Pixel: a digital lab assistant to integrate biological data in multi-omics projects

Thomas Denecker[1,*], William Durand[2,*], Julien Maupetit[2,*], Charles Hébert[3], Jean-Michel Camadro[4], Pierre Poulain[4,*,δ] and Gaëlle Lelandais[1,*,δ]

[1] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Gif-sur-Yvette cedex, France
[2] TailorDev SAS, Clermont-Ferrand, France
[3] BIOROSETICS, Houilles, France
[4] Institut Jacques Monod (IJM), UMR7592, CNRS et Université Paris Diderot, Paris, France
* These authors contributed equally
δ Corresponding authors: pierre.poulain@univ-paris-diderot.fr ; gaelle.lelandais@u-psud.fr

ABSTRACT

Background. In biology, high-throughput experimental technologies, also referred as "omics" technologies, are increasingly used in research laboratories. Several thousands of gene expression measurements can be obtained in a single experiment. Researchers are routinely facing the challenge to annotate, store, explore and integrate all the biological information they have at their disposal. We present here the Pixel web application (Pixel Web App), an original digital assistant to help people involved in a multi-omics biological project.

Methods. The Pixel Web App is built with open source technologies and hosted on the collaborative development platform GitHub (https://github.com/Candihub/pixel). It is written in Python using the Django framework and stores all the data in a PostgreSQL database. It is developed in the open and licensed under the BSD 3-clause license. The Pixel Web App is also heavily tested with both unit and functional tests, a strong code coverage and continuous integration provided by CircleCI. To ease the development and the deployment of the Pixel Web App, Docker and Docker Compose are used to bundle the application as well as its dependencies.

Results. The Pixel Web App offers researchers an intuitive way to annotate, store, explore and mine their multi-omics results. It can be installed on a personal computer or on a server to fit the needs of many users. In addition, anyone can enhance the application to better suit their needs, either by contributing directly on GitHub (encouraged) or by extending Pixel on their own. Unlike other bioinformatics platforms like Galaxy, the Pixel Web App does not provide any computational programs to analyze the data. Still, it allows to rapidly integrate existing results and thus holds a strategic position in the management of research data.

Introduction

In biology, high throughput (HT) experimental technologies - also referred as “omics” - are routinely used in an increasing number of research teams. Financial costs associated to HT experiments have been considerably reduced in the last decade [1] and the trend in HT sequencing (HTS) is now to acquire benchtop machines designed for individual research laboratories (for instance Illumina NextSeq500 or Oxford Nanopore Technologies MinION, [2]). The number of HT applications in biology has grown so rapidly in the past decade that it is hard to not feel overwhelmed [3][4]. It seems possible to address in any organism, any biological question through an “omics” perspective, providing the right HT material and method are found. If HTS is often put at the forefront of “omics” technologies (essentially genomics and transcriptomics, [5]), other technologies must be considered. Mass spectrometry (MS) for instance, enables HT identification and quantification of proteins (proteomics). Metabolomics and lipidomics are other derived applications of MS to characterize quantitative changes in small-molecular weight cellular components [6]. Together, they all account for complementary “omics area” with the advantage to quantify distinct levels of cellular components (transcripts, proteins, metabolites, etc.).

Integration of datasets issued from different HT technologies (termed as multi-omics datasets) represents a challenging task from a statistical and methodological point of view [7]. It implies the manipulation of two different types of data. The first type is the “primary data”, which correspond to raw experimental results. It can be FASTQ files for sequencing technology [8] or mzML files for MS [9]. These files can be stored in public repositories such as SRA [10], GEO [11], PRIDE [12] or PeptideAtlas [13]. Analyses of primary data rely on standard bioinformatics protocols that for instance, perform quality controls, correct experimental bias or convert files from a specific format to another. A popular tool to analyse primary data is Galaxy [14], which is an open web-based platform. "Secondary data" are produced upon analysis of primary data. It can be the counts of reads per genes for HTS results or the abundance values per proteins for MS results. In multi-omics datasets analysis, combining secondary data is essential to answer specific biological questions. It can be typically, the identification of differentially expressed genes (or proteins) between several cell growth conditions from transcriptomics (or proteomics) datasets, or the identification of cellular functions that are over-represented in a list of genes (or proteins). In that respect, secondary data can be analysed and re-analysed within a multitude of analytical strategies, introducing the idea of data analysis cycle. The researcher is thus constantly facing the

challenge to properly annotate, store, explore, mine and integrate all the biological data he/she has at his/her disposal in a multi-omics project. This challenge is directly related to the ability to extract as much information as possible from the produced data, but also to the crucial question of doing reproducible research.

A Nature's survey presented in 2016 indicates that more than 70% of the questioned researchers already experienced an impossibility to reproduce published results, and more than half of them were not able to reproduce their own experiments [15]. This last point is intriguing. If experimental biology can be subjected to random fluctuations hardly difficult to control, computational biology should not. Running the same software on the same input data is expected to give the same results. In practice, replication in computational science is harder than people generally think (see [16] as an illustration). It requires to adopt good practices for reproducible-research on a daily basis, and not only when the final results are about to be published. Initiatives to improve computational reproducibility exists [17]–[21], and today it is clear that the data alone are not enough to sustain scientific claims. Comments, explanations, software source codes and tests are prerequisites to ensure that an original research can be replicated by anyone, anytime, anywhere.

We developed the Pixel web application (Pixel Web App) with these ideas in mind. It acts as a digital lab assistant to help the researchers involved in a multi-omics biological project, to collaboratively mine and integrate their HT data. The Pixel Web App does not perform analysis on the primary data. It is focused on annotation, storage and exploration of secondary data (see **Figure 1**). These explorations represent critical steps to answer biological questions and need to be carefully annotated and recorded to be further exploited in the context of new biological questions. The Pixel Web App helps the researcher to specify necessary information required to replicate multi-omics results. We added an original hierarchical system of tags, which allows to easily explore and select multi-omics results stored in the system and to integrate them for new interpretations. The Pixel Web App can be installed on any individual computer (for a single researcher for instance), or on a web server for collaborative work between several researchers or research teams. The entire software has been developed with high quality programming standards and complies to major rules of open-source development [22]. The Pixel project is available on GitHub at https://github.com/Candihub/pixel, where full source code and detailed documentation are provided. We present in this article the Pixel Web App design and implementation. We

provide a case study with the integration of results issued from transcriptomics and proteomics experiments performed in the pathogenic yeast *Candida glabrata*.

