NIH National Institute on Drug Abuse
Advancing Addiction Science

Genes and Addiction
NIDA Center for GWAS in Outbred Rats

PipeRat

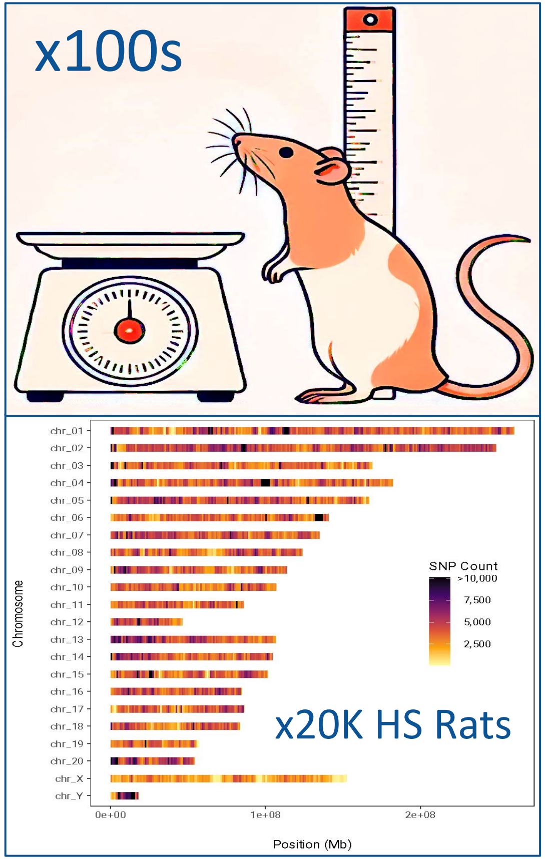
A high-throughput python package to perform and visualize large-scale genetic association analysis

Thiago Sanches, Apurva Chitre, Oksana Polleskaya, Elaine Keug,
Benjamin Johnson, Gravilla Ang, Montana Lara, Abraham Palmer

Department of Psychiatry
University of California San Diego

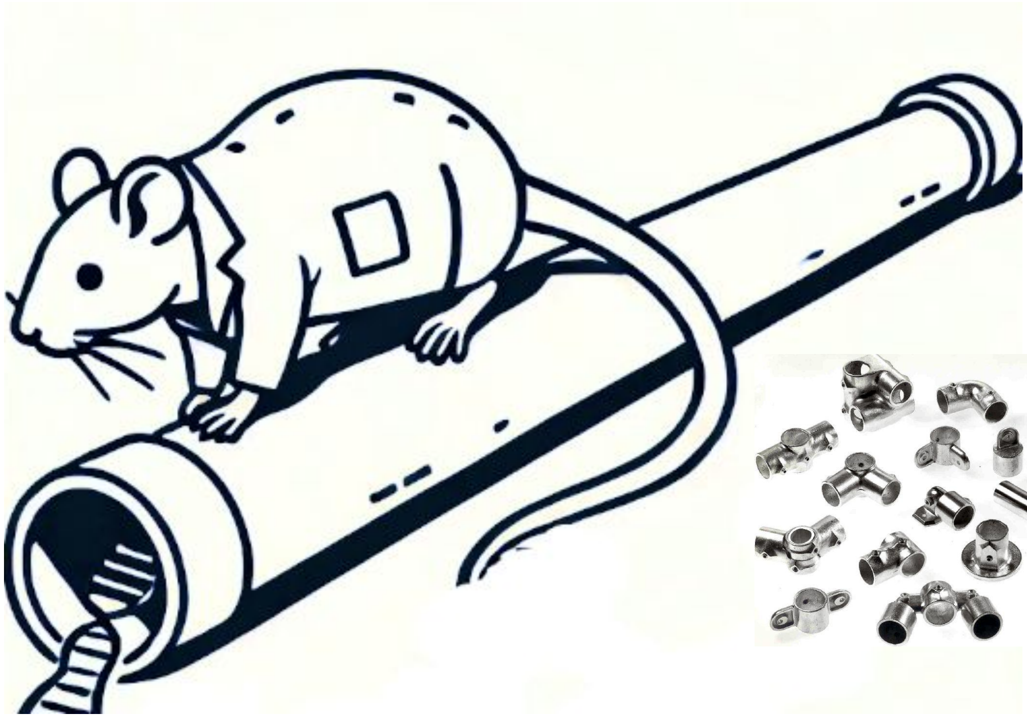
The NIDA Center of Excellence for Genetics, Genomics, and Epigenetics of Substance Use Disorders in Outbred Rats centralizes

- 10s of projects
- 100s of traits
- 10000s of genotypes
- Standardize data analysis



What is the PipeRat?

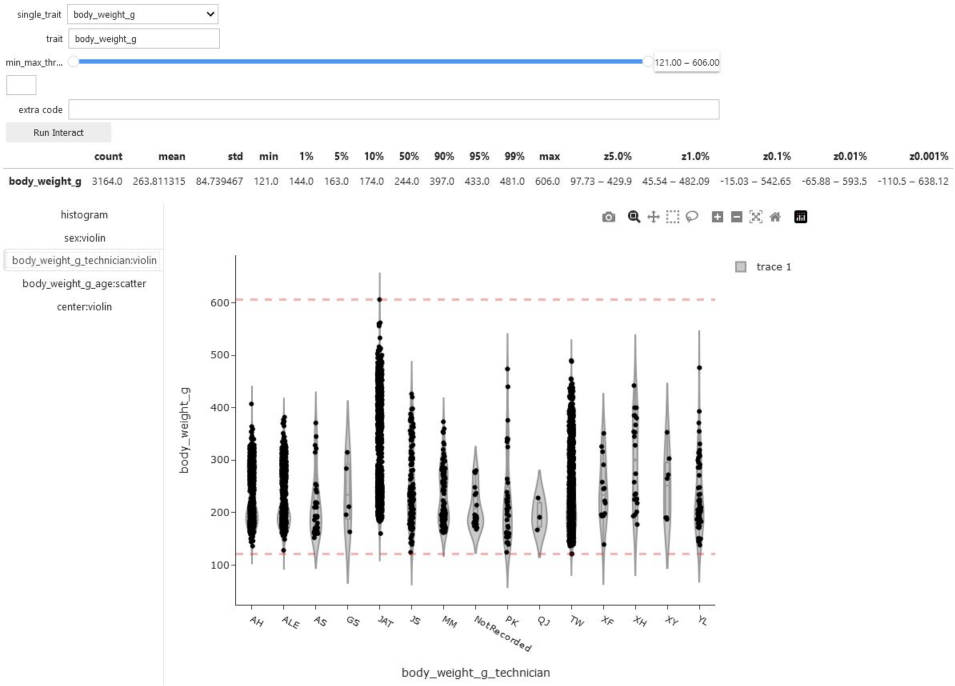
1. Is a python package
    a. To Curate traits
    b. To do Genetic Analysis
        i. for any species
    c. To visualize the results
        i. Manuscript ready figures
    d. To find target genes
    e. To facilitate future analysis



Data curation

- We interactively visualize data with collaborators to curate, filter and log the changes

```
qc = interactive_QC(raw_data=df,
                    data_dictionary=datadic)
qc.QC()
```



single_trait: body_weight_g

trait: body_weight_g

min_max_thr...: 121.00 – 606.00

extra code:

Run Interact

	count	mean	std	min	1%	5%	10%	50%	90%	95%	99%	max	z5.0%	z1.0%	z0.1%	z0.01%	z0.001%
body_weight_g	3164.0	263.811315	84.739467	121.0	144.0	163.0	174.0	244.0	397.0	433.0	481.0	606.0	97.73 – 429.9	45.54 – 482.09	-15.03 – 542.65	-65.88 – 593.5	-110.5 – 638.12

