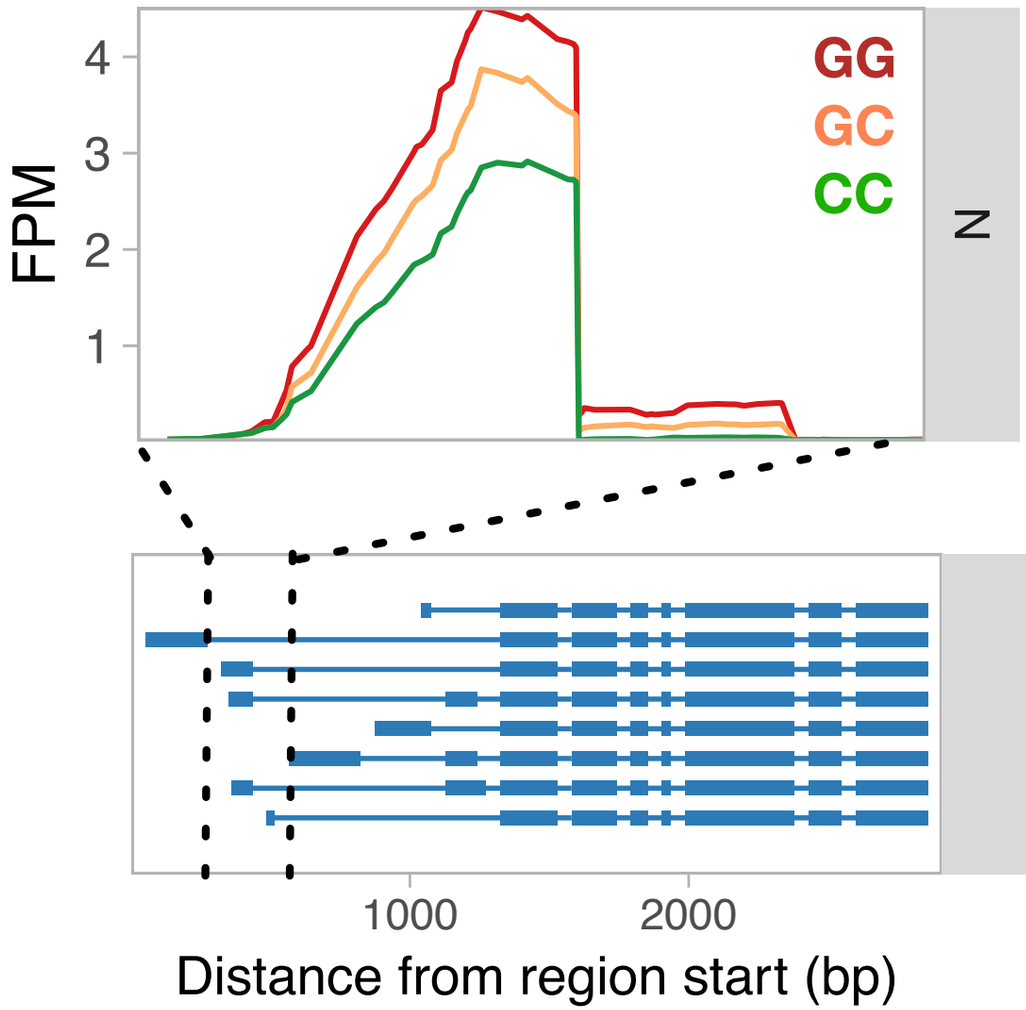
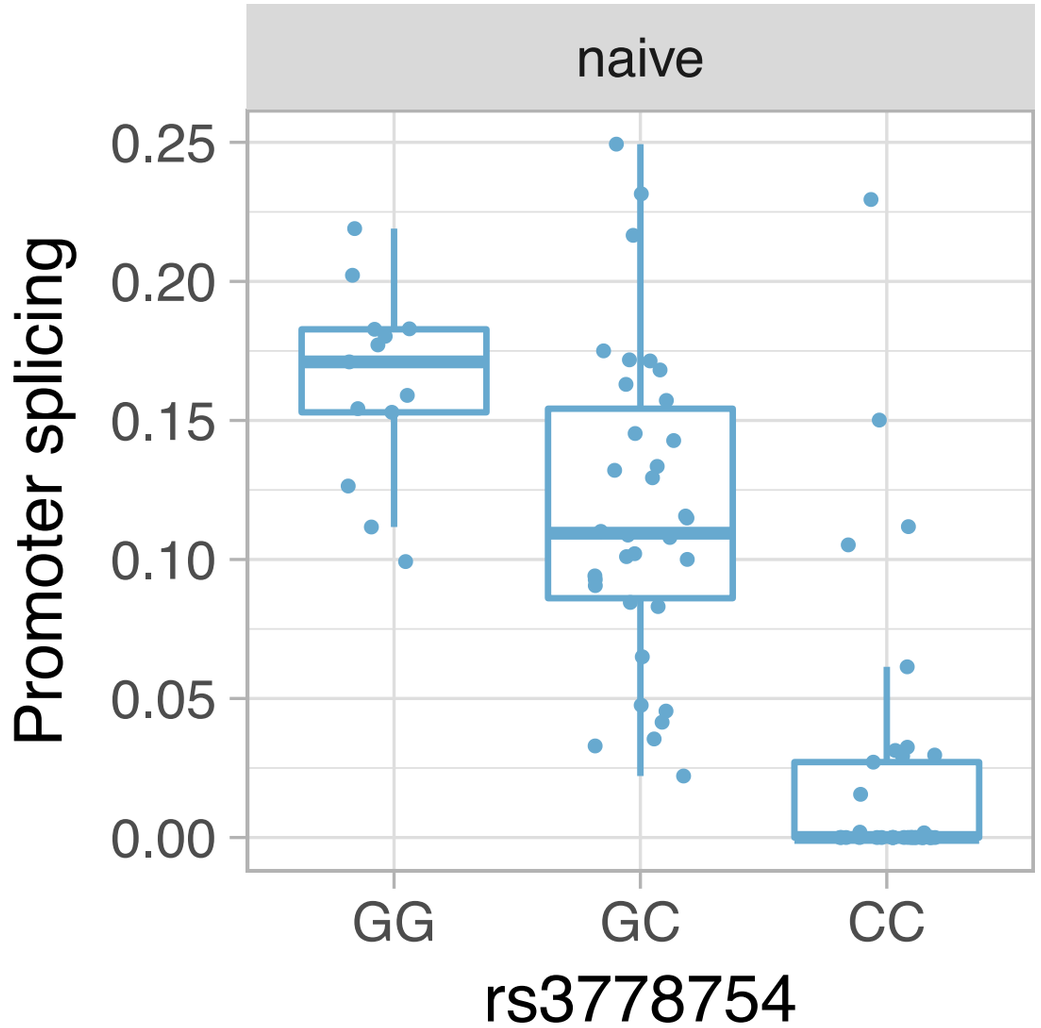
Annotation preprocessing improves the detection of genetic effects on transcript usage

Kaur Alasoo
University of Tartu
6 December 2017

@kauralasoo
http://kauralasoo.net

Alternative 5’ splice site in the first exon of IRF5

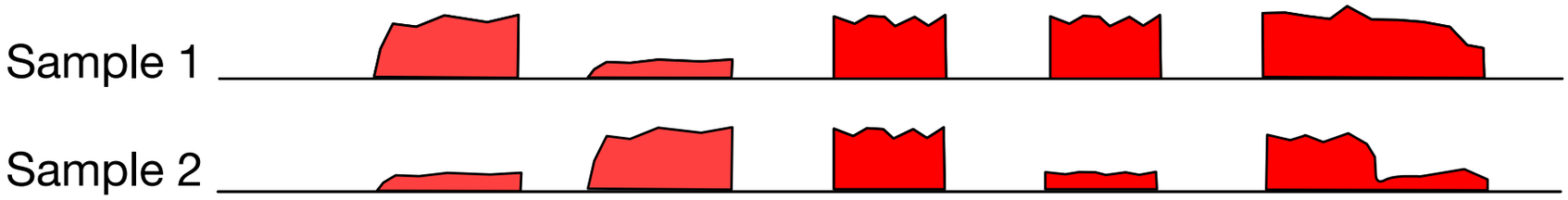




Transcript usage quantitative trait locus (QTL)

