RNAbrowse: RNA-Seq De Novo Assembly Results Browser

Jérôme Mariette[1]*, Céline Noirot[1], Ibounyamine Nabihoudine[1], Philippe Bardou[2], Claire Hoede[1], Anis Djari[2], Cédric Cabau[2], Christophe Klopp[1,2]

1 Plate-forme bio-informatique Genotoul/Biométrie et Intelligence Artificielle, INRA, Castanet-Tolosan, France, **2** Plate-forme SIGENAE/Génétique Cellulaire, INRA, Castanet-Tolosan, France

Abstract

Transcriptome analysis based on a de novo assembly of next generation RNA sequences is now performed routinely in many laboratories. The generated results, including contig sequences, quantification figures, functional annotations and variation discovery outputs are usually bulky and quite diverse. This article presents a user oriented storage and visualisation environment permitting to explore the data in a top-down manner, going from general graphical views to all possible details. The software package is based on biomart, easy to install and populate with local data. The software package is available under the GNU General Public License (GPL) at http://bioinfo.genotoul.fr/RNAbrowse.

Citation: Mariette J, Noirot C, Nabihoudine I, Bardou P, Hoede C, et al. (2014) RNAbrowse: RNA-Seq De Novo Assembly Results Browser. PLoS ONE 9(5): e96821. doi:10.1371/journal.pone.0096821

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

Received January 28, 2014; **Accepted** April 11, 2014; **Published** May 13, 2014

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: support.genopole@toulouse.inra.fr

Introduction

The massive sequencing cost decrease has attracted a large community of new users, some of them studying organisms for which the reference genome sequence is still not available. When trying to understand mechanisms taking place at the gene level they usually would start with a de novo transcriptome assembly approach. Software packages such as Trinity [1] or Oases [2] are mature enough to produce reliable contigs from short reads. The analyses performed on the assembled contigs generate a large amount of heterogeneous results including variations, functional annotation and expression measurements. The processing steps of these pipelines are usually shared inside the community but the parameters, the tools and the reference databases used are specific. The results are often provided throught a WEB server including a BLAST query form and download links.

Cbrowse [3] is a WEB environment presenting this kind of results throught graphical views and query forms. The functional annotation part is not implemented yet and the query possibilities are very limited. The Galaxy [4] engine provides users with an interface to create and track workflow executions. It already embeds RNA-Seq analysis and assembly components. However, none of them offers user-friendly query and vizualisation features designed for RNA-Seq de novo annotation.

In its last version (0.8), biomart [5] is developed as an easily extensible query infrastructures which can be specialized in the presentation of focused data types. On top of the database and beside the proposed query forms it is possible to add new pages as plug-ins in order to present data in a user-friendly way.

Results

RNAbrowse permits sequencing facilities and, even small, bioinformatic teams to give a user-friendly access to RNA-Seq de novo results, helping biologists to analyse and extract meaningful information from their data.

RNAbrowse includes two components: a web-based user interface and an administration command line tool presented here-after.

Implementation and installation

RNAbrowse is based on a newly defined biomart schema specialized in RNA-Seq de novo assembly. It includes two marts, storing contig and variant data, and some extra tables for meta-data. The WEB-based user interface is implemented as a biomart plugin using the latest jquery and highcharts features. To perform all third parts requests, the software takes advantage of the biomart API and implements several biomart processors to prepare the data before presenting it.

The command line administration tool is based on a workflow execution environment called jflow. It is developed in Python. The installation requires a standard unix server and a MySQL database. The downloadable installation archive, provided on the project web-site, has only to be uncompressed before setting up a first instance. In order to ease the testing, a minimal and a complete example dataset are included in the archive.

WEB interface

The web interface was designed to display the data starting with general views before zooming into detailed ones. It is organised in four levels. The first, called the instance level, corresponds to the home page of the web-site presenting all the available projects. A project gathers the data of an assembly. The second level is entered after having chosen a project. The introduction page contains a picture of the assembled species, a description text and a table including all the analysis processing steps with software names, parameters and versions in order to ease materials and methods writing.

The third level is accessed through the menu bar items at the top of the project page and presents general informations about contigs and variants. It also includes the download page. The contigs (Figure 1) and variants (Figure 2) overview pages include a set of graphics showing general statistics, containing for example the contig length histogram, alignment rate bar-chart, mostly represented species in the functional annotation histogram or variant types distribution as a pie-chart. These graphs enable rapid assembly validation and comparison. For contigs, the next section includes detailed information about the different sequenced libraries and provides access to tools such as Venn diagrams (Figure 3) and a differential digital display. The last section corresponds to the favourite contigs or variants table which can be updated by the user while exploring detailed views. The forth level sections gathers all information on a contig or a variant and provides multiple tools to analyse them further.

To illustrate these features we could imagine a simple use case in which a user would like to find all contigs corresponding to a gene for which the sequence is available for another species. This can be done by an alignment versus the contigs using the blast query form (Figure 4) or by a name or description search using the biomart form. The user can then add the found contigs to the favourite table. For each contig, the sequence can be extracted to perform a multiple alignment in order to check if different splice forms have been assembled. All possible open reading frames can be sought (Figure 5). Annotations can be graphically displayed in jbrowse [6]. It is also possible to graphically verify if the expression levels along the contig are conserved between libraries (Figure 6).

All generated graphics can directly be printed from the web-page or downloaded in JPG, PNG, PDF, SVG formats or as a CSV file. Tables can be sorted on all columns, text searched and exported to the clipboard or as CSV files. The download page files are organised as topics. It is possible to download a single file, all the files of a topic or to get the text file containing all the URLs of a topic in order to download the files from the command line. The Download section gathers raw files but also processed ones, enabling users to perform further analyses.

A demo site containing a small set of results is accessible at: http://ngspipelines.toulouse.inra.fr:9012/.