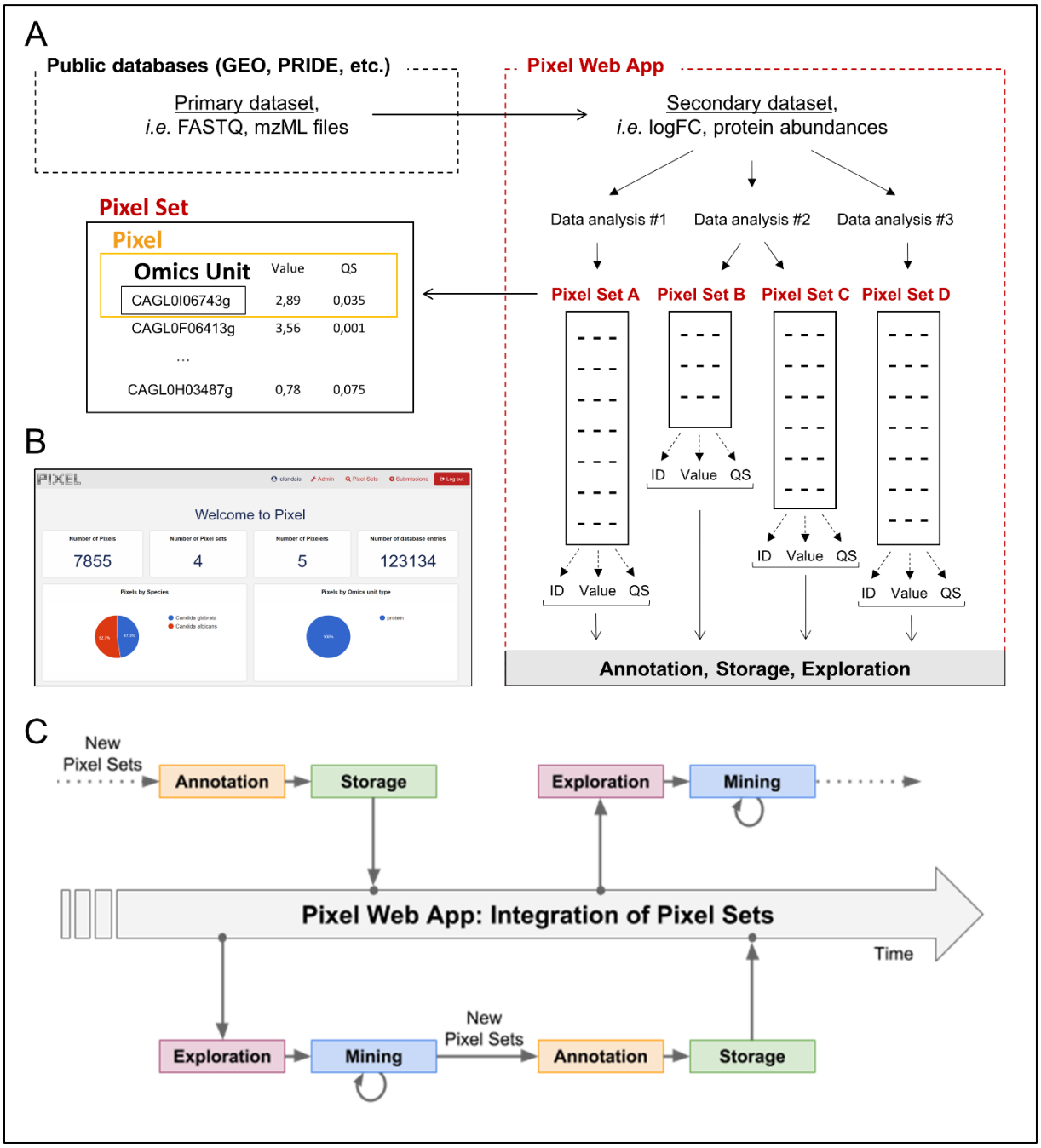

Figure 1: Dataset flow through the Pixel Web App. (A) Different types of datasets, which are managed in a multi-omics biological project. Primary and secondary datasets are two types of information arising from HT experimental technologies (see the section **Introduction**). Only secondary data and their associated Pixel Sets are stored in the Pixel Web App. Note that several Pixel Sets can emerge from multiple secondary data analyses. They comprise quantitative values (Value) together with quality scores (QS) for several hundred of different "Omics Units" elemesnts (for instance mRNA or proteins, see the main text). Omics Units are identified with a unique identifier (ID). (B) Screenshot of the home page of the Pixel web interface. (C) Schematic representation of the data analysis cycles that surrounds the integration of Pixel Sets in the Pixel Web App (see the main text).

Material and Methods

Stack overview

The Pixel Web App provides researchers an intuitive way to annotate, store, explore and mine their secondary data analyses, in multi-omics biological projects. It is built upon mainstream open source technologies (see **Figure 2**). Source code is hosted on the collaborative development platform GitHub[1] and continuous integration is provided by CircleCI[2]. More precisely, the Pixel Web App uses the Python Django framework. This framework is based on a model-template-view architecture pattern, and data are stored in a PostgreSQL[3] database. We have built a docker image for the Pixel Web App. Other containers, Nginx (to serve the Django application) and PostgreSQL rely on official docker images. Each installation / deployment will result in the creation / execution of three docker instances: one for the Pixel Web App, one for the PostgreSQL database and one for the Nginx web server. In case of multiple installations, each trio of docker instances is fully isolated, meaning that data are not shared across multiple Pixel Web App installations.

Technical considerations

- Docker images

The Pixel Web App is built on containerization paradigm (see **Figure 2**). It relies on Docker[4], *i.e.* a tool which packages an application and its dependencies in an image that will be run as a container. Docker helps developers to build self-contained images to run a software. These images are downloaded on the host system and used to build the Pixel Web App.

- Minimal configuration and dependencies

The Pixel Web App can be deployed on Linux and MacOS operating systems (OS). Deployment on Windows is possible, but this situation will not be described here. Minimal requirements are: *(i)* 64 bits Unix-based OS (Linux / MacOS), *(ii)* Docker community edition > v18, *(iii)* Internet access (required in order to download the Docker images) and *(iv)* [optional] a web server (Apache or Nginx) configured as a reverse proxy.

[1] https://github.com/
[2] https://circleci.com/
[3] https://www.postgresql.org/
[4] https://www.docker.com/

