WAVE™
LIFE SCIENCES

Data Science in Production: Making Data Science Accessible for Domain Experts and Creating Real Business Value

Kenneth A. Longo, Ph.D.
2020-11-19

KNIME Fall Summit 2020

A problem and a solution

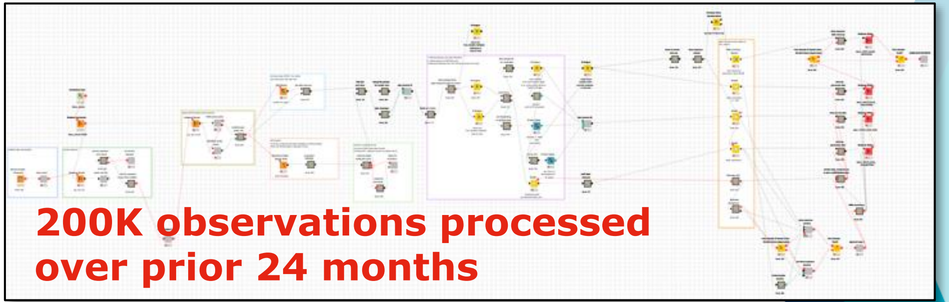
- **Problem**
  - How can we quickly and reliably develop, test and deploy software that uses modern tools of data science and answers key business questions?
  - Is there a platform that provides end-to-end software development tools for data science backend, data connectivity, UI, UX?
- **Solution**
  - KNIME Analytics Platform, Server & WebPortal
  - Provides access to 1) robust data science tools and 2) a means for productionizing solutions for broad use, delivery via browser.

KNIME Server pipeline & Webportal GUI for qPCR

The **qPCR workflow** for **mRNA measurement** is a guided analytic process for scientists that analyses 384-well bioassay data, combining:

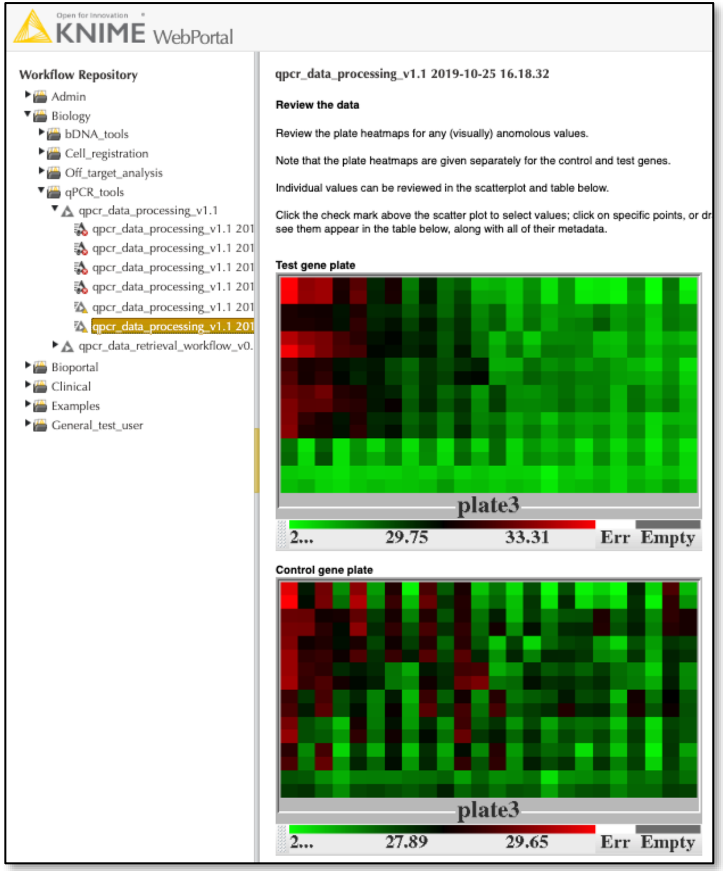
- **Ready-made features of KNIME**
  - **Components** with interactive JS graphs and tables for **guided metadata selection** & examining data
  - **Context properties** extraction (who, what, when...)
  - **Controlled metadata** = flow variables = protocol def.
- **Advanced statistical features (R)**
  - Linear mixed effect (LME) models
  - Robust fitting, outlier handling
  - R:ggplot2 graphics
- **AWS & Db connectivity**
  - Metadata Selection, compound validation & Db write

