Arabidopsis structural annotation is not perfect!

PEPI-IBIS-GT Annot + MERIT GT Annotation

17 mars 2025

Véronique Brunaud

IPS2 - GNet team

IPS2 Institute of Plant Sciences Paris - Saclay

université PARIS-SACLAY

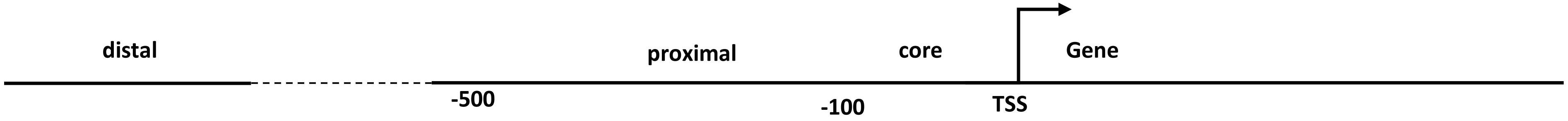
Background

The regulation of transcription

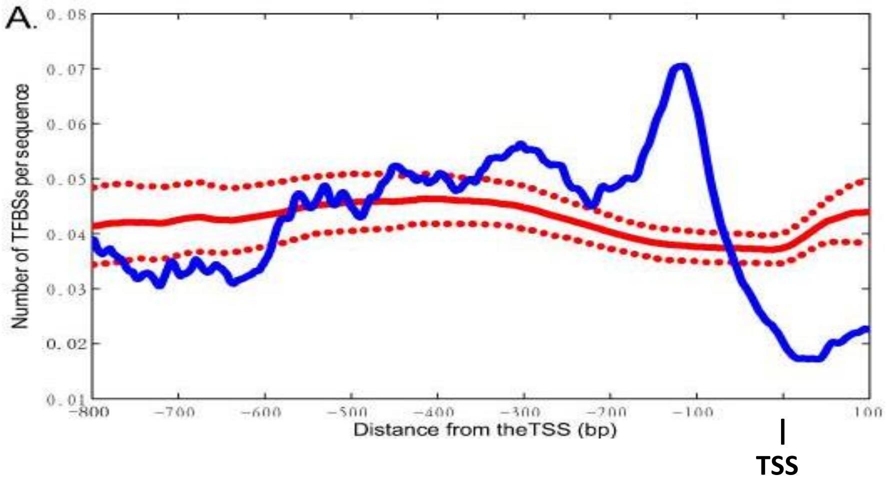
Margot thesis is on detection of transcription factor binding site (TFBS) associated to stress gene responses using in silico method PLMdetect

PLMdetect: detection of enriched motifs in gene proximal regions

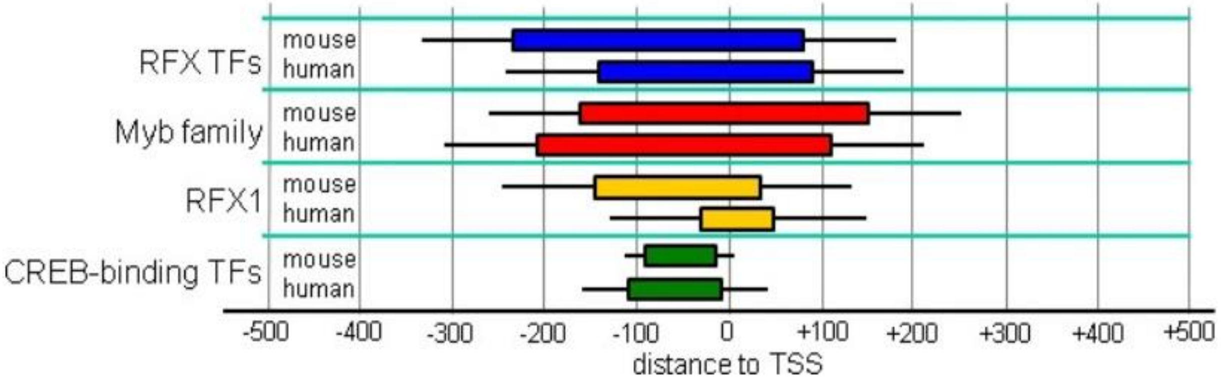
→ need “correct” UTR/bounds of genes



Gene proximal regions enriched in TFBS=CRE



In yeast since 2004 (Harbison et al.) and in 2010 Lin et al.



In 2013, Vandenbon et al. in mice and humans

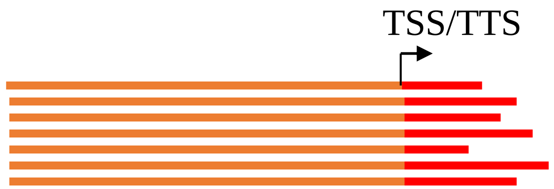
In Arabidopsis, the 2016 article by Yu et al. confirms this enrichment of 86% of TFBS in the [-1000, +200] region

The TFBS are very short DNA sequences from 5 to 20 bases

→ **PLMdetect focuses the prediction of theses motifs in these proximal regions**

PLMdetect usage

Extraction of gene proximal regions
[-1000, TSS,+500] or [-500,TTS,+1000]

