Theses and Dissertations

May 2018

Quantitative Phenotype Analysis to Identify, Validate and Compare Rat Disease Models

Yiqing Zhao
University of Wisconsin-Milwaukee

Recommended Citation

QUANTITATIVE PHENOTYPE ANALYSIS TO IDENTIFY, VALIDATE AND

COMPARE RAT DISEASE MODELS

by

Yiqing Zhao

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

May 2018

ABSTRACT

QUANTITATIVE PHENOTYPE ANALYSIS TO IDENTIFY, VALIDATE AND COMPARE RAT DISEASE MODELS

by

Yiqing Zhao

The University of Wisconsin-Milwaukee, 2018
Under the Supervision of Professor Mary Shimoyama

Introduction

The laboratory rat has been widely used as an animal model in biomedical research. There are many strains exhibiting a wide variety of phenotypes. Capturing these phenotypes in a centralized database provides researchers with an easy method for choosing the appropriate strains for their studies. Current resources such as NBRP and PhysGen provided some preliminary work in rat phenotype databases. However, there are drawbacks in both projects: (1) small number of animals (6 rats) used by NBRP; (2) NBRP project is a one-time effort for each strain; (3) PhysGen web interface only enables queries within a single study – data comparison and integration not possible; (4) PhysGen lacks a data standardization process so that the measurement method, experimental condition, and age of rats used are hidden. Therefore, there is a need for a better data integration and visualization method in order to provide users with more insights about phenotype differences across rat strains. The Rat Genome Database (RGD) PhenoMiner tool has provided the first step in this effort by standardizing and integrating data from individual studies as well as NBRP and PhysGen.

Methods

Our work involved the following key steps: (1) we developed a meta-analysis pipeline to automatically integrate data from heterogeneous sources and to produce expected ranges (standardized phenotype ranges) for different strains, and different phenotypes under different experimental conditions; (2) we created tools to visualize expected ranges for individual strains and strain groups; (3) we clustered substrains into different sub-populations according to phenotype correlations.

Results

We developed a meta-analysis pipeline and an interactive web interface that summarizes and visualizes expected ranges produced from the meta-analysis pipeline. Automation of the pipeline allows for updates as additional data becomes available. The interactive web interface provides the researchers with a platform for identifying and validating expected ranges for a variety of quantitative phenotypes. In addition, we performed a preliminary cluster analysis that enables researchers to examine similarities of strains, substrains, and different sex or age groups of strains on a multi-dimensional scale by using multiple phenotype features.

Conclusion

The data resources and the data mining and visualization tools will promote an understanding of rat disease models, guide researchers to choose optimal strains for their research needs, and encourage data sharing from different research hubs. Such resources also help to promote research reproducibility. Data produced and interactive platforms created in this project will continue to provide a valuable resource for Translational Research efforts.

To
my parents,
families and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

RGD   Rat Genome Database

RDO   Rat Disease Ontology

GO   Gene Ontology

PO   Pathway Ontology

CHEBI   Chemical Entities of Biological Interest

RSO   Rat Strain Ontology

GWAS   Genome-Wide Association Studies

NBRP   National Bio Resource Project

QTL   Quantitative Trait Loci

CMO   Clinical Measurement Ontology

MMO   Measurement Method Ontology

XCO   Experimental Condition Ontology

eQTL   Expression Quantitative Trait Loci

EST   Expressed Sequence Tag

ML   Maximum Likelihood

REML   Restricted Maximum Likelihood

MOM   Method-of-Moments

ACKNOWLEDGMENTS

1. Introduction

1. 1   Model Organisms

Model organisms are important tools in biomedical research. Studies using model organisms have the potential to reveal the molecular mechanisms underlying disease (1-5) in human. The large-scale comparative analysis of phenotype and genotype data in model organisms can further reveal novel associations between genotypes and diseases (6-10). Such analysis traditionally has not been done as extensively in human.

Rattus norvegicus, or the laboratory rat has been widely used as an animal model for physiology, immunology, neoplasia, pharmacology, toxicology, nutrition and behavior research for over 160 years (11). The rat genome sequence project completed in 2004 (12) has greatly transformed the research paradigm, creating exceptional opportunities for identifying genes and pathways contributing to disease phenotypes in rats. Results generated from rat studies can then be translated to human. With the integrated use of genetic mapping, gene expression and computational analysis, researchers were able to expand their focus from monogenic rat traits to polygenic traits, including left ventricular mass (13), heart failure (14), mammary cancer (15), neuroinflammation (16), and glomerulonephritis (17, 18). A large number of rat strains have been bred to exhibit the phenotypes of common diseases, either spontaneously, or through the application of dietary, environmental or other conditions.

In order to leverage the power of the rat for such studies, a clear understanding of the phenotypic profiles of individual rat strains and commonly used control strains is needed. Phenotype refers to the observable morphological, physiological and behavioral characteristics of an individual under certain contexts of a study environment (19). Many

phenotypic characteristics can appear or disappear, or increase or decrease in severity throughout the lifespan of an individual. Phenotypic variation is an expression of genotype, or the sum of an individual's genetic makeup and environmental exposure. Thousands of human diseases are associated with phenotypic and genetic variations. Phenotypes observed in rats are often similar to those observed for particular human diseases and researchers will choose particular strains as models of the disease based on these observations. However, these choices are often based on a single previous experiment, the researchers' familiarity with or accessibility of the strain, or the fact that it is commonly seen by the community as a model for a particular disease. In addition, due to constraints in resources, individual investigators often focus on a limited number of phenotypes in a given strain, recording values for these few without recording a comprehensive phenotype profile of that strain.

Statistical analysis comparing phenotype values between strains is also commonly done in a single experiment. Unlike physicians in the clinic, rat researchers have not had the benefit of comprehensive expected (normal or abnormal) ranges for quantitative phenotype measurements for individual strains or for commonly used control strains based on multiple studies. The availability of statistically determined quantitative phenotype profiles for a wide range of rat strains would provide researchers with the data necessary for selecting optimal strains for their studies and help identify strains with profiles that closely mimic that of humans with particular diseases. The use of diverse panels of strains, both in phenotype and genotype, is increasing as a means to represent the diversity of human populations. Access to comprehensive quantitative phenotype profiles and comparisons with expected ranges, will facilitate the assembly of such strain panels.

1. 2   Current Resources

There have been several attempts to integrate quantitative phenotype data for model organisms such as mouse and rat to provide researchers with a view of data across experiments. Current resources include: 1) the Rat Phenome Project of the National BioResource Project for Rat in Japan (NBRP); 2) the PhysGen and PhysGen knockout program; 3) the Mouse Phenome Database; 4) the Rat Genome Database PhenoMiner Project. The Rat Phenome Project by NBRP in Japan, and the PhysGen program and PhysGen Knockout program are some of the most comprehensive rat phenotype measurement studies that have been conducted. However, both projects have some limitations (will be discussed in the following paragraphs). The Mouse Phenome Database is a good example in terms of data curation and visualization. Their tools provide different visualizations to view phenotype measurements in one dataset. However, the tools don't allow comparing or integrating different phenotype measurements across different studies. The PhenoMiner project is the forerunner of our current project. It integrates quantitative phenotype records from multiple experiments using standardized data formats and vocabularies for the sample used, the phenotype measured, how it was measured and under what experimental conditions, making it easier to query and compare data from multiple studies. Standardizing these four components using the Rat Strain Ontology (RSO), Clinical Measurement Ontology (CMO), Measurement Method Ontology (MMO) and Experimental Condition Ontology (XCO) was an important step in providing comprehensive phenotype profiles for individual strains and across strains and to create a foundation on which expected ranges for particular phenotypes could be determined.

