MicroScope-an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data

David Vallenet, Eugeni Belda, Alexandra Calteau, Stéphane Cruveiller, Stefan Engelen, Aurélie Lajus, François Le Fèvre, Cyrille Longin, Damien Mornico, David Roche, et al.

D636–D647 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
doi:10.1093/nar/gks1194

MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data

David Vallenet[1,2,3,*], Eugeni Belda[1,2,3], Alexandra Calteau[1,2,3], Stéphane Cruveiller[1,2,3], Stefan Engelen[1], Aurélie Lajus[1,2,3], François Le Fèvre[1,2,3], Cyrille Longin[1,2,3], Damien Mornico[1,2,3], David Roche[1,2,3], Zoé Rouy[1,2,3], Gregory Salvignol[1,2,3], Claude Scarpelli[1], Adam Alexander Thil Smith[1,2,3], Marion Weiman[1,2,3] and Claudine Médigue[1,2,3,*]

[1]CEA, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057 Evry, France, [2]CNRS-UMR 8030, Laboratoire d'Analyse Bioinformatique pour la Génomique et le Métabolisme, 2 rue Gaston Crémieux, 91057 Evry, France and [3]UEVE, Université d'Evry, boulevard François Mitterrand, 91025 Evry, France

Received September 28, 2012; Revised and Accepted October 29, 2012

ABSTRACT

MicroScope is an integrated platform dedicated to both the methodical updating of microbial genome annotation and to comparative analysis. The resource provides data from completed and ongoing genome projects (automatic and expert annotations), together with data sources from post-genomic experiments (i.e. transcriptomics, mutant collections) allowing users to perfect and improve the understanding of gene functions. MicroScope (http://www.genoscope.cns.fr/agc/microscope) combines tools and graphical interfaces to analyse genomes and to perform the manual curation of gene annotations in a comparative context. Since its first publication in January 2006, the system (previously named MaGe for Magnifying Genomes) has been continuously extended both in terms of data content and analysis tools. The last update of MicroScope was published in 2009 in the Database journal. Today, the resource contains data for >1600 microbial genomes, of which ~300 are manually curated and maintained by biologists (1200 personal accounts today). Expert annotations are continuously gathered in the MicroScope database (~50 000 a year), contributing to the improvement of the quality of microbial genomes annotations. Improved data browsing and searching tools have been added, original tools useful in the context of expert annotation have been developed and integrated and the website has been significantly redesigned to be more user-friendly. Furthermore, in the context of the European project Microme (Framework Program 7 Collaborative Project), MicroScope is becoming a resource providing for the curation and analysis of both genomic and metabolic data. An increasing number of projects are related to the study of environmental bacterial (meta)genomes that are able to metabolize a large variety of chemical compounds that may be of high industrial interest.

INTRODUCTION

MicroScope [originally MaGe (1)] was first tailored for biologists who did not have access to proper computing infrastructure to perform efficient annotation and analysis of the bacterial genomes they had had sequenced at the French National Sequencing Center (CEA/DSV/Institut de Genomique/Genoscope) (http://www.genoscope.cns.fr). It has now become a prokaryotic annotation system widely used by the microbiologist community in France and in others countries. In contrast to organism-specific annotation systems reviewed in (2), MicroScope enables curation in a rich comparative genomic and metabolic context, the annotated genome being compared with all public complete genomes available at the NCBI RefSeq section (3). Comparison of MicroScope with other related systems is discussed in (1,4). Its primary feature, synteny map visualization (i.e. conservation of local gene order), facilitates the correct identification of orthologs in complex cases (gene duplication or fusion events, local rearrangements and translocations) and thus allows improvement of the final annotation quality. Recently,

*To whom correspondence should be addressed. Tel: +33 1 60 87 84 59; Fax: +33 1 60 87 25 14; Email: cmedigue@genoscope.cns.fr
Correspondence may also be addressed to David Vallenet. Tel: +33 1 60 87 84 53; Fax: +33 1 60 87 25 14; Email: vallenet@genoscope.cns.fr

MicroScope/MaGe has been identified as one of the most useful platform for handling complex functional relationships in the context of expert gene annotation (5).

The MicroScope platform is made of three main components (1,4): (i) a process management system that performs data updates and orchestrates the execution of bioinformatics methods as workflows; (ii) a data management system built for the Prokaryotic Genome DataBase (PkGDB) genome database together with the MicroCyc pathway database; and (iii) the MaGe web interface that allows users to explore and edit annotation data. Public genome sequence information (primary data) is imported from NCBI's RefSeq resource (3). A first automatic syntactic annotation workflow is used for gene, repeat and non-coding RNA predictions; missing genes or incorrectly predicted genes in public databanks are identified using the MICheck procedure (6). A second workflow, including >20 computational methods, follows and updates functional and relational analyses with new genomic annotations, primary databank releases and new software versions (4). The results of these analysis tools are stored in specific relational tables within PkGDB (1), together with the primary data used as inputs [UniProt (7), InterPro (8), COG (9), etc.]. All these results can be queried using the 'Search by Keywords' functionality. Genes are then associated with a secondary functional annotation, initially based on protein sequence similarity results obtained with closely related and curated genome sequences. This generated set of functional annotations involving enzymatic activities is then the starting point for metabolic pathway reconstructions using the PathoLogic software (10) and the MetaCyc pathway reference database (11). The predicted metabolic network of each prokaryotic genome integrated into PkGDB is computed and stored in a Pathway/Genome Database (PGDB), which is connected to the MaGe interface. These metabolic data sets, together with the mapping of the predicted enzymatic functions onto the KEGG metabolic maps (12), are made available through the 'Metabolic tools' menu of the web interface. The database architecture supports integration of automatic and manually curated annotations and records a history of all the modifications. At least one-third of the 120 current projects (gathering 866 public bacterial genomes and 794 private ones) contain genomes undergoing an active in-depth annotation. Currently, MicroScope has an average of 250 active accounts per month among the 1200 registered users, as well as $\sim$1700 monthly unique visitors that use the database for its large set of exploration functionalities.

Here, we present the major improvements and changes made to MicroScope including the expansion of the database and the redesign of the website to extend functionality for querying, exploring, comparing, sorting or selecting all genomic and metabolic information from the database. In addition, original tools have also been developed to guide metabolic data expertise. With these new enhancements, MicroScope is becoming one of the most comprehensive bacterial genome resources to systematically reconstruct and curate complete representations of metabolic processes/networks from genome annotations.

DATA CONTENT GROWTH AND CURATION

As previously described, the MicroScope database, named PkGDB, uses the open source MySQL relational database management system for storing and accessing genomic data from prokaryotes (both publicly available and newly sequenced genomes), together with the results of the prediction tools implemented in the pipelines (4).

MicroScope is used either for the annotation of novel genomes or for the curation of previously annotated genomes (i.e. re-annotation projects). The latest release of PkGDB (as of September 2012) contains 120 projects gathering 1660 bacterial and archaeal genomes, 866 of which are available in public databases; this represents a 4-fold increase in the number of species in the database since 2009. Original annotations are stored in PkGDB together with the automated annotations generated by the MicroScope annotation pipelines. Since the last MicroScope release (4), several new prediction tools have been added to the structural and functional annotation pipelines: gene prediction using the Prodigal software (13), similarity searches in the FigFam functional protein families (14) and prediction of non-ribosomal proteins using the 2metdb method (15). The results of the analysis tools are stored in specific relational tables: this information is made available in the gene editor (i.e. in the context of the expert annotation process); moreover, the 'Search by keywords' interface has been updated to allow scientists to query all the results of the new precomputed methods. A large part of these are updated in PkGDB at regular intervals to take into account databases growth and new expert annotations.

