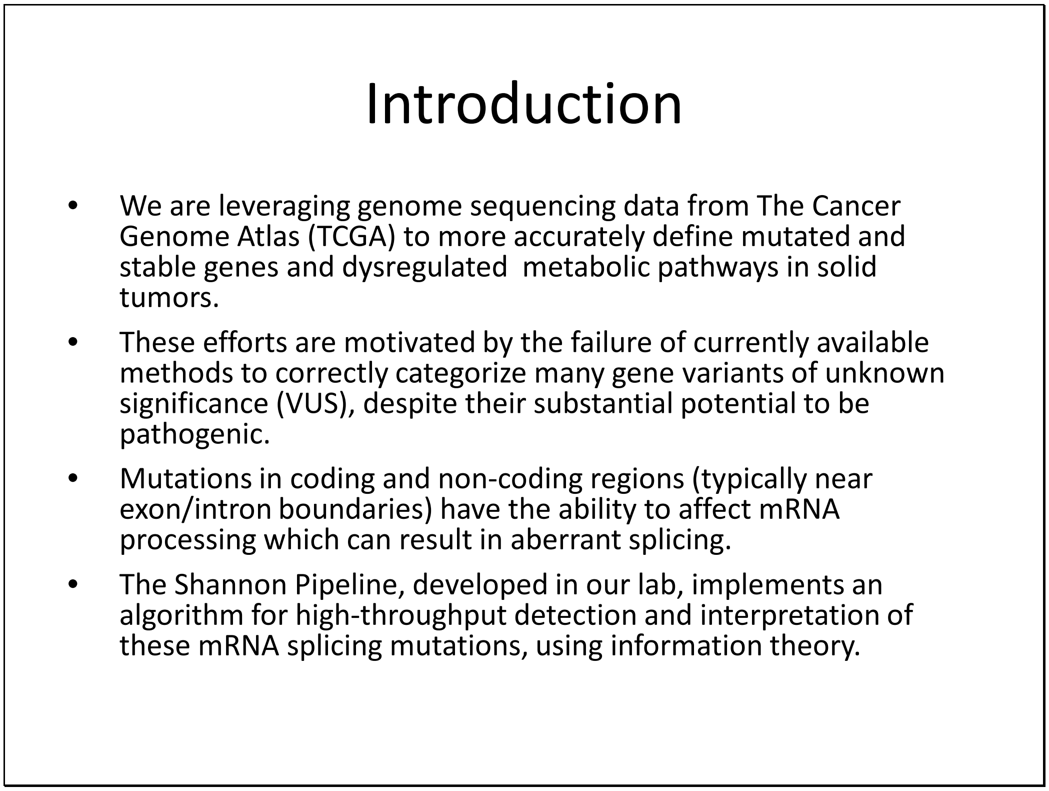
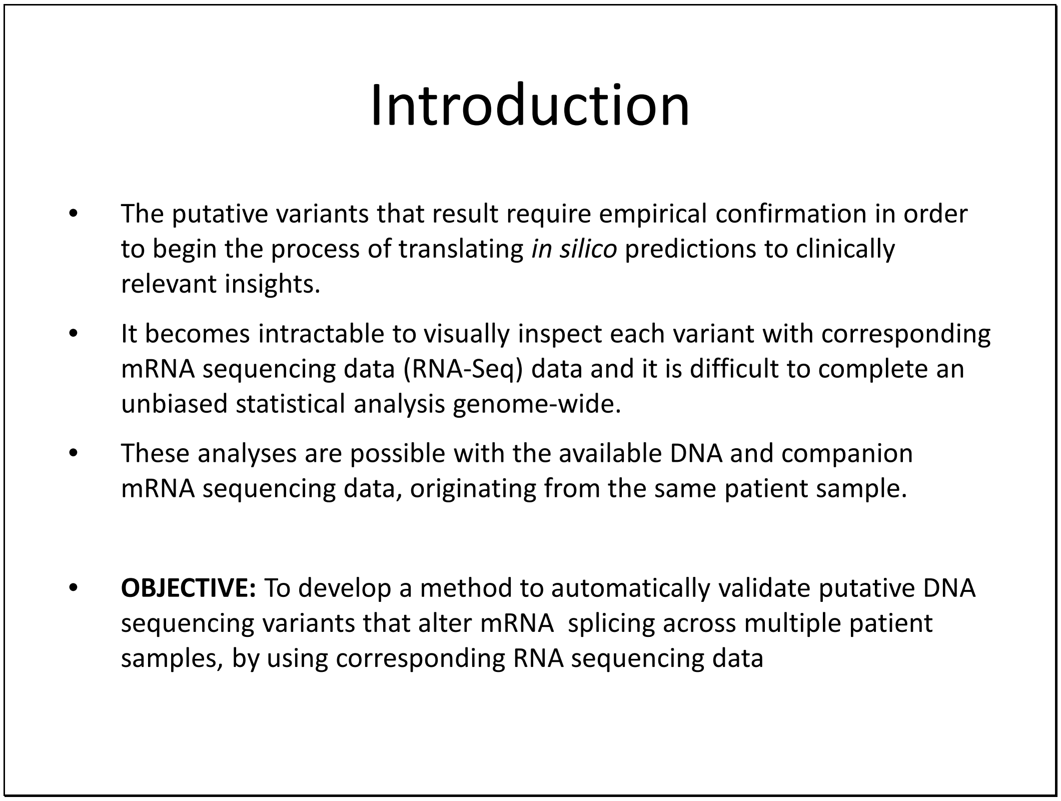
**Validation of predicted mRNA splicing mutations using high-throughput transcriptome data**

