Uncovering and characterizing splice variants associated with survival in lung cancer patients

Sean West, Sushil Kumar, Surinder K. Batra,

Hesham Ali, and Dario Ghersi

University of Nebraska Medical Center

University of Nebraska Omaha

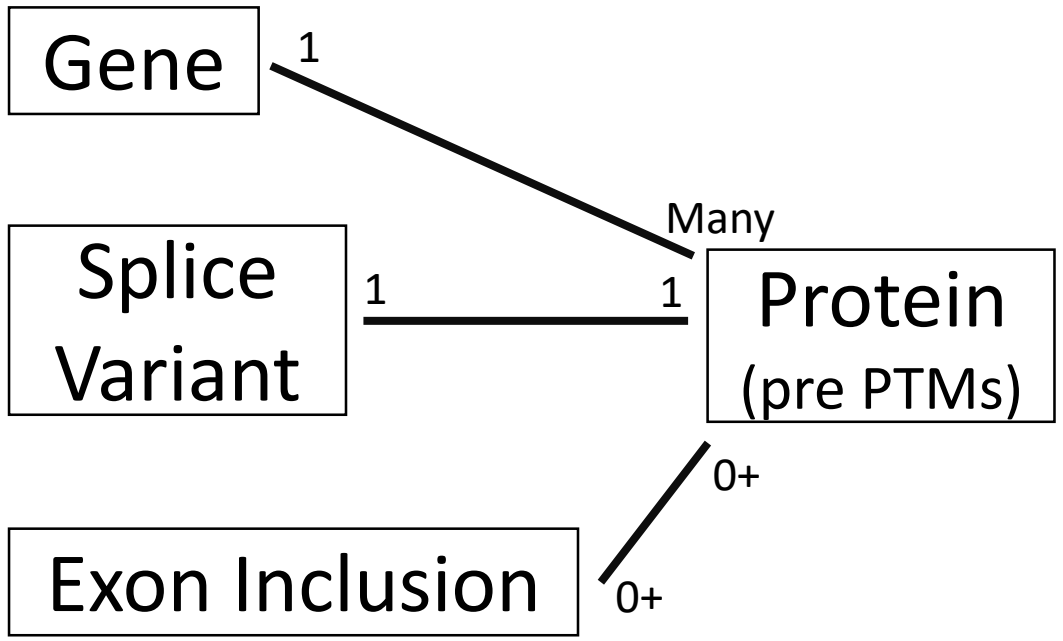
Molecular targeted therapy

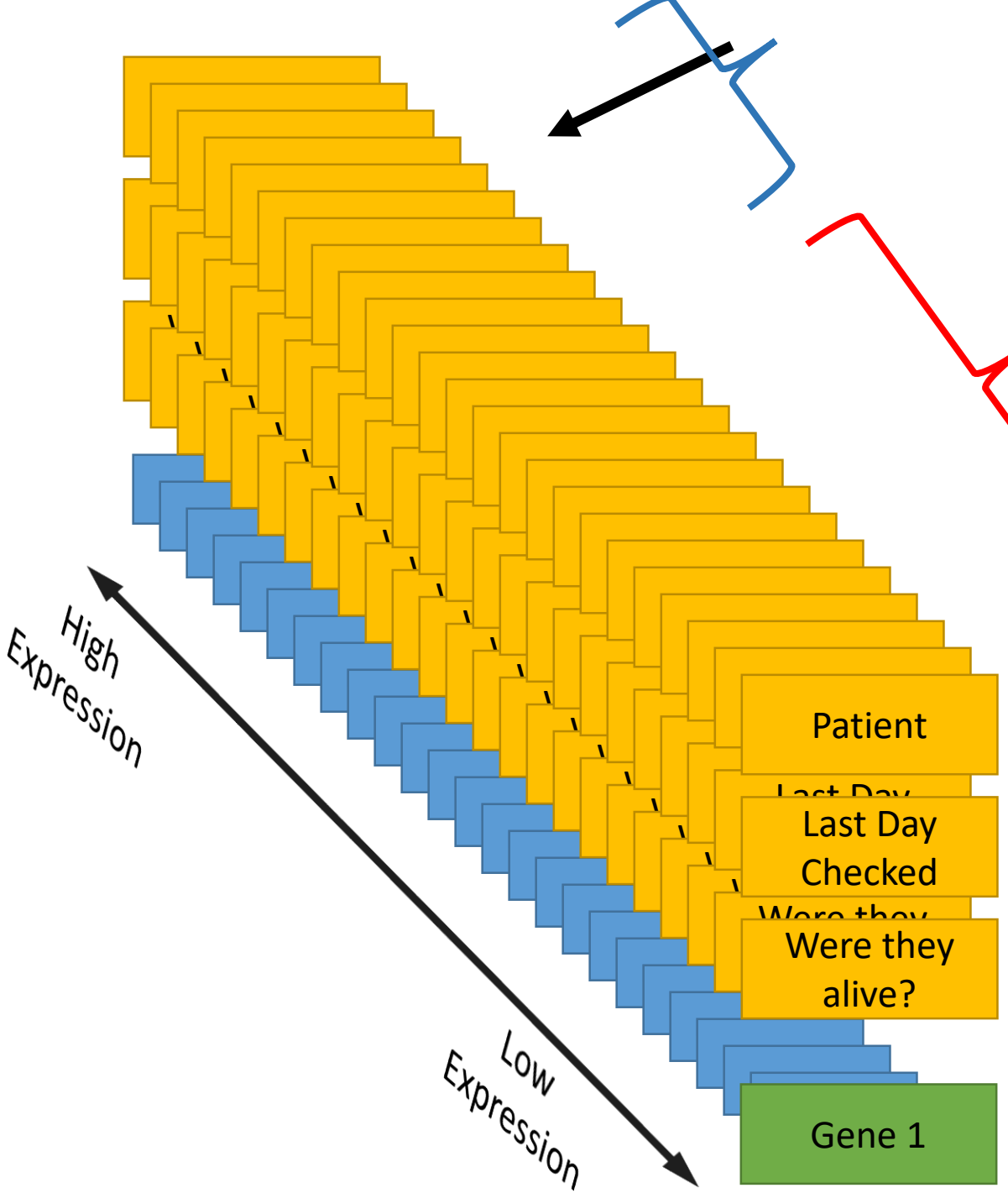
- Small molecules to inhibit specific pathways important for cancer cell survival and invasiveness.*
- MTT benefits from highly-specific molecular mechanisms.
- Methods to target splicing in cancer have recently been proposed.**

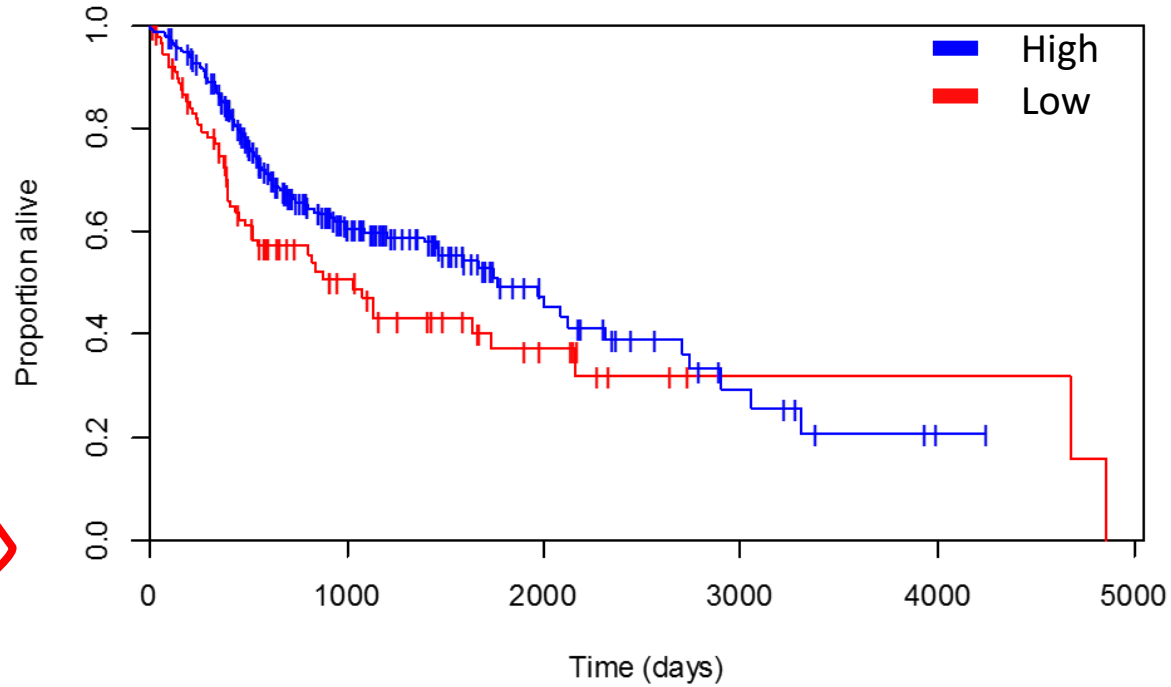
*Lee YT, Tan YJ, Oon CE. Molecular targeted therapy: Treating cancer with specificity. European journal of pharmacology. 2018
**Lee SCW, Abdel-Wahab O. Therapeutic targeting of splicing in cancer. Nature Medicine. 2016;doi:10.1038/nm.4165

Splice variant level analysis

- Genes frequently produce multiple gene products.
- Exon junctions frequently cannot be assigned to individual protein products.
- Isoforms have a naturally interpretable alternative splicing program.
- Domains can be accurately predicted for isoforms.

