PhenoGen Informatics
The site for quantitative genetics of the transcriptome.

PhenoGen: Building a REST API

Spencer Mahaffey, Paula L. Hoffman PhD, Laura Saba PhD, Boris Tabakoff PhD
Skaggs School of Pharmacy and Pharmaceutical Sciences and School of Medicine, University of Colorado Anschutz Medical Campus, Aurora CO 80045

University of Colorado Anschutz Medical Campus

Abstract

The PhenoGen website (https://phenogen.org) has been operating for the last 16 years providing a resource for quantitative transcriptome data for RI panels of rats and mice. Recent development efforts have focused on providing RNA-Seq data with user friendly tools on the web interface. PhenoGen currently has more than 5TB of raw and 35TB of processed sequencing data on ribosome depleted totalRNA and smallRNA for 43 strains of the Hybrid Rat Diversity Panel in whole brain, heart, liver, and kidney, much of which is accessible through the web interface. The representational state transfer application programming interface (REST API) will provide direct access to data and tools to make functions accessible for advanced users or enable programmatic access for other services. The REST API is the next step in the move to a new architecture, which along with the new domain name, is now hosted in the cloud to provide robust performance and more flexible data access. The API is implemented using AWS API Gateway connected to python AWS Lambda functions for modularity, scalability, and flexibility. As many programming languages support REST APIs, this will enable direct access to the RNA-Seq and microarray data in other environments for reuse and further analysis. The goal is to provide a central resource for versions of the data set with access to various steps in the processing pipeline, and to provide programmatic access to PhenoGen's unique set of data and tools that are centered around analyzing differences in and regulation of the transcriptome across the HRDP and across tissues. The API will extend beyond data to provide meta-data related to documenting how versions of the dataset were generated, to generating images that summarize expression, QTLs, alternative splicing, among others.
PhenoGen is supported by NIAAA (R24AA013162), NIDA (P30DA044223), Banbury Fund, and computing resources from the National Supercomputing Center & Dedicated Research Network, UNLV.

Background

The full HRDP when completed will be a combination of recombinant inbred (HXB/BXH and FXLE/LEXF) and classic inbred strains of rats that optimizes genetic mapping power and precision. It will provide a genetically stable and diverse population lending itself to multi-omics studies. As can be seen in table 1 to the right, currently about half of the final 96 HRDP strains have RNA sequencing of ribosomal RNA depleted total RNA data from two tissues. The PhenoGen website (https://phenogen.org) provides data and tools to apply both phenotype-driven analyses and gene driven analyses to understand the contribution that a genetic network can make to a phenotype. The site provides graphics and tables to explore genetic predisposition of the HRDP population to any new phenotype measured on all or part of the population. The tools used for this range from gene co-expression networks to miRNA target predictions to gene or transcript expression of novel and annotated genes/transcripts. A brain and liver transcriptome reconstruction has been completed across the currently available HRDP strains. The REST API will support retrieving these data programmatically to enable use of the data on other sites or by advanced users developing their own tools/pipelines.

With these tools you can explore questions like:

- What tissues express this gene/transcript?
- What region of the genome is associated with regulating expression?
- Do any miRNAs target this gene?
- What genes have expression that is correlated with expression of this gene? What are the eQTLs for this module? Do any miRNAs target genes in the module?

1.

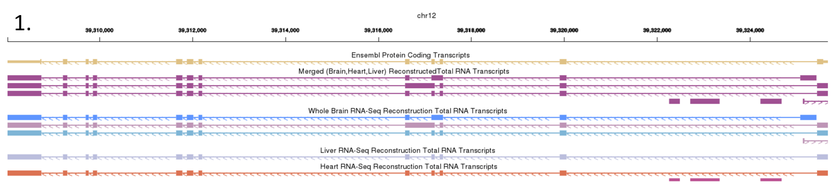


Figure 1: Browser image showing the reconstructed transcripts for each tissue. The browser is focused on providing access to data/tools. The following images and data are all directly accessible from the browser without navigating away from the browser page.

2a. 2b. 2c.

