Detecting transcriptional regulation despite incomplete annotation

Arnaud Liehrmann

JOBIM - July 2, 2026



Biological question

☞ What are the transcriptome differences between two biological conditions?

- molecular markers of a specific condition ?

Biological question

☞ What are the transcriptome differences between two biological conditions?

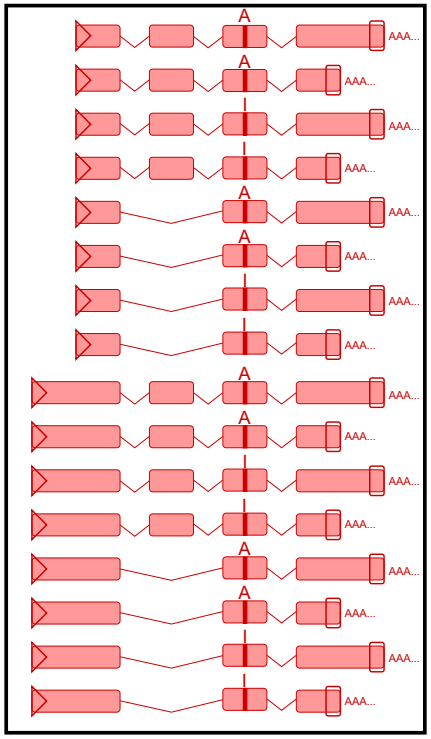
- molecular markers of a specific condition ?

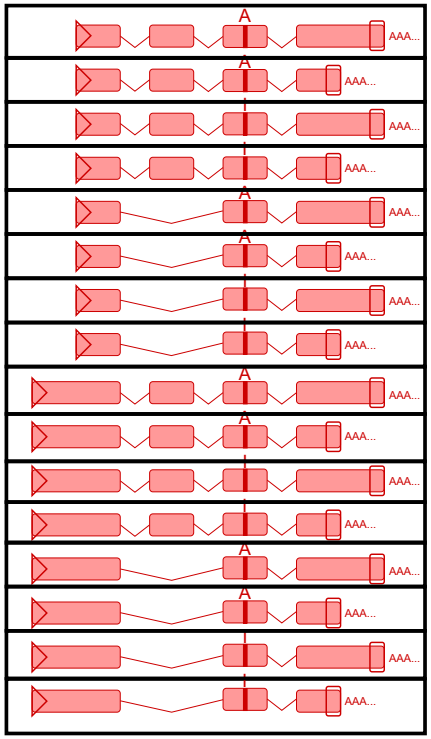
RNA-Seq experiment

- numerous analysis strategies (gene/events/isoforms)

Transcriptome analysis: two extreme views

Gene-level and isoform-level analysis



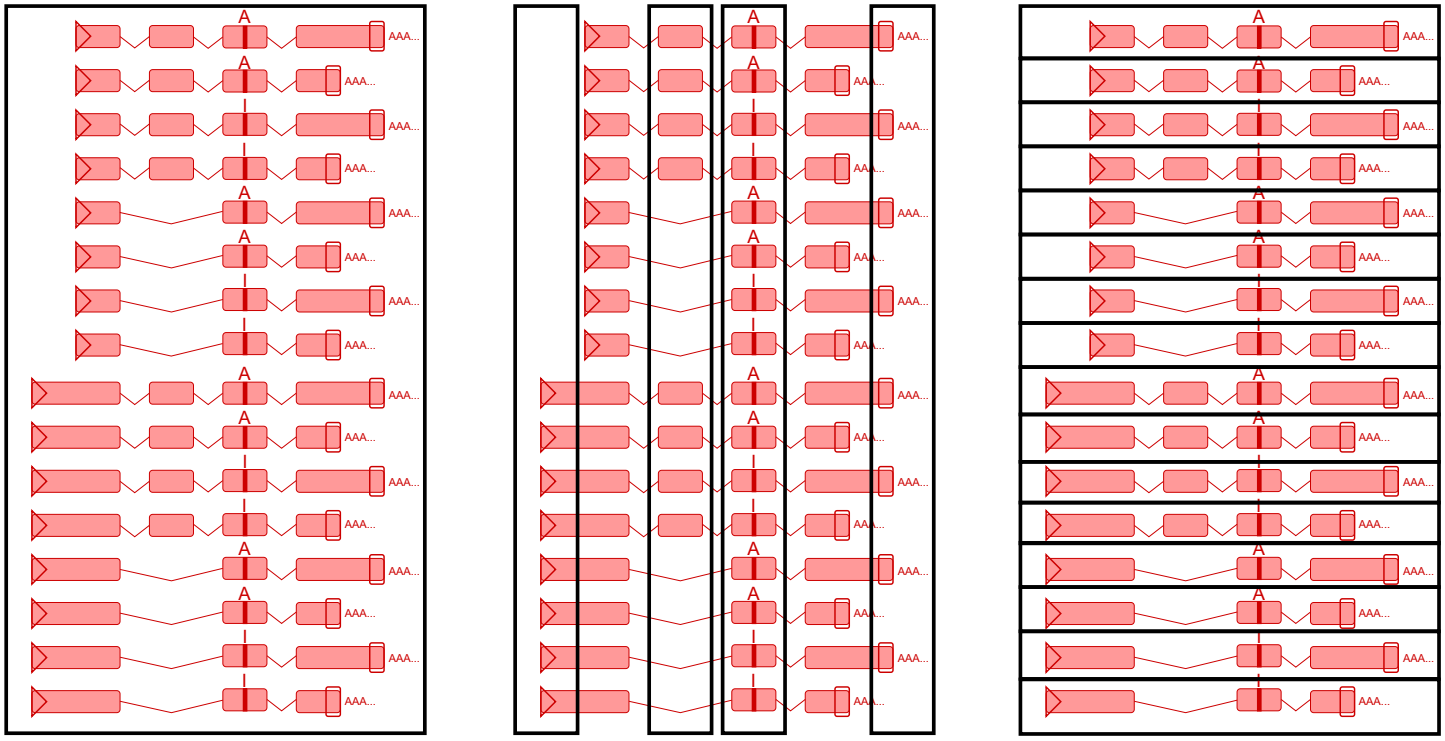


gene-level: miss post-transcriptional regulations [Rhinn et al., 2012]

isoform-level: technically, statistically and biologically more challenging [Soneson et al., 2019, Karlebach et al., 2022]

Transcriptome analysis: the correct scale

Event-level analysis



- should detect post-transcriptional regulations
- [events] <<< [isoforms] $\sim 2^{[\text{events}]}$

Biological question

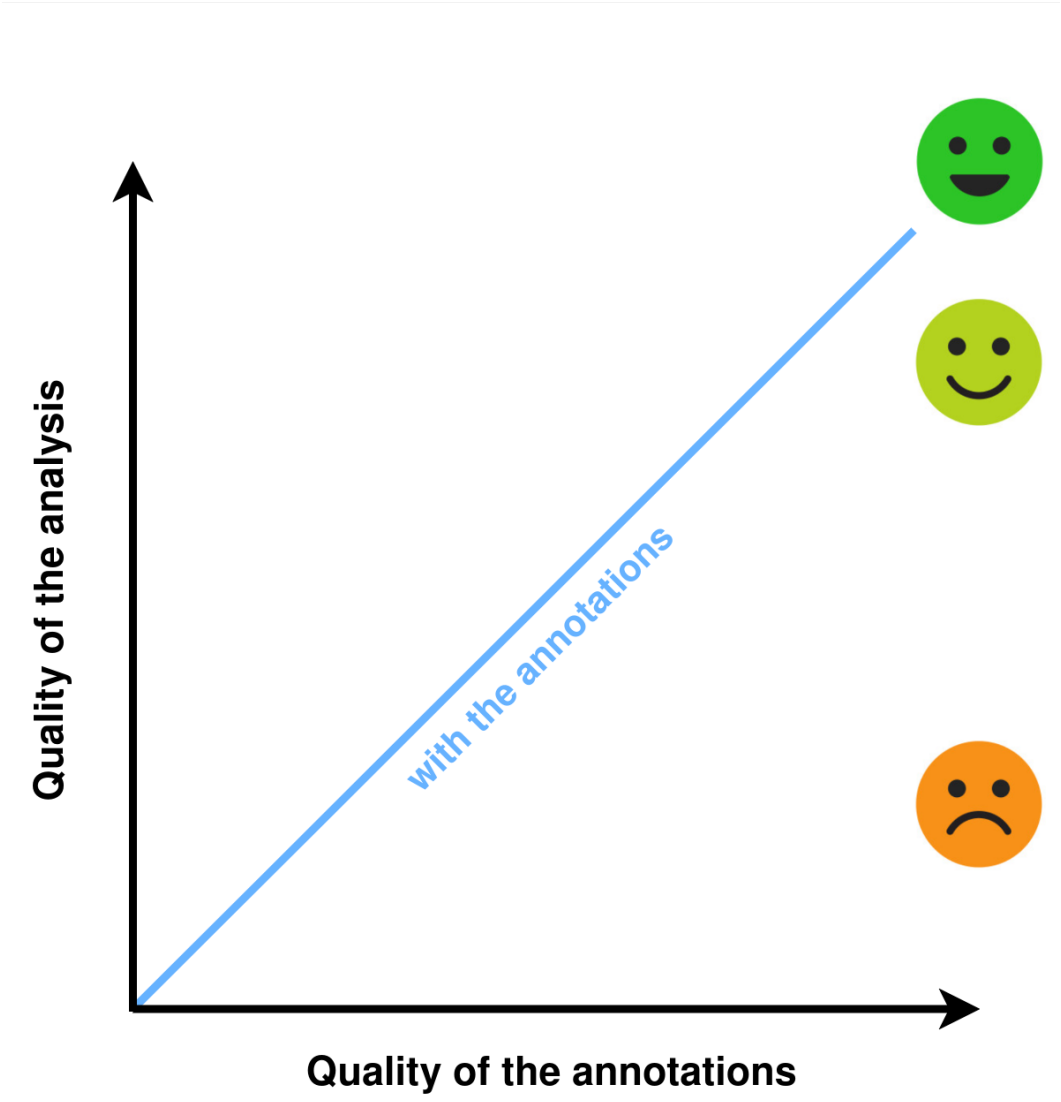
☞ What are the transcriptome differences between two biological conditions?

- molecular markers of a specific condition ?

