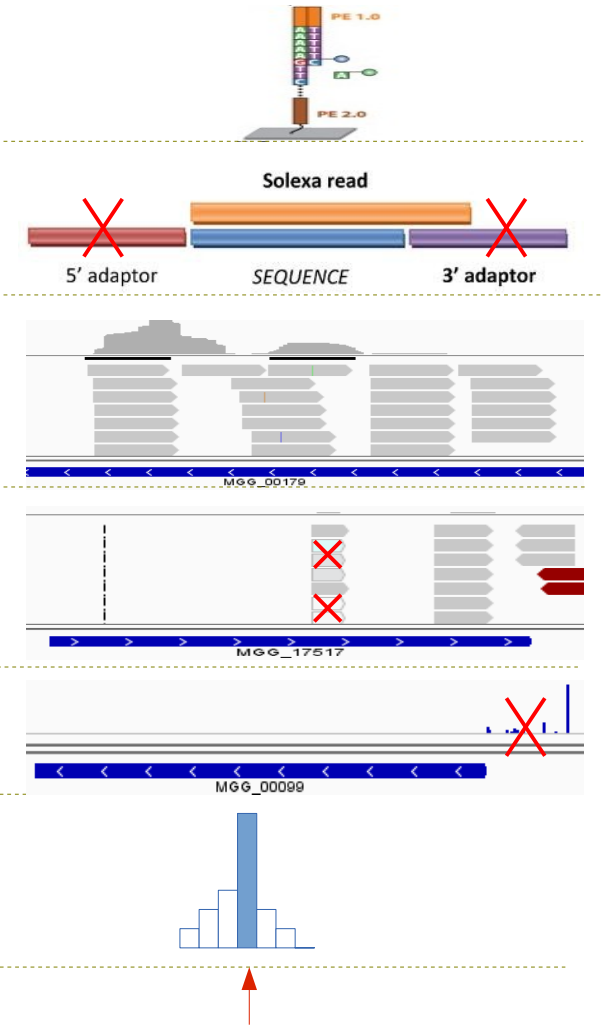
# Sequencing resume

- 2 strains (*WT, Δrbp35*) x 4 conditions (*CM, MM, -N, -C*) x 3 replicates
- 4751592 – 11517077 total reads database
- ~62% - ~82% 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

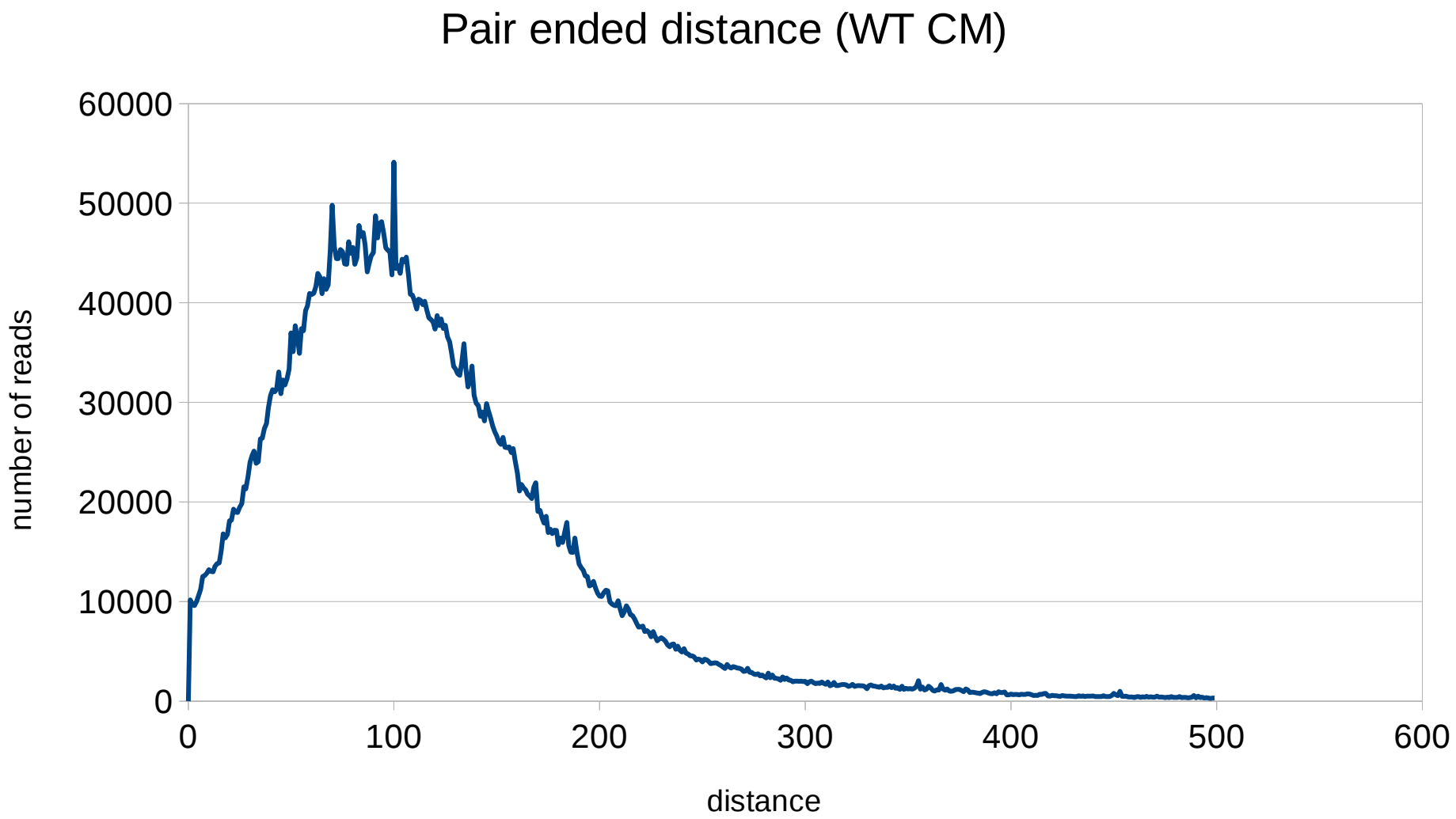
| Program and details | Procedure | % of data left |
|---|---|---|
| | Illumina sequencing | 100% |
| **fasta-mcf**, *adapter trimming* | Data prepping | 96% |
| **gsnap**, *default parameters* | Alignment | 96% |
| • *Low quality mapping (<30))*<br>• *High A/T content (>90%)*<br>• *Internal priming* | 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 (at least 5 supporting reads) at 33 bp distance* | poly(A) sites detection | 27% |




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

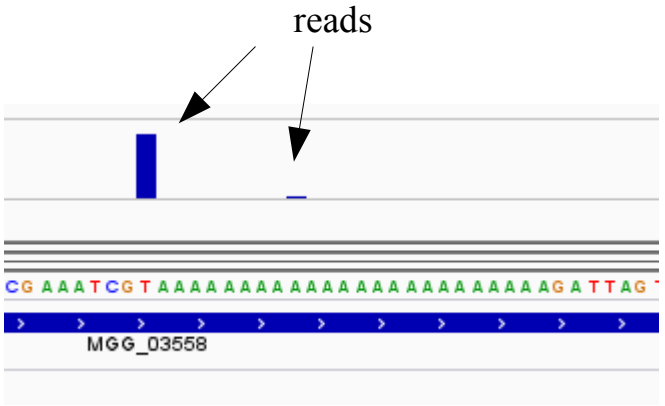
- whole gene expression ~7000 reads
- poly(A) site expression ~3000 reads

# Pair ended reads distance

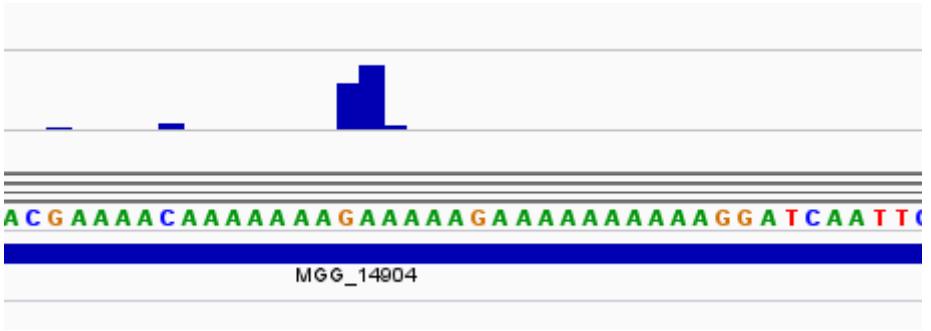


# ~2.5% of poly(A) sites are internal priming

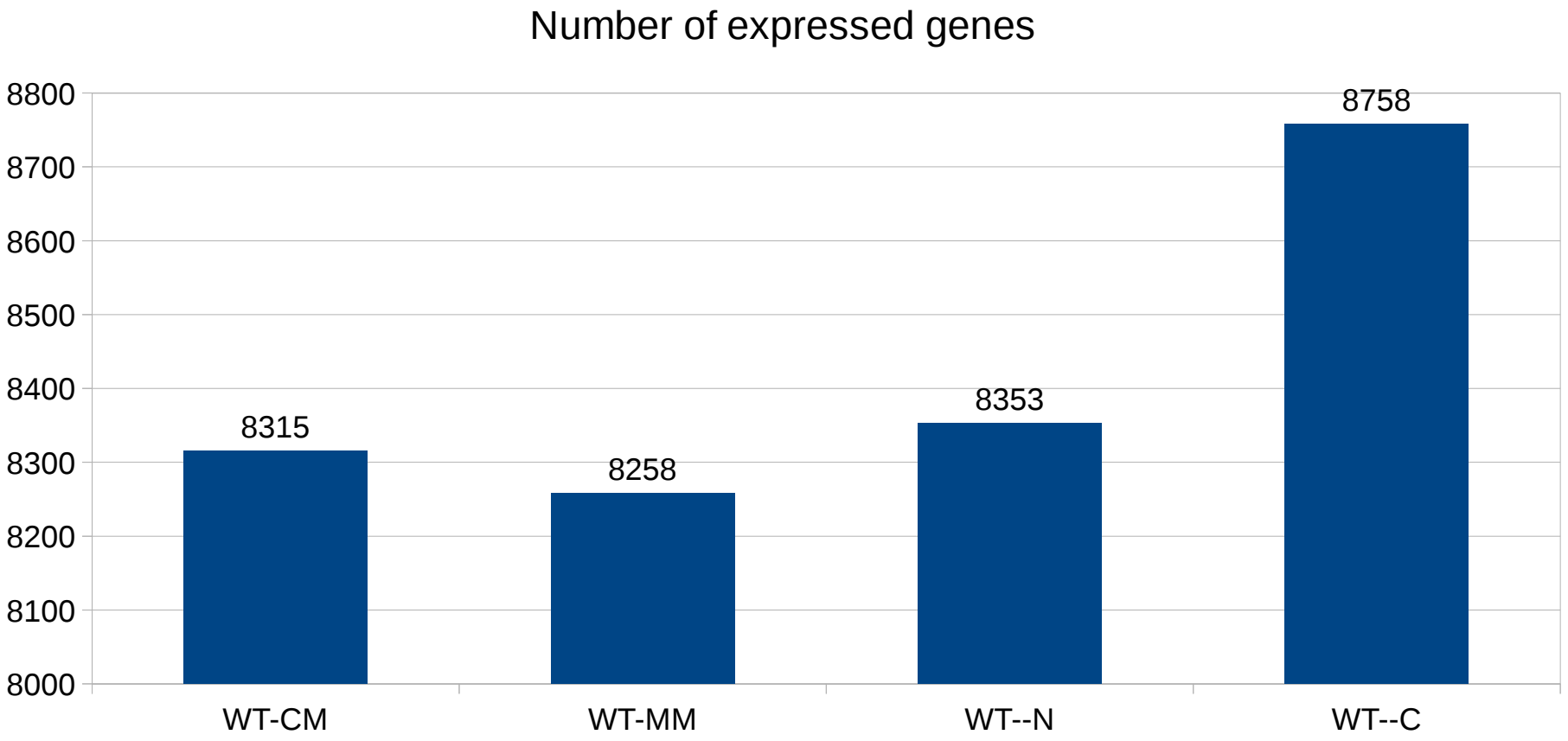
- Some poly(A) sites are just a side effect of poly(A) ***genomic*** regions




stretches of As



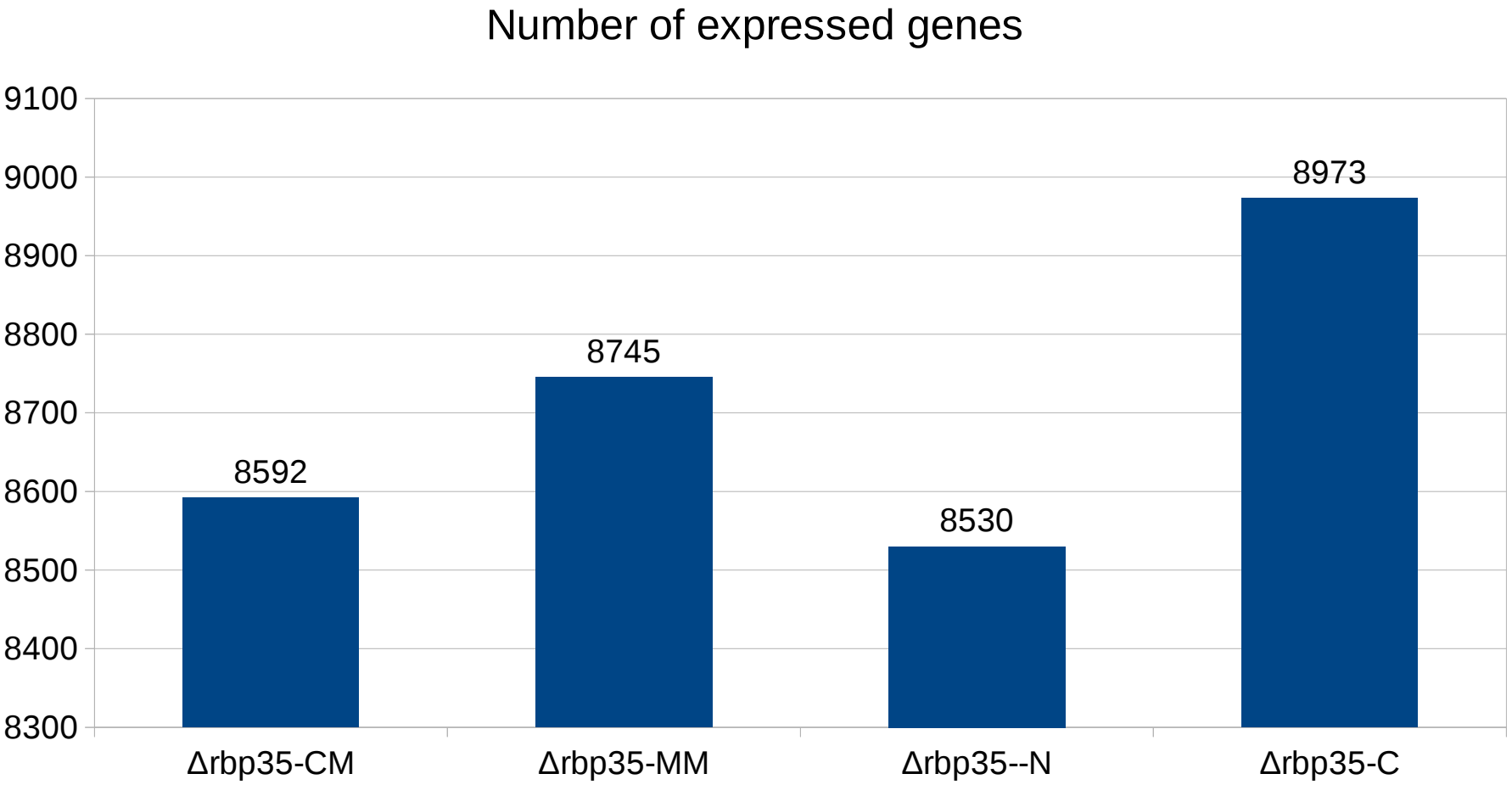
# ~8500 genes are expressed, out of a total of 13218 annotated genes (WT)




- 7662 genes are expressed in every condition (WT only)
- 3979 genes are never expressed (WT only)

*A gene is considered as expressed when has at least 10 supporting reads in a least 2 replicates*

# ~8500 genes are expressed, out of a total of 13218 annotated genes (Δrbp35)

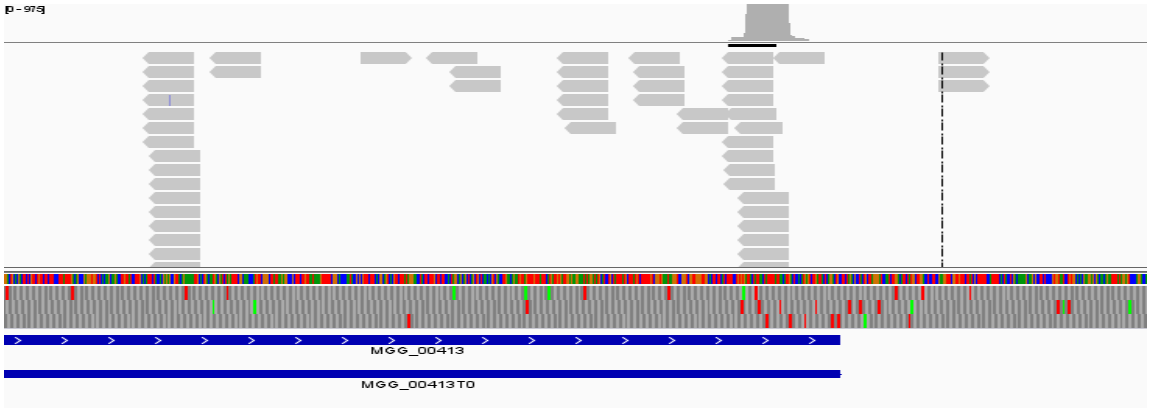



- 7993 genes are expressed in every condition (*Δrbp35* only)
- 3757 genes are never expressed (*Δrbp35* only)

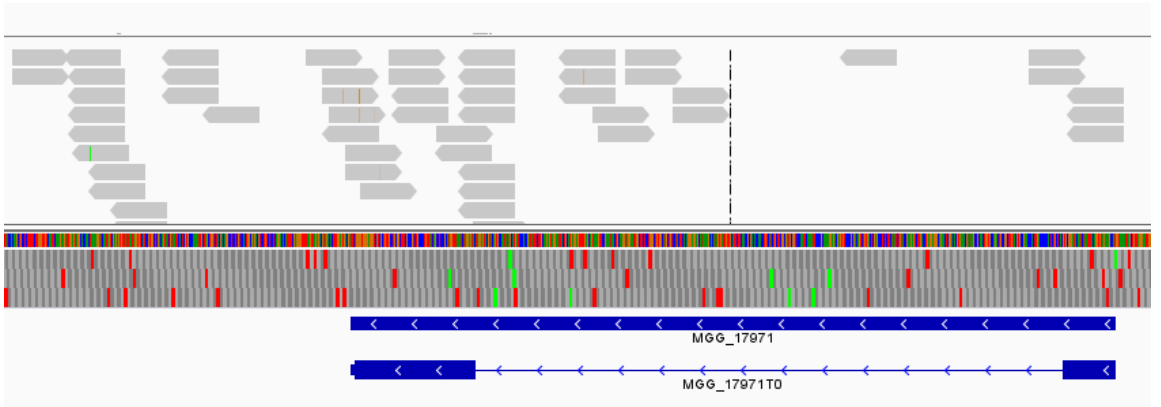
*A gene is considered as expressed when has at least 10 supporting reads in a least 2 replicates*

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

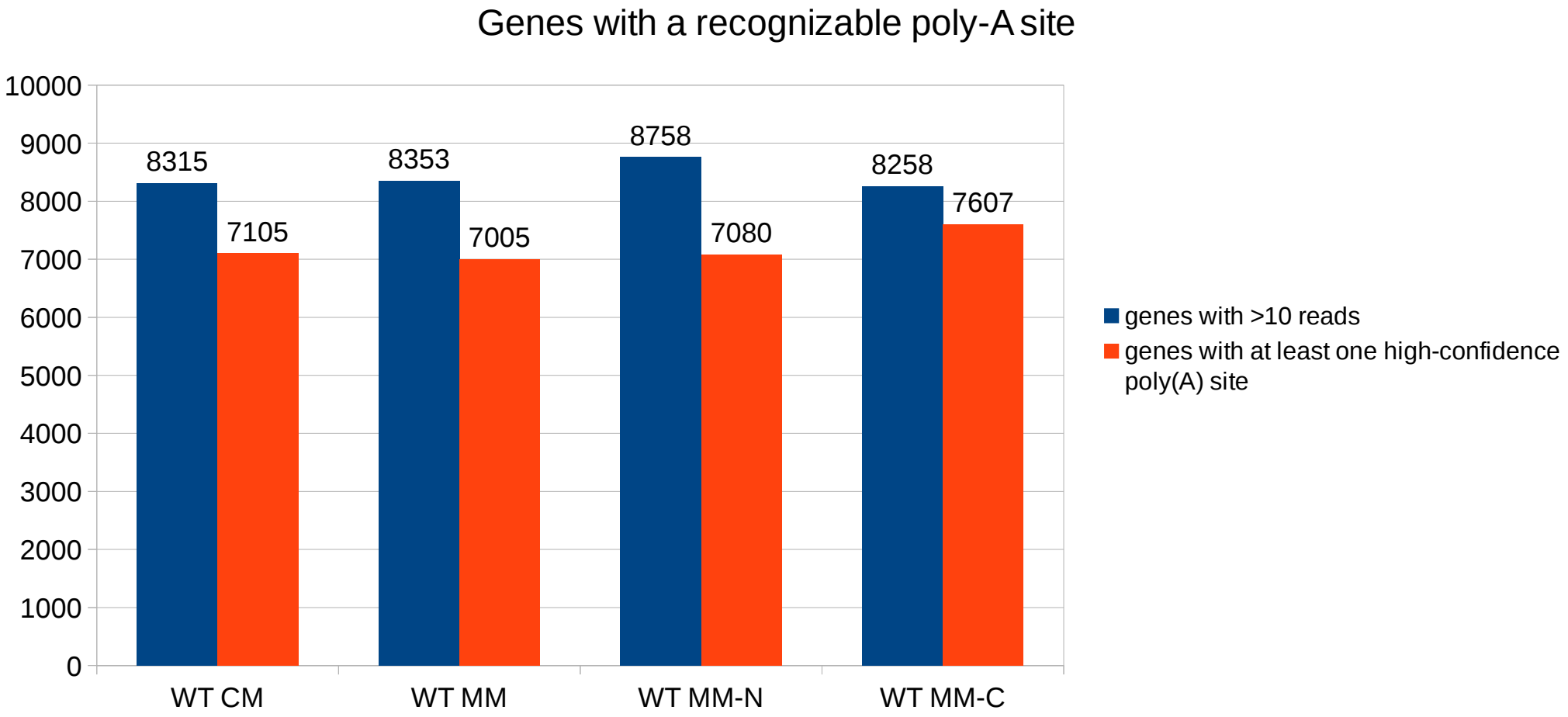
*Expressed gene <u>with</u> a recognizable poly(A) site:*



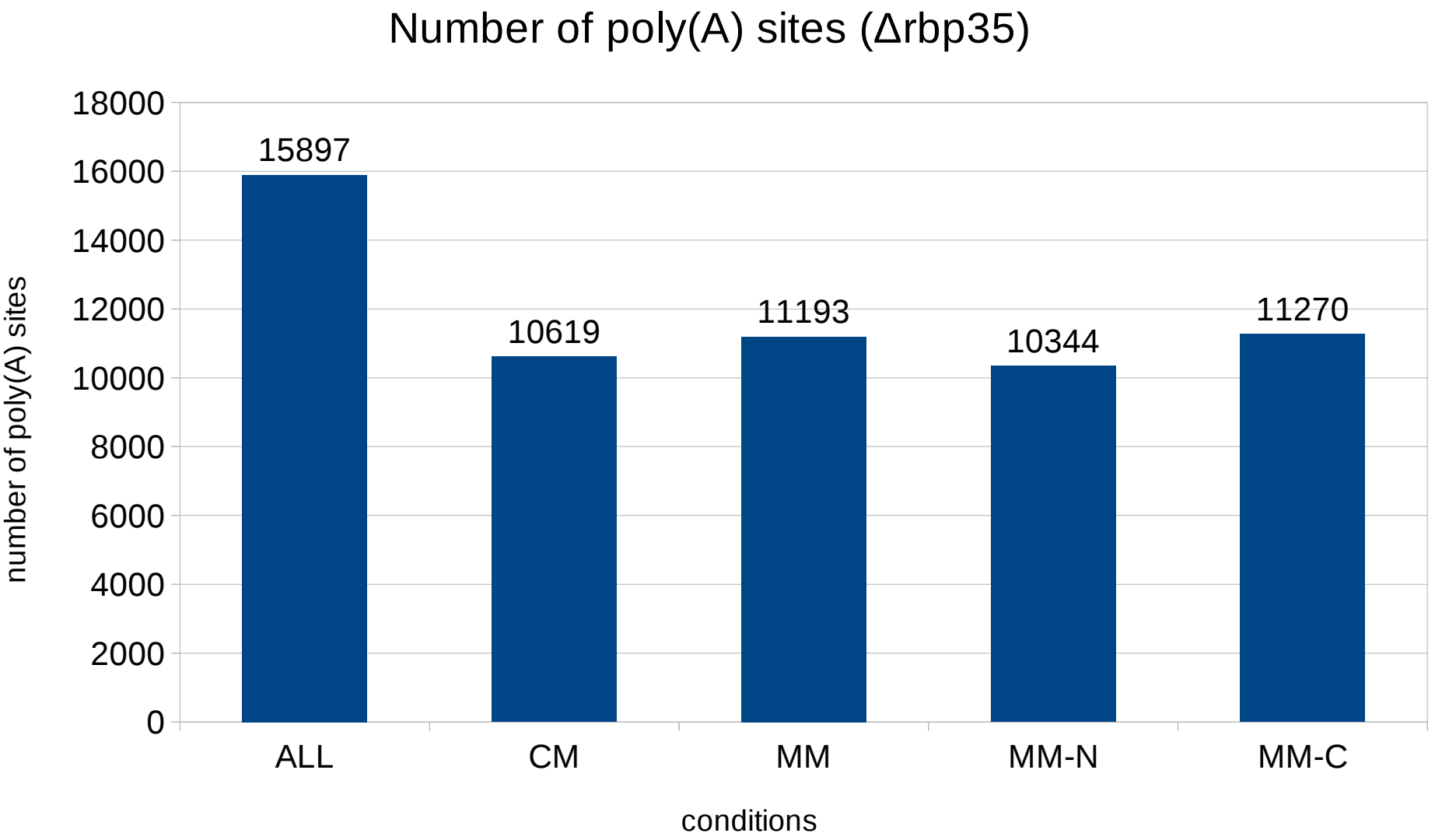

*Expressed gene <u>without</u> a recognizable poly(A) site:*



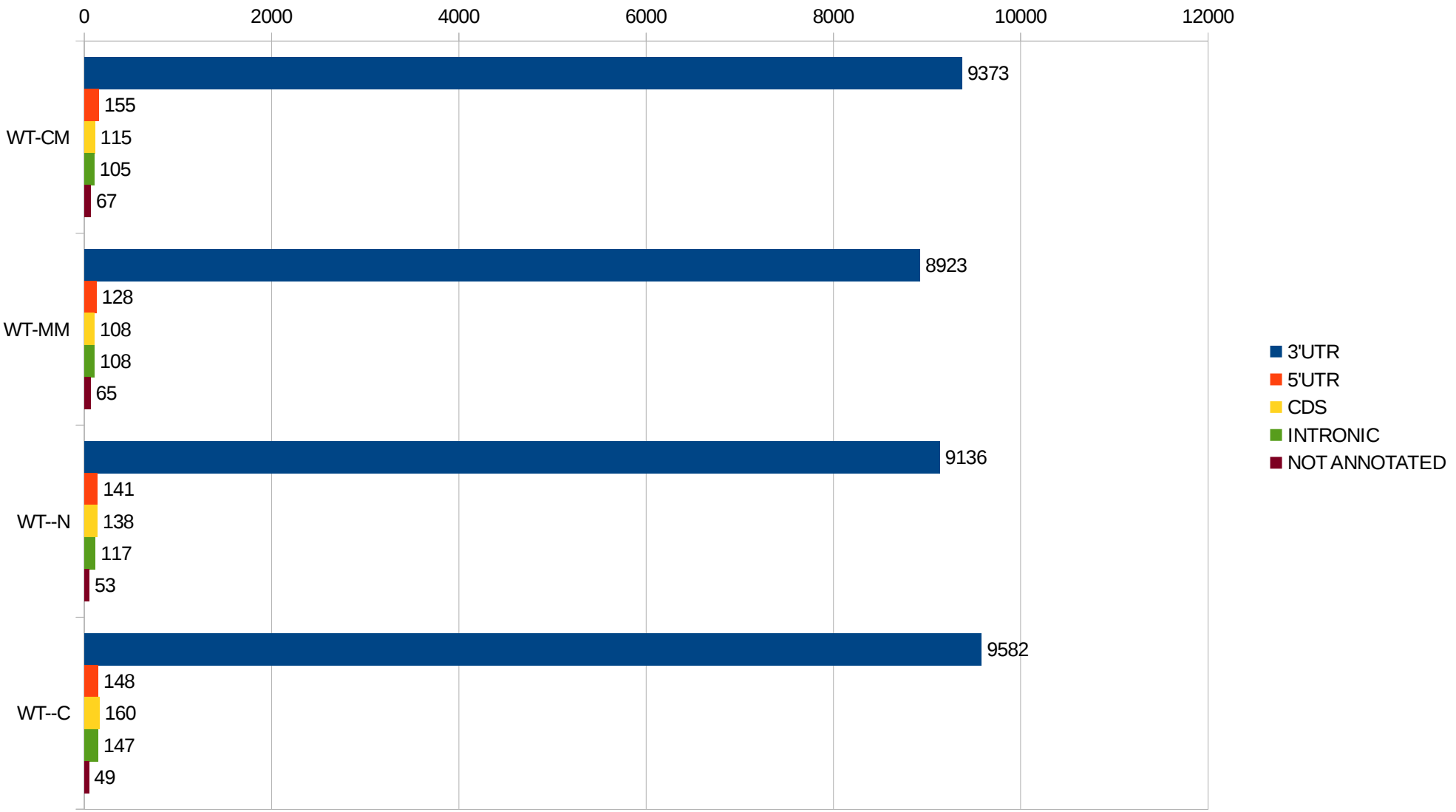
# ~85% of genes expressed have a recognizable poly(A) site



