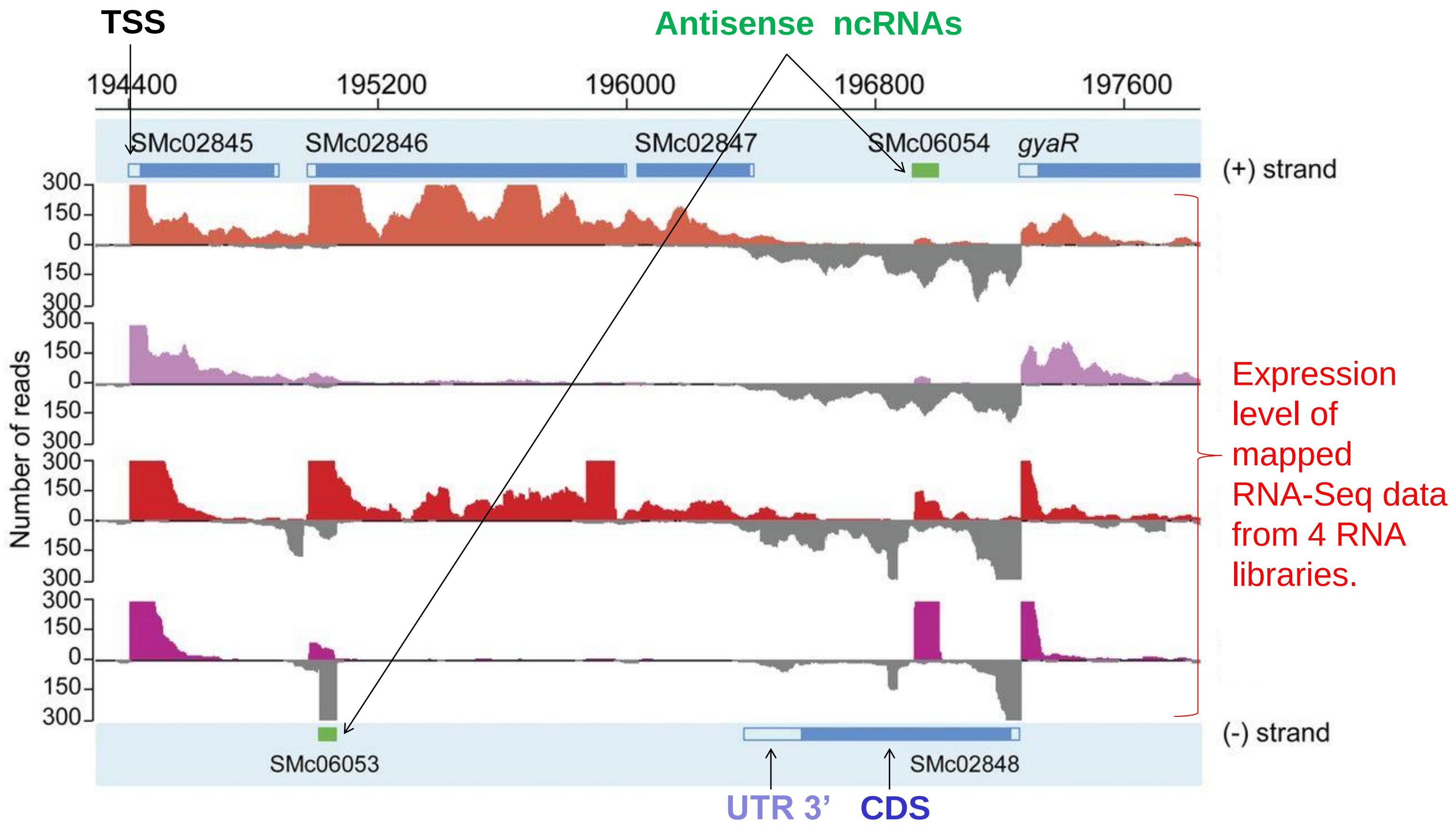
EuGene-PP
Automatic and comprehensive annotation Pipeline of Prokaryotic genome with oriented RNAseq