Open for Innovation
KNIME
R
amazon web services
JS



200K observations processed over prior 24 months

Perform interactive visual examination of the data (1)



Workflow Repository

- Admin
- Biology
  - bDNA_tools
  - Cell_registration
  - Off_target_analysis
  - qPCR_tools
    - qpcr_data_processing_v1.1
      - qpcr_data_processing_v1.1 201
      - qpcr_data_processing_v1.1 201
      - qpcr_data_processing_v1.1 201
      - qpcr_data_processing_v1.1 201
      - qpcr_data_processing_v1.1 201
      - qpcr_data_processing_v1.1 201
    - qpcr_data_retrieval_workflow_v0.
- Bioportal
- Clinical
- Examples
- General_test_user

qpcr_data_processing_v1.1 2019-10-25 16.18.32

Review the data

Review the plate heatmaps for any (visually) anomolous values.

Note that the plate heatmaps are given separately for the control and test genes.

Individual values can be reviewed in the scatterplot and table below.

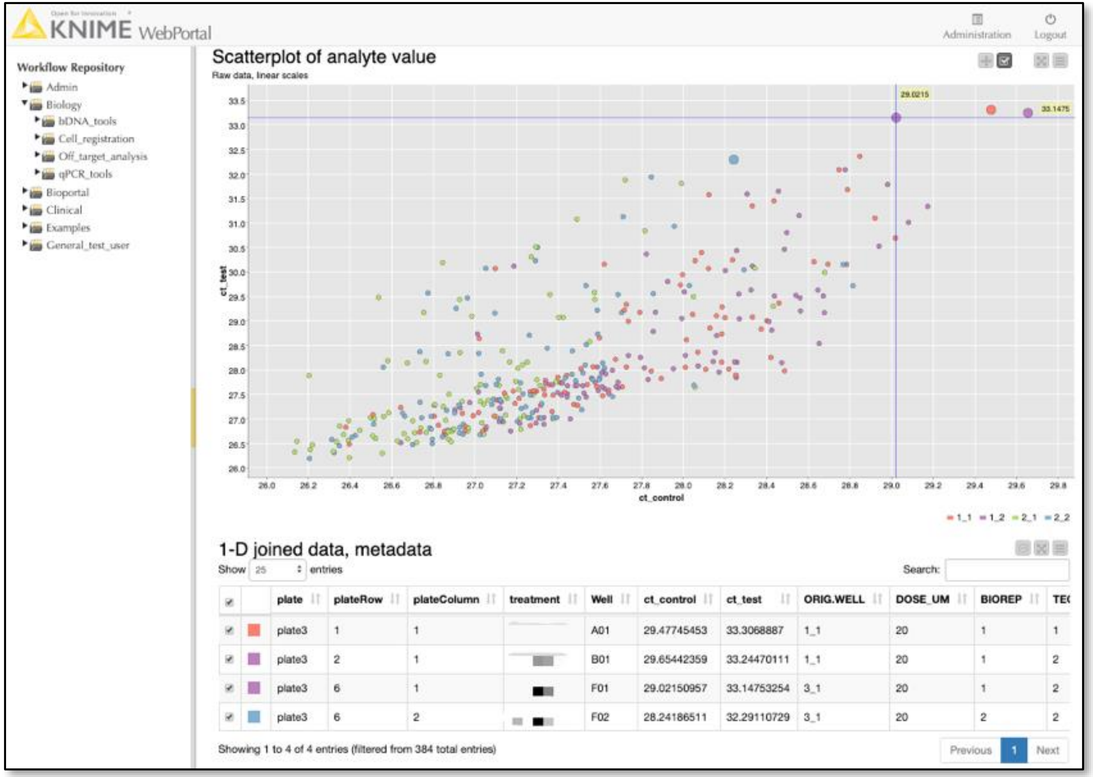
Click the check mark above the scatter plot to select values; click on specific points, or dr see them appear in the table below, along with all of their metadata.

