Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D517–D528***
doi: 10.1093/nar/gkw1101

MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes

David Vallenet[1,*], Alexandra Calteau[1], Stéphane Cruveiller[1], Mathieu Gachet[1], Aurélie Lajus[1], Adrien Josso[1], Jonathan Mercier[1], Alexandre Renaux[1], Johan Rollin[1], Zoe Rouy[1], David Roche[1], Claude Scarpelli[2] and Claudine Médigue[1,*]

[1]UMR 8030, CNRS, Université Évry-Val-d'Essonne, CEA, Institut de Génomique - Genoscope, Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, F-91000 Évry, France and [2]CEA, Institut de Génomique - Genoscope, Laboratoire d'Informatique Scientifique, F-91000 Évry, France

Received September 22, 2016; Revised October 25, 2016; Editorial Decision October 26, 2016; Accepted October 27, 2016

ABSTRACT

The annotation of genomes from NGS platforms needs to be automated and fully integrated. However, maintaining consistency and accuracy in genome annotation is a challenging problem because millions of protein database entries are not assigned reliable functions. This shortcoming limits the knowledge that can be extracted from genomes and metabolic models. Launched in 2005, the MicroScope platform (http://www.genoscope.cns.fr/agc/microscope) is an integrative resource that supports systematic and efficient revision of microbial genome annotation, data management and comparative analysis. Effective comparative analysis requires a consistent and complete view of biological data, and therefore, support for reviewing the quality of functional annotation is critical. MicroScope allows users to analyze microbial (meta)genomes together with post-genomic experiment results if any (i.e. transcriptomics, re-sequencing of evolved strains, mutant collections, phenotype data). It combines tools and graphical interfaces to analyze genomes and to perform the expert curation of gene functions in a comparative context. Starting with a short overview of the MicroScope system, this paper focuses on some major improvements of the Web interface, mainly for the submission of genomic data and on original tools and pipelines that have been developed and integrated in the platform: computation of pan-genomes and prediction of biosynthetic gene clusters. Today the resource contains data for more than 6000 microbial genomes, and among the 2700 personal accounts (65% of which are now from foreign countries), 14% of the users are performing expert annotations, on at least a weekly basis, contributing to improve the quality of microbial genome annotations.

INTRODUCTION

In the era of high-throughput sequencing technologies, deciphering genome sequence of an organism or an environmental sample that contains multiple organisms is a routine task. Indeed, genomics is entering in the 'Big data' era (1), and understanding the biological meanings encoded in genomes remains a challenging task. To address the need for 'quick' microbial genome annotation, several automatic pipelines have been developed. These include the RAST (2) and PATRIC (3) web servers that provide genome annotation services in under a day, or the command line software Prokka (4) that produces similar results in a few hours and which is useful when throughput and privacy issues are critical. For metagenome analysis, MG-RAST (5) and EBI metagenomics (6) offer similar services on shotgun sequencing reads. Last year, the NCBI has released a new version of their Prokaryotic Genome Annotation Pipeline capable of analyzing more than 2000 prokaryotic genomes per day (7). Today, the vast majority of genome sequences receives only fully-automatic annotation, mainly based on sequence similarity, i.e. a process which can lead to the introduction and propagation of poor annotation and errors (8). Resources such as HAMAP (9) and NCBI RefSeq (10) have been developed with the aim of gradually increasing the quality and completeness of microbial genome annotation. However, curation efforts remain restricted to large and widespread protein families and these resources can-

[*]To whom correspondence should be addressed. Tel: +33 1 60 87 84 53; Fax: +33 1 60 87 25 14; Email: vallenet@genoscope.cns.fr
Correspondence may also be addressed to Claudine Médigue. Tel: +33 1 60 87 84 59; Fax: +33 1 60 87 25 14; Email: cmedigue@genoscope.cns.fr
Present address: Alexandre Renaux, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK.

not replace expert curations by specialist biologists. Thus, several systems have been developed to generate automatic annotations and provide graphical facilities for subsequent manual review of the predictions (i.e. GenDB (11), SABIA (12), The Integrated Microbial Genomes, IMG (13)). Some of them facilitate community-based curation: this is the case of the SEED (14) which focuses on expert annotation of metabolic pathways as a whole, or the 'wiki-based' systems dedicated for example to gene (Gene Wiki (15)), protein (WikiProteins (16)) and pathway (WikiPathways (17)) annotation. Although these open systems are useful to add and share annotations (which could also be problematic without proper quality control and privacy, at least in a short period of time), few of them (e.g. IMG and SEED) have followed the integrated system approach offering an environment for deeper and extended comparative analysis and expert annotation of gene function in a single system (18).

At the French Genomic Institute (CEA), which is part of the France Génomique infrastructure (https://www.france-genomique.org), the MicroScope platform (originally MaGe (19)) supports systematic and efficient revision of microbial genome annotation and data management, as well as functionalities for comparative genomics, metabolic analysis and transcriptomic analysis (http://www.genoscope.cns.fr/agc/microscope). Compared to the genome annotation resources mentioned above, MicroScope enables collaborative curation in a rich comparative genomic context and offers numerous integrated tools which guide the expert curation process of gene function. The content of the MicroScope data warehouse increases with the submitted projects (genome (re)annotation, analysis of transcriptomic data and evolved strains) which are carried out in close collaboration with microbiologists working on reference species, or in the context of free of charge annotation services (i.e. only part of the >70 000 bacterial genomes available in GenBank/EMBL databanks today are integrated into the MicroScope platform).

For more than 10 years, several analysis tools and functionalities have been developed to ease the automatic and expert annotation process (20) and to systematically reconstruct and curate complete metabolic networks from genome annotations (21). After a short overview of the MicroScope system, this paper focuses on some major novelties since 2013: improvements of the Web interface for the submission of genomic data and integration of original tools and pipelines for the computation of pan-genomes and the prediction of biosynthetic gene clusters. Statistics about the MicroScope database growth content and the MicroScope users are also presented together with some selected use cases leading to information sharing and publications. Finally, we conclude by ongoing work and future directions.

OVERVIEW OF THE MICROSCOPE PLATFORM

The MicroScope platform comprises three main components: several analysis pipelines organized in a workflow management system, a relational database for storing and accessing genomic and metabolic data, and Web graphical interfaces that allow users to explore data, edit annotations and use analysis tools (Figure 1). Technical details of the data analysis and management infrastructure are given in (22).

Workflow management system

Depending on the nature of the submitted input data, distinct workflows are processed. For genome sequences, pipelines for the structural, functional and relational annotation orchestrate more than 25 external/internal bioinformatics software; these include the identification of missing genes in public genomes, a sequence similarity and gene neighborhood based annotation, metabolic network reconstruction and biosynthetic gene cluster prediction. Other workflows manage RNA-seq data pipelines for quantitative transcriptomics and re-sequencing data pipelines for the identification of single nucleotide polymorphisms (SNPs) and Insertion/Deletion events in evolved strains (Figure 1).

These pipelines are organized in a robust automated task management system using the jBPM framework (Java Business Process Management; http://www.jbpm.org) which allows us to handle simultaneously the analysis of several new microbial genomes. Indeed, genomes are continuously integrated into MicroScope in batches of hundred of new genomic sequences (see the 'DATA GROWTH CONTENT AND MICROSCOPE USERS' section). Moreover, a large part of these analyses are updated at regular intervals to take into account primary databases growth and new expert annotations.

Knowledge database: PkGDB (Prokaryotic Genome DataBase)

The results of these analysis tools, together with the primary data used as inputs, are stored in specific relational tables within PkGDB: a database based on the open source MySQL relational database management system. The PkGDB architecture supports integration of automatic and human-curated functional annotations and records a history of all the modifications. Technical improvements were made to reduce the PkGDB loading time of new results generated by the various pipelines. The InnoDB table engine is now used to directly insert new data in production, minimizing performance and concurrency issues. This improvement has resulted in a decrease of genome integration time in MicroScope (by a factor of 2 to 10 depending on the size of the table to modify). Indeed, inserts were previously made by a costly table copy/modification/reindexation procedure on a dedicated server.

