GeneNetwork Web-Services And Data Uploading

Felix Lisso

Date TBD

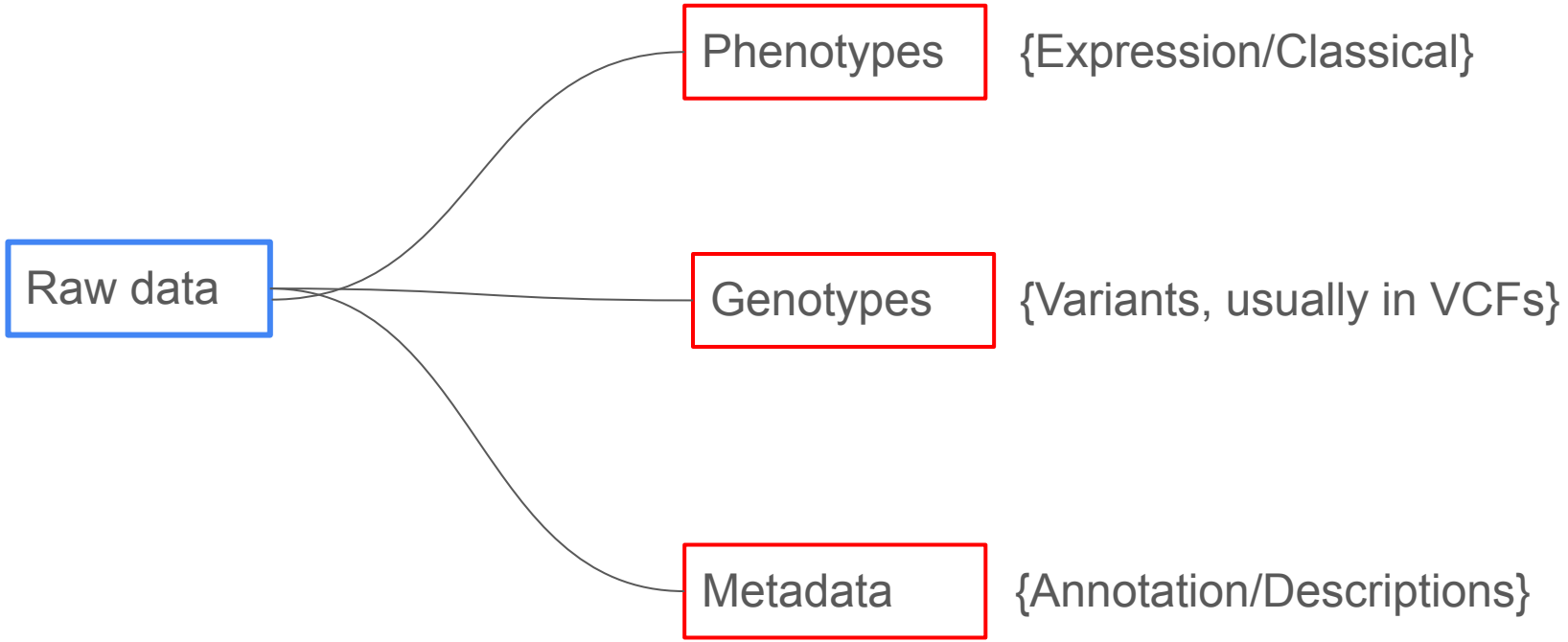
A bit on GeneNetwork

- A Web-Service for System Genetics
- Integrating large/diverse datasets (molecular, genotypes, and phenotypes)
- Curated data on model organisms

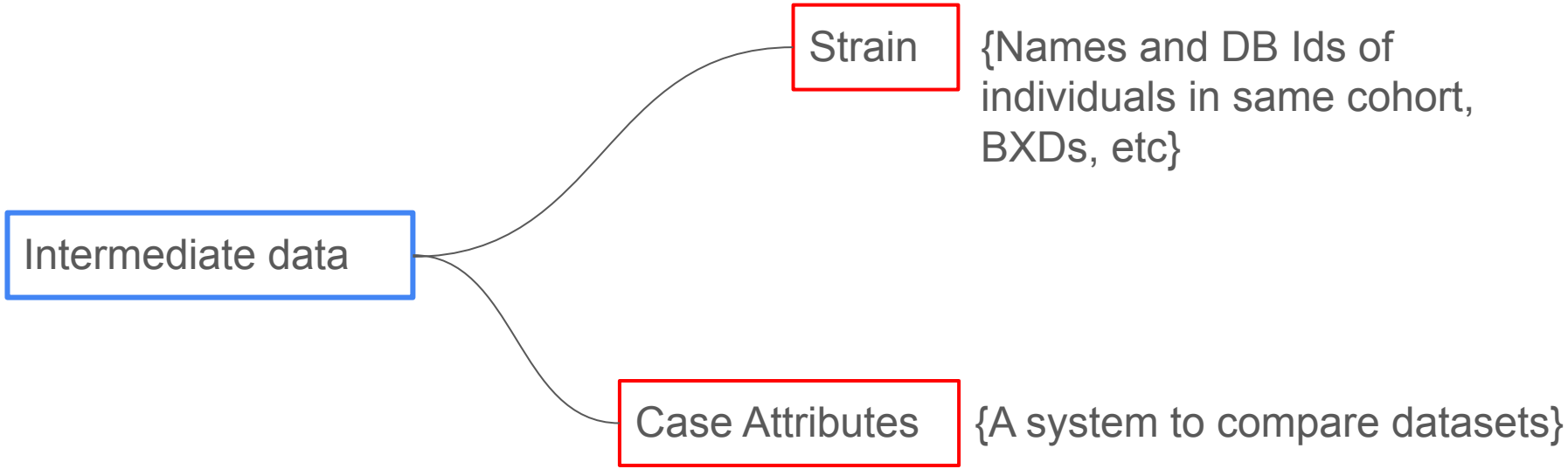
Why GeneNetwork?

