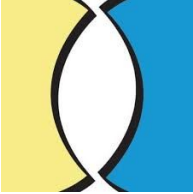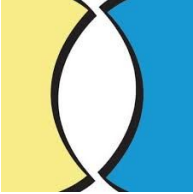# GeneNetwork Web-Services And Data Uploading

Felix Lisso

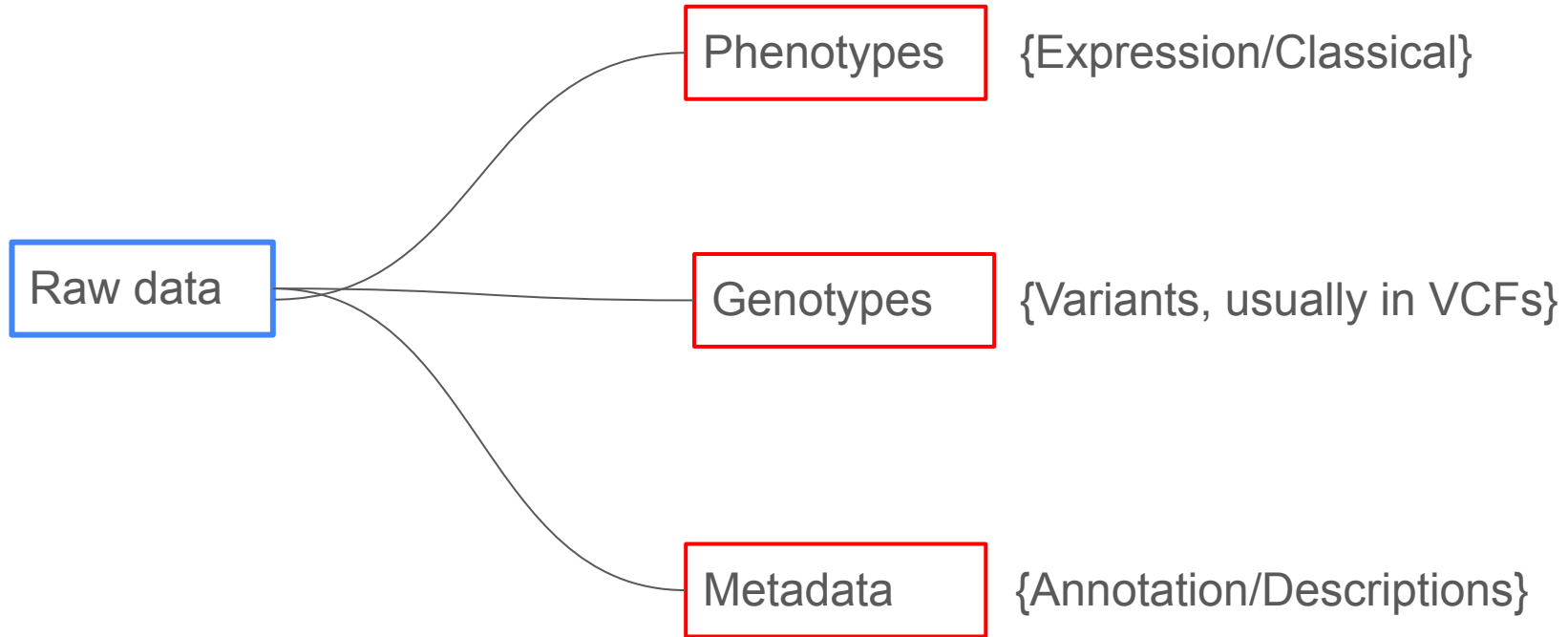Date TBD

# 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?

Raw data

Phenotypes {Expression/Classical}

Genotypes {Variants, usually in VCFs}

Metadata {Annotation/Descriptions}

# What data GN2 expects?

Intermediate data

Strain — {Names and DB Ids of individuals in same cohort, BXDs, etc}
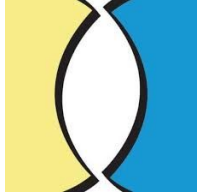
Case Attributes — {A system to compare datasets}
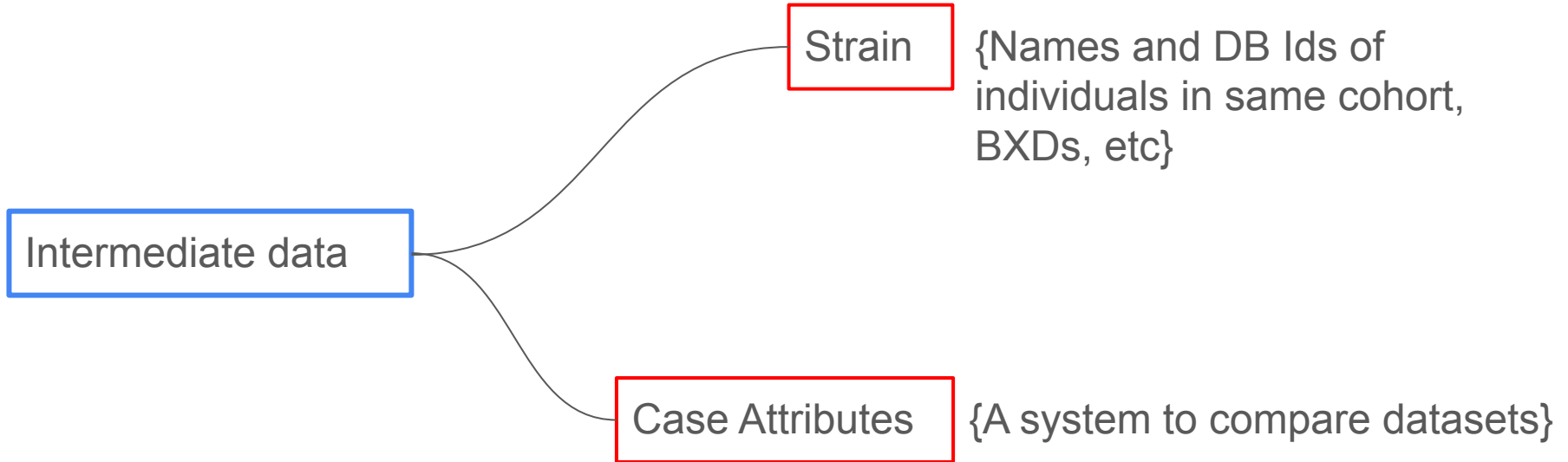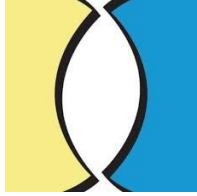
# Quality checks prior to Uploading
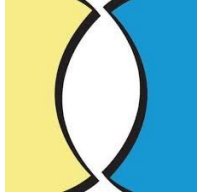
- For phenotypes:
    - Log2 transformation and normalization (expression data mostly)
    - Numerical data
    - Empty cells to be filled with NAs string
    - At most, 6 decimal places
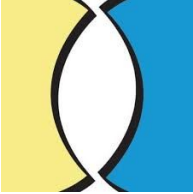
# Quality checks prior to Uploading

- 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

# Processing Input files (for C.elegans)

**Genotypes**

| 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)**

# Processing Input files (for C.elegans)

**Genotypes**

- 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)

**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)

**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

# Processing Input files (for C.elegans)

**Genotypes**

| 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

# Processing Input files (for C.elegans)

## Genotypes



| | | | | BXD1 | BXD2 | BXD5 | BXD6 | BXD8 | BXD9 | BXD11 | BXD12 | BXD13 | BXD14 | BXD15 | BXD16 | BXD18 | BXD19 | BXD20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # 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 so | | | | | | | | | | | | | | | | | | |
| # 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 subst | | | | | | | | | | | | | | | | | | |