Metabolic pathway predictions of the MicroScope genomes are stored in the MicroCyc resource (http://www.genoscope.cns.fr/agc/microcyc) which gathers instances of Pathway Genome DataBases (PGDBs) of the BioCyc system (23). The MicroCyc web server is connected to the MaGe graphical interfaces (Figure 1). To ease the comparison of metabolic networks from several organisms, relational tables have been designed in PkGDB to store information of these MicroCyc PGDBs, together with the KEGG metabolic pathways and modules (24).

The size of PkGDB is today 500 GB for databanks and genome data, and 22TB for the computational results.

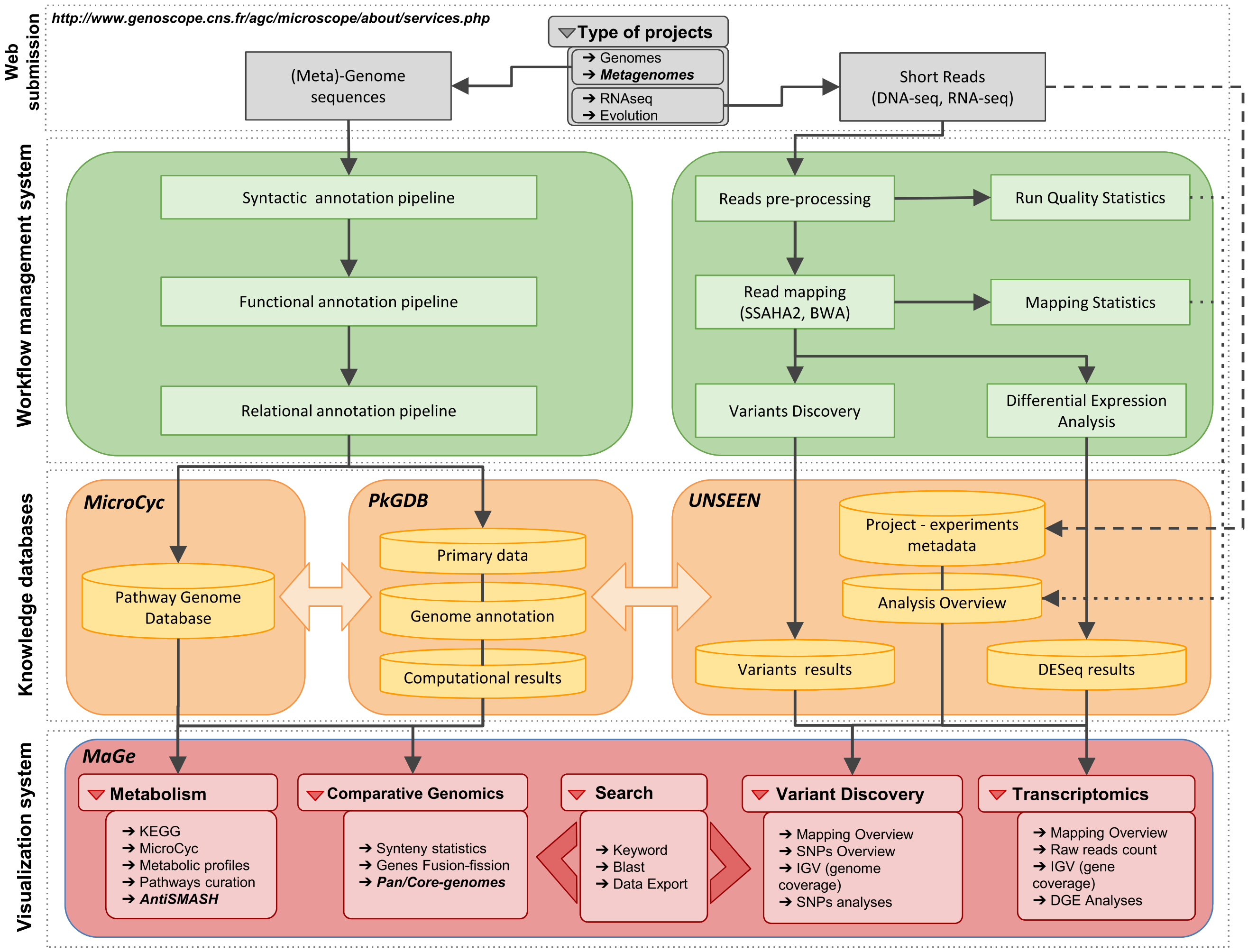


Figure 1. Schematic representation of the MicroScope pipelines for processing genome and short read data. The input (meta-)genome sequences or short reads (DNA-seq or RNA-seq) are submitted via a dedicated Web interface and subsequently processed through several automatic analysis pipelines made up of homemade and third party computational tools. Those latter are organized in a workflow management system including syntactic, functional and relational annotation pipelines, analysis of gene expression based on RNA-seq data and Variant Discovery relying on DNA-seq data. The results of these automatic workflows are stored and organized in various interconnected relational databases. The central database, called PkGDB (Prokaryotic Genome DataBase), stores all the results from the genome annotation pipelines while MicroCyc and UNSEEN (Unified Nextgen SEquENcing data knowledge base) store genome-scale metabolic networks that have been reconstructed using the PathoLogic algorithm and the results from the processing of short read data, respectively. MicroScope users can browse sequences, annotations and analysis results through dedicated Web graphical interfaces directly connected to the databases. Components with major enhancements done since the previous publication of MicroScope are indicated in bold italic text.

Visualization system: MaGe (Magnifying Genome)

All this data is made accessible to users through the MicroScope Web interface (http://www.genoscope.cns.fr/agc/microscope) via secure or anonymous connection as appropriate. Developed using the Apache/PHP server-based language, it consists of numerous dynamic web pages containing textual and graphical representations for accessing and querying data. Several useful graphical applications, such as Artemis (25), MeV (26) and IGV (27), are also available in the MicroScope GUI through plugged Java applications.

One of the main objectives of the platform is to allow biologists to make relevant assessments of the predicted gene functions using (i) a gene annotation editor giving access to the results of each method applied, together with links to several useful public resources, (ii) functionalities for querying and browsing the available data (i.e. each time a new method is integrated into PkGDB, a corresponding data set is added in the 'Search by Keywords' functionality and in the gene editor) and (iii) synteny results and metabolic network predictions, the combination of which can be very helpful in formulating hypotheses on the biological function of non-annotated genes (22) (Figure 1). To facilitate the exploration and curation process, the MicroScope analysis tools are organized in a menu bar which is detailed in Vallenet *et al.* (21).

MICROSCOPE ENHANCEMENTS SINCE 2013

A new MicroScope Web interface for user data integration

Integration of new genomic data in MicroScope is open and free of charge for the worldwide community of microbiologists. To standardize integrated data and make user submission fully automated, we have developed a dedicated Web interface (https://www.genoscope.cns.fr/agc/microscope/about/services.php). Four types of services are provided that correspond to the integration of (i) newly sequenced or publicly available genomes (ii) genome assemblies (bins) from metagenomic samples (iii) RNA-seq data for quantitative transcriptomics (iv) DNA-seq data to identify genomic variations in evolved strains. Using this Web interface, users can easily upload their fasta (genome assemblies) or fastq (RNA-seq or DNA-seq reads) files, complete the requested metadata (sequencing procedure, genome and experimental properties), and finally, approve the term of services. Users are then informed by e-mail about the progress of their integration request. To keep confidentiality, data access through the MicroScope Web interface is restricted to the user that made the request, with the exception of public genomes from databanks that remain publicly available. At any moment, users may grant access to collaborators using the 'Access Right Management' functionality.