Coby Viner, Stephanie N. Dorman, Ben C. Shirley, Peter K. Rogan



# Introduction

- We are leveraging genome sequencing data from The Cancer Genome Atlas (TCGA) to more accurately define mutated and stable genes and dysregulated metabolic pathways in solid tumors.
- These efforts are motivated by the failure of currently available methods to correctly categorize many gene variants of unknown significance (VUS), despite their substantial potential to be pathogenic.
- Mutations in coding and non-coding regions (typically near exon/intron boundaries) have the ability to affect mRNA processing which can result in aberrant splicing.
- The Shannon Pipeline, developed in our lab, implements an algorithm for high-throughput detection and interpretation of these mRNA splicing mutations, using information theory.



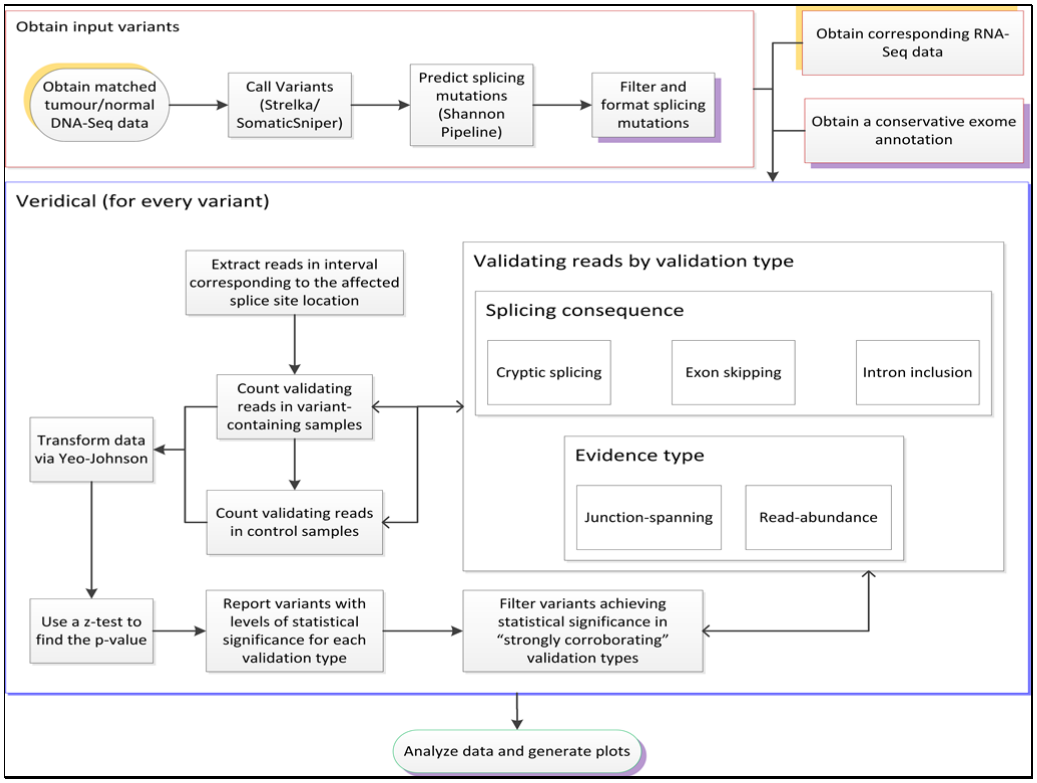
