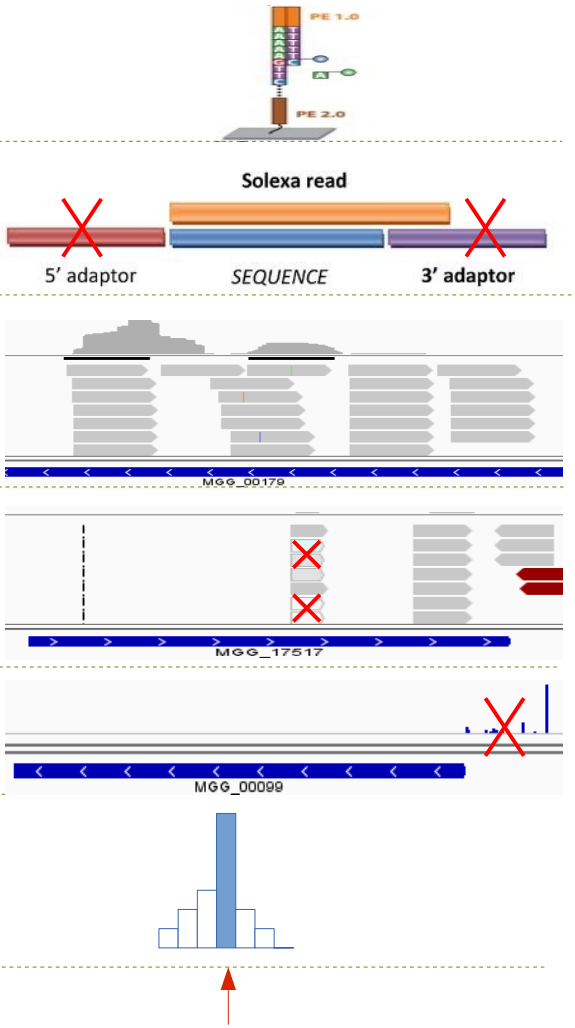
Sequencing resume

- 2 strains (*WT, 2D4*) x 4 conditions (*CM, MM, -N, -C*) x 3 replicates
- 4751592 – 11517077 total reads database
- ~63% - ~83% successfully mapped reads
- 43 bp mean read length
- ~92% - ~98% replicates correlation
- ~100bp mean pair ended distance
- ~400x coverage per poly(A) site*

** assuming an amount of 22000 mRNA molecules per cell*

Workflow

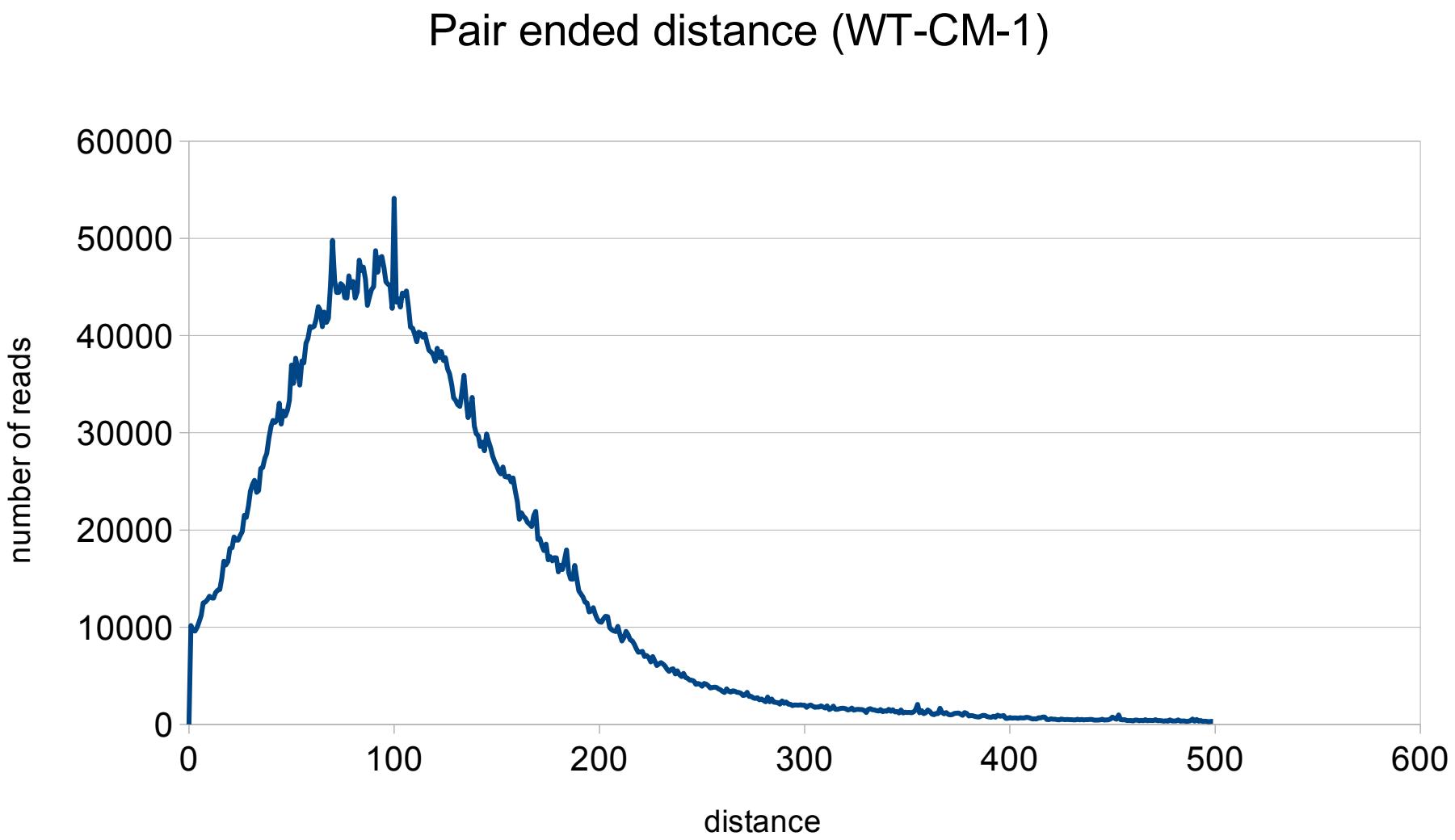
Program and details	Procedure	% of data left
	Illumina sequencing	100%
fasta-mcf, *adapter trimming*	Data prepping	96%
gsnap, *default parameters*	Alignment	96%
• *Low quality mapping (<30))* • *High A/T content (>90%)* • *Internal priming*	Filtering	75%
Assign reads to known features, Including all intergenic space up to 400 bp from gene end	Assignment	65%
Clustering of most significant poly(A) sites (p-value < 0.05 over global gene expression) at 33 bp	poly(A) sites detection	27%



*Example: **Actin,** single cut poly(A):*

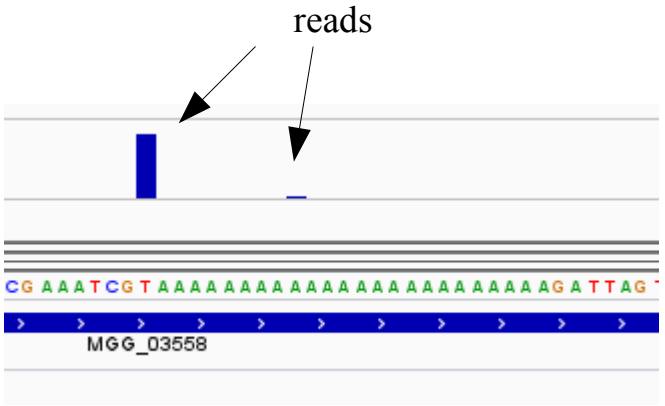
- whole gene expression ~7000 reads
- poly(A) site expression ~3000 reads

Pair ended reads distance

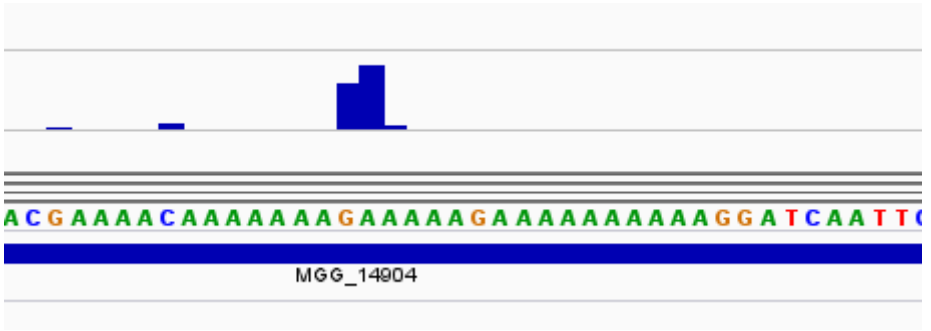


~2.5% of poly(A) sites are internal priming

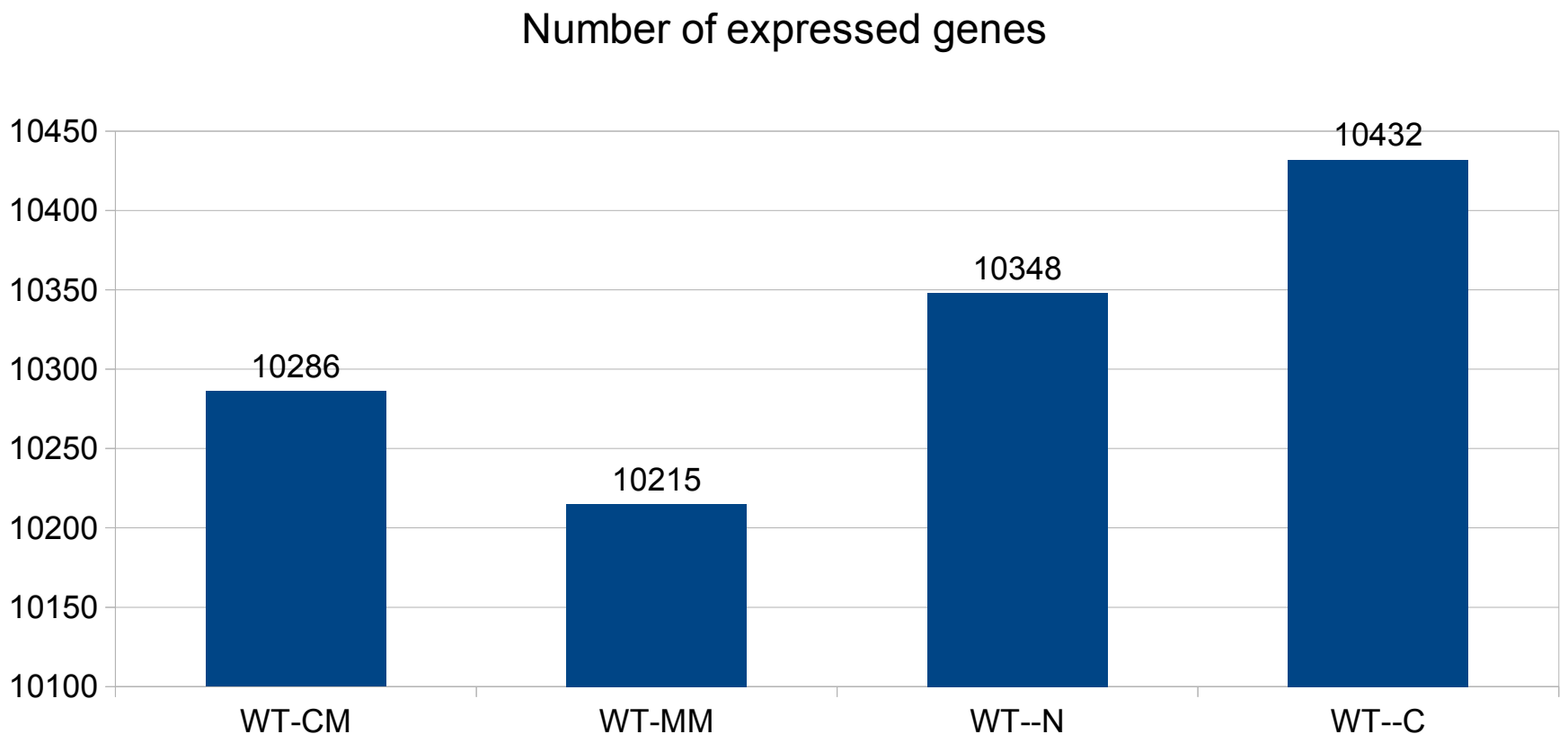
- Some poly(A) sites are just a side effect of poly(A) ***genomic*** regions



stretches of As

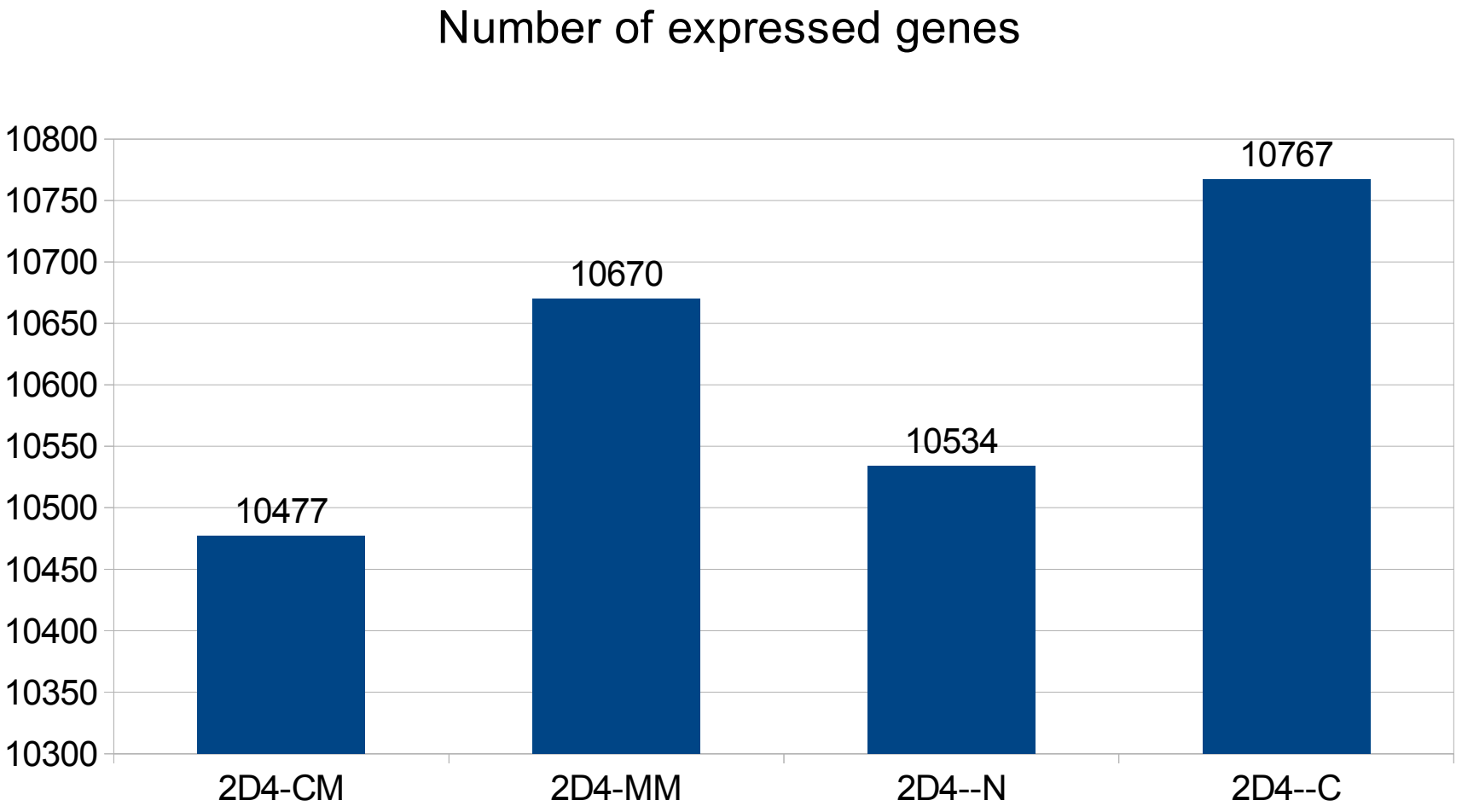


~10000 genes are expressed, out of a total of 13218 annotated genes (WT)



- 9546 genes are expressed in every condition (WT only)
- 2166 genes are never expressed (WT only)

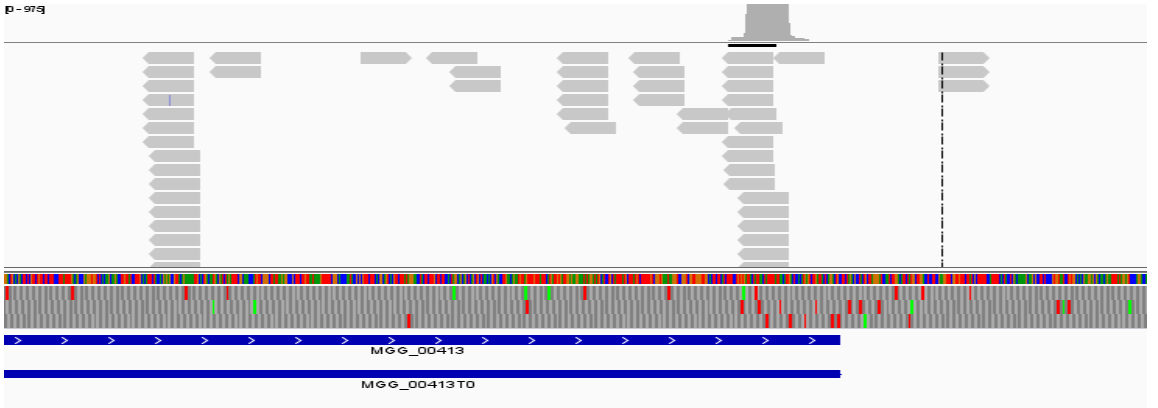
~10000 genes are expressed, out of a total of 13218 annotated genes (2D4)



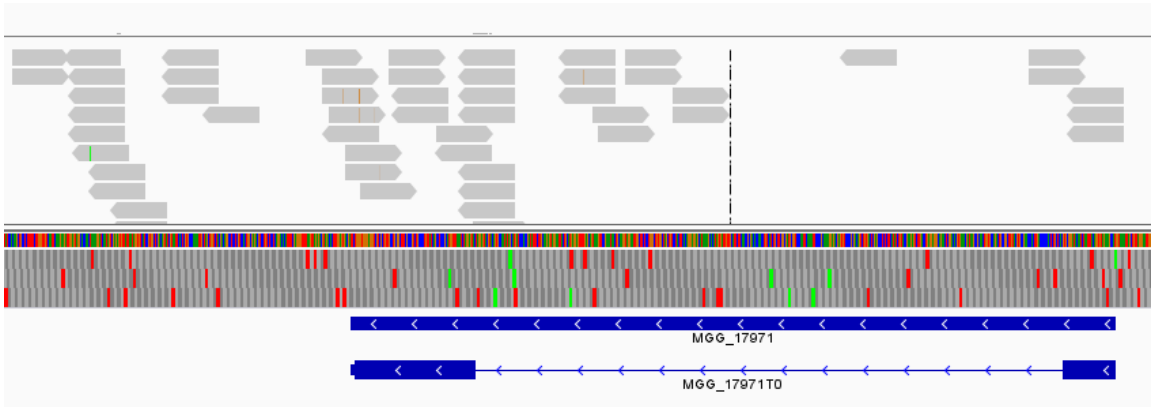
- 9918 genes are expressed in every condition (2D4 only)
- 1964 genes are never expressed (2D4 only)

Not every expressed gene has a recognizable poly(A) site

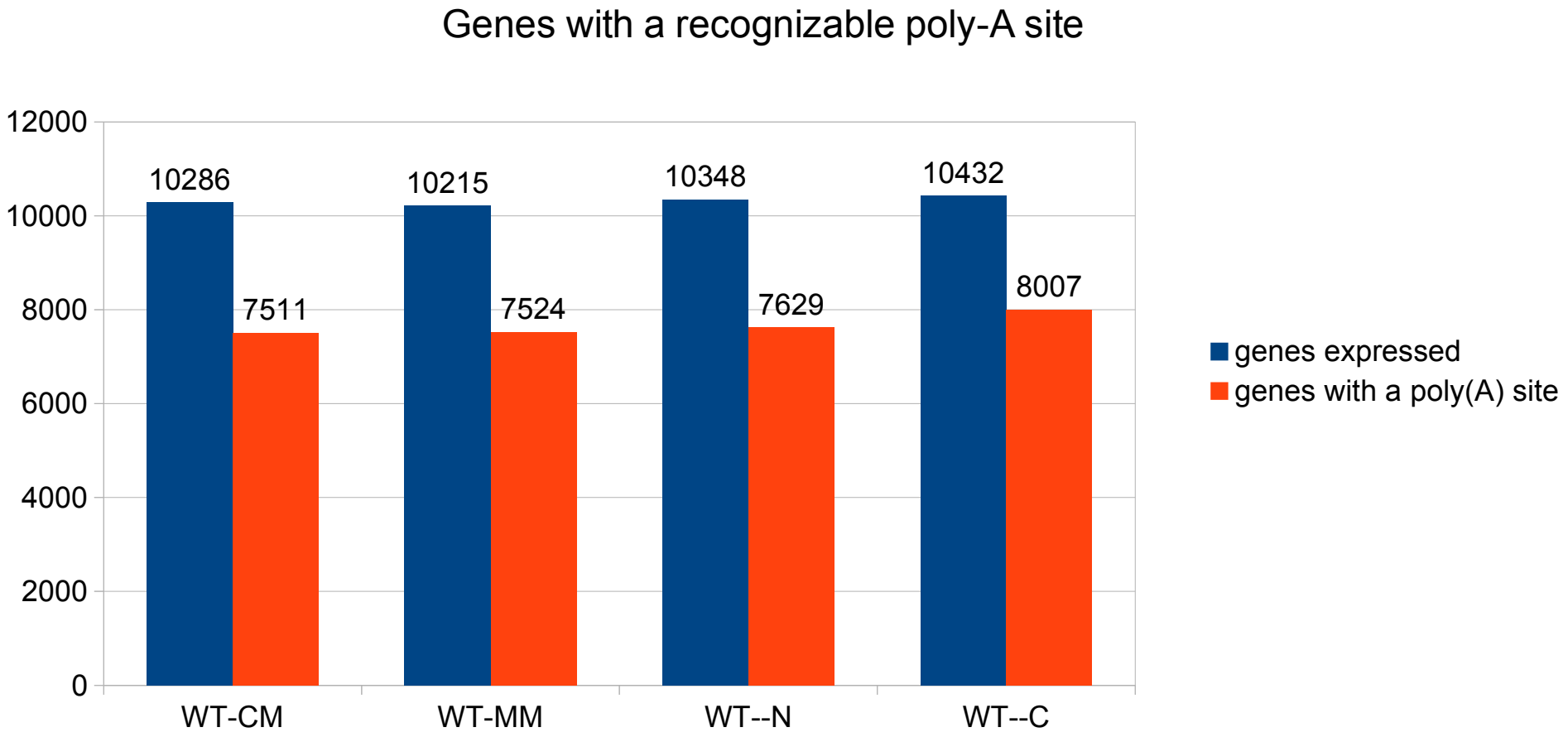
Expressed gene <u>with</u> a recognizable poly(A) site:



