

Breakout Session 2: Track B

Migration of Core Applications from the NIDDK information Network (dkNET)

Dr. Jeffrey Grethe

PI, NIDDK Information Network (dkNET), University of California at San Diego



Migration of Core Applications from the NIDDK information Network (dkNET)

Breakout Session 2 – Track B

Jeffrey S. Grethe, Ph.D.

FAIR Data Informatics Laboratory

University of California San Diego, School of Medicine

January 17, 2024



What is dkNET?

- dkNET provides a single point of access to information about diverse research resources, including **data, information, materials, tools, funding opportunities, literature, services, events, news,** and **projects** that advance the mission of the NIDDK.
- dkNET provides tools and services in support of **rigor and reproducibility**, built around the Research Resource Identifier (**RRID**) and the **FAIR data principles** (Findable, Accessible, Interoperable, Re-usable).

<https://dknet.org>

The screenshot shows the dkNET website header with the logo and the title "dkNET: Connecting Researchers to Resources". Below the header, there are four main service tiles:

- Resource Reports**: "Is my antibody specific? Who else is using my software tools? Answer these questions and more using Research Resource Identifiers (RRIDs) and Digital Object Identifier (DOIs). Tools | Cell lines | Antibodies | Organisms | Plasmids | Biosamples | Protocols"
- Discovery Portal**: "Search across 100s of biomedical databases for... Funding | Images | Phenotypes | Literature | and more"
- Authentication Reports & FAIR Data**: "View resources on how to comply with NIH's new policies on authentication of key biological resources, using our authentication reports, and making data FAIR. Authentication reports | Research data management | Suggested data repositories"
- Hypothesis Center**: "Analyze diverse 'omics data to generate or test research hypotheses – powered by the Signaling Pathways Project."



Hypothesis Center - A Powerful FAIR Platform for Biomedical Researchers

- Lowering the barrier of entry to bioinformatics resources and workflows
- Provide information to the DK community on computational resources
- A hub for big data and hypothesis generation, bringing together a collection of online tools
- Detailed tutorials guide researchers in using these resources (Am I using the tools correctly? feedback from ASCB)

The screenshot shows the Hypothesis Center website. At the top left is the DKnet logo. The main heading is "Hypothesis Center" with a navigation link for "Home / Hypothesis Center". Below this is a main content area with a featured article titled "Hypothesis Center - A Powerful FAIR Platform for Biomedical Researchers". The article text describes dkNET as a hub for big data and hypothesis generation, mentioning the Signaling Pathways Project (SPP) and Mouse Metabolic Phenotyping Centers (MMPC). It includes a "GET STARTED" section with a dropdown menu for "WHAT HYPOTHESIS ARE YOU DEVELOPING OR WHAT INFORMATION ARE YOU LOOKING FOR?". To the right, there's a "New Resources and Tutorials Added!" section featuring logos for HIRN RESOURCE BROWSER, appyters, and TYPE 1 DIABETES KNOWLEDGE PORTAL. Below this is a question: "Do you want to learn how to find gene sets and signaling pathways relevant to mouse phenotypes?" with a "Start Here" button and a small image of a brain. The main content area also features two articles: "Signaling Pathways Project Analysis Generates Hypotheses Around Host Responses to SARS-CoV-2 Infection in COVID-19" and "Researchers Study Metabolic Rate and Response to Obesity Analyzing MMPC Energy Expenditure Data". The bottom section is titled "CURRENTLY AVAILABLE CORE RESOURCES" and lists Signaling Pathways Project (SPP), Human Islet Research Network (HIRN), and Type 1 Diabetes Knowledge Portal (T1DKP). It also includes a "CURRENTLY AVAILABLE CORE RESOURCES" section with logos for Mouse Metabolic Phenotyping Centers (MMPC) and Appyters. At the very bottom, there's an "OVERVIEW OF HYPOTHESIS CENTER RESOURCES" section with links for "Resources" and "Available tutorials".



Hypothesis Center - Signaling Pathways Project



The Signaling Pathways Project

A multi-omics knowledgebase for cellular signaling pathways

Consensome [Download Results](#)

Category: Receptors
Class: Nuclear receptors
Family: Thyroid hormone receptors
Species: House Mouse
Physiological System: All
Organ: All

Consensomes are list of genes ranked according to a meta-analysis of their differential expression in publicly archived transcriptomic datasets involving perturbations of a specific signaling pathway in a given biosample category. Consensome are intended as a guide to identifying those genes most consistently impacted by a given pathway in a given tissue context.

Calculated across X data points from Y experiments in Z data points

Show 50 entries

Target	Gene Name
Bcl3	B cell leukemia/lymphoma 3
Idh3a	isocitrate dehydrogenase 3
Mmd	monocyte to macrophage differentiation
Stat5a	signal transducer and activator of transcription 5A
Ndr3	N-myc downstream regulated 3
Trp53inp2	transformation related protein 53 interacting protein 2
Ces1f	carboxylesterase 1F
Eph1	epoxide hydrolase 1, membrane bound

Regulation Report [Download Results](#)

Transcriptomics

Display by: Category Down Up Currently displaying 300 out of 300 data points.