# Introduction

- The putative variants that result require empirical confirmation in order to begin the process of translating *in silico* predictions to clinically relevant insights.
- It becomes intractable to visually inspect each variant with corresponding mRNA sequencing data (RNA-Seq) data and it is difficult to complete an unbiased statistical analysis genome-wide.
- These analyses are possible with the available DNA and companion mRNA sequencing data, originating from the same patient sample.

- **OBJECTIVE:** To develop a method to automatically validate putative DNA sequencing variants that alter mRNA splicing across multiple patient samples, by using corresponding RNA sequencing data

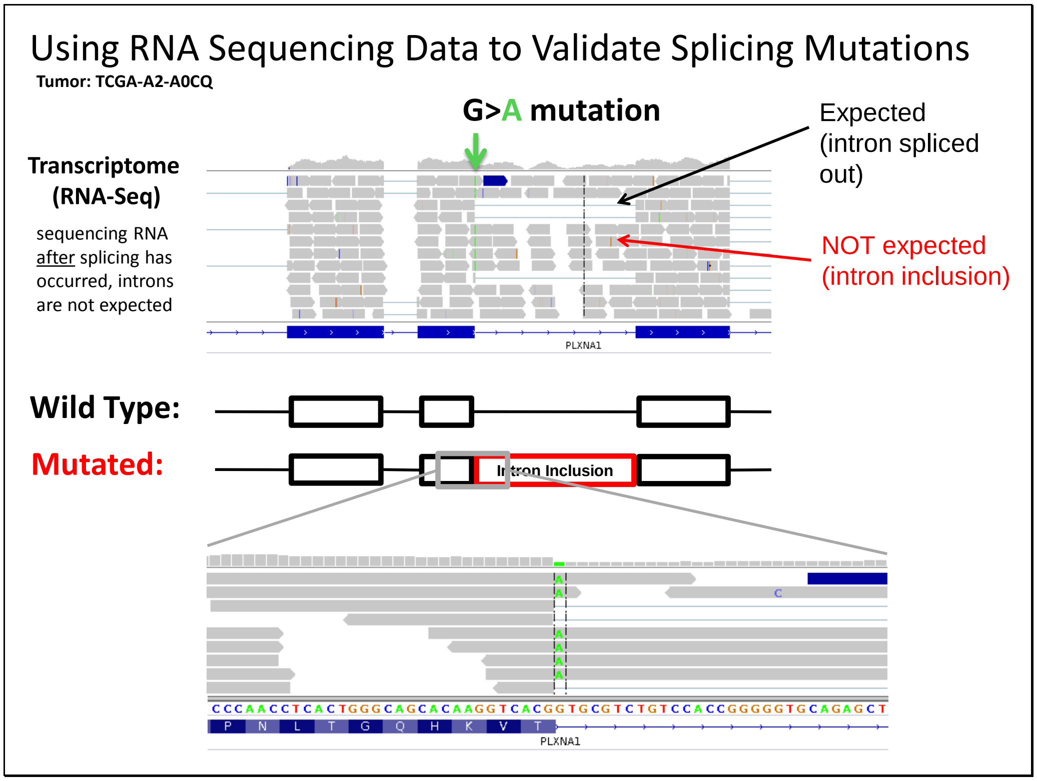
**Validation of predicted mRNA splicing mutations using high-throughput transcriptome data**