- Database storage to a number of the biological datasets
- QC checks prior to data uploads
- Contain tools for analysis (GWAS, mapping, statistics, etc)
- Data can be downloaded for research use

What data GN2 expects?



What data GN2 expects?



Quality checks prior to Uploading

- For Genotypes:
  - Marker Ids (e.g rsxxxxx ids for snps, custom derivatives, etc)
  - Chromosomes
  - Variant positions (used to calculate physical maps{in base pairs})
  - Information on genetic maps

Quality checks prior to Uploading

- For phenotypes:
  - Log2 transformation and normalization (expression data mostly)
  - Numerical data
  - Empty cells to be filled with NAs string
  - At most, 6 decimal places

Processing Input files (for C.elegans)

Genotypes raw inputs (taken from Snoek et al files)

ID	WN1001	WN1002	WN1003	WN1004	WN1005	WN1006	WN1007
I_6313_SDP_14	1	-1	1	1	1	1	1
I_6314_SDP_14	1	-1	1	1	1	1	1
I_17860_SDP_JU1511	1	-1	-1	-1	0.5	-1	-1
I_17885_SDP_JU1511	1	-1	-1	-1	0.5	-1	-1

Map file (Variants)

name	Chromosome	position
I_6313_SDP_14	I	6313
I_6314_SDP_14	I	6314
I_17860_SDP_JU1511	I	17860
I_17885_SDP_JU1511	I	17885

Marker file (IDs, Xsomes, Positions)

- **The two files are Snoek's raw datasets from which genotype info for GN is obtained**
- **They need to be processed to get the final genotype file for GN2**

Processing Input files (for C.elegans)

Genotypes (data processing protocol for C.elegans)

- A command line tool is available to convert input geno data into GeneNetwork required format

  “ ./Celegans_geno_processor.py geno_file marker_file output_file”

- Where:
  - “./Celegans_geno_processor.py”, is the python tool
  - geno_file=”path to Snoek_mapfile.txt”
  - marker_file=”path to Snoek_markerfile.txt”
  - output_file=”path to where the output file will be stored”
- More information on the tool is found in this link: Celegans_protocol_section01

Processing Input files (for C.elegans)

Genotypes (C.elegans processed file from the protocol)

Chr	Locus	cM	Mb	WN1001	WN1002	WN1003	WN1004	WN1005	WN1006
I	I_6313_SDP_14	0.006313	0.006313	B	A	B	B	B	B
I	I_6314_SDP_14	0.006314	0.006314	B	A	B	B	B	B
I	I_17860_SDP_JU1511	0.01786	0.01786	B	A	A	A	H	A
I	I_17885_SDP_JU1511	0.017885	0.017885	B	A	A	A	H	A

Output file

- The output file should have a .geno extension (i.e Snoek_genotype.geno)
- The variants are encoded to letters to facilitate mapping analyses in GN2 (Part of QC standards in GN2)
- Metadata will be added on top of the output file (yet to be part of the protocol automation)

Processing Input files (for C.elegans)

Genotypes (an example file with metadata on top of it)

```
# File name: BXD_experimental_DGA_7_Dec_2021
# Metadata: Please retain this header information with file.
# Data Source and Contact: GeneNetwork at http://www.genenetwork.org/webqtl/main.py?FormID=sharinginfo&GN_AccessionId=600 . Please contact Robert W. Williams (rwilliams@uthsc.edu or la
# Citation: Please see http://www.genenetwork.org/reference.html For information on the BXD family of strains please see http://www.genenetwork.org/reference.html and Wang X, Pandey AK, Mullig
# Date Modified: March 22, 2022 by Rob Williams, David Ashbrook, and Danny Arends to remove excessive cross-over events in strains BXD42 (Chr9), BXD81 (Chrs1, 5, 10), BXD99 (Chr1), and BX
# File entered and modified in GN database on Mar 22, 2022 by Arthur Centeno and Suheeta Roy. Adding additional 37 new BXD F1s
# Coordinates of Markers and Assembly: Megabase and basepair positions of markers before 2017 used mm9 NCBI Build 37 coordinates. In Jan 2017 we converted to mm10, GRCm38. Note that sc
# Genotypes: This file provides consensus genotypes for 198 BXD strains and for the two progenitor strains and the reciprocal F1s. Of the 198 BXD strains, 191 are independent, whereas 7 are subs
# Material and Cases: 155 BXD strains and substrains are available as live stock from JAX or UTHSC (Jan 2017). 148 are independent strains (excludes 7 substrains). Most samples were were obtai
# Breeding: BXD1 to BXD30 generated by BA Taylor at the Jackson Laboratory starting in 1971 from F2 stock. BXD31 and BXD32 generated by BA Taylor at the Jackson Laboratory in the late 1970s
# Errors and Corrections: This file contains genotyping errors and imprecision in the locations of recombination events. Locations of most of the 10000 recombinations are typically accurate to better t
# cM position: cM positions were estimated by Danny Arends, Dec 2016. One recombination on both chromosomes (e.g., a switch from B to D or D to B on both chromosomes) is equivalent to about
# Acknowledgments  We thank Lu Lu for generating many of these strains (BXD43 and higher). We thank Casey Bohl, Melinda McCarty, and Jesse Ingels for supporting the UTHSC colony and prepa
# Funding:  This work and GeneNetwork have been funded by The UTHSC Center for Integrative and Translational Genomics, and grants from NIGMS (R01GM123489), NIAAA (U01AA16662), NIDA
# Funding for The GeneNetwork: NIAAA (U01AA13499, U24AA13513), NIDA, NIMH, and NIAAA (P20-DA 21131), NCI MMHCC (U01CA105417), and NCRR (U24 RR021760)
# Column Heads: Chr = chromosome, Locus=marker name, Mb_mm9= megabase position mouse genome assembly mm9, cM_Cox and cM_BXD are centimorgan positions of markers, nRecP = nur
#
@name:BXD
@type:riset
@mat:B
@pat:D
@het:H
@unk:U
```

