly, even the two versions of a model organism such as *Escherichia coli* K12 (substrain MG1655) that has been recently extensively reannotated in cooperative works [24,25] display significant differences. Of the 4295 gene-encoding proteins, only 4130 are matching (including 10 inexact matches), and both databases differ in their interpretation of some genes as being described as pseudo-CDS: 23 in RefSeq versus 24 in Genome Reviews. In fact, the structural identification of putative pseudogenes in *E. coli* K12 has been previously described (see [26] and references inside) but it is surprising that there is still disagreement even for these *E. coli* K12 pseudogenes.

As we were writing this paper, UniProtKB began to add /db_xref to RefSeq and Genome Reviews protein_id. However, we observed that rather often the same SwissProt file has cross-references to multiple RefSeq and Genome Reviews protein_id. This is why we think that CorBank is - presently - the only software publishing unambiguous mapping of RefSeq, Genome Reviews, and UniProt identifiers of a protein.

Conclusions

Data dependencies inherent to the annotation process by homology make genome data predestined for propagated errors [1-4]. Thus, data cleansing is a necessity for genome data after the data is produced. However, such cleansing is uneasy since it is often impossible to find the correct solution right away. Instead, there often exists a set of alternative solutions. Accordingly, RefSeq and Genome Reviews appear to have diverged in looking for correct solutions when performing credibility checking on the

INSD crude data. Credibility checking is a very important step for genome data production since the correctness of data is crucial before it is used within other processes such as annotation of newly sequenced genomes by homology to previously annotated genomes. However, such independent efforts made by both automatic and manual procedures [5,6] led to increasingly divergent reannotated data as shown in this work. Clearly, the time has come to enable curators of both genome databases to establish a minimum of dialogue. Whenever it would be technically possible, a useful compromise may be found where each database publishes the point of view of the other one. We acknowledge that such a harmonization effort looks rather complicated to be done. However, it would be very helpful for the whole community.

Methods

Comparing copies of the same genomes in curated databases

The whole genomic sequences present in RefSeq [5] and Genome Reviews [6] were downloaded at their respective FTP sites [27,28]. A first script allows matching respective downloaded files for the same genome. This script creates a mapping list between the replicons (chromosomes and plasmids) of the genome databases RefSeq and Genome Reviews. It links each respective genome identifier by using their common INSD identifier. A recognizable label, based on the association of its short name, NCBI tax_id and its INSD identifier, is associated to each matched replicon, e.g. PYRHO_70601_BA000001 for *Pyrococcus horikoshii* OT3 [17]. CorBank further compiles for each analyzed species the respective number of perfect and imperfect matches, and the sequences that are specific to a genome database as detailed below and in Fig. 1.

Detecting perfect matches between copies of the same gene in RefSeq and Genome Reviews

We built hash tables where each amino acid sequence is a key that indexes its encoding CDS (Fig. 1, yellow part). Each time the same key is found for the two versions of the same gene sequence made possible to cross-reference the respective protein identifiers in RefSeq [5], Genome Reviews [6], and UniProt [11] as shown on Fig. 2 (Table C).

Estimating similarity of partially identical sequences

In a second step (Fig. 1, blue part), CorBank is detecting all imperfect matches using an alignment-free comparison [for a review, see 16]. We used a word approach as initially proposed by Blaisdell [29] and further documented by Zharkikh and Rzhetsky [30] to measure the similarity between sequences without any alignment. The distribution of the frequency of words of length L ($L = 10$ residues) in each amino acid sequence was computed for both copies of the same gene. These L -uplets are the respective signature of the sequence. The measure of the similarity between both copies of the same sequence is based on the Euclidean distance d^E that separates them:

$$d_L^E(X, Y) = \sum_{i=1}^K (c_{L,i}^X - c_{L,i}^Y)^2$$

The vectors c_L^X and c_L^Y represent word counts for the versions X and Y of the amino acid sequences encoded by the same gene in the respective RefSeq and Genome Reviews versions and K is the number of different L -uplets possible for the L -length. These X and Y copies are expected to share a largely common part but are of unequal sizes, one copy having an extension of variable size. To exclude any bias due to too large extensions, we stated that the maximum value of the distance d that separates

two unequal copies of the same sequence could not be less than the difference between their respective numbers of amino acids.

