

James R. Brown, M.Sc., Ph.D.

Founder, JRBrown Bio Consulting LLC

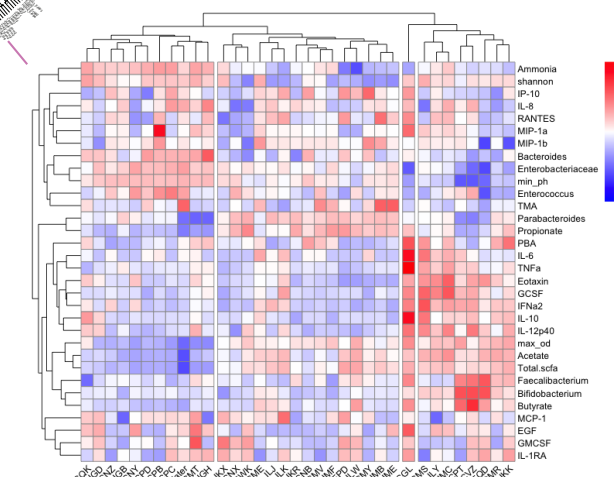
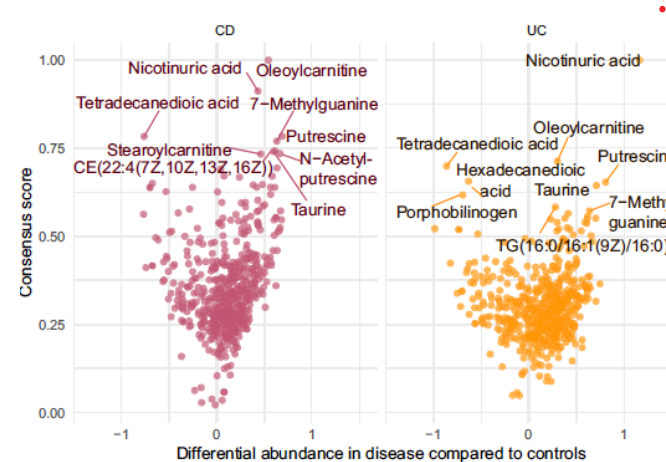
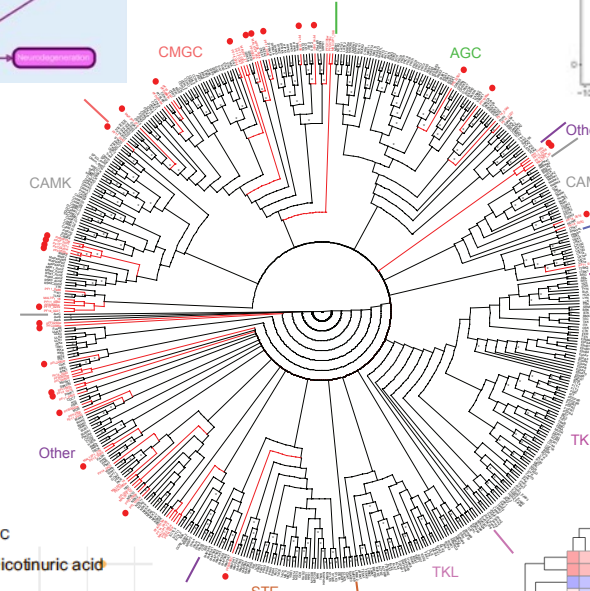
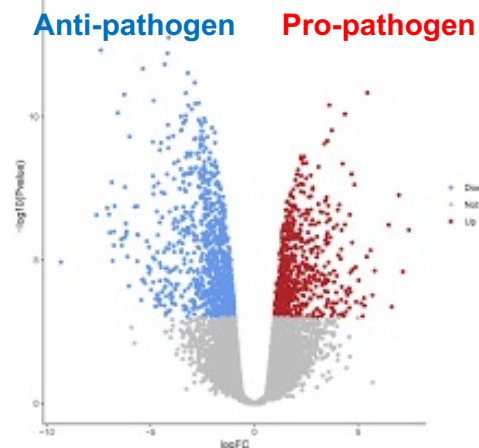
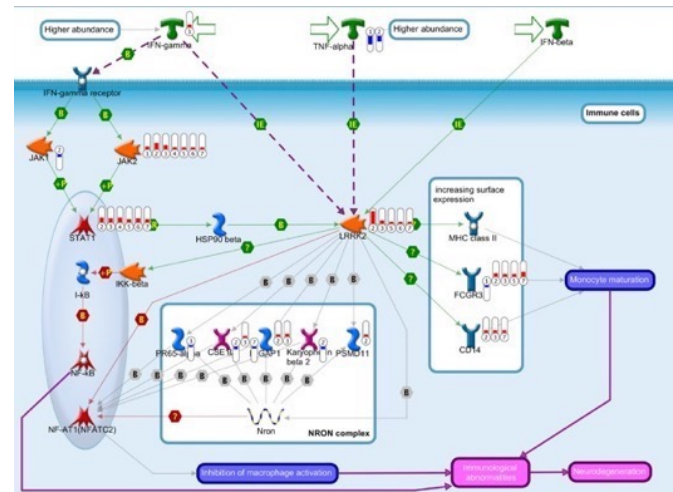
Visiting Scholar, Drexel University

Courtesy Professor, U. Florida

Former affiliations: GSK , Kaleido Biosciences, Novasenta

# Computational Approaches to Drug Target Discovery and Validation

March 12, 2024  
University of Puerto Rico –  
Medical Sciences Campus





# Introduction

## Biography – James (Jim) R. Brown



VP, Head of Computational Sciences

Executive Director,  
Head of Computational Biology  
& Integrated Data Sciences

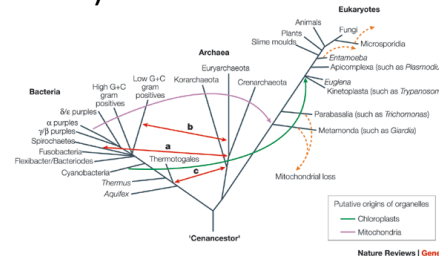


MRC Post-doctoral Fellow,  
Molecular Evolution, Tree of Life  
(W.F. Doolittle lab)

Director & Senior Fellow,  
Computational Biology,  
Infectious Diseases



Ph.D. Molecular Genetics,  
Bioinformatics



Marine Technologist,  
Canada Centre for Inland Waters



B.Sc. Marine Biology



M.Sc. Quantitative  
Ecology, Aquaculture

### Computational Biology BioPharma Experience (1996-present)

- Leader of computational teams for drug discovery
- Mainly infectious diseases, microbiome & oncology but also respiratory, neurological, metabolic & immune diseases.
- Active in public science (~118 papers; NIH panels, ad boards)
- **PR-INBRE EAC member since 2001**

- Visiting Scholar (G. Rosen Lab)
- Courtesy Faculty, U. Florida
- JRBrown Bio Consulting



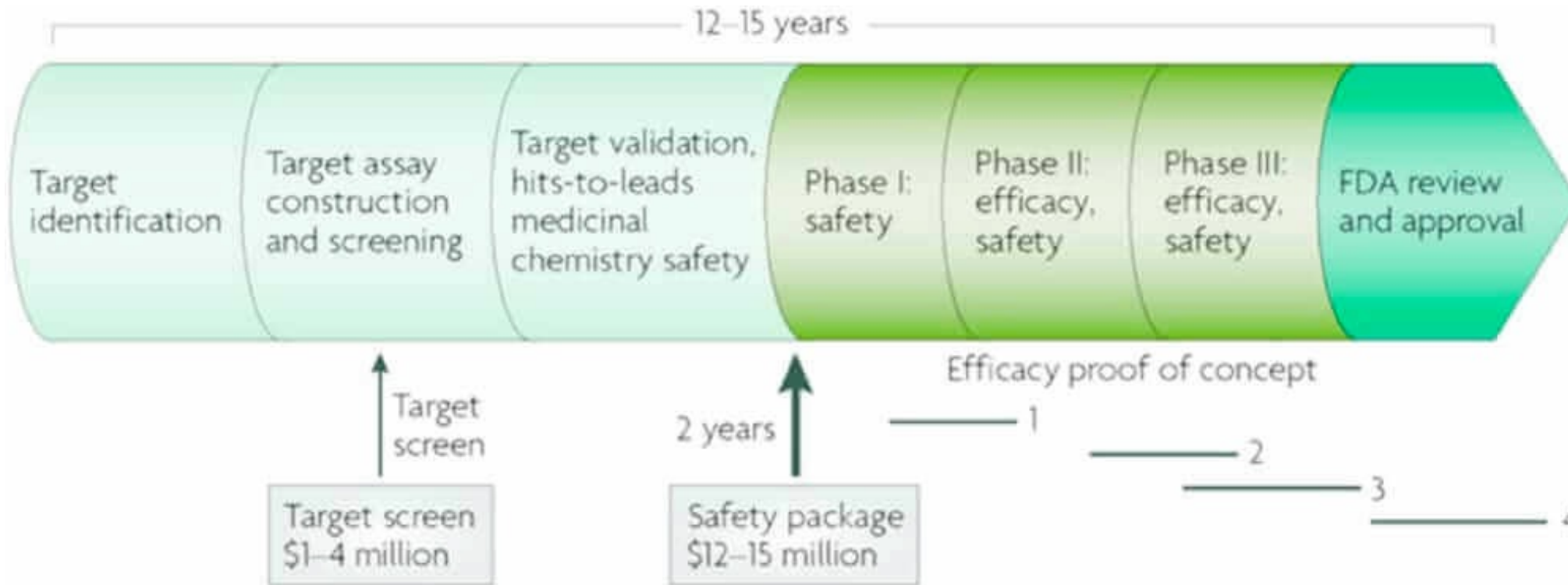


# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism databases
7. Cancer relevant databases
8. Concluding remarks & discussion



# Key Stages and Timelines in Drug Development



- Estimated R&D costs per drug range from \$133M to \$6B.\*
- Clinical phases are the most costly stages
- Need to “fail early” and “fail fast”

Roses. 2008. Nature Rev Drug Disc. 7:807. <https://www.nature.com/articles/nrd2593>

Nature Reviews | Drug Discovery

\* Rennane et al. 2022. Inquiry. 58: 004695802. <https://journals.sagepub.com/doi/10.1177/00469580211059731>



# Challenges of Drug Discovery and Development

- From 2017-2022, among 10 major therapy areas the top two as measured by proportion of clinical trials are oncology (24 %) and infectious disease (12 %).
- Oncology drugs also have the lowest clinical trial success rate (3.4%).
- Vaccines for infectious diseases have the highest clinical trial success rate (33.4%).
- Main reasons for drug failures:
  1. Efficacy\*
  2. Safety\*
  3. Commercial / financial
- \* Can be partially addressed by computational approaches

## Probability of Success<sup>2</sup> by Clinical Trial Phase and Therapeutic Area

	<i>P1 to P2</i>	<i>P2 to P3</i>	<i>P3 to Approval</i>	<i>Overall</i>
<i>Oncology</i>	57.6	32.7	35.5	3.4
<i>Metabolic/Endocrinology</i>	76.2	59.7	51.6	19.6
<i>Cardiovascular</i>	73.3	65.7	62.2	25.5
<i>Central Nervous System</i>	73.2	51.9	51.1	15.0
<i>Autoimmune/Inflammation</i>	69.8	45.7	63.7	15.1
<i>Genitourinary</i>	68.7	57.1	66.5	21.6
<i>Infectious Disease</i>	70.1	58.3	75.3	25.2
<i>Ophthalmology</i>	87.1	60.7	74.9	32.6
<i>Vaccines (Infectious Disease)</i>	76.8	58.2	85.4	33.4
<i>Overall</i>	66.4	48.6	59.0	13.8
<i>Overall (Excluding Oncology)</i>	73.0	55.7	63.6	20.9

Source: Chi Heem Wong, Kien Wei Siah, Andrew W Lo. "Estimation of clinical trial success rates and related parameters." *Biostatistics* 20(2): April 2019, Pages 273-286. Published online: 31 January 2018. DOI: <https://doi.org/10.1093/biostatistics/kxx069>

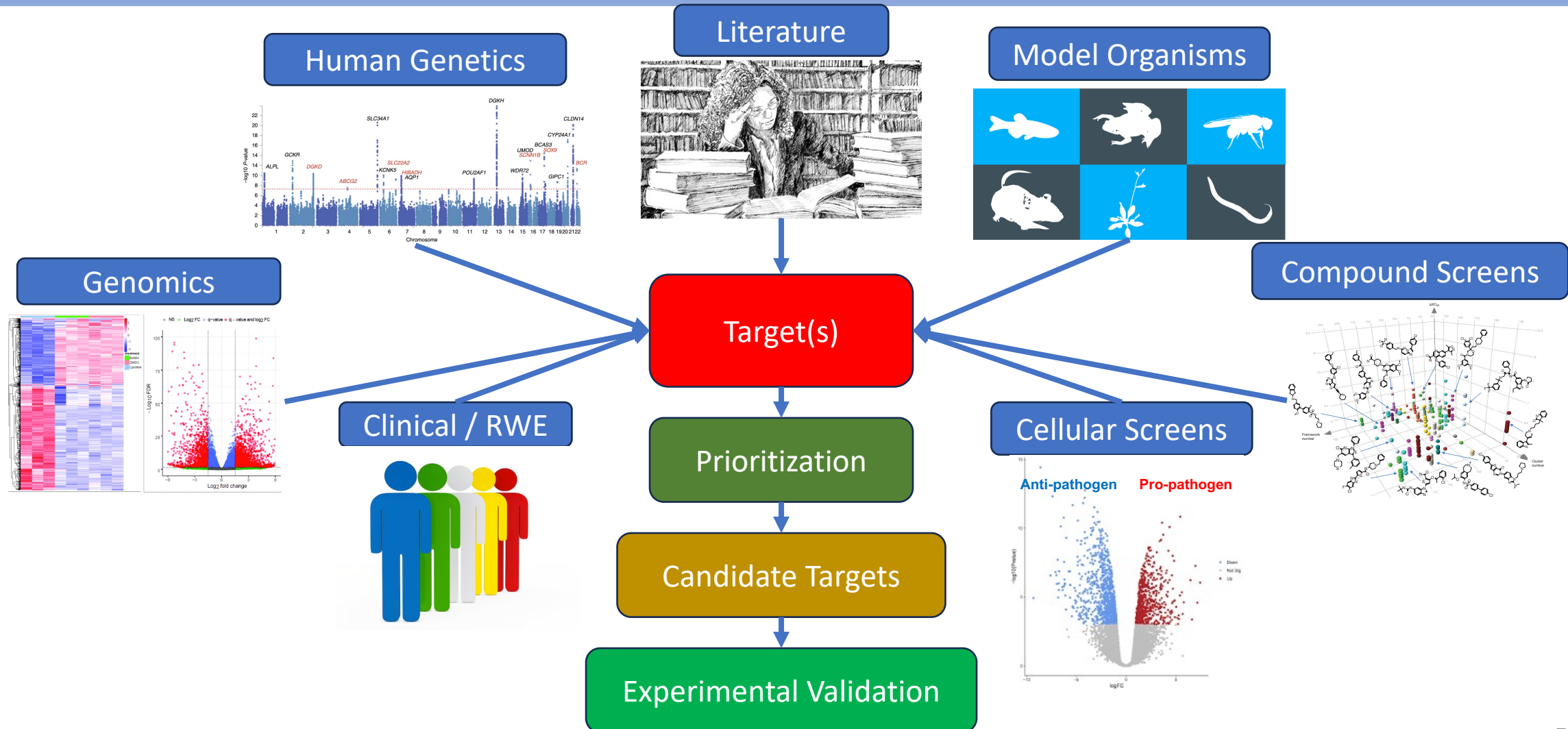


# Key Features of a “Good” Drug Target

Feature	Considerations
Addresses An Unmet Medical Need	<ul style="list-style-type: none"><li>• Target has a causal role in the disease</li><li>• Target is expressed in disease tissue</li><li>• Linked to human disease epidemiology and any potential sub-type cohorts are known</li></ul>
Efficacy	<ul style="list-style-type: none"><li>• Modulating the target has the potential to change the disease phenotype through a known mechanism of action (MOA)</li><li>• Target linked to disease related pathways</li><li>• Human genetic phenotypes exist that might inform about target-disease associations</li><li>• Understanding potential target redundancies and other drug-resistance mechanisms</li><li>• Having informative pre-clinical in vitro, ex vivo and in vivo models for clinical translation</li></ul>
Druggability	<ul style="list-style-type: none"><li>• Target gene, transcript or protein can be modulated in the desired direction and intensity</li><li>• Known drug modalities for modulating the target (i.e., small molecule, mAb, vaccine, siRNA, protein degradation, cell-gene therapy, CRISPR, etc.)</li><li>• For any particular modality, the target is accessible and therapeutic dosing is tolerable and efficacious</li></ul>
Safety	<ul style="list-style-type: none"><li>• Assessments of potential off-target effects of the drug</li><li>• Existence of suitable pre-clinical models for toxicity testing</li><li>• An understanding of potential genetic factors that could impact drug tolerance and safety</li></ul>



# Sources of Novel Therapeutic Targets: Finding the Best Candidate





# Target Evidence: Multiple Approaches

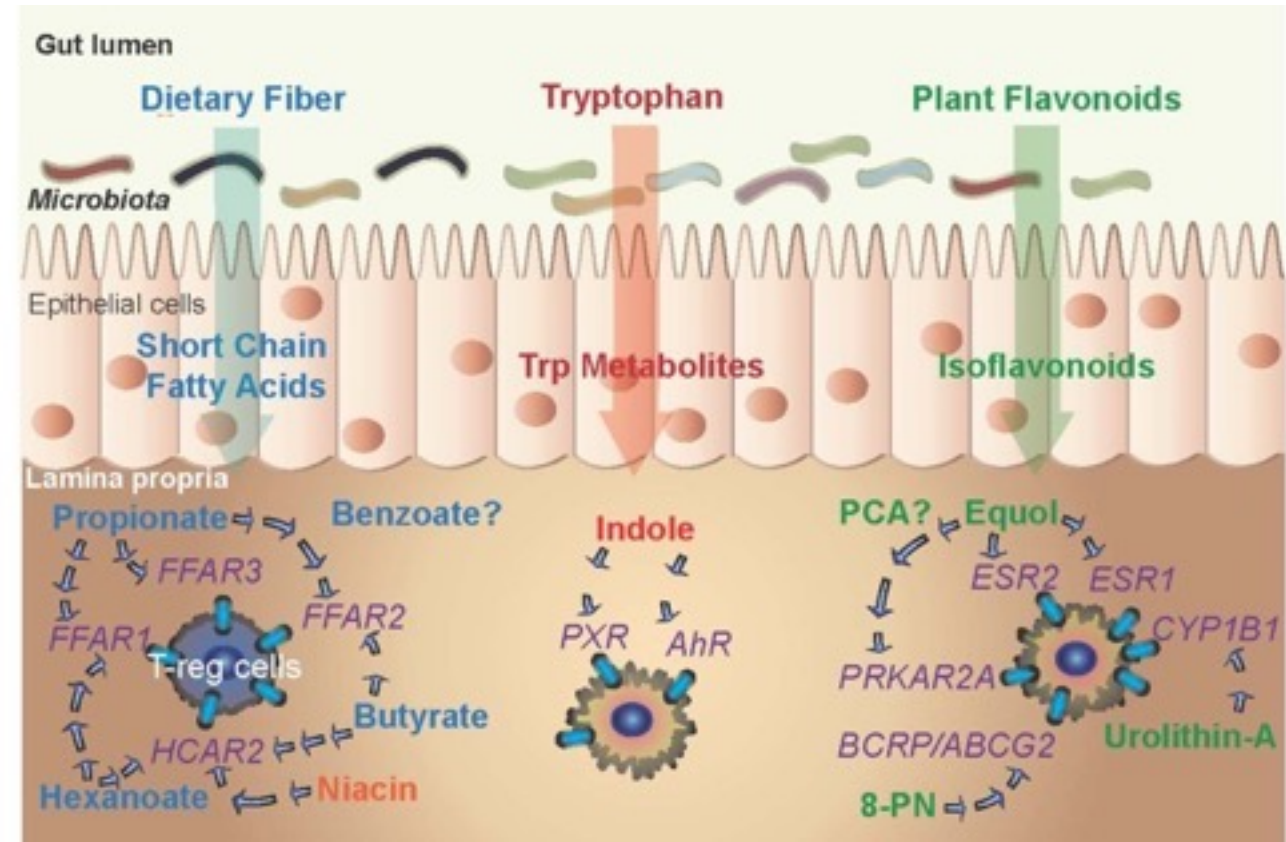
- Importance of in silico approaches for prioritizing therapeutic targets
  - De-risking and prioritization before devoting time and resources for lab studies
  - Precision medicine – targets specific to certain patient sub-populations
  - Increasing the probability of clinical success
- Key areas for target evidence
  - Human genetics – disease-to-gene linkages in genome-wide association studies (GWAS)
  - Genomics – expression of the target in diseased tissues
  - Known drugs or tool compounds
  - Clinical trial status
  - Model organism phenotypes
  - Large scale genome-wide gene knock-out and/or overexpression datasets
- Tremendous growth in genomics technologies, databases, analytical tools and query interfaces



# Example: The Apothecary Within – Targeting Human-Microbial Crosstalk



- Microbiome metabolism of dietary fibers generates many diverse metabolites with positive immuno-modulatory effects.
- Metabolites are advantageous starting points for drug discovery:
  - Known modulators of host immunity (i.e., Cohen et al. 2017. *Nature* 549:48).
  - Well-tolerated as endogenous molecules.
  - Evolutionary optimized metabolite-receptor pairing for selectivity and specificity.
  - Many successfully launched drugs have “metabolite-like” properties (Dobson et al. 2009 *Drug Discovery Today* 14:31).
- Challenge: Low-throughput of current experimental approaches to identify potential metabolite ligand-receptor linkages.
- *Can we accelerate the discovery of useful metabolite-protein ligand pairings via in silico hypothesis generation?*
  - *Then test/validate some predictions with in vitro cellular assays.*