1)   The Rat Phenome Project by NBRP in Japan

The NBRP Rat Phenome Project in Japan (http://www. anim. med. kyoto-u. ac. jp/nbr/phenome. aspx) (20) has comprehensively phenotyped more than 250 rat strains on 109 parameters in 7 categories: locomotor activity, neurobehavior, blood pressure, blood chemistry, hematology, urology, and anatomy. The measurements for each phenotype were conducted on a group of 6 rats - male or female, aged 5 to 10 weeks, for each strain and results presented as sample means. For some phenotypes, such as body weight, measurements were taken at three different time points (5, 6, and 10 weeks) and for others at a single time point. The NBRP website provides tables which show measurements for a single strain or tables and bar charts that show measurements across strains for a single phenotype. They also indicate the research category in which the strain is commonly used (e. g. diabetes, immunology, cardio-hypertension).

The benefit of using the NBRP Rat Phenome Project methodology is that by constructing measurement groups with the same number of animals (6 rats), and on both sexes, researchers could compare the same phenotype across strains (between study variance can be controlled by using the same number of animals). In addition, conclusions about sex differences in phenotypes will be easier to draw since the measurements were done under the same conditions and at the same age.

However, there are some drawbacks to this project: 1) the number of animals (6 rats) they used was small, resulting in relatively large within study variance; 2) the measurement method used by NBRP may not be available in other labs. As a result, NBRP phenotype measurements may be biased and make it hard for researchers to compare their own results with NBRP measurements; 3) their project is a one-time effort for each strain. However, even inbred rat strains can drift in their genetic make-up or physiological

characteristics. This again makes it hard for researchers to compare their own results with NBRP measurements. Thus, it is essential to continuously measure rat phenotypes and then compare or combine old measurement values with new ones.

2)   The PhysGen and PhysGen Knockout Program

The PhysGen (http://pga. mcw. edu/) Program for Genomic Application at the Medical College of Wisconsin has produced large-scale phenotype data using a variety of inbred, consomic and knockout rat strains. The PhysGen Program developed two panels of consomic rats using the SS/JrHsdMcwi, the FHH/EurMcwi and the BN/NHsdMcwi strains (21). Comprehensive characterization (434, 845 physiological data points) of these consomic strains, each carrying a chromosome from the sequenced Brown Norway strain, allowed for immediate mapping of traits to a particular chromosome without the need for genetic crosses (22). The PhysGen Knockout Program collaborated with several labs to generate mutant rats through different mutagenesis protocols (23). Inbred, mutant and consomic rat strains were characterized on 213 mainly cardiovascular phenotypes (24). Utilizing comparative genomic tools and the available PhysGen rat models, in vivo studies have been conducted to investigate the role of mutant genes in cardiovascular and metabolic diseases. The advantage of The PhysGen Program is that it created a federated database with curated measurements on rats from different laboratories and studies on different rat strains (inbred, mutant and consomic). The PhysGen Program developed web tools (**Figure 1**) that enable querying of experiments for a specific phenotype. The PhysGen website provides visualization of individual phenotype results across multiple strains with statistical analysis. It also provides strain profiles which summarize both general and phenotype data for individual strains (**Figure 2**) (25).

Phenotype Data Status

The number to the right of each strain indicates the number of rats of this strain used in the protocol stated in the first row. The row "# of PHENOTYPES" indicates the total number of phenotypes measured for each protocol. The column " TOTAL" is the total number of rats used from a given strain.

Click on a protocol name to view the phenotypes measured. Click on a strain name to see its description. To examine and explore data for a given protocol, select that protocol's radio button and press "Download Phenotype data". To visualize and analyze data, select that protocol's radio button and press "Visualization & Statistics". To view the T-test table for each protocol, select that protocol's radio button and press "T-test table".

* Rats in progress

➢ Please **click here** for a statement about important diet effects.

➢ Please **click here** for an explanation about important changes in the Renal Protocol.

| Download phenotype data | | Visualization & Statistics | | T-test table | |

STRAIN \ PROTOCOL	BIOCHEMISTRY ◯	CARDIAC ◯	HISTOLOGY ◯	LUNG ◯	RENAL_A ◯	RENAL_B ◯	RESPIRATORY ◯	VASCULAR ◯	TOTAL
BN	70	86	12	104	33	52	64	85	506
SS	264	291	12	337	63	76	153	331	1527
FHH	160	208	14	195	49	66	108	177	977
LEW	40	40	8	40	18	*	45	100	291
CDF	63	40	6	64	20	*	39	60	292
CD®IGS	36	40	6	45	18	*	29	60	234
GH	64	66	6	111	36	*	35	60	378
LE/BluGill	44	41	*	62	*	16	28	55	246
SHR	58	67	6	71	38	*	58	57	355
SD/Hsd	62	56	6	45	40	*	45	78	332
WKY	43	39	6	50	17	*	45	60	260
SS-1BN/Mcwi	25	20	4	43	*	40	20	22	174
SS-2BN/Mcwi	40	45	8	51	*	20	36	65	265
SS-3BN/Mcwi	34	39	2	52	*	16	23	39	205
SS-4BN/Mcwi	85	60	4	69	*	15	27	50	310
SS-5BN/Mcwi	30	39	2	40	*	18	19	39	187
SS-6BN/Mcwi	36	39	4	39	*	21	31	60	230
SS-7BN/Mcwi	45	72	4	47	*	22	27	44	261
SS.BN-(D8rat163-D8rat81)/Mcwi	48	66	4	49	*	35	29	40	271

Rat Select Option for CARDIAC PROTOCOL

Select phenotypes

* Push "Ctrl" to make multiple choices.

```
1. body weight (kg)
2. pre-ischemic heart wet weight (g)
3. Heart weight expressed as % of body weight
4. pre-ischemic heart rate (beats/min)
5. pre-ischemic coronary flow rate (ml/min/g)
6. pre-ischemic left ventricle diastolic pressure (mmHg)
7. pre-ischemic left ventricle systolic pressure (mmHg)
8. pre-ischemic left ventricle developed pressure (mmHg)
9. pre-ischemic enzyme leakage (IU/g wet wt)
10. ischemic time to onset of contracture (sec)
```

| Do Analysis | Change protocol |

Figure 1 The PhysGen Web Interface for Querying Experiments of a Phenotype (selected "CARDIAC PROTOCOL")

Figure 2 The PhysGen PGA website provides access to strain and phenotype data for more than 70 strains. Users can query by phenotype or strain. Phenotype queries (lower left) return data for all strains for a single phenotype. Strain profile data for a particular strain across all phenotypes (lower right) (25)

However, there are several drawbacks to the data as presented: 1) the web interface only enables queries within a certain protocol or experiment e. g. BIOCHEMISTRY, CARDIAC, RENAL, RESPIRATORY. However, some phenotypes were measured in multiple protocols. For example, "heart rate" was measured in both CARDIAC and RENAL which make phenotype-based comparison and integration difficult; 2) it lacks a data standardization process so that the measurement method, experimental condition and age of rats used remains hidden

behind those data points. A user needs to refer to the protocol to gain this information. This again makes it difficult to truly compare data from various sources and across experiments.

3)   Mouse Phenome Database

The mouse phenome project offers another example of a coordinated effort in phenotyping, data standardization as well as data federation. In 2001, the Mouse Phenome Project was launched to complement mouse genome sequencing efforts by promoting new phenotyping initiatives under standardized conditions and to integrate the data in a central public database, the Mouse Phenome Database (MPD) (https://phenome. jax. org).

The advantage of MPD is that it has huge collection of mouse research dataset, most of them contributed from the mouse research community. All phenotype related data were organized and displayed by dataset (**Figure 3**). The visualization of each dataset is by default shown as scatter plot with error bar as mean and standard deviation range (**Figure 4**). In addition, it also shows the Measurement Summary, ANOVA, and Q-Q Plot (**Figure 5**).