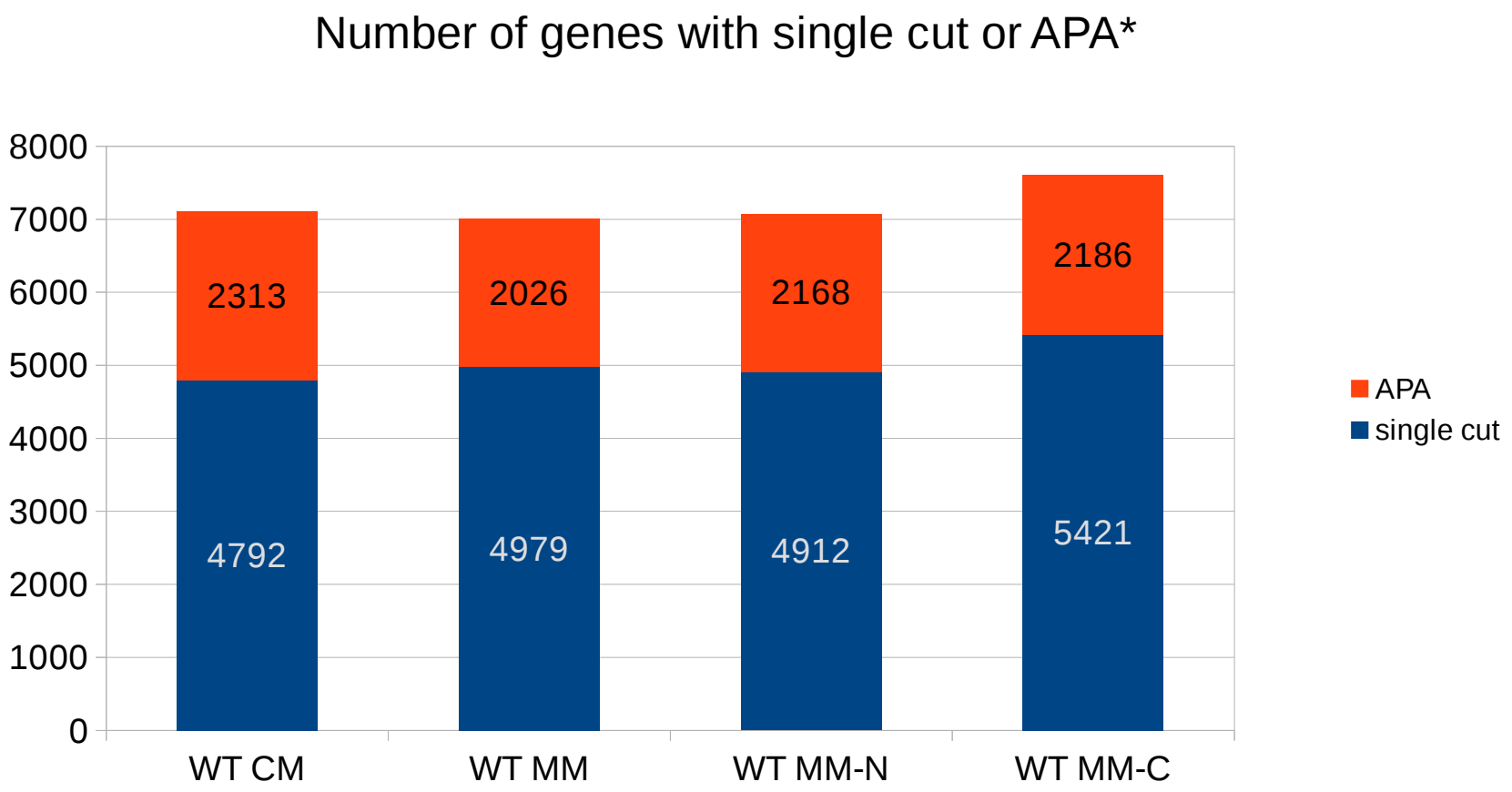
# ~15000 poly(A) site could be assigned to annotated genes (*Δrbp35*)



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

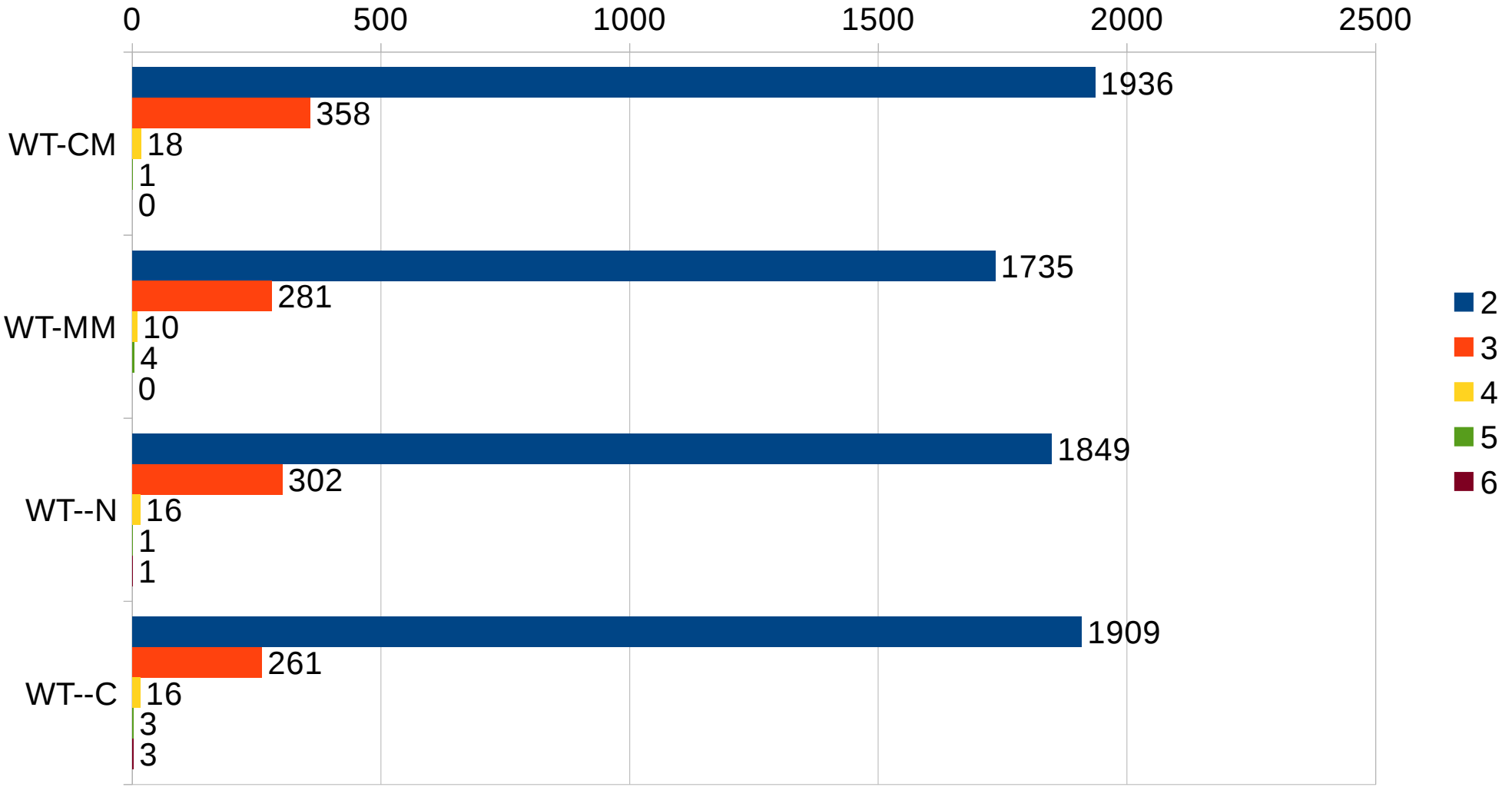


# ~30% of genes are alternatively polyadenilated

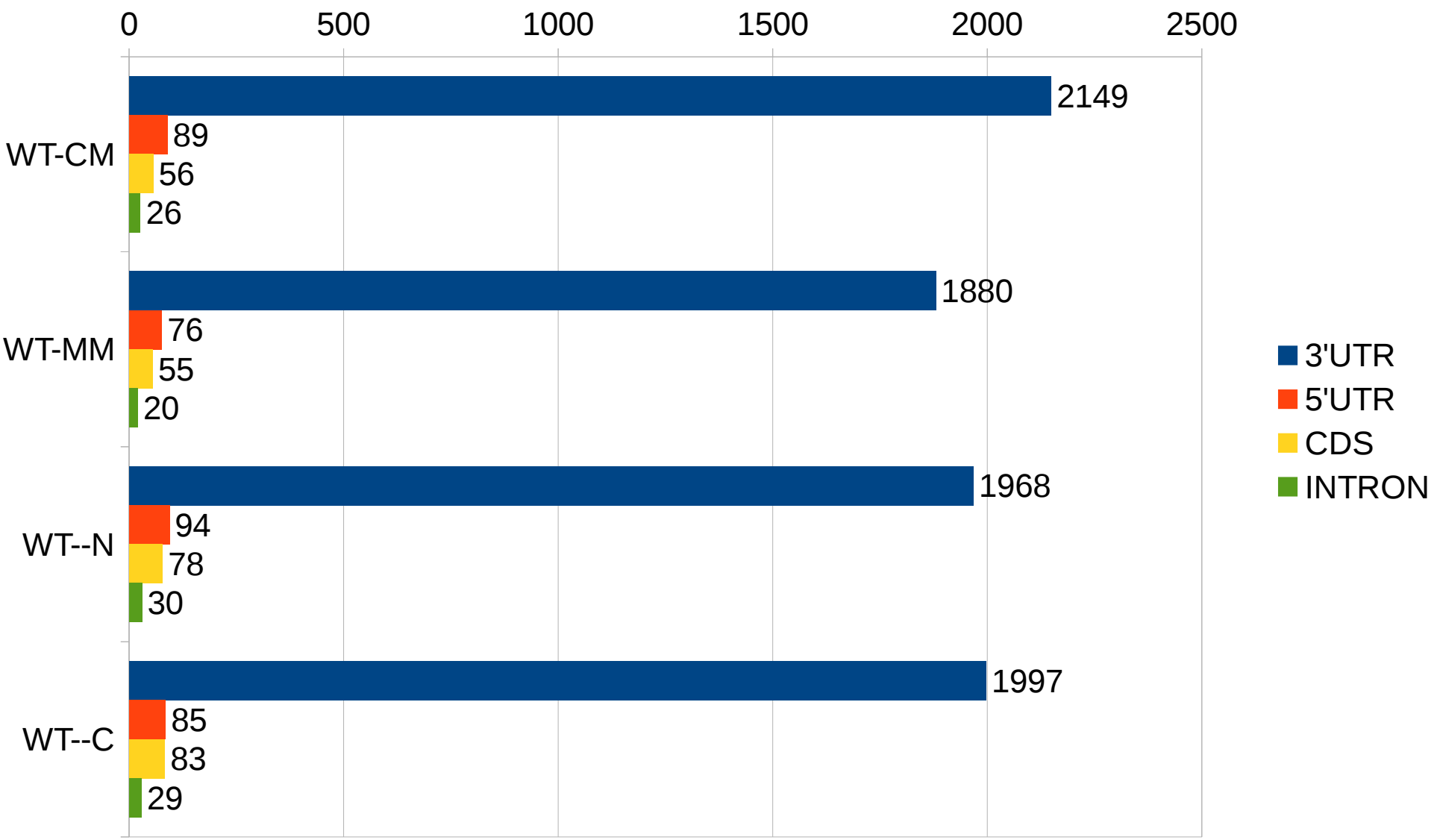



** calculated over the global number of expressed genes with a recognizable poly(A) site*

# >80% of APA is composed of two cleveage sites



# >90% of APA are tandem poly(A) sites in the 3'UTR



# 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
- 81 overlapping annotated genes antisense
- 63 matching CPA-sRNA sequences
- 16 matching retrotransposons
- 7 located in telemeric avirulence regions

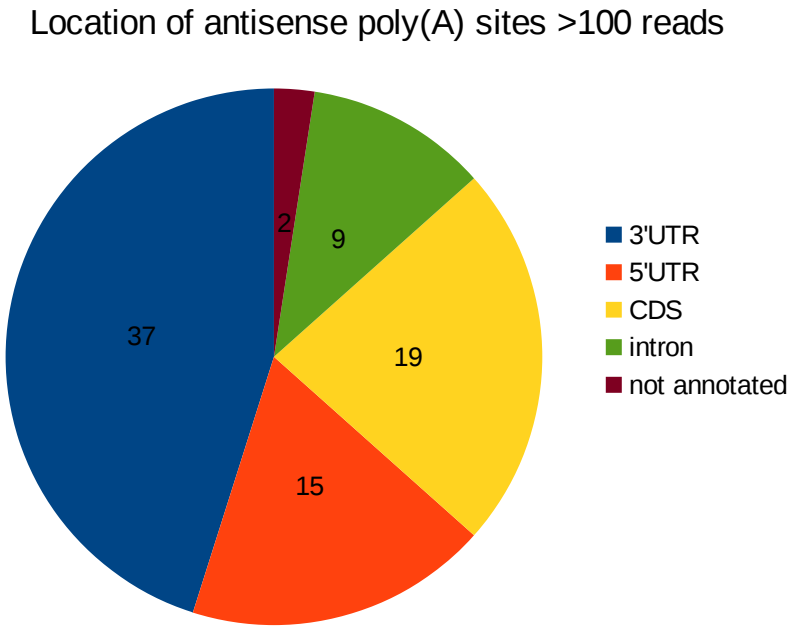
# 3165 orphan poly(A) sites expressed in WT (>10 reads)

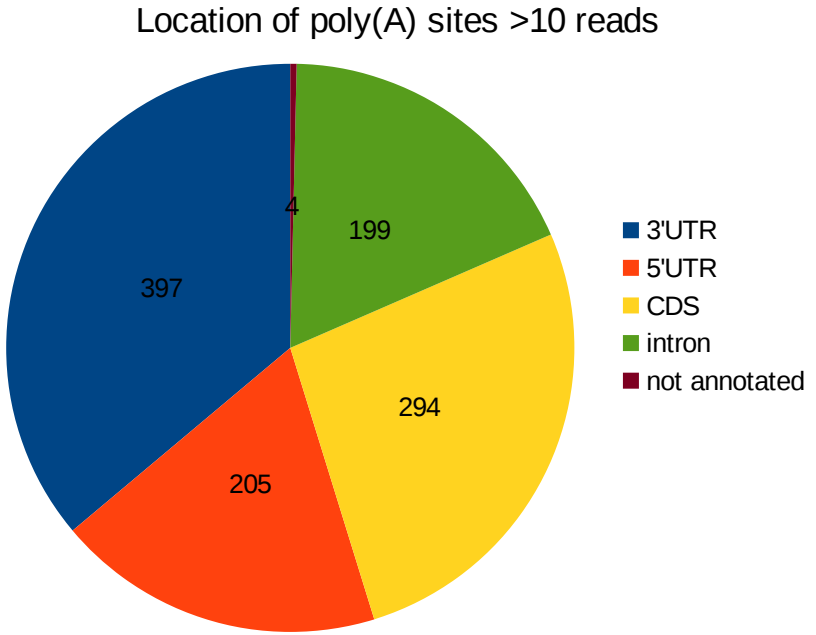
- 102 hits against other gene copies in *M.oryzae*
- 438 hits against Uniprot nt/nr database
- 10 hits against Rfam(ncRNA) database
- 1098 overlapping annotated genes antisense
- 253 matching CPA-sRNA sequences
- 129 matching retrotransposons
- 57 located in telemeric avirulence regions

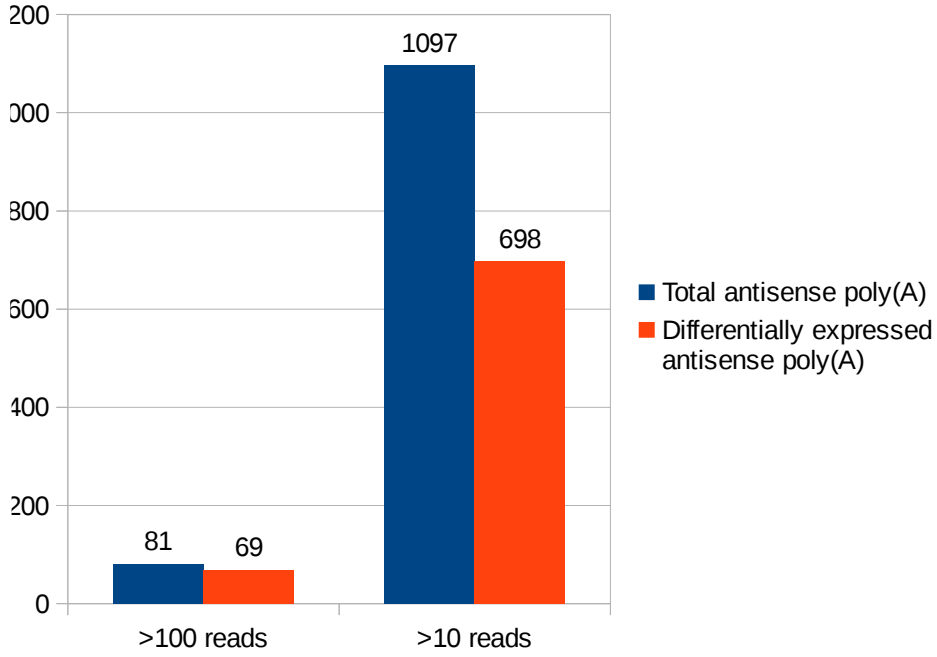
# Orphans differentially expressed in WT

| | |
|---|---|
| (>100 reads) | |
| CM → MM-C | 167 |
| CM → MM | 36 |
| CM → MM-N | 51 |
| MM → MM-C | 129 |
| MM → MM-N | 0 |
| | |
| (>10 reads) | |
| CM → MM-C | 1499 |
| CM → MM | 177 |
| CM → MM-N | 285 |
| MM → MM-C | 1110 |
| MM → MM-N | 0 |

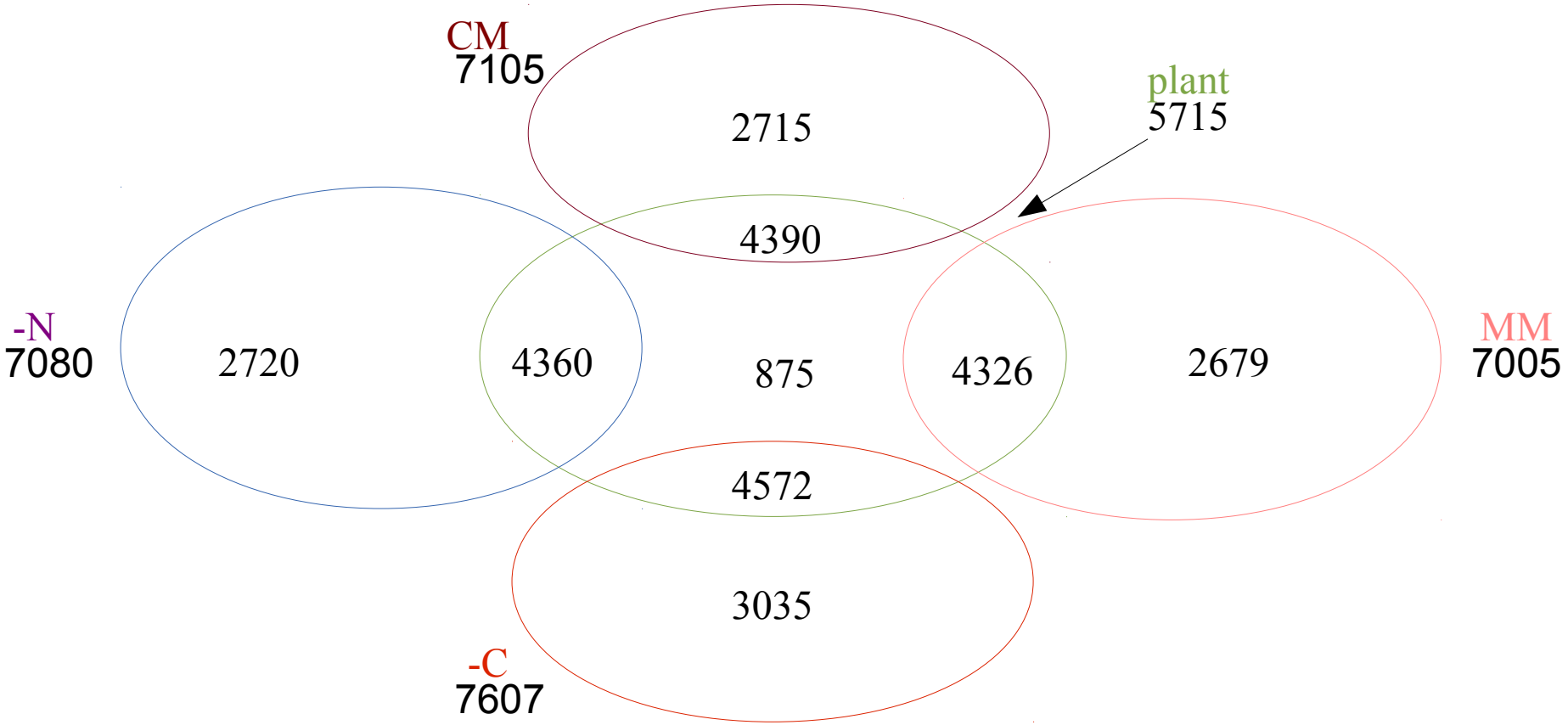
# Antisense poly(A) are usually located in the 3'UTR, most of antisense poly(A) are differentially expressed in any condition