Chr	Locus	cM	Mb	BXD1	BXD2	BXD5	BXD6	BXD8	BXD9	BXD11	BXD12	BXD13	BXD14	BXD15	BXD16	BXD18	BXD19	BXD20
1	rsm10000000001	0	3.00149	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs31443144	0.11	3.010274	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs6269442	0.21	3.492195	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs32285189	0.32	3.511204	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs258367496	0.43	3.659804	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs32430919	0.53	3.777023	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs36251697	0.64	3.812265	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs30658298	0.75	4.430623	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs31879829	0.85	4.518714	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D
1	rs36742481	0.96	4.776319	B	B	D	D	D	B	B	D	B	B	D	D	B	D	D

Metadata
(file taken from BXD genotype datasets)

Processing Input files (for C.elegans)

Phenotype raw input (taken from Snoek et al files)

Ids	WN1001	WN1002	WN1003	WN1004	WN1005	WN1006	WN1007
trait_1	3.169925	3.169925	3.169925	1.584963	3.321928	NA	3.906891
trait_10	5.60733	5.622052	NA	5.67948	5.665336	5.622052	NA
trait_11	5.693487	5.761551	NA	5.748193	5.761551	5.748193	NA
trait_12	NA	6.574909	6.828665	6.2062	6.51307	6.222392	6.51307
trait_13	NA	4.564785	5.97728	6.811642	6.52878	5.672425	5.992466
trait_14	NA	5.235216	5.624491	6.442943	6.539159	5.672425	5.874469
trait_15	NA	1	1	0.222392	3.584963	0.415037	0.415037
trait_16	NA	2.5025	0.736966	1.736966	2.222392	2.584963	3.369234
trait_17	NA	0.415037	0.736966	1.938599	2.415037	1.222392	0.736966
trait_18	0.329753	0.489885	0.580061	0.517542	0.546592	0.327543	0.447016
trait_19	4.552746	4.530968	4.57344	4.581711	4.542171	4.539816	4.520901
trait_2	0	-0.053242	NA	0	NA	-1.800877	NA
trait_20	9.905086	9.943321	10.001901	9.986126	9.970178	9.895575	9.92251
trait_21	5.352264	5.412782	5.427941	5.40429	5.427941	5.355792	5.400879
trait_3	4.371559	4.102658	4.288359	4.32553	4.383704	4.087463	4.099295
trait_4	4.269482	4.199672	4.193772	4.510962	4.451211	4.139551	4.266787
trait_5	NA	0.257011	NA	1.459432	0	-0.678072	1.077243
trait_6	-inf	-inf	NA	-inf	-1	-inf	NA
trait_7	-inf	-inf	NA	-inf	-1	-inf	NA
trait_8	5.554589	5.577429	NA	5.60733	5.592457	5.592457	NA
trait_9	5.636625	5.665336	NA	5.665336	5.707359	5.707359	NA

Processing Input files (for C.elegans)

Phenotypes

- A command line tool is available to convert input pheno data into GeneNetwork required format

  “ ./Celegans_phenotype_processor.py pheno_file output_file”

- Where:
  - “./Celegans_phenotype_processor.py”, is the python tool
  - pheno_file=”path to Snoek_phenotype.txt”
  - output_file=”path to where the output file will be stored”
- More information on the tool is found in this link: Celegans_protocol_section02

Processing Input files (for C.elegans)

Phenotypes (C.elegans processed file from the protocol)

WN_ID	WN001	WN002	WN003	WN004
17_Rodriguez_2012_trait_1	204.333333	274.818182	240.916667	290.333333
SE	x	x	x	x
N	x	x	x	x
17_Rodriguez_2012_trait_2	12.372685	5.842593	10.578125	14.561921
SE	x	x	x	x
N	x	x	x	x
17_Rodriguez_2012_trait_3	54.852419	112.589358	66.874996	119.882013
SE	x	x	x	x
N	x	x	x	x
17_Rodriguez_2012_trait_4	3.134638	1.110597	4.271893	5.971903
SE	x	x	x	x
N	x	x	x	x
17_Rodriguez_2012_trait_5	154.454545	167.3	165.833333	197.909091
SE	x	x	x	x
N	x	x	x	x

A snapshot of a processed phenotype file

- Notice the “SE’ and ‘N’
  - SE short for Standard Error, shows how accurate the average value is representative of the actual population (in this case, the replicates)
  - N to represent number of replicates per sample

Processing Input files (for C.elegans)

Metadata

- Information explaining the data provided, its essence, structure, and use
- Usually provided as part of supplementary information as used on a published research paper
- Or, customary to the data providers
- For phenotypes, there are several columns that are expected to be present for the metadata to be complete
- For more details, visit: Celegans_protocol_section04
- And for an example file format, visit: Phenotype descriptions

Uploading Protocols are important ……..