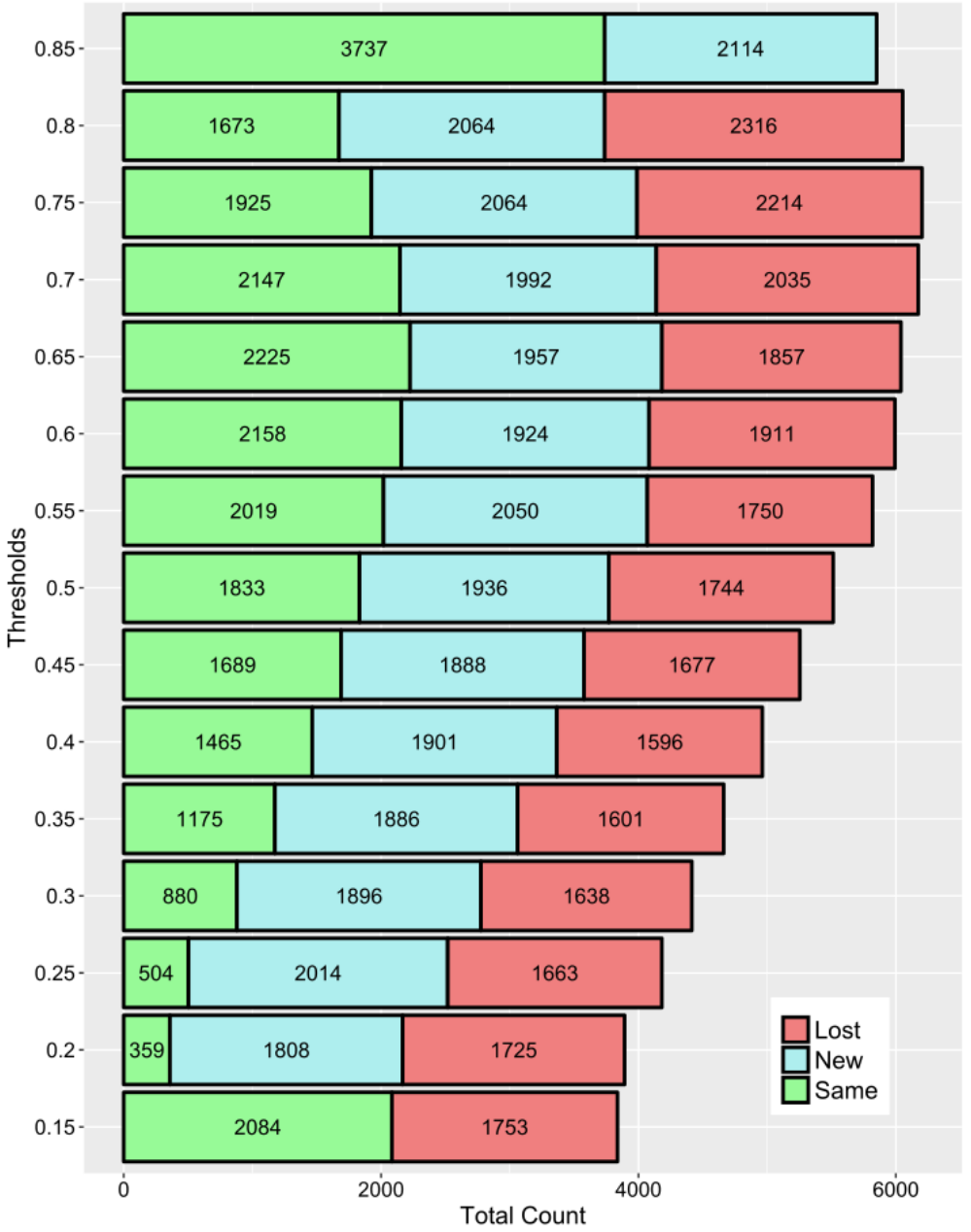



High Expression
Low Expression
Patient
Last Day Checked
Were they alive?
Gene 1

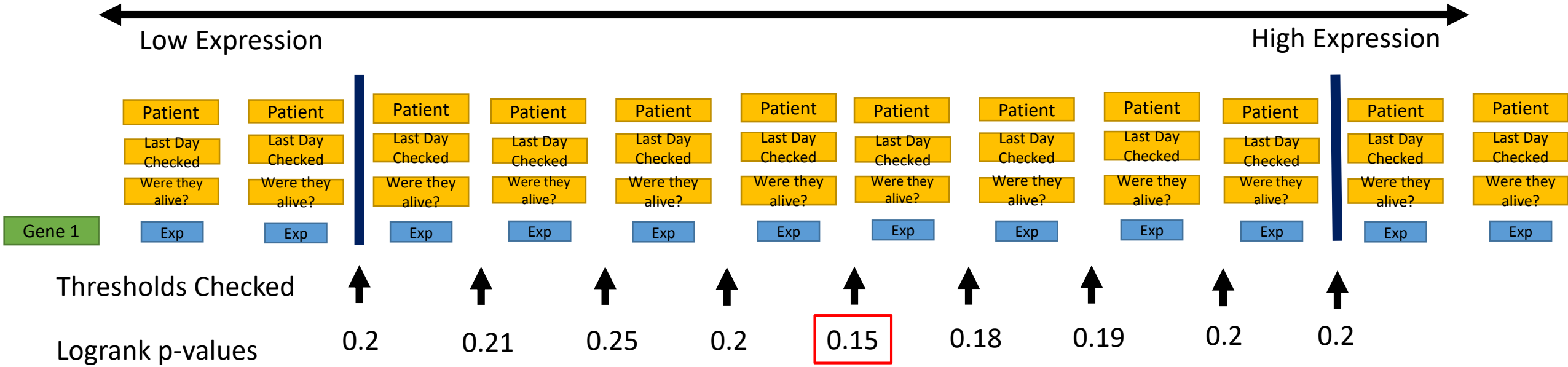

High
Low
Proportion alive
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Time (days)

Choosing a single threshold

- Most studies choose a single threshold to split into *low*/*high* expression.
- This produces a valid, uniform p-value distribution.
- Choosing a single threshold is not robust to small changes in the threshold.



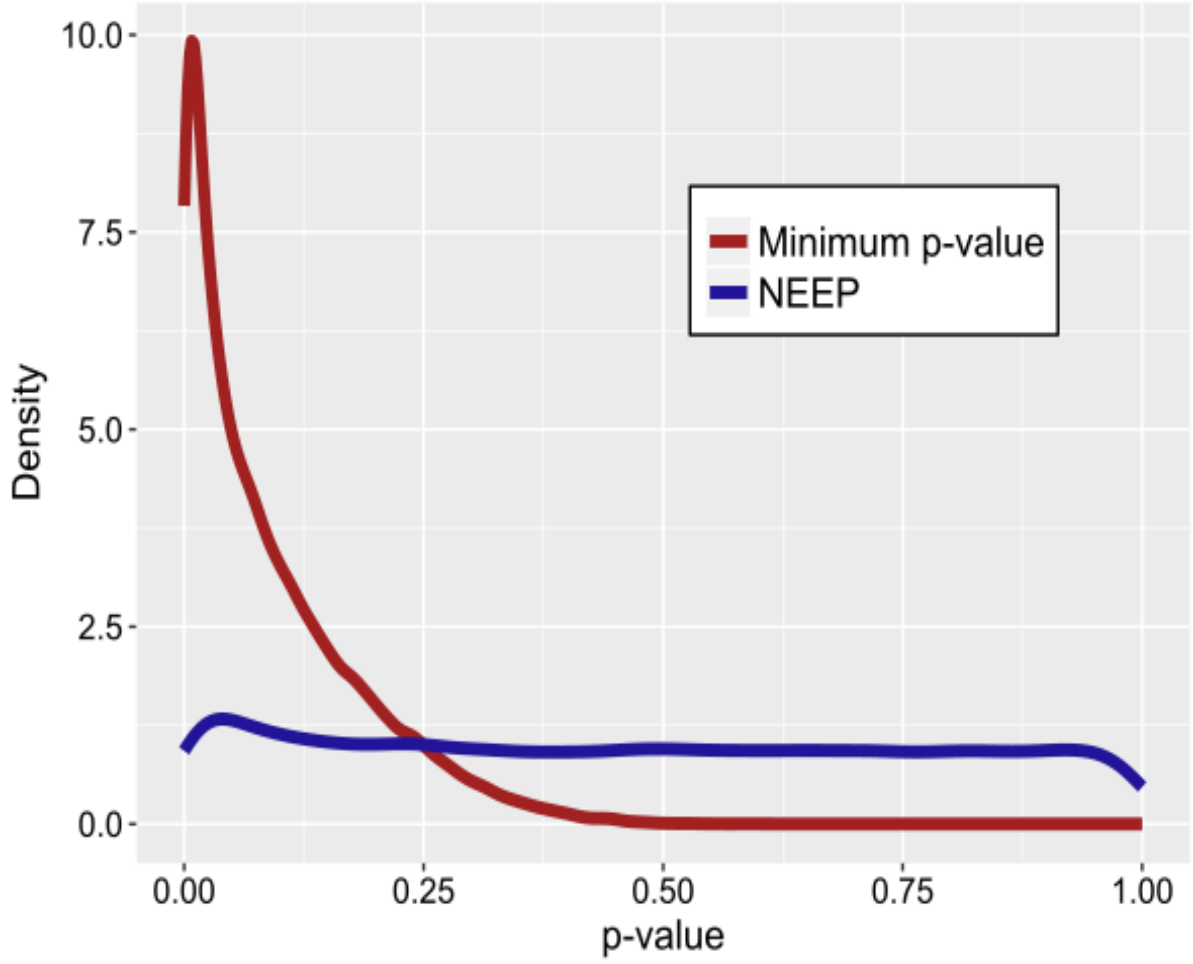
Choosing the minimum p-value within a range of thresholds increases reliability*