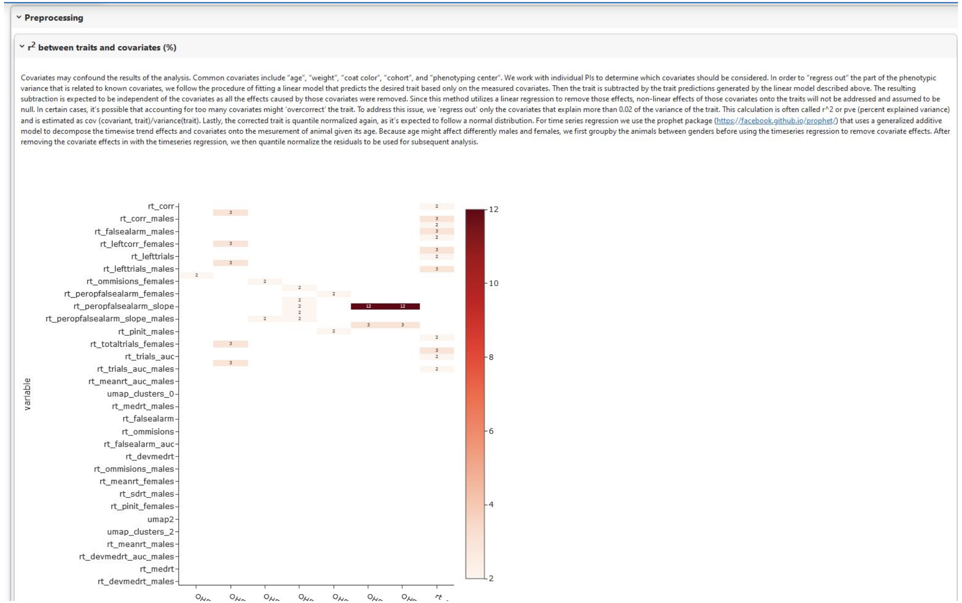
Regressing out covariates

- Autoremoves covariates with low PVE (2% in our case)
- Allows regressing out sites|sexes separately
- Allows time series regressing out
- Output always a normal distribution

```
pipe.regressout_groupby(data_dictionary=dic,
                                   groupby_columns=groupby_animals)

                        or

pipe.regressout_timeseries(data_dictionary= dic,
                                   groupby_columns=groupby_animals)
```



Preprocessing

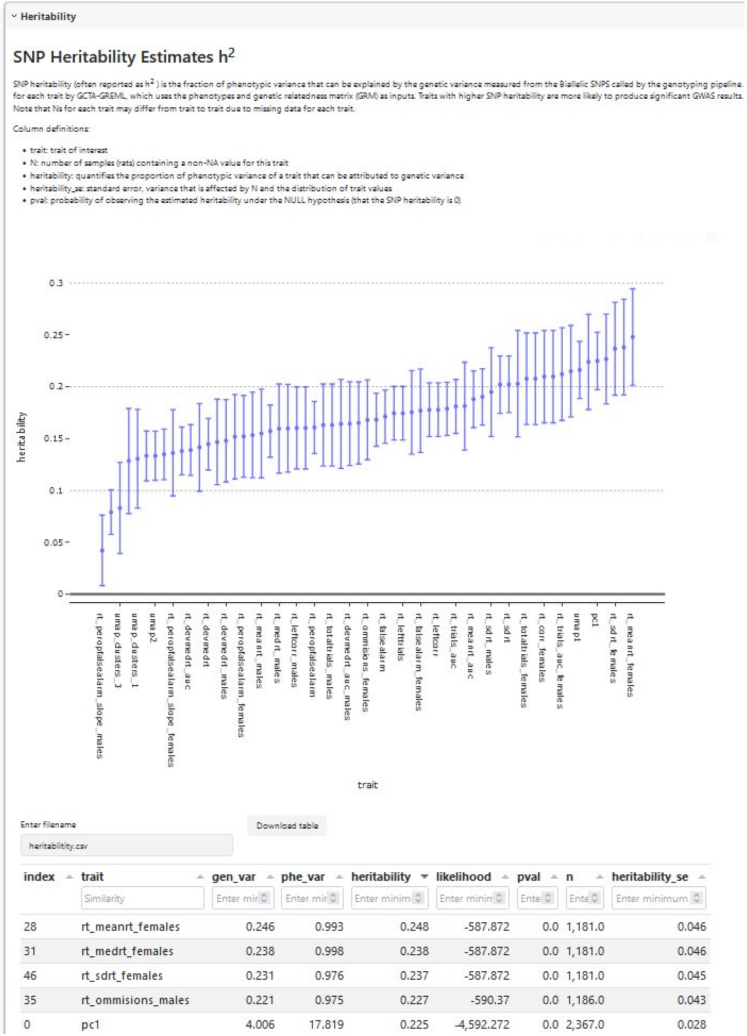
r² between traits and covariates (%)

Covariates may confound the results of the analysis. Common covariates include "age", "weight", "coat color", "cohort", and "phenotyping center". We work with individual PIs to determine which covariates should be considered. In order to "regress out" the part of the phenotypic variance that is related to known covariates, we follow the procedure of fitting a linear model that predicts the desired trait based only on the measured covariates. Then the trait is subtracted by the trait predictions generated by the linear model described above. The resulting subtraction is expected to be independent of the covariates as all the effects caused by those covariates were removed. Since this method utilizes a linear regression to remove those effects, non-linear effects of those covariates onto the traits will not be addressed and assumed to be null. In certain cases, it's possible that accounting for too many covariates might 'overcorrect' the trait. To address this issue, we 'regress out' only the covariates that explain more than 0.02 of the variance of the trait. This calculation is often called r^2 or pve (percent explained variance) and is estimated as cov (covariant, trait)/variance(trait). Lastly, the corrected trait is quantile normalized again, as it's expected to follow a normal distribution. For time series regression we use the prophet package (https://facebook.github.io/prophet/) that uses a generalized additive model to decompose the timewise trend effects and covariates onto the mesurement of animal given its age. Because age might affect differently males and females, we first groupby the animals between genders before using the timeseries regression to remove covariate effects. After removing the covariate effects in with the timeseries regression, we then quantile normalize the residuals to be used for subsequent analysis.

Heritability

- How much genetic relatedness explains the variance in the traits
- GCTA for estimation
- Plotly for plotting

```
pipe.SubsetAndFilter()
pipe.generateGRM()
pipe.snpHeritability()
```



˅ Heritability

SNP Heritability Estimates h^2

SNP heritability (often reported as h^2) is the fraction of phenotypic variance that can be explained by the genetic variance measured from the Biallelic SNPS called by the genotyping pipeline. for each trait by GCTA-GREML, which uses the phenotypes and genetic relatedness matrix (GRM) as inputs. Traits with higher SNP heritability are more likely to produce significant GWAS results. Note that Ns for each trait may differ from trait to trait due to missing data for each trait.

Column definitions:

- trait: trait of interest
- N: number of samples (rats) containing a non-NA value for this trait
- heritability: quantifies the proportion of phenotypic variance of a trait that can be attributed to genetic variance
- heritability_se: standard error, variance that is affected by N and the distribution of trait values
- pval: probability of observing the estimated heritability under the NULL hypothesis (that the SNP heritability is 0)

Enter filename

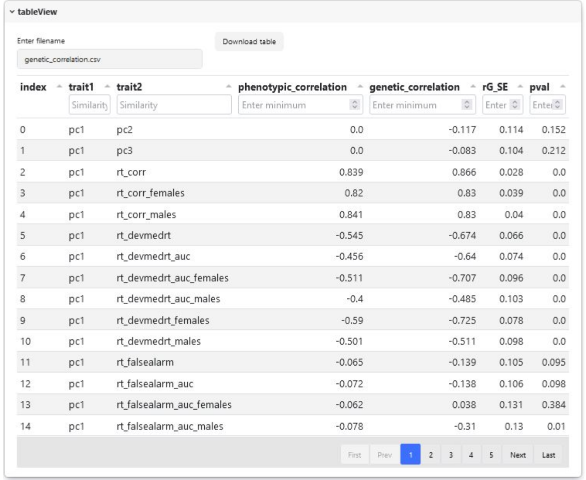
heritablitity.csv

index	trait	gen_var	phe_var	heritability	likelihood	pval	n	heritability_se
	Similarity	Enter mir	Enter mir	Enter minim	Enter minin	Ente	Ente	Enter minimum
28	rt_meanrt_females	0.246	0.993	0.248	-587.872	0.0	1,181.0	0.046
31	rt_medrt_females	0.238	0.998	0.238	-587.872	0.0	1,181.0	0.046
46	rt_sdrt_females	0.231	0.976	0.237	-587.872	0.0	1,181.0	0.045
35	rt_ommisions_males	0.221	0.975	0.227	-590.37	0.0	1,186.0	0.043
0	pc1	4.006	17.819	0.225	-4,592.272	0.0	2,367.0	0.028

