Databases and ontologies

The ReproGenomics Viewer: a multi-omics and cross-species resource compatible with single-cell studies for the reproductive science community

Thomas A. Darde[1,†], Estelle Lecluze[1,†], Aurélie Lardenois[1], Isabelle Stévant[2], Nathan Alary[1], Frank Tüttelmann[3], Olivier Collin[4], Serge Nef[2], Bernard Jégou[1], Antoine D. Rolland[1] and Frédéric Chalmel[1,*]

[1] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35000 Rennes, France, [2] Department of Genetic Medicine and Development, University of Geneva, 1211 Geneva, Switzerland, [3] Institute of Human Genetics, University of Münster, Münster, Germany, [4] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA/INRIA) - GenOuest platform, Université de Rennes 1, F-35042 Rennes, France,

[*]To whom correspondence should be addressed.

[†]These authors contributed equally to this work

Abstract

Motivation: Recent advances in transcriptomics have enabled unprecedented insight into gene expression analysis at a single-cell resolution. While it is anticipated that the number of publications based on such technologies will increase in the next decade, there is currently no public resource to centralize and enable scientists to explore single-cell datasets published in the field of reproductive biology.

Results: Here, we present a major update of the ReproGenomics Viewer (RGV), a cross-species and cross-technology web-based resource of manually-curated sequencing datasets related to reproduction. The redesign of RGV's architecture is accompanied by significant growth of the database content including several landmark single-cell RNA sequencing datasets. The implementation of additional tools enables users to visualize and browse the complex, high-dimensional data now being generated in the reproductive field.

Availability and implementation: The ReproGenomics Viewer resource is freely accessible at http://rgv.genouest.org. The website is implemented in Python, JavaScript, and MongoDB, and is compatible with all major browsers. Source codes can be downloaded from https://github.com/fchalmel/RGV.

Contact: frederic.chalmel@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Accepted manuscript

1 Introduction

Advances in ultra-high-throughput sequencing technologies, including increased accuracy, diverse applications, and decreased overall cost have significantly contributed to their democratization (Sims *et al.*, 2014). As these techniques became more popular, a myriad of datasets were published, and their corresponding raw sequence read data made available from public repositories, such as the Sequence Read Archive (Kodama *et al.*, 2012). While several databases devoted to reproductive biology were

developed to provide access to relevant microarray-based transcriptomics data (Lee *et al.*, 2010; Hsueh and Rauch, 2012; Luk *et al*., 2015; Hua *et al.*, 2015; Lardenois *et al.*, 2010; Schuster *et al.*, 2016; Lee *et al*., 2009; Zhang *et al.*, 2013), until recently no resources were able to manage ultra-high-throughput sequencing data. To deal with these data, in 2015 we introduced the ReproGenomics Viewer (RGV) (Darde *et al*., 2015), a web-based genomic resource for researchers of the reproductive science community. Its aim was to centralize and offer easy access to the published sequencing datasets (e.g., RNA-seq, ChIP-seq, and MNase-seq) that have accumulated in this field by overcoming the standard technical issues regarding data format, technology, and cross-species comparison. The system was based on implementation of a 'JBrowse genome browser' (Buels *et al.*, 2016). It also included unique features such as the conversion of genome coordinates between species for the direct comparison of data acquired in different organisms. The ReproGenomics Viewer was thus not only a multi- but also a cross-species resource for comparing genomics data.

More recently, the rapid emergence of novel technologies enabling the study of biological systems at a single-cell resolution (Linnarsson and Teichmann, 2016) has again presented serious concerns for the storage, mining, and visualization of such complex datasets (Raja *et al*., 2017). Several dedicated databases and webservers have been set up to host (Cao *et al*., 2017; Abugessaisa *et al*., 2018) and to explore (Lang *et al.*, 2015; DeTomaso and Yosef, 2016; Zhu *et al.*, 2017; Weinreb *et al*., 2018) single-cell RNA-sequencing (scRNA-seq) datasets.