To ease data integration and comparative studies, standardization of contextual data about genome sequences is essential. Thus, we try to follow as much as possible INSDC (International Nucleotide Sequence Database Collaboration) and GSC (Genomic Standards Consortium; http://gensc.org/) guidelines in MicroScope. For metagenomes, a dedicated form which follows the MIMS specifications (Minimum Information about a Metagenome Sequence (28)) is available since July 2016 in the 'Delivery of Service' Web interface. Submitting a metagenome in Microscope is as straightforward as submitting a complete/draft single genome. After giving a name and a free text description of their metagenomic sample, users are invited to select the type of environment (e.g. soil, air, water, human-associated, plant-associated) and to complete the associated fields (e.g. collection date, environment biome, geographic location, etc). These fields might be variable depending on the origin of the selected metagenomic sample (i.e. they are dynamically loaded and displayed upon metagenome type selection). Like in IMG/M, only assembled metagenomic data can be analyzed in MicroScope running the pipelines used for isolate genomes. These developments are quite new and submitted metagenomic data is still being analyzed by their owners. They will become publicly available as soon as the work is published. Beside, tools to add and query metadata on genomes and metagenomes will continue to be improved in MicroScope. Indeed, the PkGDB data model is flexible enough to store predefined descriptors, like MIMS or the ones defined by users. We plan, for example, to allow users to add organism phenotypes like antibiotic resistance and growth on specific media.

Data integration, service continuity and data conservation (backups) are currently provided free of charge. MicroScope services follow the quality management system of our laboratory (ISO 9001:2008 and NF X50-900:2013 standards).

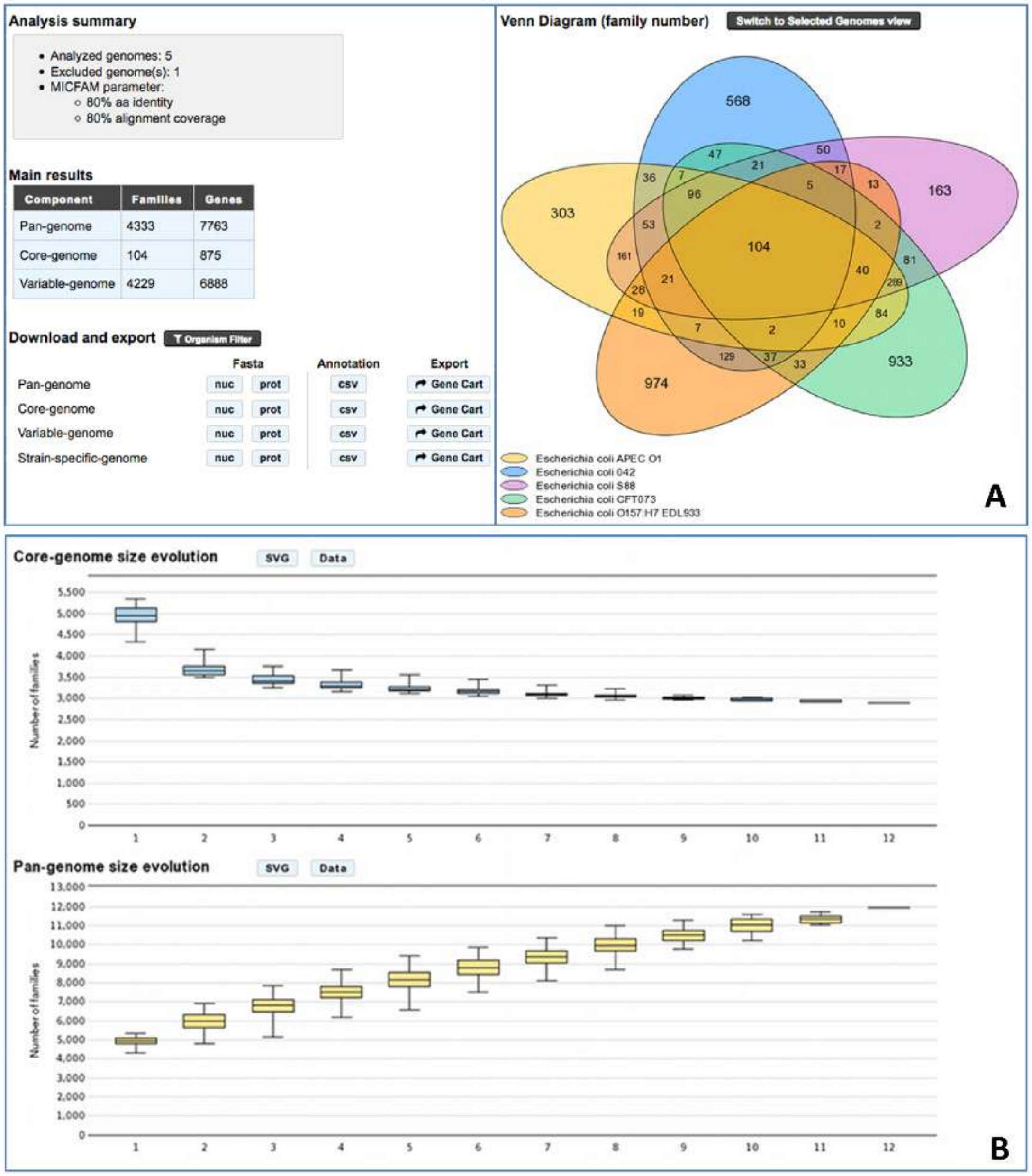
Dynamic computation of pan-genomes

A new functionality has been integrated in MicroScope to dynamically compute pan-genomes according to a selection of organisms. This tool is based on MicroScope gene/protein families (MICFAMs) that are computed incrementally each time a new genome is integrated into the platform. MICFAMs classify proteins in homolog groups using a single-linkage clustering (SiLiX software (29)) of protein sequence pairs sharing at least 80% of amino-acid identity and 80% of alignment coverage. A second set of MICFAMs is also available with a relaxed identity parameter of 50%. Using the pan-genome Web form, users may select up to 200 genomes and, optionally, another set of genomes to exclude (i.e. protein families of the pan-, core- or variable genomes of these genomes will be removed from the final result). According to the user request, corresponding MICFAMs (i.e. present in at least one selected organism) are extracted from the PkGDB database and encoded as vectors of presence/absence. Bitwise operations on these binary vectors are then applied to quickly compute the pan-genome components of the selected organisms. The result page displays the size of each pan-genome component in number of families and genes, and provides links to download protein sequences and annotations. Family counts are also displayed on a Venn diagram when the number of input genomes is below six (Figure 2A). Finally, box plot graphs could be displayed to visualize the evolution of the core-genome and pan-genome size according to the number of considered genomes (i.e. for more than 10 selected genomes, ∼1000 combinations are sampled among the total combination distribution to limit the combinatorial explosion) (Figure 2B). Pan-genome results can be exported in various text formats or saved in a gene cart for additional analyses using other MicroScope functionalities like the 'Search by Keywords' one (20).

Biosynthetic gene cluster annotation

New functionalities dedicated to secondary metabolite prediction have been implemented lately in the MicroScope platform. Prediction relies on the integration of the antiSMASH (antibiotics and Secondary Metabolite Analysis Shell) program, which enables rapid genome-wide identification, annotation and analysis of secondary metabolite Biosynthesis Gene Clusters (BGCs) in bacterial and fungal genomes. It integrates (and cross-links with) a large number of *in silico* secondary metabolite analysis tools that have been published earlier (30,31). AntiSMASH detects more than 40 types of BGCs covering a large range of known secondary metabolite compound classes that are produced by non-ribosomal peptide synthetases (NRPS), polyketide synthase (PKS) or ribosomally synthesized peptide products (RiPP). Moreover, predicted clusters are compared to the recently published reference database MIBiG (Minimal Information about a Biosynthetic Gene cluster), which aims at developing a genomic standard for experimentally characterized BGC of all types (NRPS, PKS or RiPP) and from all phyla (bacteria, fungi or plants; (32)). This database provides a comprehensive and standardized specification for BGC annotations and gene cluster-associated metadata.