Genetic correlation

- Correlation between traits through the GRM
- GCTA greml bivariate
- Correlation without overlap

```
pipe.genetic_correlation_matrix()
```

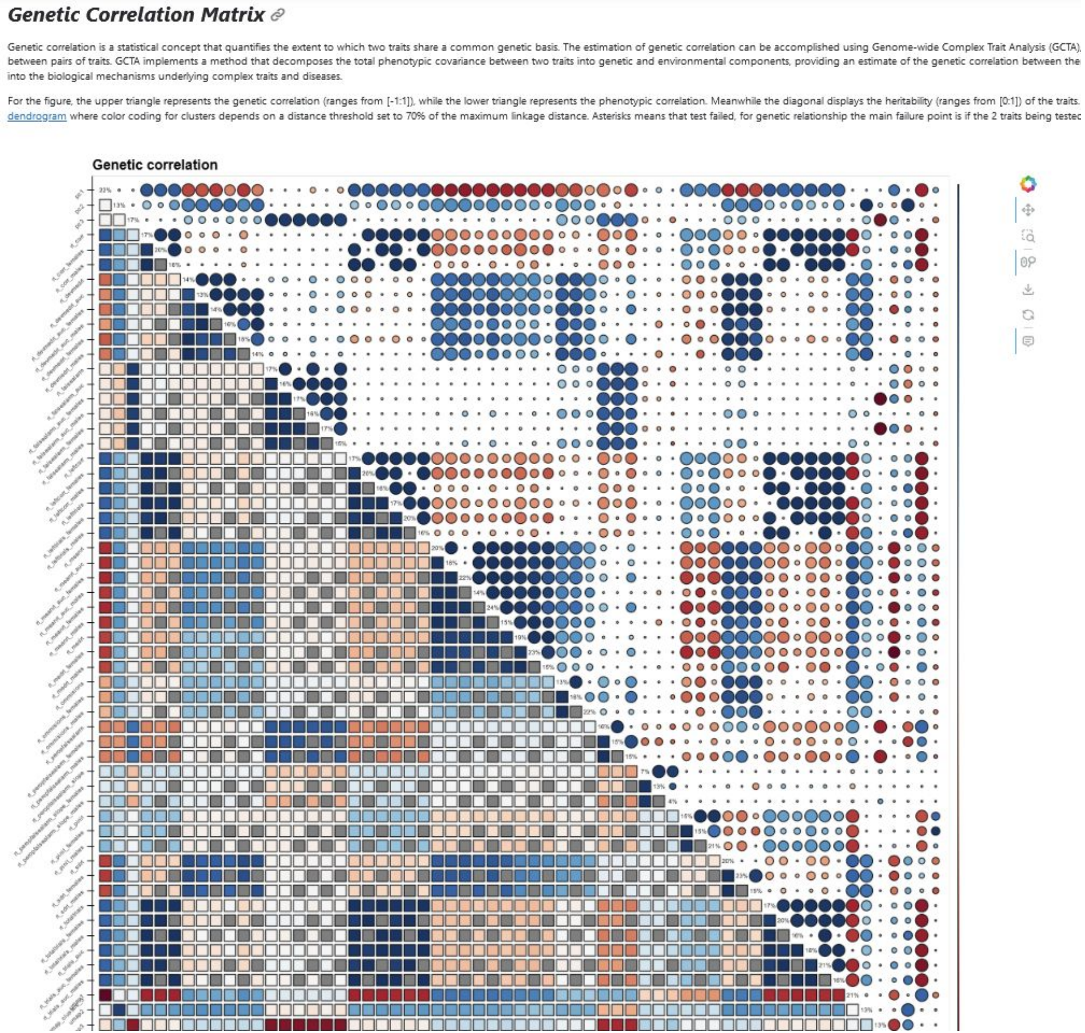


tableView

Enter filename: genetic_correlation.csv

Download table

index	trait1	trait2	phenotypic_correlation	genetic_correlation	rG_SE	pval
	Similarity	Similarity	Enter minimum	Enter minimum	Enter	Ente
0	pc1	pc2	0.0	-0.117	0.114	0.152
1	pc1	pc3	0.0	-0.083	0.104	0.212
2	pc1	rt_corr	0.839	0.866	0.028	0.0
3	pc1	rt_corr_females	0.82	0.83	0.039	0.0
4	pc1	rt_corr_males	0.841	0.83	0.04	0.0
5	pc1	rt_devmedrt	-0.545	-0.674	0.066	0.0
6	pc1	rt_devmedrt_auc	-0.456	-0.64	0.074	0.0
7	pc1	rt_devmedrt_auc_females	-0.511	-0.707	0.096	0.0
8	pc1	rt_devmedrt_auc_males	-0.4	-0.485	0.103	0.0
9	pc1	rt_devmedrt_females	-0.59	-0.725	0.078	0.0
10	pc1	rt_devmedrt_males	-0.501	-0.511	0.098	0.0
11	pc1	rt_falsealarm	-0.065	-0.139	0.105	0.095
12	pc1	rt_falsealarm_auc	-0.072	-0.138	0.106	0.098
13	pc1	rt_falsealarm_auc_females	-0.062	0.038	0.131	0.384
14	pc1	rt_falsealarm_auc_males	-0.078	-0.31	0.13	0.01



Genetic Correlation Matrix

Genetic correlation is a statistical concept that quantifies the extent to which two traits share a common genetic basis. The estimation of genetic correlation can be accomplished using Genome-wide Complex Trait Analysis (GCTA), between pairs of traits. GCTA implements a method that decomposes the total phenotypic covariance between two traits into genetic and environmental components, providing an estimate of the genetic correlation between the into the biological mechanisms underlying complex traits and diseases.

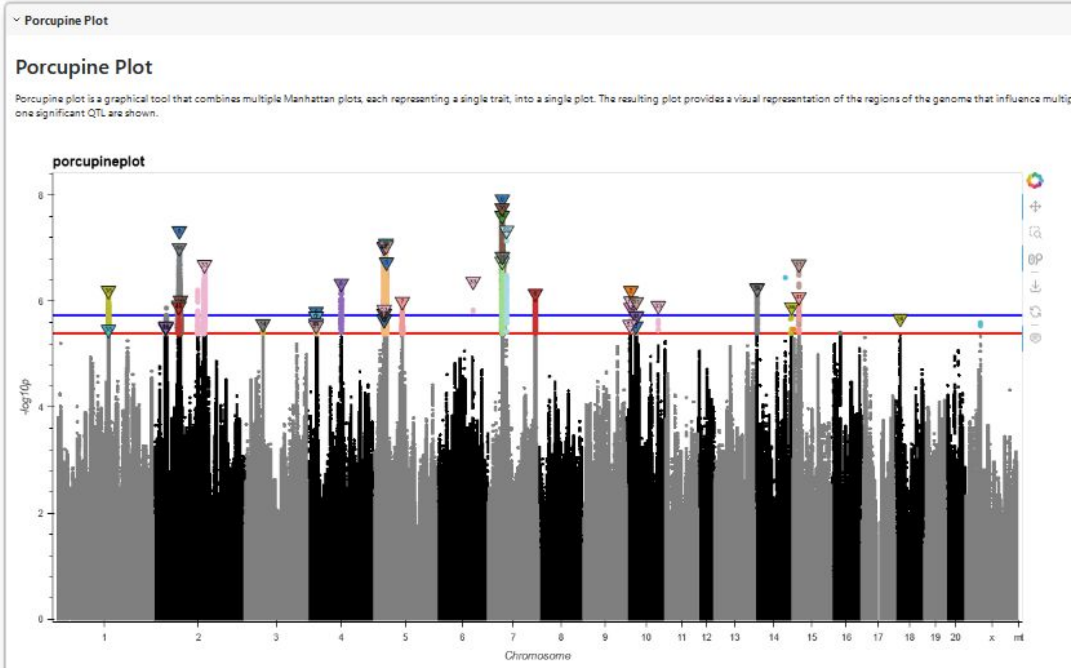
For the figure, the upper triangle represents the genetic correlation (ranges from [-1:1]), while the lower triangle represents the phenotypic correlation. Meanwhile the diagonal displays the heritability (ranges from [0:1]) of the traits. dendrogram where color coding for clusters depends on a distance threshold set to 70% of the maximum linkage distance. Asterisks means that test failed, for genetic relationship the main failure point is if the 2 traits being testec

