ResponderID™: The Biology-First Machine Learning Platform for Biomarker Discovery

Daniel Pointing[1], Rafael Rosengarten[1], LukaAusec[1], Mark Uhlik[2] and Laura Benjamin[2]

[1] Genialis, Inc. | [2] OncXerna Therapeutics, Inc.

ABSTRACT

Rarely has a complex gene signature been successfully translated from conception to a predictive biomarker algorithm ready for clinical practice. Genialis' ResponderID™ is a biomarker discovery platform that espouses a biology-first framework to model underlying disease biology. The platform consists of software and data/signature assets, and has been validated through the development of the Xerna™ TME panel. The Xerna™ TME Panel is a novel machine learning based transcriptomic biomarker designed to predict therapeutic response to various targeted therapies across multiple cancers. ResponderID continues to support the commercialization and market acceptance of the Xerna™ TME Panel.

CONCLUSIONS

ResponderID is a validated platform that enables the development and commercialization of predictive biomarkers. This platform combines proprietary technology, data assets and signatures with a comprehensive understanding of the underlying biology.

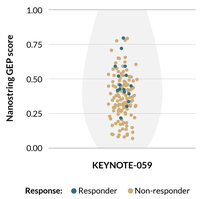
The platform's biology-first approach typically yields biomarkers that are applicable to a class of therapies rather than a single agent. Increasing the clinical utility for drug developers and diagnostic organizations but also putting patients and people first

The Xerna™ TME Panel is OncXerna's proprietary pan-cancer multi-MOA transcriptomic signature that was modeled using the ResponderID platform. This panel has been tested on over 10,000 patient samples across more than 10 solid tumor types and a range of NGS platforms. The Xerna™ TME Panel provides much more phenotypic information compared to standard of care immunotherapy and anti angiogenic biomarkers, typically with superior predictive performance.

ResponderID

Background

Traditional approaches to biomarker development rely on the modeling of drug response directly on clinical cohorts with outcomes data. This presents a number of challenges as clinical data is small and seldom has enough responders to power biomarker discovery. Resulting biomarkers are often overfitted, cannot be applied to new datasets and poorly differentiate those that are going to and not going to respond to treatment. *Figure 1* visualizes this style of approach with the 18 gene T-Cell-inflamed gene expression biomarker scores and responses for a cohort of gastric cancer patients treated with immune checkpoint inhibitors.



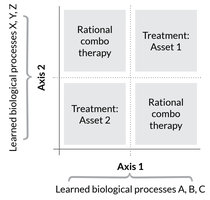
◀ *Figure 1.*
Example of a traditional biomarker which directly models response. *Biomarker score is based on the 18 gene inflammation expression signature. Patients of the Keynote-059 study are colored by response to pembrolizumab. Responders were not easily distinguishable from non responders using this biomarker.*

Biomarker Conceptualization

The ResponderID marries the use of software and NGS data with complex data science techniques to develop predictive machine learning classifiers based on attainable gene-expression values.

Implementing an approach that identifies and abstracts the biology that is most likely to respond to a given treatment modality increases the utility for clinicians and diagnostics organizations - not fixating on outcome data for any given drug means that the biomarker can likely be utilized on a class of treatments rather than a single agent.

ResponderID implementation starts with technical feasibility which evaluates biospecimen preparation, the sequencing technology, and the range of signal from gene expression signatures of interest. *Figure 2* shows a cartesian plane created by the intersection of two known biologies. This gives you four phenotype definitions, each with its own respective therapeutic hypothesis.

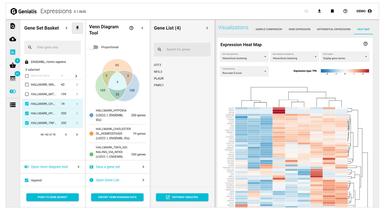


▲ *Figure 2.*
Phenotype landscape. *A schematic of how the intersection of multiple learned biological processes can predict a response to treatment or inform clinical decisions.*

Genialis Expressions

A cloud based repository for NGS data that enables consistent processing, annotation and visualization is essential to foster cross functional team collaboration for remote teams. Genialis Expressions is the technology layer that produces versioned machine learning ready data at scale.

Validated bioinformatic and quality workflow ensures the uniform sequencing and processing of data and assists in defining the analytical requirements for a downstream assay. FAIR data management facilitates annotation and curation, and supports reproducibility and reporting.

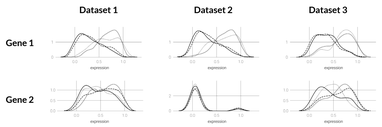


▲ *Figure 3.*
Genialis Expressions. A screenshot of the modules for interactive data visualization.

