An Analysis of MDM4 Alternative Splicing and Effects Across Cancer Cell Lines

Kevin Hu
Mentor: Dr. Mahmoud Ghandi

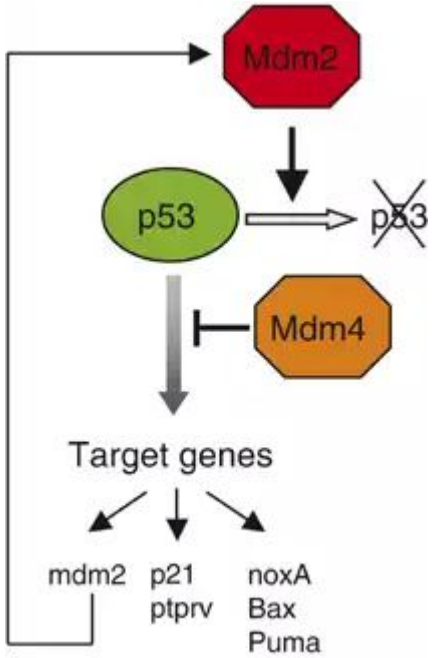
7th Annual MIT PRIMES Conference
May 20-21, 2017

Outline

- Introduction
  - MDM4
  - Isoforms
- Methodology
- Results
- Conclusion
- Extensions

MDM4

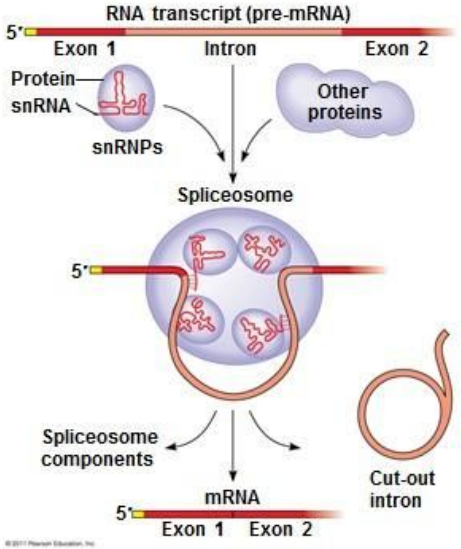
- Short for “Mouse double minute 4 homolog,” also called HDMX, MDMX, MRP1, MDM2-like p53-binding protein
- Suppresses the p53 tumor-suppressor.
- Binds to and activates transcriptional activity of E2F1
- In mice, knockdown of MDM4 results in embryonic death in 7.5-8.5 days post-conception due to p53 overactivity.
- MDM2 and MDM4 are structurally similar and are both inhibitors of p53; whether or not they work in concert or are nonoverlapping is not clear.



(Marine et al. 2006)

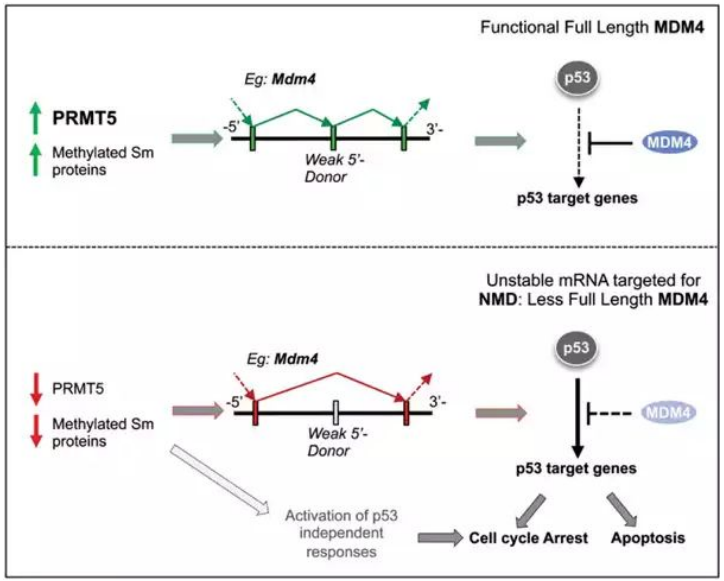
Alternative Splicing

- In eukaryotes, the initial mRNA transcript is first processed prior to translation into the final protein product.
- *Introns* are removed using a ribozyme called the *spliceosome*, and *exons* are joined together, to produce different *isoforms*.
- By splicing the RNA, cells are able to gain an additional level of control over gene expression (in addition to transcription factors, methylation, etc.)
- The locations between individual exons and introns are called splicing sites, or junctions.