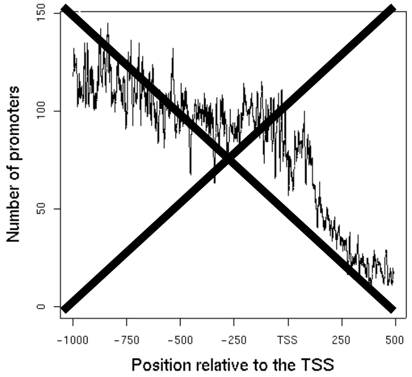


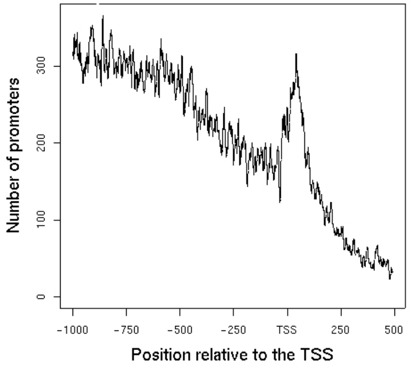
1- Aligned regions from a gene list (i.e. DEG)

PLMdetect

2- a list of DNA motifs (TFBS known, or new motifs)

Motif occurences





3- PLM are significant motif over-represented

4- PLM features

→ From promoters of genes to Preferentially Located Motifs (PLMs)

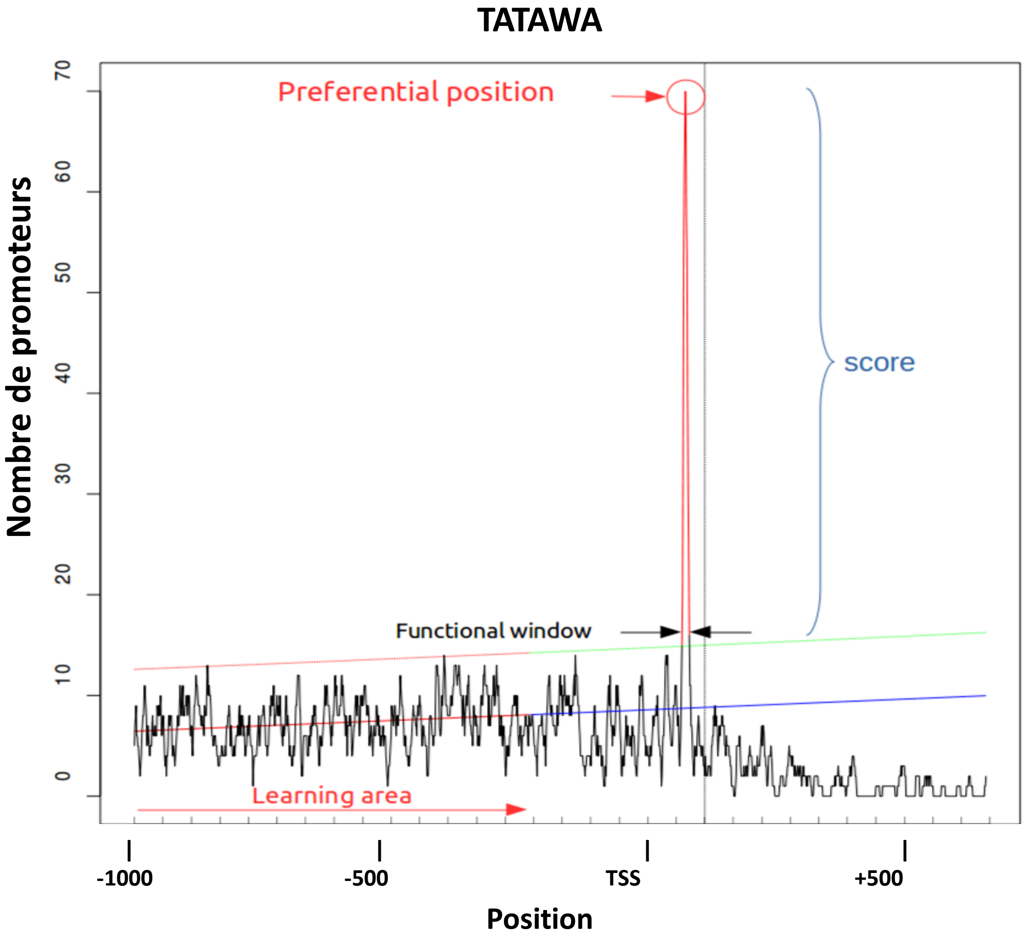
PLMdetect - Method

→ Detect Preferentially Located Motif (PLM)