In response to the incoming wave of studies at a single-cell resolution in reproductive biology, we would like to introduce the new release of RGV. Single-cell transcriptome profiling has already contributed substantially to deepening our understanding of reproductive processes in mammals, including gonad development (Stévant *et al*., 2018; Han *et al.*, 2018), spermatogonial stem cell biology (Li *et al*., 2017; Guo *et al*., 2017), and spermatogenesis (Lukassen *et al.*, 2018). The novel architecture of RGV has been totally redesigned and now offers several visualization tools for browsing complex high-dimensional sequencing data including single-cell genomic data. The database content has also grown dramatically and now covers 11 biological topics related to reproduction in 9 species, including 5 landmark scRNA-seq studies. To the best of our knowledge, RGV provides the most comprehensive resource of manually-curated ultra-high-throughput sequencing data in the field of reproduction currently available and thus enables researchers to explore this invaluable source of information.

2 Methods

2.1 Data curation and processing

As previously described (Darde *et al.*, 2015), the backbone of the ReproGenomics Viewer relies on a six-step pipeline to consistently download, curate, organize and process data within the system (Fig. 1, panels A-D). Briefly, *i*) PubMed, SRA, and GEO identifiers assigned to a given study are manually extracted (Fig. 1, panel A). These database entries are then used to mine other information relevant for describing the datasets, such as the species and the technologies. Manual curation is further required to standardize sample names across studies and to gather specific information about, for example, whether or not the RNA sequencing protocol is directional. Next, *ii*) raw data (sra files) are downloaded from the NCBI SRA repository. The sequencing reads (fastq files) are then *iii*) aligned to the appropriate reference genome with STAR v2.0 (Dobin and Gingeras, 2015) (Fig. 1, panel B). After read mapping *iv*), the deepTools suite (Ramírez *et al*., 2016) is used to convert the alignment files (bam files) into standard coverage tracks (bigWig files). During this process, coverage values are scaled by using a constant factor corresponding to the maximum number of mapped reads in all samples of a given dataset, divided by the number of mapped reads in the corresponding sample. The next step involves *v*) cross-species conversion, during which the genome coordinates of the scaled coverage tracks are converted to the other reference genomes indexed in the RGV system with CrossMap (Zhao *et al.*, 2014), based on the pairwise whole genome alignment files (chain files) provided by the UCSC genome browser (Kuhn *et al.*, 2013) (Fig. 1, panel C). Finally, and only for transcriptomic datasets (Fig. 1, panel D), *vi*) expression at the gene level is accurately quantified for transcriptome studies by applying the StringTie suite (cuffquant and cuffnorm tools) (Pertea *et al*., 2015) on the alignment files (bam files). A principal component analysis is also performed with the FactoMineR package implemented in R (Lê *et al*., 2008) to project samples onto the first two components.

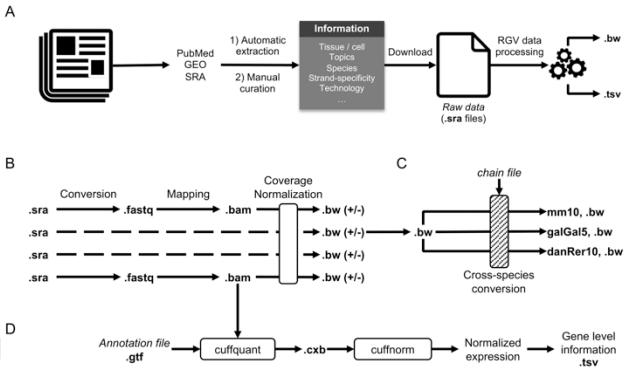


Fig. 1. The RGV data processing pipeline. A) A schematic diagram of the strategy used to organize and process each individual sample in the RGV system from the published datasets. The information used to structure and organize the data – the Pubmed, GEO, and SRA IDs – are manually extracted from the publication first before automatic extraction and manual curation (species name, biological topic, technology, samples, etc). Raw data (.sra files) are next downloaded from the NCBI SRA archive and processed (see below, B-D), leading to the generation of two types of files: first, an indexed binary (bigWig or .bw) file for each sample, to enable fast remote access to the data in the genome browser; second, a tabulation-separated value (.tsv) file for each transcriptomic study, used to summarize gene expression levels in the violin and scatter plot visualization tools. **B-D)** A schematic diagram of the RGV data processing workflow. **B)** After download, the SRA Toolkit is used to convert raw data (.sra) files into .fastq files, which are further mapped on their corresponding genome sequences with STAR v2.0. Next, alignment (.bam) files are converted into standard coverage (.bw) files with the deepTools suite, with a scaling factor used to normalize the coverage of all samples within a given study. When strand information is available, two .bw files per sample (+/-) are generated, i.e. one for each strand of the chromosomes. **C)** A cross-species conversion is performed next, with the CrossMap tool and the chain files from the UCSC genome browser. This process allows the visualization of data from any species on a selected reference genome (human, mouse, chicken, and zebrafish in the current version). **D)** Specifically, for transcriptomic data, expression is quantified at the gene level with the StringTie suite on individual .bam files, by providing the reference annotation GTF file and using the cuffquant and cuffnorm tools. This procedure generates the gene expression matrix (.tsv file) suitable for the violin plot and the scatter plot visualization tools.