Test gene plate

Control gene plate

- Assess plate heatmaps for:
  - Obvious data deformations
  - Data loss
  - Outliers
  - Systematic column- or row-based effects
- Users can request assistance or insights by sending DS the job link

Perform interactive visual examination of the data (2)



Workflow Repository

- Admin
- Biology
  - bDNA_tools
  - Cell_registration
  - Off_target_analysis
  - qPCR_tools
- Bioportal
- Clinical
- Examples
- General_test_user

Scatterplot of analyte value

Raw data, linear scales

1-D joined data, metadata

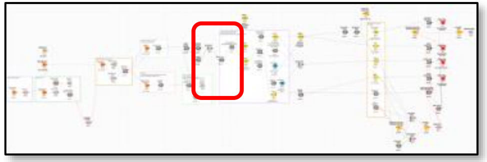
Show 25 entries

Search:

☑		plate	plateRow	plateColumn	treatment	Well	ct_control	ct_test	ORIG.WELL	DOSE_UM	BIOREP	TE
☑	🟥	plate3	1	1		A01	29.47745453	33.3068887	1_1	20	1	1
☑	🟪	plate3	2	1		B01	29.65442359	33.24470111	1_1	20	1	2
☑	🟪	plate3	6	1		F01	29.02150957	33.14753254	3_1	20	1	2
☑	🟦	plate3	6	2		F02	28.24186511	32.29110729	3_1	20	2	2

Showing 1 to 4 of 4 entries (filtered from 384 total entries)

Previous 1 Next

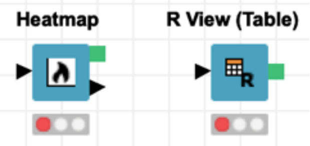


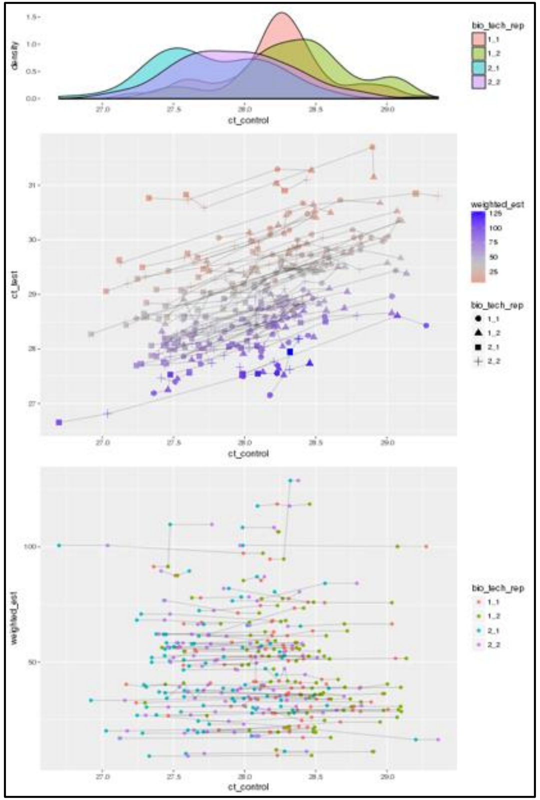
- Interactively view scatterplot of test and control gene raw data and its connected metadata in tabular form using KNIME js.nodes
- Generate initial impressions of data quality and likely outliers, reasons behind, etc

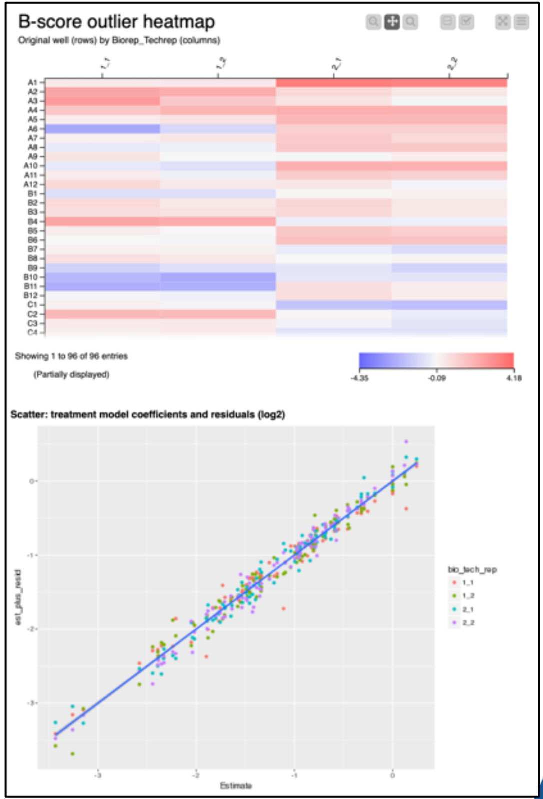
Scatter Plot
Table View

Review model fit & outlier adjustments

- Examine how the rLME model dealt with outliers & impacted the final parameters
- Differentiate between technical and biological variation
- Uses combination of:
  - R:ggplot2 graphics
  - KNIME js.nodes for interactive visualization of outliers