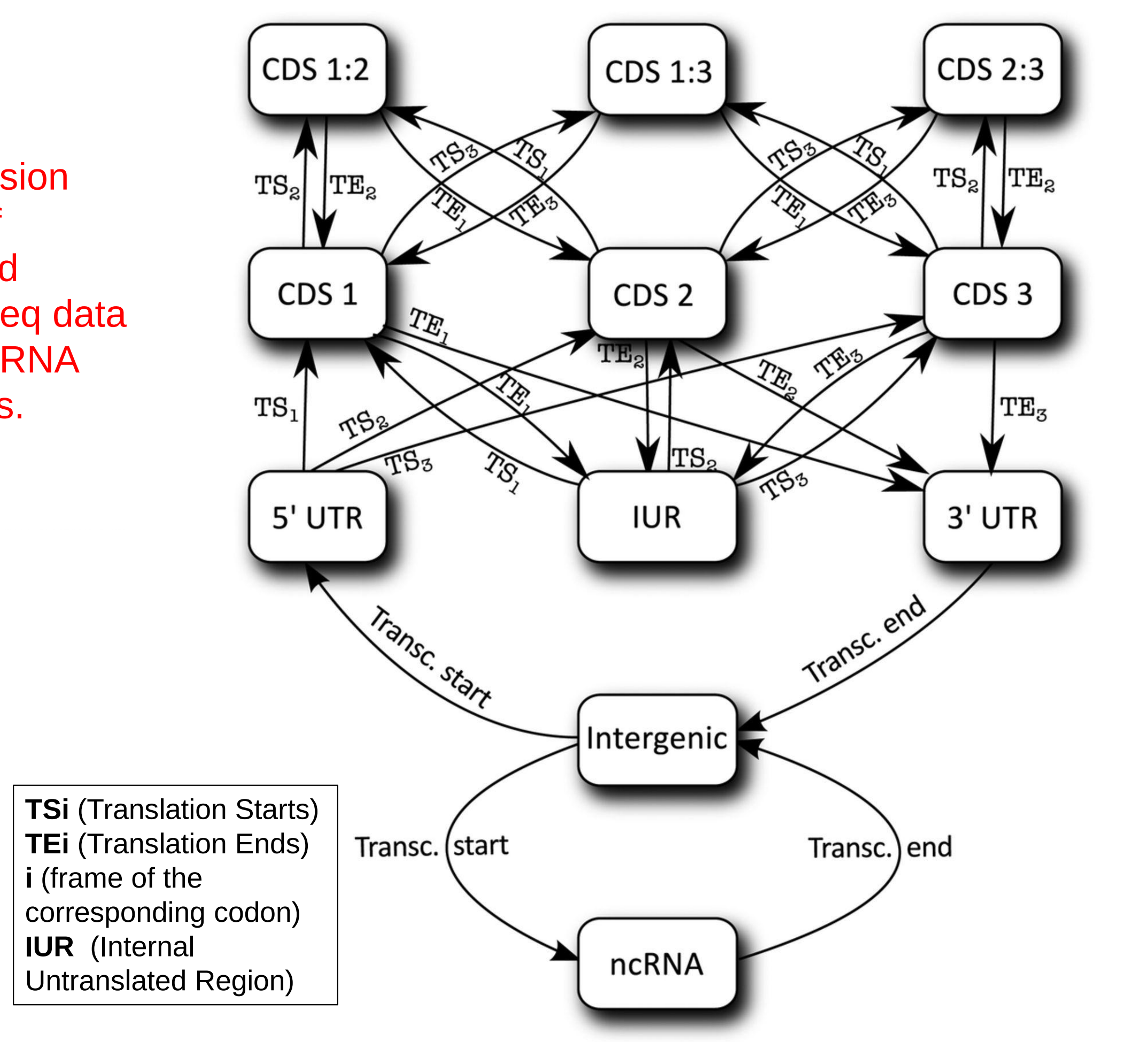
Erika SALLET[1], Emmanuel COURCELLE[1], Thomas FARAUT[2], Jérôme GOUZY[1] and Thomas SCHIEX[3]
[1]Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441 INRA, UMR2594 CNRS, Castanet-Tolosan, F-31326, France.
[2] Laboratoire de Génétique Cellulaire, UMR 444, INRA ENVT Castanet-Tolosan, F-31326, France
[3] Mathématiques et Informatique Appliquées Toulouse (MIAT), UR875 INRA, Castanet-Tolosan, F-31326, France.
Erika.Sallet@toulouse.inra.fr, Jerome.Gouzy@toulouse.inra.fr, Thomas.Schiex@toulouse.inra.fr

With the new generation of sequencing (NGS) technologies, bacterial and archeal genome projects now combine deep genomic sequencing with a variety of transcriptome libraries (see [1] for example). The transcribed sequences generated by deep sequencing can contribute to prokaryotic genome annotation by the elucidation of gene structural features, including transcription start sites (TSSs), 5' and 3' UnTranslated regions (UTRs), and the identification of non-coding RNA (ncRNA) genes. In the recent sequencing of bacterial and archeal genomes, the annotation has still been done manually due to the lack of appropriate tools to integrate RNA-Seq data [2]. Indeed, most existing prokaryotic gene finders [3] or higher level bacterial annotation system [4] are based on genomic sequence analysis and do not take into account available expression data in the structural prediction.
Here, we present **EuGene-PP (EuGene-Prokaryote Pipeline),** a fully automatic and generic bacterial annotation pipeline capable of producing a qualitatively enriched structural genome annotation.



