LineUp

Identifying Deleterious Mutations Using Protein Domain Alignment

Brady Bernard

Shmulevich Group

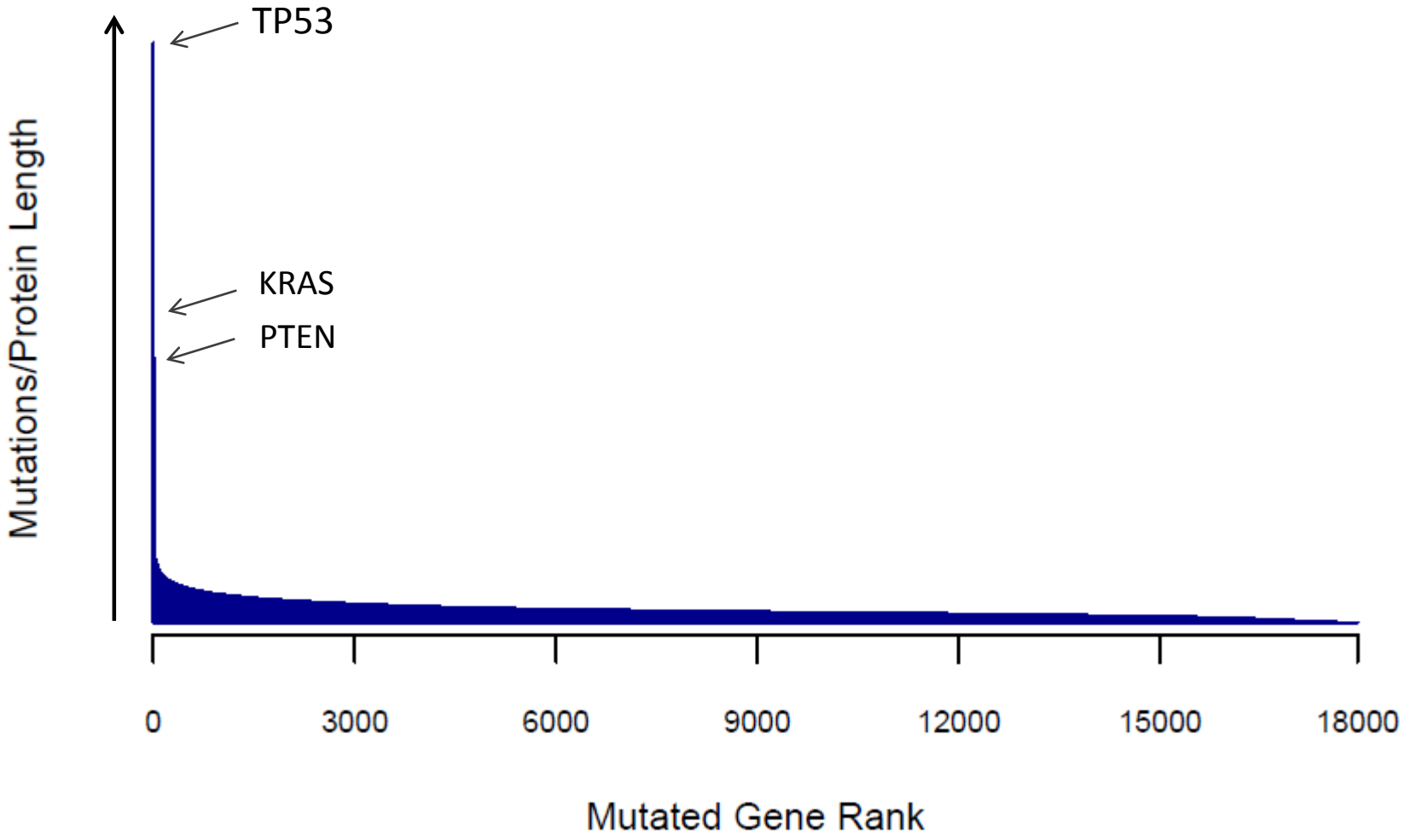
The Cancer Genome Atlas

Institute for Systems Biology

Revolutionizing Science. Enhancing Life.

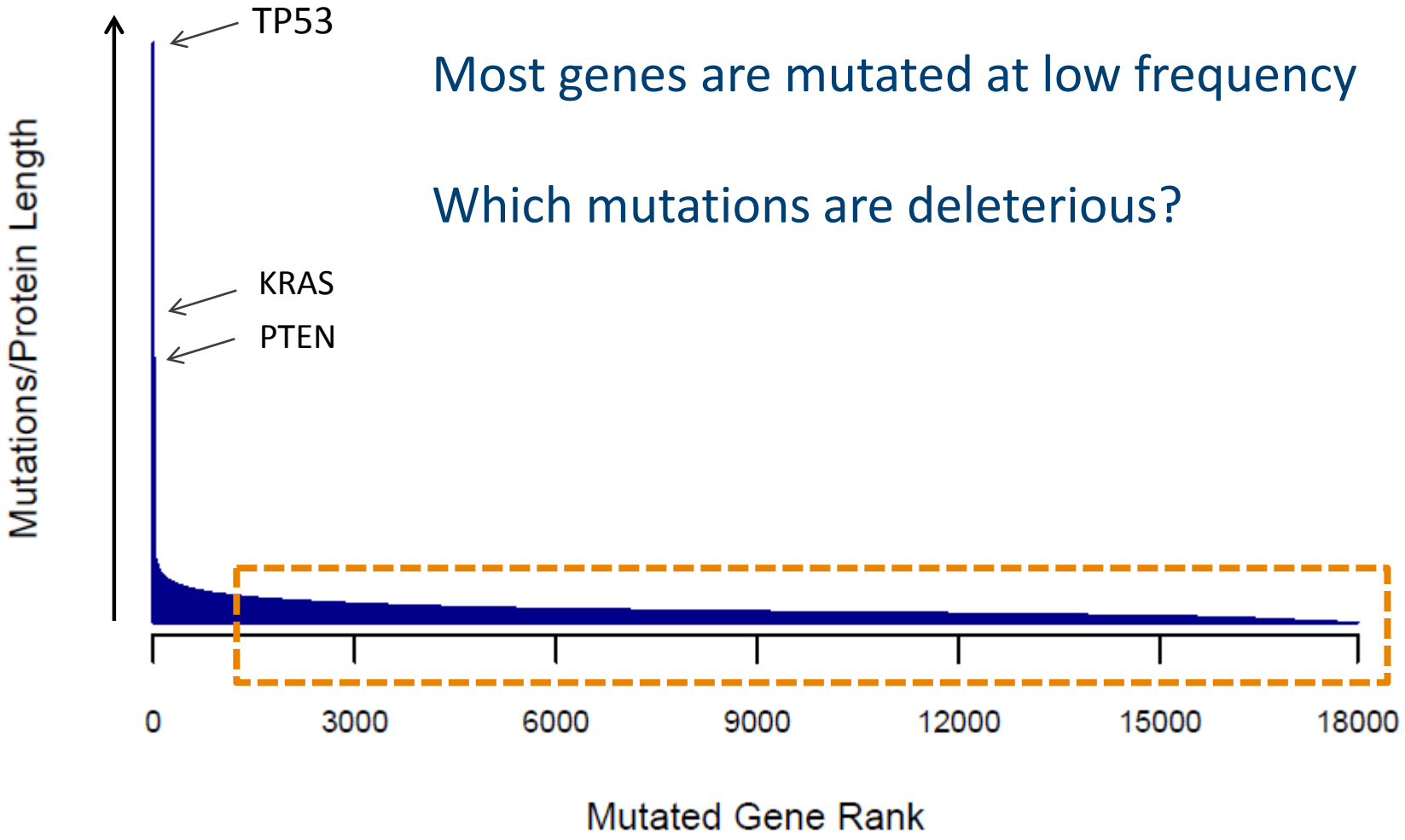
Normalized mutation frequencies

Twenty tumor types



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Normalized mutation frequencies
Twenty tumor types

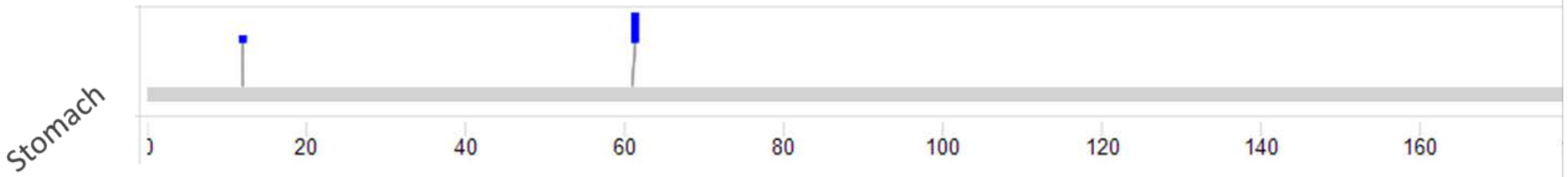


Most genes are mutated at low frequency

Which mutations are deleterious?

