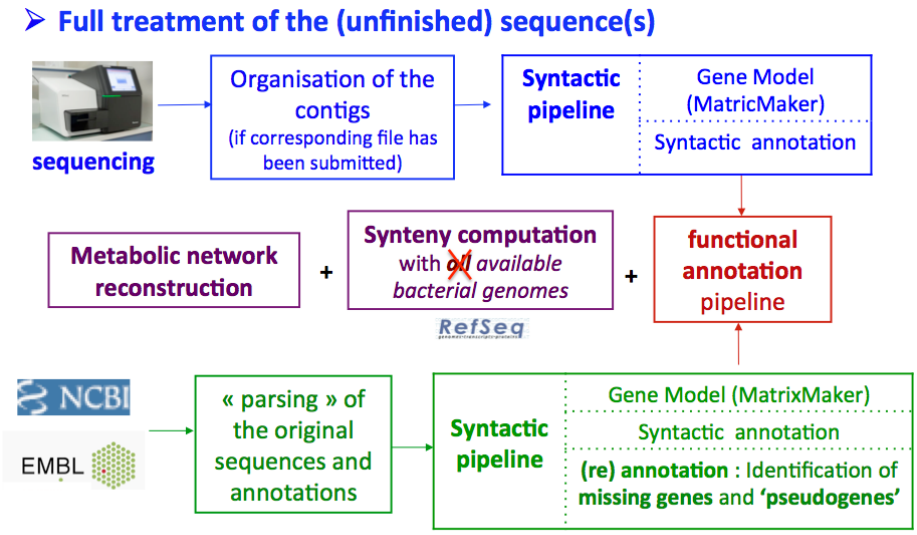
2014 MicroScope User Survey Results

In addition to the results gathered here, we present an analysis of several important points addressed in this survey:

Why using the MicroScope platform (*i.e.*, instead of other tools like the RAST server)?

First of all, if you just need a quick annotation of your bacterial genome (*i.,e.*, gene prediction and functional assignation of gene function using blast on protein families (COG, FIGfam, etc), and/or domain databanks, and/or generalist databanks, you don't need MicroScope.
User interest in the platform relies on:

1) The power of the automated analyses organized in the MicroScope workflows (**25 workflows to date, gathering 62 different software**, 15 of which have been developed in the LABGeM team) – The annotation workflow could be summarized as follow:



2) The availability of the results of each tool in the PkGDB database (**about 2 GB per genome**) and in the Gene Editor, results which can be queried using the "Search by keyword" functionality available in the 'Search/Export' menu.
3) The tools and graphical interfaces allowing users to explore the annotation and to perform expert curation of gene function in a collaborative way.

=> Indeed, our main objective (and the reason why MicroScope is developed) is to **offer an integrated environment for the exploration and the expertise of bacterial genome annotations**. In this context, our four-day training course (https://www.genoscope.cns.fr/agc/microscope/training) focuses on the use of the various functionalities implemented into the MicroScope platform.

Several required functionalities are already implemented in the platform

This survey (and, in a lesser extend, the 2012 survey) showed that some MicroScope functionalities are not well known or even ignored. We are currently working on the on-line tutorial to update its content and to make it easier to use. In addition a FAQ page will be prepared and updated according to the regular questions that have been asked using our contact address: mage@genocope.cns.fr.

From the last survey, the following functionalities should be highlighted:

Download data ('Search/Export' Menu), this functionality allows users to retrieve the sequence(s), annotations (in the most common files format), and the metabolic network (BioCyc format) of one specific organism. Moreover, non-coding regions or a precise region (defined or not by a CDS and its flanking sequence) can be extracted (e.g., to design specific oligos).
We are fully aware that data export is really crucial for the integration of Microscope with further/other analyses that are user-specific: thus, all the interactive tables (*i.e.*, results of a query/tool) can be downloaded using the first icon available in the top table row. Using a tab-delimited format, the content of the table is copied in a buffer and can then be pasted in an excel file for example.
Finally, results of a query and/or of a method can be exported into a gene cart (see the next point).
We have started to work on the development of an API to query data in the context of a European project (www.microme.eu) and we will continue in this direction.

