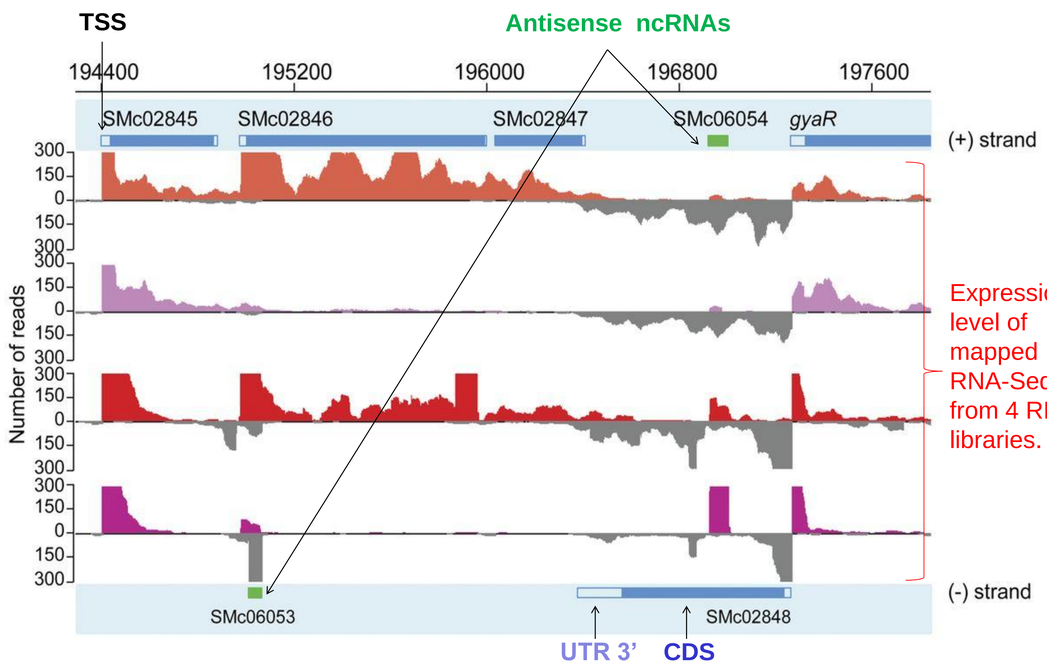
EuGene-PP
A next-generation automated annotation pipeline for prokaryotic genomes

Erika SALLET[1], Jérôme GOUZY[1] and Thomas SCHIEX[2]
[1]Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441 INRA, UMR2594 CNRS, Castanet-Tolosan, F-31326, France.
[2] Mathématiques et Informatique Appliquées Toulouse (MIAT), UR875 INRA, Castanet-Tolosan, F-31326, France.
Erika.Sallet@toulouse.inra.fr

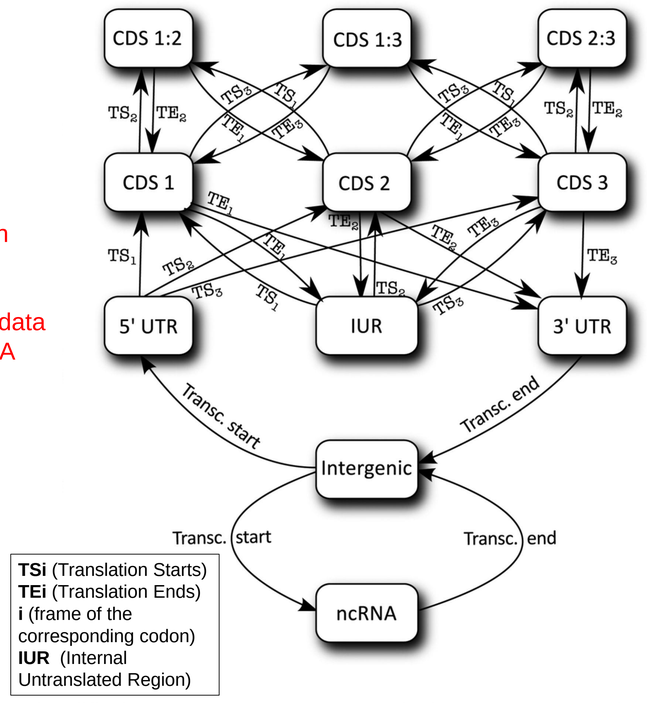
With the new generation of sequencing technologies, bacterial genome projects now combine deep genomic sequencing with a variety of transcriptome libraries. The transcribed sequences can contribute to genome annotation by the elucidation of gene structural features, including transcription start sites (TSSs), untranslated regions (UTRs) and the identification of non-coding RNA (ncRNA) genes. Existing prokaryotic gene finders are either *ab initio* gene finders that identify only coding regions (CDS) [1,2] or purely RNA-Seq-based gene finders predicting transcripts and are much less effective than their *ab initio* competitors for CDS prediction [3]. Reconciling conflicting predictions is a tedious work, which is incompatible with the growing prokaryotic genome sequencing rate.
Here, we present EuGene-PP (EuGene-Prokaryote Pipeline) [4], a fully automatic and generic bacterial annotation pipeline capable of producing a qualitatively enriched structural genome annotation.