# 875 genes expressed in plant are never expressed in vitro

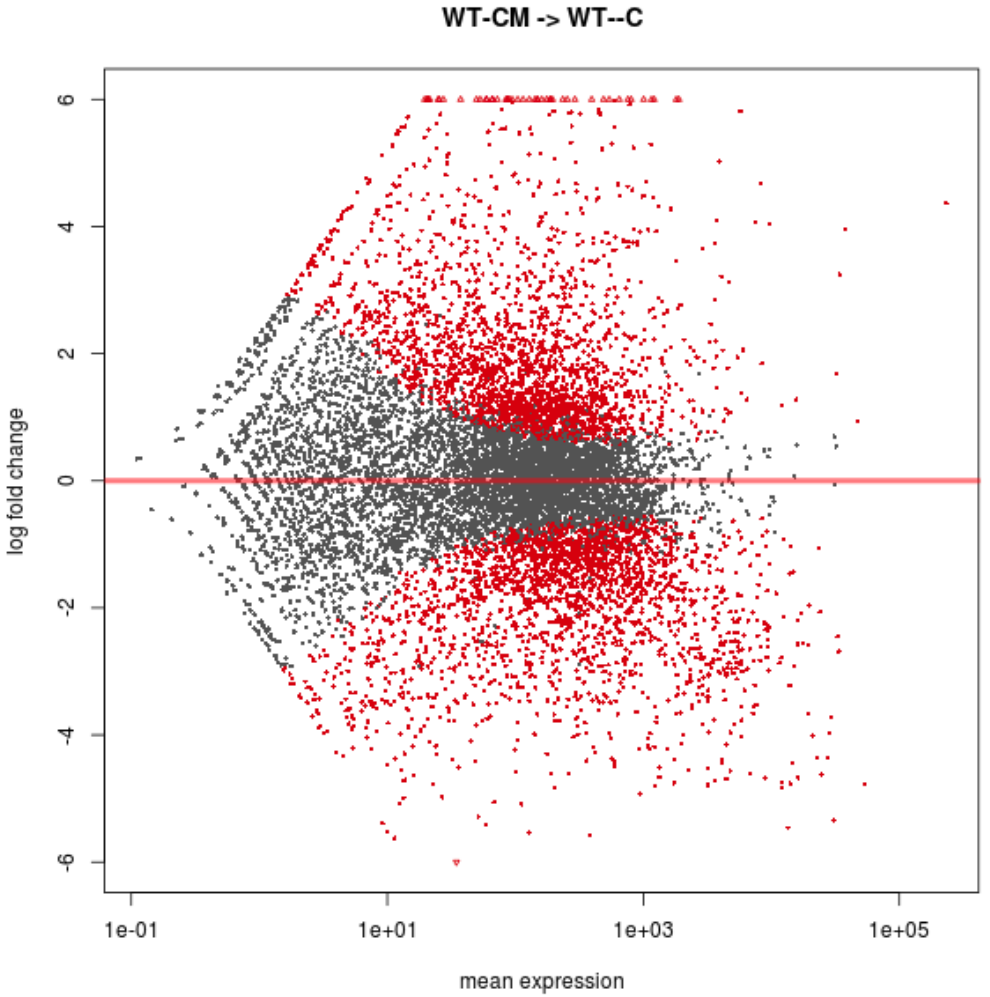


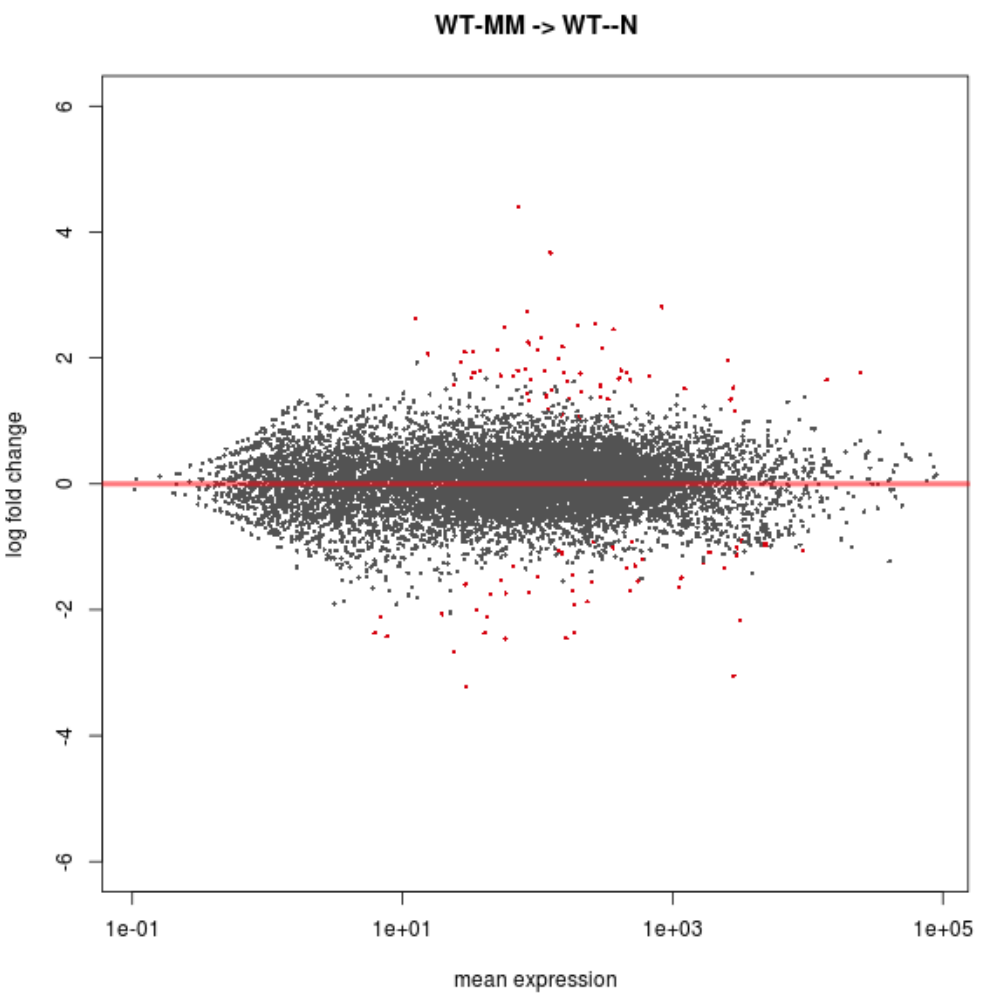

Sage + mosquera = 5715 genes
875 of these last ones. never found in our experiment

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

| DIFFERENTIALLY EXPRESSED GENES IN THE WT | | | |
|---|---|---|---|
| | DOWN | UP | TOTAL |
| CM → MM | 314 | 559 | *873* |
| CM → MM-N | 630 | 874 | *1504* |
| CM → MM-C | 2307 | 2342 | *4649* |
| MM → MM-N | 48 | 59 | *107* |
| MM → MM-C | 1882 | 1589 | *3471* |

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



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

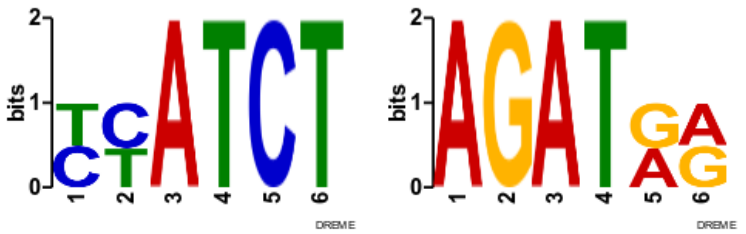
## TOP 20 HIGHEST DIFFERENTIALLY EXPRESSED GENES CM → -C

| gene | log2foldChange | description |
|---|---|---|
| MGG_09072 | 8.1538523515 | Alcohol oxidase |
| MGG_00244 | 8.007087145 | 115-hydroxyprostaglandin dehydrogenase |
| MGG_07210 | 7.9942521054 | Putative uncharacterized protein |
| MGG_09607 | 7.4595914704 | Maltose permease MAL31 |
| MGG_01367 | 7.2171221376 | Putative uncharacterized protein |
| MGG_07253 | 7.1375757829 | Putative uncharacterized protein |
| MGG_11289 | 7.0894568362 | Putative uncharacterized protein |
| MGG_08937 | 7.0375631939 | Quinate permease |
| MGG_15267 | 6.9971932885 | Putative uncharacterized protein |
| MGG_03793 | 6.898822123 | 82,3-dihydroxybenzoic acid decarboxylase |
| MGG_06828 | 6.872238681 | Putative uncharacterized protein |
| MGG_05941 | 6.8217314329 | Maltose permease MAL31 |
| MGG_02245 | 6.7008914884 | Endoglucanase type F |
| MGG_00659 | 6.6866120353 | Glucan 1,3-beta-glucosidase |
| MGG_10663 | 6.4805171445 | cAMP-regulated D2 protein |

| gene | log2foldChange | description |
|---|---|---|
| MGG_17996 | -5.6066608795 | no_description |
| MGG_07973 | -5.3026451979 | Surface protein 1 |
| MGG_08019 | -5.156698481 | F-box domain-containing protein |
| MGG_06234 | -5.0149833139 | Putative uncharacterized protein |
| MGG_04258 | -4.9786465581 | Putative uncharacterized protein |
| MGG_01952 | -4.8319947186 | Putative uncharacterized protein |
| MGG_17706 | -4.7704849827 | Putative uncharacterized protein |
| MGG_10456 | -4.7234757359 | Putative uncharacterized protein |
| MGG_09015 | -4.700976561 | Putative uncharacterized protein |
| MGG_17103 | -4.6464513062 | Putative uncharacterized protein |
| MGG_08360 | -4.6275010826 | Putative uncharacterized protein |
| MGG_17677 | -4.5723814478 | Putative uncharacterized protein |
| MGG_11608 | -4.5509909646 | Laccase-2 |
| MGG_05344 | -4.5356873581 | SnodProt1 |
| MGG_07966 | -4.5297712533 | Phosphate transporter |

# Up-regulated genes in nitrogen starvation show the typical GATA motif in the promoter region

**1. YYATCT**




Submit

Download

**Details**

| Positives | Negatives | P-value | E-value | Unerased E-value |
|---|---|---|---|---|
| 463/955 | 234/700 | 5.1e-10 | 0.0000028 | 0.0000028 |

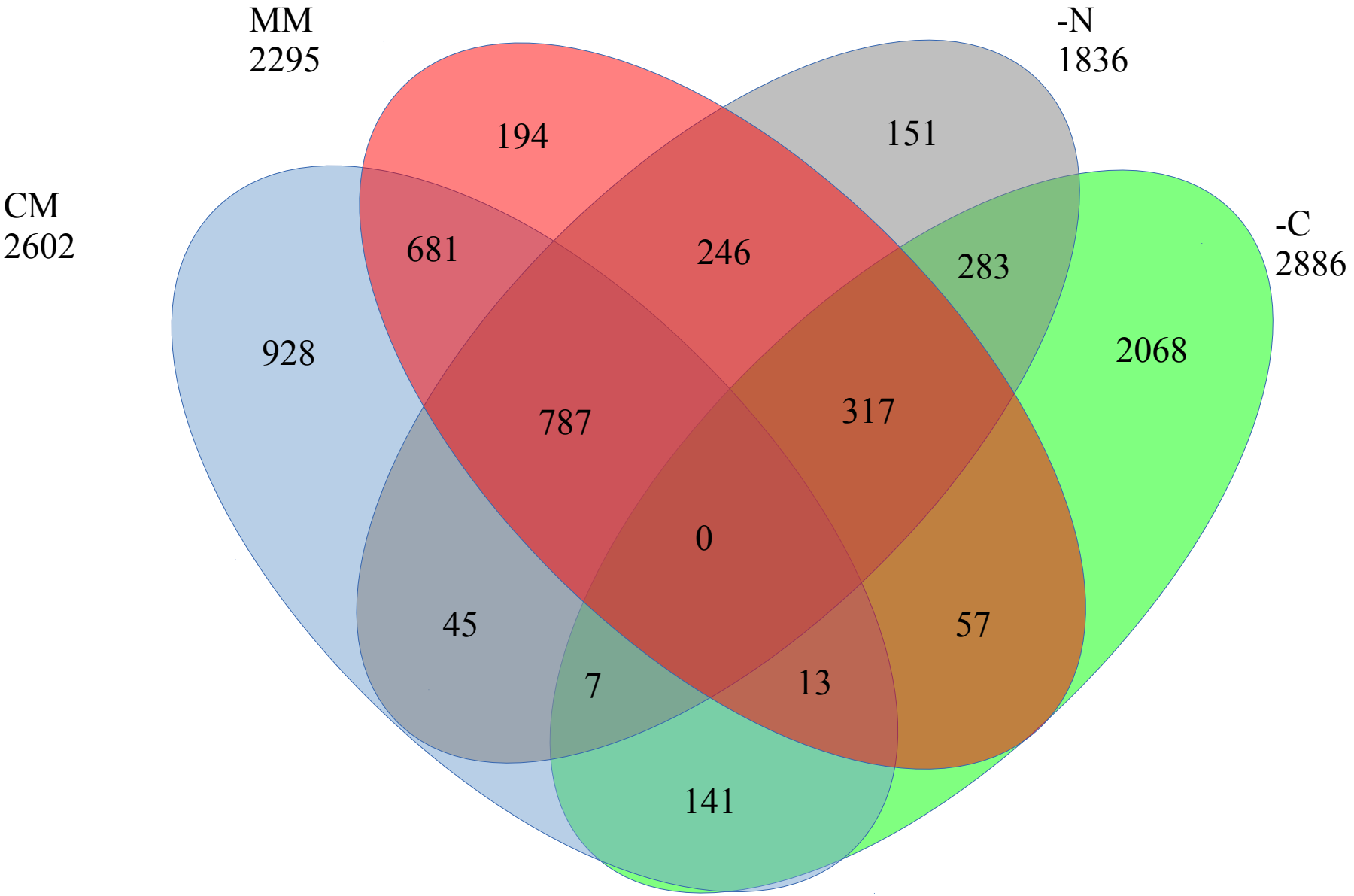
**Enriched Matching Words**

| Word | Positives | Negatives | P-value | ▼E-value |
|---|---|---|---|---|
| CTATCT | 122/955 | 49/700 | 7.3e-5 | 4.1e-1 |
| TTATCT | 150/955 | 67/700 | 1.4e-4 | 8.0e-1 |
| TCATCT | 146/955 | 74/700 | 3.1e-3 | 1.7e+1 |
| CCATCT | 163/955 | 88/700 | 6.8e-3 | 3.8e+1 |

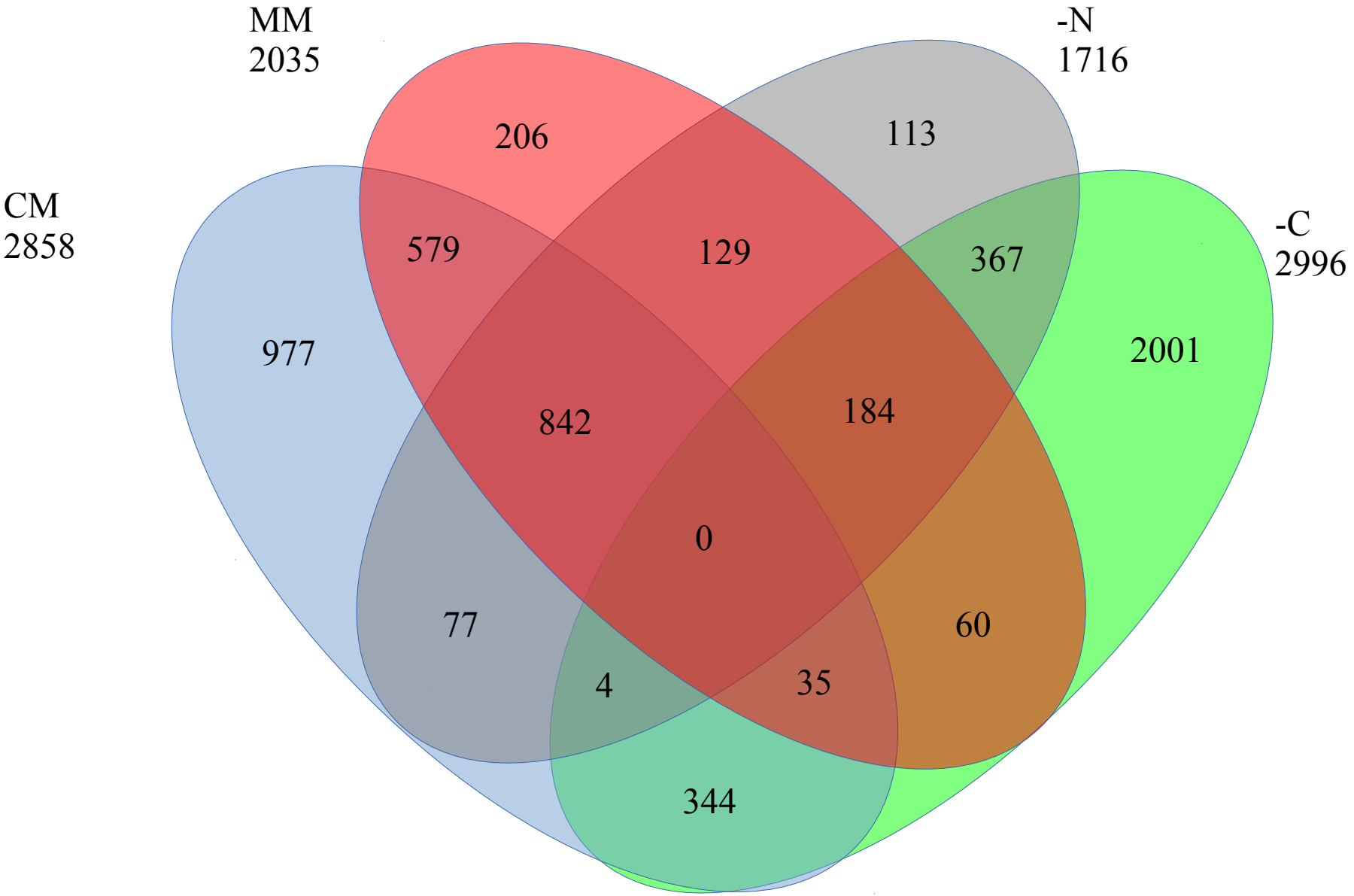
- *None of the known GATA-binding transcription factors is found to be significantly up-regulated in our 12h NS experiment*
- *NUT1, the important nitrogen-related TF is found to be generally down-regulated in Carbon-starvation*
- *30 of the 51 top up-regulated genes listed in www.ncbi.nlm.nih.gov/pubmed/16731015 are confirmed in our experiment*

- *Only two Transcription factors, MGG_05829 and MGG_01486, probably related with purine Regulation, are found to be up-regulated in MM → -N, MGG_05829 is down-regulated in the mutant*

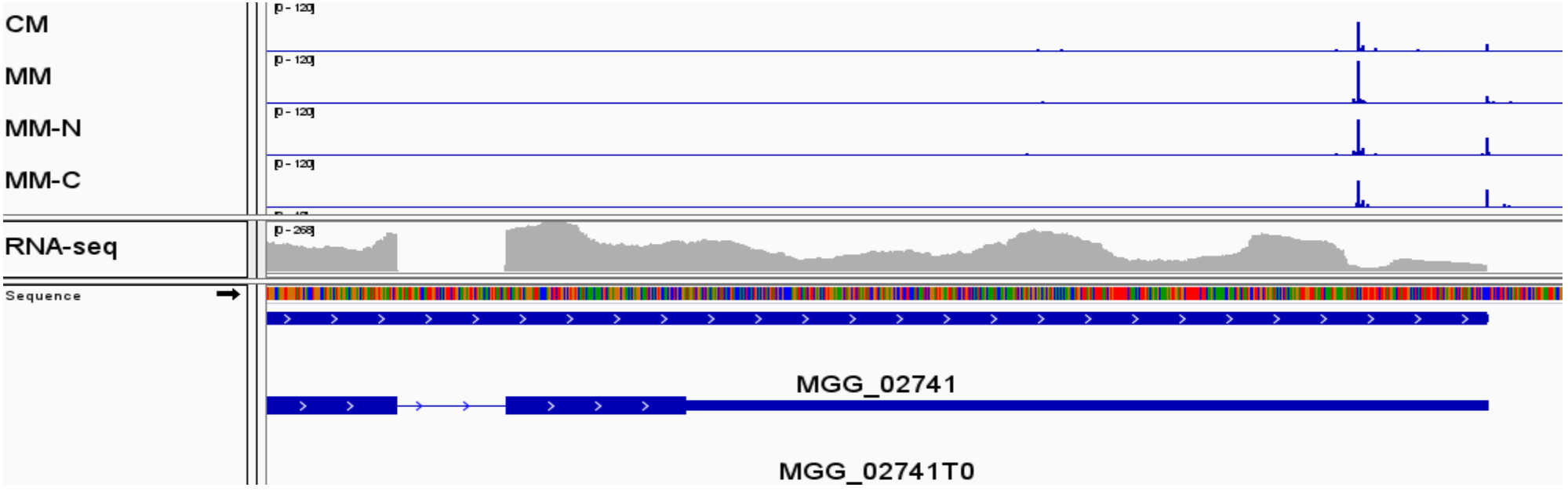
# Up-regulated genes between conditions

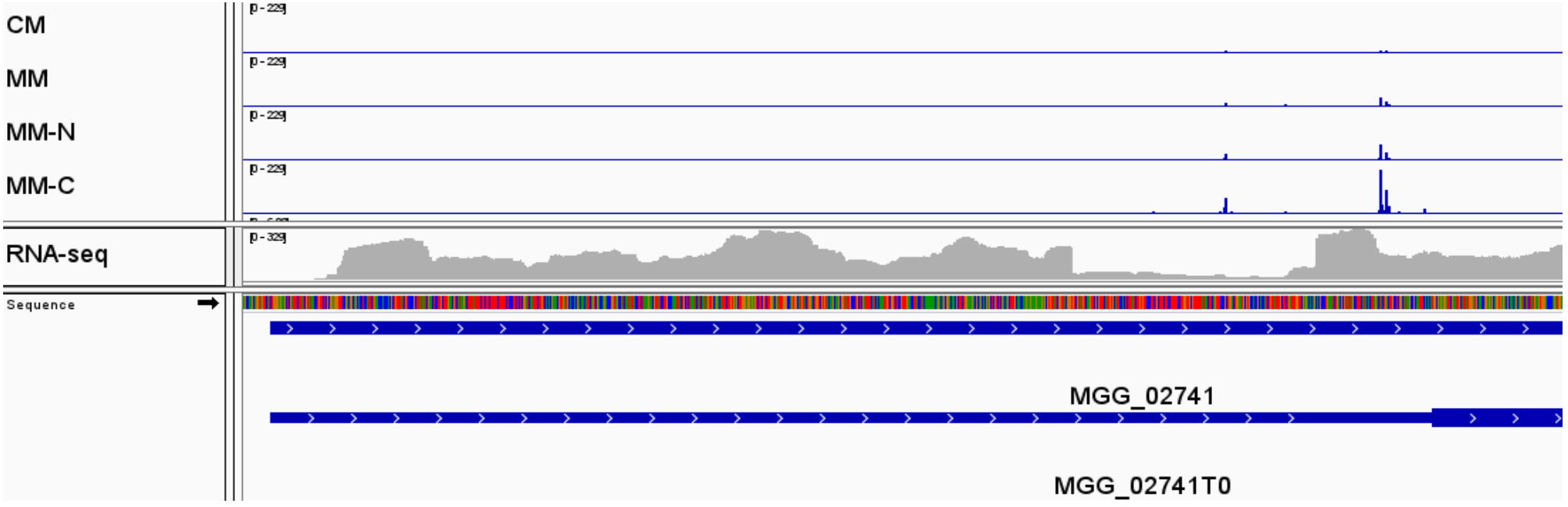


# Down-regulated genes between conditions

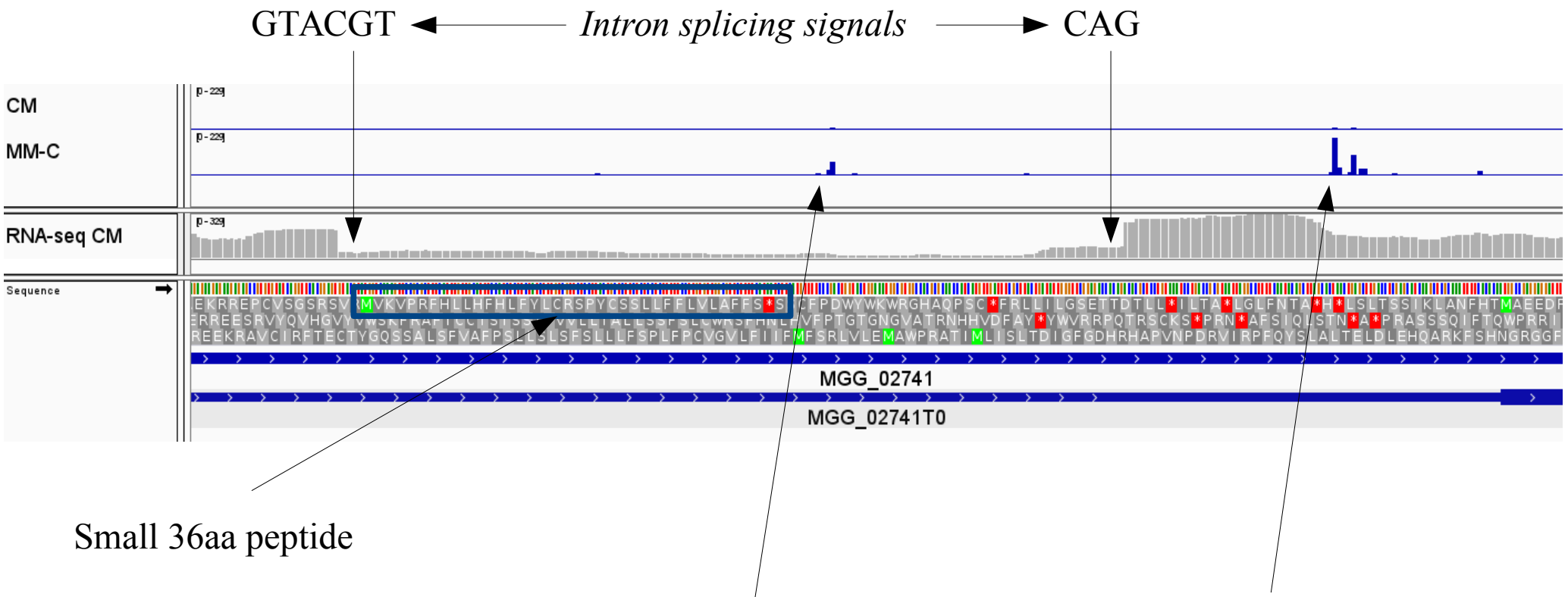


# RBP35 shows different polyadenilation in each medium, with strong differencies in MM-C



# RBP35 shows an alternative polyadenylated 5'UTR, putatively enconding a small peptide

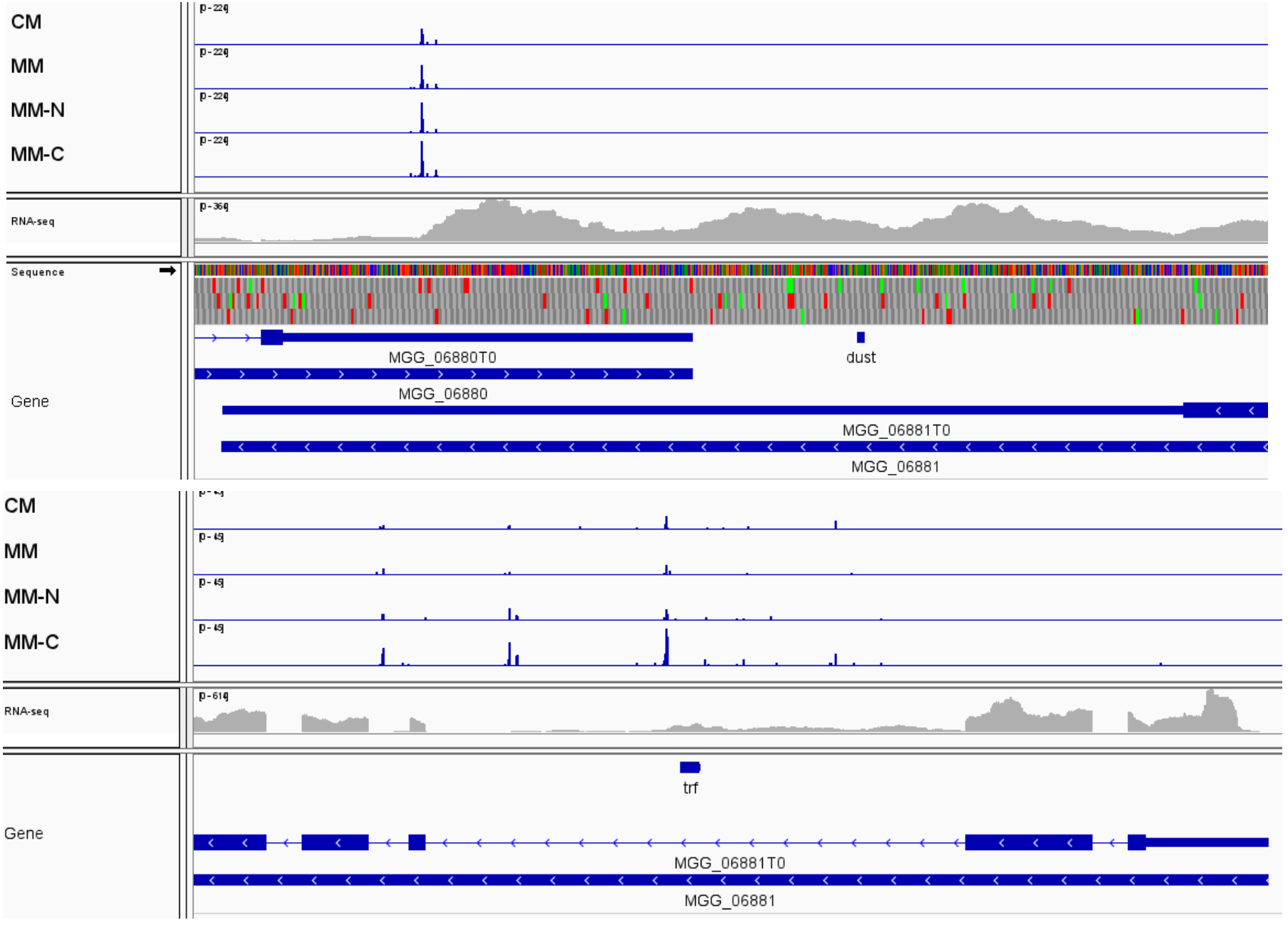



Small 36aa peptide

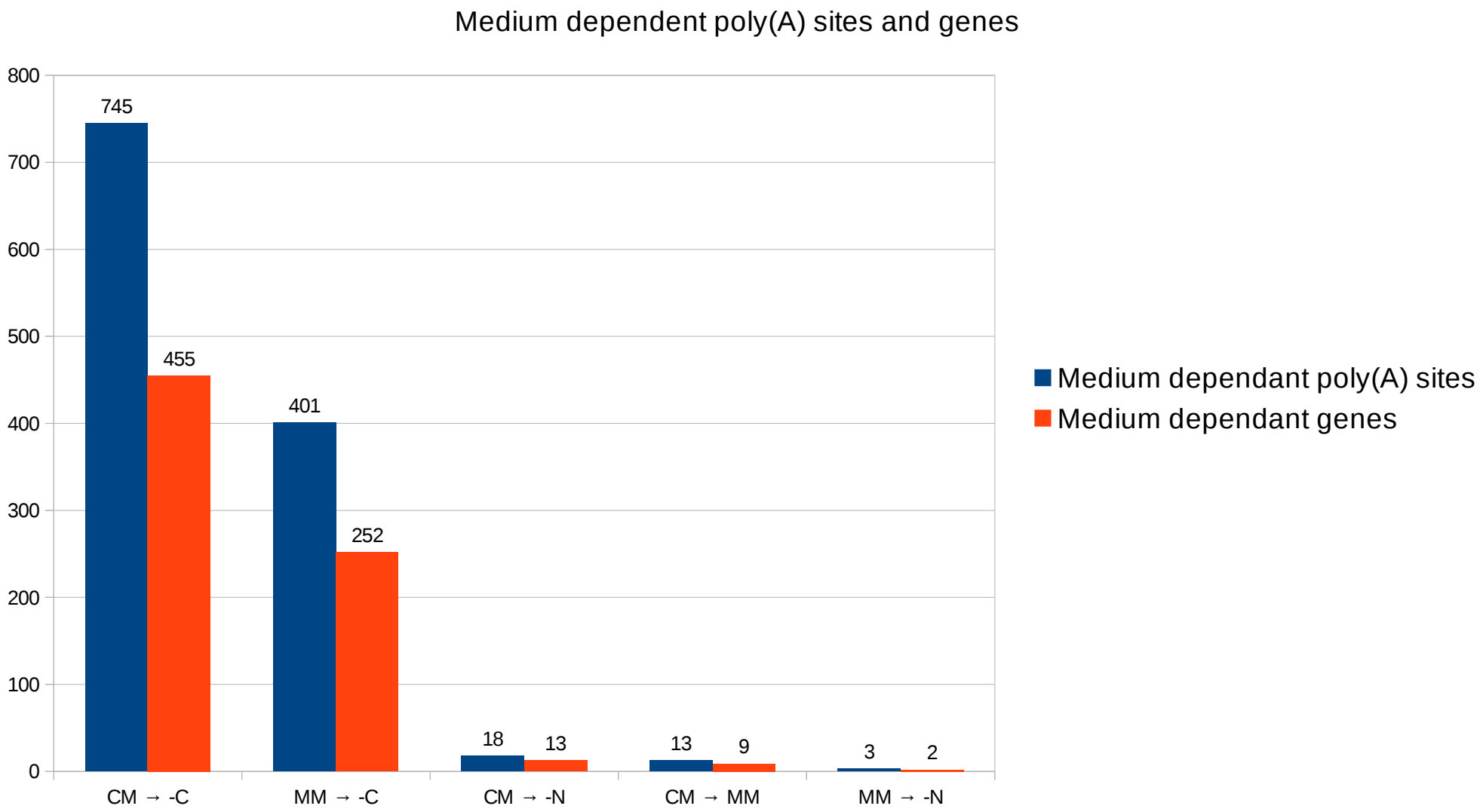
Poly(A) site used in absense of splicing, when the small petide **is** transcribed