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

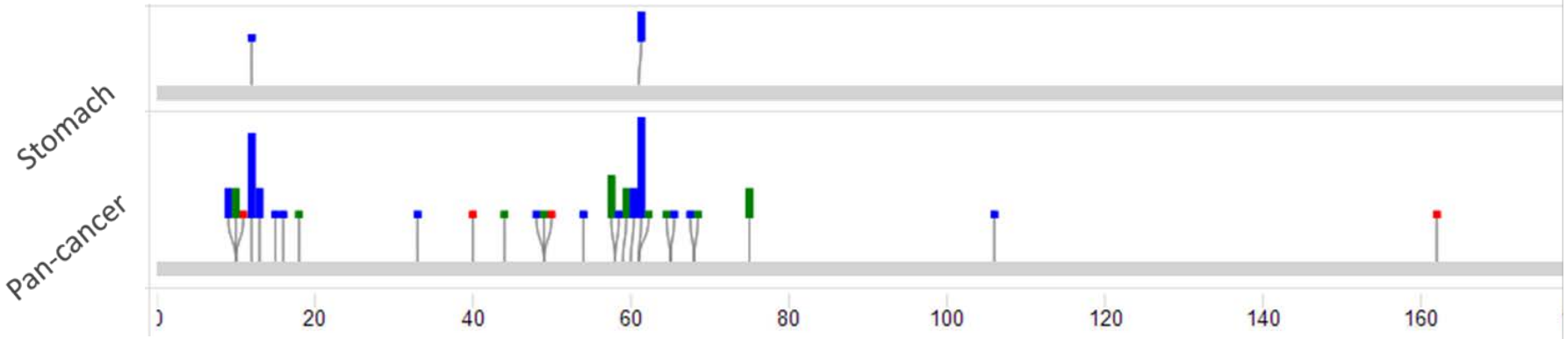
Example protein mutation spectrum



$\log_2$ scaled counts per position

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

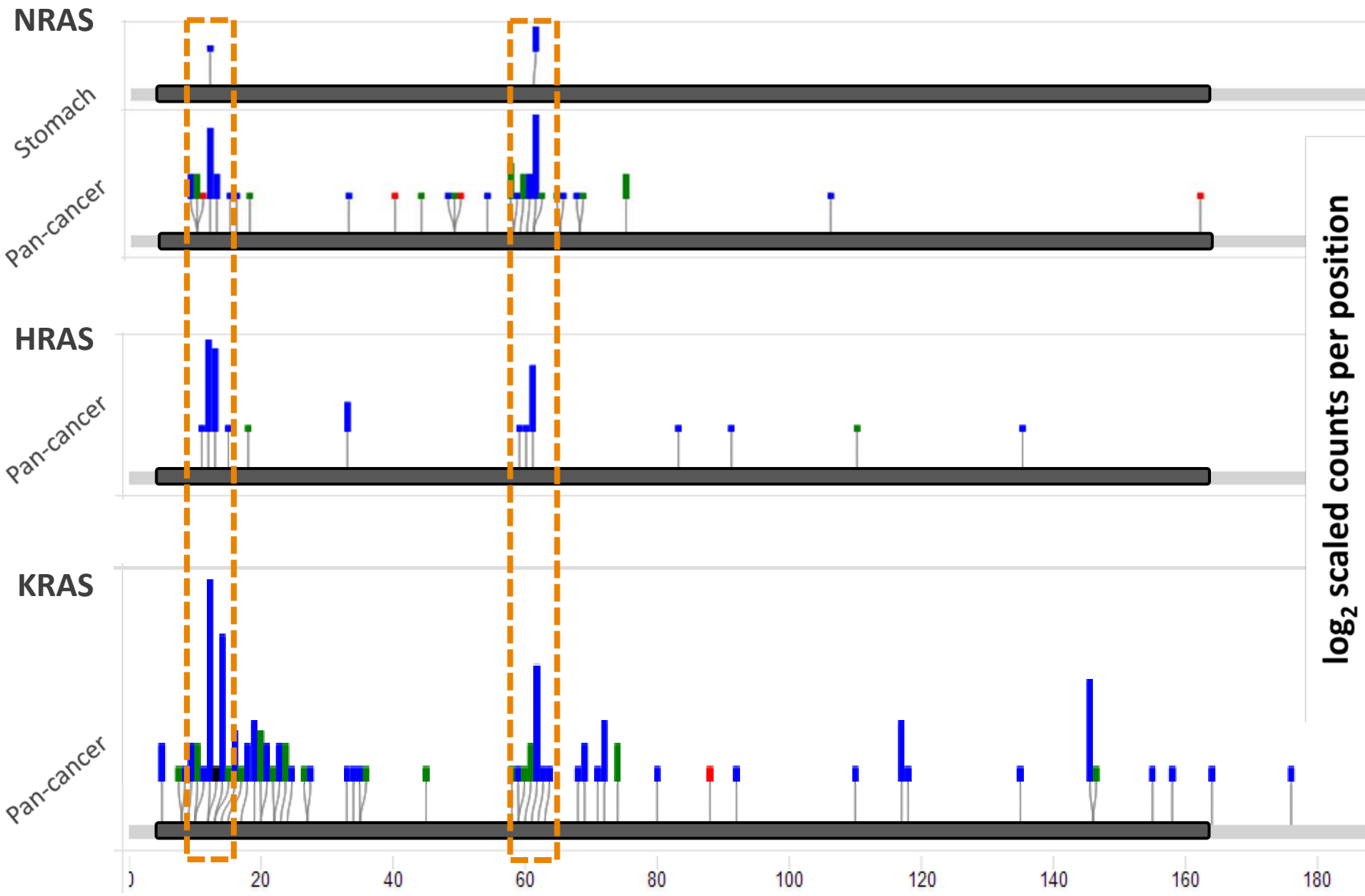
Pan-cancer mutation spectrum



$\log_2$ scaled counts per position

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Pan-cancer Ras domain mutation spectra