Figure 3 Mouse Phenome Database Browse by Phenotype (Systolic Blood Pressure)

Phenotype measure: CGDpheno3   systolic_BP

ID, description, units	MPD:43925 **systolic_BP** systolic blood pressure [mmHg]
Data set, strains	CGDpheno3 CC diallel w/par 24 strains sex: both age: 10wks
Procedure	tail cuff
Ontology mappings	

Click here for diallel cross view
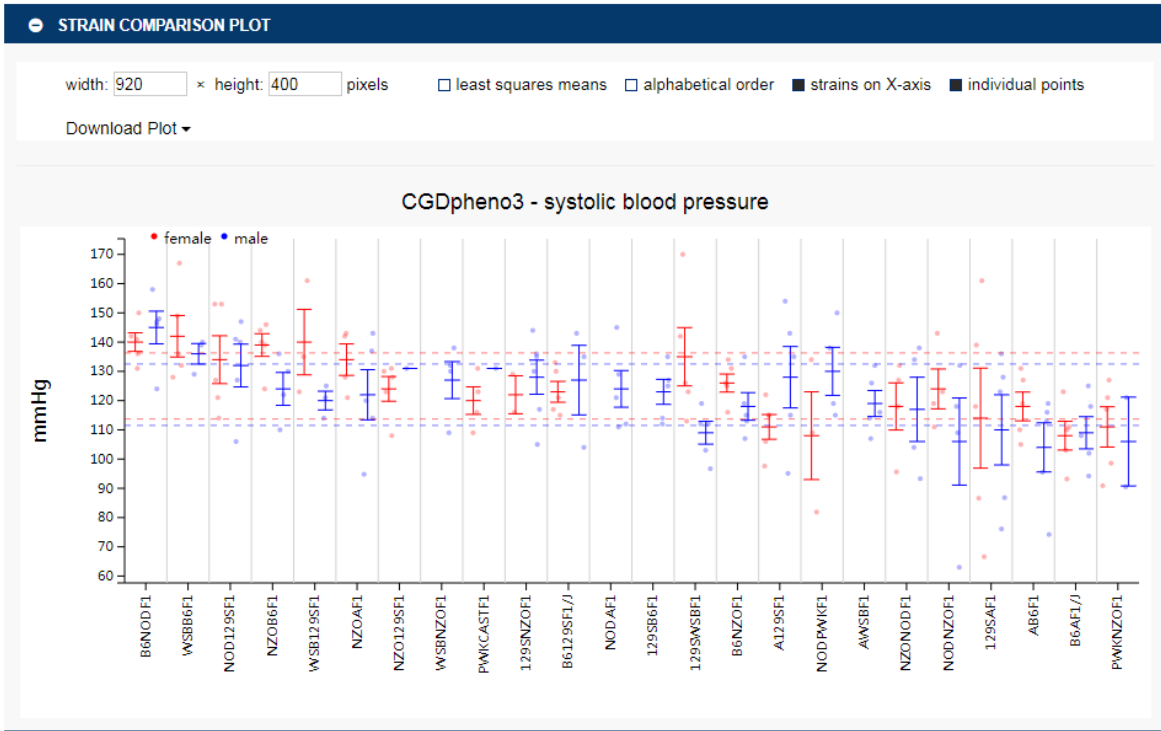
Figure 4 Visualization of Phenotype Measurement in one MPD Dataset

Measure Summary	Female	Male
Number of strains tested	20 strains	24 strains
Mean of the strain means	125 mmHg	122 mmHg
Median of the strain means	124 mmHg	124 mmHg
SD of the strain means	± 11.3	± 10.5
Coefficient of variation (CV)	0.0907	0.0864
Min–max range of strain means	108 – 142 mmHg	104 – 145 mmHg
Mean sample size per strain	4.6 mice	4.2 mice

● ANOVA, Q-Q NORMALITY ASSESSMENT

ANOVA summary   Help

Factor	DF	Sum of squares	Mean sum of squares	F value	p value (Pr>F)
sex	1	701.7815	701.7815	2.3745	0.1259
strain	17	15289.322	899.3719	3.0431	0.0002
sex:strain	17	5599.9744	329.4103	1.1146	0.3478
Residuals	122	36056.6052	295.5459		

Q-Q normality assessment based on residuals
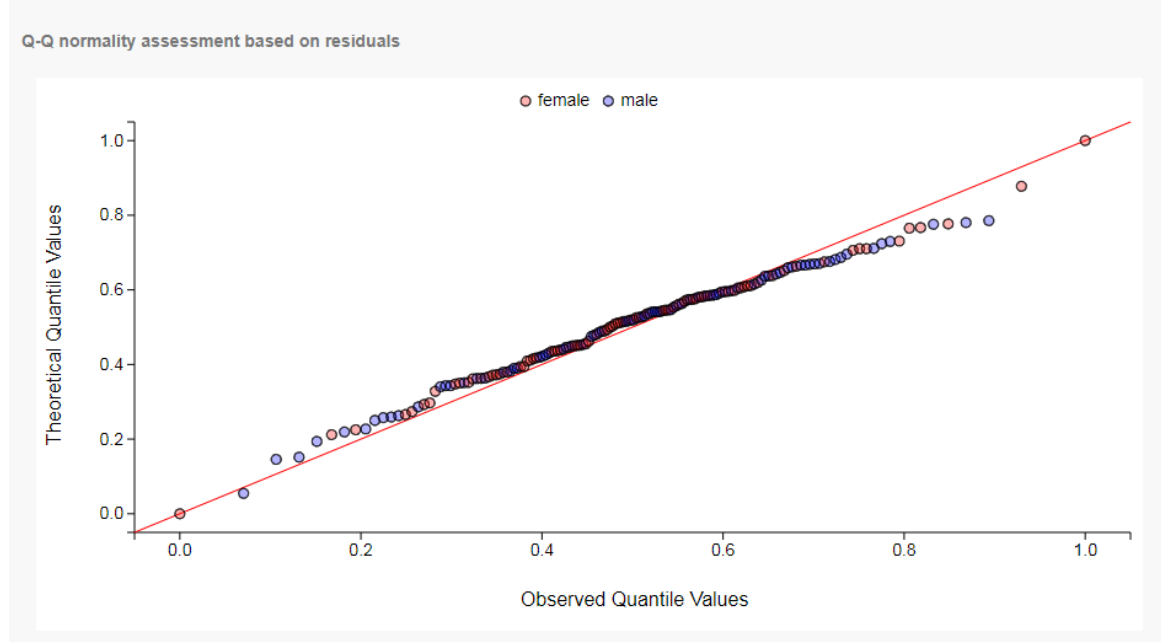
● STRAIN MEANS (UNADJUSTED)

● LEAST SQUARES MEANS (ADJUSTED)

● LEAST SQUARES MEANS (ADJUSTED), SEXES COMBINED

Figure 5 Visualization of Phenotype Measurement in on MPD Dataset (continued)