RNA-seq data highlight complex and dense genome structure (overlapping genes and/or ncRNA) requiring **a strand specific annotation**

We recently adapted the eukaryotic gene finder EuGene[5] to the specific requirements of gene identification in prokaryotes. We used this extended EuGene version to annotate the genome of the bacteria *Sinorhizobium meliloti* (*Sm*) strain 2011. This raw annotation was then submitted to manual checking, leading to the prediction of 6 308 CDSs as well as 1 940 ncRNAs[6]. Based on this experience we developed EuGene-PP to propose a prokarotic fully automatic annotation pipeline.

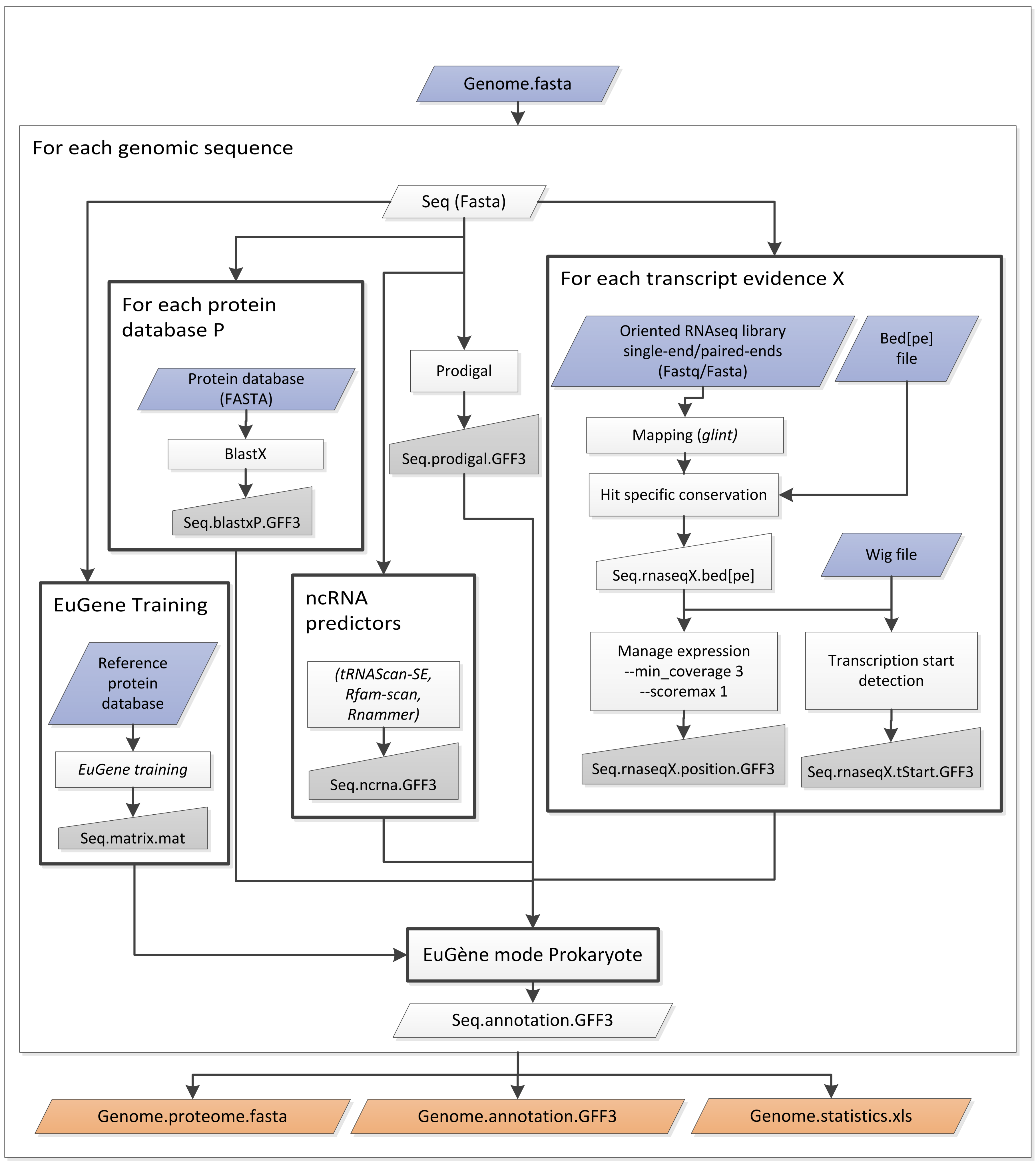


Simplified automaton represented the EuGene prokaryotic gene model

Feature type	Number predicted by EuGene-PP	Variation compared with the reference annotation
CDS	**6 621**	**+4.96%**
Identical (start-end)	5 670 (89.89%)	
90% Overlap (*)	6 154 (97.56%)	
New	283	
Removed	34	
ncRNA	**1 986**	**+2.37%**
90% Overlap (*)	1286 (66.29%)	