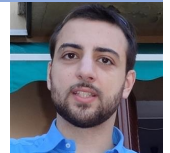
Saha et al. 2016. *Drug Discovery Today* 21:692



# The Human Microbiome Project 2 (HMP2)

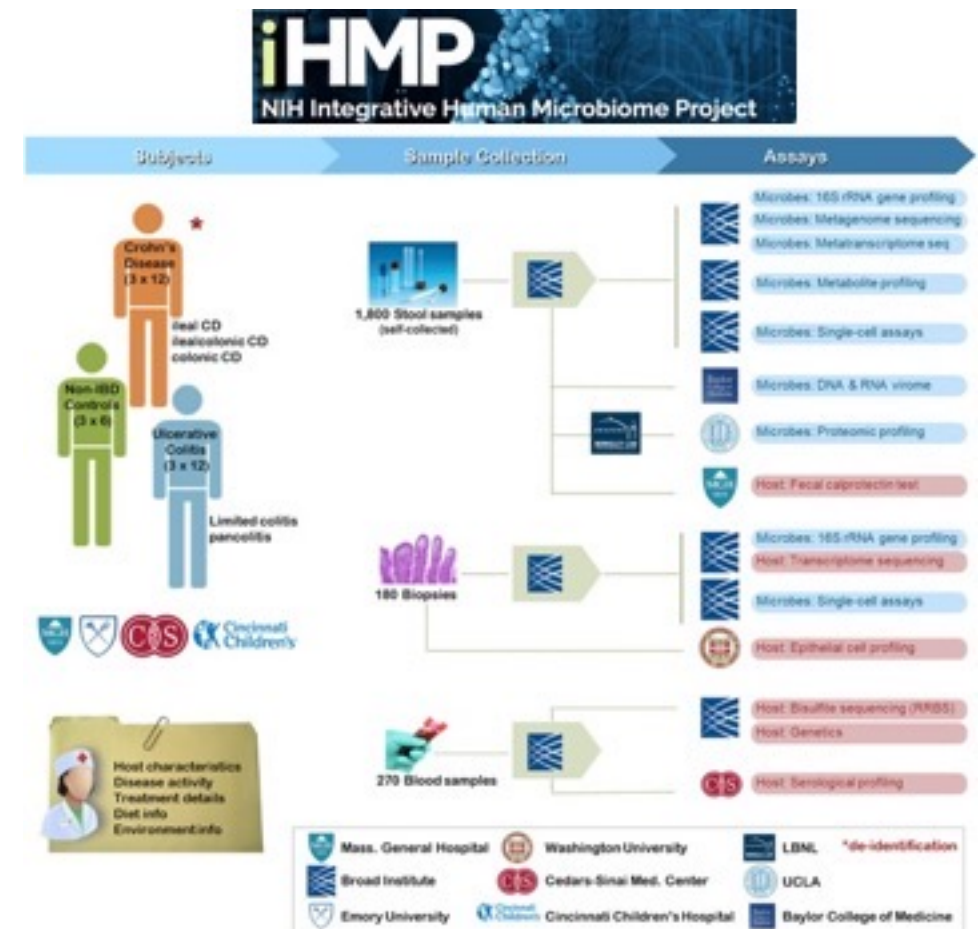


Dr. Andrea Nuzzo,  
Early Talent PDF;  
Assoc. Dir., GSK



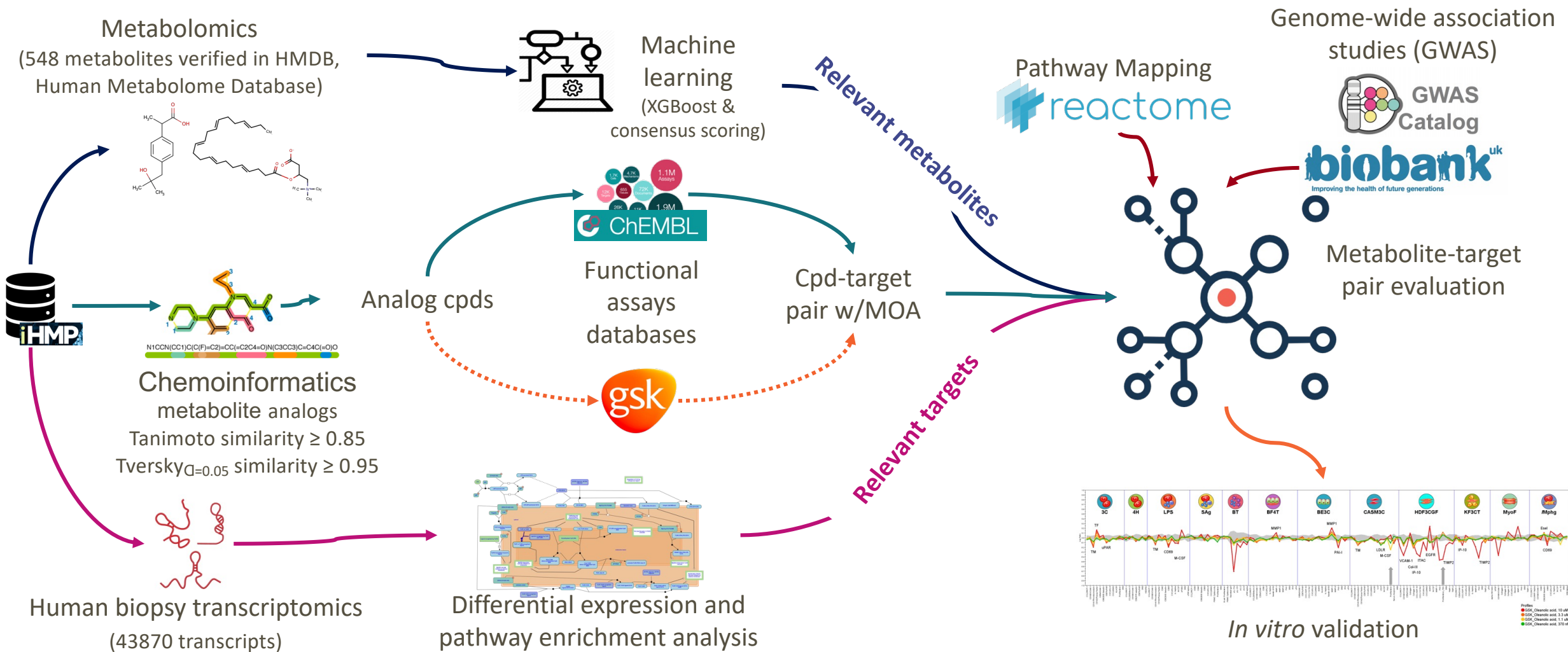
- Inflammatory bowel disease (IBD) patients:
  - CD: Crohn's disease
  - UC: Ulcerative colitis
- Multi-omics longitudinal assays:
  - Human host genetics (though underpowered for GWAS)
  - RNASeq from human biopsies
  - Metagenome, metatranscriptome, metaproteome & stool metabolome

	Controls (nonIBD)	Crohn's disease (CD)	Ulcerative colitis (UC)	Tot
Participants	26	49	30	105
Metagenomic samples	429	750	459	1638
Metabolomic samples	135	265	146	546
RNAseq samples	51	127	74	252





# Computational and In vitro Validation Workflow



\* identified in the Human Metabolome Database [HMDB]

Nuzzo...Brown. 2021. Commun. Biol. (Nature). 4:288



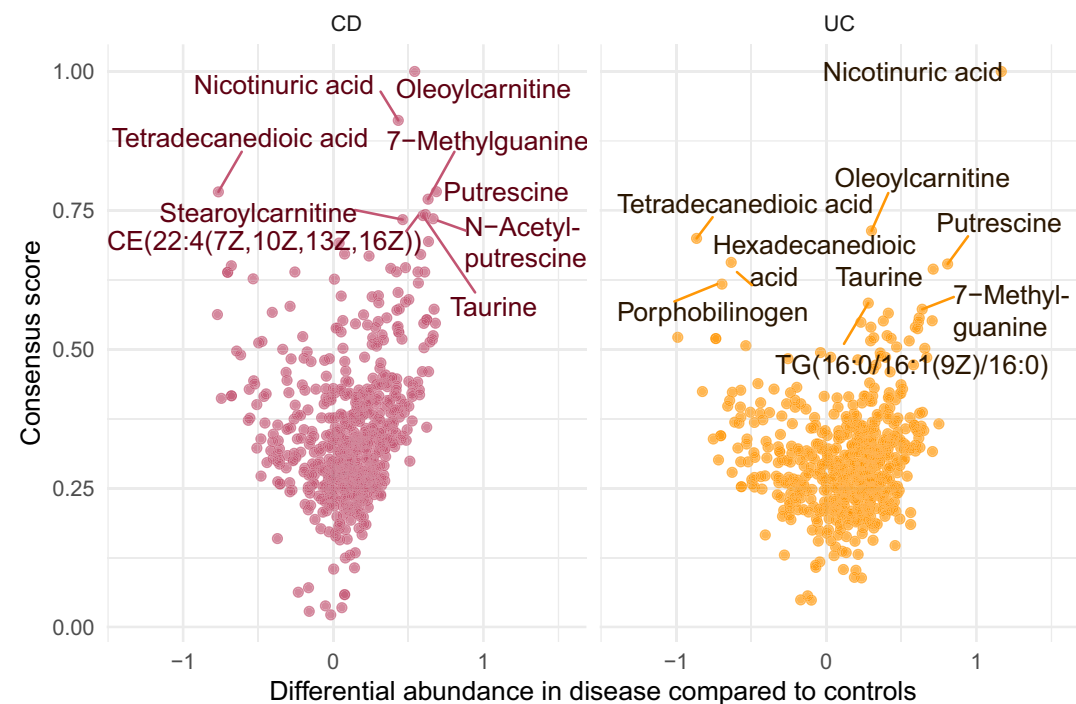
# Metabolomics and Transcriptomics in IBD Samples



## Metabolites

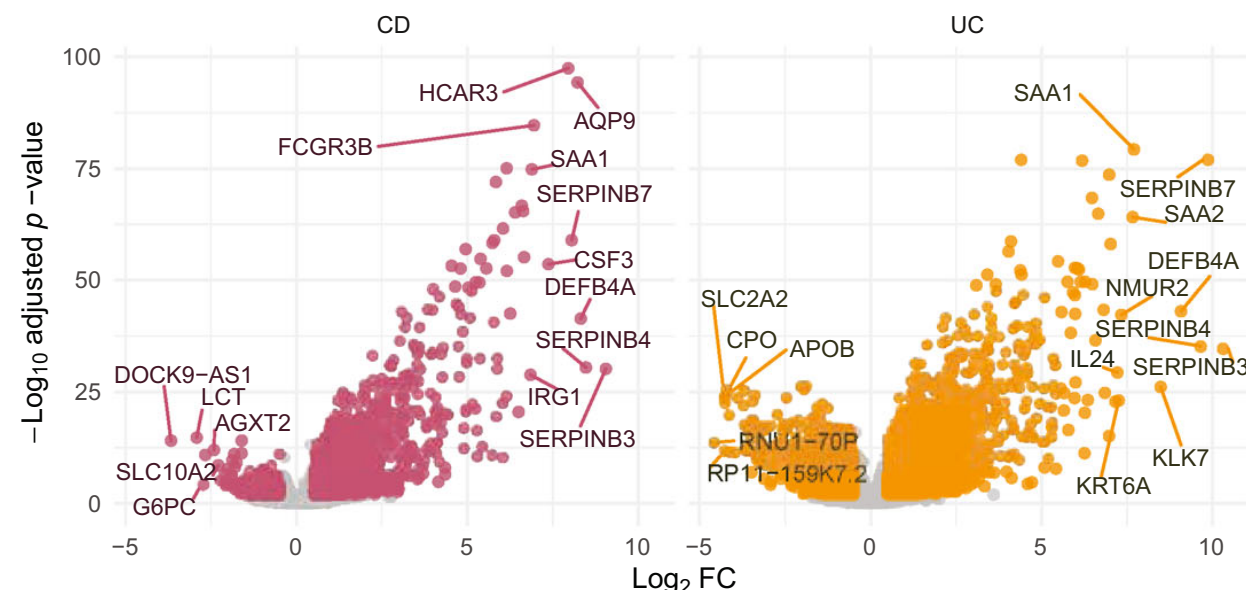
After ML analyses, top quartile (n = 192) to downstream analysis)

diagnosis



## Differentially Expressed Genes (DEGs)

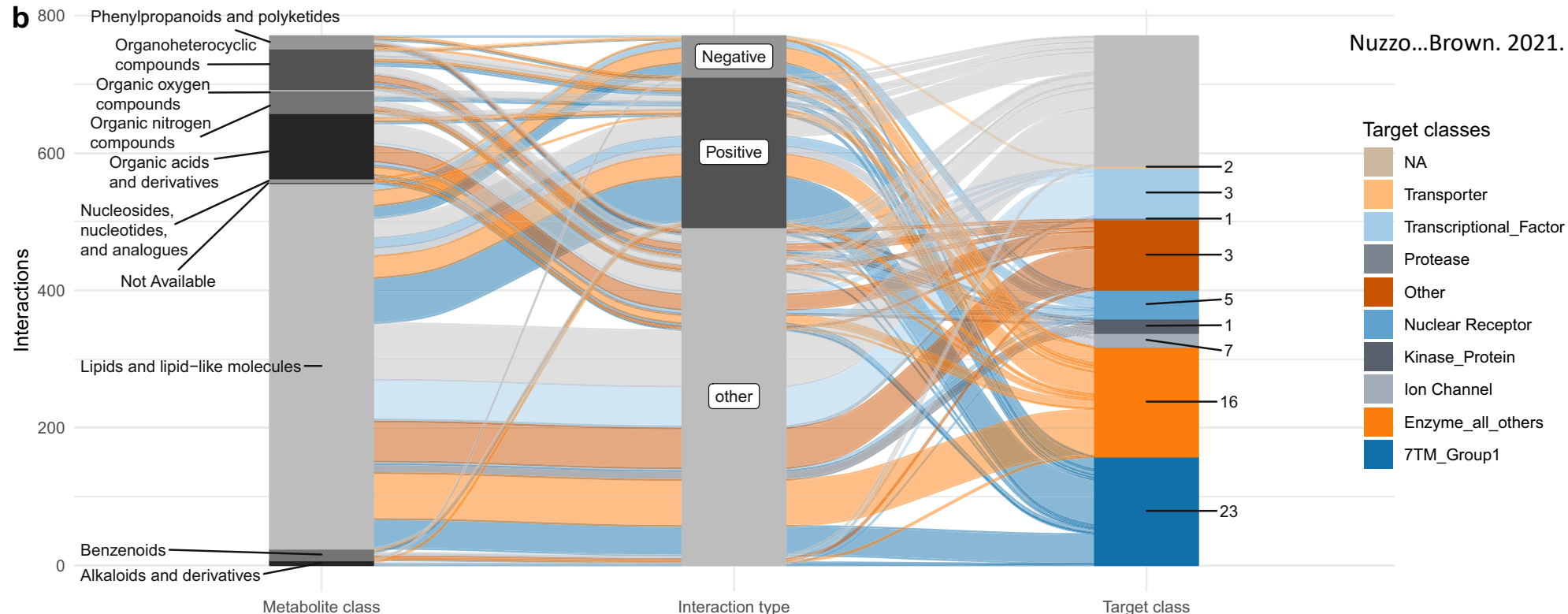
Total DEGs n = 2107 of which 820 DEGs shared between CD & UC



- Differential abundance of metabolites and gene RNA-seq in CD and UC patients compared to non-IBD subjects
- Prioritized known metabolites reported in the Human Metabolome Database.
- Gene transcripts were aligned to Genome Reference Consortium Human Build 37 (GRCh37).



# Connecting Metabolites and Drug Targets



Nuzzo...Brown. 2021. Commun. Biol. (Nature). 4:288

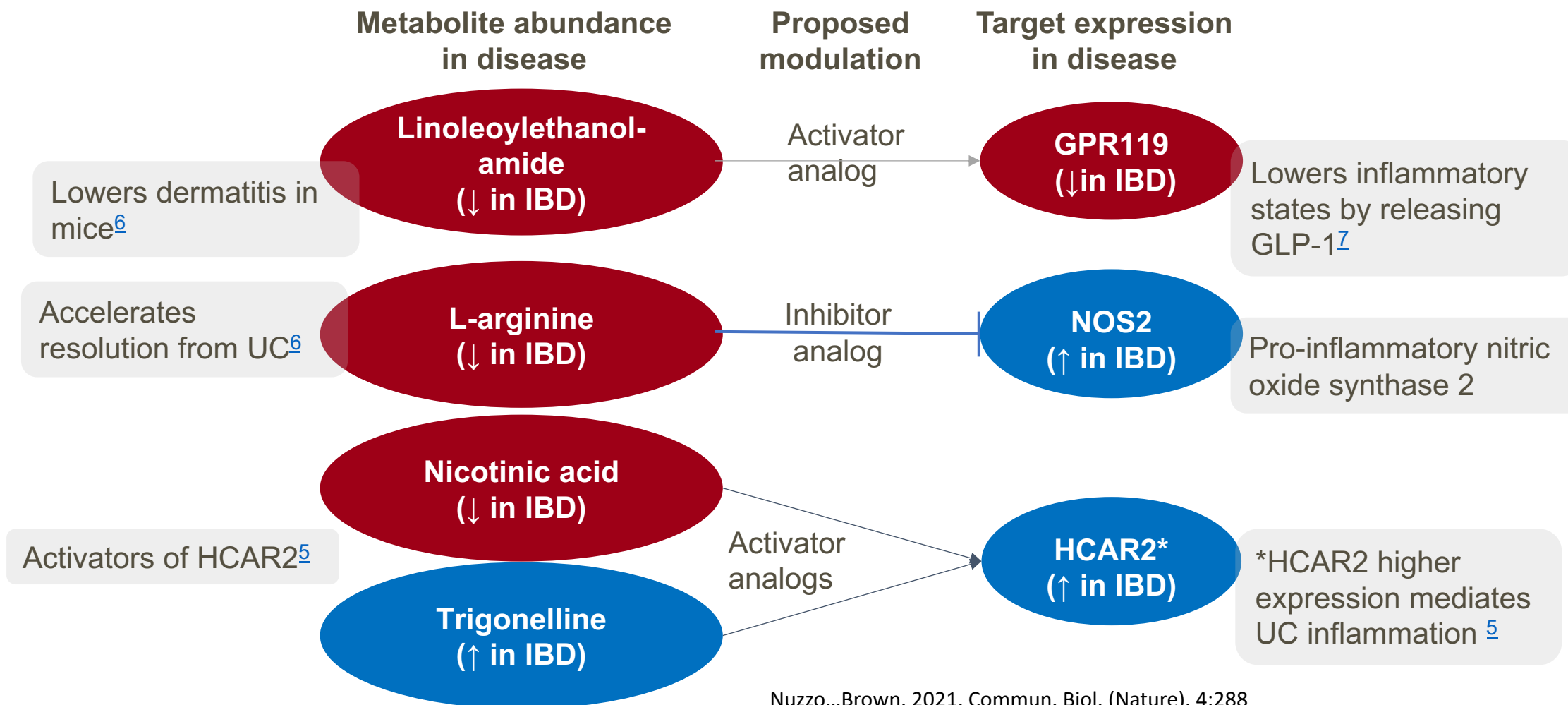
- After filtering, 135 metabolites provisionally connected to 80 perspective proteins.
- Distribution of connections between metabolite classes, modulation type and drug target classes (numbers represent unique targets per drug target class [  $n = 61$ ]). Some genes and metabolites have multiple interactions)
  - Filtered for metabolite-protein pairs with high binding affinity (i.e., either  $pIC_{50}$  or  $pEC_{50}$  values  $\geq 5.5$ )
  - Highly pleiotropic metabolites and targets ( $\geq 20$  predicted interactions) were removed.