Poly(A) site used after splicing, when The small petide **is not** transcribed
*Up-regulated poly(A) site in MM-C*

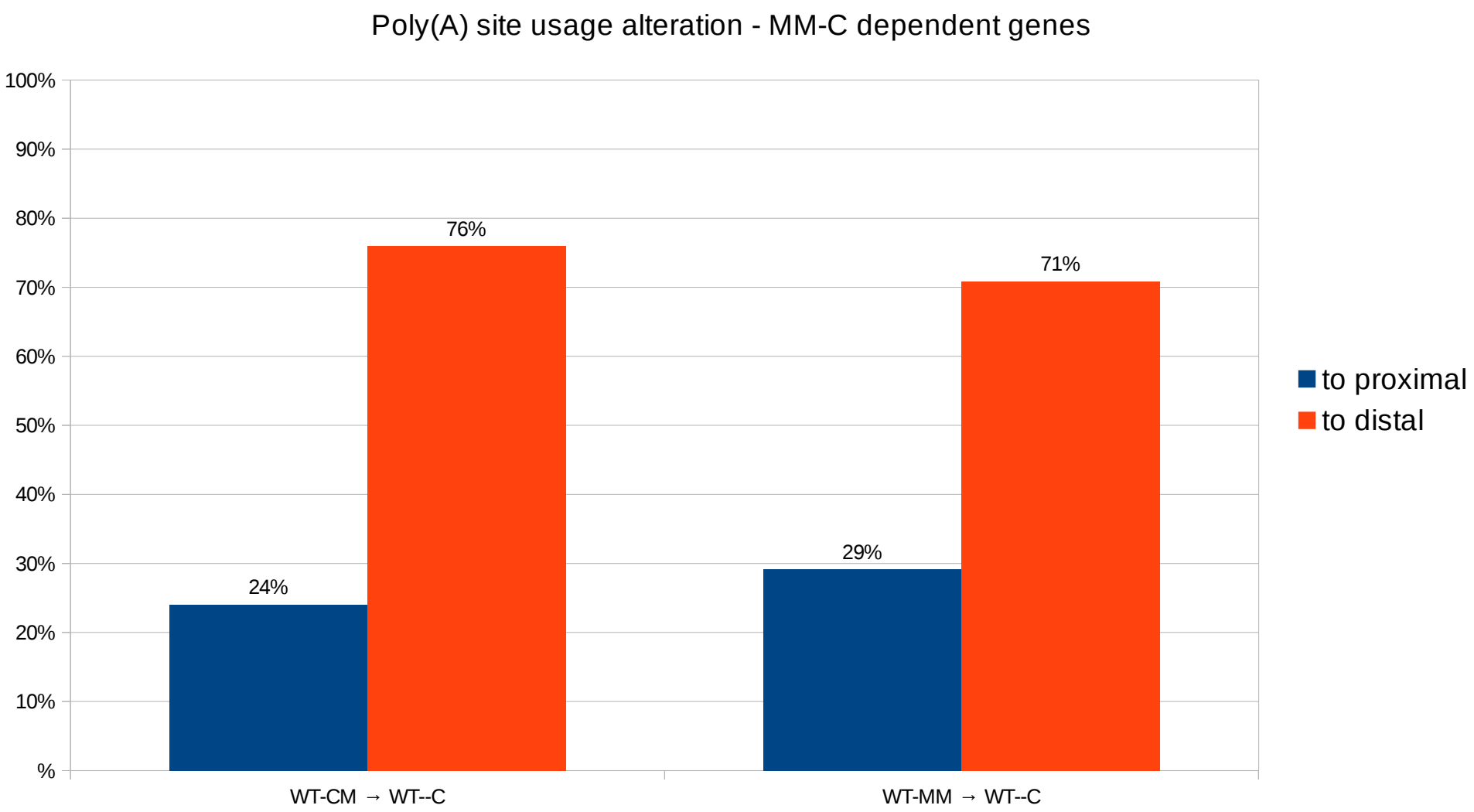
# HRP1 shows an up-regulated intronic poly(A) site in MM-C



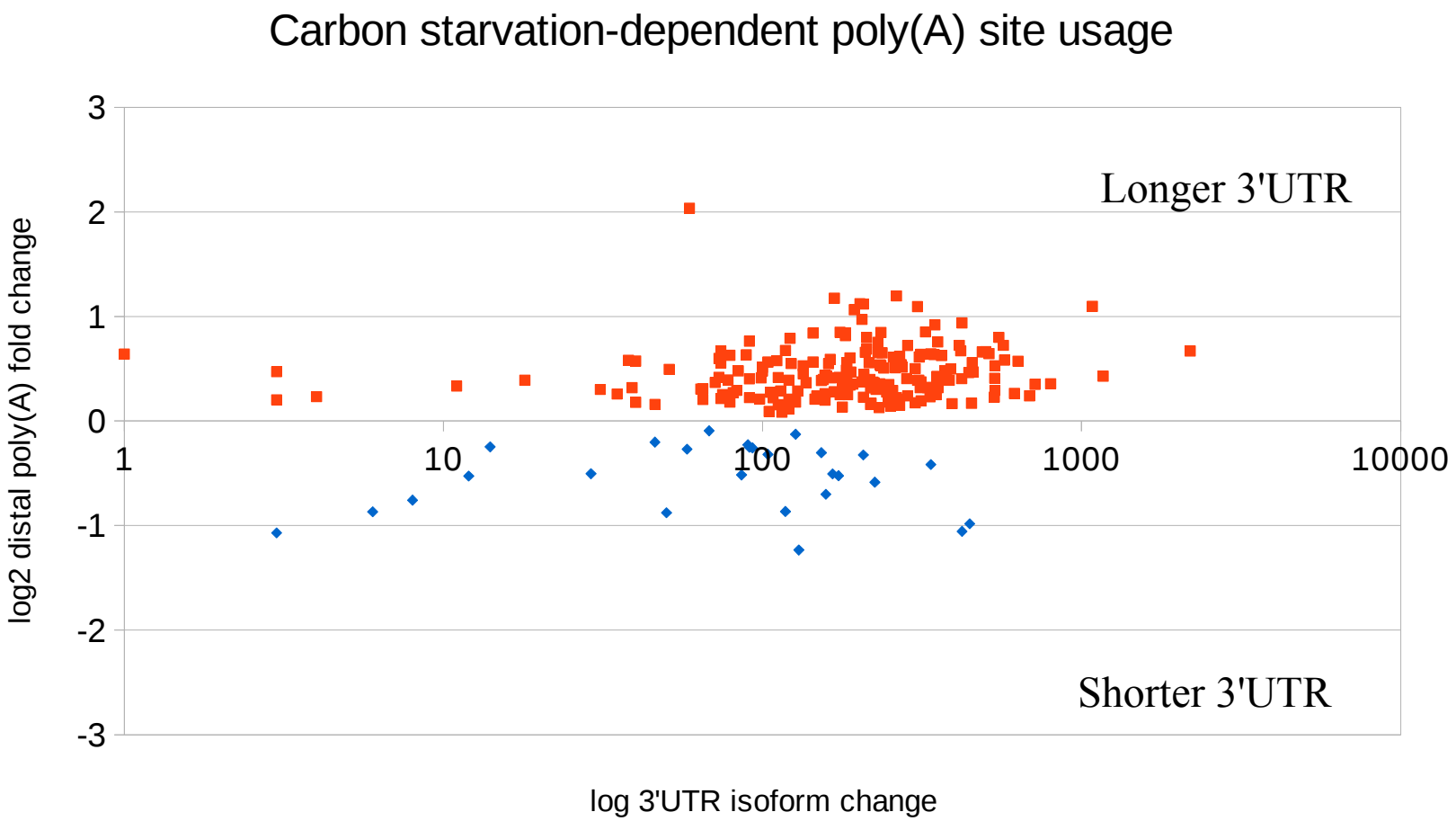
# Carbon starvation affects a great number of poly(A) sites



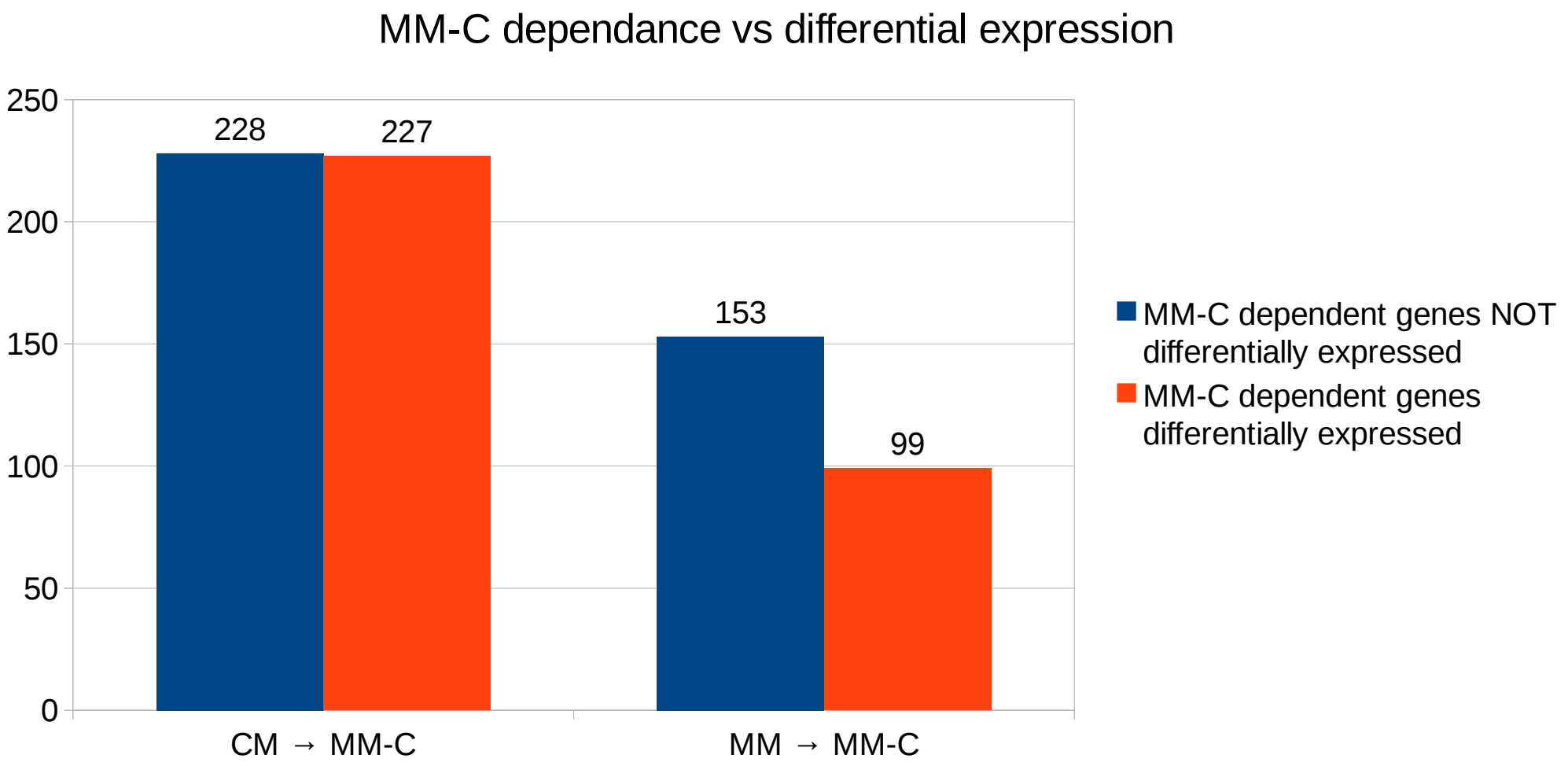
# Carbon starvation affects poly(A) sites usage, preferring distal cuts



# Carbon starvation affects poly(A) sites usage, preferring distal cuts

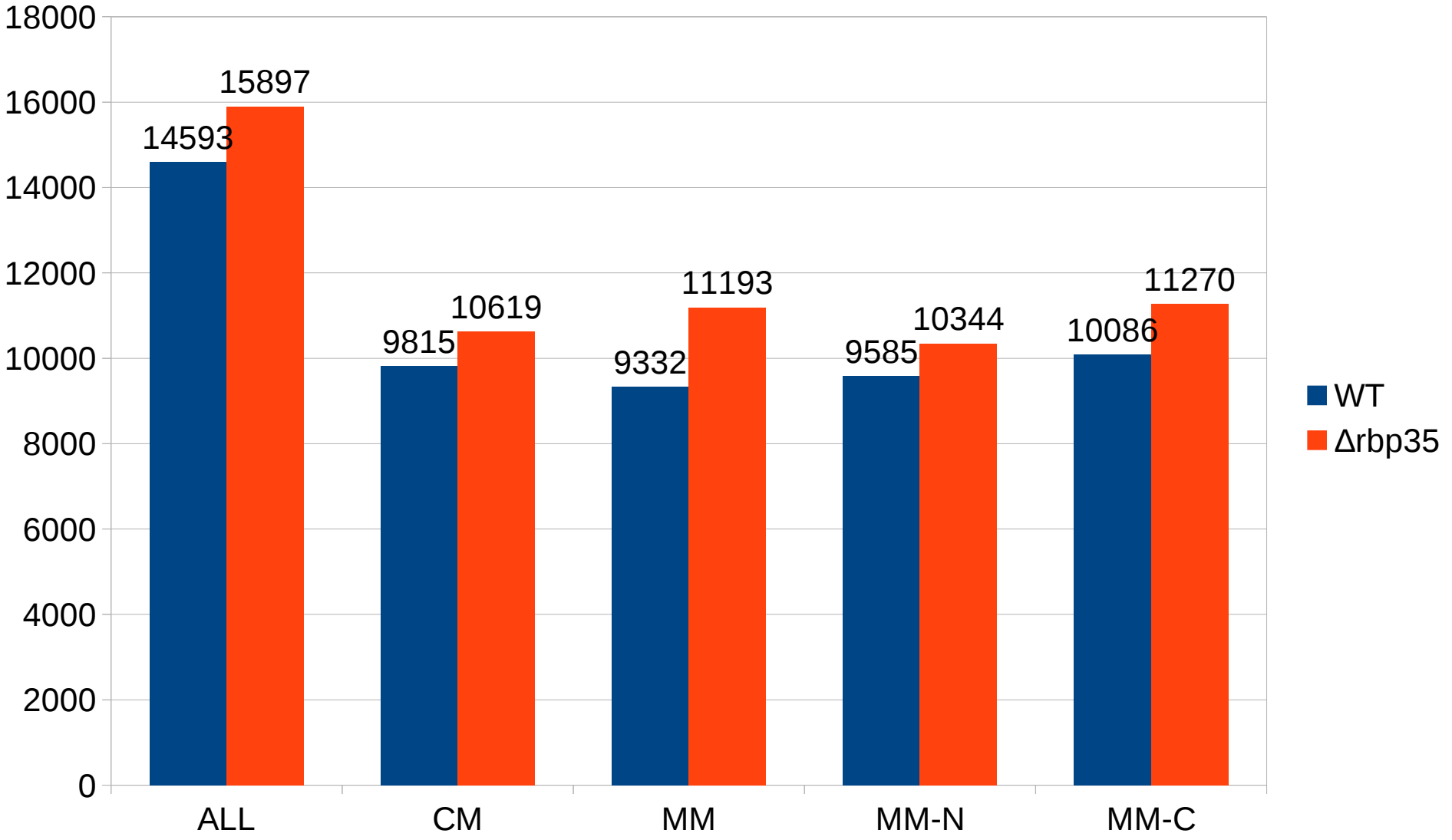


# MM-C dependent genes are usually differentially expressed

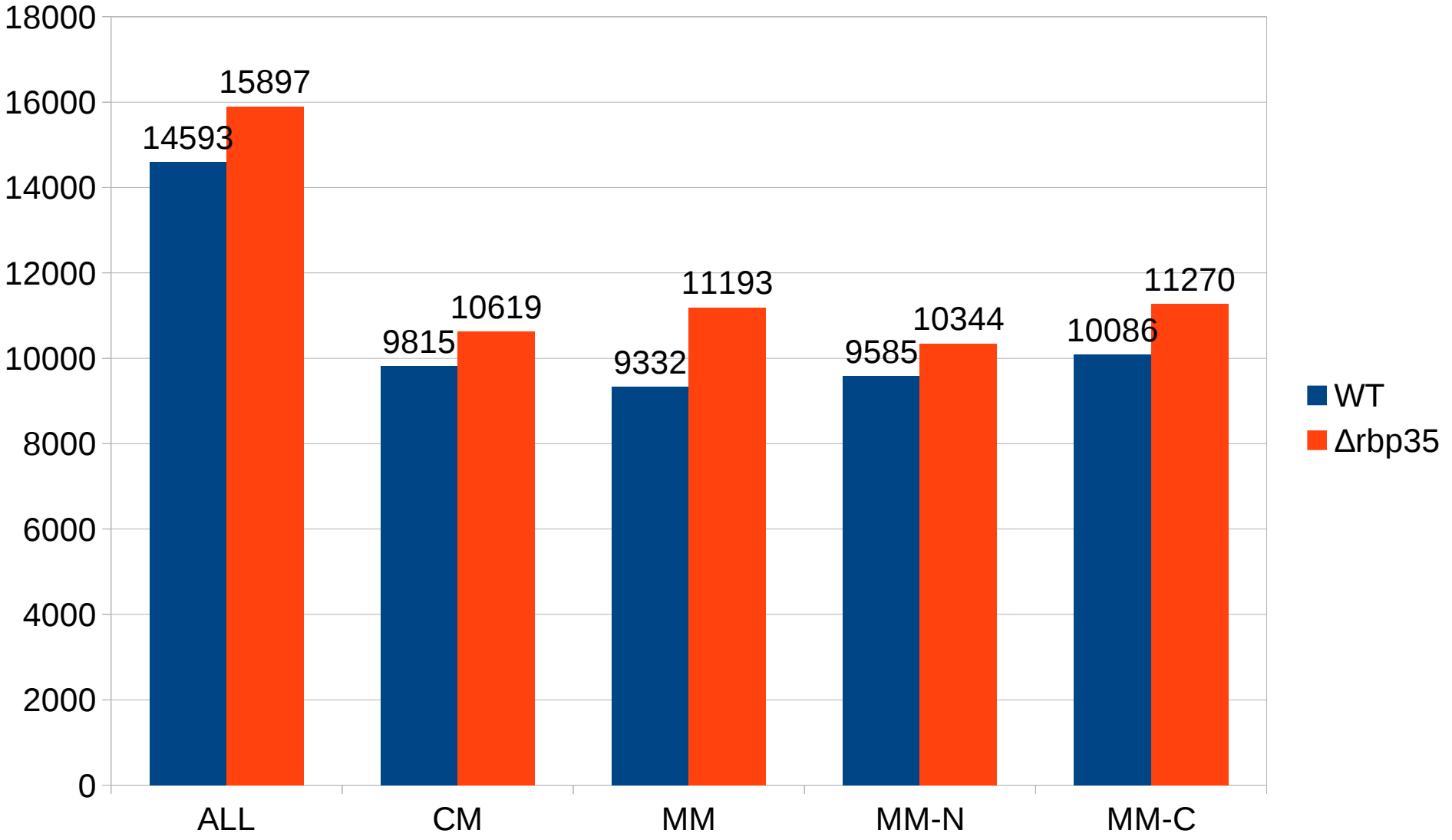



*Differentially expressed gene are equally distributed between up & down regulated (data not shown)*

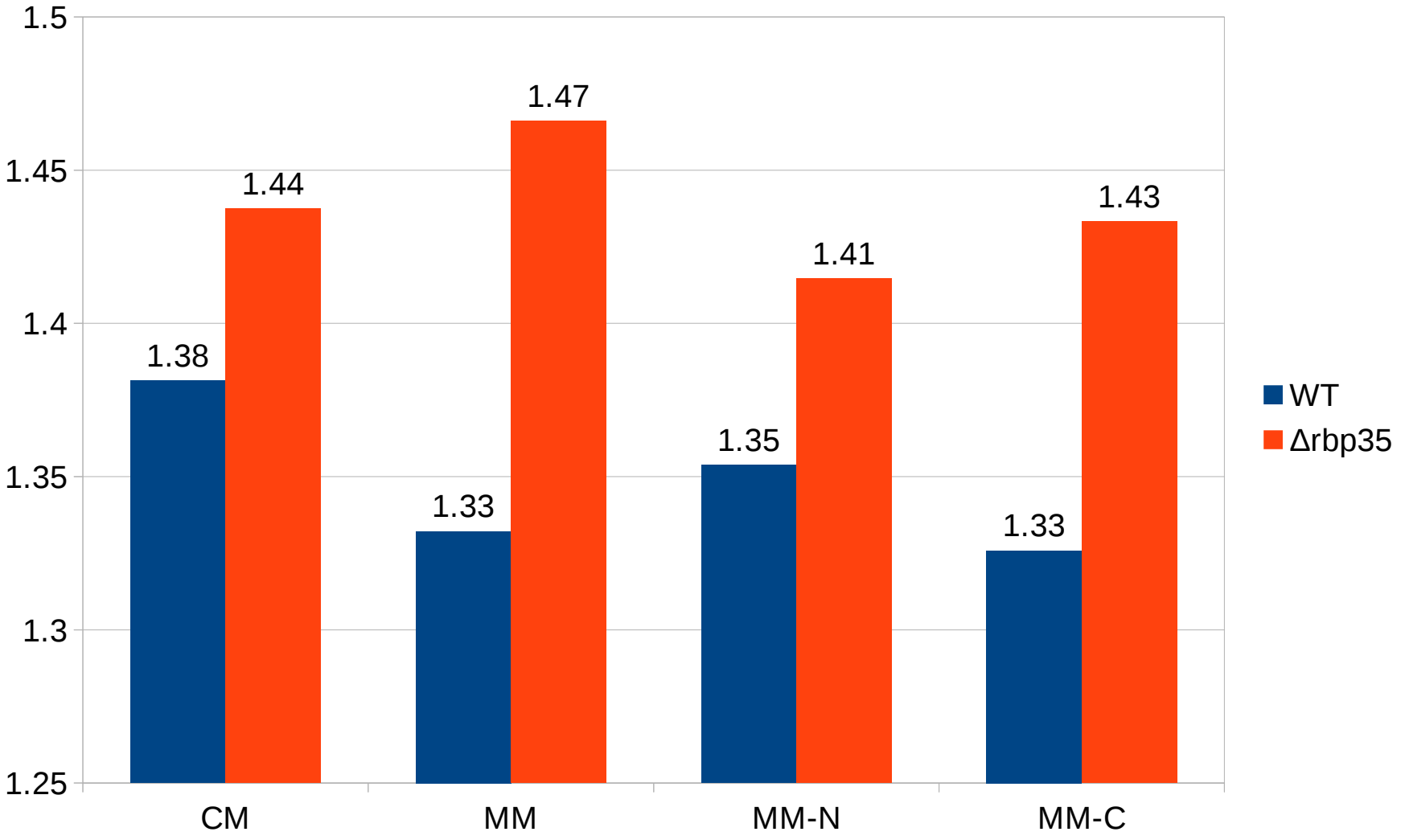
# *Δrbp35* affects poly(A) sites number



# *Δrbp35* affects poly(A) sites number



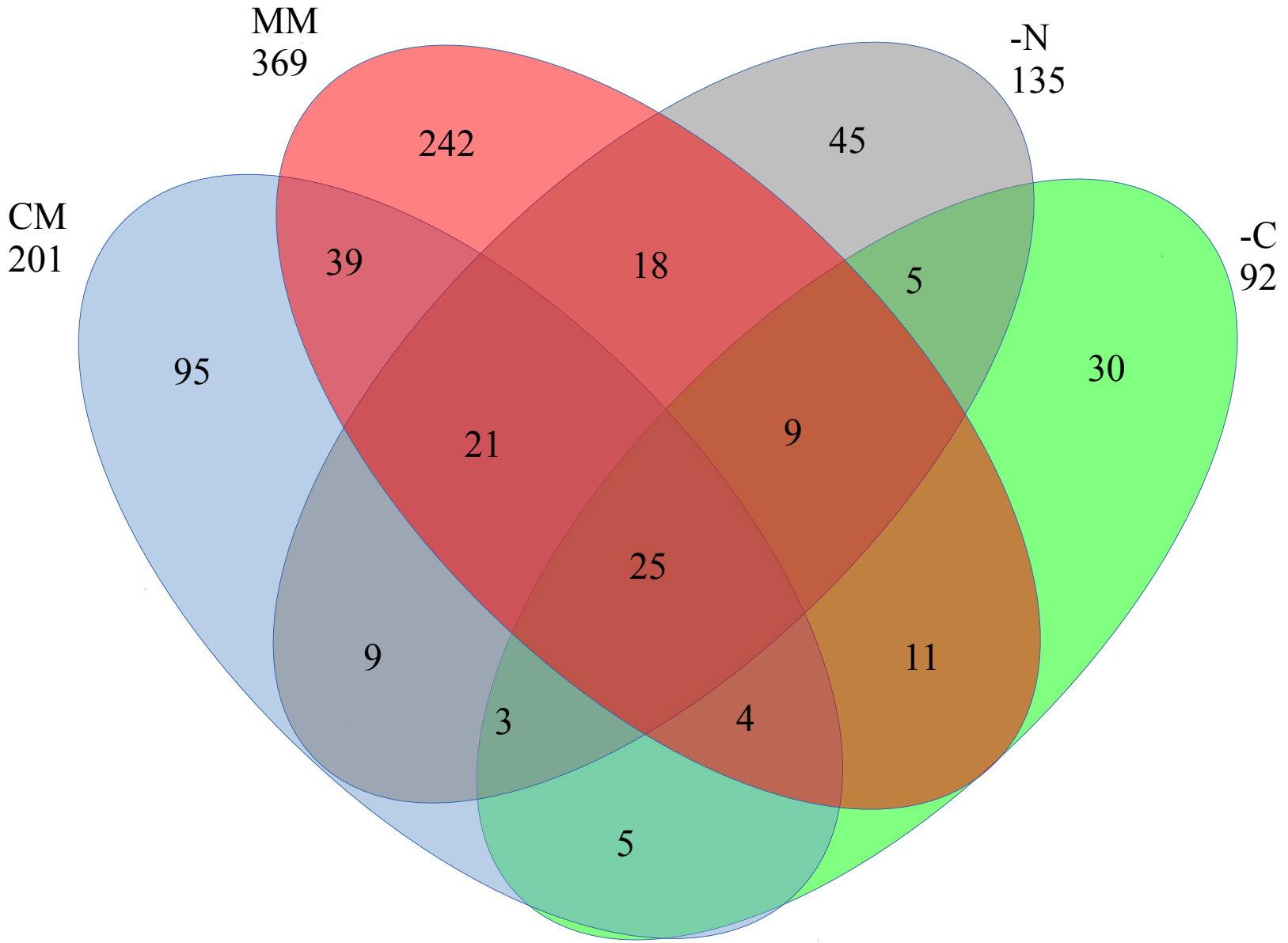
# *Δrbp35* affects number of cut sites per gene



