COV2HTML: visualization and analysis tool of Bacterial NGS data for biologists

M. Monot[1], M. Orgeur[2], E. Camiade[1], C. Brehier[1] and B. Dupuy[1]

(1) Laboratoire de pathogenèse des bactéries anaérobies, Institut Pasteur, Paris. (2) Unité de Pathogénomique Mycobactérienne, Institut Pasteur, Paris

Abstract

All NGS technologies pose several challenges in terms of visualization of the mapping coverage of the NGS data and in the results analysis. Ideally, biologists should have a “plug and play” software for such studies. Thus we developed COV2HTML, an accessible visualization and analysis tool for biologists, which consists in two highly specialized softwares. The first one (i) calculates the genome coverage from the heavy mapping file, 1-10 Go, and (ii) extracts genes information from annotation files to create light result files of 1 Mo. The second software is a visualization tool dedicated to study the mean coverage of genes or their promoter region. One or two conditions with a maximum of 4 replicates per condition or up to 8 experiments of diverse types can be visualized and analysed by our interface. The analysis is managed by filters that use two criteria: gene coverage level and fold change. The strength of the COV2HTML programs is to easily analyse and share data without software installation, login or a long training period.

Implementation

COV2HTML is a tool for biologists that allows coverage visualization of the NGS data before to be analysed. In order to ease both data loading and processing, COV2HTML uses an own coverage format instead of directly handling huge alignment map or coverage file. Thus, when mapped against the reference genome, the NGS data are converted by the tool MAP2COV provided with the visualization interface.

MAP2COV: format data

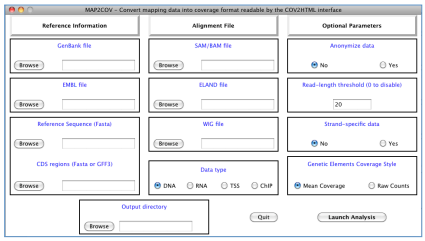


Figure 1. MAP2COV Tcl/Tk graphical interface. Three parts to fill out (i) Reference information: GenBank, EMBL file or Genome and Annotation files; (ii) Alignment File: SAM/BAM, Eland or Wig files; Data type: DNA, RNA, TSS, ChIP; (iii) Optional Parameters: Anonymize, Read-length, Strand-specific or Genetic Elements Coverage Style.

File Size	Mapping file 1-10 Go	Coverage file ~10 Mo	MAP2COV file ~1 Mo
Read Informations			
Sequence	yes	no	no
Quality	yes	no	no
SNP	yes	no	no
Coverage Data			
Genomic	no	yes	yes
Genetic elements	no	no	yes
Genome Annotation			
Genes	no	no	yes
Intergenic regions	no	no	yes

Table 1. Comparison of NGS data files composition. The mapping file contains read Informations. The coverage files contain the genomic coverage. The MAP2COV file contains the genomic and genetic elements coverage and their annotation.

COV2HTML: Visualization & Analysis

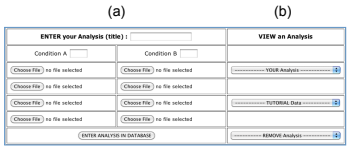


Figure 2. COV2HTML connexion.php web page. a) An analysis is specified by a title, a label attached to condition A and B (optional) and up to 8 files using the « Choose File » button. To enter analysis in database press the action button. b) Analyses stored in database. YOUR Analysis (User’s analysis), TUTORIAL Data and a Remove Analysis button.

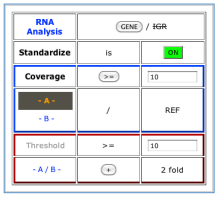


Figure 4. COV2HTML Gene or IGR Filter box. Example of a 2 conditions RNA-seq experiment. From top to bottom: Experiment type and filters switch button from Genes to IGR analysis; Standardization ON/OFF: Switch button which balance the mean coverage of each sample (Figure 5A); One condition panel (blue rectangle): Filter genetic elements by their coverage. Two conditions comparison panel (red rectangle): Filter fold change differences by setting the minimum coverage value assigned to genetic elements before ratio calculation (threshold); Filters results are displayed below the filter box (Figure 3).

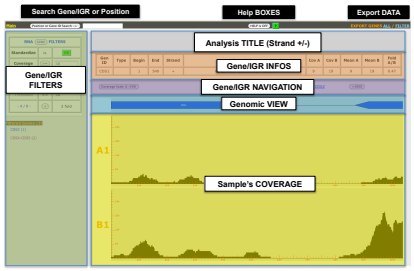
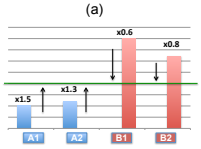


Figure 3. COV2HTML visualization.php. [Grey box] Analysis title and genomic strand. [Orange box] Selected gene or IGR informations and coverage for each replicate, the mean value per conditions and the ratio of conditions A/B. [Yellow box] The analysis condition and replicate genomic coverage views. One graphical bar (one pixel) represents a genomic base coverage. [Blue box] Gene and IGR in a genomic view are in blue color except rRNA and tRNA which are in red. [Purple box] Navigation tools within genome. Current gene or IGR surround with link to the previous and next one. A Y-axis zoom button, to change coverage scale from 1-250 to 1-2500 (10x). [Green box] Gene or IGR filters to analyse coverage data. [Black boxes] Others tools: search for a position or a gene ID; activate help boxes; export all or filtered results in a comma-separated value file.