Ras domain structure alignment

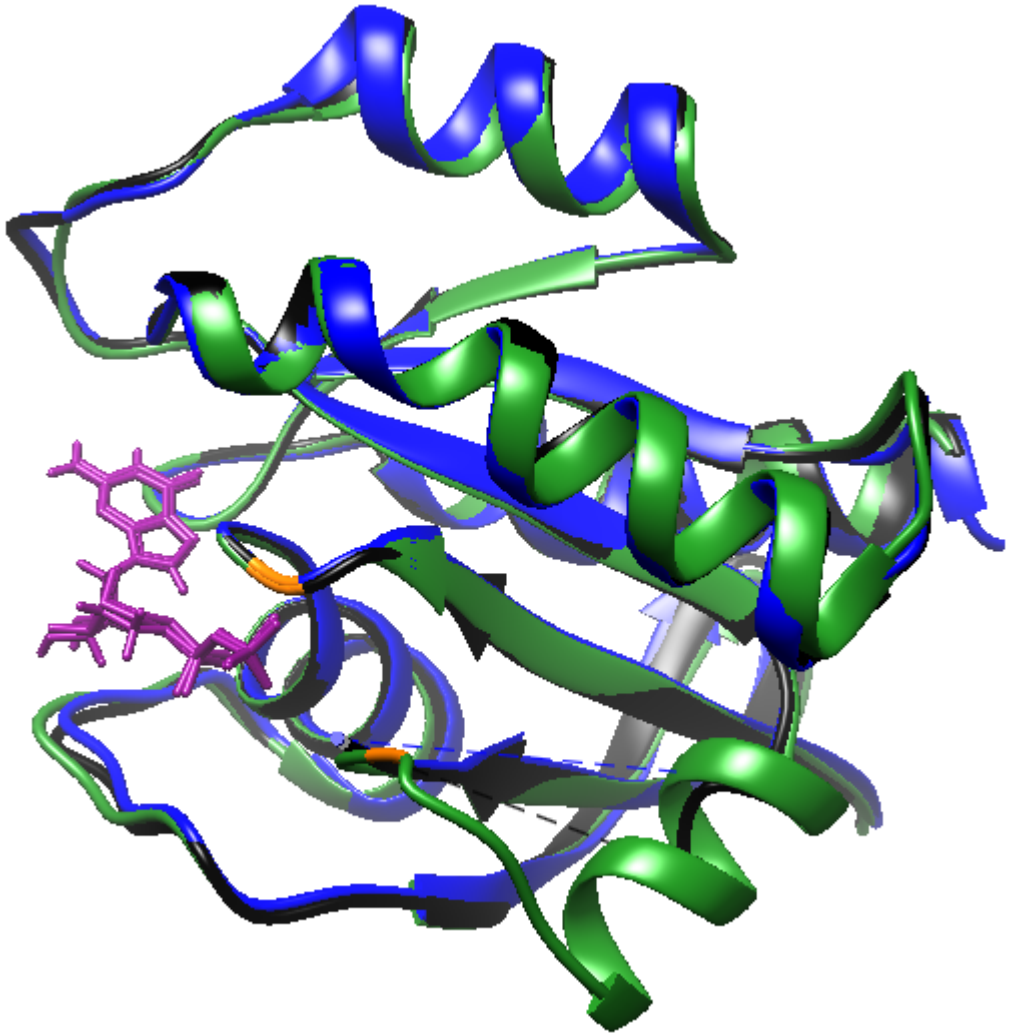
KRAS

NRAS

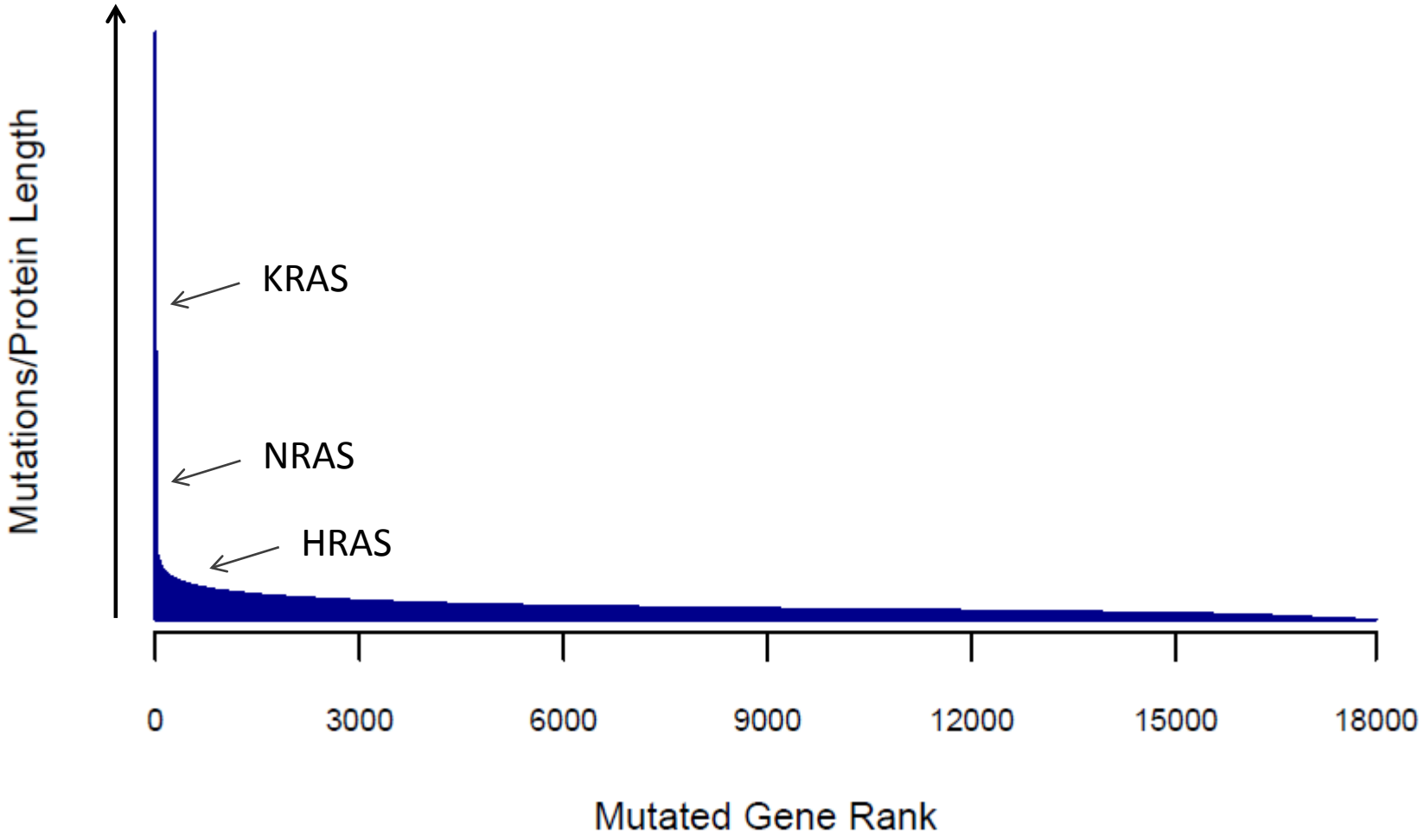
HRAS

GDP

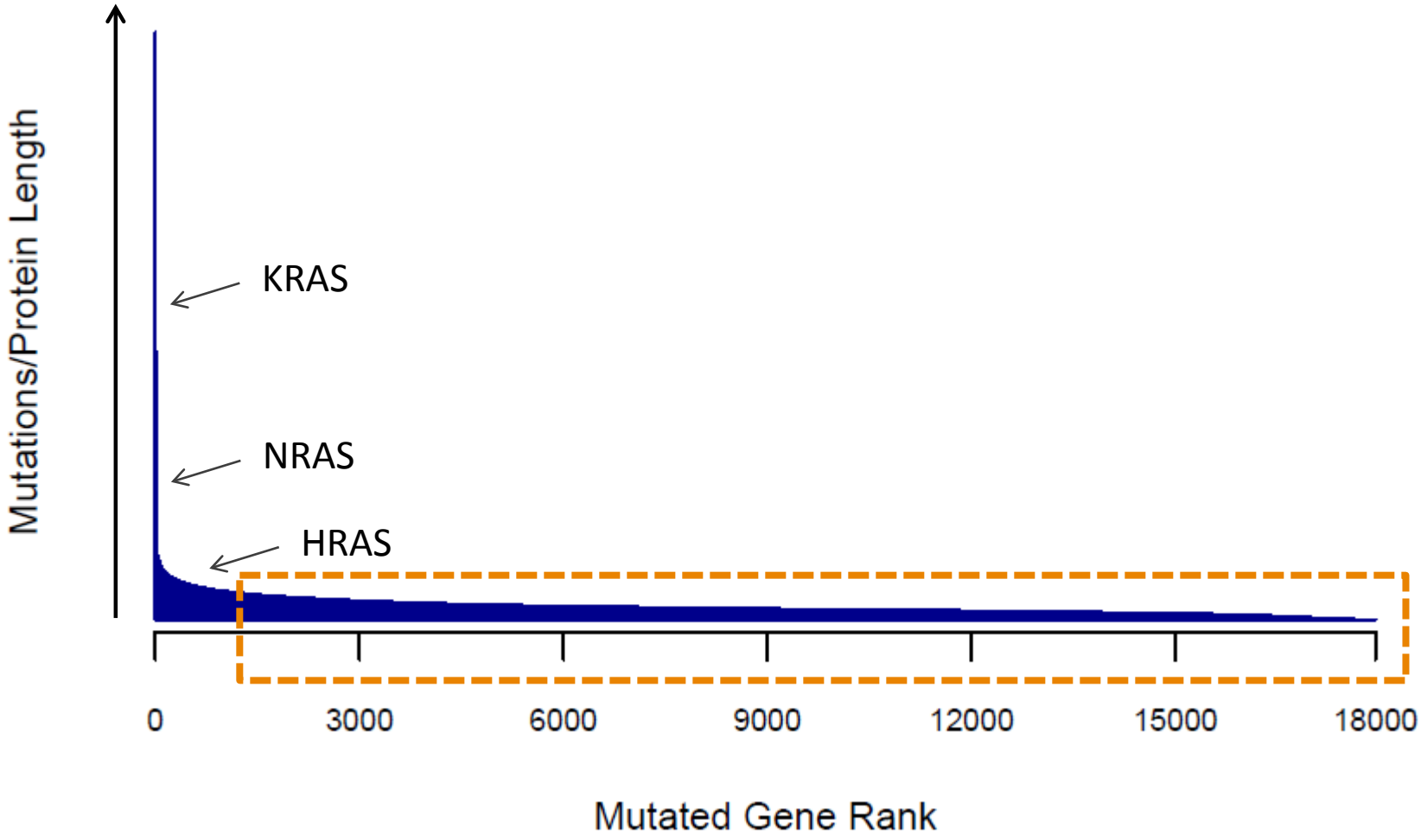
Hotspots



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.



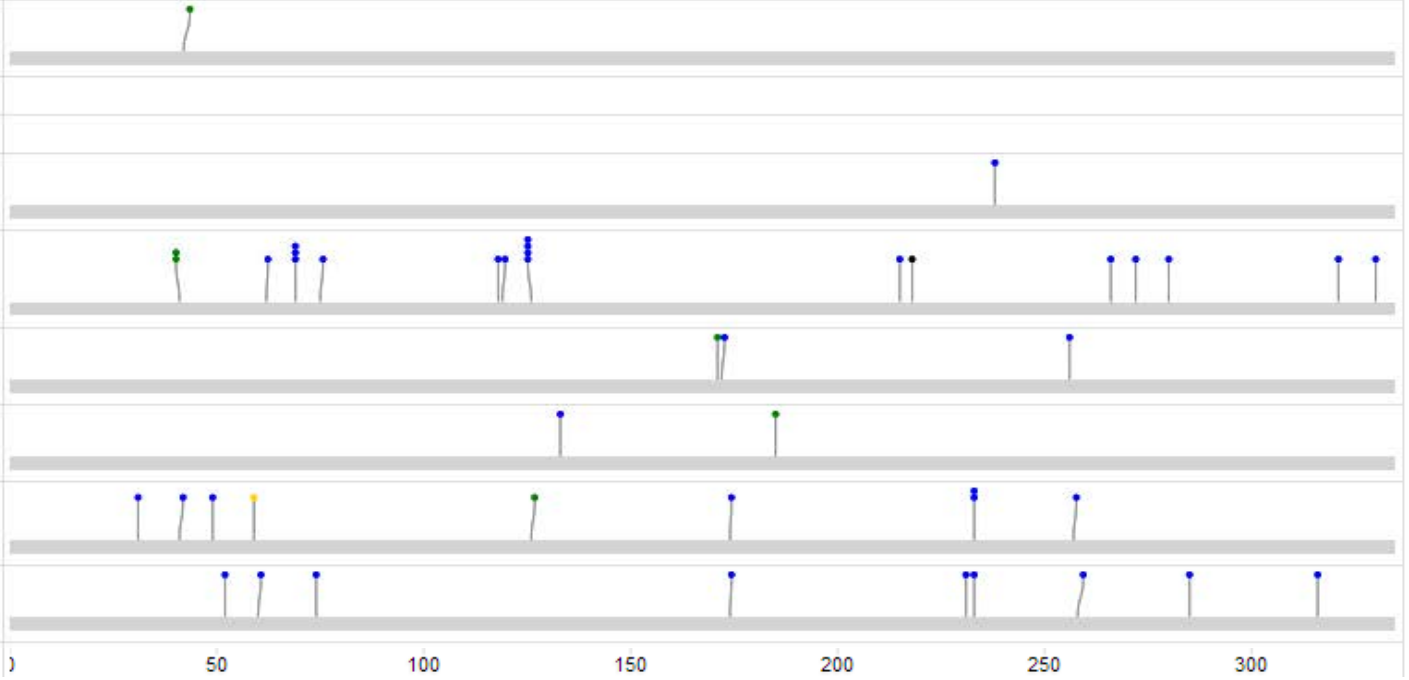
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Typical pan-cancer mutation spectrum

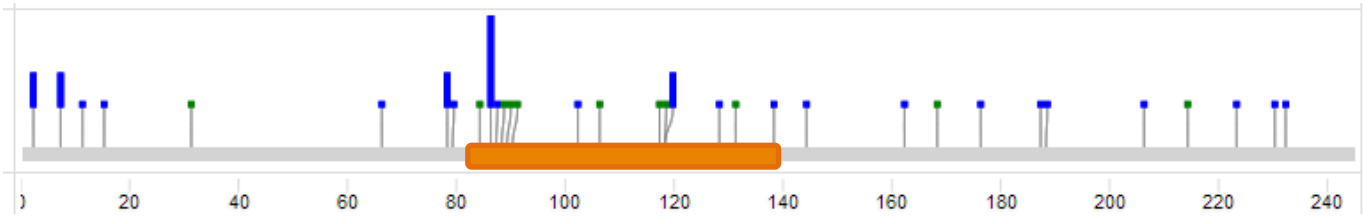
Tumor Type	MutSig Rank	Samples (#)
BLCA	9638	1
BRCA	952	0
GBM	14085	0
KIRC	2051	1
LUAD	925	14
LUSC	9703	2
OV	3421	2
STAD	1118	8
UCEC	2211	7