The MicroScope's Gene Editor allows individual scientists or groups of scientists to review and curate the functional annotation of microbial genomes using public knowledge created during annotation of others genomes, by making comparative analysis in terms of gene order, conservation gene content, function capabilities and metabolic consistency. An integrative strategy allows annotators to quickly browse functional evidence, tracking the history of an annotation and checking the gene context conservation with orthologous genes associated to experimentally demonstrated biological function (4). Since the last MicroScope update, we registered about 50 000 manual annotations a year (3700 per month). Indeed, among the 1660 microbial genomes integrated in PkGDB, 445 contain at least one manually annotated gene. This set of genomes can be divided into four categories depending on the percentage of genes having been manually curated (Figure 1): (i) expertly curated genomes, for which a comprehensive and regular improvement of gene annotation has been performed (i.e. between 80 and 100% of the genes are curated); (ii) adequately curated genomes (i.e. between 30 and 80%); (iii) crudely curated genomes (between 5 and 30%); and (iv) genomes with only a few specific functions curated ($<5\%$). The second and the third categories often correspond to genomes for which a large majority of the gene annotations are transferred from reference genomes; thus, only strain-specific genes are manually annotated. Interestingly, almost 60 microbial genomes (14% of 445

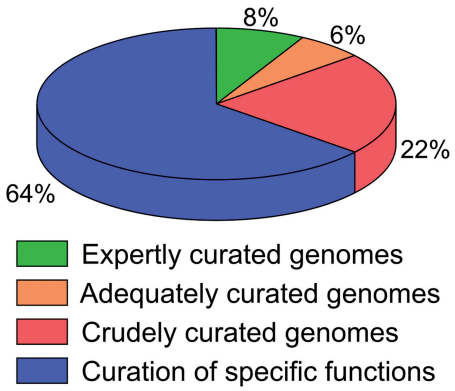

Figure 1. Curation status of MicroScope genomes with at least one gene manually curated. Status is defined by dividing the number of manually annotated genes (at least once) by the total number of genes in the genome: (i) expertly curated genomes: between 80 and 100%; (ii) adequately curated genomes: between 30 and 80%; (iii) crudely curated genomes: between 5 and 30%; and (iv) genomes with curation of specific functions: <5%.

genomes) fall in the 'expertly curated'/'adequately curated' categories (Figure 1) and include several original environmental bacteria such as *Thiomonas sp.* 3As (16), *Azospirillum lipoferum* 4B (17) and *Magnetospirillum gryphiswaldense* MSR-1 (18). In some cases, the curation process was also based on phenotypic data produced in the context of specific projects: AcinetoScope [*Acinetobacter baylyii* ADP1, (19)], NeisseriaScope [*Neisseria menigitidis*, (20)] and RhizoScope [Bradyrhizobium ORS278, (21)], amongst others. The list of (re)annotated public microbial genomes integrated in PkGDB is given in the Supplementary Table S1: a total of 277 680 manual curations based on >47 400 bibliographic references from peer-reviewed international journals have been performed on these bacterial species (Supplementary Table S1). These expert annotations, in particular those corresponding to model organisms, contribute largely to improving the quality of the automatic annotations of new microbial organisms by limiting the percolation of annotation errors (i.e. genomes annotated using old genomes as a reference which, in turn, have been annotated based on even older more out of date genomes).

As shown in Figure 2, although the number of MicroScope users having a personal account has increased significantly since 2011, the number of expert annotations stored in PkGDB is clearly decreasing, reaching only 25 000 at the end of September 2012. Indeed, in addition to the new type of project handled by the system (i.e. for evolution and transcriptomic projects; Figure 2), the MicroScope platform is also (and in some cases, exclusively) used for the set of analysis tools pertaining to microbial genomics and metabolism which have been recently integrated and made available through the new Web interface.

WEB SITE EVOLUTION AND TOOLS EXPANSION

New website design

To make all the PkGDB data and the comparative analysis tools widely accessible to the scientific community, we have developed the new web site structure shown in Supplementary Figure S1. The MicroScope analysis tools and graphical interfaces are organized as several categories (main tool bar in the black rectangle): The 'MaGe' and 'Genomic tools' menus contain functionalities needed for the curation process and for the summary of the automatic and expert annotation (Genome overview, COG functional classification, etc.). The 'Comparative genomics' menu gathers tools, which compare the query genome to others microbial genomes (Supplementary Figure S1): Gene phyloprofile, Genomic island prediction, Line plot using synteny results, computation of fusion/fission events and statistics about synteny results using the PkGDB and the RefSeq genome resources (3). The 'Metabolism' menu initially contained tools allowing the user to browse KEGG or MicroCyc metabolic data and to compare metabolic pathways across several organisms (4). This menu has recently been extended to several new tools dedicated to the curation of metabolic data (see section 'Metabolic data curation tools'). The 'Searches' menu allows the user to perform Blast searches in PkGDB and to use the 'Search by keywords' interface, which allows for the identification of genes and functions of interest using a variety of selection filters (queries can be made on the annotations and on the pre-computed results). Sequence and annotation data available in PkGDB can be downloaded in standard file formats (Genbank, EMBL, tabulated, etc.; 'Export' menu); this now includes the PGDBs computed with the Pathologic software (4,10), which can be then loaded directly into any local installation of BioCyc (11). We also added some functionalities to extract non-coding DNA sequences and/or DNA fragments. Several new tools, presented in the 'Analysis of experimental data' section, have been developed and organized in the 'Experimental Data' menu (see later in the text). Finally, using the 'User Panel' menu, various MicroScope settings can be modified (i.e. the selected genomes in the synteny maps, the set of favourite organisms to work with, etc.). In addition, the main coordinator of a project is now able to manage the personal accounts related to his/her project (i.e. remove/add accounts, setup access rights on sequences).

Gene carts in MicroScope handle lists of genes resulting from analysis tools and/or queries (4). The 'Gene carts' interface allows one to perform and combine various gene list operations: merge, intersect, differentiate of two gene carts, extract gene/protein sequences and run multiple alignments via the plugged Jalview software (22). MicroScope users, which are involved in the same project(s), can now exchange results from their analyses/queries through the upload of gene carts (in an XML file format). In addition, the 'Search by keywords' interface has been extended to use data from one or several gene carts in the context of a query.