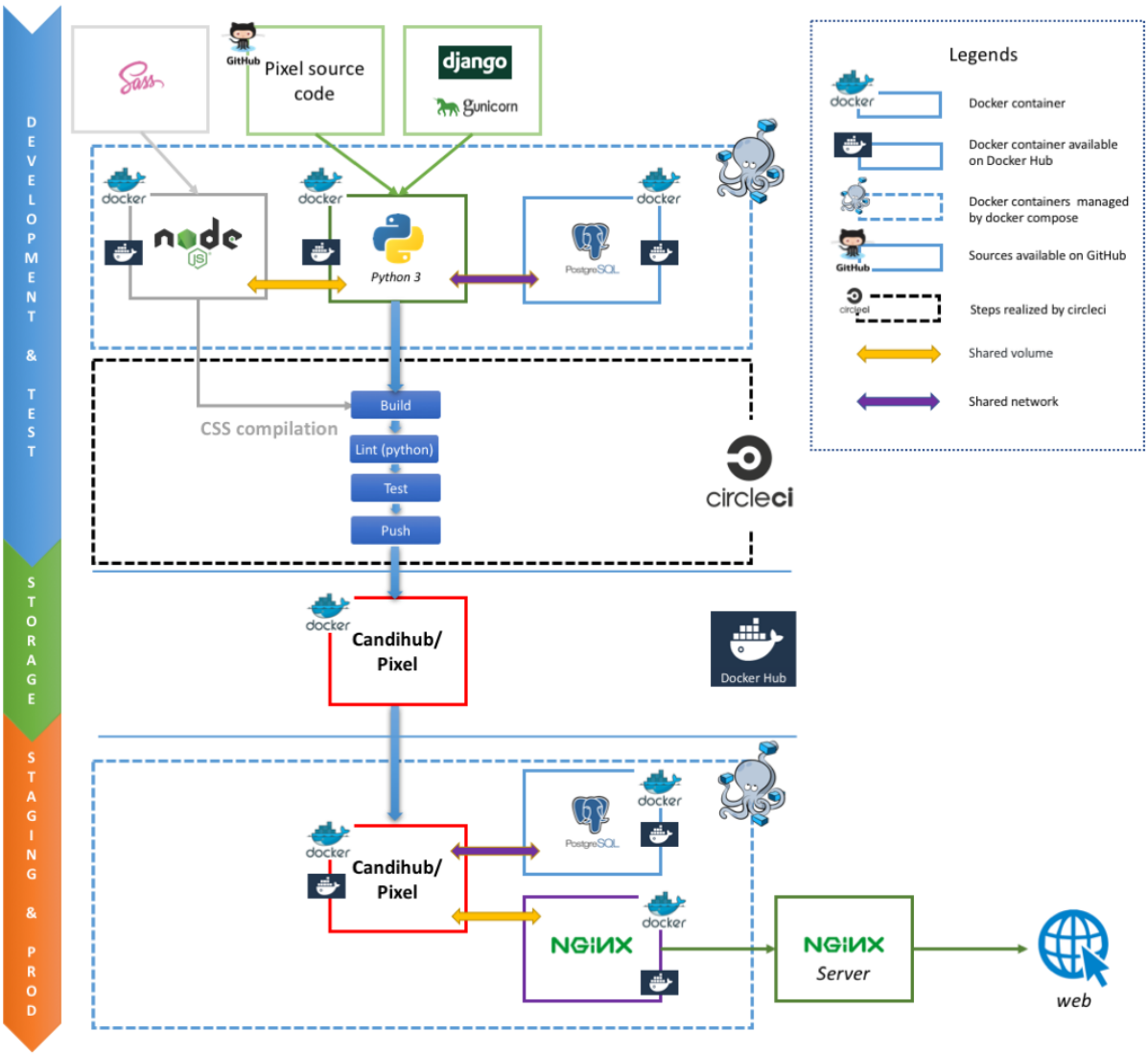

Figure 2: Stack overview of the Pixel Web App. Open source solutions used to develop Pixel are shown here. They are respectively used for the software development and test (blue section), the data storage (green section) and the web application for both staging and production (orange section).

Installation

A step-by-step tutorial to deploy the Pixel Web App can be found in the project repository[5] together with a deploy script. To summarize, this script runs the following steps:

- Pull a tagged image of Pixel (web, see docker-composer file),
- Start all instances (web, db and proxy) recreating the proxy and web instances. Collect all static files from the Django app. These files will be served by the proxy instance.
- Migrate the database schema if needed (to preserve existing data).

[5] https://github.com/Candihub/pixel/blob/master/docs-install/how-to-install.md

Note that further technical considerations and full documentation can be found on GitHub repository associated to the Pixel project[6].

Results

Definition of terms: Omics Unit, Pixel and Pixel Set

In the Pixel Web App, the term "Omics Unit" refers to any cellular component, from any organism, which is of interest for the user. The type of Omics Unit depends on the HT experimental technology (transcriptomic, proteomic, metabolomic, etc.) from which primary and secondary datasets were collected and derived (**Figure 1**A). In this context, classical Omics Units can be transcripts or proteins, but any other cellular component can be defined as, for instance, genomic regions with "peaks" in case of ChIPseq data analyses [23]. A “Pixel” refers to a quantitative measurement of a cellular activity associated to a single Omics Unit, together with a quality score (see **Figure 1**A). Quantitative measurement and quality score are results of statistical analyses performed on secondary datasets, *e.g.* search for differentially expressed genes [24]. A set of Pixels obtained from a single secondary data analysis of HT experimental results is referred as a “Pixel Set” (see **Figure 1**A). Pixel Sets represent the central information in the Pixel Web App and functionalities to annotate, store, explore and mine multi-omics biological data were designed according to this concept (see below).

Functionalities to annotate, store, explore and integrate Pixel Sets

Pixel Sets are obtained from secondary data analyses (see **Figure 1**A). Their manipulation with the Pixel Web App consists in *(i)* their annotation, *(ii)* their storage in a database, *(iii)* their exploration and *(iv)* their integration (or mining, see **Figure 1**C). This represents a cycle of multiple data analyses, which is essential in any multi-omics biological project. These different steps are detailed in the following.

- <u>Annotation of Pixel Sets</u>

Annotation of Pixel Sets consists in tracking important details of Pixel Set production. For that, Pixel Sets are associated with metadata, *i.e.* supplementary information linked to the Pixel Sets. We defined minimal information necessary for relevant annotations of Pixel Sets (see **Figure 3**). "Species", "Strain", "Omics Unit Type" and "Omics Area" are mandatory information that must be specified *before* a new Pixel Set submission (highlighted in blue, **Figure 3**). They refer to general information related to the multi-omics biological project on

[6] https://github.com/Candihub/pixel/tree/master/docs

which the researcher is working on: *(i)* the studied organism and its genetic background (Species and Strain, *e.g. Candida glabrata* and ATCC2001), *(ii)* the type of monitored cellular components (Omics Unit Type, *e.g.* mRNA, protein) and *(iii)* the nature of the experimental HT technology (Omics Area, *e.g.* RNA sequencing, mass spectrometry). All Omics Units must be declared in the Pixel Web App before new Pixel Set submission. They must be defined with a short description and a link to a reference database. "Experiment" and "Analysis" are Pixel Set mandatory information, input during the submission of new Pixel Sets in the Pixel Web App (highlighted in orange, **Figure 3**). They include respectively the detailed description of the experimental strategy that was applied to generate primary and secondary data sets (Experiment) and the detailed description of the computational procedures that were applied to obtain Pixel Sets from secondary data set (Analysis). Information regarding the researcher who performed the analyses is referred as "Pixeler".