*Sehgal V, Seviour EG, Moss TJ, Mills GB, Azencott R, Ram PT. Robust selection algorithm (RSA) for multi-omic biomarker discovery; integration with functional network analysis to identify miRNA regulated pathways in multiple cancers. PloS one. 2015;10(10):e0140072.

Minimum logrank p-value as a statistical test

- Always choosing the minimum p-value within a range results in a skewed distribution.
- Statistical tests should be uniform under the null hypothesis.
- p-value adjustment assumes a uniform distribution of p-values.



Null-based Empirical Estimation of P-values (NEEP)

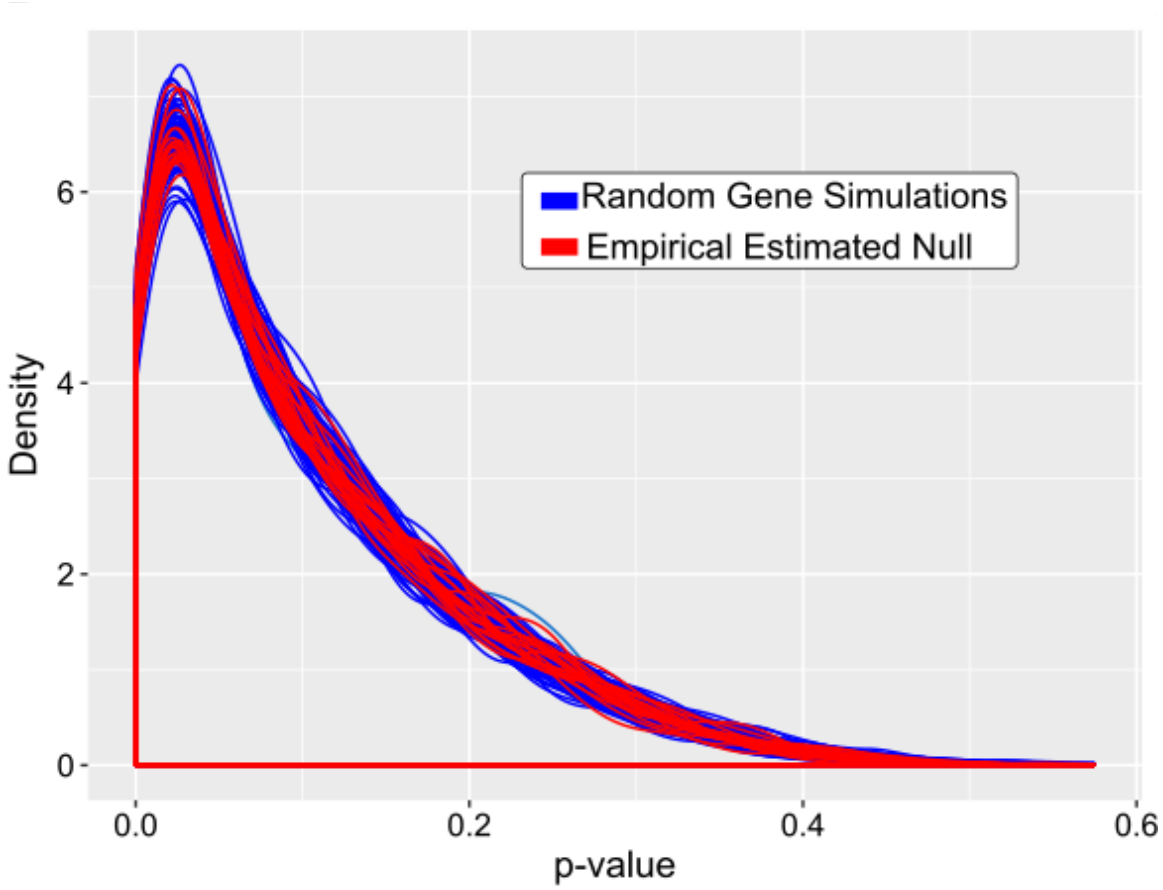
- We can empirically determine the true p-values by using the true Null p-value distribution.

$$\text{size of Null} = \sum_{i=\lfloor l*s \rfloor}^{\lfloor h*s \rfloor} \binom{S}{i} = \sum_{i=64}^{367} \binom{432}{i}$$

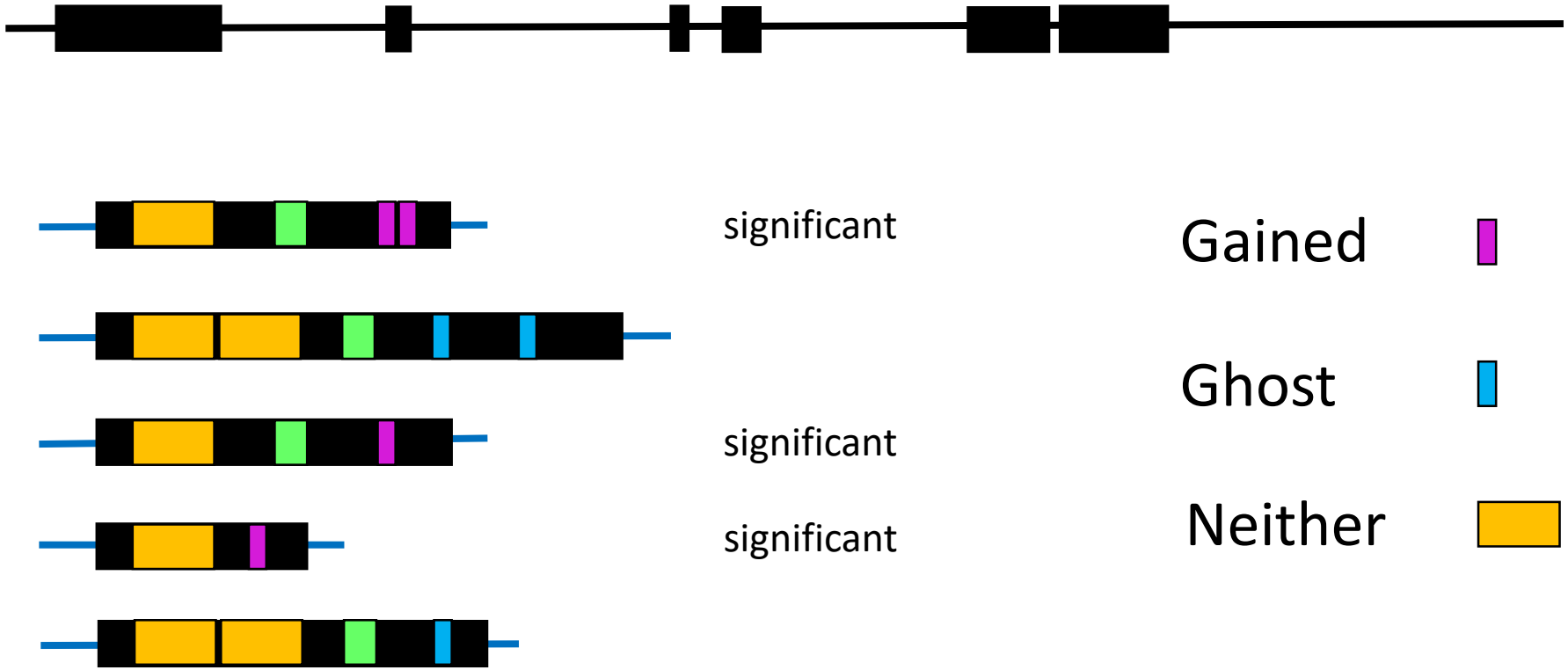
s is number of samples; l is low threshold; h is high threshold

- We build an estimated null distribution with 1M Monte Carlo simulations by randomly permuting patients.