Administration tools and extensions

The administration interface is a command line tool with which the environment can be set up. It includes database creation, data formatting and loading and web server managing. The loading process requiers standard file formats, provided by most commonly used tools. During the database upload phase, datafile format compliance is checked. It is possible to set up a minimum environment with just three files: a contig Fasta file, the corresponding annotation file and an alignment file. The reads quantifications can be provided as an input file or calculated during the upload process. The number of input files being potentially quite important for large projects, a configuration file template is provided. It can be used as a loading script parameter.

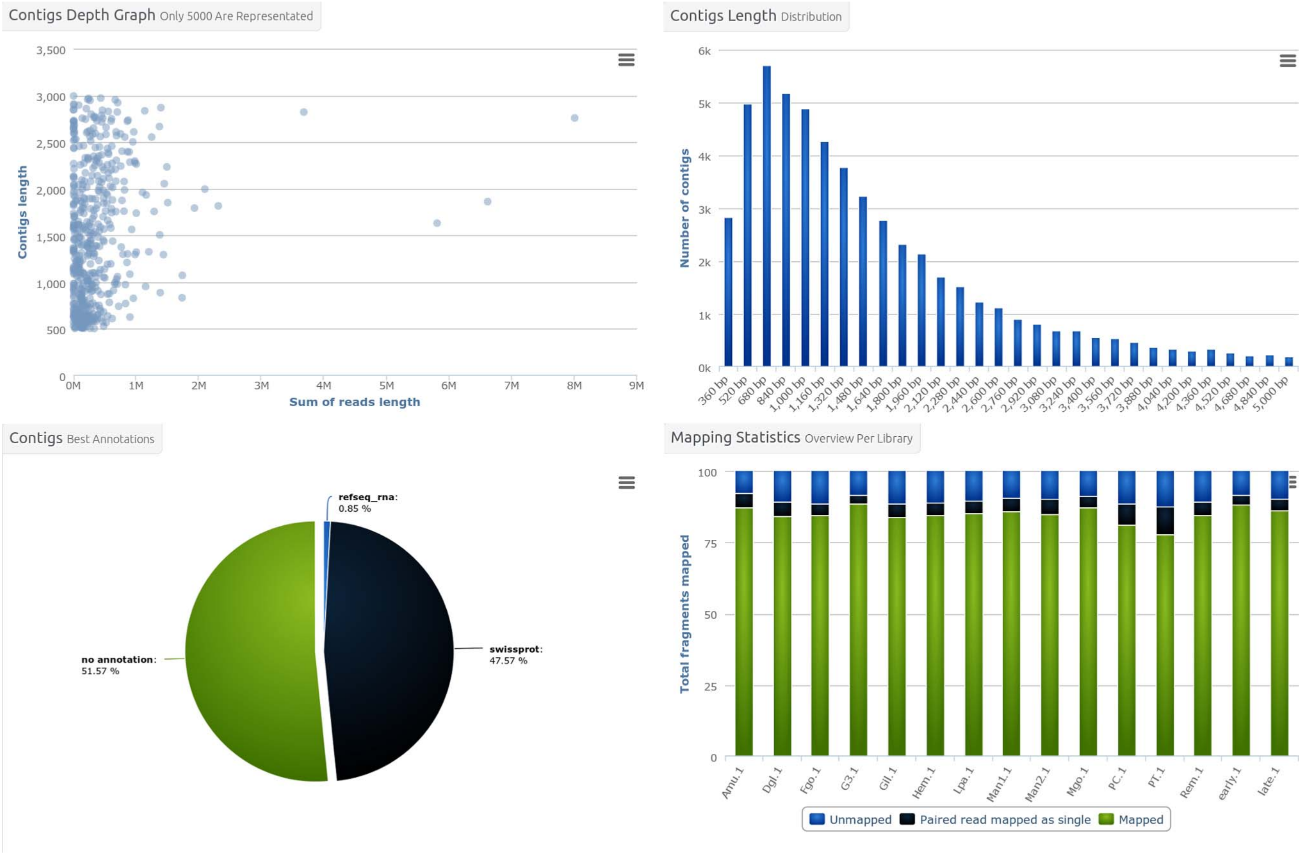
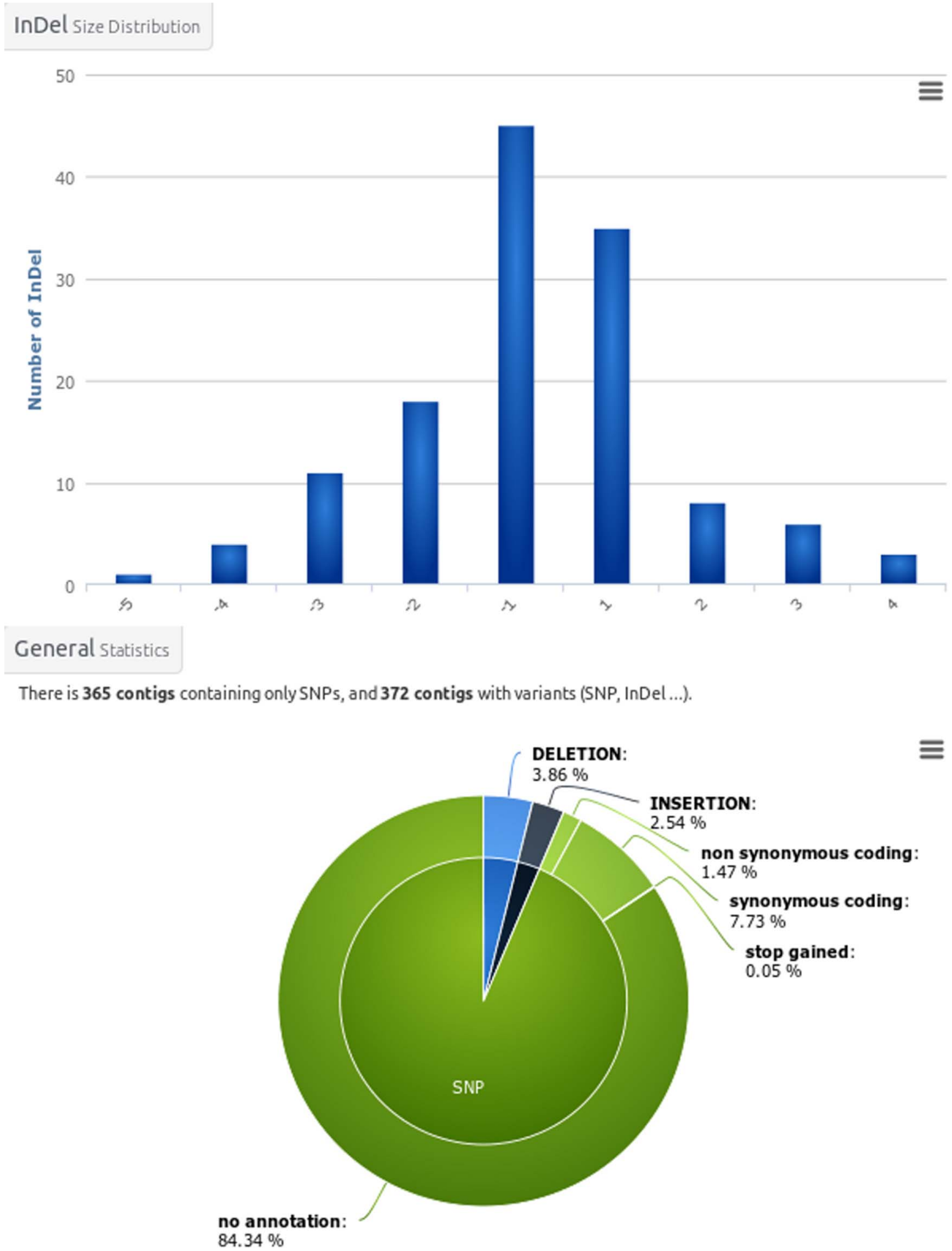