# Metabolite Co-directionality with Target Gene Expression



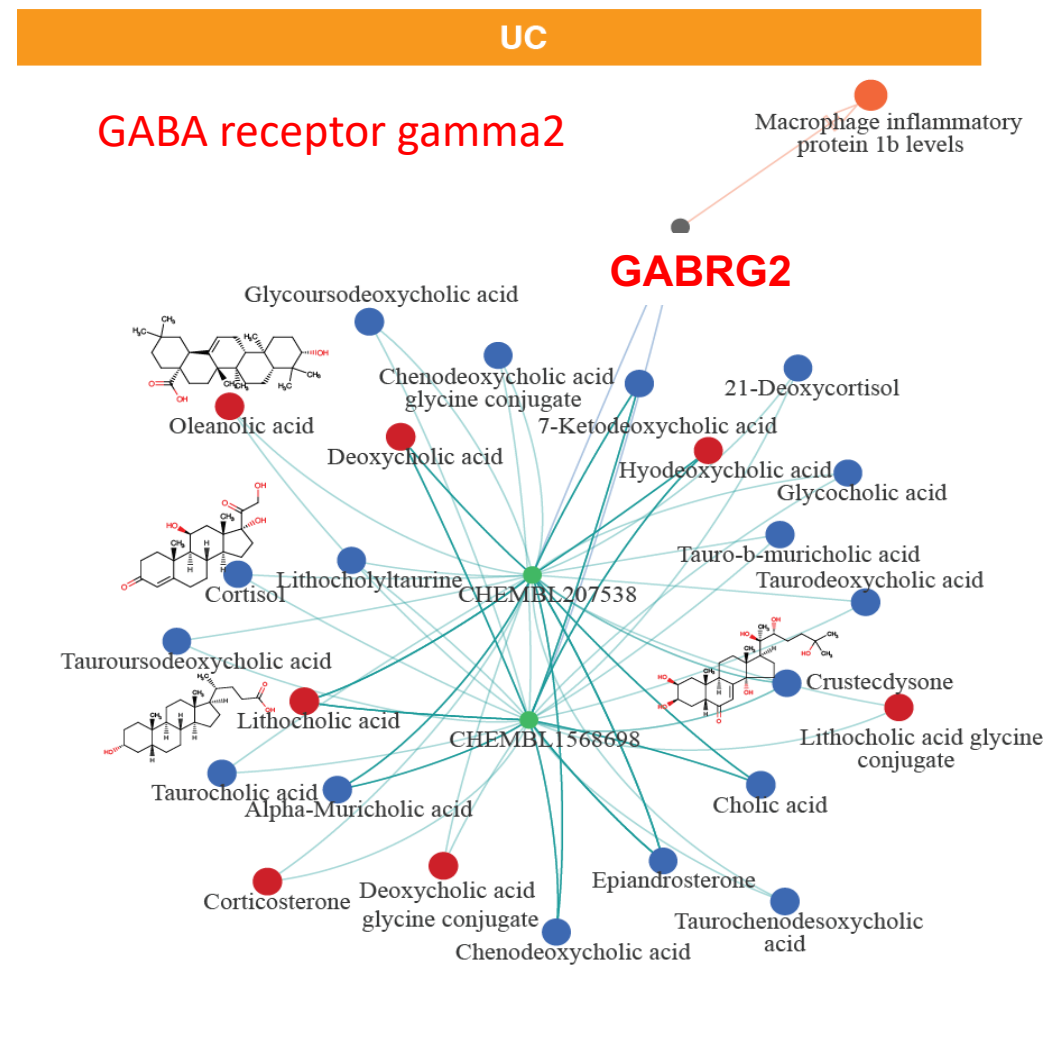
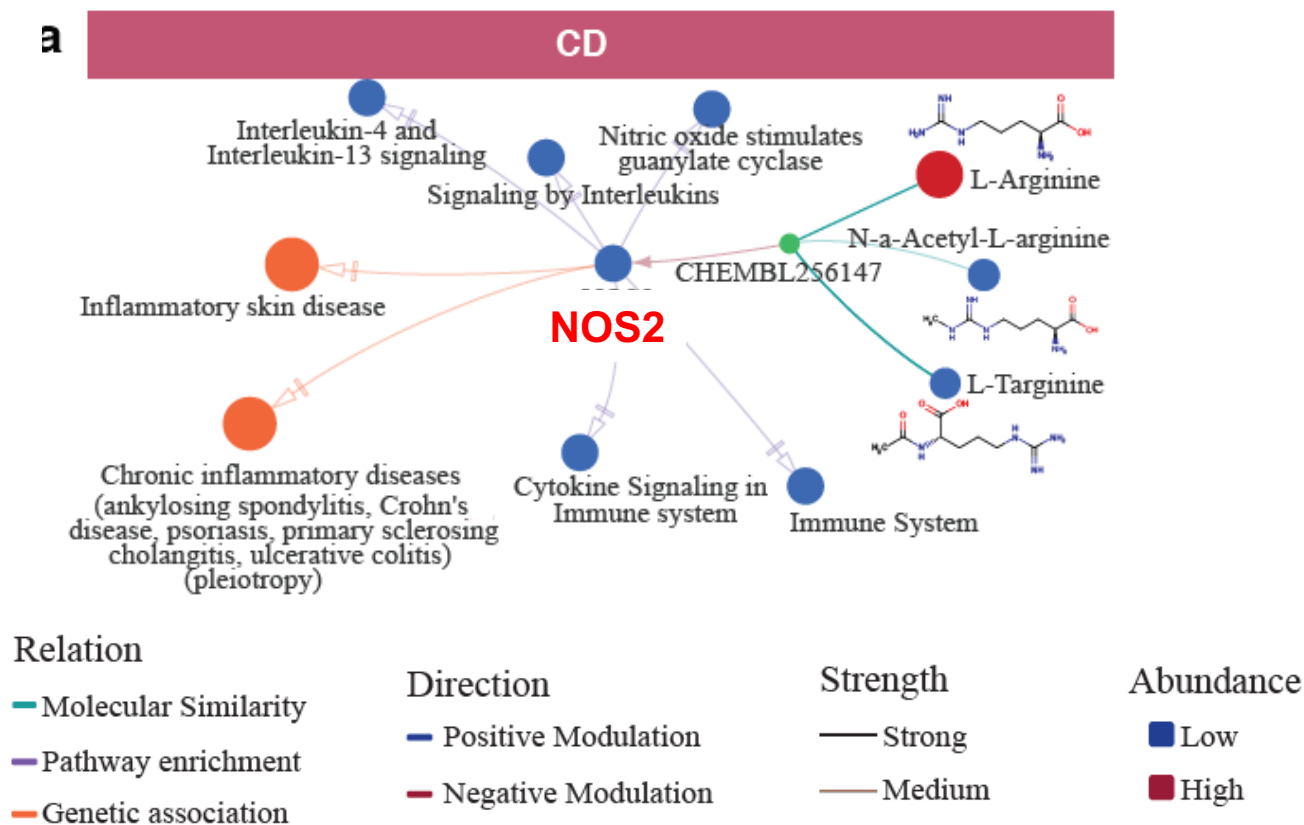
- Reversing transcriptomic disease signature using candidate modulators





# Linkages to Disease Genetics

- Metabolites passing thresholds and tractable targets with genetic evidence (GWAS and IBD-specific genetic studies)
- Retrieved 808 genes with genetic associations to IBD
- Identified 464 potential pairings between genetic targets with metabolite modulators, 13 with known modulation mechanisms



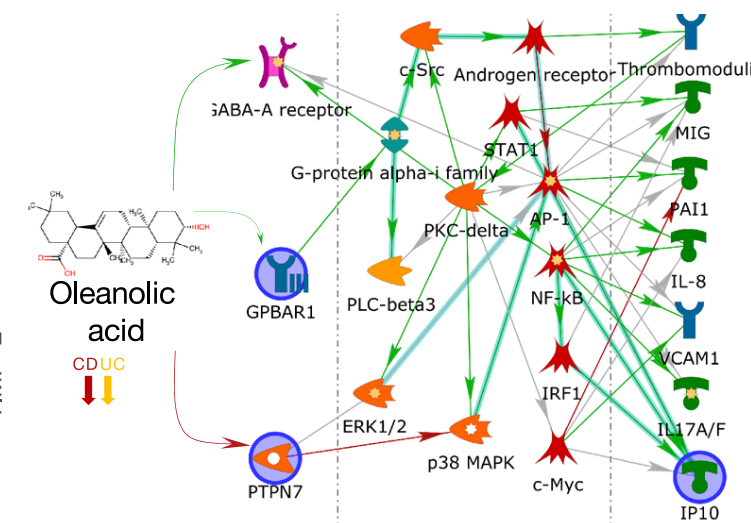
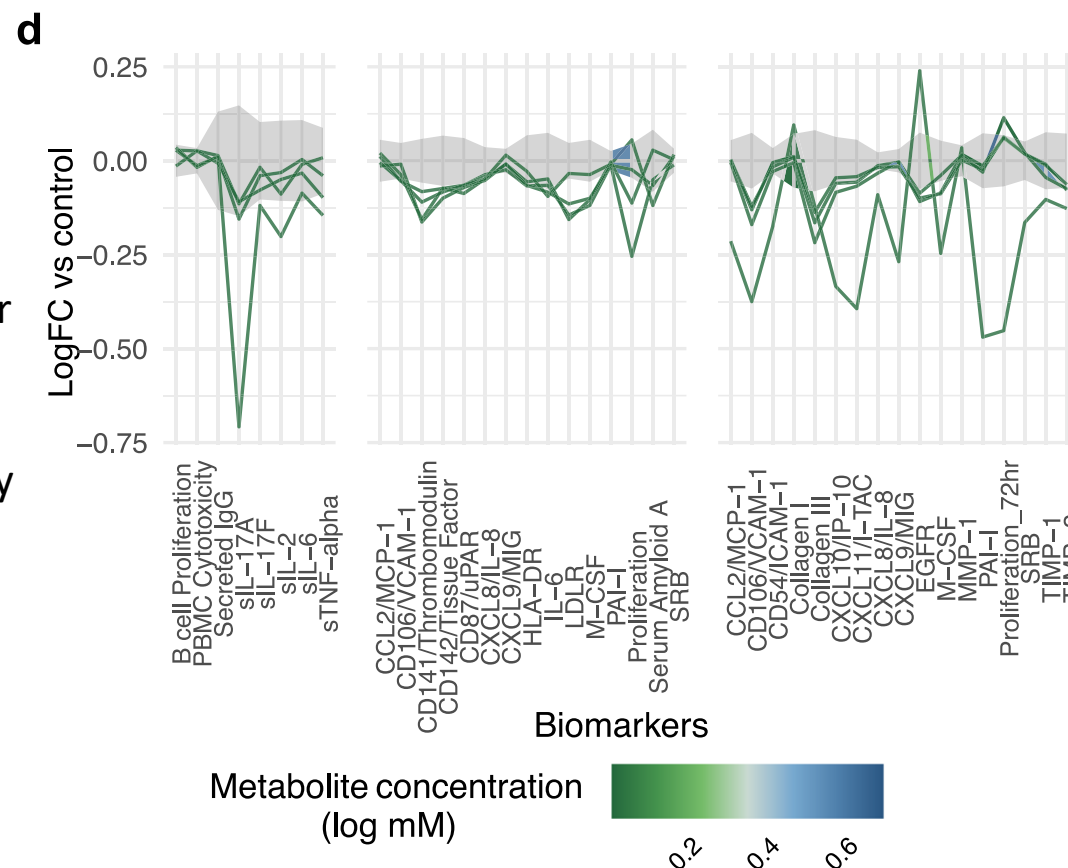


# in vitro Validation Assays for Selected Metabolites



Nuzzo...Brown. 2021. Commun. Biol. (Nature). 4:288

- Selected 11 metabolites for profiling in human primary cell-based phenotypic assays (BioMAP® Diversity PLUS panel)
- 8 metabolites showed significant modulation of immune biomarkers in one or more cellular systems.
- Summary
  - 135 metabolites provisionally connected to 80 different targets in IBD
  - 983 potential metabolite-target interactions identified
  - Immuno-modulating metabolites and targets are potential starting points for drug discovery
  - In vitro assays lend support to this approach



- Oleanolic acid (OA) showed activity in T-cell dependent B-cell activation (BT), coronary artery smooth muscle (CASM3C), fibroblasts (HDF3CGF) assays
- OA is a connected ligand of GABRG2, PTPN7 and GPBAR1



# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism databases
7. Cancer relevant databases
8. Concluding remarks & discussion



# Integrative vs Specialized Biomedical Databases and Interfaces

- **Integrative biomedical databases – consolidate multiple specialized databases:**

- NCBI\*: <https://www.ncbi.nlm.nih.gov/> (gene viewer)
- Open Targets\*: <https://www.opentargets.org/>
- ENSEMBL: <https://useast.ensembl.org/index.html>

\* Demos covered in this workshop

- **Specialty resources:**

- Human Genetics – GWAS Catalogue\*: <https://www.ebi.ac.uk/gwas/>
- Mouse phenotypes – Mouse Phenome Database (Jackson Lab): <https://phenome.jax.org/>
- RNA expression – GTEX\*: <https://gtexportal.org/home/> ; Single Expression Atlas: <https://www.ebi.ac.uk/gxa/sc/release-notes.html>
- Protein Atlas – protein expression\*: <https://www.proteinatlas.org/>
- Pathways – Reactome\*: <https://reactome.org/> ; WikiPathways: <https://www.wikipathways.org/> ; IntAct: <https://www.ebi.ac.uk/intact/home>
- Protein annotations – UniProt\*: <https://www.uniprot.org/>
- Protein-protein interactions – String\*: <https://string-db.org/>
- Metabolomics – The Human Metabolome Database\*: <https://hmdb.ca/>

- **Clinical trials and tool compounds:**

- Clinical trials -- ClinicalTrials.gov\*: <https://clinicaltrials.gov/>
- Drugs and targets – DrugBank\*: <https://go.drugbank.com/>
- Bioactive molecules and interactions: ChEMBL\*: <https://www.ebi.ac.uk/chembl/>

- **Cancer:**

- The Cancer Genome Atlas Program (TCGA)
- Integrative data-sources for cancer functional genomics – Xenabrowser\*: <https://xenabrowser.net/>
- Cancer dependency map -- DepMap\*: <https://depmap.org/portal/>
- Cancer cell lines – Cancer Cell Line Encyclopedia (CCLE) \*: <https://sites.broadinstitute.org/ccle/>



# NCBI Gene: General Gene Info

- NCBI “gene” is a good starting point: <https://www.ncbi.nlm.nih.gov/>
- NOS2 as an example: <https://www.ncbi.nlm.nih.gov/gene/4843> ← Gene ID #4843

NIH National Library of Medicine  
National Center for Biotechnology Information

Log in

Gene   [Advanced](#) [Help](#)

Full Report ▾ Send to: ▾ [Hide sidebar >>](#)

**NOS2** nitric oxide synthase 2 [ *Homo sapiens* (human) ]

Gene ID: 4843, updated on 22-Jan-2024

Summary

**Official Symbol** NOS2 provided by HGNC

**Official Full Name** nitric oxide synthase 2 provided by HGNC

**Primary source** HGNC:HGNC:7873

**See related** [Ensembl:ENSG00000007171](#) [MIM:163730](#); [AllianceGenome:HGNC:7873](#)

**Gene type** protein coding

**RefSeq status** REVIEWED

**Organism** [Homo sapiens](#)

**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

**Also known as** NOS; INOS; NOS2A; HEP-NOS

**Summary** Nitric oxide is a reactive free radical which acts as a biologic mediator in several processes, including neurotransmission and antimicrobial and antitumoral activities. This gene encodes a nitric oxide synthase which is expressed in liver and is inducible by a combination of lipopolysaccharide and certain cytokines. Three related pseudogenes are located within the Smith-Magenis syndrome region on chromosome 17. [provided by RefSeq, Jul 2008]

**Expression** Biased expression in small intestine (RPKM 10.3), appendix (RPKM 7.9) and 5 other tissues [See more](#)

**Orthologs** [mouse](#) [all](#)

NEW

[Try the new Gene table](#)

[Try the new Transcript table](#)

Gene name aliases

[Download Datasets](#)

**Table of contents**

[Summary](#)

[Genomic context](#)

[Genomic regions, transcripts, and products](#)

[Expression](#)

[Bibliography](#)

[Phenotypes](#)

[Variation](#)

[HIV-1 interactions](#)

[Pathways from PubChem](#)

[Interactions](#)

[General gene information](#)

[Markers, Related pseudogene\(s\), Homology, Gene Ontology](#)

[General protein information](#)

[NCBI Reference Sequences \(RefSeq\)](#)

[Related sequences](#)

[Additional links](#)



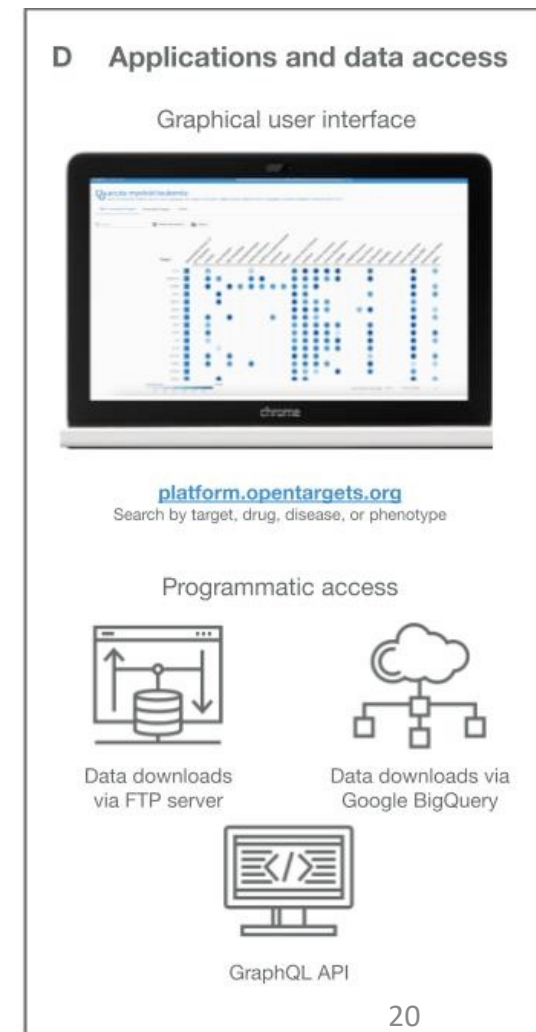
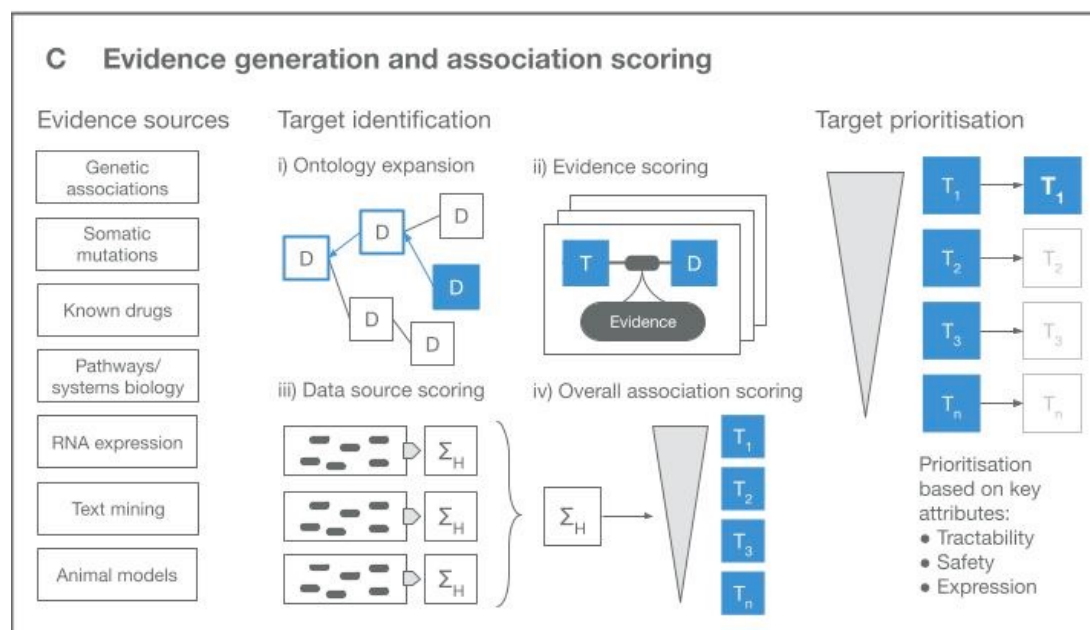
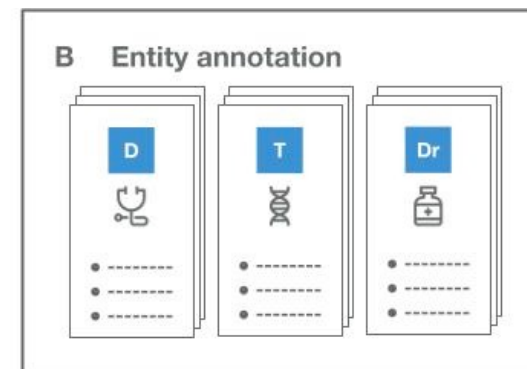
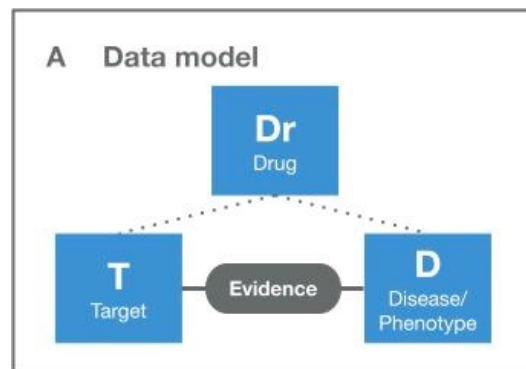
# Open Targets Platform

<https://platform.opentargets.org/>



Open Targets

- Identifying evidence implicating drug targets with diseases or phenotypes is a major challenge.
- Open Targets (OT) is a public-private partnership between the EMBL / EBI and several pharma companies.
- The OT Platform organizes public data-sources in order to enhance open-source target discovery and exploration.
- Four main entities in OT:
  - Data model
  - Entity annotation
  - Evidence and association scoring
  - Applications and data access



<https://platform-docs.opentargets.org/getting-started>





### A. The OT Data Model focuses on three main entities:

- A. Target understood as any candidate for drug binding molecule
- B. Disease or Phenotype including any disease indications, phenotypes, measurements, biological processes and other relevant traits.
- C. Drug molecules that can act as medicinal products.