Known Isoforms of MDM4

- MDM4-FL is the normal and most commonly found form in healthy cells.
- MDM4-S is a smaller variant generated by the skipping of the 6th exon.
- MDM4-S results in a protein with only the N-terminal of the p53-binding domain followed a frameshift caused by exon skipping.
- Higher expression of MDM4-S in mice is correlated with higher levels of p53, suggesting that MDM4-S acts indirectly to decrease the amount of MDM4-FL.



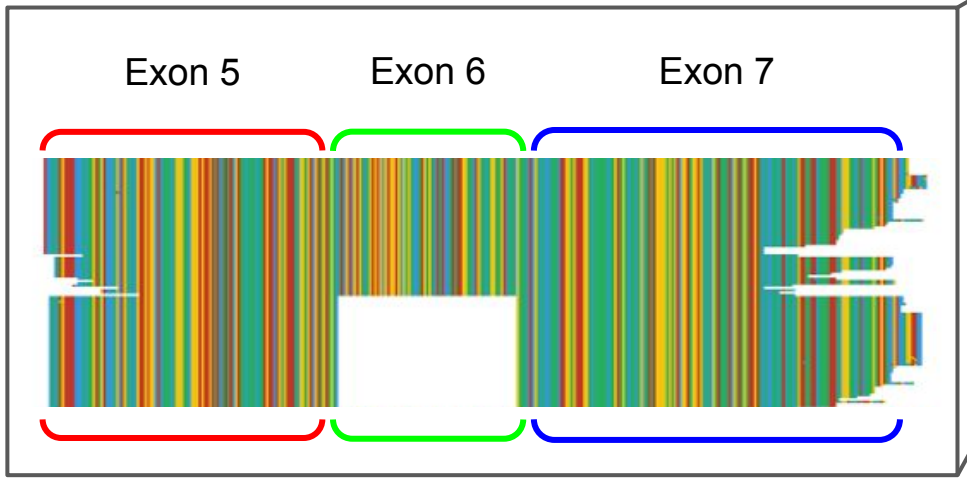
(Bezzi et al. 2013)

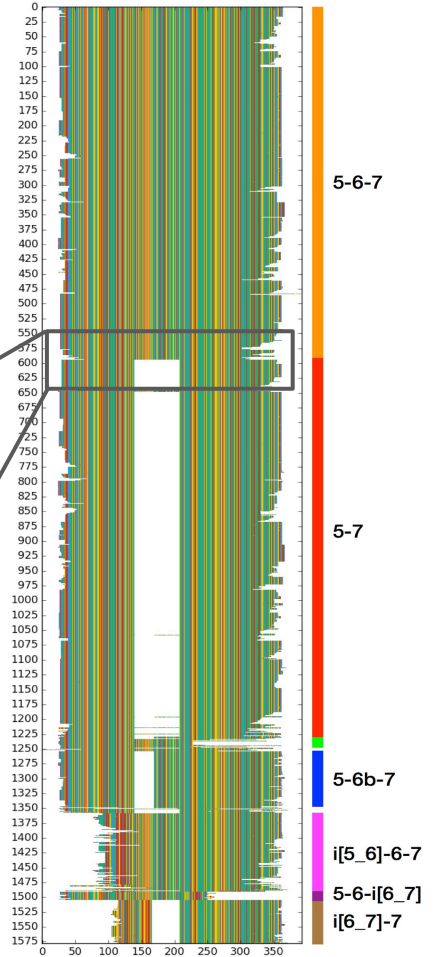
Raw Data

- Extracted MDM4-specific *k-mers*, or short nucleotide sequences, from raw RNA-sequencing data from the Cancer Cell Line Encyclopedia (CCLE) database.
- Filtered for sequences with a high match (>30 12-mer match) to MDM4 exons 5, 6, and 7.
- Summarized in FASTQ files containing individual RNA-seq reads from Each cancer cell line had two files of RNA-seq reads.
- Total of 1,019 human cancer cell lines.