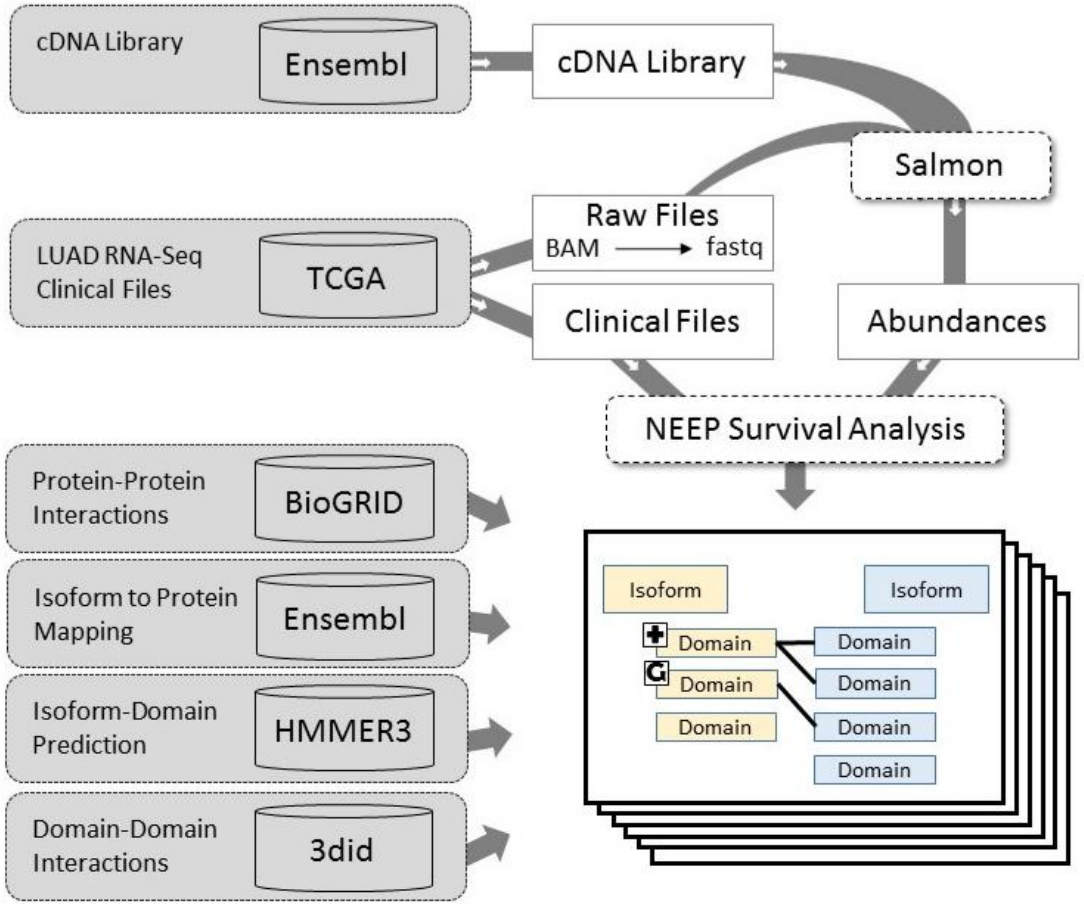
$$\text{true p} = \frac{\#\text{Null} < \text{p}}{\text{Total Null}}$$

