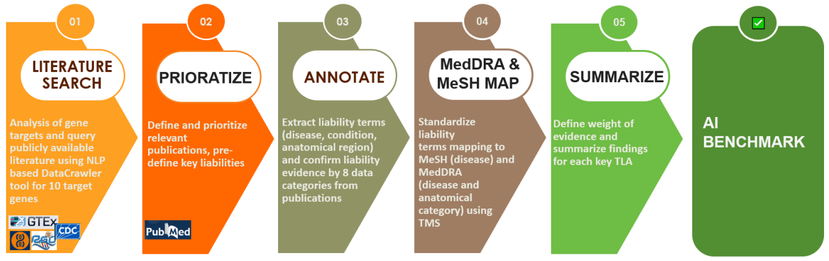
Harnessing Human Omics Data and ML, Combined with Human Insights, to Revolutionize Gene Target Safety Evaluation

Evan H. Baugh, PhD[1,†]; Alena Fedarovich, PhD[2,†]; Brian Mautz, PhD[1,†]; Bing Zhou[2]; Dzmitry Fedarovich[2]; Brad Farrell[2]; Cindy J. Grondin[2]; Julio Molineros, PhD[1]; Shuwei Li, PhD[1,*]; David Smith, BS[3,*]

1: Data Sciences & Digital Health, Johnson & Johnson Innovative Medicine, 2: Rancho Biosciences, 3: Preclinical Science & Translational Safety, Johnson & Johnson Innovative Medicine

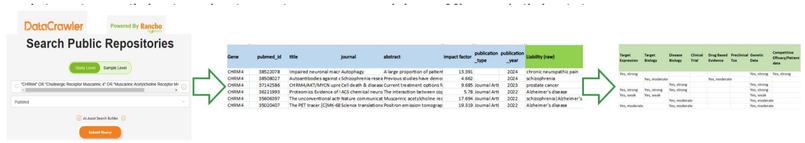
Introduction

- Evaluation of gene target safety is essential in the development of new treatments, with safety concerns accounting for 24-30% of drug failures
- Human multi-Omic data aids the assessment of likely Adverse Event (AE) risks, though can be difficult to interpret, enhanced with Machine Learning (ML) and Artificial Intelligence (AI)
- Integrating these findings and assessing literature requires multidisciplinary expertise – and is complicated for genes lacking comprehensive data
- We conducted a collaborative investigation evaluating the corroboration between a recently developed ML algorithm (TRAP, see Poster 3457) alongside expert-curated insights and AE assessments
- Rancho's internally developed Natural Language Processing-based DataCrawler tool was used to find and annotate publications for 12 safety liabilities, empowering expert curation
- TRAP risk scores were compared to Rancho human expert curation, identifying

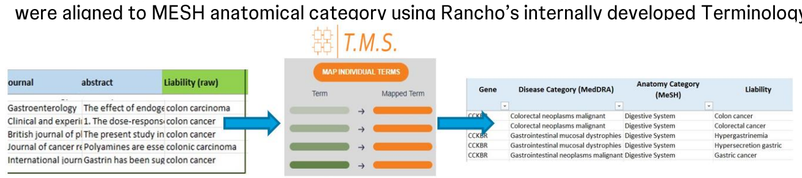


Rancho DataCrawler and Expert Assessments

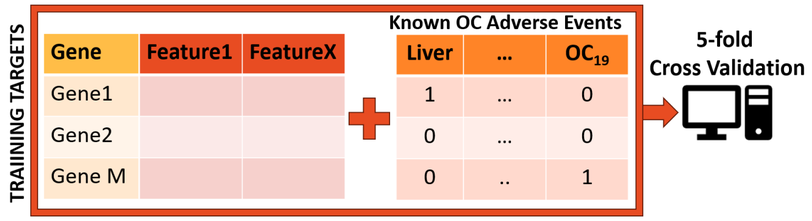
The analysis of gene targets included identifying gene aliases, tissue expression and associated diseases from various data repositories (GeneCards, GTEx Portal, Rat Genome Database (RGD), Centers for Disease Control and Prevention (CDC)). The collected information was cross referenced and used for literature search in PubMed using Rancho's internally developed Natural Language Processing (NLP)-based DataCrawler tool to find and annotate publications reporting any findings relevant to up to 12 key safety liabilities of the selected genes. The tool's output table was further used to define and prioritize the most relevant publications based on available information, and perform manual annotation for target expression, target biology, disease biology, genetic data, drug-based evidence, clinical



Each of three Rancho Data Integrity Experts worked on a set of 3-4 genes. High literature diversity per gene impacted annotation and TLA assessment time with an average per gene manual safety evaluation of 35.6 hours ± 6.7 hours (SD). The final QC and standardization of the results was performed by a designated scientist. To prepare annotated results for further analysis and validation of the ML outcomes extracted from various sources liability terms were cleaned, harmonized, and mapped to existing ontologies. Disease/condition terms were aligned to MedDRA PT and HLGT, while terms of organ class related to a disease were aligned to MESH anatomical category using Rancho's internally developed Terminology



TRAP – Target Risk Assessment Profile



- TRAP is a novel ML method for assessing association of genes to risk concerns for 19 Organ Categories (OC) with MedDRA codes and output confidence scores (0-1)
- Trained on OFF-X AE entries using 254 curated Human Omics features – including: gene constraint metrics, variant counts, essentiality scores, tolerance from UK BioBank/gnomAD, Phenome-wide association studies, gene expression(TargetMiner/GTEx), gene-gene interaction networks
- Average report generation time was 4.75min (range 1.76-9.72min)