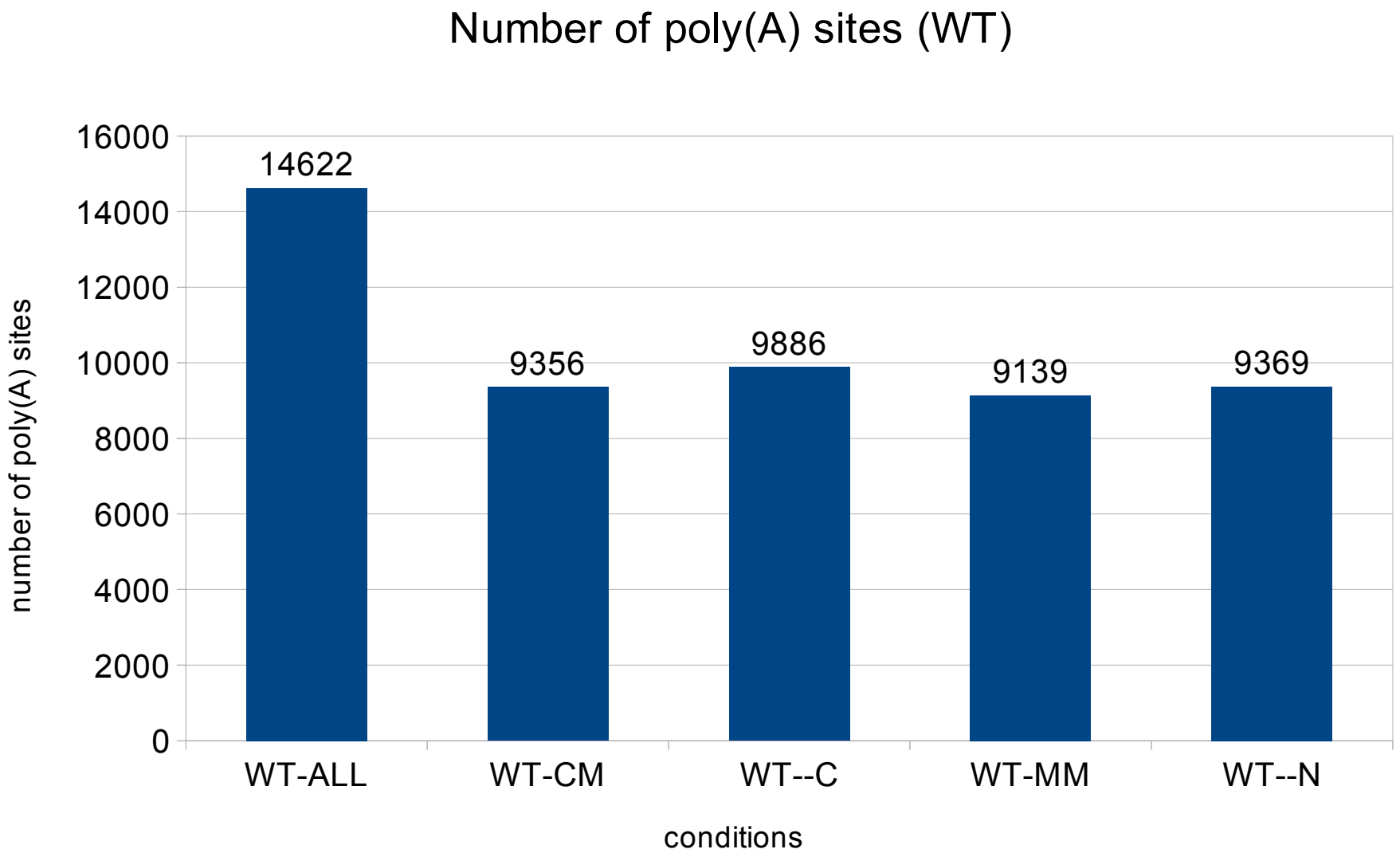
Expressed gene <u>without</u> a recognizable poly(A) site:



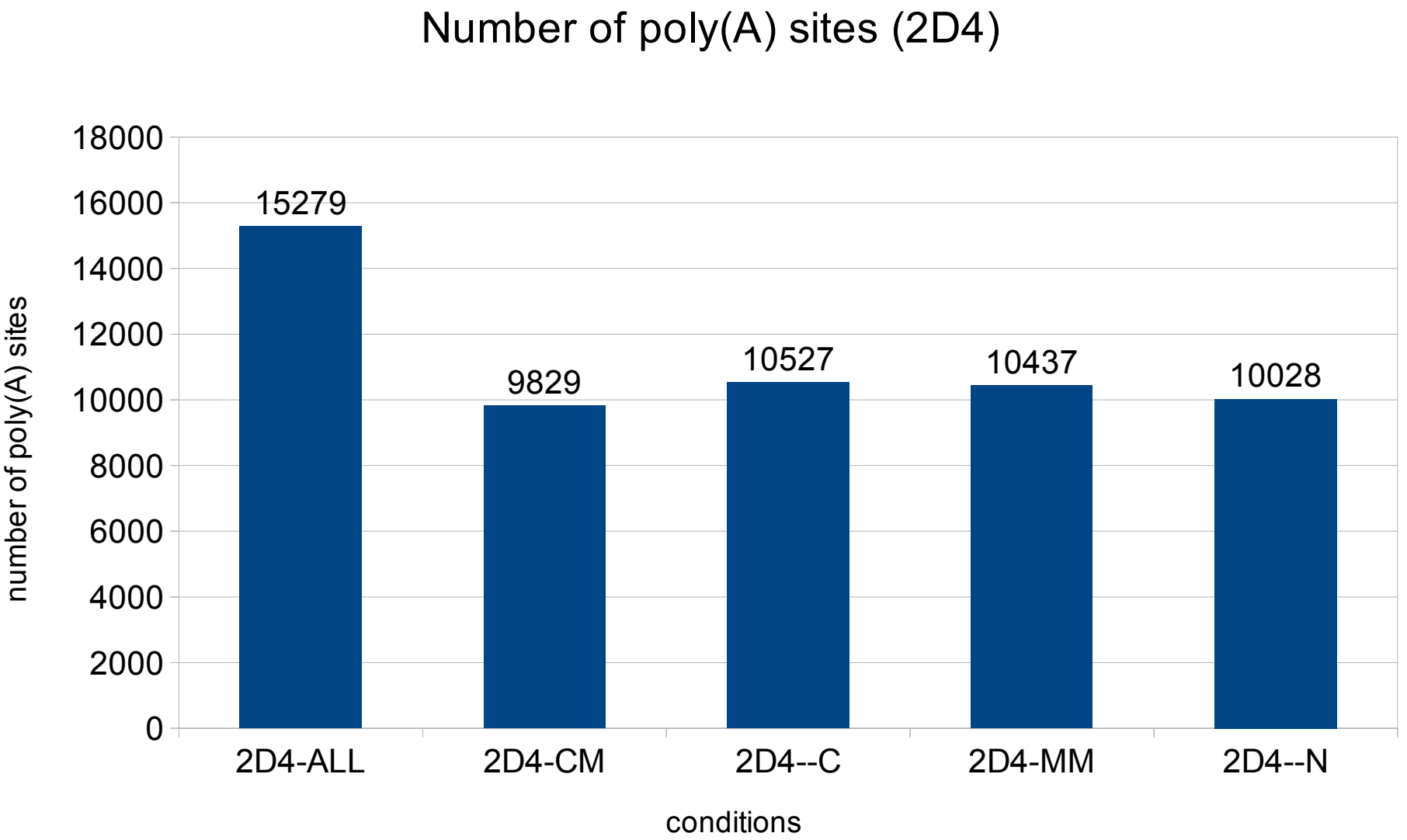
Only ~75% of genes expressed have a recognizable poly(A) site



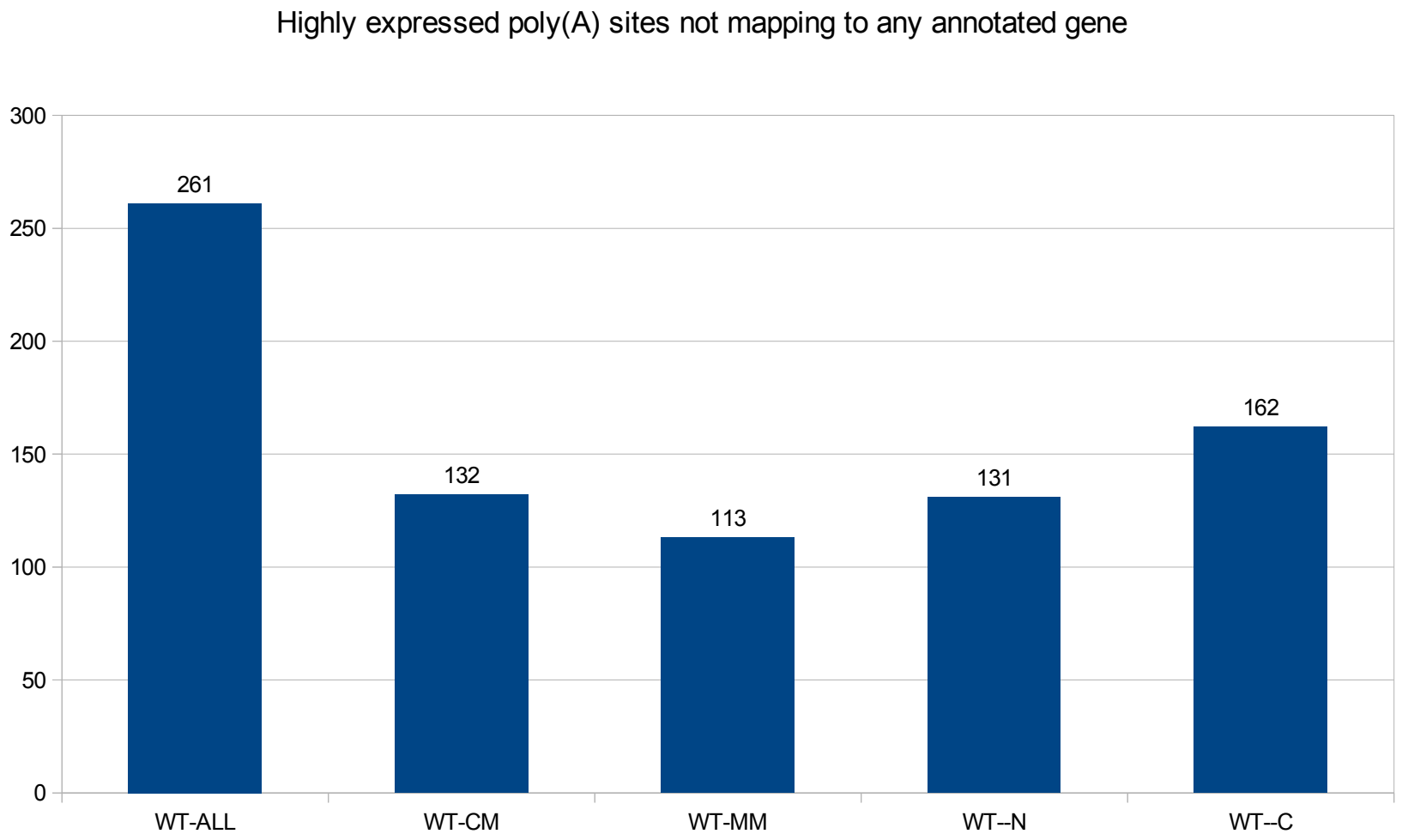
~15000 poly(A) site could be assigned to annotated genes



~15000 poly(A) site could be assigned to annotated genes (2D4)



261 highly expressed (>100 reads) poly(A) sites could not be assigned to any annotated gene



261 orphan poly(A) sites highly expressed in WT (>100 reads)

- 14 hits against other gene copies in *M.oryzae*
- 44 hits against Uniprot nt/nr database
- 4 hits against Rfam(ncRNA) database
- 82 overlapping annotated genes antisense
- 63 matching CPA-sRNA sequences
- 4 matching telomeric regions (TelRNAs?)
- 4 matching retrotransposons
- 2 matching AVR genes

3165 orphan poly(A) sites expressed in WT (>10 reads)

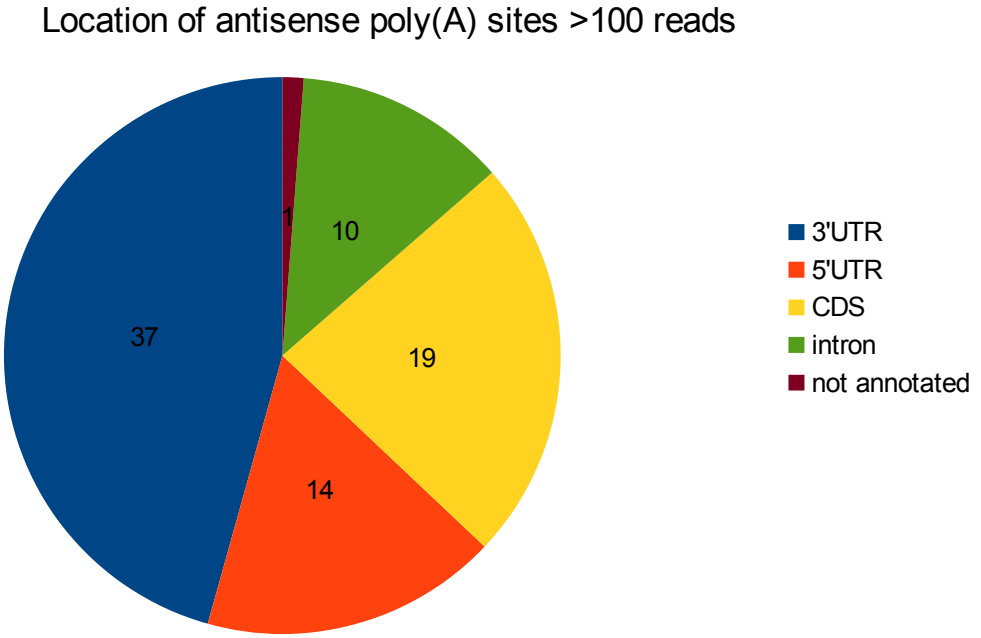
- 102 hits against other gene copies in *M.oryzae*
- 438 hits against Uniprot nt/nr database
- 12 hits against Rfam(ncRNA) database
- 1098 overlapping annotated genes antisense
- 253 matching CPA-sRNA sequences
- 56 matching telomeric regions (TelRNAs?)
- 27 matching retrotransposons
- 27 matching AVR genes

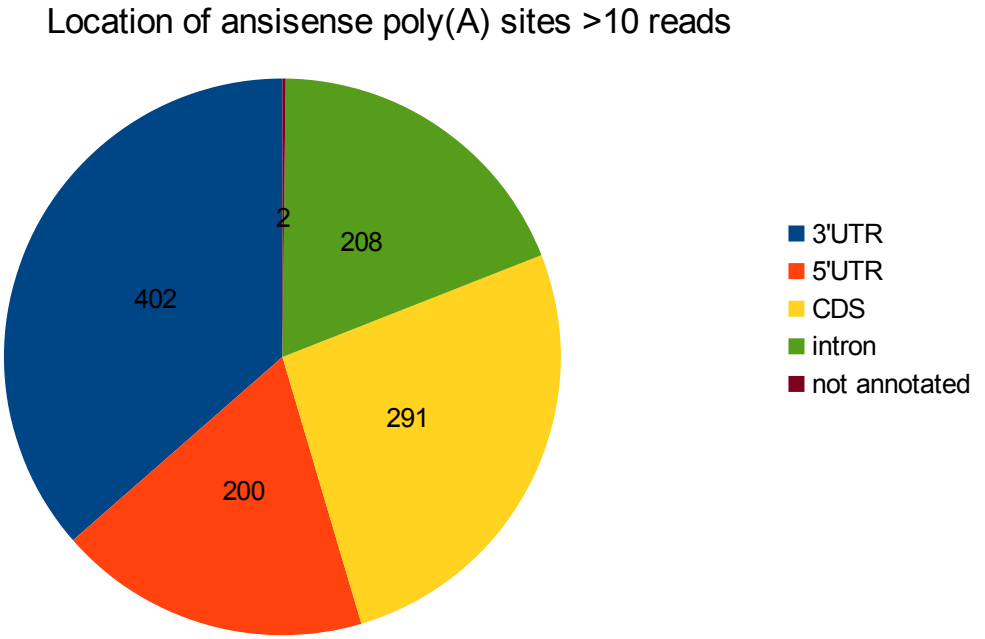
Orphans differentially expressed

(>100 reads)	
CM -> -C	146
CM -> MM	27
CM -> -N	42
MM -> -C	106
MM -> -N	3
-N -> -C	87

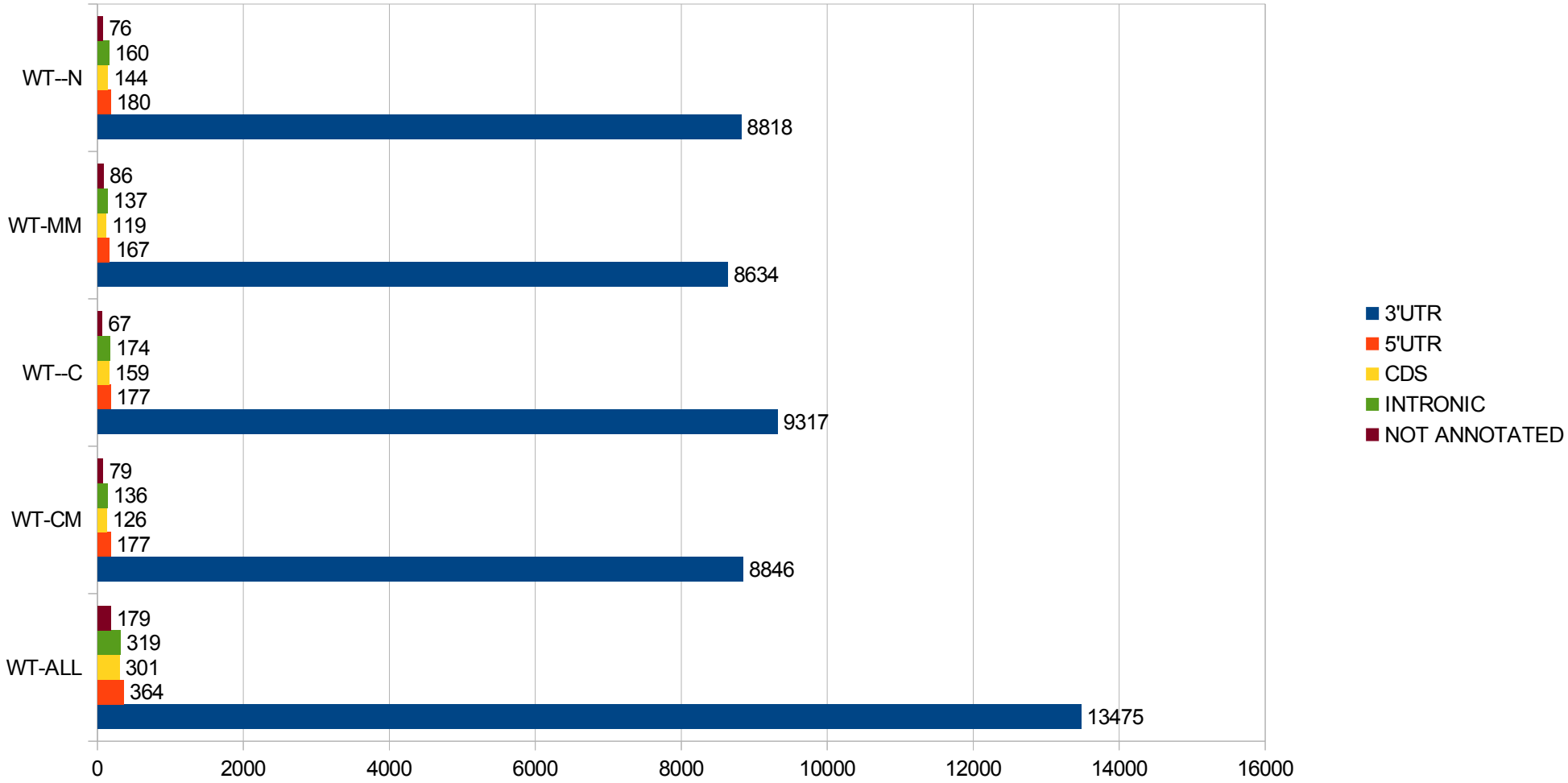
(>10 reads)	
CM -> -C	712
CM -> MM	134
CM -> -N	201
MM -> -C	506
MM -> -N	20
-N -> -C	380