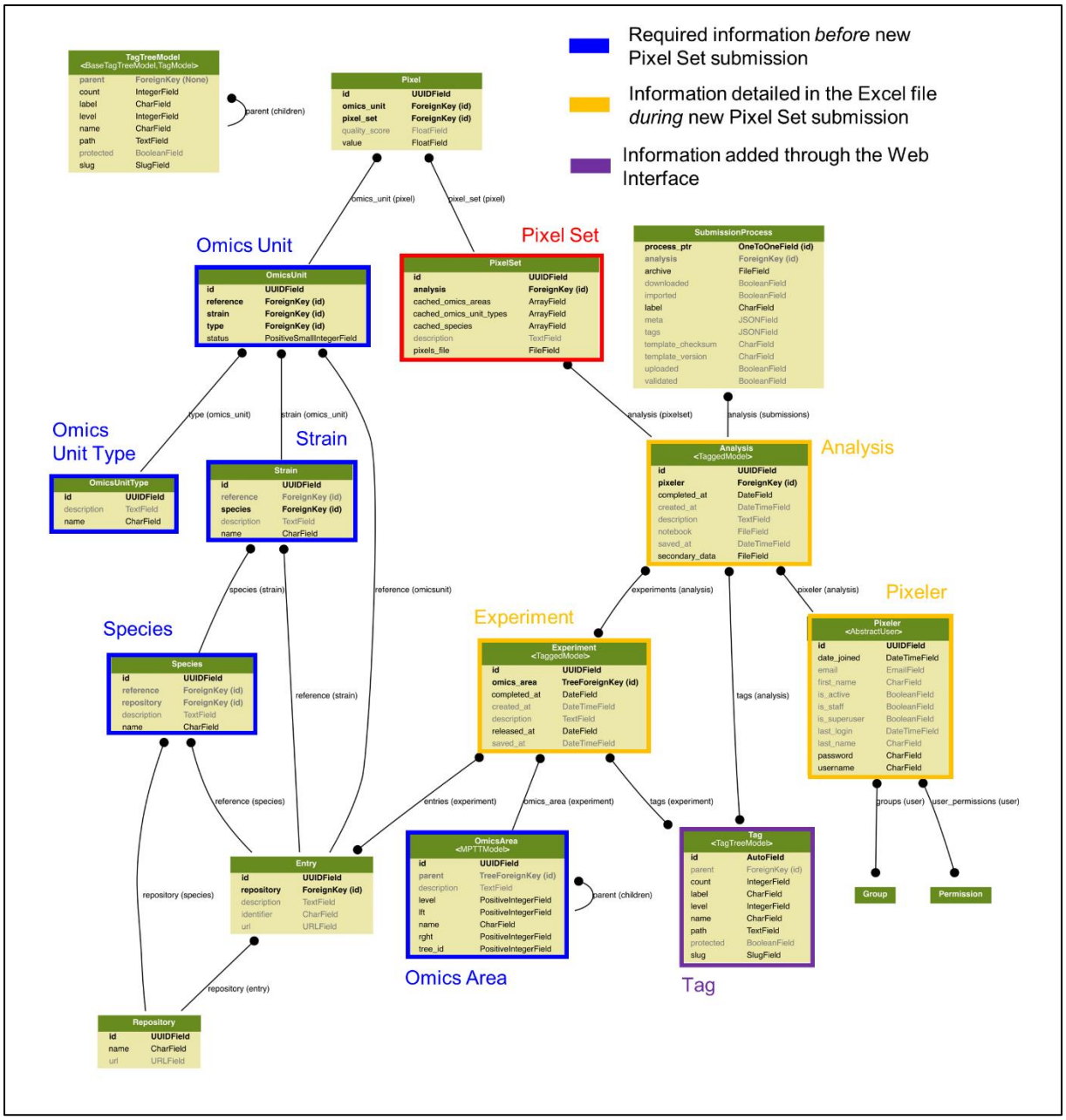

Figure 3: Data modelling in the Pixel Web App. The Pixel Set is the central information (see **Figure 1**A), the corresponding table in the model is highlighted in red. Information that is required *before* Pixel Set import in the Pixel Web App is surrounded in blue, whereas information required *during* Pixel Set import is highlighted in orange. Other tables are automatically updated during the Pixel Web App data analysis life cycle (see **Figure 1**C). Enlarge version of this picture together with full documentation is available online[7].

- <u>Storage of Pixel Sets in the database</u>

Importation of new Pixel Sets in the Pixel Web App requires the user to follow a workflow for data submission. It corresponds to six successive steps that are explained below (**Figure 4**A).

[7] https://github.com/Candihub/pixel/blob/master/docs/pixel-db.pdf

1. The "Download" step consists in downloading a template Excel file from the Pixel Web App (see **Figure 4**B). In this file, multiple-choice selections are proposed for "Species", "Strain", "Omics Unit Type" and "Omics Area" fields. These choices reflect what is currently available in the database and can be easily expanded. User must fill other annotation fields related to the "Experiment", "Analysis" and "Pixeler" information. The Excel file is next bundled into a ZIP archive with the secondary data file (in tab-separated values format), the user notebook (R markdown[8] or Jupyter notebook[9] for instance) that contains the code used to produce the Pixel Sets from the secondary data file.
2. The "Upload" step consists in uploading the ZIP file in the Pixel Web App.
3. The step "Meta" consists in running an automatic check of the imported file integrity (md5sum checks are performed, Excel file version is verified, etc.). Note that no information is imported in the database at this stage, but a careful inspection of all Omics Units listed in the submitted Pixel Sets is done. This is why Omics Units need to be pre-registered in the Pixel Web App (see previous section).
4. In "Annotation" step, the annotations of Pixel Sets found in the Excel file (see **Figure 4**C) are controlled and validated by the user.
5. Next, the "Tags" step is optional. It gives the opportunity to the user to add tags to the new Pixel Sets (see **Figure 4**C), that could be helpful for further Pixel Set explorations (see next section).
6. The final step "Import archive" consists in importing all Pixel Sets in the database, together with annotations and tags.

Note that the procedure of importing meta data as an Excel file has been inspired from the import procedure widely used in GEO [11].

