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.

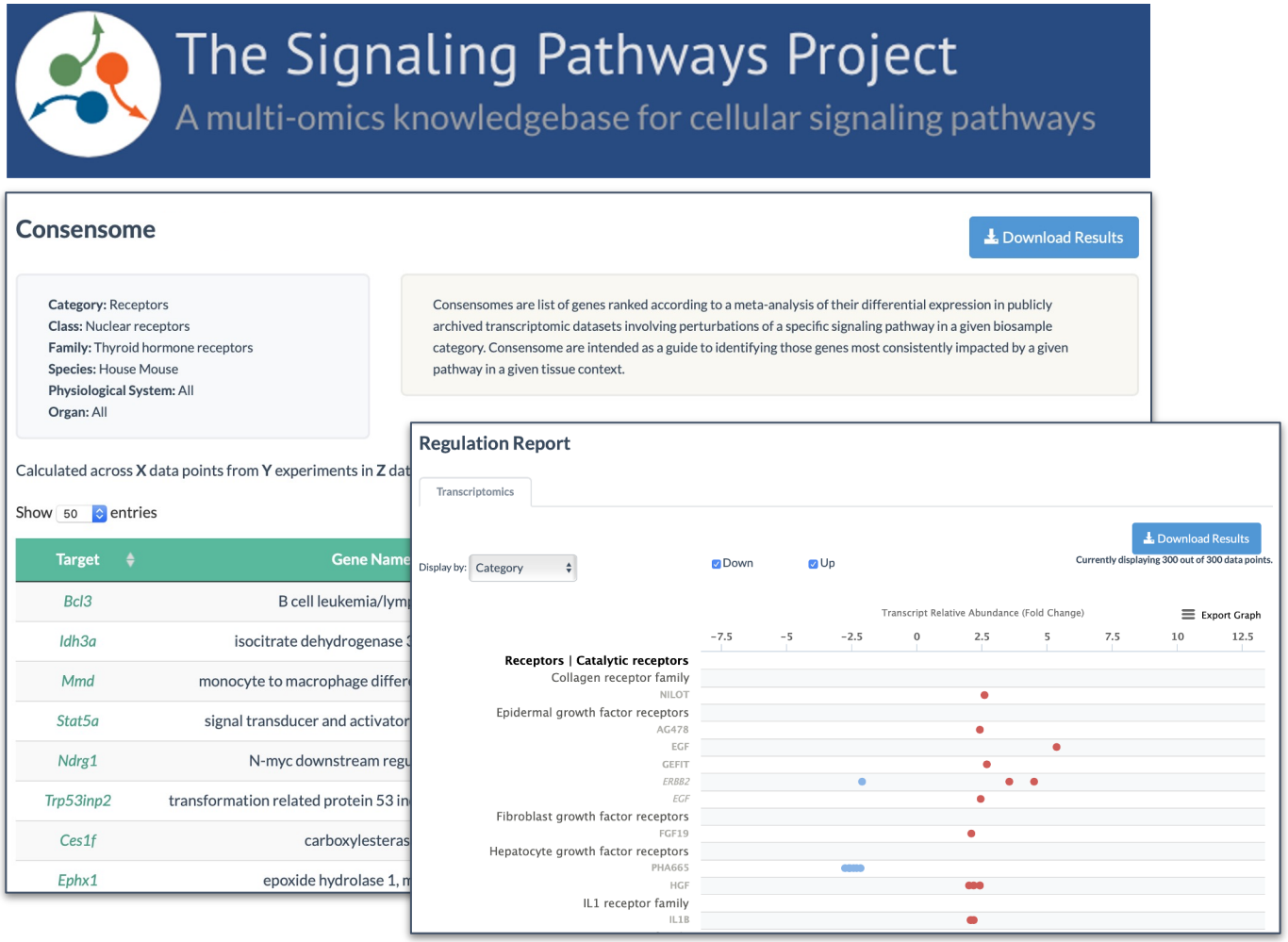
- dkNET provides tools and services in support of **rigor and reproducibility**, built around the Research Resource Identifier (**RRID**) and the **FAIR data principles** (Findable, Accessible, Interoperable, Re-usable).

https://dknet.org

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)