Antisense poly(A) are usually located in the 3'UTR

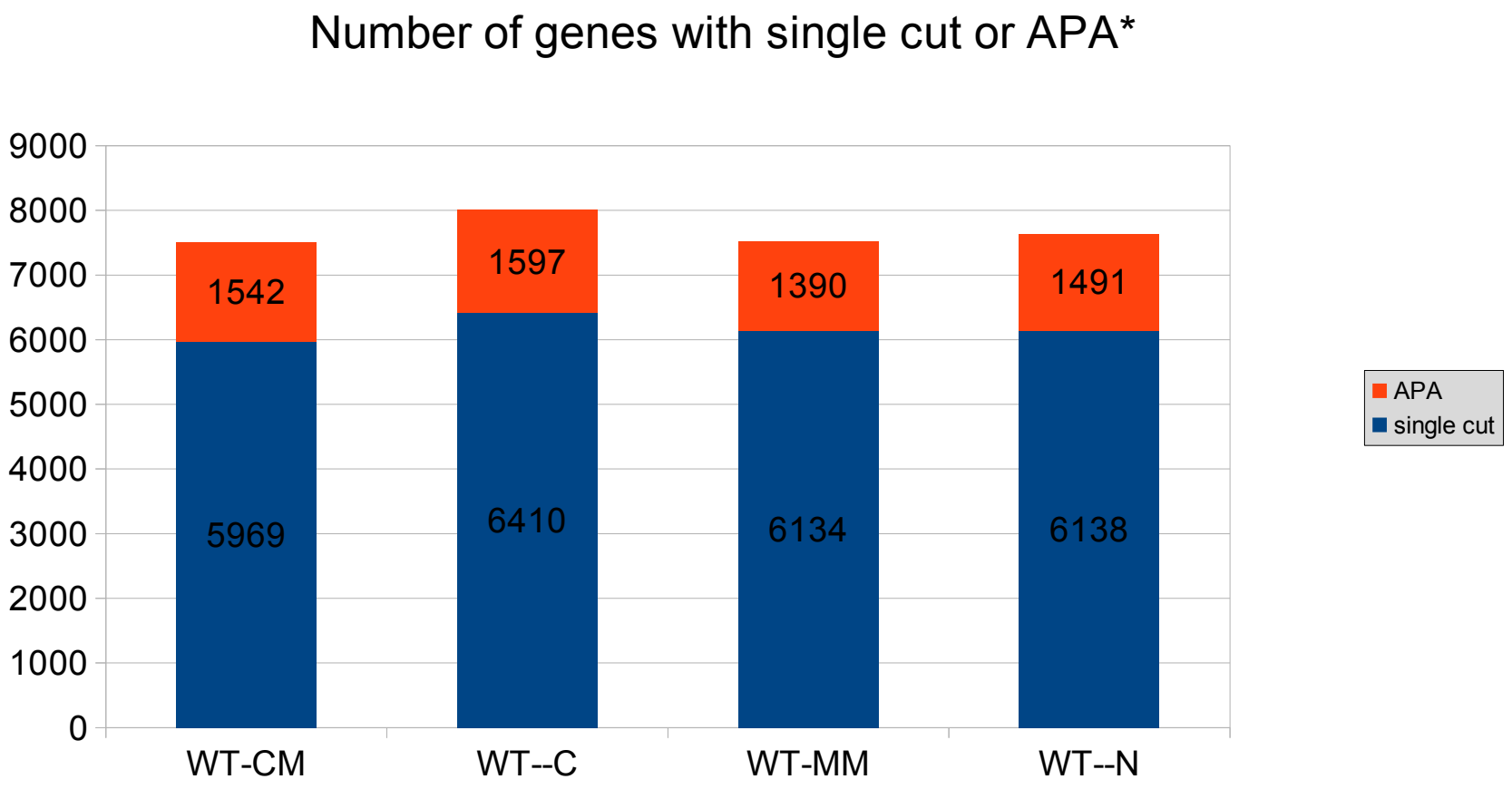




>90% of poly(A) sites are located in the 3'UTR

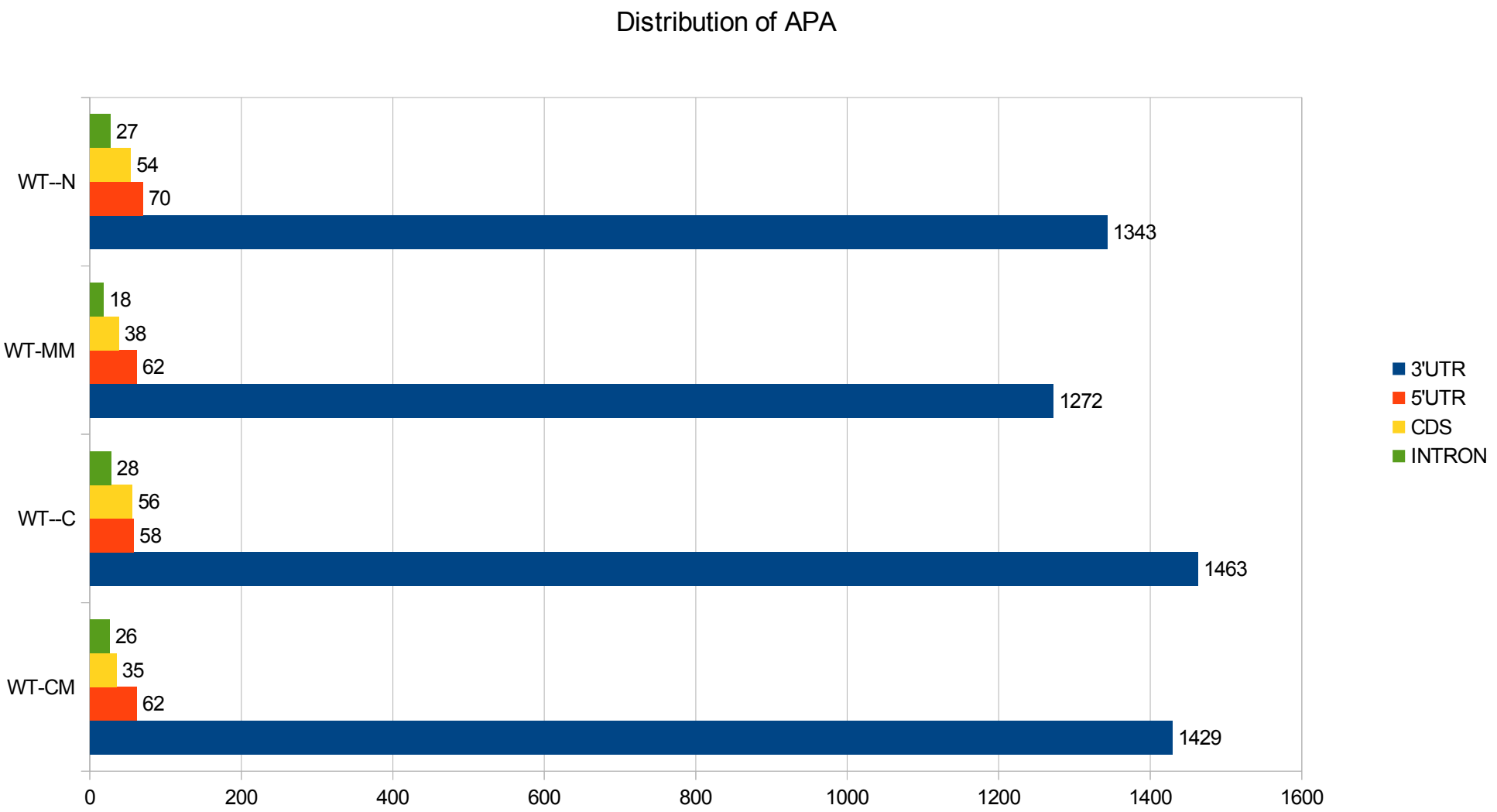


~20% of genes are alternatively polyadenilated

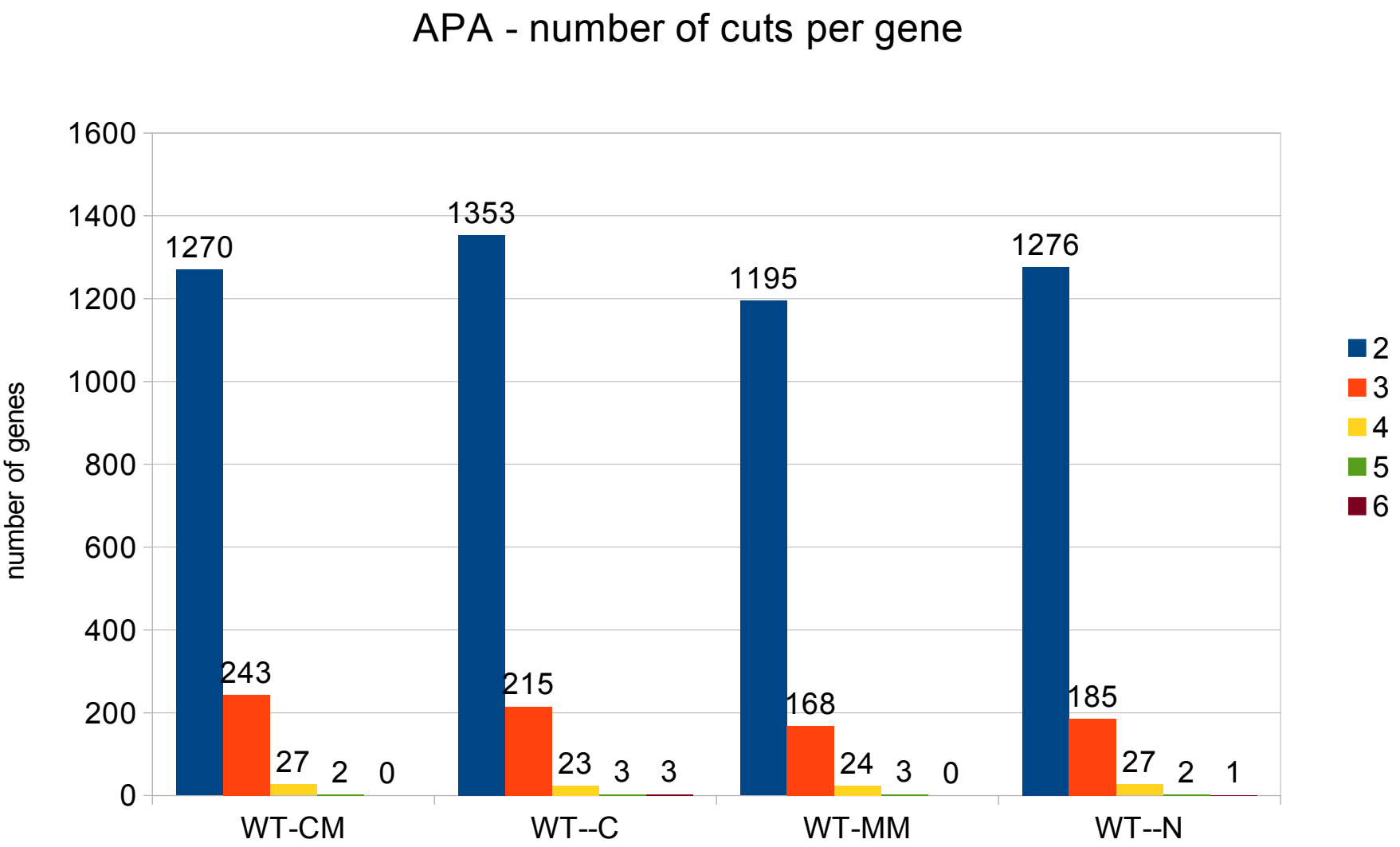


** calculated over the global number of expressed genes with a recognizable poly(A) site*

>90% of APA is located in the 3'UTR



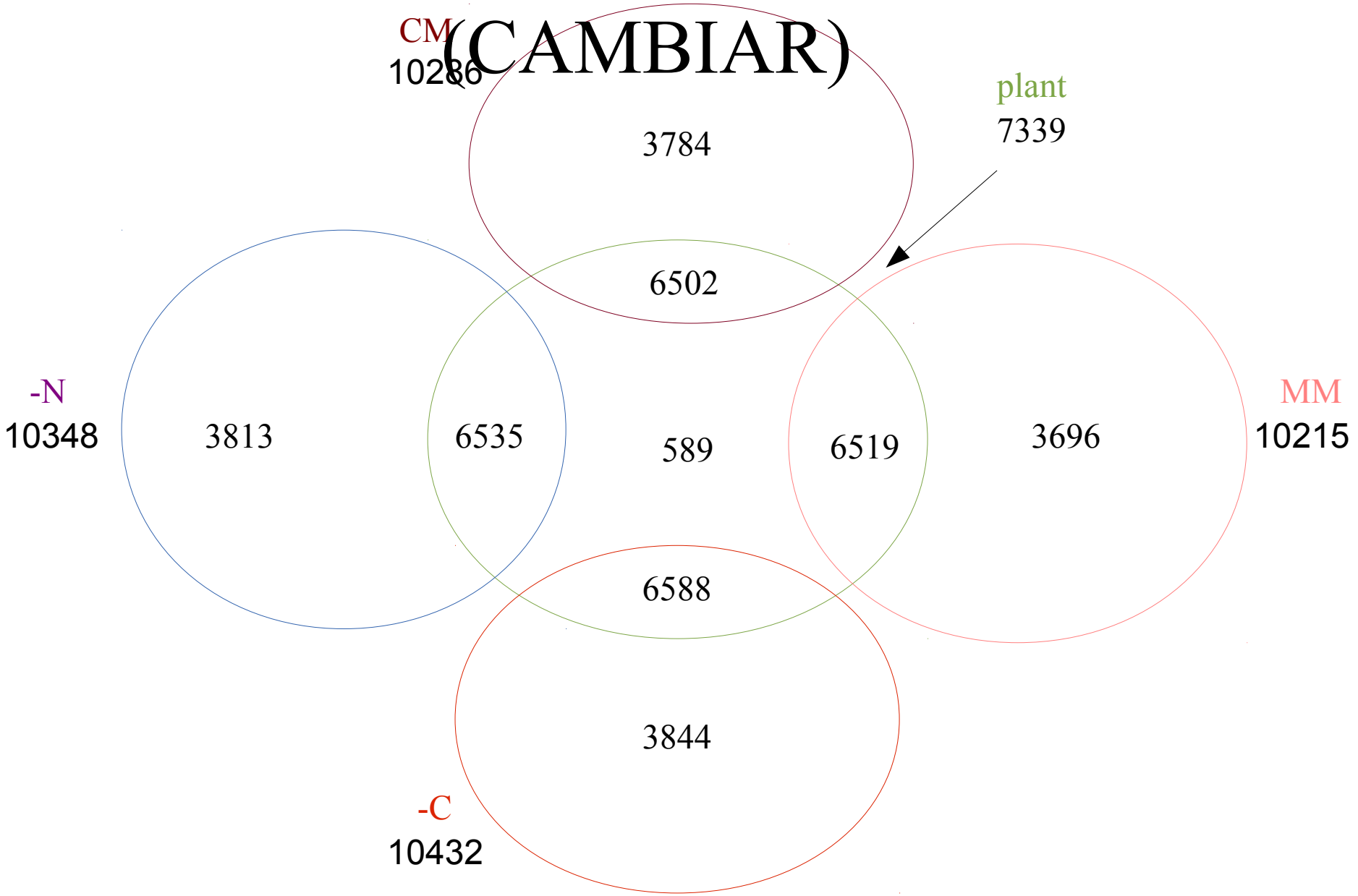
>80% of APA is composed of two cleveage sites



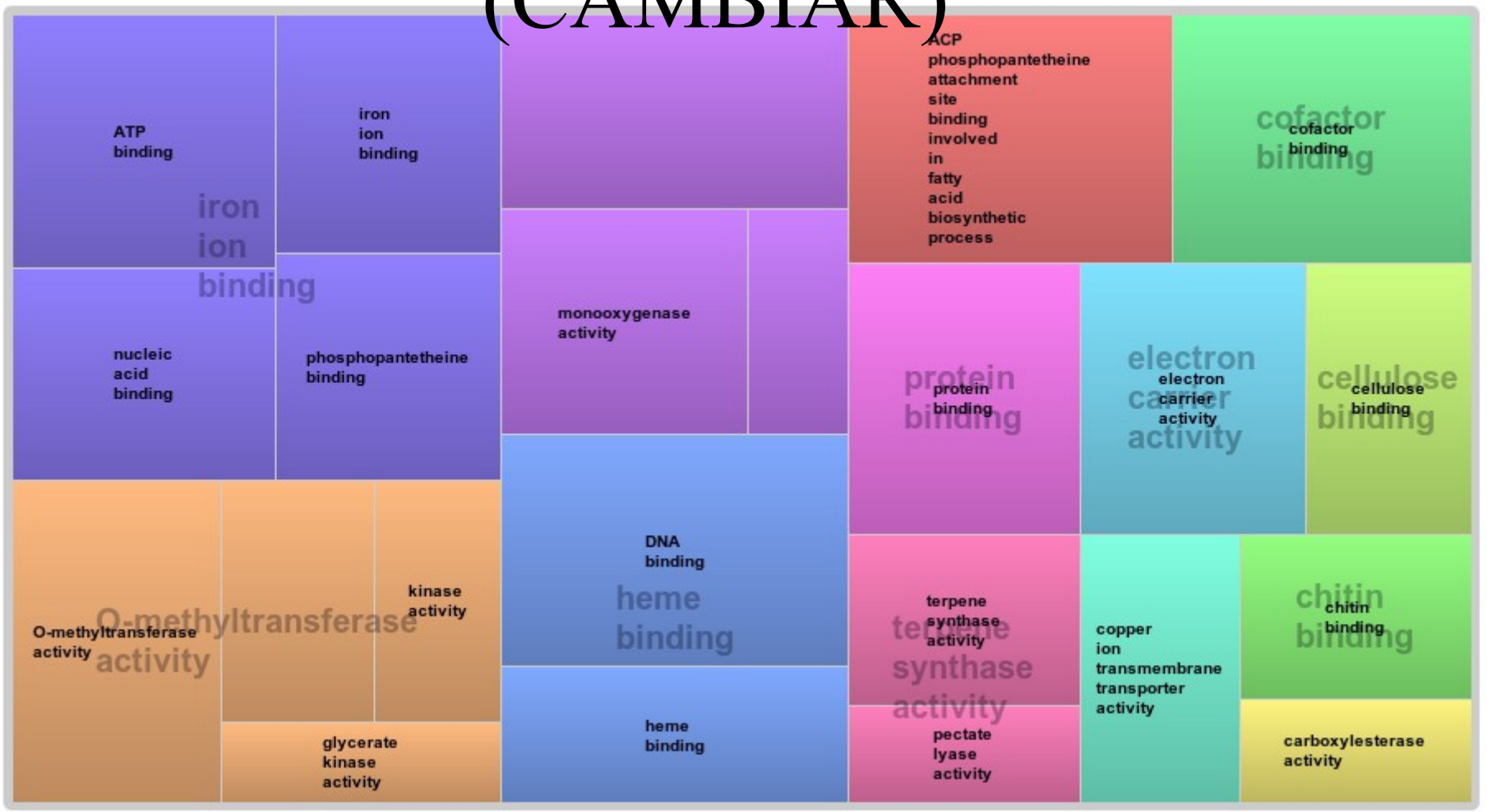
Gene expression between WT conditions

589 genes expressed in plant are never expressed in vitro

(CAMBIAR)



Plant specific genes functional analysis (CAMBIAR)

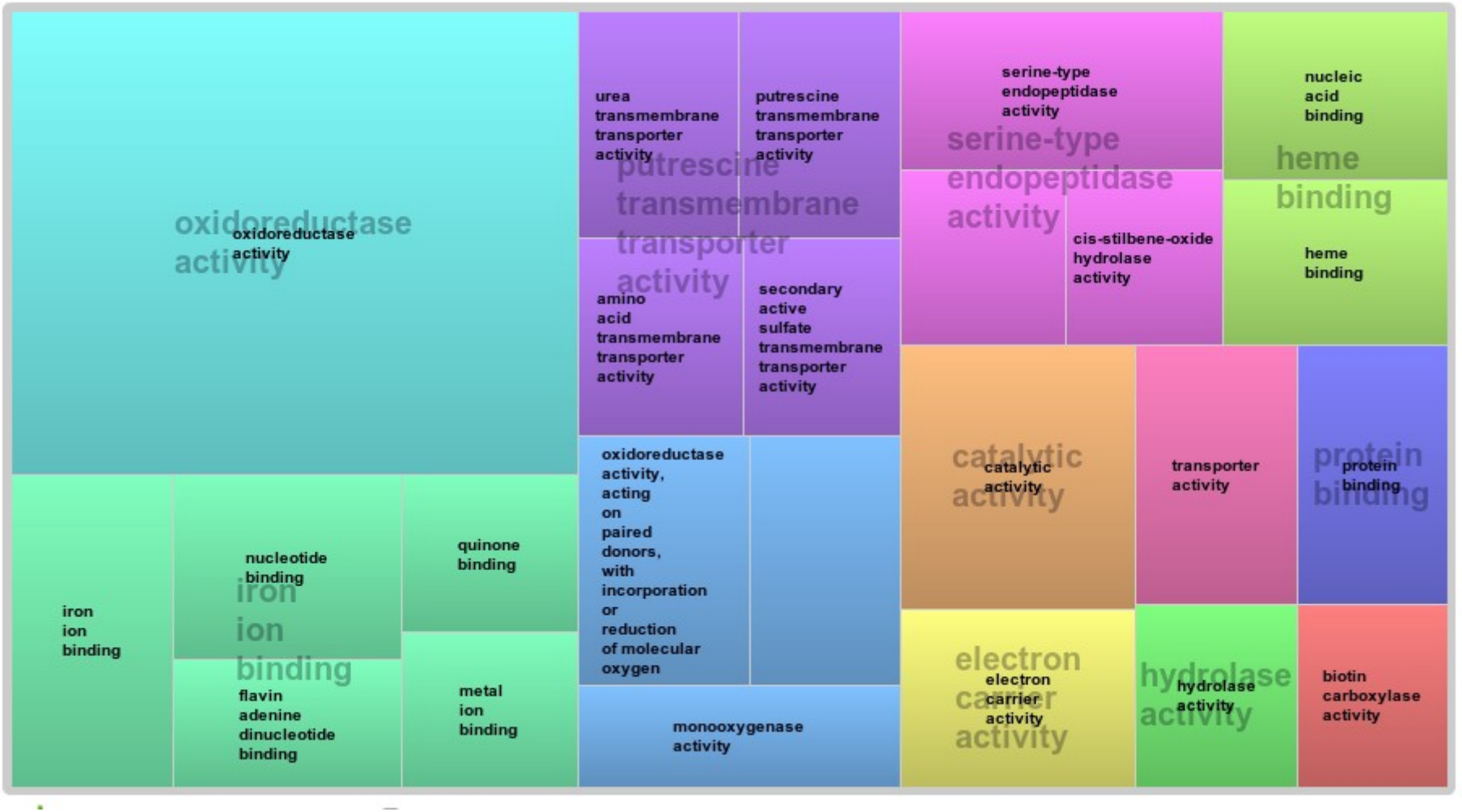


p-value = 0.05

The CM → -C condition presents the highest number of differentially expressed genes, while MM → -N the lowest