Currently more analysis tools are available for comparing measurement results within study (**Figure 6**). Correlation analysis (**Figure 7**) is available is phenotype measurement was conducted multiple times with at different time point (or with some other condition variation criteria). Various ways to compare measurements within a single dataset is also available (**Figure 8**). A pivot table (**Figure 9**) shows the quantitative results in a more concise manner.
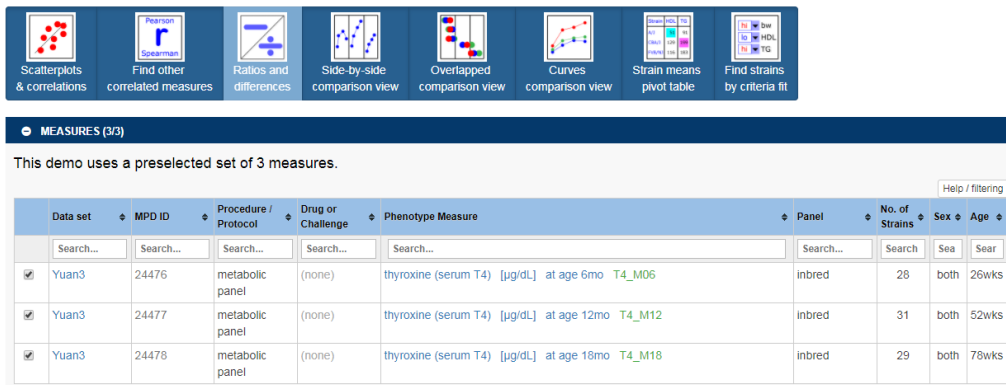
Figure 6 More Analysis Tools in MPD
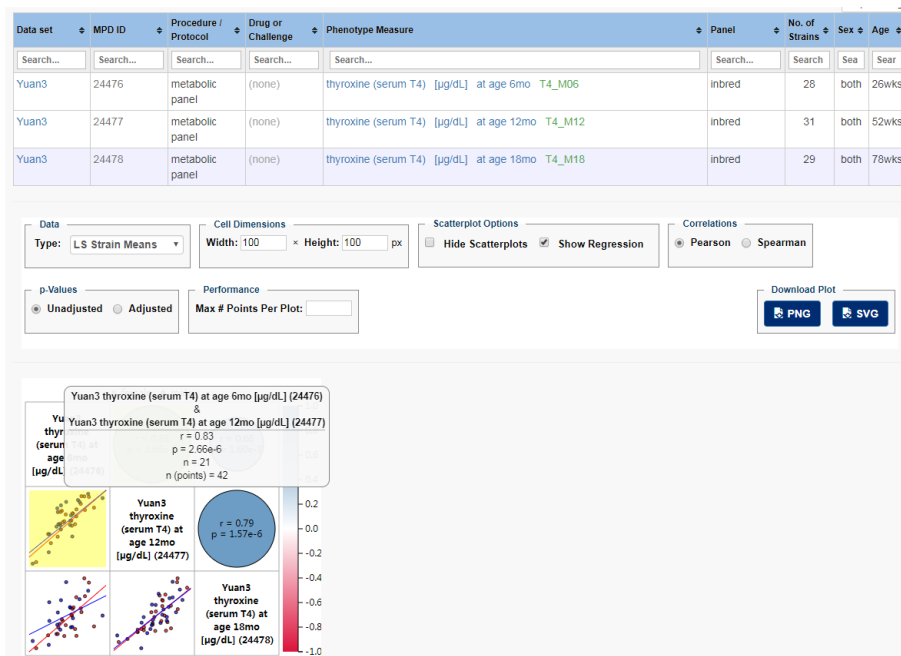
Figure 7 Scatter Plot and Correlation Analysis
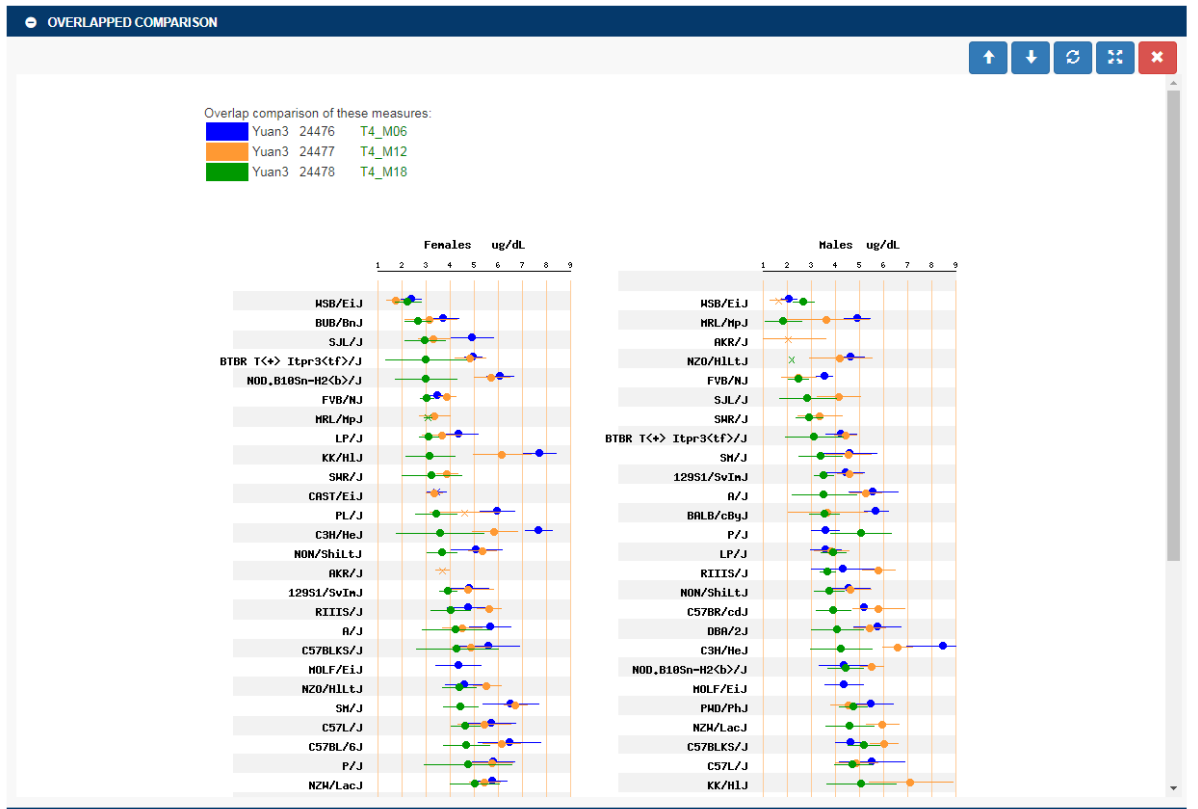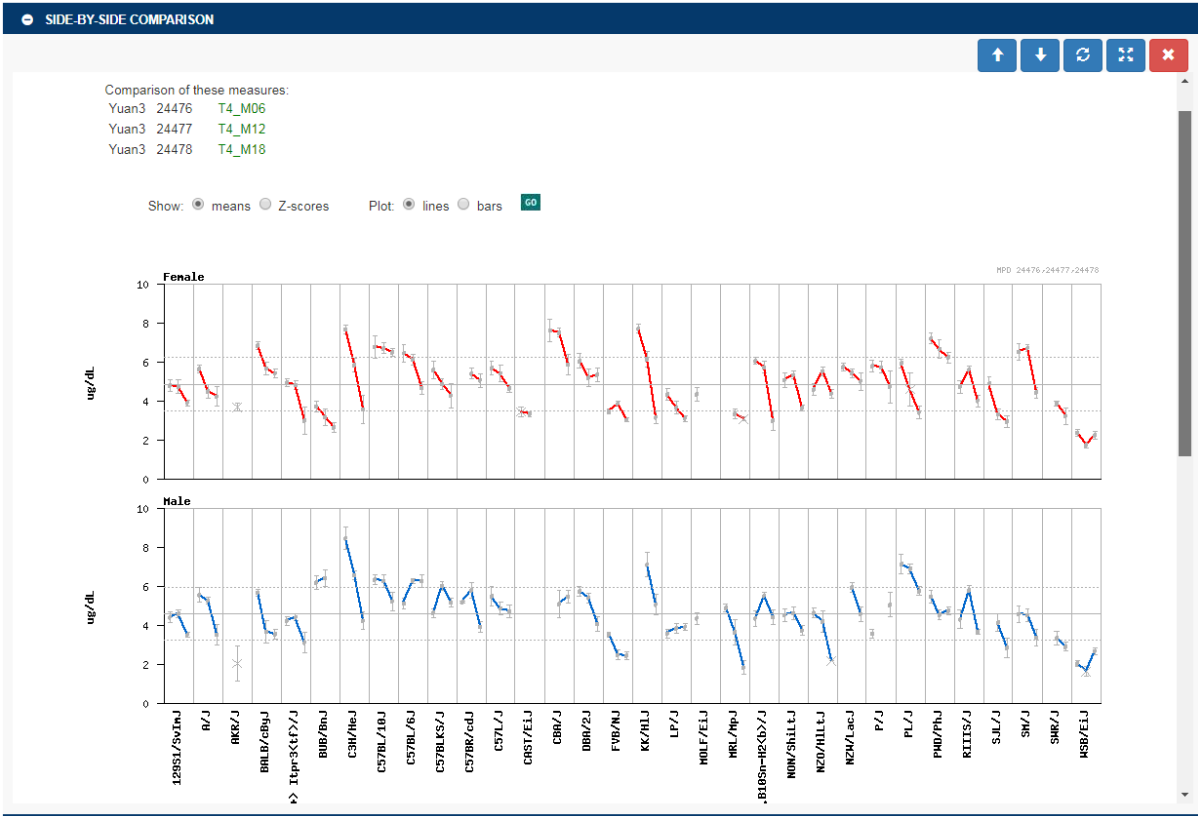
Figure 8 Side-by-Side (up) and Overlapped (bottom) Comparison

