

RNAbrowse: RNA-Seq de novo assembly results browser.

Jérôme J. Mariette, Céline Noirot, Ibouniyamine Nabihoudine, Philippe Bardou, Claire Hoede, Anis Djari, Cédric Cabau, Christophe C. Klopp

▶ To cite this version:

Jérôme J. Mariette, Céline Noirot, Ibouniyamine Nabihoudine, Philippe Bardou, Claire Hoede, et al.. RNAbrowse: RNA-Seq de novo assembly results browser.. PLoS ONE, 2014, 9 (5), 7 p. 10.1371/journal.pone.0096821 . hal-02636622

HAL Id: hal-02636622 https://hal.inrae.fr/hal-02636622

Submitted on 27 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CrossMark

RNAbrowse: RNA-Seq De Novo Assembly Results Browser

Jérôme Mariette¹*, Céline Noirot¹, Ibounyamine Nabihoudine¹, Philippe Bardou², Claire Hoede¹, Anis Djari², Cédric Cabau², 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

Copyright: © 2014 Mariette et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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 metadata. 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 webpage 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.

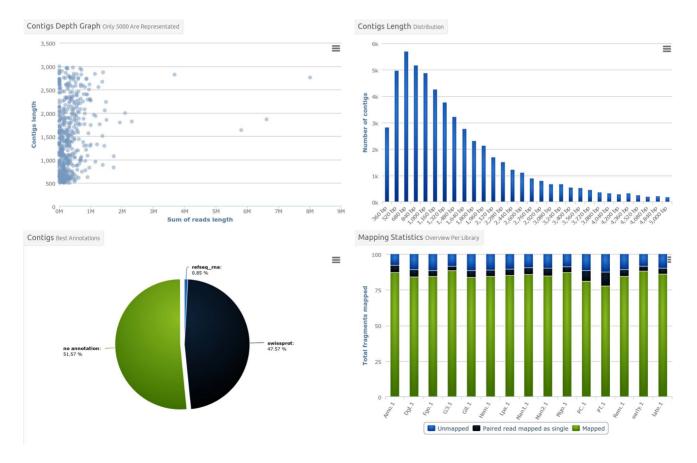
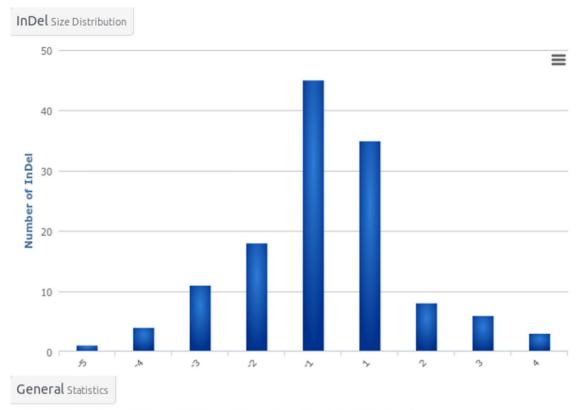


Figure 1. Contigs overview figures. From the main contig page, graphics synthesizing informations on the contig set can be displayed. The presented graphs are contigs depth versus size plot (top-left), contigs length distribution (top-right), contig best annotations pie chart (bottom-left) and libraries 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





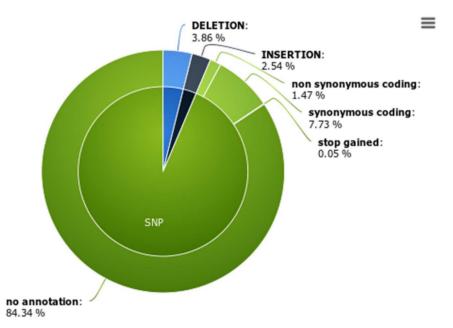


Figure 2. Variants overview figures. From the main variant page, graphics synthesizing informations on the variant set can be displayed. The presented graphs are InDels size distribution (top) and Indels annotations pie chart (bottom). doi:10.1371/journal.pone.0096821.g002

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.

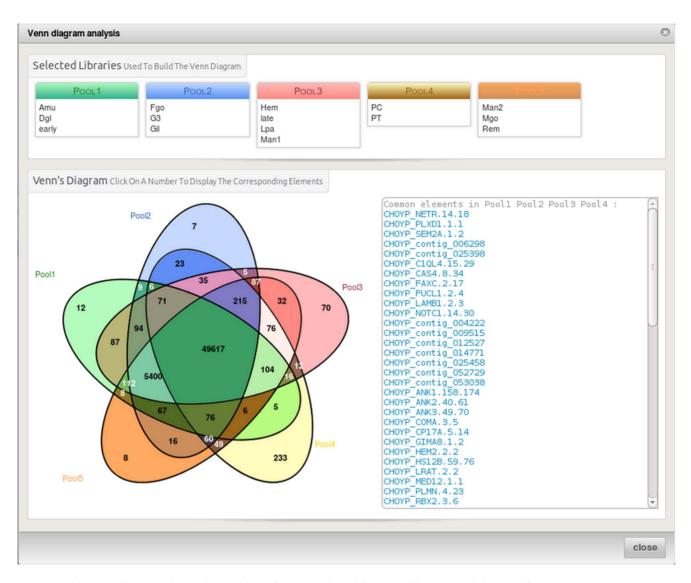


Figure 3. The Venn diagram shows the number of contigs shared between libraries and the specific ones. It has been built using libraries alignment data. To build a new diagram the user has to select the libraries he wants to have in each pool (from two to five). If the user clicks on a figure in the 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.