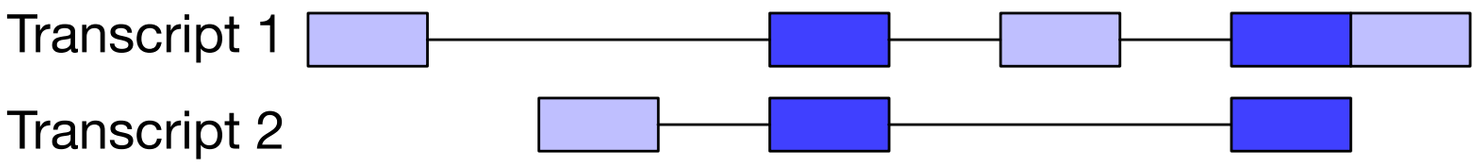
Ideal world

Scenario A: annotations match expressed transcripts



RNA-seq read coverage

Expressed transcripts

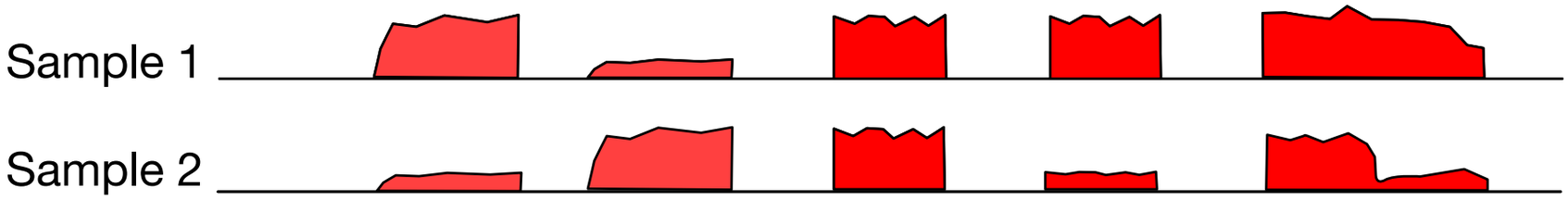


	Sample 1	Sample 2
Transcript 1	80%	20%
Transcript 2	20%	80%

True transcript usage

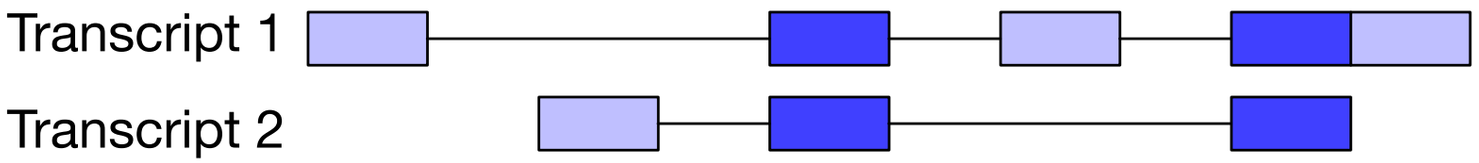
Ideal world

Scenario A: annotations match expressed transcripts



RNA-seq read coverage

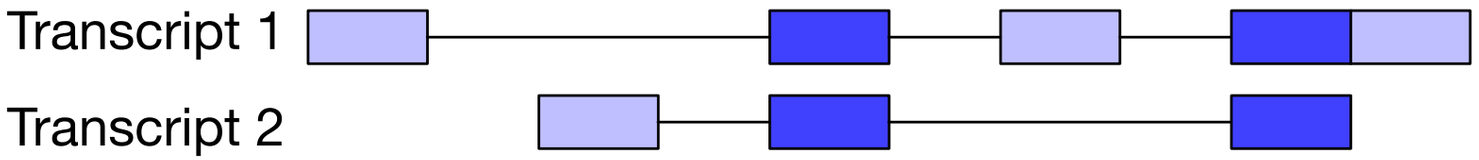
Expressed transcripts



	Sample 1	Sample 2
Transcript 1	80%	20%
Transcript 2	20%	80%

True transcript usage

Annotated transcripts

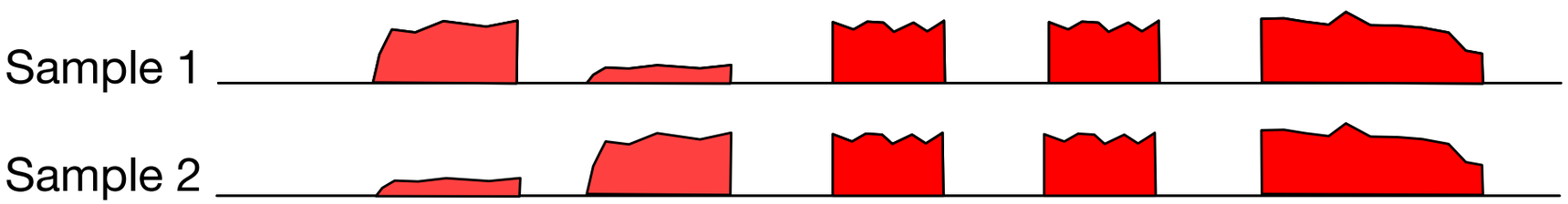


Transcript 1	80%	20%
Transcript 2	20%	80%

Estimated transcript usage

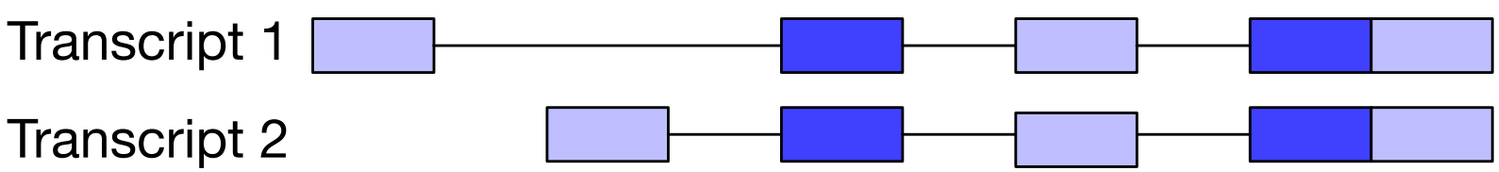
Real world

Scenario B: annotations differ from expressed transcripts



RNA-seq read coverage

Expressed transcripts

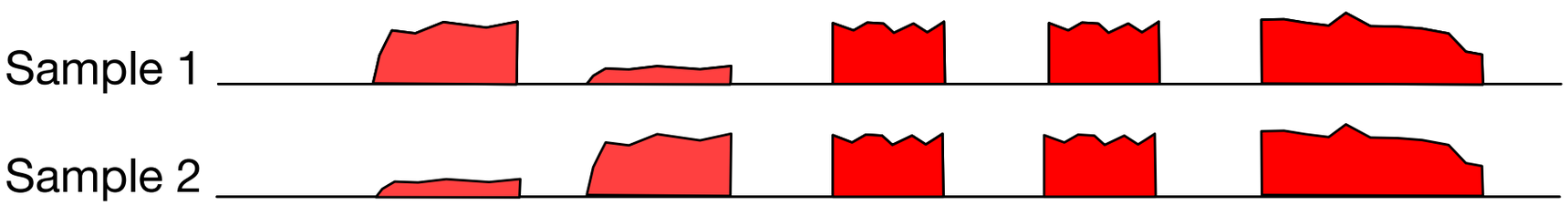


	Sample 1	Sample 2
Transcript 1	80%	20%
Transcript 2	20%	80%

True transcript usage

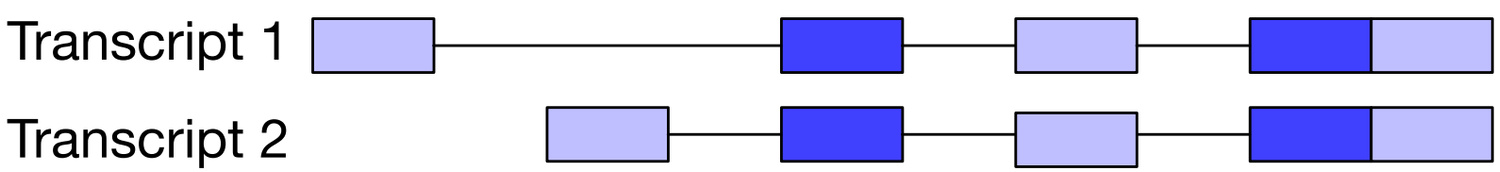
Real world

Scenario B: annotations differ from expressed transcripts



RNA-seq read coverage

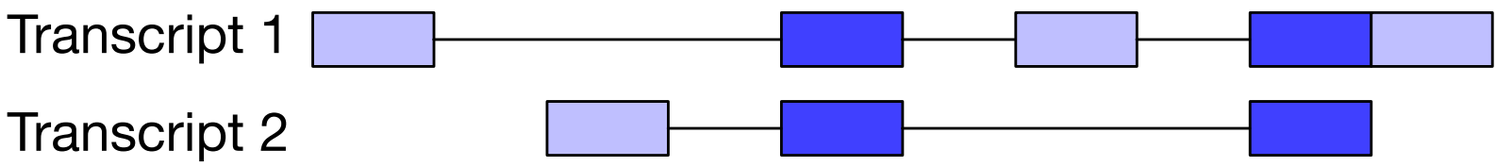
Expressed transcripts



	Sample 1	Sample 2
Transcript 1	80%	20%
Transcript 2	20%	80%

True transcript usage

Annotated transcripts

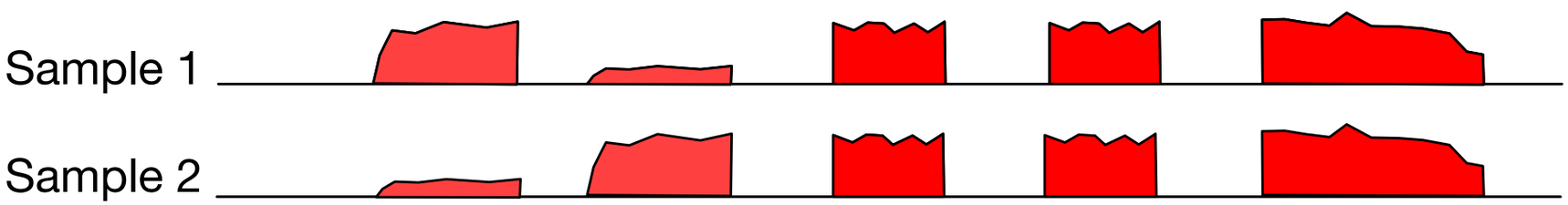


	Sample 1	Sample 2
Transcript 1	93%	73%
Transcript 2	7%	27%

Estimated transcript usage

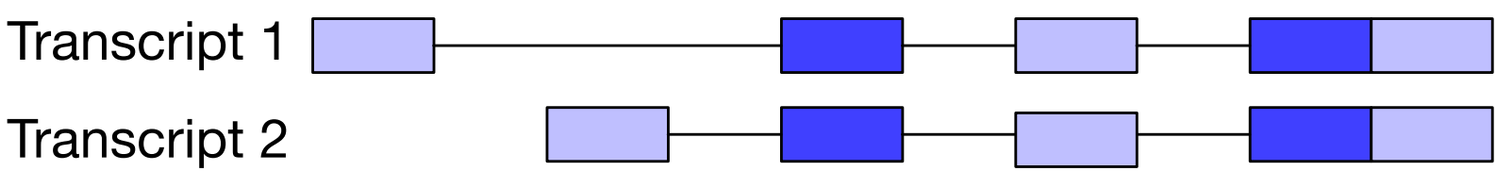
Real world

Scenario B: annotations differ from expressed transcripts



RNA-seq read coverage

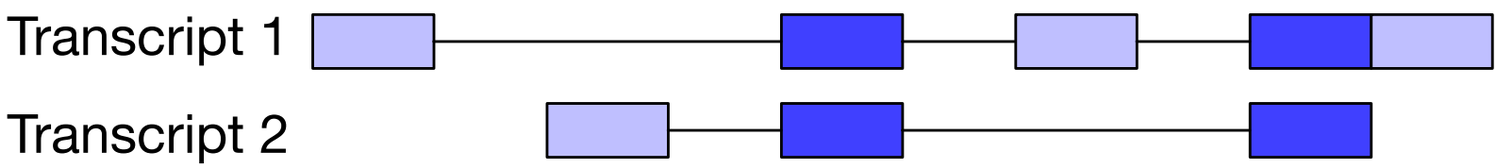
Expressed transcripts



	Sample 1	Sample 2
Transcript 1	80%	20%
Transcript 2	20%	80%

True transcript usage

Annotated transcripts

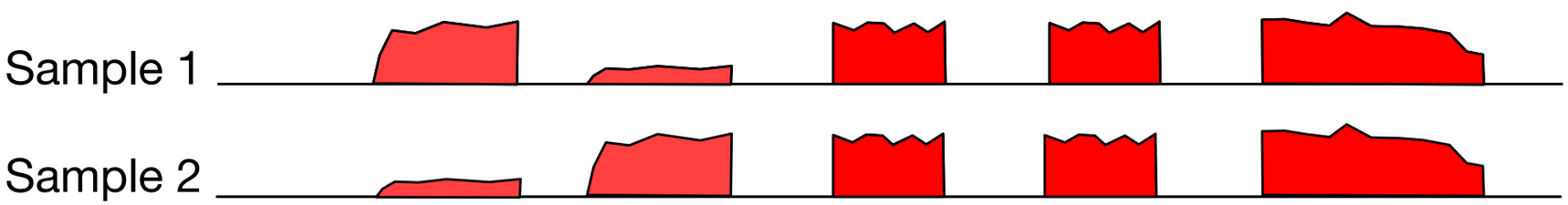


93%	73%	**Estimated transcript usage**
7%	27%	

Biased estimates!