We performed a fully automatic annotation of *Sm* genome with EuGene-PP. The table compares these results with the reference annotation[6]. Most of the CDS differences are due to the edition of the translation starts.
(*) CDS 90% Overlap = The number of CDS that overlap at least 90% of a CDS of the reference annotation (and reciprocally)

EuGene-PP annotation process



EuGene-PP has a simple fully automatic use, minimal requirements :
- a directory with genomic sequences,
- a directory with evidence files (fastq, fasta, wig, bed format allowed)
- a key/value configuration file

```
>ls –R inputdir
    inputdir/data:
       Sm_1_seq_GGK-37.fastq.xz        Sm-GGK21.ope.1.fastq.gz
       Sm_2_seq_GGK-37.fastq.xz        Sm-GGK21.ope.2.fastq.gz
    inputdir/genome:
       seq1.fna    seq2.fna

>egn-prok.pl --indir $PWD/inputdir --outdir $PWD/outdir --cfg egnpp.cfg

>ls –R outdir
    seq1.gff3  seq2.gff3  sequences.gff3 sequences_prot.fna
    sequences.general_statistics.xls  sequences.statistics_per_gene.xls
```

All training procedures required for gene finding are performed inside EuGene-PP. The pipeline is able to manage genomes with peculiar replicons (e.g: strong GC% bias compared to the rest of the genome)

EuGene-PP integrates various sources of evidence
- ➢ High throughput strand-specific RNA-Seq data
- ➢ Intrinsic information provided by coding potential (Interpolated Markov Models)
- ➢ Stop and Start codon analysis (using a dedicated RBS alignment tool)
- ➢ Similarities with known proteins (SwissProt by default)
- ➢ Gene prediction results:
    - ➢ High quality CDS predictions (Prodigal [3])
    - ➢ ncRNA predictions (tRNAscan-SE, RNAmmer and Rfam-scan software)

Time consuming task are parallelized via Paraloop software [7] (SGE cluster, multiprocessor system).
It takes 12 hours to annotate the *S meliloti* genome (6.7Mb) with 19 RNAseq libraries (~476M reads)

EuGene-PP is written in Perl and is distributed under CeCILL license. It encapsulates the C++ annotation tool EuGene (Artistic license). EuGene-PP will be soon available at http://eugene.toulouse.inra.fr

Acknowledgements
This work was supported by the grant ANR-08-GENO-106 "SYMbiMICS". This research was done in the Laboratoire des Interactions Plantes-Microorganismes, part of the Laboratoire d'Excellence (LABEX) entitled TULIP (ANR-10-LABX-41).

References

1. Weissenmayer, B. A., Prendergast, J. G. D., Lohan, A. J. and Loftus, B. J., Sequencing illustrates the transcriptional response of *Legionella pneumophila* during infection and identifies seventy novel small non-coding RNAs, *Plos One*, 6, e17570. 2011.
2. Richardson, E. J. and Watson, M., The automatic annotation of bacterial genomes, *Brief Bioinform*., 14, 1-12. 2013.
3. Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. and Hauser, .J., Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, 11, 119. 2010.
4. Aziz Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E. and Tobes, R., BG7: A new approach for bacterial genome annotation designed for next generation sequencing data, *PLoS One*, 7, e49239. 2012.
5. Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., Sterk, L., van de Peer, Y., Rouzé, P., Schiex, T. Genome Annotation in Plants and Fungi: EuGène as a Model Platform. *Current Bioinformatics*, Volume 3, Number 2, p. 87-97, 2008
6. Sallet, E., RouxB., Sauviac L., Jardinaud, F., Carrère, S., Faraut, T., de Carvalho-Niebel, F., Gouzy, J., Gamas , P., Capela, D., Bruand, Caude and Schiex, T. Next Generation Annotation of prokaryotic genomes with EuGene-P: application to *Sinorhizobium meliloti*. *DNA Research*, 2013.
7. http://lipm-bioinfo.toulouse.inra.fr/paraloop