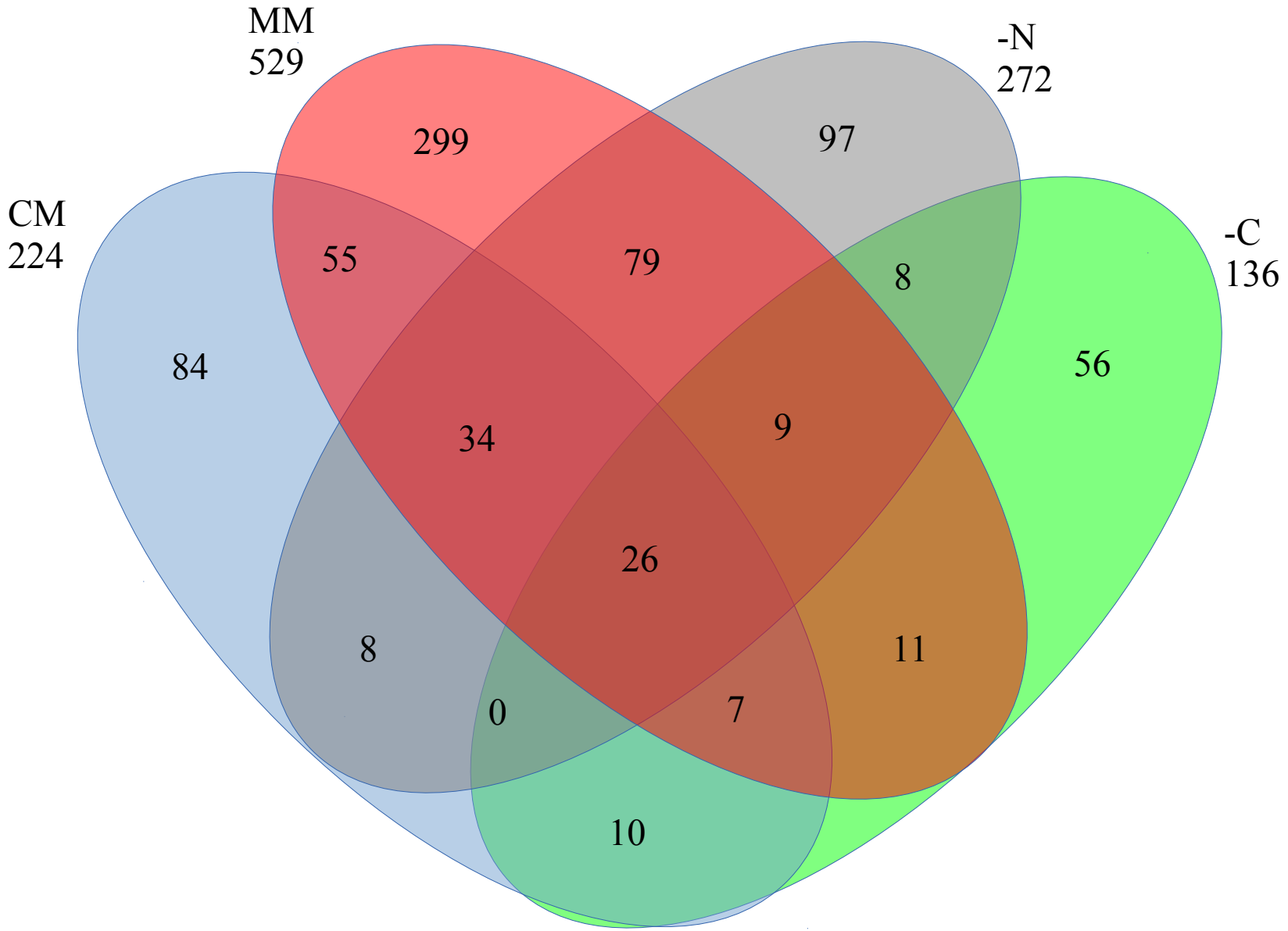
# MM is the most affected condition in Δ*rbp35*, MM-C the least affected

| | *down regulated genes* | *up regulated gene* | *total* |
|---|---|---|---|
| WT → *Δrbp35* CM | 224 | 201 | 425 |
| WT → *Δrbp35* MM | 529 | 369 | 898 |
| WT→ *Δrbp35* MM-N | 272 | 135 | 407 |
| WT→ *Δrbp35* MM-C | 136 | 92 | 228 |

# Up-regulated genes WT → Δ*rbp35*



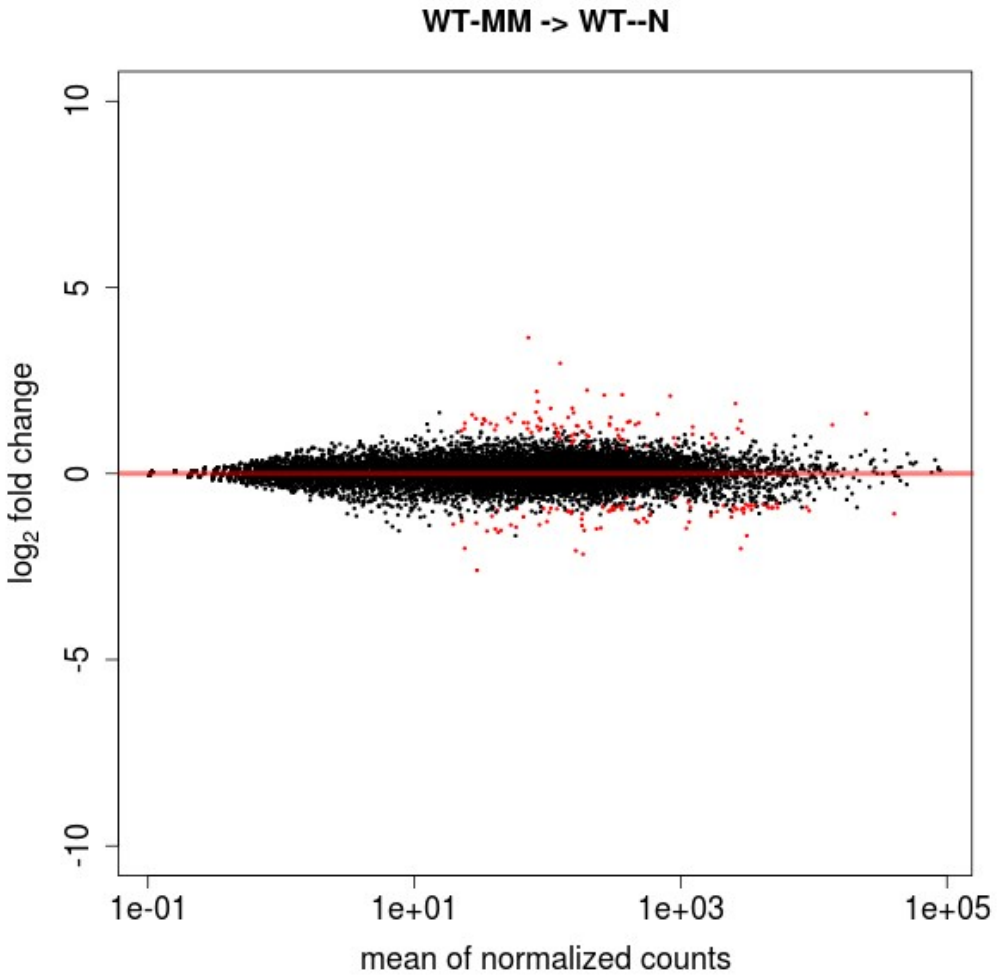
# Down-regulated genes WT → Δ*rbp35*

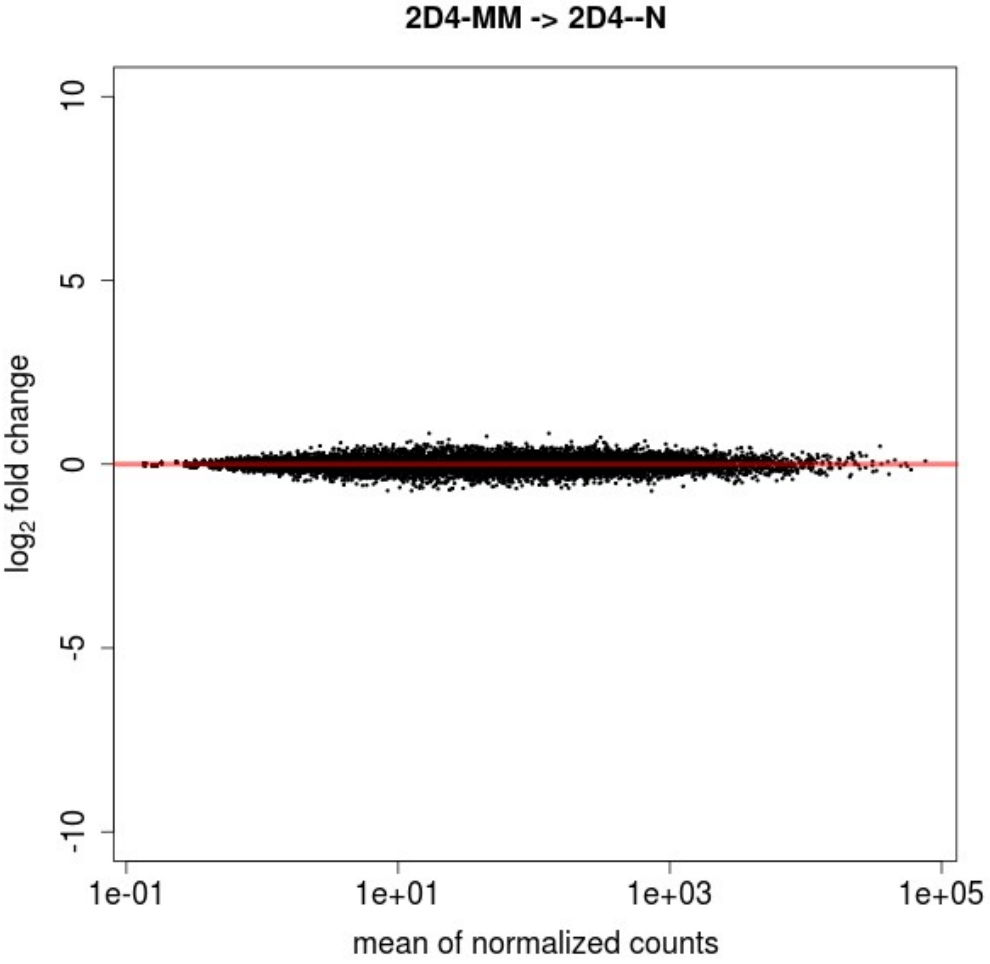


# *Δrbp35* appears to inhibit medium recognition in MM-N

| DIFFERENTIALLY EXPRESSED GENES IN Δrbp35 | | | |
|---|---|---|---|
| | | | |
| | DOWN | UP | TOTAL |
| CM → MM | 508 | 405 | 913 |
| CM → MM-N | 461 | 404 | 865 |
| CM → MM-C | 1241 | 1136 | 2377 |
| MM → MM-N | 0 | 0 | 0 |
| MM → MM-C | 475 | 493 | 968 |

# *Δrbp35* appears to inhibit medium recognition in MM-N



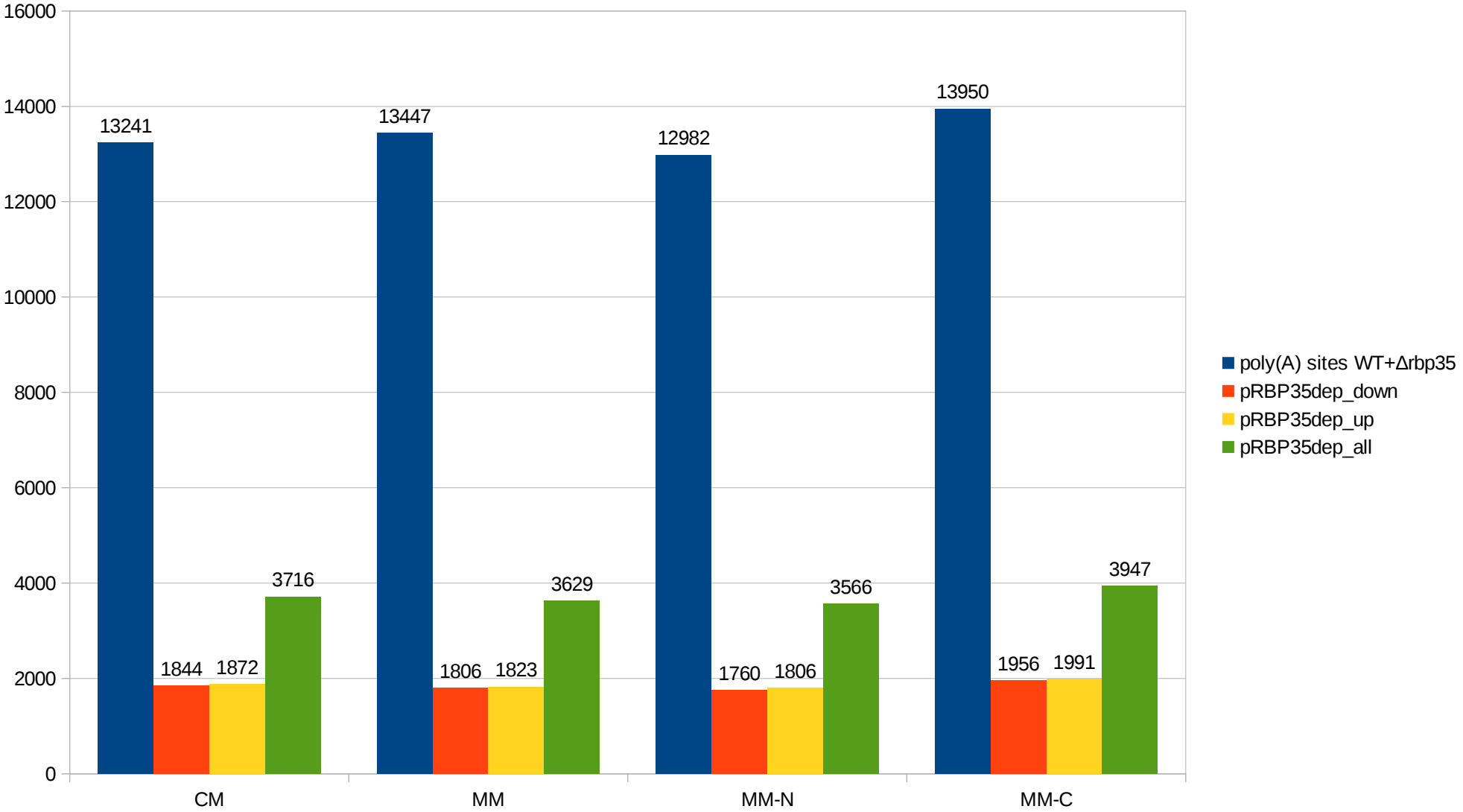
# Terminology

- **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):

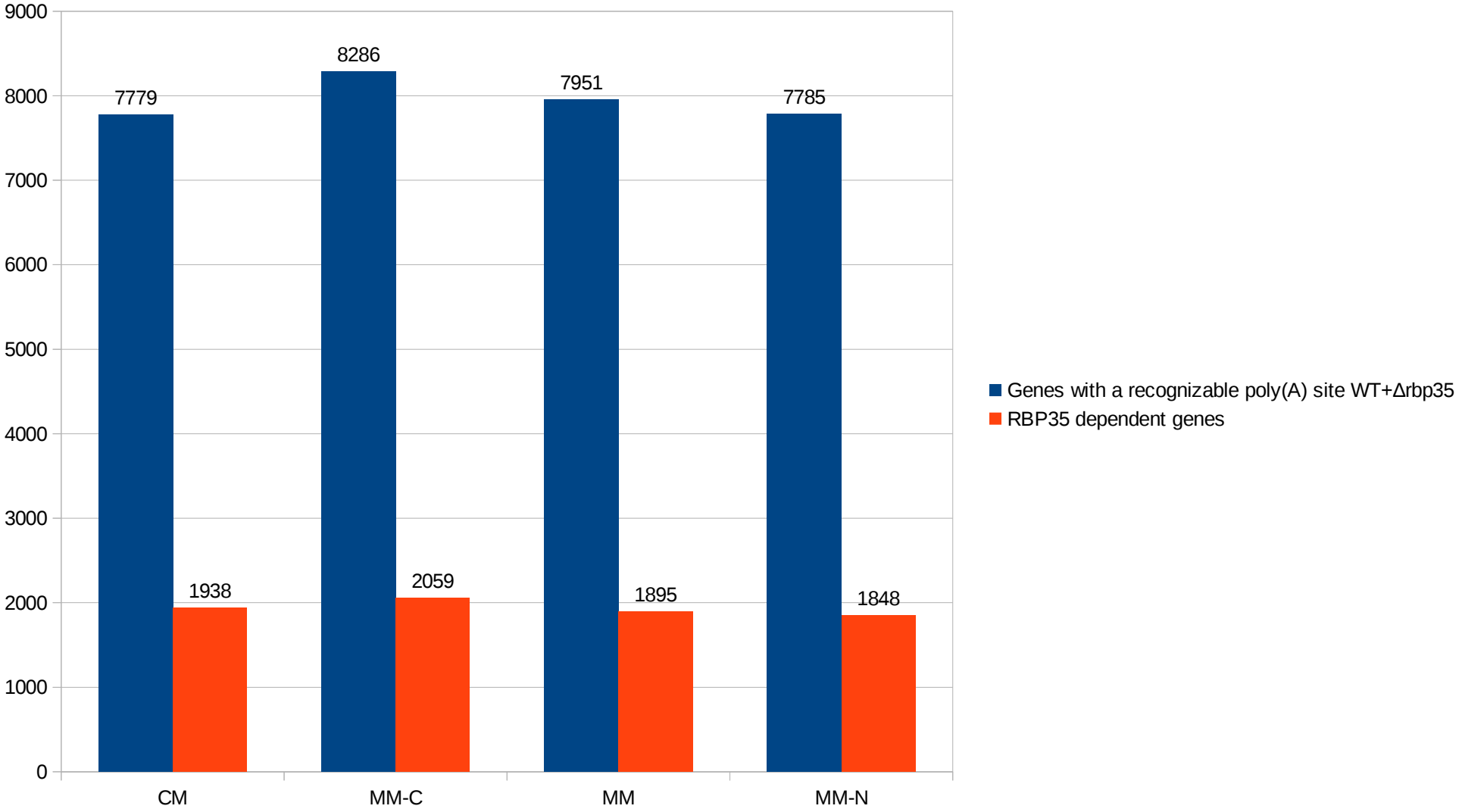

- **pRBP35dep_up**: an up-regulated RBP35 dependent poly(A):


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

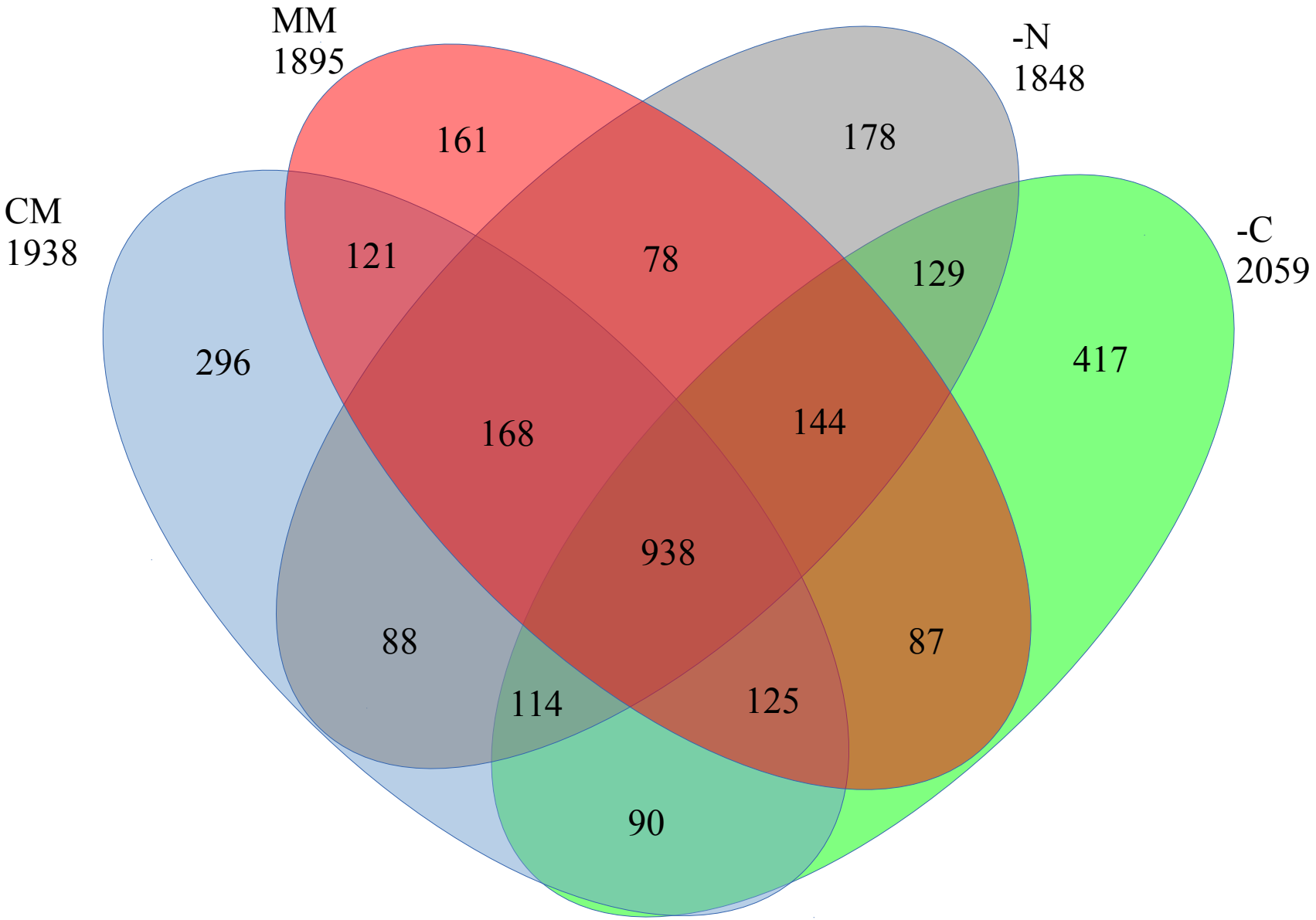
# 26%-28% of poly(A) sites are dependent from RBP35 in all media