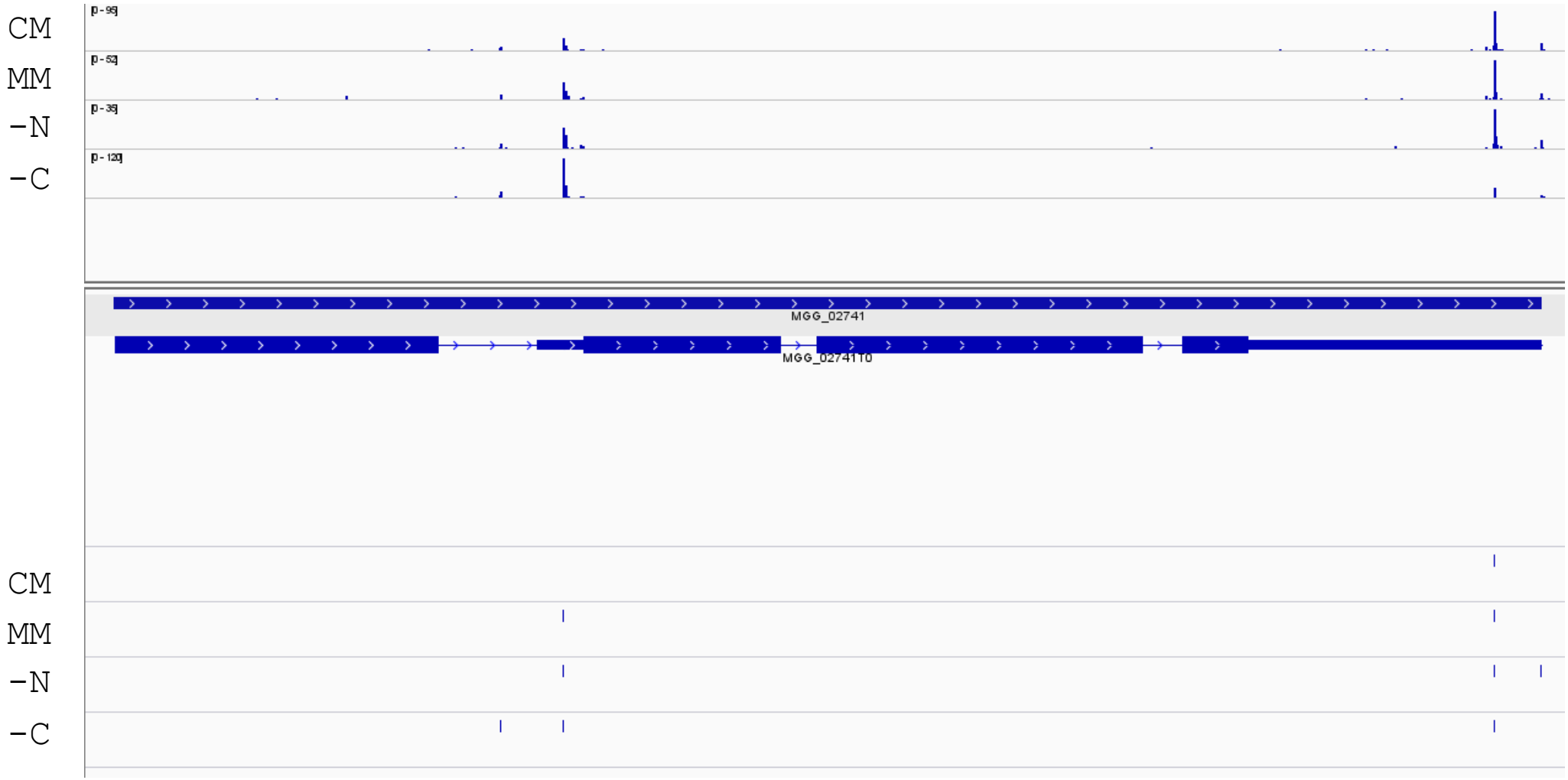
DIFFERENTIALLY EXPRESSED GENES IN THE WT			
	DOWN	UP	TOTAL
CM → MM-C	1586	1617	3203
CM → MM	151	285	436
CM → MM-N	177	347	524
MM → MM-C	601	608	1209
MM → MM-N	1	5	6
MM-N → MM-C	216	313	529

CM → MM functional analysis



p-value = 0.1

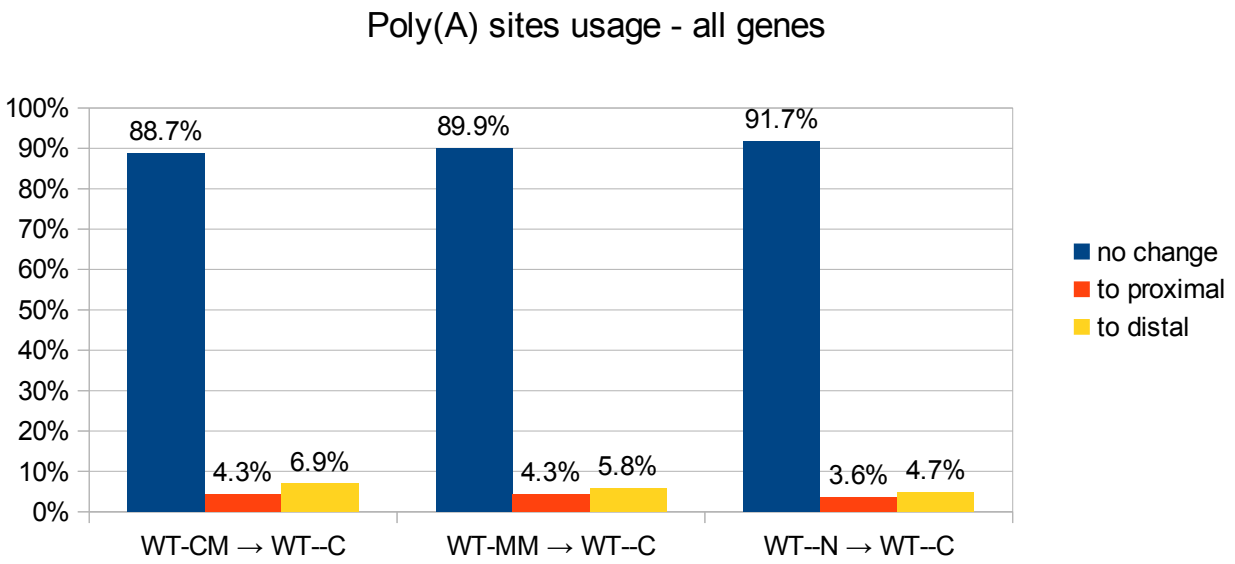
rbp35 shows different polyadenilation in each medium, with strong differencies in MM-C

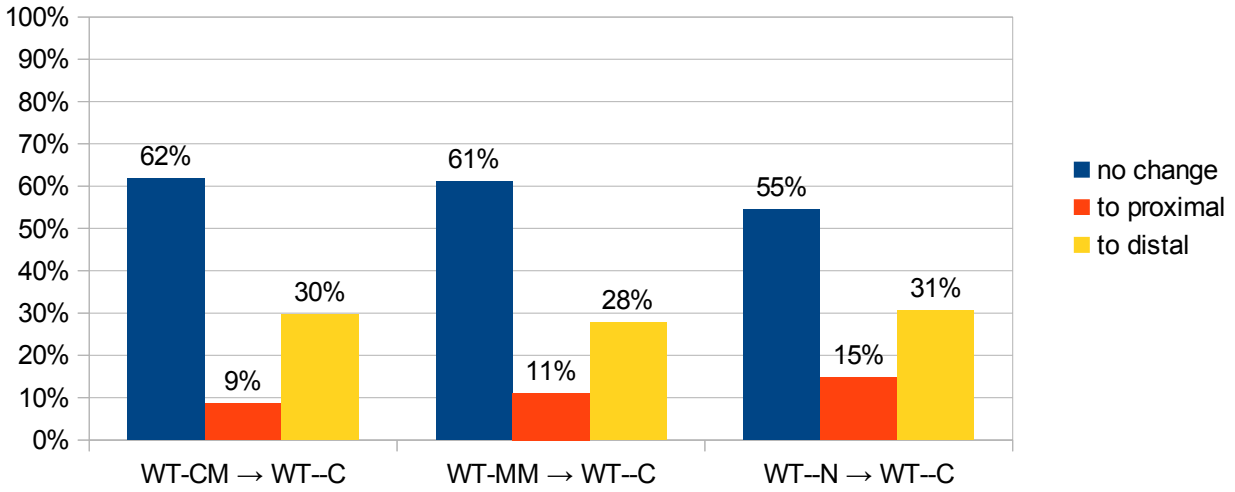


MM-C affects a great number of poly(A) sites

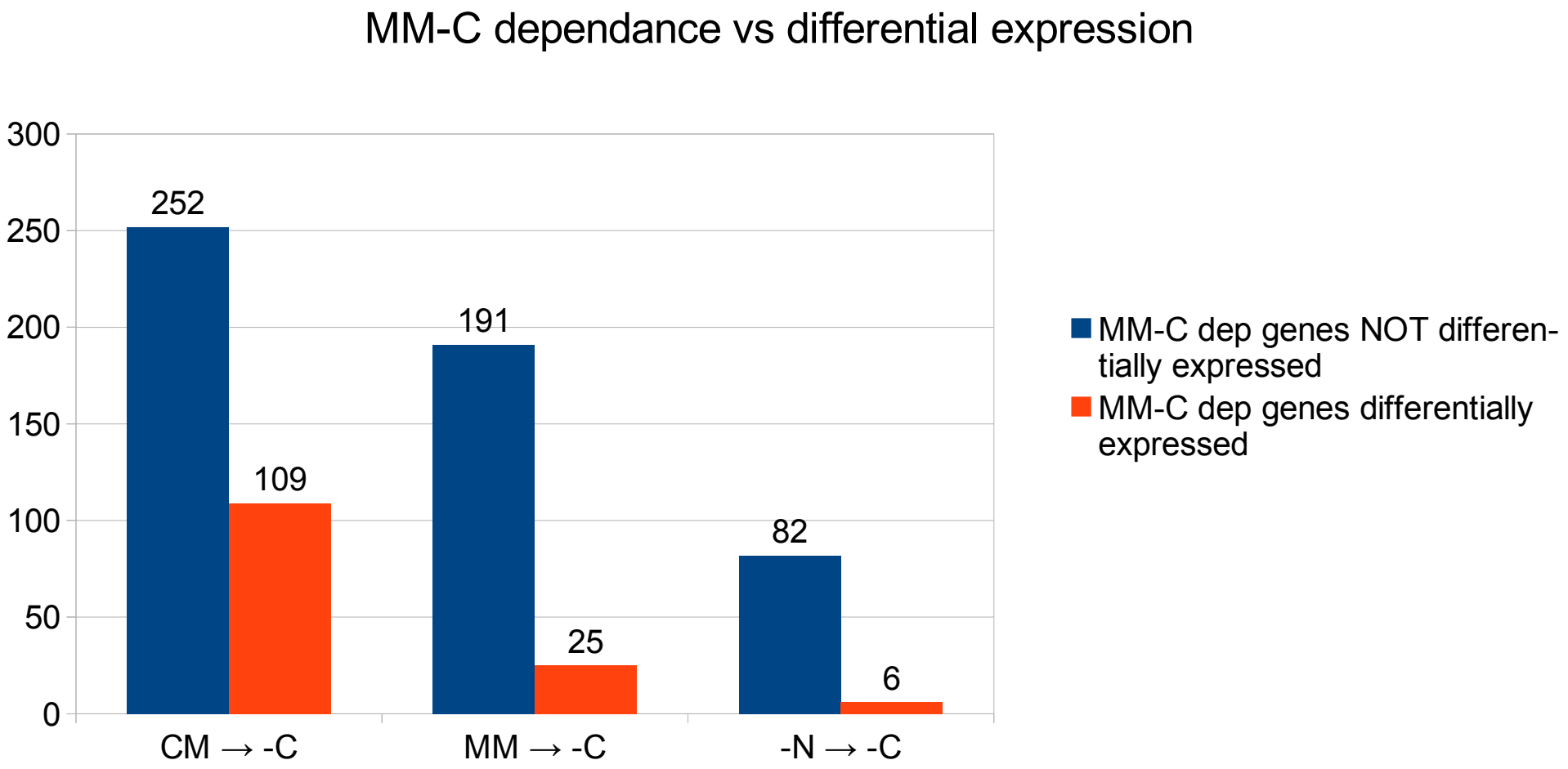
Medium change	poly(A) sites affected
CM → MM	6
CM → -N	9
CM → -C	**361**
MM → -N	0
MM → -C	**216**
-N → -C	**88**

MM-C affects poly(A) sites usage, preferring distal cuts

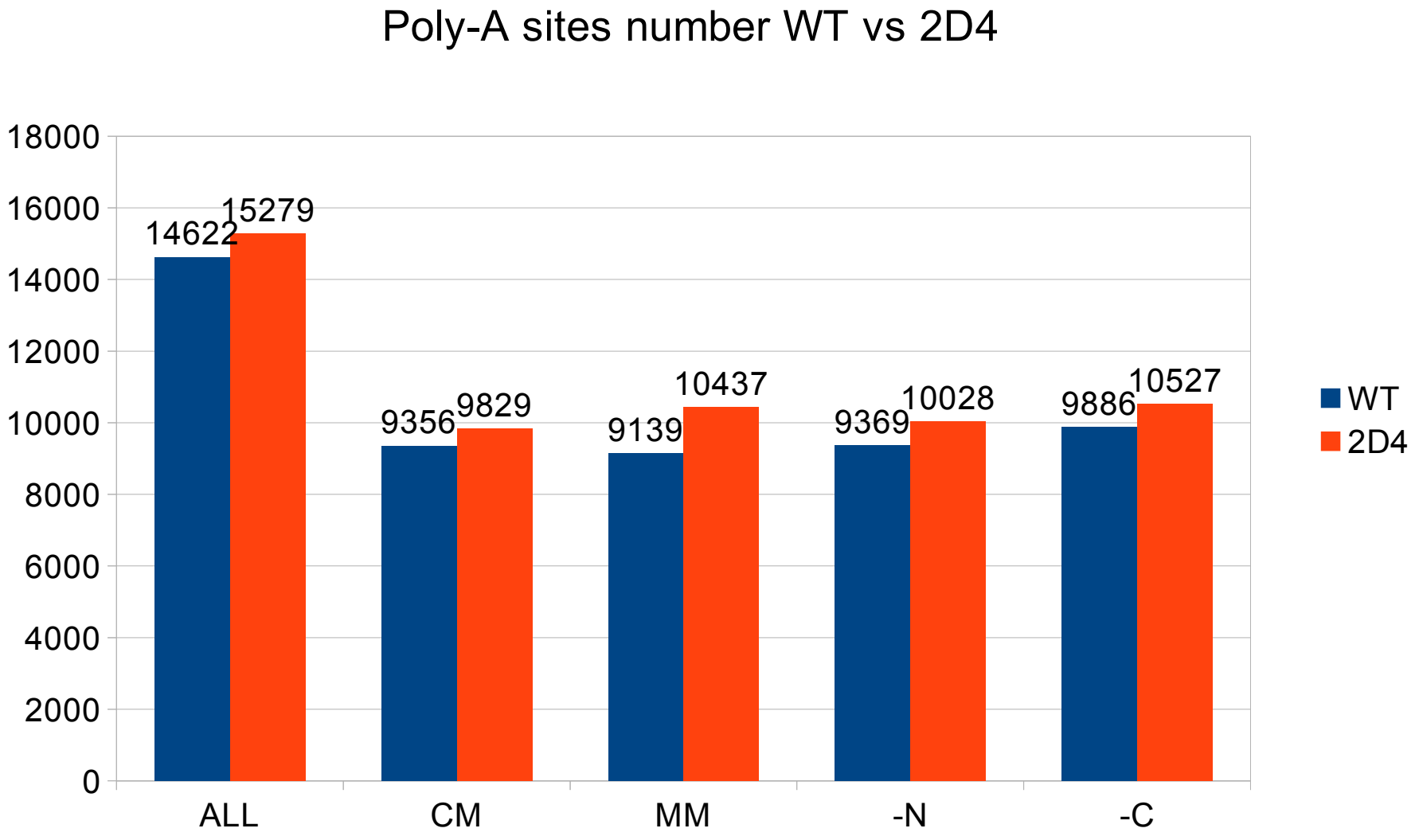




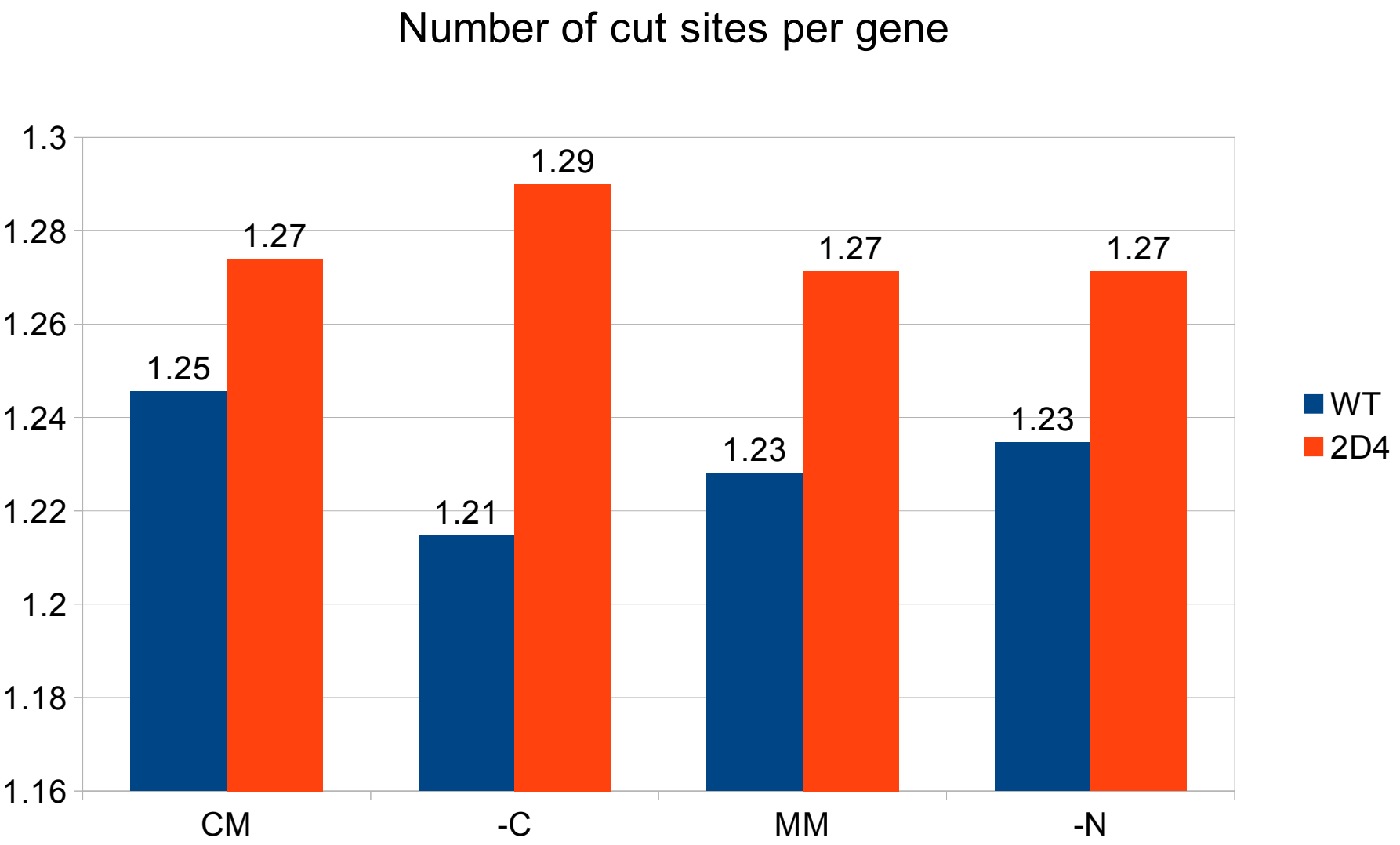
MM-C dependent genes are not always differentially expressed