"Gene carts" functionality ('User Panel' Menu): the content of all the created gene carts can be visualized using the "Gene carts" interface, which also allows users to make intersection, union, difference of several gene carts. Corresponding nucleic and/or protein sequences can be downloaded in a fasta file format, and multiple alignments of sequences can be performed using the JalView application (http://www.jalview.org/).

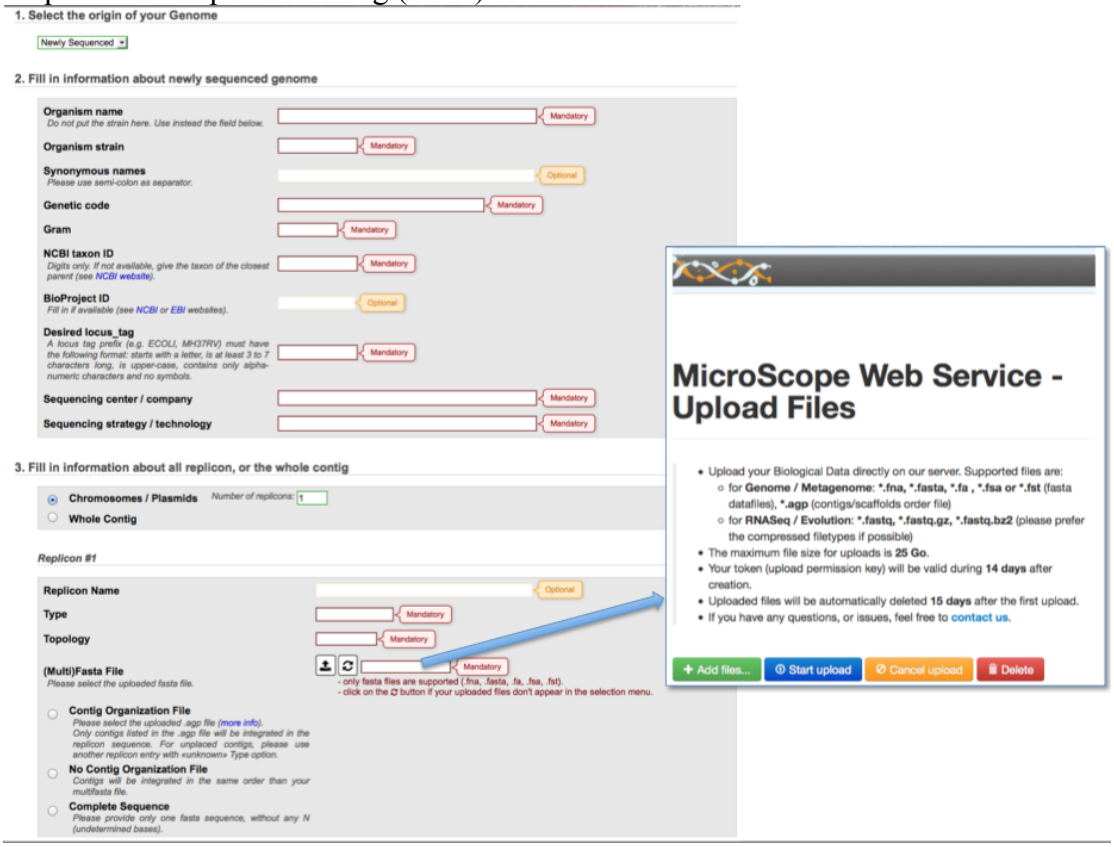
"Pan/core genome" functionality ('Comparative Genomics' Menu): from a set of selected strains (up to 200) the core and the variable genome of each strain are dynamically computed and can be downloaded (fasta or csv file format) or exported into a Gene cart.