In MicroScope, users can get a complete view of the BGCs predicted in one organism by using the 'antiSMASH' section of the Metabolism menu. Each antiSMASH cluster and its genomic context can be explored in a dedicated visu-



Analysis summary

- Analyzed genomes: 5
- Excluded genome(s): 1
- MICFAM parameter:
  - 80% aa identity
  - 80% alignment coverage

Main results

Component	Families	Genes
Pan-genome	4333	7763
Core-genome	104	875
Variable-genome	4229	6888

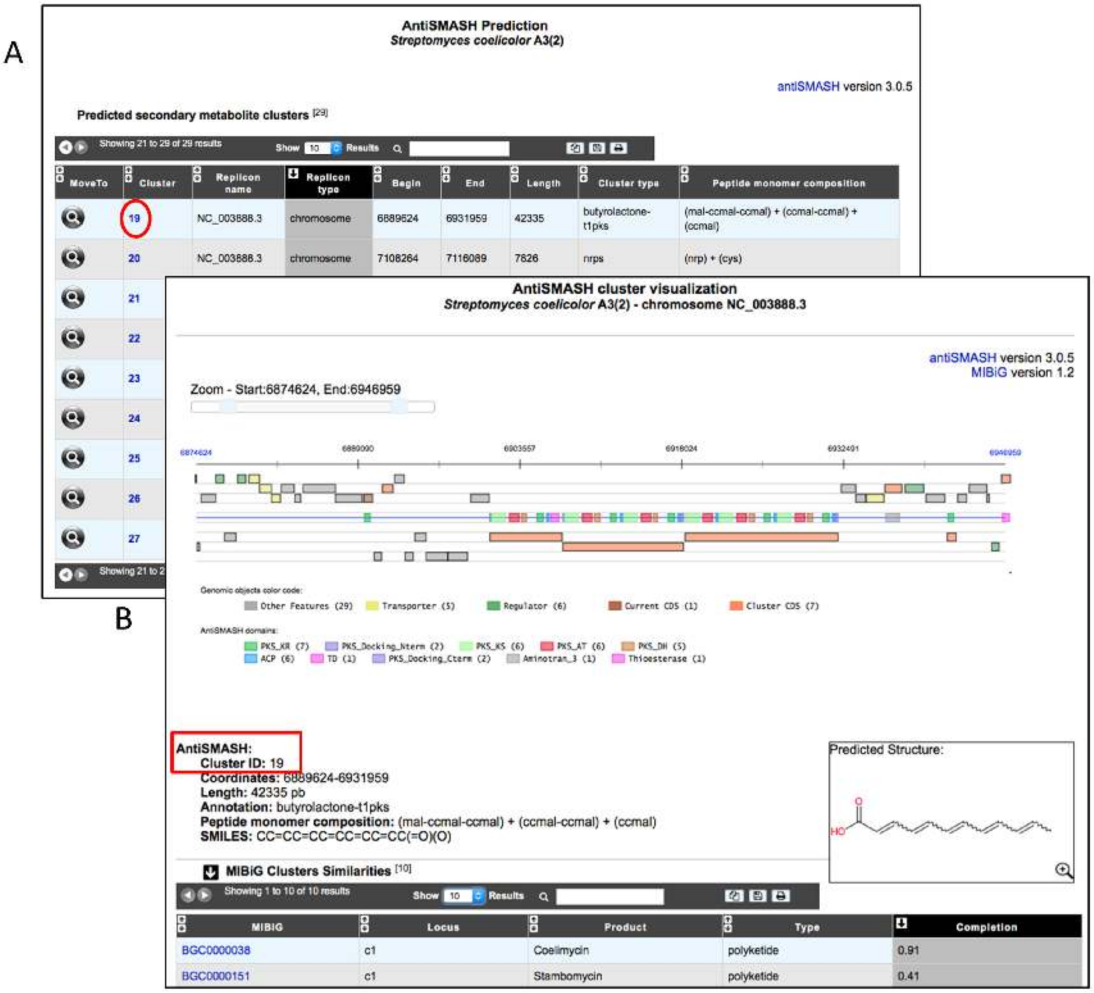
Download and export

	Fasta		Annotation	Export
Pan-genome	nuc	prot	csv	Gene Cart
Core-genome	nuc	prot	csv	Gene Cart
Variable-genome	nuc	prot	csv	Gene Cart
Strain-specific-genome	nuc	prot	csv	Gene Cart

Figure 2. Results of the dynamic computation of pan-genomes. (**A**) Five *Escherichia coli* pathogenic strains (APEC 01, O42, S88, CFT073 and O157:H7) have been selected with the exclusion of *E. coli* K-12 (see text) to compute the number of genes and MICFAM families shared by the five strains (core-genome), by less than five strains (variable-genome) and found in only one strain (strain-specific genes). Values on the Venn diagram represent the numbers of families shared by the compared genomes. The center ($n = 104$) corresponds to the core-genome (gene families shared by the five *E. coli* pathogenic strains) minus the excluded families, i.e. the ones found in the *E. coli* K-12 (commensal strain). (**B**) The two box-plot graphs have been obtained with a selection of 12 *E. coli* strains and show core- and pan-genome size evolution (see text). The last values correspond to the core-genome and the pan-genome effective sizes for the 12 genomes.

alization window showing also a graphical representation of the gene domain composition (Figure 3). For NRPS/PKS cluster types, the predicted peptide monomer composition and its corresponding SMILES formula are specified. The predicted chemical structure is displayed as well using the SMILES Depictor web service (https://cdkdepict-openchem.rhcloud.com/depict.html). Below the graphical representation of the predicted antiSMASH cluster, a summary of MIBiG Clusters similarities, BGC gene composition as well as tailoring cluster similarities are given. This last item relies on a knowledge database provided with antiSMASH about tailoring clusters already described in known BGCs and associated with publications. The results of antiSMASH are also available in the gene editor window and can be queried using the 'Search by keywords' functionality.

In order to fully exploit the results of the different tools dedicated to genomic regions study (e.g. antiSMASH or RGPfinder), we are currently working on the development of a specific editor. At present time, the gene editor is only



AntiSMASH Prediction
***Streptomyces coelicolor* A3(2)**

antiSMASH version 3.0.5

Predicted secondary metabolite clusters [29]

Showing 21 to 29 of 29 results

MoveTo	Cluster	Replicon name	Replicon type	Begin	End	Length	Cluster type	Peptide monomer composition
🔍	19	NC_003888.3	chromosome	6889624	6931959	42335	butyrolactone-t1pks	(mal-ccmal-ccmal) + (ccmal-ccmal) + (ccmal)
🔍	20	NC_003888.3	chromosome	7108264	7116089	7826	nrps	(nrp) + (cys)
🔍	21							
🔍	22							
🔍	23							
🔍	24							
🔍	25							
🔍	26							
🔍	27							

AntiSMASH cluster visualization
***Streptomyces coelicolor* A3(2) - chromosome NC_003888.3**

antiSMASH version 3.0.5
MIBiG version 1.2

AntiSMASH:
Cluster ID: 19
Coordinates: 6889624-6931959
Length: 42335 pb
Annotation: butyrolactone-t1pks
Peptide monomer composition: (mal-ccmal-ccmal) + (ccmal-ccmal) + (ccmal)
SMILES: CC=CC=CC=CC=CC=CC(=O)(O)

MIBiG Clusters Similarities [10]

Showing 1 to 10 of 10 results

MIBIG	Locus	Product	Type	Completion
BGC0000038	c1	Coelimycin	polyketide	0.91
BGC0000151	c1	Stambomycin	polyketide	0.41

Figure 3. New MicroScope Biosynthetic gene cluster and analysis functionalities. (**A**) BGC predictions for an organism can be accessed from the 'antiSMASH' section of the 'Metabolism' menu. It gives access to a table summarizing BGC predictions for all replicons of a studied organism. Here, 29 clusters have been predicted by antiSMASH in *Streptomyces coelicolor* A3(2) among which 2 are located on a plasmid. All individual BGC predictions can then be explored by clicking on the cluster numbers. (**B**) Visualization of the cluster number 19. Genes found in the predicted cluster are colored in brown and their domain composition are indicated as well. The BGC is drawn surrounded by its genomic context. Putative transporter and regulator genes are highlighted in yellow and green, respectively. Basic cluster characteristics are indicated below the visualization section such as: coordinates, length and proposed antiSMASH annotation. In case of NRPS/PKS BGC type, peptide monomer composition is indicated with the corresponding chemical structure encoded in SMILES. The BGC visualized here is very similar to MIBiG BGC0000038 cluster (completion score of 0.91) and is indeed responsible for the biosynthesis of the coelimycin, an unusual polyketide alkaloid.

dedicated to the annotation of simple genomic objects (e.g. protein genes, rRNA or tRNA). Its evolution will allow users to annotate complex objects such as gene clusters (operons, BGCs, genomic islands, etc.), CRISPR regions, secretion systems or phages. Thus annotators will be able to perform expert annotations of these specific regions using cross references to thematic databases.