In a third step (Fig. 1, pink part), CorBank is further analyzing all imperfect matches to define the location of the difference between both paired copies of the same gene. CorBank is first searching if the difference takes place on either the N-terminal side or the C-terminal one. In rare cases, the difference is located elsewhere, including in the common segment of both copies that could differ for only one residue. The Additional file 1 details all encountered cases, including the very rare ones.

Authors' contributions

SDD inspired using a word approach. SDD and MB developed together the CorBank program. MB set up the present CorBank website and also made in-depth analysis of the data obtained with CorBank. BL initiated the work, participated in the data analysis and wrote the draft manuscript. All authors read and finalized the whole version of the manuscript.

Acknowledgements

We thank Olivier Lespinet and Frédéric Lemoine for helpful discussions and critical reading of the manuscript and the three Reviewers for their constructive and useful comments. This work was funded by the CNRS (UMR 8621) and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS_NV_10).

References

1. Bork P, Bairoch A. **Go hunting in sequence databases but watch out for the traps.** *Trends in Genetics*, 1996, 12: 425-427
2. Brenner, SE. **Errors in genome annotation.** *Trends Genet*, 1999, 15: 132–133

3. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA. **Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications.** *EMBO Rep.* 2005, 6: 397-399
4. Ouzounis CA, Karp P.D. **The past, present and future of genome-wide re-annotation.** *Genome Biology*, 2002, 3: comment2001.1-2001.6
5. Pruitt KD, Tatusova T, Maglott DR. **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res.*, 2007, 35: 61–65
[\[http://www.ncbi.nlm.nih.gov/RefSeq/\]](http://www.ncbi.nlm.nih.gov/RefSeq/)
6. Sterk P., Kersey P.J., Apweiler R. **Genome Reviews: Standardizing Content and Representation of Information about Complete Genomes.** *OMICS*, 2006, 10: 114-118 [\[http://www.ebi.ac.uk/GenomeReviews/\]](http://www.ebi.ac.uk/GenomeReviews/)
7. Maglott D, Ostell J, Pruitt KD, Tatusova T. **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res.* 2007;35: D26-31
[\[http://www.ncbi.nlm.nih.gov/sites/gquery\]](http://www.ncbi.nlm.nih.gov/sites/gquery)
8. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki C, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. **CDD: a conserved domain database for interactive domain family analysis.** *Nucleic Acids Res.* 2007;35: D237-40
[\[http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml\]](http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)
9. The Gene Ontology Consortium. **Gene Ontology: tool for the unification of biology.** *Nature Genet.* 2000, 25: 25–29
[\[http://www.geneontology.org/index.shtml\]](http://www.geneontology.org/index.shtml)

10. Mulder NJ, Apweiler R. **The InterPro database and tools for protein domain analysis.** *Curr Protoc Bioinformatics* 2008, Chapter 2:Unit 2.7
[<http://www.ebi.ac.uk/interpro/>]
11. The UniProt Consortium. **The Universal Protein Resource (UniProt).** *Nucleic Acids Res.*, 2007, 35: D193-197 [<http://www.expasy.org/sprot/>]
12. HOGENOM [<http://pbil.univ-lyon1.fr/databases/hogenom.php>]
13. H.M. Berman, K. Henrick, H. Nakamura. **Announcing the worldwide Protein Data Bank.** *Nature Structural Biology* 2003, 10: 980
[<http://www ww pdb.org/>]
14. CorBank [<http://www.corbank.u-psud.fr/>]
15. BLASTClust [<http://www.ncbi.nlm.nih.gov/blast/docs/blastclust.html>]
16. Vinga S, Almeida J. **Alignment-free sequence comparison-a review.** *Bioinformatics*, 2003, 19: 513-523
17. D. Frishman, A. Mironov, H.-W. Mewes, and M. Gelfand. **Combining diverse evidence for gene recognition in completely sequenced bacterial genomes.** *Nucleic Acids Research*, 1998, 26: 2941–2947
18. Kawarabayasi Y., Sawada M., Horikawa H., Haikawa Y., Hino Y., Yamamoto S., Sekine M., Baba S., Kosugi H., Hosoyama A. et al. **Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3.** *DNA Res.*, 1998, 5: 55-76
19. Cohen G.N., Barbe V., Flament D., Galperin M., Heilig R., Lecompte O., Poch O., Prieur D., Querellou J., Ripp R., et al. **An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*.** *Mol. Microbiol.*, 2003, 47: 1495-1512