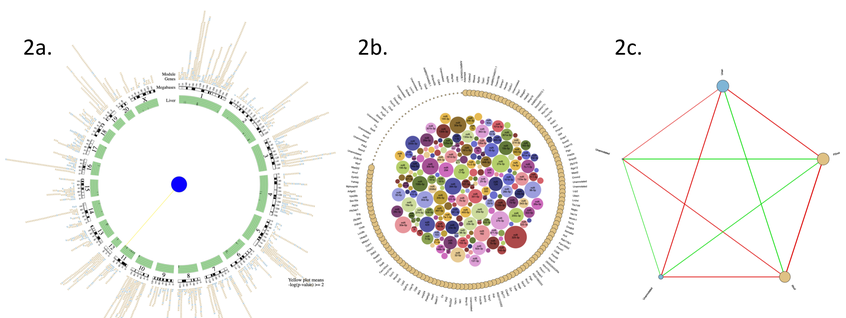


Figure 2a-c: Weighted Gene Co-expression Network Aanalysis(WGCNA) Module graphics. 2a. (Array-Liver) shows module gene locations and Eigengene eQTLs. 2b. (Array-Liver) shows miRNAs targeting genes in the module with size of miRNA related to number of genes targeted. 2c. (Seq-Brain) shows the correlations between genes in the module with green(+)/red(-) correlations

3a. 3b.

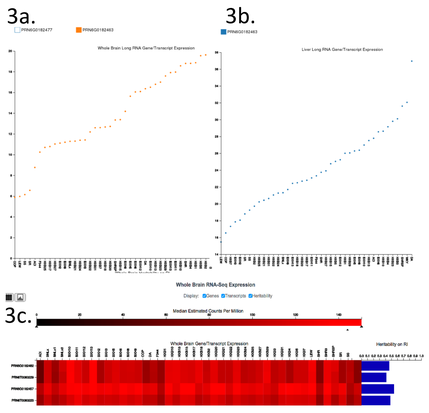


Figure 3a-c: Brain and Liver RNA-Seq expression summaries 3a. Whole Brain gene level expression across strains. 3b. Liver gene level expression across strains. 3c. Gene/transcript expression across strains in a region.

4.

Figure 4: Above is a Circos plot of Brain and Liver RNA-Seq eQTLs for P2rx4. This shows eQTLs for the gene in both tissues. It can be customized to include a specific -log(p-value) threshold and to only include specific chromosomes.

HRDP RNA-Seq Data Available

Table 1:Brain and Liver Transcriptome Data Collected (RNA-Seq of small RNA and total long RNA)

HXB1	HXB21	BXH6	LEW/SsNHsd
HXB2	HXB22	BXH8	SHRSP
HXB3	HXB23	BXH9	WKY/NCrl
HXB4	HXB24	BXH10	ACI/SegHsd
HXB5	HXB25	BXH11	Dark Agouti
HXB7	HXB26	BXH12	SS/JrHsd
HXB10	HXB27	BXH13	F344/NHsd
HXB13	HXB29	BN-Lx/Cub*	F344/Stm
HXB15	HXB31	SHR/Ola*	LE/Stm
HXB17	BXH2	SR/JrHsd	
HXB18	BXH3	COP/CRL	
HXB20	BXH5	F344/NCr	

*Heart Left Ventricle is also available for the HXB/BXH RI panel parents

Table 2: Summary of RNA-Seq Data Collected

Strains	Tissue	Sex	Number of Biological Replicates Per Strain	Number of Paired-End Reads (rRNA-depleted total RNA)	Number of Single- End Reads (small RNA)
SHR and BNLx	Brain	male	3	645 million	96 million
SHR and BNLx	Brain	female	4	982 million	297 million
SHR and BNLx	Liver	male	3	583 million	342 million
SHR and BNLx	Heart	male	4	790 million	300 million
45 HRDP Strains	Brain	male	4(RI) 3(Inbred)	17.0 billion	5.0 billion
45 HRDP Strains	Liver	male	4 (3-UCD 1-Euratrans)	15.8 billion	5.1 billion
30 HRDP Strains	Heart	male	1(RI) 5(Parentals)	5.5 billion	-

Example API functions for Gene Level Data

https://rest.phenogen.org/genes
- Gene Report provides a quick summary of data related to the gene.

https://rest.phenogen.org/genes/idLookup
- Many of the tools require specific ID types. You can use this function to start with any supported ID and look up the required types. Required parameters (Gene ID, [List of target ID types]). Without target list returns a list of target types.

https://rest.phenogen.org/genes/eQTLs
- eQTL look up gene level eQTLs with specific tissues/pvalue cutoffs/data types(RNA-Seq or Array)/genomeVersion.

https://rest.phenogen.org/genes/WGCNA
- Look for WGCNA modules containing the gene of interest. Then look at connectivity between genes in the module, module Eigengene QTLs, miRNA targeting of contained genes, Gene Ontology summaries, and other modules correlated to this module.

https://rest.phenogen.org/genes/expression
- Retrieve normalized RNA-Seq/Array expression of gene level data across strains or additionally transcript level for RNA-Seq (future development will expand to raw data, individual sample level expression).

https://rest.phenogen.org/genes/miRNATarget
- Retrieve multiMiR results for a gene or miRNA. multiMiR provides miRNA target data from 11 different miRNA-Gene target databases including both validated and prediction sources. MultiMiR is already available as an R package but this will provide functionality accessible outside R.