2.2 Data storage, management, and retrieval

The ReproGenomics Viewer database is based on MongoDB (https://www.mongodb.com), a free, open-source, and cross-platform document-oriented database program. This NoSQL database technology

Accepted manuscript

provides relevant features for RGV, such as flexible storage of massive and rapidly changing types of data, data replication, and JavaScript compatibility.

2.3 Web interface

The web interface is implemented using two web frameworks, Pyramid (https://trypyramid.com/) and AngularJS (https://angularjs.org/). Pyramid embeds many features, such as a REST API, a JSON renderer, and compatibility with SMTP servers. AngularJS, on the other hand, is an open JavaScript framework that extends traditional HTML vocabulary; it allows implementation of readable and quickly developable web environments. To handle website traffic and to provide data security, scalability, and deployment, every component of RGV (including website server, MongoDB database, and Elasticsearch server) are hosted on separate individual virtual machines, with Docker (https://www.docker.com) as a container system.

2.4 Visualization tools

The ReproGenomics Viewer provides three ways to display datasets, including a genome browser and two gene-level visualization tools: the violin plot and the scatter plot.

The **genome browser** offers a full-featured and highly flexible display for genomic annotations and data via a high-performance, dynamic web interface powered by the implementation of a JBrowse web server (Buels *et al.*, 2016). This system is fully compatible with a large spectrum of data types (fasta, gff, bam, bedGraph, wig, bigWig). The modular design of Jbrowse made it possible to implement three plugins to add new features to the system: *i*) *multibigwig* (https://github.com/elsiklab/multibigwig), which allows users to plot multiple samples from the same dataset on a single track, thereby facilitating data visualization; *ii*) *Bookmarks-JBrowse* (https://github.com/awilkey/bookmarks-jbrowse/), which enables them to save selected tracks and options into a list of bookmarks and to add descriptions and/or comments; *iii*) *Screen Shot* (https://github.com/bhofmei/jbplugin-screenshot), which lets users save a high-quality screenshot by using a dedicated dialog box with options including file format, size, quality, and track configuration.

The two gene-level visualization tools were built with Plotly.js, a high-level, declarative charting library based on the famous web graphical library D3.js. Like box plots, **the violin plot** is a method for representing the distribution of quantitative data (here, gene expression) across several groups of samples by using quartiles, but it also features a kernel density estimation (probability density) of the underlying distribution. The **scatter plot** enables exploration of complex datasets and the efficient visualization of a single gene's expression pattern following the projection of all samples in a lower-dimensional space.

3 Results

3.1 Overview and current content

The ReproGenomics Viewer website has a minimalist and user-friendly web interface. Its responsive design meets all modern web standards and enables users to display RGV on all their devices (desktops, tablets, and smartphones). Briefly, the home page and the top navigation bar offer rapid access to the visualization tools that allow scientists to explore all sequencing datasets listed in the 'Studies' tab. In the past few months, the RGV database content has grown qualitatively and quantitatively through the numerous requests from researchers of the reproductive science community. As of July 2018, the system embeds 2'970 samples from 79 published studies that were carefully curated before data processing and integration (Supplementary Table 1). Altogether these studies cover 11 biological topics related to the biology of reproduction in 9 species and involve 14 "-omics" technologies, including ATAC-seq and single-cell RNAseq (Table 1).

Table 1. RGV content evolution.