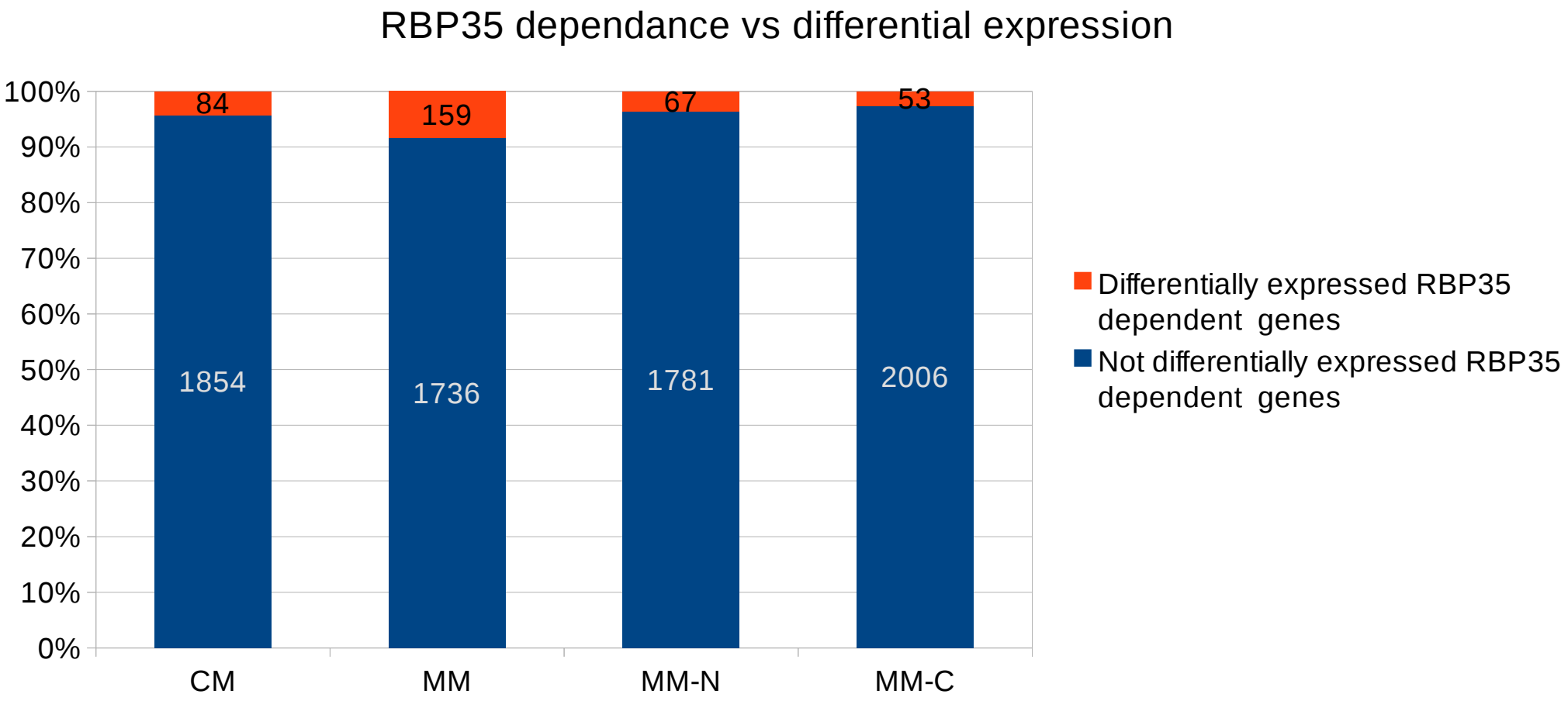
# ~25% of genes are dependent from RBP35 in all media



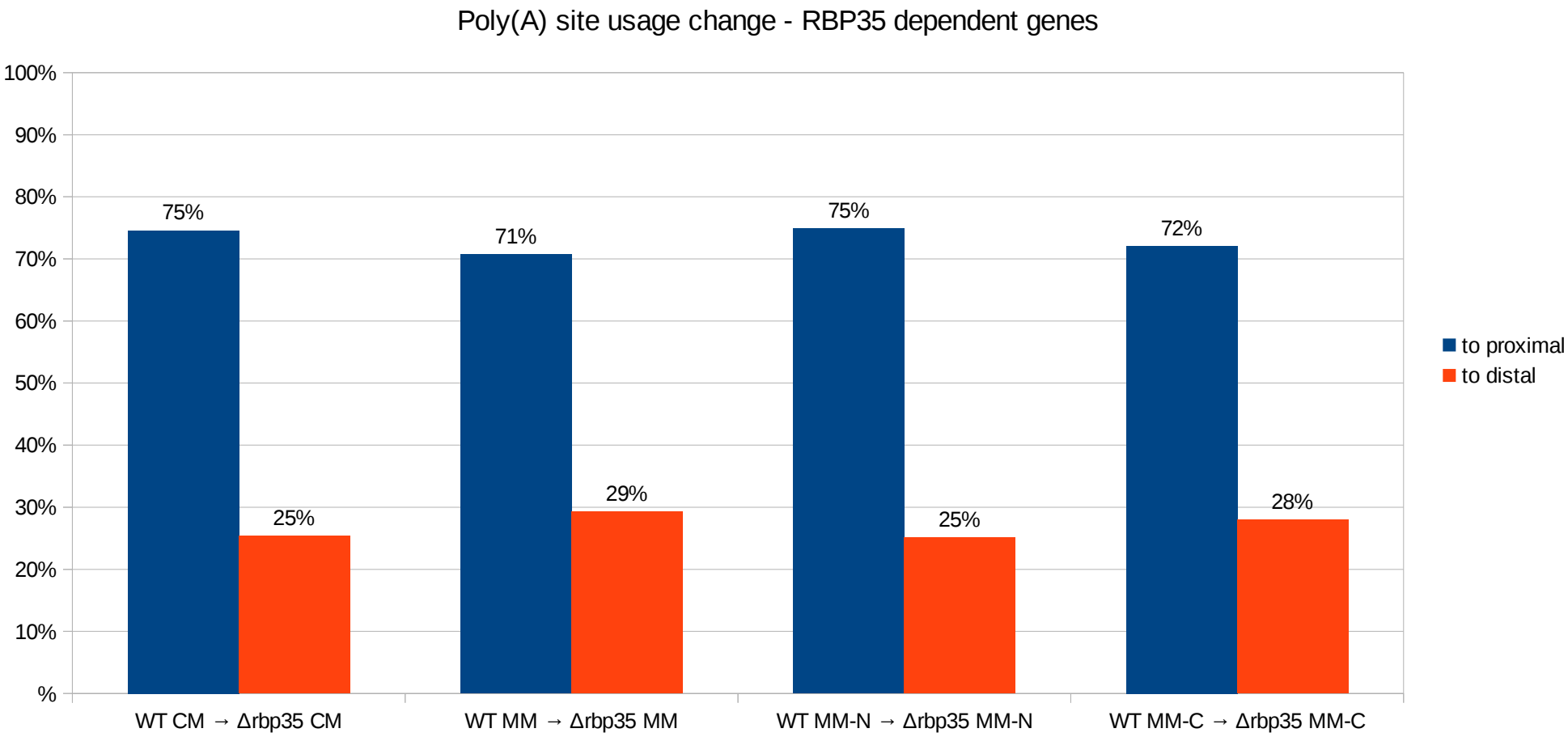
# RBP35 dependant genes



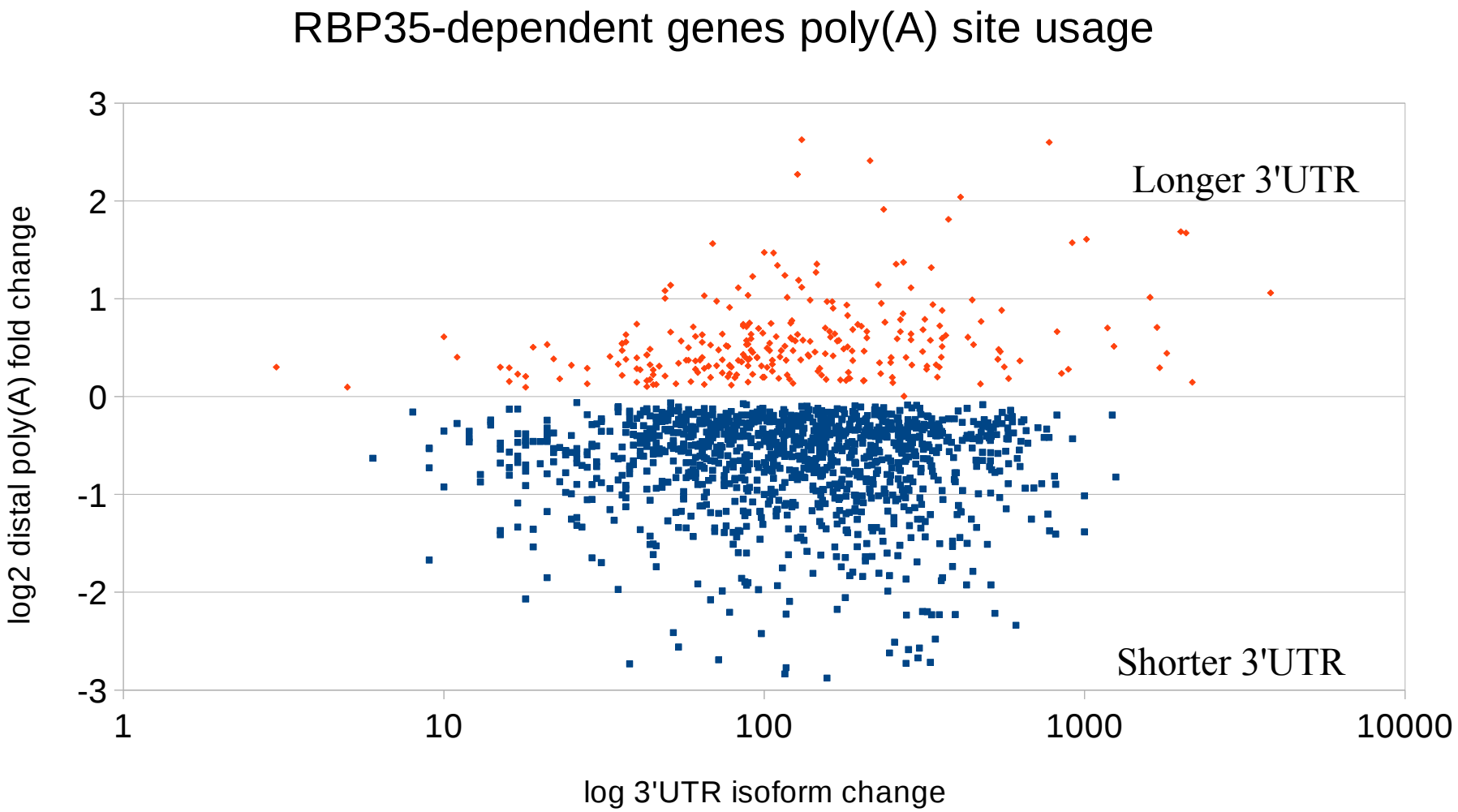
# There is no correlation between RBP35 dependance and differential expression



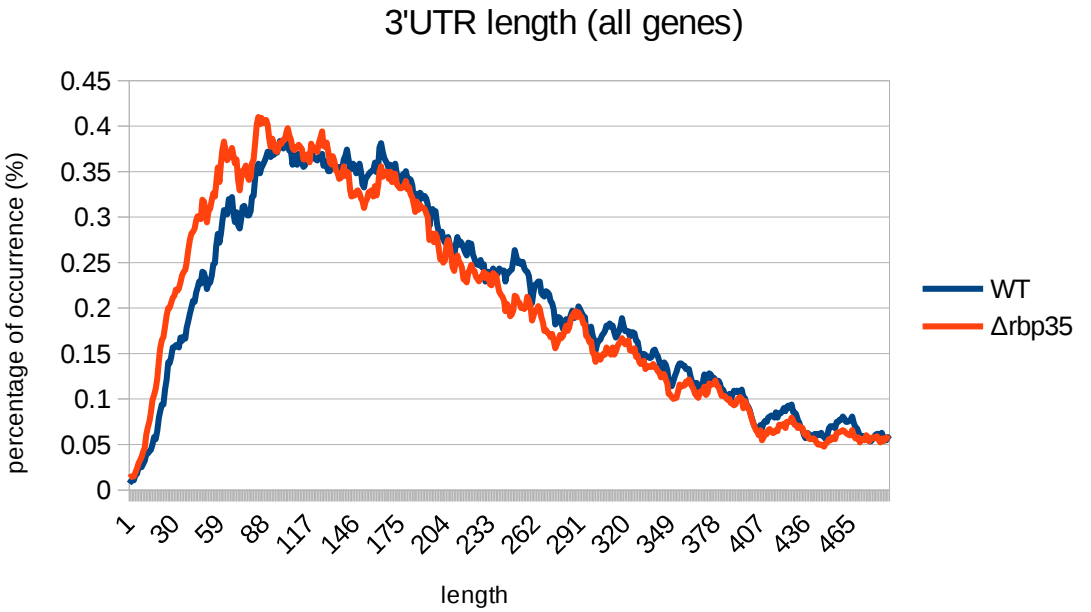
# *Δrbp35* affects poly(A) sites usage, preferring proximal cuts

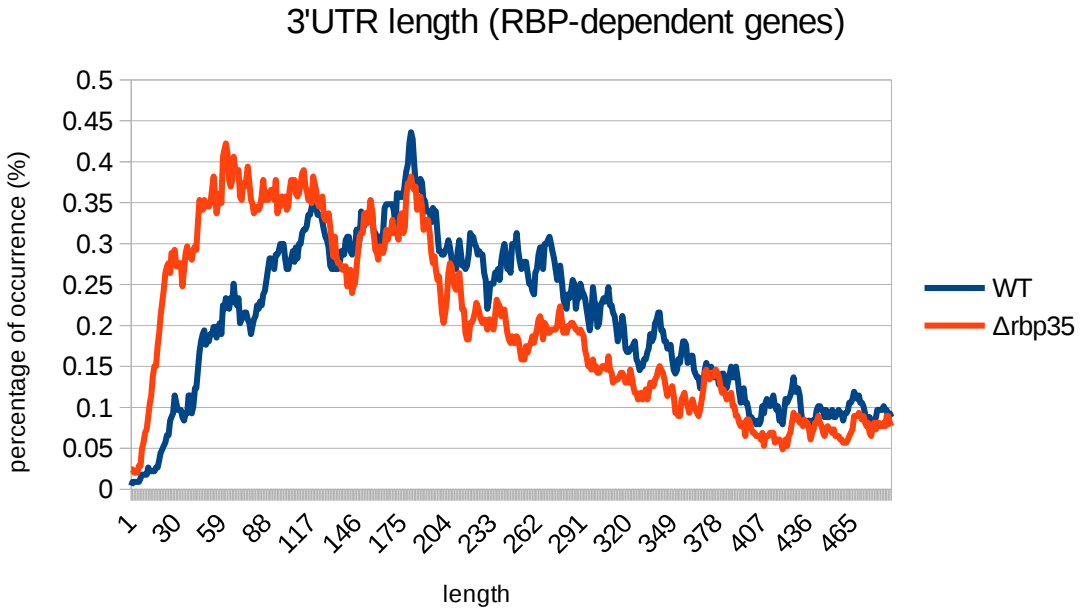


# *Δrbp35* affects poly(A) sites usage, preferring proximal cuts

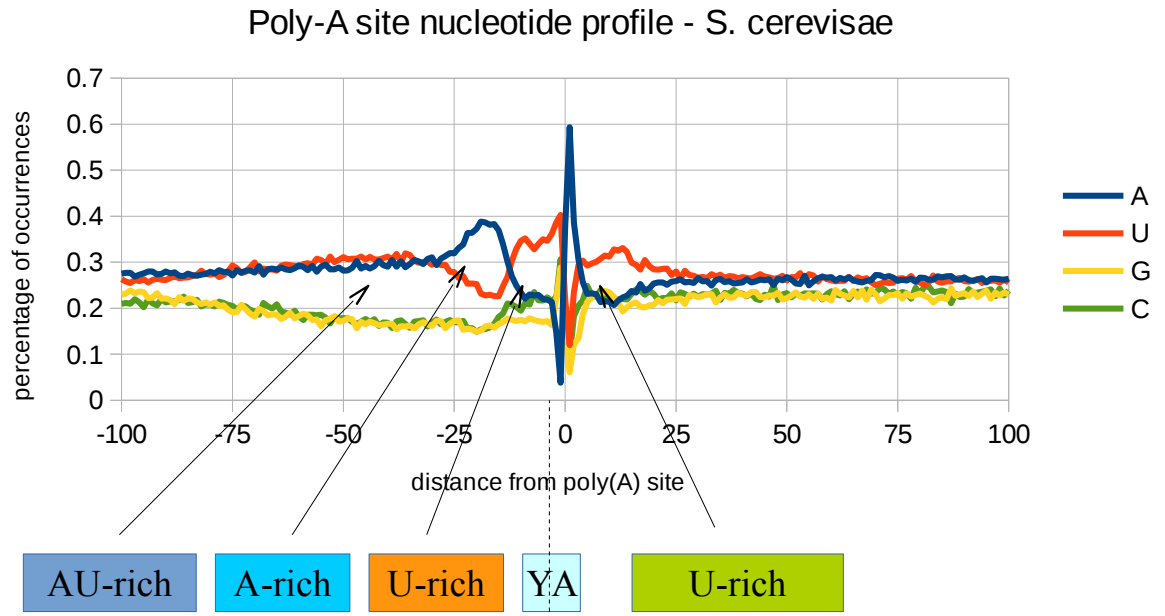


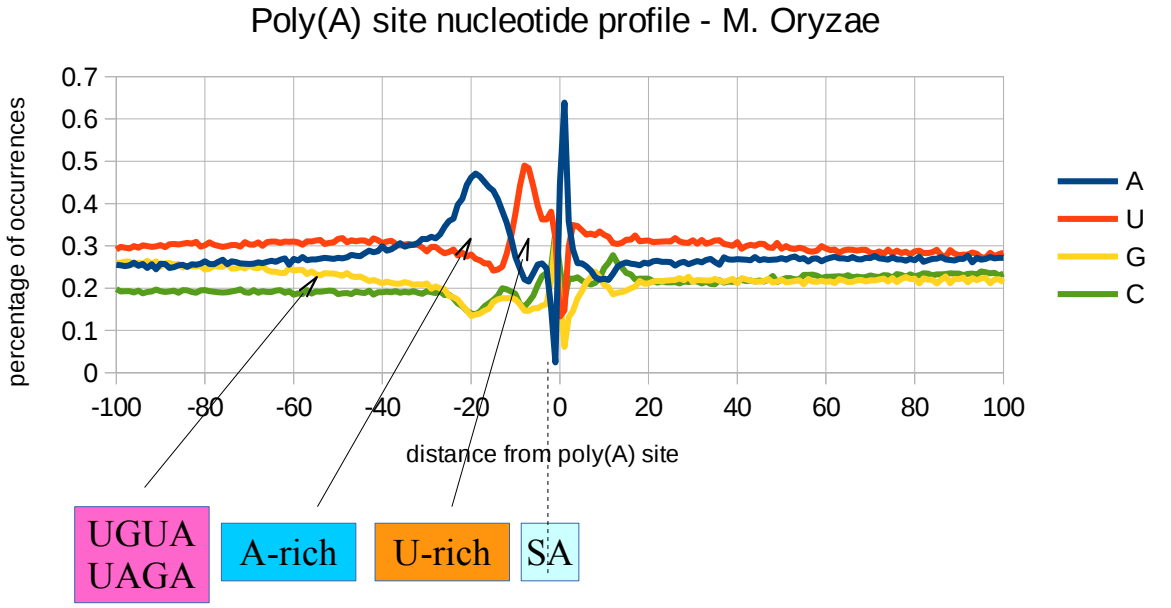
# *Δrbp35* affects 3'UTR length



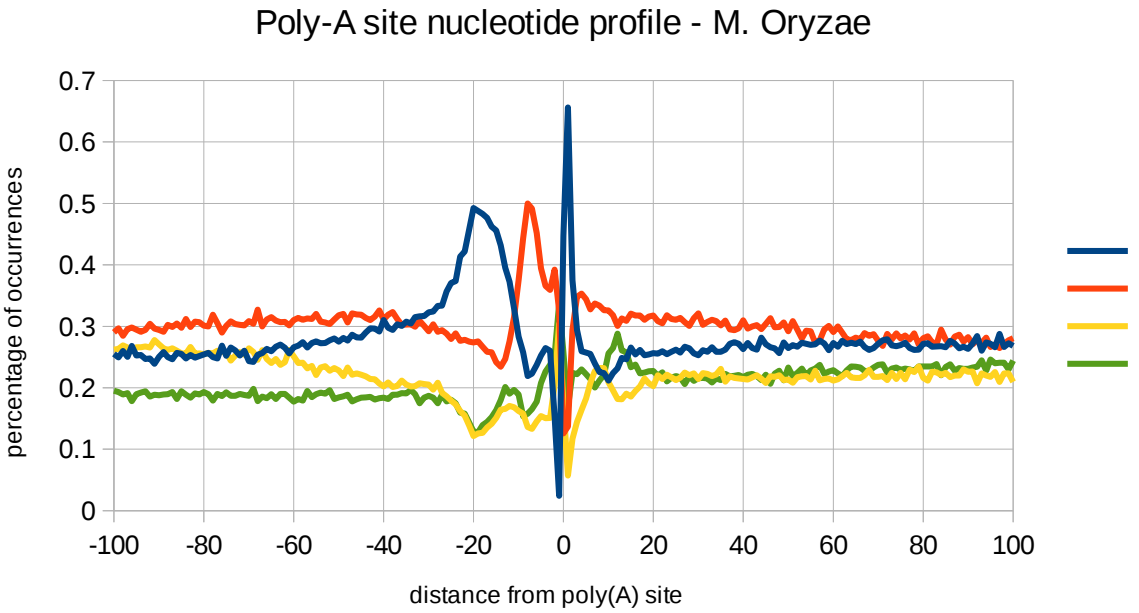



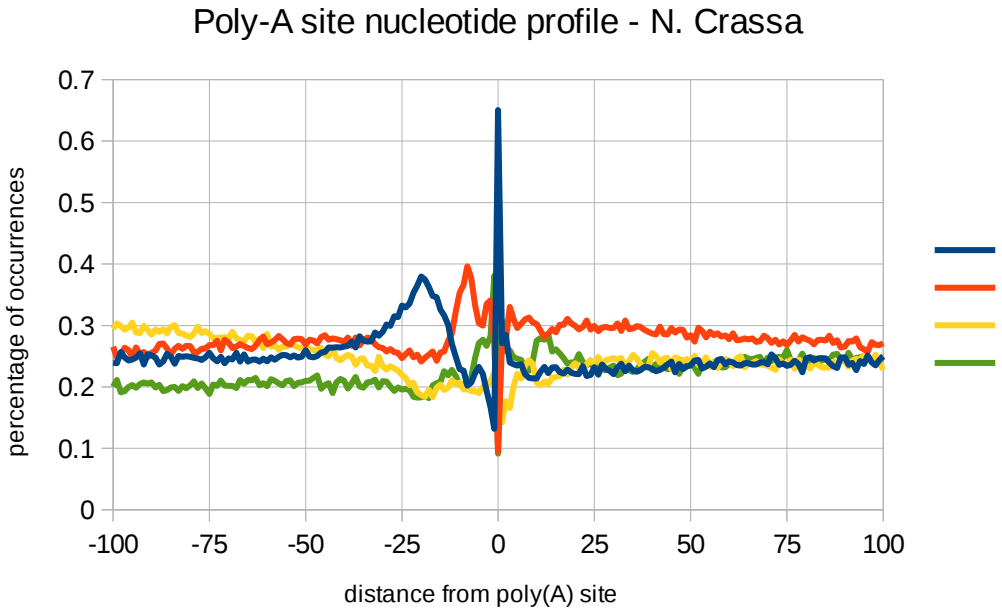
# Nucleotides profile of poly(A) sites slighty differs from *S.cerevisae*

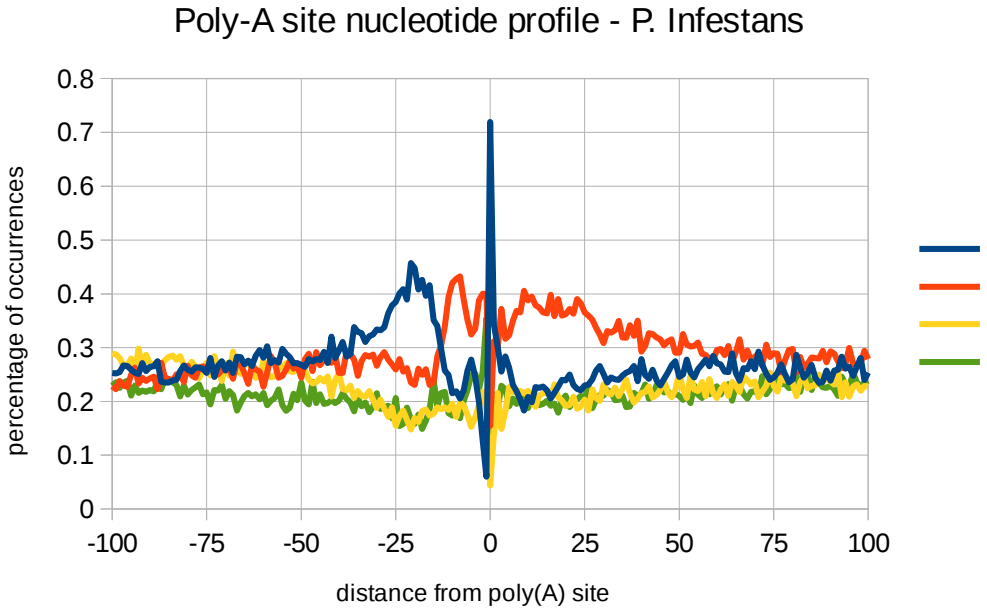


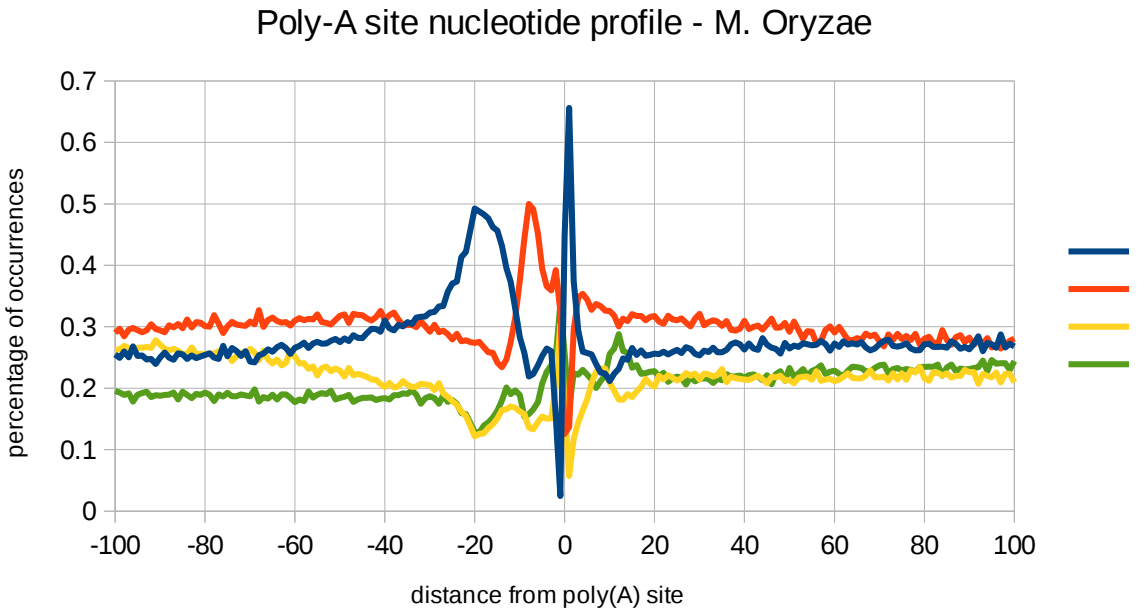
# Nucleotides profile of poly(A) sites resembles *N. crassa*