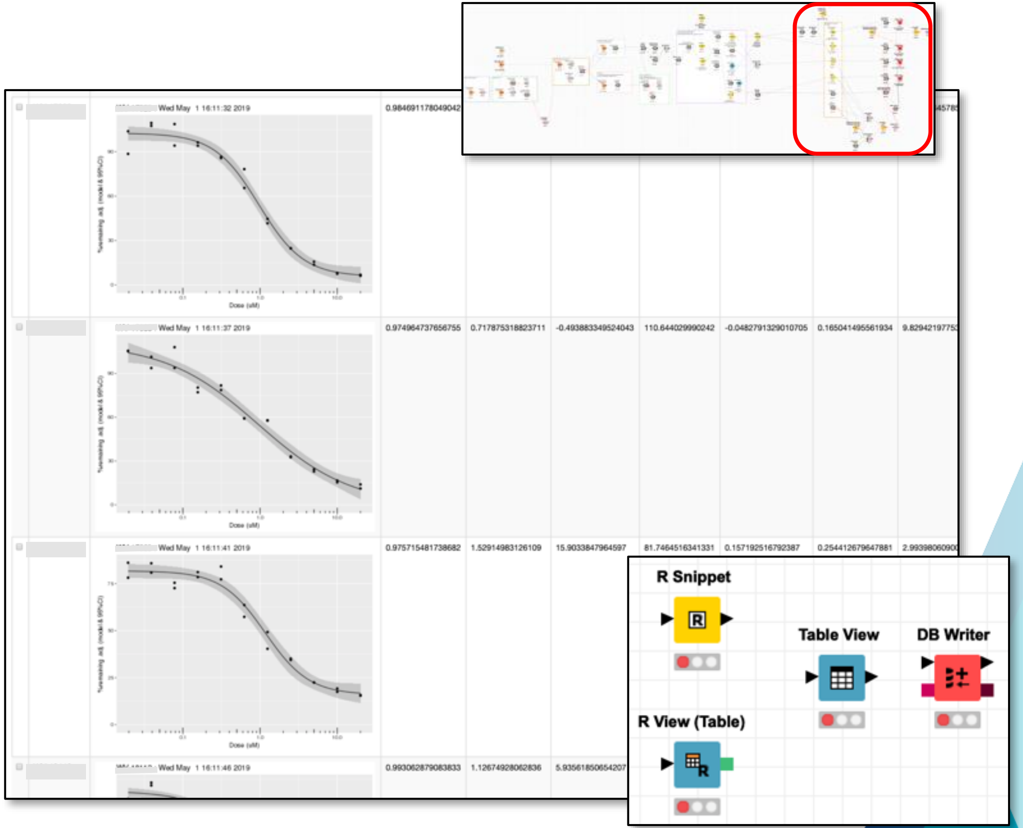




WAVE™ LIFE SCIENCES

Dose-response Kinetics and Database Commit

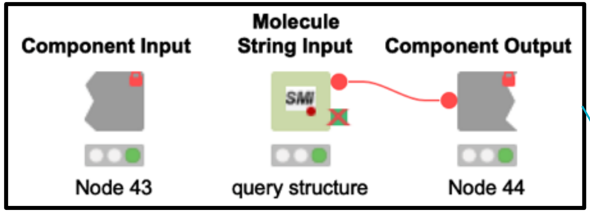
- 4-parameter dose-response kinetics and graphing using *R:drc* and *R:ggplot2*
- If-switch on > n-independent doses
- Review and QC using JS Table View
  - Graphs
  - Parameters
  - Confidence intervals
  - Goodness-of-fit stats
- Commit to AWS-resident Db