	2015	2018
Number of studies	24	79
Number of samples	274	2970
Number of tissues/cells	26	64
Number of biological topics	2	11
Number of technologies	5	14

3.2 The genome browser tool: interactive exploration of large genomic datasets

The genome browser from RGV offers fluid and intuitive genomic navigation among chromosomes of nine species and thus allows multiple users to process data simultaneously. All datasets are formatted and configured for visual display at varying levels of resolution; users can pan and zoom efficiently over a genomic sequence region and simply turn genomics tracks on or off. The background of each track name is color-coded so that samples belonging to the same experimental condition share a similar color.

The full set of available tracks is listed in the faceted track selector tool that enables dynamic queries of RGV's large datasets by the successive application of a series of intersecting filters. This is made possible by the indexation of all study metadata into a CSV file, which facilitates access to information related to biological topics, tissues, age, and technologies (Fig. 2, panel A).

As previously described, one key feature of the RGV data processing workflow is the use of pair-wise alignment files provided by UCSC to perform cross-species conversion of genome coordinates. This allows the direct comparison of experiments in different organisms and thus makes the genome browser not only a multi-species but also a cross-species tool for comparing reproductive genomics data. In addition, the improved processing workflow considers the directionality of the library preparation kits; it can therefore display genomic information associated with the forward and reverse strands on two independent tracks per sample (one for each strand) when applicable (Fig. 2, panel B).

Importantly, the genome browser of RGV also integrates genetic variants known to be associated with reproductive disorders and fertility issues from the ClinVar database (Landrum *et al.*, 2018), as well as the genomic distribution of sequence conservation scores (PhastCons score) provided by the UCSC genome browser (Kuhn *et al.*, 2013).

It is also noteworthy that users can directly upload their individual coverage (bigWig) and annotation (gtf, gff3, and vcf) files into the genome browser by using the option 'Open' in the 'File' tab. These uploaded data are accessible only during the user session and remain totally private.