### B. Entity annotations

- A. Target tractability assessment
- B. Target safety
- C. Baseline expression
- D. Molecular interactions
- E. Clinical signs and symptoms
- F. Pharmacovigilance
- G. Bibliography

### C. Evidence generation and association scoring

- A. Target-disease evidence
- B. Target-disease associations

### D. Applications and data access

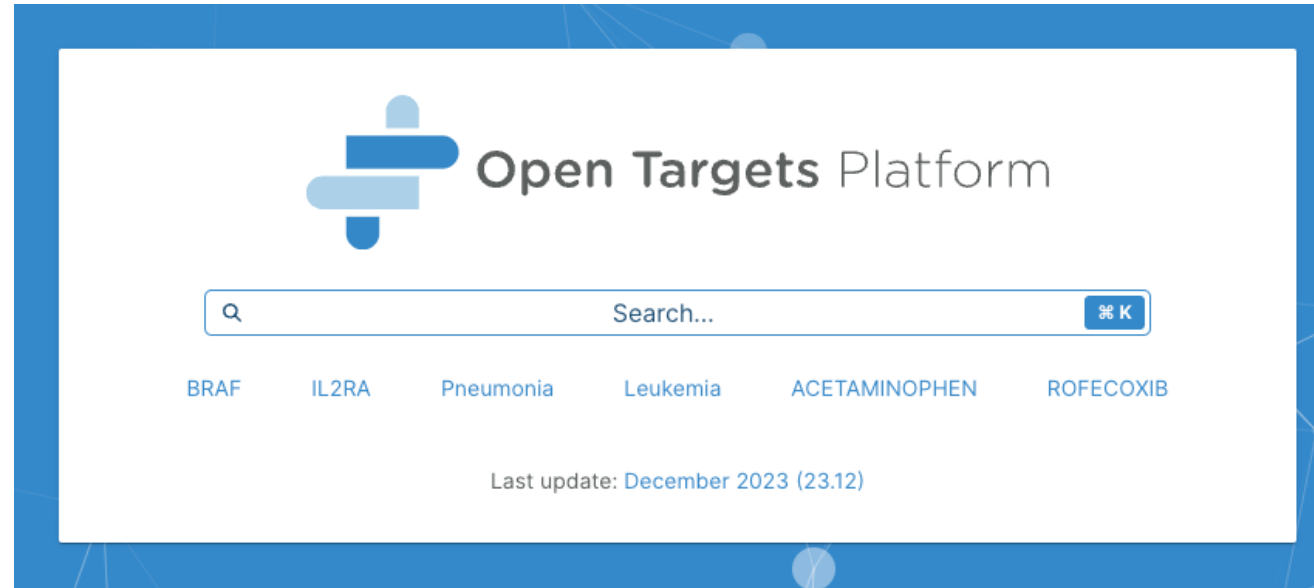
- A. Web interface
- B. Data access – programmatic



# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism databases
7. Cancer relevant databases
8. Concluding remarks & discussion





- Query entry for “gene”, “disease” or “drug”
- Caveat – OT is human **“non-communicable”** disease centric
  - *Infectious diseases and pathogen genomics are not represented*
- Example using gene “PDCD1” (alias PD1) which encodes “Programmed cell death protein 1”
  - One of the most successful targets for cancer immuno-therapies (i.e., Merck’s Pembrolizumab [Keytruda] )



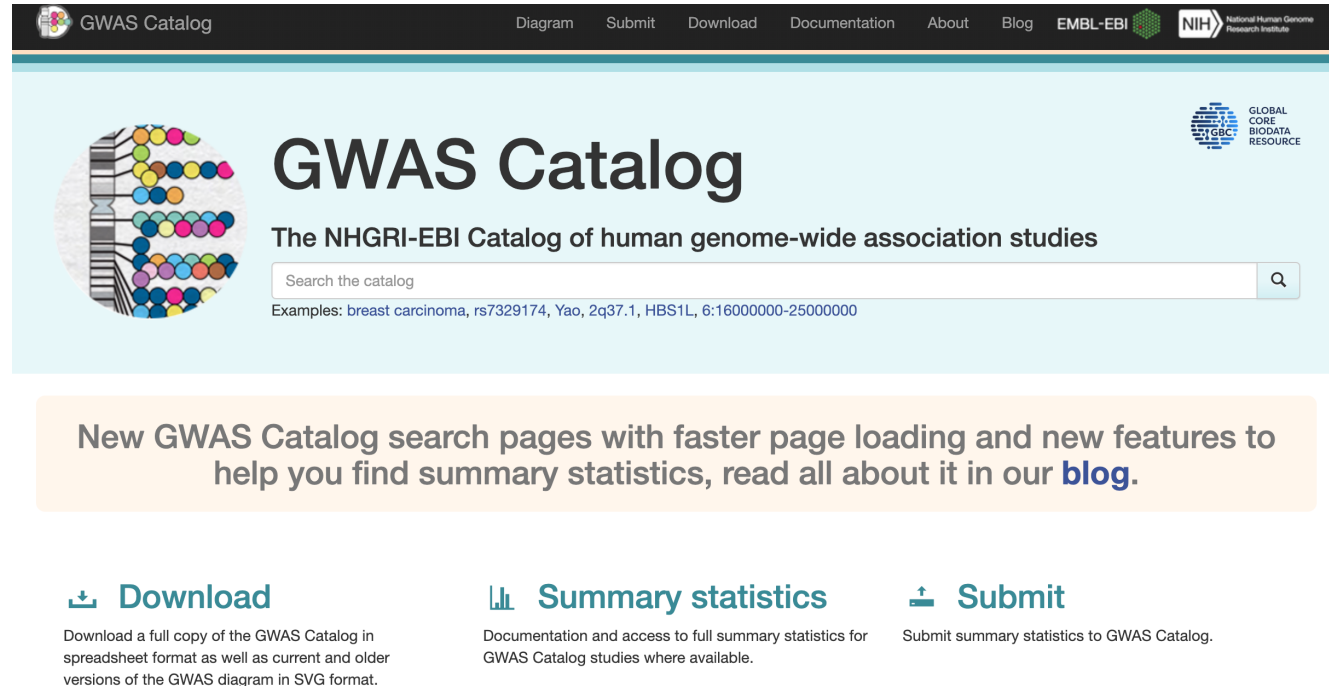


- Try and contrast the results of the 3 types of initial search queries:
  - Gene name: *(PDCD1)*  
<https://platform.opentargets.org/target/ENSG00000188389/associations>
  - Drug name: *(Pembrolizumab)*  
<https://platform.opentargets.org/drug/CHEMBL3137343>
  - Disease: *(melanoma)*  
[https://platform.opentargets.org/disease/EFO\\_0000756/classic-associations](https://platform.opentargets.org/disease/EFO_0000756/classic-associations)
- Evidence and association scoring – approximation to prioritize and sort evidence
- Note different sources associated with each query type.
- Take a deeper dive:
  - “Associated targets”: Use evidence specific filters
  - “Profile”: explore features for gene, drug and disease searches



# Human Genetics: GWAS Catalogue

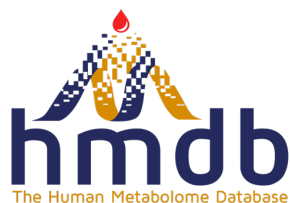
- Some studies suggest that targets with genetic evidence for disease have a two-fold greater probability for clinical success (Nelson et al.. 2015. Nature Genetics 47:856)
- GWAS Catalogue is a central repository of genome-wide association studies: <https://www.ebi.ac.uk/gwas/home>
- Flexible search queries include disease, SNP ID, study author, human chromosome localization, gene name and genomic coordinates
- Example using gene name “NOS2”:  
<https://www.ebi.ac.uk/gwas/search?query=NOS2>
- Available data:
  - **Associations** of DNA variants (variant and risk allele) with Traits
  - **Studies** behind the data
  - **Traits** summary



The screenshot shows the GWAS Catalog homepage. At the top is a navigation bar with links: Diagram, Submit, Download, Documentation, About, Blog, EMBL-EBI, and NIH. The main header features the GWAS Catalog logo (a circular diagram of chromosomes) and the text "GWAS Catalog" and "The NHGRI-EBI Catalog of human genome-wide association studies". Below this is a search bar with the placeholder "Search the catalog" and a search icon. Examples of search queries are provided: "breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000". A yellow banner below the search bar reads: "New GWAS Catalog search pages with faster page loading and new features to help you find summary statistics, read all about it in our [blog](#)." At the bottom, there are three sections: "Download" (with a download icon) with the text "Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.", "Summary statistics" (with a bar chart icon) with the text "Documentation and access to full summary statistics for GWAS Catalog studies where available.", and "Submit" (with an upload icon) with the text "Submit summary statistics to GWAS Catalog."



# Other Sources of Drug-Target Information



- Drug Bank <https://go.drugbank.com/>
  - Open source knowledgebase for 500,000+ drugs and drug products
  - Query searches for drugs, targets, pathways and indications
  - Large scale data downloads are free for academic research; surcharges for commercial use.
  - Example, gene: **PDCD1**
- The Human Metabolome Database <https://hmdb.ca/>
  - Open source knowledgebase for human metabolites and their interactions
  - Query searches for metabolites, diseases, proteins pathways and reactions indications
  - Example, metabolite: **L-tryptophan** ; gene: **IDO1**
- ChEMBL <https://www.ebi.ac.uk/chembl/>
  - Extensive and well-curated reference db for bioactive molecules
  - Query searches for drugs, genes, proteins, tissue, compound structure
  - Example, gene: **NOS2**, UniProt ID: **P35228**
- ClinicalTrials.gov <https://clinicaltrials.gov/>
  - Database of global clinical trials – targets with launched drugs are the most validated targets!
  - Clinical query terms. Not directly linked to gene name or id.
  - Example, gene: **EFGR**, **PD-1**



# Exercise 1: Human Centric Databases

- Try a few searches for one or more of the platforms
  - You can use the suggested example queries or try your own favorite genes, drugs, metabolites and/or disease:
1. Searching Opentargets: <https://platform.opentargets.org/>
    1. Gene name: (PDCD1)
    2. Drug name: (Pembrolizumab)
    3. Disease: (melanoma)
  2. Search GWAS Catalogue and contrast the number of coding variants for PDCD1 vs NOS2
  3. Search Drug Bank <https://go.drugbank.com/>
    1. Example, gene: PDCD1
  4. Search The Human Metabolome Database <https://hmdb.ca/>
    1. Example, metabolite: L-tryptophan ; gene: IDO1
  5. Search ChEMBL <https://www.ebi.ac.uk/chembl/>
    1. Example, gene: NOS2 UniProt ID: P35228
  6. Search ClinicalTrials.gov <https://clinicaltrials.gov/>
    1. Example, gene: EFGR



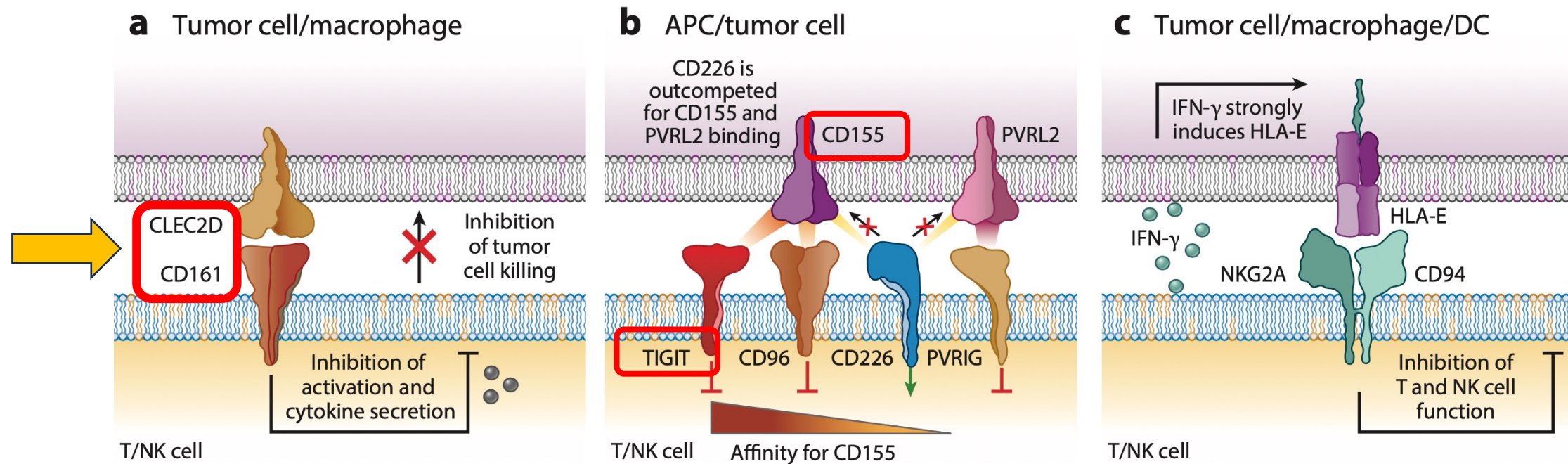
# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism databases
7. Cancer relevant databases
8. Concluding remarks & discussion



# Deep Dive Into a Target Hypothesis

- Example: Oncology immunotherapy (OI) is a highly active area for novel therapeutics.
- Modulation of T cell and natural killer (NK) responses by inhibiting any immune suppressor mechanisms of the tumor cell is an important strategy.
- Inhibition of CLEC2D and CD161 (KLRB1) interaction could re-activate T cell and NK cell killing of tumor cells.



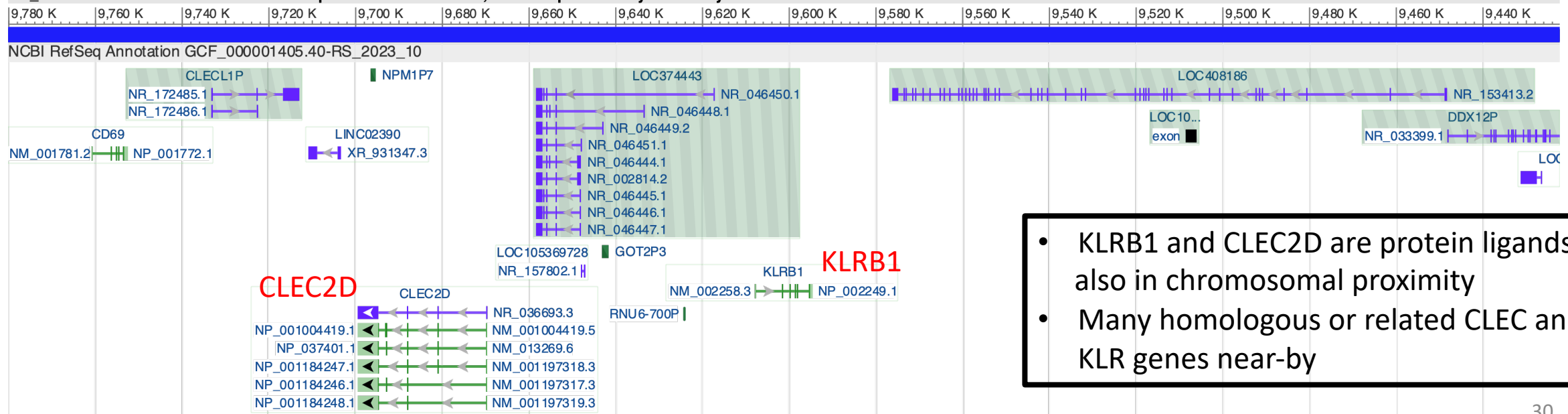
Inhibitory receptors shared by T cells and NK cells along with tumor cell receptor partners.



# CD161 (KLRB1) and CLEC2D Genomic Organization

- **CD161** now called **KLRB1** (killer cell lectin like receptor B1): <https://www.ncbi.nlm.nih.gov/gene/3820>
- Genomic regions, transcripts, and products
  - Genome browser allows visualization of customize “tracks” for mapping features onto the gene.
  - Defaults is gene-centric view with ClinVar variants; RNA-seq exon coverage, RNA-seq intron features
  - More options under “Tools” and “Tracks” .
  - Zoom in “+” to level of nucleotides or Zoom out “-” to exploring neighboring genes and features

NC\_000012.12:9422301..9780160 Homo sapiens chromosome 12, GRCh38.p14 Primary Assembly



- KLRB1 and CLEC2D are protein ligands, also in chromosomal proximity
- Many homologous or related CLEC and KLR genes near-by



# Gene Level Transcript Variants: Alternative Transcripts Which Might Encode Protein Isoforms

- Human CLEC2D encodes 5, possibly 6, mRNAs which results in potential protein isoforms with different AA lengths
- <https://www.ncbi.nlm.nih.gov/datasets/gene/id/29121/products/>

## Transcripts and Proteins

**CLEC2D – C-type lectin domain family 2 member D**

*Homo sapiens* (human)

Download ▾		Select columns							
<input checked="" type="checkbox"/>	Gene ID	Gene Symbol	Transcript	Length (nt)	Protein	Length (aa)	Protein name	Isoform	
<input checked="" type="checkbox"/>	29121	CLEC2D	NM_013269.6	5277	NP_037401.1	191	C-type lectin d...	1	
<input checked="" type="checkbox"/>	29121	CLEC2D	NM_001004419.5	5359	NP_001004419.1	194	C-type lectin d...	2	
<input checked="" type="checkbox"/>	29121	CLEC2D	NM_001197317.3	5166	NP_001184246.1	154	C-type lectin d...	3	
<input checked="" type="checkbox"/>	29121	CLEC2D	NM_001197318.3	5173	NP_001184247.1	132	C-type lectin d...	4	
<input checked="" type="checkbox"/>	29121	CLEC2D	NM_001197319.3	5062	NP_001184248.1	95	C-type lectin d...	5	
<input checked="" type="checkbox"/>	29121	CLEC2D	NR_036693.3	5248					




















# Open Targets Platform

<https://platform.opentargets.org/>



Open Targets

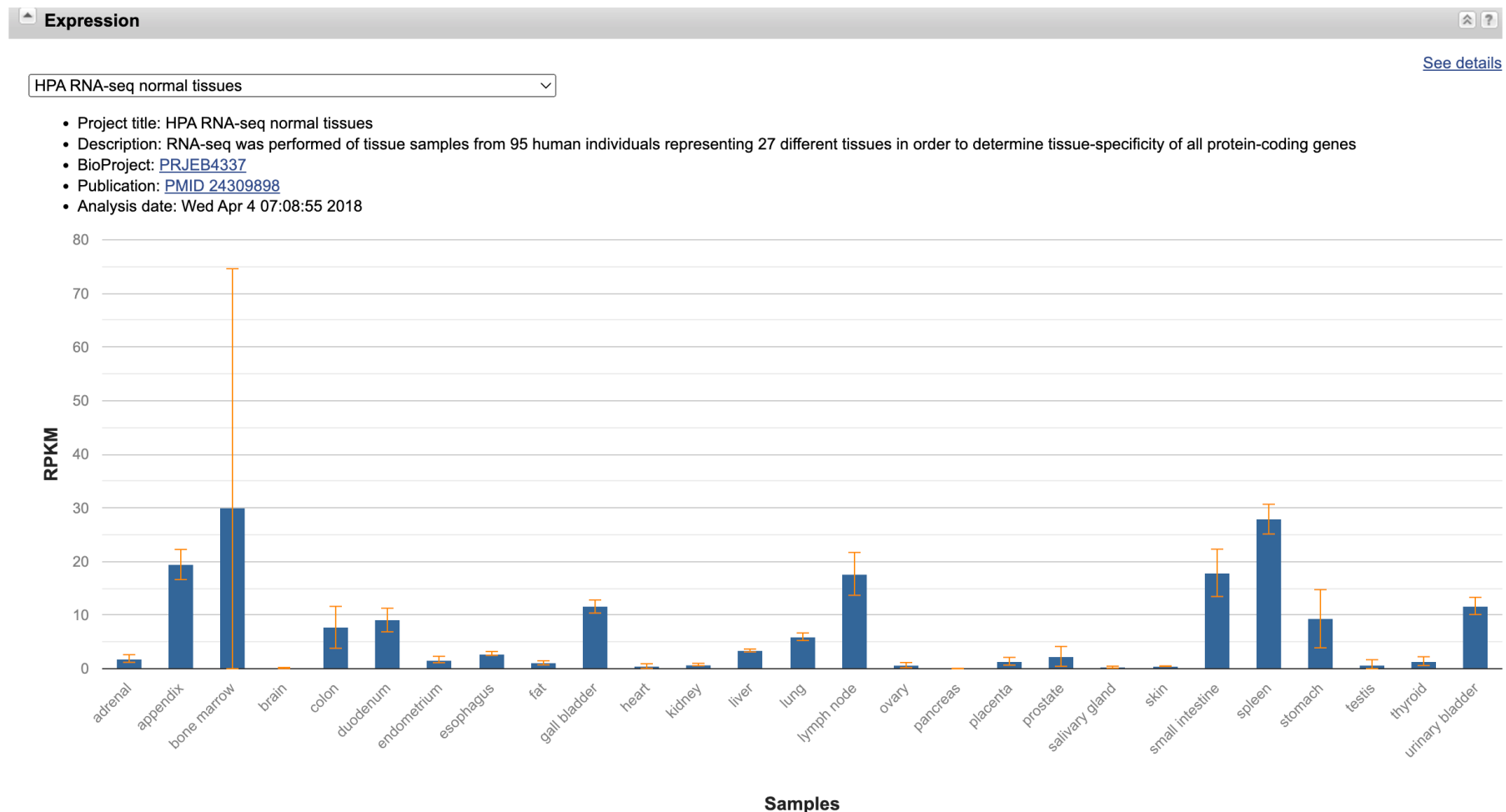
- Gene name: *(KLRB1)*  
<https://platform.opentargets.org/target/ENSG00000111796/classic-associations>
- Gene name: *(CLEC2D)*  
<https://platform.opentargets.org/target/ENSG00000069493/classic-associations>
- Associated Diseases Tab
- Profile Tab – Summaries compiled from multiple sources:

 Known Drugs	 Tractability	 Safety	 Pharmacogenetics	 Chemical Probes	 Baseline Expression
 Cancer DepMap	 Subcellular Location	 Gene Ontology	 Genetic Constraint	 ProtVista	 Molecular Interactions
 Pathways	 Cancer Hallmarks	 Mouse Phenotypes	 Comparative Genomics	 Bibliography	



# Tissue Specific Gene Expression

- KLRB1 killer cell lectin like receptor B1 [ Homo sapiens (human) ] <https://www.ncbi.nlm.nih.gov/gene/3820>
- RNA-seq data from tissue samples taken from 95 human individuals representing 27 different tissues





# Genotype-Tissue Expression (GTEx) Project



- The most definitive gene expression db is GTEx: <https://gtexportal.org/home/>
- (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation.
- Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq..
- Remaining samples are available from the GTEx Biobank upon request.
- The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.
- Can browser and search all data by:
  - Gene
  - Genetic variant
  - Tissue
  - GTEx histology images



# Bulk Tissue Gene Expression (mRNA): CLEC2D



<https://gtexportal.org/home/gene/CLEC2D> Filters: Subset = none; Scale = log; Tissue sort; Median sort. Outliers = on