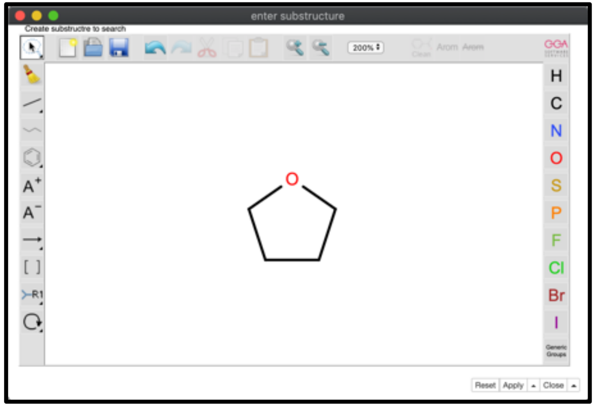


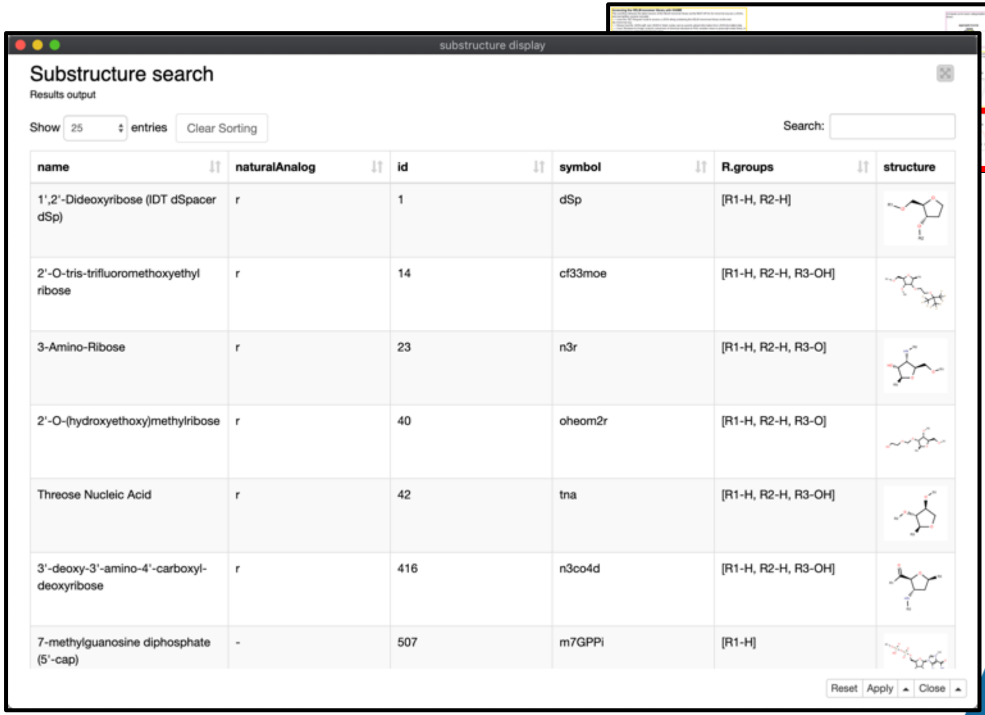
□	[redacted]	[redacted] Wed May 1 16:11:32 2019	0.984691178049042						
□	[redacted]	[redacted] Wed May 1 16:11:37 2019	0.974964737656755	0.717875318823711	-0.493883349524043	110.644029990242	-0.0482791329010705	0.165041495561934	9.82942197753
□	[redacted]	[redacted] Wed May 1 16:11:41 2019	0.975715481738682	1.52914983126109	15.9033847964597	81.7464516341331	0.157192516792387	0.254412679647881	2.99398060900
□	[redacted]	[redacted] Wed May 1 16:11:46 2019	0.993062879083833	1.12674928062836	5.93561850654207				

WAVE™
LIFE SCIENCES

Cheminformatics: substructure search







Substructure search

Results output

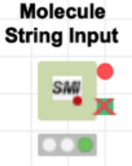
name	naturalAnalog	id	symbol	R.groups	structure
1',2'-Dideoxyribose (IDT dSpacer dSp)	r	1	dSp	[R1-H, R2-H]	
2'-O-tris-trifluoromethoxyethyl ribose	r	14	cf33moe	[R1-H, R2-H, R3-OH]	
3-Amino-Ribose	r	23	n3r	[R1-H, R2-H, R3-O]	
2'-O-(hydroxyethoxy)methylribose	r	40	oheom2r	[R1-H, R2-H, R3-O]	
Threose Nucleic Acid	r	42	tna	[R1-H, R2-H, R3-OH]	
3'-deoxy-3'-amino-4'-carboxyl-deoxyribose	r	416	n3co4d	[R1-H, R2-H, R3-OH]	
7-methylguanosine diphosphate (5'-cap)	-	507	m7GPPi	[R1-H]	