Accepted manuscript

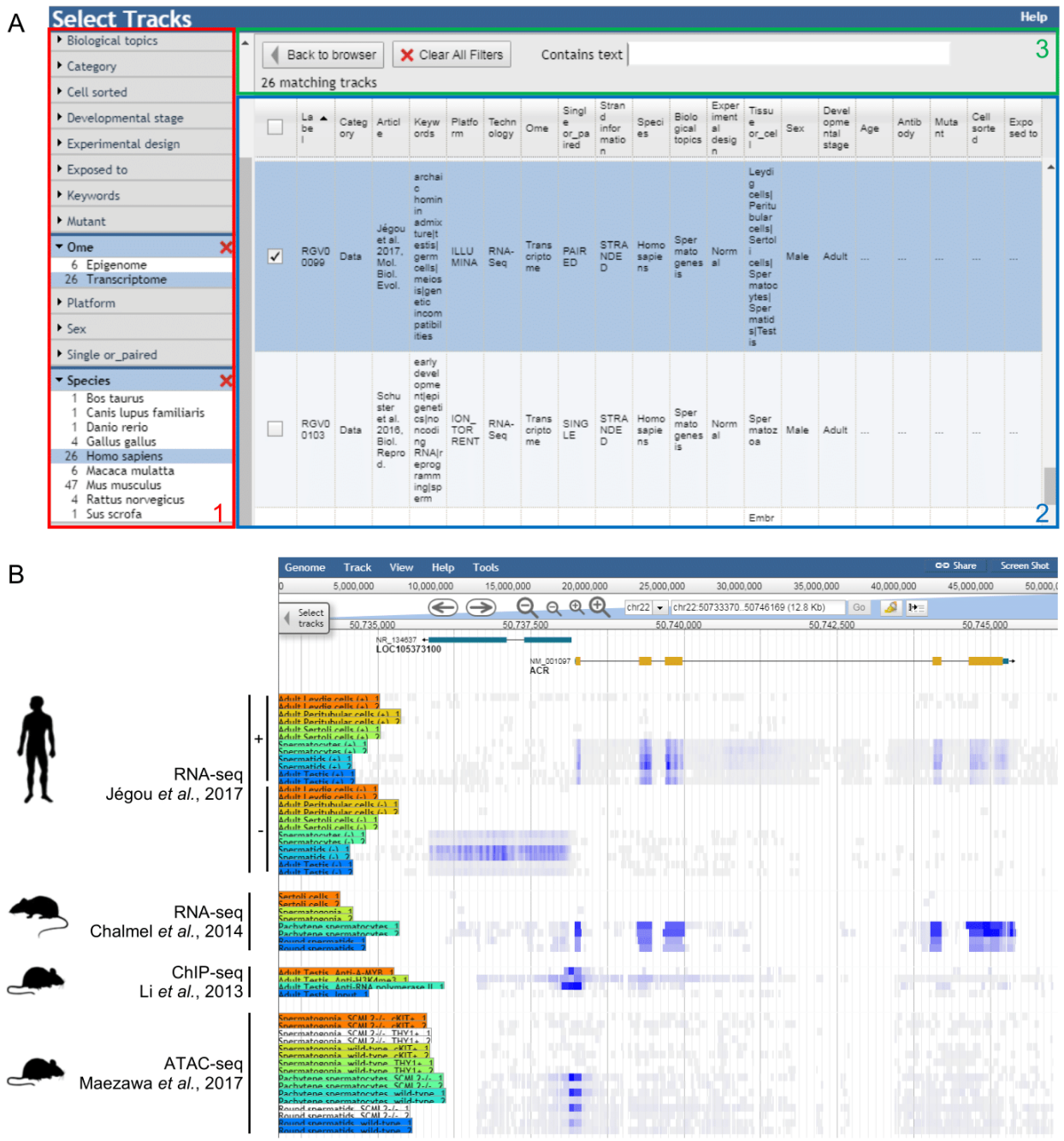


Fig. 2. The RGV genome browser. A) *JBrowse faceted track selector*. The track(s) to be displayed can be selected by applying successive filters to the track metadata. Data can be searched using (1) the topics-related panel and/or can be queried with (2) the text search engine. (3) Studies and data matching the search criteria are displayed in the track list panel for further selection. **B)** *Genome browser overview*. A screen capture from the RGV genome browser, showing the genomic environment of the human acrosin gene (ACR; Gene ID: 49), is presented together with the RefSeq annotation (green and orange boxes correspond to untranslated and coding exons, respectively, and the dark line represents introns) in the human genome (hg38). The RNA-seq analysis of human testicular cells from Jégou *et al.* was selected (Jégou *et al.*, 2017), showing the expression of ACR from the plus strand ("+" tracks) in spermatocytes and spermatids, and that of the LOC105373100 lncRNA (Gene ID: 105373100) from the minus strand ("-" tracks) in spermatids. To illustrate the cross–species and cross-technology capabilities of RGV, three studies in rats and mice were also selected. The RNA-seq data of rat testicular cells demonstrate the conserved expression of ACR in spermatocytes and spermatids (Chalmel *et al.*, 2014), while ChIP-seq data for A-MYB, H3K4me3, and RNA polymerase II in the adult mouse testis (Li *et al.*, 2013) and ATAC-seq data in wild-type and Scml2$^{-/-}$ mouse spermatogonia, spermatocytes, and spermatids (Maezawa *et al.*, 2018) suggest that A-MYB plays a role, independent of SCML2, in regulating ACR expression in meiotic and post-meiotic germ cells by binding to its proximal promoter.