- Help with automation in data processing
- Varying nature of the raw datasets
- Guidelines to help individuals cross-check their datasets before uploading
- Saves time

How to upload data to GeneNetwork

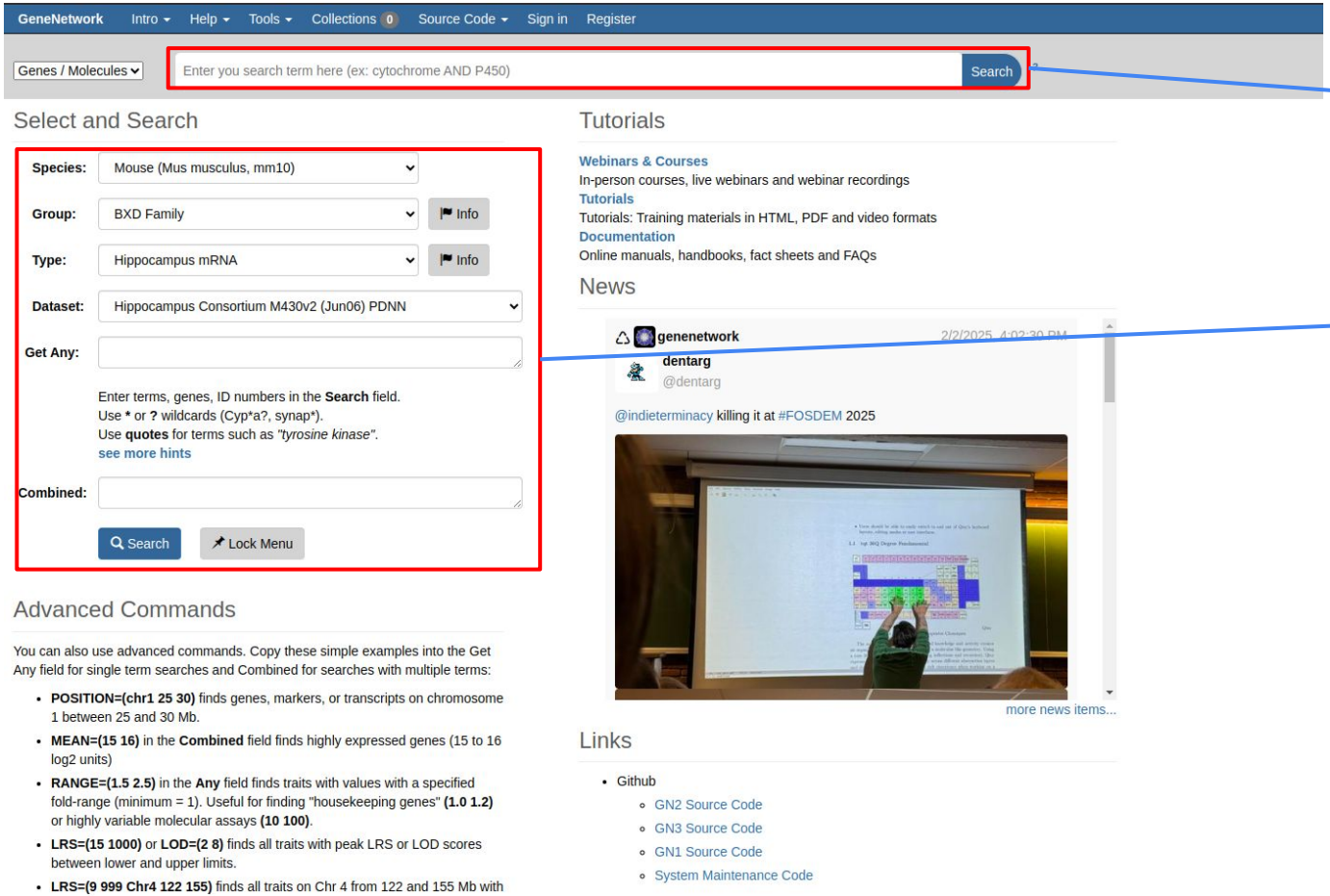
- A quick demonstration on how to use the uploader (currently being optimized)
  - Walk you through the process of Uploading new datasets using the uploader.
- The uploader will simplify steps involved in uploading data manually to GeneNetwork, therefore making it easy for even the non programmers to manage putting their data to GN.
- Link: https://staging-uploader.genenetwork.org/

TO BE EDITED ONCE THE UPLOADER IS BACK ONLINE !!!

How to use GeneNetwork

- Let’s get into a quick demonstration on how to use the Web Server:
  - Navigating menu search and global search
  - Mapping analyses (GEMMA, Rqtl2, etc)
- A brief explanation on the GeneNetwork homepage, then on with the live demo
- Link to gn2: https://genenetwork.org/

GENENETWORK HOMEPAGE (search options)



Global Search

Menu Search

GENENETWORK HOMEPAGE (search options)

GeneNetwork Intro Help Tools Collections 0 Source Code Sign in Register

Genes / Molecules | Enter you search term here (ex: cytochrome AND P450) | Search | ?

Select and Search

Species: Mouse (Mus musculus, mm10)

Group: BXD Family — Info

Type: Hippocampus mRNA — Info

Dataset: Hippocampus Consortium M430v2 (Jun06) PDNN

Get Any:

Enter terms, genes, ID numbers in the **Search** field.
Use * or ? wildcards (Cyp*a?, synap*).
Use **quotes** for terms such as *"tyrosine kinase"*.
see more hints

Combined:

Search | Lock Menu

Advanced Commands

You can also use advanced commands. Copy these simple examples into the Get Any field for single term searches and Combined for searches with multiple terms:

- **POSITION=(chr1 25 30)** finds genes, markers, or transcripts on chromosome 1 between 25 and 30 Mb.
- **MEAN=(15 16)** in the **Combined** field finds highly expressed genes (15 to 16 log2 units)
- **RANGE=(1.5 2.5)** in the **Any** field finds traits with values with a specified fold-range (minimum = 1). Useful for finding "housekeeping genes" **(1.0 1.2)** or highly variable molecular assays **(10 100)**.
- **LRS=(15 1000)** or **LOD=(2 8)** finds all traits with peak LRS or LOD scores between lower and upper limits.
- **LRS=(9 999 Chr4 122 155)** finds all traits on Chr 4 from 122 and 155 Mb with