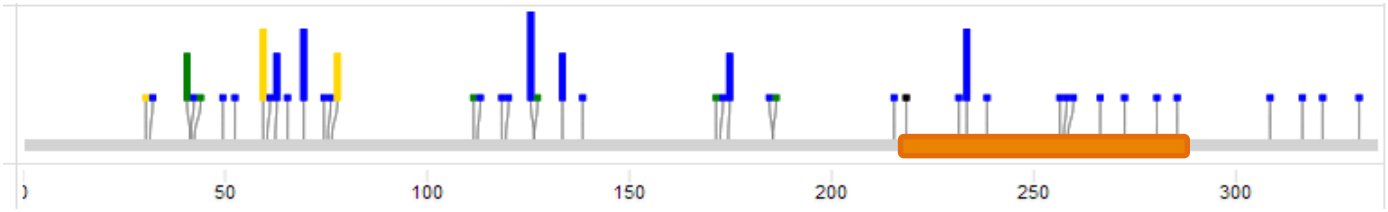


Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

LineUp Approach

Gene A
Pan-cancer



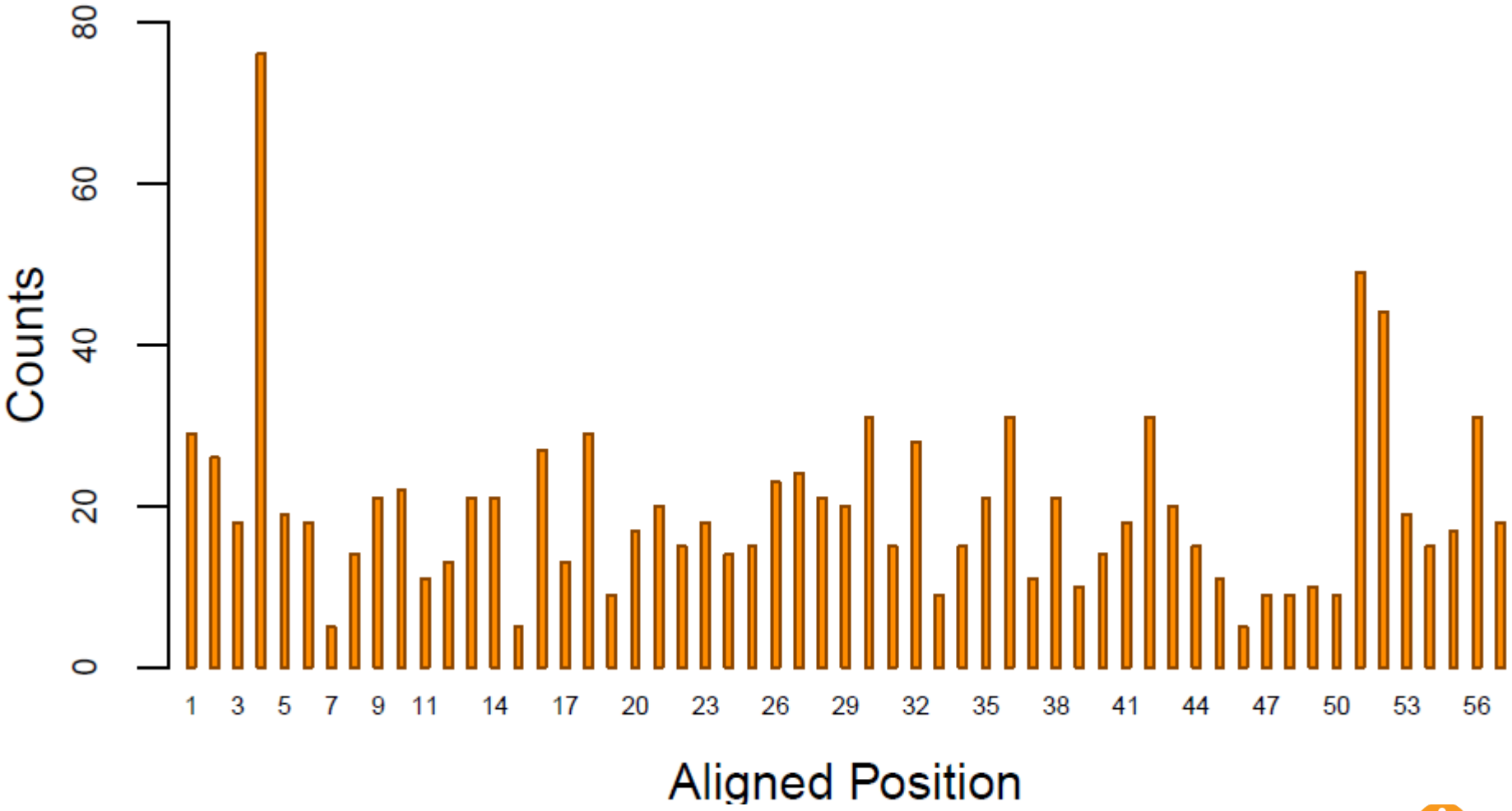


Gene B
Pan-cancer

- Align sequences from all matching domains across all tumor types
- Evaluate missense mutation frequencies
- Applies to ~40% of missense mutations

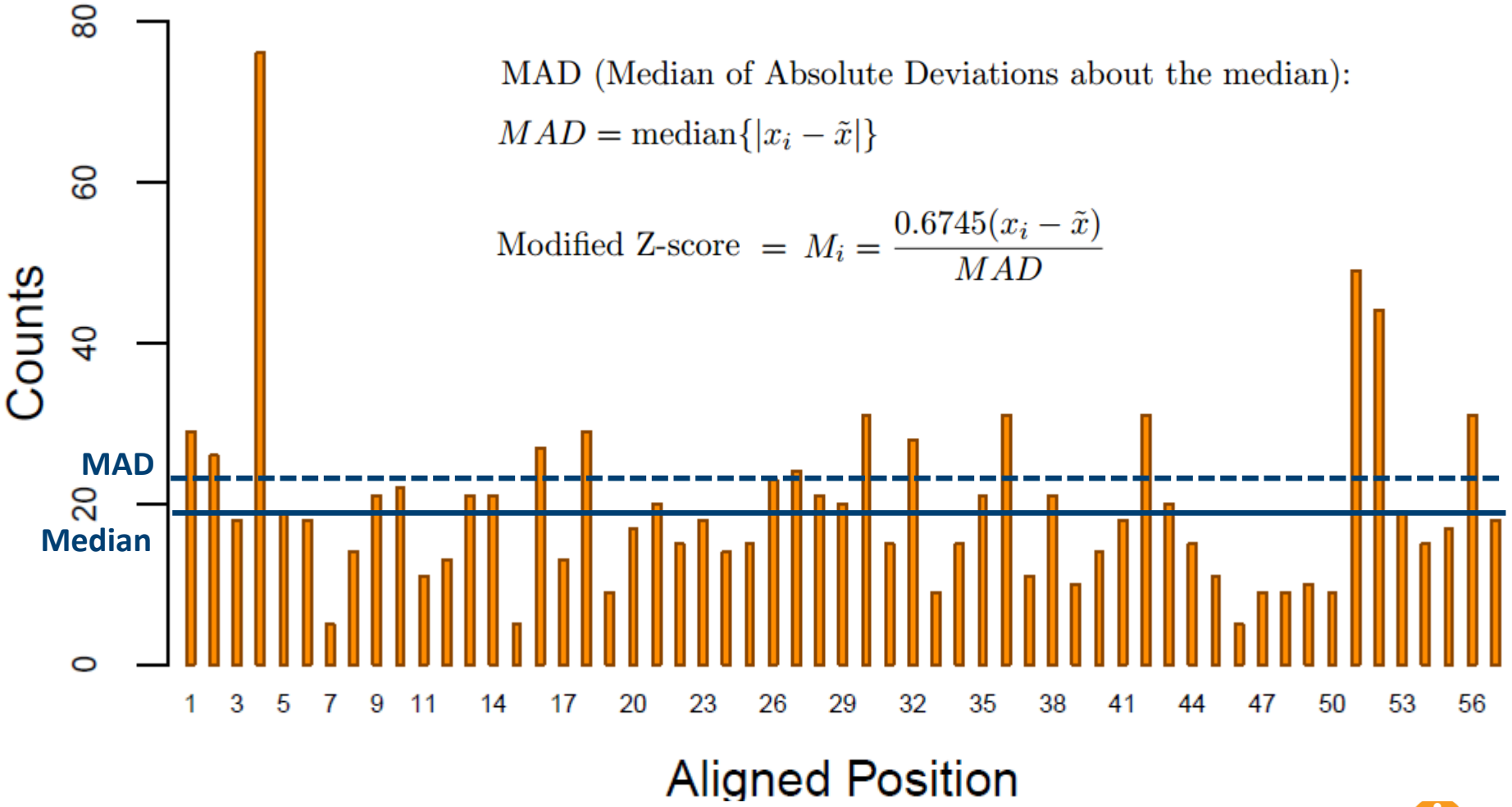
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Mutations per position for two hundred aligned Homeobox domains across twenty tumor types