De novo RNA-seq Assembly

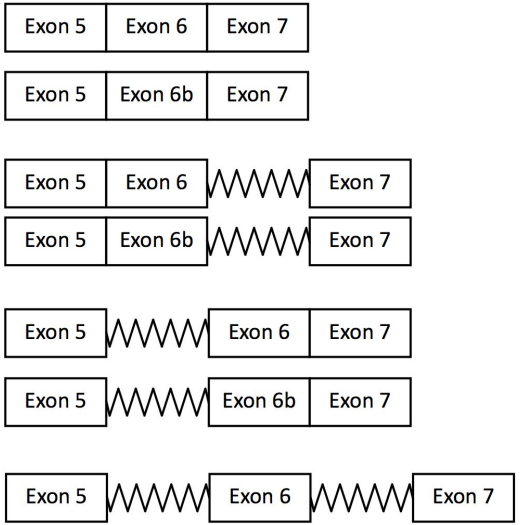
- Used *Trinity* to align and assemble raw RNA-seq reads per cell line into contiguous sequences.
- Used *MAFFT* to align Trinity sequences to view isoforms across all cell lines.
- Found evidence of an obscure isoform with “exon 6b”





Junction 16-mers

- After identifying isoforms in the Trinity/MAFFT alignments, created a list of 16-mers specific to splicing sites (junctions).
- Junctions were mapped to isoforms using a matrix model:
  - Isoform 5-6-7 contains junctions 5_6 and 6_7
  - Isoform 5-7 contains junction 5_7
  - Isoform 5-6b-7 contains junctions 5_6b and 6b_7
  - Etc.