20. Robb F.T., Maeder D.L., Brown J.R., DiRuggiero J., Stump, M.D., Yeh R.K., Weiss R.B. and Dunn D.M. **Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology.** *Meth. Enzymol.*, 2001, 330: 134-157
21. Fukui T, Atomi H, Kanai T, Matsumi R, Fujiwara S, Imanaka T. **Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes.** *Genome Res.*, 2005, 15:352-363
22. Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R, Hermjakob H. **The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases.** *BMC Bioinformatics.*; 2007, 8: 401
[\[http://www.ebi.ac.uk/Tools/picr/\]](http://www.ebi.ac.uk/Tools/picr/)
23. Smith M., Kunin V., Goldovsky L., Enright A.J., Ouzounis C.A. **MagicMatch -- crossreferencing sequence identifiers across databases.** *Bioinformatics*, 2005, **21**, 3429-3430 [<http://cgg.ebi.ac.uk/cgi-bin/magicmatch/magic.pl>]
24. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, et al. ***Escherichia coli* K-12: a cooperatively developed annotation snapshot – 2005.** *Nucleic Acids Res.*, 2006, 34: 1–9
25. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, et al. **Multidimensional annotation of the *Escherichia coli* K-12 genome.** *Nucleic Acids Res*, 2007, doi:10.1093/nar/gkm740

26. Ochman H, Davalos LM. **The nature and dynamics of bacterial genomes.**
Science, 2006, 311: 1730-1733
27. FTP NCBI [<ftp://ftp.ncbi.nih.gov/refseq/>]
28. FTP EBI [ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews]
29. Blaisdell BE. **A measure of the similarity of sets of sequences not requiring sequence alignment.** *Proc. Natl Acad. Sci. USA*, 1986, 83: 5155–5159
30. Zharkikh AA, Rzhetsky, A. **Quick assessment of similarity of two sequences by comparison of their L-tuple frequencies.** *Biosystems* 1993, 30: 93-111

Figures

Figure 1 - The different steps of the CorBank program

The main steps of the pipeline of Perl scripts are distinguished by different colors.

The process of cross-referencing exact matching of the RefSeq and Genome Reviews versions of the same gene is indicated in yellow. The identification of inexact matches of genes that display a different structural annotation in both databases is made by the blue steps. Finally, disclosing the nature of the detected structural differences is made by the pink steps.

Figure 2 - Differentiating exact and inexact matches

A partial view of the output of the CorBank program obtained when comparing the two versions of the genome of the archeon *Pyrococcus horikoshii* OT3 is detailed in several tables. Table A recapitulates the respective database information about this species and its computed label. Table B shows a summary of the data obtained using CorBank to find what is either common to both databases or specific of each one. Table C illustrates a few instances of exact matches. Table D exemplifies a few inexact matches with detailed configuration of the difference in the structural

annotations of each copy of the same gene. The definitions of these inexact configurations are given in the Additional file 1.

Figure 3 - Differentiating exact and inexact matches, following

Table E illustrates a few instances of genes found uniquely in RefSeq. Table F exemplifies a few genes specific to Genome Reviews. Table G lists the pseudo-CDS specific to Genome Reviews. Table H re-evaluates the data presented in Table B after identifying by their positions the pseudogenes and pseudo-CDS specific to RefSeq and Genome Reviews, respectively and assessing their exactitude.

Tables

Table 1 - The reannotated copies of the same genome in independently curated databases^a are predominantly divergent

copies of the same genome sequence in both annotated databases ^a with identical number of genes	all CDS matching exactly	
	<i>NO</i>	<i>YES</i>
<i>NO</i>	260 (40.5%)	321 (50%)
<i>YES</i>	10 (1.5%)	50 (8%)

^a RefSeq (Release 30) and Genome Reviews (Release 94.0) of July 2008

Table 2 - Complete distributions of the divergences of curated databases^a in the case of closely related species

analyzed species	Comparing CDS in RefSeq Release 30 (RS) and Genome Reviews Release 94.0 (GR) databases ^a							
	<i>total number</i>		<i>matches</i>				<i>specific to</i>	
	<i>RS</i>	<i>GR</i>	<i>total</i>	<i>exact</i>	<i>inexact</i>	<i>by location</i>	<i>RS</i>	<i>GR</i>
P. horikoshii	1955	2076	1806	1680	124	0	149	270
P. furiosus	2125	2065	2065	1942	115	8	60	0
P. abyssi	1896	1786	1783	1715	68	0	113	3
T. kodakarensis	2306	2306	2306	2303	2	1	0	0

