BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 18 2014, pages 2659–2661
doi:10.1093/bioinformatics/btu366

Genome analysis Advance Access publication May 30, 2014

# EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes

Erika Sallet[1,*], Jérôme Gouzy[1] and Thomas Schiex[2]

[1]Laboratoire Interactions Plantes Micro-organismes (LIPM) UMR441/2594, INRA/CNRS, F-31320 and [2]INRA, Unité de Mathématiques et Informatique Appliques de Toulouse, UR 875, Castanet-Tolosan F-31326, France

Associate Editor: Inanc Birol

**ABSTRACT**

**Summary:** It is now easy and increasingly usual to produce oriented RNA-Seq data as a prokaryotic genome is being sequenced. However, this information is usually just used for expression quantification. EuGene-PP is a fully automated pipeline for structural annotation of prokaryotic genomes integrating protein similarities, statistical information and any oriented expression information (RNA-Seq or tiling arrays) through a variety of file formats to produce a qualitatively enriched annotation including coding regions but also (possibly antisense) non-coding genes and transcription start sites.

**Availability and implementation:** EuGene-PP is an open-source software based on EuGene-P integrating a Galaxy configuration. EuGene-PP can be downloaded at eugene.toulouse.inra.fr.

**Contact:** erika.sallet@toulouse.inra.fr

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

Received on January 14, 2014; revised on May 7, 2014; accepted on May 23, 2014

## 1 INTRODUCTION

Prokaryotic genome sequencing and expression quantification using RNA-Seq or tiling arrays are becoming routine. However, existing prokaryotic gene finders are either *ab initio* gene finders that identify only coding regions (Delcher *et al.*, 2007, Hyatt *et al.*, 2010) or purely RNA-Seq-based gene finders predicting transcripts (Martin *et al.*, 2010) and are much less effective than their *ab initio* competitors for Coding DNA Sequence (CDS) prediction (Zickmann *et al.*, 2014). Reconciling conflicting predictions is a tedious work, which is incompatible with the growing prokaryotic genome sequencing rate.

There is a need for new prokaryotic gene finders that would directly integrate all available information to produce an enriched and precise structural annotation identifying CDS but also (possibly antisense) non-coding genes and transcription start sites (TSSs), avoiding tedious reconciliation. We have shown in Sallet *et al.* (2013) that this can be accomplished by finely adapting conditional random field-based integrative eukaryotic gene finding technology (Foissac *et al.*, 2008) to prokaryotic specificities (overlapping genes, operons). The resulting software, EuGene-P, simultaneously exploits statistical properties of sequences, existing annotations, similarities to proteins and oriented RNA-Seq data to produce an enriched annotation with a better delineation of functional genomic elements, and therefore improved expression quantification.

## 2 APPROACH

To facilitate the application of EuGene-P, we designed a fully automatic pipeline, described in Figure 1, that allows any user to directly apply EuGene-P starting just from genomic sequences and oriented sequence-based expression data (RNA-Seq or tiling array). The resulting Perl-based EuGene-PP(ipeline) has no parameter to tune (by default), accepts a variety of protein and expression datasets of different types under most usual formats and feeds EuGene-P with the following:

- Markov models of coding regions trained on regions with strong similarities with a reference protein databank.
- Regions of similarity with different protein databanks.
- A set of CDS predictions produced by a reliable self training *ab initio* gene finder. We use Prodigal (Hyatt *et al.*, 2010).
- A set of predicted non-coding RNA genes (ncRNA). We use tRNAscan-SE (Lowe and Eddy, 1997), rfam_scan (Griffiths-Jones, 2005) and RNAmmer (Lagesen et al., 2007).
- A set of thresholded and rescaled profiles of measured expression on each strand along the genome (either RNA-Seq or tiling arrays) showing transcription.
- A set of potential transcription start sites, defined as points of sudden increase in expression identified by the derivative of a smoothed version of the expression profile (Sallet *et al.*, 2013).