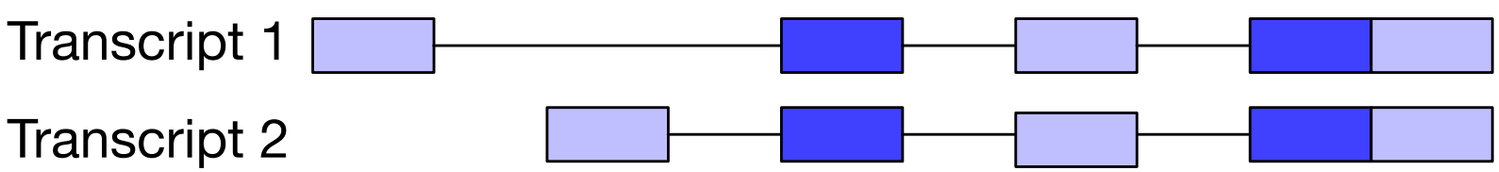
Real world

Scenario B: annotations differ from expressed transcripts



RNA-seq read coverage

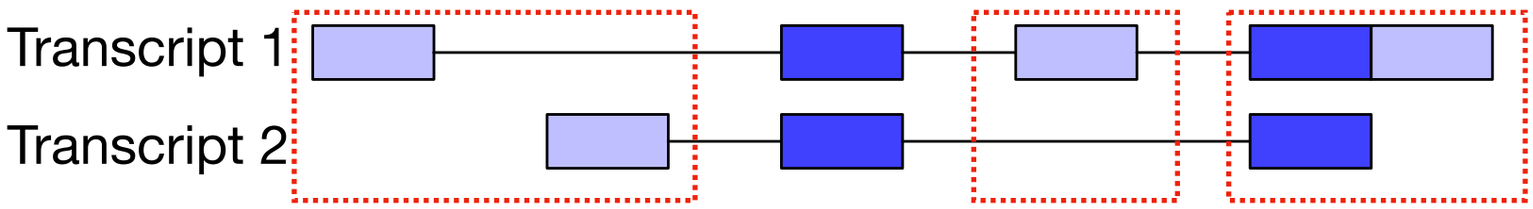
Expressed transcripts



	Sample 1	Sample 2
Transcript 1	80%	20%
Transcript 2	20%	80%

True transcript usage

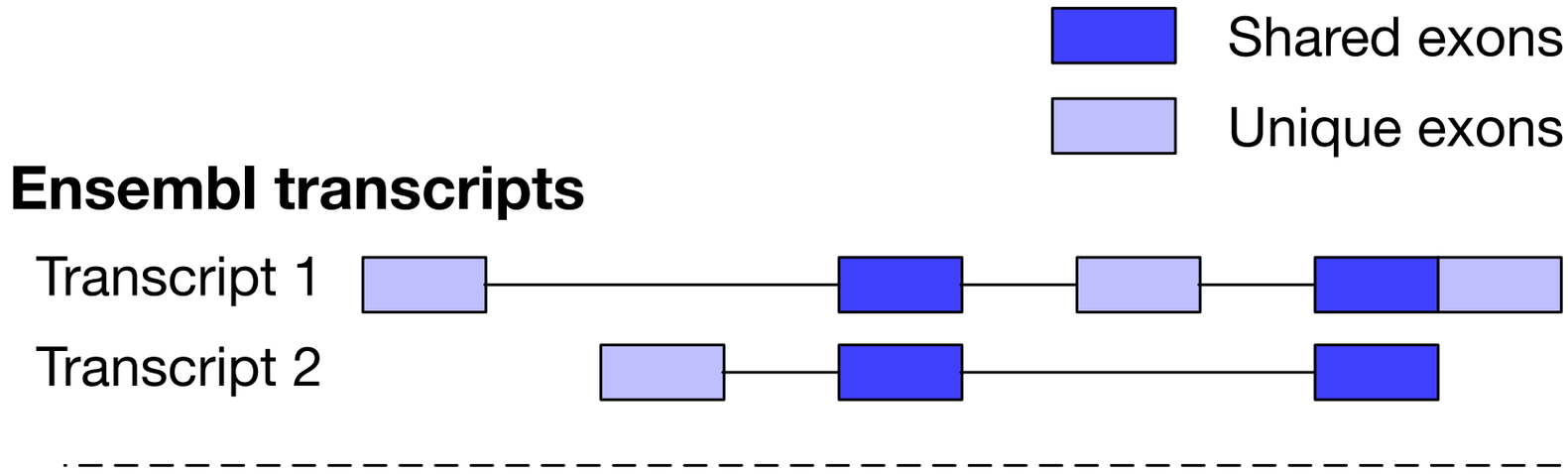
Annotated transcripts

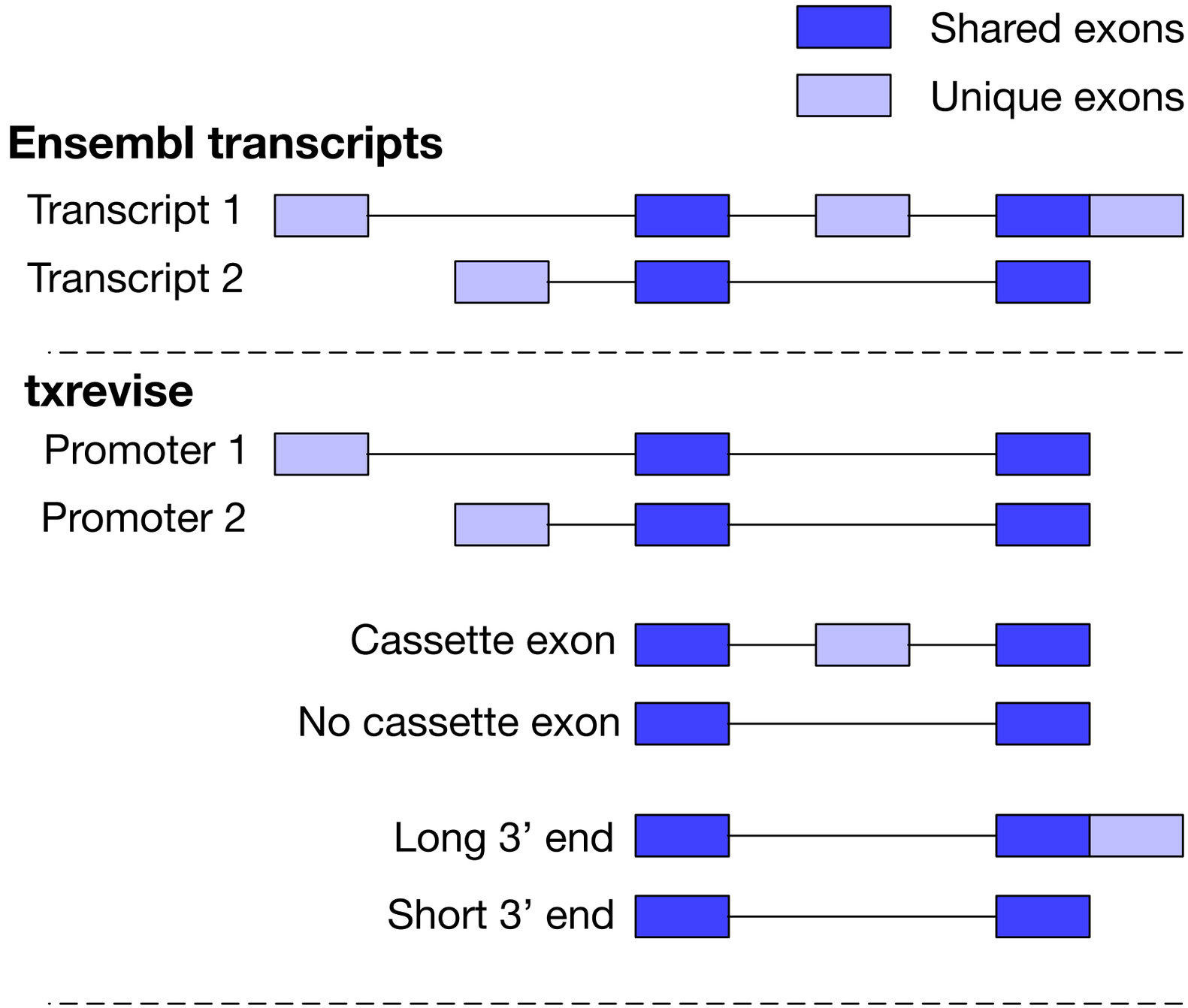


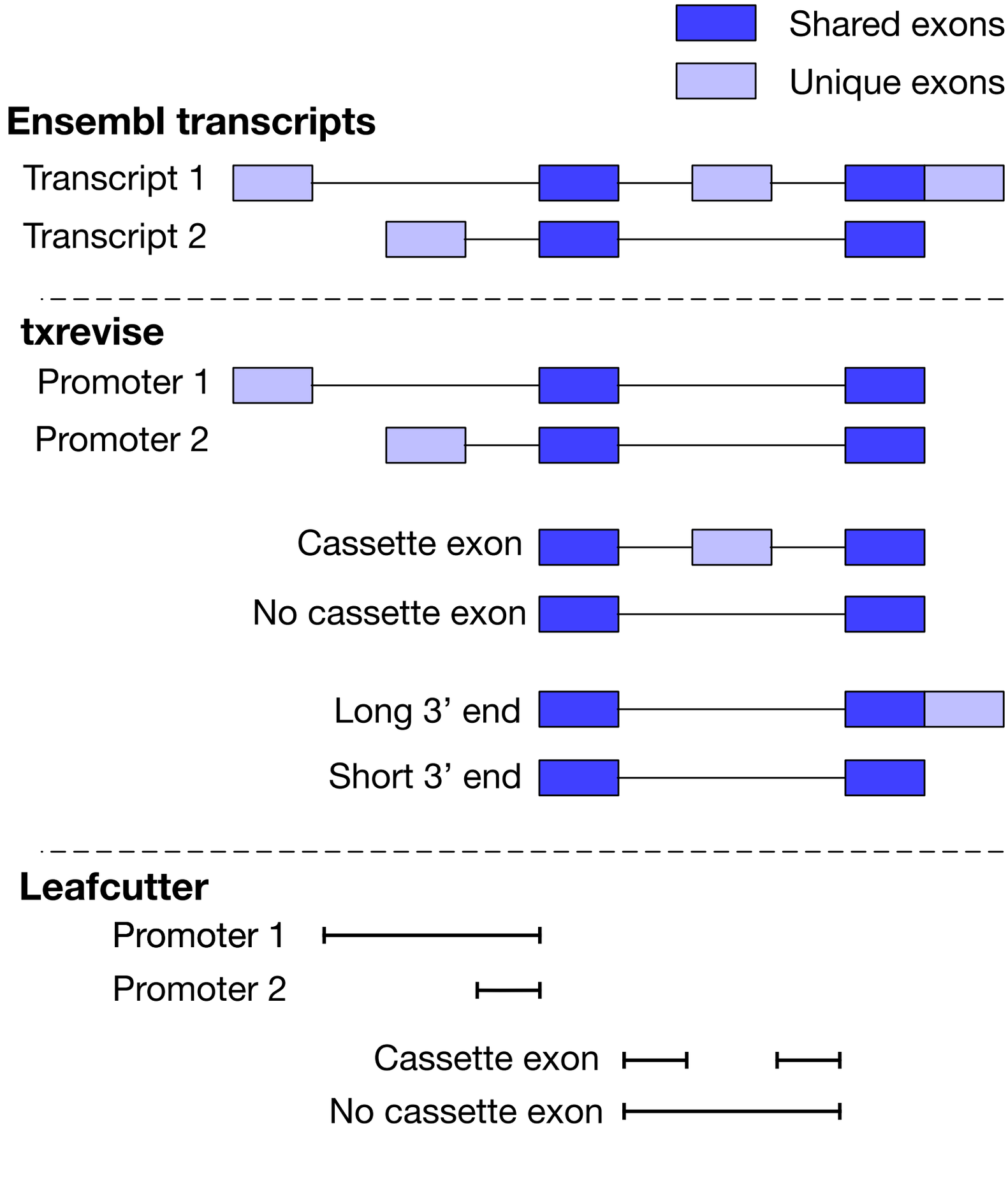
93%	73%	**Estimated transcript usage**
7%	27%	

Which event is driving the signal?

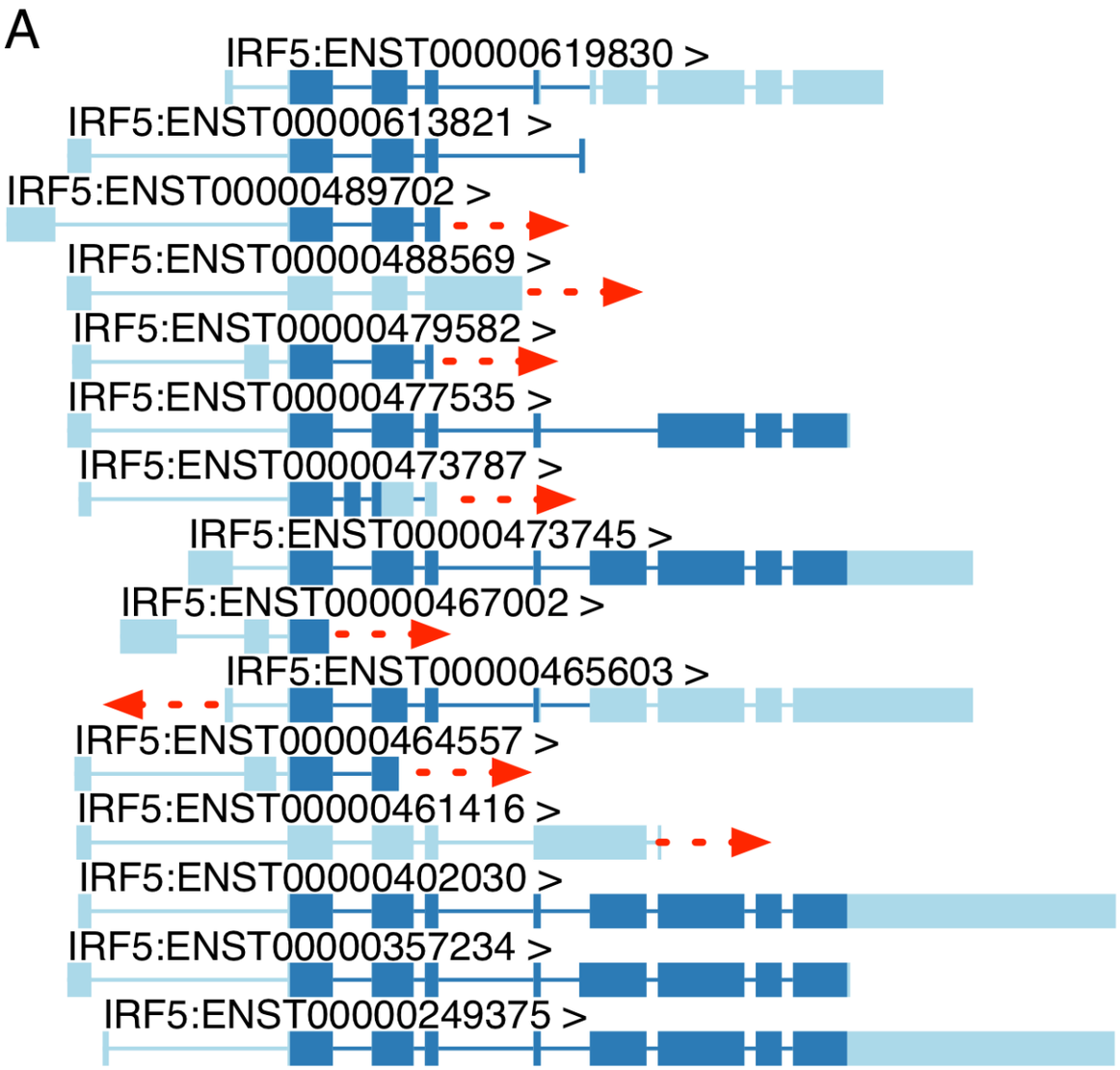
Biased estimates!

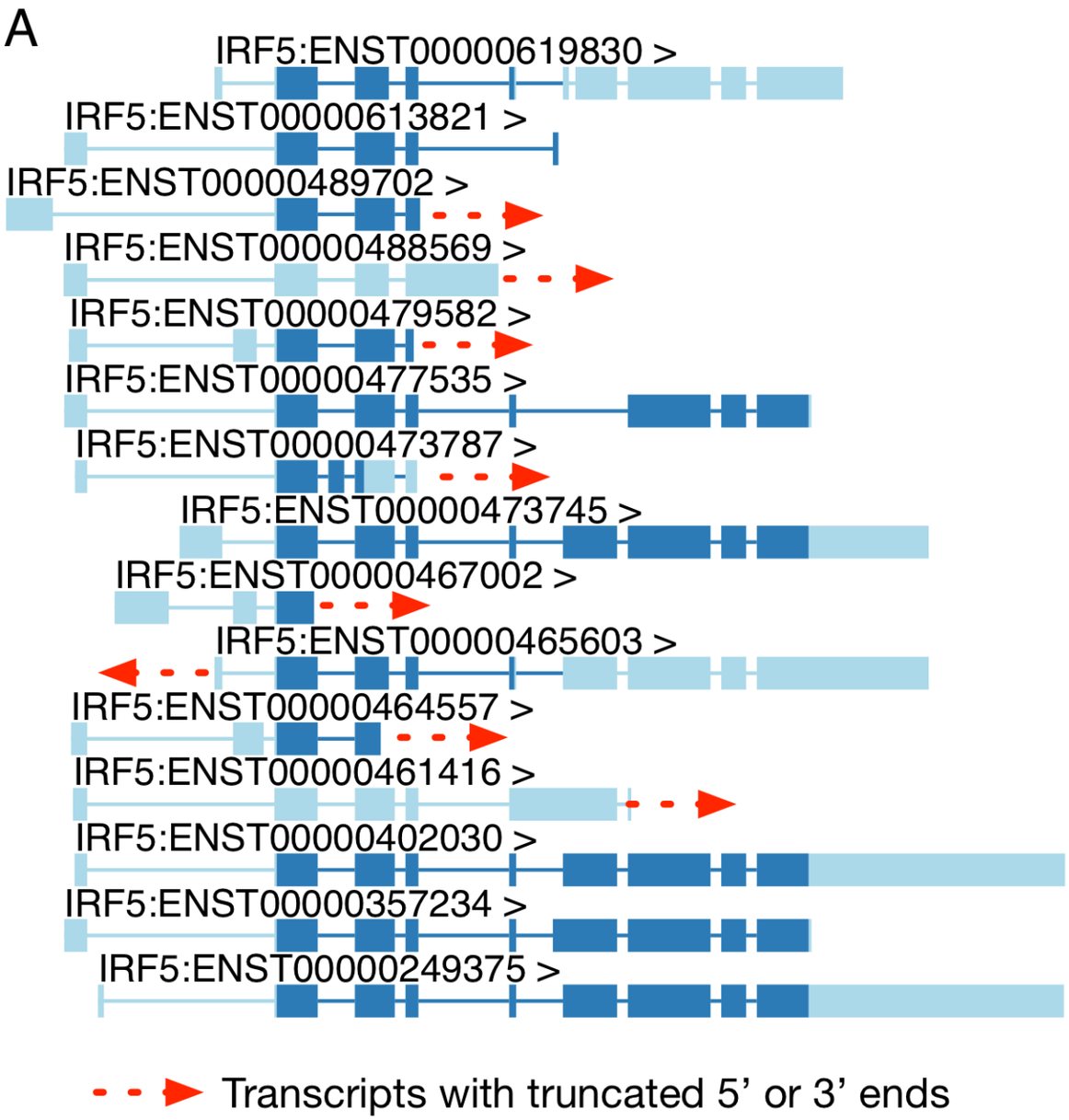

Shared exons
Unique exons
Ensembl transcripts
Transcript 1
Transcript 2


Shared exons
Unique exons
Ensembl transcripts
Transcript 1
Transcript 2
txrevise
Promoter 1
Promoter 2
Cassette exon
No cassette exon
Long 3’ end
Short 3’ end


Shared exons
Unique exons
Ensembl transcripts
Transcript 1
Transcript 2
txrevise
Promoter 1
Promoter 2
Cassette exon
No cassette exon
Long 3’ end
Short 3’ end
Leafcutter
Promoter 1
Promoter 2
Cassette exon
No cassette exon