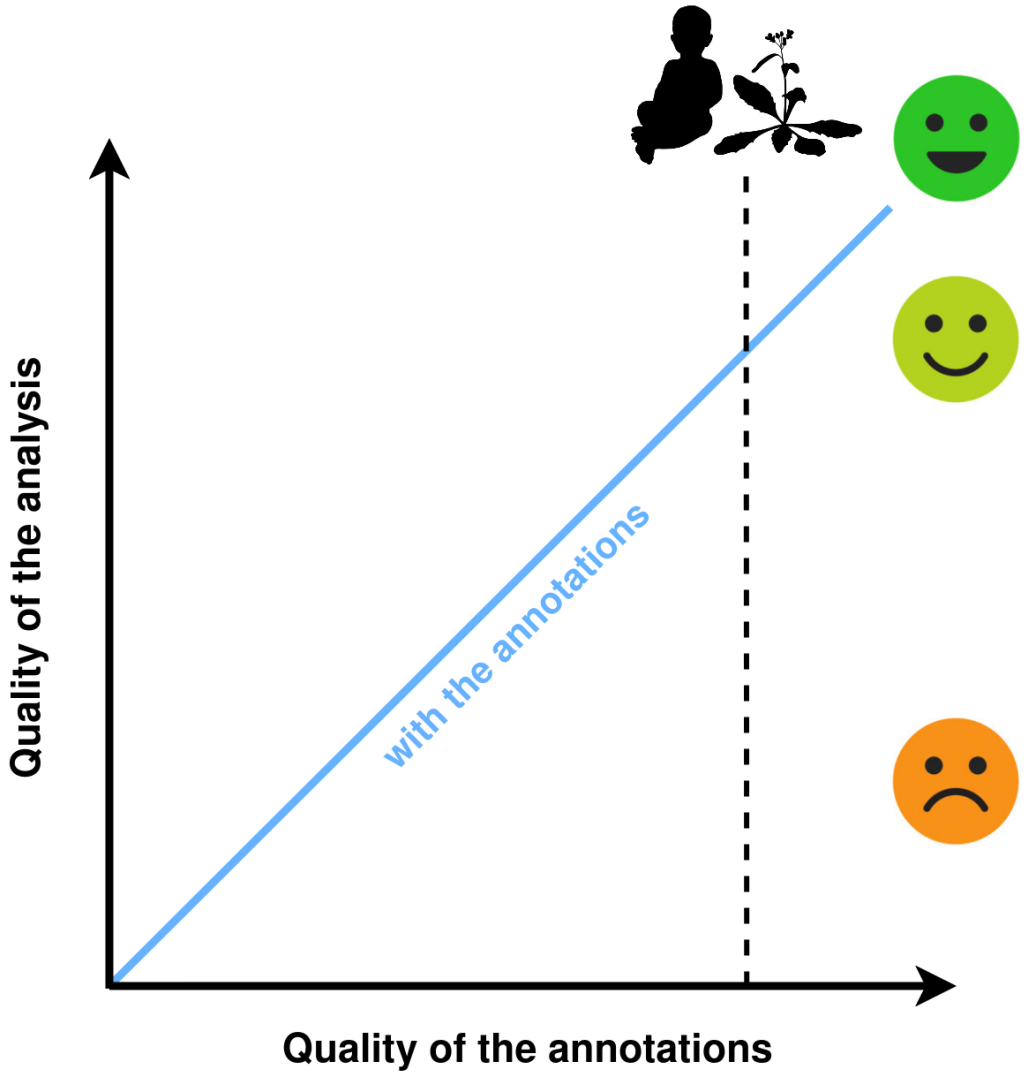
RNA-Seq experiment

- numerous analysis strategies (gene/events/isoforms)
- highly dependent on annotations quality (incomplete)