(b)

Coverage A	Coverage B	Threshold	Ratio	Fold Change
2	30	1	2 / 30	10
		10	10 / 30	3

Figure 5. COV2HTML Filter box options. (a) Standardization: A balance point is calculated using all replicates (green line). A correction factor is applied to genes or IGR of each replicate. (b) The minimum coverage value assigned to a gene or IGR is set by the threshold.

Results

To test COV2HTML, we looked in the NGS publications of the 2011’s year NGS data. The first one is a RNA-seq analysis performed in *Campylobacter jejuni*, which consists of a non stranded RNA-seq comparison of 2 conditions with 2 replicates per condition. The second publication contains stranded TSS and RNA-seq data on the Archaea *Sulfolobus solfataricus* recovered from multiple conditions. We showed that even with a simple criteria and without further manual analysis, COV2HTML is able to recover the main results of the two articles

Analysis of RNA-seq data

***Campylobacter jejuni* ΔrpoN / WT**

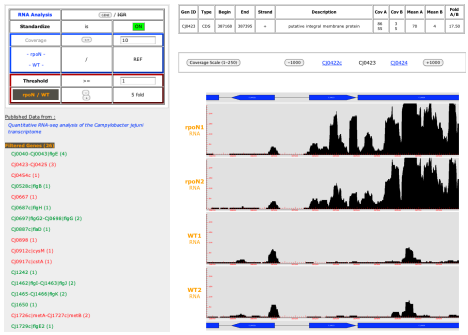


Figure 6. To validate the relevance of COV2HTML for the biologists uses, we tested the program by using the NGS data of a *C. jejuni* RNA-seq analysis recently published, which is a standard comparison of 2 experimental conditions (wild-type and mutant strains) with 2 replicates. In this study the authors showed that expression of 27 genes was significantly altered in the mutant compared to the wild-type strain with a minimum fold change of 5. With these criteria (Table 2), we were able to recover 26 out of 27 genes of the publication data whatever the coverage calculation style.

	PUBLICATION	COV2HTML	
Coverage style	Raw counts	Raw counts	Mean coverage
Normalization	RPKM	ON	ON
Threshold	No	1	1
Fold Change	5 minimum	5	5
Fold Change publication & COV2HTML			
Cj0040	0,00	0,00	0,01
Cj0041	0,02	0,02	0,04
Cj0042	0,01	0,01	0,02
Cj0043	0,02	0,02	0,03
Cj0243c	0,09	0,08	*** **0,25** ***
Cj0423	16,9	16,1	17,5
Cj0424	12,7	12,6	13,2
Cj0425	18,0	20,6	24,3
Cj0454c	6,00	5,92	10,5
Cj0528c	0,16	0,16	0,16
Cj0667	5,91	6,03	6,39
Cj0687c	0,04	0,03	0,03
Cj0697	0,07	0,06	0,06
Cj0698	0,15	0,15	0,15
Cj0887c	0,15	0,15	0,15
Cj0898	5,21	5,00	5,34
Cj0912c	6,60	6,42	6,68
Cj0917c	6,03	6,30	6,65
Cj1242	0,03	0,03	0,03
Cj1462	0,03	0,03	0,04
Cj1463	0,03	0,03	0,03
Cj1465	0,11	0,11	0,09
Cj1466	0,07	0,07	0,07
Cj1650	0,17	0,19	0,15
Cj1726c	5,04	*** **4,92** ***	5,09
Cj1727c	5,42	5,12	5,37
Cj1729c	0,01	0,01	0,01

Table 2. RNA-seq analysis of *C. jejuni* comparison between published results and COV2HTML. Results from the publication on *C. jejuni* RNA-seq experiment are compared to an analysis of authors data using COV2HTML. The coverage style, normalization method, threshold and fold change are indicated for each. Fold change of genes that are not present in COV2HTML analyses are given in bold red surround by stars.

Visualization of TSS data

	PUBLICATION	COV2HTML	COMMON (%)
Transcriptional Start Sites			
Strand +	463	755	354 (76%)
Strand -	497	817	409 (82%)

Table 3. TSS analysis comparison between published results and COV2HTML. In COV2HTML the filter used was « coverage > 10 ».

Combined analysis RNA-seq/TSS

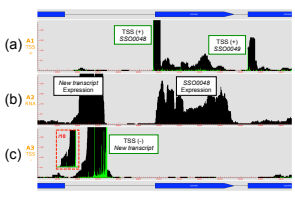


Figure 7. COV2HTML combined analysis. Three coverages from two experiments done on the same bacterial reference, RNA-seq and TSS, were mixed in one view. a) Strand plus of transcriptional start site (TSS+); b) non-stranded RNA-seq to detect gene Expression; c) Strand minus of transcriptional start site (TSS-). A dashed red insert show the 10x Y-scale zoom of TSS region.

Conclusion

As more and more NGS data become accessible online, biologists are attracted to exploit diversely the experimental results. Therefore, COV2HTML has been designed as an easy and accessible interface to simplify analysis of NGS data recovered from lab experiments or from databases. Furthermore information sharing is enhanced by a new COV2HTML tiny coverage format. A web version is currently accessible at http://mmonot.eu/COV2HTML/. This website is free and open to all users without any login requirement.