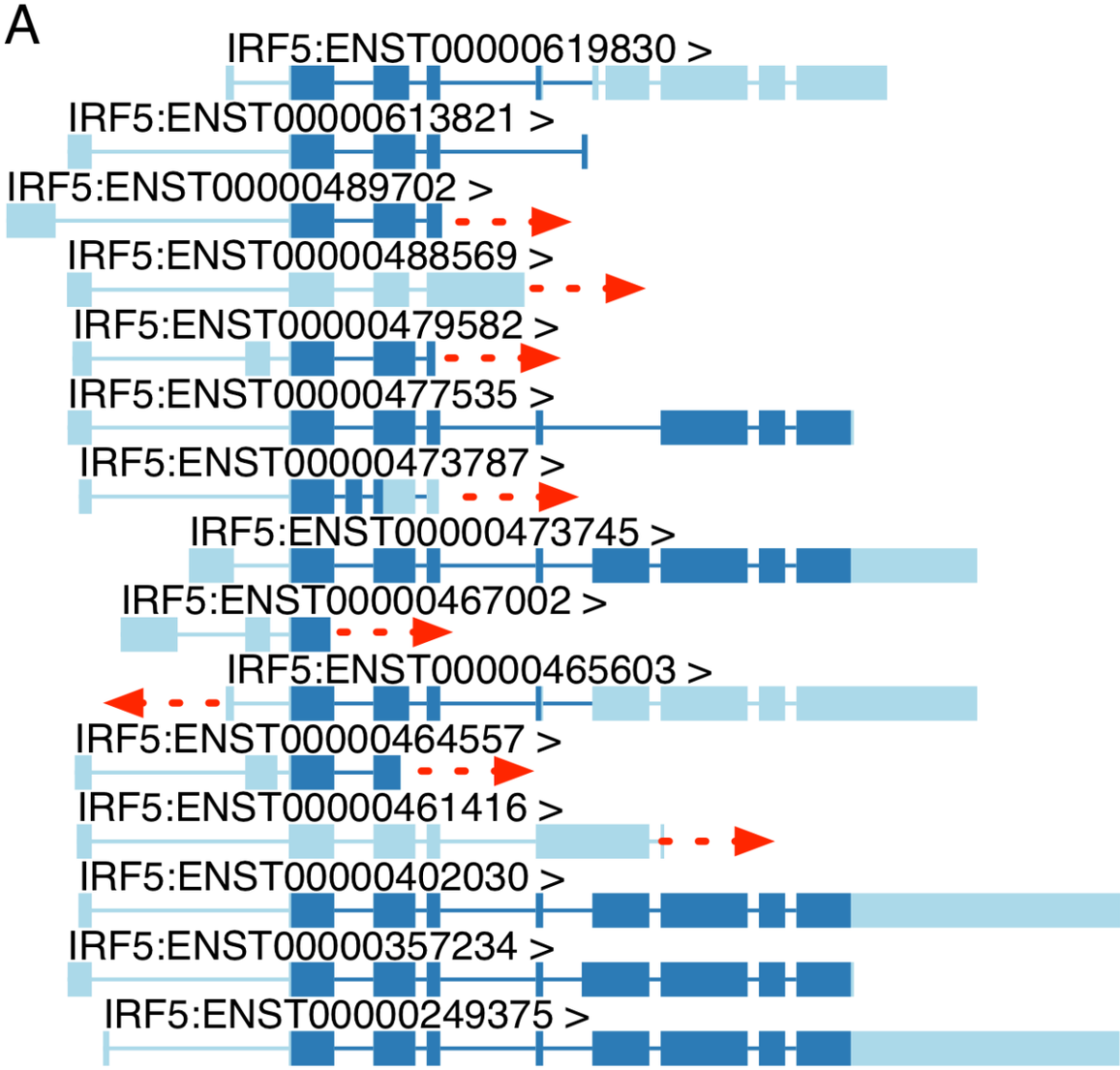
A


IRF5:ENST00000619830 >
IRF5:ENST00000613821 >
IRF5:ENST00000489702 >
IRF5:ENST00000488569 >
IRF5:ENST00000479582 >
IRF5:ENST00000477535 >
IRF5:ENST00000473787 >
IRF5:ENST00000473745 >
IRF5:ENST00000467002 >
IRF5:ENST00000465603 >
IRF5:ENST00000464557 >
IRF5:ENST00000461416 >
IRF5:ENST00000402030 >
IRF5:ENST00000357234 >
IRF5:ENST00000249375 >
Transcripts with truncated 5’ or 3’ ends

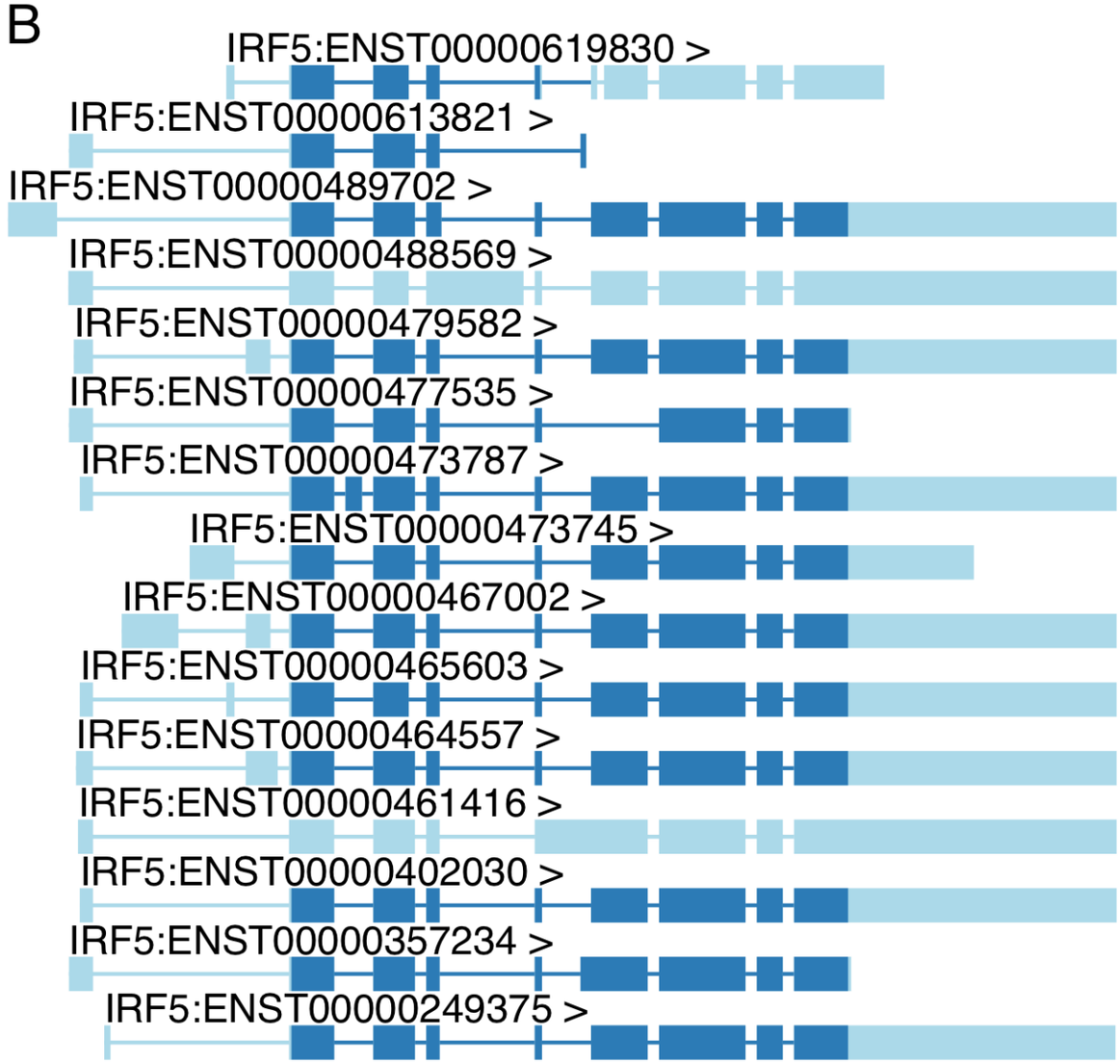


58% of the transcripts are truncated!

A

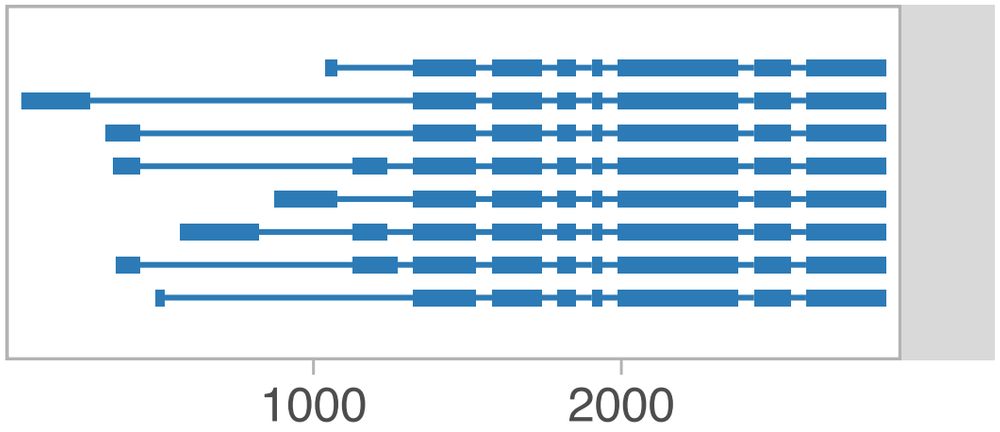


B

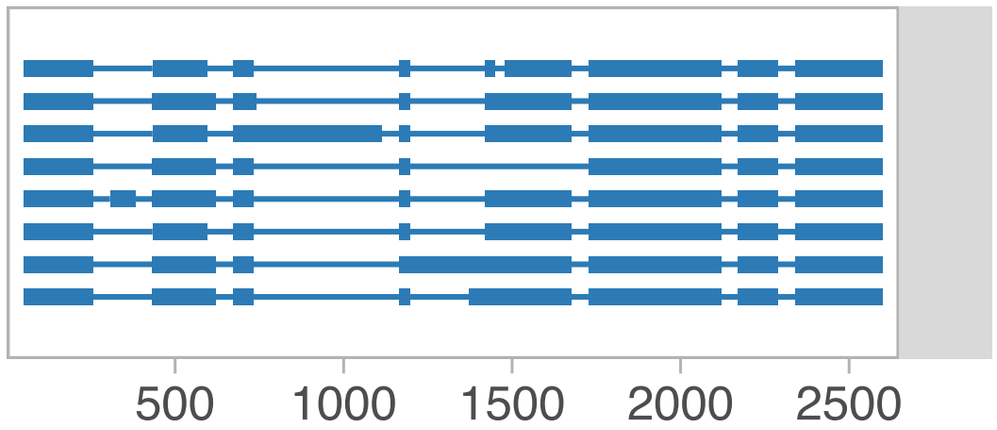


58% of the transcripts are truncated!