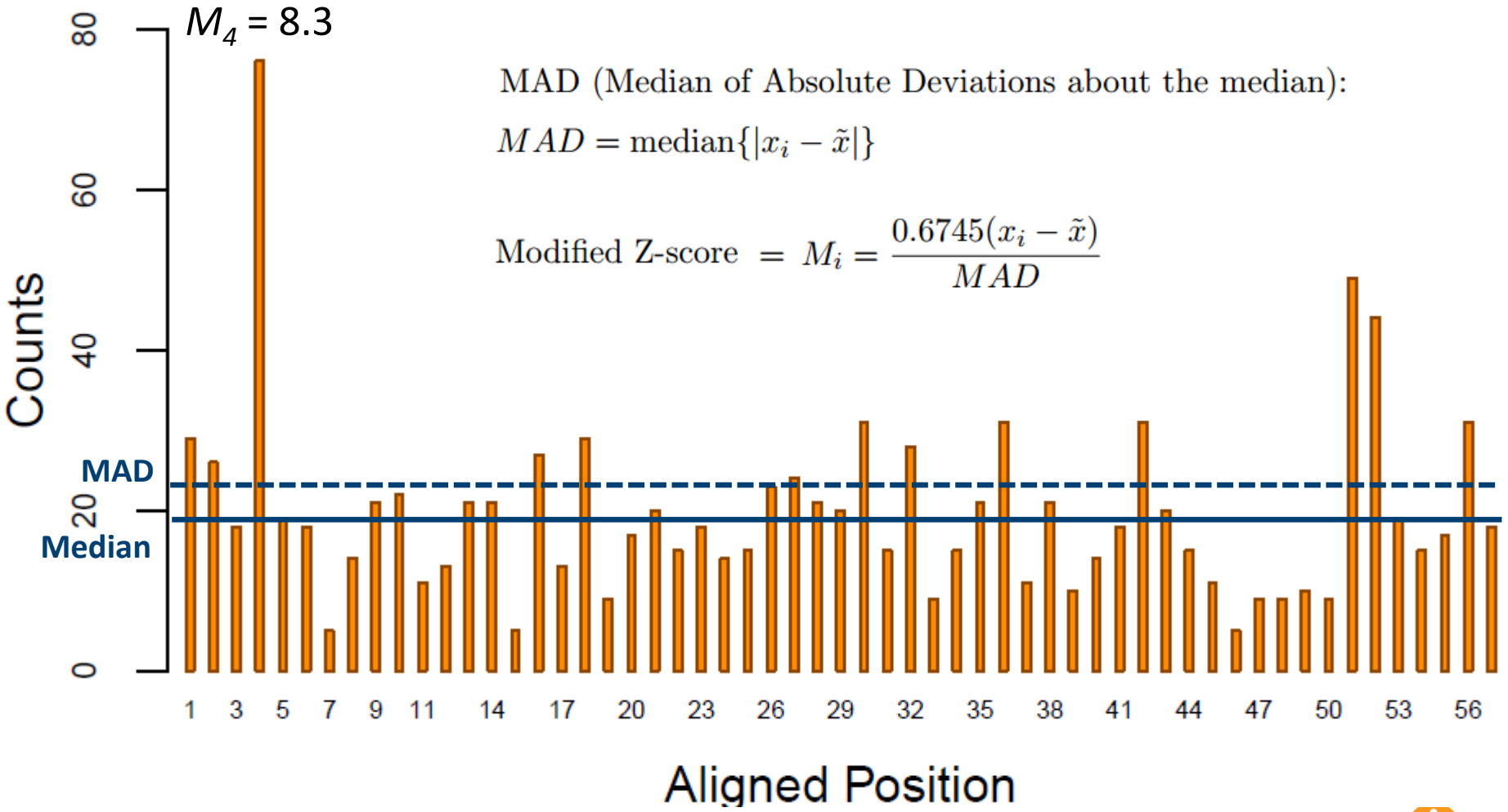
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Identifying outliers: Modified Z-score



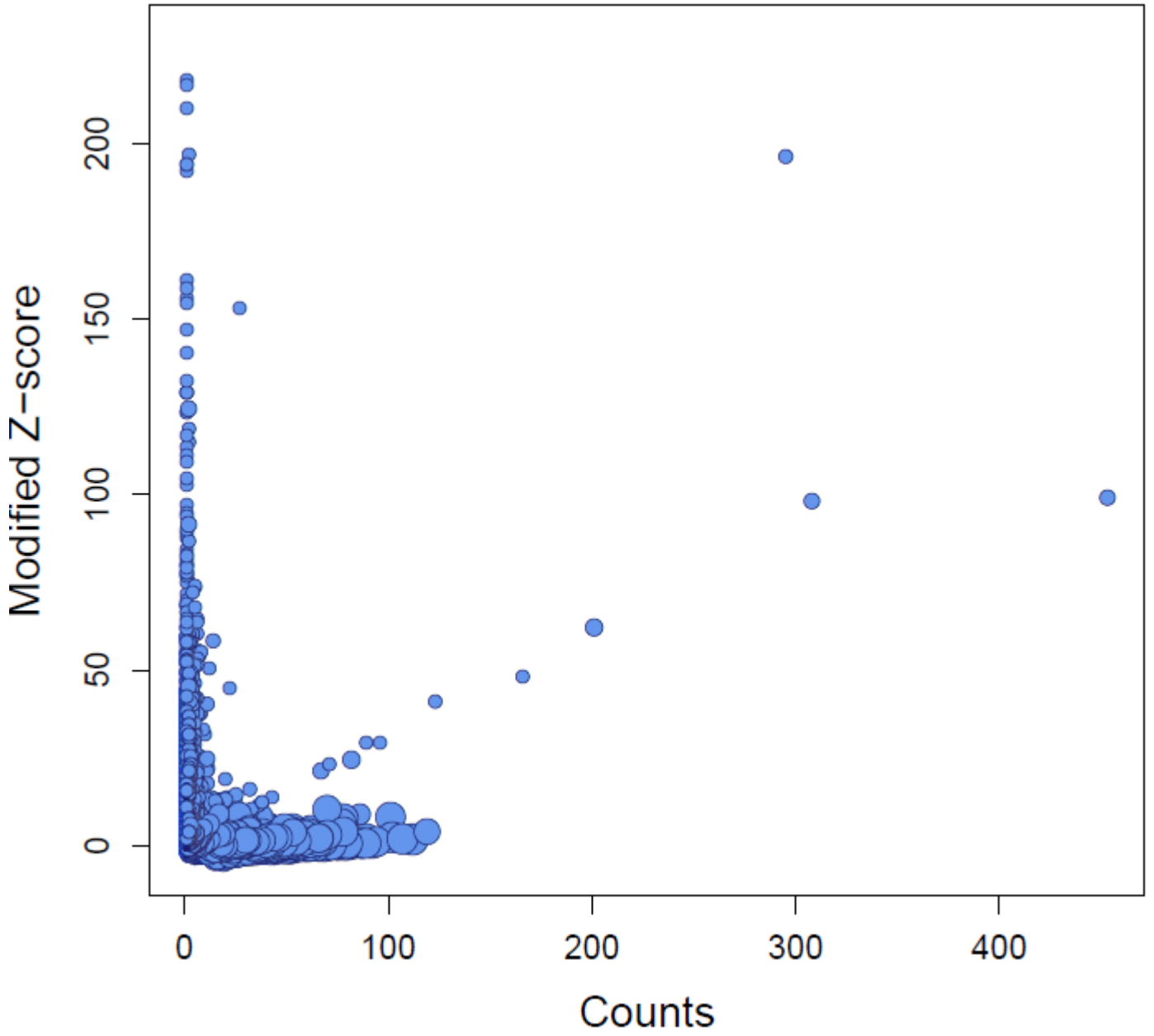
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Identifying outliers: Modified Z-score