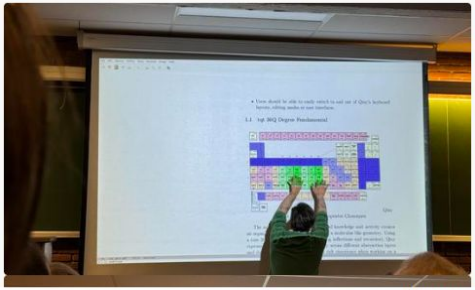
Tutorials

Webinars & Courses
In-person courses, live webinars and webinar recordings
Tutorials
Tutorials: Training materials in HTML, PDF and video formats
Documentation
Online manuals, handbooks, fact sheets and FAQs

News

genenetwork 2/2/2025, 4:02:30 PM
dentarg
@dentarg

@indieterminacy killing it at #FOSDEM 2025

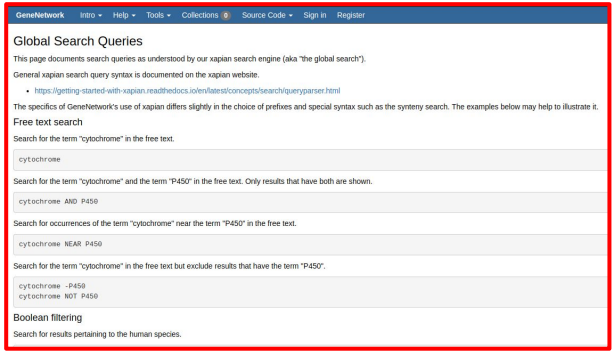


more news items...

Links

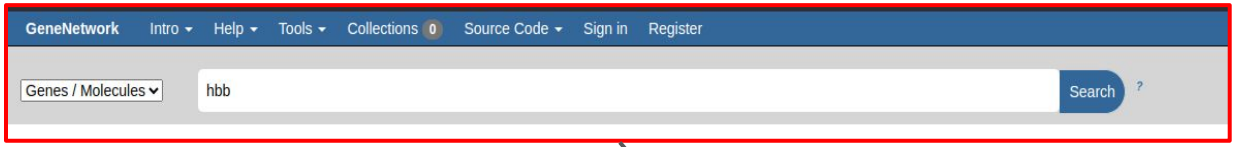
- Github
  - GN2 Source Code
  - GN3 Source Code
  - GN1 Source Code
  - System Maintenance Code

Global Search Syntax



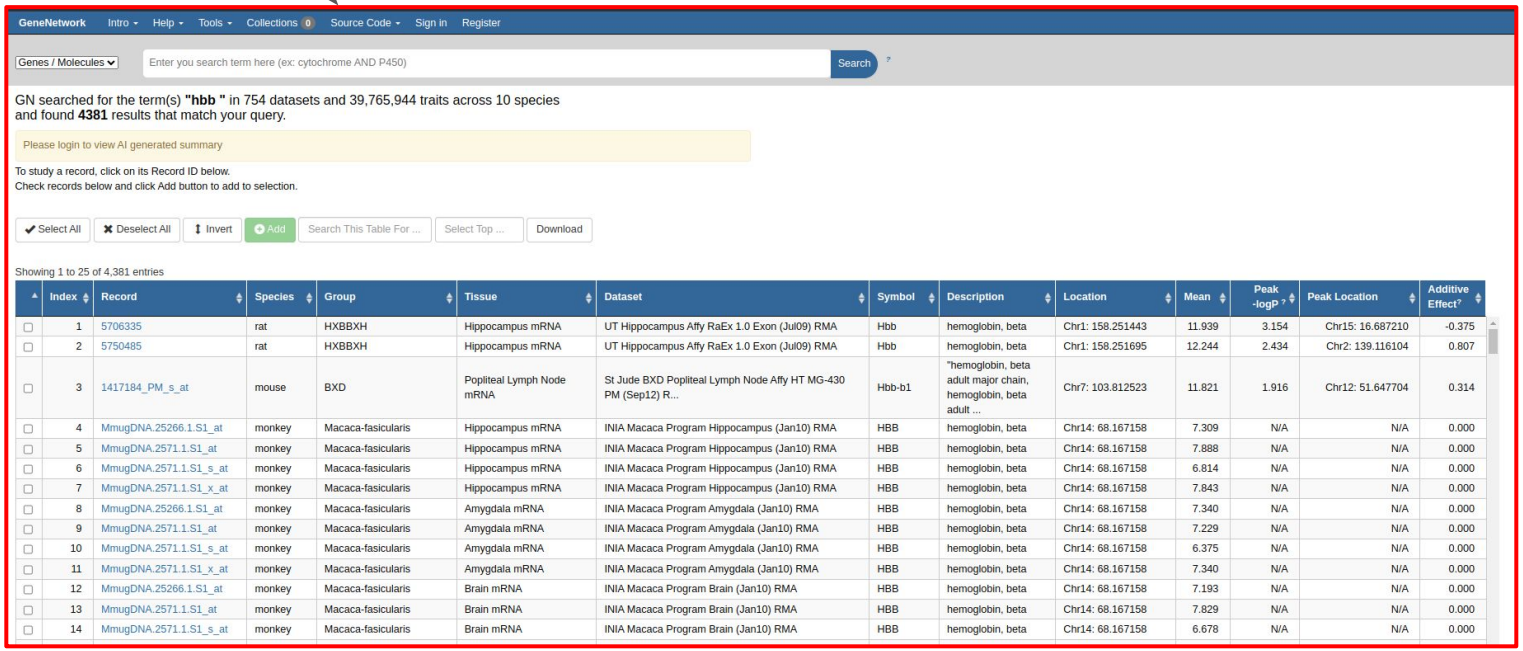
Menu search hacks

Global Search Results



Search query, ‘hbb’