DATA GROWTH CONTENT AND MICROSCOPE USERS

Since the last update (21), the number of genomes integrated into MicroScope has increased threefold. Presently, an average of 6 genomes per day are requested for integration in the platform. The resource contains data for more than 6000 microbial genomes of which ∼2600 are publicly available. Furthermore, expert annotations are continuously gathered in the MicroScope database. Indeed, ∼35 000 annotations are made in a year and more than 370 000 genes have been manually reviewed so far. This curation effort has an important impact on the overall annotation quality: (i) 36% of these annotations describe precise molecular functions that are supported by direct or indirect (i.e. from homology relationships) experimental evidences and linked to more than 38 000 publications (ii) 17% are genes of known function that were automatically annotated as putative or unknown function (iii) 21% are associated with proper chemical reactions (i.e. ∼80 000 genes are involved in ∼160 000 Gene-Protein-Reaction associations in MicroCyc metabolic networks). Biologists generally focused their annotations on some proteins/functions of interest, but MicroScope contains also several genomes that were extensively curated (i.e. ∼100 genomes got more than 1000 curated genes). MicroScope annotations are submitted to INSDC databanks when the genomes get published. Since 2013, MicroScope was cited in more than 350 publications. All the data can also be easily downloaded via the Web interface ('Search/Export→Download Data' functionality). Moreover, we provide a RESTful API to access programmatically public genome data. A dynamic documentation of the API is available at http://rest-microscope.genoscope.cns.fr.

At present time, MicroScope counts more than 2700 personal accounts. The number of registered users has increased by 74% since 2013 with more than 70% of new accounts being held by scientists outside of France. These last years, the platform has even widened its international popularity with 44% of new accounts outside Europe. In total, 65% of MicroScope accounts are held by non-french scientists. It is also important to note that many international projects are conducted through the platform involving users from distant geographic areas (Figure 4). Although authentication isn't required to navigate in MicroScope, it allows users to annotate genes and save data on their personal session. On average per month, we count 360 active accounts (i.e. the user logged in at least once in the month) and 2200 authentications among a total of ∼1700 monthly unique visitors.

Training sessions dedicated to the MicroScope platform are organized at least twice a year in collaboration with the University of Evry. We provide a four and a half-day training 'Annotation and analysis of prokaryotic genomes using the MicroScope platform' which aims to introduce new users to genome exploration, expert annotation and comparative genomics using the various MicroScope functionalities. Since 2016, we also provide an advanced course for former trainees so that they can remain up-to-date on recent developments. More specifically, we focus on transcriptomics analyses and metabolic pathway curation. Since 2008, more than 400 users from 20 countries have been trained and 13 external sessions have been organized in France and abroad (Tunisia, Denmark, Germany, Switzerland, Spain, Netherlands, China). More information is available on our website: http://www.genoscope.cns.fr/agc/microscope/training.

SEVERAL USE CASES LEADING TO INFORMATION SHARING AND PUBLICATIONS

Added-value of the genome re-annotation of reference species

The general tendency of improvement in genome-annotation comes from our increasing knowledge of genomes and from progress in methods and bioinformatics software. Indeed, missed genes and wrong functional annotations impact different types of analyses (i.e. phylogenetic analysis, identification of core genome and strain or species-specific genes). Using a large range of MicroScope functionalities (21), two reference bacterial genomes have been (and are still currently) re-annotated since 2013.

Bacillus subtilis 168 is the only model organism for the Firmicutes clade that has been continuously annotated in depth. Its genome has been re-sequenced and re-annotated (33), and in the mid-2013 we have published a novel curation work in which new genomic objects have been included (i.e. toxin/antitoxin genes, small RNA genes) and the metabolic network has been entirely updated. Seven new metabolic pathways and six new pathway variants that were not presents in the other existing repository (KEGG, MetaCyc) were identified (34).

More recently, we have re-sequenced the *Pseudomonas putida* KT2440 genome and used the MicroScope suite of state-of-the-art genomic analysis tools to revisit the functional and metabolic information encoded in its chromosomal sequence. We identified 242 new protein-coding genes and re-annotated the functions of 1548 genes, which are linked to almost 4900 PubMed references. We also predicted catabolic pathways for 92 compounds (carbon, nitrogen, and phosphorus sources) that could not be accommodated by the previously constructed metabolic models (35). Coupled with BIOLOG® Phenotype Microarray data, this allowed us to improve considerably the outcome of a system biology approach where a model metabolism of the organism could be matched with experimental data. Our approach pinpointed a specific deficiency in our knowledge: we need to considerably improve explicit identification of transport systems. This should be a major task for the immediate future of studies with the *P. putida chassis*, but also for other *chassis* as well.

In MicroScope, gene products that are enzymes or transporters are explicitly linked to the biochemical reactions of the RHEA reaction resource (36), while novel compound entries have been created in the ChEBI database (37). The updated *B. subtilis* 168 and *P. putida* KT2440 genome sequences and annotations were deposited at the International Nucleotide Sequence Database Collaboration (accession numbers AL009126.4 and AE015451.2, respectively) and are the new reference data for these species.

Whole genome SNP-based analysis to decipher evolution of prokaryotic organisms

Starting with the idea that genome sequences represent snapshots of a genome that is constantly evolving, full genome re-sequencing data can provide information on the genetic basis of genome adaptation. Management of evolution projects (genome re-sequencing data from phylogenetically related species or clones of the same species at different generation time) is available in the 'Variant Discovery' menu of the MicroScope platform (21). Pipeline for the identification of SNPs/InDels has been successfully used in the context of several collaborative projects of which the study of natural evolution of *Mycobacterium tuberculosis* and the longest experimental evolution performed on *Escherichia coli* to date.

Understanding the evolution and the pathoadaptation of *M. tuberculosis,* the causative agent of the human tuberculosis agent, was the aim of the first project. *Mycobacterium canettii* and five relevant strains of smooth tubercle bacilli (STB) were selected for deep genomic sequencing and analyzed with our pipeline using *M. tuberculosis* H37Rv as reference. Although tuberculosis-causing mycobacteria share a highly conserved core-genome, SNP based network tree showed that STB are evolutionarily early branching and have much higher rates of genetic variation compared to *M. tuberculosis*. These results, combined with large events

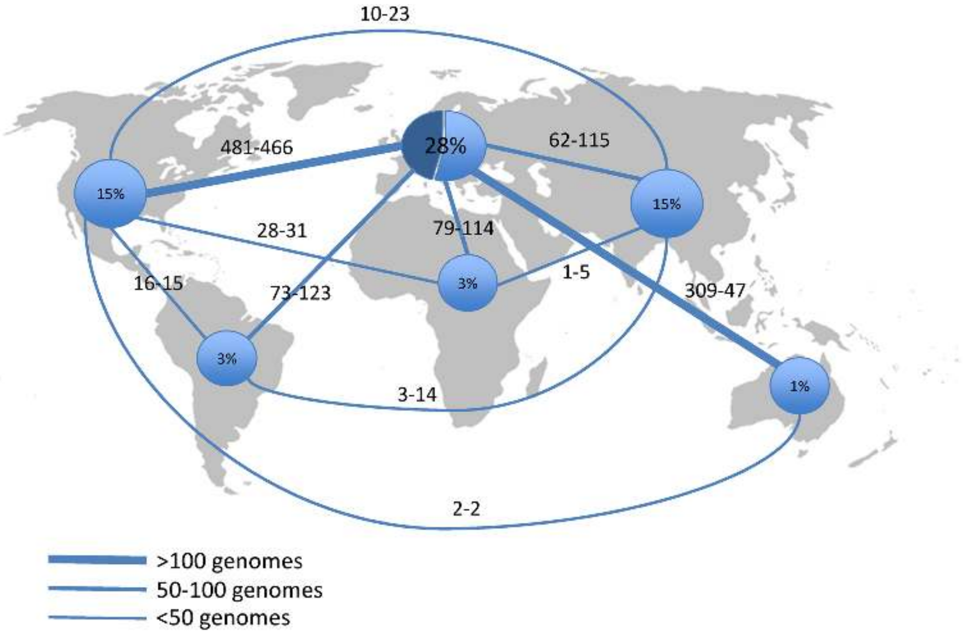


Figure 4. Worldwide distribution of MicroScope user accounts and data interaction. The percentage of MicroScope user accounts are represented as blue circles with their respective percentages on six geographic areas: Europe, North America, South America, Africa, Asia, Oceania. Regarding Europe, 45% of these accounts belong to users from France as indicated by the dark blue part. The lines between circles represent scientific interaction between users. The numbers next to each line indicate the number of shared genomes (on the left) and the number of annotators working together (on the right).