Hypothesis Center - Signaling Pathways Project



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

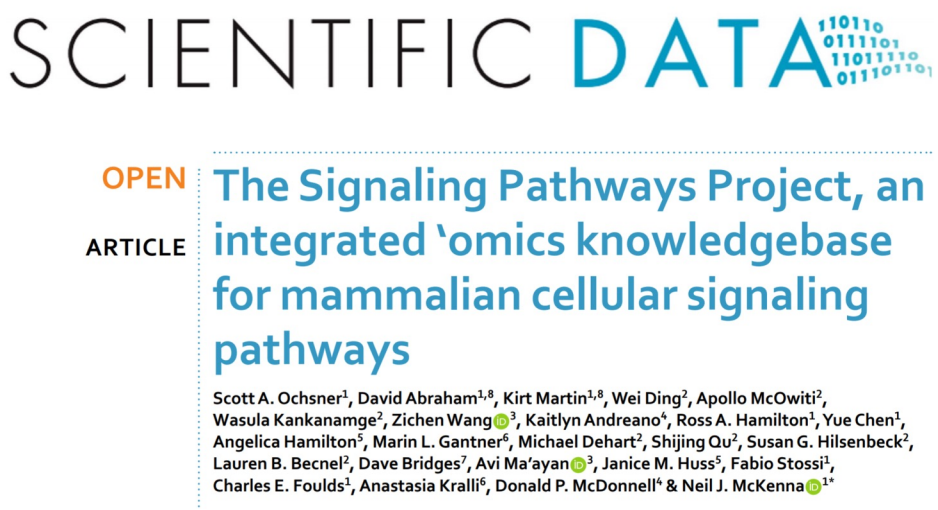
- **Find genes** with important roles in receptors, enzymes, organs and tissues
- **Define signaling pathways** relevant to a single gene or a regulation

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 ho
Medicine. The application seeks support to mi
allow the investigators of dkNET to further enh
and to make it more finable and accessible to t

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

Signaling Pathways Project



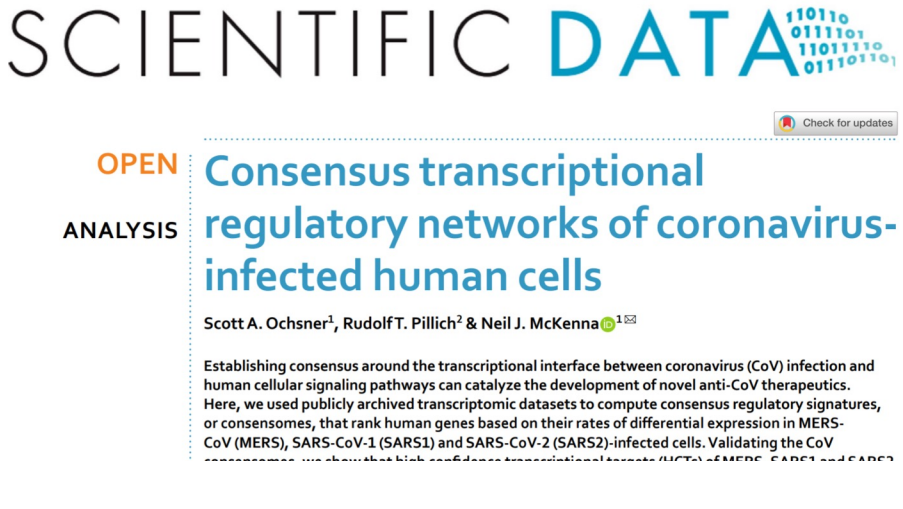
SCIENTIFIC DATA

OPEN

ARTICLE

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

Scott A. Ochsner[1], David Abraham[1,8], Kirt Martin[1,8], Wei Ding[2], Apollo McOwiti[2], Wasula Kankanamge[2], Zichen Wang[3], Kaitlyn Andreano[4], Ross A. Hamilton[1], Yue Chen[1], Angelica Hamilton[5], Marin L. Gantner[6], Michael Dehart[2], Shijing Qu[2], Susan G. Hilsenbeck[2], Lauren B. Becnel[2], Dave Bridges[7], Avi Ma'ayan[3], Janice M. Huss[5], Fabio Stossi[1], Charles E. Foulds[1], Anastasia Kralli[6], Donald P. McDonnell[4] & Neil J. McKenna[1*]



SCIENTIFIC DATA

Check for updates

OPEN

ANALYSIS

Consensus transcriptional regulatory networks of coronavirus-infected human cells