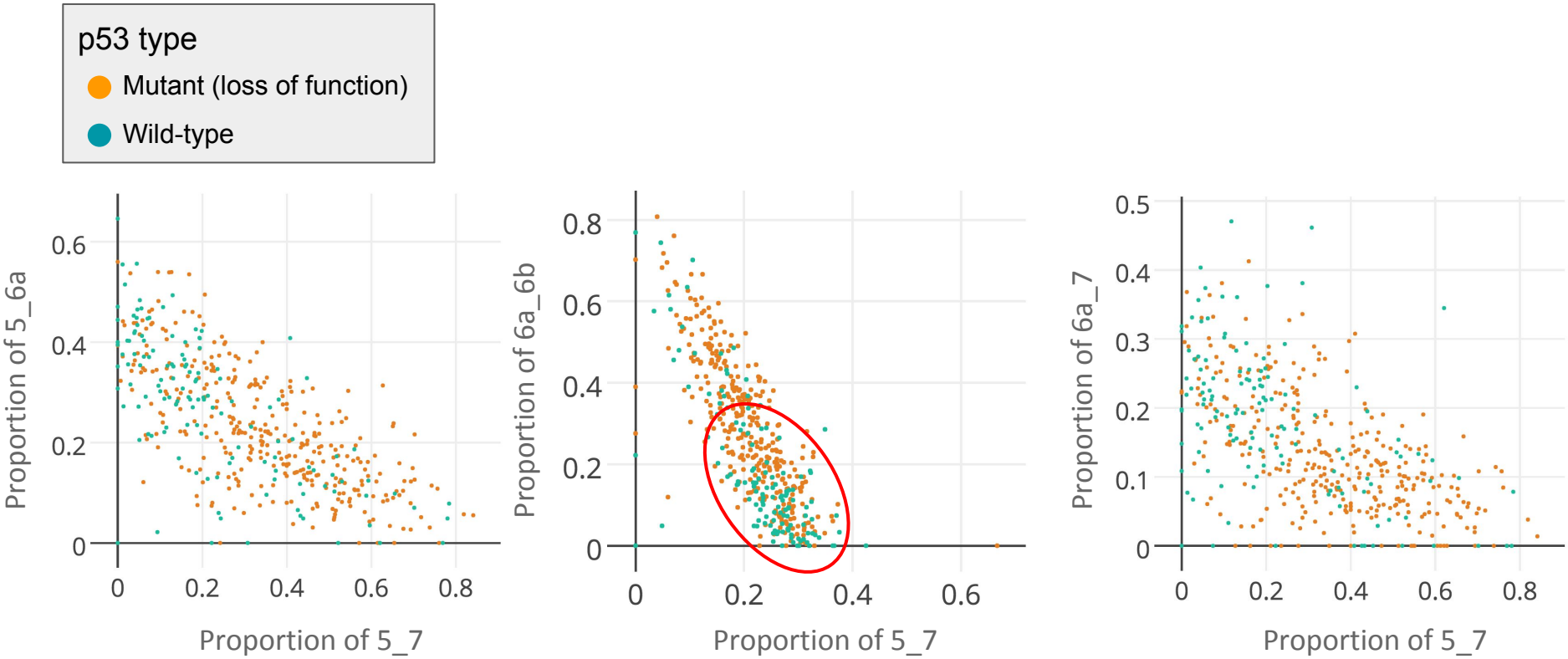
Junction 16-mers

Junction 16-mer counts

Cell lines

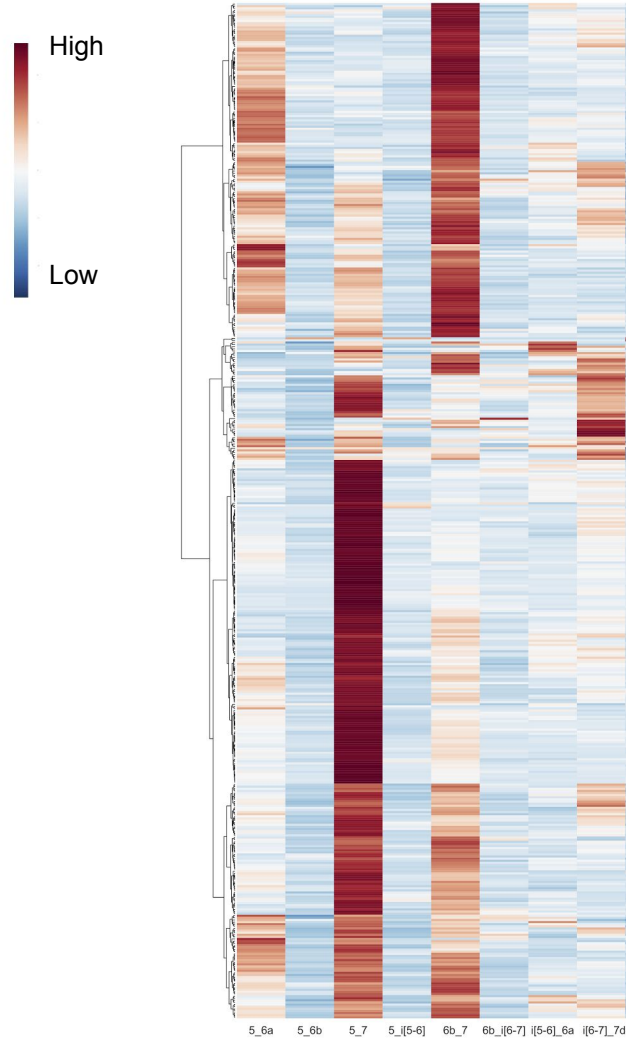
cell_line	5_6a	5_6b	5_7	5_i[5-6]	6a_6b	6b_7	6b_i[6-7]	i[5-6]_6a	i[6-7]_7
22RV1_PROSTATE	66	6	30	18	170	124	38	34	46
2313287_STOMACH	60	2	2	4	74	72	4	4	14
253J_URINARY_TRACT	12	2	8	0	10	16	0	2	2
253JBV_URINARY_TRACT	20	4	34	2	44	36	0	4	8
42MGBA_CENTRAL_NERVOUS_SYSTEM	4	0	18	0	2	4	0	0	6
5637_URINARY_TRACT	14	0	22	0	14	12	0	0	0
59M_OVARY	10	0	22	2	10	16	0	0	8
639V_URINARY_TRACT	12	4	28	2	14	22	2	8	6
647V_URINARY_TRACT	10	0	14	0	8	10	2	0	2
697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	92	10	42	8	116	134	4	18	16
769P_KIDNEY	12	6	10	0	22	38	2	2	6
786O_KIDNEY	0	0	10	4	4	2	4	0	4
8305C_THYROID	4	0	10	2	10	16	0	0	0
8505C_THYROID	4	0	18	0	6	8	0	0	0
8MGBA_CENTRAL_NERVOUS_SYSTEM	22	2	18	8	56	48	12	26	38
A101D_SKIN	8	2	30	2	22	10	4	4	8
A1207_CENTRAL_NERVOUS_SYSTEM	4	2	8	0	18	16	6	6	4
A172_CENTRAL_NERVOUS_SYSTEM	20	2	18	0	38	38	0	4	10
A204_SOFT_TISSUE	14	2	0	0	26	18	0	6	4
A2058_SKIN	26	6	18	4	58	46	4	18	32
A253_SALIVARY_GLAND	12	4	16	4	8	22	0	0	6
A2780_OVARY	48	10	24	10	96	92	12	20	12