RNA-Seq data highlight complex and dense prokaryotic genome structure



Prokaryotic gene model

We adapted the eukaryotic gene finder EuGene[5] to the specific requirements of gene identification in prokaryotes: possibly overlapping genes, operon structure, possibly antisense ncRNA.



TSi (Translation Starts)
TEi (Translation Ends)
i (frame of the corresponding codon)
IUR (Internal Untranslated Region)

Simplified automaton represented the EuGene prokaryotic gene model[6]

➢ We compared the annotation produced by EuGene-PP with a curated annotation of *Bacillus subtilis* [7]. We used rfam_scan to produce a set of 207 reference ncRNA genes. We applied EuGene-PP using a selection of 59 tiling-arrays data and removing all inputs from rfam_scan, RNAmmer or tRNAscan-SE.

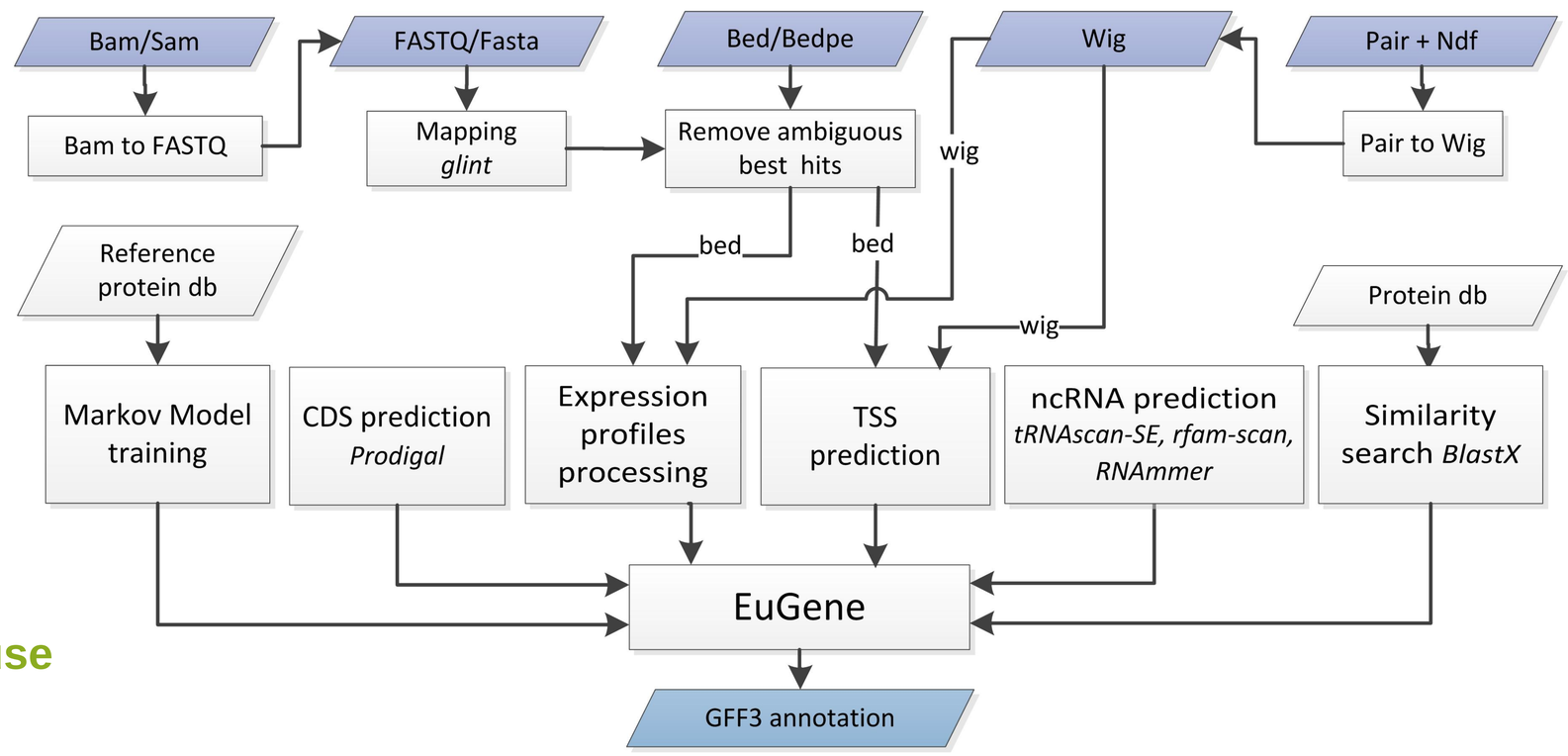
	EuGene-PP	Nicolas *et al.*
Shared CDS	**97%**	
Predicted ncRNA	2492	1600
Number of reference ncRNA covered on 50% of their length by predicted ncRNA	98	71
Number of reference ncRNA with a reciproqual hit covering at least 50% of both regions	55	66