Search results for hbb



Index	Record	Species	Group	Tissue	Dataset	Symbol	Description	Location	Mean	Peak -logP	Peak Location	Additive Effect
1	5706335	rat	HXBBXH	Hippocampus mRNA	UT Hippocampus Affy RaEx 1.0 Exon (Jul09) RMA	Hbb	hemoglobin, beta	Chr1: 158.251443	11.939	3.154	Chr15: 16.687210	-0.375
2	5750485	rat	HXBBXH	Hippocampus mRNA	UT Hippocampus Affy RaEx 1.0 Exon (Jul09) RMA	Hbb	hemoglobin, beta	Chr1: 158.251695	12.244	2.434	Chr2: 139.116104	0.807
3	1417184_PM_s_at	mouse	BXD	Popliteal Lymph Node mRNA	St Jude BXD Popliteal Lymph Node Affy HT MG-430 PM (Sep12) R...	Hbb-b1	"hemoglobin, beta adult major chain, hemoglobin, beta adult ...	Chr7: 103.812523	11.821	1.916	Chr12: 51.647704	0.314
4	MmugDNA.25266.1.S1_at	monkey	Macaca-fasicularis	Hippocampus mRNA	INIA Macaca Program Hippocampus (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.309	N/A	N/A	0.000
5	MmugDNA.2571.1.S1_at	monkey	Macaca-fasicularis	Hippocampus mRNA	INIA Macaca Program Hippocampus (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.888	N/A	N/A	0.000
6	MmugDNA.2571.1.S1_s_at	monkey	Macaca-fasicularis	Hippocampus mRNA	INIA Macaca Program Hippocampus (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	6.814	N/A	N/A	0.000
7	MmugDNA.2571.1.S1_x_at	monkey	Macaca-fasicularis	Hippocampus mRNA	INIA Macaca Program Hippocampus (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.843	N/A	N/A	0.000
8	MmugDNA.25266.1.S1_at	monkey	Macaca-fasicularis	Amygdala mRNA	INIA Macaca Program Amygdala (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.340	N/A	N/A	0.000
9	MmugDNA.2571.1.S1_at	monkey	Macaca-fasicularis	Amygdala mRNA	INIA Macaca Program Amygdala (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.229	N/A	N/A	0.000
10	MmugDNA.2571.1.S1_s_at	monkey	Macaca-fasicularis	Amygdala mRNA	INIA Macaca Program Amygdala (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	6.375	N/A	N/A	0.000
11	MmugDNA.2571.1.S1_x_at	monkey	Macaca-fasicularis	Amygdala mRNA	INIA Macaca Program Amygdala (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.340	N/A	N/A	0.000
12	MmugDNA.25266.1.S1_at	monkey	Macaca-fasicularis	Brain mRNA	INIA Macaca Program Brain (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.193	N/A	N/A	0.000
13	MmugDNA.2571.1.S1_at	monkey	Macaca-fasicularis	Brain mRNA	INIA Macaca Program Brain (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	7.829	N/A	N/A	0.000
14	MmugDNA.2571.1.S1_s_at	monkey	Macaca-fasicularis	Brain mRNA	INIA Macaca Program Brain (Jan10) RMA	HBB	hemoglobin, beta	Chr14: 68.167158	6.678	N/A	N/A	0.000

Menu Search for C.elegans

GeneNetwork Intro Help Tools Collections 0 Source Code Sign in Register

Genes / Molecules | Enter you search term here (ex: cytochrome AND P450) | Search ?

Select and Search

Species: Nematode (Caenorhabditis elegans, ce10)

Group: Multiparent Advanced Generation InterCross (M Info

Type: Traits and Cofactors Info

Dataset: MAGIC_Lines Phenotypes

Get Any:

Enter terms, genes, ID numbers in the **Search** field.
Use ***** or **?** wildcards (Cyp*a?, synap*).
Use **quotes** for terms such as *"tyrosine kinase"*.
see more hints

Combined: *

Search Lock Menu

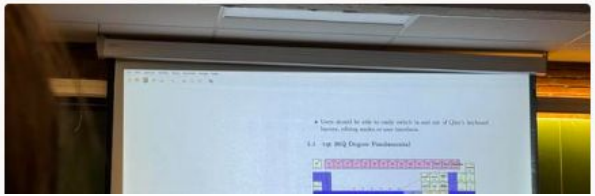
Tutorials

Webinars & Courses
In-person courses, live webinars and webinar recordings
Tutorials
Tutorials: Training materials in HTML, PDF and video formats
Documentation
Online manuals, handbooks, fact sheets and FAQs

News

genenetwork 2/2/2025, 4:02:30 PM
dentarg
@dentarg

@indieterminacy killing it at #FOSDEM 2025



Search Results for C.elegans (Traits records)

GeneNetwork Intro Help Tools Collections 0 Source Code Sign in Register

Genes / Molecules | Enter you search term here (ex: cytochrome AND P450) | Search

We searched MAGIC_Lines Phenotypes
to find all records in the dataset.
21 records found

Correlations | Networks | WebGestalt | BNW | GeneCup | WGCNA | CTL Maps | MultiMap | Partial Correlations