Comparative genomics tools

Taxon synteny view

Genomes may be compared in terms of gene content via the 'Gene Phyloprofile' tool, which uses pre-computed homologies and synteny groups (1). A new visualization mode has been added to represent synteny conservation

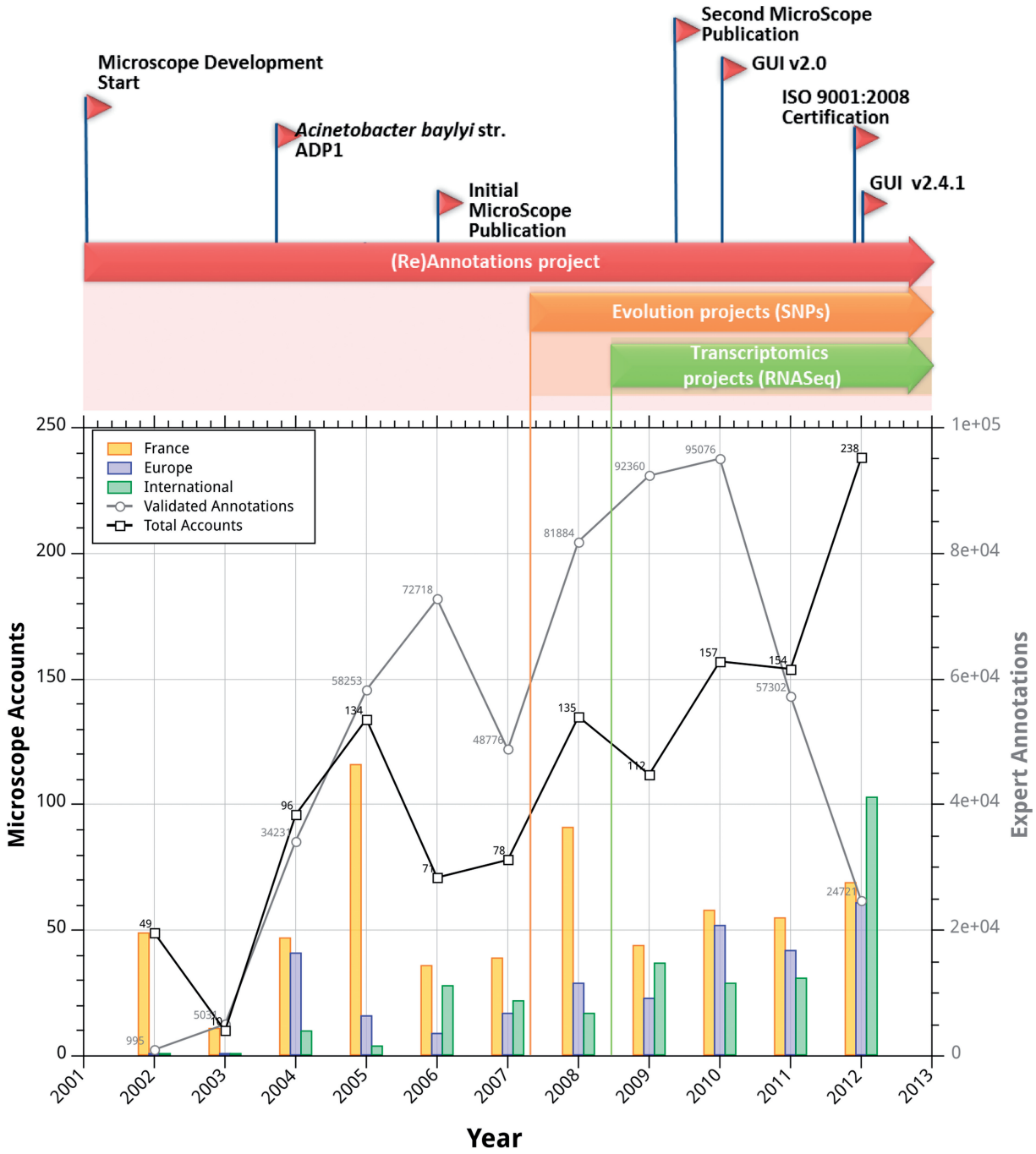

Figure 2. MicroScope usage trend and types of managed projects.

grouped by taxonomic level (i.e. phylum, class, order, family or species). In this new 'taxon-synteny' mode, each line of the synteny map refers to a taxon and colored boxes represent synteny conservations in at least one organism of the corresponding taxon. These tags have a size identical to that of the reference genes on the genome browser, and a colour gradient represents the percentage of genomes in synteny (see Supplementary Figure S1 for an example).

Searching for co-evolving genes

To analyse the co-variation of the *dnaE* and *polC* genes in bacterial genomes, we developed a generic algorithm that can be used as a tool for uncovering co-evolving genes in general (23). The method uses the fact that functionally linked genes may have similar phylogenetic occurrence vectors (24). It computes phylogenetic profiles using gene neighborhood information. The full rationale of the strategy is described in (23). This novel phylogenetic profile method has been integrated into the MicroScope's gene editor. It allows the user to compute co-evolution scores of any target gene against all genes of the studied organism. In addition, this tool has also been added in a MicroCyc interface [that gives the list of gene-reaction associations predicted by the Pathologic algorithm

involved in a given MetaCyc pathway (11)]. This functionality helps searching for candidate genes for missing gene-reaction associations by computing an average co-evolution score using all known genes of a given pathway. As shown in Figure 4C, the method is able to find an interesting gene candidate (STM2090), also co-localized with the two other genes involved in the 'CDP-2,6-dideoxyhexose biosynthesis'. This prediction is strengthened by other approaches that allowed the user to correct/improve the functional annotation of the STM2090 *Salmonella typhimurium* gene (see section 'Using the new MicroScope functionalities: a case study' and Figure 4).

Analysis of experimental data

High-Throughput Sequencing technologies increased not only the flow of genome sequencing projects but also paved the way to new applications of sequencing in both population genetics and functional genomics. To deal with these new types of experimental data, we set up two MicroScope extensions following the three-way architecture of the platform: background analysis pipelines, a database for storing results produced by the pipeline and Graphical User Interfaces (GUI) for the exploration and analysis of data. The pipelines of the data from evolution projects (i.e. clones of the same species at different generation time) and from transcriptomic projects (i.e. RNAseq experiments) rely on a common base: the two first steps consist in the pre-processing of reads (i.e. discarding low quality reads, trimming low quality bases at 3′-ends, etc.) and in their mapping onto a reference genome.

Evolution projects

The problem of discriminating between true mutations that have occurred during evolution and sequencing errors (in the reference sequence or in the new sequence data) is addressed in a 'home made' pipeline called SNiPer, which is based on the ssaha2 software (25). The detected events are stored in a relational database, the model of which allows one to (i) maintain an experimental tracking (i.e. identification of the sequencing run and the technology used); (ii) maintain an evolutionary tracking (i.e. lineages, ancestors, generation time, etc.); and (iii) infer impacts of detected mutations on the reference genomic sequence thanks to the connectivity with existing annotations stored in PkGDB. To efficiently explore this large amount of data (~160 000 SNP/inDel events currently identified for all projects), several GUIs have been developed to address three kinds of analyses: a comparative mode, a parallelism mode and a graphical mode (Figure 3). All these tools provide links to other MicroScope data, enabling exploration and analyses of results in a broader context.

Transcriptomic projects

We designed a database model to store information on experimental conditions, sequencing runs, transcript coverage along a genome, expression levels computed for all genomic objects and the results of statistical tests in case of differential expression analysis. This model supports to run the analysis pipeline on several projects in parallel, as well as the integration of the corresponding RNA-Seq data with all other MicroScope data. The developed GUI allows users to explore most of the RNA-Seq results online, to combine and analyse them using other tools from MicroScope (explore annotations, highlight metabolic pathways, search for orthologs of differentially expressed genes, etc.) and to download results locally. More specifically, raw and normalized expression levels can be displayed for any genomic object on any experimental condition, and all appropriate pairwise comparisons of experimental conditions can be directly queried from the interface. In addition, transcript coverage over genomes are displayed together with genome annotations using the Integrative Genomics Viewer software (26) directly available from the MicroScope web interface, and expression levels can be automatically loaded into the Multi-experiment Viewer software (27) for further data analysis, such as clustering or gene-set enrichment analyses.