Permanent links to other tools (in the "Gene Editor"): when results are based on other tools/resources (Uniprot, PRIAM, COG, FigFam, InterPro, Metacyc, etc) the links to these other tools/resources are always available in the table, which summarizes the results obtained with each method.
Moreover, permanent links are available with KEGG and BioCyc resources ('Metabolism' Menu).

Adding/uploading bacterial genomes (public or private)*:* at present time, upload of a new bacterial genome (public or private) is performed via a new (free) MicroScope service:
https://www.genoscope.cns.fr/agc/microscope/about/services.php

⇨ the automatic annotation process shown in the Figure above (first section of this document) will be performed (including the comparison with all public genomes available in the MicroScope database).

Sequences are uploaded using (multi)fasta files:



Draft genomes can also be uploaded with an AGP file for contig organization. Separation between scaffolds and contigs will then be represented by black and blue bars in the MaGe cartographic representation ('Genome Browser').

Working with pseudogenes: most of the tools are able to work with pseudogenes. That's obviously the case for the synteny computation, but also for specific tools like core/pan genome computation or for the comparison of metabolic pathways in several organisms ("Metabolic phyloprofile" in the 'Metabolism' Menu). In these latter tools, you can choose if you want to consider pseudogenes during the comparison process using a dedicated check box.

Searching for regulatory motifs ("Blast & pattern searches" functionality in the 'Search/Export' Menu): this interface uses a fasta sequence OR a Prosite pattern format as input.

Linking gene expression and metabolic pathways ('Transcriptomics' Menu) Results of the "Differential Expression Analysis" are listed in a table containing, at the bottom, several options: "Export to Gene Cart", "Launch MeV" (clustering of genes according to their level of expressivity in various experimental conditions), "Launch IGV" (read coverage on the genomic sequence), and "MicroCyc overview" which

shows genes that are up- or down-regulated on the metabolic pathways predicted with Pathologic/MetaCyc for the studied organism (MicroCyc PGDB).
Similarly, we plan to add colour coding for fold increase or decrease on the KEGG metabolic maps.

About "missing" bacterial genomes in PkGDB and synteny computations

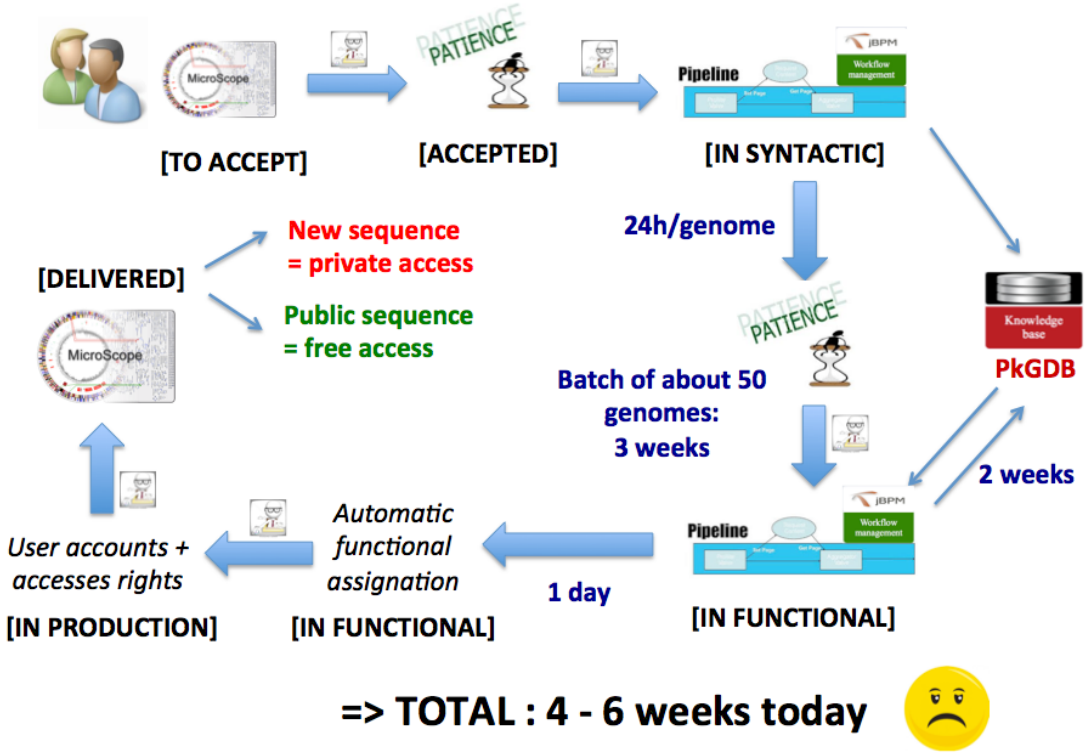
Since the opening of the MicroScope services, integration of bacterial and archaeal genomes (newly sequenced genomes or available genomes in public databank) is closely linked to our user projects/requests (https://www.genoscope.cns.fr/agc/microscope/about/services.php). If, 10 years ago, it was easy to add available close bacterial genomes to a new sequenced strain, this is clearly not the case today. There are about 50,000 sequenced prokaryotic genomes, most of them being in permanent draft status (between 3,000-3,500 are completely finished).