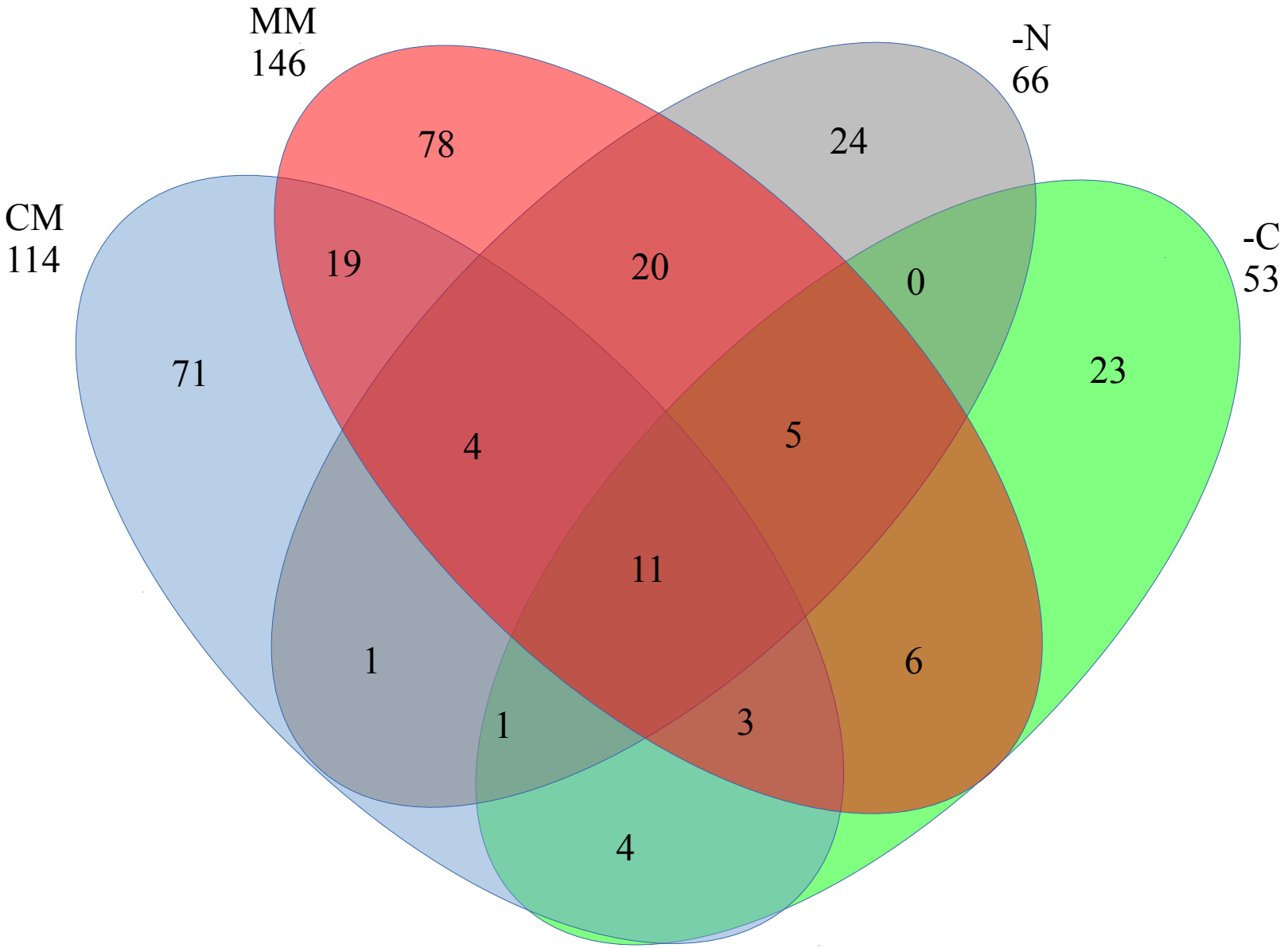
Δrbp35 affects poly(A) sites number



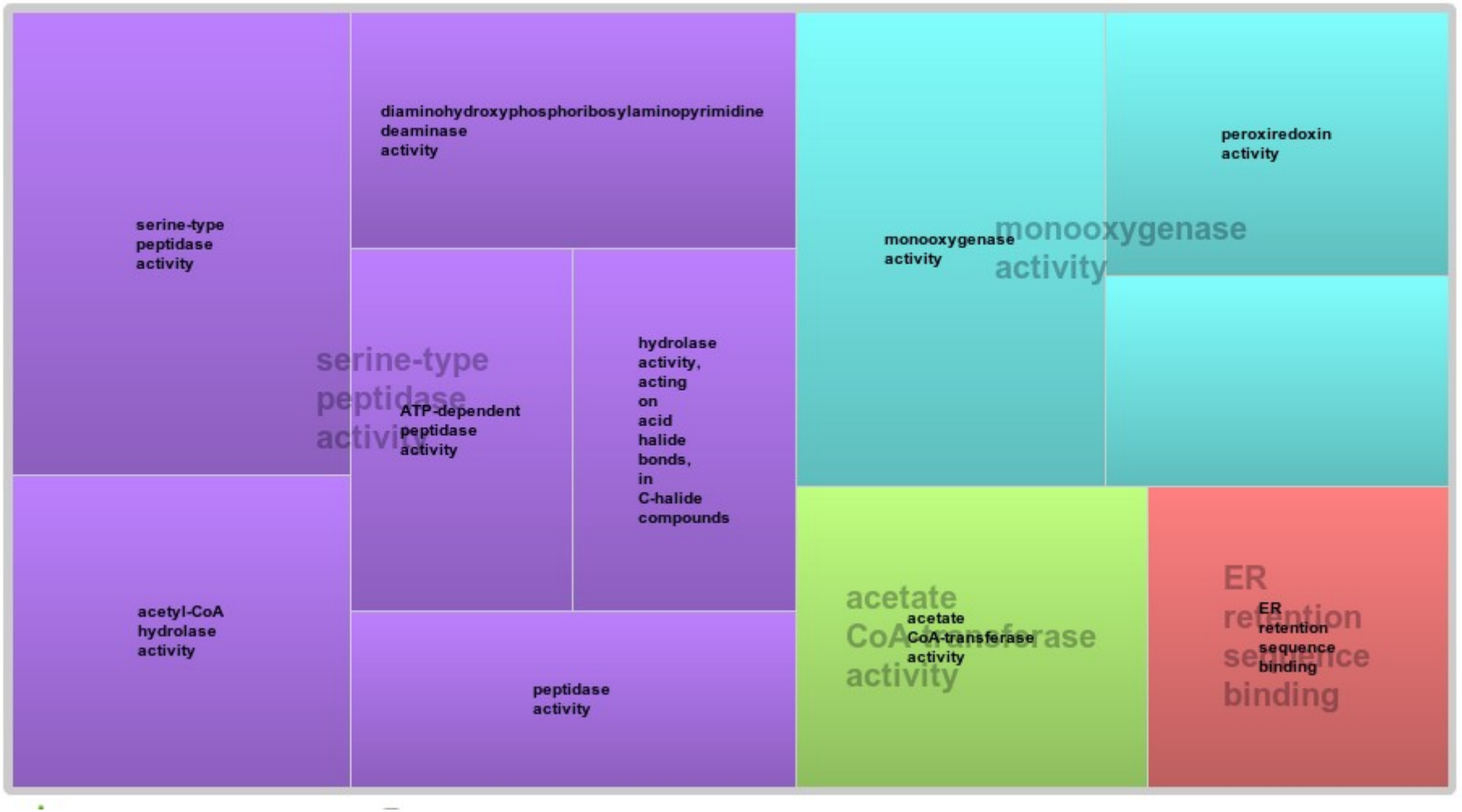
Δrbp35 affects number of cut sites per gene



Differentially expressed genes WT → Δ*rbp35*



WT → 2D4 functional analysis



p-value = 0.05

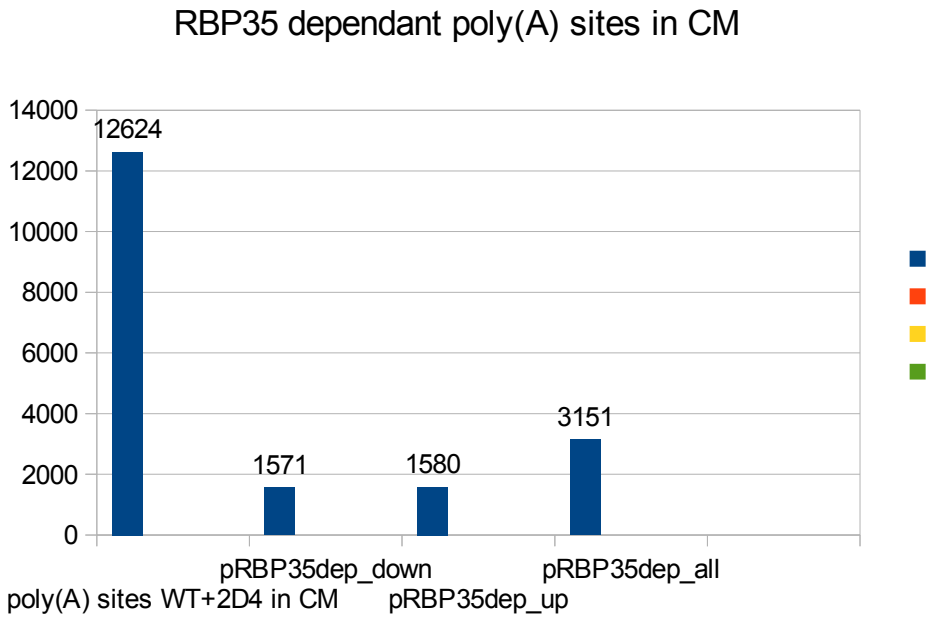
Terminology

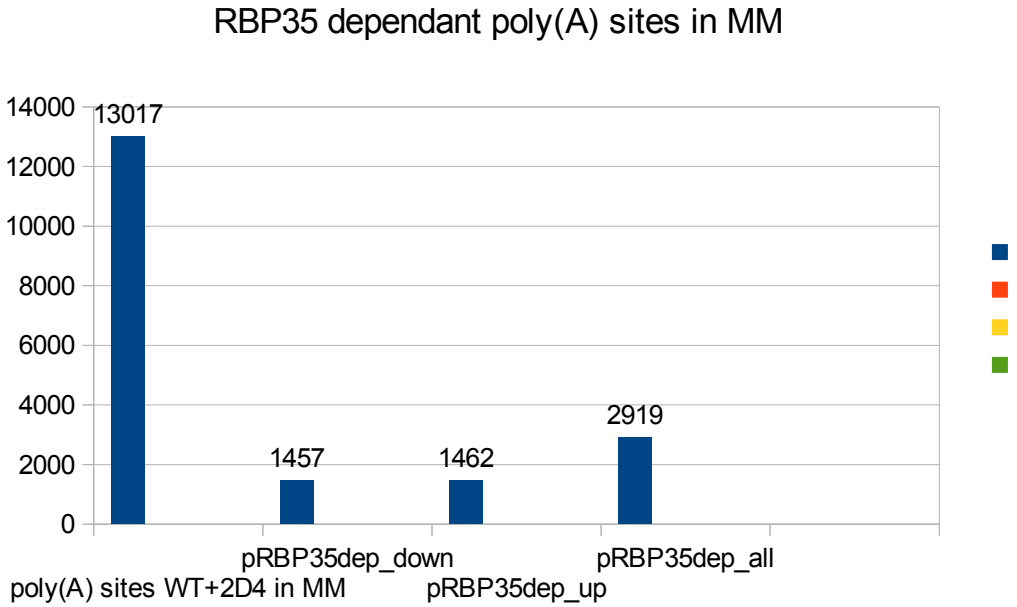
- **pRBP35dep:** poly(A) sites that show a differential expression between wild-type and Δrbp35. We call it “RBP35 dependent poly(A) sites”
- **pRBP35dep_down**: a down-regulated RBP35 dependent poly(A):

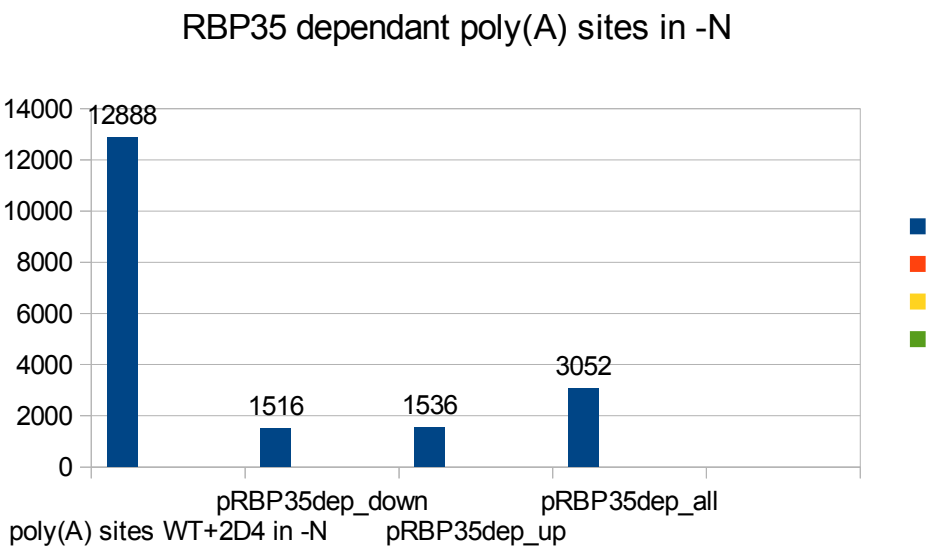
- **pRBP35dep_up**: an up-regulated RBP35 dependent poly(A):

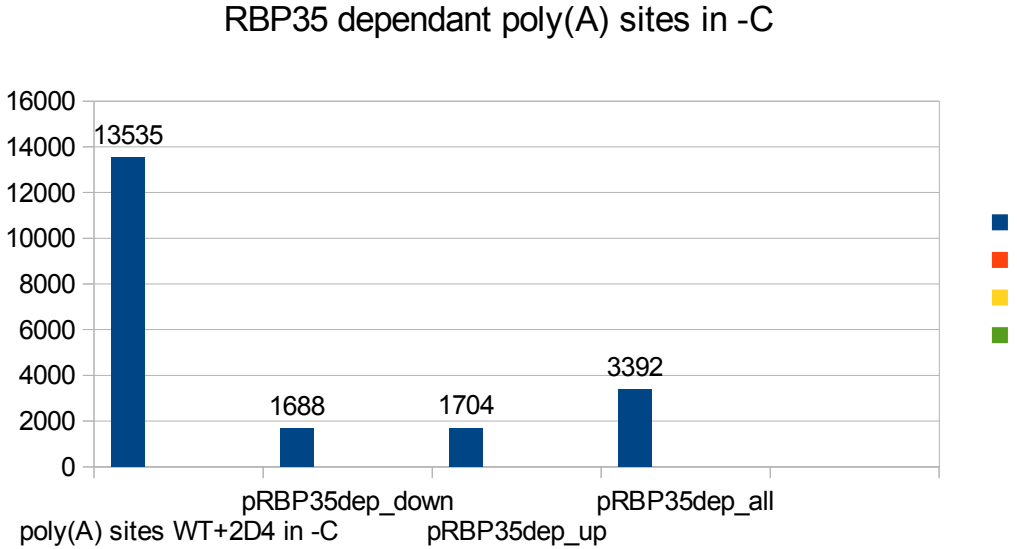
- A gene is defined “RBP35 dependent gene” (or simply **RBP35dep**) when one or more of its poly(A) belong to the previous groups