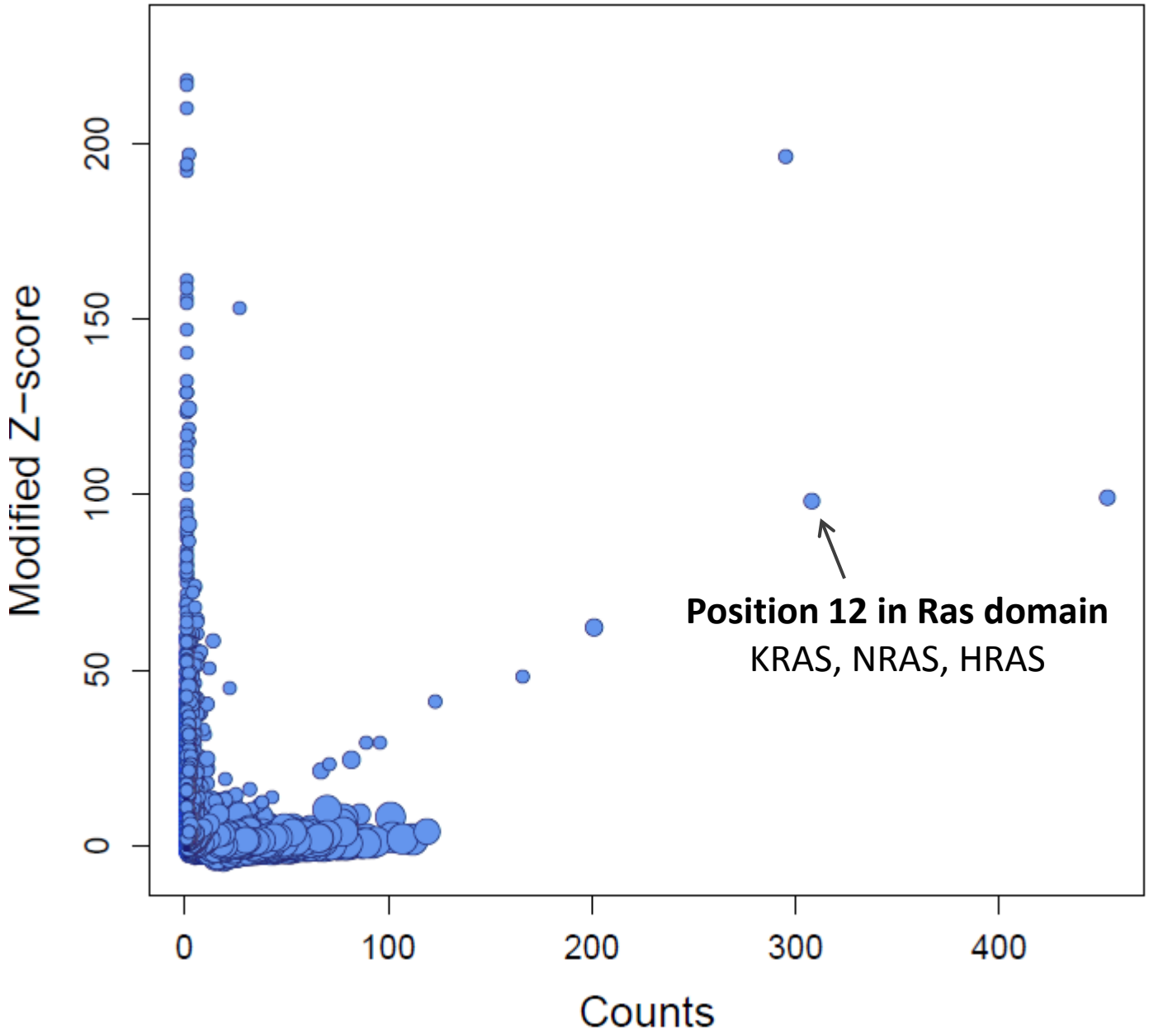


Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

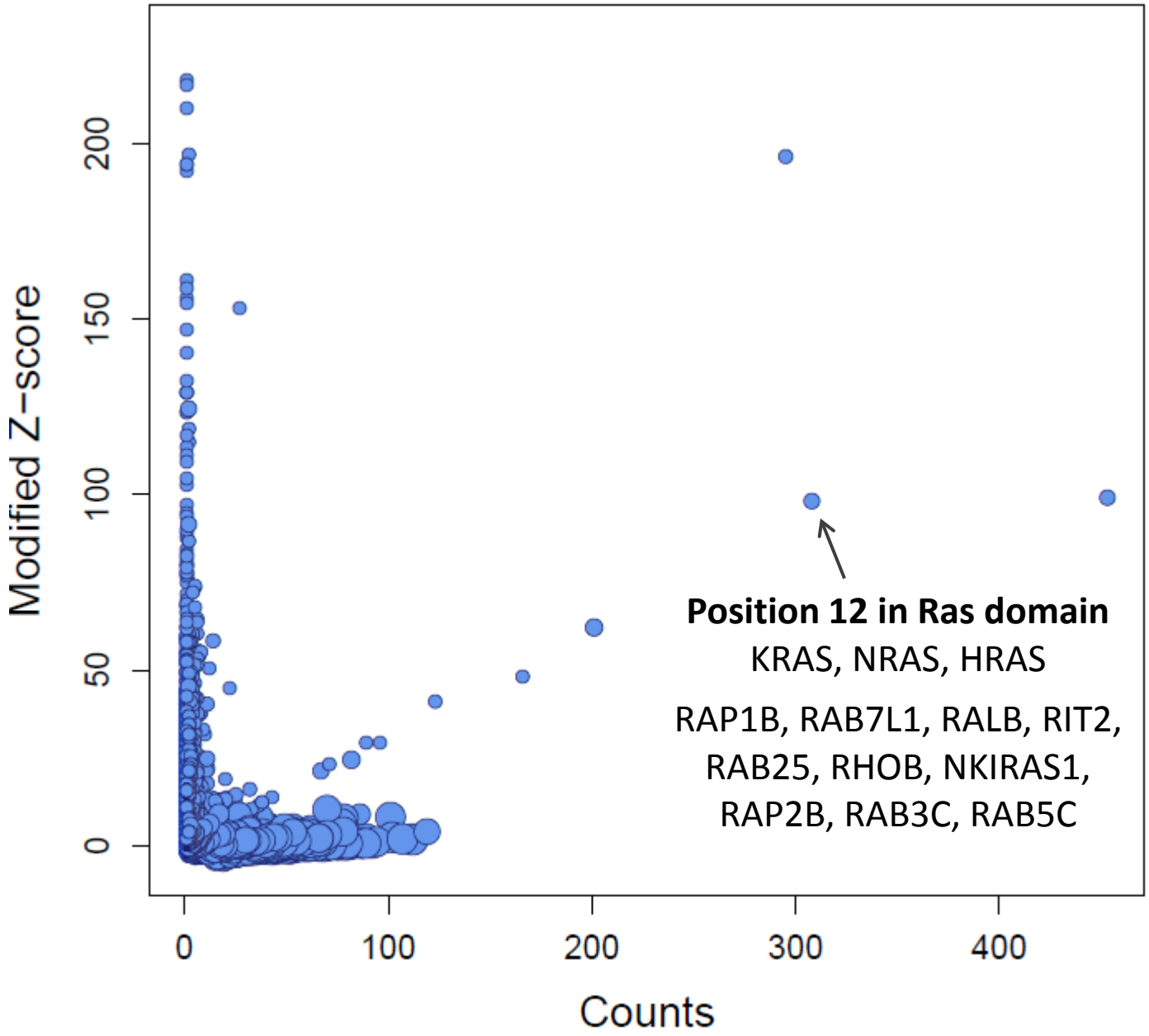
All positions within all domains



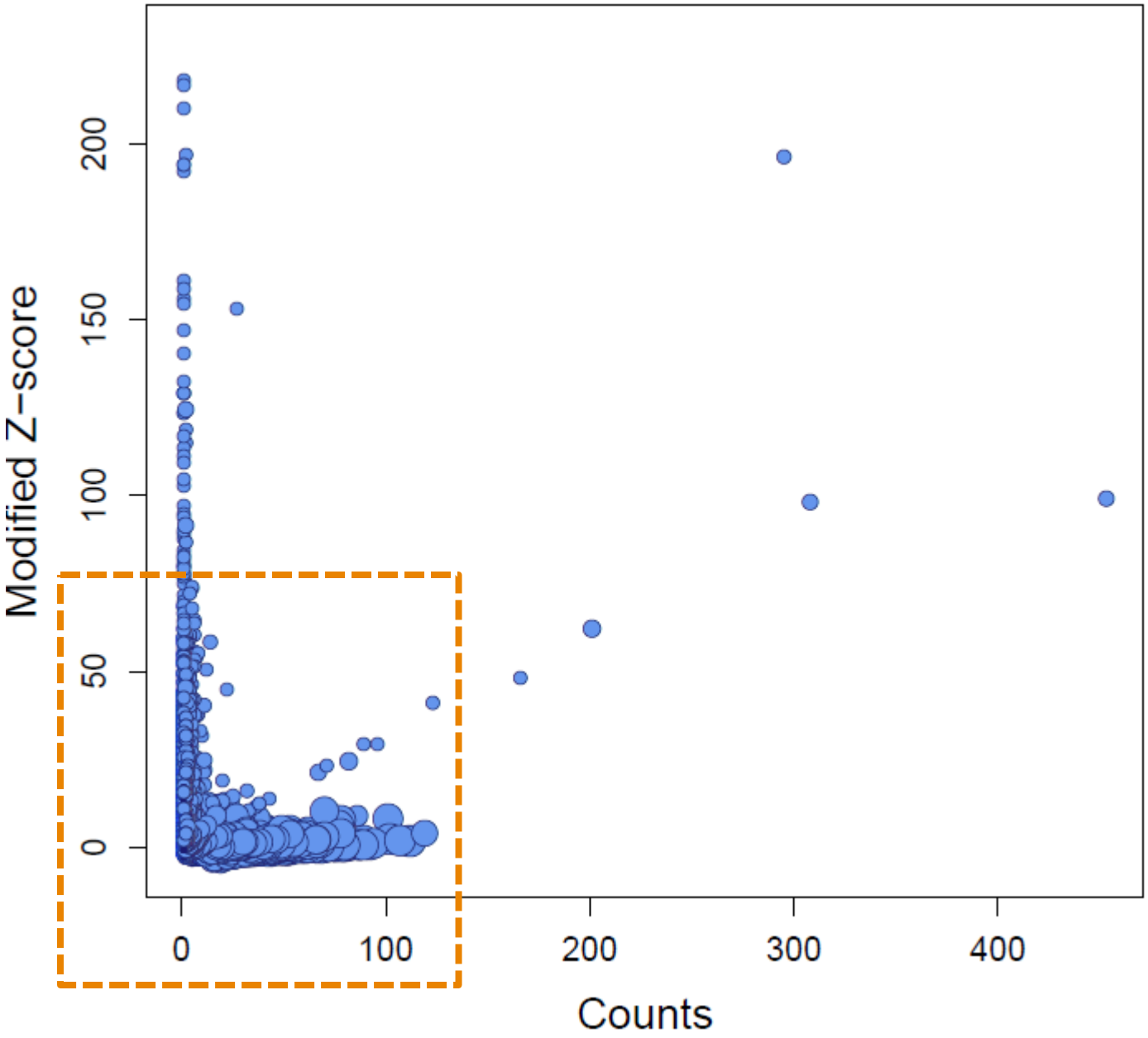
High Modified Z-score and counts for Ras domain



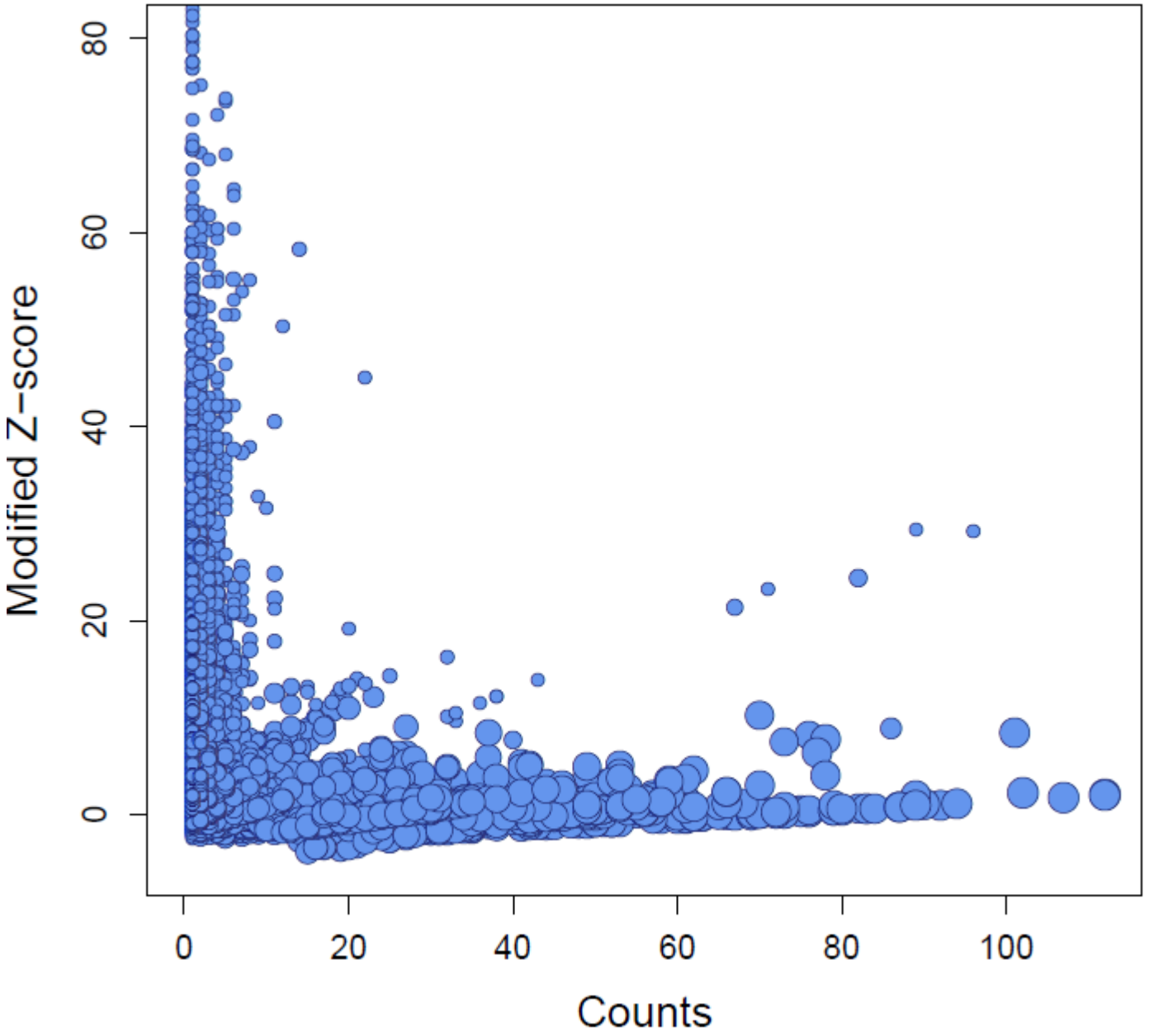
Multiple other Ras genes mutated at the same position