Genetic correlation

GWAS

- Using GCTA in parallel using dask
- Holoviews for plotting billions of points in seconds

```
pipe.fastGWAS()
pipe.addGWASresultsToDb(researcher=researcher,
                        round_version=round_version,
                        gwas_version=gwas_version)
pipe.callQTLs()
pipe.porcupineplot()
```



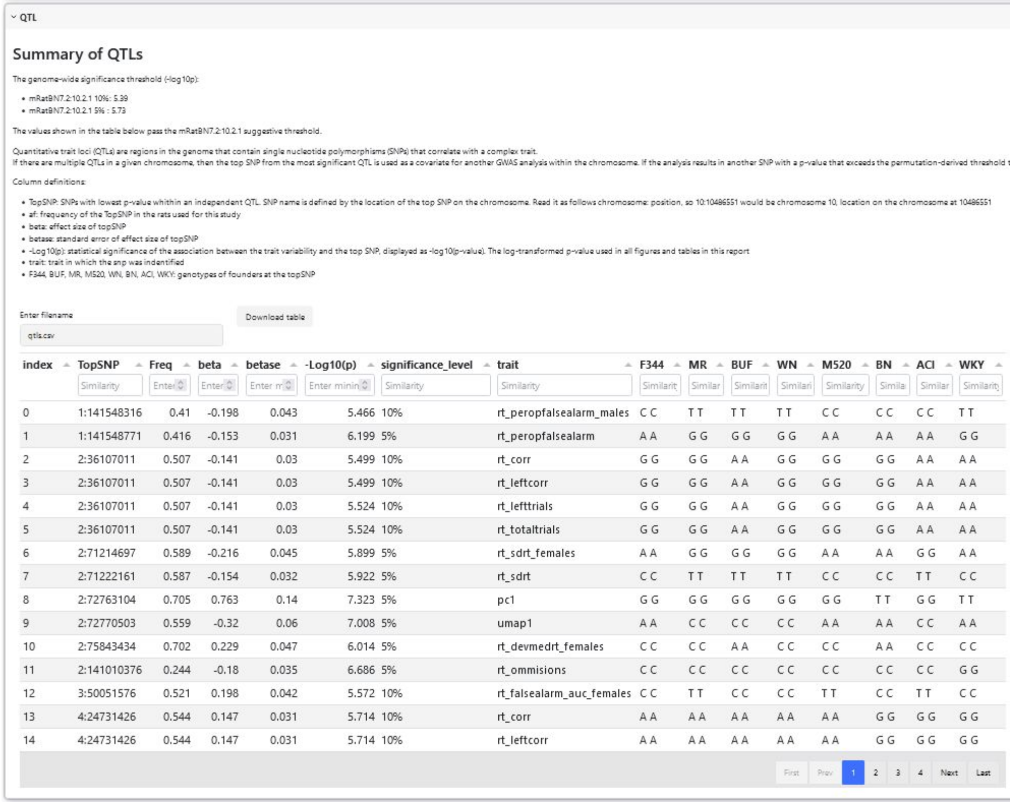
Porcupine Plot

Porcupine plot is a graphical tool that combines multiple Manhattan plots, each representing a single trait, into a single plot. The resulting plot provides a visual representation of the regions of the genome that influence multip
one significant QTL are shown.

GWAS

- Using GCTA in parallel using dask
- Holoviews for plotting billions of points in seconds

```
pipe.fastGWAS()
pipe.addGWASresultsToDb(researcher=researcher,
                        round_version=round_version,
                        gwas_version=gwas_version)
pipe.callQTLs()
pipe.porcupineplot()
```



Summary of QTLs

The genome-wide significance threshold (-log10p):

- mRatBN7.2:10.2.1 10%: 5.39
- mRatBN7.2:10.2.1 5% : 5.73

The values shown in the table below pass the mRatBN7.2:10.2.1 suggestive threshold.

Quantitative trait loci (QTLs) are regions in the genome that contain single nucleotide polymorphisms (SNPs) that correlate with a complex trait.
If there are multiple QTLs in a given chromosome, then the top SNP from the most significant QTL is used as a covariate for another GWAS analysis within the chromosome. If the analysis results in another SNP with a p-value that exceeds the permutation-derived threshold t

Column definitions:

- TopSNP: SNPs with lowest p-value whithin an independent QTL. SNP name is defined by the location of the top SNP on the chromosome. Read it as follows chromosome: position, so 10:10486551 would be chromosome 10, location on the chromosome at 10486551
- af: frequency of the TopSNP in the rats used for this study
- beta: effect size of topSNP
- betase: standard error of effect size of topSNP
- -Log10(p): statistical significance of the association between the trait variability and the top SNP, displayed as -log10(p-value). The log-transformed p-value used in all figures and tables in this report
- trait: trait in which the snp was indentified
- F344, BUF, MR, M520, WN, BN, ACI, WKY: genotypes of founders at the topSNP