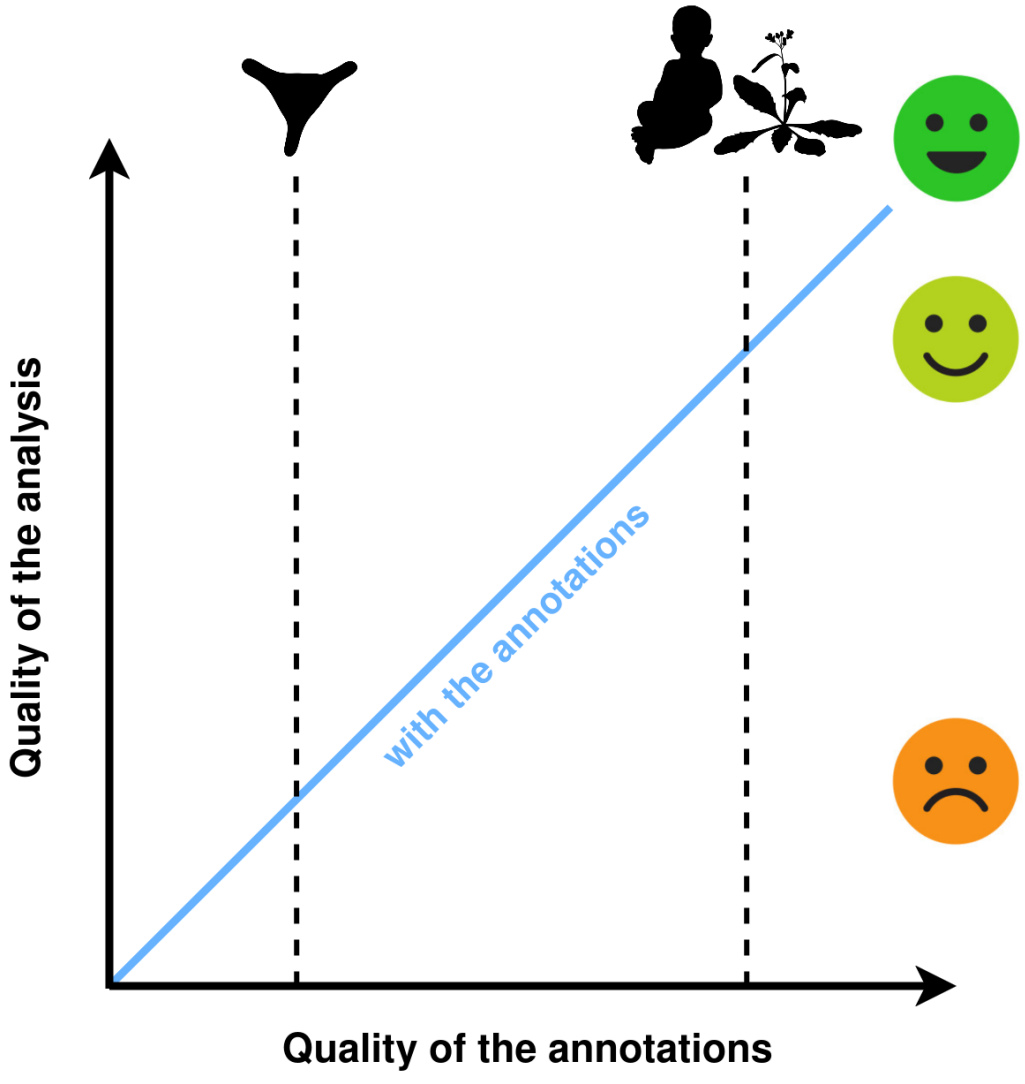
Better Annotations, Better Analyses



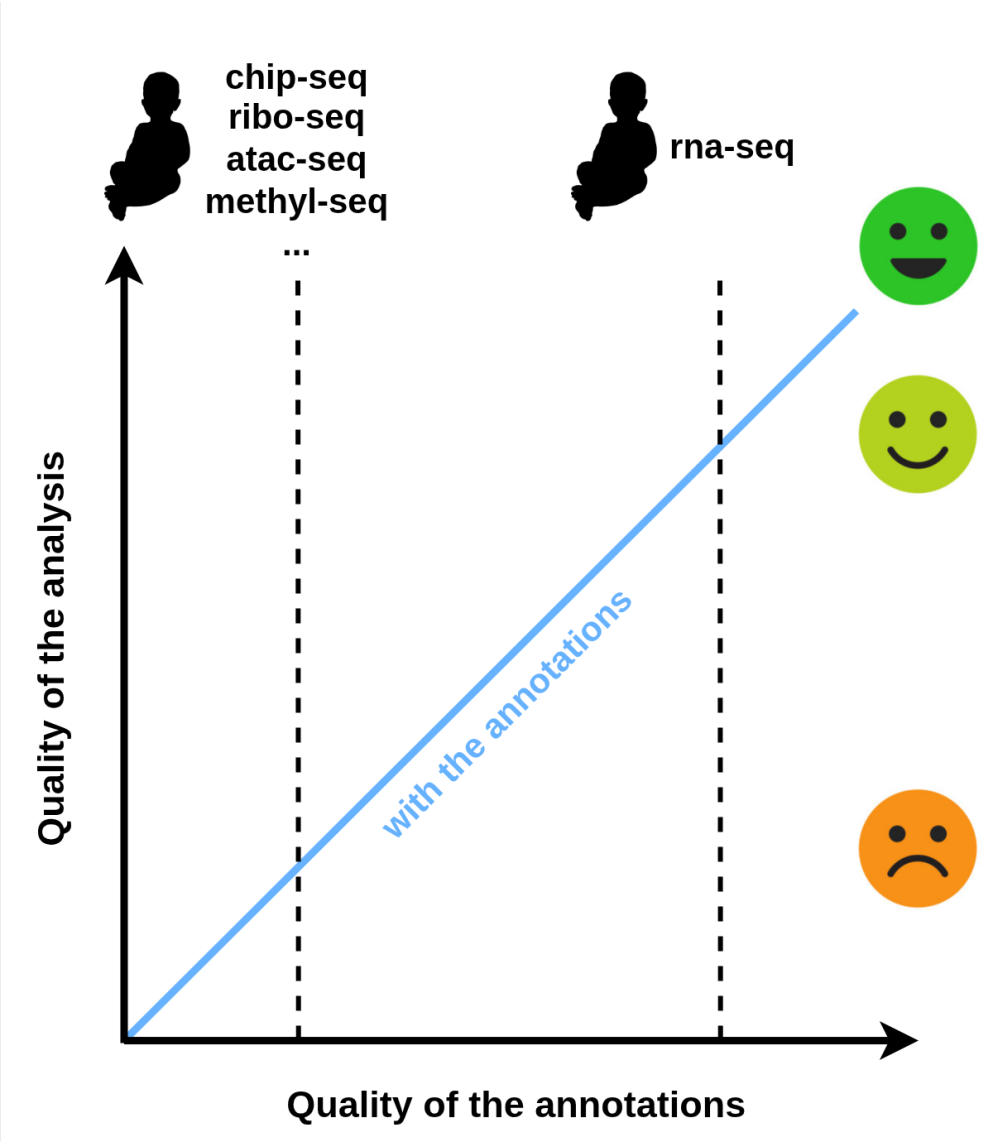
Better Annotations, Better Analyses



Better Annotations, Better Analyses

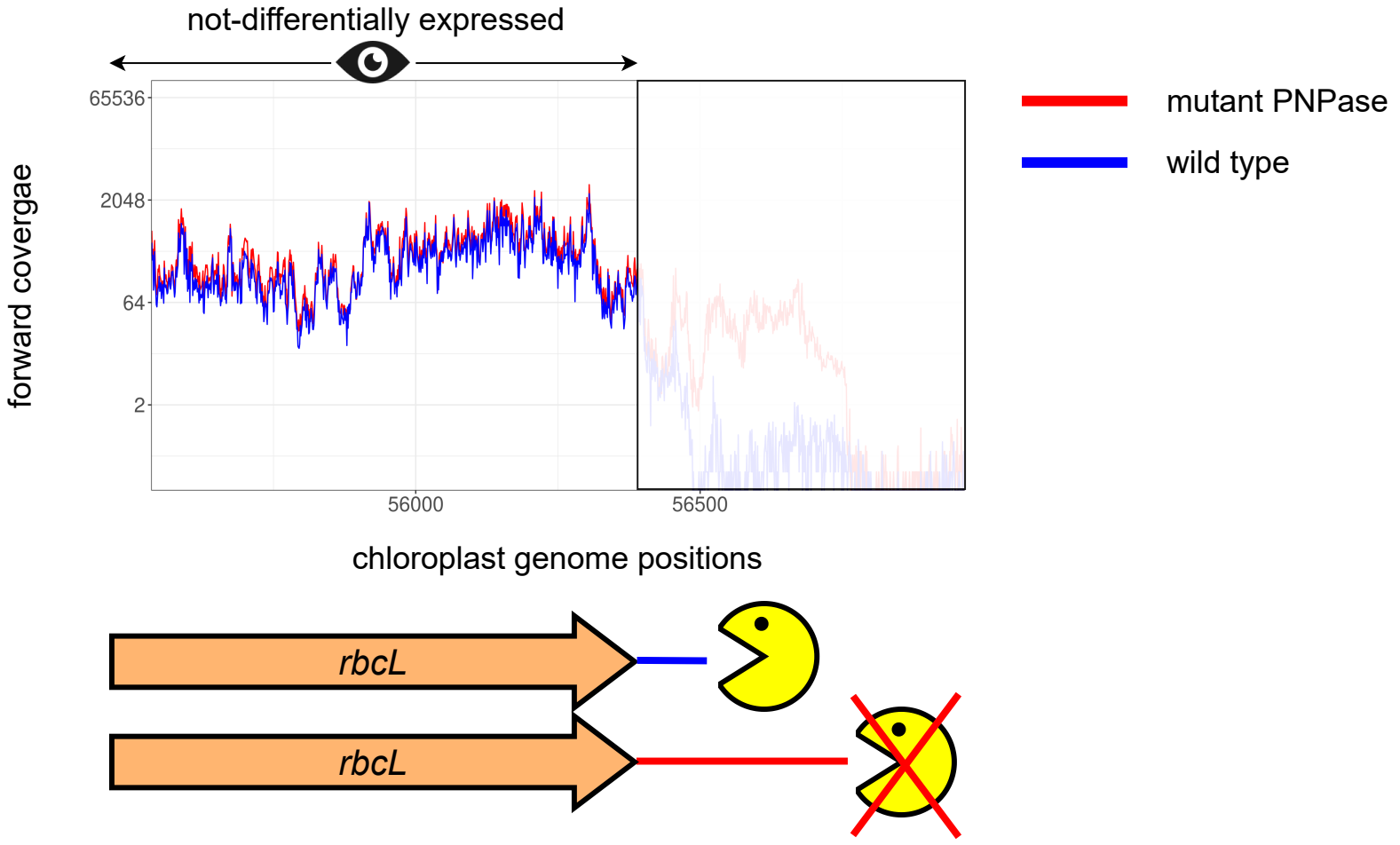


Better Annotations, Better Analyses



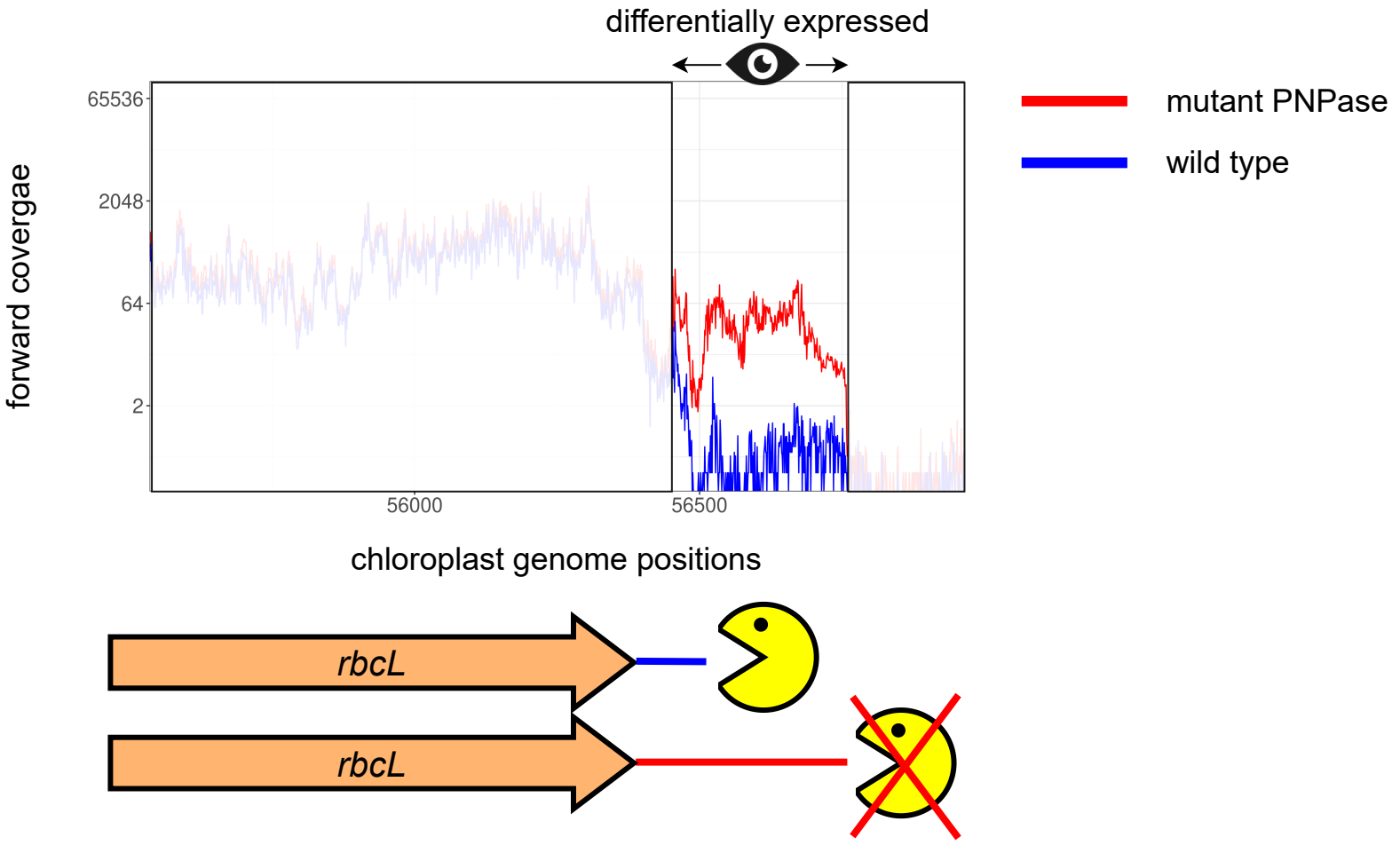
Differential expression analysis within annotations

The annotation of the *rbcL* gene does not capture the 3' extension observed in the *Arabidopsis thaliana* mutant deficient in PNPase activity