## Bulk tissue gene expression for CLEC2D (ENSG00000069493.14)

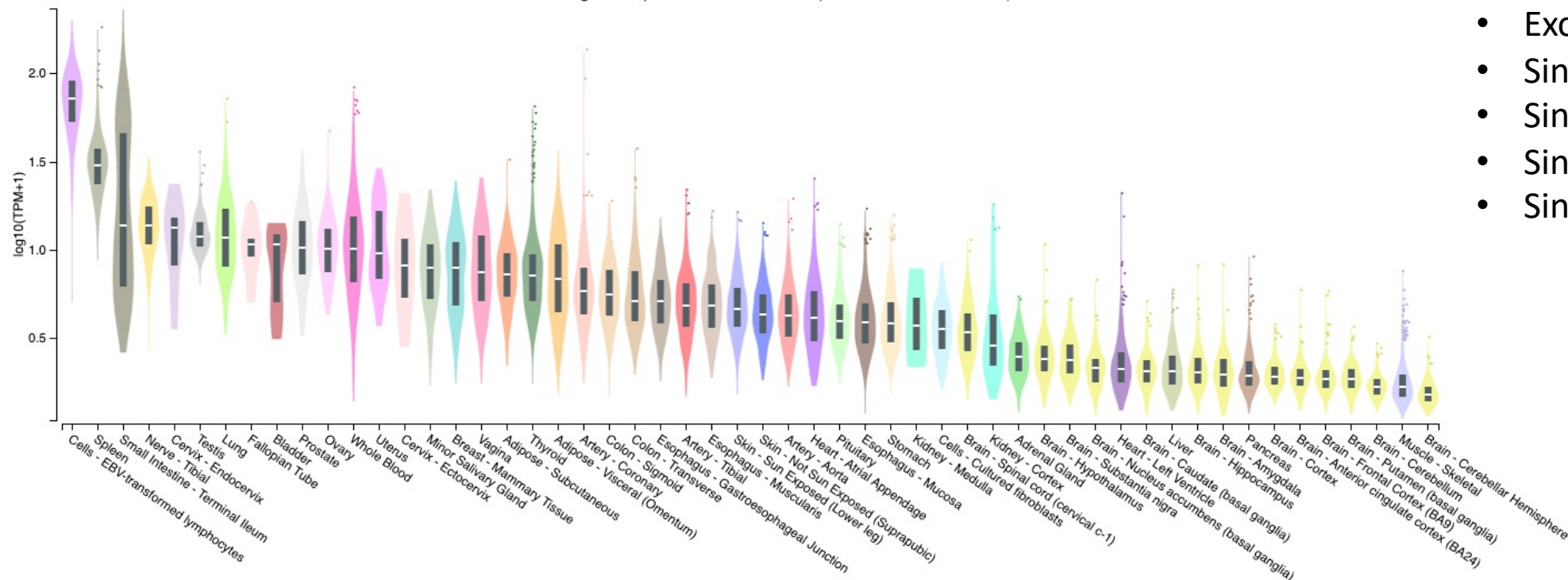
Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

Data processing and normalization ⓘ

Download the plot

SUBSET   SCALE   TISSUE SORT   MEDIAN SORT   OUTLIERS

Bulk tissue gene expression for CLEC2D (ENSG00000069493.14)



- Bulk Tissue Expression
- Single Cell Expression
- Exon Expression
- Single-Tissue eQTLs
- Single-Tissue sQTLs
- Single-Tissue ieQTLs
- Single-Tissue isQTLs

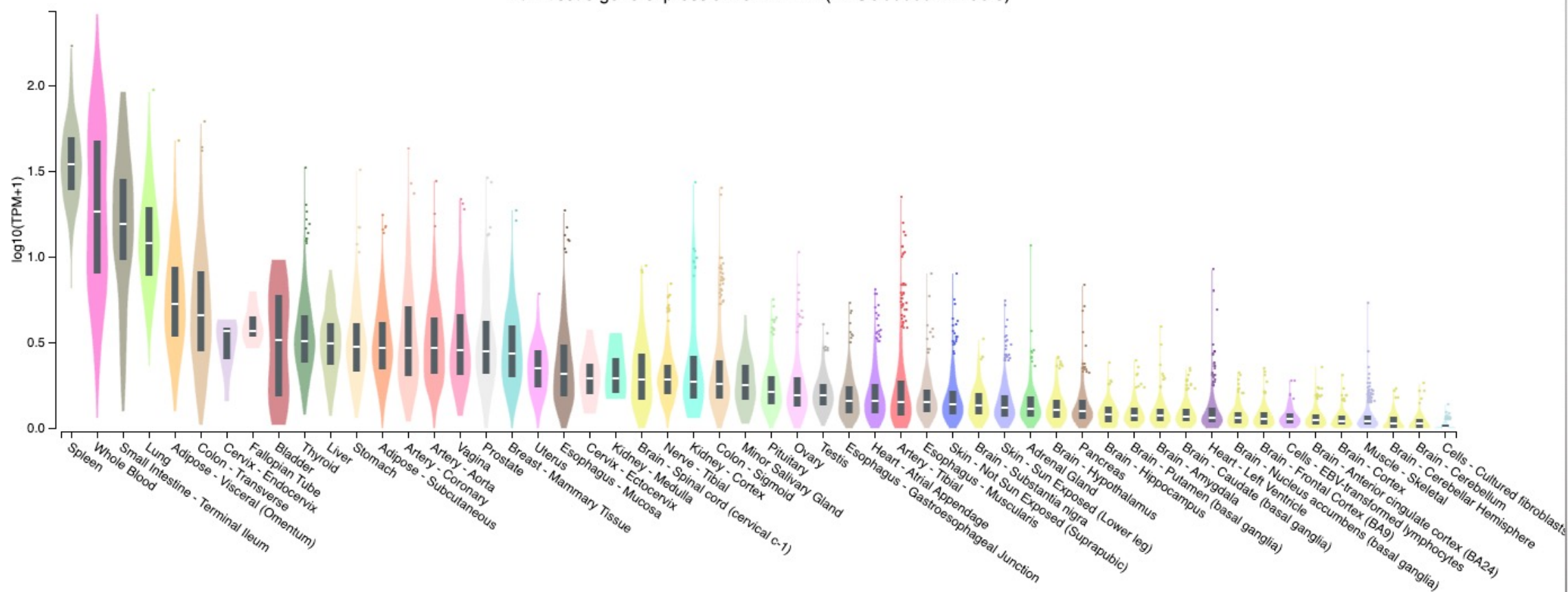


# Bulk Tissue Gene Expression (mRNA): KLRB1



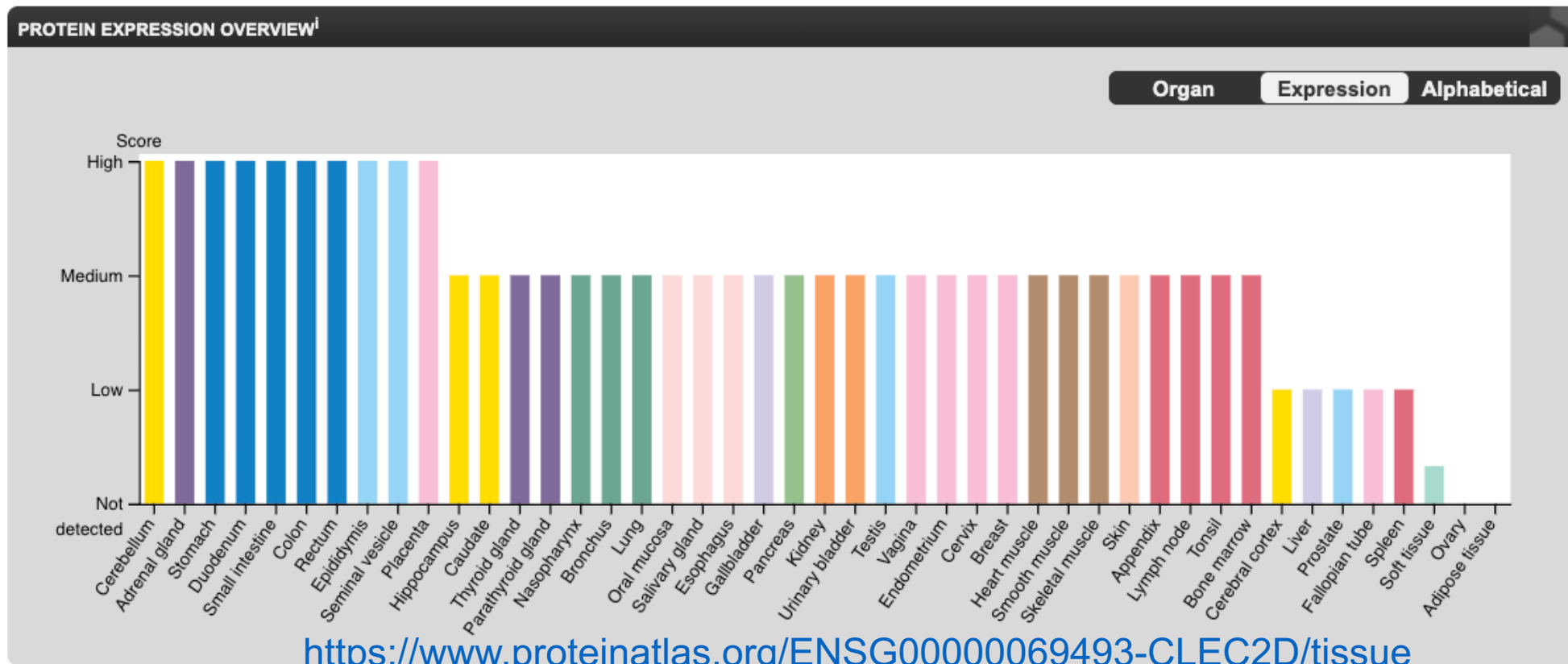
<https://gtexportal.org/home/gene/KLRB1> Filters: Subset = none; Scale = log; Tissue sort; Median sort. Outliers = on

Bulk tissue gene expression for KLRB1 (ENSG00000111796.3)





- The Human Protein Atlas (HPA) maps all the human proteins in cells, tissues, and organs using an integration of various 'omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology. <https://www.proteinatlas.org/>
- The HPA has 12 separate sections – gene names as the initial query.
- CLEC2D example: <https://www.proteinatlas.org/search/CLEC2D>





## Exercise 2: Multi-omics Databases

- Try a few searches for one or more of the ‘omics platforms
  - You can use the shown examples or try your own favorite genes:
1. For a given gene, look-up its genomic structure, gene expression and transcript variants using NCBI, “gene”: <https://www.ncbi.nlm.nih.gov/gene/>
  2. For the same gene, contrast the results in Open Targets: <https://platform.opentargets.org/>
  3. Reconstruct tissue specific expression using GTEx: <https://gtexportal.org/home/>
    1. Produce a figure where tissue expression levels are plotted by log values and ordered from high to low.
  4. Produce a plot of proteomics expression by organs / tissues using The Human Protein Atlas: <https://www.proteinatlas.org/>



# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism databases
7. Cancer relevant databases
8. Concluding remarks & discussion



- UniProt <https://www.uniprot.org/>
  - A comprehensive and well-curated resource for protein sequence and functional information
- Query entry point can be protein name, gene name, species, organism or protein ID
- Extensive entry information
- Customize columns allows to select and re-order data for downloading
- Multiple tools enabling sequence searches
  - BLAST
  - Align
  - Peptide Search
  - ID mapping
- Example, Genes: **KLRB1 & CLEC2D:**
- <https://www.uniprot.org/uniprotkb?query=KLRB1>
- <https://www.uniprot.org/uniprotkb?query=CLEC2D>



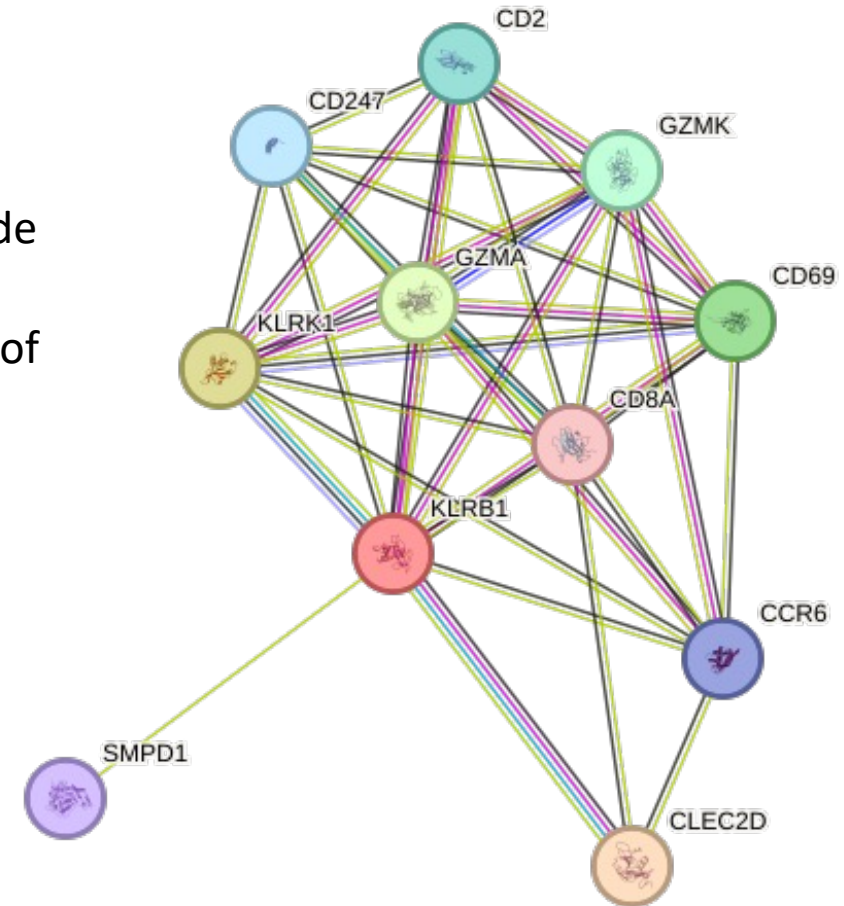
# Protein Interaction Databases



STRING

IntAct

- Receptor-ligand relationships, also called protein-protein interactions (PPIs), have a central role in all cellular functions.
- Dysfunctional PPIs are associated with many diseases – thus PPIs are potential therapeutic targets.
- Most integrative databases (i.e., NCBI, Open Targets & Uniprot) include PPIs evidence in their gene / protein annotations.
- These PPI annotations are pulled from one or more primary sources of curated or predicted PPIs such as:
- STRING-db <https://string-db.org/>
  - Protein-Protein Interaction Networks & Functional Enrichment Analysis
  - v12.0: 12535 organisms; 59.2 Mln proteins; >20 Bln Interactions
  - Query search by gene or protein name
  - Multi-data formats for downloading (graphical, tabular, etc.).
  - Example, gene: **KLRB1**
    - Select “Homo sapiens”
    - Cluster view -- Click-on “More”; “Legend”; “Viewers”; “Export”
- IntAct <https://www.ebi.ac.uk/intact/>
  - All interactions are derived from literature curation or direct user submissions
- Biogrid <https://thebiogrid.org/>
  - Includes Open Repository of CRISPR Screens (ORCS)





# Pathway Mapping and Analyses



- Reactome: <https://reactome.org/>
  - Open source, curated pathway database
  - Query by gene, protein, metabolite, pathway name or ID
  - Example, gene: **KLRB1**
- Pathway ontology
- Subcellular location
- Links to pathway map
- Also see:

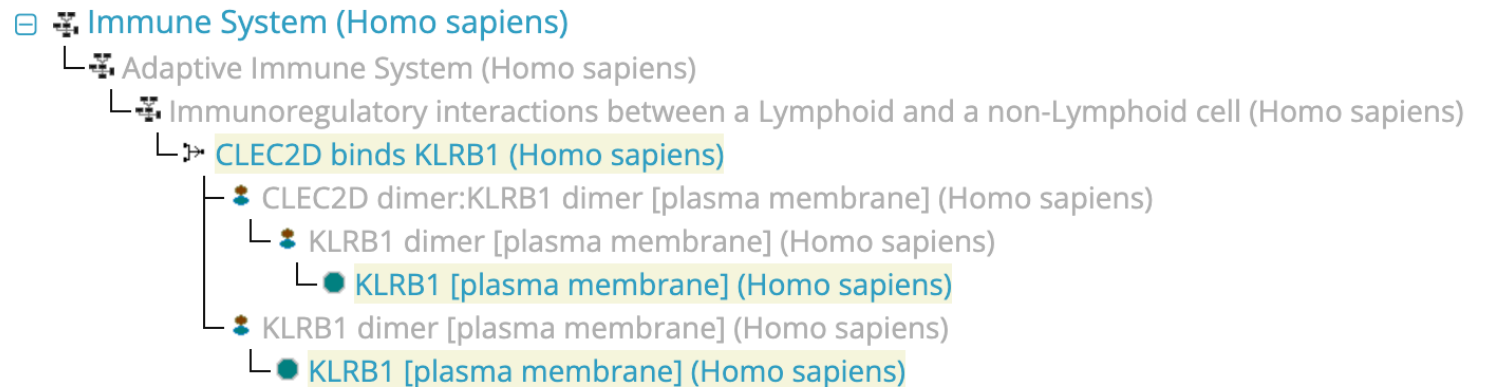
- KEGG Pathways:   
<https://www.genome.jp/kegg/>

- WikiPathways:  WIKIPATHWAYS  
<https://www.wikipathways.org/>

## ● KLRB1 [plasma membrane]

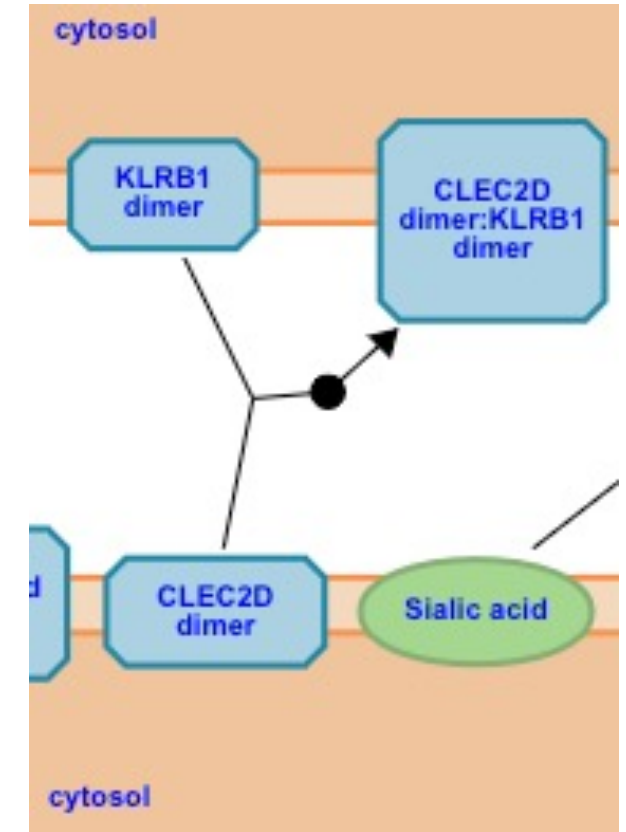
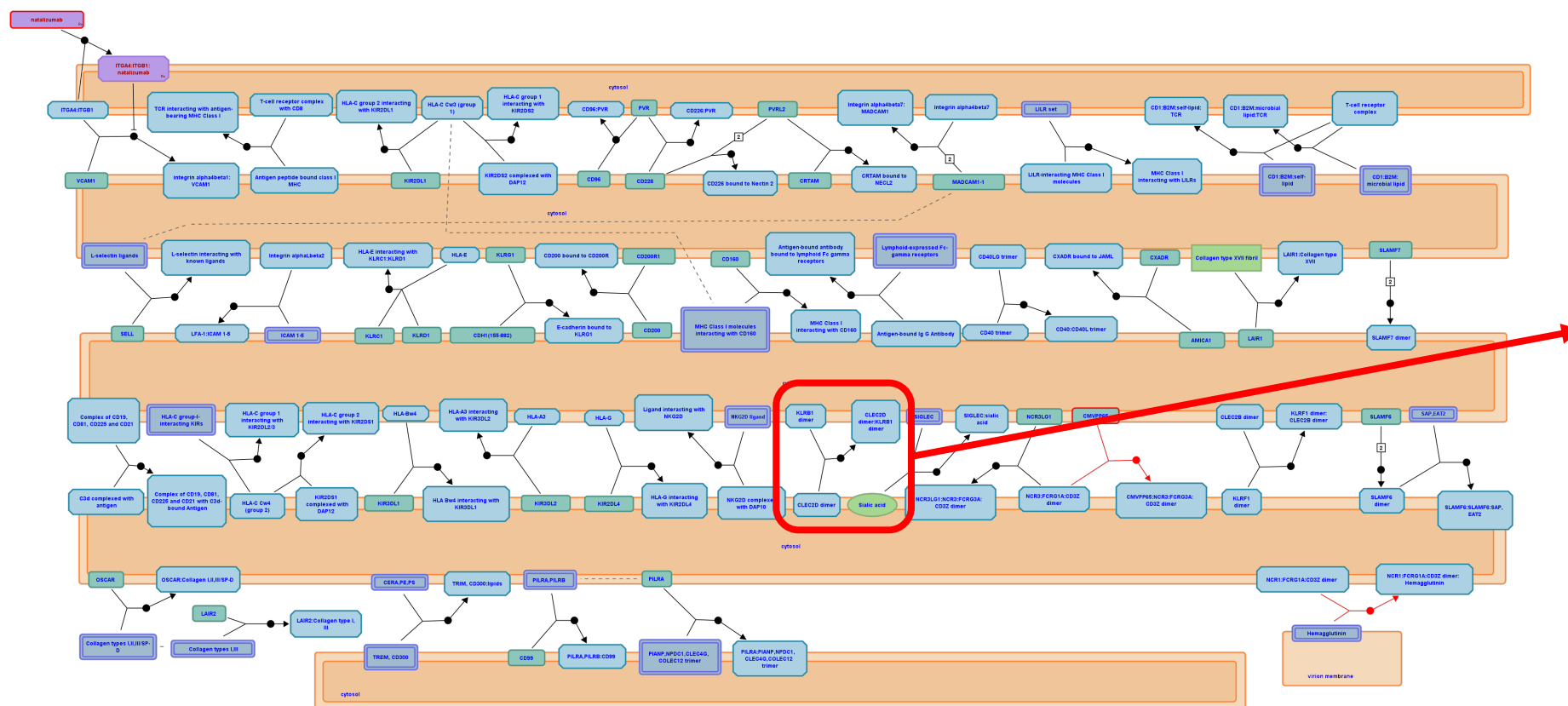
Stable Identifier	R-HSA-2132225
Type	Protein [EntityWithAccessionedSequence]
Species	Homo sapiens
Compartment	<a href="#">plasma membrane</a>
Synonyms	Killer cell lectin-like receptor subfamily B member 1, KLRB1_HUMAN