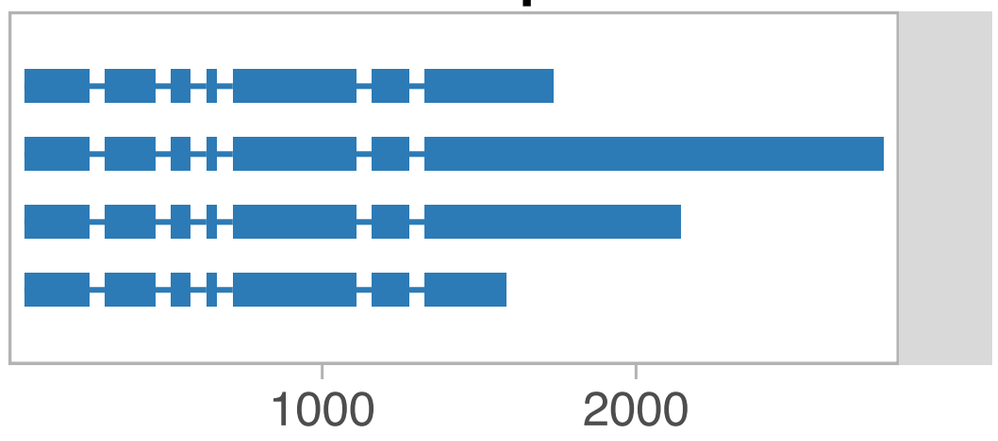
Alternative transcript starts



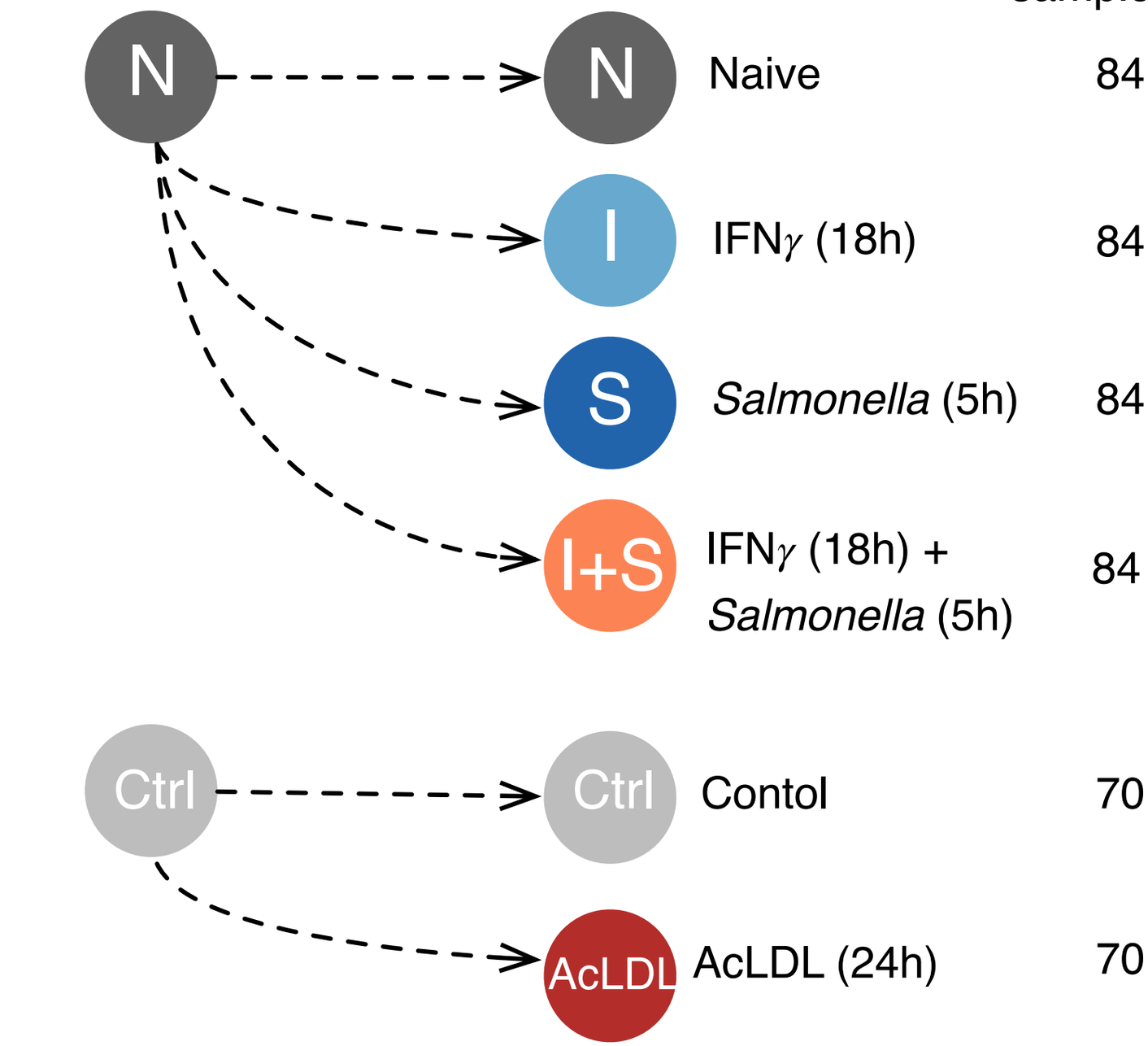
Alternative middle sections



Alternative transcript ends

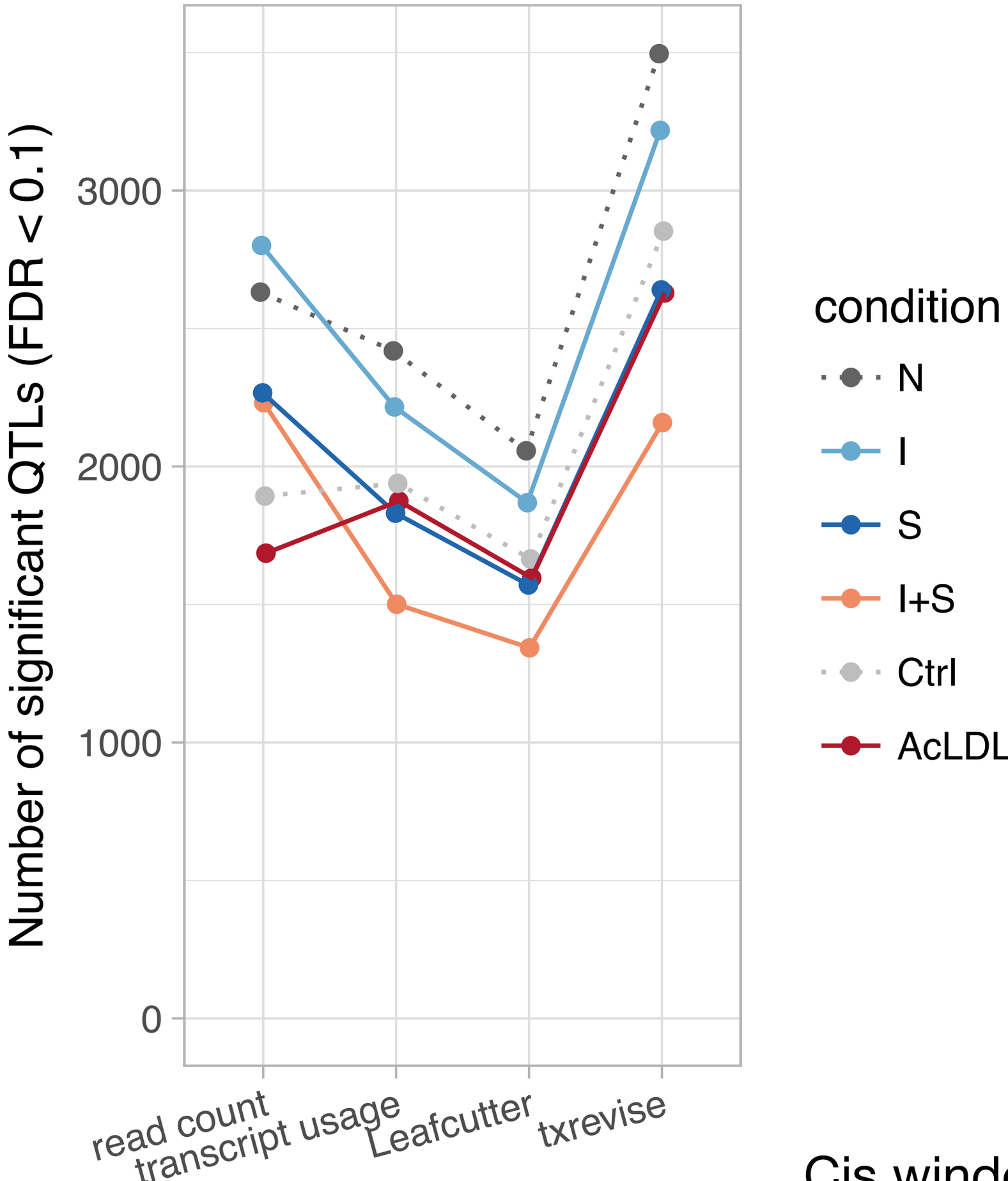


Human macrophages



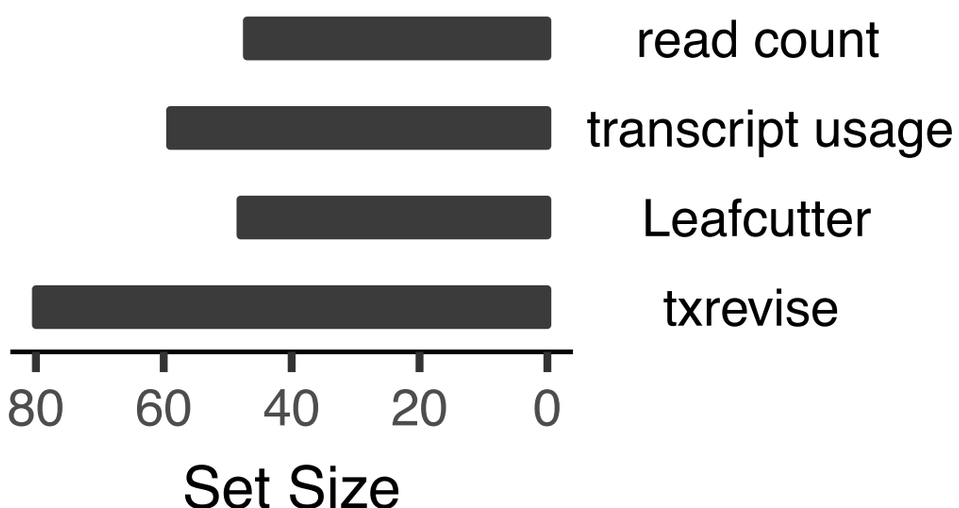
Quantification strategies

- **read count** - STAR alignment + featureCounts
- **transcript usage** - Ensembl transcripts + Salmon
- **txrevise** - transcript events + Salmon
- **Leafcutter** - exon-exon junction read counts

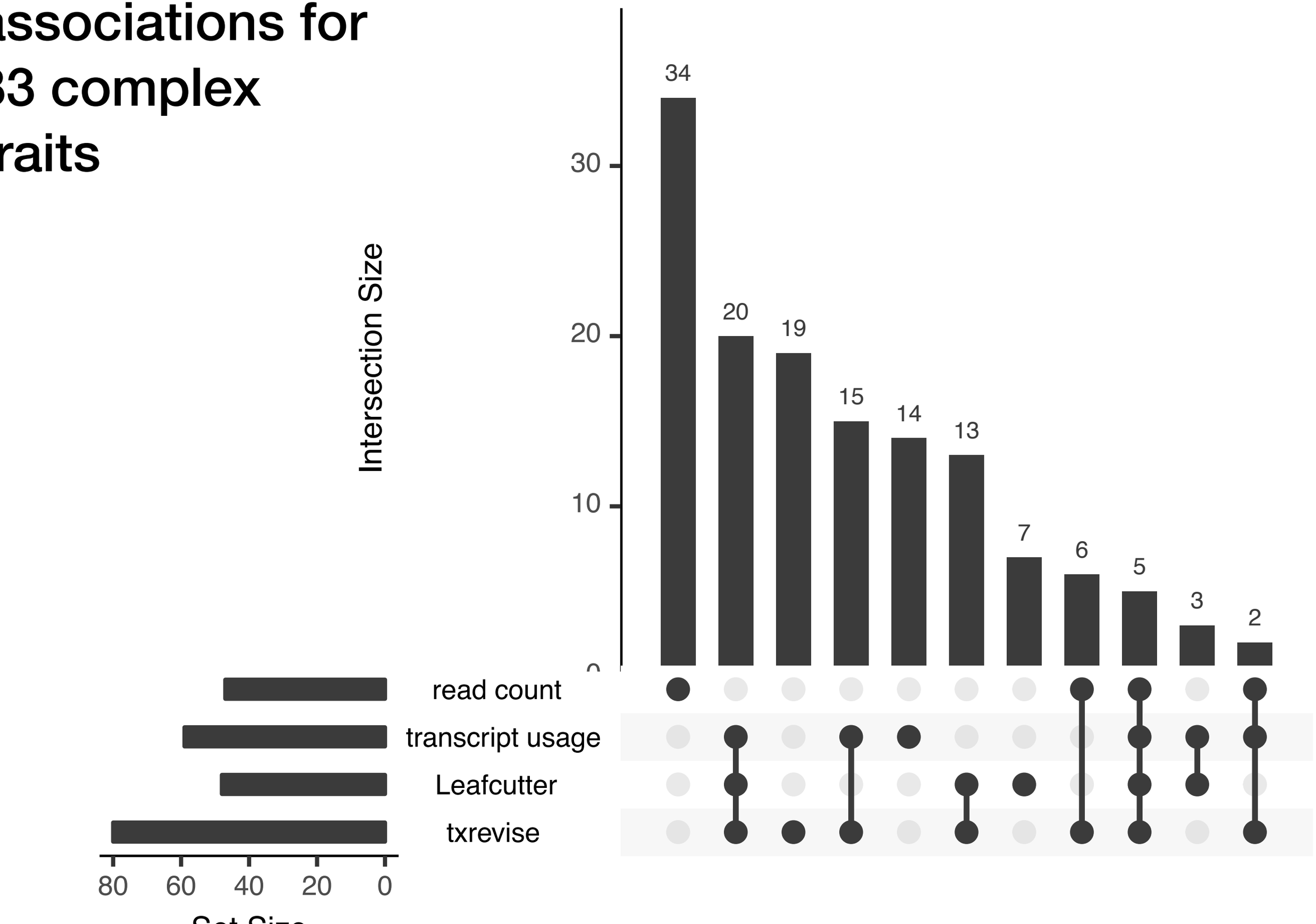


Cis window: +/- 100kb

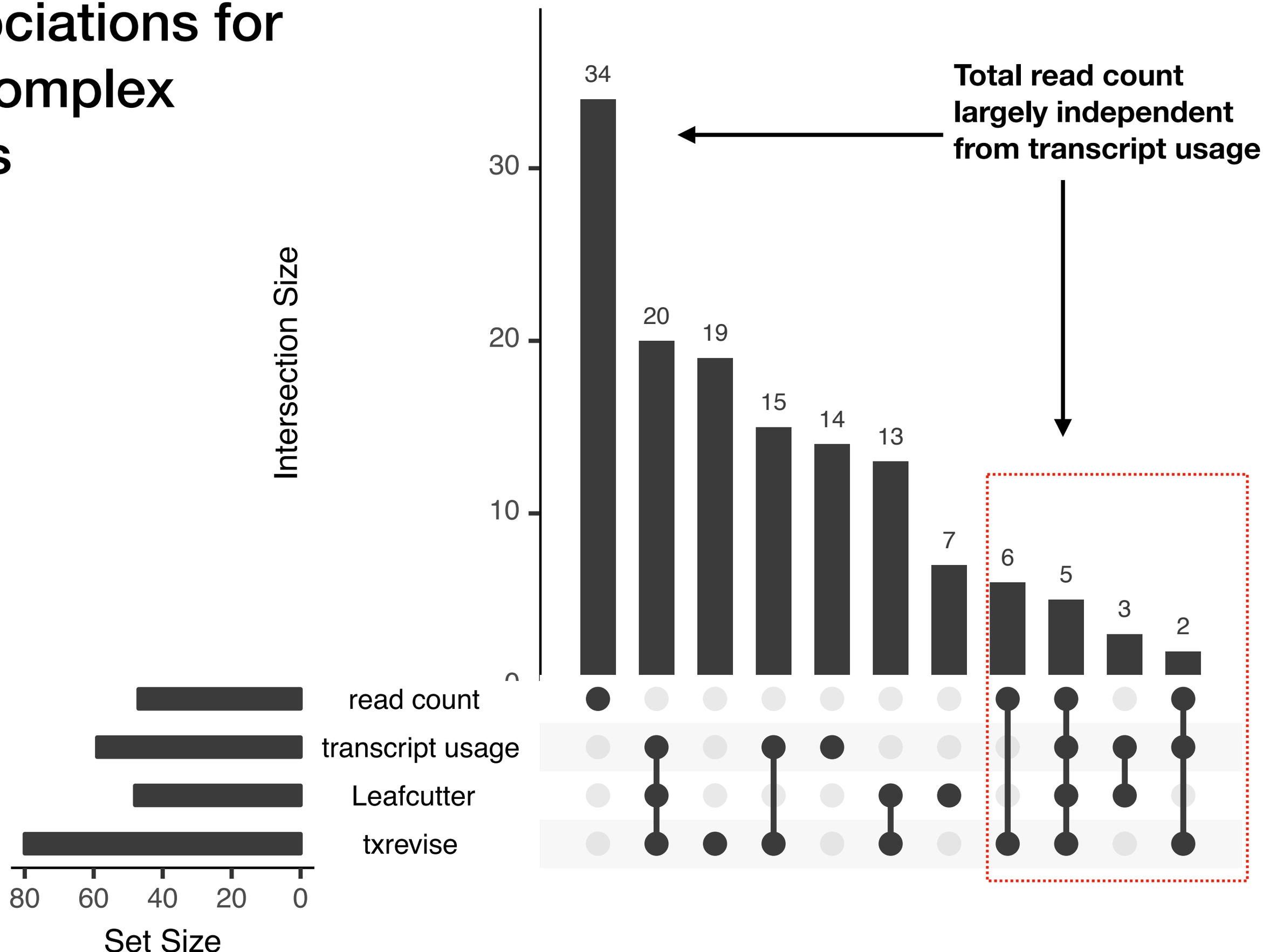
Overlap with associations for 33 complex traits



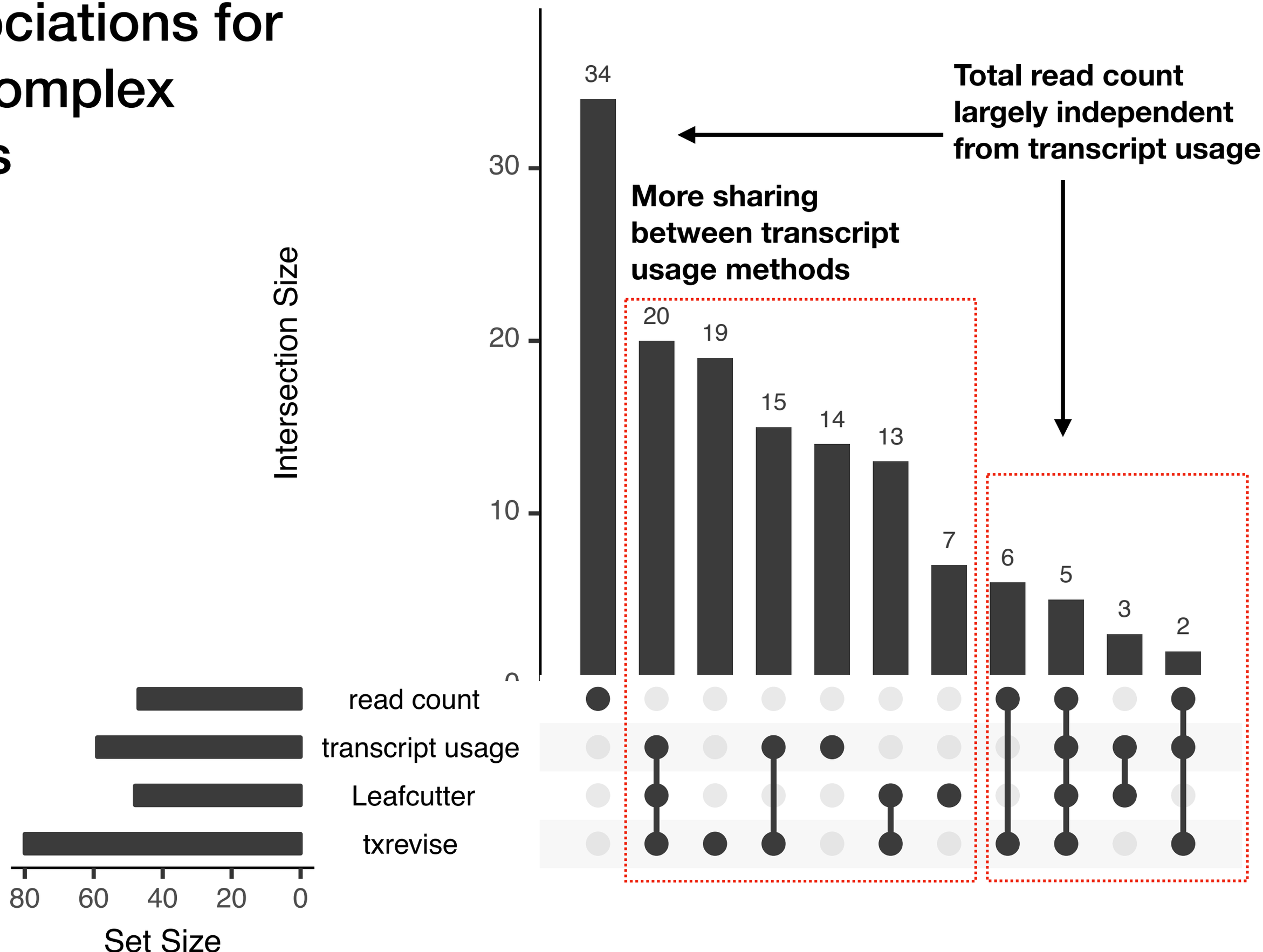
Overlap with associations for 33 complex traits