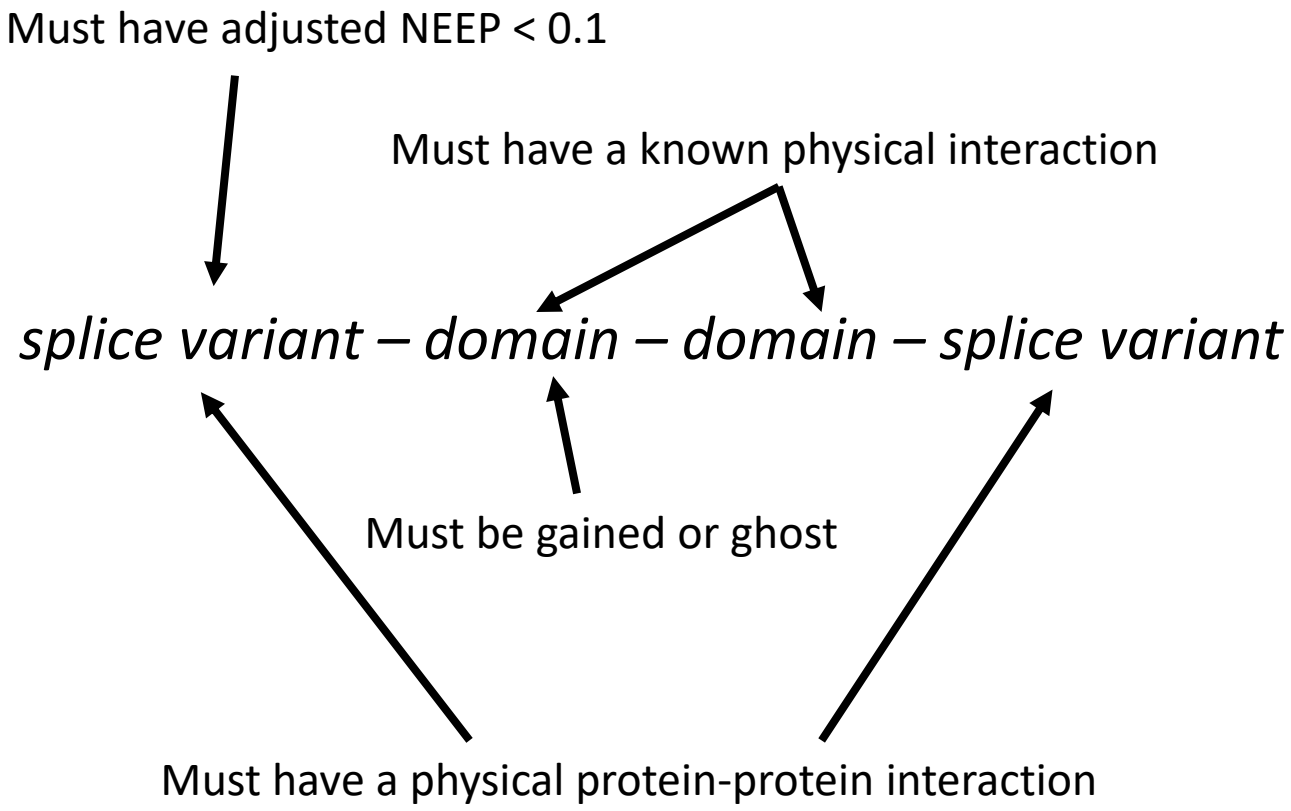


“Gained” and “ghost” domains



Construction of Multi-granular graphs

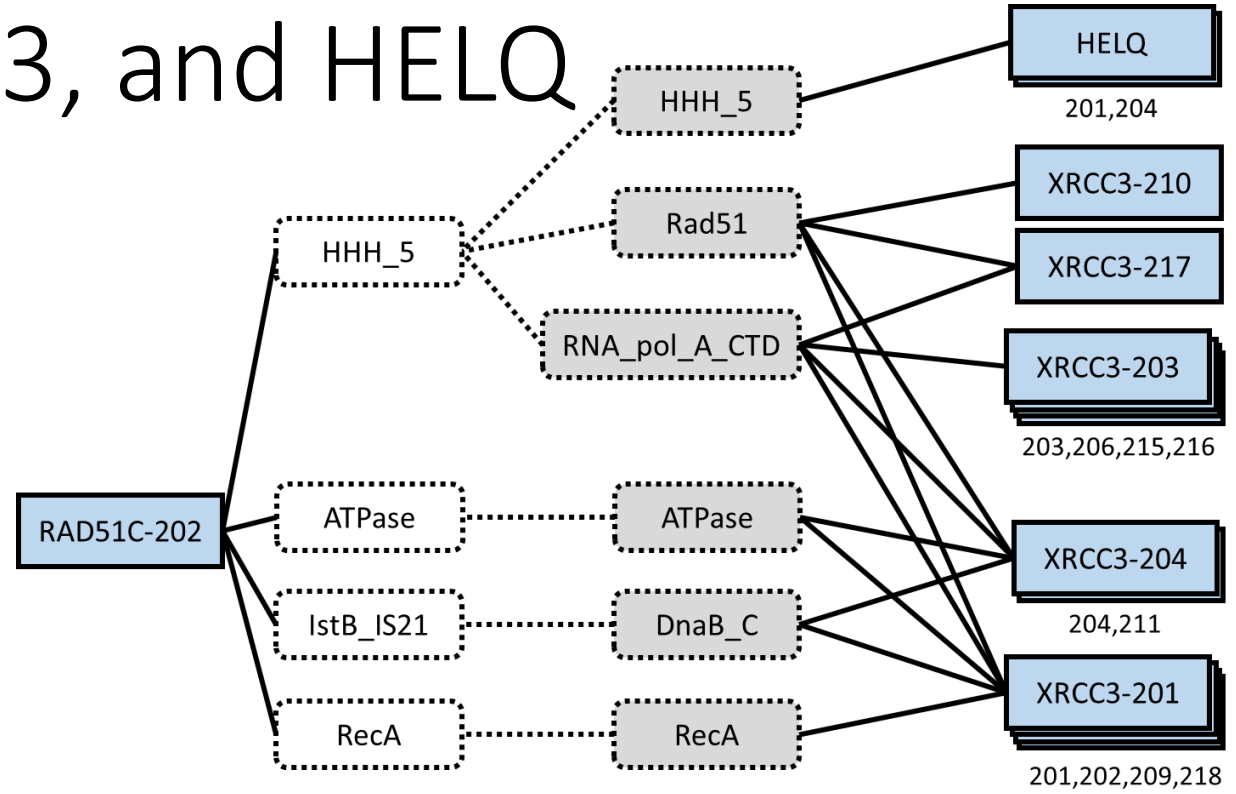




Case studies of MGGs

RAD51C-202 loses ability to bind to the Holliday junction, XRCC3, and HELQ

- RAD51C-202 is missing 2 exons.
  - Exon 1 binds to the Holliday Junction.
  - Exon 2 binds to XRCC3.
- RAD51C and XRCC3 form the CX3 complex which regulates genetic recombination and DNA repair.*
- RAD51C and HELQ form the BCDX2 complex.**
- Disruption of BCDX2 increases susceptibility to DNA-interstrand crosslinks.***



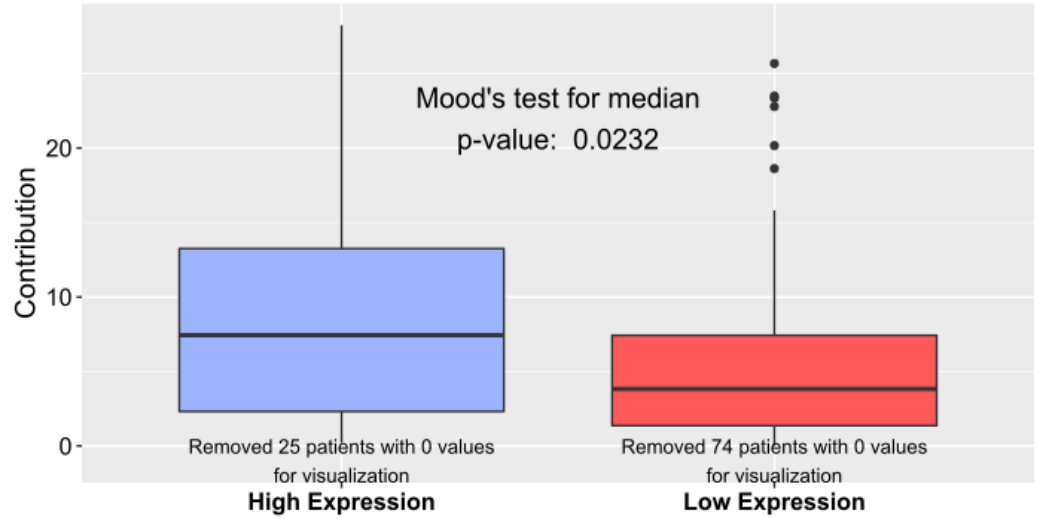
*Liu Y, Tarsounas M, O'Regan P, West SC. Role of RAD51C and XRCC3 in genetic recombination and DNA repair. Journal of Biological Chemistry. 2007;282(3):1973–1979

**Pelttari LM, Kinnunen L, Kiiski JI, Khan S, Blomqvist C, Aittom¨aki K, et al. Screening of HELQ in breast and ovarian cancer families. Familial cancer. 2016;15(1):19–23

***ichi Takata K, Reh S, Tomida J, Person MD, Wood RD. Human DNA helicase HELQ participates in DNA interstrand crosslink tolerance with ATR and RAD51 paralogs. Nature communications. 2013;4:2338

RAD51C-202 expression is associated with a mutation pattern associated with homologous recombination.

- We identified significant COSMIC mutation patterns.
  - Signature 3 is attributed to the “failure of DNA double-strand break-repair by homologous recombination”.*
  - Signature 3 strength was significantly related to RAD51C-202 expression.
- To rule out smoking as a cause, we compared signature 3 strength with 3 smoking variables. None were significant.



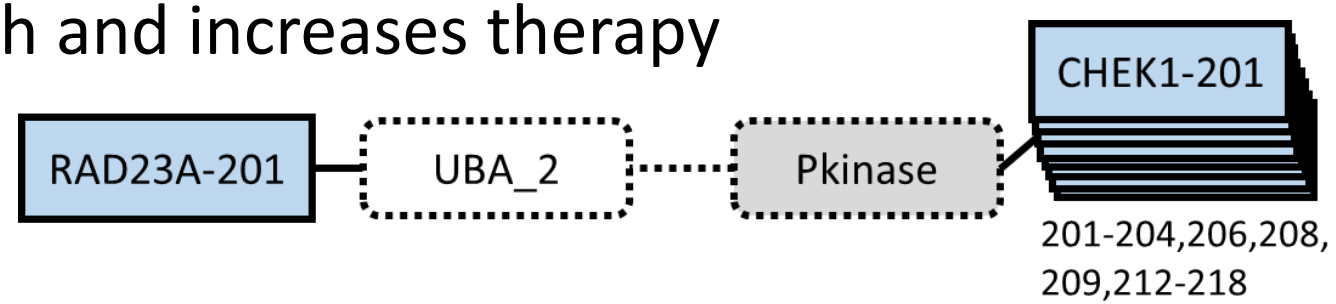
Smoking Variable	Data Type	Statistical Test	p-value
Cigarettes Per Day	Metric	Welch Two Sample T-test	0.086
Years Smoked	Integer	Wilcoxon Rank Sum Test	0.098
Is Smoker	Binary	2-Sample Test for Equal Proportions	0.383

	High – Low (Mean Contribution)	Signature Description from COSMIC
Signature 2	11.032	Activity of AID/APOBEC deaminases
Signature 3	3.864	DNA double-strand break-repair by homologous recombination
Signature 11	0.809	Similar mutation pattern to alkylating agents
Signature 12	2.323	Unknown
Signature 16	3.194	Unknown
Signature 17	-0.016	Unknown
Signature 23	0.533	Unknown

* https://cancer.sanger.ac.uk/cosmic/signatures

RAD23A-201 loses interaction with CHECK1 proteins

- RAD23A-201 is missing the UBA_2 domain.
  - This domain inhibits ubiquitinization of RAD23A
- RAD23A works with the proteasome to target proteins for degredation.*
- Inhibition of the PPI between RAD23A and CHEK1 may extend the lifespan of CHEK1.
- CHEK1 promotes tumor growth and increases therapy resistance.**