### Locations in the PathwayBrowser





# Pathway Mapping and Analyses



<https://reactome.org/PathwayBrowser/#/R-HSA-198933>



# Exercise 3: Protein Characterization and Interactions Databases

- Try a few searches for one or more of the protein-focused platforms
  - You can use the shown examples or try your own favorite genes:
1. For a given protein, look-up its record and select about 4 or 5 different records (could be the homologs in different species) using UniProt: <https://www.uniprot.org/>
    1. Create a customized table of several features then download the CSV formatted file (readable by MS Excel)
  2. Create a simple protein interaction network using STRING-db: <https://string-db.org/>
  3. For a given gene, look-up its associated pathways in Reactome: <https://reactome.org/>
    1. If you have time, compare differences in output using WikiPathways (<https://www.wikipathways.org/>), IntAct (<https://www.ebi.ac.uk/intact/>) and/or Biogrid (<https://thebiogrid.org/>)



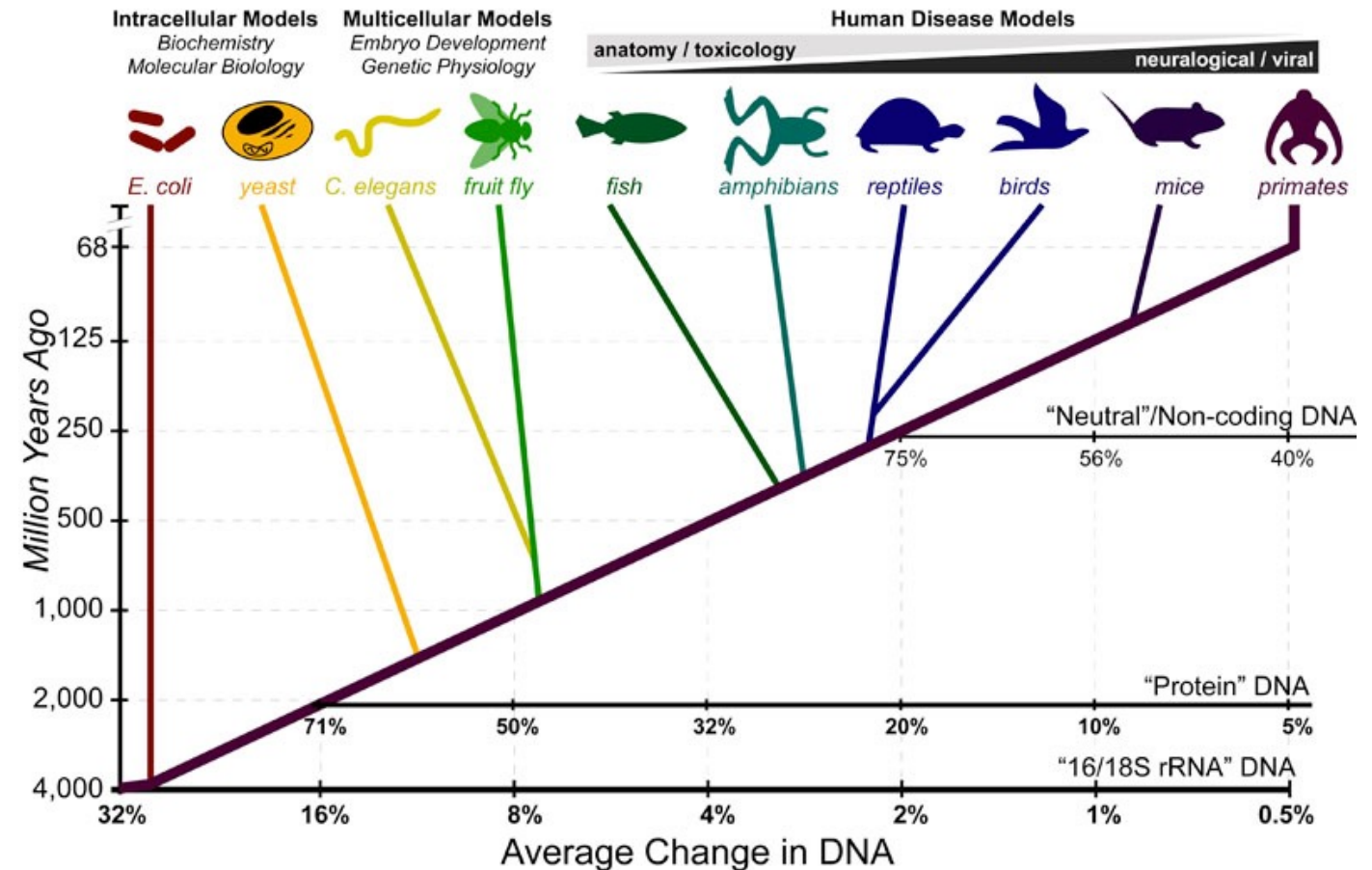
# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism databases
7. Cancer relevant databases
8. Concluding remarks & discussion



# Comparative Genomics and Phylogenomics

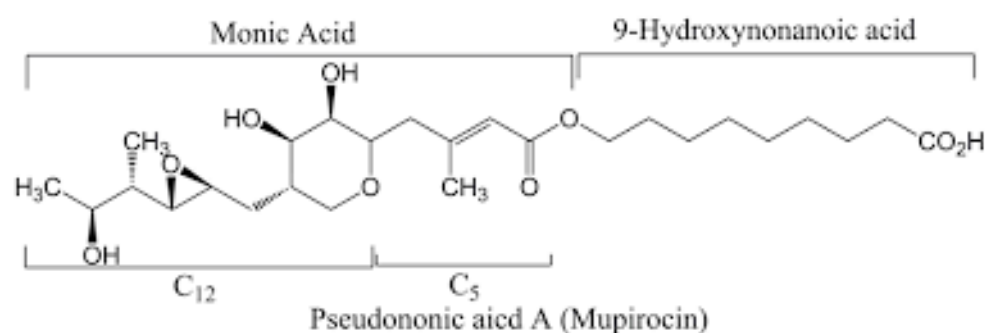
- Key questions in drug discovery which might be addressable by comparative genomics
- Inter-species homology: How similar is the targeted protein between human and model organism species?
  - Disease translation from preclinical in vivo model organisms through to humans
  - Interpretation of drug efficacy and safety in sentinel species (mouse, rat, dog, NHPs)
  - For infectious diseases, evaluate target variation across highly mutable pathogens
- Intra-species homology: Within the human genome, do any other proteins have significant sequence similarity to the target protein?
  - Design of counter screens to reduce off-target effects and increase drug targeting specificity
  - Identify potential target and pathway redundancies which might impact drug efficacy



<https://www.practicallyscience.com/model-organisms-and-dnas-molecular-clock/>



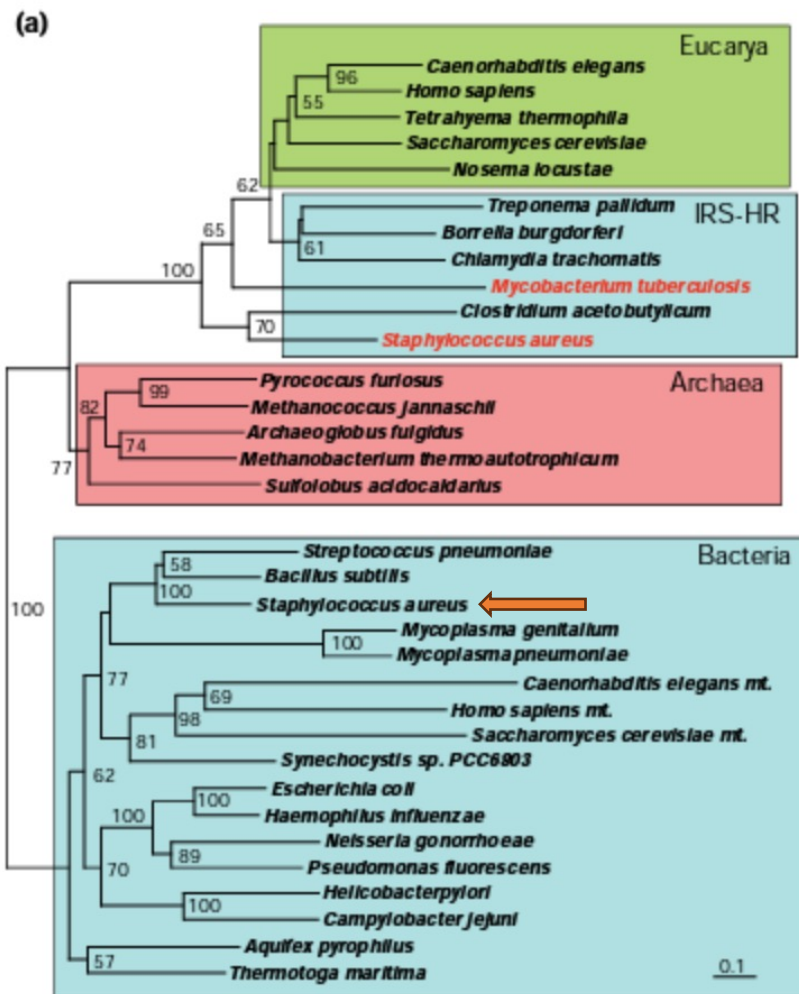
# The Quest for Novel Antibiotics



- In the late 1990's SmithKlineBeecham (later GSK) launched a genomics-based approach for the discovery of novel targets for antibiotics
- Bactroban is a highly successful topical antibiotic
- The compound pseudomonic acid (Mupriocin™) is a specific inhibitor of bacterial isoleucyl-tRNA synthetase (IleRs), one of 20 amino-acyl tRNA synthetases (AA-tRS).
- In late 1990's, GSK(fSB) had a new initiative focused on developing novel inhibitors of other AAtRSs for oral and/or IV delivered antibacterials.



# Trans-Domain Horizontal Gene Transfer (HGT)

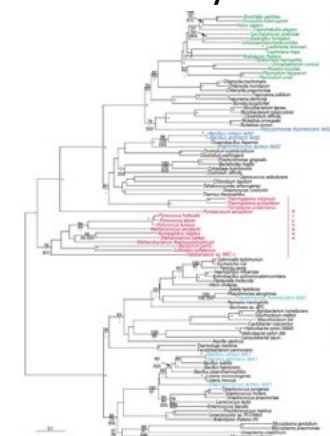


<b>Bacteria</b>	
<i>T. maritima</i>	EKMLDTLDVWIDSCASFEYIT-----TKREDHPFPLDMYLEGSDQHRG
<i>A. pyrophilus</i>	RKEEDILDVWFDSGCSHASV-----IRPLGFEKADLYLEGSQDHRG
<i>C. jejuni</i>	EKVYDILDVWFDSGCTSTNAVL-----NSCLYDACEKRASMYLEGSQDHRG
<i>H. pylori</i>	EKIMHILDVWFDSGCTFFKAVL-----EDYHCEKQSPSDVILEGSQDHRG
<i>M. genitalium</i>	HKEIDTLDVWFDSGSSYNVL-----EINKYCSIADLYLEGSQDHRG
<i>M. pneumoniae</i>	KKETDITLVWFDSGCTSYNVL-----ISNKLNFADLYLEGSQDHRG
<i>Synechocystis</i>	RKGEDTMDVWFDSGSSWAAVA-----NAKNRPLKYPVDMYLEGSQDHRG
<i>P. fluorescens</i>	DKISDTLDVWFDSGCTTHHVL-----RCSHPMGHETGPRADLYLEGSQDHRG
<i>S. cerevisiae</i> mt.	CRSQDTMDVWFDSGSSWSVIKDFYEKSLKSLKPLSPYQVCLGSDQHRG
<i>C. elegans</i> mt.	EKNTDINDVWLDSCLAHHAAR-----DNDTEREHVADVLEGVQDHRG
<i>H. sapiens</i> mt.	VPCQDILDVWFDSGTSWSYV-----LPCPDQADLYLECKDQLCG
<i>H. influenzae</i>	RKVPTDLDVWFDSGCTSYSSV-----ANRLEFNCQDIDMYLEGSQDHRG
<i>E. coli</i>	VKVPDITLDVWFDSGCTHSSVV-----DVRPEFACHAADMYLEGSQDHRG
<i>N. gonorrhoeae</i>	DKLPDITMDVWFDSGCTHSSV-----KQREELWPAADLYLEGSQDHRG
<i>B. subtilis</i>	TKEQDINDVWFDSGSSHQAVL-----EERDDLVRPADLYLEGSQDHRG
<i>S. pneumoniae</i>	KKETDINDVWFDSGSSWNGV-----VNRPELTYPADLYLEGSQDHRG
<i>S. aureus</i>	TKETDINDVWFDSGSSHRCVL-----ETRPELSFPADMYLEGSQDHRG
<b>IRS-HR</b>	
<i>S. aureus</i>	SRVEEVIDVWFDSGSMPPFAQHHPYFD-NQKIFNQHFPAADFAEGVDQTRG
<i>C. acetobutylicum</i>	TRTEEVIDVWFDSGSMPPFAQLHPYFE-NKEVFENTFPAQFISEAVDQTRG
<i>M. tuberculosis</i>	RRIPDVLVWFDSGSMPPYAQVHPYFE-NLDWFQGHYPGDFIVEYIQDTRG
<i>T. pallidum</i>	RRVPEVLDCWFESGAMPYAQVHPYFE-HATDFERYFAHFISEGLDQTRG
<i>B. burgdorferi</i>	IRTSSEVLDCWFESGAMPYASNHPYFT-NEINFKNIFPADFAEGLDQTRG
<i>C. trachomatis</i>	RRIPYVFDVWFDSGAMPYAQVHPYFE-RAEETACFPADFAEGLDQTRG
<b>Eucarya</b>	
<i>S. cerevisiae</i>	KRIEEVFDVWFESGSMPPYASQHPYFE-NTEKFDERVPANFISEGLDQTRG
<i>T. thermophila</i>	RRIDEVFDVWFESGSMPPYQGHYPFSMNEEEFSKRFPAADFAEGIDQTRG
<i>H. sapiens</i>	HRISEVFDVWFESGSMPPYAQVHPYFE-NKREFEDAFPAADFAEGIDQTRG
<i>C. elegans</i>	KRVSEVFDVWFESGSMPPYAQVHPYFE-NRKIFEDNFPADFAEGIDQTRG
<i>N. locustae</i>	RRIEEVFDVWFESGSMPPYAQVHPYFE---CDNLCLPADFAEGVDQTRG
<b>Archaea</b>	
<i>M. thermoautotrophicum</i>	KRTPDVLVWIDVSCVAGWAALHYPRE--KELFSEWFFYDFITEGHQDTRG
<i>M. jannaschii</i>	KRVPDVLVWFDSGLAPYASIGV-----KELKADFIIEGHQDQVTK
<i>P. furiosus</i>	RRVKDVVDVWFDSGLASWASLCYPR--NKELFEKLWPAADFAEGEDQVTK
<i>A. fulgidus</i>	RRVPDVLVWFDSGVASWCSIAAYPL--RKDKFEELWPAADFAEGEDQVTK
<i>S. acidocaldarius</i>	RRISDVADVWFDSGVAFFASLCQDW--RKR--WSELGPVLDVLEGHQDQTRG
	518 566

Current Biology

- Genomes of certain key pathogens (i.e., Staph, anthrax) harbor two copies of IleRS
  - A bacterial-like IleRS – mupirocin-sensitive
  - An eukaryote-like IleRS – mupirocin-high resistant (IRS-HR)
- IRS-HR loci was not evident in available published Staph genomes but found in > 30% of clinical isolates used by GSK.

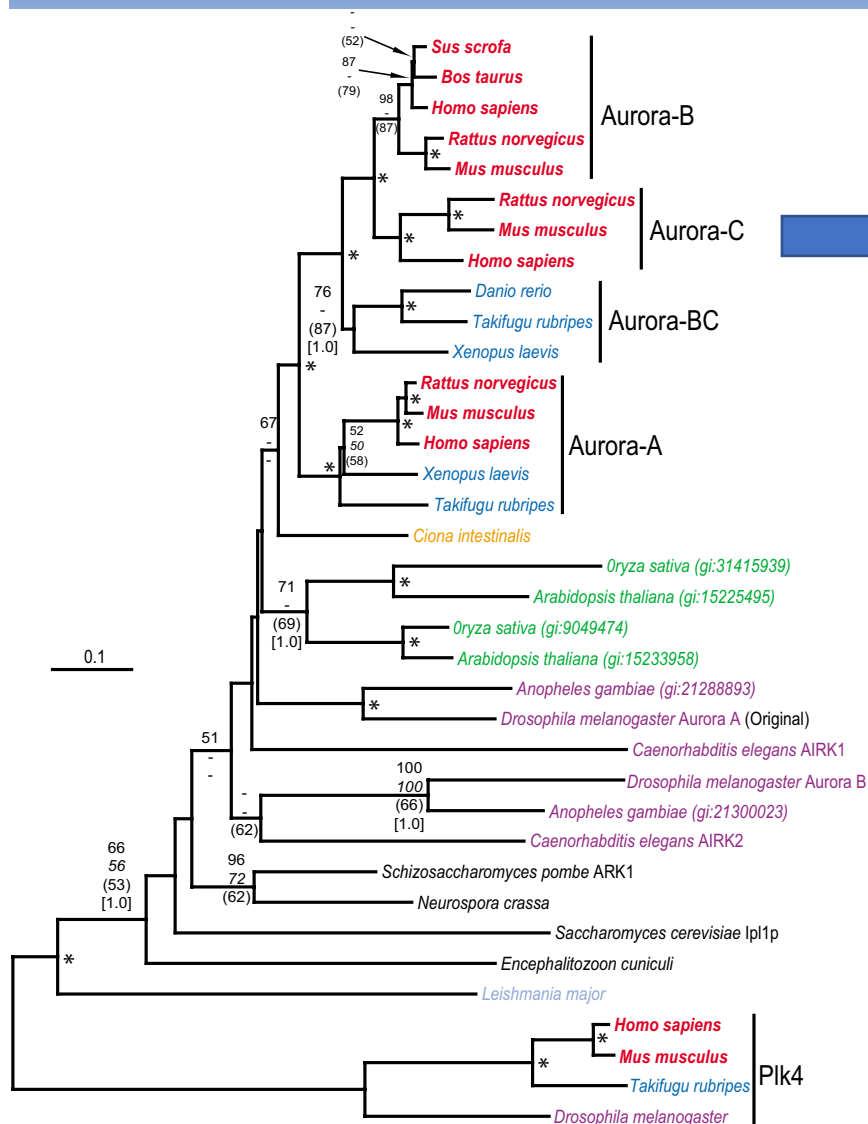
- Subsequently, similar HGT events found for AAtRSs (Brown et al. 2003. EMBO Reports. 4:692)



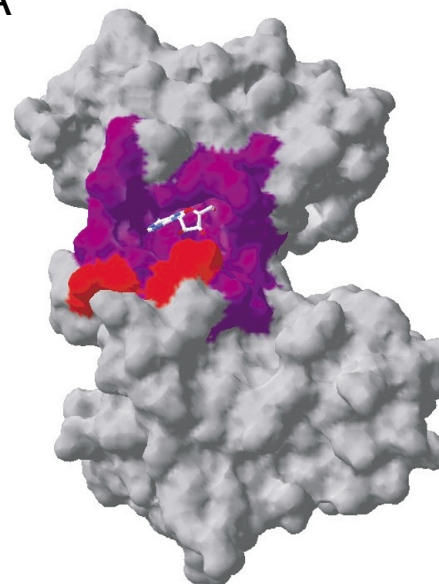
- In 2022, GSK announced positive results for Ph2a trial of GSK30336656 an inhibitor of *Mycobacterium tuberculosis* LeuRS.



# Targeting Aurora Kinases in Cancer: Evolutionary Factors



A



B

AuroraA	FEIGRPLGKGFNVYLAREKQSKFIALKVLFKAQLEKAGV	174
AuroraB	FEIGRPLGKGFNVYLAREKSHFIVALKVLFKSQIEKEGV	118
AuroraC	FEIGRPLGKGFNVYLARKESHFIVALKVLFKSQIEKEGL	84
	* * * * *	
AuroraA	EHQLRREVEIQSHLRHPNILLRLYGYFHDATRVYLILEYAPLG	216
AuroraB	EHQLRREVEIQSHLRHPNILLRLYNYFYDRRIYLIILEYAPRG	160
AuroraC	EHQLRREVEIQSHLRHPNILLRLYNYFHDARRVYLILEYAPRG	126
	* * * * *	
AuroraA	TVYRELQKLSKFDEQRTANLYNRIANALSYCHSKRVIHRDIK	258
AuroraB	ELYKELQKSCTFDEQRTATIMEELADALMYCHGKKVIHRDIK	202
AuroraC	ELYKELQKSEKLEQRTATIEEVADALTYCHDKKVIHRDIK	168
	* * * * *	
AuroraA	PENLLGSAGELKIADFGWSVHAPSSRRITLCGLDYLPPPM	300
AuroraB	PENLLGLKGLKIADFGWSVHAPSLRRTKTCGLDYLPPPM	244
AuroraC	PENLLGFRGEVKIADFGWSVHTPLPERKTCGLDYLPPPM	210
	* * * * *	
AuroraA	IEGRMHDEKVDLWSLGVLCYEFVLGKPPPEANTYQETYKRIS	342
AuroraB	IEGRMHNEKVDLWLCIGVLCYELLVGNPPFESASHNETYRRIV	286
AuroraC	IEGRTYDEKVDLWLCIGVLCYELLVGYPFESASHSETYRRIL	252
	* * * * *	
AuroraA	RVEFTFPDFVTEGARDLISRLKHNPSQRPMLREVLEHPW	382
AuroraB	KVDLFKFPASVPTGAQDLISKLLRHNPSERLPLAQVSAHPW	326
AuroraC	KVDVRFPLSMPLGARDLISRLRYQPLERLPLAQILKHPW	292

- Aurora Ser/Thr kinases are key regulators of mitotic chromosome segregation.
- Targeting specific Aurora family members (Aurora A, B or C) is a goal for cancer therapies.
- Phylogenomic analyses suggests Aurora-A occurs throughout eukaryotes while Aurora B and C evolved via two gene duplications, first in vertebrates and, second in mammals.
- Structurally, the druggable ATP-binding domain of Aurora A differs by only 3 amino acids to those of of Aurora B and C, which have identical domains to each other.
- Complicates the development of specific ATP inhibitors for each Aurora family but supports inhibition AurA alone or AurB plus AurC.