[8] https://rmarkdown.rstudio.com/
[9] http://jupyter.org/

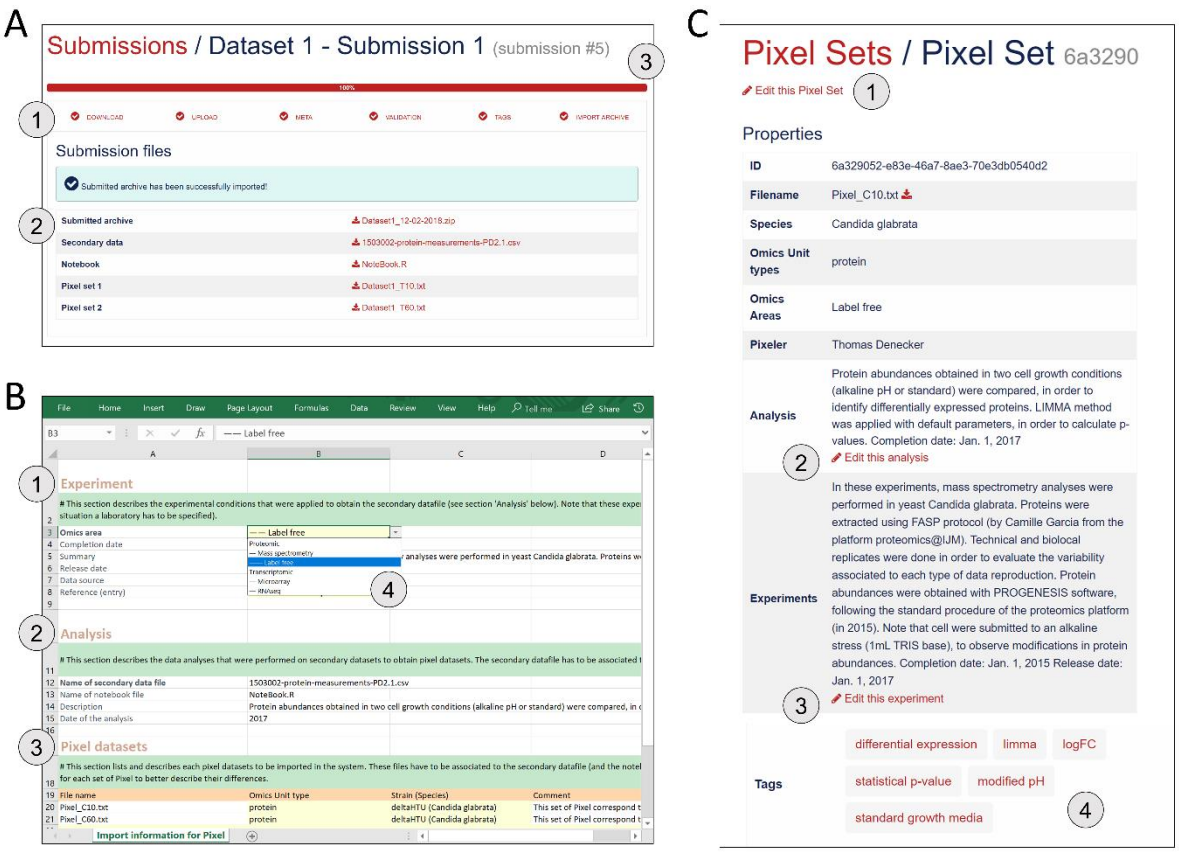

Figure 4: Procedure to import new Pixel Sets in the Pixel Web App. (A) New data-sets are submitted following a dedicated workflow that comprised 6 successive actions named "Download", "Upload", "Meta", "Validation", "Tags" and "Import archive" (see 1). Several files are required (see 2): the secondary data from which the Pixel Sets were calculated, the notebook in which the procedure to compute Pixel Sets from secondary data is described and the Pixel Set files (2 files in this example). A progression bar allows the user to follow the sequence of the submission process. (B) Excel spreadsheet in which annotations of Pixel Sets are written. Information related to the Experiment (see 1), the Analysis (see 2) and the Pixel datasets (see 3) is required. Note that this file must be downloaded at the first step of the submission process ("Download", see A), allowing several cells to be pre-filled with annotations stored in the database (see 4 as an illustration, with Omics area information). (C) All information filled in the Excel file (see B) is extracted and can be modified anytime through a dedicated web page as shown here. User can edit the Pixel Set (see 1), edit the analysis (see 2), edit the experiment (see 3) and add "Tags" (see 4). The Tags are of interest to further explore Pixel Sets in the Pixel Web App.

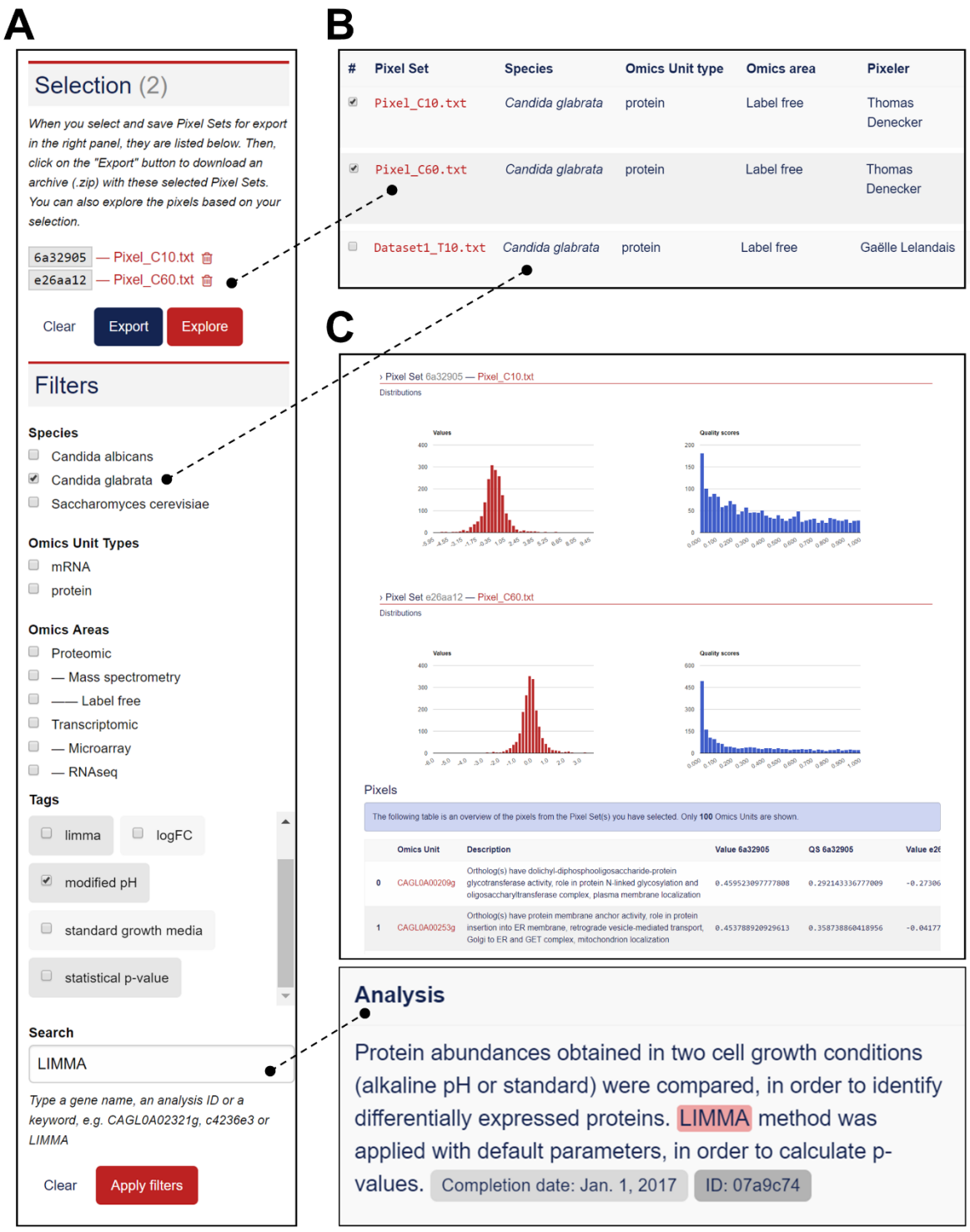

Figure 5 : Functionalities to explore the Pixel Sets stored in the Pixel Web App. (A) Screenshot of the exploration menu available *via* the web interface. (B) Screenshot of the table that comprises all Pixel Sets, which match the filter criteria (see A). Particular Pixel Sets can be selected here (for instance “Pixel_C10.txt” and “Pixel_C60.txt”). They will therefore appear in the "Selection" list (see A). (C) Screenshot of the web interface that gives detailed information for the selected subset of Pixel Sets (see A). Distribution of values and quality scores are shown and individual Omics Unit are listed at the bottom of the page.

- Exploration of Pixel Sets

The Pixel Web App aims to help researchers to mine and integrate multiple Pixel Sets stored in the system. We developed a dedicated web interface to explore all the Pixel Sets stored in a particular Pixel instance (see **Figure 5**). The upper part named "Selection" lists a group of Pixel Sets selected by the user for further explorations (**Figure 5**A). The middle part named "Filters" lists the Pixel database contents regarding the Species, Omics Unit Types, Omics Areas and Tags annotation fields. The user can select information (*Candida glabrata* and modified pH here), search and filter the Pixel Sets stored in the database (**Figure 5**B). The lower part is a more flexible search field in which keywords can be type. These keywords are search in the Analysis and Experiment detailed description fields as illustrated here with LIMMA. The web interface also comprised detailed information for the selected subset of Pixel Sets with for instance, distributions of values and quality scores and a list of individual Omics Unit shown at the bottom of the page (**Figure 5**C). Note that tags have been implemented to offer to the user a versatile yet robust annotation of Pixel Sets. They are defined during the import process, but they can be modified at any time through the Pixel web interface. Once searched, matching Pixel Sets are gathered in a table that can be exported.