Example API workflow for P2rx4

1. Look up required Gene IDs using gene symbol P2rx4
https://rest.phenogen.org/genes/idLookup
Parameters={gene:"P2rx4",target:"Ensembl ID,PhenoGen ID"}
Result={content:[{target:"Ensembl ID" , id:"ENSRONG00000001300"} , {target:"Ensembl ID", id:"ENSRNOT00000001752"} , {target="PhenoGen ID", id:"PRN6G0182463"}, …]}

2. Get gene report
https://rest.phenogen.org/genes
Parameters={gene:"ENSRONG00000001300"}
Result={content:{geneSymbol:"P2rx4",location:"chr12:39,308,025-39,325,673",description:"purinergic receptor P2X 4",strand:"-",exonicVariants:[{strain:"SHR/OlaPrin",SNPs:7,INDELs:2}],transcripts:[{trxID:"ENSRNOT00000001752"},{trxID:"PRN6T0080232",match:"Transcript Match: ENSRNOT00000001752 12 Perfect Exon Matches"}…],…}}

3. Look at gene expression
https://rest.phenogen.org/genes/expression
Parameters={gene:"PRN6G0182463",type:"seq",level:"gene",genome:"rn6",version:"3.0",tissue:"Brain"}
Result={content:{values:[{gene:"PRN6G0182463", expression:[{strain:"ACI", mean: 11.7861}, {strain:"BNLx",mean: 13.3047} ,…], transcripts: [transcript:"" ,expression: [{},{},…]]}]}}

4. Look at gene eQTLs from RNA-Seq and/or Arrays
https://rest.phenogen.org/genes/eQTLs
Parameters={gene:"PRN6G0182463"}
Result={Gene: "PRN6G0182463", DataSource: "seq", P-valueCutoff: 2, Tissues: "Brain,Liver", GenomeVersion: "rn6",
eQTLs: [{SNP_ID:"SHRSPc67b05_r1_216", CHROMOSOME:"15", SNP_START:88429288, TISSUE:"Whole Brain", PVALUE:2.2335871529},…] }

5. Look at WGCNA
https://rest.phenogen.org/genes/WGCNA
Parameters={gene:"PRN6G0182463"}
Result={modules:[{datasetID:6,module:"lightblue1.3",geneMatchCount:1,geneCount:5}]}
https://rest.phenogen.org/genes/WGCNA
Parameters={datasetID:6 ,module:"lightblue1.3",type:"connectivity"}
Result={module:{geneList: [{gene:"PRN6G0182622"} ,{gene:" PRN6G0182463"}, {gene"ENSRNOG0000000042607"} ,…], links:[{id1:" PRN6G0182622", id2:" PRN6G0182463",value:0.85},…]}}

6. Save a Gene List and continue analysis (future feature)
POST -> https://rest.phenogen.org/geneList
Parameters={UID:"7Z2049-832D033457-02A394",Name:"P2rx4 module genes", Organism:"Rn", geneList: [{gene:"PRN6G0182622"} ,{gene:" PRN6G0182463"}, {gene"ENSRNOG0000000042607"},…] }
Result={status:"Successfully saved gene list: P2rx4 module genes",id:432049}

Example R code – Gene eQTLs

Get eQTLs for P2rx4 from RNA-Seq data using default parameters. If desired, additional parameters can be specified in the query=list()
Required parameters:
gene – gene id - phenogen id for RNA-Seq or Transcript cluster id for Arrays
Optional parameters:
type(seq or array) – specifies if you want eQTLs calculated from array expression data or RNA-Seq expression data – default seq
cutoff – specifies the –log(p-value) cutoff – default 2.0
tissues – comma separated list of tissues – default Brain,Liver(seq) Brain,Heart,Liver,BAT(array)
genomeVersion – genome version to use – default rn6(rat) mm10(mouse)

```r
#Load httr for http requests
library(httr)
#Load jsonlite to parse json
library(jsonlite)

#Call the API at https://rest.phenogen.org/genes/eQTL?gene=PRN6G0182463
result<-GET("https://rest.phenogen.org/", path = "genes/eQTL", query = list(gene = "PRN6G0182463"))

# Get JSON(JavaScript Object Notation) content
# JSON provides a way to encode data with two main structures key, value pairs and lists
# It is easy to write/parse and can be nested to create a hierachical structure
# initially JSON will be the default format, future development will allow request to
# receive HTML/CSV/Txt formatted output
result_text<-content(result,"text")

#Parse JSON and flatten(remove the hierachical structure)
result_json <- fromJSON(result_text, flatten = TRUE)

#Get the eQTL list and save as a data.frame for further analysis
result_df <- as.data.frame(result_json['eQTLs'])
```