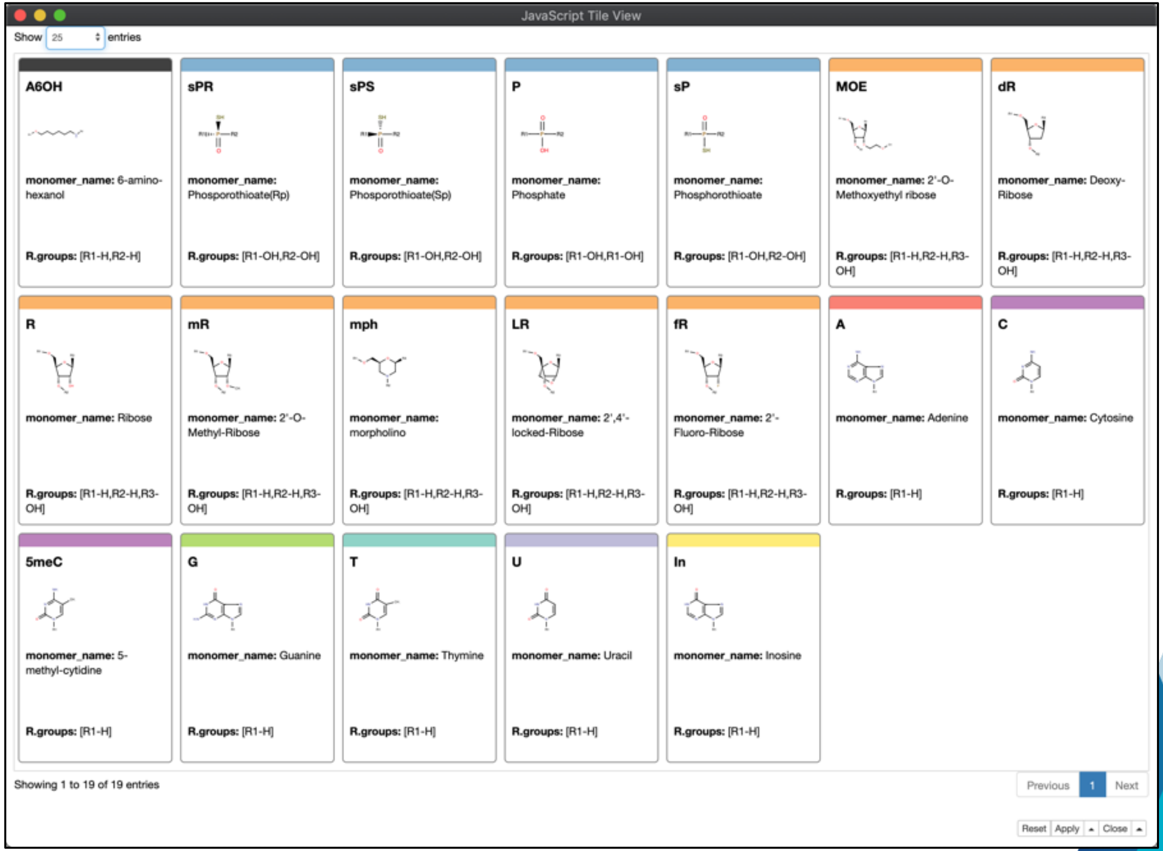
- …

WAVE™ LIFE SCIENCES

Chemical display, categorization, selection

- **Composed** our own monomer library
  - **Used exisiting** minimal set (providing majority coverage)
  - **Supplemented** with proprietary monomers created using the *Molecule String Input* node, complete with R-groups
- Data Science/Chemistry **collaboration** framework!





Show 25 entries

A6OH
monomer_name: 6-amino-hexanol
R.groups: [R1-H,R2-H]

sPR
monomer_name: Phosporothioate(Rp)
R.groups: [R1-OH,R2-OH]

sPS
monomer_name: Phosporothioate(Sp)
R.groups: [R1-OH,R2-OH]

P
monomer_name: Phosphate
R.groups: [R1-OH,R1-OH]

sP
monomer_name: Phosphorothioate
R.groups: [R1-OH,R2-OH]

MOE
monomer_name: 2'-O-Methoxyethyl ribose
R.groups: [R1-H,R2-H,R3-OH]

dR
monomer_name: Deoxy-Ribose
R.groups: [R1-H,R2-H,R3-OH]