Coby Viner, Stephanie N. Dorman, Ben C. Shirley, Peter K. Rogan



Obtain input variants
Obtain matched tumour/normal DNA-Seq data
Call Variants (Strelka/ SomaticSniper)
Predict splicing mutations (Shannon Pipeline)
Filter and format splicing mutations
Obtain corresponding RNA-Seq data
Obtain a conservative exome annotation
Veridical (for every variant)
Extract reads in interval corresponding to the affected splice site location
Count validating reads in variant-containing samples
Count validating reads in control samples
Transform data via Yeo-Johnson
Use a z-test to find the p-value
Report variants with levels of statistical significance for each validation type
Filter variants achieving statistical significance in "strongly corroborating" validation types
Validating reads by validation type
Splicing consequence
Cryptic splicing
Exon skipping
Intron inclusion
Evidence type
Junction-spanning
Read-abundance
Analyze data and generate plots

Workflow diagram depicting the analyses undertaken by Veridical to analyze splicing mutations in TCGA breast carcinoma data. Items outlined in red comprise the input data for Veridical. TCGA input data is indicated by an orange shadow. The steps involving perl programs have a purple shadow.



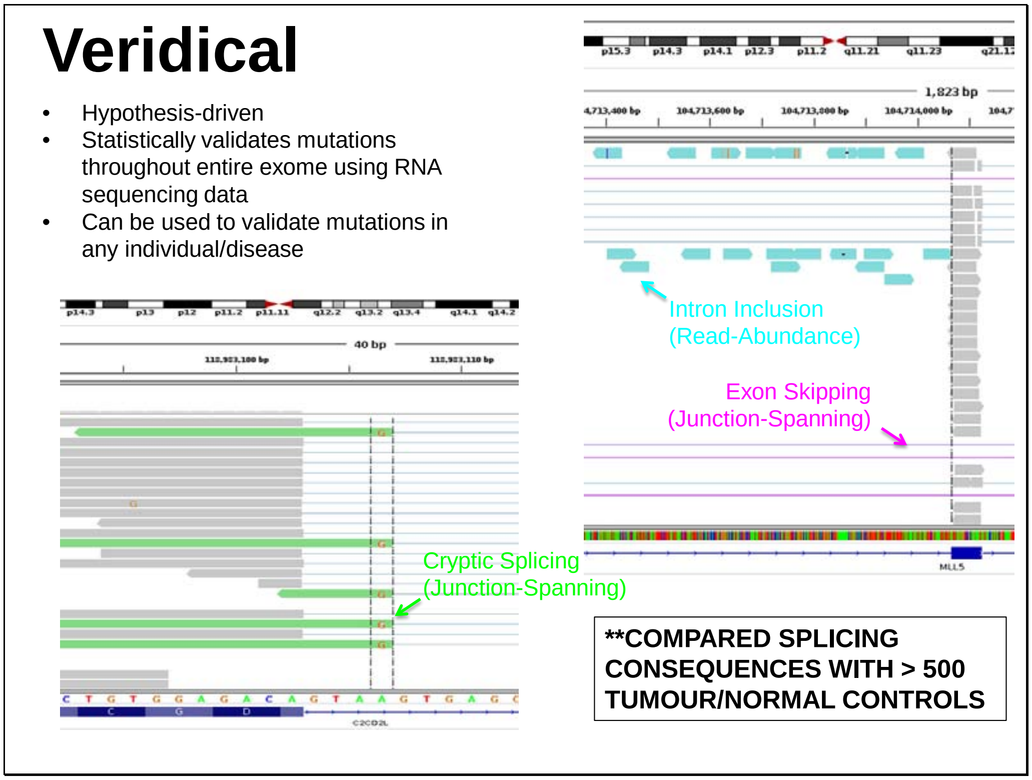
## Using RNA Sequencing Data to Validate Splicing Mutations

**Tumor: TCGA-A2-A0CQ**

**Transcriptome (RNA-Seq)**

sequencing RNA after splicing has occurred, introns are not expected

G>A mutation
Expected (intron spliced out)
NOT expected (intron inclusion)
PLXNA1
Wild Type:
Mutated:
Intron Inclusion
CCCAACCTCACTGGGCAGCACAAGGTCACGGTGCGTCTGTCCACCGGGGGTGCAGAGCT
P N L T G Q H K V T
PLXNA1