# Inter-Species Orthologs and Paralogs

**OrthoDB** v11

- OrthoDB – Hierarchical catalog of orthologs: <https://www.orthodb.org/>
- KLRB1 Orthologs: <https://www.orthodb.org/?ncbi=3820> (originally found in NCBI Gene record: <https://www.ncbi.nlm.nih.gov/gene/3820> )
- CLEC2D Orthologs: <https://www.orthodb.org/?ncbi=29121>
- Potential impact on in vivo translational studies: Four homologs in rodents vs only one in human and other primates.
  - Explore mouse phenotypes associated with each gene for an indication of functional similarity
- Best practice is to confirm using BLAST > Multiple sequence alignments (MSAs) > phylogenetic analyses.

## Homo sapiens (Human)

[KLRB1](#) ([Q12918](#)) Killer cell lectin-like receptor subfamily B member 1 » 225 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

### Orthologs in example species

### Get Ortholog Groups

Pan troglodytes - group [13709at9604](#) at Hominidae level

[KLRB1](#) ([A0A6D2XVA3](#)) A0A6D2XVA3\_PANTR » 225 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

Macaca mulatta (rhesus macaque;rhesus macaques;rhesus monkeys) - group [89220at9443](#) at Primates level

[KLRB1](#) ([A0A1D5R310](#)) C-type lectin domain-containing protein » 227 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

Rattus norvegicus (brown rat;rat;rats) - group [183878at314146](#) at Euarchontoglires level

1 [Klr1b1b](#) ([A4KWA1](#)) Killer cell lectin-like receptor subfamily B member 1B allele A » 223 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

2 [Klr1b1f](#) ([Q63378](#)) Killer cell lectin-like receptor subfamily B member 1F » 217 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

3 [Klr1b1](#) ([Q0ZUP0](#)) Killer cell lectin-like receptor subfamily B member 1 » 214 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

4 [Klr1b1a](#) ([B7TYL0](#)) Klr1b1a » 223 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

Mus musculus (mouse) - group [183878at314146](#) at Euarchontoglires level

1 [Klr1b1](#) ([A0A1U9W1A8](#)) Klr1b1 » 243 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

2 [Klr1b1f](#) ([I3QI43](#)) Klr1b1f » 217 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

3 [Klr1b1c](#) ([E9Q3U6](#)) Klr1b1c » 271 Q [IPR016187](#) [16186](#) [01304](#) [33992](#)

4 [Klr1b1a](#) ([B7ZN67](#)) Klr1b1a » 233 Q [IPR016187](#) [16186](#) [33992](#) [01304](#)

5 [Klr1b1b](#) ([A0A1U9W1A4](#)) Klr1b1b » 223 Q [IPR016187](#) [16186](#) [33992](#) [01304](#)



Mouse Gene Phenotypes  
<https://phenome.jax.org/>



# Exercise 4: Inter-Species Orthologs and Paralogs

1. Try searching for gene orthologs for a particular gene via NCBI gene: <https://www.ncbi.nlm.nih.gov/gene/>
  1. You can use the shown examples or your own favorite genes (can be non-human).
  2. Scroll down to the section called, “General Gene Information”
  3. Compare outputs from OrthoDB and NCBI Ortholog
  4. Any potential incidences of gene duplication or loss?
2. Look up mouse phenotypes using The Mouse Phenome Database: <https://phenome.jax.org/>
  1. Works best if you have the mouse gene name – can be retrieved using the Ortholog databases



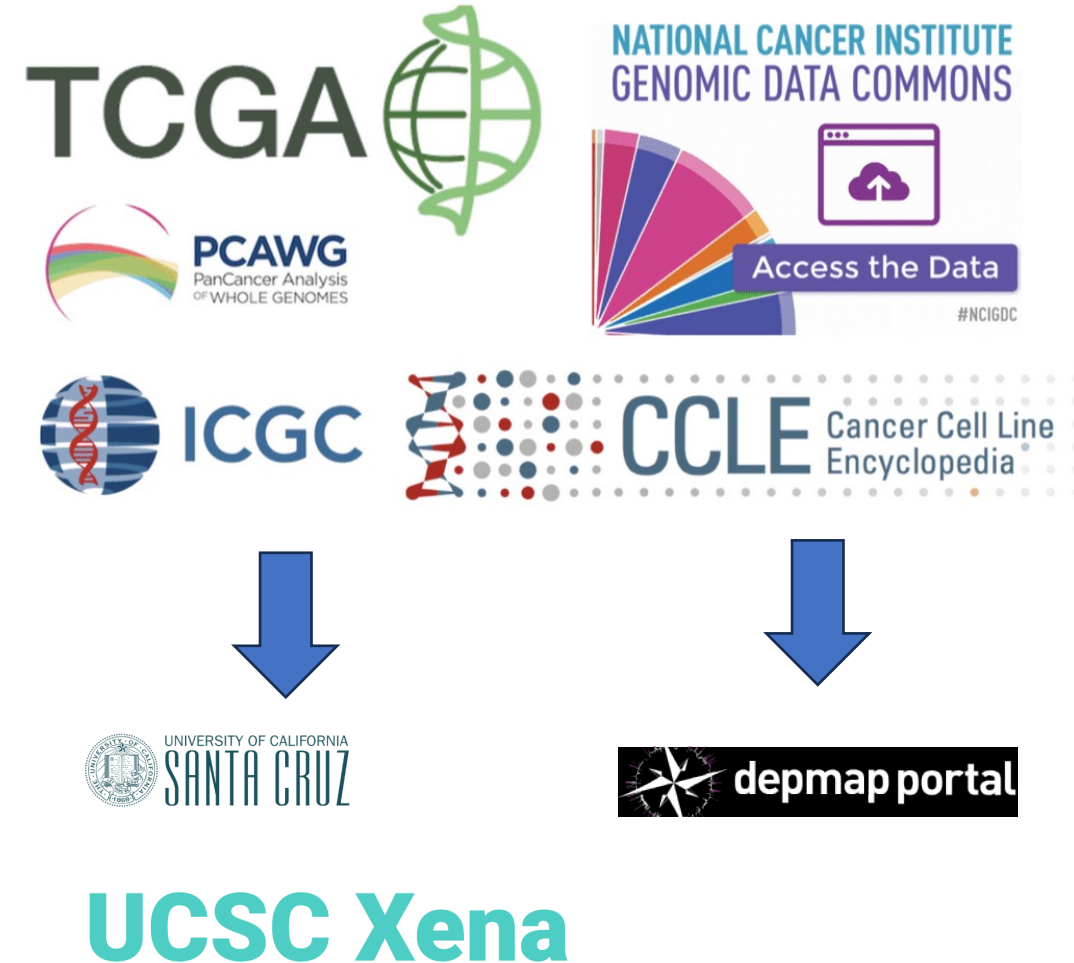
# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism database and strategies
7. Cancer relevant databases
8. Concluding remarks & discussion



# Cancer Genomics Resources

- The Cancer Genome Atlas (TCGA)
  - Active from 2006 to 2018 – data available via Genomic Data Commons and **Xena browser** (facilitates analyses as well)
  - Concluding project is called The Pan-Cancer Analysis of Whole Genomes (PCAWG)
- International Cancer Genomics Consortium (ICGC)
  - ICGC Data Portal closing down June 2024 but data available via Xena
- Cancer Cell Line Encyclopedia (CCLE)
- Catalogue of Somatic Mutations in Cancer (COSMIC)
- **Integrative cancer genomics resources (to be discussed)**
  - **UCSC Xena browser** for clinical cancer genomics
  - **Depmap portal** for cell-types and cancer dependencies





# Cancer Nomenclature

- **Mutation types – coding and non-coding regions**
  - SNPs (single nucleotide polymorphisms) and small INDELs (nucleotide insertions or deletions)
  - Copy number variants (CNVs)
  - Gene fusions
  - Large structural variants
- **Gene-, Transcript-, Exon-, Protein-, LncRNA-, and miRNA-expression**
- **Epigenetics – DNA methylation**
- **Synthetic lethality** – Pairs of genes for which an aberration in either gene alone is non-lethal, but co-occurrence of the aberrations is lethal to the cell
- **Cancer types**
  - TCGA coding and abbreviations: <https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations>



# The Cancer Genome Atlas Program (TCGA)



- TCGA: <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>
- Initiated in 2006, molecularly characterized 20K+ primary cancer and matched normal samples for 33 cancer types.
- Program closed in 2018 but data remains available via the Genomic Data Commons – Data Portal:  
<https://portal.gdc.cancer.gov/>

## Genomic Data Commons Data Portal

### Harmonized Cancer Datasets

A repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.

[Explore Our Cancer Datasets](#)

### Data Portal Summary

[Data Release 39.0 - December 04, 2023](#)



79  
Projects



69  
Primary Sites



44,451  
Cases



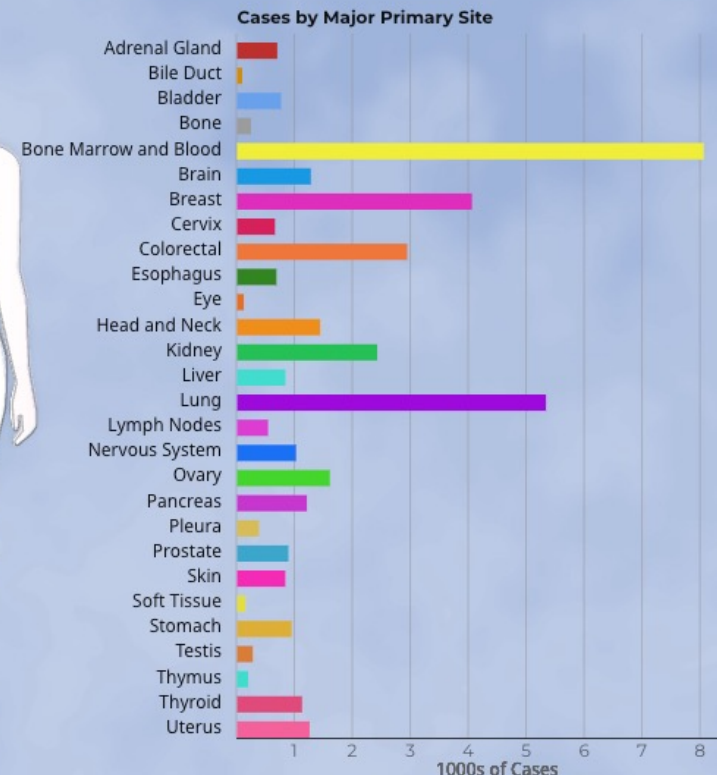
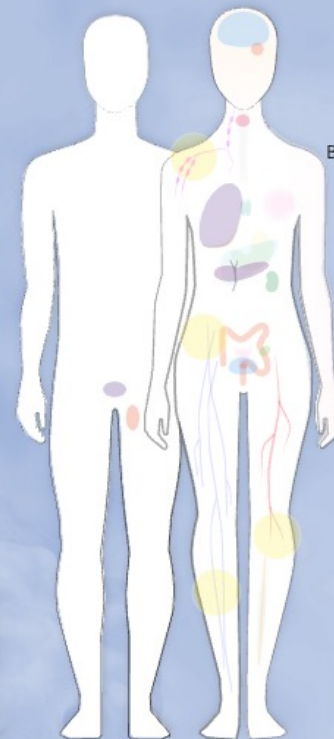
986,114  
Files



22,534  
Genes



2,930,136  
Mutations





# Xena Functional Genomics Explorer



- Xena: <https://xena.ucsc.edu/>
- Graphical interface to multiple cancer genomics data-types
  - Both on-line and downloadable desktop versions
- TCGA
  - TCGA Pan-Cancer Atlas (PANCAN) – Recommended for most analysis
  - TCGA data from Genomic Data Commons (GDC)
  - TCGA & GTEX data from the UCSC RNA-seq Recompute Compendium
  - Legacy TCGA data
- International Cancer Genome Consortium (ICGC)
- Pan-Cancer Analysis of Whole Genomes (PCAWG) study
- GDC
- MET500 (metastatic cancer study)
- CCLE
- Pediatric data:
  - KidsFirst
  - Target
  - Treehouse Consortium
- Can add and view your own data

Xena supports a wide variety of data types including:

- SNPs and small INDELs
- Large structural variants
- Segmented copy number, gene-level copy number
- Gene-, Transcript-, Exon-, Protein-, LncRNA-, and miRNA-expression
- DNA methylation (genes and probes)
- Phenotype, clinical data
- Signature scores, classifications, derived parameters

<https://ucsc-xena.gitbook.io/project/public-data-we-host>



# Xena Functional Genomics Explorer



- Xena: <https://xena.ucsc.edu/>
- Recommend reviewing tutorials and walkthroughs
- Python and R APIs
- Install a local hub to analyze your own data
- Goldman, M.J., Craft, B., Hastie, M. et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol (2020).

<https://doi.org/10.1038/s41587-020-0546-8>



## UCSC Xena

*See the bigger picture*

An online exploration tool for public and private, multi-omic and clinical/phenotype data

Launch Xena

### Tutorials and walkthroughs

Don't know where to start? Jump in with one of our tutorials or "How do I ..." walkthroughs

Tutorials

Overview

Analysis

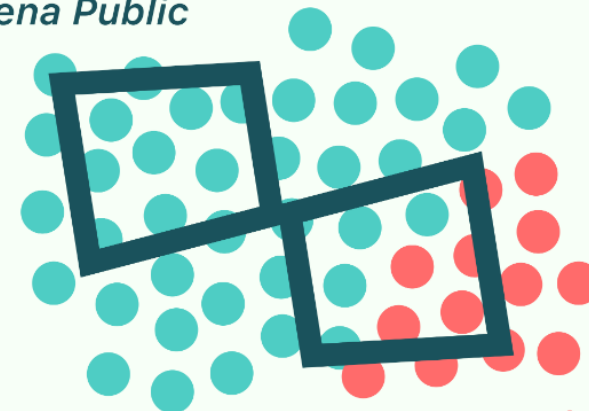
Tutorials

What's New

Cite Us

Subscribe

*Xena Public*



*Xena Private*



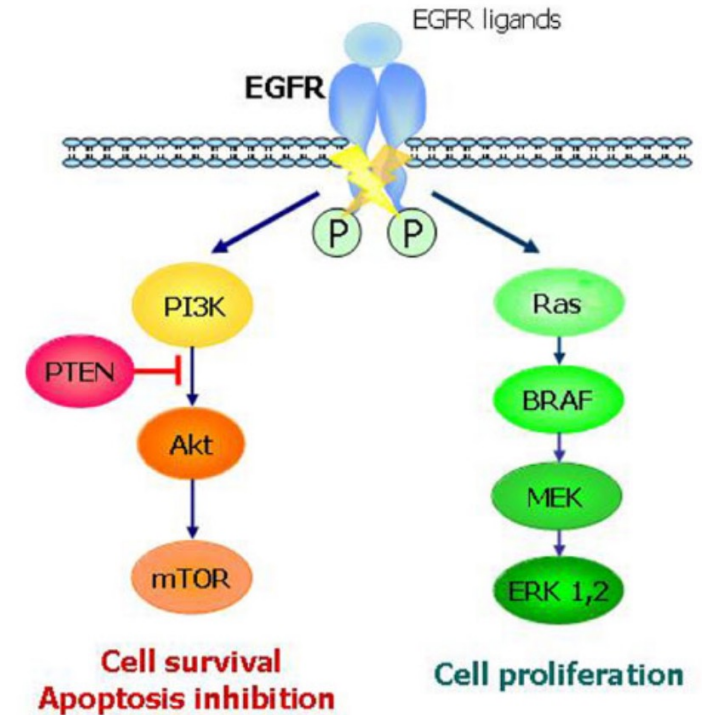
# Xena: Exploring EGFR Mutations in Lung Cancer



- Xena: <https://xena.ucsc.edu/>
- Do samples that have aberrations in EGFR have statistically higher expression than those without aberrations?
- Is there a survival difference between these two groups?
- Is there a gender difference in the occurrences of EGFR aberrations?

## Overview of *EGFR* in Lung Cancer

- *EGFR* aberrations (mutations or amplifications) are present in 10–35% of Lung Adenocarcinoma patients
- *EGFR* inhibitors are currently being used in the clinic
- Aberrations are more common in women



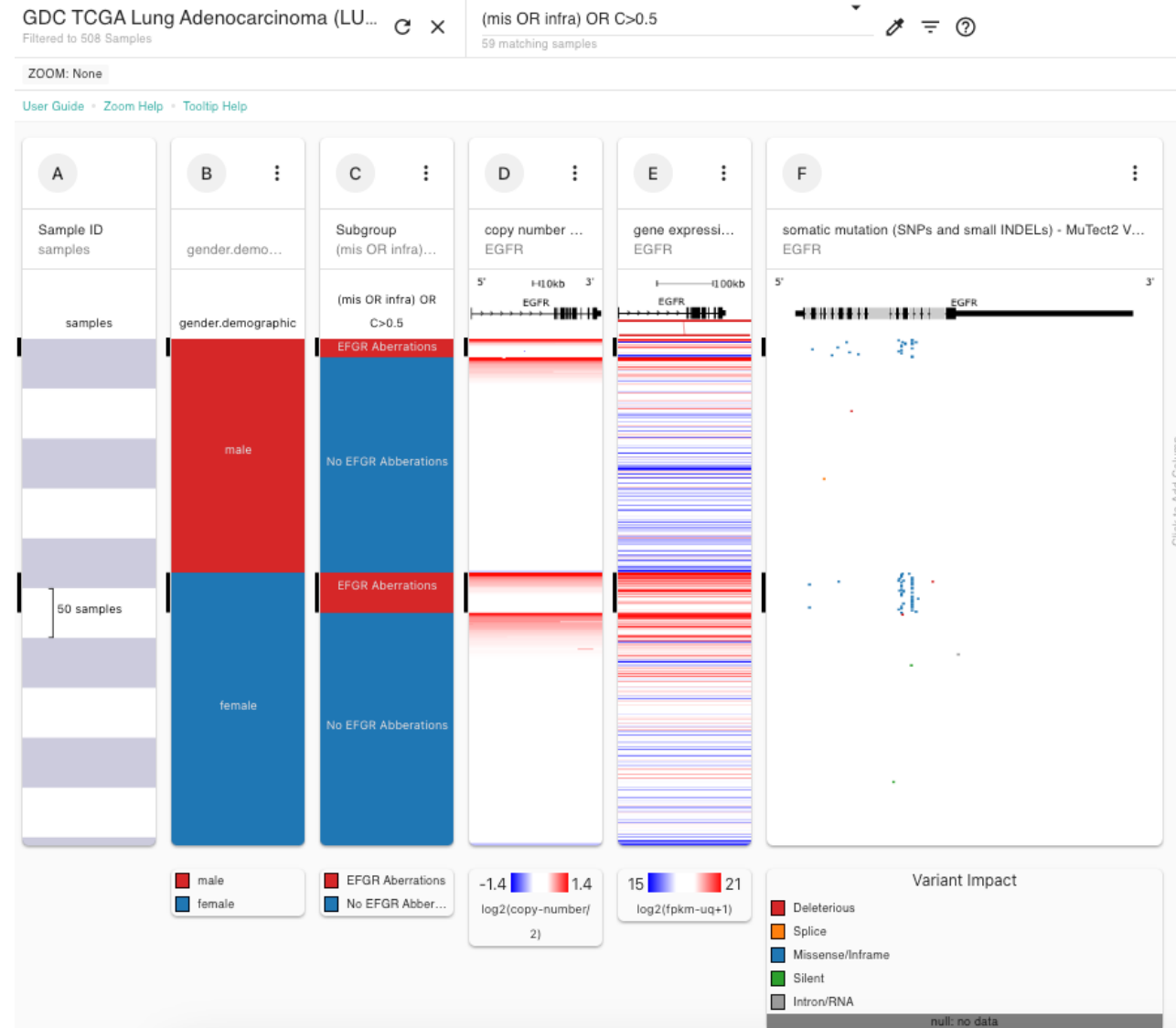
Saletti, et al 2015



# Xena: Building a Visual Spreadsheet



- Xena: <https://xena.ucsc.edu/>
- Create a visual spreadsheet
  1. Load study sample data: 'GDC TCGA Lung Adenocarcinoma (LUAD)'
  2. Variable: Genotype 'EGFR'; Gene Expression, Copy Number, and Somatic Mutation data
  3. Filter-out 'null' samples
  4. Add new subgroups:
    1. EGFR Aberrations ('(mis OR infra) OR C:>0.5')
    2. No EGFR Aberrations
  5. Add 'Gender.demographic'
  6. Screenshot of completed data-table
  7. Generate Kaplan-Meier survival plots
  8. Generate box or violin comparison plots
  9. Differential gene expression