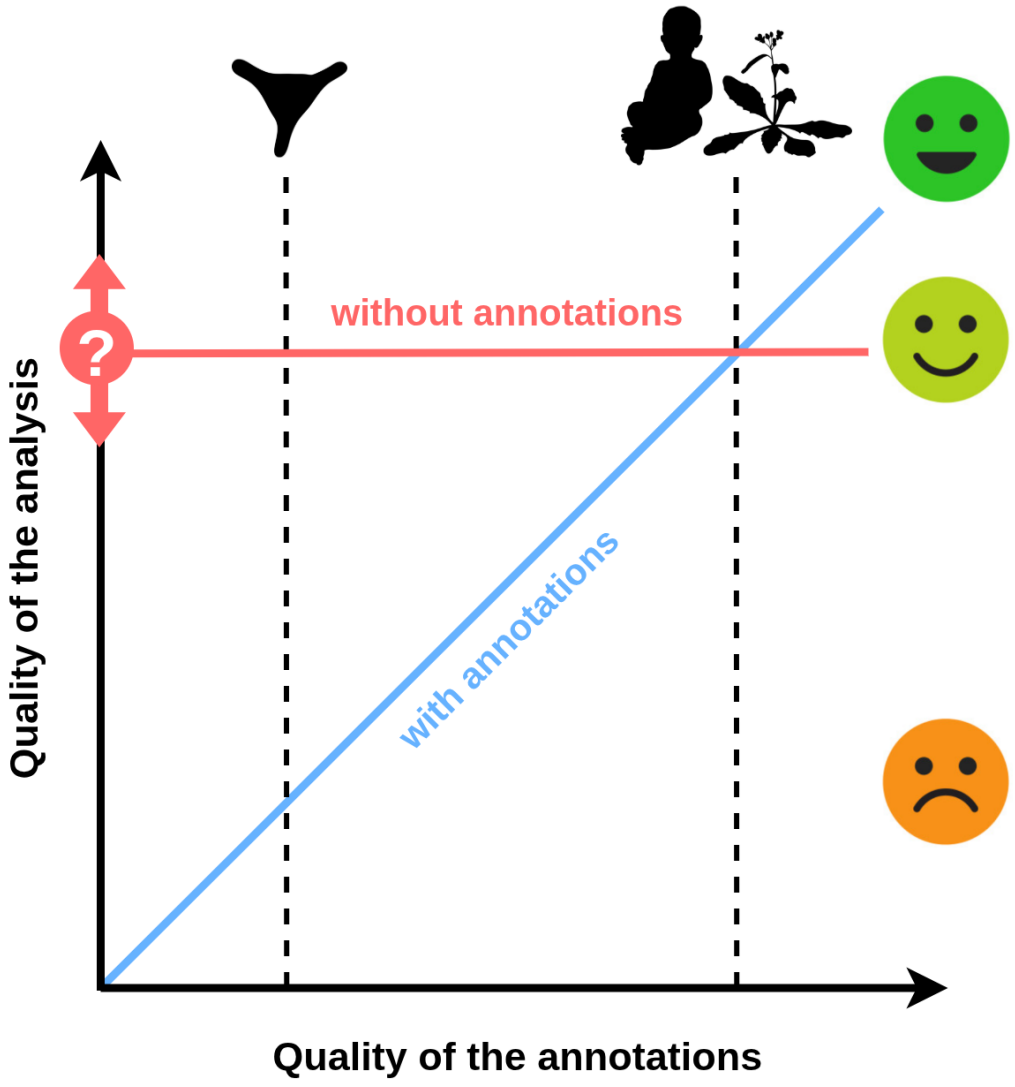


Differential expression analysis within annotations

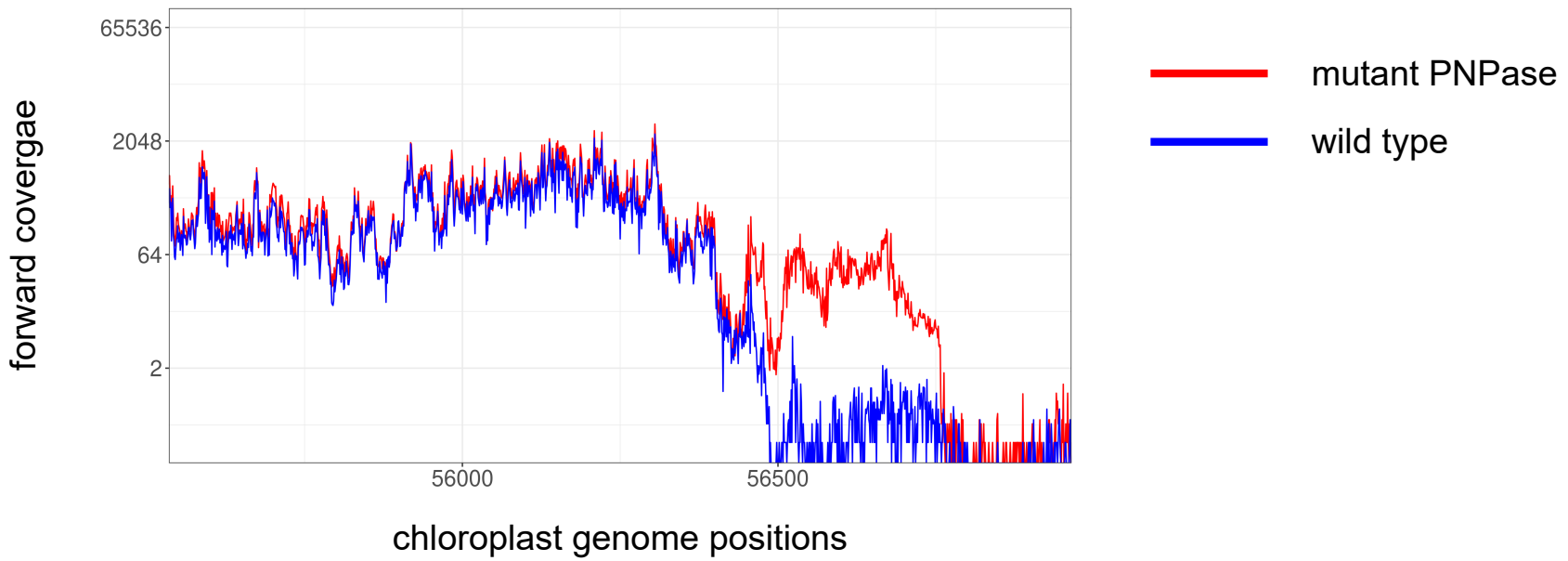
The annotation of the *rbcL* gene does not capture the 3' extension observed in the *Arabidopsis thaliana* mutant deficient in PNPase activity



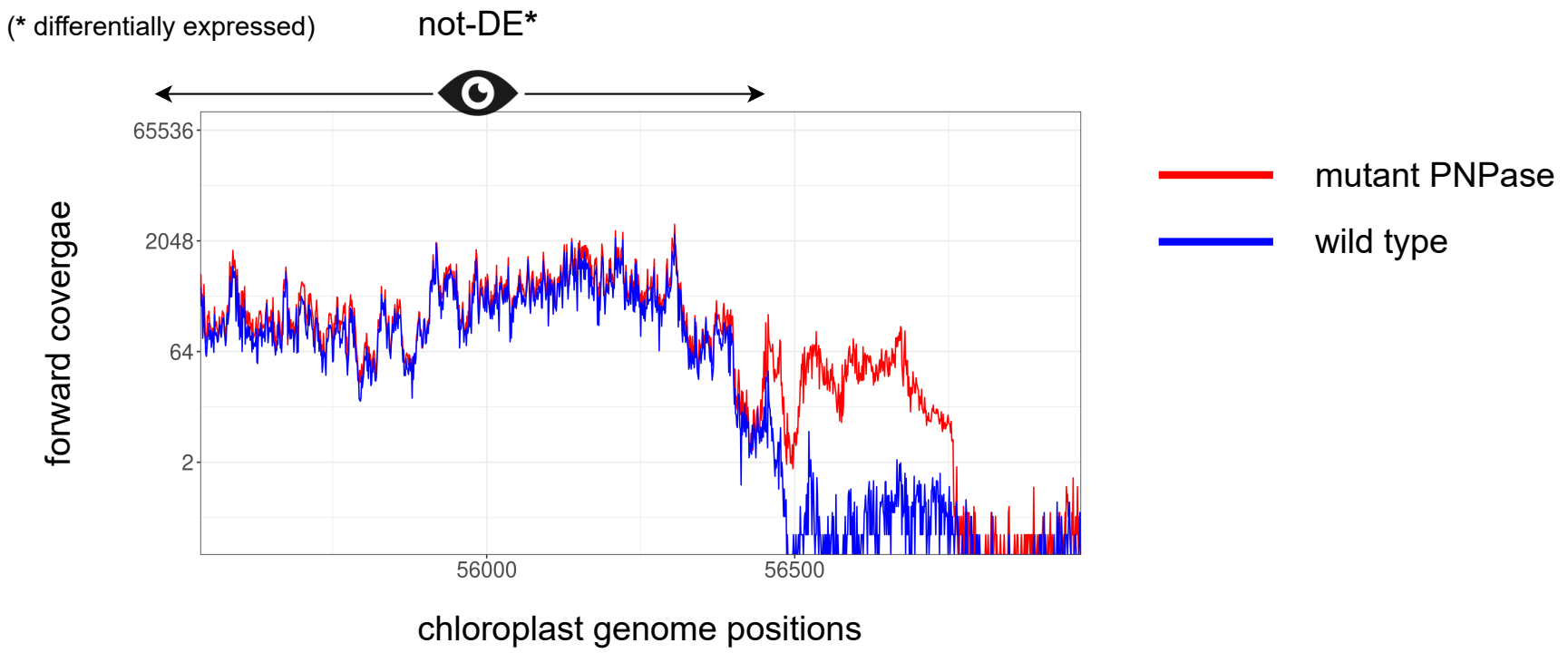
How well can we perform without annotations?



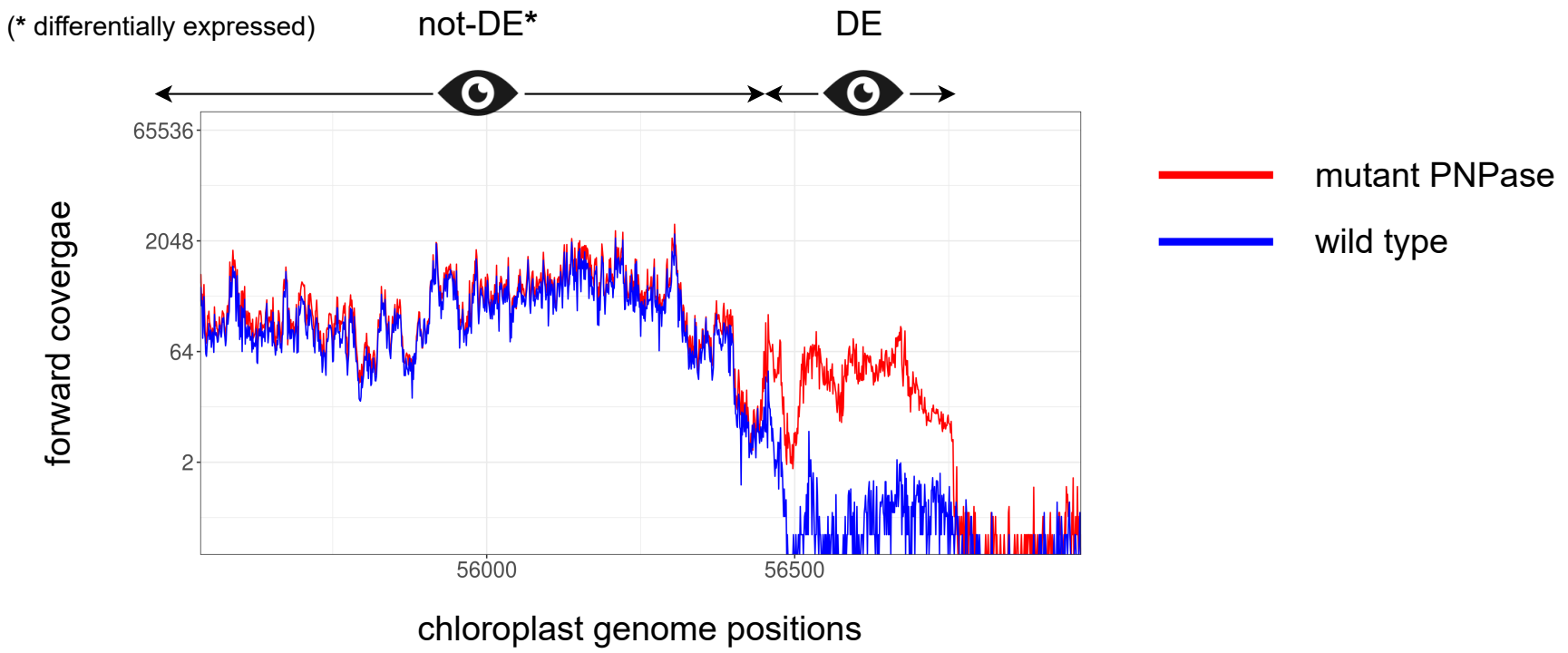
Summarize the signal into regions that are easier to interpret