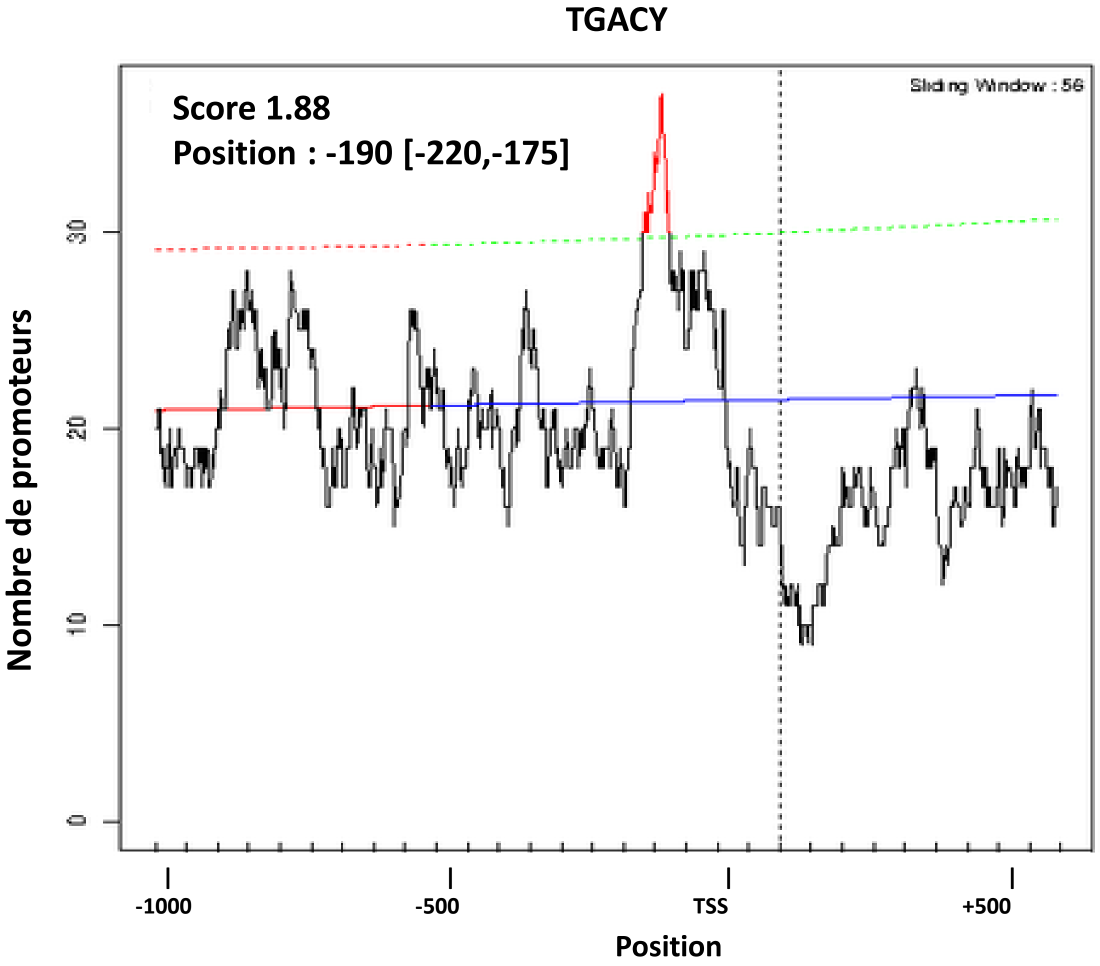
→ PLM have preferential position compared to TSS/TTS

- ❑ Define a reference for motif occurrences, region [-1000, -500]
- ❑ Calculate mean line and confidence interval

→ **PLM = Peak and area above this interval**



PLMdetect – PLM exemple



PLM TGACY (WRKY-box) defined by :

- ✓ Preferential windows: -190 [-220,-175]
- ✓ List of genes with PLM in these preferential win
- ✓ Score : 1.88 (peak size) and graphic output

Background

Method developed by Margot shows that the positions of these PLMs (known as TFBS) match with corresponding TF (peaks of ChIP/DAP-seq)

- First the method was developed with annotation from TAIR10

→ List of ~300 PLM

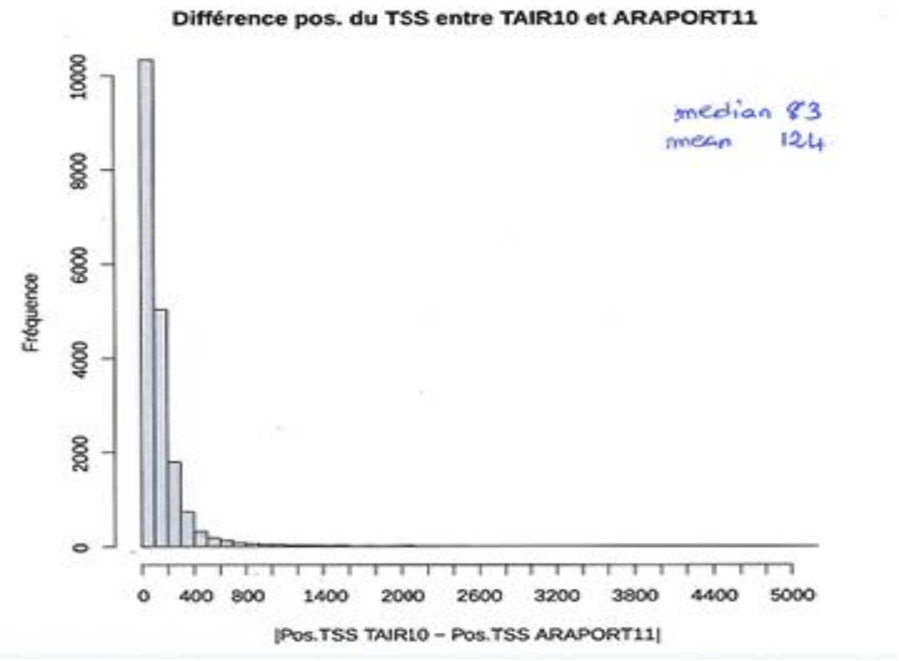
→She moved to last annotation ARAPORT11

Background

Margot thesis is on detection of TFBS associated to stress gene responses via PLMdetect, *in silico* method

She moved to last annotation ARAPORT11

- ➢ Problem : the PLM (Preferred Located Motifs) previously enriched close to the TSS have not been fully retrieved !
- ➢ She found a difference of 100 bases on mean between TSS definitions from TAIR10 and ARAPORT11



median 83
mean 124

Annotation of Arabidopsis

Araport11: a complete reannotation of the Arabidopsis thaliana reference genome Article 2016 (Cheng et al.)