EuGene-PP can run using just FASTA genomic sequences and expression data (provided as oriented Single/Pair-end reads in either FASTQ/FASTA format, mapped reads in Bam/Sam, Bed or Wig format, or tiling array data in ndf/pair files). Protein databanks for similarity detection are configurable and may include organism-specific proteomes. EuGene-PP is provided with a Galaxy configuration Goecks *et al.* (2010) to deploy EuGene-PP through a Web interface.

The probabilistic model used by EuGene-P (Sallet *et al.*, 2013) integrates all this information and its own Ribosome Binding Site predictions to segment the genome in possibly overlapping coding regions, untranslated regions (UTR) and non-coding genes. The integration of expression data leads to more reliable

*To whom correspondence should be addressed.



 

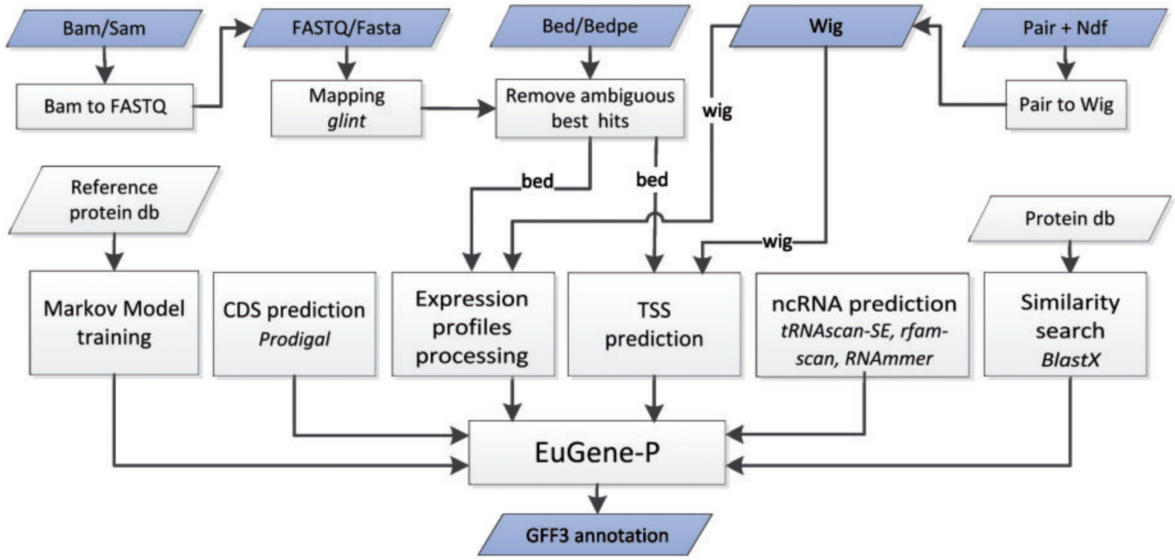


**Fig. 1.** A diagram describing the input and formats accepted by EuGene-PP and how information is prepared for EuGene to produce a final GFF3 annotation

**Table 1.** For each annotation, we report the percentage of shared stops, the number of ncRNA genes (Rfam and predicted), the total number of bases represented, the number of reference genes covered on >50% of their length by predicted genes (cover) or with a reciproqual hit covering at least 50% of both regions (recip.)

| Source | Shared | Number of ncRNA | | Size | > 50% | > 50% |
|---|---|---|---|---|---|---|
| | CDS | Rfam | pred. | (kbp) | cover | recip. |
| *B.subtilis* | | | | | | |
| EuGene-PP | 97% | 207 | 2492 | 817 | 98 | 55 |
| Nicolas *et al*., 2012 | | 207 | 1600 | 503 | 71 | 66 |
| *S.avermilitis*, (Moody *et al*., 2013) | | | | | | |
| EuGenePP | 95% | 162 | 166 | 20 | 56 | 34 |
| *E.coli*, (Li *et al*., 2013) | | | | | | |
| EuGenePP | 96% | 263 | 145 | 20 | 97 | 61 |
| *S.enterica*, (Kröger *et al*., 2013) | | | | | | |
| EuGenePP | 96% | 290 | 3456 | 299 | 146 | 86 |