Comparison Between TRAP and Rancho Associations

- TRAP OC assessments were harmonized to Rancho expert-curated findings, yielding 62 unique gene-OC pairs assessed by Rancho across 10 genes

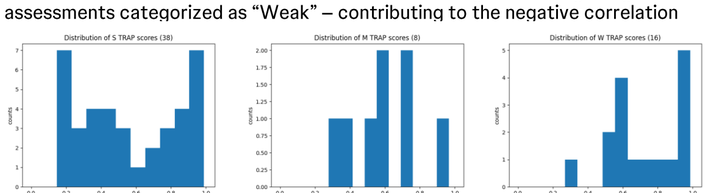
gene	Rancho OC
AR	6
CCKBR	3
CHRM4	8
EDNRA	7
HTR5A	3
KCNA1	6
MTNR1A	11
P2RX1	9
CCKAR	7
TSPO	2

OC	Rancho genes
Blood and lymphatic system disorders	2
Cardiac disorders	**5**
Ear and labyrinth disorders	1
Endocrine disorders	**4**
Eye disorders	1
Gastrointestinal disorders	**4**
Hepatobiliary disorders	2
Immune system disorders	2
Metabolism and nutrition disorders	3
Musculoskeletal and connective tissue disorders	3
Neoplasms, benign, malignant, and unspecified	**10**
Nervous system disorders	**9**
Psychiatric disorders	2
Renal and urinary disorders	**4**
Reproductive system and breast disorders	3
Respiratory, thoracic, and mediastinal disorders	**4**
Skin and subcutaneous tissue disorders	1
Vascular disorders	2

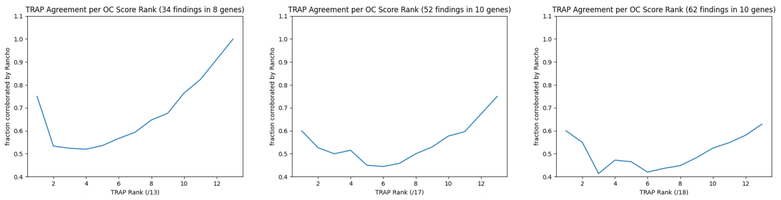
- Rancho expert assessed gene OC associations were ranked and quantified based on the total number of publications found supporting that association including a breakdown of evidence from target expression, target biology, disease biology, genetic data, etc.
- Comparisons were made between Rancho expert qualitative ranking, quantitative ranking by associated literature counts, and various transformations of TRAP scores

correlation rules	correlation	p-value
Rancho lit. count vs TRAP score	-0.27	0.037
Rancho lit. rank vs TRAP rank-scaled	-0.31	0.014
restricted OC (29 pairs), Rancho lit. count vs TRAP rank bins	-0.63	2.30E-04
Rancho lit. rank excluding 'Weak' vs TRAP rank bins	0.22	0.14

- Most comparisons between Rancho literature counts vs TRAP scores yielded ***negative*** correlations, indicating TRAP quantitative ranking does not indicate higher inclusion in the literature
- Rancho findings were primarily ranked as "Strong" (38/62) based on the number of publications discovered, however TRAP scores were highest among the gene-OC assessments categorized as "Weak" – contributing to the negative correlation



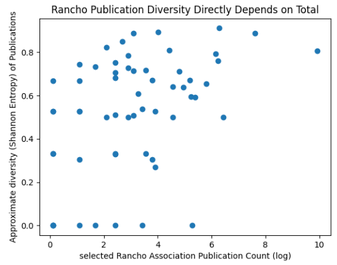
TRAP Corroborates Top-ranked Rancho findings

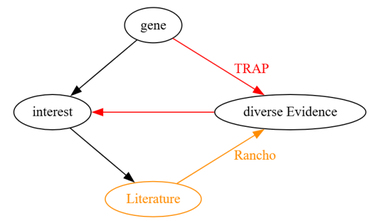


- TRAP OC scores were ranked and compared against Rancho expert findings
- Liability resolution was limited between:
  - 'Nervous system' vs 'Psychiatric'
  - 'Metabolism' vs 'Endocrine'
  - 'Neoplasms'
- Excluding these cases, the top ranked TRAP association was found by Rancho for 75% of the investigated genes (6/8)
- Corroboration decreased for higher ranks but remained above 50%, indicating a high overlap/TRAP sensitivity
- Inclusion of these difficult liability categories decreases ranking agreement

Literature vs Omics?

- Literature annotations suggest the diversity of available Evidence depends on the overall interest/popularity of gene targets
- Approximated diversity as the Shannon Entropy among literature Evidence categories found by Rancho correlated with the total number of publications found, Spearman rank correlation = 0.40, p-value=1.8e-3





- This correlation confounds comparisons between TRAP scores vs Literature counts – TRAP scores *require* diverse Human Omics Evidence, while targets with more Literature coverage consequently obtain more diverse Evidence within the Literature

Conclusions

- Literature and Omics –driven approaches provide distinct but complementary insights
- These case comparisons demonstrated high agreement between these automated and expert-curated assessments, when ambiguous liability distinctions were excluded
- Anti-correlation between TRAP risk scores and Rancho literature-based rankings demonstrate the limitations of both methods – while also indicating non-random trends
- TRAP is significantly faster than expert curation, producing reports within minutes but without links to associated literature
- Automated methods of risk assessment provide an informative 'first pass' and distinct insights, indicating utility as complementary analyses to expert literature evaluation

Alena Fedarovich, PhD
Harmonization and Knowledge Mining
Data Integrity Department

Email: alena.Fedarovich@ranchobiosciences.com

Evan H. Baugh, PhD
Population Analytics and Insights
Data Science Analytics and Insights AI/ML
Data Science and Digital Health
Email: ebaugh1@its.jnj.com