Strain means pivot table
Sort by column strain (alphabetical) ▼ ☐ Descending   GO
Highlight outliers yes ▼   N must be ≥ 3 ▼

Female

Strain	T4_M06 [µg/dL] MPD:24476	T4_M12 [µg/dL] MPD:24477	T4_M18 [µg/dL] MPD:24478
129S1/SvlmJ	4.83	4.78	3.94
A/J	5.68	4.52	4.26
AKR/J		[3.70]	
BALB/cByJ	6.87	5.70	5.44
BTBR T+ ltpr3tf/J	4.97	4.85	3.02
BUB/BnJ	3.75	3.20	2.68
C3H/HeJ	7.70	5.87	3.60
C57BL/10J	6.81	6.73	6.53
C57BL/6J	6.49	6.16	4.70
C57BLKS/J	5.62	4.90	4.31
C57BR/cdJ		5.44	5.09
C57L/J	5.72	5.44	4.67
CAST/EiJ	3.47	3.36	
CBA/J	7.65	7.53	5.90
DBA/2J	6.08	5.22	5.38
FVB/NJ	3.50	3.89	3.06
KK/HIJ	7.74	6.17	3.19
LP/J	4.38	3.69	3.14
MOLF/EiJ	4.37		

Male

Strain	T4_M06 [µg/dL] MPD:24476	T4_M12 [µg/dL] MPD:24477	T4_M18 [µg/dL] MPD:24478
129S1/SvlmJ	4.44	4.62	3.54
A/J	5.59	5.29	3.55
AKR/J		2.07	
BALB/cByJ	5.71	3.70	3.57
BTBR T+ ltpr3tf/J	4.26	4.44	3.14
BUB/BnJ	6.23	6.45	
C3H/HeJ	8.49	6.60	4.26
C57BL/10J	6.38	6.31	5.26
C57BL/6J	5.15	6.31	6.30
C57BLKS/J	4.67	6.05	5.20
C57BR/cdJ	5.23	5.82	3.95
C57L/J	5.53	4.89	4.75
CBA/J		5.13	5.50
DBA/2J	5.77	5.44	4.10
FVB/NJ	3.56	2.51	2.48
KK/HIJ		7.14	5.09
LP/J	3.61	3.87	3.94
MOLF/EiJ	4.39		
MRL/MpJ	4.92	3.67	1.86

Figure 9 Strain Pivot Table

In addition to MPD's own analysis tool, **Figure 10** shows another tool developed by International Mouse Phenotype Consortium using MPD data (26). Compared to previous tools that aim to facilitate candidate gene select for GWAS studies, this tool displays an informative summary chart of gene-phenotype relationships. There are also labels of phenodeviance, homozygosity/heterozygosity, and sexual dimorphism available. Again, those labels are assigned by curators and cannot be fully explored or validated by researchers.
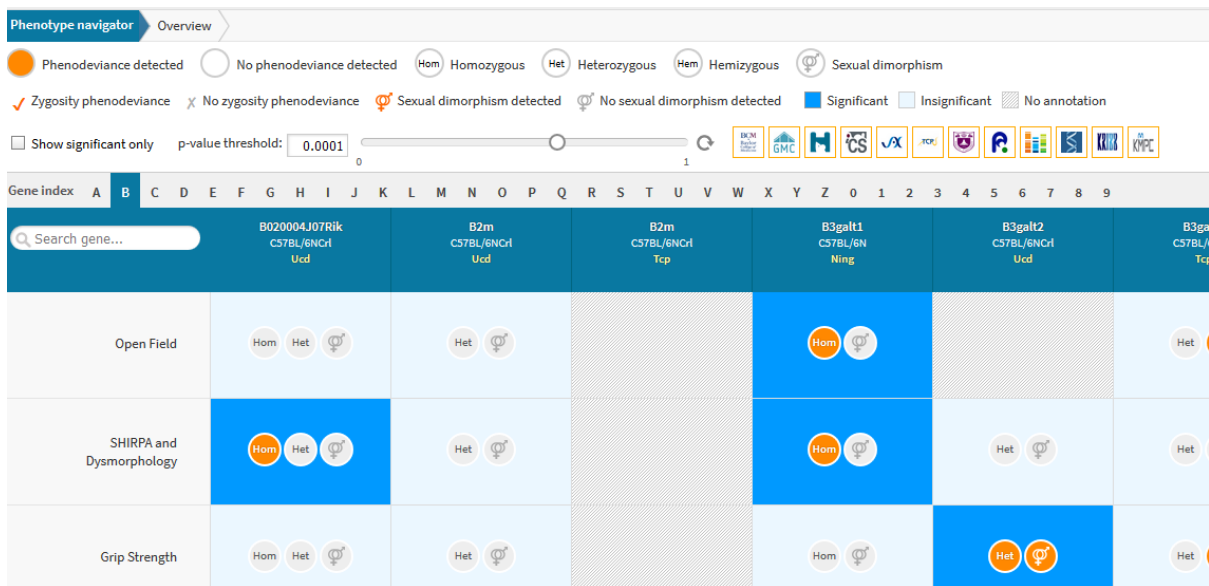
Figure 10 "Phenoview": a Tool that Shows Gene-Phenotype Relationships (26)

However, existing problem with analysis tools provided by MPD is that users can only query by free text keywords (**Figure 11**). This might be an obstacle for users not familiar with their database and ontology to start exploring the database.
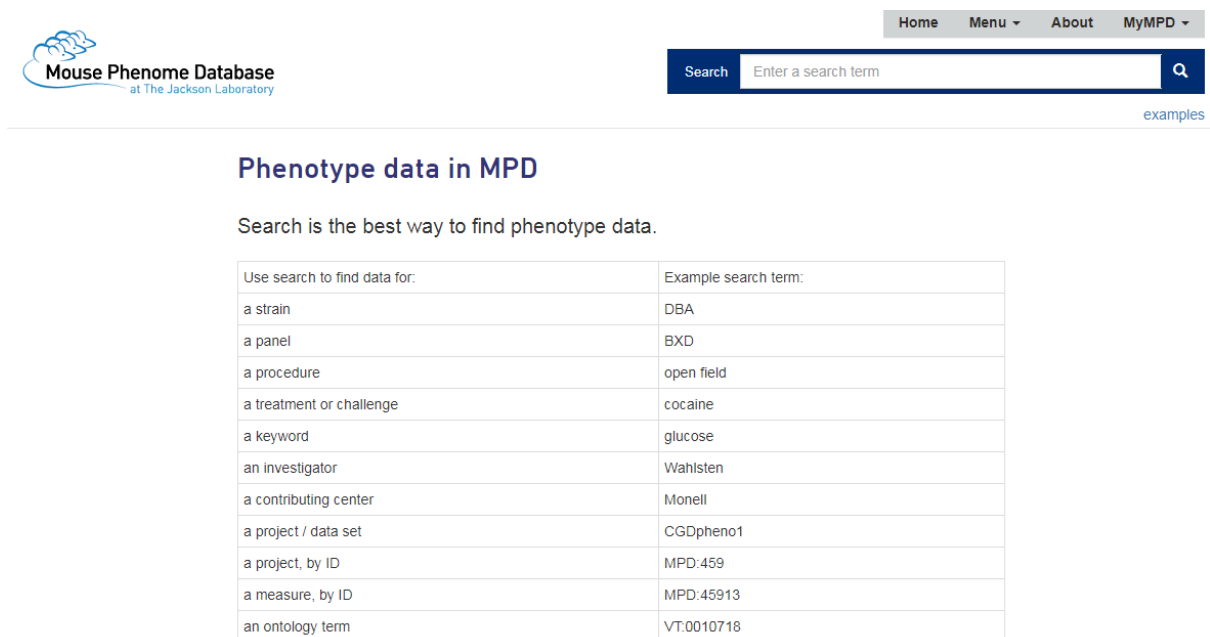
Figure 11 Landing Page of MPD for Viewing Phenotype Data