- The TAIR10 genome annotation was informed by *ab initio* gene models, EST sequences from Sanger platforms, and two RNA-Seq datasets available at that time (Lamesch *et al*. 2012) → This annotation is oriented to detect coding genes
- Araport11 annotation (cheng et al. 2016) used 113 RNA-Seq datasets partitioned into 11 groups according to their tissue or organ of origin → oriented new genes, new isoformes and non-coding genes.

Protein-coding genes	TAIR10	ARAPORT11	Change
Total number of loci (coding protein)	27 416	27 655	+239
Number of transcript isoforms	35 386	48 359	+12 973
Total number of loci	33 602	38 194	**+4592**

Annotation of Arabidopsis

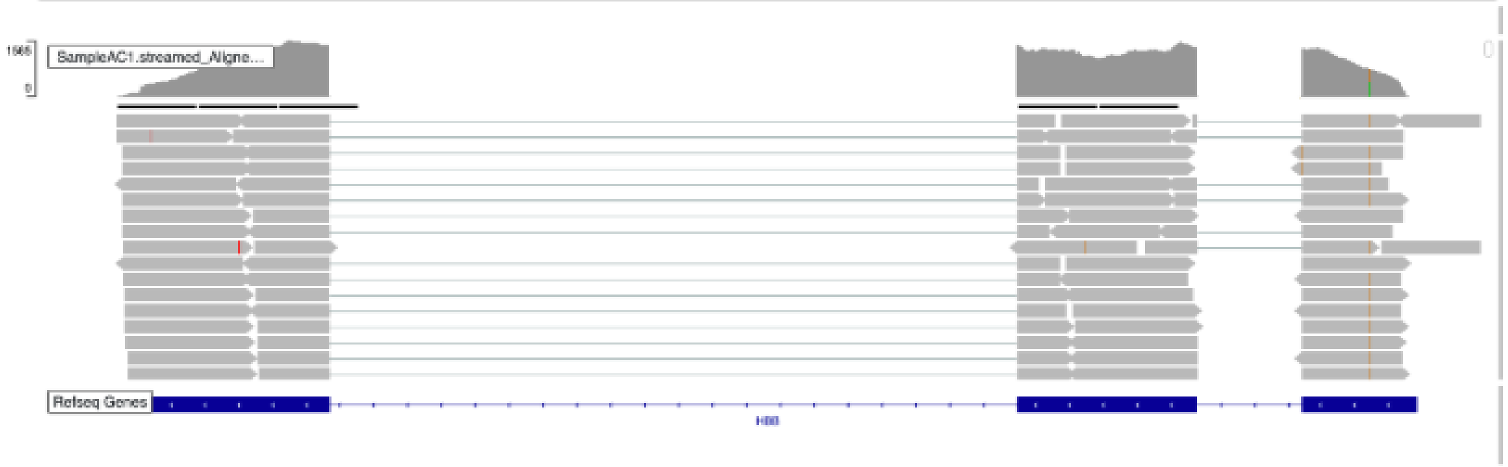
Article : Araport11: a complete reannotation of the Arabidopsis thaliana reference genome Article paru en 2016 (Cheng et al.)

Protein-coding genes	TAIR10	ARAPORT11	Change
Total number of loci	27 416	27 655	+239
Number of transcript isoforms	35 386	48 359	+12 973
Number of loci with changes in UTR(s)	–	–	**+21 298 (77%)**
Total number of loci (with non-coding)	33 602	38 194	+4592

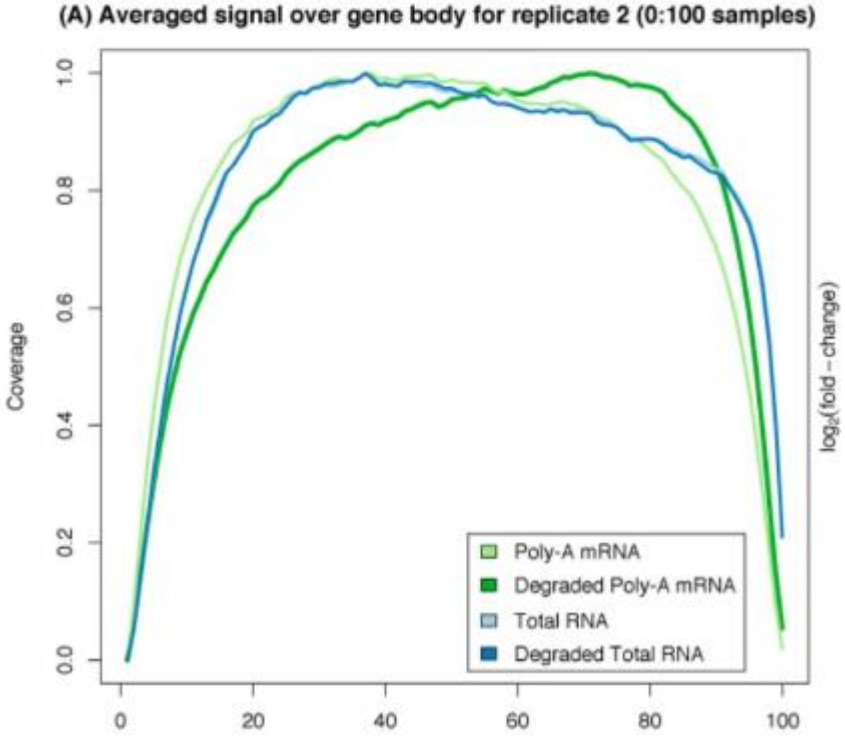
→ ARAPORT11 annotation is therefore greatly improved compared to gene detection