A case study in the pathogenic yeast *Candida glabrata*

The yeast *Candida glabrata* (*C. glabrata*) is a fungal pathogen of human [25]. It has been reported as the second most frequent cause of invasive infections due to Candida species, *i.e.* candidemia, arising especially in patients with compromised immunity (HIV virus infection, cancer treatment, organ transplantation, etc.). Candidemia remains a major cause of morbidity and mortality in the healthcare structures [26], [27]. The genome of *Candida glabrata* has been published in 2004 [28]. Its size is 12.3 Mb with 13 chromosomes and is composed of ~5200 coding regions. Our research team is familiar with functional genomic studies in *C. glabrata*. In collaboration with experimental biologists, we published in the past ten years half dozen of articles, in which HT technologies were used [29]–[33]. In our lab, the Pixel Web App is installed locally and store all the necessary genomics annotations to manage any multi-omics datasets in this species.

As a case study, we decided to present how the Pixel Web App can be helpful to answer a specific biological question with only a few mouse clicks. As a biological question, we wanted to identify the genes in the entire *C. glabrata* genome: *(i)* which are annotated as involved in the yeast pathogenicity and *(ii)* for which the expression is significantly modified in response to an environmental stress induced by alkaline pH. Indeed, during a human host

infection, *C. glabrata* has to face important pH fluctuations (see [34]–[36] for more detailed information). Understanding the molecular processes that allow the pathogenic yeast *C. glabrata* to adapt extreme pH situations is therefore of medical interest to better understand host-pathogen interaction [36].

In a paper published in 2015, Linde *et al.* provided a detailed RNAseq based analysis of the transcriptional landscape of *C. glabrata* in several growth conditions, including pH shift experiments [37]. The primary dataset (RNAseq fastq files) is available in the Gene Expression Omnibus [11] under accession number GSE61606. The secondary dataset (log2 Fold Change values) is available in Supplementary Table S1 on the journal website[10]. A first Pixel Set (labelled A) was created from this secondary dataset, annotated and imported into our Pixel Web App instance, following the procedure previously described. The associated ZIP archive is provided as supplementary data along with the all the details related to the experiment set up and the analysis. The Pixel Set A thus illustrates how publicly available data can be managed with the Pixel Web App. In our laboratory, we performed mass spectrometry experiments that also include pH shift (unpublished data). Secondary dataset issued from these experiments leads to the Pixel Set B. Pixel Sets A and B comprise 5,253 Pixels and 1,879 Pixels (**Figure 6**).

Transcriptomics (Pixel Set A) and proteomics (Pixel Set B) are interesting complementary multi-omics information that can be easily associated and compared with the Pixel Web App. In that respect, tags allowed to rapidly retrieve them using the web interface, applying the keywords "Candida glabrata" and "alkaline pH" (**Figure 6**, Step 1). As we wanted to limit the analysis to the *C. glabrata* genes potentially involved in the yeast pathogenesis, a filter could be used to only retain the Omics Units for which the keyword "pathogenicity" is written in their description filed (see **Figure 6**, Step 2). As a result, a few numbers of Pixels were thus selected, respectively 17 in Pixel Set A and 6 in Pixel Set B. The last step consists in integrating the mRNA and protein information (see **Figure 6**, Step 3). For that a table comprising the multi-pixel sets can be automatically generated and easily exported (this table is provided as supplementary data). We present **Table 1** five genes for which logFC values were obtained both at the mRNA and the protein levels, and for which statistical p-values were significant (< 0.05). Notably two genes (CAGL0I02970g and CAGL0L08448g, lines 3 and 5 in **Table 1**) exhibited opposite logFC values, *i.e.* induction was observed at the mRNA

[10] https://academic.oup.com/nar/article/43/3/1392/2411170

level whereas repression was observed at the protein levels. Such observations can arise from post-translational regulation processes or from possible experimental noise, which could explain approximative mRNA or protein quantifications. In both cases, further experimental investigations are required. The three other genes (CAGL0F04807g, CAGL0F06457g and CAGL0I10516g, underlined in grey **Table 1**) exhibited multi-omics coherent results and significant inductions were observed at the mRNA and protein levels. Again, further experimental investigations are required to fully validated these observations. Still, it is worth noting that the gene CAGL0F04807g, is described as "uncharacterized" in the Candida Genome Database [11]. Considering that logFC values for this gene are particularly high (> 1), such an observation represents a good starting point to refine the functional annotation of this gene, clearly supporting the hypothesis that is has a role in the ability of *C. glabrata* to deal with varying pH situations.

[11] http://www.candidagenome.org/cgi-bin/locus.pl?locus=CAGL0F04807g&organism=C_glabrata_CBS138