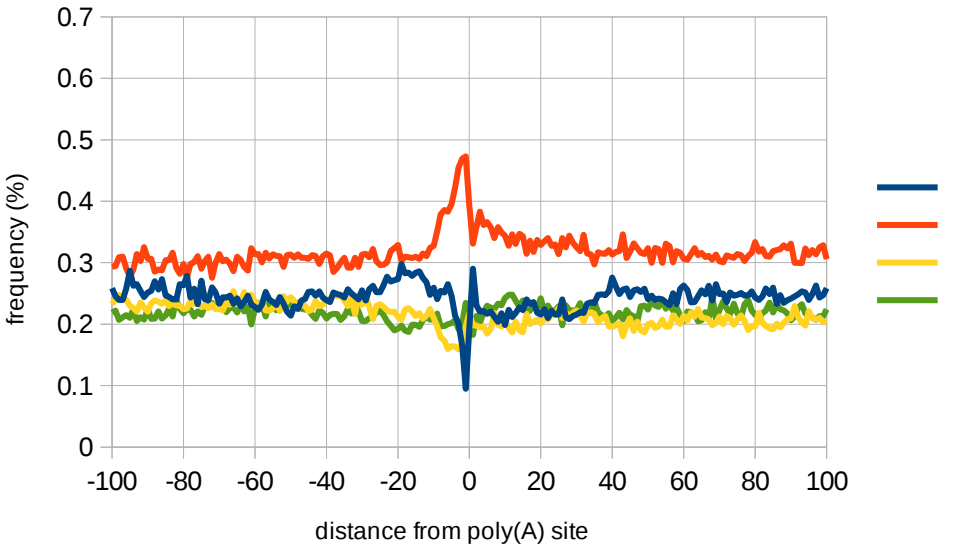


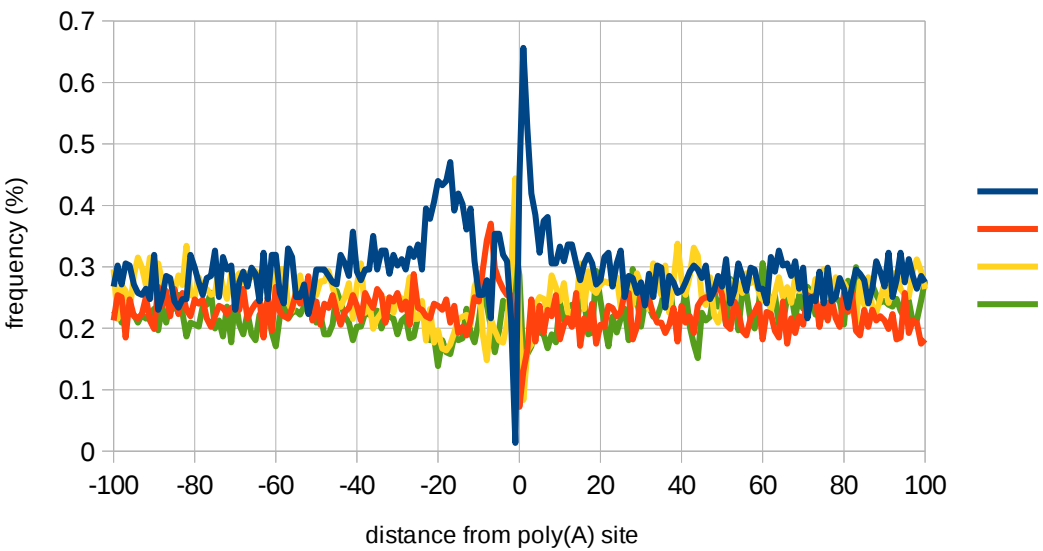



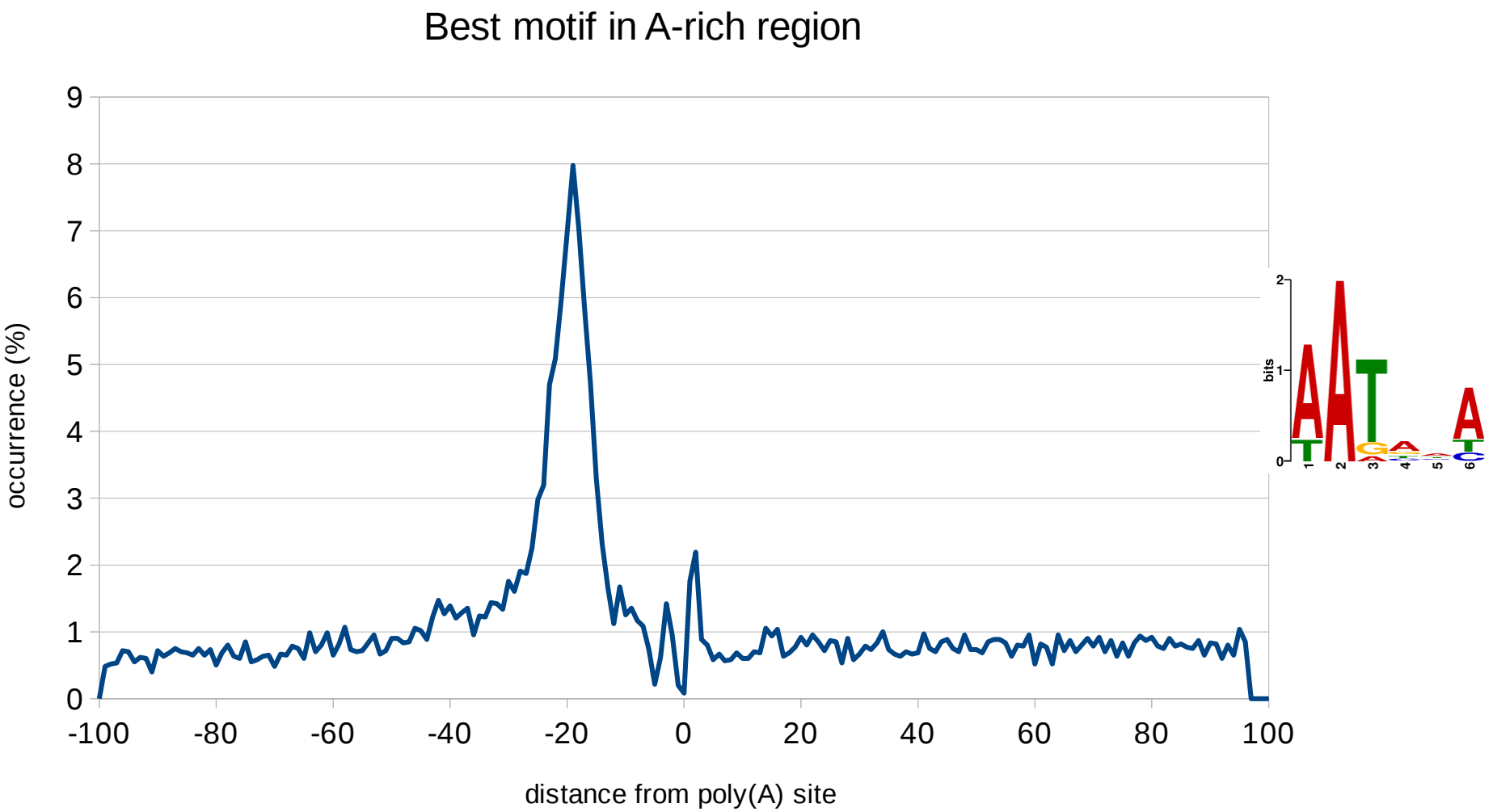
# Nucleotides profile of poly(A) of ncRNA and CDS poly(A) is different



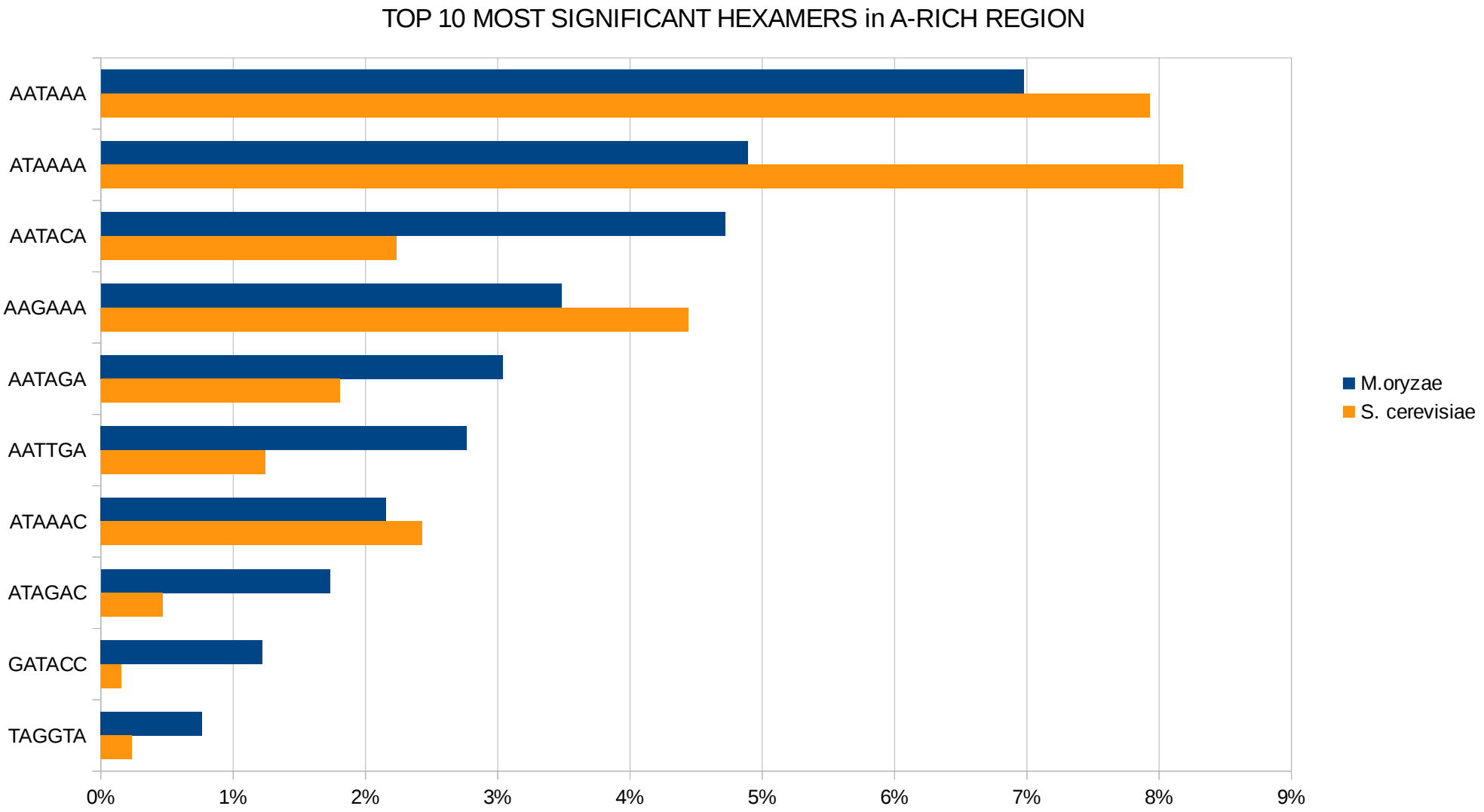



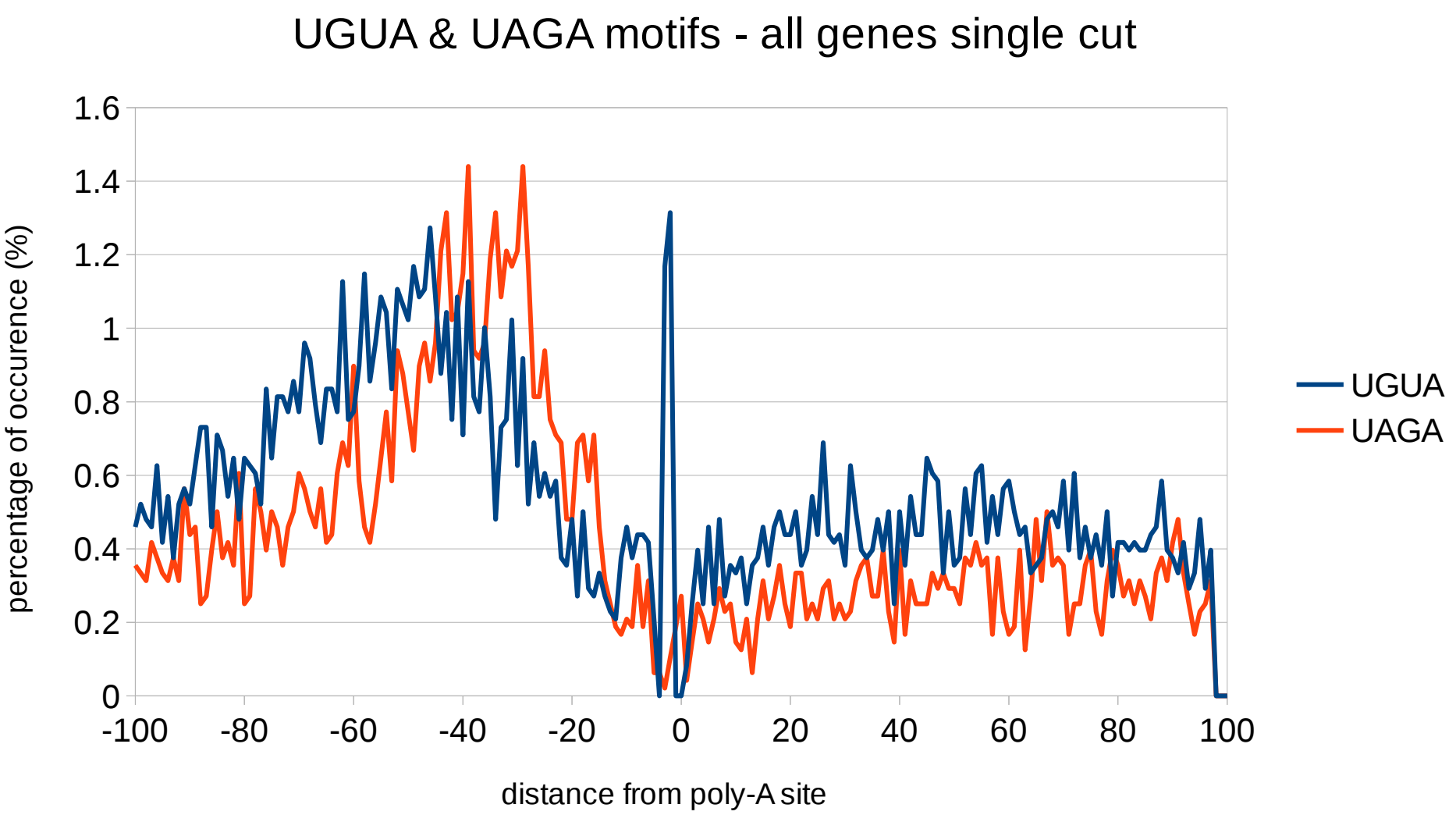
# The A-RICH region is located -30 -10 bp upstream



# The canonical metazoan AAUAAA motif has a frequency of only 7% in *M. oryzae*



# UAGA & UGUA motifs



# Polyadenilation signals in common genes

***MPG1***

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

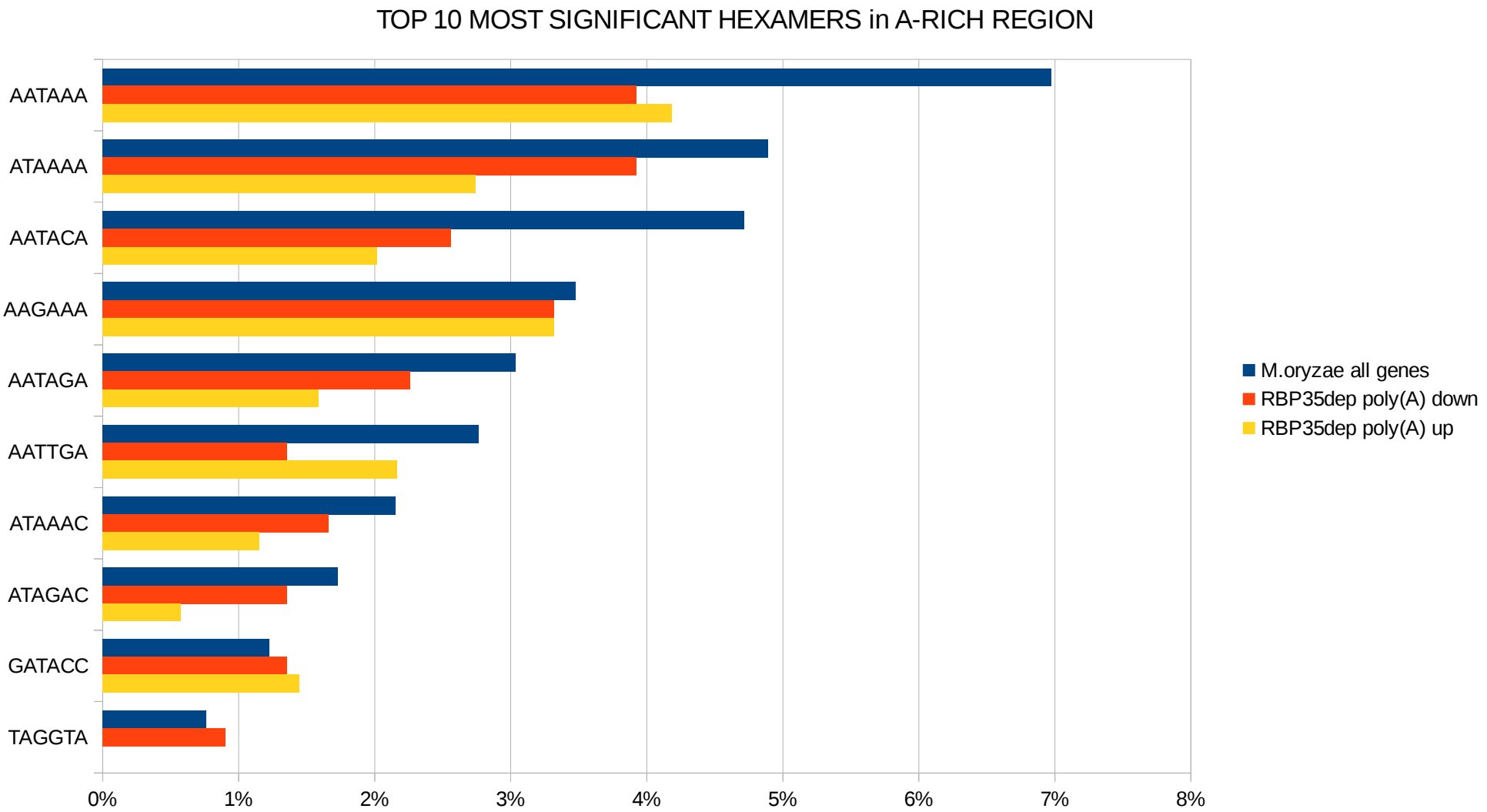
***PMK1***

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

***MST12***

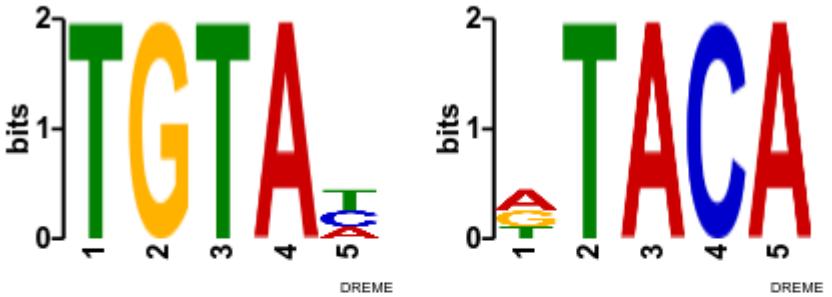
...CAGUGGCAUAAAAUCACAAAAUCUU**UAGA**AAGAUCAC**AGAAAA**CCUUUUGUC**CA**...

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



# 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 |
|---|---|---|---|---|
| 1836/3667 | 927/3667 | 6.6e-108 | 5.3e-103 | 5.3e-103 |

**Enriched Matching Words**

| Word | Positives | Negatives | P-value | ▼E-value |
|---|---|---|---|---|
| TGTAT | 922/3667 | 426/3667 | 1.3e-51 | 1.1e-46 |
| TGTAC | 706/3667 | 326/3667 | 4.0e-38 | 3.2e-33 |
| TGTAA | 535/3667 | 278/3667 | 4.3e-22 | 3.5e-17 |
| TGTAC | 465/3667 | 286/3667 | 2.9e-12 | 2.4e-7 |

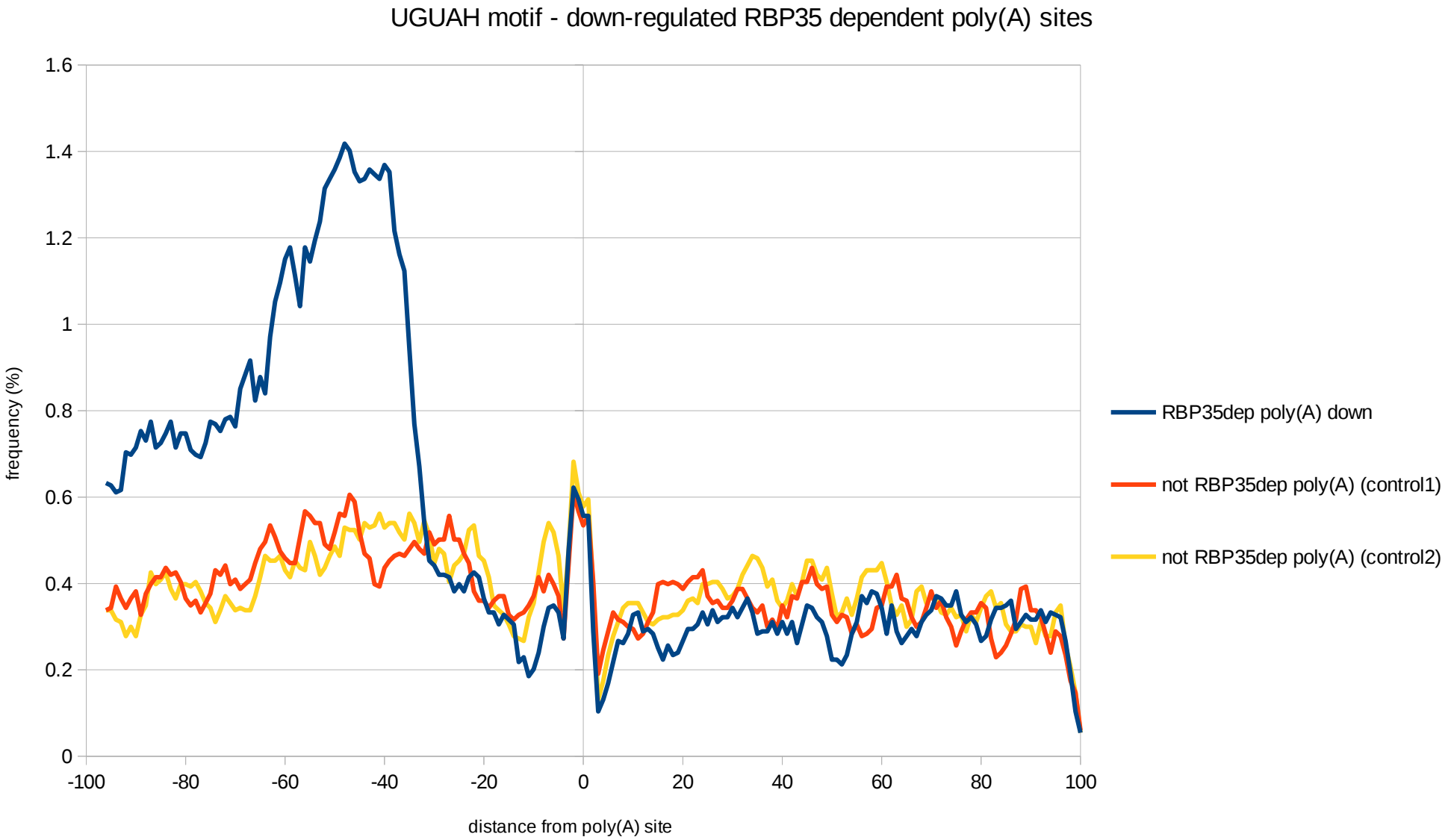
- Output of DREME, pRBP35dep as positive sequences list versus not-pRBP35dep negative list

# UGUAH motif – pRBP35dep vs not pRBP35dep

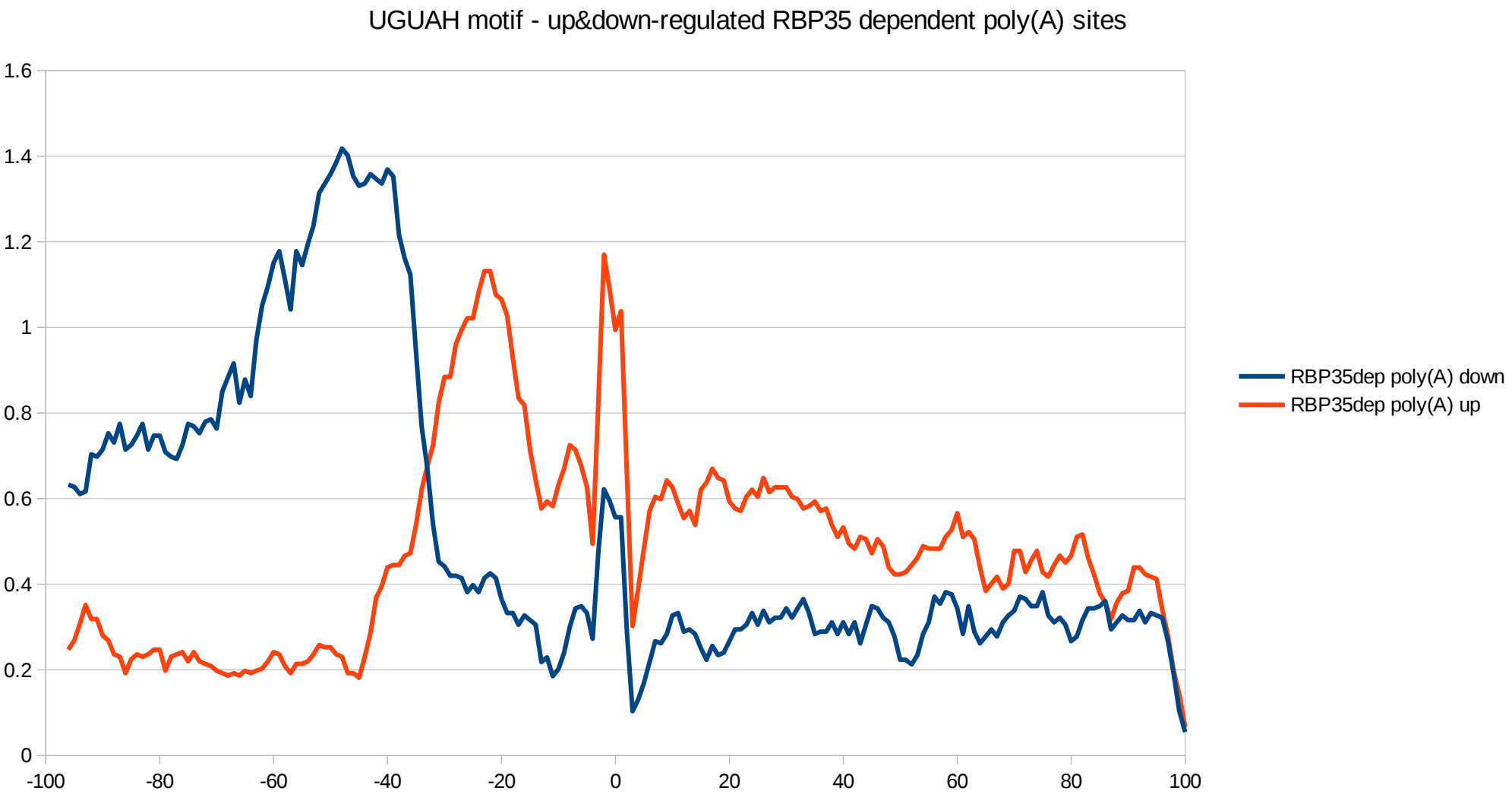
- In the first graph, we want to show how poly(A) sites dependent from *RBP35* display a different profile for the UGUAH motif in the respect to "regular" poly(A) sites