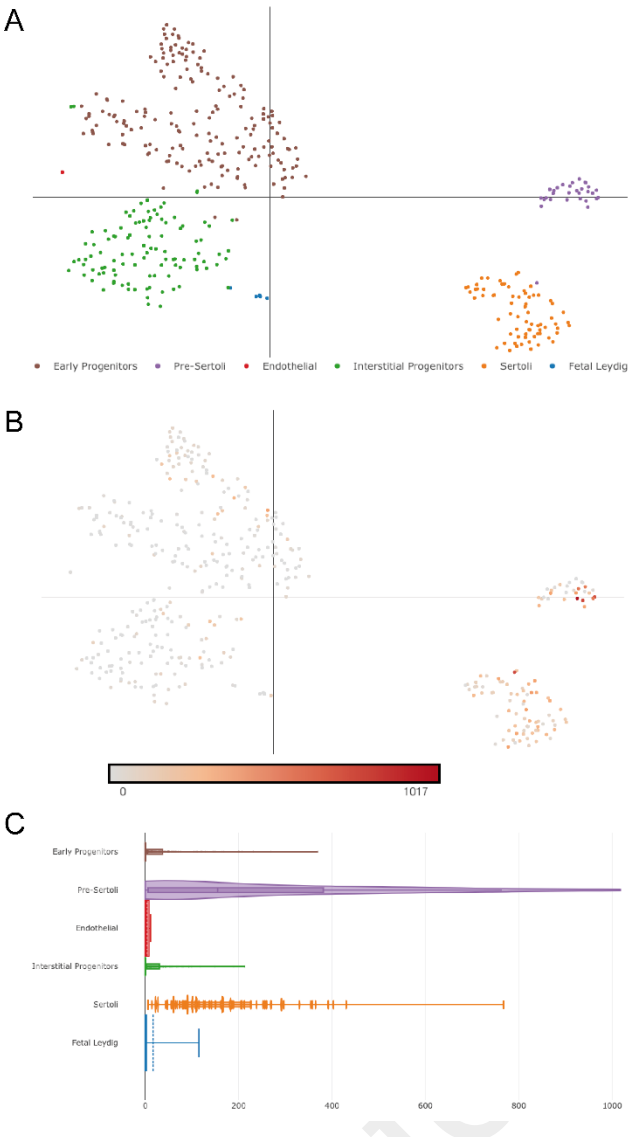


Fig. 3. The RGV visualization tools. Large datasets that may contain hundreds to thousands of cells or samples can be browsed and investigated through both the scatter plot (**A** and **B**) and the violin plot (**C**) visualization tools, which are also compatible with the bulk transcriptomic datasets hosted in RGV. The single-cell RNA sequencing analysis of mouse sf1+ fetal testicular cells is taken as an example (Stévant *et al.*, 2018). **A)** Cells are displayed according to the T-distributed Stochastic Neighbor Embedding (t-SNE) coordinates as originally published and can be color-coded according to different classes of cells defined by the authors (cell types, developmental stages, etc). **B)** The expression levels of selected genes (Sox9 in this example) can then be displayed within each cell, which are color-coded according to the scale bar. **C)** Finally, the expression of candidate genes (Sox9 again in this example) within classes of cells (cell types, developmental stages, etc) can be summarized with violin plots; kernel density estimation makes it possible to display the full distribution of expression values.

3.3 Gene-level visualization tools: towards shifting paradigms in reproductive biology with single-cell sequencing

The current release of RGV embeds two new visualization tools: the violin plot and the scatter plot. Both are part of the current graphical arsenal used by the "single-cell" community for visualizing the structure of high-dimensional data (Lang *et al.*, 2015; DeTomaso and Yosef, 2016; Cao *et al.*, 2017; Weinreb *et al.*, 2018).

These tools allow users to display quantitative (e.g. gene expression or chromatin accessibility) or qualitative (e.g. copy number variations) information in extremely large datasets that may contain several thousand samples or cells.

For this purpose, the scatter plot (and the violin plot to a lesser extent) relies especially on dimensionality reduction methods such as the t-distributed Stochastic Neighbor Embedding approach (t-SNE) (Van Der Maaten and Hinton, 2008). As of July 2018, five scRNA-seq datasets relevant to male and female gonad development (Han *et al.*, 2018; Stévant *et al.*, 2018), primordial germ and adult spermatogonial stem cells (Li *et al.*, 2017; Guo *et al.*, 2017), and spermatogenesis (Lukassen *et al.*, 2018) have already been indexed in RGV, and other unpublished studies are underways in the process of being submitted (Stévant *et al.*, in preparation; Lardenois *et al.*, in preparation). Of note is that although both the scatter plot and the violin plot were initially designed for single-cell studies, they are also fully compatible with and available for all "bulk" RNA-seq studies hosted in RGV.

Studies of interest can be selected through a dataset selector in the form of a table to which users can apply successive criteria (Fig. 3, panel A). In the scatter plots, the spatial distribution of all cells in a selected study is, by default, color-coded according to the subpopulation classes to which they belong (as published in the corresponding study) (Fig. 3, panel B). Users are then invited to select one or more genes of interest so that each individual cell is color-coded according to its relative expression level (Fig. 3, panel C). This feature could thus help users to identify novel markers specific for particular cell subpopulations, or simply to query their favorite genes.

3.4 RGV documentation and data download

The website provides a 'Tutorial' section containing several media types to allow researchers to become rapidly able to use the ReproGenomics Viewer. Moreover, all scripts and data generated or used are freely available through the 'Download' page and the GitHub repository (https://github.com/fchalmel/RGV). The genome browser tool implemented in RGV lets users download each available individual track by choosing a track of interest and then clicking on 'Save track data'.