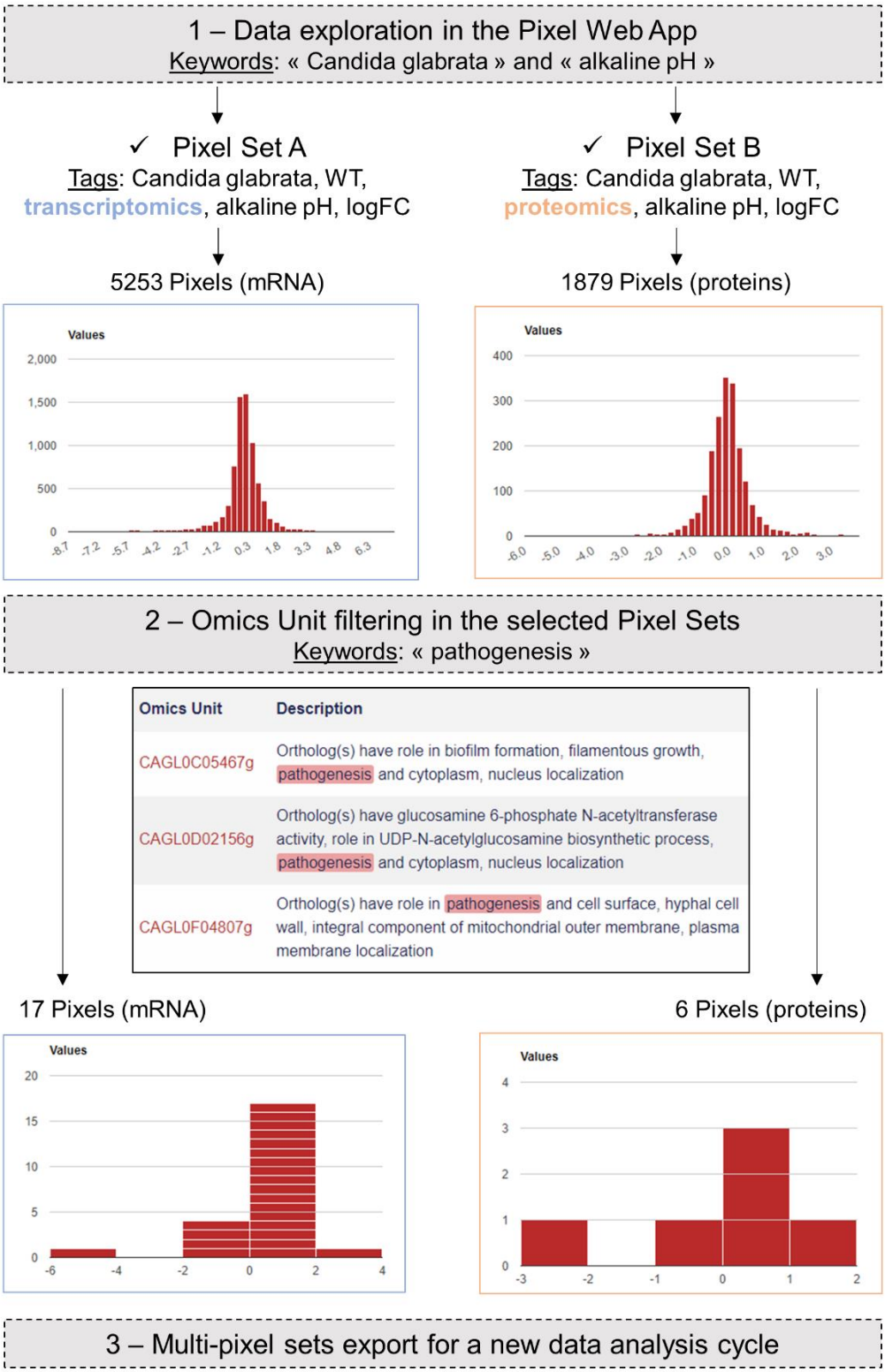

Figure 6: Case study in the pathogenic yeast *Candida glabrata*. Our Pixel Web App was explored with the keywords "Candida glabrata" and "alkaline pH". Two Pixel Sets were thus identified because of their tags. Two other tags were identical between the two Pixel Sets ("WT" and "logFC"), indicating that *(i) C. glabrata* strains are the same, *i.e.* Wild Type, and *(ii)* Pixel values are of the same type, *i.e.* log Fold Change. Notably Pixel Set A is based on transcriptomics experiments (RNAseq, see the main text), whereas Pixel Set B is based on proteomics experiments (mass spectrometry, see the

main text). Omics Unit were next explored using the keyword "pathogenesis" resulting in the identification of 17 Pixels (respectively 6 Pixels) in transcriptomics (respectively proteomics) results. They were combined and exported from the Pixel Web App, hence starting a new data analysis cycle.

Omics Unit	**Description**	**A**	**B**	**A (QS)**	**B (QS)**
1. CAGL0F04807g	Ortholog(s) have role in **pathogenesis** and cell surface, hyphal cell wall, integral component of mitochondrial outer membrane, plasma membrane localization	1,09	1,81	2,23E-19	7,31E-05
2. CAGL0F06457g	Ortholog(s) have role in fungal-type cell wall organization or biogenesis, mitochondrial outer membrane translocase complex assembly, **pathogenesis**, phospholipid transport, protein import into mitochondrial outer membrane	0,30	0,19	4,14E-02	2,65E-01
3. CAGL0I02970g	Ortholog(s) have delta14-sterol reductase activity and role in cellular response to drug, ergosterol biosynthetic process, filamentous growth of a population of unicellular organisms in response to biotic stimulus, **pathogenesis**	0,90	-2,64	4,65E-16	2,19E-05
4. CAGL0I10516g	Ortholog(s) have role in fungal-type cell wall organization, pathogenesis and cytoplasm, eisosome, integral component of plasma membrane, membrane raft localization	1,50	0,57	8,29E-60	1,16E-02
5. CAGL0L08448g	Ortholog(s) have role in actin cytoskeleton organization, eisosome assembly, negative regulation of protein phosphorylation, negative regulation of sphingolipid biosynthetic process and **pathogenesis**, more	1,67	-0,57	1,77E-75	7,04E-03

Table 1: Detailed information regarding the Omics Unit identified in the *C. glabrata* case study. The two first column give Omics Unit information as described in the Candida Genome Database [38]. All the description fields comprise the keyword "pathogenesis" (in bold). LogFC values measured in transcriptomic (Pixel Set A) and proteomic (Pixel Set B) experiments are shown in the third and fourth columns. Quality scores (QS) are following logFC values. They are p-values coming from the differential analysis of logFC replicates. The entire table of multi-pixel sets is available in supplementary data.

Software Availability

Pixel is released under the open-source 3-Clause BSD license (https://opensource.org/licenses/BSD-3-Clause). Its source code can be freely downloaded

from the GitHub repository of the project: https://github.com/Candihub/pixel. In addition, the present version of Pixel (4.0.4) is also archived in the digital repository Zenodo (https://doi.org/10.5281/zenodo.1434316).

Discussion / Conclusion

In this article, we introduced the principle and the main functionalities of the Pixel Web App. With this application, our aim was to develop a tool to support on a daily basis, the biological data integration in our multi-omics research projects. It is our experience that research studies in which HT experimental strategies are applied, require much more time to analyse and interpret the data, than to experimentally generate the data. Testing multiple bioinformatics tools and statistical approaches is a critical step to fully understand the meaning of a biological dataset and in this context, the annotation, the storage and the ability to easily explore the all results obtained in a laboratory can be the decisive steps to the success of the entire multi-omics project.

The data modelling around which the Pixel Web App was developed, has been conceived to find a compromise between a too detailed and precise description of the data (which could discourage the researchers of systematically use the application after each of their analyses) and a too short and approximate description of the data (which could prevent the perfect reproduction of the results by anyone). Also, a particular attention has been paid to allow heterogeneous data, *i.e.* different Omics Unit Type quantified in different Omics Area, to be stored in a coherent and flexible way. Unlike other bioinformatics platforms like Galaxy, the Pixel Web App does not provide any computational programs to analyse the data. Still, it allows to explore existing results in a laboratory and to rapidly combine them for further investigations (using for instance the Galaxy platform or any other data analysis tool).

Therefore, the Pixel Web App holds a strategic position in the data management in a research laboratory, *i.e.* as the starting point but also at the final point of all new data explorations. It also ensures data analysis reproducibility and gives a constant feedback regarding the frequency of the data analysis cycles; the nature of the import and export data sets as well as full associated annotations. It is thus expected that the content of different Pixel Web App instance will evolve with time, according to the type of information stored in the system and the scientific interests of a research team. This will be the case for our Pixel Web App instance[12] (from which the case study was obtained), which presently (July 2018) stored more

[12] https://pixel.candihub.eu

than 20,000 pixels, arising from transcriptomics (microarray and RNAseq technologies) or proteomics (mass spectrometry) technologies applied in two different pathogenic yeasts *Candida glabrata* and *Candida albicans*.

The Pixel Web App is freely available to any interested people. The initial installation on a personal workstation required IT support from a bioinformatician, but once this is done, all administration tasks can be performed through the Web Interface. This is of interest for user with a few technical skills. We chose to work exclusively with open source technologies and our GitHub repository is publicly accessible[13]. We thus hope that the overall quality of the Pixel Web App source code and documentation will be guaranteed over time, through the shared contributions of other developers.

References

[1] E. C. Hayden, "The $1,000 genome," *Nature*, vol. 507, no. 7492, p. 294, 2014.