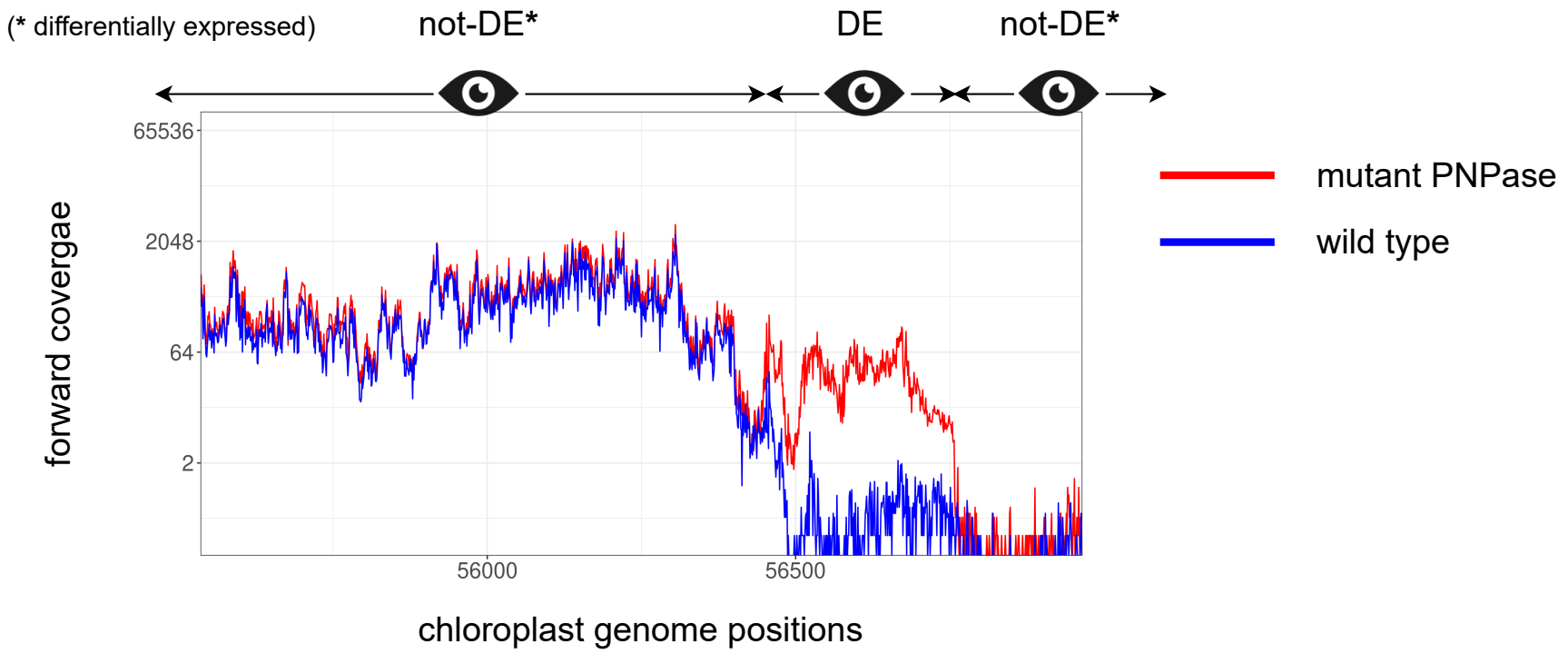
Summarize the signal into regions that are easier to interpret



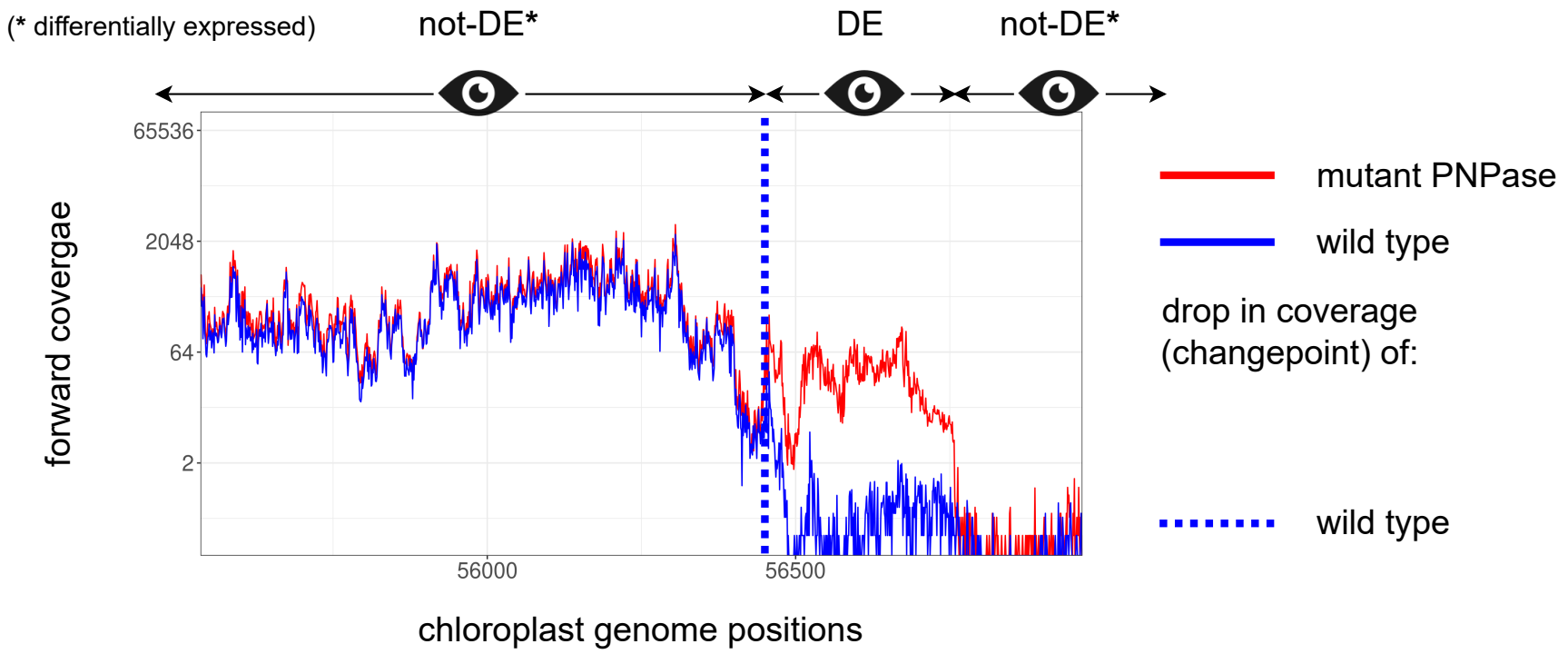
Summarize the signal into regions that are easier to interpret



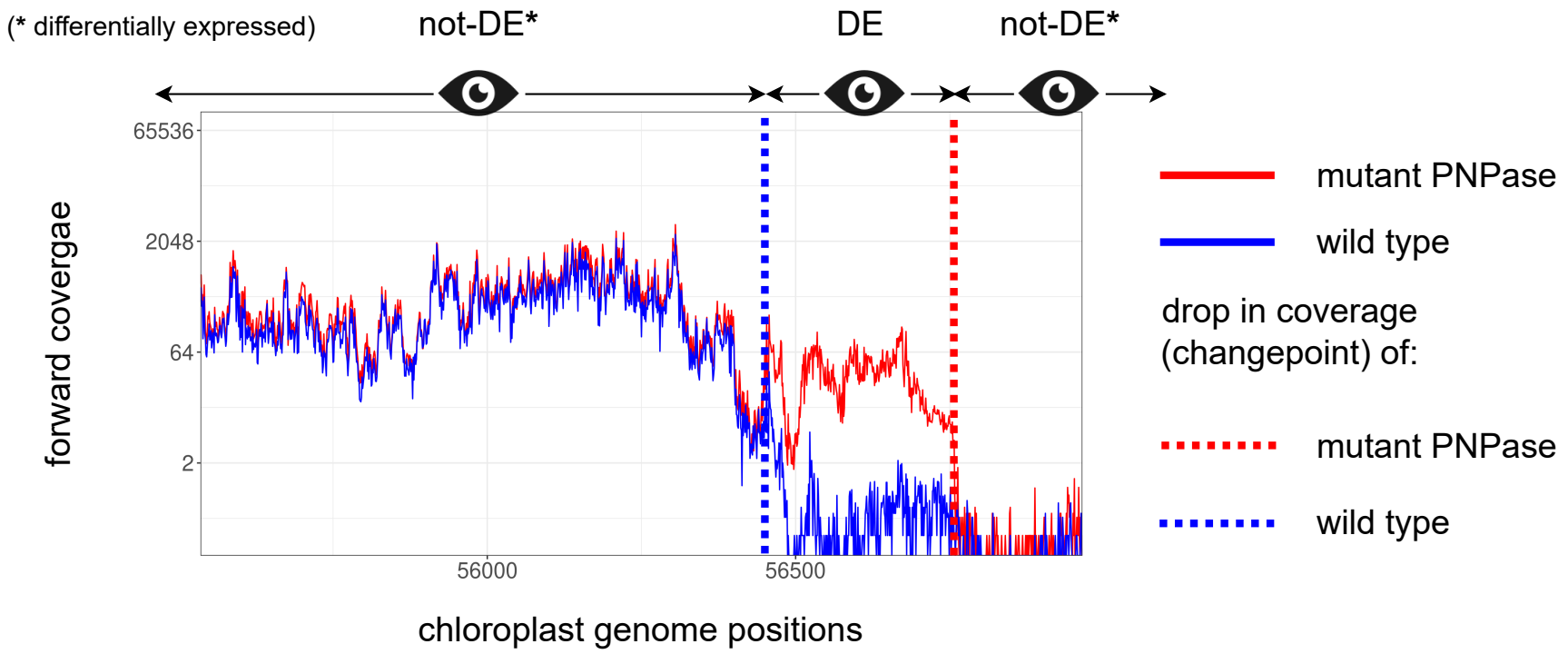
Summarize the signal into regions that are easier to interpret



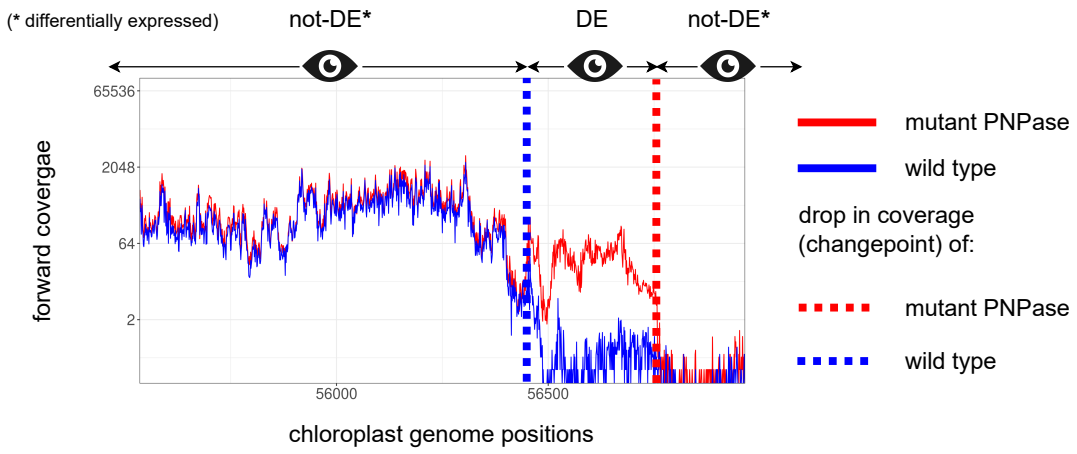
Summarize the signal into regions that are easier to interpret