→ The annotation pipeline: « Union of the independently generated PASA annotation updates of the 11 tissues was created using a Python script (annotation.consolidate) (Tang et al., 2015) **to collapse the isoforms sharing the same splicing structure within a given locus while allowing for variation in UTR. The representative for each locus was identified as the isoform encoding the longest CDS »**

RNA-seq : coverage with short-reads



→ The coverage of gene extremities is not complete



doi: 10.1093/nar/gkw1063

Articles highlighting the differences in annotation between TAIR10 and ARAPORT11

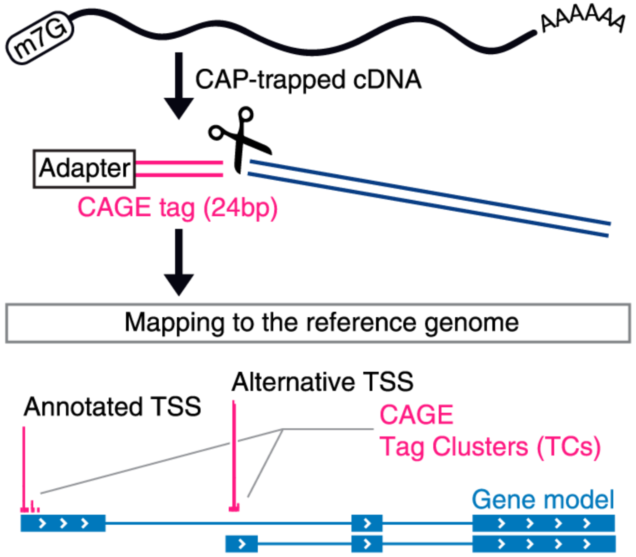
Characterization of Arabidopsis Thaliana Promoter Bidirectionality and Antisense RNAs by Inactivation of Nuclear RNA Decay Pathways. (Thieffry et al. 2020)

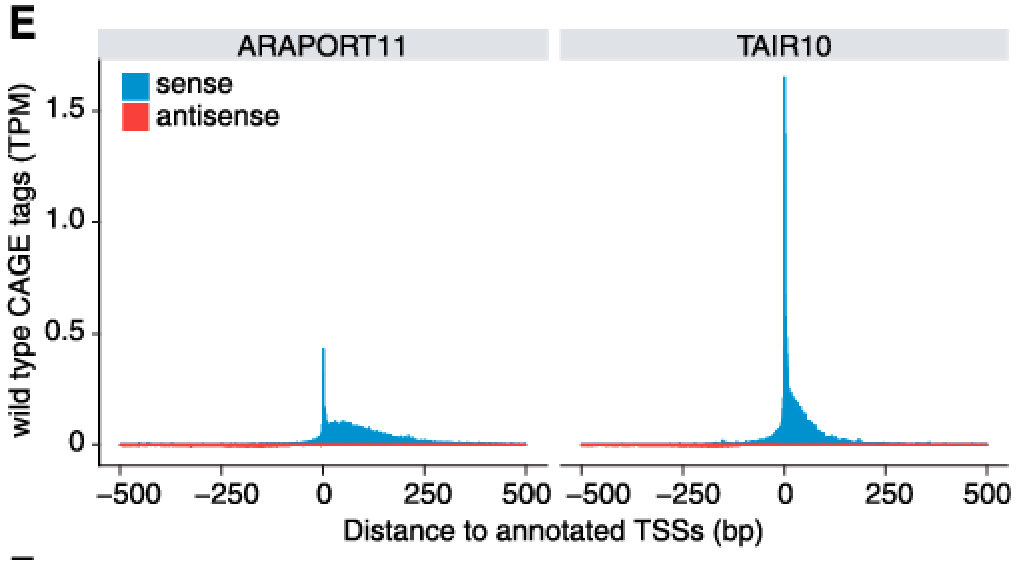
TrancriptomeReconstructoR: data-driven annotation of complex transcriptomes (Ivanov et al. 2021)

→ 2 publications supported by experimental methods to improve annotation and detection of gene bounds

Articles highlighting the differences in annotation between TAIR10 and ARAPORT11

Characterization of Arabidopsis Thaliana Promoter Bidirectionality and Antisense RNAs by Inactivation of Nuclear RNA Decay Pathways. (Thieffry et al. 2020)





→ As expected, the vast majority of TCs fell into annotated promoter regions (ARAPORT11 or TAIR10), although ARAPORT11 promoters accounted for a smaller fraction of TCs than TAIR10

→The distribution of CAGE signal around annotated TSSs was much broader (Figure 1E), pointing to possible inaccuracies in TSS annotation in ARAPORT11