In addition, it only allows users to view records within one study at a time. This limits the ability to compare and further integrate phenotype measurements across different studies

for the same phenotype. MPD have not attempted to integrate phenotype data statistically. Hence no summary or expected range of phenotype data for a specific strain is available for reference. This hinders users' ability to fully explore phenotype differences across mouse strains. Therefore, there is an urgent need to: 1) develop a more user friendly query interface for users to browse data in the database (by providing ontologies for query, showing summaries of data entry counts, etc.); 2) develop a pipeline or protocol for evidence-based phenotype assignment utilizing quantitative phenotype data; 3) develop a tool that supports visualization of integrated quantitative phenotype data in a systematic manner; 4) increase research reproducibility by enhance data transparency and interactive data exploration by researchers.

4)   Rat Genome Database and PhenoMiner

The Rat Genome Database (rgd. mcw. edu) is the most comprehensive data repository and informatics platform for the laboratory rat (27). RGD maintains and updates data about genes, transcripts, variants and provides functional annotations for disease, pathways, drug/chemical-gene interactions, gene function and biological processes. In addition to genomic data, RGD also curates and integrates data on strains, QTLs and experimental phenotype measurements across hundreds of strains. Data is curated from various sources: published literature, submitted by individual researchers, and acquired through bulk data pipelines from other public repositories (27, 28).

As an initial step in developing phenotype profiles for individual strains, RGD created the PhenoMiner project (29) to integrate quantitative phenotype data from individual research projects as well as the PhysGen Program for Genomic Application (24), and the NBRP Rat Phenome Project in Japan (http://www. anim. med. kyoto-u. ac. jp/nbr/phenome. aspx) (20).

Each of these large-scale projects were created to present their own quantitative phenotype data while PhenoMiner was created to integrate data from all of these large scale projects as well as published results from the literature and data directly submitted from investigators.

RGD uses multiple ontologies to standardize and integrate data from many resources including the rat strain ontology (RSO) (31), clinical measurement ontology (CMO), measurement method ontology (MMO) and experimental condition ontology (XCO) (32, 33). The database structure (**Figure 12**) includes 11 tables to standardize and link information on sample measured, type of phenotype measurement and value, measurement method and the experimental conditions under which the measurement was made. The database structure provides the ability to record study information, link multiple experiments to a single study and link multiple phenotype records to a single experiment. The EXPERIMENT_RECORD table includes information about the quantitative data for a particular record (value, standard deviation, standard error and measurement units) as well as the Sample ID, Clinical Measurement Ontology ID, Measurement Method Ontology ID, and Experimental Conditions Group ID. From EXPERIMENT_RECORD table and links to the SAMPLE (rat strain, age, sex, number of animals in the experiment), CLINICAL_MEASUREMENT (Clinical Measurement Ontology table), MEASUREMENT_METHOD (Measurement Method Ontology Term, measurement site, apparatuses and measurement duration), CONDITION_GROUP (data on the diet and special treatment of rats and ordinality and duration of each condition, e. g. "controlled sodium content drinking water (1 %) (between 9 and 12 days)", "controlled sodium content diet (0. 3 %) (for 12 days) then controlled sodium content diet (2 %) (for 24 days)").
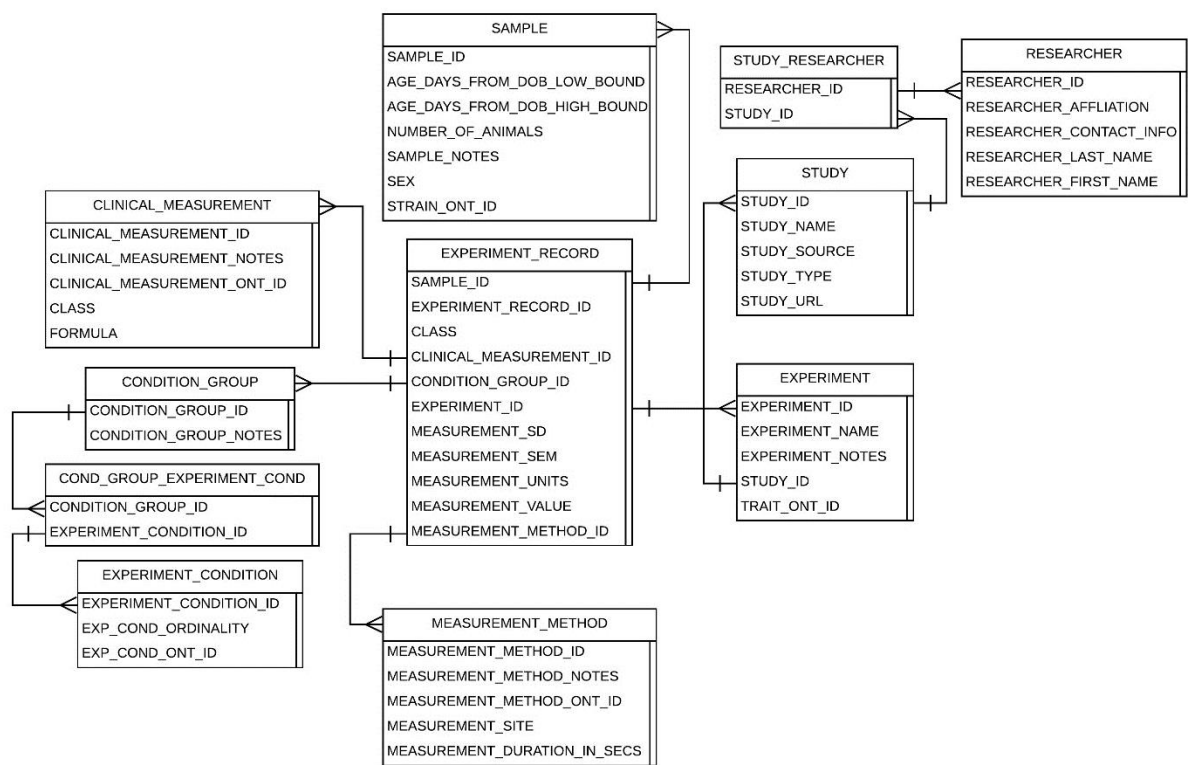
Figure 12 Database Structure of PhenoMiner

The creation of this structure and the use of multiple ontologies to standardize four components of a quantitative phenotype record has resulted 60, 000 entries (28). A summary of quantitative phenotype records in RGD is listed in **Table 1**.