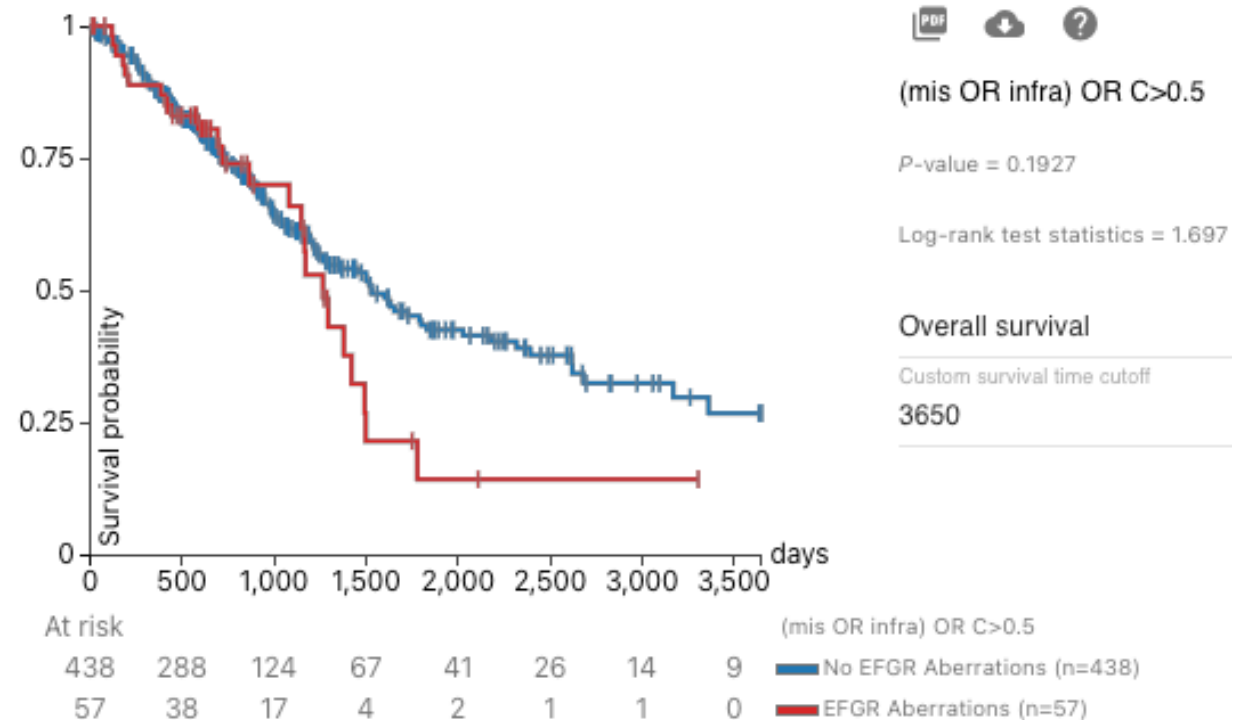


# Xena Functional Genomics Explorer



- Xena: <https://xena.ucsc.edu/>
- Share your tables via Bookmarks (bookmarks are only guaranteed for 3 months):  
<https://xenabrowser.net/?bookmark=1eb2cbadfe4a36d0d1dd47d18d2c24cc>
- Differential gene expression and pathway enrichment analyses between EFGR +/- aberrations datasets (access by clicking on 3-dots Subgroup col. C):  
<http://analysis.xenahubs.net/3e53da9a2084901f83dc09510c9e65b09086ac2e/>
- Alternative to archive analysis results:
  - Download data & plots
  - Create your own local data-hub by download and installing a local copy of Xena
- See Xena's Advanced Tutorials for more information

Kaplan Meier Subgroup



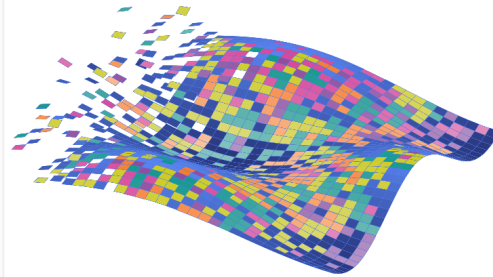


# DepMap (Dependencies Map)

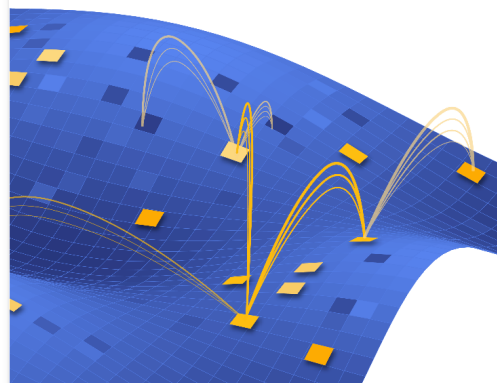


- DepMap: <https://depmap.org/portal/home/#/>
  - Builds on the original Cancer Cell Line Encyclopedia (CCLE) project, which characterized 1000 cell line models. To date, more than 2000 models have been collected: <https://sites.broadinstitute.org/ccle/>
- Data explorer: <https://depmap.org/portal/interactive/>

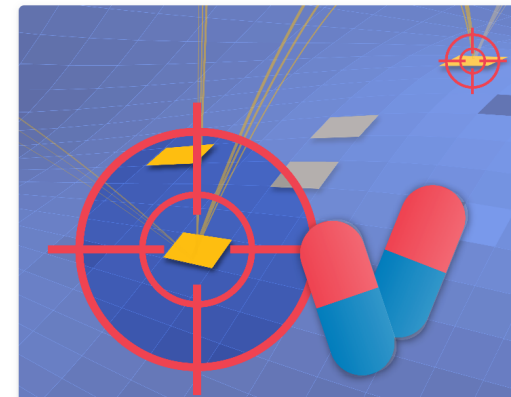
## Our mission



Represent all human cancer genetic and molecular diversity



Identify and understand the landscape of cancer dependencies

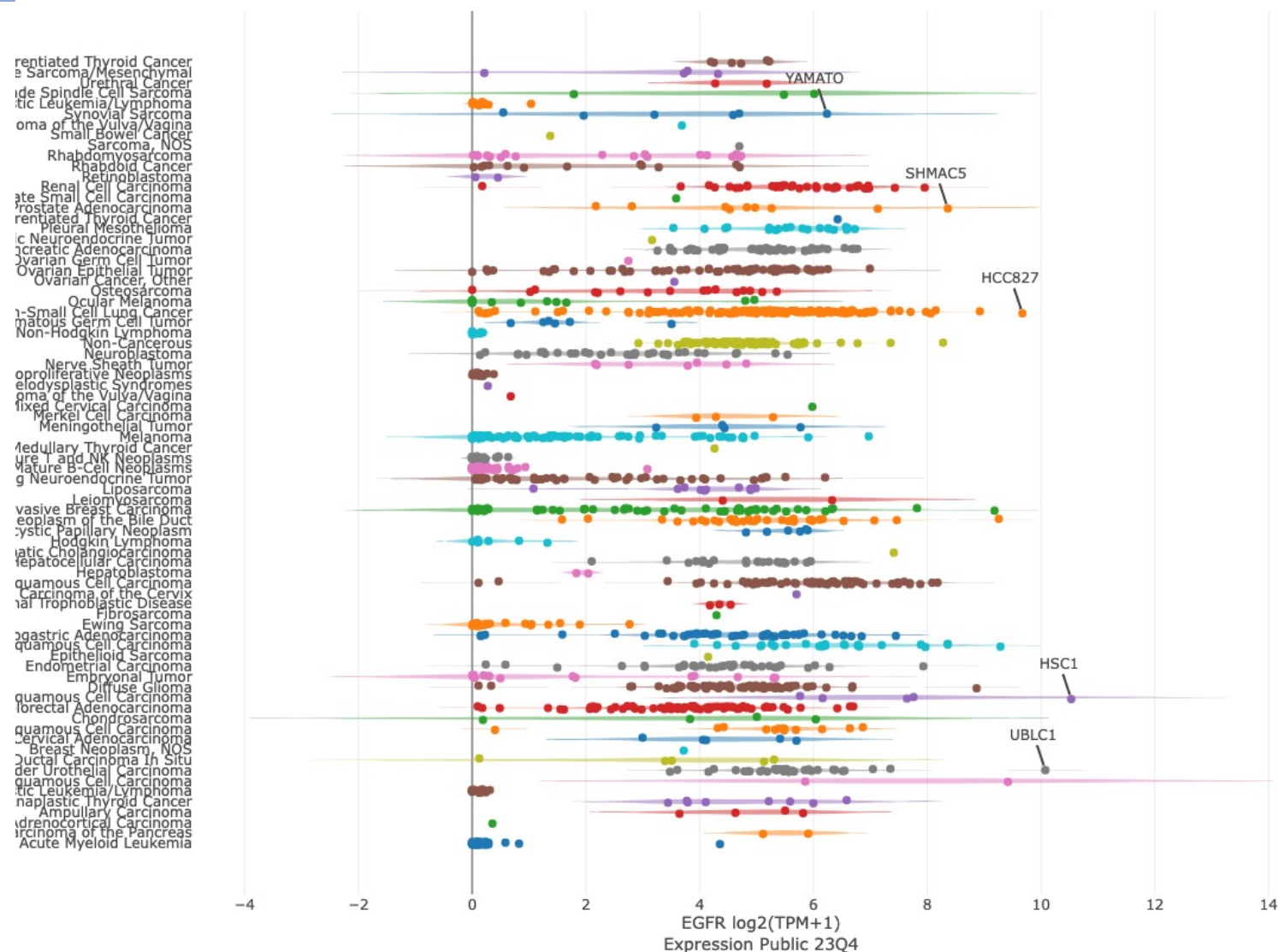


Create a resource for drug target and biomarker discovery



# DepMap Example: Genes EGFR

- DepMap Data explorer:  
<https://depmap.org/portal/interactive/>
  - X – Axis
  - Select gene: EGFR
  - Select dataset: Expression public 23Q4
  - View options: Group by primary disease
  - Add cell-line labels via click
  - Downloadable data & figures





# DepMap Example: EGFR and GRB2 Co-dependency

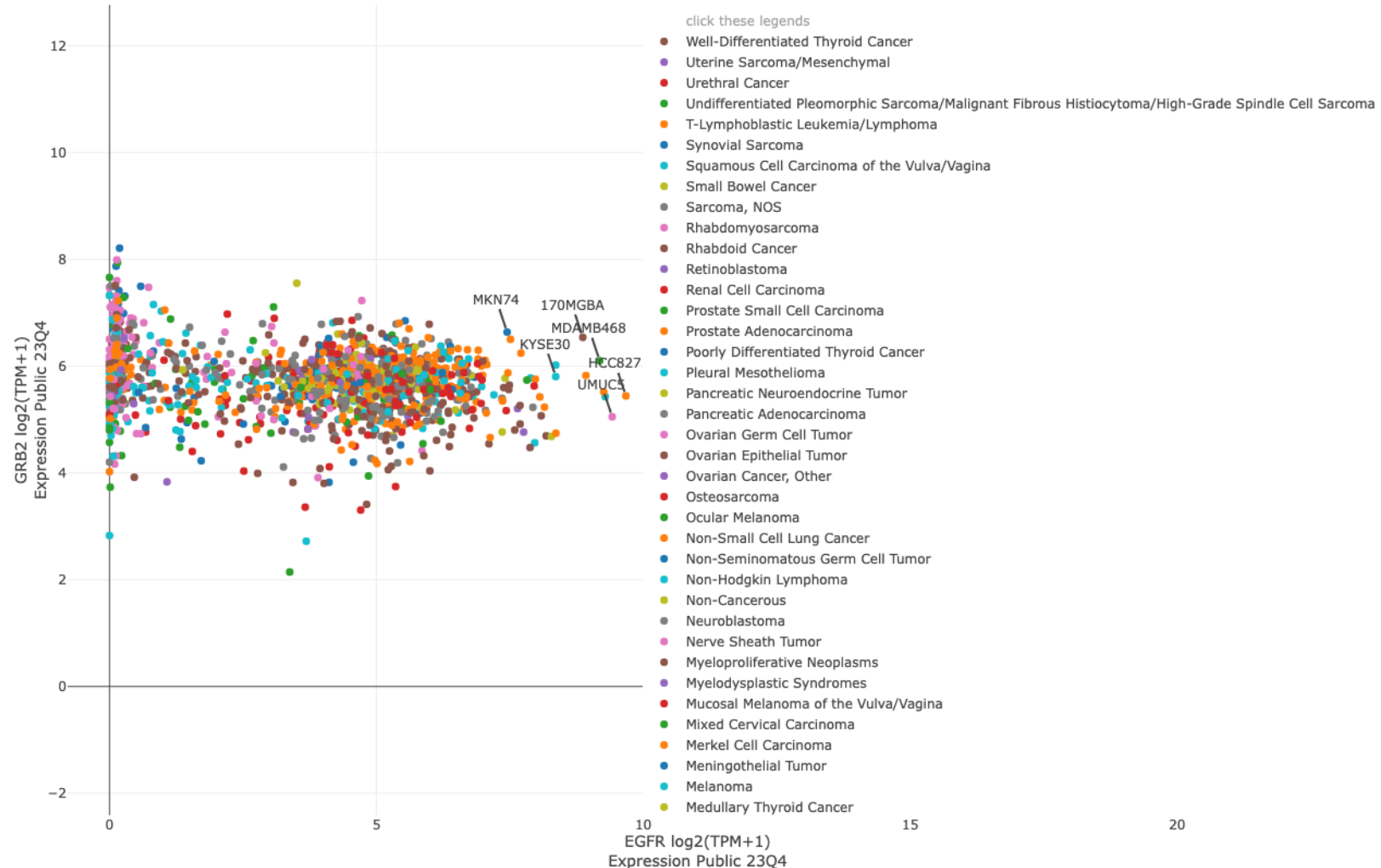
- EGFR overview:  
<https://depmap.org/portal/gene/EGFR?tab=overview>
- CRISPR Gene effects summary suggests that EGFR and GRB2 are co-dependent
- Score of “0” is equivalent to the gene not being essential
- Whereas a score of “-1” corresponds to the mean of all essential genes.
- Several cell-lines have values < -1 for both genes.





# DepMap Example: Genes EGFR and GRB1 Co-expression

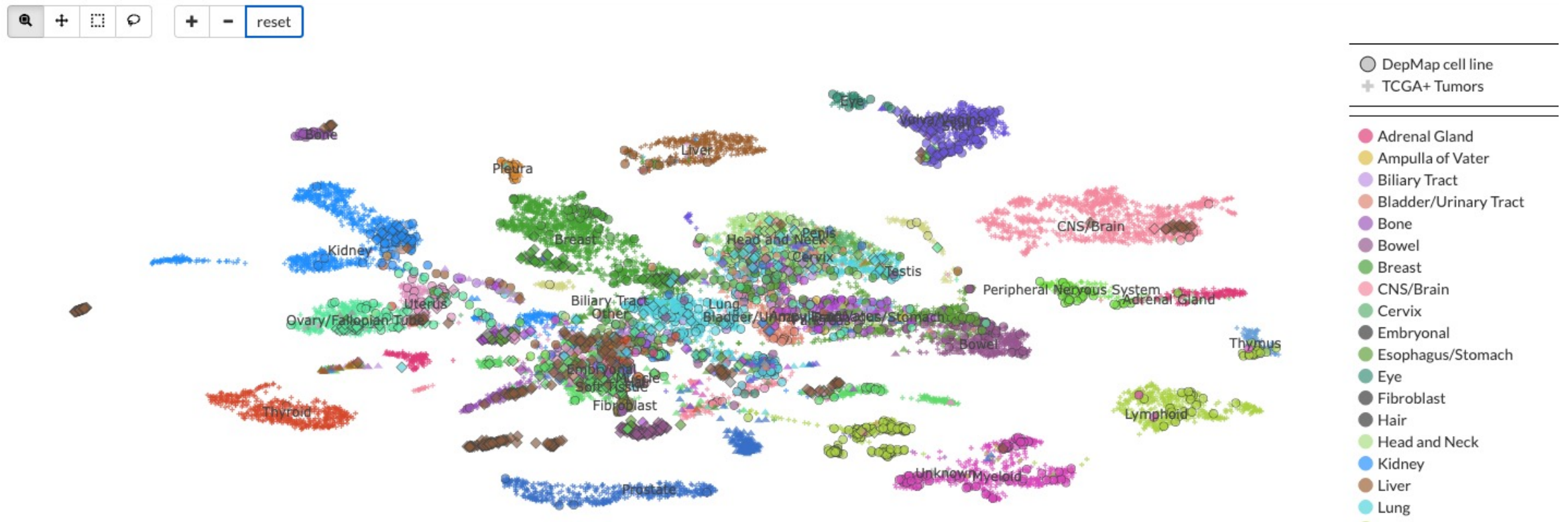
- DepMap Data explorer:  
<https://depmap.org/portal/interactive/>
- EGFR & GRB1 co-expression:
  - X – Axis
  - Select gene: EGFR
  - Select dataset: Expression public 23Q4
  - Y – Axis
  - Select gene: GRB1
  - Select dataset: Expression public 23Q4
  - View options: Group by primary disease





# DepMap: Celligner – Tumor + Cell Line Model Alignment

- DepMap: <https://depmap.org/portal/celligner/>
- Integrated CCLE and tumor expression datasets with calculated distance metrics for overall similarity
  - Rank cell lines for selected tumors
  - Find most similar tumors for a given cell line





# Catalogue Of Somatic Mutations in Cancer (COSMIC)



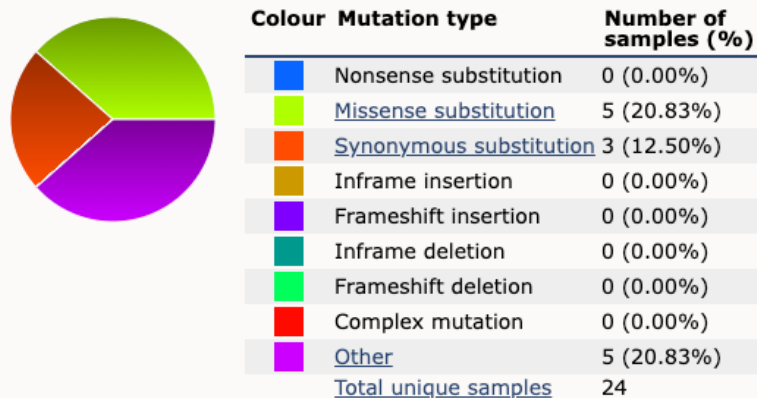
- Cosmic: <https://cancer.sanger.ac.uk/cosmic> (outdated?)
- Initial COSMIC Search results for gene CLEC2D: <https://cancer.sanger.ac.uk/cosmic/search?q=CLEC2D>
- Gene view link: <https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=CLEC2D>

## Mutation distribution

This section displays a series of charts that show the distribution of different types of mutations for CLEC2D. [Show more](#)

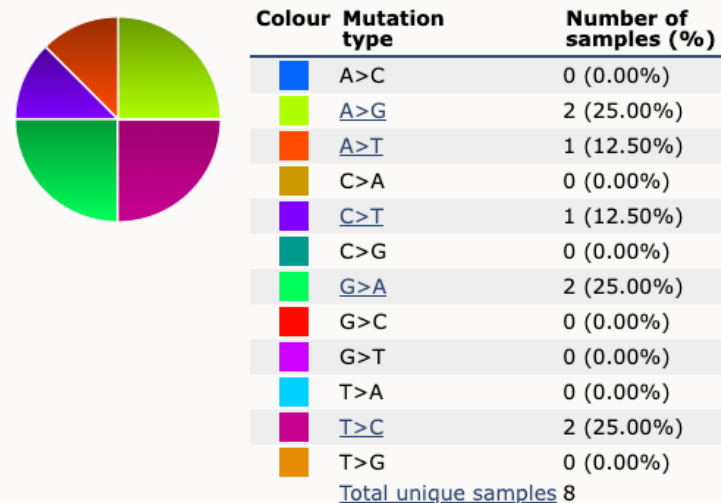
### Summary

An overview of the types of mutation observed.



### Substitutions

A breakdown of the observed substitution mutations.



### Deletions

There are no observed deletion mutations.

### Insertions

There are no observed insertion mutations.



# Exercise 4: Cancer Databases

1. Recreate the example of EGFR mutations in lung adenocarcinoma using Xena: <https://xena.ucsc.edu>
  1. Hint: this is Basic Tutorial Section 1: <https://ucsc-xena.gitbook.io/project/tutorials/basic-tutorial-section-1>
2. For the genes EGFR and GRB1, recreate CRISPR co-dependency and co-expression plots using DepMap data explorer: <https://depmap.org/portal/interactive/>



# Outline

1. Overview of drug research and development
2. Integrative biomedical databases
3. Human centric data (genetics, clinical trials, drug and tool compounds)
4. Multi-omics evidence databases
5. Protein characterization and interactions databases
6. Comparative genomics and model organism database and strategies
7. Cancer relevant databases
8. Concluding remarks & discussion



# Future Directions

- Further innovations in functional genomics assays (i.e. single-cell genomics, spatial genomics, CRISPR)
- Growth of clinical datasets with deep genomic analyses and precision medicine focus
- The future of AI/Machine Learning and Drug Target Discovery/Validation
  - Entering a new era of AI enabled target discovery
  - Large language models (LLMs) trained on diverse chemical, biological and clinical datasets
  - Understanding feature selection and the underlying drivers of AI model predictions could be insightful
  - Applications of AI to multi-omics analyses are exciting yet still evolving
    - Bzdok et al. 2024. Neuron. Data science opportunities of large language models for neuroscience and biomedicine  
<https://doi.org/10.1016/j.neuron.2024.01.016>
    - Ren et al. 2024 Nature Biotechnology. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models  
<https://doi.org/10.1038/s41587-024-02143-0>
  - Cautionary notes on applying machine learning to Clinical Prediction (Chekroud et al. 2024. Science 383:164.  
<https://www.science.org/doi/10.1126/science.adg8538> )



# Concluding Remarks

- Caveats
  - Critically evaluate any results and cross check using multiple sources.
  - Be mindful of the lag time between discovery in the literature and incorporation in public databases.
  - Many databases are very human or mammalian centric:
    - Pathogens have their own resources as does the microbiome.
  - These web-tools are initial starting points. Leading towards greater more in-depth computational biology analyses such as phylogenomic analyses of orthologs or differentially expressed gene analyses.
  - Most of these databases have excellent free tutorials as well as helpful community blogs and discussion groups.
  - For any computational hypothesis, it is essential to have experimental and/or clinical validation.
- *Thank you!!*
- *Questions?*
- *I am available for 1x1 meetings today and tomorrow (E-mail: [jb4633@drexel.edu](mailto:jb4633@drexel.edu) )*