- 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

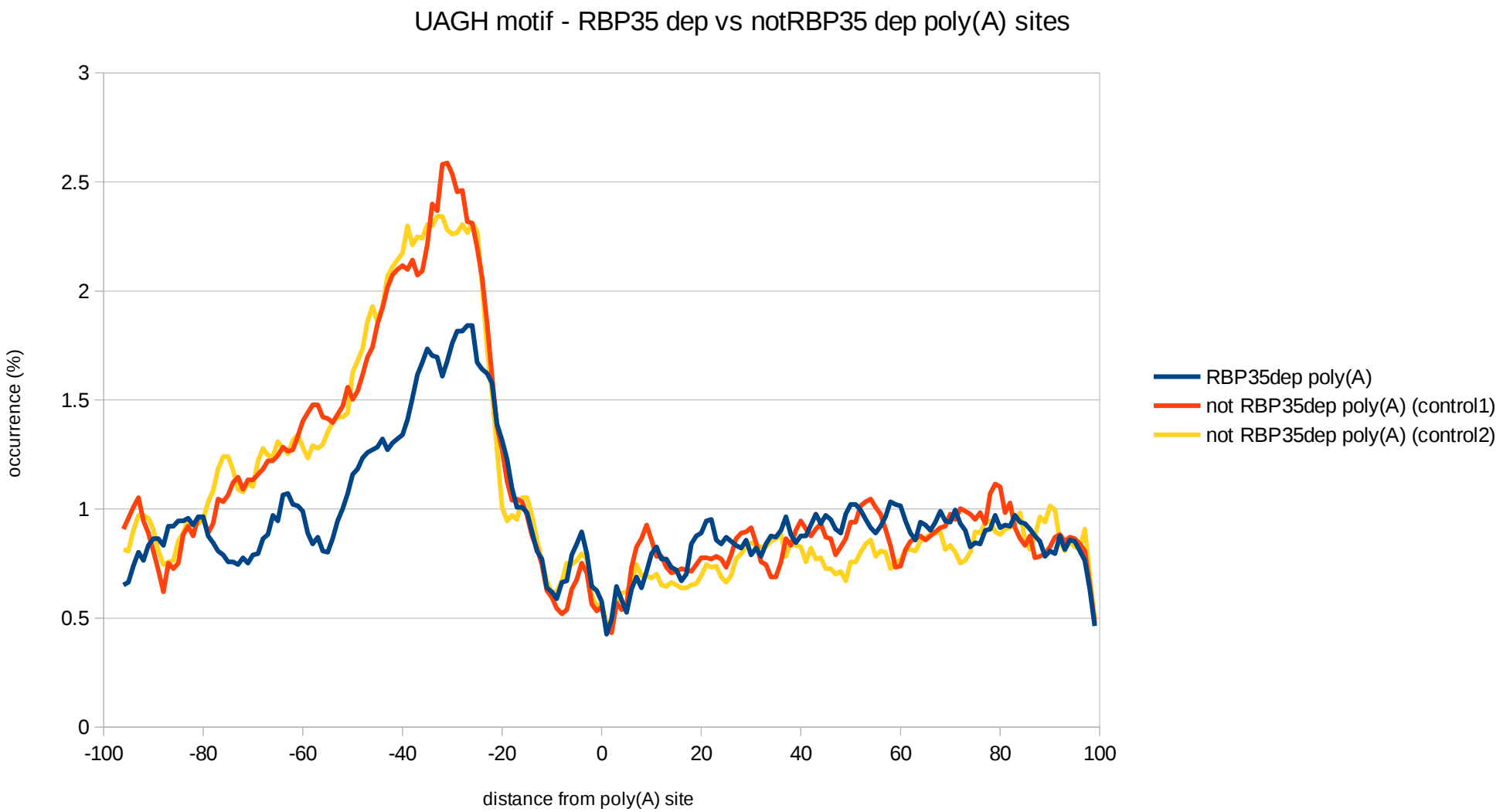
# UGUAH is enriched at -45 in poly(A) sites dependent from *RBP35*



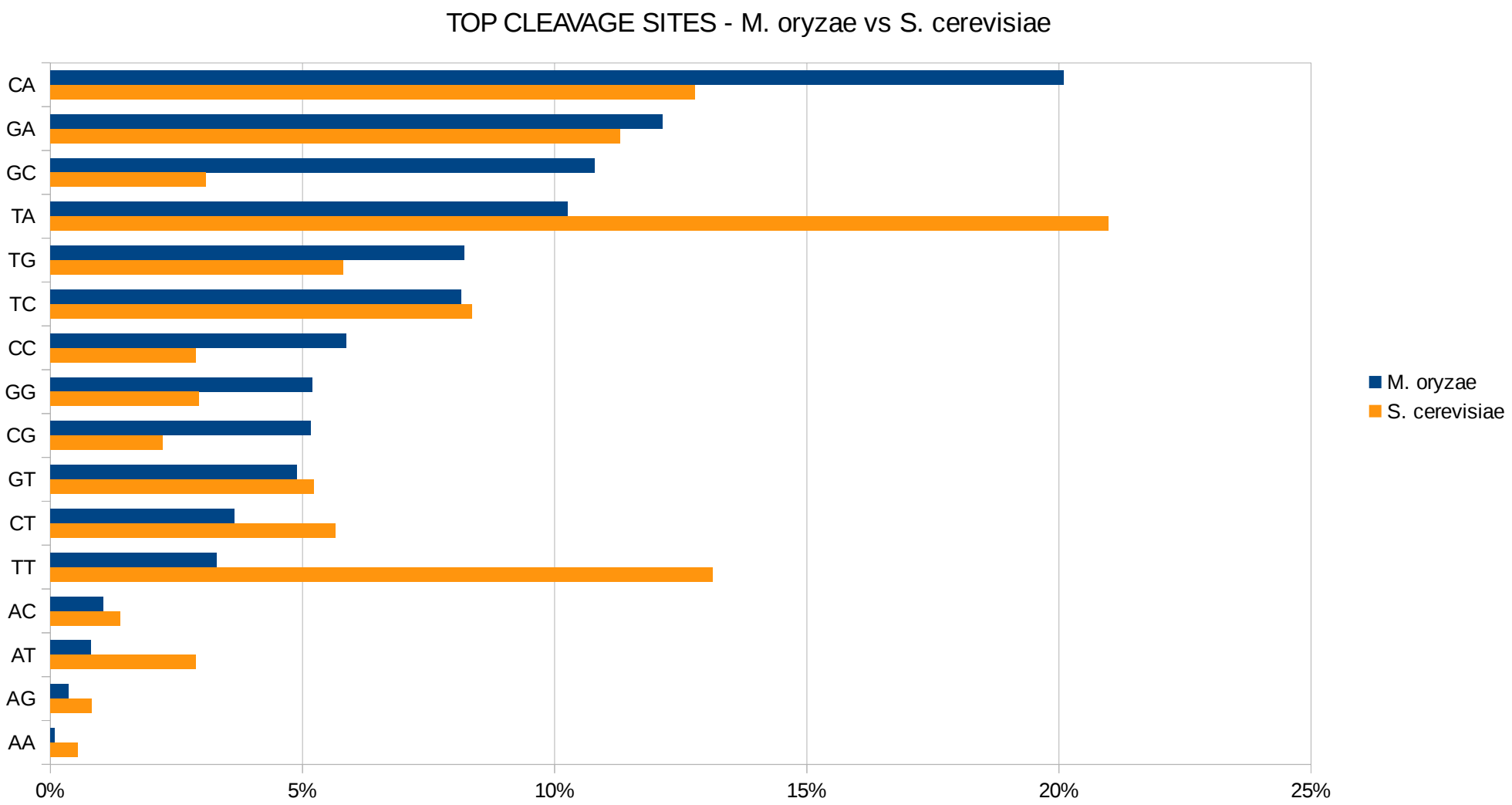
# UGUAH motif – RBP35 dependent poly(A) sites (up vs down regulated)



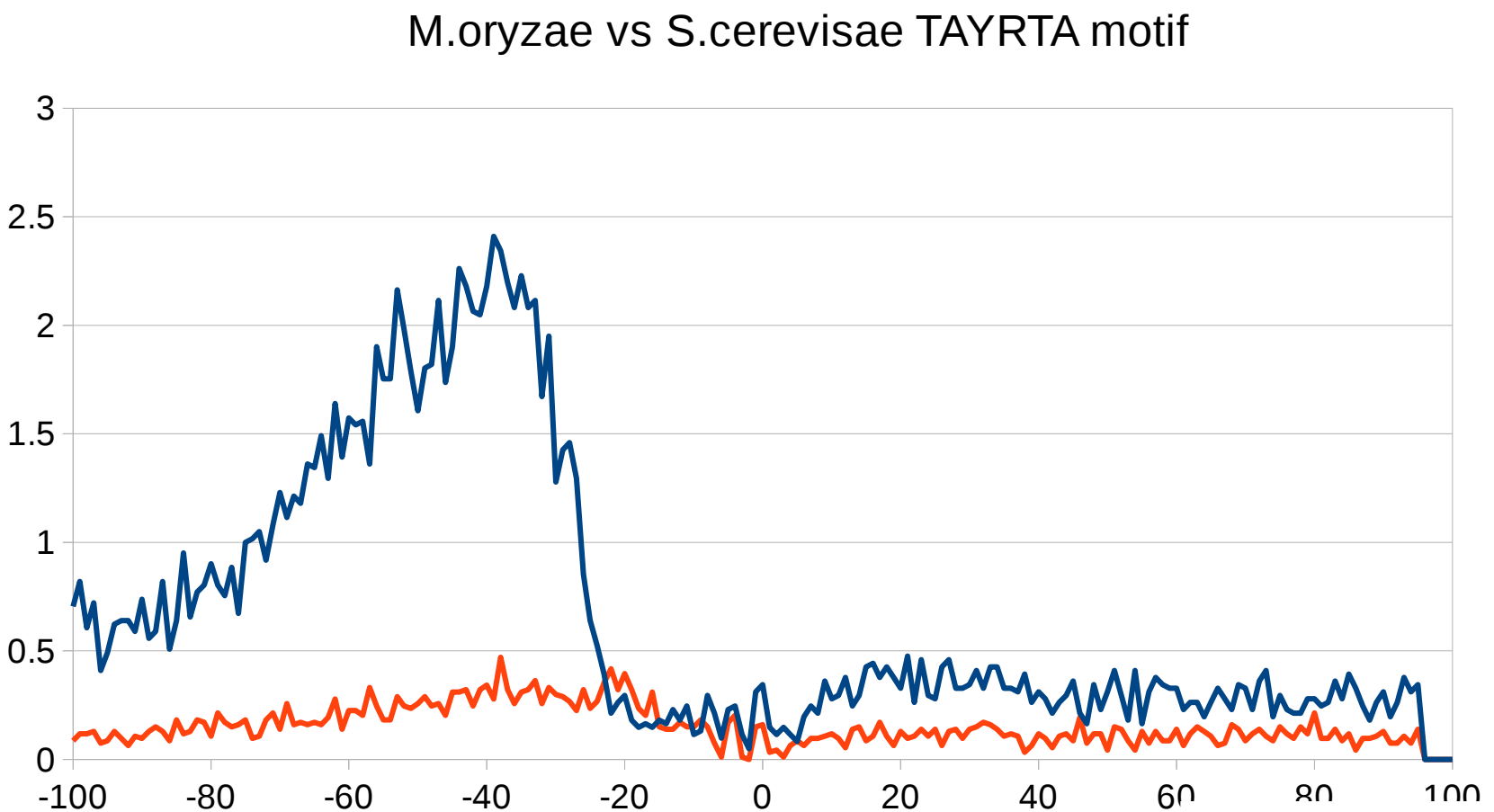
# UAGH is impoverished at -35 in poly(A) sites dependent from *RBP35*



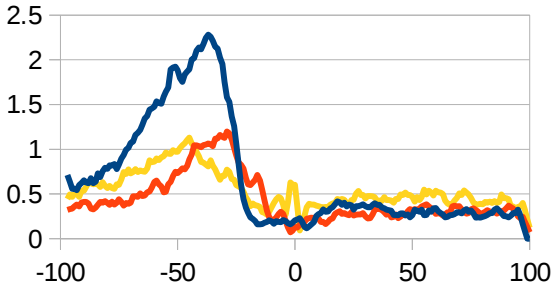
# *M. oryzae* prefers SA as cutting-site instead of YA



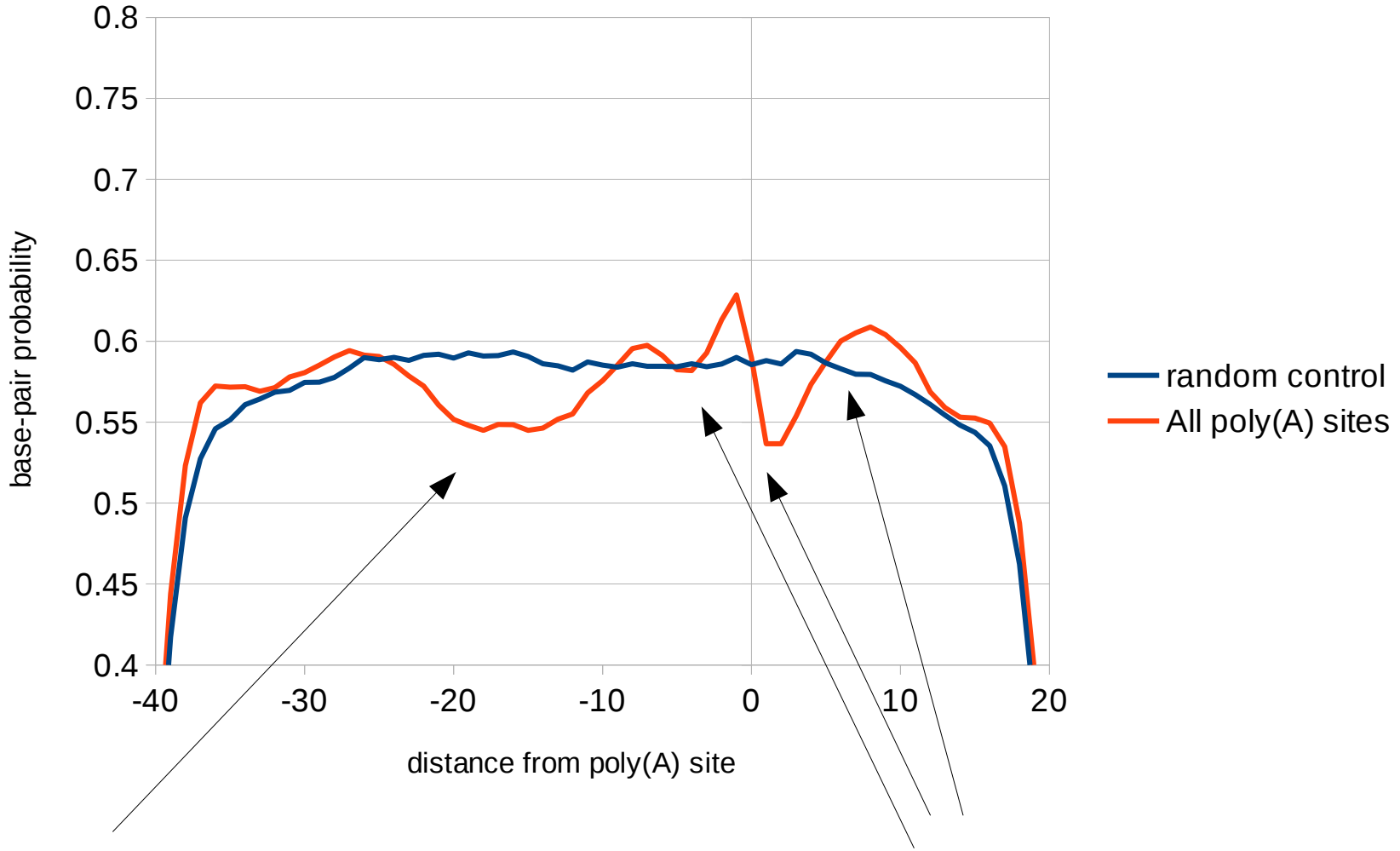
# The HRP1 binding motif TAYRTA from *S.cerevisae* in not found in *M.oryzae*




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



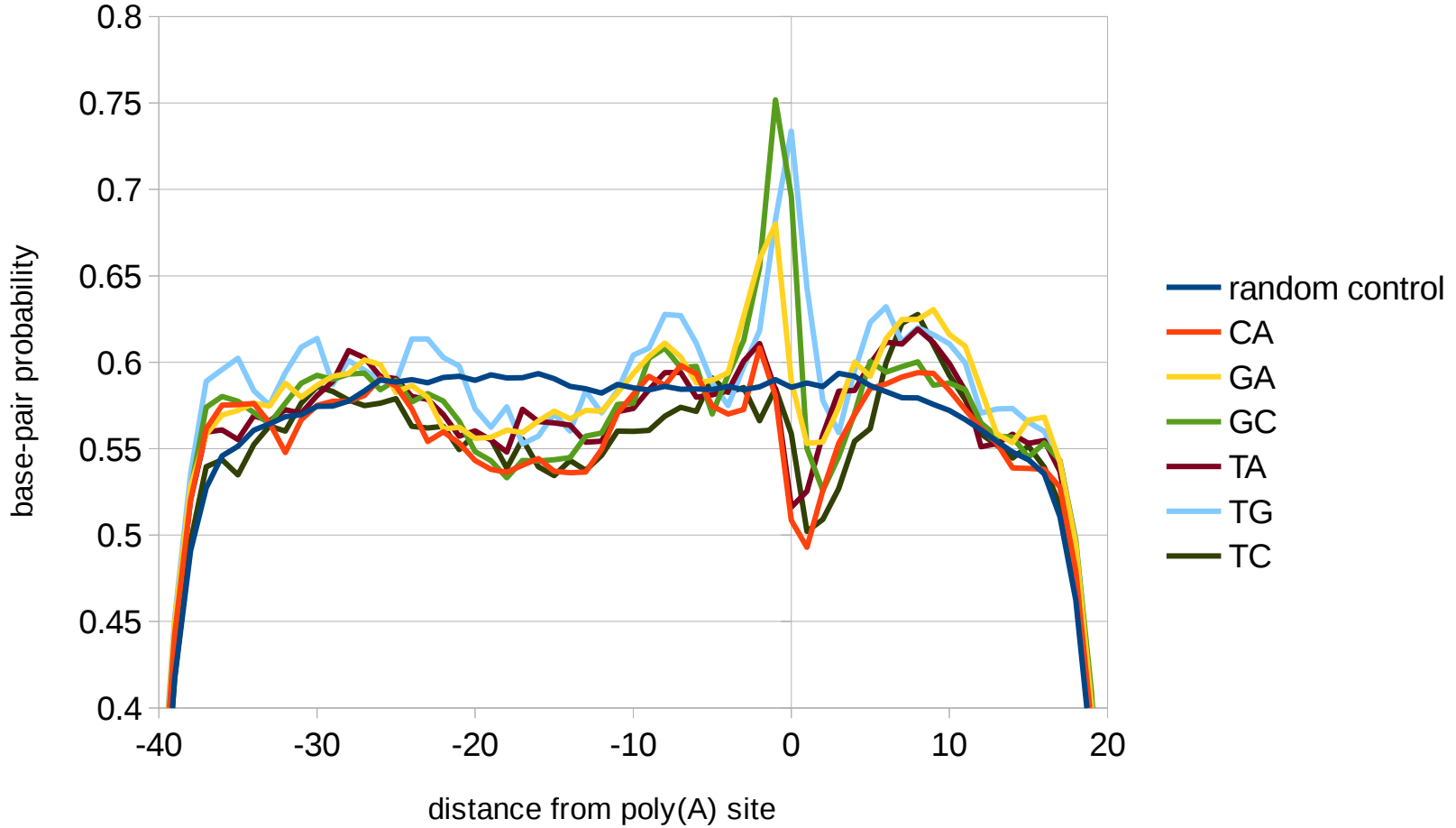
# The polyadenylation site region has a defined structure




*The A-rich region is usually not structured*

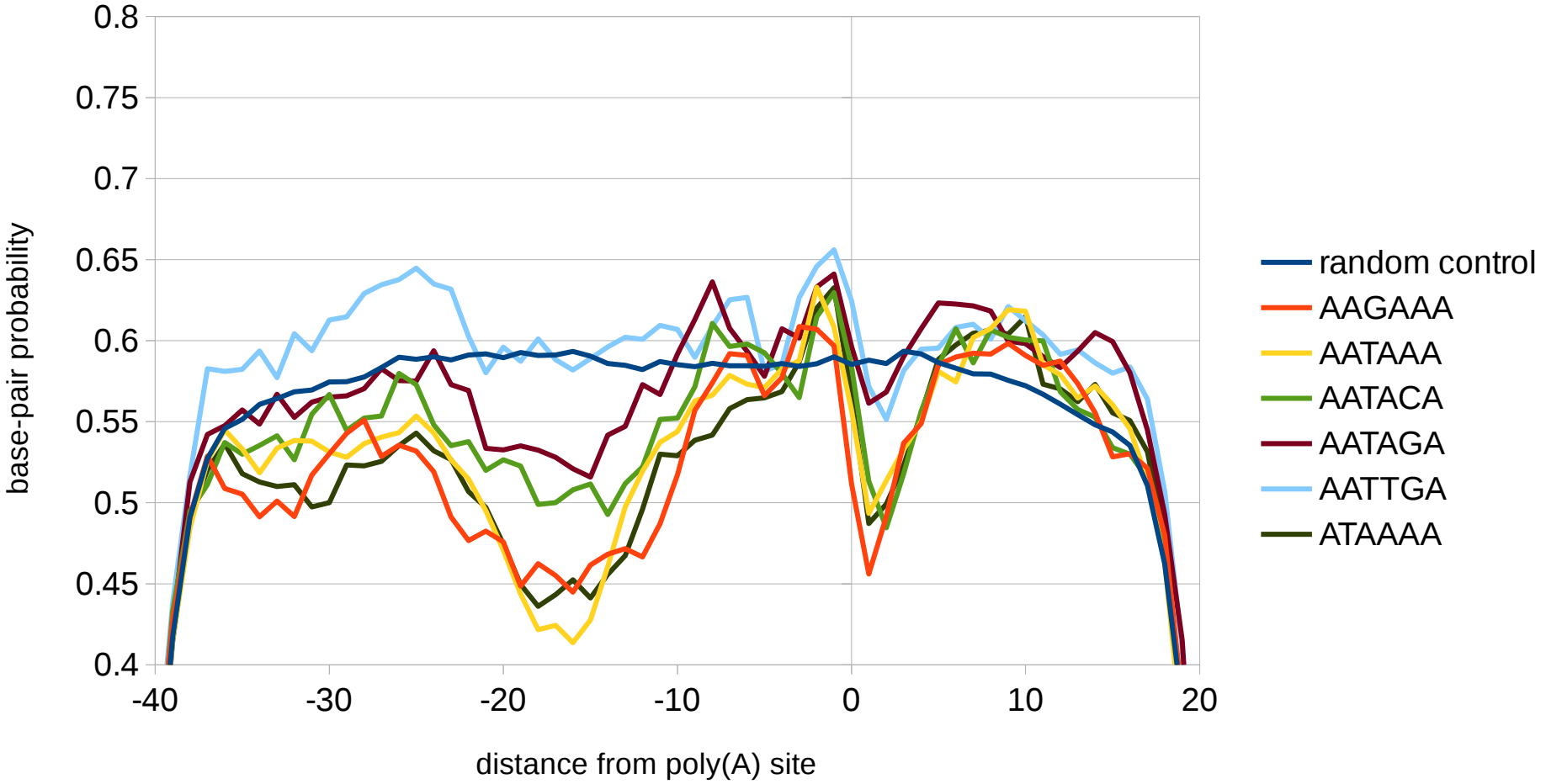
*The polyadenilation site is usually structured and located in a hairpin-loop*

# The polyadenylation site region has a defined structure



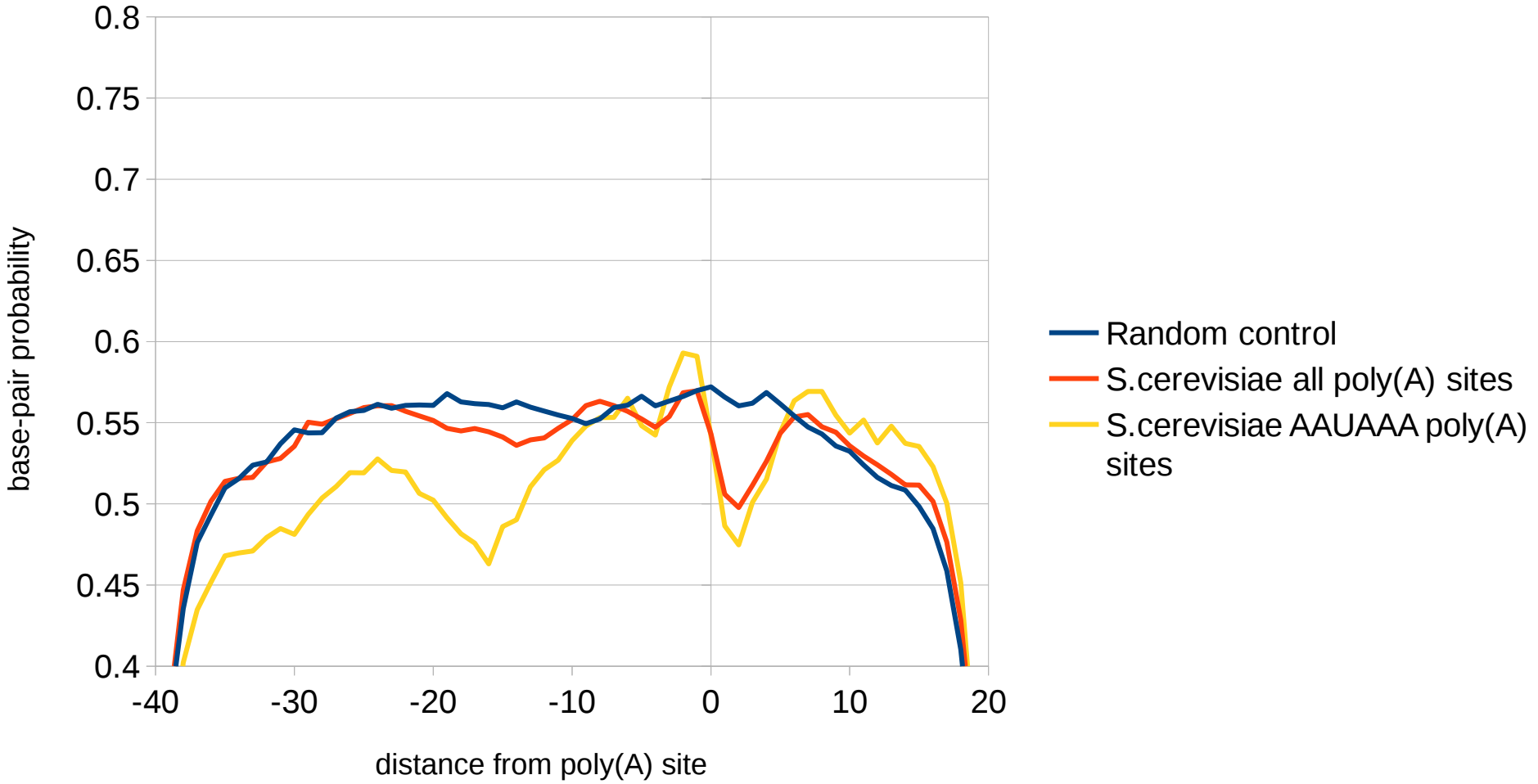

*Different cutsites have different base pais probabilites, with TG and CG the most structured. The most common poly(A) site CA has a average conformation*

# The polyadenylation site region has a defined structure




*Different A-rich motifs results in different degrees of conformation, with AAUAAA the most unstructured*

# The polyadenylation site region has a defined structure




*In S.cerevisiae, the poly(A) site is not clearly structured*