R
monomer_name: Ribose
R.groups: [R1-H,R2-H,R3-OH]

mR
monomer_name: 2'-O-Methyl-Ribose
R.groups: [R1-H,R2-H,R3-OH]

mph
monomer_name: morpholino
R.groups: [R1-H,R2-H,R3-OH]

LR
monomer_name: 2',4'-locked-Ribose
R.groups: [R1-H,R2-H,R3-OH]

fR
monomer_name: 2'-Fluoro-Ribose
R.groups: [R1-H,R2-H,R3-OH]

A
monomer_name: Adenine
R.groups: [R1-H]

C
monomer_name: Cytosine
R.groups: [R1-H]

5meC
monomer_name: 5-methyl-cytidine
R.groups: [R1-H]

G
monomer_name: Guanine
R.groups: [R1-H]

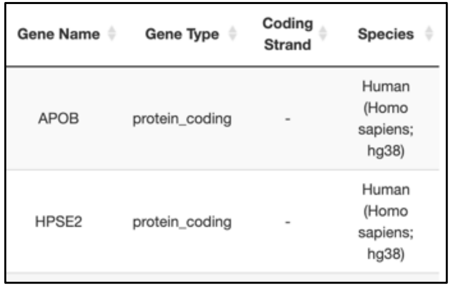
T
monomer_name: Thymine
R.groups: [R1-H]

U
monomer_name: Uracil
R.groups: [R1-H]

In
monomer_name: Inosine
R.groups: [R1-H]

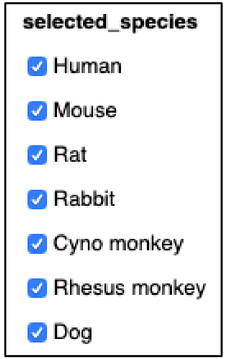
Showing 1 to 19 of 19 entries

Bioinformatics: gene-species alignment network



Gene Name	Gene Type	Coding Strand	Species
APOB	protein_coding	-	Human (Homo sapiens; hg38)
HPSE2	protein_coding	-	Human (Homo sapiens; hg38)

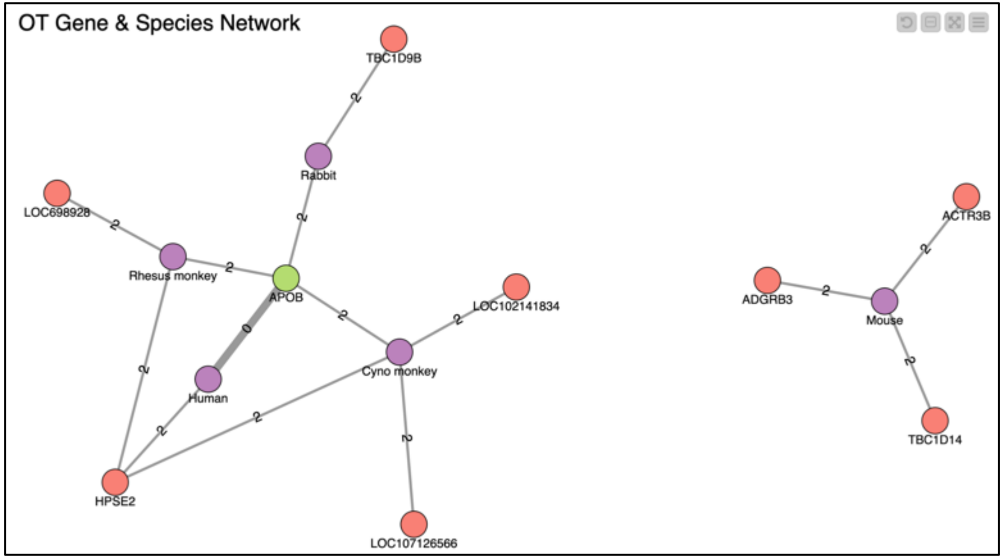
Bowtie gene-species alignment table w/ mismatch count



selected_species

- ☑ Human
- ☑ Mouse
- ☑ Rat
- ☑ Rabbit
- ☑ Cyno monkey
- ☑ Rhesus monkey
- ☑ Dog

Species list



- Feeds from the bowtie alignment table; users select species of interest
- Webportal view using Network Viewer JS node
- Establishes relationship of sequence with multiple species, inversely edge-weighted by mismatch count, including off-target hypotheses