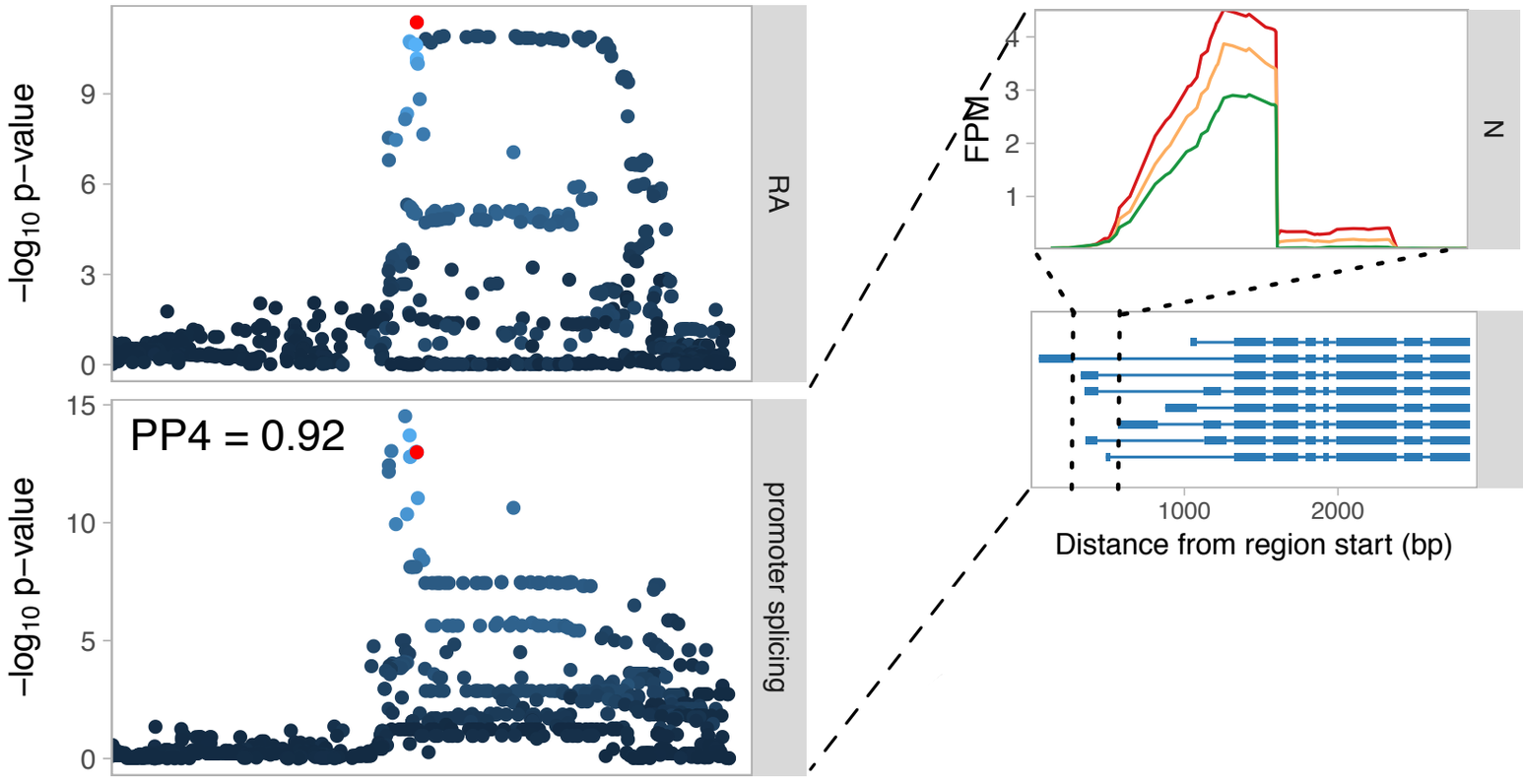


Overlap with associations for 33 complex traits



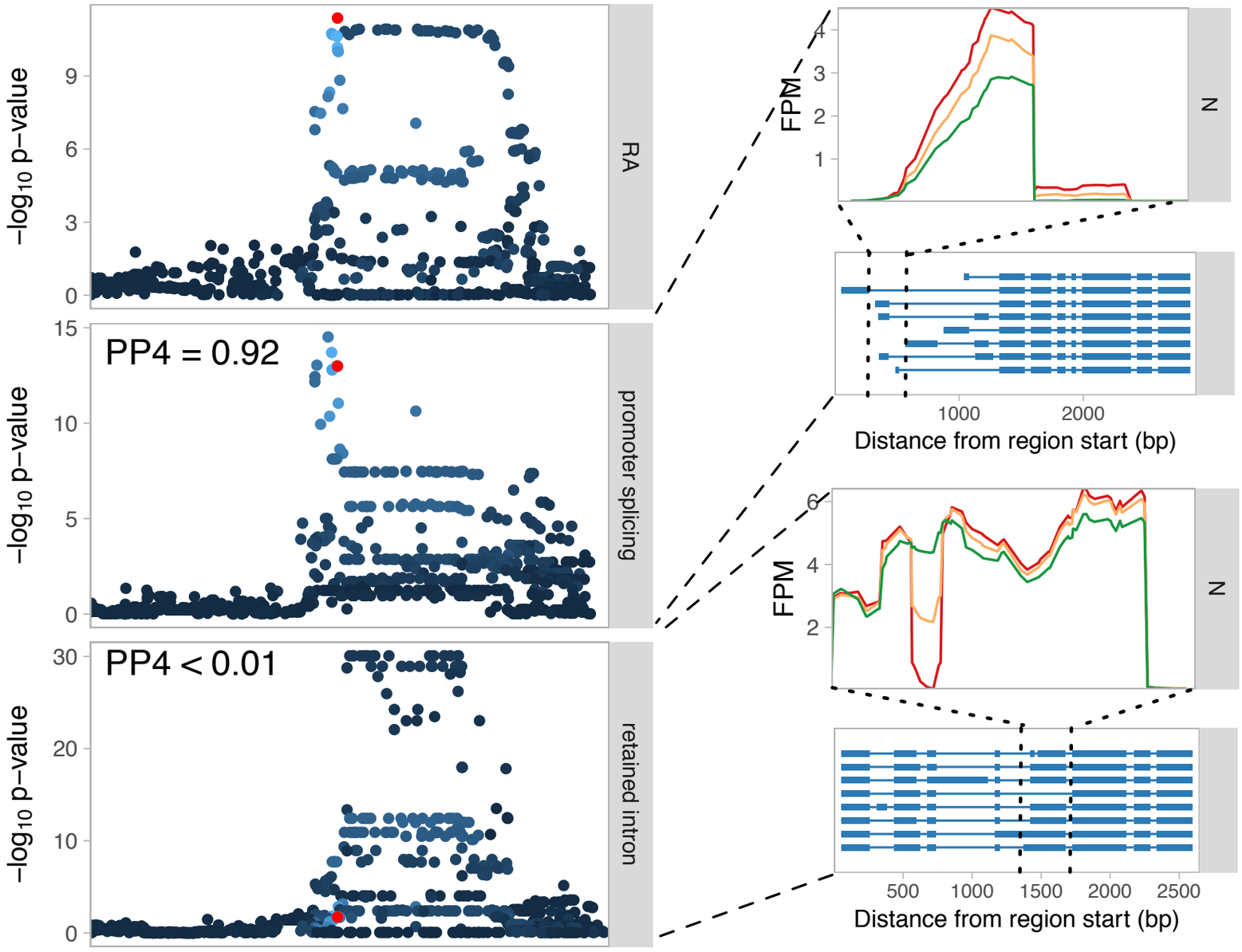
Overlap with associations for 33 complex traits





Detected by:

- ✗ read count
- ✗ transcript usage
- ✓ Leafcutter
- ✓ txrevise

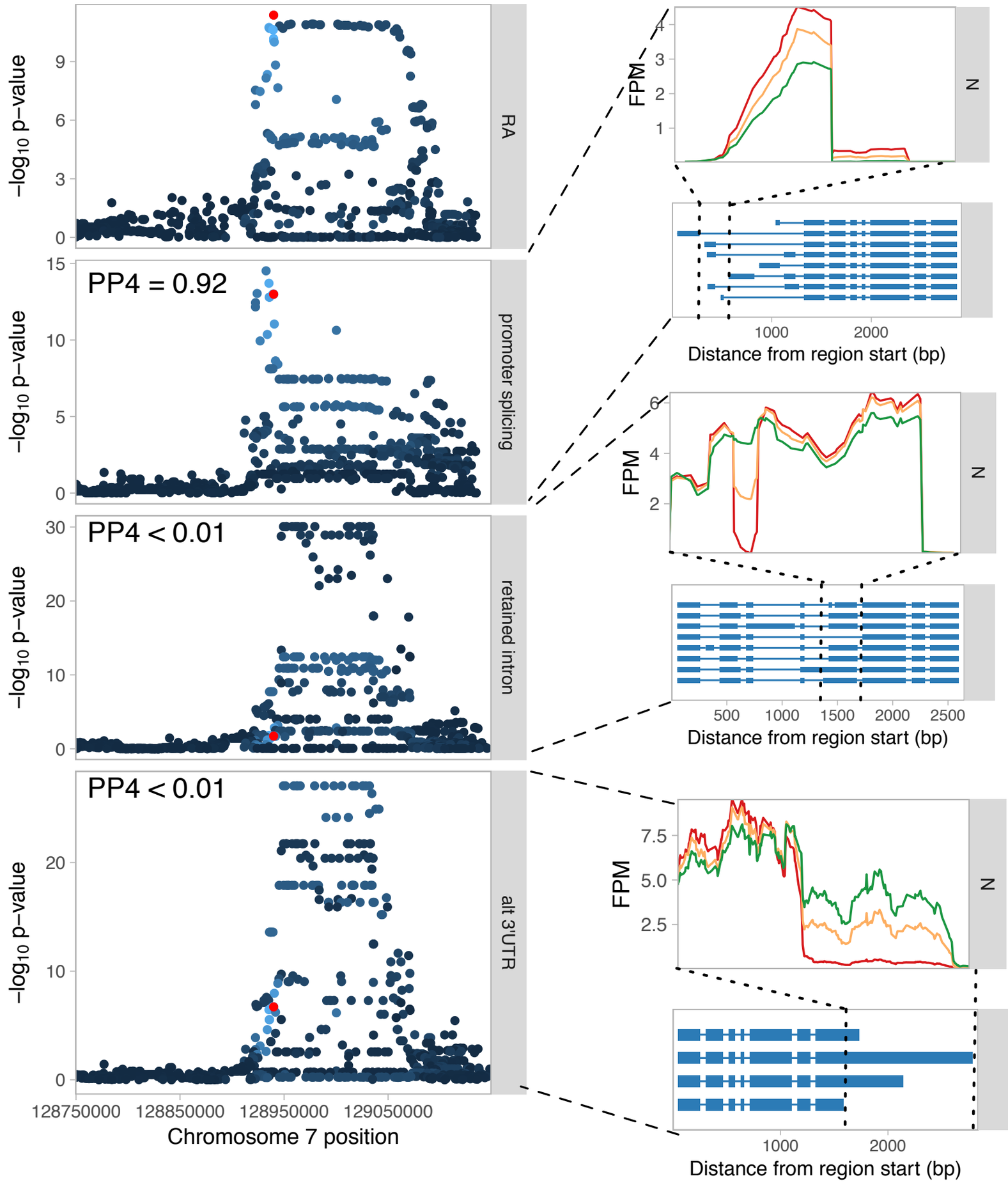


Detected by:

- ✗ read count
- ✗ transcript usage
- ✓ Leafcutter
- ✓ txrevise

Detected by:

- ✗ read count
- ✗ transcript usage
- ✗ Leafcutter
- ✓ txrevise



Detected by:

- ✗ read count
- ✗ transcript usage
- ✓ Leafcutter
- ✓ txrevise

Detected by:

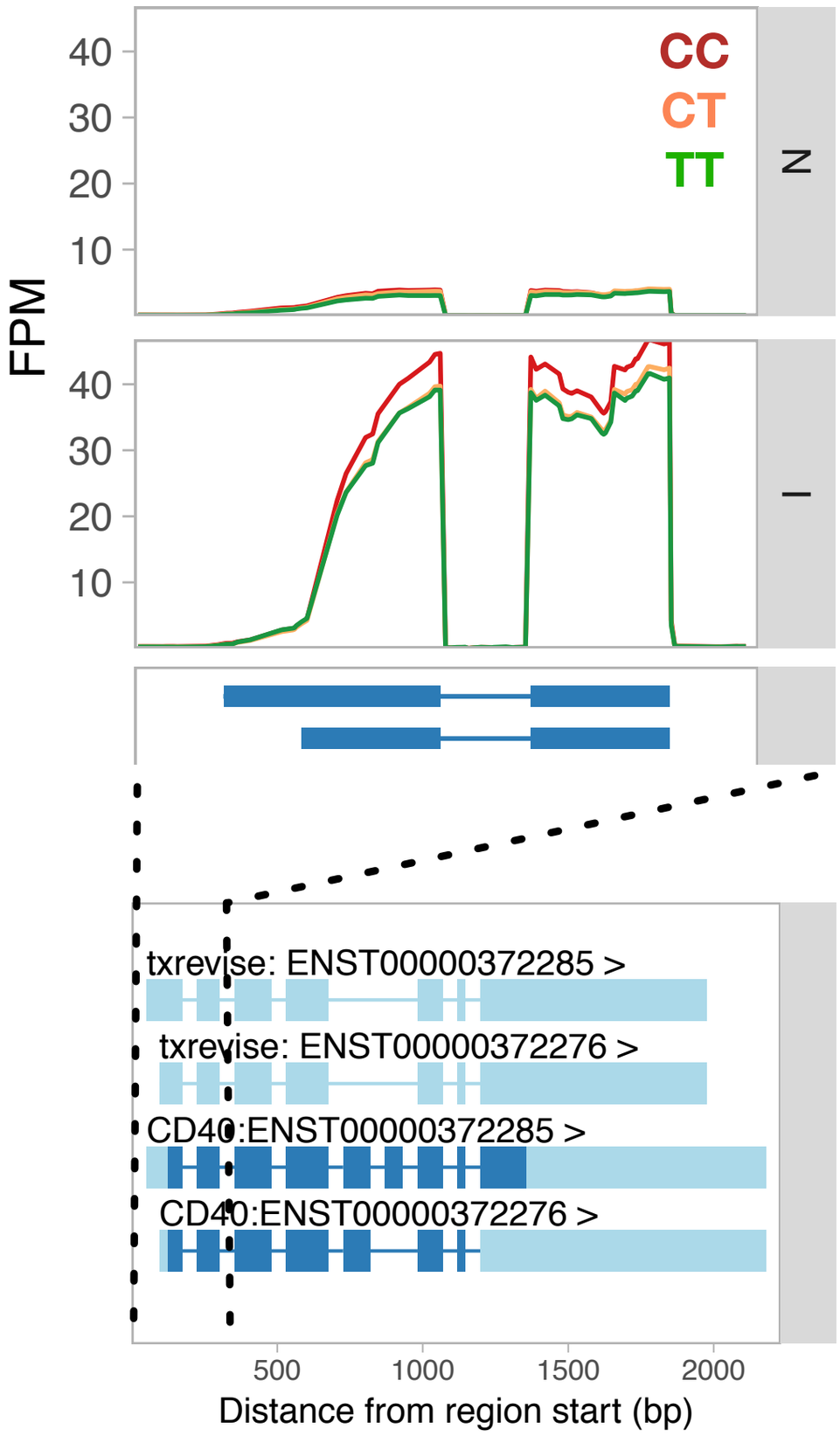
- ✗ read count
- ✗ transcript usage
- ✗ Leafcutter
- ✓ txrevise

Detected by:

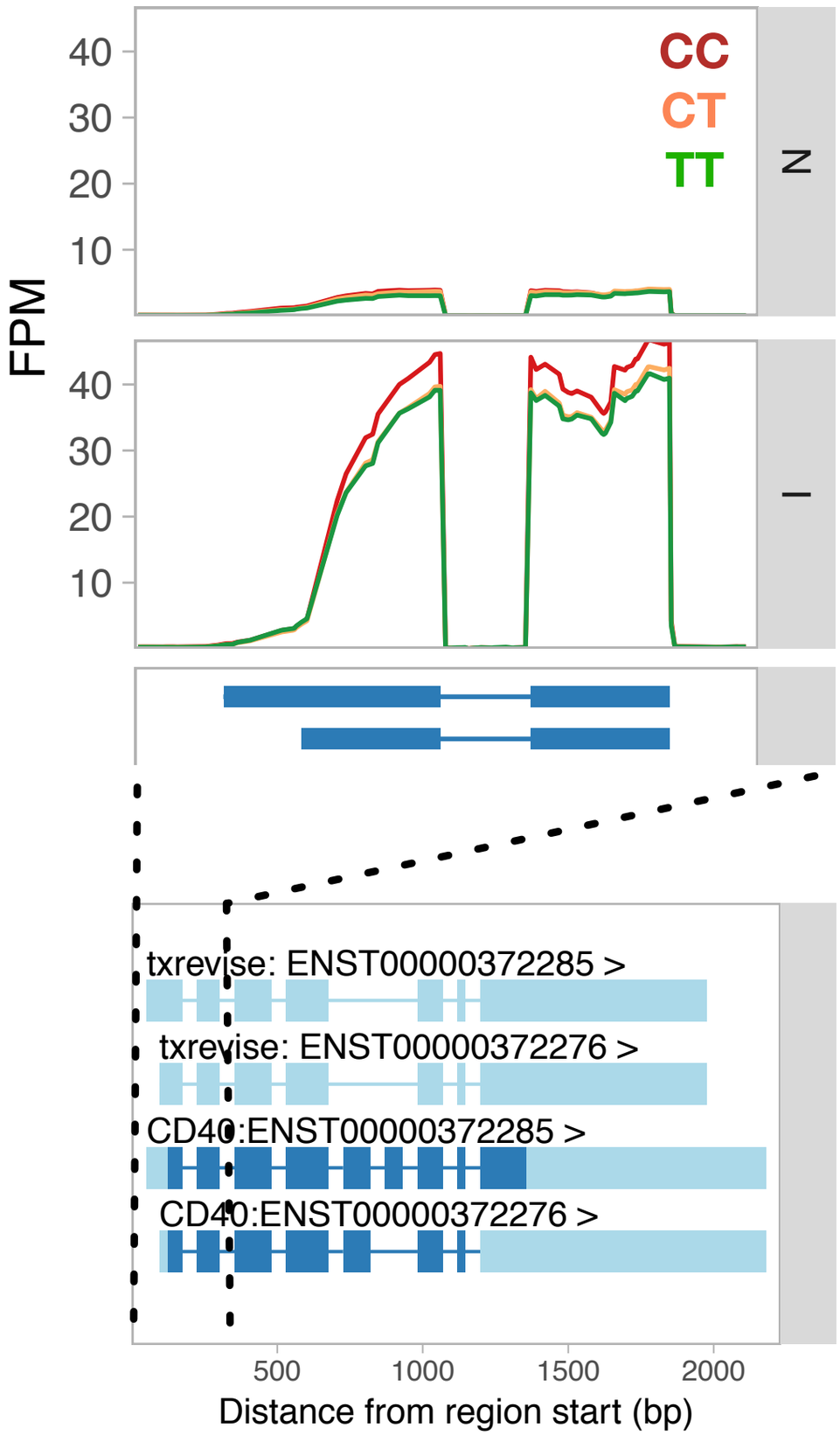
- ✗ read count
- ✓ transcript usage
- ✗ Leafcutter
- ✓ txrevise

Response QTLs: genetic effects that appear after stimulation

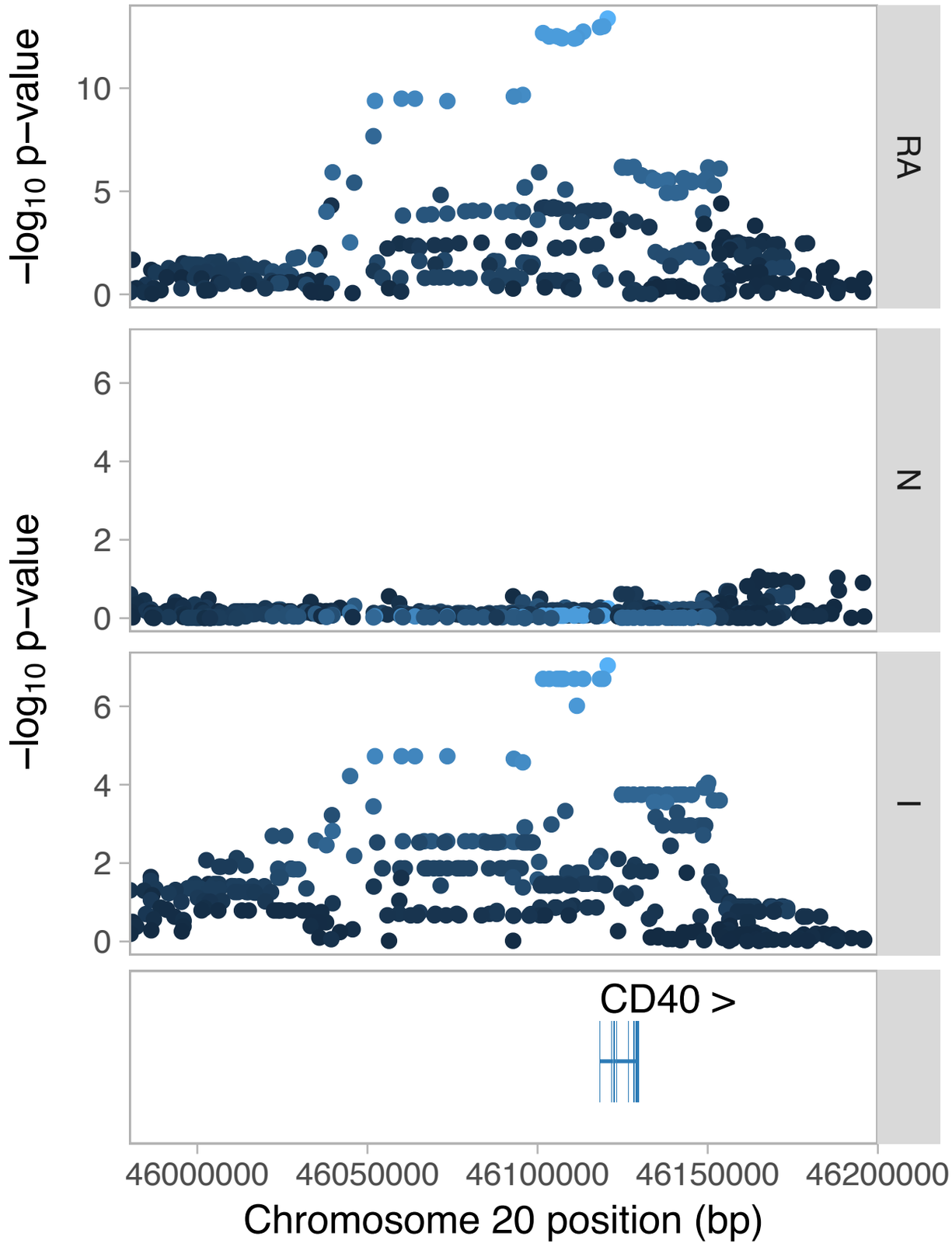
IFNg-specific response QTL at the 5'UTR of CD40

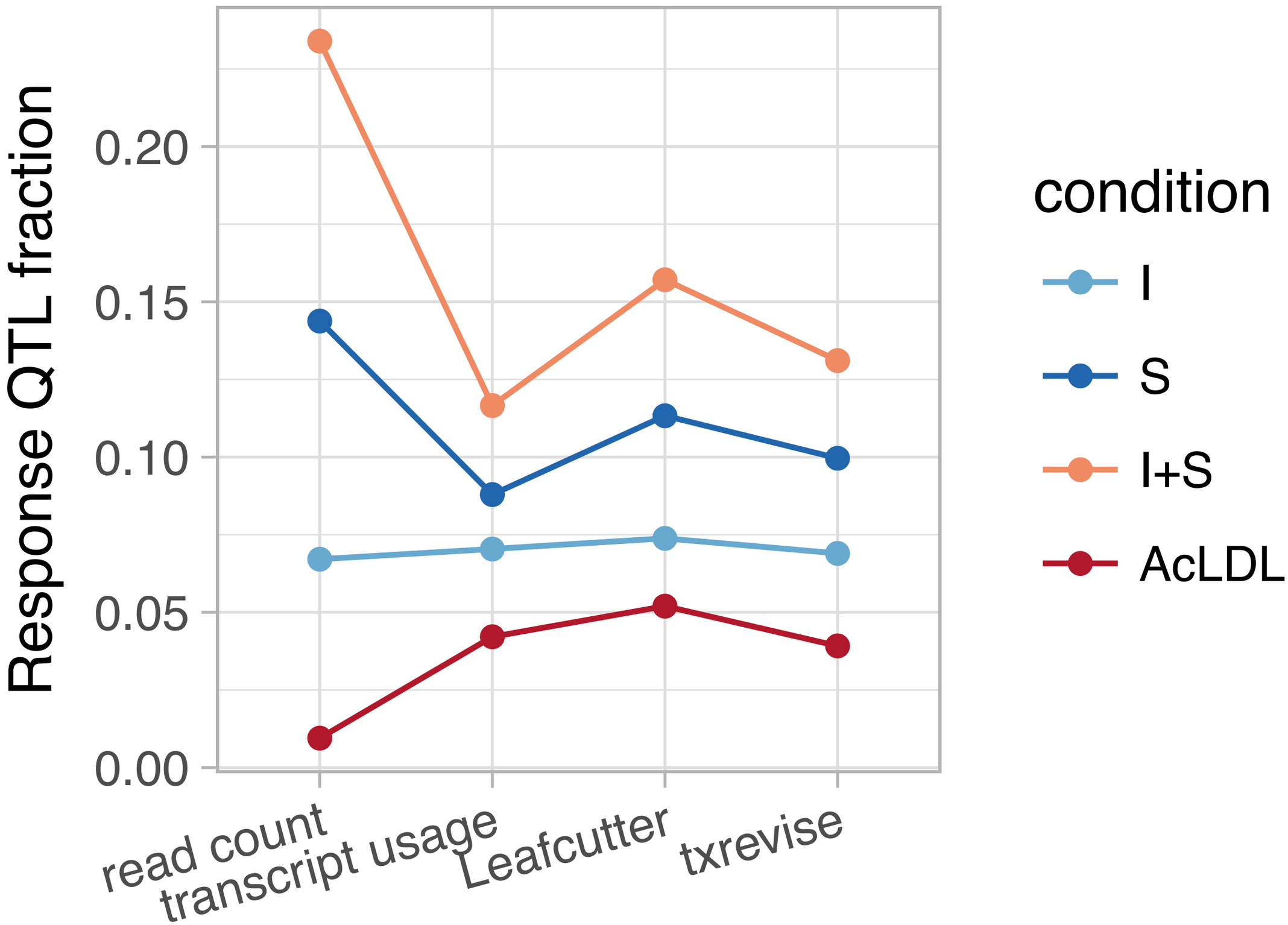


IFNg-specific response QTL at the 5'UTR of CD40



Colocalisation with rheumatoid arthritis (RA) GWAS hit



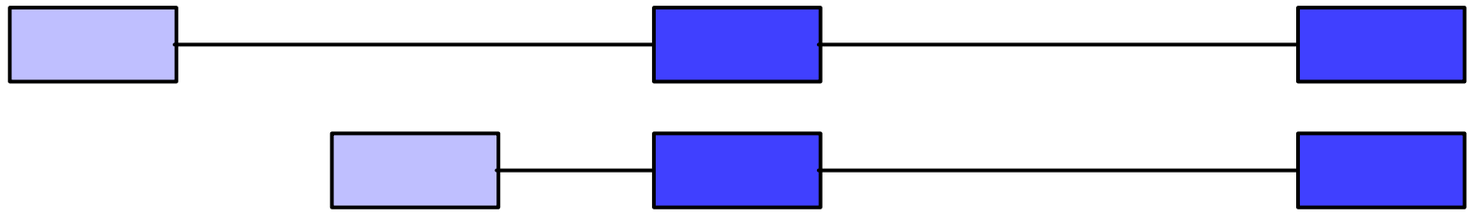

Response QTL fraction
0.20
0.15
0.10
0.05
0.00
read count
transcript usage
Leafcutter
txrevise
condition
I
S
I+S
AcLDL