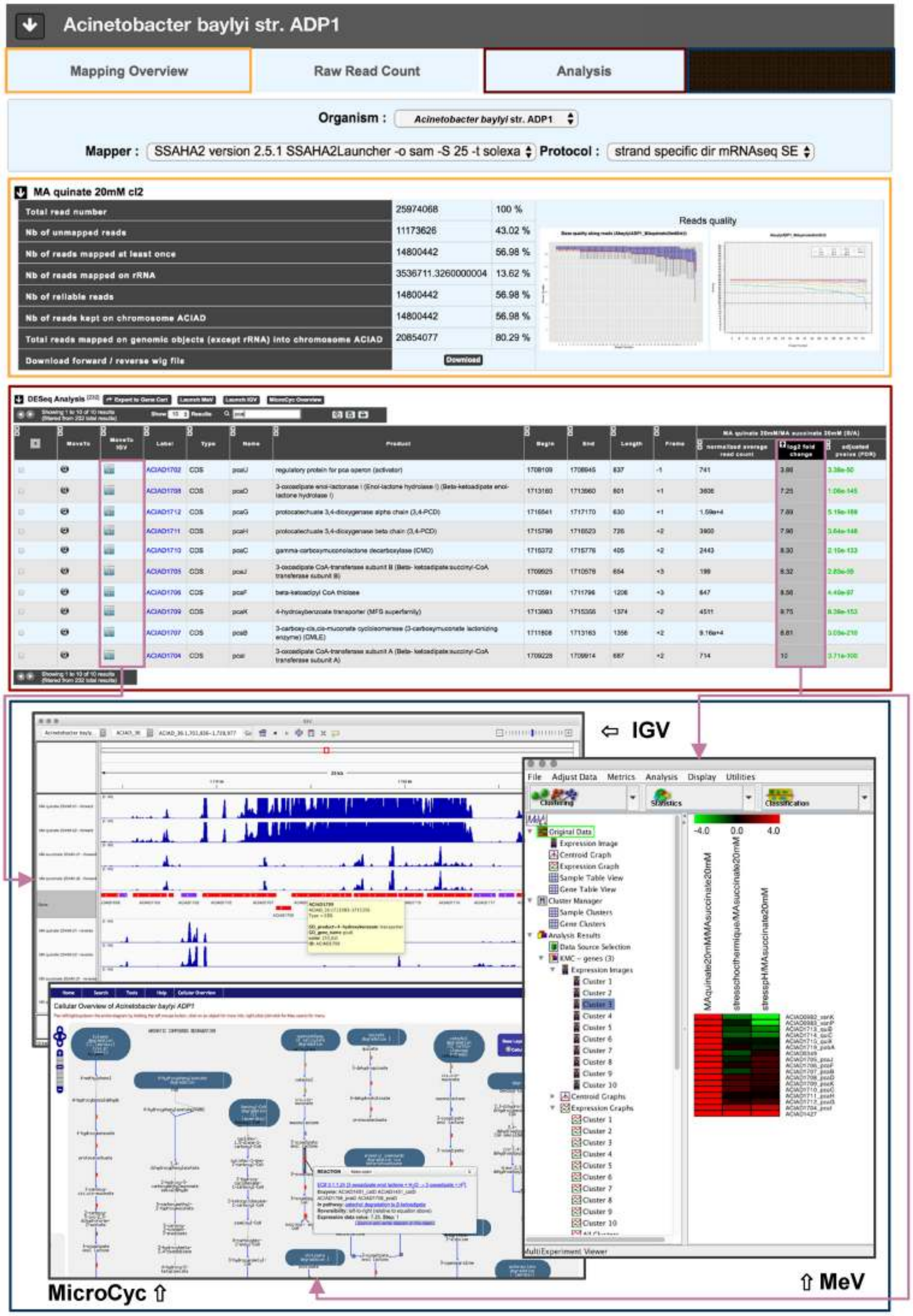
and recombination analyses, led to the conclusion that *M. tuberculosis* evolved from STB by gain of persistence and virulence mechanisms (38).

The long term evolution project aims at directly inferred point-mutation rate based on the accumulation of synonymous substitutions on the sequences of *E. coli* genomes from a 50 000-generation evolution experiment. By sequencing and reconstructing the genetic history of the *E. coli* lineages, several very nice findings have been published. For instance in 2014, it has been shown that mutations all affected regulatory genes and collectively caused substantial metabolic changes. Indeed, three mutations together were sufficient to produce the frequency-dependent fitness effects that allowed one *E. coli* new lineage to invade and stably coexist with the other (39).

Joint usage of RNA-seq and metabolomics data to decipher metabolic features

While transcriptomic data actually help to refine structural gene annotations, they might reveal novel transcriptionally active regions in genome whose sequences harbor genes missed by *ab initio* gene finding algorithms. Within the Microscope platform, RNA-seq data are analyzed using a standard three-steps procedure: (i) quality control of raw reads (ii) reads mapping (iii) differential gene expression analysis. The developed GUI allows users to explore and combine these results using other tools from MicroScope (explore annotations, highlight metabolic pathways, search for homologs of differentially expressed genes, etc.). The 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.

This MicroScope functionality is illustrated in Figure 5, using the results from the *Acinetobacter baylyi* ADP1 genome annotation and transcriptome analysis. This data was obtained in the context of a multi-omics approach combining RNA-seq transcriptomics, LC/MS-based metabolomics, systematic phenotyping of the complete collection of single-gene deletion mutants of *A. baylyi* ADP1 (40) and human expertise on its gene function using the MicroScope platform. This soil bacterium is able to metabolize a wide range of plant secondary metabolites as the sole source of carbon and energy. Shifting the carbon source from succinate to quinate induced not only a specific transcriptional response necessary to catabolize the new carbon source, but also a tremendous upheaval of the transcription pattern affecting the expression of more than 12 % of the total number of genes, most of them being of unknown function (41). These perturbations were ultimately reflected in the metabolome, in which the concentration of about 50% of the LC/MS-detected metabolites was impacted. Moreover, numerous unidentified metabolites were present only in quinate-grown cells suggesting that these exclusive compounds are synthesized by upregulated genes of unknown function, these latter being probably involved in unsuspected enzymatic reactions that await discovery. For this purpose, integrative capabilities of the MicroScope platform (comparative tools, queries, etc.) can be used to refine and improve gene annotation in a metabolic context (21,22).