Today, we manage 10 evolution projects gathering a total of 250 evolved clones of nine reference genomes; only some of these projects are publicly available to date (28,29). Moreover, 17 transcriptomic projects have also been integrated four of which can be explored using the MicroScope guest access.

METABOLIC DATA CURATION TOOLS

As described in Vallenet *et al.* (4), the quality of the homology-based reconstruction of metabolic networks performed with Pathway Tools depends highly on annotation quality, metabolic database completeness and the criterion for assessing the presence of a pathway. This automatic process is a good starting point to get an overview of the metabolic capabilities of the studied genomes, but many false-positive pathways are predicted. In the context of the Microme FP7 European project (http://www.microme.eu), we focused our effort on the improvement of the network reconstruction and on the development of new MicroScope tools dedicated to the curation of microbial metabolism, both essential points to improve genome-scale models of microbial organisms (11).

Automatic genome-scale metabolic network reconstructions

Each genome stored in PkGDB is processed by an in-house workflow mainly based on Pathologic (version 16.0), one component of the Pathway Tools software suite (30), to generate a genome-scale metabolic network reconstruction. Pathologic uses a two-step process to build PGDBs from genome annotations. The 'Reactome' projection step associates genes with metabolic reactions from MetaCyc [i.e. the BioCyc reference pathways, (11)]. Owing to wrongly formatted or unspecific Enzyme Commission (EC) numbers, or to insufficiently explicit textual annotations, Pathologic may in some cases either over-predict or miss enzymatic reactions. We thus enhanced the matching procedure by directly using the official MetaCyc reaction frame identifier as the functional annotation. Our export

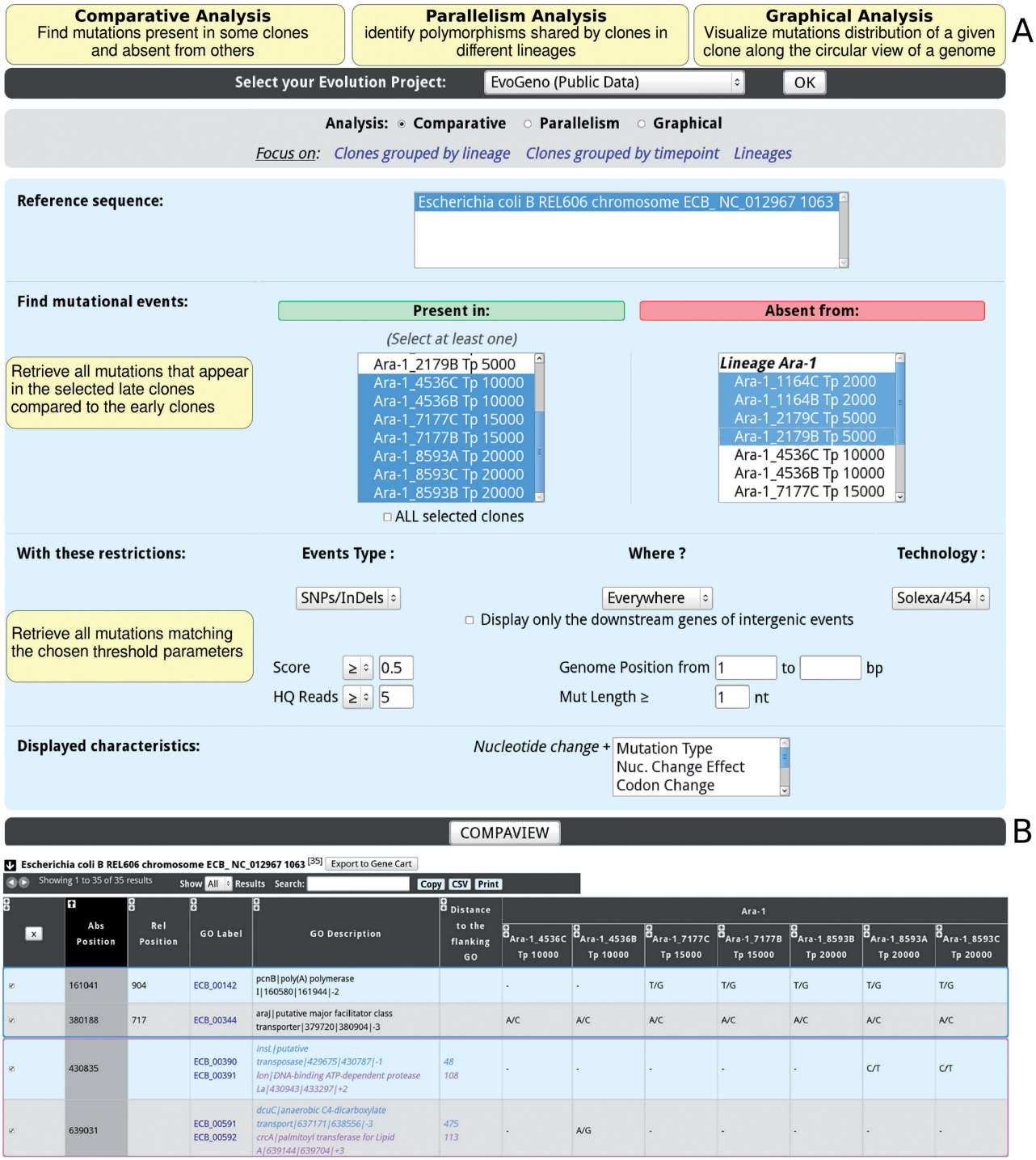

Figure 3. GUI for exploring evolution data stored in MicroScope. The 'comparative mode' allows the user to compare the mutation content of several clones/organisms. It is shown here using public data from the long-term evolution experiment of *E. coli* B (26). (**A**) The query interface allows the user to compare the mutational events (SNPs/InDels) between several clones of the Ara-1 lineage. The mutation content of the selected clones is retrieved using the parameters indicated in the 'With these restrictions' part of the query interface. (**B**) The outcome of the analysis is presented in a table reporting all the genomic objects of the reference genome which contains mutations (blue frame) and the mutations found in intergenic regions (pink frame). In the example shown here, the pcnB gene of *E. coli* B REL606 has undergone a T→G transversion (H302N) at position 904, which appeared at the 15 k generation; it seems to have been subsequently fixed in the population, as it is still present at 20 k generation. In the query interface shown in panel A, the 'parallelism mode' allows the user to analyse the dynamics of mutations found in clones/lineages over time. The 'graphical mode' enables the visualization of mutations along the chromosome of the reference organism and eases the detection of mutation hot spots. All these tools provide links to the main MicroScope database, enabling exploration and analyses of results in a broader context.

procedure from the MicroScope PkGDB to Pathologic input format directly associates the genes to MetaCyc reaction identifiers if any; alternatively, the procedure also exports the partial EC numbers and/or the common gene name product. The set of predicted catalysed reactions in the studied genome is used in the pathway projection step to infer metabolic pathways stored in the MetaCyc resource. The precise rules applied for inferring pathways can be found in (30). This tailor-made reconstruction strategy has been used to build and analyse the genome-scale metabolic networks of 29 *Escherichia coli* strains (8 commensal and 21 pathogenic strains, including six Shigella strains) (31). Evaluation of the benefits of our reconstruction strategy showed that (i) the genome annotation quality directly impacted the number of matched reactions and improved the completion of predicted pathways and (ii) the curation done on a reference metabolic network can be efficiently adapted to closely related organisms (31). Indeed, using a unified source of genome annotations and a common reconstruction process for all metabolic networks limits the biases originating from the reconstruction process, thus making the networks reliably comparable.