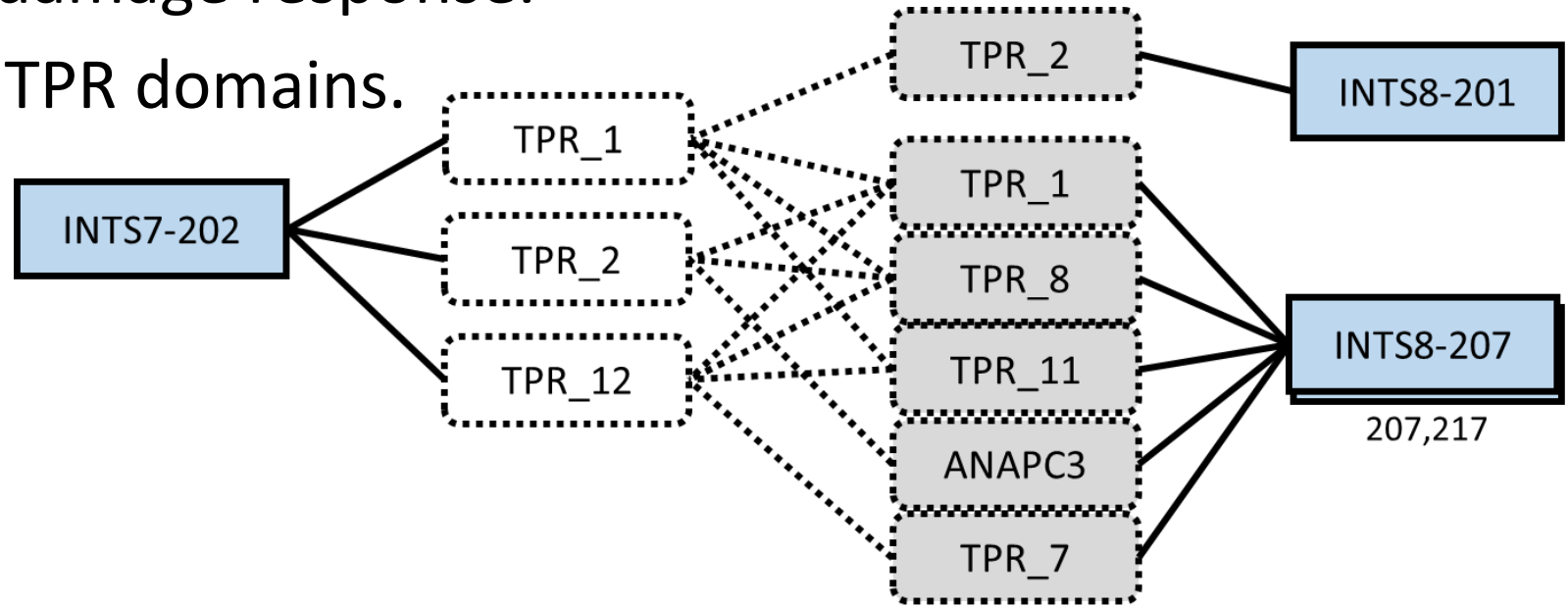


*Wade SL, Auble DT. The Rad23 ubiquitin receptor, the proteasome and functional specificity in transcriptional control. Transcription. 2010;1(1):22–26

**Zhang Y, Hunter T. Roles of Chk1 in cell biology and cancer therapy. International journal of cancer. 2014;134(5):1013–1023

INTS7-202 has altered structure and number of TPR domains

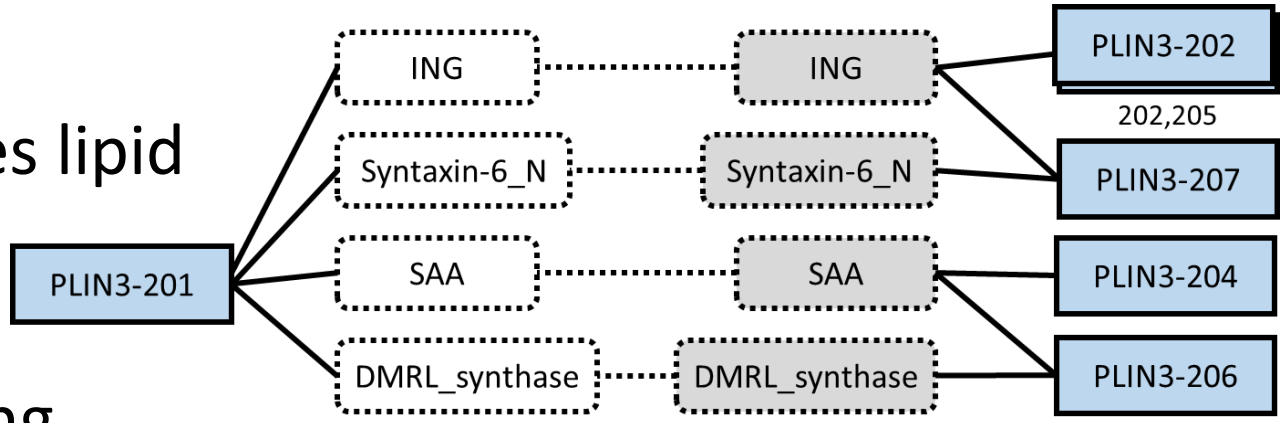
- INTS proteins are members of the Integrator complex (INT).
- INT is involved in mediation of RNA polymerase II and DNA damage response.*
- INTS7-202 still contains TPR domains.



*Federico A, Rienzo M, Abbondanza C, Costa V, Ciccodicola A, Casamassimi A. Pan-cancer mutational and transcriptional analysis of the integrator complex. International journal of molecular sciences. 2017;18(5):936.

PLIN3-201 cannot form homo-dimers

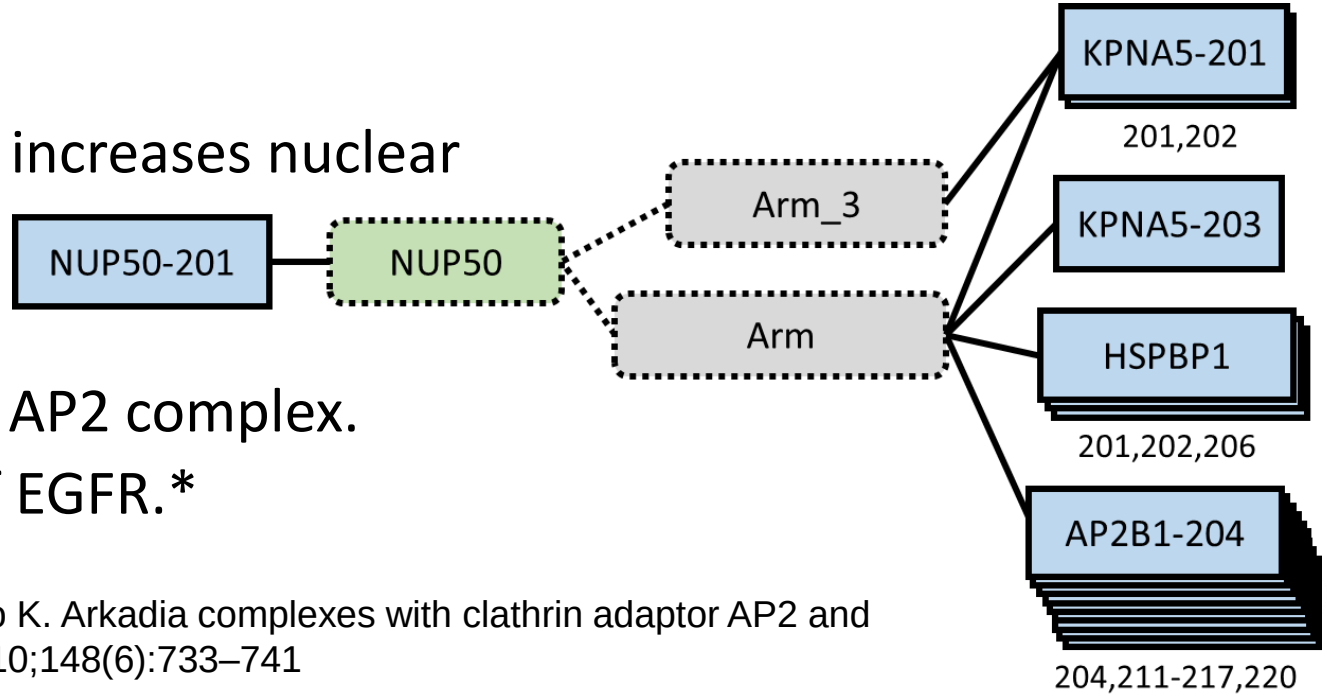
- PLIN3-201 has not been associated with cancer.
- Down-regulation of PLIN3 reduces lipid droplets*.
- Lipid droplet formation** and oscillation*** are observed in lung cancer.
- Active versions of PLIN3 form trimers *in vitro* during migration of PLIN3 to the lipid droplet.



