BioInsight, a customized application for biomarker exploration and discovery

Brian Frost[1], Hansen Han[2], Martin Weigl[3], Dale Erikson[2], Oltion Champari[3], Dominic Girardi[3], Daniela Moitzi[3*], Lin Tang[1*]
[1] Schödinger Inc., San Diego, CA. [2] Schrödinger Inc., New York, NY. [3] DataVisyn Inc., Linz, Austria

Introduction

Cell lines and other preclinical models are routinely used for drug screening studies. Genomic profiles are largely available for these models. Genomic features derived from multi-omic profiles are often associated with the screening results to identify therapeutic biomarkers. These features include gene-level genomic data related to gene expression, mutation, copy number, protein expression, and gene dependency. There is also a growing body of knowledge on composite genomic features that could serve as candidate biomarkers. These include pathway activities, gene signatures, tumor mutation burden (TMB), chromosomal instability (CIN) status, MSI status, tumor subtypes. Although information on these genomic features are available in public portals or in publications, there is currently no central repository that integrates these genomic features with pharmacogenomic screening studies across model types or internal and external resources, making it challenging to explore and discover biomarkers in a comprehensive and unified manner.

Here, we designed and built a customized application that integrates large-scale pharmacogenomic screening studies from public and internal resources. We also collected the multi-omic profiles and composite genomic features for the preclinical models. This allows for therapeutic biomarker discovery, exploration and validation between different studies and preclinical model systems. The data scheme and framework proposed here will also foster the future growth of knowledge in both the genomic biomarker and screening study spaces.

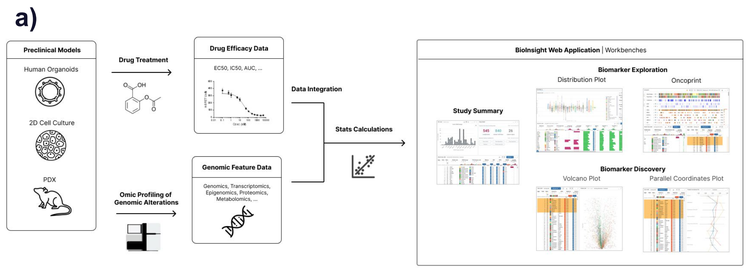
Data scheme

The BioInsight database is structured in three parts: genomic datasets, screening datasets, and calculated statistics. BioInsight links omic data with compound screening data by their shared preclinical models (Figure 1a). By doing so, the application supports multiple sets of omic and drug screening studies and enables users to analyze both public and internal data. To handle cases of compounds which can be screened in multiple studies, we use the term “singlet” to represent a unique study and compound combination. For our initial purposes, we ingested CTD2[1] screening dataset with corresponding genomic feature dataset adopted from DepMap[2].

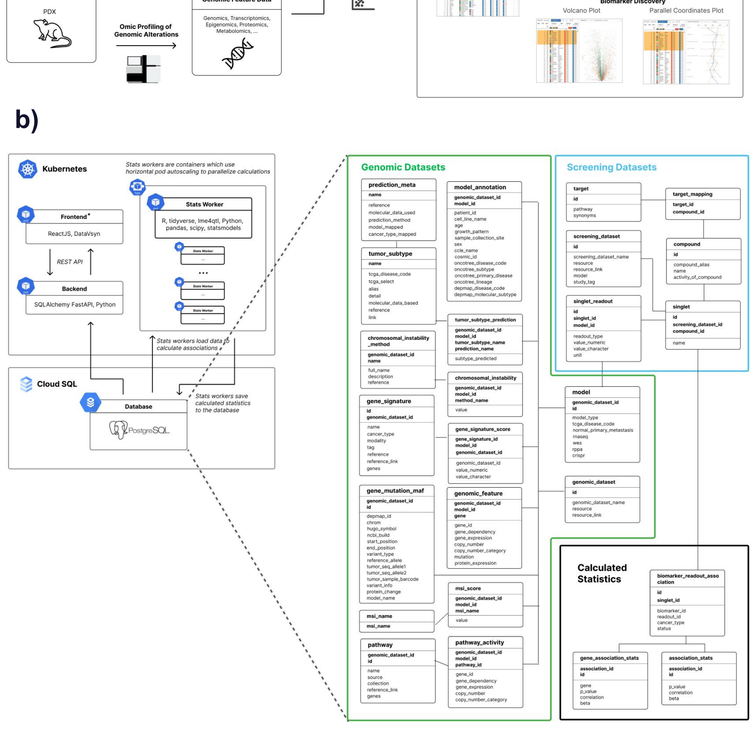
By deploying the application via Kubernetes and using Horizontal Pod Autoscaling (HPA), we are able to highly parallelize statistics calculations of associations between genomic features and compound efficacy (Figure 1b).