Figure 1. Contigs overview figures. From the main contig page, graphics synthesizing informations on the contig set can be displayed. The presented graphs are contigs depth versus size plot (top-left), contigs length distribution (top-right), contig best annotations pie chart (bottom-left) and libraries mapping bar chart (bottom-right). All graphics can be printed or downloaded in four different formats (PNG, JPEG, PDF, SVG).
doi:10.1371/journal.pone.0096821.g001



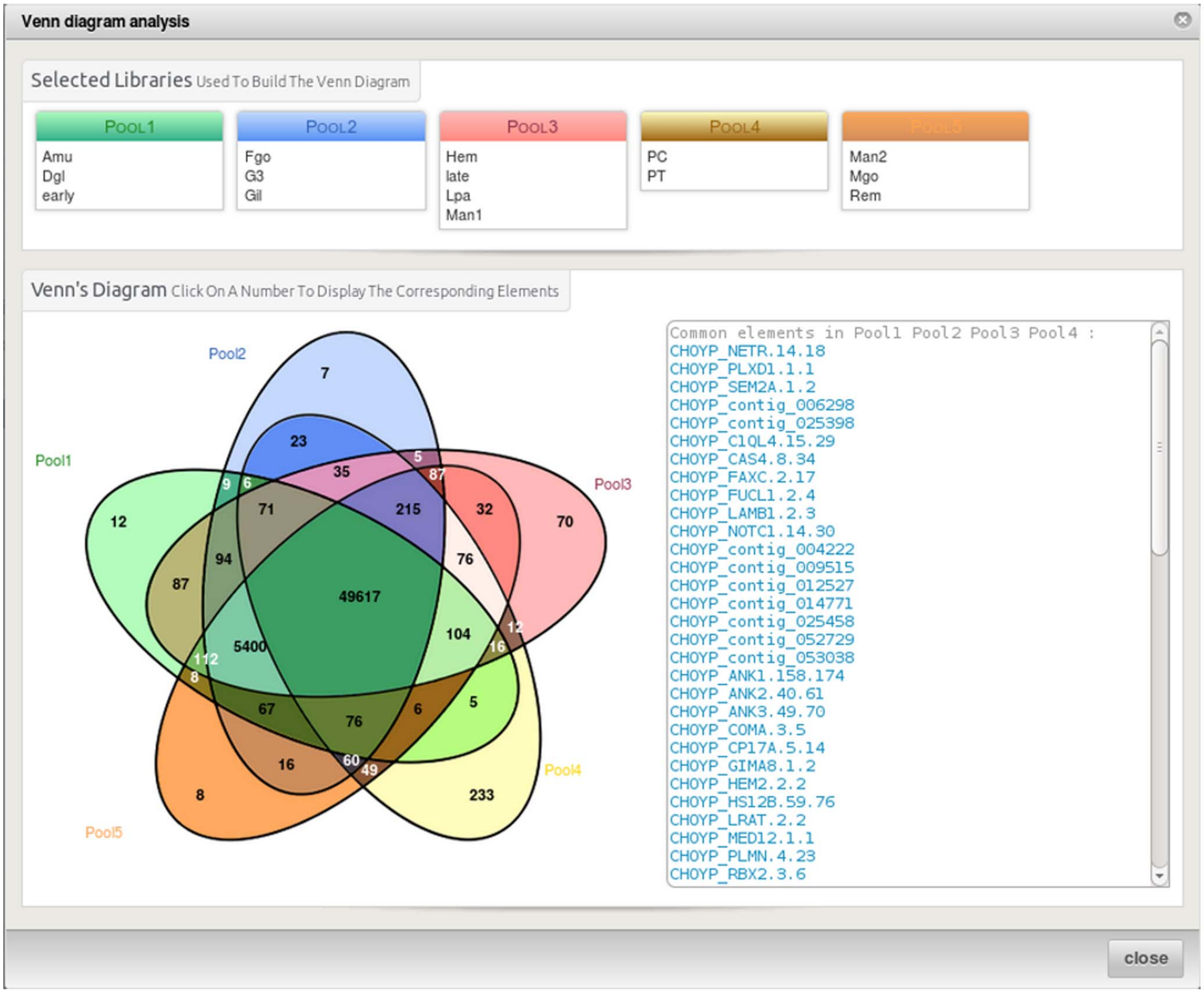
There is **365 contigs** containing only SNPs, and **372 contigs** with variants (SNP, InDel ...).