To keep the current efficiency of MicroScope and of the dynamic queries performed on PkGDB, we do not plan to integrate all these prokaryotic genomes. Following user requests, our aim is much more to focus on representative strains and a customized selection of other very close strains depending of the organism of interest. Also, we work on a new data model to represent the core/pan genome of a given species avoiding the redundancy in functional annotation computations and result storage of the genes belonging to the core genome. Together, new graphical views will be developed to navigate in the pan-genome.

Actual slow process in integrating bacterial genomes and/or updating blast results

Integration of prokaryotic genomes in MicroScope is clearly a too long process at present time (between 4 and 6 weeks) as the IT infrastructure of the MicroScope platform is shared with the other needs of the Genoscope centre (the overall process of genome integration into MicroScope, and the global timing of each step, is given in the figure below). There are clearly two bottlenecks:

- Functional annotation workflows for the batch of about 100 genomes: the process should evolve toward an "on demand" analysis of each submitted genome (see below)
- Integration of analysis results into the PkGDB database: technical improvements are currently addressed (*i.e.*, switch to another database engine).



Indeed to face the challenge of Big Data in genomics and continue to efficiently annotate and compare prokaryotic genomes, an IT evolution of the platform is absolutely required. Especially it should increase the flexibility in scale and cost for the needs of computation and storage, and offer a rapid annotation service. In the context of National projects, we are starting to design a version of the MicroScope platform using Cloud technologies to progressively switch into a Software as a Service (SaaS) distribution mode. This work focuses around the design of virtual appliances for the three components of the platform (*i.e.* the workflows of the production system, databases and Web graphical user interfaces) and the choice and implementation of the best HPC Cloud solution for highly flexible on demand computing capabilities of the service.

One aim of these technological developments is to deploy multiple instances of the MicroScope platform on internal and external infrastructures using a shared software framework that will enable common developments on the core MicroScope modules and specific developments on dedicated tools (*e.g.* for microbial pathogens). To start, the LABGeM team plan to closely work in strong collaboration with scientists at Pasteur Institute (C3BI) and with the Institut Français de Bioinformatique (IFB) core team. The latter already propose to use virtualization technologies to design and standardize bioinformatics pipelines.

What about the future MicroScope economic model?

To financially help the MicroScope platform, collaborative grant requests seem to be the preferred solution (see). However, today this kind of project represents less than 5% of the total integrated bacterial genomes. Financial involvement of our institutes to guarantee the sustainability of the resources has been suggested too.

This important point (sustainability, in general, of resources dedicated to important services for the research community) is currently addressed in French National infrastructures (*i.e.,* “France Genomique” and “Institut Français de Bioinformatique”) and European infrastructure (*i.e*., ELIXIR). We are confident that a common solution will be found rapidly.