4 Conclusion and perspectives

We report a new version of the ReproGenomics Viewer resource intended to contribute to the reproductive science community by facilitating the centralization and harmonization of the sequencing data we have together accumulated. The system includes a unique workflow for sequencing data management, curation, processing, organization, and comparison across species. With the recent democratization of single-cell genomics, the number of these studies will increase rapidly in the near future. To deal efficiently with this new information, we have complemented existing features with two novel visualization tools, the scatter plot and the violin plot, which enable a more in-depth investigation of these highly complex datasets. This logical transition from "bulk" to single-cell approaches is also accompanied by additional major changes to the system, including a new user-friendly and powerful project selector that makes dataset retrieval and mining easier. Since the original publication (Darde *et al.*, 2015), the database content has also increased dramatically – there are now 12 times as many samples as in 2015, covering four times as many biological topics related to reproductive biology. To the best of our knowledge, the ReproGenomics Viewer

is the most comprehensive and the only cross-species resource of sequencing data in this field.

We plan to keep improving the features and possibilities already offered by the ReproGenomics Viewer. For instance, in the short term, a lasso select tool will be added to the scatter plot visualization tool to generate additional plots based on manually-defined cell subpopulations. A false-color heatmap visualization tool will also be implemented to display the expression pattern of several selected genes within the same graphical representation. In the mid-term, we will allow the creation of private sessions so that users can upload and display their own processed data in the RGV system. Finally, in the long term, we intend to develop a social community toolbox in RGV that we hope will facilitate and stimulate collaborative work in our research field.

It goes without saying that we also intend to gather even more datasets from a wide range of species to cover additional biological topics relevant to reproductive biology. In our continuous effort to help researchers make the most of the technological breakthroughs in the life sciences, we also intend to make RGV compatible with other technologies, such as proteomics and bisulfite sequencing experiments. Other genetic information related to reproductive disorders (such as GWAS and Quantitative trait loci information from diverse sources and diverse model organisms) will also be added to the genome browser tool, as we already plan to integrate data from the GWAS catalog database (Welter *et al*., 2014).

To keep the resource up-to-date, we strongly encourage scientists to request new datasets and types of data to be added to the system. Indeed, in view of the community's demand for direct and easy access to processed data, we think that making raw data available to the community may be insufficient. We also encourage scientists to upload their own reprogenomic datasets to RGV before submission for publication. This could create a win-win situation for all parties, since submitted studies would benefit from enhanced visibility, while the community would take advantage of a continually updated resource. In this context we would like to draw attention to an important issue, which does not concern only RGV, but all (omics) databases in life sciences. Researchers often cite the studies hosted by the databases, but they forget to cite the databases themselves. These citations, however, are essential for the teams that develop and maintain such resources.

The ReproGenomics Viewer is a valuable resource for and supported by researchers in the field of reproductive biology. This system may help scientists and clinicians to overcome the data format and species issues so that they can compare their own datasets with relevant published studies. The very positive feedback of scientists, at international conferences notably, encourages us to maintain our effort and to improve RGV further in terms of both its content and its functionalities. The modular design of the RGV could be applied to other biological domains studied in several model organisms. In that sense, we believe that the ReproGenomics Viewer resource can serve as a reference database in the life sciences.

Acknowledgements

The ReproGenomics Viewer database is supported, built, and maintained by the Research Institute for Environmental and Occupational Health (IRSET), the French School of Public Health (EHESP), and the GenOuest BioInformatics core facility. We thank the Institut national de la santé et de la recherche médicale (Inserm) and the Université de Rennes 1 for supporting this work.

Funding

This work was supported by the French agency for food and safety [ANSES n° EST-13-081 and n° EST-17-256 to F.C.]; the French National Research Agency [ANR n° 16-CE14-0017-02 and n°18-CE14-0038-02 to F.C]; the Fondation pour la recherche médicale [FRM n° DBI20131228558 to F.C.], the Swiss National Science Foundation [SNF n° CRS115_171007 to B.J.], the DFG Clinical Research Unit 'Male Germ Cells: from Genes to Function' [CRU 326 to F.T.], the French National Institute of Health and Medical Research (Inserm), the University of Rennes 1 and the French School of Public Health (EHESP).