Summarize the signal into regions that are easier to interpret



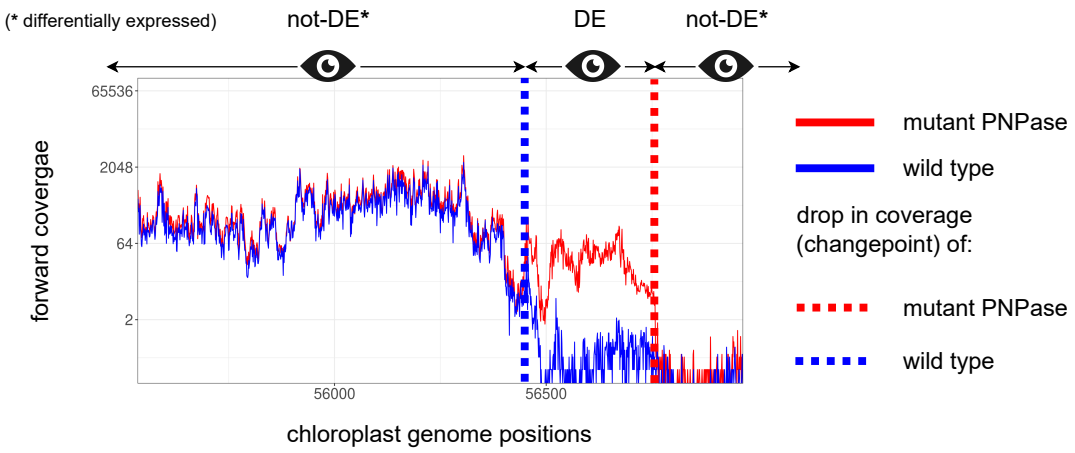
Summarize the signal into regions that are easier to interpret



The multiple changepoint detection problem [Fearnhead and Rigaill 2019; Truong et al., 2020]

1. estimate the number of changepoints
2. estimate the positions of changepoints

Summarize the signal into regions that are easier to interpret



The multiple changepoint detection problem [Fearnhead and Rigaill 2019; Truong et al., 2020]

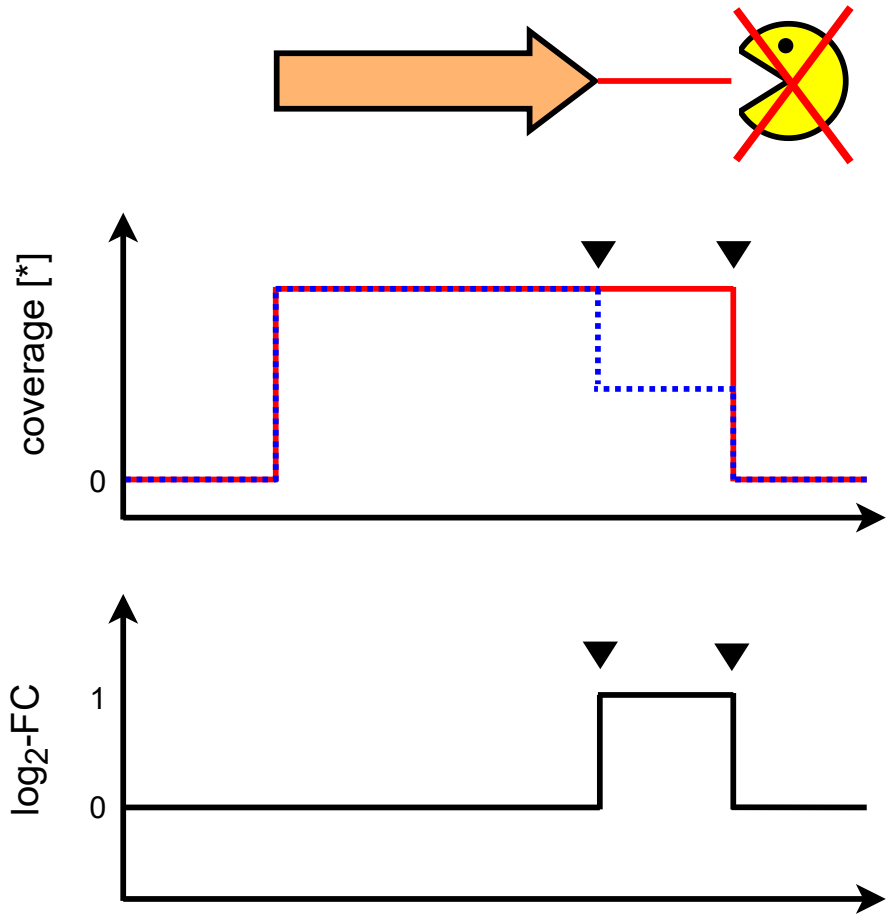
1. estimate the number of differentially expressed regions
2. estimate the positions of differentially expressed regions

What signal should we segment ?

Per-base $\log_2$-FC *"difference of coverages on the log scale"*

1) scales with transcription differences between two conditions

[*Collado-Torres et al., 2017; Mirauta et al., 2014]



changepoint :

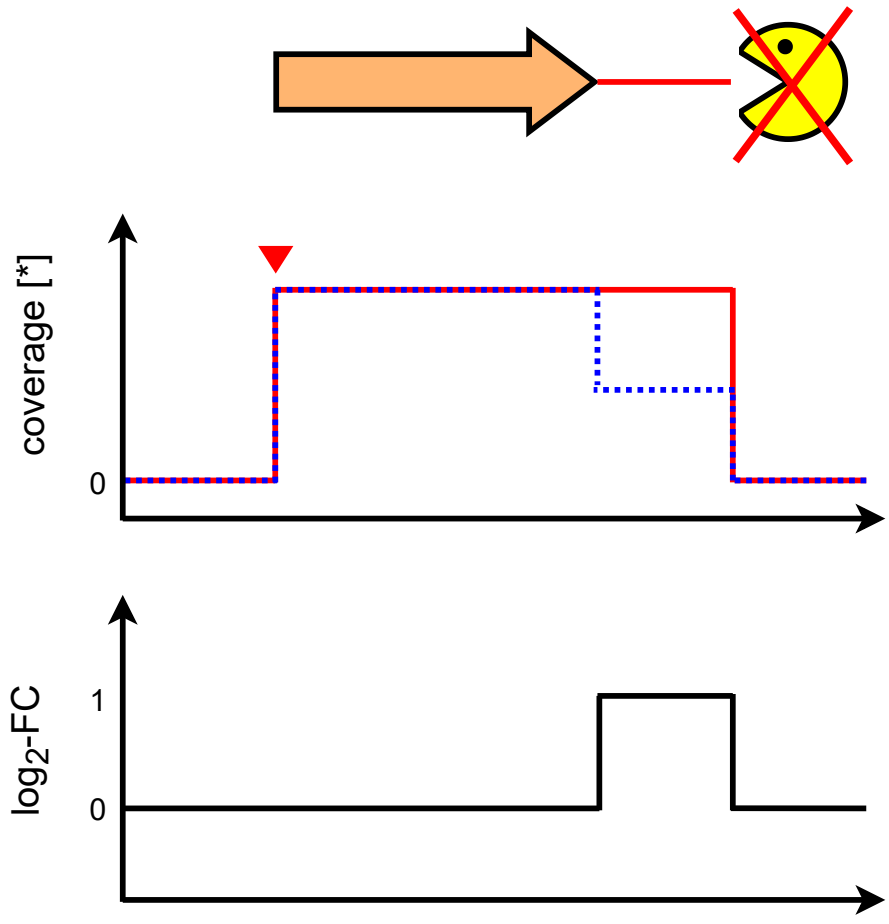
▼ matches transcription differences

What signal should we segment ?

Per-base $\log_2$-FC *"difference of coverages on the log scale"*

2) robust to biological/technical variations in the coverage

[*Collado-Torres et al., 2017; Mirauta et al., 2014]



changepoint :

▼ does not match transcription differences

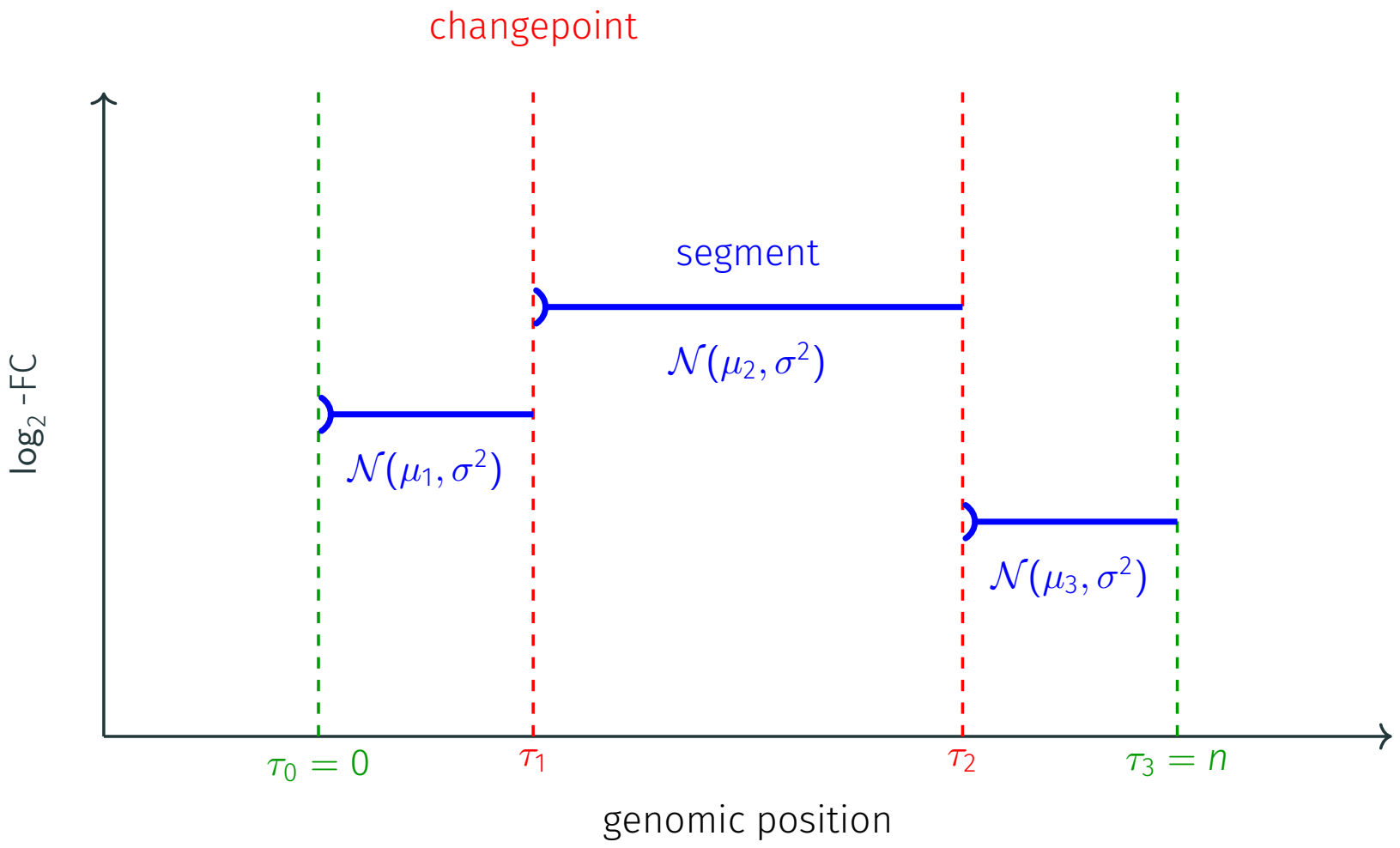
Which methodology should we use?