22%-25% of poly(A) sites are dependent from RBP35 in all media

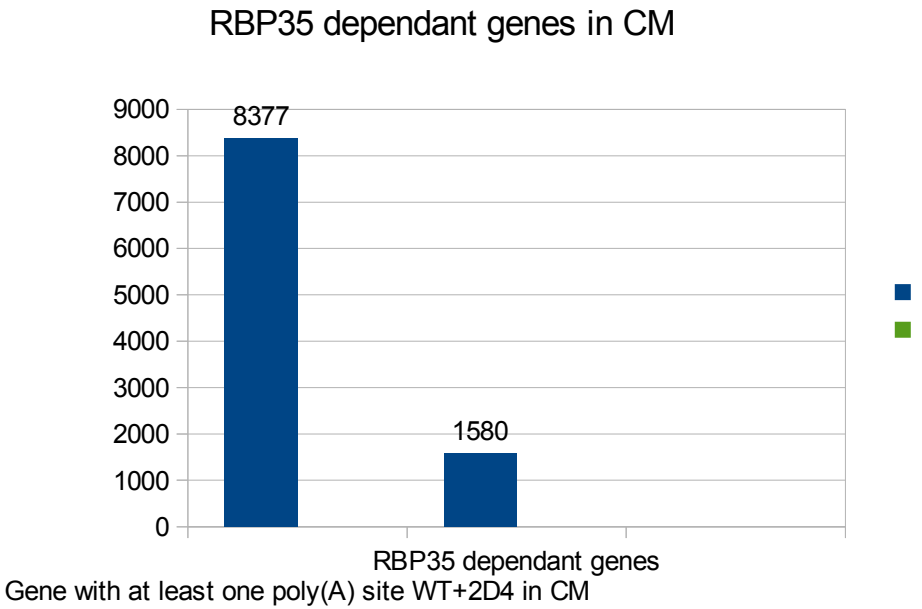


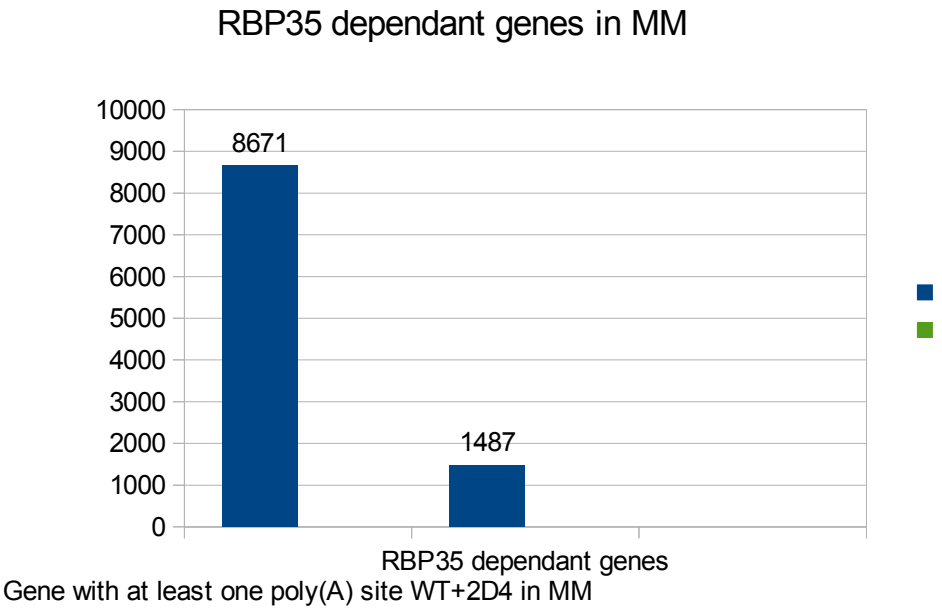


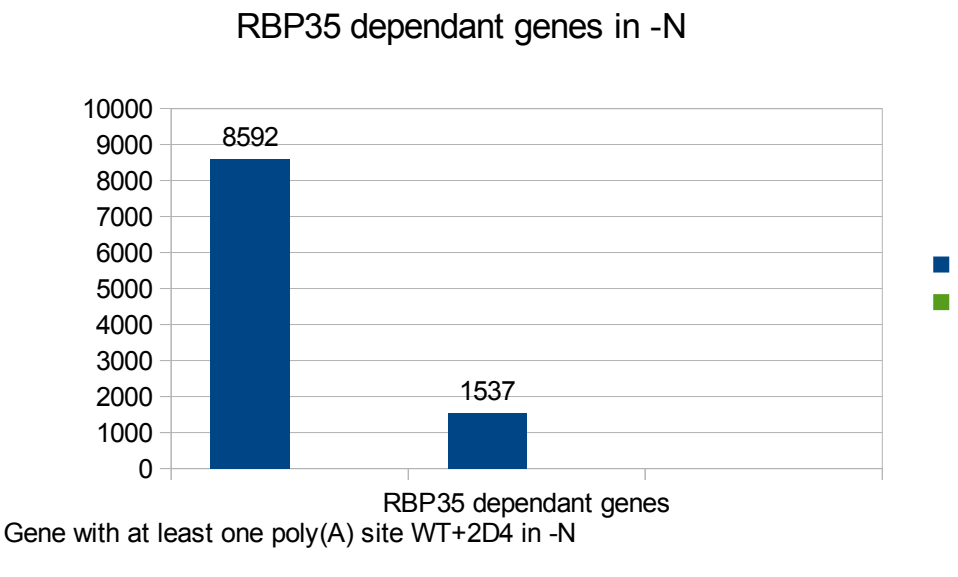


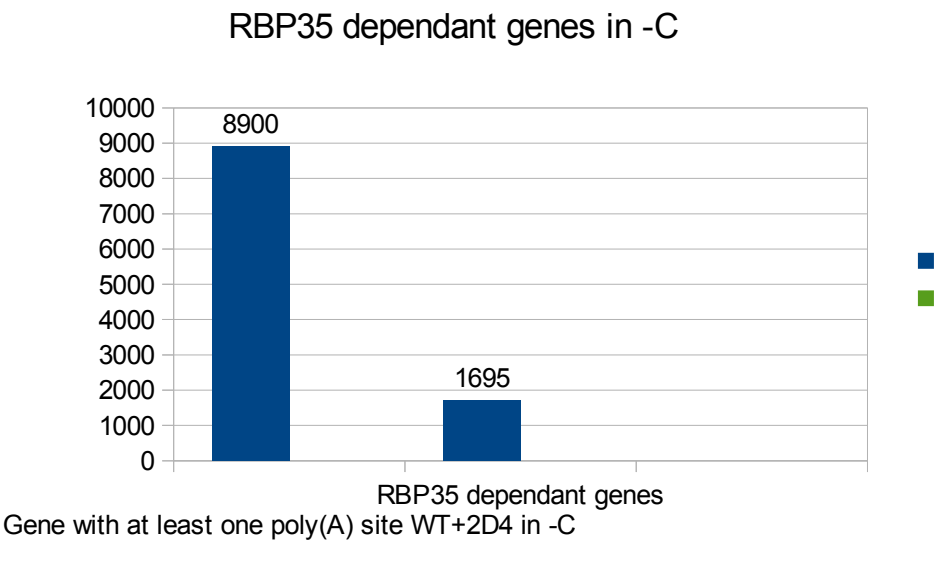


17%-19% of genes are dependent from RBP35 in all media

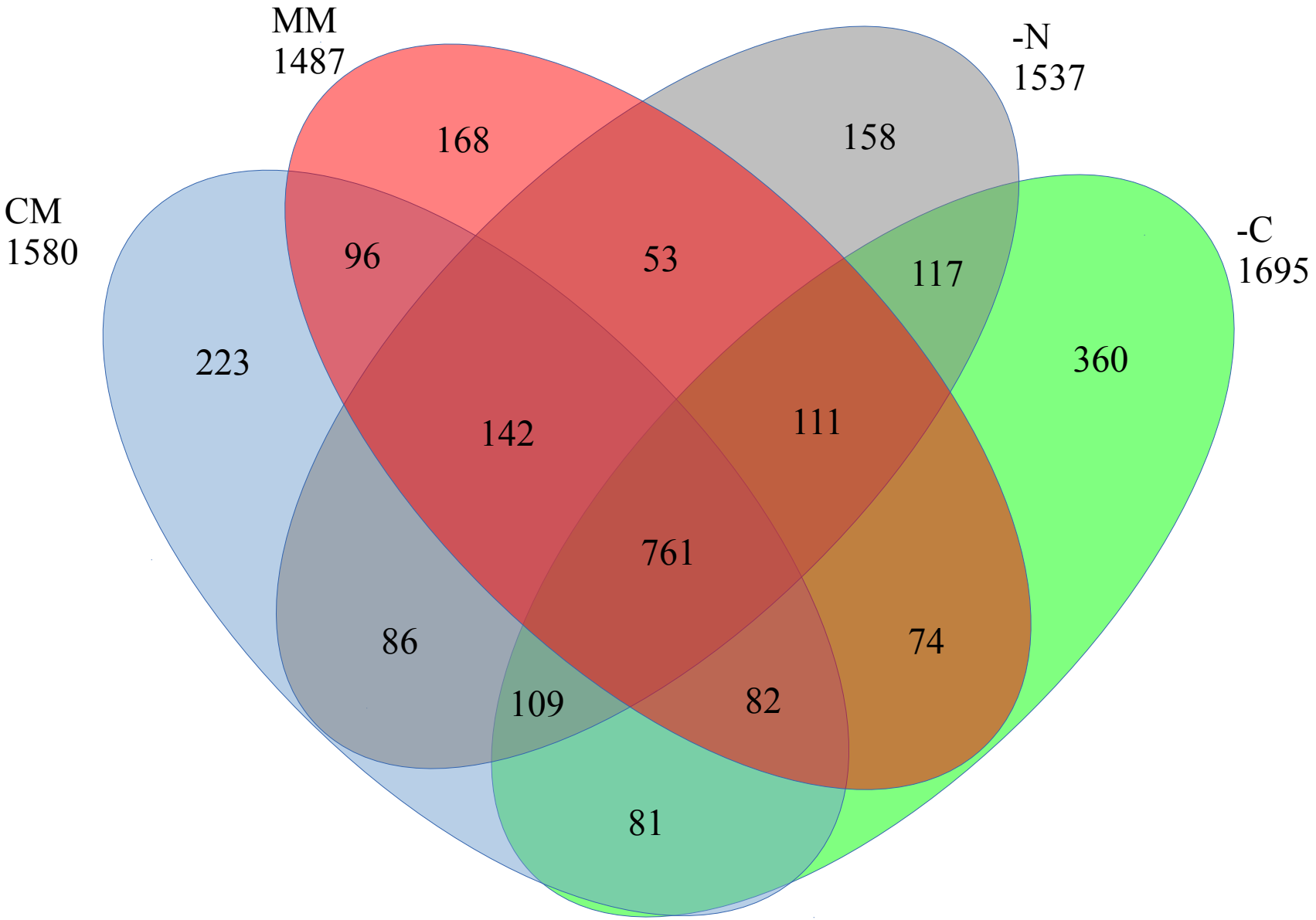




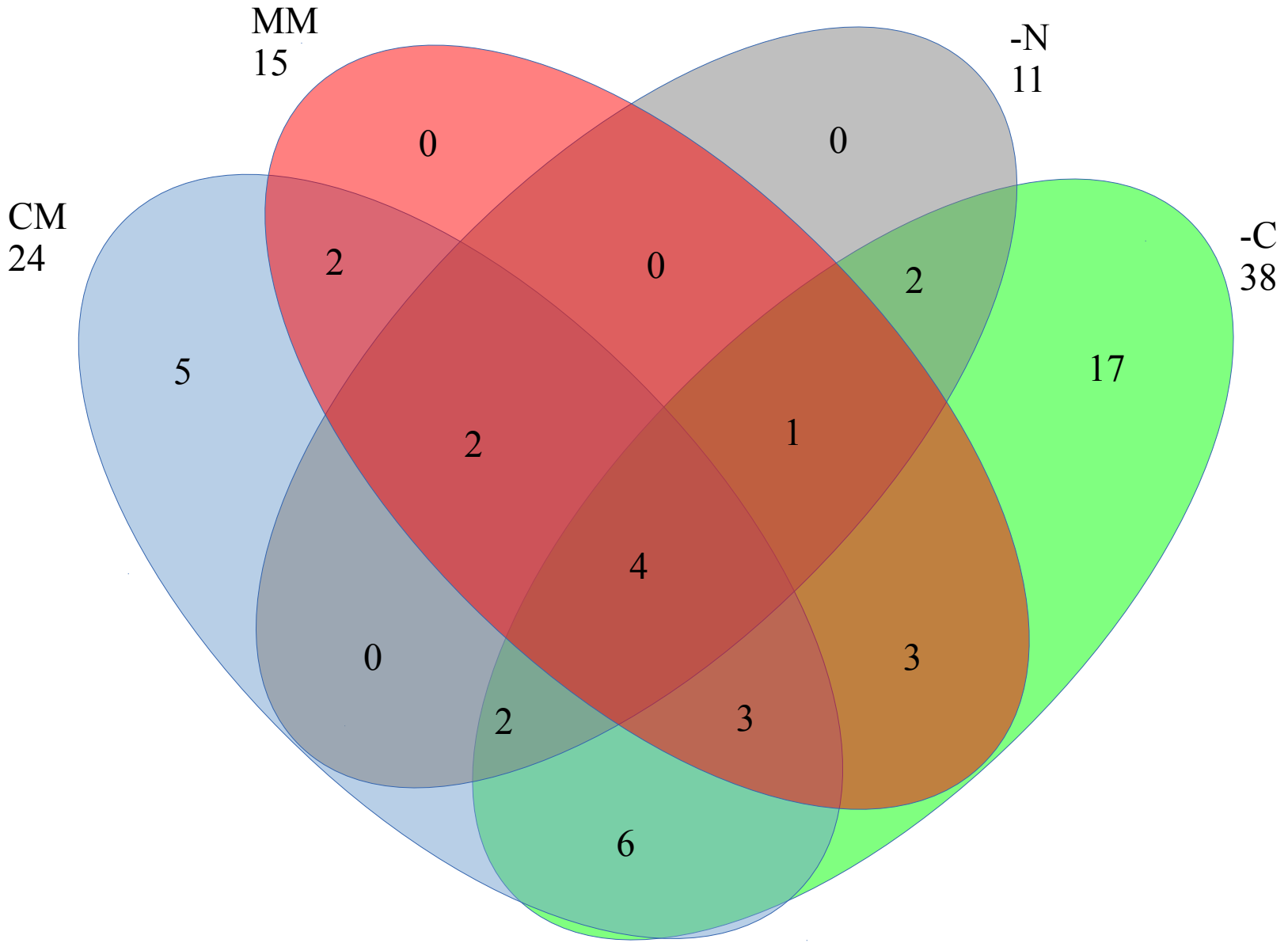




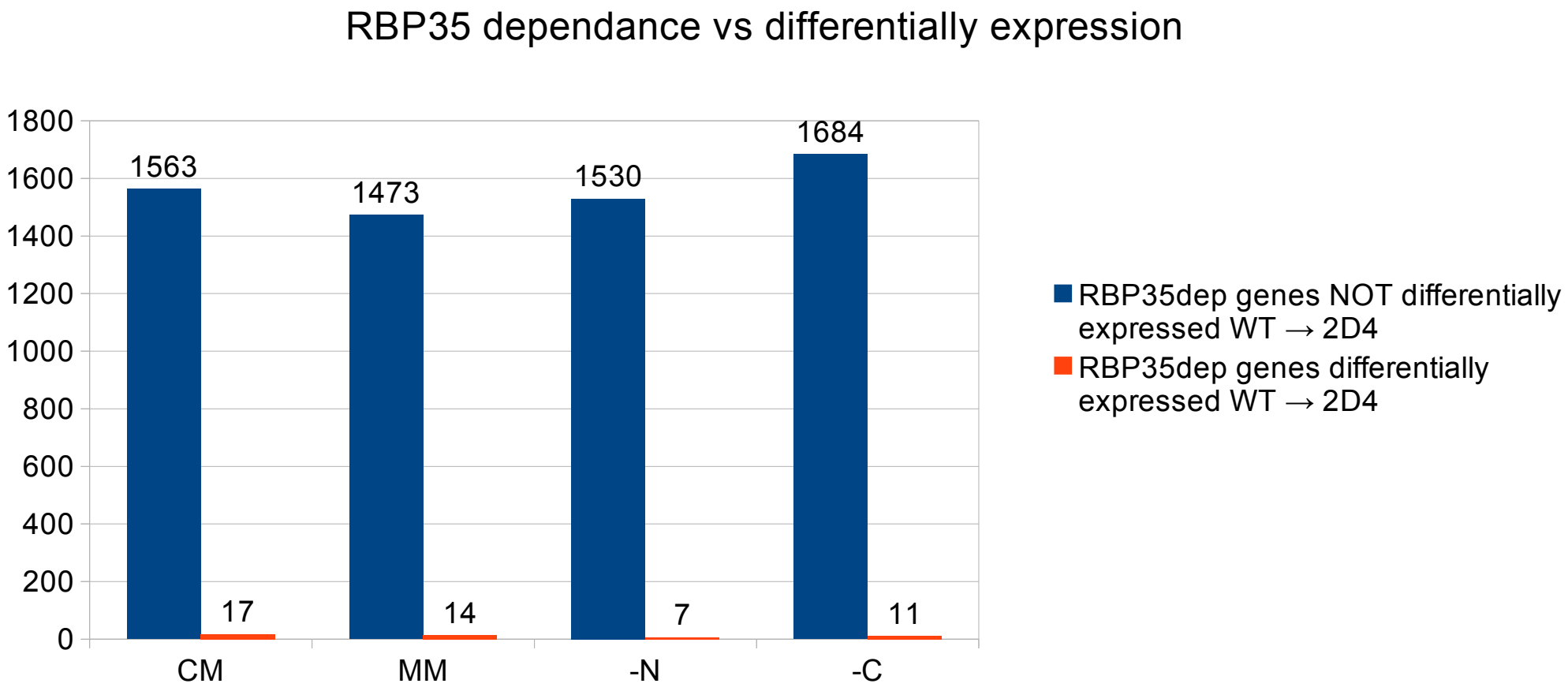
RBP35 dependant genes



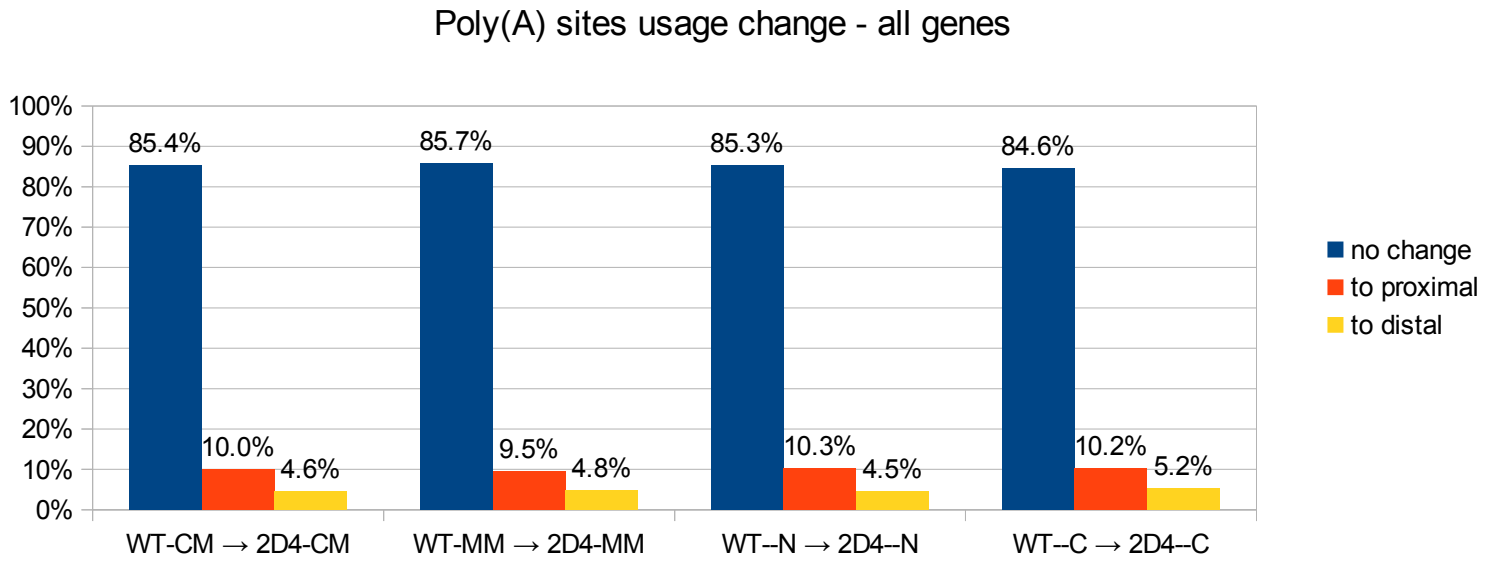
RBP35 dependant **Pathogenic** genes (CAMBIAR)

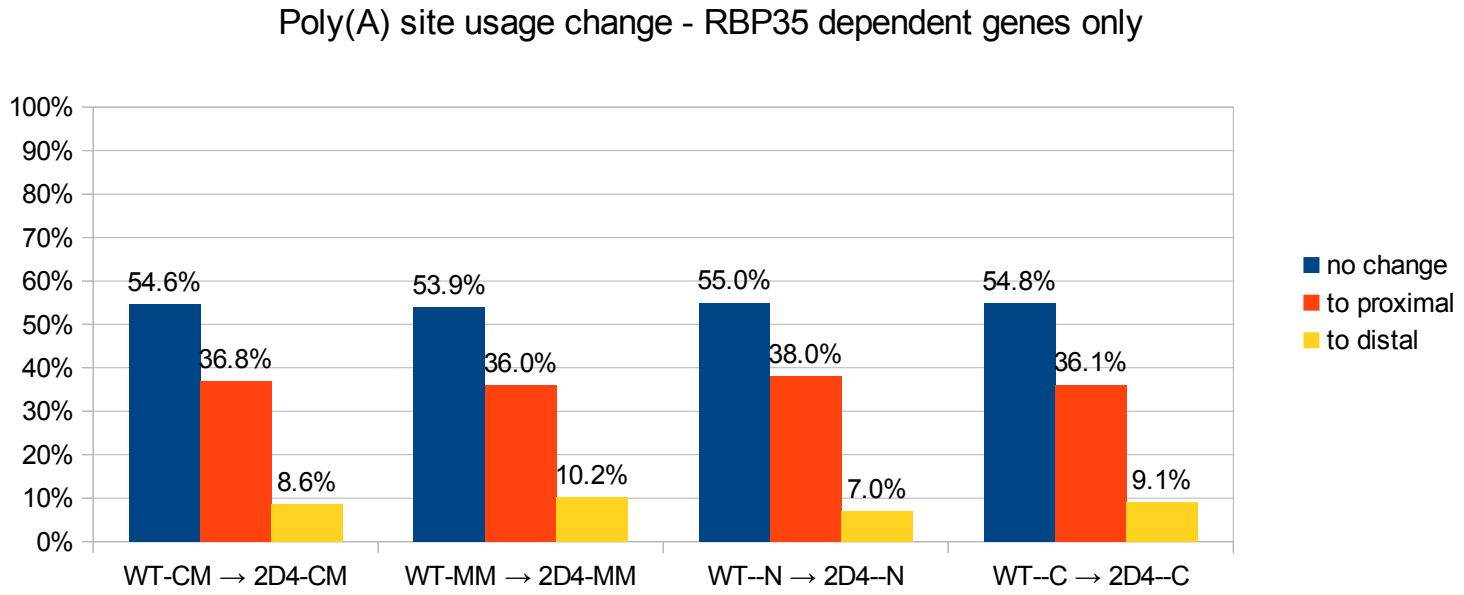


There is no correlation between RBP35 dependance and differential expression

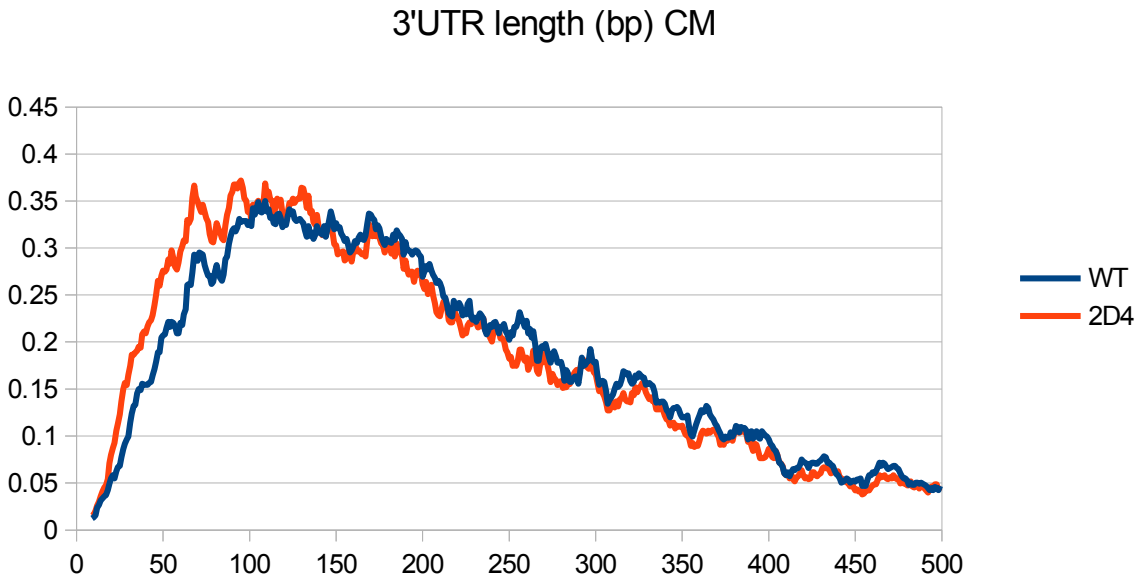


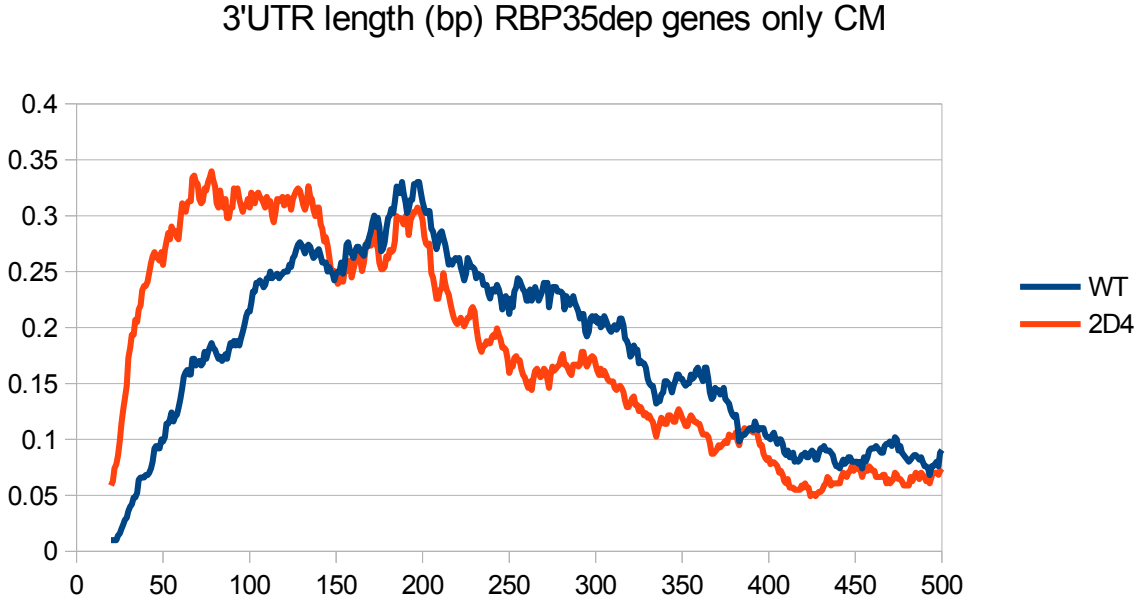
Δrbp35 affects poly(A) sites usage, preferring proximal cuts