The reconstructed metabolic networks are accessible via the MicroScope web interfaces and are also stored in the MicroCyc repository (http://www.genoscope.cns.fr/agc/microcyc), which contains to date >1600 complete metabolic networks (>800 are publicly available). The reconstruction strategy is re-run every day taking into account new expert annotations saved in PkGDB. It provides a list of pathways that need to be asserted as existing or not, and a list of pathway holes (i.e. enzymatic reactions which are not linked to any gene). Two dedicated user interfaces have been developed to ease the expert curation process, and our recently published CanOE strategy (32) has also been integrated in the MicroScope workflow to guide the user in the process of pathway hole filling.

Finding gene candidates for orphan enzymes: the CanOE strategy

In the last update of MicroScope, we described the 'Pathway Synteny' functionality that allowed annotators to retrieve groups of genes in a given organism sharing conserved syntenies and encoding–enzymes involved in a same metabolic pathway (4). The listed results could be used to quickly check for reaction-hole candidate coding genes among the conserved miss-annotated genes of a given group. We have now developed a new tool named Candidate genes for Orphan Enzymes (CanOE) that allows one to automatically analyse the conservation of both genomic and metabolic contexts to provide candidate genes for orphan enzymatic reactions. Details on the CanOE algorithm can be found in (32). This strategy has been applied to all prokaryotic genomes stored in the MicroScope platform, and the results (i.e. genomic metabolons and ranked candidate genes for orphan enzymatic activities) were stored in PkGDB. MicroScope users can query these results in various ways (see the online tutorial for CanOE). An illustration of the use of CanOE is given in Figure 4A and corresponds to the search for *S. typhimurium* LT2 orphan reactions. Four candidate genes have been found for local orphan reactions (i.e. orphan reactions in the target organism), and one candidate gene is associated to a global orphan reaction (i.e. orphan reaction in all known organisms). As explained in the section 'Using the new MicroScope functionalities: a case study' (see later in the text), this last candidate gene is the starting point of the manual curation of the 'CDP-3,6-dideoxyhexose biosynthesis' pathway in *S. typhimurium*.

Curation of metabolic pathways

To help the user in the curation of false-positive predicted pathways and incomplete pathways (i.e. gene-reaction associations are missing), a new 'Pathway curation' interface, available in the 'Metabolism' menu of the MicroScope platform (Supplementary Figure S1), has been developed.

For each predicted pathway x in the reference organism, a pathway completion is computed that corresponds to the number of reactions from pathway x that are found in the genome divided by the total number of reactions in pathway x, as described in the MetaCyc resource (this value ranges between 0 = absence of the pathway, and 1 = complete pathway). The developed interface displays the list of predicted pathways, using the MetaCyc pathway hierarchy, together with the completion value and the number of reactions in the corresponding pathway (see Figure 4B). The users can curate each pathway by assigning different pathway statuses, which reflect the degree of functionality of any predicted pathway: (i) predicted (default status); (ii) validated (the pathway is known to be functional in the organism under study and thus, all the reactions of the pathway should be present); (iii) variant needed (the organism under study metabolizes the compound in a way different from that of the corresponding pathway described in MetaCyc), (iv) unknown (not enough evidence to select another status), (v) non-functional (inactive pathway, i.e. 'pseudogenization' events have affected some enzymatic steps); and (vi) deleted (false-positive predictions of the pathway projection algorithm). Figure 4B illustrates the use of this curation interface for *S. typhimurium* LT2: among the four listed predicted pathways (default status shown in orange), one is missing a reaction (PWY-5833, completion value = 0.67). The proposed candidate gene for this reaction (i.e. STM2090) is found both by CanOE and the 'co-evolved genes' methods. The complete curation process is described in section 'Using the new MicroScope functionalities: a case study', and leads to the 'validated' status for the 'CPD-3,6 dideoxyhexose biosynthesis' in *S. typhimurium* LT2 (Figure 5B).

Curation of Gene-Protein-Reaction associations in the Gene Editor

An essential step in the curation of metabolic data is the curation of associations between genes coding for enzymatic activities and the biochemical reactions catalysed by these enzymes [isozymes, multifunctional enzymes and protein complexes (33)]. The MicroScope Gene Editor