Anticorrelation between 5_7 and MDM4-FL junctions



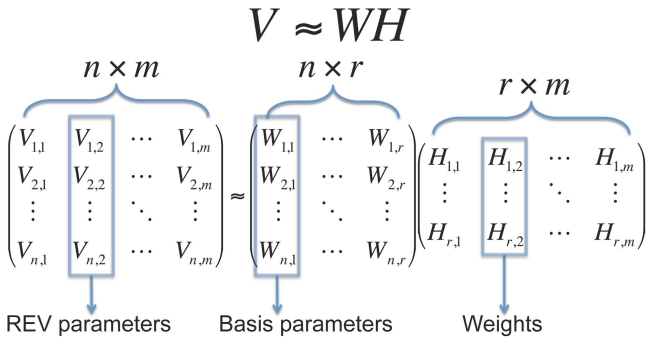
Junction Count Clustering

- Normalize to proportion of total junction counts in a specific cell line.
- Cluster using hierarchical method.
- 3 main junction profiles are evident:
  - 5_7 and 6b_7 (both isoforms expressed)
  - 5_7 (MDM4-S only)
  - 5_6a and 6b_7 (MDM4-FL only)

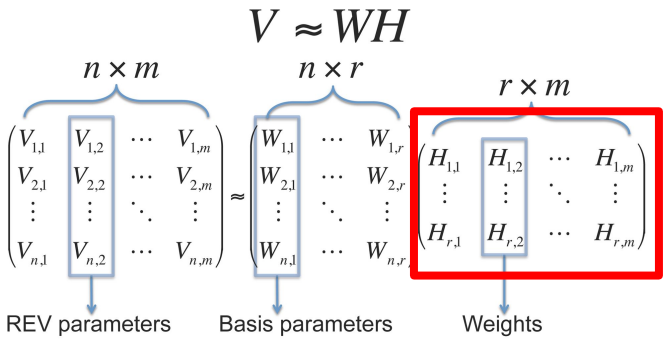


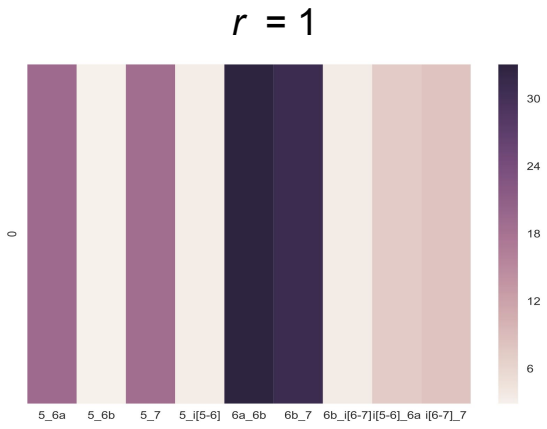
Nonnegative Matrix Factorization

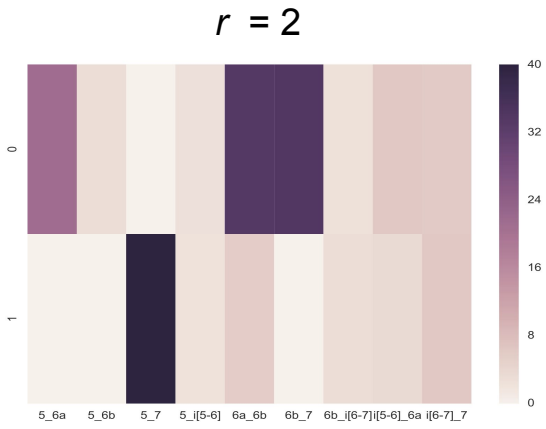
- Break matrix *V* of *n* samples by *m* features into components *W* and *H* with intermediate dimension *r*.
- In this case, separate 1,019 × 9 matrix (1,019 cell lines by 9 distinct junction counts) into 1,019 × *r* and *r* × 9 matrices (*r* isoforms).
- *H* is the matrix mapping the isoforms (rows) to the junctions (columns).
- *W* is the isoform estimate (columns) per each cell line (rows).
- Ultimately unsupervisedly estimates the frequencies of each isoform.

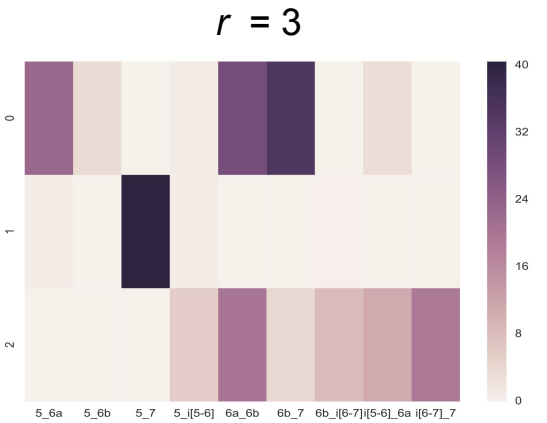


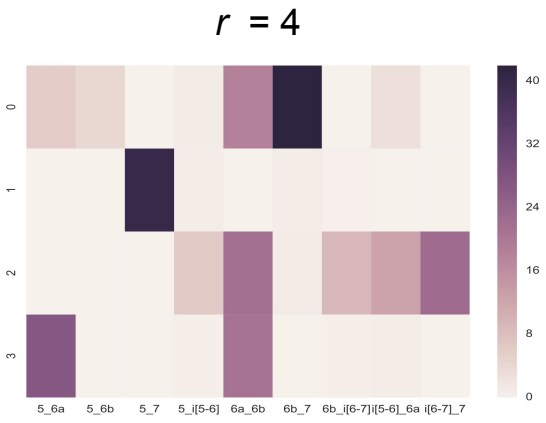
Weight matrices











Future Extensions

- Functional analysis of the 5-6b-7 and 5-7 isoforms.
- Find enrichment of isoforms in specific cancer types.
- Explore the correlation between MDM4 and MDM2 activity and isoforms.

Acknowledgements

Thank you:

- The MIT PRIMES program, for making this research possible
- Dr. Mahmoud Ghandi for his generous support and guidance
- My parents for their support