Category	Gene	Transcript Relative Abundance (Fold Change)
Receptors Catalytic receptors	Collagen receptor family	
	NILOT	
	Epidermal growth factor receptors	
Epidermal growth factor receptors	AG478	~2.5
	EGF	~5.5
	GEFIT	~2.5
	ERBB2	~-2.5
	EGF	~3.5
Fibroblast growth factor receptors	FGF19	~2.5
	Hepatocyte growth factor receptors	
Hepatocyte growth factor receptors	PHA665	~-2.5
	HGF	~2.5
IL1 receptor family		~2.5
	IL1B	~2.5

SPP simplifies data mining of 'omics data, connects bench researchers to FAIR data to allow them to easily interrogate the data to generate hypotheses

- **Find genes** with important roles in receptors, enzymes, organs and tissues
- **Define signaling pathways** relevant to a single gene or a regulation



STRIDES Goals

The Hypothesis Center allows researchers to extract information across transcriptomic and ChIP-Seq datasets, providing a powerful meta-analysis platform that surveys across millions of FAIRly biocurated omics data points to make high-confidence connections between genomic targets, their upstream regulatory pathways, and disease states. Future plans include expansion to more data types. The HC is comprised of two primary components: 1) [The Signaling Pathways Project \(SPP\) knowledgebase](#) and 2) Research resource information and resolution services. Through the tools in the HC, a researcher can find a list of target genes for receptors, enzymes and transcription factors, obtain a snapshot of the dynamic regulatory programs of the cells under study, and utilize this information to formulate new hypotheses. Researchers can then further, for example, prioritize molecules and design experiments to test the novel hypothesis by targeting these molecules. The HC is designed to be user-friendly for a wide spectrum of scientists including bench scientists with little computer programming skills. It is currently housed at an institutional server at the Baylor College of Medicine. The application seeks support to migrate the HC to the public cloud. The migration will allow the investigators of dkNET to further enhance its functionality utilizing the cloud environment, and to make it more findable and accessible to the community.

Initial Phase: Transition of SPP database and application server to cloud environment






Signaling Pathways Project


SCIENTIFIC DATA 

OPEN
ARTICLE

The Signaling Pathways Project, an integrated 'omics knowledgebase for mammalian cellular signaling pathways


Scott A. Ochsner¹, David Abraham^{1,8}, Kirt Martin^{1,8}, Wei Ding², Apollo McOwiti², Wasula Kankanamge², Zichen Wang , Kaitlyn Andreano⁴, Ross A. Hamilton¹, Yue Chen¹, Angelica Hamilton⁵, Marin L. Gantner⁶, Michael Dehart², Shijing Qu², Susan G. Hilsenbeck², Lauren B. Becnel², Dave Bridges⁷, Avi Ma'ayan , Janice M. Huss⁵, Fabio Stossi¹, Charles E. Foulds¹, Anastasia Kralli⁶, Donald P. McDonnell⁶ & Neil J. McKenna ^{1*}

SCIENTIFIC DATA 

 Check for updates

OPEN
ANALYSIS

Consensus transcriptional regulatory networks of coronavirus-infected human cells

Scott A. Ochsner¹, Rudolf T. Pillich² & Neil J. McKenna ^{1,2*}

Establishing consensus around the transcriptional interface between coronavirus (CoV) infection and human cellular signaling pathways can catalyze the development of novel anti-CoV therapeutics. Here, we used publicly archived transcriptomic datasets to compute consensus regulatory signatures, or consensomes, that rank human genes based on their rates of differential expression in MERS-CoV (MERS), SARS-CoV-1 (SARS1) and SARS-CoV-2 (SARS2)-infected cells. Validating the CoV

Research article  JHEP|Reports

A human liver chimeric mouse model for non-alcoholic fatty liver disease

Authors

Beatrice Bissig-Choisat, Michele Alves-Bezerra, Barry Zorman, Scott A. Ochsner, Mercedes Barzi, Xavier Legras, Diane Yang, Malgorzata Borowiak, Adam M. Dean, Robert B. York, N. Thao N. Galvan, John Goss, William R. Lagor, David D. Moore, David E. Cohen, Neil J. McKenna, Pavel Sumazin, Karl-Dimiter Bissig