Figure 2. Variants overview figures. From the main variant page, graphics synthesizing informations on the variant set can be displayed. The presented graphs are InDels size distribution (top) and Indels annotations pie chart (bottom).
doi:10.1371/journal.pone.0096821.g002

Depending on the provided inputs, RNAbrowse processing steps can be quite time consuming. Thus, in a production environment, the tool can be set up to use different schedulers such as condor, sge, moab, workqueue or mpi-queuein in order to parallelize the loading. If the loading process fails, a recovery command is available to rerun it.

The website provided to the biologists can be secured using the biomart Open-ID or jetty realm features.

Users with some programming knowledge can add new graphics by loading the corresponding data in the analysis table and writing a javascript module.


Venn diagram analysis
Selected Libraries Used To Build The Venn Diagram
Venn's Diagram Click On A Number To Display The Corresponding Elements
Pool2
Pool1
Pool3
Pool4
Pool5
7
23
5
35
87
9
6
71
215
32
12
70
94
76
87
49617
104
12
16
5400
112
8
67
76
6
5
16
60
49
8
233
close

POOL1	POOL2	POOL3	POOL4	POOL5
Amu Dgl early	Fgo G3 Gil	Hem late Lpa Man1	PC PT	Man2 Mgo Rem

```
Common elements in Pool1 Pool2 Pool3 Pool4 :
CHOYP_NETR.14.18
CHOYP_PLXD1.1.1
CHOYP_SEM2A.1.2
CHOYP_contig_006298
CHOYP_contig_025398
CHOYP_C1QL4.15.29
CHOYP_CAS4.8.34
CHOYP_FAXC.2.17
CHOYP_FUCL1.2.4
CHOYP_LAMB1.2.3
CHOYP_NOTC1.14.30
CHOYP_contig_004222
CHOYP_contig_009515
CHOYP_contig_012527
CHOYP_contig_014771
CHOYP_contig_025458
CHOYP_contig_052729
CHOYP_contig_053038
CHOYP_ANK1.158.174
CHOYP_ANK2.40.61
CHOYP_ANK3.49.70
CHOYP_COMA.3.5
CHOYP_CP17A.5.14
CHOYP_GIMA8.1.2
CHOYP_HEM2.2.2
CHOYP_HS12B.59.76
CHOYP_LRAT.2.2
CHOYP_MED12.1.1
CHOYP_PLMN.4.23
CHOYP_RBX2.3.6
```

Figure 3. The Venn diagram shows the number of contigs shared between libraries and the specific ones. It has been built using libraries alignment data. To build a new diagram the user has to select the libraries he wants to have in each pool (from two to five). If the user clicks on a figure in the graph the list of corresponding contigs will appear in the list box on the right hand side. Clicking on the contig names redirects to the corresponding page.

doi:10.1371/journal.pone.0096821.g003

Discussion

RNAbrowse is a simple and efficient solution to give access to RNA-Seq de novo results on the Internet. It includes many features that help biologists to analyse and extract biologically meaningful information from their data. The installation and user manuals as well as a general documentation are available on the project website: http://bioinfo.genotoul.fr/RNAbrowse. The software has been designed to be easily extended by developpers having some biomart insights.

Acknowledgments

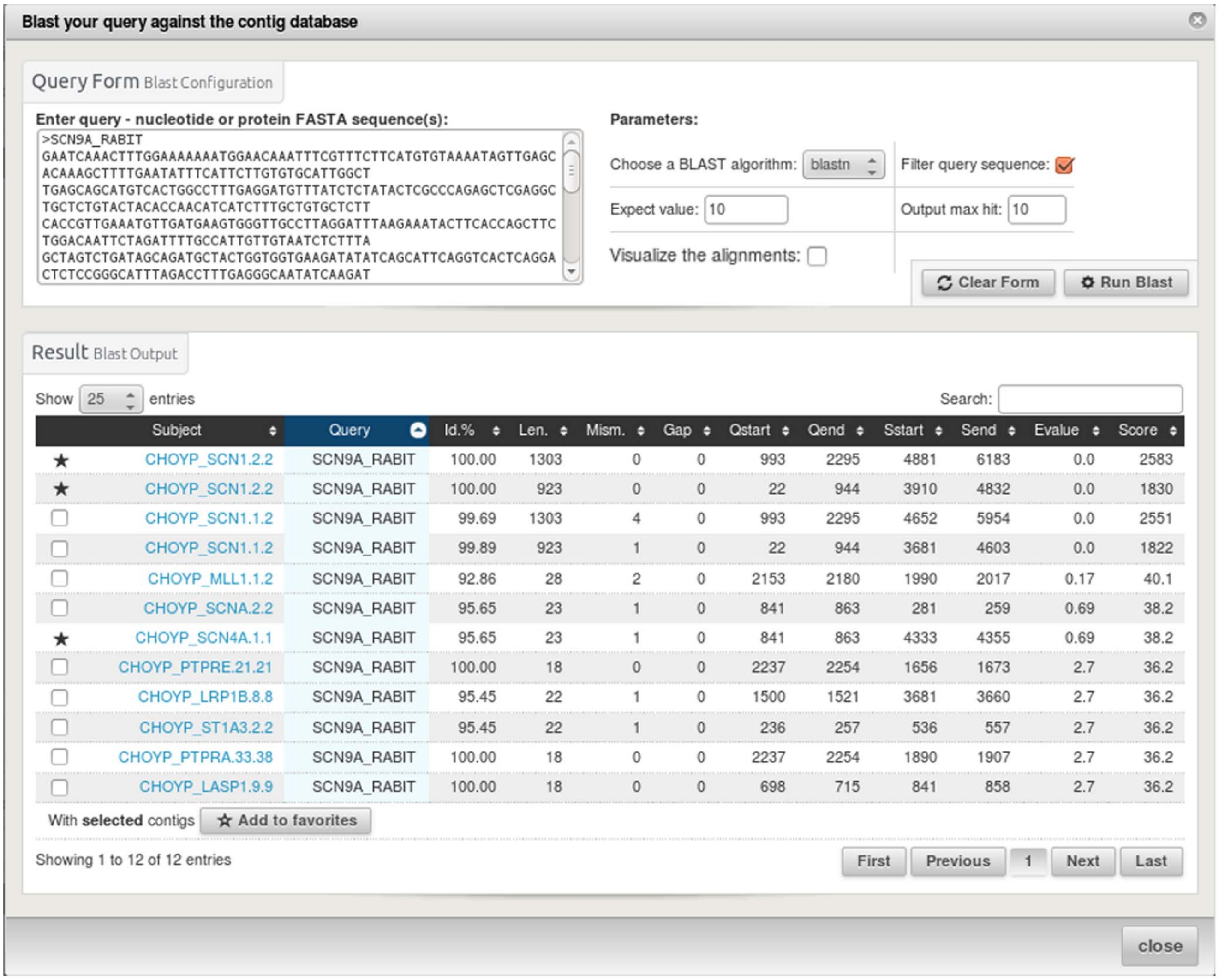
We would like to acknowledge all our users for providing us useful feedback on the system and for pointing out features worth developing.