➢ We annotated the genome of *Sinorhizobium meliloti* strain 2011[6]. The ncRNA predictions (1876 genes) cover a large fraction of already characterized 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.

EuGene-PP annotation process

Integration of various sources of evidence

➢ Oriented sequence-based expression data (RNA-Seq or tiling array data)
➢ Intrinsic information provided by coding potential (Interpolated Markov Models)
➢ Similarities with known proteins (Swiss-Prot by default)
➢ Gene prediction:
  ➢ High quality CDS predictions (Prodigal [2])
  ➢ ncRNA predictions (tRNAscan-SE, RNAmmer and rfam-scan)
➢ Start codon analysis (using own Ribosome Binding Site predictor)



Simple fully automatic use

Minimal requirements :
➢ a directory with genomic sequences
➢ a directory with expression data
  Allowed formats : fastq, fasta, bam/sam, wig, bed[pe], pair+NDF (NimbleGen Design File)
➢ 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
    seq1-2.gff3      seq1-2.general_statistics.xls
    seq1-2_prot.fna  seq1-2.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)

Time consuming task are parallelized via Paraloop software [8] (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 the CeCILL license. It encapsulates the C++ annotation tool EuGene (Artistic license). It is provided with a Galaxy configuration to deploy EuGene-PP through a web interface.

EuGene-PP is 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. Delcher, A.L., Bratke, K.A., Powers, E.C., Salzberg, S.L. Identifying bacterial genes and endosymbiont DNA with Glimmer *Bioinformatics* 23:673-679, 2007
2. Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. and Hauser, L.J., Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, 11, 119. 2010.
3. Zickmann, F., Lindner, M.S., Renard, B.Y. GIIRA-RNA-seq driven gene finding incorporating ambiguous reads. *Bioinformatics* 30:606-613. 2014
4. Sallet, E., Gouzy, J., Schiex, T. EuGene-PP a next-generation automated annotation pipeline for prokaryotic genomes *Bioinformatics* 30 (18): 2659-2661. 2014
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. Nicolas, P. *et al*. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* 335:1103-1106 2012
8. http://lipm-bioinfo.toulouse.inra.fr/paraloop