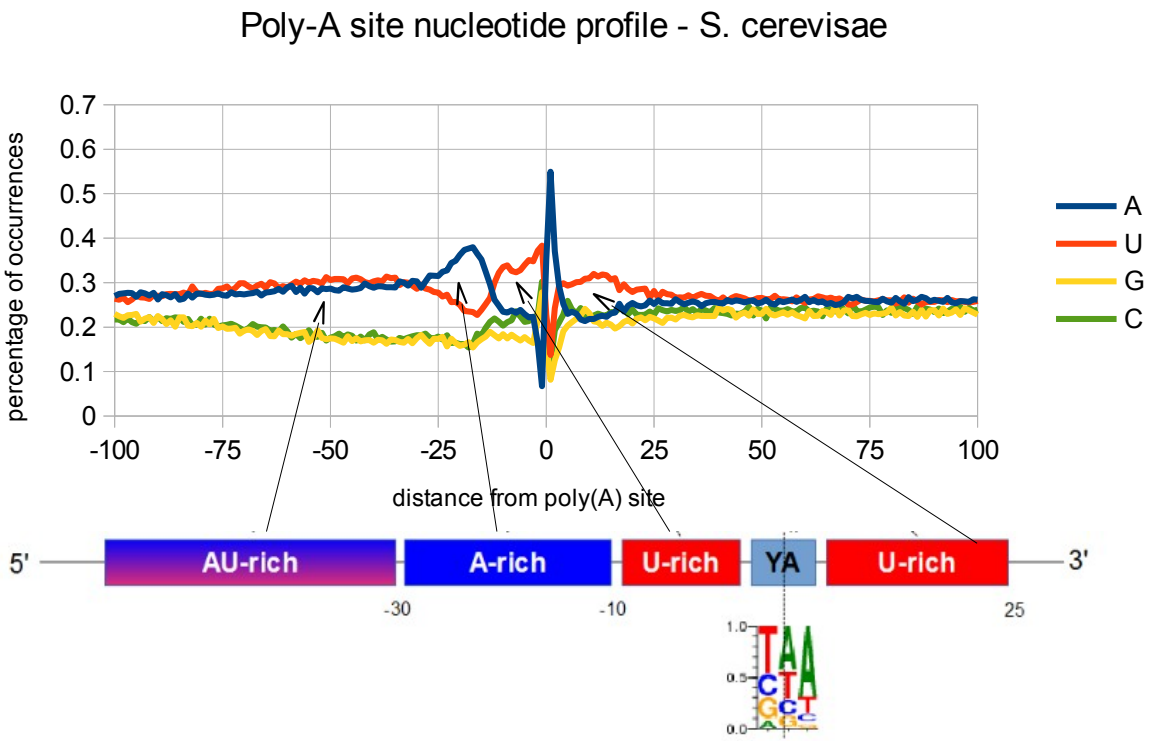


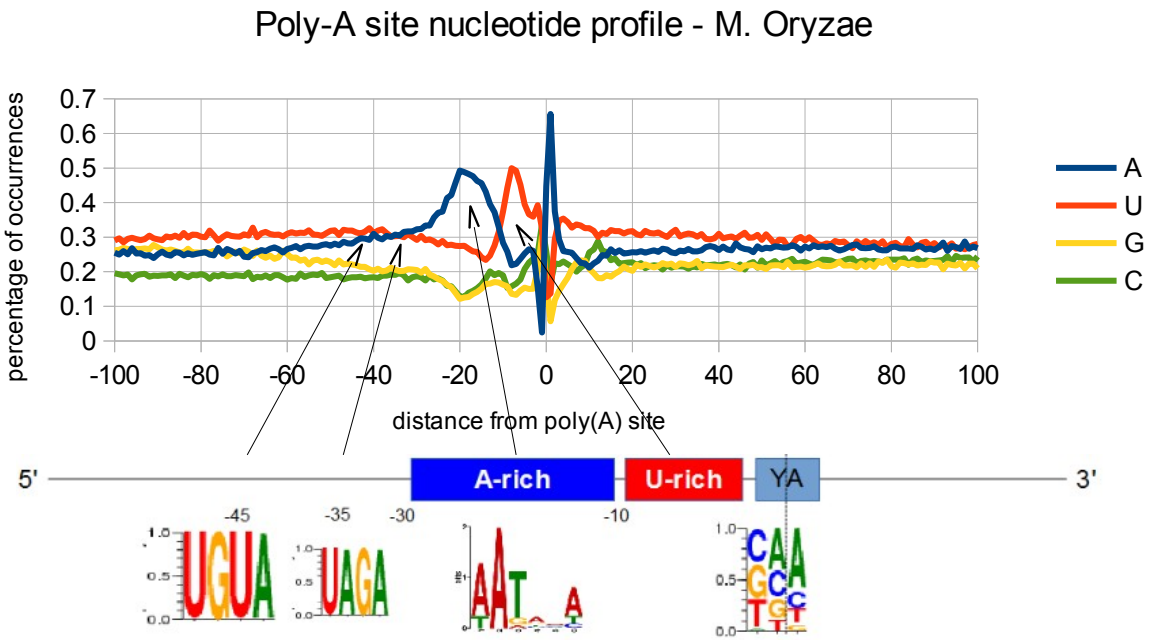
Δrbp35 affects 3'UTR length



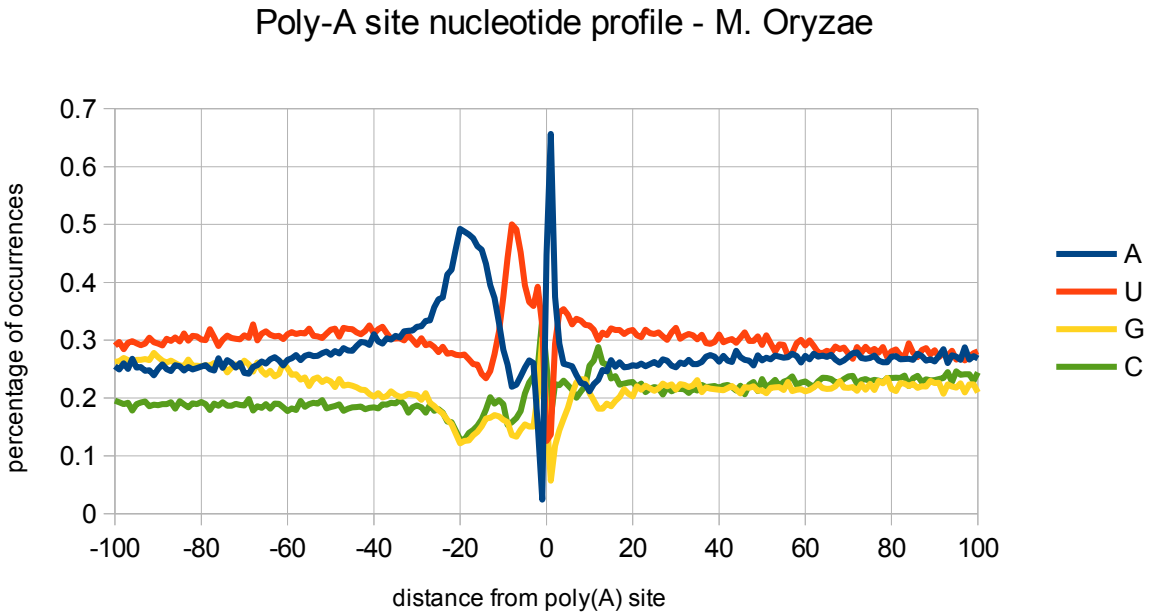


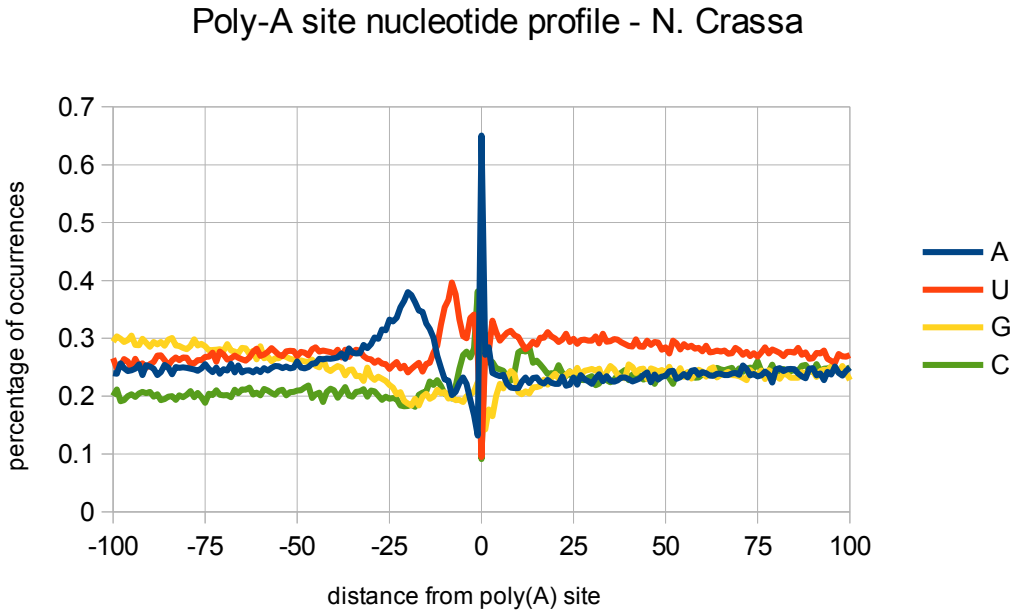
Nucleotides profile of poly(A) sites slighty differs from *S.cerevisae*

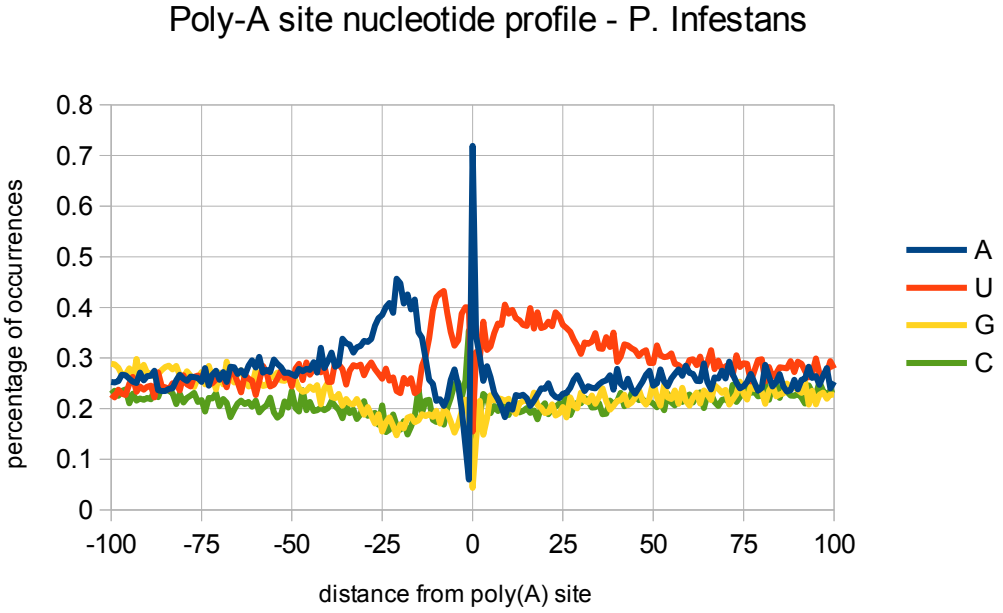




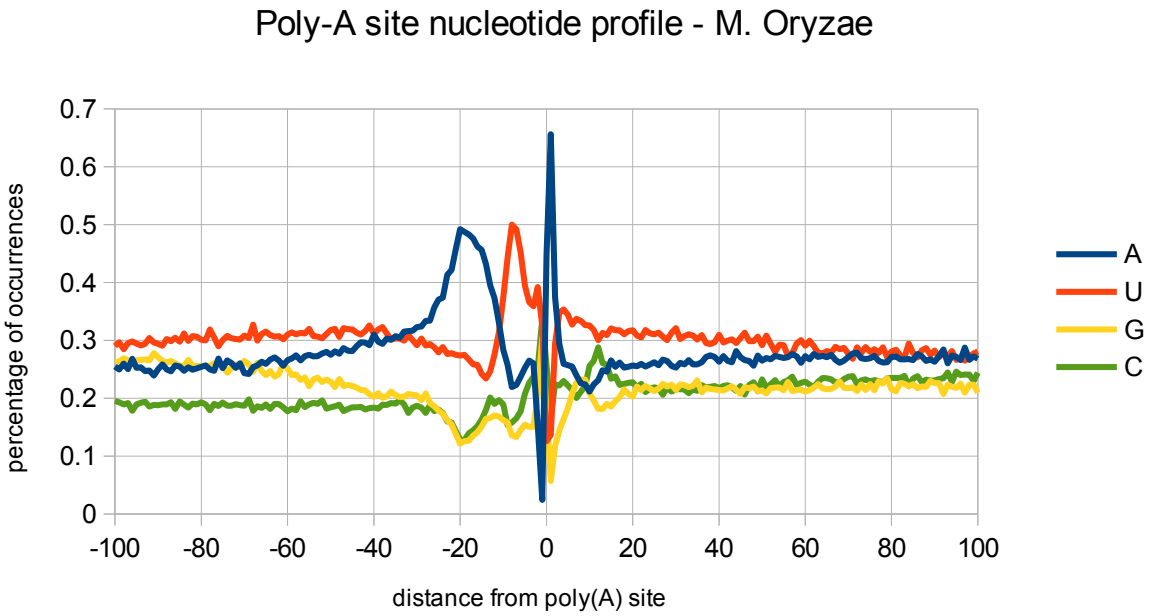
Nucleotides profile of poly(A) sites resembles *N. crassa*

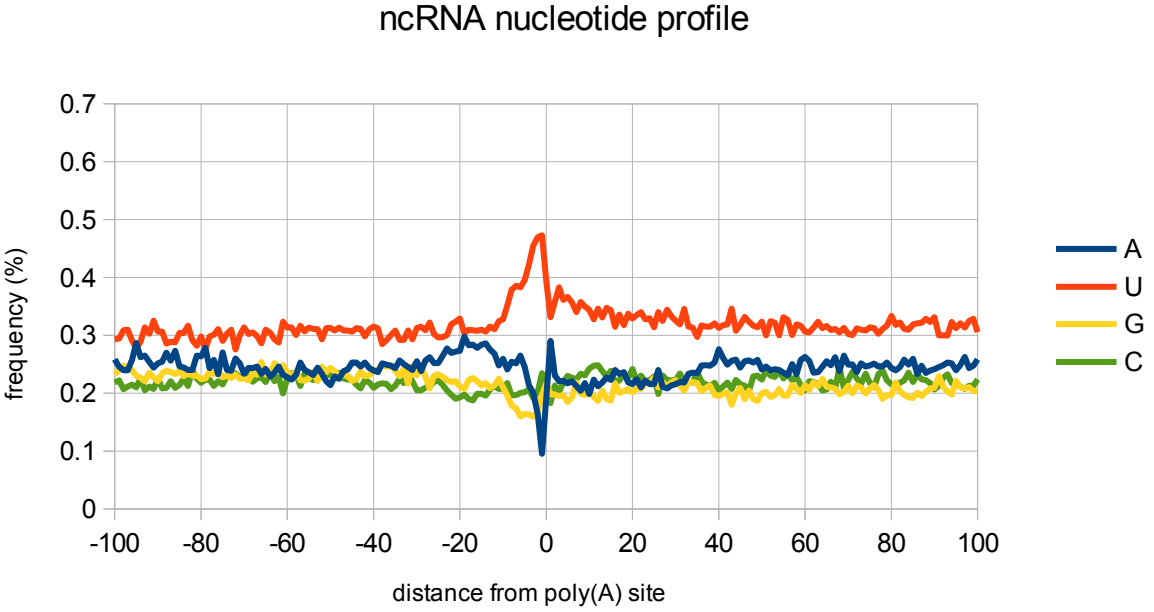




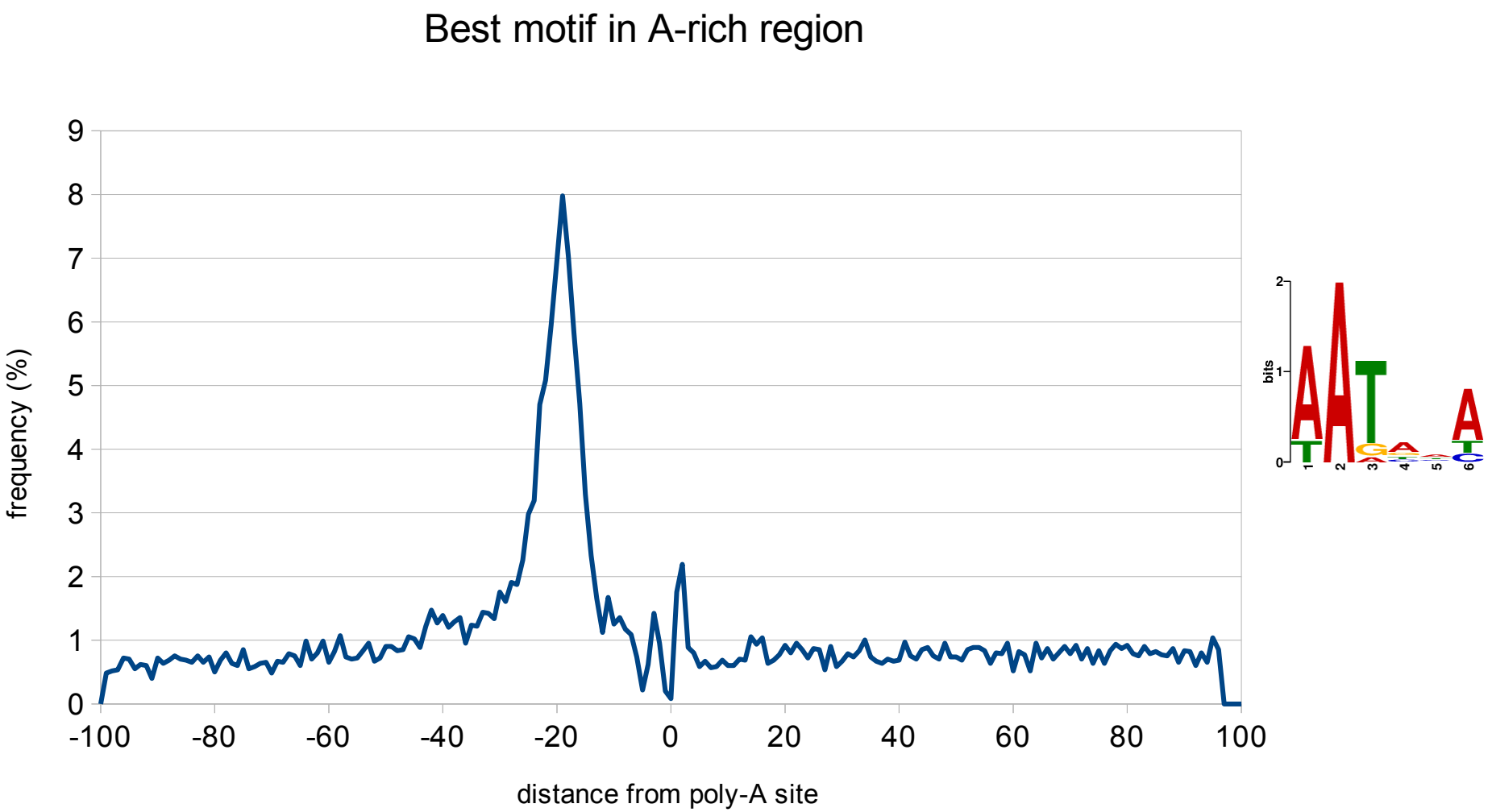


Nucleotides profile of poly(A) of ncRNA is different

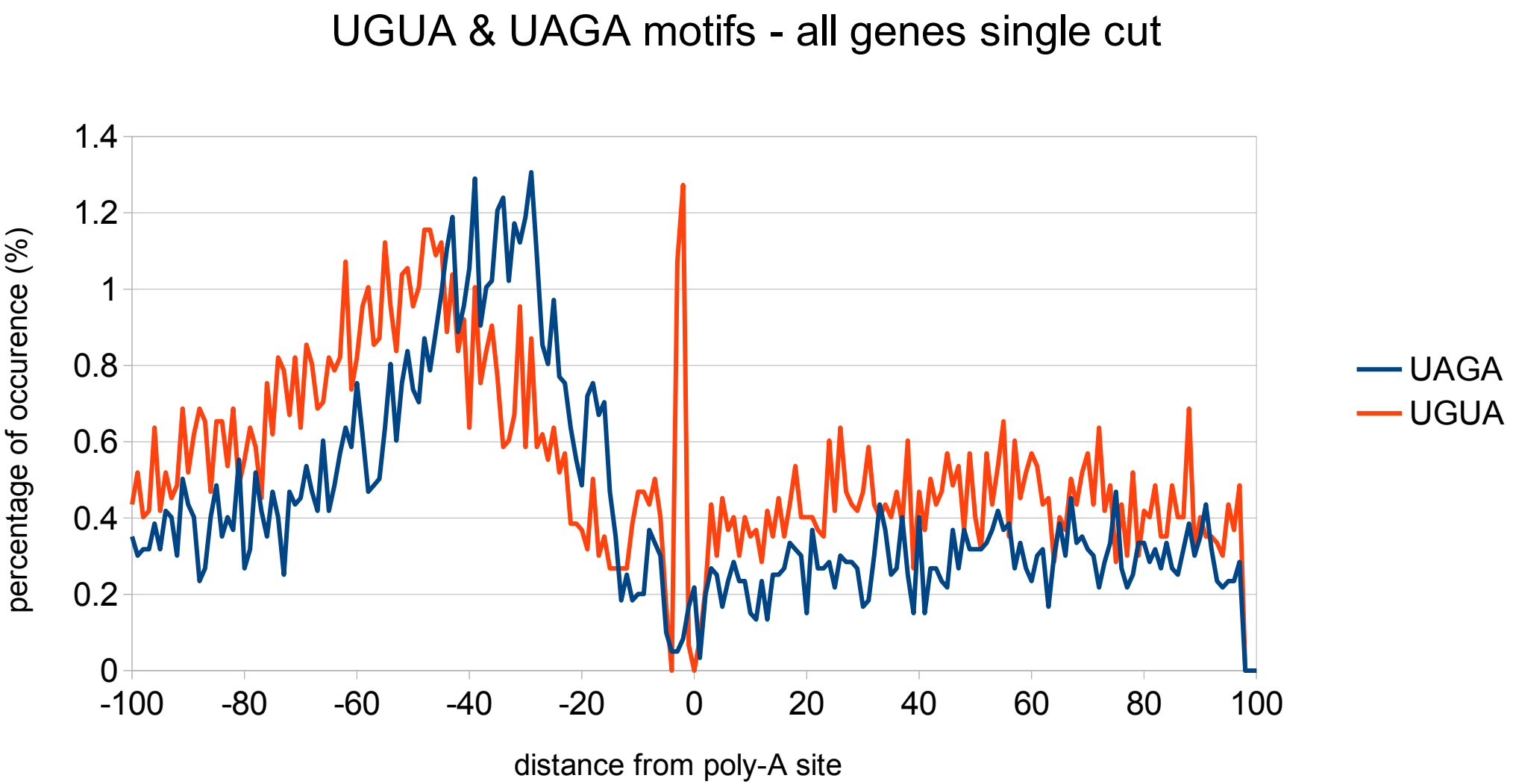




The A-RICH region is located -30 -10 bp upstream



UAGA & UGUA motifs



Polyadenilation signals in common genes

MPG1

...GG**UAGA**GAAGUCUCUUCUCGUUCCACUCAUUU**AAUAAA**ACCCCUUCCAGACC**UA**...

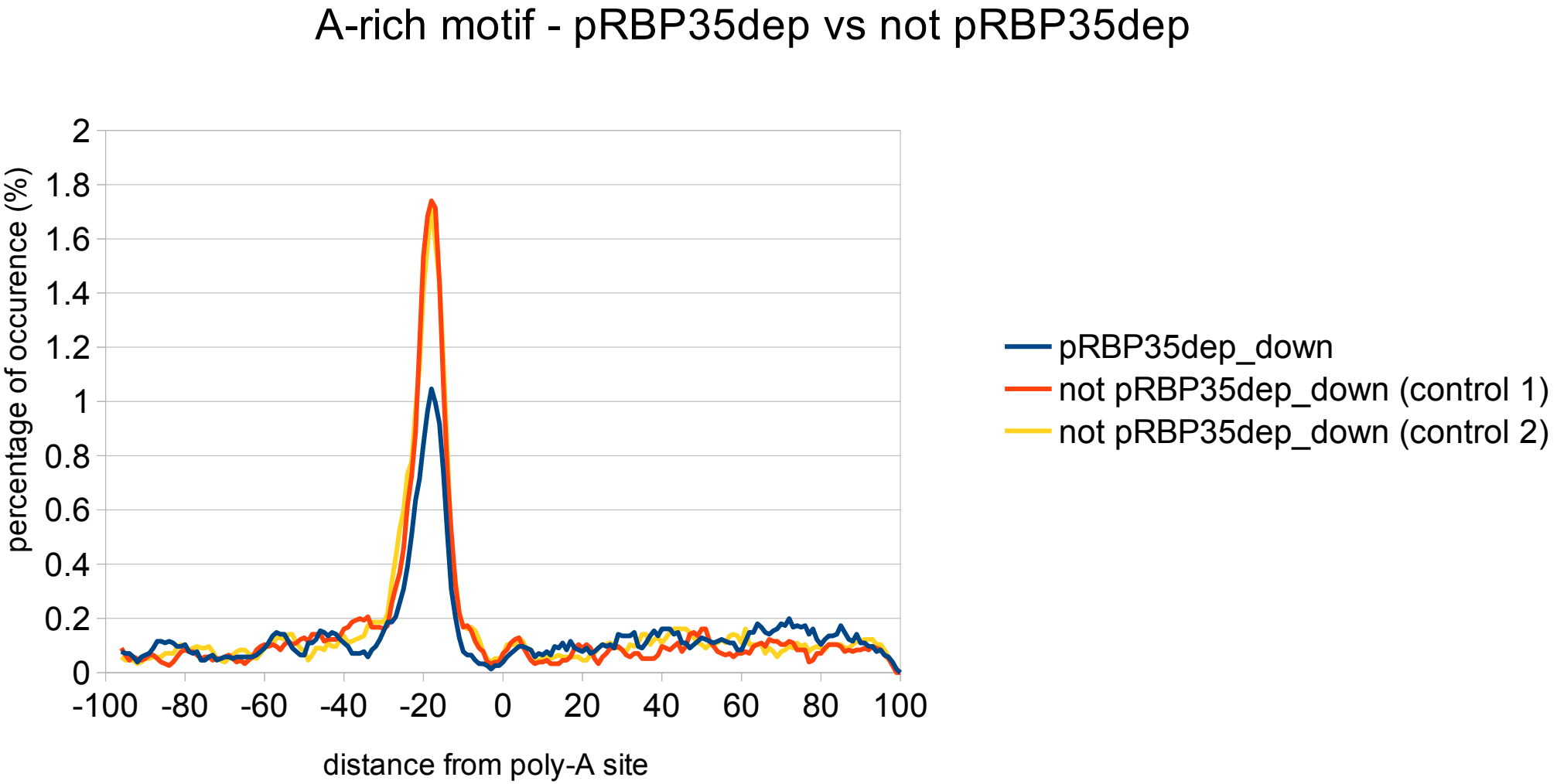
PMK1

...CGUU**UAGA**AUGUGCAGGAGACACGAGUGGGAAAAUG**AAUACA**UGGAUGCCAG**CA**...