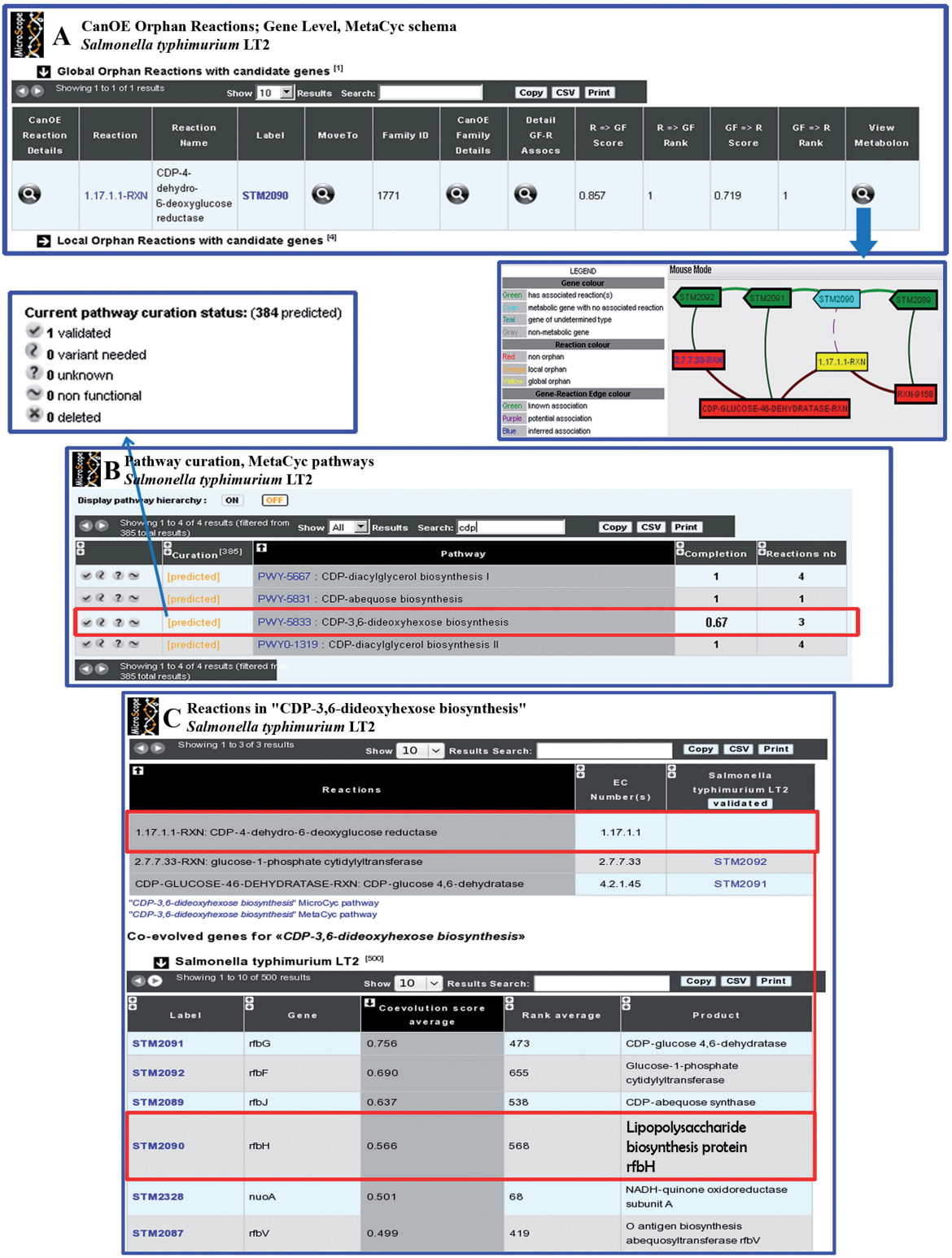

Figure 4. Metabolic data curation strategy in MicroScope platform applied to CDP-3,6-dideoxyhexose biosynthesis in *S. typhimurium* LT2. (**A**) CanOE results for the single global orphan reaction identified in *S. typhimurium* LT2 corresponding to CDP-4-dehydro-6-deoxyglucose reductase activity (1.17.1.1-RXN). This reaction is not linked to any gene but is located in a conserved genomic and metabolic context (genomic Metabolon) with other genes of the CDP-3,6-dideoxyhexose biosynthesis pathway. The CanOE strategy proposes a potential association between this global orphan reaction (yellow box) and STM2090, a metabolic gene with no associated reaction (blue box). (**B**) Pathway curation interface displaying the CDP-3,6-dideoxyhexose biosynthesis pathway. The pathway is incomplete in the initial pathway projection process. By clicking on the pathway identifier (PWY-5833), the reaction table of the CDP-3,6-dideoxyhexose biosynthesis pathway is displayed, showing a single pathway hole that corresponds to the global orphan reaction 1.17.1.1-RXN identified by CanOE (**C**). The same interface also allows the user to display the list of co-evolved genes with other genes of the pathway. In this table, the STM2090 gene proposed by CanOE as possibly associated with 1.17.1.1-RXN is associated to the third best average co-evolution score after the STM2092 and STM2091 genes (which are associated to the other steps of the pathway).

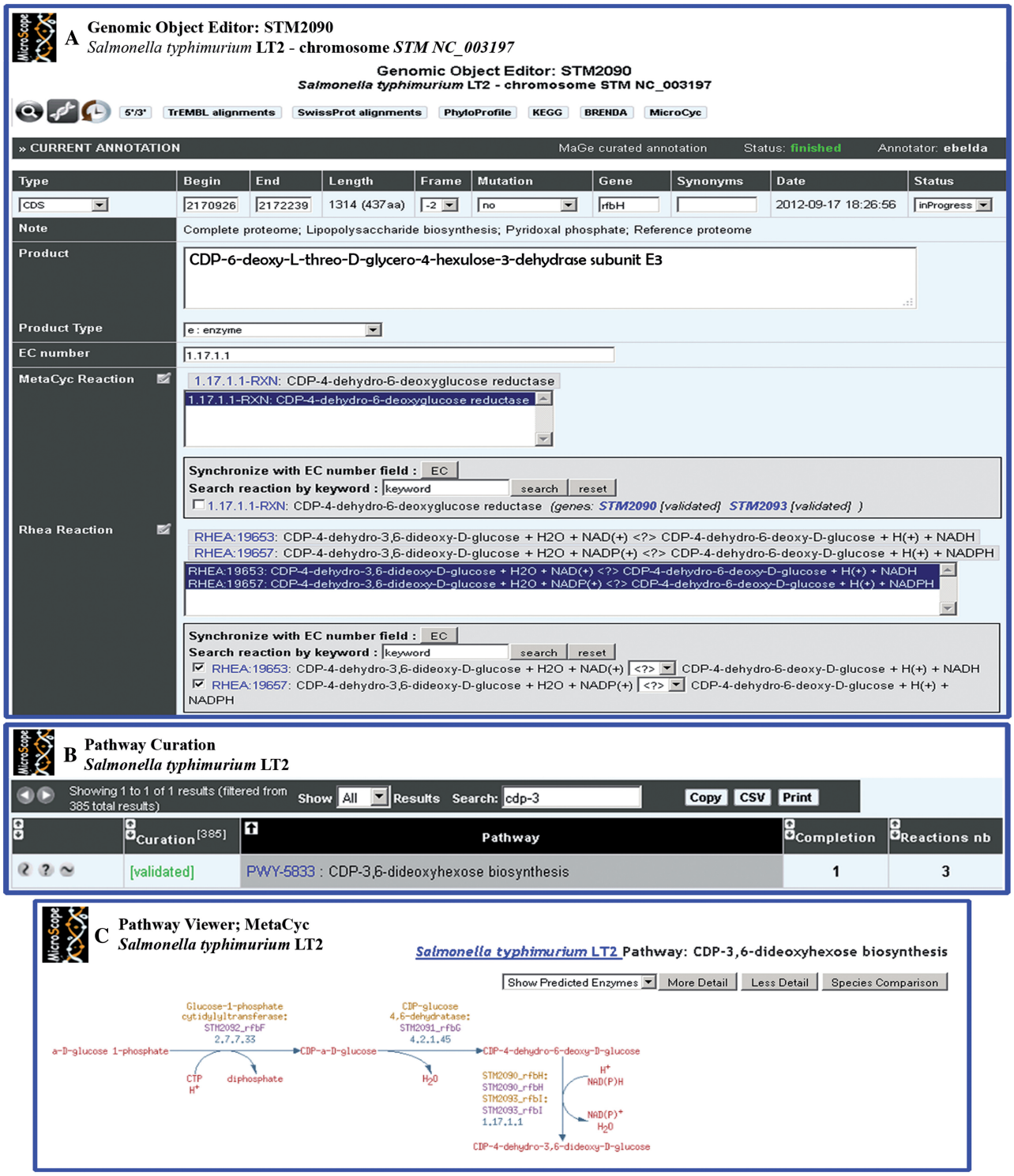

Figure 5. Metabolic data curation strategy in MicroScope platform applied to CDP-3,6-dideoxyhexose biosynthesis in *S. typhimurium* LT2. (**A**) Gene annotation interface of STM2090 in MicroScope. Curation includes the update of the Product and EC number annotation fields and also the manual validation of the corresponding reactions in MetaCyc (1.17.1.1-RXN) and RhEA (RHEA: 19653 and RHEA: 19657) repositories, which can be retrieved automatically from the EC number annotation or using keyword searches. (**B**) Display of the pathway curation interface of MicroScope for CDP-3,6-dideoxyhexose biosynthesis pathway after the curation process. The pathway appears complete (Completion = 1), and it has been assigned to the 'validated' status. (**C**) Display of the complete *S. typhimurium* LT2 CDP-3,6-dideoxyhexose biosynthesis pathway in the specific MicroCyc pathway interface of MicroScope.

has recently evolved to allow association of the gene being annotated to two main enzymatic reactions resources: MetaCyc (11) and RhEA (34). RhEA is the primary resource data for reactions in the Microme project. It defines reactions, which are unique and chemically balanced at the level of mass and charge, using chemical compounds stored in the ChEBI database (35).