Changepoint in the mean of a Gaussian signal with FPOP

model $\log_2 \text{-FC}_i \sim \mathcal{N}(\mu_{(\tau_{d-1},\tau_d]}, \sigma^2)$ *i.i.d* (1)

Which methodology should we use?

The homoscedastic piecewise constant Gaussian model (graphically)



Which methodology should we use?

Changepoint in the mean of a Gaussian signal with FPOP

model $\log_2\text{-FC}_i \sim \mathcal{N}(\mu_{(\tau_{d-1},\tau_d]}, \sigma^2)$ *i.i.d* (1)

Which methodology should we use?

Changepoint in the mean of a Gaussian signal with FPOP

model $\log_2\text{-FC}_i \sim \mathcal{N}(\mu_{(\tau_{d-1},\tau_d]}, \sigma^2)$ *i.i.d* (1)

estimate τ minimize a penalized least-squares criterion

[Yao and Au 1989]

Which methodology should we use?

Changepoint in the mean of a Gaussian signal with FPOP

model $\log_2\text{-FC}_i \sim \mathcal{N}(\mu_{(\tau_{d-1},\tau_d]}, \sigma^2)$ *i.i.d* (1)

estimate τ minimize a penalized least-squares criterion

[Yao and Au 1989]

algorithms dynamic programming

[Jackson et al., 2003; Maidstone et al., 2017 (FPOP, log-linear)]

- FPOP : $\sim$ 1 sec on bacterial genome (8Mb)

Which methodology should we use?

Changepoint in the mean of a Gaussian signal with FPOP

$$\text{model} \quad \log_2\text{-FC}_i \sim \mathcal{N}(\mu_{(\tau_{d-1},\tau_d]}, \sigma^2) \;\; i.i.d \qquad (1)$$

More complex models were considered, but (1) is sufficient !

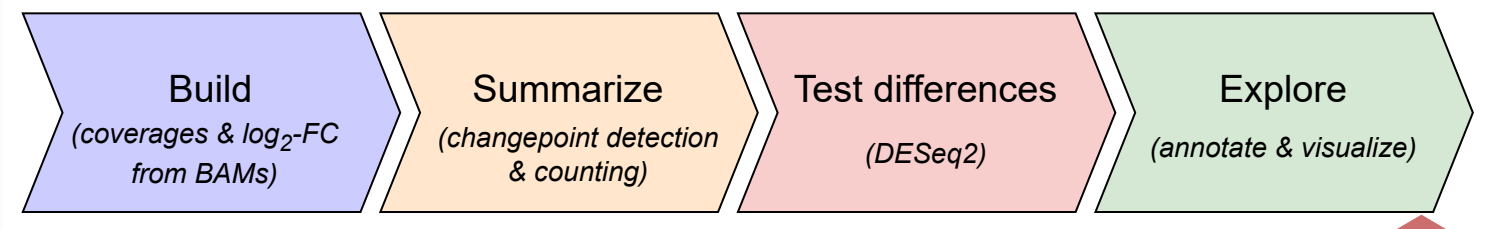
- finds all experimentally validated RNA regulations [1]

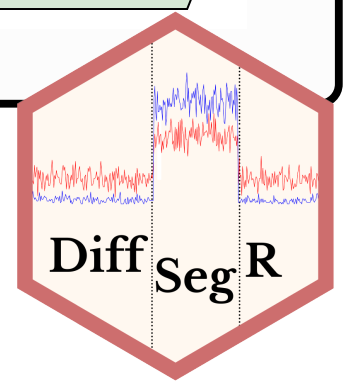
  (threshold heuristics : 13/16 at best [Tran et al., 2016; Collado-Torres et al., 2017; Zytnicki et al., 2021])
- controls the FP on blank experiments

[1] 5' & 3' extension; antisenses, introns and protected RNAs accumulation, (up/down)-regulated exons [Hotto et al., 2011, Germain et al., 2011, Ruwe et al., 2016]

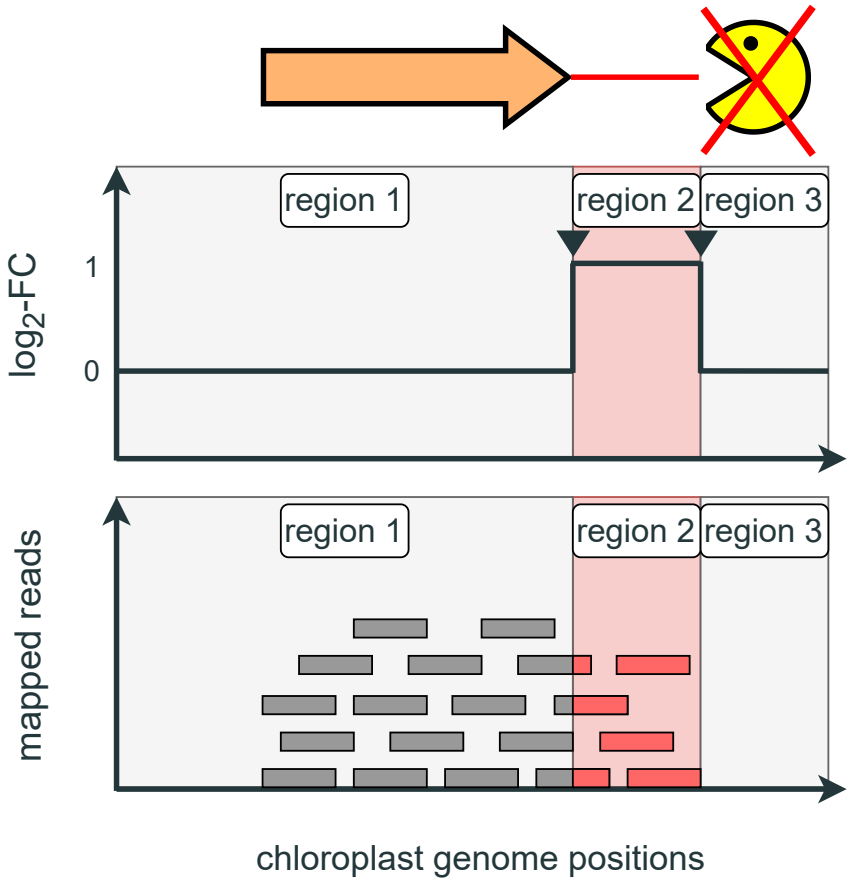
DiffSegR [Liehrmann et al., 2023]

DiffSegR segments the per-base $\log_2$-FC with FPOP to automates the discovery of differentially expressed regions along the genome.





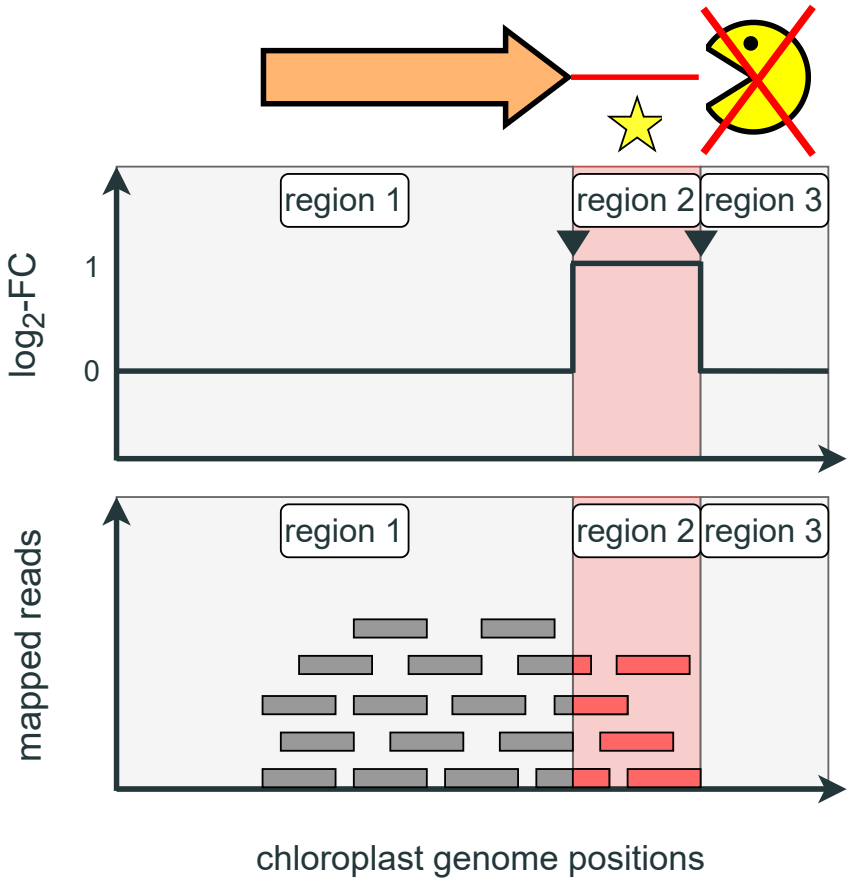
Step 2/4: Counting overlapping reads



region \ sample	mutant 1	mutant 2	wild type 1	wild type 2
region 1	100	104	106	98
region 2	100	104	53	49
region 3	0	0	0	0

counting

Step 3/4: Differential expression analysis



sample / region	mutant 1	mutant 2	wild type 1	wild type 2	adjusted p-value
region 1	100	104	106	98	1
region 2	100	104	53	49	1e-4
region 3	0	0	0	0	1

counting

- $\widehat{\log_2 -\text{FC}} =_? 0$ with **DESeq2** (*NB* generalized linear model) [Love et al., 2014]

Step 4/4: Detection of fine regulations with IGV

mutant PNPase

wild type

up-regulated (DESeq2)

not-differentially expressed (DESeq2)

Step 4/4: Detection of fine regulations with IGV

1 up-regulated exon

2 3' extension

3 5' extension

4 intron accumulation

[Walter et al., 2002; Germain et al., 2011; Hotto et al., 2011; Castandet et al., 2013]