*Note*: Both annotations show comparable quality albeit for a better behavior of the curated annotation of (Nicolas *et al*., 2012) for reciprocal hits. This is explained by split/merged genes in Eugene-PP. Also, the curated annotation includes 56 GenBank annotated ncRNA genes, each with a perfect match in the reference set. On three additional genomes, EuGene-PP recovers a similar fraction of the Rfam set.

transcripts and TSSs prediction. Prediction is performed independently on each strand, allowing for the prediction of antisense genes.

## 3 RESULTS AND DISCUSSION

In Sallet *et al*. (2013), we showed how EuGene-P performed when applied to the genome of the symbiont bacteria *Sinorhizobiummeliloti* and associated oriented RNA-Seq data. Besides its 6 308 CDS, the produced annotation contains 1876 ncRNA genes. These ncRNA predictions, with a mean length of 107 nt, cover a large fraction of already characterized or candidate ncRNA genes. Furthermore, by looking for specific RpoE2-binding sites upstream of predicted TSSs, the *S.meliloti* RpoE2 regulon could be extended by 3-fold, showing the added value of predicted TSSs.

To complete this application of EuGene-P, we decided to compare the fully automated annotation produced by EuGene-PP with a recently published curated annotation of the model bacteria *Bacillus subtilis*. This annotation is based on a number of condition-specific expression measures based on tiling arrays (Nicolas *et al*., 2012). For CDS, the two annotations were highly consistent, with >97% of shared CDSs (same STOP). This is consistent with the reliability of Prodigal *ab initio* CDS prediction. We therefore focused our evaluation on ncRNA transcript prediction. We used rfam_scan to produce a set of 207 reference ncRNA genes. We separately applied EuGene-PP using a subset of all tiling-arrays and removing all input from rfam_scan, RNAmmer or tRNAscan-SE. The comparison of the automated annotation of EuGene-PP with the curated annotation of Nicolas *et al.* (2012) with reference to this reference gene set is given in Table 1. On three additional genomes, EuGene-PP recovers a similar fraction of the reference genes.

These results also show the flexibility of EuGene-PP that exploits a variety of information sources, under most usual formats, to produce an annotation comparable with a curated semiautomated structural annotation, especially on ncRNA genes, which are still difficult to predict.

*Funding*: This work was supported by the Agence Nationale de la Recherche ANR-08-GENO-106 grant 'SYMbiMICS', in TULIP LabEx (ANR-10-LABX-41).

*Conflict of interest*: none declared.

## REFERENCES

Delcher,A.L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.

Foissac,S. *et al.* (2008) Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.*, **3**, 87–97.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Griffiths-Jones,S. (2005) Annotating non-coding RNAs with Rfam. *Curr. Protoc. Bioinform.*, **12**, 125.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Kröger,C. *et al.* (2013) An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*, **14**, 683–695.

Lagesen,K. *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.

Li,S. *et al.* (2013) Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E.coli k12 through accurate full-length transcripts assembling. *BMC Genomics*, **14**, 520.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Martin,J. *et al.* (2010) *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics*, **11** (**Suppl 3**), S10.

Moody,M.J. *et al.* (2013) Comparative analysis of non-coding rnas in the antibiotic-producing streptomyces bacteria. *BMC Genomics*, **14**, 558.

Nicolas,P. *et al.* (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*, **335**, 1103–1106.

Sallet,E. *et al.* (2013) Next-generation annotation of prokaryotic genomes with eugene-p: Application to *Sinorhizobium meliloti* 2011. *DNA Res.*, **20**, 339–354.

Zickmann,F. *et al.* (2014) GIIRA-RNA-seq driven gene finding incorporating ambiguous reads. *Bioinformatics*, **30**, 606–613.