txrevise

Start

Promoter 1

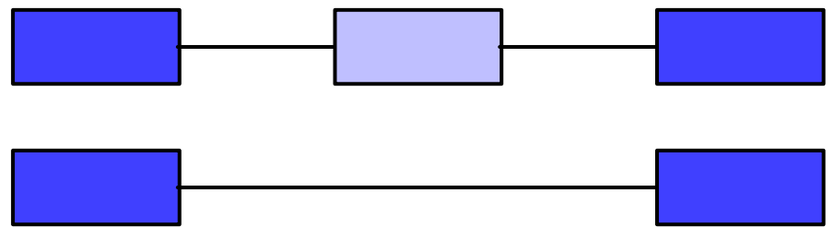
Promoter 2



Middle

Cassette exon

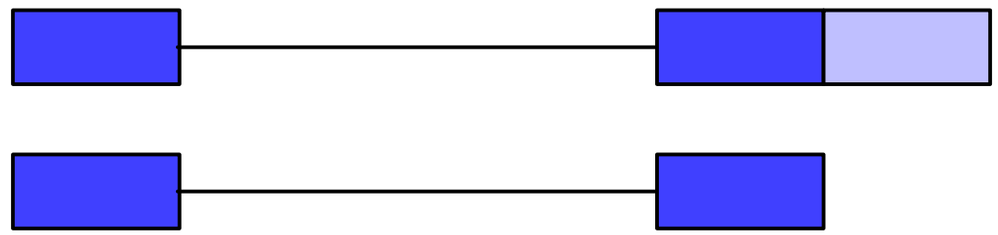
No cassette exon



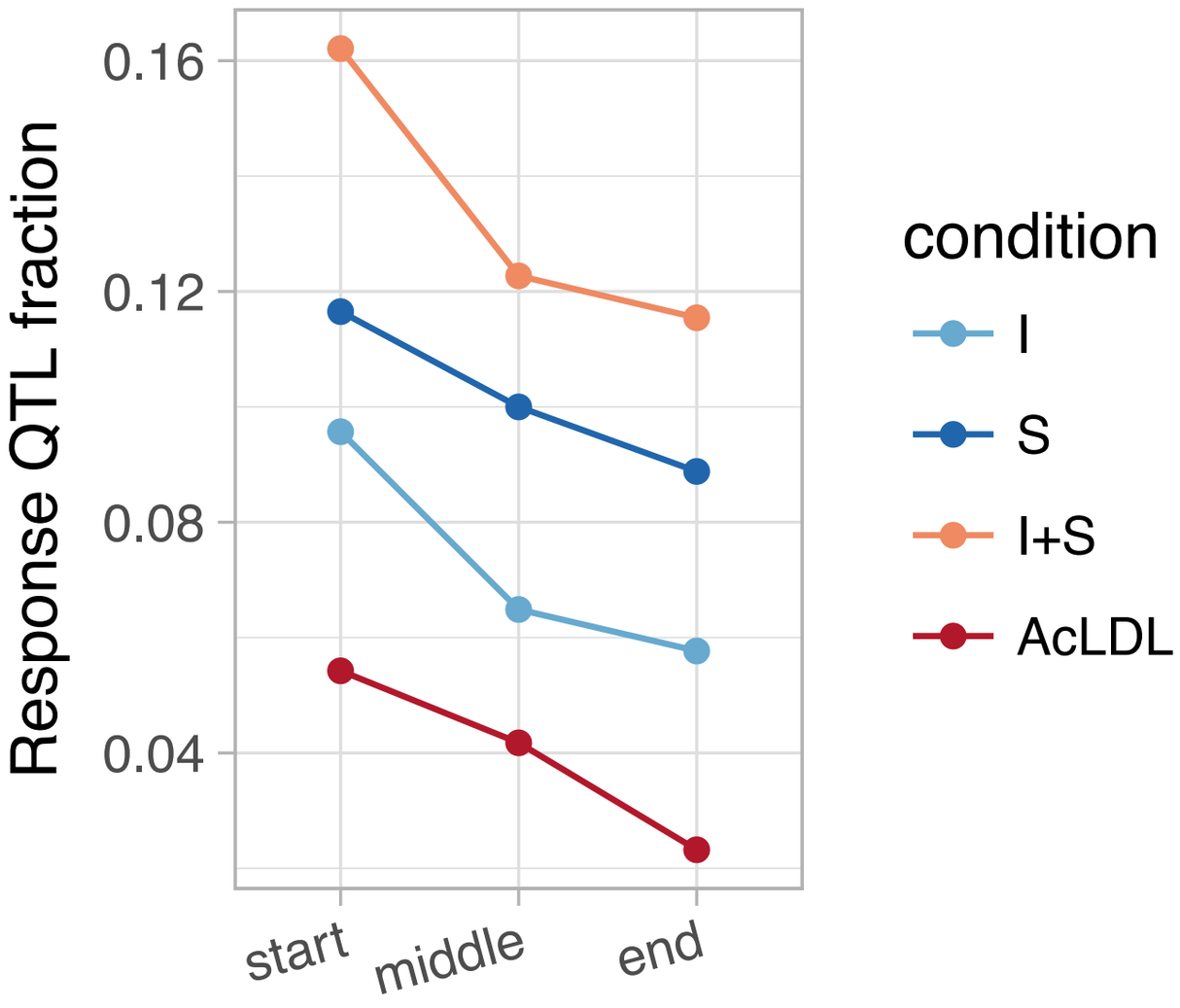
End

Long 3’ end

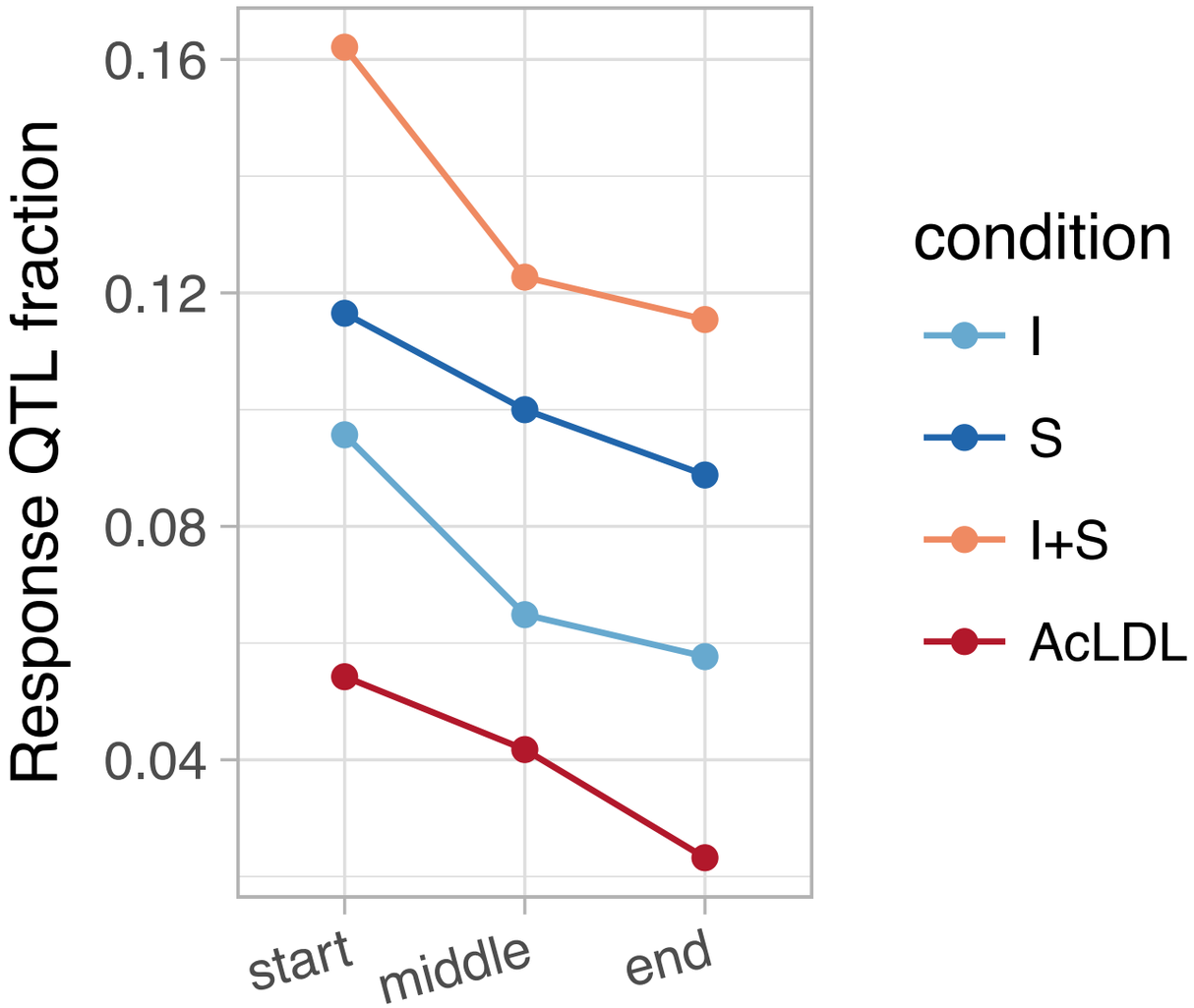
Short 3’ end

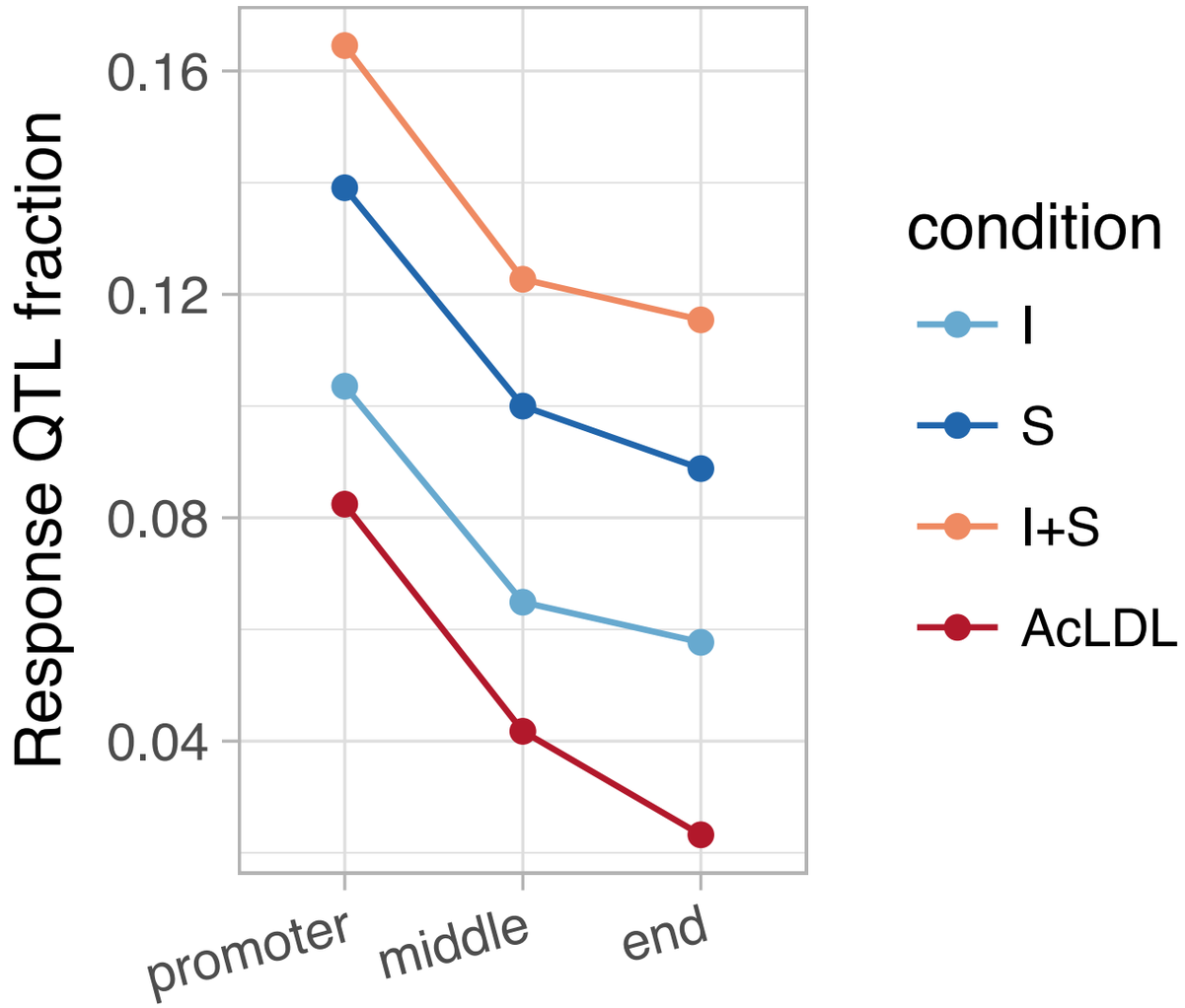


Response QTLs across the gene body



Response QTLs across the gene body





Conclusions

- **txrevise** detects additional associations missed by other methods.

- Genetics of transcript usage is largely independent from gene expression.

- Promoter QTLs are more condition-specific than other transcript usage QTLs.

References

- **txrevise:** revised transcript annotations
https://github.com/kauralasoo/txrevise

- **wiggleplotr:** RNA-seq read coverage plots
http://bioconductor.org/packages/release/bioc/html/wiggleplotr.html

- **Experimental setup and data**:
Shared genetic effects on chromatin and gene expression reveal widespread enhancer priming in immune response
http://www.biorxiv.org/content/early/2017/05/18/102392

Wellcome Trust Sanger Institute

Daniel Gaffney

Julia Rodrigues

Gordon Dougan

HipSci Project Sequencing Core Facility

University of Cambridge

Dirk Paul

Funding

wellcome trust

Mobilitas+

Eesti Teadusagentuur
Estonian Research Council