Genialis Precision Medicine

This is an extensively validated code base developed in accordance with the FDA guidance for Good Machine Learning Practice. GPM is used throughout the biomarker implementation lifecycle from commencing with the derivation of gene signatures that represent complex biological systems to the systematic mapping of independent datasets for real world validation. Some of the key methods that are readily deployed and dramatically improve the robustness of a biomarker in the clinical practise include:

- Evaluating the transferability of predictive signatures across molecular biomarker datasets
  (U.S. Patent App. Ser. No: 17/154,683) *(Figure 4)*
- System of preprocessors to harmonize disparate 'omics datasets by addressing bias and/or batch effects
  (PCT Application No. PCT/US2022/037860)



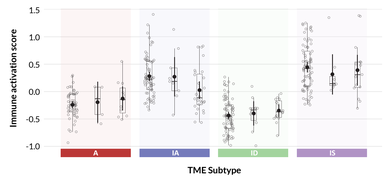
▲ *Figure 4.*
Feature transferability. *This method enables the identification of genes that are consistently and robustly expressed across different bias modalities including but not limited to tissue type and sequencing platform, thus broadening the applicability and utility of the locked model. The consistency of expressions for two genes across multiple datasets is shown, Gene 1 is similarly expressed across all datasets and phenotypes, in contrast, gene 2 is not consistently expressed in dataset 3.*

Data Catalog

Many publicly available biomedical datasets are not well suited for ML applications. An ideal dataset would measure tens of features (i.e. genes, proteins, metabolites, or morphological attributes), from hundreds or thousands of samples (e.g. patient biopsies, blood draws, cell lines, mouse models, etc). Conversely clinical datasets generally consist of too few patient samples with limited metadata that often do not follow a unified vocabulary, and these datasets typically cannot be used in unison. Genialis has onboarded, consistently annotated and aligned thousands of publicly available samples in addition to a large number of proprietary datasets to remedy this. The Genialis data catalog is an asset that allows us to readily call upon and reuse interoperable datasets to support the training, testing and validation of biomarkers.

Signatures and Algorithms

The ResponderID platform has an ever growing list of integrated public and propriety signatures and algorithms. Utilizing both public and proprietary biomarkers serves two purposes. Firstly, the biology of interest for a given program can be better mapped by understanding the correlation between multiple signatures from the public domain such as the Molecular Signature Database (MSigDB) as visualized in *Figure 5*. Secondly, it allows performance to be computed and compared for benchmarking purposes. *Table 1* below visualizes a selection of standard of care biomarkers including detected variants in genes and tumor mutational status and their associations with the Xerna™ TME panel subtypes. In a cohort of NSCLC patients, seven biomarkers (B2M, TMB high, LRP1B, TP53, RB1, JAK2, CDKN2B) showed significant associations with Xerna subtypes. Of note, all but one of these biomarkers were more common in the Immune Active (IA) subtype. The exception was CDK2NB, which was most common in the Immune Desert (ID) subtype.



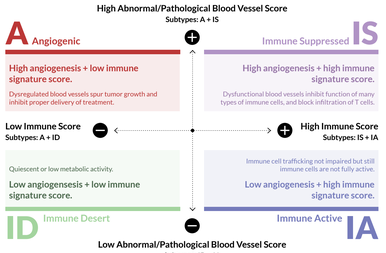
▲ *Figure 5.*
Gene set variation analysis activation (GSVA) scores for three gastric and colorectal cancer datasets using an Immune activation gene signature derived from the MsigDB across Xerna™ TME subtypes. There is a significant enrichment in activation scores within the high immune (IA & IS) TME subtypes when compared to low immune (A & ID).

Biomarker*	Total	A (n=15)	IA (n=13)	ID (n=34)	IS (n=42)	P-value Exact Test
B2M	3 (2.9%)	0 (0.0%)	3 (23.1%)	0 (0.0%)	0 (0.0%)	0.002
TMB high	28 (26.9%)	2 (13.3%)	8 (61.5%)	12 (35.3%)	6 (14.3%)	0.004
LRP1B	4 (3.8%)	0 (0.0%)	3 (23.1%)	1 (2.9%)	0 (0.0%)	0.01
TP53	57 (54.8%)	4 (26.7%)	10 (76.9%)	23 (67.6%)	20 (47.6%)	0.02
RB1	15 (14.4%)	0 (0.0%)	4 (30.8%)	8 (23.5%)	3 (7.1%)	0.02
JAK2	5 (4.8%)	0 (0.0%)	2 (15.4%)	3 (8.8%)	0 (0.0%)	0.04
CDKN2B	4 (3.8%)	0 (0.0%)	0 (0.0%)	4 (11.8%)	0 (0.0%)	0.04
KRAS	21 (20.2%)	2 (13.3%)	1 (7.7%)	4 (11.8%)	14 (33.3%)	0.07
ERBB2	3 (2.9%)	2 (13.3%)	0 (0.0%)	1 (2.9%)	0 (0.0%)	0.07

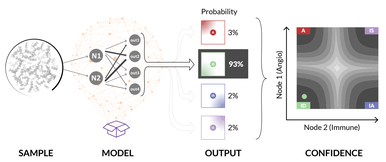
◀ *Table 1.*
The number (%) of patient samples carrying selected genetic markers with actionable alterations across Xerna™ TME subtypes.