^a versions of May 2008

Table 3a - Top ten organisms having the highest number of CDS specific to RefSeq (RS) database

rank	organism	Total			matches		specific to	
		<i>RS</i>	<i>GR</i>	<i>total</i>	<i>inexact</i>	<i>by location</i>	<i>RS</i>	<i>GR</i>
RS1/GR7	Pyrococcus horikoshii OT3	1955	2076	1806	124	0	149	270
RS2	Neisseria meningitidis Z2491	2049	1991	1897	37	26	120	68
RS3/GR4	Xanthomonas oryzae pv. oryzae KACC10331	4144	4540	4030	497	2	114	510
RS4	Pyrococcus abyssi GE5	1896	1796	1783	68	0	113	3
RS5/GR6	Shewanella oneidensis MR-1	4467	4779	4364	34	1	103	415
RS6/GR8	Escherichia coli O157:H7 str. Sakai	5318	5461	5227	391	2	87	232
RS7	Deinococcus radiodurans R1	3181	1303	3099	91	1	82	4
RS8	Pyrococcus furiosus DSM 3638	2125	2065	2065	115	8	60	0
RS9	Lactococcus lactis subsp. lactis II1403	2321	2266	2263	68	0	58	3
RS10	Thermoplasma volcanium GSS1	1499	1526	1444	351	1	55	82

The organisms are sorted by their respective rank that is computed as the number of CDS that are found only in RefSeq database (Release 30). The organism names standing in the top ten list of both databases (Tables 3a and 3b) are in bold.

Table 3b - Top ten organisms having the highest number of CDS specific to Genome Reviews (GR) database

rank	organism	Total		matches			without sequence or specific to	
		<i>RS</i>	<i>GR</i>	<i>total</i>	<i>inexact</i>	<i>by location</i>	<i>RS</i>	<i>GR</i>
GR1	Mycobacterium leprae TN	1605	2723	1605	77	1	0	1118
GR2	Orientia tsutsugamushi str. Boryong (Seoul National University)	1182	2143	1182	3	0	0	961
GR3	Orientia tsutsugamushi str. Boryong (Kitasato University)	1562	2085	1562	6	0	0	523
GR4/RS3	Xanthomonas oryzae pv. oryzae KACC10331	4144	4540	4030	497	2	114	510
GR5	Acinetobacter baumannii ATCC 17978	3368	3807	3368	77	0	0	439
GR6/RS5	Shewanella oneidensis MR-1	4467	4779	4364	34	1	103	415
GR7/RS1	Pyrococcus horikoshii OT3	1955	2076	1806	124	0	149	270
GR8/RS6	Escherichia coli O157:H7 str. Sakai	5318	5461	5227	391	2	87	232
GR9	Prochlorococcus marinus subsp. pastoris str. CCMP1986	1717	1935	1714	4	2	3	221
GR10	Prochlorococcus marinus str. MIT 9312	1810	1962	1810	10	0	0	152

The organisms are sorted by their respective rank that is computed as the number of CDS that are found only in Genome Reviews database (Release 94.0). The organism names standing in the top ten list of both databases (Tables 3a and 3b) are in bold.