References

- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Fulllength transcriptome assembly from rna-seq data without a reference genome. Nat Biotech 29.
- 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.
- Li P, Ji G, Dong M, Schmidt E, Lenox D, et al. (2012) Cbrowse: a sam/bambased contig browser for transcriptome assembly visualization and analysis. Bioinformatics 28: 2382–2384.

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.

- 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.
- Zhang J, Haider S, Baran J, Cros A, Guberman JM, et al. (2011) Biomart: a data federation framework for large collaborative projects. Database 2011.
- Westesson O, Skinner M, Holmes I (2013) Visualizing next-generation sequencing data with jbrowse. Briefings in Bioinformatics 14: 172–177.

Query	Form Blast Configuration											
Enter qu	iery - nucleotide or protein	FASTA sequence(s):		Paramet	ters:						
ACAAAG	_RABIT AACTTTGGAAAAAAATGGAACAAA CTTTTGAATATTTCATTCTTGTGT GCATGTCACTGGCCTTTGAGGATG	Choose a BLAST algorithm: blastn 🔹 Filter query sequence: 🧭										
CACCGT	GTACTACACCAACATCATCTTTGC TGAAATGTTGATGAAGTGGGTTGC ATTCTAGATTTTGCCATTGTTGTA	CTTAGGATTTAAGAAA	TACTTCACCA	GCTTC	Expect v	alue: 10			Output n	nax hit: 10		
GCTAGT	CTGATAGCAGATGCTACTGGTGGT GGGCATTTAGACCTTTGAGGGCAA	GAAGATATATCAGCAT	TCAGGTCACT	CAGGA	Visualiz	e the ali	gnments:		C	Clear Form	n 🗘 Ri	un Blas
≷esult	Blast Output											
Show 2	5 🌲 entries								Se	arch:		
	Subject ¢	Query 🕒	ld.% ¢	Len. ¢	Mism. 🗢 🤇	Gap 🔹	Qstart ¢	Qend ¢ Ss	start ¢	Send ¢	Evalue 🕈	Score
*	CHOYP_SCN1.2.2	SCN9A_RABIT	100.00	1303	0	0	993	2295	4881	6183	0.0	258
*	CHOYP_SCN1.2.2	SCN9A_RABIT	100.00	923	0	0	22	944	3910	4832	0.0	183
	CHOYP_SCN1.1.2	SCN9A_RABIT	99.69	1303	4	0	993	2295	4652	5954	0.0	255
	CHOYP_SCN1.1.2	SCN9A_RABIT	99.89	923	1	0	22	944	3681	4603	0.0	182
	CHOYP_MLL1.1.2	SCN9A_RABIT	92.86	28	2	0	2153	2180	1990	2017	0.17	40.
	CHOYP_SCNA.2.2	SCN9A_RABIT	95.65	23	1	0	841	863	281	259	0.69	38.
*	CHOYP_SCN4A.1.1	SCN9A_RABIT	95.65	23	1	0	841	863	4333	4355	0.69	38.
	CHOYP_PTPRE.21.21	SCN9A_RABIT	100.00	18	0	0	2237	2254	1656	1673	2.7	36.3
	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 se	elected contigs Add to	favorites										
Showing	1 to 12 of 12 entries							First	Prev	rious 1	Next	Last
								-	_			

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	Query Form													
 Name: LOC100703663_ORENI.1.1 Length: 2724 GC%: 47.06 A: 777 (28.52%) T: 665 (24.41%) 	Extract region from: 1 to: 2724 Get frame translation: current all	Reverse complement: Define the space separator: codon (75bp by row)												
 C: 661 (24.27%) G: 621 (22.8%) Others: 0 (0%) 	Search the ORFs in the frame: 3 🛟 Search nucleotide sequence:													
Longest ORF: 3 (1524bp) View														
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 AGA GGC GAG AAA AGG TGG AAG GAG	GTC CTT TTT CTG TTT TCA CAC 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 TCC AAT CAC ATT TCC ACA AAC GTA TAC AGC TGA TCT GGT TGC CTT TGG	TGT CAG TTG CGA TCA ACA GTT ACC AAT CCT TGT CGA TTA ACG 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	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	
	CAT																								1350	
	TGA																								1425	
	AGG																								1500	
	ACA																								1575	
	ATT																								1650	
	AGC																								1725	
	GAA																								1800	
	CGC																								1875	
	GCA																								1950	
	AAG																								2025	
	ACT																								2100	
	TCT																								2175	
	CAT																								2250	
	AGT																								2325	
	CAT																								2400	
	TTG																								2475	
	TAT																								2550	
	AAC																								2625 2700	
	GTA AAA							AAT	CIG	UCA	UTG	UTT	TTT	GAG	CCG	AAT	TTA	ACA	CAA	ATA	AAA	AAA	TAC	TAC	2700	
AAC	AAA	919	IIC	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



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