Acinetobacter baylyi str. ADP1

Mapping Overview | Raw Read Count | Analysis

Organism : Acinetobacter baylyi str. ADP1

Mapper : SSAHA2 version 2.5.1 SSAHA2Launcher -o sam -S 25 -t solexa   Protocol : strand specific dir mRNAseq SE

MA quinate 20mM cl2

Total read number	25974068	100 %
Nb of unmapped reads	11173626	43.02 %
Nb of reads mapped at least once	14800442	56.98 %
Nb of reads mapped on rRNA	3536711.3260000004	13.62 %
Nb of reliable reads	14800442	56.98 %
Nb of reads kept on chromosome ACIAD	14800442	56.98 %
Total reads mapped on genomic objects (except rRNA) into chromosome ACIAD	20854077	80.29 %
Download forward / reverse wig file	Download	

DESeq Analysis [232]   Export to Gene Cart   Launch MeV   Launch IGV   MicroCyc Overview

Showing 1 to 10 of 10 results (filtered from 232 total results)   Show 10 Results

	MoveTo	MoveTo IGV	Label	Type	Name	Product	Begin	End	Length	Frame	MA quinate 20mM/MA succinate 20mM (B/A)		
											normalized average read count	log2 fold change	adjusted pvalue (FDR)
□			ACIAD1702	CDS	pcaU	regulatory protein for pca operon (activator)	1708109	1708945	837	-1	741	3.88	3.38e-50
□			ACIAD1708	CDS	pcaD	3-oxoadipate enol-lactonase I (Enol-lactone hydrolase I) (Beta-ketoadipate enol-lactone hydrolase I)	1713160	1713960	801	+1	3606	7.25	1.06e-145
□			ACIAD1712	CDS	pcaG	protocatechuate 3,4-dioxygenase alpha chain (3,4-PCD)	1716541	1717170	630	+1	1.59e+4	7.83	5.19e-189
□			ACIAD1711	CDS	pcaH	protocatechuate 3,4-dioxygenase beta chain (3,4-PCD)	1715798	1716523	726	+2	3900	7.90	3.64e-148
□			ACIAD1710	CDS	pcaC	gamma-carboxymuconolactone decarboxylase (CMD)	1715372	1715776	405	+2	2443	8.30	2.15e-133
□			ACIAD1705	CDS	pcaJ	3-oxoadipate CoA-transferase subunit B (Beta- ketoadipate:succinyl-CoA transferase subunit B)	1709925	1710578	654	+3	199	8.32	2.83e-55
□			ACIAD1706	CDS	pcaF	beta-ketoadipyl CoA thiolase	1710591	1711796	1206	+3	647	8.56	4.40e-97
□			ACIAD1709	CDS	pcaK	4-hydroxybenzoate transporter (MFS superfamily)	1713983	1715356	1374	+2	4511	8.75	8.38e-153
□			ACIAD1707	CDS	pcaB	3-carboxy-cis,cis-muconate cycloisomerase (3-carboxymuconate lactonizing enzyme) (CMLE)	1711808	1713163	1356	+2	9.16e+4	8.81	3.03e-210
□			ACIAD1704	CDS	pcaI	3-oxoadipate CoA-transferase subunit A (Beta- ketoadipate:succinyl-CoA transferase subunit A)	1709228	1709914	687	+2	714	10	3.71e-100

Showing 1 to 10 of 10 results (filtered from 232 total results)

Figure 5. Example of transcriptomic analysis using MicroScope: the *Acinetobacter baylyi* ADP1 pca/qui operon. *A. baylyi* ADP1 metabolic response when quinate is provided as the sole source of carbon and energy has been surveyed using RNA-seq experiments. The main route for the degradation of aromatic compounds to enter central metabolism is the ß-ketoadipate pathway through catechol and protocatechuate. Quinate is an alicyclic compound that feeds into the protocatechuate branch and, in ADP1, it is catabolized by 14 genes, 11 of which encode enzymatic activities (*pca* and *qui* genes). This figure shows the search for differentially expressed genes in the two conditions (quinate versus succinate) and the results confirmed the involvement of the genes known to be involved in quinate degradation (middle frame). Using the Integrative Genomics Viewer (IGV) software (directly available from MicroScope by clicking the 'launch IGV' button), transcripts coverage over ADP1 genome is displayed, together with genome annotations (lower frame). The pca/qui operon is clearly over expressed in the two experiments (duplicates) when *A. baylyi* ADP1 is grown on quinate instead of succinate. Differentially expressed genes listed in middle frame can be loaded into the Multi-experiment Viewer ('Launch MeV' button in the middle frame) software to perform either hierarchical or non-hierarchical clustering analyses. In this example, *pca* and *qui* genes belong to the same cluster, showing that they share the same transcriptional profile (lower frame). Finally up and downregulated genes may be mapped onto the AcinetoCyc metabolic network ('MicroCyc Overview' button in the middle frame). In the current use case, all but one gene belonging the pca/qui operon are highlighted in orange indicating that the whole Quinate Degradation I, Superpathway of salicylate degradation and Catechol Degradation III pathways are activated in quinate condition (lower frame).

ONGOING WORK AND FUTURE DIRECTIONS

Grools: reactive graph reasoning for genome annotation

When curating gene/protein annotations, biologists try to assign precise molecular functions by evaluating various bioinformatics predictions and taking into account background knowledge on the studied organism. This laborious task may lead to inconsistencies and uncertainties in the annotations due to the lack of experimental results and/or the difficulty to find a correct trade-off between sensitivity and specificity of predictions. To help biologists in this slow expert curation process, we are developing an expert system named Grools (Genomic rule object-oriented logic system) that evaluates the whole microbial genome annotations through biological processes like metabolic pathways. These processes are represented as a graph of biological prior knowledge that could be extracted from generalist databases like UniPathway (42) or Genome Properties (43). Using genome annotations and experimental observations (e.g. growth phenotypes from BIOLOG®, compound auxotrophy) as input, Grools applies a reactive graph reasoning to propagate knowledge and highlights confirmed, contradictory and missing annotations. This system is currently validated in the context of several ongoing projects and will be integrated in MicroScope as an annotation companion to guide biologists during the curation process.

MicroCloud: toward a Software as a Service (SaaS)

To face the challenge of Big Data in genomics and continue to efficiently annotate and compare prokaryotic genomes, a technological evolution of the MicroScope platform is absolutely required. Especially, it should increase the flexibility in scale and cost for the need of computation and storage and offers an 'on-demand' microbial genome annotation service. In the context of the French Institute of Bioinformatics (IFB) infrastructure, we are currently designing a version of the MicroScope platform using Cloud technologies to progressively switch into a Software as a Service (SaaS) distribution model. The work is focusing around the choice of the best High Throughput Computing (HTC) Cloud solution and the design of virtual appliances for the three components of the platform (i.e. the production system, the databases and the Web graphical user interface). The automatic annotation process will take advantage of a set of analysis tools and the know-how included in the MicroScope workflows. Moreover, the deployment of custom instances (i.e. with only organisms of interest) will provide faster analyses than the current single MicroScope instance in which the comparison of thousands of genomes is computationally expensive. This flexible distribution of MicroScope through Virtual Machines on the Cloud (i.e. distribution of pre-configured data analysis pipelines with all software dependencies for their execution) should ease the process of integrating NGS technologies and bioinformatic predictions, and thus, the diagnostics analysis.

GenomeOnRails: representation and visualization of pangenomes

Genomics approaches in microbiology now use thousands of genomes to analyze a given species in different environmental or medical contexts. From a more technical point of view, the existence of so many genomes poses a crucial computational problem to manage comparative genomics. So far, the emphasis has focused on the computation of coregenomes and the identification of SNPs. This approach resulted in remarkably detailed information about, for example, epidemic strains. Nevertheless, it gives very little information on the adaptive potential of bacteria as most adaptive events take place in the variable genome (i.e. genomic regions that are exchanged between strains by horizontal gene transfer). As recently suggested (44), a consensus representation of multiple genomes would provide a better analysis framework than using individual reference genomes. In this way, we are working, through the GenomeOnRails project, in a formal representation of a consensus pan-genome using a graph model where nodes represent gene families and edges chromosomal neighborhood information. We plan to integrate this kind of representation in MicroScope to facilitate comparative analysis and data visualization of thousands of strains.

Development of a MicroScope instance dedicated to metagenomics data analysis

At present time, only set of contigs or even reconstructed 'pseudo genomes' from environmental samples (MAGs for Metagenome-assembled genomes) can be analyzed in MicroScope with associated metadata following the MIMS standard (see 'MICROSCOPE ENHANCEMENTS SINCE 2013' section). We plan to provide additional services whose pipelines are currently developed in the context of the IFB national infrastructure. One of them will allow users to quantify the relative abundance of organisms in a metagenomic sample by mapping reads on MAGs already available in MicroScope. This tool will provide a dynamic view of a microbial community through time and in various environmental or laboratory conditions. However, one limit of this approach is the comprehensiveness of the reference MAGs which could prevent from correctly capturing the overall biodiversity of the studied samples. Indeed, for complex metagenomes, it could be difficult to define reference MAGs representative of one unique species since the genome fractions of strains accounting for variable regions may be highly different. A pan-genomic representation of MAGs, introduced in the previous paragraph, will therefore help to tackle this issue by agglomerating core regions while retaining the genomic diversity of each species in a single reference.