Case study: Xerna™ TME Panel

The Xerna™ TME Panel is a novel diagnostic assay that uses formalin-fixed paraffin-embedded (FFPE) tissue-derived RNA gene expression data based on ~100 genes to classify patients into dominant biologies of the tumor microenvironment (TME). Built on the ResponderID platform the input gene signature represents angiogenic and immunogenic properties of stromal biology *(Figure 5)*, and the machine learning neural network that comprises the Xerna™ TME Panel algorithm *(Figure 6)* has learned interactions between these critical processes. The Xerna™ TME Panel can be used to classify a patient's tumor microenvironment in multiple solid cancers *(Figure 7)* along an immune and angiogenic axis, resulting in one of four TME phenotypes Angiogenic (A), Immune Active (IA), Immune Desert (ID), and Immune Suppressed (IS). Each class, or combinations of classes, can be predictive of treatment outcome with various targeted therapies, including anti-angiogenics and checkpoint inhibitors, as well as novel drugs targeting the intersection of these biologies.

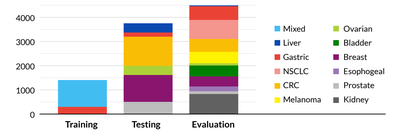


▲ *Figure 6.*
The four TME subtypes and biology that defines them described



▲ *Figure 7.*
Gene expression data (SAMPLE) is analyzed through a machine learning-based artificial neural net algorithm (MODEL) which provides a single subtype designation (OUTPUT) that is based on probabilized scores of the TME subtypes (CONFIDENCE). Shown here (Figure 6) is a representation of a single tissue sample analyzed by the model with its subtype designation "ID" and visualization on a latent space plot highlighted as a green circle.

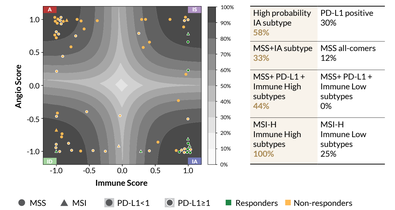
Number and Type of Samples Analyzed using the Xerna TME Panel



▲ *Figure 8.*
Xerna™ TME Panel works pan-caner. Approximately 10,000 patient samples have had Xerna TME subtypes computed across twelve major disease types

Xerna™ TME Panel is potentially predictive of response to immune checkpoint inhibitors *(Figure 9)*. Most responders are found to be immune-high, with the IA subtype significantly enriched for response compared to standard of care biomarkers *(Figure 9a)*.

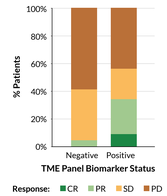
Gastric cancer patient cohort (N=73) treated with immune checkpoint inhibitor (ICI) monotherapy



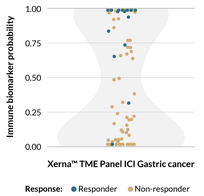
Best overall response rates comparison (%)

High probability IA subtype 58%	PD-L1 positive 30%
MSS+IA subtype 33%	MSS all-comers 12%
MSS+ PD-L1 + Immune High subtypes 44%	MSS+ PD-L1 + Immune Low subtypes 0%
MSI-H Immune High subtypes 100%	MSI-H Immune Low subtypes 25%

▲ *Figure 9a.*
Latent space plot of Xerna™ TME calls for samples from the Gastric Immune cohort. The latent space is a two-dimensional representation of the two neurons in the hidden layer of the model, with neuron 1 consisting of genes linked with immunogenic biology as the x-axis (Immune Score) and neuron 2 similarly with genes synonymous with angiogenesis as the y-axis (Angio Score) Glyphs are shaped according to their MSS/MSI status, outlined according to their PD-L1 CPS score status, and color-coded according to their best response. Contours represent different levels of probability estimates for the Xerna™ TME calls. Both the high immune subtypes (IS and IA) were deemed as biomarker positive.



▲ *Figure 9b.*
Stacked bar graph showing the enrichment of responders in the biomarker positive group.



▲ *Figure 9c.*
Xerna™ TME panel biomarker outputs in this cohort are binary-like and highly enriched for responders in the immune biomarker-positive group (>0.5)