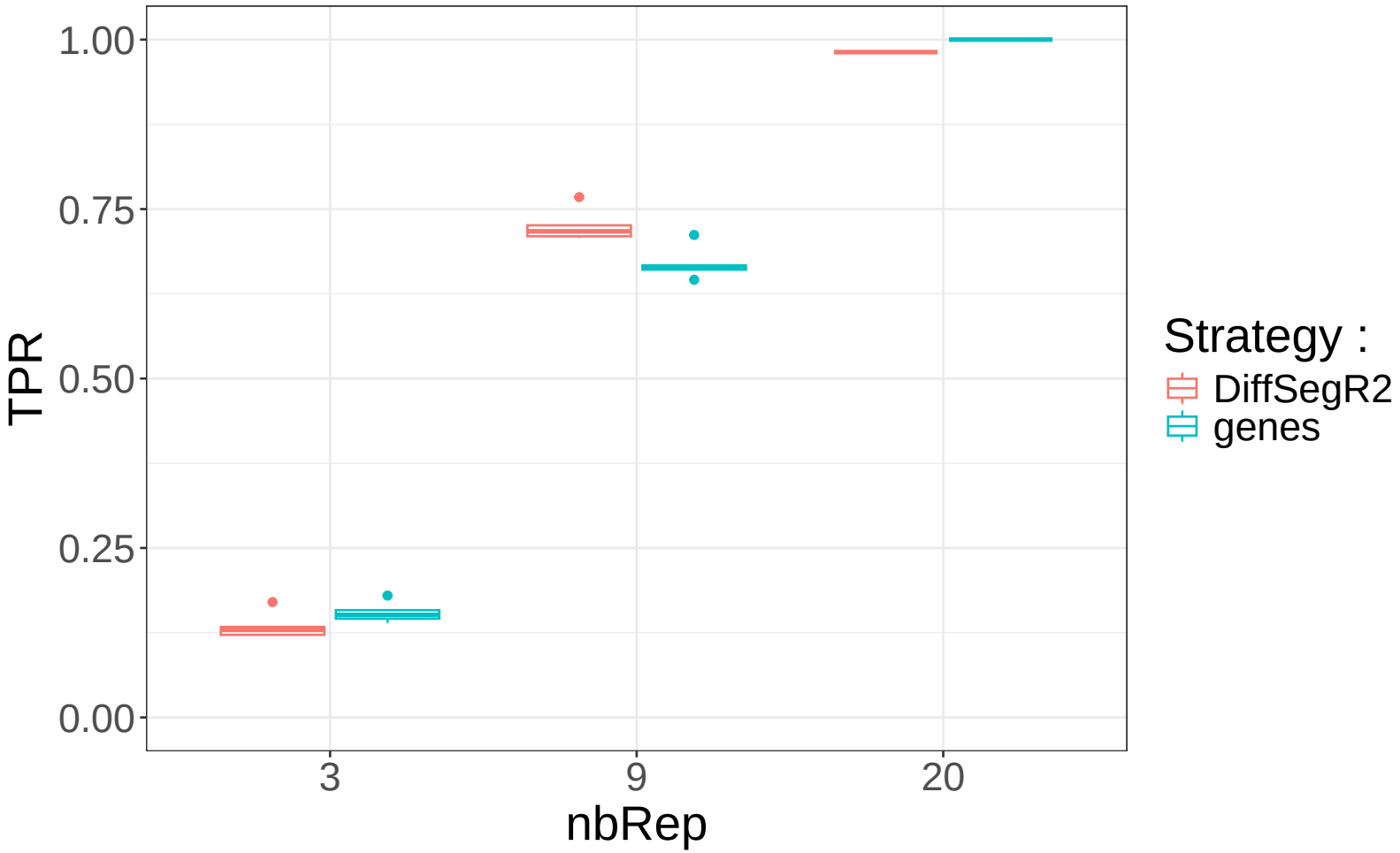
Evaluating TPR on *A. thaliana* nuclear genome [preliminary]

- Peersim dataset :
  - ambient T° vs elevated T°
  - 20 biological replicates per condition

Evaluating TPR on *A. thaliana* nuclear genome [preliminary]

- Peersim dataset :
  - ambient T° vs elevated T°
  - 20 biological replicates per condition
- Power experiment :
  - differential gene expression analysis, 20 vs 20 [gold-standard]
  - Resampling : 3 vs 3 - 9 vs 9
    - Ex : {rep 1, rep 9, rep 17} vs {rep 1, rep 9, rep 17}
  - chromosome 1 (30M bp)

DiffSegR2 performs on par with gene annotations [preliminary]

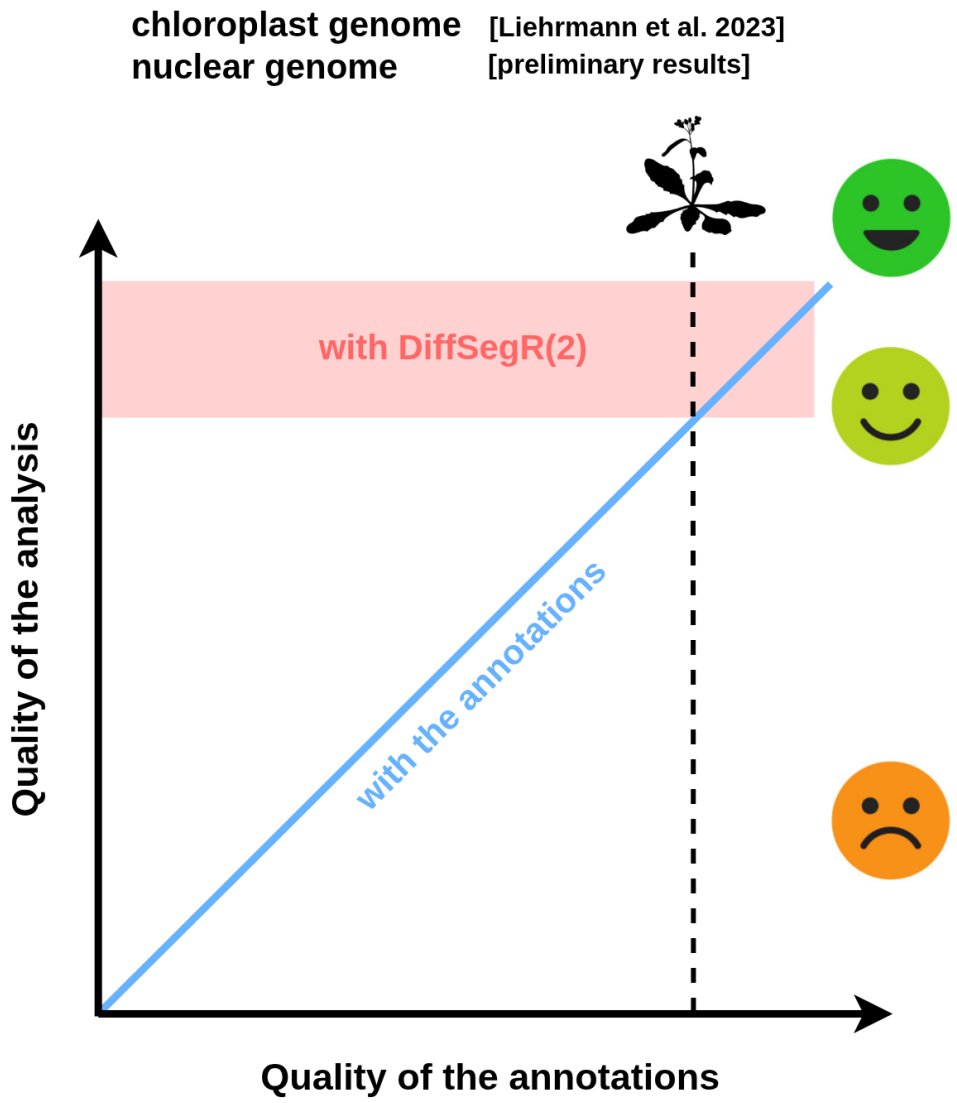


Up to 16% of segments fall outside the gold-standard

[preliminary]

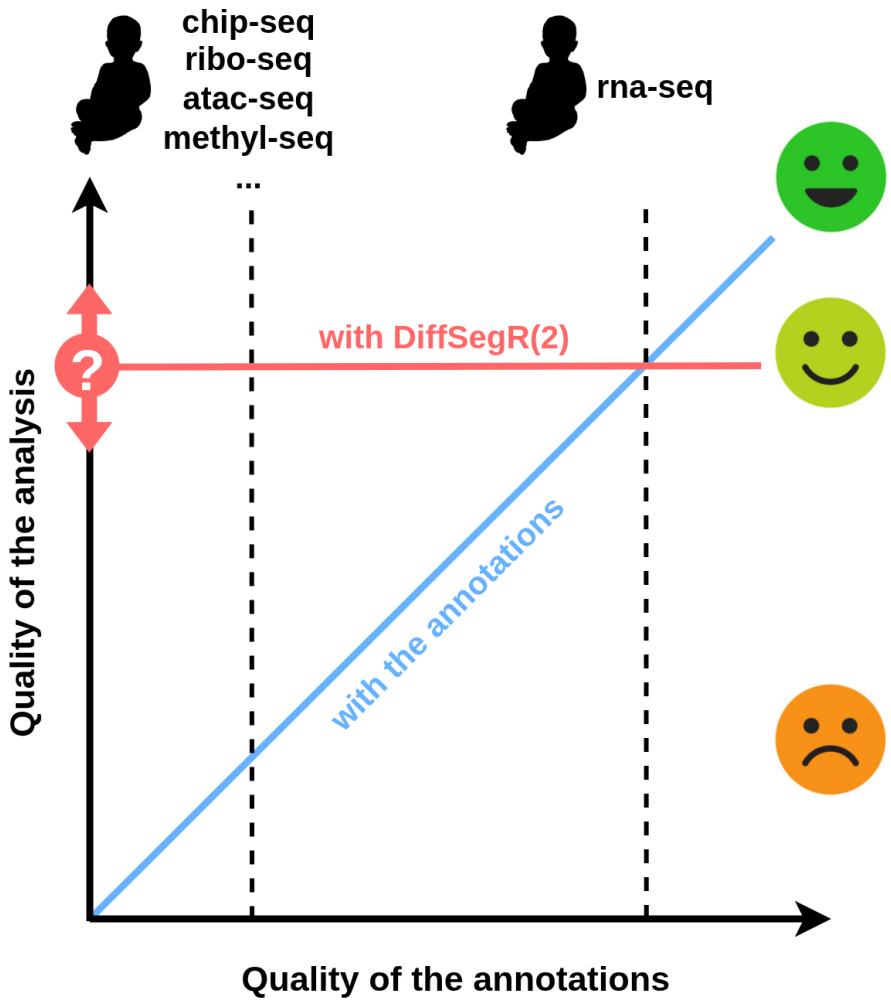
- 9% fall within genes not found in the gold-standard
- 5% fall within antisense of genes
- 2% fall within intergenic regions ($\sim$ 200 segments)

Take home message



Perspectives

- DiffSegR can analyze other experiments (RIP-Seq, [Tran et al., 2023])



Acknowledgements

Collaborators

- Guillem Rigaill
- Georgy Zaouk, M1 student
- Véronique Brunaud

Organizers

- Véronique Brunaud
- Sophie Lemoine
- Erwan Corre