Articles highlighting the differences in annotation between TAIR10 and ARAPORT11

TrancriptomeReconstructoR: data-driven annotation of complex transcriptomes (Ivanov et al. 2021)

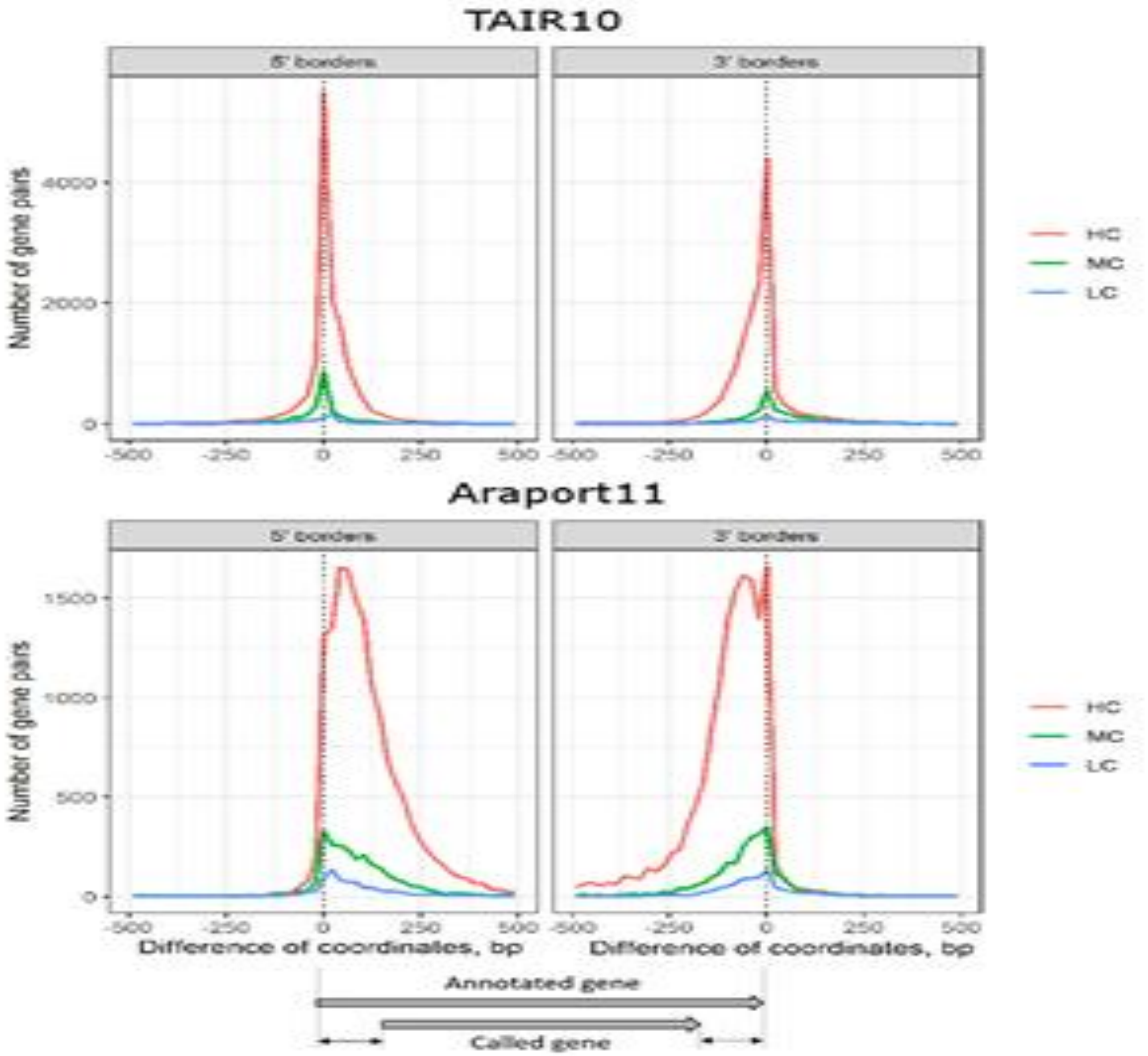
- Limits of classical RNA-seq: Poorly defined gene ends with mapping decrease at gene edges.- Definition of isoforms is complicated in short RNA-seq- Not sure how to distinguish steady-state RNA from non-coding RNA, for example.
- Long RNA-seq (ONT) is better but still has a high error rate (rq true in 2021, largely improve in 2024) but bias towards 3' (RNA fragmentation, polyA selection etc.)
- **Experimental methods: CAGE-seq (5’),PAT-seq (3’)**
- a *de novo* gene and transcript model construction pipeline **TranscriptomeReconstructoR** which takes three datasets as input: (i) full-length RNA-seq (e.g. ONT Direct RNA-seq) to resolve splicing patterns; (ii) 5′ tag sequencing (e.g. CAGE- seq) to detect TSS; (iii) 3′ tag sequencing (e.g. PAT-seq) to detect polyadenylation sites (PAS).
- Finally, transcripts are divided into **High Confidence (HC**), Medium Confidence (MC) and Low Confidence (LC) groups, depending on the support from TSS and PAS datasets

Articles highlighting the differences in annotation between TAIR10 and ARAPORT11

TrancriptomeReconstructoR: data-driven annotation of complex transcriptomes (Ivanov et al. 2021)