Table 1 Summary of PhenoMiner Data in RGD

Phenotype Category	Number of Records	Number of Strains
alimentary/gastrointestinal measurement	8	5
blood measurement	19491	472
body morphological measurement	10503	777
body movement/balance measurement	98	13
body temperature	349	63
cardiovascular measurement	18334	951
cell measurement	6318	316
chemical response/sensitivity measurement	2285	98

consumption measurement	339	19
disease population measurement	426	163
disease process measurement	502	85
endocrine/exocrine system measurement	64	20
exudate measurement	216	4
growth measurement	23	9
immune system measurement	3408	311
liver/biliary measurement	739	236
mortality/survival measurement	52	21
musculoskeletal system measurement	202	14
nervous system measurement	697	242
organ measurement	12037	790
renal/urinary measurement	4883	634
reproduction measurement	62	6
respiratory system measurement	4923	241
tissue composition measurement	139	24
tumor measurement	220	77

The advantages of PhenoMiner are: 1) it standardizes quantitative phenotype records for rat strains, clinical measurements, measurement methods and experimental conditions using ontologies; 2) this standardization allowed for the integration of data from large scale and small scale phenotype projects, 3) users can query and retrieve data from multiple experiments and visualize results, 4) users can also download retrieved data. In this paper (29), authors demonstrated the use of RGD visualization to compare cardiovascular and renal phenotypes of SS, SS congenics and SS mutants under salt-induce hypertension model. In a similar manner, we did a query and the results are display as below (**Figure 13**).
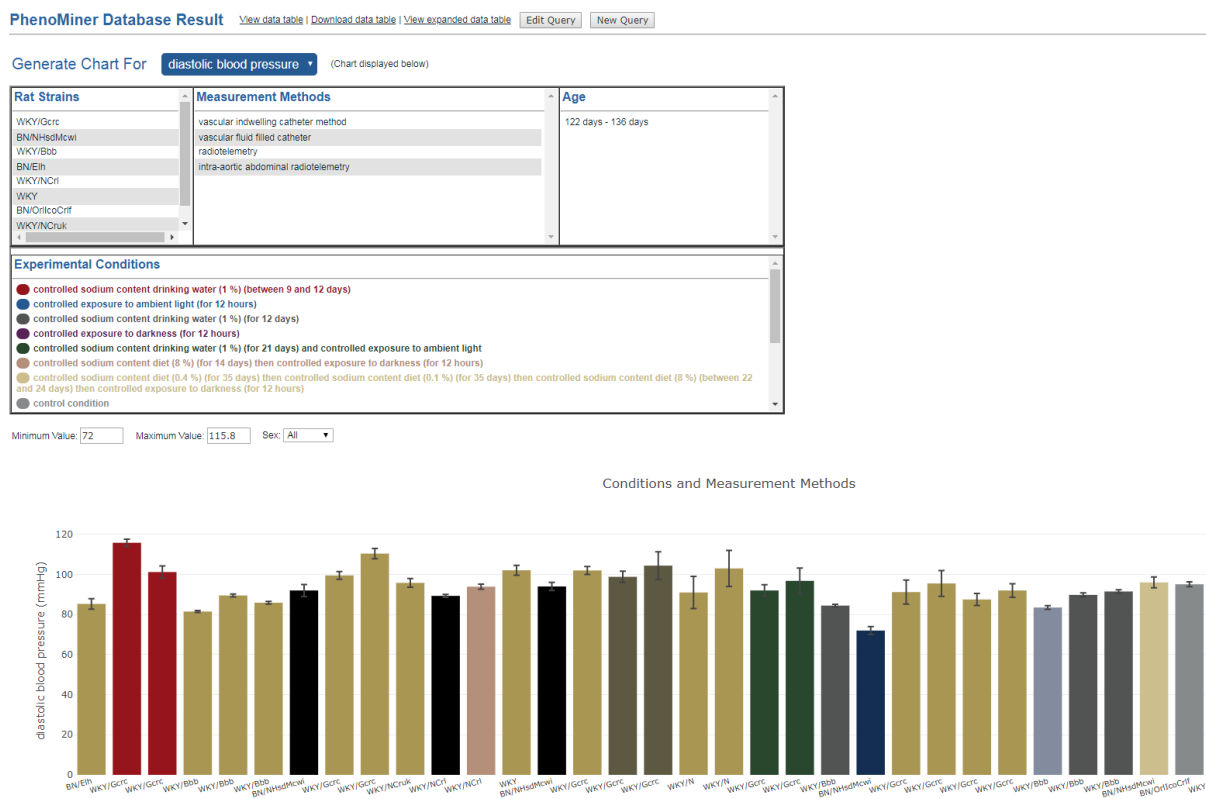
Figure 13 RGD PhenoMiner Visualization of Individual Records for Query={Strain="BN"&"WKY", Phenotype="diastolic blood pressure"}

While systematic data integration and visualization in PhenoMiner enabled qualitative comparisons across experiments and conclusions to be drawn, the drawback of the current PhenoMiner portal is its limited ability for statistical integration of data. Further quantitative analysis using a standardized statistical tool would provide more insights in understanding

20

rat strains in terms of disease models. Moreover, an integrated analysis of overall strain phenotype measurement is desirable for researchers to better understand cross strain differences. Currently researchers often choose strains for use as disease models using data points from a limited number of experiments or based on availability, prior use or familiarity. The availability of statistically determined expected ranges for quantitative phenotypes for multiple individual strains, for those often used as controls and for rat in general would improve the ability of investigators to choose appropriate strains for their studies. One example of the need for statistically determined expected ranges was illustrated with the commonly used outbred CD-SD rats often seen to be less responsive to estrogenic substances than F344 inbred rat strains for various estrogen-sensitive endpoints (34, 35). The authors showed that the commonly held view of the distinct phenotype characteristics of two strains did not hold true when the results from a collection of experiments were aggregated. This case illustrates why it would be desirable to have a statistically determined expected range as a quantitative reference for analyzing measurement data across different experiments. Such references would help researchers choose better strain models for their research objective and assist them in examining potential factors that might cause measurement variation. The expected range would also provide a standard interpretation of experimental results from different laboratories.

1. 3   Motivation and Aims

Expected ranges for different phenotypes for individual rat strains as well as across multiple strains are currently not available. Nor is there a tool to help researchers to select an optimal strain as a disease model. Therefore, the motivation for our work is to take advantage of the huge volume of quantitative phenotype data in the Rat Genome Database

to establish expected ranges for different rat strains. By comparing expected ranges for different rat strains, researchers will be able to choose the optimal disease model and control for their investigations. To achieve this goal, our work involve the following key steps: (1) establish a standardized phenotype range for different rat strains using the meta-analysis method, (2) create tools to mine and visualize data for individual strains and across strains, (3) cluster substrains into different sub-populations according to phenotype correlations.

In the first step, we conducted a meta-analysis to effectively synthesize archived phenotype data in the PhenoMiner database, stratify each population based on strain (inbred/outbred/congenic/transgenic/mutant), gender, age, . and produce comparable expected ranges of important physiological phenotypes (such as heart weight, systolic blood pressure). Statistical tests will also be performed to assess differences between different strains of certain phenotypes. The result from this work will greatly benefit researchers using rat models in determining a proper strain, age, gender and all relevant parameters for their studies.

Tools was developed to allow users to search, retrieve and visualize expected ranges for a variety of phenotypes for a single strain or across strains for a single or multiple phenotypes. Next, we integrated quantitative expected ranges from the previous steps as essential phenotype features for rat disease models and use those features to cluster strains into subtype clusters. The clustering result will provide in-depth insight of the substrain variation in rats. It can be useful for researchers when they are designing experiments and trying to pick a substrain suitable for their specific research objective.