[2] N. Blow, "A sequencer in every lab," *Biotechniques*, vol. 55, no. 6, p. 284, 2013.

[3] J. Hadfield and J. Retief, "A profusion of confusion in NGS methods naming," *Nat. Methods*, vol. 15, no. 1, pp. 7–8, 2018.

[4] "The data deluge," *Nat. Cell Biol.*, vol. 14, no. 8, pp. 775–775, Aug. 2012.

[5] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-Throughput Sequencing Technologies," *Mol. Cell*, vol. 58, no. 4, pp. 586–597, 2015.

[6] R. Smith, A. D. Mathis, D. Ventura, and J. T. Prince, "Proteomics, lipidomics, metabolomics: A mass spectrometry tutorial from a computer scientist's point of view," *BMC Bioinformatics*, vol. 15, no. Suppl 7, 2014.

[7] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Front. Genet.*, vol. 8, no. JUN, pp. 1–12, 2017.

[8] P. J. a Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.," *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–71, Apr. 2010.

[9] L. Martens *et al.*, "mzML—a Community Standard for Mass Spectrometry Data," *Mol. Cell. Proteomics*, vol. 10, no. 1, p. R110.000133, Jan. 2011.

[10] R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration, "The sequence read archive.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D19-21, Jan. 2011.

[11] E. Clough and T. Barrett, "The Gene Expression Omnibus Database.," *Methods Mol. Biol.*, vol. 1418, pp. 93–110, 2016.

[12] L. Martens *et al.*, "PRIDE: The proteomics identifications database," *Proteomics*, vol. 5, no. 13, pp. 3537–3545, Aug. 2005.

[13] https://github.com/Candihub/pixel

[13] F. Desiere *et al.*, “The PeptideAtlas project,” *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D655–D658, Jan. 2006.

[14] E. Afgan *et al.*, “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W3–W10, Jul. 2016.

[15] M. Baker, “1,500 scientists lift the lid on reproducibility.,” *Nature*, vol. 533, no. 7604, pp. 452–4, May 2016.

[16] O. Mesnard and L. A. Barba, “Reproducible and Replicable Computational Fluid Dynamics: It’s Harder Than You Think,” *Comput. Sci. Eng.*, vol. 19, no. 4, pp. 44–55, 2017.

[17] N. P. Rougier *et al.*, “Sustainable computational science: the ReScience initiative,” *PeerJ Comput. Sci.*, vol. 3, p. e142, Dec. 2017.

[18] N. A. Vasilevsky, J. Minnier, M. A. Haendel, and R. E. Champieux, “Reproducible and reusable research: are journal data sharing policies meeting the mark?,” *PeerJ*, vol. 5, p. e3208, Apr. 2017.

[19] V. Stodden, P. Guo, and Z. Ma, “Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals.,” *PLoS One*, vol. 8, no. 6, p. e67111, Jun. 2013.

[20] V. Stodden, J. Seiler, and Z. Ma, “An empirical analysis of journal policy effectiveness for computational reproducibility.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 11, pp. 2584–2589, Mar. 2018.

[21] R. D. Peng, “Reproducible research in computational science.,” *Science*, vol. 334, no. 6060, pp. 1226–7, Dec. 2011.

[22] M. Taschuk and G. Wilson, “Ten simple rules for making research software more robust.,” *PLoS Comput. Biol.*, vol. 13, no. 4, p. e1005412, Apr. 2017.

[23] J. Merhej, A. Frigo, S. Le Crom, J.-M. Camadro, F. Devaux, and G. Lelandais, “bPeaks: a bioinformatics tool to detect transcription factor binding sites from ChIPseq data in yeasts and other organisms with small genomes,” *Yeast*, vol. 31, no. 10, pp. 375–391, Oct. 2014.

[24] F. Seyednasrollah, A. Laiho, and L. L. Elo, “Comparison of software packages for detecting differential expression in RNA-seq studies,” *Brief. Bioinform.*, vol. 16, no. 1, pp. 59–70, Jan. 2015.

[25] M. Bolotin-Fukuhara and C. Fairhead, “Candida glabrata: a deadly companion?,” *Yeast*, vol. 31, no. 8, pp. 279–88, Aug. 2014.

[26] M. Pfaller *et al.*, “Epidemiology and outcomes of candidemia in 3648 patients: Data from the Prospective Antifungal Therapy (PATH Alliance) registry, 2004-2008,” *Diagn. Microbiol. Infect. Dis.*, vol. 74, no. 4, pp. 323–331, 2012.

[27] D. L. Horn *et al.*, “Epidemiology and Outcomes of Candidemia in 2019 Patients: Data from the Prospective Antifungal Therapy Alliance Registry,” *Clin. Infect. Dis.*, vol. 48, no. 12, pp. 1695–1703, 2009.

[28] B. Dujon *et al.*, “Genome evolution in yeasts.,” *Nature*, vol. 430, no. 6995, pp. 35–44, Jul. 2004.

[29] A. Thiébaut *et al.*, “The CCAAT-Binding Complex Controls Respiratory Gene Expression and Iron Homeostasis in Candida Glabrata,” *Sci. Rep.*, vol. 7, no. 1, 2017.

[30] J. Merhej *et al.*, “A Network of Paralogous Stress Response Transcription Factors in the Human Pathogen Candida glabrata,” *Front. Microbiol.*, vol. 7, no. May, pp. 1–16, 2016.

[31] J. Merhej *et al.*, “Yap7 is a Transcriptional Repressor of Nitric Oxide Oxidase in Yeasts, which arose from Neofunctionalization after Whole Genome Duplication,” *Mol. Microbiol.*, 2015.

[32] C. Goudot, C. Etchebest, F. Devaux, and G. Lelandais, “The Reconstruction of Condition-Specific Transcriptional Modules Provides New Insights in the Evolution of Yeast AP-1 Proteins.,” *PLoS One*, vol. 6, no. 6, p. e20924, Jan. 2011.

[33] G. Lelandais, V. Tanty, C. Geneix, C. Etchebest, C. Jacq, and F. Devaux, “Genome adaptation to chemical stress: clues from comparative transcriptomics in Saccharomyces cerevisiae and Candida glabrata.,” *Genome Biol.*, vol. 9, no. 11, p. R164, Jan. 2008.

[34] S. Brunke and B. Hube, “Two unlike cousins: *Candida albicans* and *C. glabrata* infection strategies,” *Cell. Microbiol.*, vol. 15, no. 5, pp. 701–708, May 2013.

[35] A. Ullah, M. I. Lopes, S. Brul, and G. J. Smits, “Intracellular pH homeostasis in Candida glabrata in infection-associated conditions.,” *Microbiology*, vol. 159, no. Pt 4, pp. 803–13, Apr. 2013.

[36] J. Linde *et al.*, “Defining the transcriptomic landscape of Candida glabrata by RNA-Seq.,” *Nucleic Acids Res.*, vol. 43, no. 3, pp. 1392–406, Feb. 2015.

[37] J. Linde *et al.*, “Defining the transcriptomic landscape of Candida glabrata by RNA-Seq,” *Nucleic Acids Res.*, vol. 43, no. 3, pp. 1392–1406, 2015.

[38] M. S. Skrzypek, J. Binkley, G. Binkley, S. R. Miyasato, M. Simison, and G. Sherlock, “The Candida Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D592–D596, 2017.