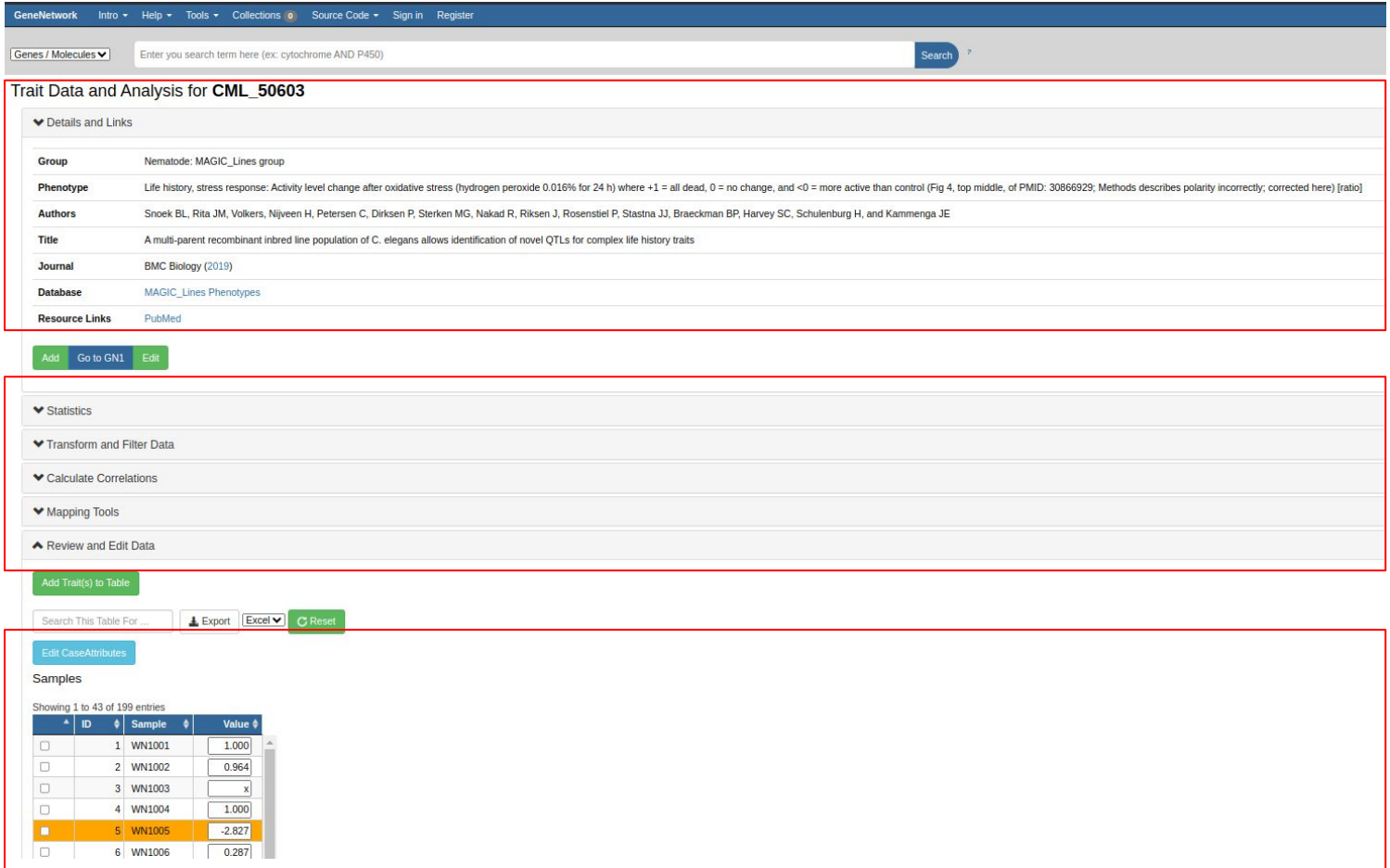
Search For... | Add | Select All | Select Rows (1-5, 11) | Download | Invert | Deselect

Show/Hide Columns: Description | Mean | Authors | Year | Peak -logP | Peak Location | Effect Size

Showing 1 to 21 of 21 entries

	Index	Record	Description	Mean	Authors	Year	Peak -logP	Peak Location	Effect Size
☐	1	CML_50602	Life history, stress response: Survival percentage of young adults worms (L4 stage) after exposure to high heat (heat shock) for 10 hours at 35 deg C (see Fig. 4 of PMID: 30866929, top left) [%]	11.321	Snoek BL, Rita JM, et al.	2019	3.1	ChrX: 12.450218	1.464
☐	2	CML_50603	Life history, stress response: Activity level change after oxidative stress (hydrogen peroxide 0.016% for 24 h) where +1 = all dead, 0 = no change, and <0 = more active than control (Fig 4, top midd...	0.223	Snoek BL, Rita JM, et al.	2019	7.2	ChrX: 16.975054	-0.491
☐	3	CML_50604	Cofactor, metadata: Mean Lifespan on OP50 [lsp]	18.184	Snoek BL, Rita JM, et al.	2019	4.6	Chr3: 4.986884	0.951
☐	4	CML_50605	Cofactor, metadata: Mean Lifespan on DR [lsp.dr]	19.399	Snoek BL, Rita JM, et al.	2019	5.5	ChrX: 5.237889	-0.927
☐	5	CML_50606	Cofactor, metadata: difference in lifespan [dr.eff]	1.254	Snoek BL, Rita JM, et al.	2019	2.4	Chr3: 4.986884	-0.764
☐	6	CML_50607	Cofactor, metadata: Occurrence of males on OP50 [males.op50]	0.236	Snoek BL, Rita JM, et al.	2019	3.4	Chr1: 14.569925	-0.140

Mapping results for Trait “CML_50603”



Metadata

Analysis tools

Case attributes

Mapping results for Trait “CML_50603” (GEMMA)

Add | Go to GN1 | Edit

- Statistics
- Transform and Filter Data
- Calculate Correlations
- Mapping Tools