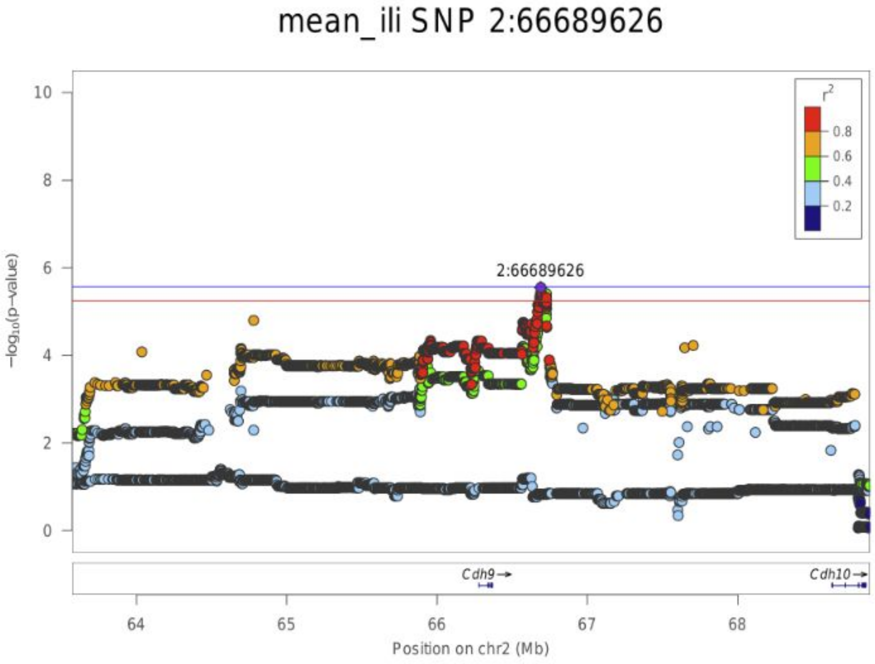
Enter filename: qtls.csv

index	TopSNP	Freq	beta	betase	-Log10(p)	significance_level	trait	F344	MR	BUF	WN	M520	BN	ACI	WKY
0	1:141548316	0.41	-0.198	0.043	5.466	10%	rt_peropfalsealarm_males	C C	T T	T T	T T	C C	C C	C C	T T
1	1:141548771	0.416	-0.153	0.031	6.199	5%	rt_peropfalsealarm	A A	G G	G G	G G	A A	A A	A A	G G
2	2:36107011	0.507	-0.141	0.03	5.499	10%	rt_corr	G G	G G	A A	G G	G G	G G	A A	A A
3	2:36107011	0.507	-0.141	0.03	5.499	10%	rt_leftcorr	G G	G G	A A	G G	G G	G G	A A	A A
4	2:36107011	0.507	-0.141	0.03	5.524	10%	rt_lefttrials	G G	G G	A A	G G	G G	G G	A A	A A
5	2:36107011	0.507	-0.141	0.03	5.524	10%	rt_totaltrials	G G	G G	A A	G G	G G	G G	A A	A A
6	2:71214697	0.589	-0.216	0.045	5.899	5%	rt_sdrt_females	A A	G G	G G	G G	A A	A A	G G	A A
7	2:71222161	0.587	-0.154	0.032	5.922	5%	rt_sdrt	C C	T T	T T	T T	C C	C C	T T	C C
8	2:72763104	0.705	0.763	0.14	7.323	5%	pc1	G G	G G	G G	G G	G G	T T	G G	T T
9	2:72770503	0.559	-0.32	0.06	7.008	5%	umap1	A A	C C	C C	C C	A A	A A	C C	A A
10	2:75843434	0.702	0.229	0.047	6.014	5%	rt_devmedrt_females	C C	C C	A A	C C	C C	A A	C C	C C
11	2:141010376	0.244	-0.18	0.035	6.686	5%	rt_ommisions	C C	C C	C C	C C	C C	C C	C C	G G
12	3:50051576	0.521	0.198	0.042	5.572	10%	rt_falsealarm_auc_females	C C	T T	C C	C C	T T	C C	T T	C C
13	4:24731426	0.544	0.147	0.031	5.714	10%	rt_corr	A A	A A	A A	A A	A A	G G	G G	G G
14	4:24731426	0.544	0.147	0.031	5.714	10%	rt_leftcorr	A A	A A	A A	A A	A A	G G	G G	G G

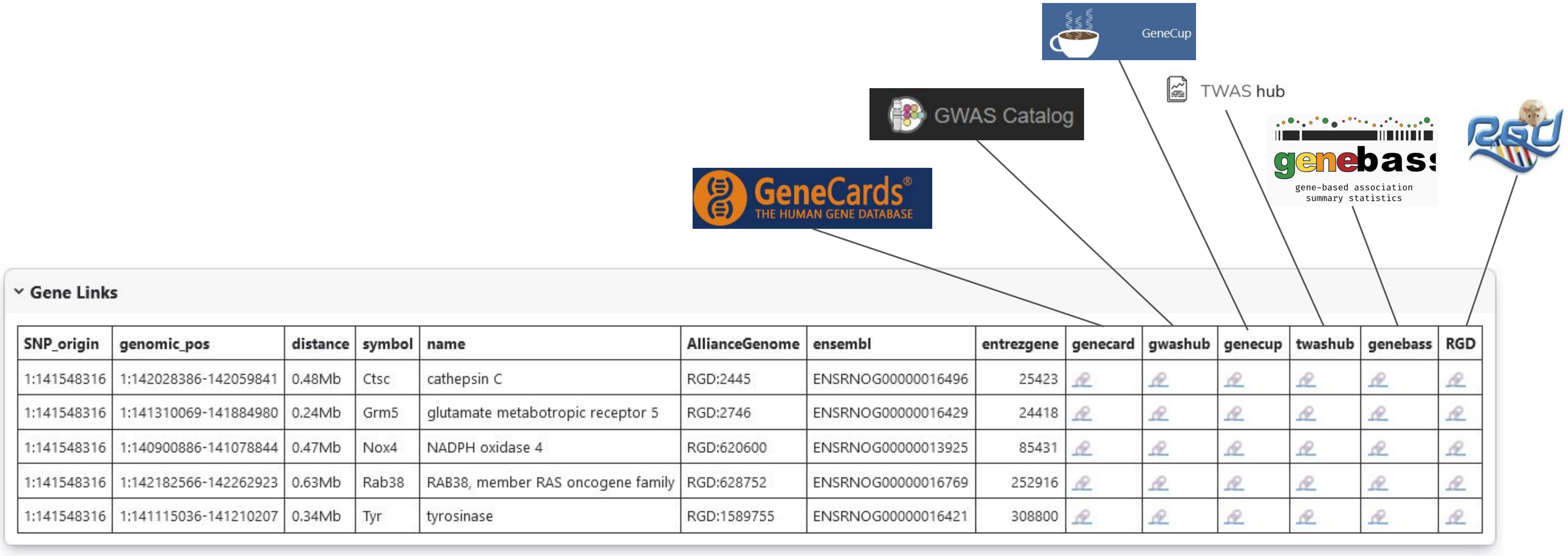
Zooming in each QTL

- Colocalization of QTL topsnps and eQTLs | sQTLs

```
pipe.locuszoom()
```



Resources for genes in QTL



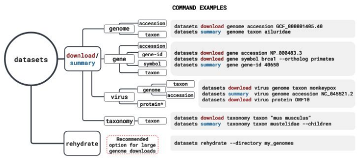
Gene Links

SNP_origin	genomic_pos	distance	symbol	name	AllianceGenome	ensembl	entrezgene	genecard	gwashub	genecup	twashub	genebass	RGD
1:141548316	1:142028386-142059841	0.48Mb	Ctsc	cathepsin C	RGD:2445	ENSRNOG00000016496	25423	🔗	🔗	🔗	🔗	🔗	🔗
1:141548316	1:141310069-141884980	0.24Mb	Grm5	glutamate metabotropic receptor 5	RGD:2746	ENSRNOG00000016429	24418	🔗	🔗	🔗	🔗	🔗	🔗
1:141548316	1:140900886-141078844	0.47Mb	Nox4	NADPH oxidase 4	RGD:620600	ENSRNOG00000013925	85431	🔗	🔗	🔗	🔗	🔗	🔗
1:141548316	1:142182566-142262923	0.63Mb	Rab38	RAB38, member RAS oncogene family	RGD:628752	ENSRNOG00000016769	252916	🔗	🔗	🔗	🔗	🔗	🔗
1:141548316	1:141115036-141210207	0.34Mb	Tyr	tyrosinase	RGD:1589755	ENSRNOG00000016421	308800	🔗	🔗	🔗	🔗	🔗	🔗

Annotations

- Uses VEP offline for annotation
- GTF and GFF files from NCBI datasets

```
pipe.annotQTL()
```



COMMAND EXAMPLES

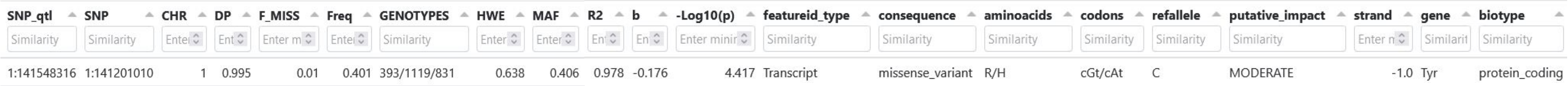
```
datasets download genome accession GCF_000001405.40
datasets summary  genome taxon ailuridae

datasets download gene accession NP_000483.3
datasets download gene symbol brca1 --ortholog primates
datasets summary  gene gene-id 40650

datasets download virus genome taxon monkeypox
datasets summary  virus genome accession NC_045521.2
datasets download virus protein ORF10

datasets download taxonomy taxon "mus musculus"
datasets summary  taxonomy taxon mustelidae --children

datasets rehydrate --directory my_genomes
```