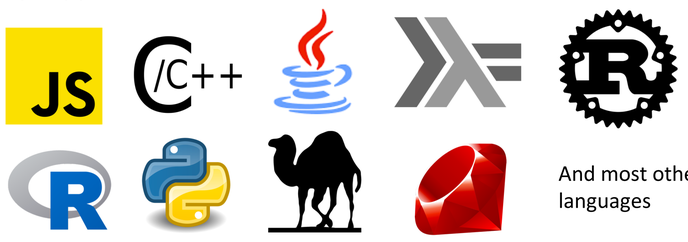
result_df - eQTL results table

	eQTLs.SNP_ID	eQTLs.CHROMOSOME	eQTLs.SNP_START	eQTLs.TISSUE	eQTLs.NegLogPvalue
1	SHRSPc67b05_r1_216	14	88429288	Whole Brain	2.2335871529
2	SHRSPc37e06_r1_616	15	104783292	Whole Brain	2.0525662781
3	DahlSa03h04_r1_204	17	25304710	Liver	2.1030881592
4	WKYOa42h01_r1_538	2	12017378	Whole Brain	2.0122443833
5	Cpn_1201127302	1	213743164	Whole Brain	2.0666645899
6	SHRSPa24f09_r1_1037	2	28298336	Liver	2.4732731327
7	gko-16g21_rp2_b1_529	20	20498812	Whole Brain	2.0461437018
8	SHRSPc04b05_r1_675	2	2609022	Whole Brain	2.7525177393
9	J488183	14	88137210	Whole Brain	2.2335871529
10	SHRSPc42h08_r1_389	2	11068172	Whole Brain	2.3821612523
11	WKY-G-j-49h07_f1_819	2	8939880	Whole Brain	2.5867002359
12	WKY-G-i-40d02_f1_330	2	12605368	Whole Brain	2.0886825577

Returned JSON includes all parameters(including defaults) used for the search.
{Gene: "PRN6G0182463",
DataSource: "seq",
P-valueCutoff: 2,
Tissues: "Brain,Liver",
GenomeVersion: "rn6",
eQTLs: [{SNP_ID:"SHRSPc67b05_r1_216", CHROMOSOME:"15",
SNP_START:88429288, TISSUE:"Whole Brain", NegLogPvalue:2.2335871529},…] }

Language Support

Provided over HTTP so any language with HTTP client support will easily support the REST API.



And most other languages

Roadmap

Timeline to rollout various API functions/data:

June 2019 – Genes
September 2019 – Genome/Transcriptome Data Browser Track/Table data
October 2019 – IPFS for file distribution, REST API for metadata and file lookup
November 2019 – Gene Lists and Run analyses/tools
December 2019
- HRDP v4 – will include Brain/Liver from LE-Stm/F344-Stm parents of the FXLE/LEXF RI panel.
- CirRNA predictions/array data(See related poster in this session "Array and Computational Methods to investigate a role for circRNAs in complex traits in the Hybrid Rat Diversity Panel")
- REST API select return format(JSON/HTML/Txt)

Spring 2020 – Graphics
- Circos Plots
- Browser Images
- WGCNA Module Visualizations

New Advanced Tools to be developed with REST API as backend and then build web UI using the REST API.
- Compare Modules/Tissues
- Run supported third party programs to analyze data
- Find QTL overlap from QTLs from various sources.

Conclusion

Phenogen.org provides extensive data on the HRDP transcriptome, which at the gene level is available through the REST API. Continued development will make remaining data and tools available to access programmatically in addition to the web user interface. This will include generating some of the graphics available through the web interface or creating gene lists from other function calls and proceeding to analyze the gene lists. This will provide a comprehensive set of tools to access data and perform analyses on the site. The API interface will allow for development of software development kits (SDKs) in other languages to further capitalize on the data sets and tools developed for PhenoGen. This will further make data readily available for other resources providing unique complimentary tools.

Acknowledgements

This project was supported by NIAAA (R24AA013162, NIDA (P30DA044223), and Banbury Fund.
Computing Resources: National Supercomputing Center & Dedicated Research Network, UNLV

Contact Information

Boris Tabakoff, PhD.
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado Denver Anschutz Medical Campus
Email: Boris.Tabakoff@CUAnschutz.edu
Phone: (303) 724-3668