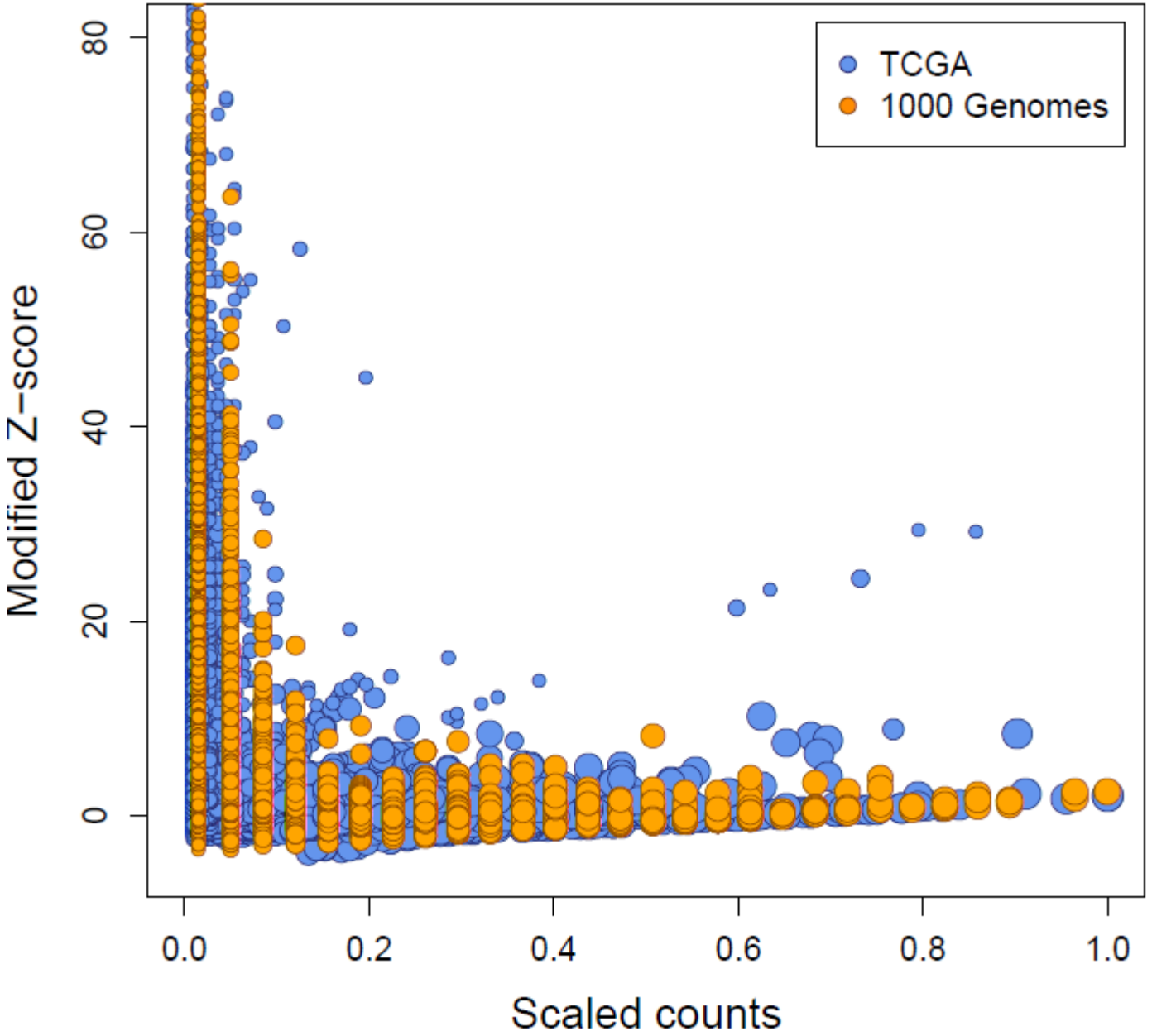
Investigating genes with low mutation frequency



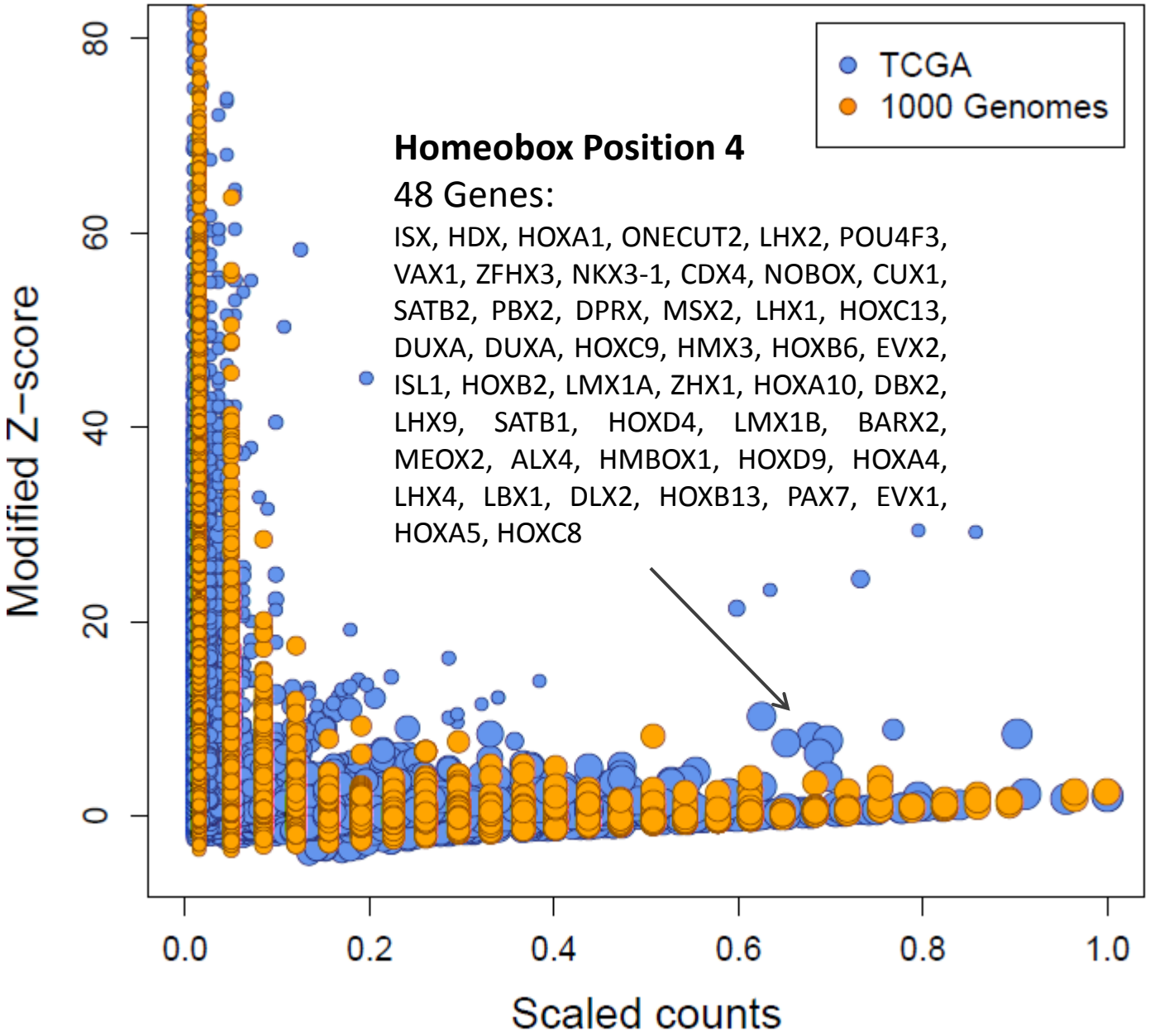
Investigating genes with low mutation frequency