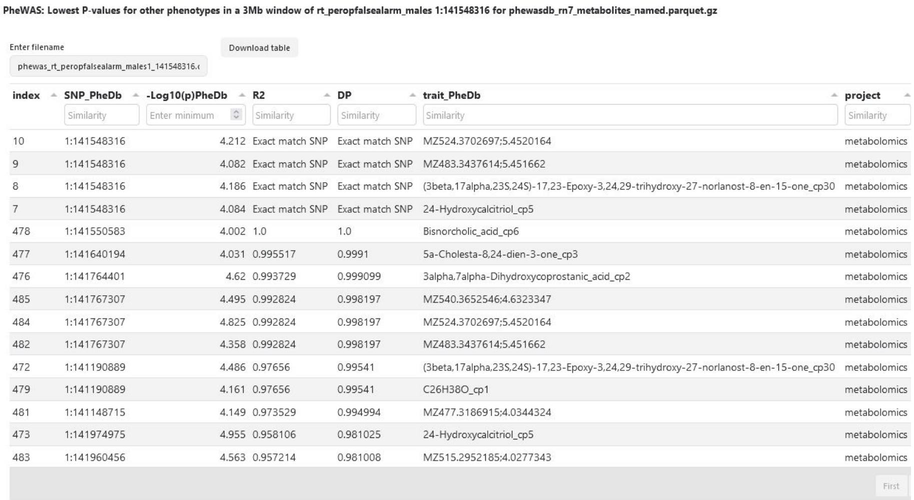
SNP_qtl	SNP	CHR	DP	F_MISS	Freq	GENOTYPES	HWE	MAF	R2	b	-Log10(p)	featureid_type	consequence	aminoacids	codons	refallele	putative_impact	strand	gene	biotype
Similarity	Similarity	Ente	Ent	Enter m	Ente	Similarity	Enter	Enter	Ent	En	Enter minir	Similarity	Similarity	Similarity	Similarity	Similarity	Similarity	Enter n	Similarit	Similarity
1:141548316	1:141201010	1	0.995	0.01	0.401	393/1119/831	0.638	0.406	0.978	-0.176	4.417	Transcript	missense_variant	R/H	cGt/cAt	C	MODERATE	-1.0	Tyr	protein_coding

Phewas

Colocalization of QTL topsnps and QTLs from other traits

- Available online for all traits in t lab

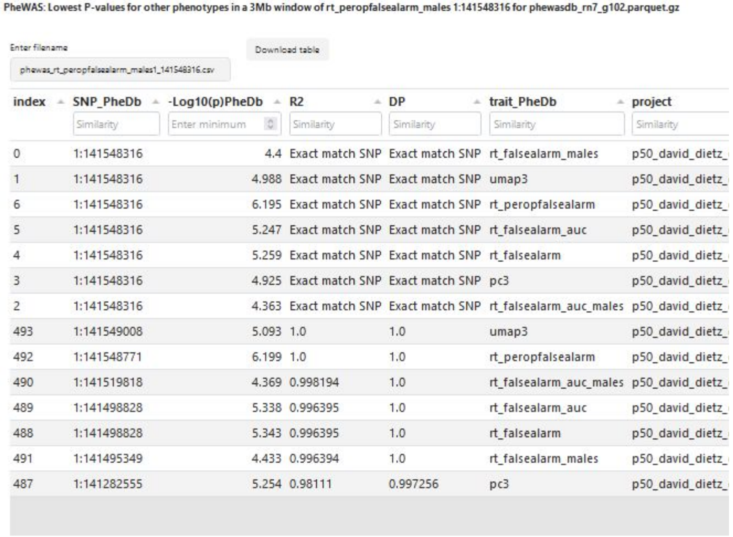
```
pipe.phewas()
```



PheWAS: Lowest P-values for other phenotypes in a 3Mb window of rt_peropfalsealarm_males 1:141548316 for phewasdb_rn7_metabolites_named.parquet.gz

Enter filename: phewas_rt_peropfalsealarm_males1_141548316.c

index	SNP_PheDb	-Log10(p)PheDb	R2	DP	trait_PheDb	project
10	1:141548316	4.212	Exact match SNP	Exact match SNP	MZ524.3702697;5.4520164	metabolomics
9	1:141548316	4.082	Exact match SNP	Exact match SNP	MZ483.3437614;5.451662	metabolomics
8	1:141548316	4.186	Exact match SNP	Exact match SNP	(3beta,17alpha,23S,24S)-17,23-Epoxy-3,24,29-trihydroxy-27-norlanost-8-en-15-one_cp30	metabolomics
7	1:141548316	4.084	Exact match SNP	Exact match SNP	24-Hydroxycalcitriol_cp5	metabolomics
478	1:141550583	4.002	1.0	1.0	Bisnorcholic_acid_cp6	metabolomics
477	1:141640194	4.031	0.995517	0.9991	5a-Cholesta-8,24-dien-3-one_cp3	metabolomics
476	1:141764401	4.62	0.993729	0.999099	3alpha,7alpha-Dihydroxycoprostanic_acid_cp2	metabolomics
485	1:141767307	4.495	0.992824	0.998197	MZ540.3652546;4.6323347	metabolomics
484	1:141767307	4.825	0.992824	0.998197	MZ524.3702697;5.4520164	metabolomics
482	1:141767307	4.358	0.992824	0.998197	MZ483.3437614;5.451662	metabolomics
472	1:141190889	4.486	0.97656	0.99541	(3beta,17alpha,23S,24S)-17,23-Epoxy-3,24,29-trihydroxy-27-norlanost-8-en-15-one_cp30	metabolomics
479	1:141190889	4.161	0.97656	0.99541	C26H38O_cp1	metabolomics
481	1:141148715	4.149	0.973529	0.994994	MZ477.3186915;4.0344324	metabolomics
473	1:141974975	4.955	0.958106	0.981025	24-Hydroxycalcitriol_cp5	metabolomics
483	1:141960456	4.563	0.957214	0.981008	MZ515.2952185;4.0277343	metabolomics



PheWAS: Lowest P-values for other phenotypes in a 3Mb window of rt_peropfalsealarm_males 1:141548316 for phewasdb_rn7_g102.parquet.gz

Enter filename: phewas_rt_peropfalsealarm_males1_141548316.csv

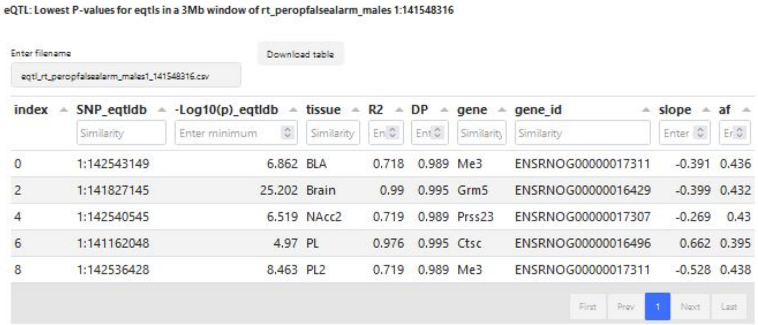
index	SNP_PheDb	-Log10(p)PheDb	R2	DP	trait_PheDb	project
0	1:141548316	4.4	Exact match SNP	Exact match SNP	rt_falsealarm_males	p50_david_dietz_
1	1:141548316	4.988	Exact match SNP	Exact match SNP	umap3	p50_david_dietz_
6	1:141548316	6.195	Exact match SNP	Exact match SNP	rt_peropfalsealarm	p50_david_dietz_
5	1:141548316	5.247	Exact match SNP	Exact match SNP	rt_falsealarm_auc	p50_david_dietz_
4	1:141548316	5.259	Exact match SNP	Exact match SNP	rt_falsealarm	p50_david_dietz_
3	1:141548316	4.925	Exact match SNP	Exact match SNP	pc3	p50_david_dietz_
2	1:141548316	4.363	Exact match SNP	Exact match SNP	rt_falsealarm_auc_males	p50_david_dietz_
493	1:141549008	5.093	1.0	1.0	umap3	p50_david_dietz_
492	1:141548771	6.199	1.0	1.0	rt_peropfalsealarm	p50_david_dietz_
490	1:141519818	4.369	0.998194	1.0	rt_falsealarm_auc_males	p50_david_dietz_
489	1:141498828	5.338	0.996395	1.0	rt_falsealarm_auc	p50_david_dietz_
488	1:141498828	5.343	0.996395	1.0	rt_falsealarm	p50_david_dietz_
491	1:141495349	4.433	0.996394	1.0	rt_falsealarm_males	p50_david_dietz_
487	1:141282555	5.254	0.98111	0.997256	pc3	p50_david_dietz_