# Veridical

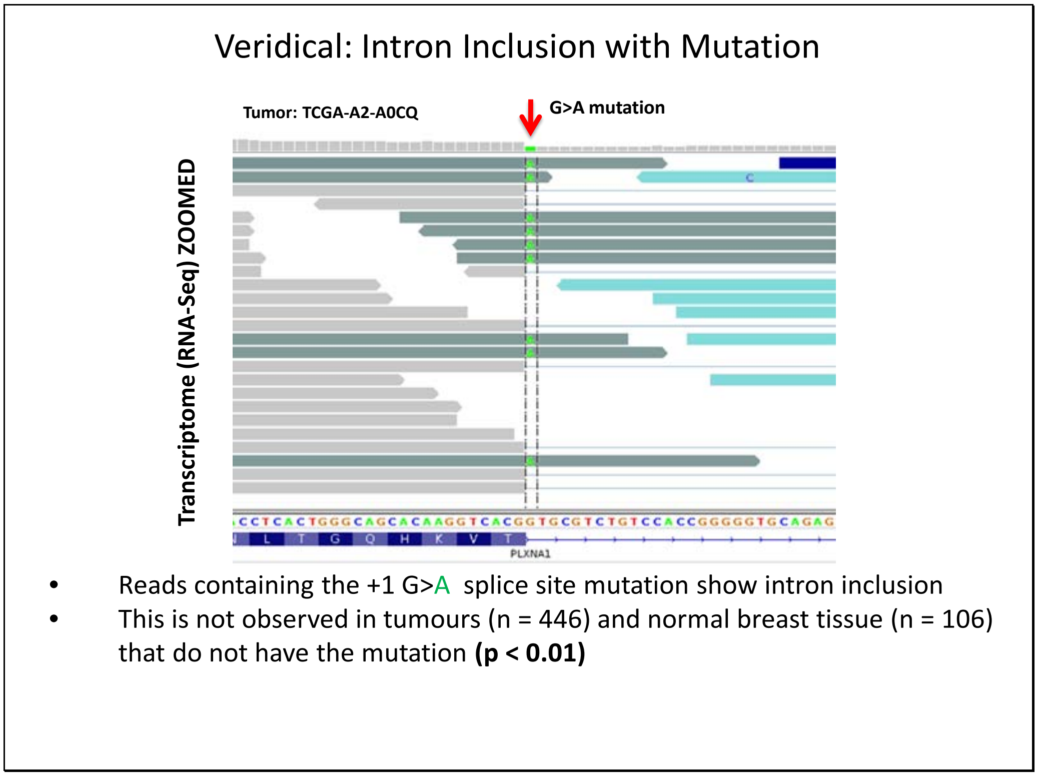
- Hypothesis-driven
- Statistically validates mutations throughout entire exome using RNA sequencing data
- Can be used to validate mutations in any individual/disease

Intron Inclusion (Read-Abundance)

Exon Skipping (Junction-Spanning)

Cryptic Splicing (Junction-Spanning)