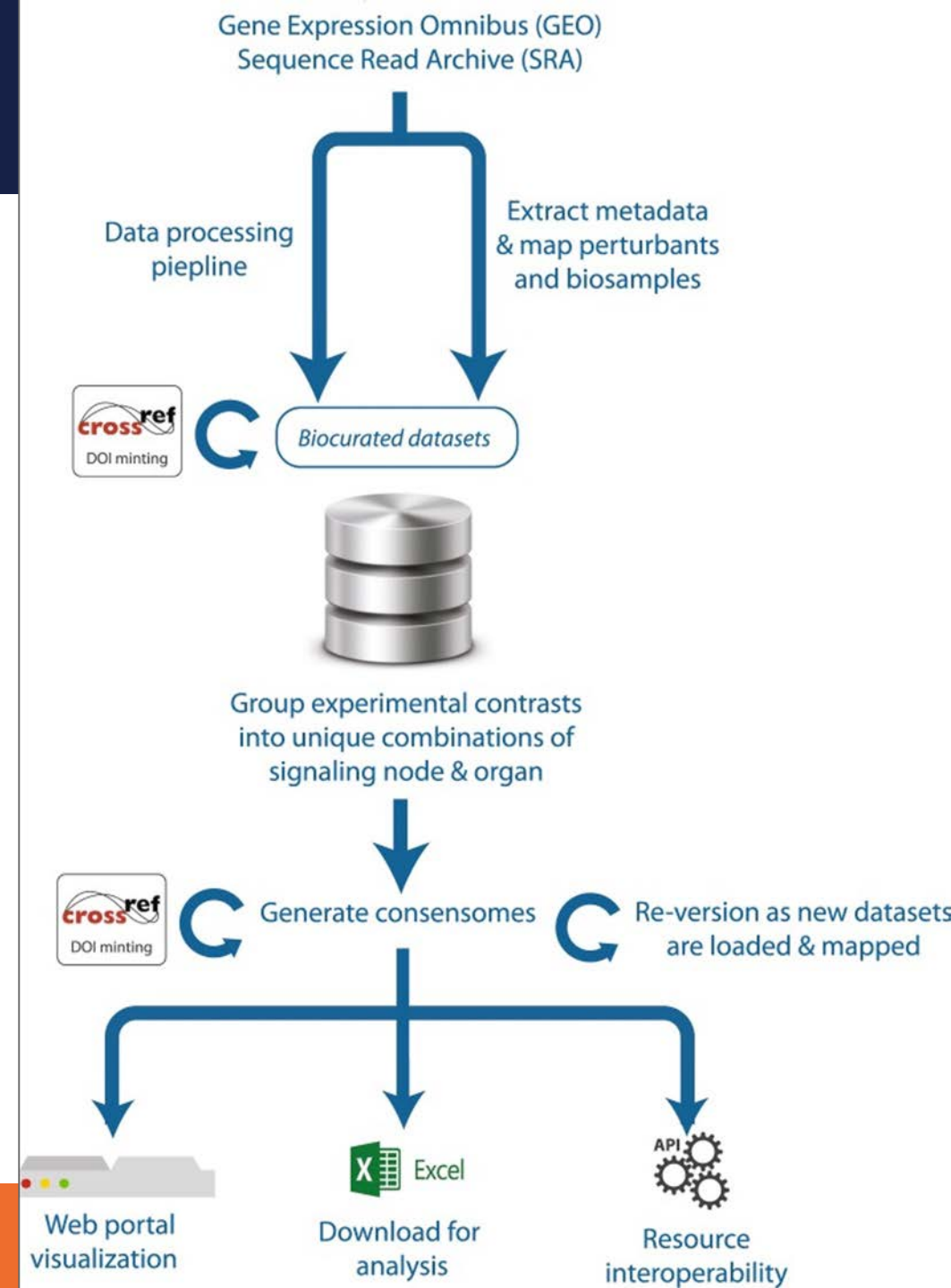
Correspondence

karldimiter.bissig@duke.edu (K.-D. Bissig).



SPP FAIR Curation

- Curation of existing datasets to improve FAIRness
- Enhanced datasets can be pre-processed to populate database
- Database content used to drive SPP web application





Current dkNET Cloud Experience

- Familiarity with AWS administration
- Foundry data aggregation pipeline deployed to AWS
 - Utilizes EC2, S3, RDS (MariaDB)
 - Data processing takes advantage of AWS auto-scaling with spot instances



SPP STRIDES Migration

Typical systems administration:

- SPP Application
 - Standard Java application server
 - Basic AWS administration
 - Java application development and deployment
- Data Processing
 - Basic AWS administration
 - Knowledge of Amazon S3



SPP STRIDES Migration

Need for specialized AWS knowledge:

- SPP Data (Oracle Database)
 - AWS “variant” of Oracle database
 - Special AWS commands for data transfer (via data pump)
 - Deeper knowledge of AWS configuration and security

Utilized AWS Support Center to complete migration

- STRIDES provides Enterprise support plan
- Provide information to navigate AWS specific Oracle implementation



Benefits of SPP STRIDES Migration

- Collaborative Development
 - Ability to provide access to cloud resources for developers from multiple sites (e.g. UCSD, BCM)
- Ability to use built-in AWS features
 - Automated backups and snapshots
 - Load balancing and fault tolerance (currently investigating)
- More options to tailor service capacity
 - Development versus Production
 - EC2 compute nodes
 - Oracle RDS options



Current SPP STRIDES Costs Overview

- Amazon RDS (Oracle)
 - 72% of total monthly cost
 - Largest cost driver due to Oracle database license [60% of RDS cost]
- Amazon EC2 (Web Application & Data Transfer nodes)
 - Expect increased cost (2-3X with fault tolerant deployment and additional processing)
- Other Costs (e.g. detailed monitoring, data transfer)