- 85% and 61% of HC genes had at least 90% overlap with TAIR10 and Araport11 genes, respectively.
- Notably, the called 5′ and 3′ gene borders were systematically shifted downstream and upstream, respectively, from the genomic positions predicted by Araport11

The TAIR10 annotation is better to define gene bounds



Consequences –To do list

- For Margot's thesis: generate the 5' promoters to be studied to launch PLMdetect (done) - annotation of TAIR10 + new ARAPORT11 genes

  → **23724 genes with 5’UTR can be processed**

  → **PLMs are retrieved**

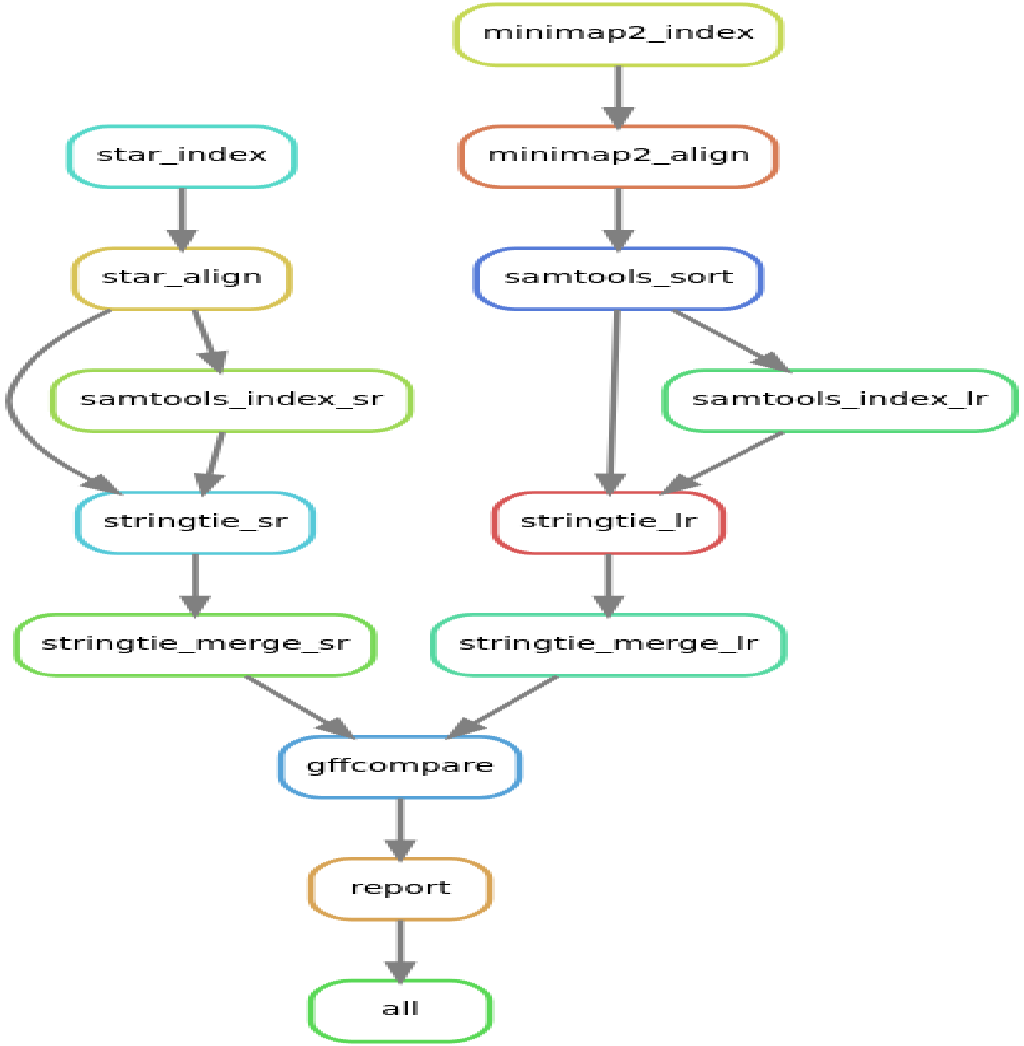
- Review the annotation of gene bounds via RNA-seq: SPS-bioinfo project with Hannah → **e2annot**
  - Pipeline done which can be used for all species
  - But some pb of gene fusion → so tune parameters and filter results