Two new fields 'MetaCyc reactions' and 'RhEA reactions' have been added in the Gene Editor (see Figure 5A). The annotator can use the search modules (in grey background; Figure 5A) to retrieve MetaCyc and/or RhEA reactions using the EC number(s) available in the corresponding field of the Gene Editor, or using keywords entered in the 'Search reaction by keyword' area. The result of this query is a multiple selection list of reaction(s) organized as follows: the identifier of the reaction, the name of the reaction and, in parenthesis, the gene(s) in the target genome already associated to the corresponding reaction, if any. These Gene-Reaction (G-R) associations in the genome may have three statuses: (i) 'validated' for associations that have been manually validated; (ii) 'annotated' for reactions that have been transferred from an homologous gene of a closely related genome; and (iii) 'predicted' for G-R associations that have been matched by Pathologic using EC numbers or product names. Such information is very helpful to assert if the enzyme acts as an isoenzyme or as a protein complex. In the example shown in Figure 5A, curation of the STM2090 gene in *S. typhimurium* led to the association of this gene with the 1.17.1.1-RXN MetaCyc reaction (using generic NAD(P) coenzymes) and with two RhEA reactions: RHEA: 19653 (using the NAD coenzyme) and RHEA: 19657 (using the NADP coenzyme).

The PkGDB resource already stores 2800 validated G-R associations such as those of *E. coli* K-12 and *Bacillus subtilis* 168. It also gathers 32 224 transferred and 1 482 039 predicted GR associations.

Using the new MicroScope functionalities: a case study

An example of the use of these different curation tools and interfaces is illustrated in Figure 4 and 5. In the context of the curation of *S. typhimurium* LT2 metabolic data, results of the CanOE strategy were explored to search for predicted gene associations to orphan reactions. Figure 4A shows that only one global orphan reaction (i.e. a reaction without any known coding genes in all MicroScope organisms) corresponding to the EC number 1.17.1.1 is proposed to be associated with a *S. typhimurium* candidate gene, namely STM2090. Indeed, the corresponding genomic metabolon shows that this potential association (dotted purple line in the graph) is found in the genomic region STM2089-STM2092 of the *S. typhimurium* chromosome, the co-localized genes of which have known associations (shown in green) with three others reactions (2.7.7.33-RNX, CDP-GLUCOSE-46-DEHYDRATASE-RNX and RNX9158) that all belong to a same metabolic pathway. The CDP-3,6-dideoxyhexose is a basic component of the O-antigen part of lipopolysaccharides in enterobacteria, and a dominant antigenic determinant in the membranes of these organisms (36). This pathway appears incomplete in the pathway curation interface of MicrosScope, because it has a pathway completion below 1 (Figure 4B). Using the link PWY-5833, the list of reactions involved in the CDP-3,6-dideoxyhexose biosynthesis pathway is displayed (Figure 4C): the automatic projection procedure failed to associate the EC:1.17.1.1 to a *S. typhimurium* gene, leading to a pathway hole for the CDP-4-dehydro-6-deoxyglucose reductase activity (i.e. the global orphan enzymatic activity for which CanOE is able to propose a candidate gene, STM2090).

Looking at the co-evolution scores of *S. typhimurium* LT2 genes with the other genes of this pathway ('co-evolved genes' functionality, Figure 4C), it appears that the STM2090 gene has the third best score just after STM2092 and STM2091, pointing out a common evolutionary profile of presence–absence with other genes of the CDP-3,6-dideoxyhexose biosynthesis pathway. The STM2090 candidate gene was initially annotated as a 'lipopolysaccharide biosynthesis protein rfbH' (Figure 4C), and no enzymatic activity was associated to this gene. However, looking at the blast results available in the Gene Editor, it appears that STM2090 shares significant similarity with *Yersinia pseudotuberculosis* TrEMBL entries (>80% amino-acid identity over the whole protein length); the protein encodes the CDP-6-deoxy-L-threo-D-glycero-4-hexulose-3-dehydrase subunit of the CDP-4-dehydro-6-deoxyglucose reductase complex, the activity of which has been experimentally demonstrated in this organism (36). The analysis of the flanking genes in the metabolon (STM2089-STM2092) led us to identify STM2093 as the gene coding for the CDP-6-deoxy-L-threo-D-glycero-4-hexulose-3-dehydrase reductase subunit of the same complex. Thus, we can assume, with a high confidence, that the EC:1.17.1.1 is catalysed by a CDP-4-dehydro-6-deoxyglucose reductase enzyme complex encoded by the STM2090 and STM2093 genes in *S. typhimurium*. Based on these evidences, the functional annotations of STM2090 and STM02093 have been updated accordingly and associated to the EC number 1.17.1.1. The MetaCyc and RhEA reactions linked to this EC number have been validated (Figure 5A). Following this curation step, the pathway projection tool has been re-executed: the CDP-3,6-dideoxyhexose biosynthesis pathway is now complete (Figure 5C) and has been validated in the pathway curation interface of MicroScope selecting the 'validated' status (Figure 5B).

FUTURE DIRECTIONS

During these past 3 years, an important effort has been made in the improvement of MicroScope functionalities for the reconstruction and the curation of metabolic networks of microbial organisms. One important goal was to facilitate the interpretation of the genomes directly in the light of biological processes. Users are now urged to refine functional annotations in the context of the metabolic capabilities of the studied organism. Following the Microme project goals (www.microme.eu), we plan to integrate this curation process through metabolic models data and experimental data.

The automatic genome-scale metabolic models may be used to predict growth phenotypes and/or gene essentiality. Thus, the model predictions can be compared with experimental data (e.g. the standardized high-throughput phenotyping screen, BIOLOG®) to estimate the quality of the reconstructed metabolic network and to subsequently improve it by refining gene annotations. Indeed, MicroScope offers a solid upstream genome and reaction annotation infrastructure to contribute to model curation/refinement through the flagging of inconsistencies or the generation of candidate corrections.

Moreover, the integration of quantitative RNA-seq transcriptomic data in MicroScope allows users to interpret the predicted gene functions in the light of regulatory cell processes. These experiments give a dynamic view of the genome and are useful, for example, to interpret changes in metabolic fluxes that are predicted by metabolic models. In MicroScope, we plan to enhance the exploitation of RNA-seq data by developing new functionalities that will allow us to (i) to improve the prediction of transcription units, i.e. the detection of transcription start sites and the delimitation of operons using strand specific pair-end sequencing data and (ii) to predict new non-coding RNA, i.e. the detection of highly transcribed intergenic regions.

With the decreasing costs of sequencing, we observe an increasing number of genome projects, which include dozen of strains of the same organism (or related species). Thus, microbiologists have now the opportunity to perform wide pan-genomic studies to understand the genetic bases of phenotypic differences across diverse bacterial isolates. In this context, we have started developing a workflow that incrementally computes protein families over all the genomes integrated in MicroScope. These families will ease the development of procedures to perform vertical annotation of genes across multiple genomes; based on these methods, we will design new MicroScope web interfaces to dynamically compute the core and pan-genome among a selection of organisms. Moreover, we will also add a tool allowing the user to filter or select genomes based on their taxonomy.

To share the results of curated/improved annotations with the scientific community in a more simple way (for computational biologists) than downloading data, we started to develop web services. In the context of the Microme project (http://www.microme.eu), MicroScope is a repository of curated Gene-Protein-Reaction associations that will be used as seed data for automatic projection procedures developed by European Bioinformatics Institute (EBI) partner. This information has been recently made available as a web service using evidence tags between the genes and their metabolic annotation (automatic, curated). We also have developed a java client to allow connecting quickly to the web service. We thus plan to extend this procedure by designing web services that will provide a comprehensive API to the rich information stored on PkGDB.