MICROSCOPE ACCESS AND DATA AVAILABILITY

Integration of new genomic data in MicroScope is open and free of charge for academics. MicroScope data is available for download on various file formats ('Search/Export→Download Data' functionality of the Web interface). Access is also granted at the programmatic level via a RESTful API (http://rest-microscope.genoscope.cns.fr).

ACKNOWLEDGEMENTS

The authors would like to thank all MicroScope users for their feedback, which helped greatly in optimizing and improving many functionalities of the system. We also thank the entire IT system team of Genoscope for its essential contribution to the efficiency of the platform.

FUNDING

French Government 'Investissements d'Avenir' programmes, namely FRANCE GENOMIQUE [ANR-10-INBS-09-08]; INSTITUT FRANCAIS DE BOINFORMATIQUE [ANR-11-INBS-0013].

Conflict of interest statement. None declared.

REFERENCES

1. Stephens,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J., Iyer,R., Schatz,M.C., Sinha,S. and Robinson,G.E. (2015) Big data: astronomical or genomical? *PLoS Biol.*, **13**, e1002195.
2. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
3. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
4. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
5. Keegan,K.P., Glass,E.M. and Meyer,F. (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol. Biol.*, **1399**, 207–233.
6. Mitchell,A., Alex,M., Francois,B., Guy,C., Hubert,D., ten Hoopen,P., Matthew,F., Sebastien,P., Simon,P., Maxim,S. *et al.* (2015) EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, D595–D603.
7. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
8. Richardson,E.J. and Watson,M. (2013) The automatic annotation of bacterial genomes. *Brief. Bioinform.*, **14**, 1–12.
9. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S. *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
10. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The national center for biotechnology information's protein clusters database. *Nucleic Acids Res.*, **37**, D216–D223.
11. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. *et al.* (2003) GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
12. Almeida,L.G.P., Paixao,R., Souza,R.C., da Costa,G.C., Barrientos,F.J.A., dos Santos,M.T., de Almeida,D.F. and Vasconcelos,A.T.R. (2004) A system for automated bacterial (genome) integrated annotation–SABIA. *Bioinformatics*, **20**, 2832–2833.
13. Chen,I.-M.A., Markowitz,V.M., Palaniappan,K., Szeto,E., Chu,K., Huang,J., Ratner,A., Pillay,M., Hadjithomas,M., Huntemann,M. *et al.* (2016) Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system. *BMC Genomics*, **17**, 307.
14. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.-Y., Cohoon,M., de Crécy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
15. Huss,J.W. 3rd, Orozco,C., Goodale,J., Wu,C., Batalov,S., Vickers,T.J., Valafar,F. and Su,A.I. (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, **6**, e175.
16. Mons,B., Ashburner,M., Chichester,C., van Mulligen,E., Weeber,M., den Dunnen,J., van Ommen,G.-J., Musen,M., Cockerill,M., Hermjakob,H. *et al.* (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol.*, **9**, R89.
17. Kelder,T., van Iersel,M.P., Hanspers,K., Kutmon,M., Conklin,B.R., Evelo,C.T. and Pico,A.R. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
18. Markowitz,V.M., Chen,I.-M.A., Chu,K., Pati,A., Ivanova,N.N. and Kyrpides,N.C. (2015) Ten years of maintaining and expanding a microbial genome and metagenome analysis system. *Trends Microbiol.*, **23**, 730–741.
19. Vallenet,D., Labarre,L., Rouy,Z., Barbe,V., Bocs,S., Cruveiller,S., Lajus,A., Pascal,G., Scarpelli,C. and Médigue,C. (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65.
20. 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.
21. Vallenet,D., Belda,E., Calteau,A., Cruveiller,S., Engelen,S., Lajus,A., Le Fèvre,F., Longin,C., Mornico,D., Roche,D. *et al.* (2013) MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.*, **41**, D636–D647.
22. Belda,E., Vallenet,D. and Médigue,C. (2015) Accurate microbial genome annotation using an integrated and user-friendly environment for community expertise of gene functions: the microscope platform. In: McGenity,TJ, Timmis,KN and Nogales,B (eds). *Hydrocarbon and Lipid Microbiology Protocols*. Springer Protocols Handbooks, Berlin Heidelberg, pp. 141–169.
23. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
24. Kanehisa,M., Minoru,K., Yoko,S., Masayuki,K., Miho,F. and Mao,T. (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
25. Carver,T., Harris,S.R., Berriman,M., Parkhill,J. and McQuillan,J.A. (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, **28**, 464–469.
26. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.
27. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
28. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thomson,N., Allen,M.J., Angiuoli,S.V. *et al.* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.*, **26**, 541–547.
29. Miele,V., Penel,S. and Duret,L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
30. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–W346.
31. Weber,T., Blin,K., Duddela,S., Krug,D., Kim,H.U., Bruccoleri,R., Lee,S.Y., Fischbach,M.A., Müller,R., Wohlleben,W. *et al.* (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**, W237–W243.

32. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
33. Barbe,V., Cruveiller,S., Kunst,F., Lenoble,P., Meurice,G., Sekowska,A., Vallenet,D., Wang,T., Moszer,I., Medigue,C. *et al.* (2009) From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. *Microbiology*, **155**, 1758–1775.
34. Belda,E., Sekowska,A., Le Fèvre,F., Morgat,A., Mornico,D., Ouzounis,C., Vallenet,D., Médigue,C. and Danchin,A. (2013) An updated metabolic view of the Bacillus subtilis 168 genome. *Microbiology*, **159**, 757–770.
35. Belda,E., van Heck,R.G.A., Lopez-Sanchez,M.J., Cruveiller,S., Barbe,V., Fraser,C., Klenk,H.-P., Petersen,J., Morgat,A., Nikel,P.I. *et al.* (2016) The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. *Environ. Microbiol.*, **18**, 3403-3424.
36. Morgat,A., Axelsen,K.B., Lombardot,T., Alcantara,R., Aimo,L., Zerara,M., Niknejad,A., Belda,E., Hyka-Nouspikel,N., Coudert,E. *et al.* (2014) Updates in Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **43**, D459–D464.
37. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
38. Supply,P., Marceau,M., Mangenot,S., Roche,D., Rouanet,C., Khanna,V., Majlessi,L., Criscuolo,A., Tap,J., Pawlik,A. *et al.* (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of mycobacterium tuberculosis. *Nat. Genet.*, **45**, 172–179.
39. Plucain,J., Hindré,T., Le Gac,M., Tenaillon,O., Cruveiller,S., Médigue,C., Leiby,N., Harcombe,W.R., Marx,C.J., Lenski,R.E. *et al.* (2014) Epistasis and allele specificity in the emergence of a stable polymorphism in Escherichia coli. *Science*, **343**, 1366–1369.
40. 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.
41. Stuani,L., Lechaplais,C., Salminen,A.V., Ségurens,B., Durot,M., Castelli,V., Pinet,A., Labadie,K., Cruveiller,S., Weissenbach,J. *et al.* (2014) Novel metabolic features in Acinetobacter baylyi ADP1 revealed by a multiomics approach. *Metabolomics*, **10**, 1223–1238.
42. Morgat,A., Coissac,E., Coudert,E., Axelsen,K.B., Keller,G., Bairoch,A., Bridge,A., Bougueleret,L., Xenarios,I. and Viari,A. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.
43. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
44. Chan,A.P., Sutton,G., DePew,J., Krishnakumar,R., Choi,Y., Huang,X.-Z., Beck,E., Harkins,D.M., Kim,M., Lesho,E.P. *et al.* (2015) A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. *Genome Biol.*, **16**, 143.