Figure 1. **a)** Conceptual design for BioInsight **b)** Deployment scheme for the application and entity relationship diagram of the BioInsight database.

a)



b)



Genomic features that serve as therapeutic biomarkers

We collected gene-level and composite genomic features from multi-omics profiles. Together, these genomic features serve as candidate biomarkers to predict the therapeutic response of the compounds. Table 1. summarizes the genomic features collected for cancer cell line models. The genomic features defined here could be added for other preclinical models, including *ex vivo*, CDX, PDX model, if such data is available.

Table 1. Genomic features collected for cancer cell line models.

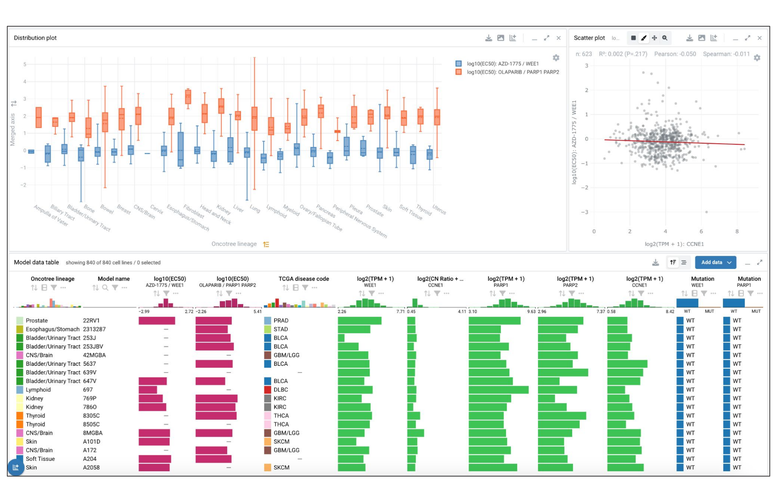
	Genomic Feature	Description	Resource / Reference
Gene-Level	Gene Expression	From RNAseq in value of log2(TPM+1)	DepMap Public 22Q4 Files
	Gene Copy Number	From WES. log2 transformed values with a pseudo-count of 1	
	Gene Mutation	Somatic point mutations	
	Protein Expression	CCLE Reverse Phase Protein Array (RPPA) data	
	Gene Dependency	CRISPR gene effect scores	
Composite	Pathway Activity Score	Calculated with GSVA from gene expression profiles. Canonical pathway definitions were adopted from MSigDB pathway signature database.	Hänzelmann S 2013 MSigDB
	Gene Signature Score	• Replication stress signature: Gene expression based replication stress signature for pan-cancer types • Leukemia stem cell signature: Gene expression based stem cell signature for leukemia	Takahashi N. et. al. 2022 Ng SWK. et. al. 2016
	Tumor Mutation Burden (TMB)	Tumor mutation burden are calculated from somatic mutation files for total mutation, total mutation per MB	
	Chromosomal Instability (CIN)	Chromosomal instability are calculated from mutation maf profile for several CIN metrics, CIN_tai, CIN_tai.modefiled, CIN_cna, CIN_base_segments, CIN_Break_points, CIN_fga	Baumbusch LO. et. al. 2013 Davidson JM. et. al. 2014 Chin SF. et. al. 2007
	Microsatellite Instability (MSI)	MSI status are adopted from DepMap model annotation with 2 original sources: CCLE_MSI, GDSC_MSI. Filters are applied to select only those biomarker levels (ie. MSI-H vs. MSI-L) with >= 5 samples and that comprised >=10% of the total samples.	DepMap Public 22Q4 Files
	Molecular Tumor Subtype	Four different computational studies that were selected for their comprehensiveness in CCLE model panel and prediction in pan-cancer manner. Cancer subtype prediction results were adopted from the four publications.	Peng D. et. al 2021 Zhang X. et. al. 2021 Salvadores M. et. al. 2020 Yu K. et. al. 2019