****COMPARED SPLICING CONSEQUENCES WITH > 500 TUMOUR/NORMAL CONTROLS**



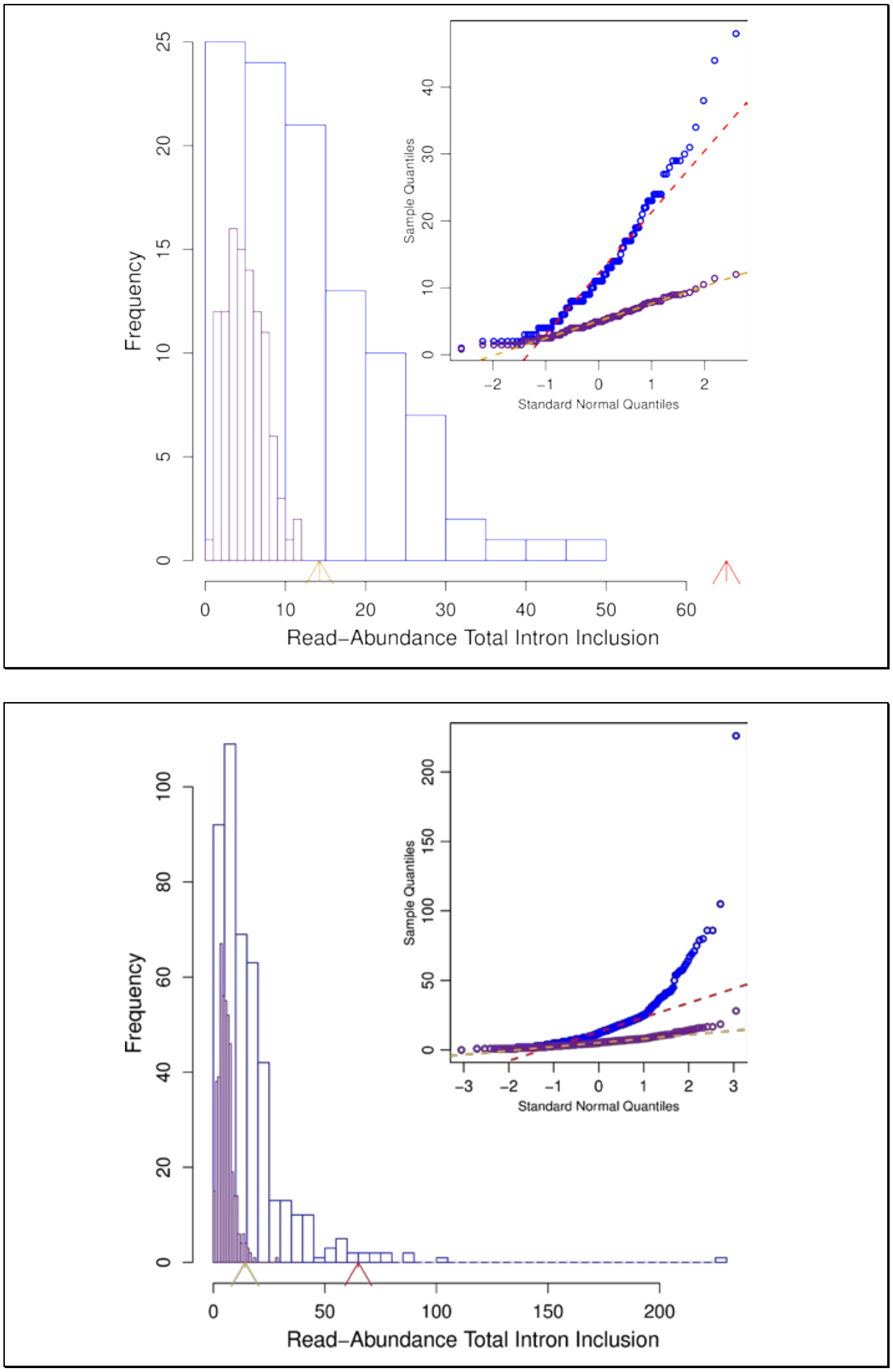
## Veridical: Intron Inclusion with Mutation