e/sQTL

Colocalization of QTL topsnps and eQTLs | sQTLs

```
pipe.eQTL()
pipe.sQTL()
```

RatGTEx Portal

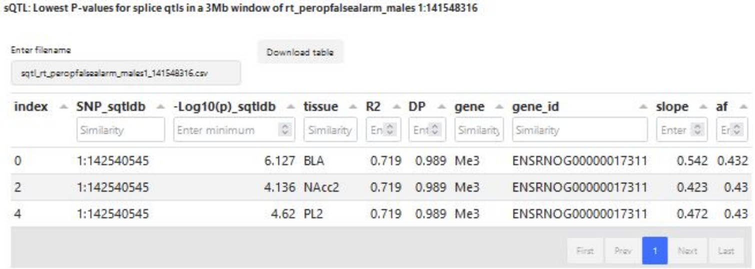


eQTL: Lowest P-values for eqtls in a 3Mb window of rt_peropfalsealarm_males 1:141548316

Enter filename

eqtl_rt_peropfalsealarm_males1_141548316.csv

index	SNP_eqtldb	-Log10(p)_eqtldb	tissue	R2	DP	gene	gene_id	slope	af
0	1:142543149	6.862	BLA	0.718	0.989	Me3	ENSRNOG00000017311	-0.391	0.436
2	1:141827145	25.202	Brain	0.99	0.995	Grm5	ENSRNOG00000016429	-0.399	0.432
4	1:142540545	6.519	NAcc2	0.719	0.989	Prss23	ENSRNOG00000017307	-0.269	0.43
6	1:141162048	4.97	PL	0.976	0.995	Ctsc	ENSRNOG00000016496	0.662	0.395
8	1:142536428	8.463	PL2	0.719	0.989	Me3	ENSRNOG00000017311	-0.528	0.438



sQTL: Lowest P-values for splice qtls in a 3Mb window of rt_peropfalsealarm_males 1:141548316

Enter filename

sqtl_rt_peropfalsealarm_males1_141548316.csv

index	SNP_sqtldb	-Log10(p)_sqtldb	tissue	R2	DP	gene	gene_id	slope	af
0	1:142540545	6.127	BLA	0.719	0.989	Me3	ENSRNOG00000017311	0.542	0.432
2	1:142540545	4.136	NAcc2	0.719	0.989	Me3	ENSRNOG00000017311	0.423	0.43
4	1:142540545	4.62	PL2	0.719	0.989	Me3	ENSRNOG00000017311	0.472	0.43

3 lines of code to download

```
git clone https://github.com/sanchestm/GWAS-pipeline.git
cd GWAS-pipeline
conda env create -n gwas -f environment.yml
```

6 lines of code to run

```
import sys
sys.path.append('***/GWAS-pipeline/')

from gwas_class_auto import *
df = pd.read_csv('raw_data')
pipeline = gwas_pipe( all_genotypes = 'P50_round2_LD_pruned_3473',
            data = df,
            project_name = 'example',
            genome_accession = 'GCF_000001895.5',
            threshold = 5.38,
            founderfile = None ,
            phewas_db = 'https://palmerlab.s3.sdsc.edu/tsanches_dash_genotypes/phewas/phewasdb_rn6.parquet.gz',
            threads = 8)
pipeline.run(round_version='genotypes_test', add_sex_specific_traits = True, clear_directories = True,
            gwas_version = '0.3.0', groupby_animals = ['center'], add_latent_space=False,
            researcher = 'user')
```

Thanks

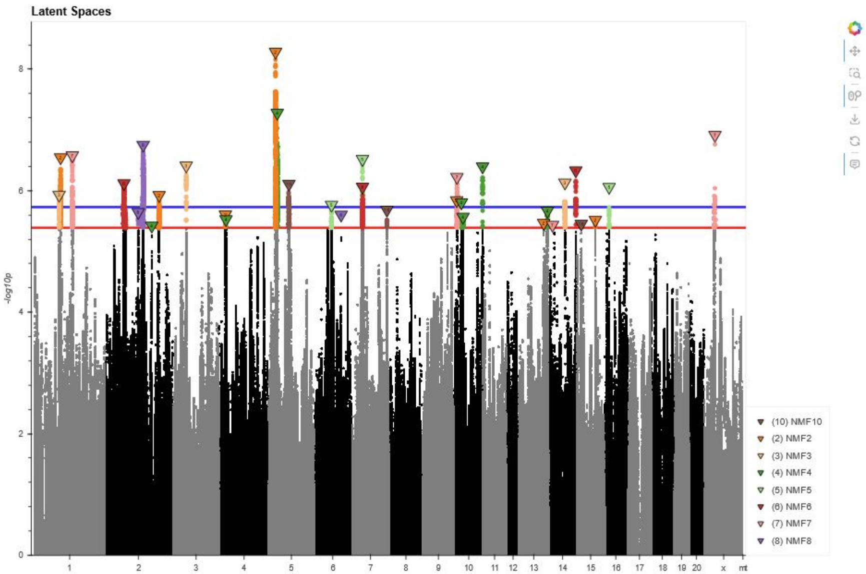
NIH National Institute on Drug Abuse
Advancing Addiction Science

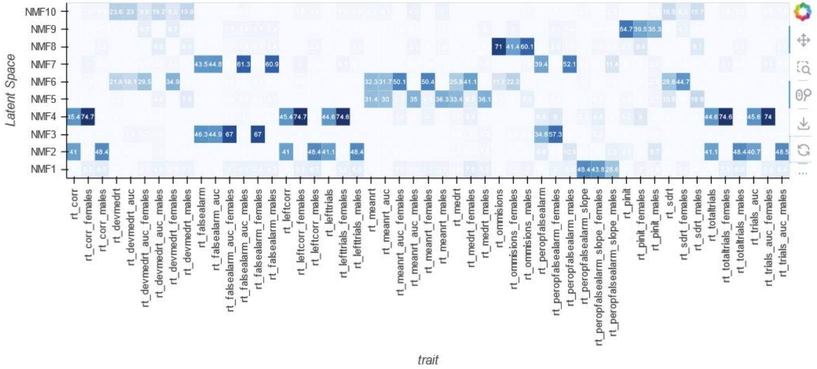
P50DA037844 | P30DA060810
Palmer Lab
All collaborators that made this possible

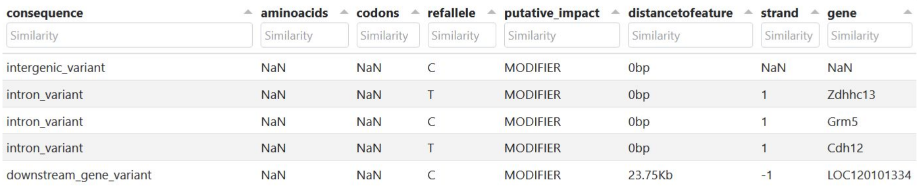
Experimental

Latent Spaces

PCA | UMAP | NMF



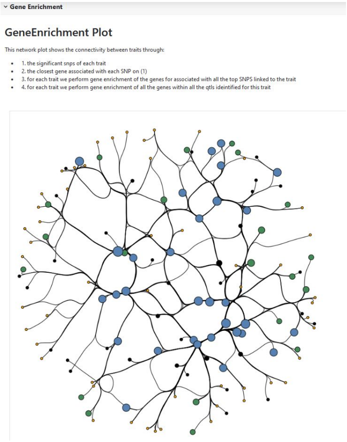




consequence	aminoacids	codons	refallele	putative_impact	distancetofeature	strand	gene
Similarity	Similarity	Similarity	Similarity	Similarity	Similarity	Similarity	Similarity
intergenic_variant	NaN	NaN	C	MODIFIER	0bp	NaN	NaN
intron_variant	NaN	NaN	T	MODIFIER	0bp	1	Zdhhc13
intron_variant	NaN	NaN	C	MODIFIER	0bp	1	Grm5
intron_variant	NaN	NaN	T	MODIFIER	0bp	1	Cdh12
downstream_gene_variant	NaN	NaN	C	MODIFIER	23.75Kb	-1	LOC120101334