Biomarker exploration

BioInsight offers robust capabilities for visually exploring integrated pharmacogenomic data using state-of-the-art visualization technologies. The core of the user interface is based on Ordino[3]. Users can seamlessly compare and visualize the different types of screening and biomarker data within BioInsight, all in a single workbench.

In addition to standard plot types to analyze and interpret the data (i.e. boxplots, violin plots, scatterplots, and bar charts), BioInsight also includes an Oncoprint visualization to view mutation/copy number alterations alongside other model features, allowing for easy identification of patterns and correlations.

Figure 2. Example workbench in BioInsight for biomarker exploration. The distribution plot shown is comparing EC50 values for AZD-1775 vs. Olaparib across multiple cancer types while the scatter plot is plotting CCNE1 gene expression against EC50 values for AZD-1775. The model data table includes all the model and pharmacogenomic data for this screening study, and allows for easy data subsetting and highlighting in the plots.



Biomarker discovery

To assist with biomarker discovery in a more complete genomic feature space, we precalculated association tests between genomic features and compound sensitivity, at both a pan-cancer and cancer-specific level. Linear regression, logistic regression, ANOVA, linear mixed models, or the Kruskal-Wallis test were used as appropriate for genomic biomarker features (continuous or categorical). To enable users to explore the large volume of precalculated results, we built biomarker discovery workbenches that include a biomarker guide panel, an Ordino-based data table and visualizations using a volcano plot or a parallel coordinates plot. The volcano plot displays effect size and statistical significance to quickly highlight the outlier biomarker(s) (not shown). The parallel coordinates plot displays the distribution of association test results for given biomarker(s) across cancer types (Figure 3). Using these tools, users are able to identify biomarkers with significant associations with compound sensitivity readout and compare them in a visually efficient manner across cancer types.

Figure 3. Examples of comparison of CCNE1 and TP53 copy number as biomarker for AZD-1775. The data table facilitates filtering to biomarkers with high association with AZD-1775 EC50. CCNE1 copy number was identified as a top biomarker showed higher association with EC50 values in Bone and Head and Neck cancers than TP53.



Conclusion

- BioInsight is a unique solution to integrate large-scale pharmacogenomic data that enables biomarker exploration, validation, and discovery across preclinical models and drug screening studies from public and internal resources.
- The application uses a data schema that integrates drug screening study, gene-level multi-omics biomarker and heterogeneous composite genomic biomarkers, including pathway activity score, gene signature, TMB, CIN, MSI and molecular tumor subtypes. This sets up the foundation for future data growth in both the genomic biomarker and screening study spaces.
- This project incorporates state-of-the-art visualization technologies with Ordino for flexible data and plot selection, and easy data comparison across screening studies and model systems.
- A light public version of BioInsight is available at www.bioinsight.app.datavisyn.io/

Reference:
1. CTD2, www.cancer.gov/ccg/research/functional-genomics/ctd2
2. DepMap, https://depmap.org
3. Marc Streit et. al. Ordino: visual analysis tool for ranking and exploring genes, cell lines, and tissue samples. *Bioinformatics, 35(17): 3140-3142, 2019.*

Scan here to access light public version of BioInsight

*Corresponding author
www.datavisyn.io
daniela.moitzi@datavisyn.io

www.schrodinger.com
lin.tang@schrodinger.com