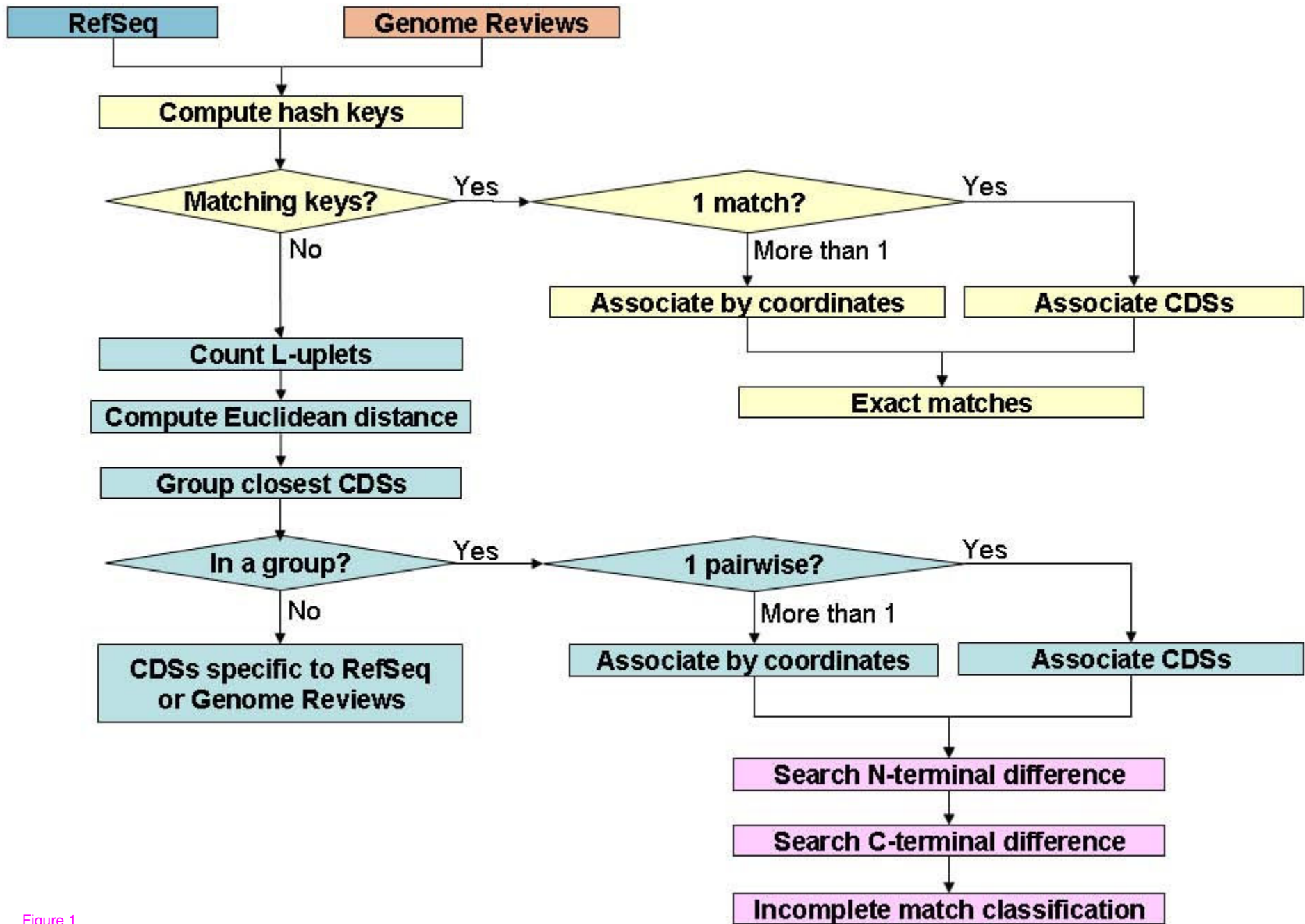


Figure 1

A

RefSeq (RS)	Genome Reviews (GR)	short name_TaxID_GenBank/EMBL/DDBJ
NC_000961.gbk	BA000001_GR.dat	PYRHO_70601_BA000001



B

total RS	total GR	matches	exact	inexact	only RS	only GR	pseudo-CDS_RS	pseudo-CDS_GR
1955	2076	1789	1667	122	166	272	0	15



C

RS ID	GR ID	Uniprot ID
NP_877946	BAA31942	O57782
NP_142024	BAA29069	O57779
NP_142025	BAA29070	O57778
NP_142027	BAA29071	O57776
NP_142028	BAA29073	O57775
NP_142029	BAA29074	O57774
NP_142030	BAA29075	O57773
NP_142031	BAA29076	O57772
NP_142032	BAA29077	O57771
NP_142033	BAA29078	O57770
NP_142034	BAA29079	O57769
NP_142035	BAA29080	O57768
NP_142036	BAA29081	O57767
NP_142037	BAA29082	O57766
NP_142038	BAA29083	O57765
NP_142040	BAA29085	O57762
NP_142041	BAA29087	O57761
NP_142042	BAA29088	O57760
NP_142043	BAA29089	O57759
NP_142044	BAA29090	O57758
NP_142045	BAA29091	O57757
NP_142046	BAA29092	O57756
NP_142047	BAA29093	O57755
NP_142048	BAA29094	O57754
NP_142051	BAA29095	O57753
NP_142052	BAA29096	O57750