*Nose F, Yamaguchi T, Kato R, Aiuchi T, Obama T, Hara S, et al. Crucial role of perilipin-3 (TIP47) in formation of lipid droplets and PGE2 production in HL-60-derived neutrophils. PLoS One. 2013;8(8):e71542.
**Bozza PT, Viola JP. Lipid droplets in inflammation and cancer. Prostaglandins, Leukotrienes and Essential Fatty Acids (PLEFA). 2010;82(4-6):243–250
***Chowdhury R, Amin MA, Bhattacharyya K. Intermittent Fluorescence Oscillations in Lipid Droplets in a Live Normal and Lung Cancer Cell: Time-Resolved Confocal Microscopy. The Journal of Physical Chemistry B. 2015;119(34):10868–10875

NUP50-201 gains interactions with KPNA5, HSPBP1, and AP2B1 proteins

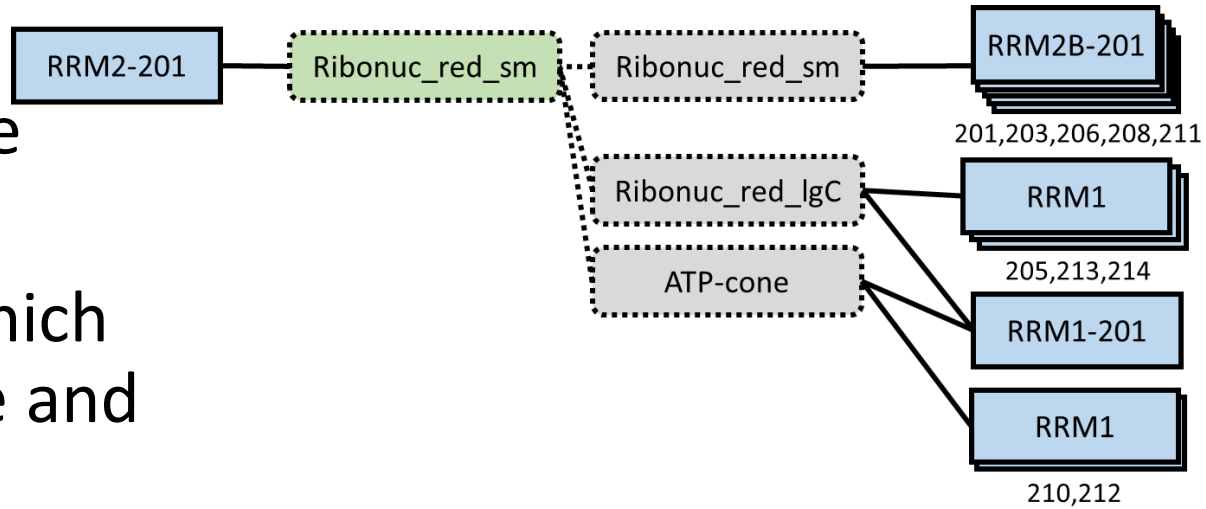
- All other isoforms of NUP50 have disrupted NUP50 domains.
- NUP50 and NUP153 form the nuclear import complex.
- Possible mechanism 1:
  - KNPA5 (importin subunit alpha-6) increases nuclear transport.
- Possible mechanism 2:
  - NUP50-201 encourages uptake of AP2 complex.
  - AP2 helps regulate endocytosis of EGFR.*



*Mizutani A, Saitoh M, Imamura T, Miyazawa K, Miyazono K. Arkadia complexes with clathrin adaptor AP2 and regulates EGF signalling. The journal of biochemistry. 2010;148(6):733–741

RRM2-201 contains domain required for ribonucleotide reductase

- RRM1 and RRM2 are previously implicated in non-small cell lung cancer survival* and therapy response.**
- RRM1 and RRM2 form a tetramer structure called the ribonucleotide reductase (RNR).
- The RNR regulates dNTP levels, which alter DNA replication error chance and increase genomic instability.***



*Boukovinas I, Papadaki C, Mendez P, Taron M, Mavroudis D, Koutsopoulos A, et al. Tumor BRCA1, RRM1 and RRM2 mRNA expression levels and clinical response to first-line gemcitabine plus docetaxel in non-small-cell lung cancer patients. PloS one. 2008;3(11):e3695

**Wang L, Meng L, wen Wang X, yuan Ma G, han Chen J. Expression of RRM1 and RRM2 as a novel prognostic marker in advanced non-small cell lung cancer receiving chemotherapy. Tumor Biology. 2014;35(3):1899–1906

***Aye Y, Li M, Long M, Weiss RS. Ribonucleotide reductase and cancer: biological mechanisms and targeted therapies. Oncogene. 2015;34(16):2011

Conclusions

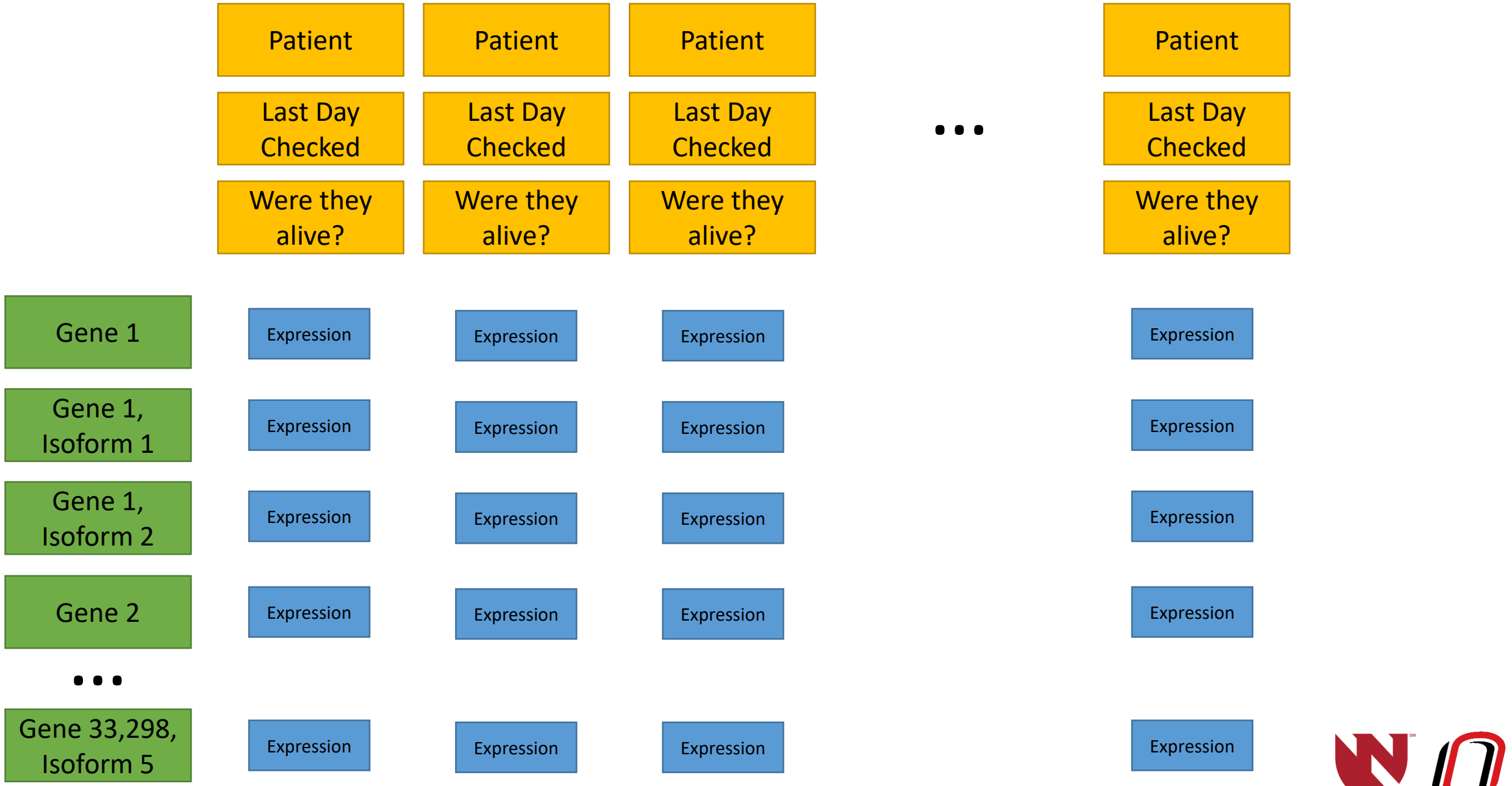
- We introduce NEEP, a method to calculate a valid *p*-value distribution for logrank tests with thresholds that maximize survival curve separation.
  - Code available at https://github.com/thecodingdoc/neep
- We construct multi-granularity graphs using information about gained and ghost domains.
  - Code available at https://github.com/scwest/SINBAD
- Biological hypotheses were generated from the MGGs and manual explored as case studies.
- These hypotheses uncovered known and novel molecular mechanisms that potentially impact lung cancer patient survival.