Experimental Graph operations

- Steiner tree
- Closest drug to network
- Gene enrichment

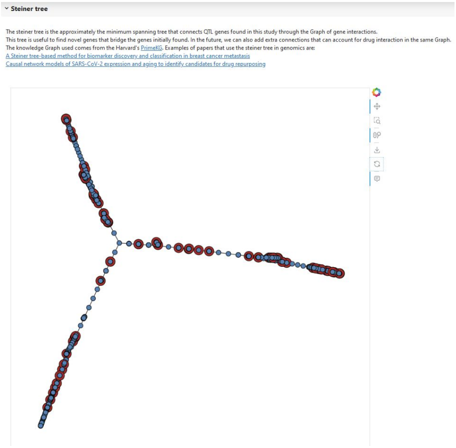


Gene Enrichment

GeneEnrichment Plot

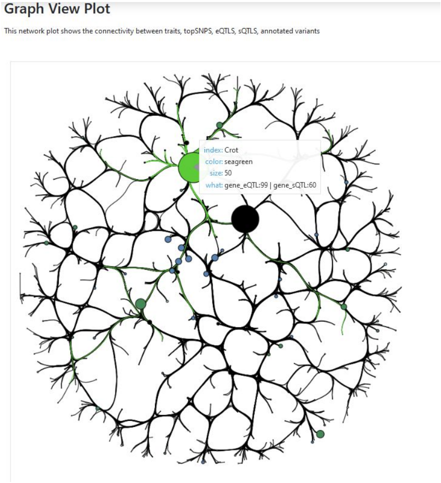
This network plot shows the connectivity between traits through:

- 1. the significant snps of each trait
- 2. the closest gene associated with each SNP on (1)
- 3. for each trait we perform gene enrichment of the genes for associated with all the top SNPS linked to the trait
- 4. for each trait we perform gene enrichment of all the genes within all the qtls ideintified for this trait



Steiner tree

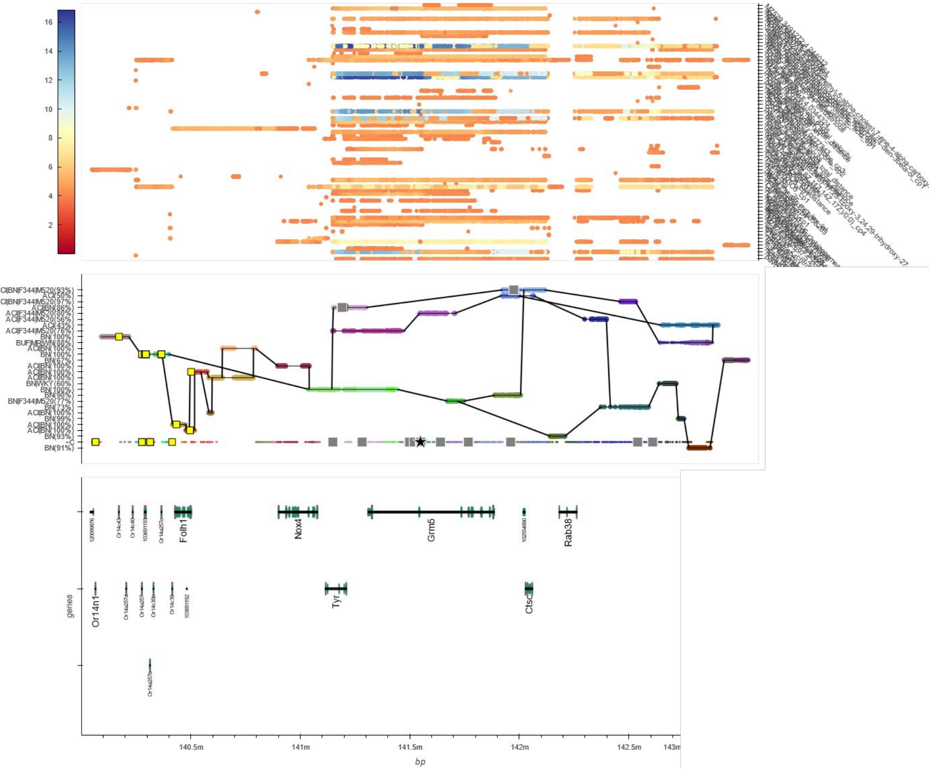
The steiner tree is the approximately the minimum spanning tree that connects QTL genes found in this study through the Graph of gene interactions.
This tree is useful to find novel genes that bridge the genes initially found. In the future, we can also add extra connections that can account for drug interaction in the same Graph.
The knowledge Graph used comes from the Harvard's PrimeKG. Examples of papers that use the steiner tree in genomics are:
A Steiner tree-based method for biomarker discovery and classification in breast cancer metastasis
Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing



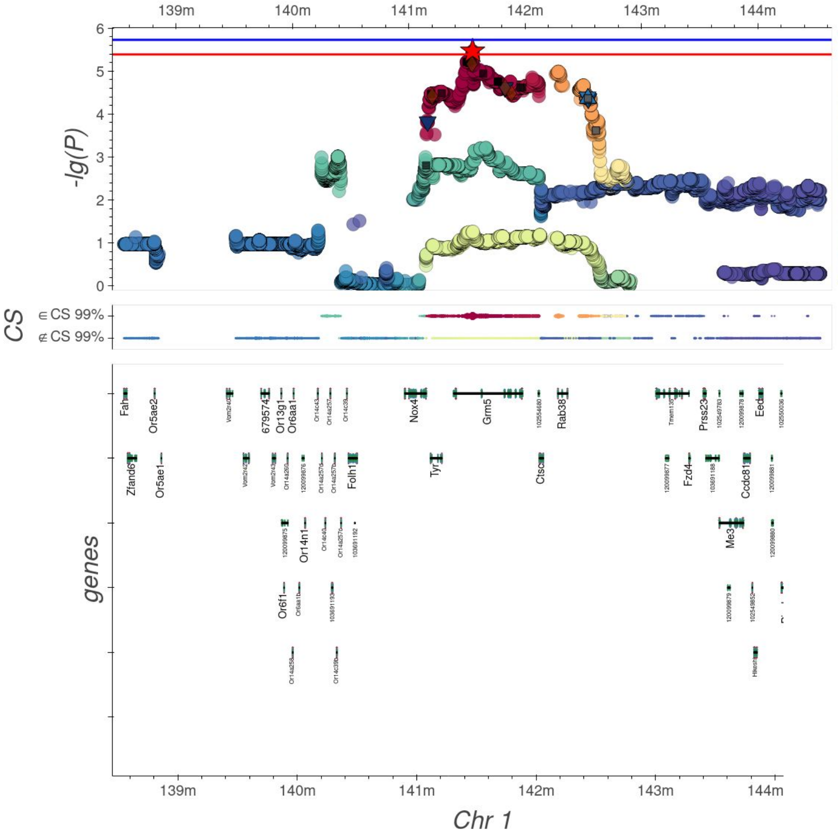
Graph View Plot

This network plot shows the connectivity between traits, topSNPS, eQTLS, sQTLS, annotated variants

index: Crot
color: seagreen
size: 50
what: gene_eQTL:99 | gene_sQTL:60



locuszoom Chr1 rt_peropfalsealarm_males 1:141548316



Combine all results into a single report