Scott A. Ochsner[1], Rudolf T. Pillich[2] & Neil J. McKenna[1]✉

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

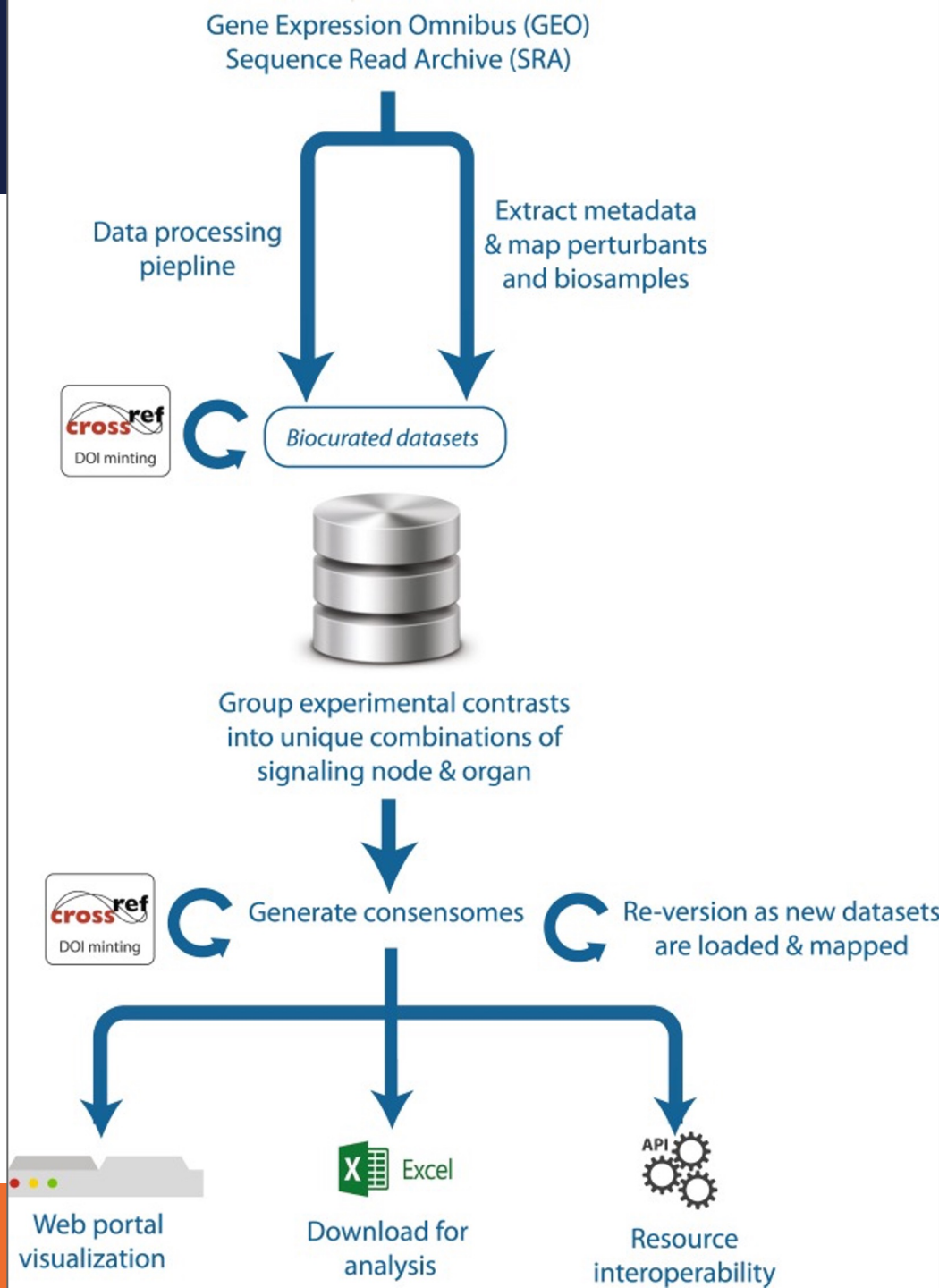
Beatrice Bissig-Choisat, Michele Alves-Bezerra, Barry Zorman, Scott A. Ochsner, Mercedes Barzi, Xavier Legras, Diane Yang, Malgorzata Borowiak, Adam M. Dean, Robert B. York, N. Thao N. Galvan, John Goss, William R. Lagor, David D. Moore, David E. Cohen, Neil J. McKenna, Pavel Sumazin, Karl-Dimiter Bissig

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)

thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
grazie
arigatō
dakujem
takk
tapadh leat
спасибо
Баярлалаа
рахмат
dank je
mersi
barka
welalin
tack
spas
faafetai lava
vinaka
blagodaram
misaotra
matondo
paldies
kiitos
dankie
hvala
mauruuru
koszönöm
enkosi
bayarlalaa
gracie
chnorakaloutioun
gracies
sulpay
taiku
djiere dieuf
tau
mamnun
хвала
asante
manana
murakoze
tenki
sobodi
dekuji
sagolun
mesi
najis tuke
ありがとう
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
merce
мерси
trugarez
didi madloba
kam sah hamnida
rahmat
감사합니다
xiexie
ευχαριστώ
akun
dankon
aciu
mahalo