**Tumor: TCGA-A2-A0CQ**

**G>A mutation**

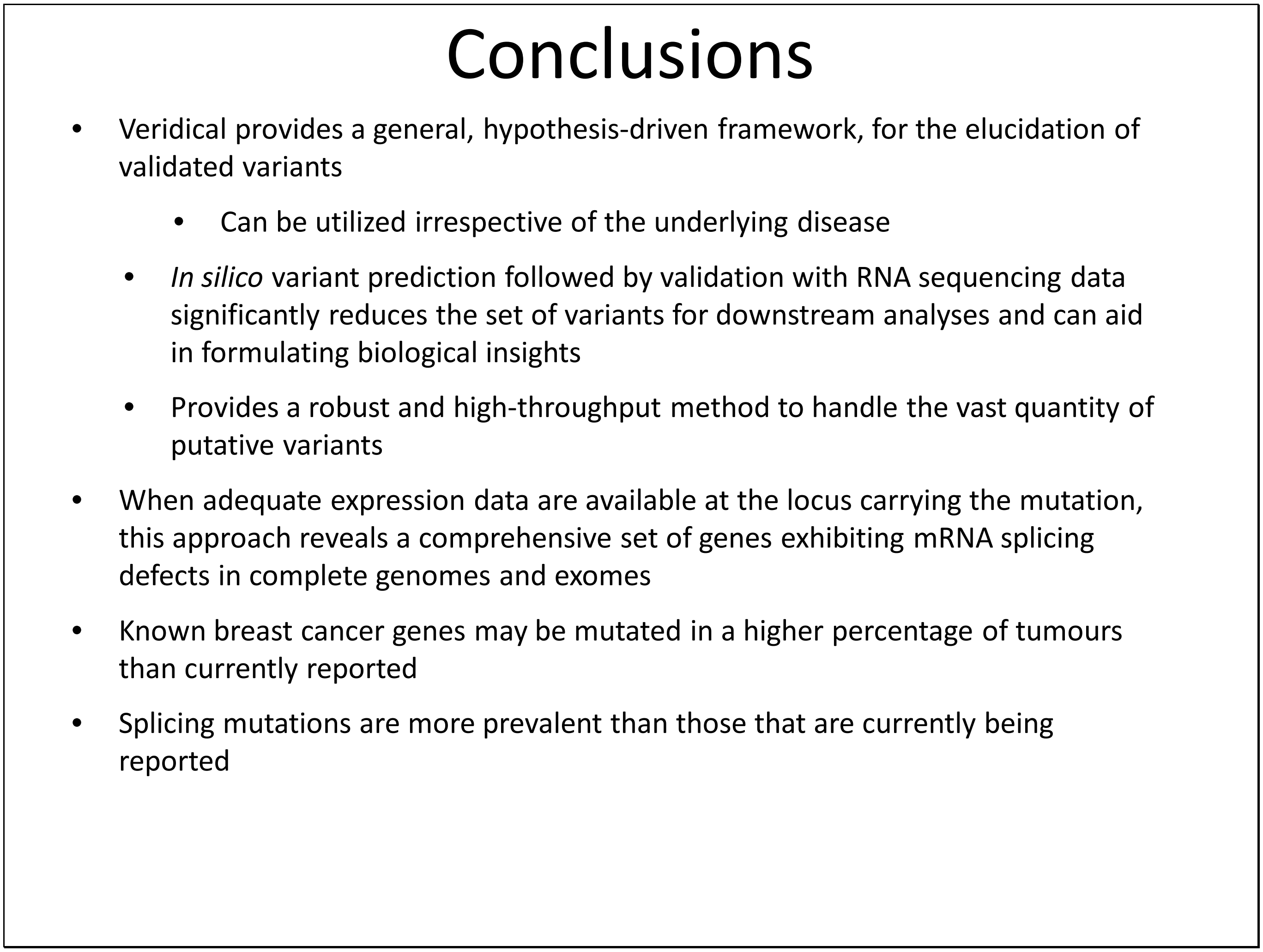
Transcriptome (RNA-Seq) ZOOMED
CCTCACTGGGCAGCACAAGGTCACGGTGCGTCTGTCCACCGGGGGTGCAGAG
N L T G Q H K V T
PLXNA1

- Reads containing the +1 G>A splice site mutation show intron inclusion
- This is not observed in tumours (n = 446) and normal breast tissue (n = 106) that do not have the mutation **(p < 0.01)**



Histogram and embedded Q-Q plots portraying the difference between untransformed and Yeo-Johnson (YJ)[1] transformed data. The plots depict intron inclusion for the inactivating mutation (chr12:83359523 G>A) within TMTC2. The arrowheads denote the number of reads in the variant-containing file, which is, in all cases, more than observed in the control samples $p < 0.01$. Blue and red plot elements correspond to untransformed data, while yellow and purple correspond to YJ transformed elements. Dotted lines in the Q-Q plots are lines passing through the first and third quantiles for a normal reference distribution. The top plot shows the distribution of read-abundance intron inclusion in normal samples, while the bottom shows the same for tumour samples.

[1] Yeo IK, Johnson RA: **A new family of power transformations to improve normality or symmetry.** *Biometrika.* 2000; **87**(4): 954–959.



# Conclusions

- Veridical provides a general, hypothesis-driven framework, for the elucidation of validated variants
    - Can be utilized irrespective of the underlying disease
  - *In silico* variant prediction followed by validation with RNA sequencing data significantly reduces the set of variants for downstream analyses and can aid in formulating biological insights
  - Provides a robust and high-throughput method to handle the vast quantity of putative variants
- When adequate expression data are available at the locus carrying the mutation, this approach reveals a comprehensive set of genes exhibiting mRNA splicing defects in complete genomes and exomes
- Known breast cancer genes may be mutated in a higher percentage of tumours than currently reported
- Splicing mutations are more prevalent than those that are currently being reported