Finally, in a near future, we can anticipate that single molecule real time sequencing technologies will give access to complete microbial genomes in few minutes. Thus, their automatic analysis should also be done in a similar timescale. We are working towards this goal, by designing new web services and interfaces, which will allow our users to independently submit their genomes to MicroScope. On the IT side, we will increase the number of possible concurrent active threads of our workflows to get a shorter restitution time on high-performance computing infrastructure (thousands of cores). This step has already started in the context of the France Genomique infrastructure (http://www.france-genomique.org, 'Invertissements d'Avenir' calls, 2011).

MICROSCOPE DATA AVAILABILITY

MicroScope's resources are available for download on various file formats. Access is also granted at the programmatic level via simple URL through a REST web service. These features are located under the 'Export' menu of the MicroScope web site.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors would like to thank all MaGe users for their feedback, which helped greatly in optimizing and improving many functionalities of the system. They also thank the entire system network team of Genoscope for its essential contribution to the efficiency of the platform, and Antoine Danchin for his careful editing of the manuscript.

FUNDING

French Ministry of research (funds allocated by the ANR PFTV 2007), the 'groupement d'Intérêt Scientifique Infrastructures en Biologie Sante et Agronomie' (GIS IBiSA), and the MICROME project, EU Framework Program 7 Collaborative Project [222886-2]. Funding for open access charge: the MICROME project, EU Framework Program 7 Collaborative Project [222886-2].

Conflict of interest statement. None declared.

REFERENCES

1. Vallenet,D., Labarre,L., Rouy,Z., Barbe,V., Bocs,S., Cruveiller,S., Lajus,A., Pascal,G., Scarpelli,C. and Medigue,C. (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65.
2. Médigue,C. and Moszer,I. (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Res. Microbiol.*, **158**, 724–736.
3. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
4. Vallenet,D., Engelen,S., Mornico,D., Cruveiller,S., Fleury,L., Lajus,A., Rouy,Z., Roche,D., Salvignol,G., Scarpelli,C. *et al.* (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. *Database*, **2009**, bap021.

5. Richardson,E.J. and Watson,M. (2012) The automatic annotation of bacterial genomes. *Brief Bioinformatics*, March 9 (doi:10.1093/bib/bbs007; epub ahead of print).
6. Cruveiller,S., Le Saux,J., Vallenet,D., Lajus,A., Bocs,S. and Médigue,C. (2005) MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res.*, **33**, W471–W479.
7. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
8. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
9. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
10. Karp,P.D., Paley,S.M., Krummenacker,M., Latendresse,M., Dale,J.M., Lee,T.J., Kaipa,P., Gilham,F., Spaulding,A., Popescu,L. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **11**, 40–79.
11. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–D753.
12. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
13. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
14. Meyer,F., Overbeek,R. and Rodriguez,A. (2009) FIGfams: yet another set of protein families. *Nucleic Acids Res.*, **37**, 6643–6654.
15. Bachmann,B.O. and Ravel,J. (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.*, **458**, 181–217.
16. Arsène-Ploetze,F., Koechler,S., Marchal,M., Coppée,J.Y., Chandler,M., Bonnefoy,V., Brochier-Armanet,C., Barakat,M., Barbe,V., Battaglia-Brunet,F. *et al.* (2010) Structure, function, and evolution of the Thiomonas spp. genome. *PLoS Genet.*, **6**, e1000859.
17. Wisniewski-Dyé,F., Borziak,K., Khalsa-Moyers,G., Alexandre,G., Sukharnikov,L.O., Wuichet,K., Hurst,G.B., McDonald,W.H., Robertson,J.S., Barbe,V. *et al.* (2011) Azospirillum genomes reveal transition of bacteria from aquatic to terrestrial environments. *PLoS Genet.*, **7**, e1002430.
18. Guo,F.F., Yang,W., Jiang,W., Geng,S., Peng,T. and Li,J.L. (2012) Magnetosomes eliminate intracellular reactive oxygen species in Magnetospirillum gryphiswaldense MSR-1. *Environ. Microbiol.*, **14**, 1722–1729.
19. de Berardinis,V., Vallenet,D., Castelli,V., Besnard,M., Pinet,A., Cruaud,C., Samair,S., Lechaplais,C., Gyapay,G., Richez,C. *et al.* (2008) A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Mol. Syst. Biol.*, **4**, 174.
20. Rusniok,C., Vallenet,D., Floquet,S., Ewles,H., Mouzé-Soulama,C., Brown,D., Lajus,A., Buchrieser,C., Médigue,C., Glaser,P. *et al.* (2009) NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen Neisseria meningitidis. *Genome Biol.*, **10**, R110.
21. Giraud,E., Moulin,L., Vallenet,D., Barbe,V., Cytryn,E., Avarre,J.C., Jaubert,M., Simon,D., Cartieaux,F., Prin,Y. *et al.* (2007) Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. *Science*, **316**, 1307–1312.
22. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
23. Engelen,S., Vallenet,D., Médigue,C. and Danchin,A. (2012) Distinct co-evolution patterns of genes associated to DNA polymerase III DnaE and PolC. *BMC Genomics*, **13**, 69.
24. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
25. Ning,Z. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
26. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotech.*, **29**, 24–26.
27. Saeed,A.I., Bhagabati,N.K., Braisted,J.C., Liang,W., Sharov,V., Howe,E.A., Li,J., Thiagarajan,M., White,J.A. and Quackenbush,J. (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134–193.
28. Wielgoss,S., Barrick,J.E., Tenaillon,O., Cruveiller,S., Chane-Woon-Ming,B., Médigue,C., Lenski,R.E. and Schneider,D. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)*, **1**, 183–186.
29. Marlière,P., Patrouix,J., Döring,V., Herdewijn,P., Tricot,S., Cruveiller,S., Bouzon,M. and Mutzel,R. (2011) Chemical evolution of a bacterium's genome. *Angew. Chem. Int. Ed. Engl.*, **50**, 7109–7114.
30. Karp,P.D., Latendresse,M. and Caspi,R. (2011) The pathway tools pathway prediction algorithm. *Stand. Genomic Sci.*, **5**, 424–429.
31. Vieira,G., Sabarly,V., Bourguignon,P.Y., Durot,M., Le Fèvre,F., Mornico,D., Vallenet,D., Bouvet,O., Denamur,E., Schachter,V. *et al.* (2011) Core and panmetabolism in *Escherichia coli*. *J. Bacteriol.*, **193**, 1461–1472.
32. Smith,A.A., Belda,E., Viari,A., Medigue,C. and Vallenet,D. (2012) The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput. Biol.*, **8**, e1002540.
33. Thiele,I. and Palsson,B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.
34. Alcántara,R., Axelsen,K.B., Morgat,A., Belda,E., Coudert,E., Bridge,A., Cao,H., de Matos,P., Ennis,M., Turner,S. *et al.* (2012) Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **40**, D754–D760.
35. de Matos,P., Adams,N., Hastings,J., Moreno,P. and Steinbeck,C. (2012) A database for chemical proteomics: ChEBI. *Methods Mol. Biol.*, **803**, 273–296.
36. Thorson,J.S., Lo,S.F., Ploux,O., He,X. and Liu,H.W. (1994) Studies of the biosynthesis of 3,6-dideoxyhexoses: molecular cloning and characterization of the asc (ascarylose) region from Yersinia pseudotuberculosis serogroup VA. *J. Bacteriol.*, **176**, 5483–5493.