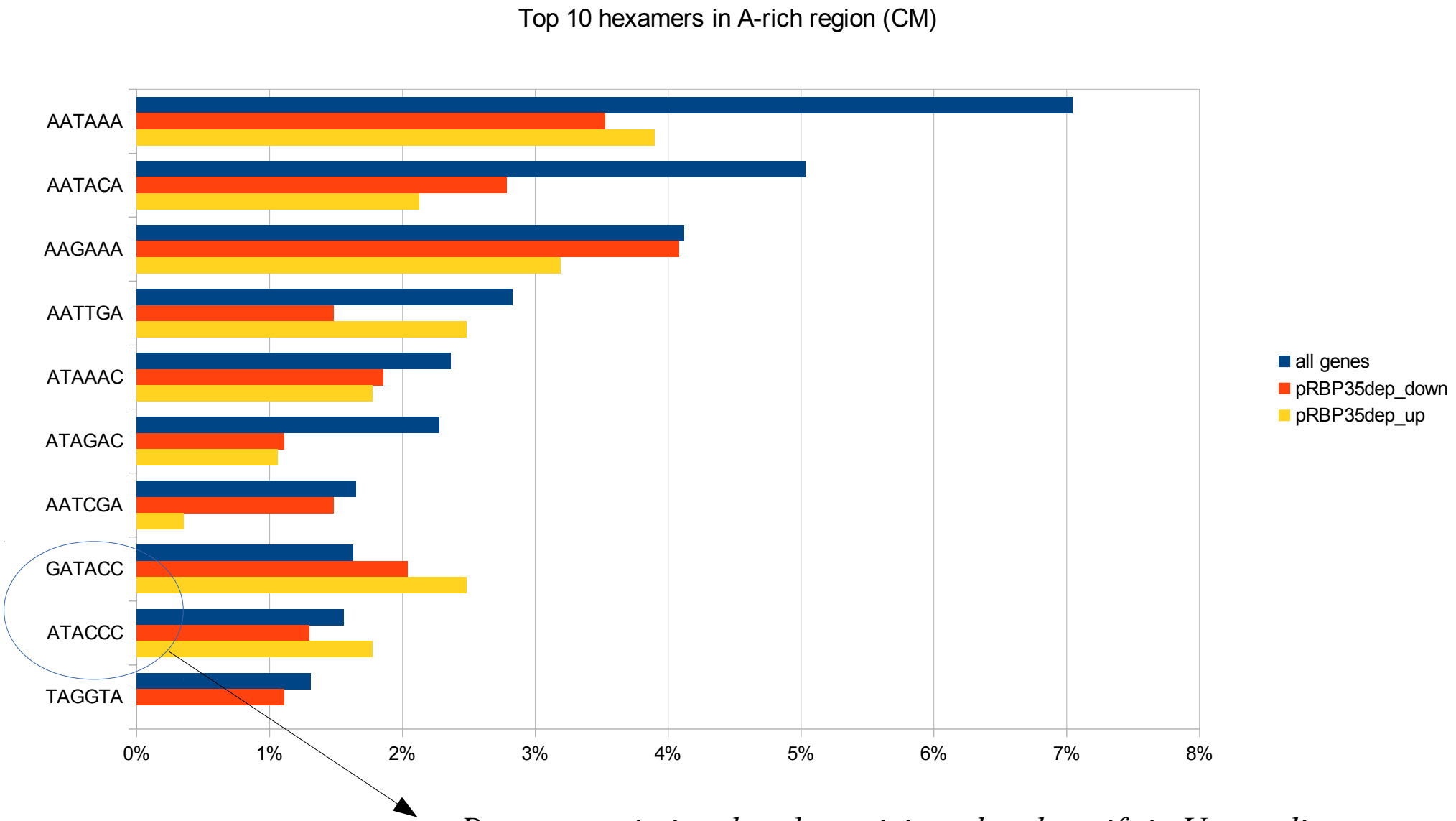
MST12

...CAGUGGCAUAAAAUCACAAAAUCUU**UAGA**AAGAUCAC**AGAAAA**CCUUUUGUC**CA**...

poly(A) sites dependent from *RBP35* show a less pronounced A-rich region



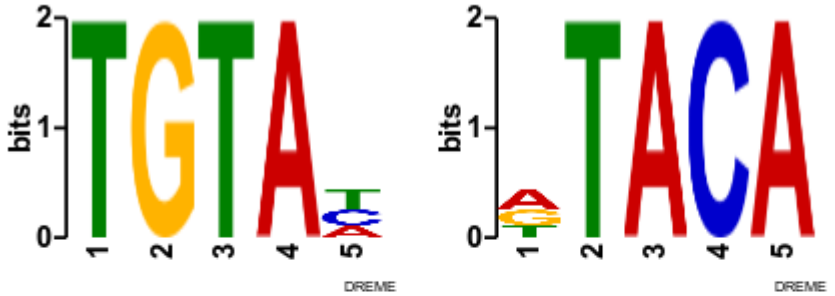
poly(A) sites dependent from *RBP35* are low in the canonical AATAAA signal



Post-transcriptional pathogenicity-related motifs in U. maydis

UGUAH motif is enriched in poly(A) sites dependent from *RBP35* down-regulated in Δrbp35, in the region -100 -30

1. TGTAH



Details

Positives	Negatives	P-value	E-value	Unerased E-value
1608/3115	815/3115	9.6e-96	7.6e-91	7.6e-91

Enriched Matching Words

Word	Positives	Negatives	P-value	▼E-value
TGTAT	802/3115	363/3115	3.6e-47	2.9e-42
TGTAC	616/3115	297/3115	7.0e-31	5.5e-26
TGTAA	472/3115	237/3115	2.4e-21	1.9e-16
TGTAC	412/3115	273/3115	1.0e-8	8.2e-4

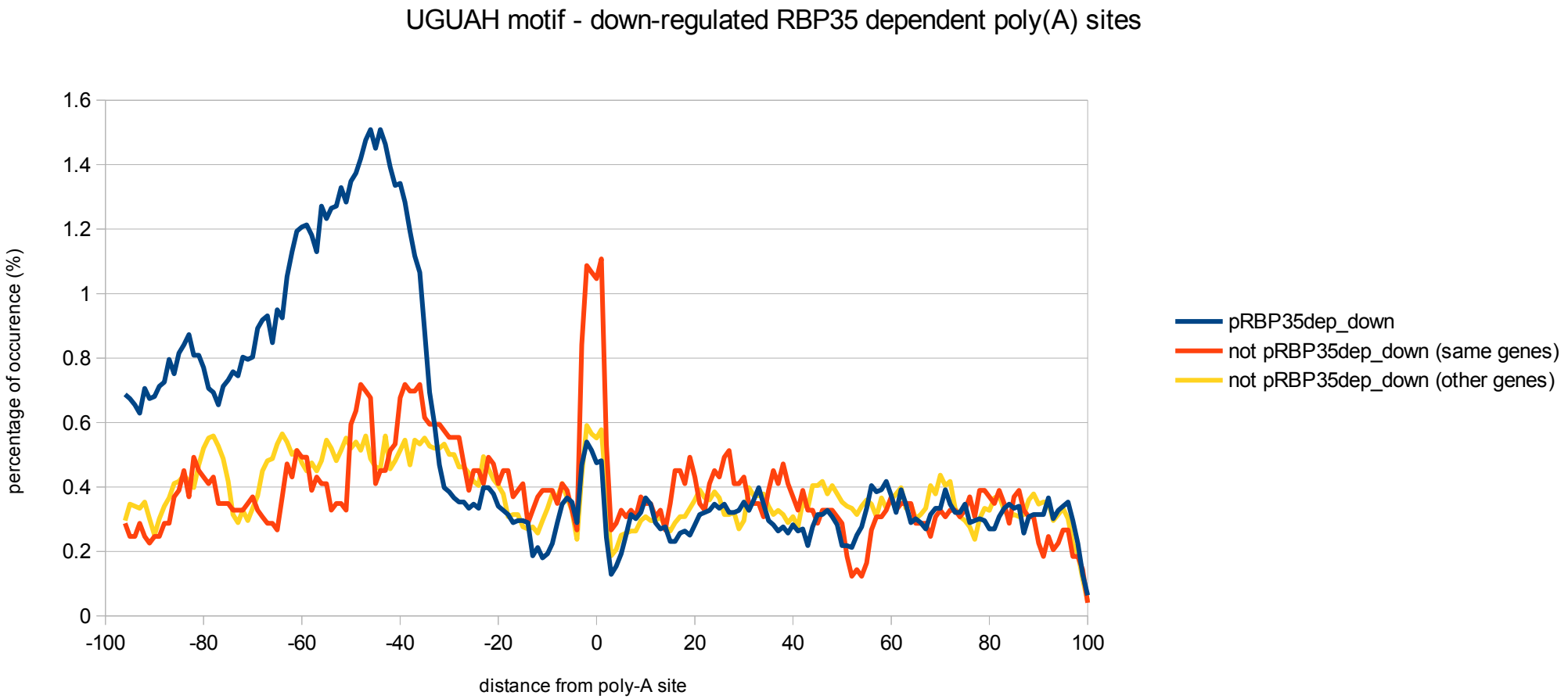
- Output of DREME, pRBP35dep as positive sequences list versus not-pRBP35dep negative list

UGUAH motif – pRBP35dep vs not pRBP35dep

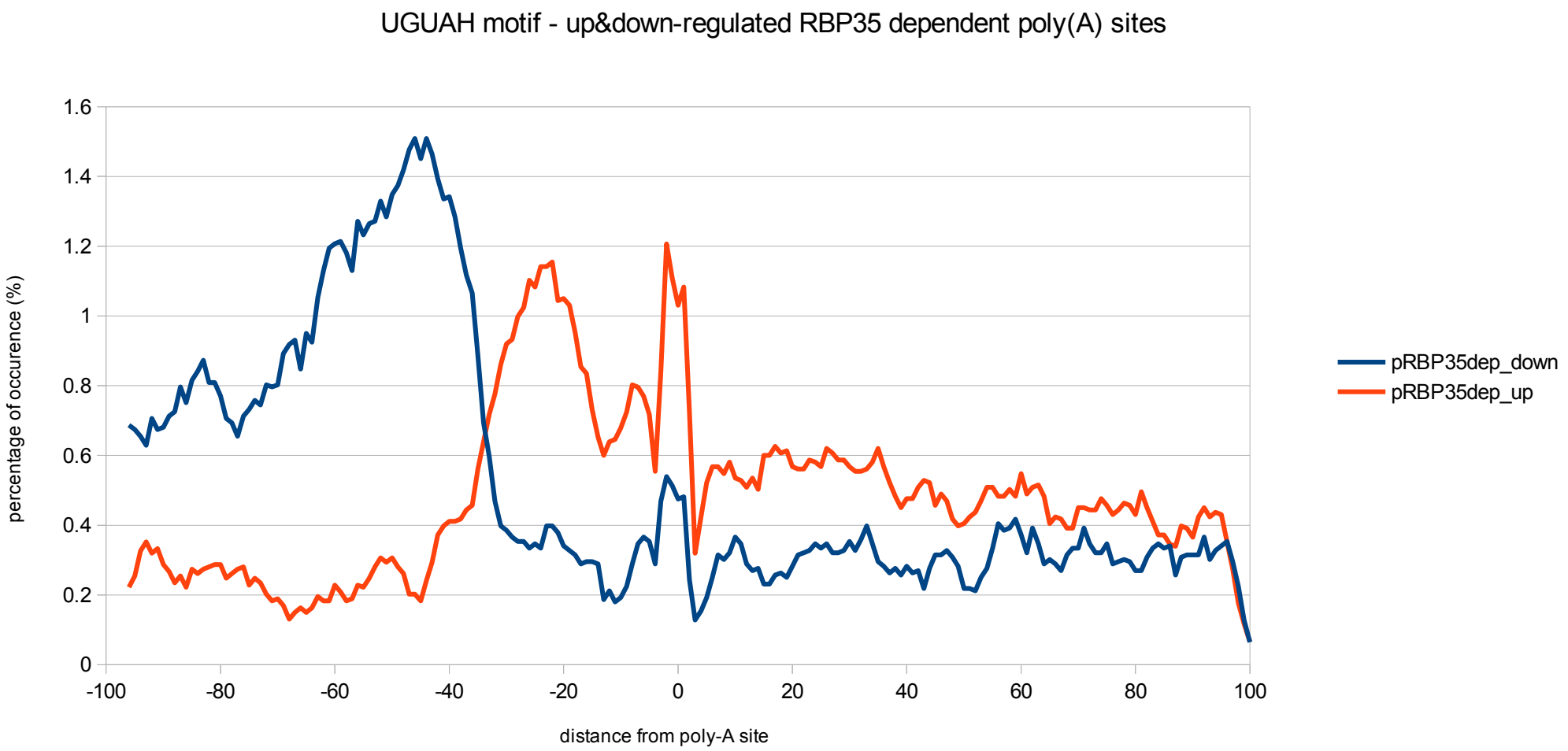
- In the first graph, we want to show how poly(A) sites dependent from *RBP35* display a different profile for the UGUAH motif in the respect to "regular" poly(A) sites

- We therefore plot down-regulated RBP35 dependent poly(A) sites against two groups of poly(A) not dependent from RBP35 of the same size, one group of poly(A) sites belonging to the same genes and one group of poly(A) sites belonging to other genes

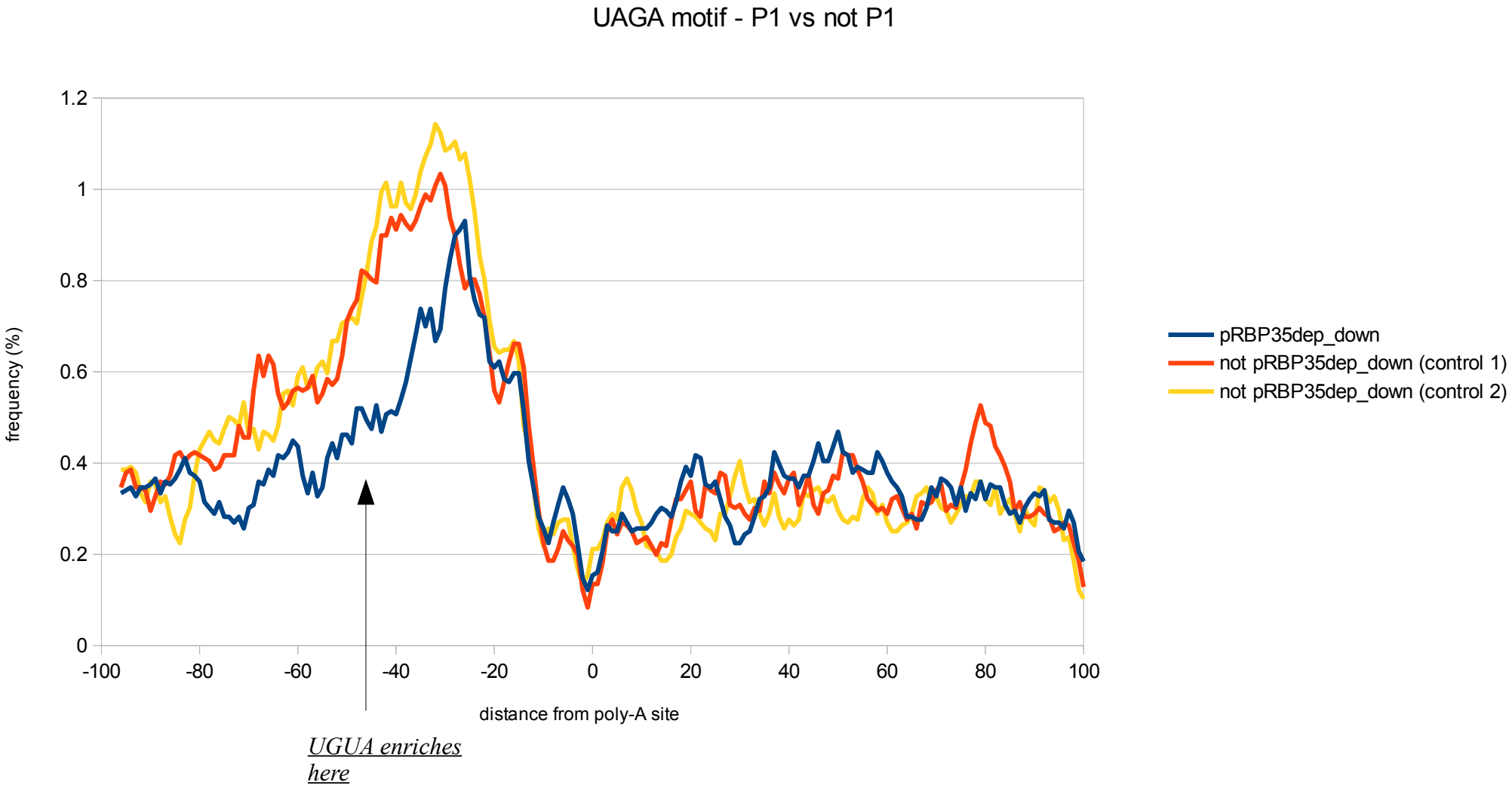
UGUAH is enriched at -45 in poly(A) sites dependent from *RBP35*



UGUAH motif – RBP35 dependent poly(A) sites (up vs down regulated)

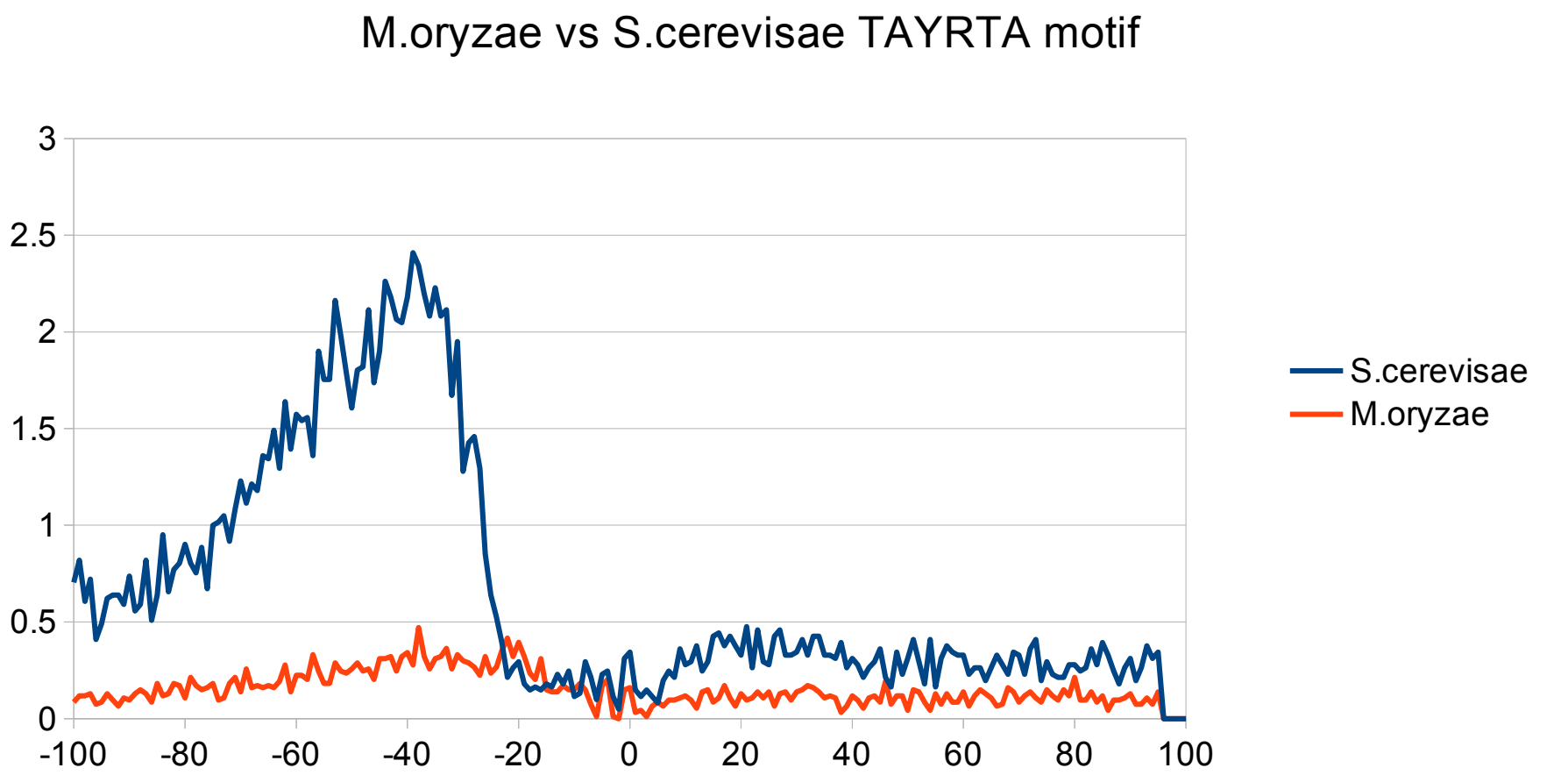
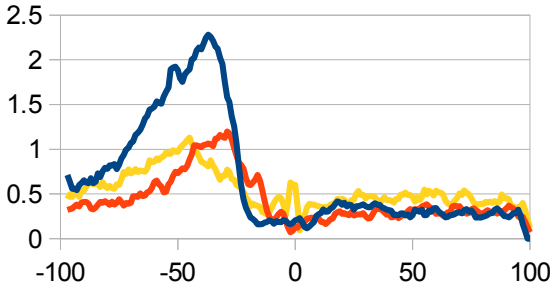


UAGA– pRBP35dep vs not pRBP35dep



UAGA is shifted in RBP35dep poly(A) sites, probably due to UGUA enrichment

The HRP1 binding motif TAYRTA from *S.cerevisae* in not found in *M.oryzae*



Maybe in M.oryzae CFIm binds to UGUA and Hrp1 binds to UAGA?