Author Contributions

Conceived and designed the experiments: JM CK. Performed the experiments: JM CN IN PB. Wrote the paper: CK JM. Tested the software: AD CH CC.

References

1. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from rna-seq data without a reference genome. Nat Biotech 29.
2. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092.
3. Li P, Ji G, Dong M, Schmidt E, Lenox D, et al. (2012) Cbrowse: a sam/bam-based contig browser for transcriptome assembly visualization and analysis. Bioinformatics 28: 2382–2384.
4. Goecks J, Nekrutenko A, Taylor J, Team TG (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11: R86.
5. Zhang J, Haider S, Baran J, Cros A, Guberman JM, et al. (2011) Biomart: a data federation framework for large collaborative projects. Database 2011.
6. Westesson O, Skinner M, Holmes I (2013) Visualizing next-generation sequencing data with jbrowse. Briefings in Bioinformatics 14: 172–177.



Blast your query against the contig database
Query Form Blast Configuration

Enter query - nucleotide or protein FASTA sequence(s):

```
>SCN9A_RABIT
GAATCAAACTTTGGAAAAAAATGGAACAAATTTCGTTTCTTCATGTGTAAAATAGTTGAGC
ACAAAGCTTTTGAATATTTCATTCTTGTGTGCATTGGCT
TGAGCAGCATGTCACTGGCCTTTGAGGATGTTTATCTCTATACTCGCCCAGAGCTCGAGGC
TGCTCTGTACTACACCAACATCATCTTTGCTGTGCTCTT
CACCGTTGAAATGTTGATGAAGTGGGTTGCCTTAGGATTTAAGAAATACTTCACCAGCTTC
TGGACAATTCTAGATTTTGCCATTGTTGTAATCTCTTTA
GCTAGTCTGATAGCAGATGCTACTGGTGGTGAAGATATATCAGCATTCAGGTCACTCAGGA
CTCTCCGGGCATTTAGACCTTTGAGGGCAATATCAAGAT
```

Parameters:

Choose a BLAST algorithm: blastn

Filter query sequence: ☑

Expect value: 10

Output max hit: 10

Visualize the alignments: □

Show 25 entries

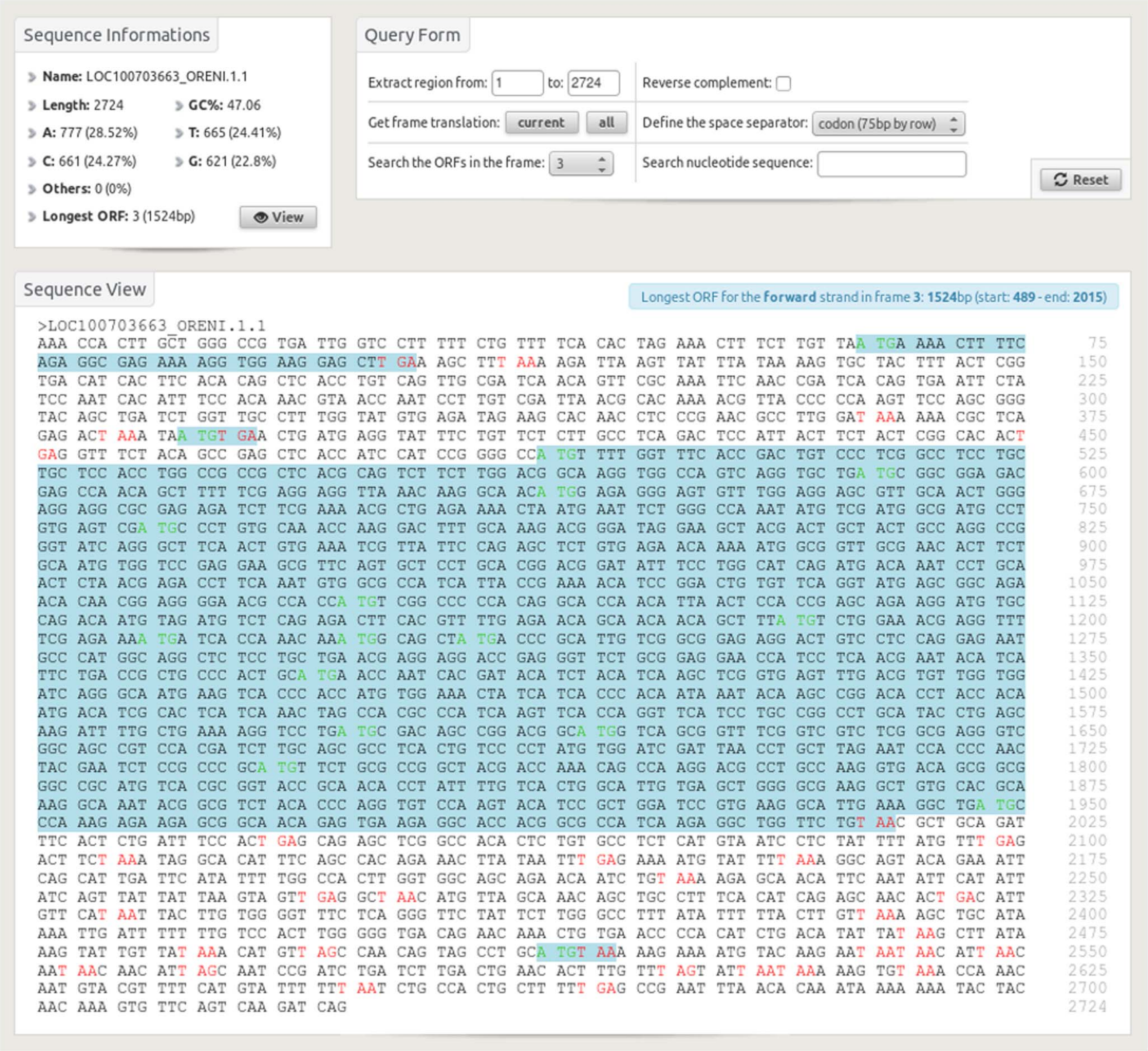
Search:

	Subject	Query	Id.%	Len.	Mism.	Gap	Qstart	Qend	Sstart	Send	Evalue	Score
★	CHOYP_SCN1.2.2	SCN9A_RABIT	100.00	1303	0	0	993	2295	4881	6183	0.0	2583
★	CHOYP_SCN1.2.2	SCN9A_RABIT	100.00	923	0	0	22	944	3910	4832	0.0	1830
□	CHOYP_SCN1.1.2	SCN9A_RABIT	99.69	1303	4	0	993	2295	4652	5954	0.0	2551
□	CHOYP_SCN1.1.2	SCN9A_RABIT	99.89	923	1	0	22	944	3681	4603	0.0	1822
□	CHOYP_MLL1.1.2	SCN9A_RABIT	92.86	28	2	0	2153	2180	1990	2017	0.17	40.1
□	CHOYP_SCNA.2.2	SCN9A_RABIT	95.65	23	1	0	841	863	281	259	0.69	38.2
★	CHOYP_SCN4A.1.1	SCN9A_RABIT	95.65	23	1	0	841	863	4333	4355	0.69	38.2
□	CHOYP_PTPRE.21.21	SCN9A_RABIT	100.00	18	0	0	2237	2254	1656	1673	2.7	36.2
□	CHOYP_LRP1B.8.8	SCN9A_RABIT	95.45	22	1	0	1500	1521	3681	3660	2.7	36.2
□	CHOYP_ST1A3.2.2	SCN9A_RABIT	95.45	22	1	0	236	257	536	557	2.7	36.2
□	CHOYP_PTPRA.33.38	SCN9A_RABIT	100.00	18	0	0	2237	2254	1890	1907	2.7	36.2
□	CHOYP_LASP1.9.9	SCN9A_RABIT	100.00	18	0	0	698	715	841	858	2.7	36.2

With **selected** contigs ☆ Add to favorites

Showing 1 to 12 of 12 entries

Figure 4. The blast query form allows to search the contigs as a database using as input sequence(s) in Fasta format. Others parameters which can be used in this query are blast program (blastn, blastx), expected value and maximum number of outputs to be shown. The blast results are shown in a table allowing to add new contigs as favorite. Ticking the alignment checkbox enables to browse the alignment results.
doi:10.1371/journal.pone.0096821.g004



Sequence Informations

- **Name:** LOC100703663_ORENI.1.1
- **Length:** 2724 **GC%:** 47.06
- **A:** 777 (28.52%) **T:** 665 (24.41%)
- **C:** 661 (24.27%) **G:** 621 (22.8%)
- **Others:** 0 (0%)
- **Longest ORF:** 3 (1524bp)

Query Form

Extract region from: 1 to: 2724

Reverse complement: □

Define the space separator: codon (75bp by row)

Search the ORFs in the frame: 3

Search nucleotide sequence:

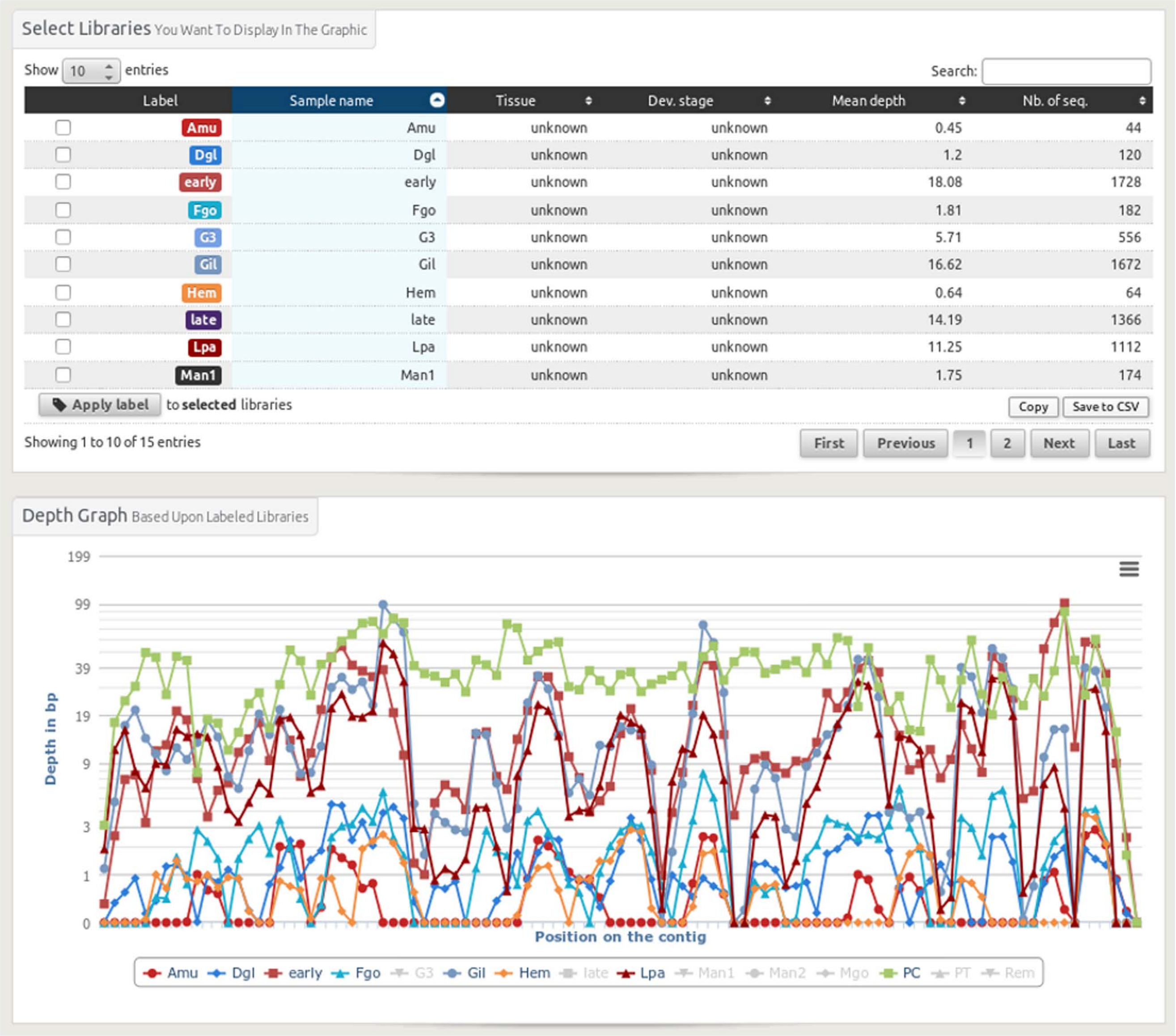
Sequence View

Longest ORF for the **forward** strand in frame **3**: **1524**bp (start: **489** - end: **2015**)

```
>LOC100703663_ORENI.1.1
AAA CCA CTT GCT GGG CCG TGA TTG GTC CTT TTT CTG TTT TCA CAC TAG AAA CTT TCT TGT TAA TGA AAA CTT TTC     75
AGA GGC GAG AAA AGG TGG AAG GAG CTT GAA AGC TTT AAA AGA TTA AGT TAT TTA TAA AAG TGC TAC TTT ACT CGG    150
TGA CAT CAC TTC ACA CAG CTC ACC TGT CAG TTG CGA TCA ACA GTT CGC AAA TTC AAC CGA TCA CAG TGA ATT CTA    225
TCC AAT CAC ATT TCC ACA AAC GTA ACC AAT CCT TGT CGA TTA ACG CAC AAA ACG TTA CCC CCA AGT TCC AGC GGG    300
TAC AGC TGA TCT GGT TGC CTT TGG TAT GTG AGA TAG AAG CAC AAC CTC CCG AAC GCC TTG GAT AAA AAA CGC TCA    375
GAG ACT AAA TAA TGT GAA CTG ATG AGG TAT TTC TGT TCT CTT GCC TCA GAC TCC ATT ACT TCT ACT CGG CAC ACT    450
GAG GTT TCT ACA GCC GAG CTC ACC ATC CAT CCG GGG CCA TGT TTT GGT TTC ACC GAC TGT CCC TCG GCC TCC TGC    525
TGC TCC ACC TGG CCG CCG CTC ACG CAG TCT TCT TGG ACG GCA AGG TGG CCA GTC AGG TGC TGA TGC GGC GGA GAC    600
GAG CCA ACA GCT TTT TCG AGG AGG TTA AAC AAG GCA ACA TGG AGA GGG AGT GTT TGG AGG AGC GTT GCA ACT GGG    675
AGG AGG CGC GAG AGA TCT TCG AAA ACG CTG AGA AAA CTA ATG AAT TCT GGG CCA AAT ATG TCG ATG GCG ATG CCT    750
GTG AGT CGA TGC CCT GTG CAA ACC AAG GAC TTT GCA AAG ACG GGA TAG GAA GCT ACG ACT GCT ACT GCC AGG CCG    825
GGT ATC AGG GCT TCA ACT GTG AAA TCG TTA TTC CAG AGC TCT GTG AGA ACA AAA ATG GCG GTT GCG AAC ACT TCT    900
GCA ATG TGG TCC GAG GAA GCG TTC AGT GCT CCT GCA CGG ACG GAT ATT TCC TGG CAT CAG ATG ACA AAT CCT GCA    975
ACT CTA ACG AGA CCT TCA AAT GTG GCG CCA TCA TTA CCG AAA ACA TCC GGA CTG TGT TCA GGT ATG AGC GGC AGA   1050
ACA CAA CGG AGG GGA ACG CCA CCA TGT CGG CCC CCA CAG GCA CCA ACA TTA ACT CCA CCG AGC AGA AGG ATG TGC   1125
CAG ACA ATG TAG ATG TCT CAG AGA CTT CAC GTT TTG AGA ACA GCA ACA ACA GCT TTA TGT CTG GAA ACG AGG TTT   1200
TCG AGA AAA TGA TCA CCA AAC AAA TGG CAG CTA TGA CCC GCA TTG TCG GCG GAG AGG ACT GTC CTC CAG GAG AAT   1275
GCC CAT GGC AGG CTC TCC TGC TGA ACG AGG AGG ACC GAG GGT TCT GCG GAG GAA CCA TCC TCA ACG AAT ACA TCA   1350
TTC TGA CCG CTG CCC ACT GCA TGA ACC AAT CAC GAT ACA TCT ACA TCA AGC TCG GTG AGT TTG ACG TGT TGG TGG   1425
ATC AGG GCA ATG AAG TCA CCC ACC ATG TGG AAA CTA TCA TCA CCC ACA ATA AAT ACA AGC CGG ACA CCT ACC ACA   1500
ATG ACA TCG CAC TCA TCA AAC TAG CCA CGC CCA TCA AGT TCA CCA GGT TCA TCC TGC CGG CCT GCA TAC CTG AGC   1575
AAG ATT TTG CTG AAA AGG TCC TGA TGC GAC AGC CGG ACG GCA TGG TCA GCG GTT TCG GTC GTC TCG GCG AGG GTC   1650
GGC AGC CGT CCA CGA TCT TGC AGC GCC TCA CTG TCC CCT ATG TGG ATC GAT TAA CCT GCT TAG AAT CCA CCC AAC   1725
TAC GAA TCT CCG CCC GCA TGT TCT GCG CCG GCT ACG ACC AAA CAG CCA AGG ACG CCT GCC AAG GTG ACA GCG GCG   1800