Conflict of Interest: none declared.

References

Abugessaisa,I. *et al.* (2018) SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res.*, **46**, D781–D787.

Buels,R. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.

Cao,Y. *et al.* (2017) scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells. *Genes (Basel).*, **8**.

Chalmel,F. *et al.* (2014) High-Resolution Profiling of Novel Transcribed Regions During Rat Spermatogenesis1. *Biol. Reprod.*, **91**, 5.

Darde,T. a. *et al.* (2015) The ReproGenomics Viewer: an integrative cross-species toolbox for the reproductive science community. *Nucleic Acids Res.*, **43**, 1–8.

DeTomaso,D. and Yosef,N. (2016) FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics*, **17**, 315.

Dobin,A. and Gingeras,T.R. (2015) Mapping RNA-seq Reads with STAR. In, *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, p. 11.14.1-11.14.19.

Guo,H. *et al.* (2017) DNA methylation and chromatin accessibility profiling of mouse and human fetal germ cells. *Cell Res.*, **27**, 165–183.

Han,X. *et al.* (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, **172**, 1091–1107.e17.

Hsueh,A.J. and Rauch,R. (2012) Ovarian Kaleidoscope database: ten years and beyond. *Biol. Reprod.*, **86**, 192.

Hua,J. *et al.* (2015) Follicle Online: an integrated database of follicle assembly, development and ovulation. *Database*, **2015**, bav036-bav036.

Jégou,B. *et al.* (2017) Meiotic Genes Are Enriched in Regions of Reduced Archaic Ancestry. *Mol. Biol. Evol.*, **34**, 1974–1980.

Kodama,Y. *et al.* (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54-6.

Kuhn,R.M. *et al.* (2013) The UCSC genome browser and associated tools. *Brief. Bioinform.*, **14**, 144–161.

Landrum,M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.

Lang,S. *et al.* (2015) SCExV: a webtool for the analysis and visualisation of single cell qRT-PCR data. *BMC Bioinformatics*, **16**, 320.

Lardenois,A. *et al.* (2010) GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. *Database (Oxford).*, **2010**, baq030.

Lê,S. *et al.* (2008) FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.*, **25**, 1–18.

Lee,T.-L. *et al.* (2009) GermSAGE: a comprehensive SAGE database for transcript discovery on male germ cell development. *Nucleic Acids Res.*, **37**, D891-7.

Lee,T.-L. *et al.* (2010) GonadSAGE: a comprehensive SAGE database for transcript discovery on male embryonic gonad development.

Accepted manuscript

Bioinformatics, **26**, 585–6.

Li,L. *et al.* (2017) Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell*, **20**, 858–873.e4.

Li,X.Z. *et al.* (2013) An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Mol. Cell*, **50**, 67–81.

Linnarsson,S. and Teichmann,S.A. (2016) Single-cell genomics: coming of age. *Genome Biol.*, **17**, 97.

Luk,A.C.-S. *et al.* (2015) GermlncRNA: a unique catalogue of long non-coding RNAs and associated regulations in male germ cell development. *Database (Oxford).*, **2015**, bav044.

Lukassen,S. *et al.* (2018) Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Sci. Rep.*, **8**, 6521.

Van Der Maaten,L. and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Maezawa,S. *et al.* (2018) Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis. *Nucleic Acids Res.*, **46**, 593–608.

Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

Raja,K. *et al.* (2017) A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. *Int. J. Genomics*, **2017**, 1–10.

Ramírez,F. *et al.* (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160-5.

Schuster,A. *et al.* (2016) SpermBase: A Database for Sperm-Borne RNA Contents. *Biol. Reprod.*, **95**, 99.

Sims,D. *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–32.

Stévant,I. *et al.* (2018) Deciphering Cell Lineage Specification during Male Sex Determination with Single-Cell RNA Sequencing. *Cell Rep.*, **22**, 1589–1599.

Weinreb,C. *et al.* (2018) SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, **34**, 1246–1248.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001-6.

Zhang,Y. *et al.* (2013) SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucleic Acids Res.*, **41**, D1055-62.

Zhao,H. *et al.* (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–7.

Zhu,X. *et al.* (2017) Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.*, **9**, 108.

Accepted manuscript