GEMMA | Haley-Knott Regression | R/qtl | Pair Scan

Chromosome All

Minor Allele ≥ 0.05

Use LOCO Yes No

Covariates
Select covariate(s) from a collection

Select | Remove | Clear

No covariates selected

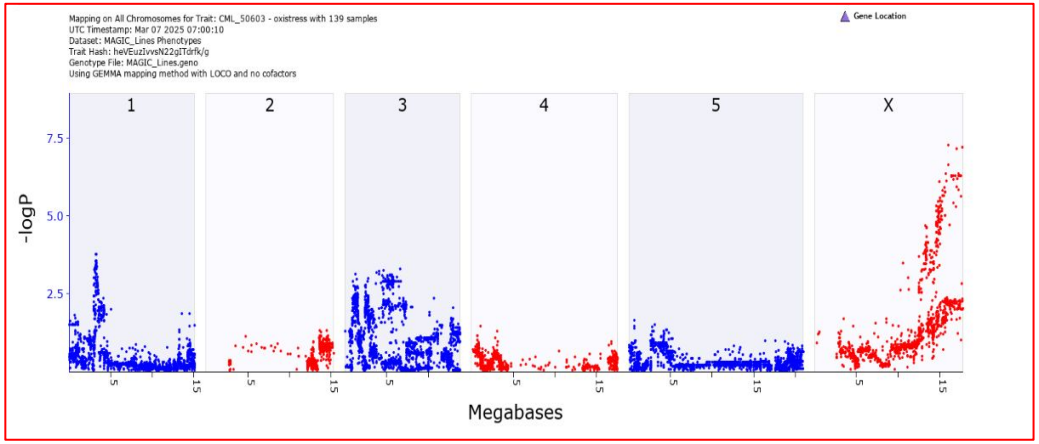
Compute

GEMMA
GEMMA maps with correction for kinship using a linear mixed model and can include covariates such as sex and age. Defaults include a minor allele frequency of 0.05 and the leave-one-chromosome-out method (PMID: 2453419, and GitHub code).

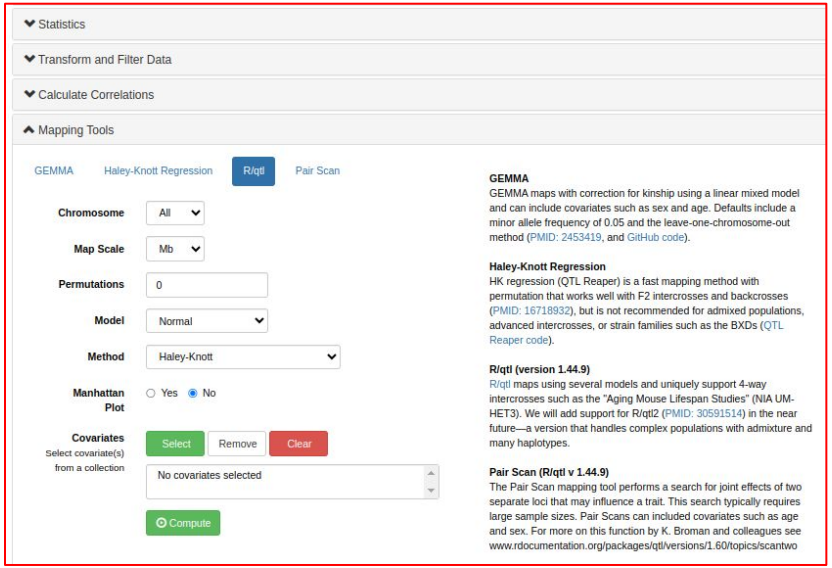
Haley-Knott Regression
HK regression (QTL Reaper) is a fast mapping method with permutation that works well with F2 intercrosses and backcrosses (PMID: 16718932), but is not recommended for admixed populations, advanced intercrosses, or strain families such as the BXDs (QTL Reaper code).

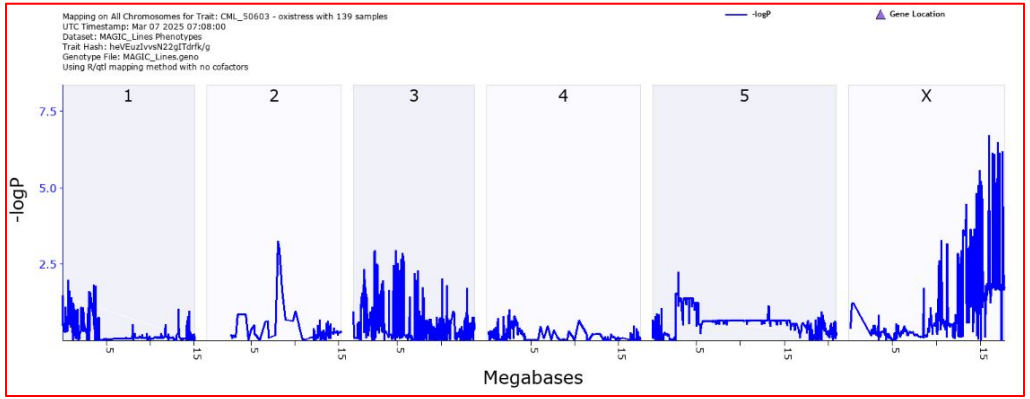
R/qtl (version 1.44.9)
R/qtl maps using several models and uniquely support 4-way intercrosses such as the "Aging Mouse Lifespan Studies" (NIA UM-HET3). We will add support for R/qtl2 (PMID: 30591514) in the near future—a version that handles complex populations with admixture and many haplotypes.

Pair Scan (R/qtl v 1.44.9)
The Pair Scan mapping tool performs a search for joint effects of two separate loci that may influence a trait. This search typically requires large sample sizes. Pair Scans can included covariates such as age and sex. For more on this function by K. Broman and colleagues see www.rdocumentation.org/packages/qtl/versions/1.60/topics/scantwo



Mapping results for Trait “CML_50603” (R/qtl)





Current challenges facing GeneNetwork….

- Uploading huge datasets
- Access and formatting metadata for data representation
    - Useful during queries
- Defining case Attributes
- Data processing protocols are custom to the data structure and organisation from source
    - Automation becomes a bit affected
    - More optimization needed time to time

Thank you for listening!