Projet e2annot: Extend Eukaryote Annotation with transcriptome data

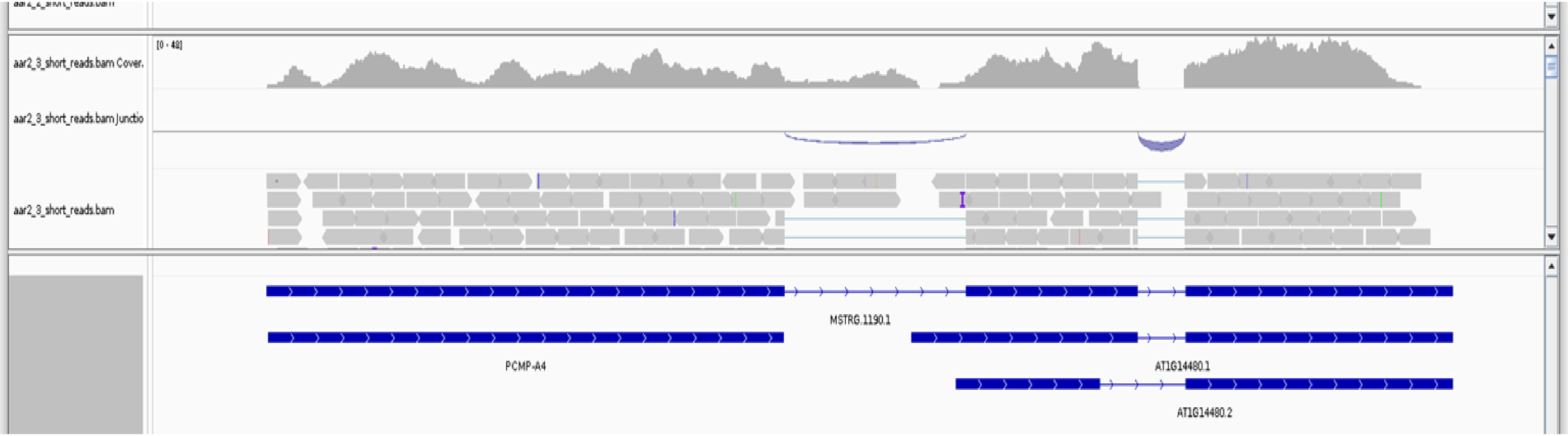
SPS-Bioinfo Project (Hannah Tomelka)

Developped a workflow (NextFlow)

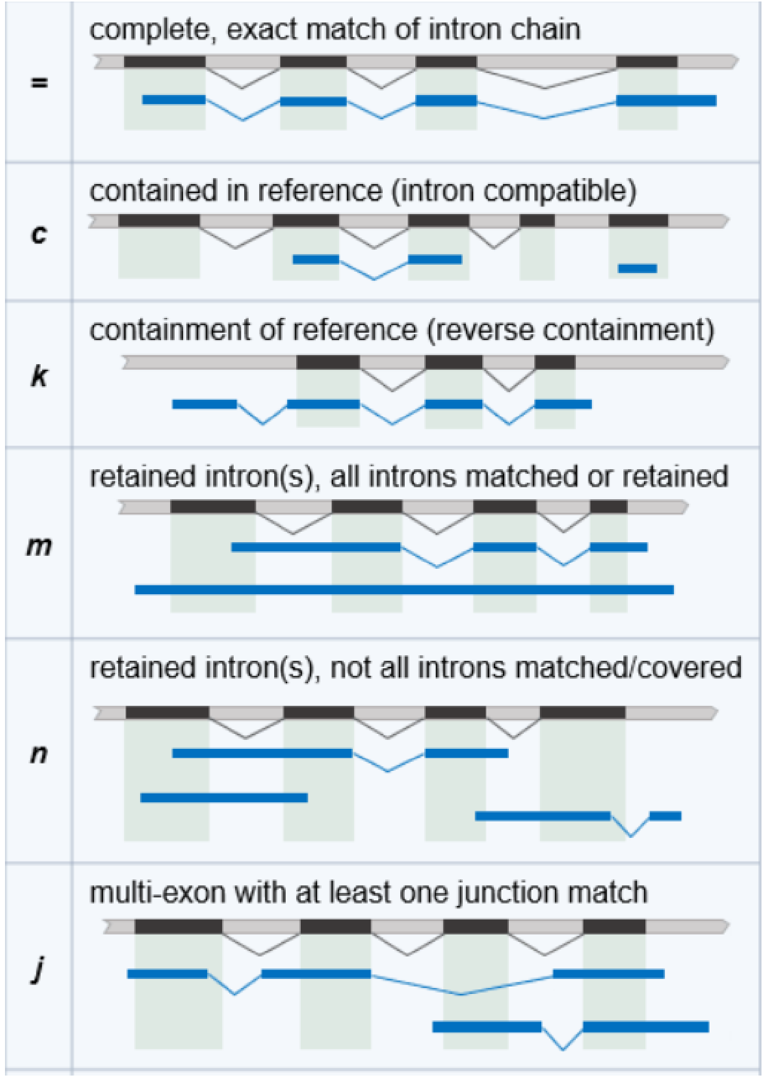
1. Using RNA-seq (Short and/or Long-reads)
2. Mapping against reference genome
3. Assembly transcripts
4. Generate GFF associated to transcripts
5. Compare GFF annotations



Projet e2annot : StringTie problem of gene fusion (reality or not!)



Projet e2annot : GFF compare class code, considering definition



=	complete, exact match of intron chain
c	contained in reference (intron compatible)
k	containment of reference (reverse containment)
m	retained intron(s), all introns matched or retained
n	retained intron(s), not all introns matched/covered
j	multi-exon with at least one junction match

Consider intron/exon exact even if extremities are notes exacts → so very interesting to explore these cases

Warning ! the comparison must contain not complete transcript

Consequences –To do list

- For Margot's thesis: generate the 5' promoters to be studied to launch PLMdetect (done) - annotation of TAIR10 + new ARAPORT11 genes

  → **23724 genes with 5’UTR can be processed**
- Review the annotation of gene bounds via RNA-seq: SPS-bioinfo project with Hannah → **e2annot**
  - Pipeline done which can be used for all species
  - But some pb of gene fusion → so tune parameters and filter results
- **A new TAIR12 version was expected 1st half of 2024 (new assembly and annotation of the genome)!**
- **Helixer ran on Arabidopsis genome**

→ **27200 annotated genes with UTR**

→ **To do list: compare with official annotation**

→ To continue…