GGC CGC ATG TCA CGC GGT ACC GCA ACA CCT ATT TTG TCA CTG GCA TTG TGA GCT GGG GCG AAG GCT GTG CAC GCA   1875
AAG GCA AAT ACG GCG TCT ACA CCC AGG TGT CCA AGT ACA TCC GCT GGA TCC GTG AAG GCA TTG AAA GGC TGA TGC   1950
CCA AAG AGA AGA GCG GCA ACA GAG TGA AGA GGC ACC ACG GCG CCA TCA AGA GGC TGG TTC TGT AAC GCT GCA GAT   2025
TTC ACT CTG ATT TCC ACT GAG CAG AGC TCG GCC ACA CTC TGT GCC TCT CAT GTA ATC CTC TAT TTT ATG TTT GAG   2100
ACT TCT AAA TAG GCA CAT TTC AGC CAC AGA AAC TTA TAA TTT GAG AAA ATG TAT TTT AAA GGC AGT ACA GAA ATT   2175
CAG CAT TGA TTC ATA TTT TGG CCA CTT GGT GGC AGC AGA ACA ATC TGT AAA AGA GCA ACA TTC AAT ATT CAT ATT   2250
ATC AGT TAT TAT TAA GTA GTT GAG GCT AAC ATG TTA GCA AAC AGC TGC CTT TCA CAT CAG AGC AAC ACT GAC ATT   2325
GTT CAT AAT TAC TTG TGG GGT TTC TCA GGG TTC TAT TCT TGG GCC TTT ATA TTT TTA CTT GTT AAA AGC TGC ATA   2400
AAA TTG ATT TTT TTG TCC ACT TGG GGG TGA CAG AAC AAA CTG TGA ACC CCA CAT CTG ACA TAT TAT AAG CTT ATA   2475
AAG TAT TGT TAT AAA CAT GTT AGC CAA CAG TAG CCT GCA TGT AAA AAG AAA ATG TAC AAG AAT AAT AAC ATT AAC   2550
AAT AAC AAC ATT AGC AAT CCG ATC TGA TCT TGA CTG AAC ACT TTG TTT AGT ATT AAT AAA AAG TGT AAA CCA AAC   2625
AAT GTA CGT TTT CAT GTA TTT TTT AAT CTG CCA CTG CTT TTT GAG CCG AAT TTA ACA CAA ATA AAA AAA TAC TAC   2700
AAC AAA GTG TTC AGT CAA GAT CAG                                                                      2724
```

Figure 5. The sequence view provides informations such as nucleotide content and longest open reading frame (ORF). The possible starts are presented in green, stops in red and ORFs in blue. The query form permits different actions on the sequence such as extraction, reverse complementation, translation in different frames, ORF presentation and text search.

doi:10.1371/journal.pone.0096821.g005



Select Libraries You Want To Display In The Graphic

Show 10 entries

Search:

Label	Sample name	Tissue	Dev. stage	Mean depth	Nb. of seq.
□ Amu	Amu	unknown	unknown	0.45	44
□ Dgl	Dgl	unknown	unknown	1.2	120
□ early	early	unknown	unknown	18.08	1728
□ Fgo	Fgo	unknown	unknown	1.81	182
□ G3	G3	unknown	unknown	5.71	556
□ Gil	Gil	unknown	unknown	16.62	1672
□ Hem	Hem	unknown	unknown	0.64	64
□ late	late	unknown	unknown	14.19	1366
□ Lpa	Lpa	unknown	unknown	11.25	1112
□ Man1	Man1	unknown	unknown	1.75	174

Apply label to **selected** libraries

Copy Save to CSV

Showing 1 to 10 of 15 entries

First Previous 1 2 Next Last

Depth Graph Based Upon Labeled Libraries

Figure 6. The contig depth view enables to visualise the coverage of the reads of the different libraries on a given contig. Each library has a defined colour in the table and the same one on the graphic. It is possible to modify the graphical layout by averaging different library depths. This is typically to be used when you work with replicates.

doi:10.1371/journal.pone.0096821.g006