D

RS ID	GR ID	Uniprot ID	Configuration	RS length	GR length
NP_142438	BAA29547	O58214	L	206	203
NP_142239	BAA29315	O57981	Ls	601	598
NP_142645	BAA29789	O58429	L	160	157
NP_142124	BAA29183	O57854	L	316	310
NP_142916	BAA30099	O58730	L	352	307
NP_142115	BAA29170	O74081	L	481	449
NP_142167	BAA29235	O57905	Ls	288	284
NP_143259	BAA30486	O50088	L	457	442
NP_142078	BAA29129	O57802	L	151	148
NP_142039	BAA29084	O57764	L	475	469
NP_142132	BAA29193	O57864	L	437	419
NP_143725	BAA31022	O59521	L	735	732
NP_143631	BAA30913	O59458	L	144	158
NP_142198	BAA29269	O57939	L	190	216
NP_143534	BAA30804	O59309	L	151	148
NP_142270	BAA29353	O58019	L	166	174
NP_142531	BAA29653	O58299	L	177	175
NP_142485	BAA29605	O58253	Ls	266	256
NP_877855	BAA30608	O59169	Ls	160	148
NP_143764	BAA31062	O59598	Ls	261	266
NP_142085	BAA29136	O57796	L	192	184
NP_143517	BAA30783	O59330	L	236	225
NP_142972	BAA30161	O74000	LRbad	52	75

E**F**

RS ID	Start	End	Locus	Product	GR ID	Uniprot ID	Start	End	Locus	Product
NP_142026 →	1,281	1,096	PH0002a	preprotein translocase subunit SecE	BAA29072 →	O57777 →	1,883	2,194	PH0004	Putative uncharacterized protein PH0004
NP_877726 →	3,258	2,965	PH0005.1n	hypothetical protein	BAA29086 →	O57763 →	15,743	16,264	PH0018	Putative uncharacterized protein PH0018
NP_142049 →	26,382	26,107	PH0026b	elongation factor 1-beta	BAA29097 →	O57752 →	28,374	28,718	PH0029	Putative uncharacterized protein PH0029
NP_142050 →	26,574	26,383	PH0026a		BAA29098 →	O57751 →	29,049	29,387	PH0030	Putative uncharacterized protein PH0030
NP_877727 →	27,507	27,884	PH0027.1n		BAA29101 →	O57748 →	30618	30941	PH0033	Putative uncharacterized protein PH0033
NP_142060 →	37,724	37,437	PH0041a	DNA-directed RNA polymerase subunit L	BAA29106 →	O57743 →	35,561	36,028	PH0038	Putative uncharacterized protein PH0038
NP_877728 →	52,311	52,562	PH0059.1n	hypothetical protein	BAA29108 →	O74077 →	37,046	37,480	PH0040	Putative uncharacterized protein PH0040
NP_877729 →	55,985	56,197	PH0065.1n	hypothetical protein	BAA29109 →	O74078 →	37,177	37,599	PH0041	Putative uncharacterized protein PH0041

B

total RS	total GR	matches	exact	inexact	only RS	only GR	pseudo-CDS_RS	pseudo-CDS_GR
1955	2076	1789	1667	122	166	272	0	15

H

total RS	total GR	matches	exact	inexact	only RS	only GR
1955	2076	1806 (15)	1680 (13)	124 (2)	149	270

protein_id	uniprot	start	end	locus	product
none	P58199	1281	1096	PH0002.1	Preprotein translocase subunit secE
none	P58748	26382	26107	PH0026.1	Elongation factor 1-beta
none	P57671	37724	37437	PH0041.1	DNA-directed RNA polymerase subunit L
none	P84142	270758	271033	PH0305.1	Acylphosphatase
none	P58189	475976	475689	PH0529.1	50S ribosomal protein L31e
none	P59734	1091926	1091675	PH1212.1	UPF0248 protein PH1212.1
none	P59937	1138710	1138958	PH1266.1	50S ribosomal protein L14e
none	P58503	1185334	1185537	PH1316.1	30S ribosomal protein S17e
none	P61124	1334400	1334200	PH1495.1	50S ribosomal protein L24e
none	P62015	1446425	1446252	PH1631.1	DNA-directed RNA polymerase subunit K
none	P58193	1545510	1545211	PH1771.1	Protein translation factor SUI1 homolog
none	P60463	1647744	1647574	PH1895.1	Preprotein translocase subunit secG
none	P58746	1661808	1662107	PH1909.1	30S ribosomal protein S24e
none	P61239	1662109	1662261	PH1910.1	30S ribosomal protein S27ae
none	P58078	1682393	1682196	PH1939.1	30S ribosomal protein S27e

G

Figure 3

Additional files provided with this submission:

Additional file 1: additional_file_1_2.pdf, 44K

<http://www.biomedcentral.com/imedia/2170953542062954/supp1.pdf>