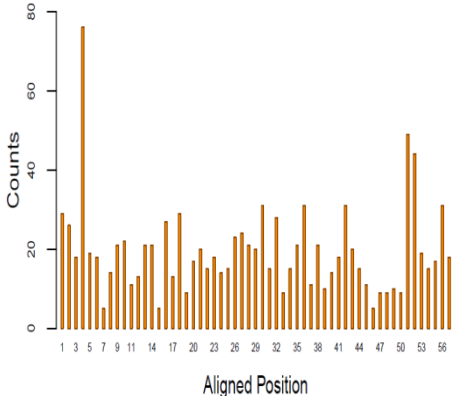


Background expectation:
Comparing TCGA to 1000 Genomes



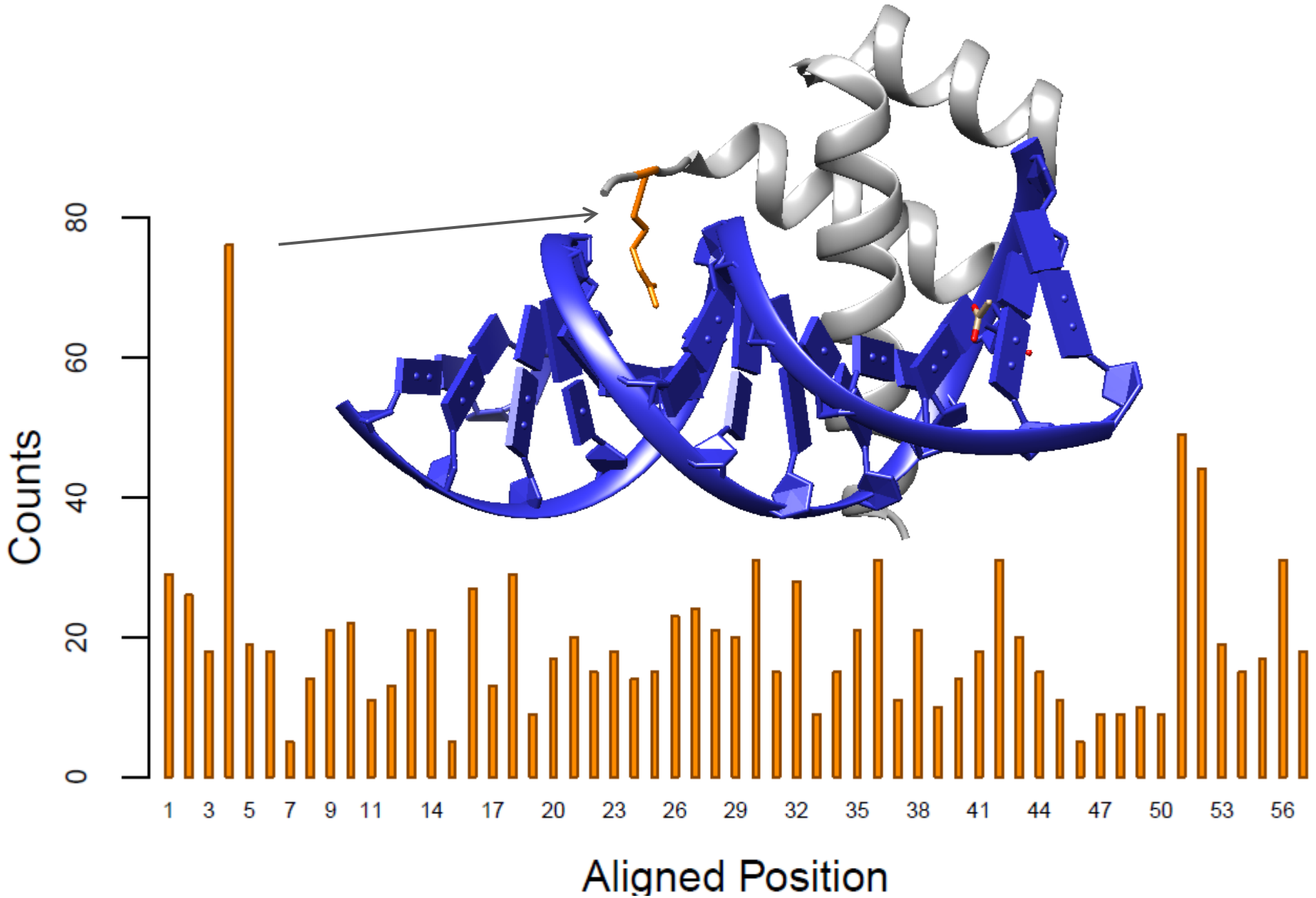
Outlier position within Homeobox domain





Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Structural interpretation



Summary and Next Steps

- We have comprehensively evaluated mutations at all positions within all domains to identify low frequency but likely deleterious mutations
- Hotspots outside of domains and mutations that broadly disrupt structure and function not addressed
  - Integration with other methods is essential
- Functional validation of low frequency events in such data sets remains challenging
- As cancer and normal genome data size increases, more robust normalization per position per domain can be achieved

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Acknowledgements

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

The Cancer Genome Atlas

National Cancer Institute

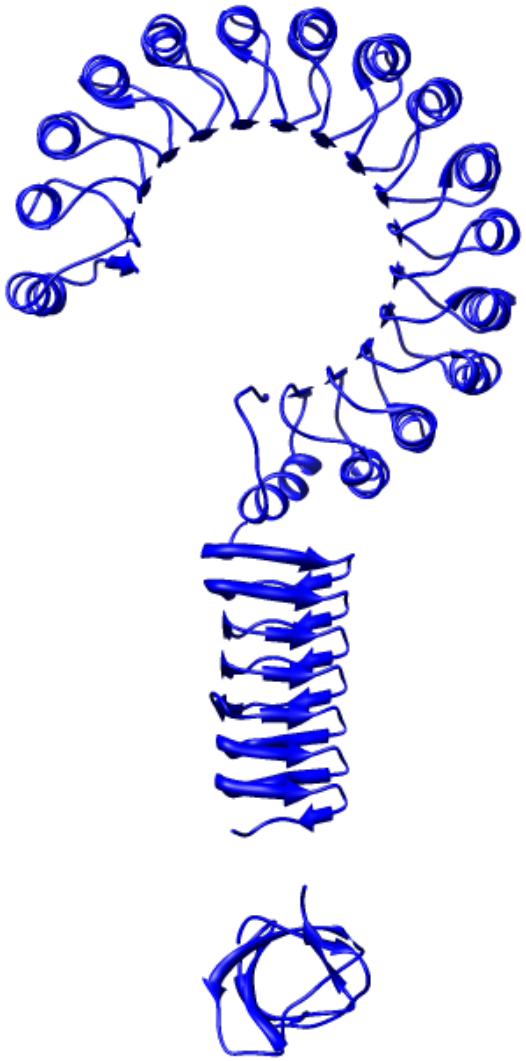
National Human Genome Research Institute

Ilya Shmulevich

Tae Denwongkun

Theo Knijnenburg

Vesteinn Thorsson
Kalle Leinonen
Richard Kreisberg
Hector Rovira
Sheila Reynolds
Lisa Iype
Natalie Tasman



2bnh: alpha-beta horseshoe

1hv9: left-handed beta helix

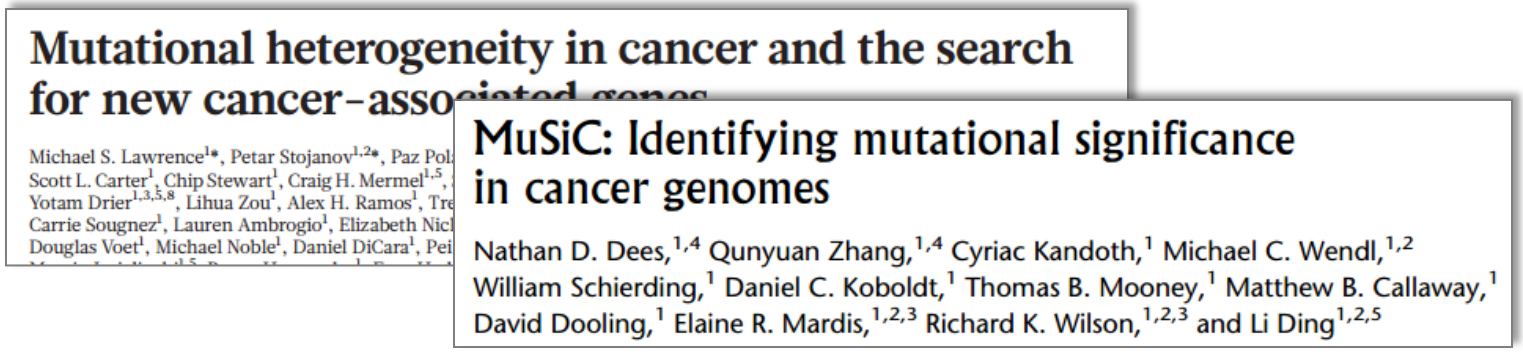
1m30: SH3-like barrel

Institute for Systems Biology

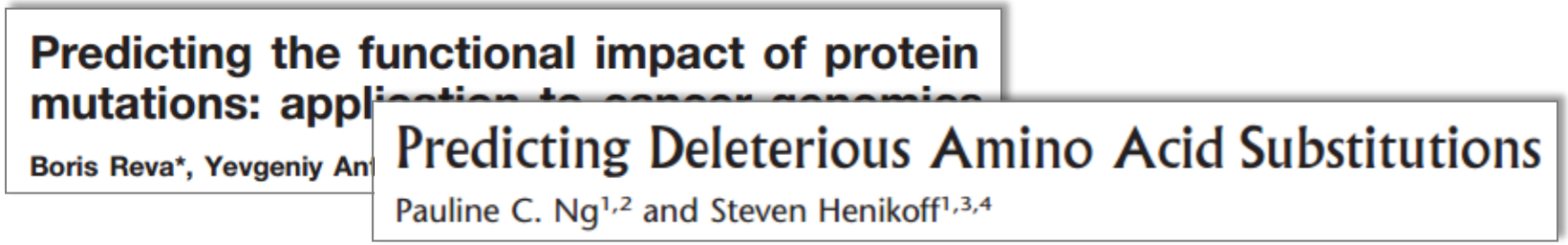
Revolutionizing Science. Enhancing Life.

Example approaches and methods

- Statistical assessment of multiple samples (MutSig, MuSiC, …)



- Sequence conservation (MutationAssessor, SIFT, …)



- Machine learning and classification (PolyPhen-2, CHASM, …)