| # 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 = num | | | | | | | | | | | | | | | | | | |
| # | | | | | | | | | | | | | | | | | | |
| @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 | B | D |
| 1 | rs31443144 | 0.11 | 3.010274 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs6269442 | 0.21 | 3.492195 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs32285189 | 0.32 | 3.511204 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs258367496 | 0.43 | 3.659804 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs32430919 | 0.53 | 3.777023 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs36251697 | 0.64 | 3.812265 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs30658298 | 0.75 | 4.430623 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs31879829 | 0.85 | 4.518714 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |
| 1 | rs36742481 | 0.96 | 4.776319 | B | B | D | D | D | B | B | D | B | B | D | D | B | B | D |

Genotype file with metadata

# Processing Input files (for C.elegans)

**Phenotypes**

| 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'
- There are at most 6 decimal places
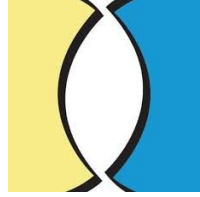
# Uploading Protocols are important ……..

- Help with automation in data processing

- Varying nature of the raw datasets

- Guidelines to help individuals cross-check their datasets before uploading

- Saves time
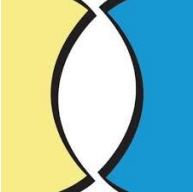
# How to upload data to GeneNetwork

- 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/
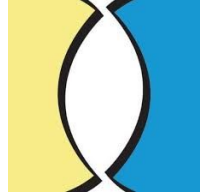
# How to use GeneNetwork

- 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/

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

**Global 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

**Menu Search**

## 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

**GENENETWORK HOMEPAGE**

**Global Search Syntax**

**Menu search hacks**

**GENENETWORK HOMEPAGE**

**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