Acknowledgements

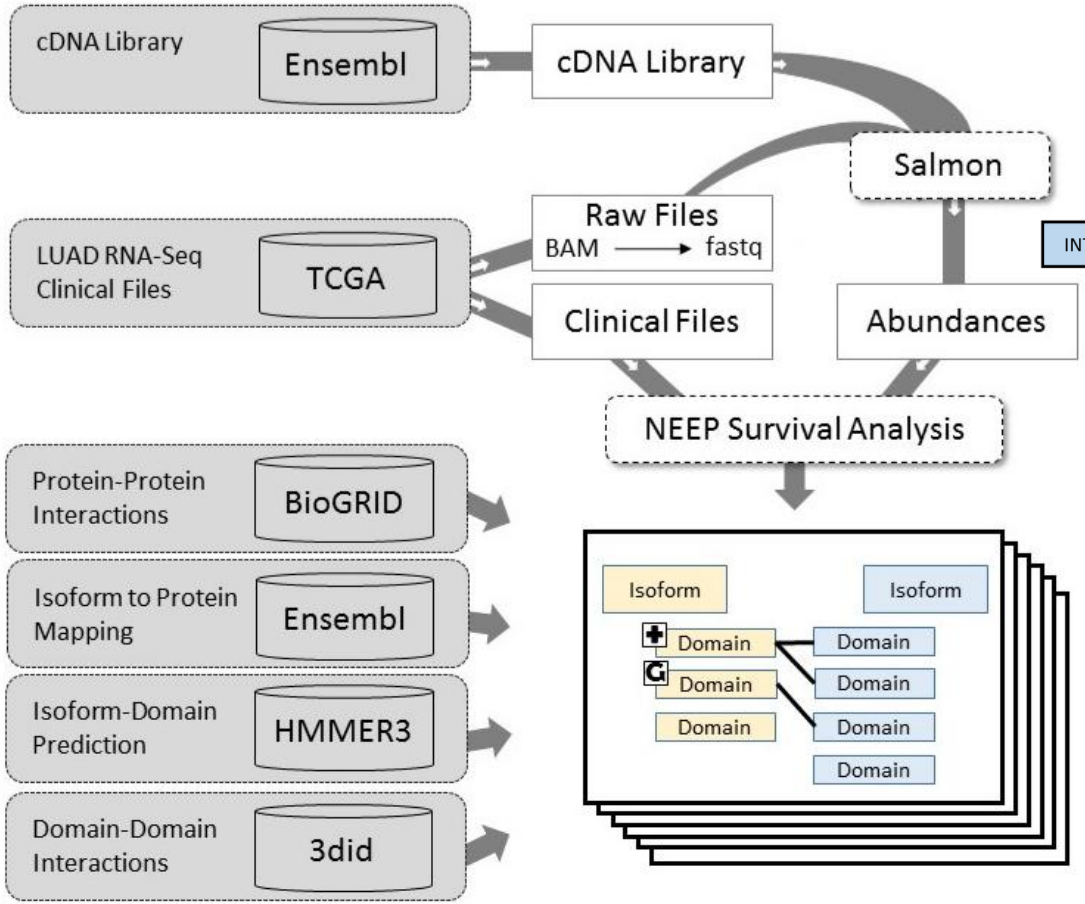
The University of Nebraska Collaboration Initiative System Science Grant

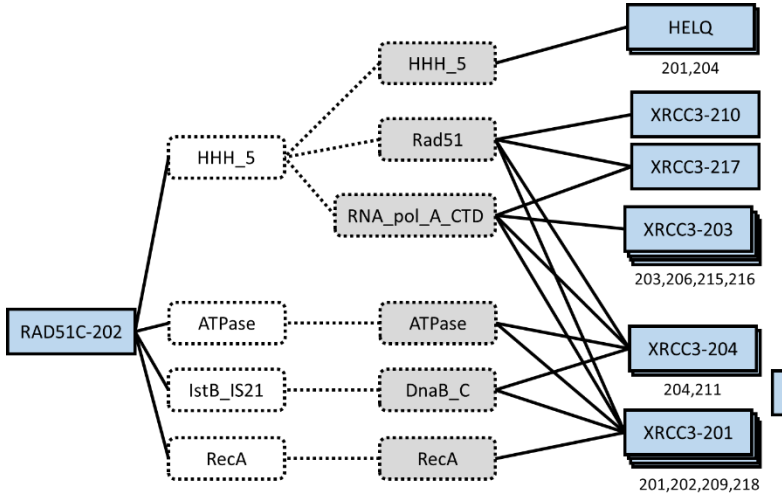
UNMC Collaborators

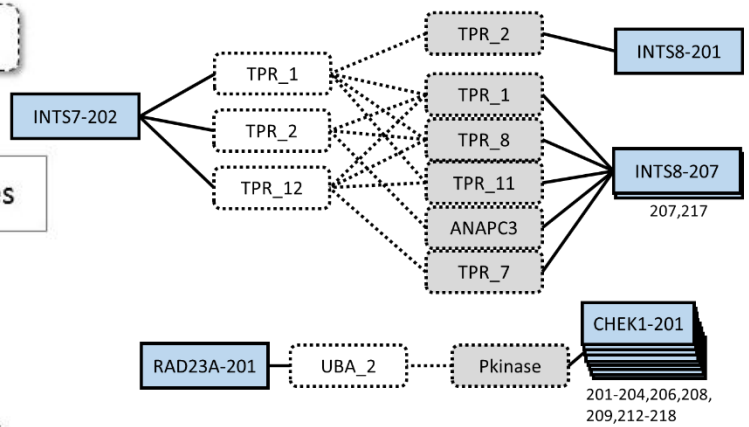
- Surinder Batra
- Sushil Kumar
- Chris Thompson
- Batra lab members

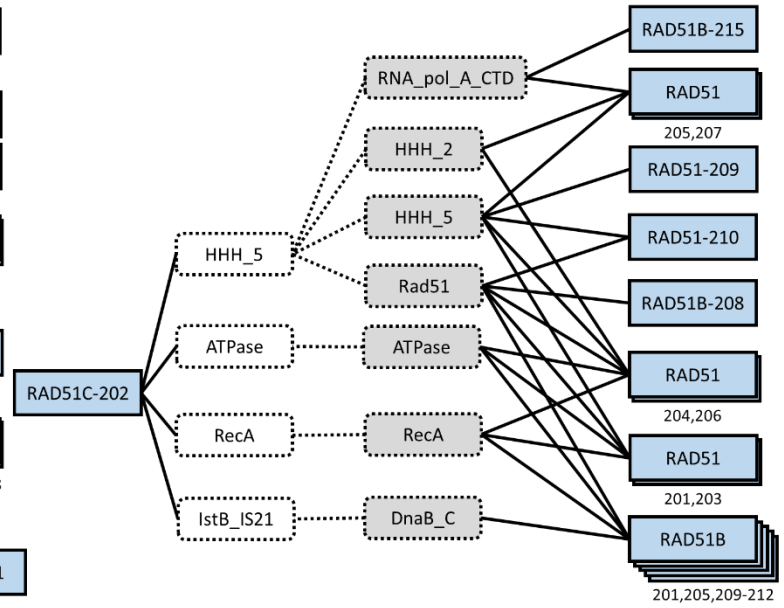

Patient
Patient
Patient
Patient
Last Day Checked
Last Day Checked
Last Day Checked
Last Day Checked
Were they alive?
Were they alive?
Were they alive?
Were they alive?
Gene 1
Gene 1, Isoform 1
Gene 1, Isoform 2
Gene 2
Gene 33,298, Isoform 5
Expression

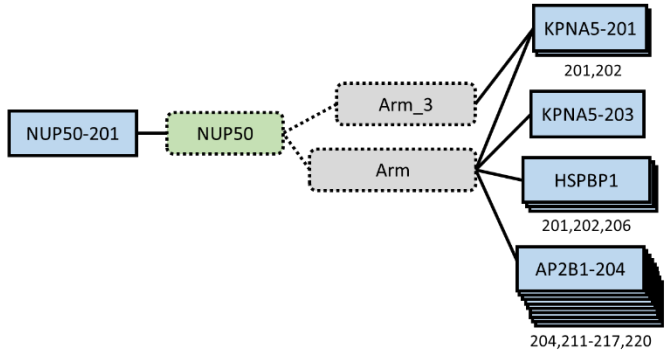
Questions?











NEEP:
https://github.com/thecodingdoc/neep

Multi-granularity graphs (SINBAD):
https://github.com/thecodingdoc/neep

Current Directions

- Construction of MGGs for other cancer types.
- Connect immunogenicity of gained or lost sequences which span exon junctions to pancreatic cancer.
- Extend methods presented here to include non-protein models.
  - i.e. RNA-binding, DNA-binding, micro-RNA

Comparison to other survival analysis statistics

- Lausen and Schumacher created a parametric estimator of the true logrank p-value.*
  - Assumes normal distribution of expression for each isoform.
  - Small estimation differences cause large changes for p-value adjustment.
- Cox-PH
  - Strongly assumes the proportional hazard assumption.
  - Less intuitive hypothesis for our problem:
    - Cox-PH Null: Variability of survival cannot be explained by the variability of the splice variant.
    - Logrank Null: Low expression patients have the same hazard rate as high expression patients.

*Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics. 1992; p. 73–85

Comparison to other survival analysis methods

SURVIV

- Cox-PH of exon inclusion levels.
- Downstream analysis of protein products require association with individual exon inclusion events.

Robust Selection Algorithm

- Produces a statistically invalid result.
- Made for miRNA.
- Medium-throughput.

NEEP (our method)

- High-throughput (only need to construct 1 Null distribution).
- Can be run for genes, isoforms, and exon junctions.

The MGGs capture more functional information than genes/isoforms alone.