2   Generating Phenotype Expected Ranges

2. 1   Background

A meta-analysis approach is a powerful tool for determining expected ranges for particular

phenotype measurements across multiple experiments. This approach involves statistical

techniques for combining measurements or findings from independent studies to draw

insights on a specific research question. It is often used to assess the effectiveness of clinical

treatments by combining data from several randomized control trials. It provides a precise

estimate of treatment effect, overcoming biases that could occur when examining a single

study and it offers a systematic synthesis of the experimental data. A recent research study

revealed that single-laboratory studies with large sample size produce results that are more

precise but less accurate and therefore less reproducible (36). By contrast, multi-laboratory

designs including as few as 2 to 4 laboratories increased coverage probability by up to 42

percentage points without a need for larger sample sizes. They also demonstrated that

within-study standardization is a major cause of poor reproducibility (36).

A systematic review methodology is essential as the first step of meta-analysis. The

objective of a systematic review is to present a balanced and impartial summary of the

existing research, enabling synthesis of all relevant studies of adequate quality (37). This

stresses the need to take great effort and care to find all the relevant studies (published and

unpublished), and to assess the methodological quality of the design and execution of each

study (38). The standardized and integrated data at RGD is a good resource of systematically

managed experimental phenotype measurements. It includes both data from published

studies from current biomedical literature as well as large scale data from rat community

repositories such as PhysGen (24), PhysGen Knockout Project (30) and the Rat Phenome Project (20).

Meta-analysis is not just a single statistical analysis; it involves a pipeline of preliminary stratification, exploratory decision making (publication bias and sensitivity analysis) before the final statistical meta-analysis can be performed. The pipeline for analyzing RGD PhenoMiner data consists of four major components (**Figure 14**). In the following sections, I will introduce the methods for each step in the pipeline followed by its corresponding results since results from each step are useful in deciding the method used in the next step. In addition to developing the algorithms for each component, a user interface was created to facilitate determination of appropriate parameters, and to dynamically implement the workflow needed for the analyses (described in further detail below).
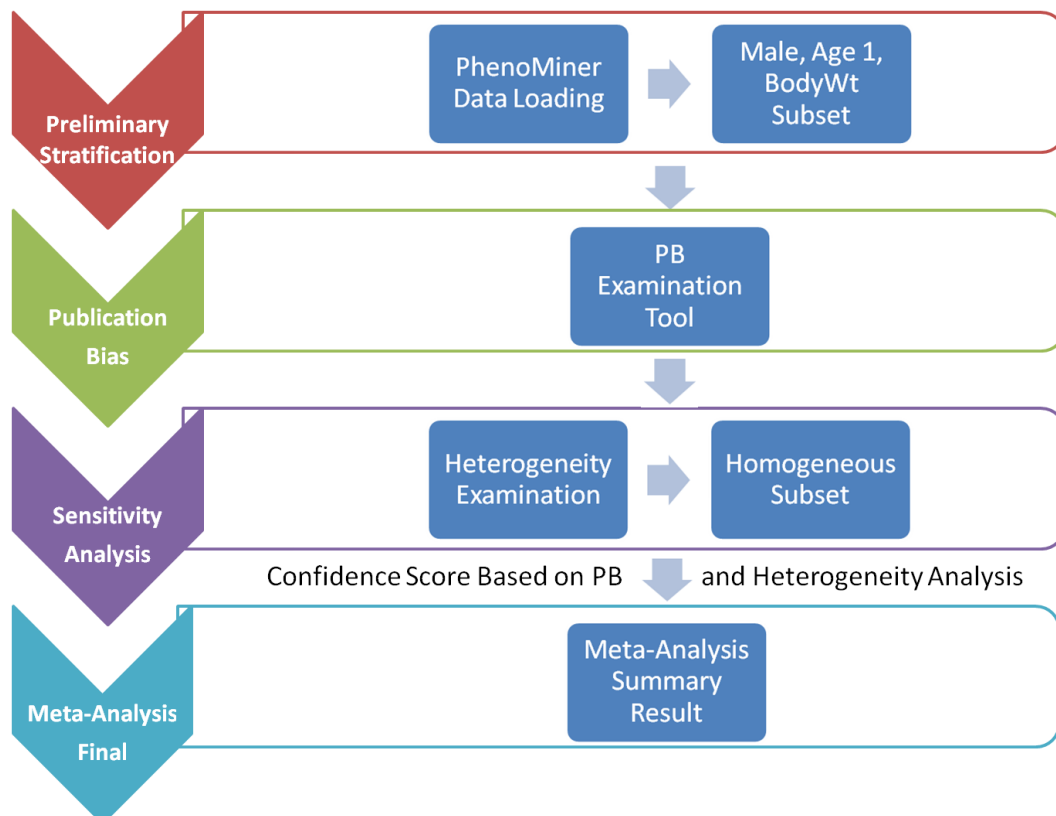
Figure 14 System Pipeline for Meta-Analysis

2. 2   Preliminary Stratification

First it was necessary to choose a subset of phenotype measurements based on preliminary stratification, which included strain, sex, age group, and phenotype measurement methods. **Figure 15** show the interface created to dynamically conduct the preliminary stratification step. Options for strain and phenotype measurement methods depend on the major phenotype under analysis. Age group divisions were decided by expert heuristic definitions of young, adult and old rats. However, for different phenotypes, young and adult divisions can vary in order to achieve a low heterogeneity score. As a result, age group division is a data-driven heuristic score with expert provided prior definition.

Figure 15 PhenoMiner Preliminary Stratification for Body Weight

2. 3   Publication Bias with Funnel Plot

Because much of the data included in this study arises from published research, one key concern is publication bias which arises because experiments with negative findings are less likely to be published than those that highlight results which support hypotheses (39). Funnel plots can be used to assess the presence of publication bias (40) by displaying the studies included in the meta-analysis in a plot of measurement value or effect size (explained in detail in the statistical analysis section) against sample size or another measure

of precision (41, 42). The expected picture should be a symmetrical inverted funnel (43). This is in accordance with the assumption that smaller studies have more chance of variability than larger studies. An asymmetric plot suggests: (1) smaller studies showing no effect might be missing or (2) small studies tend to have larger effect sizes (44). The first reveals a true publication bias while the second does not. There are also controversies over the use of funnel plots due to disputes over appropriate interpretation of asymmetry (45-47). For example, true heterogeneity in study population (due to subgroups with a different intervention effect) will lead to funnel plot asymmetry (47). In addition, chance is also critical for interpretation of funnel plot asymmetry since most meta-analyses in the biomedical field contain few studies (48). Therefore we need to examine closely before reaching a conclusion of publication bias (49).

There are two estimators (Macaskill (50) and Egger (40)) for detecting publication bias. In our pipeline, we provided two different bias detection methods: funnel plot and funnel regression test using the Egger estimator.

In 1997, Egger et al. proposed an estimator for visualizing asymmetry in the funnel plot (40). In addition to the simple visualization of asymmetry, they also used a regression test to measuring asymmetry quantitatively. The regression test is a linear regression of normalized effect size estimate (value/SD) against precision(1/SD). The assumption of the regression test is that a homogeneous set of trials (without publication bias), will regress toward a line that runs through the origin (intercept = 0), with the slope indicating the size and direction of effect (51). When the regression line runs through the origin, it indicates a symmetrical funnel plot. However, the Egger test has a relatively high false positive rate (higher type I error rate).

The Macaskill estimator was found to have a lower false-positive rate in simulations (50) than the Egger estimator, though at the expense of lower power (higher type II error rate). The Macaskill estimator is similar to the Egger estimator but it used total sample size as a measure of precision. Use of sample size reduces the correlation between the effect size and its SD thus avoids violating an assumption of regression models that an independent variable is subject to random error. A minor modification of Macaskill's test, with the inverse of the total sample size as a measure of precision, produces more balanced type I error rates (42).

In 2005, Harbord et al. developed a test that maintains the power of the Egger test while reducing the false positive rate, which is especially significant with the Egger test when there is a large measurement value, small number of observations per trial or all trials are of similar sizes (52). However, the original Egger test should be used instead of the Harbord method if there is a large imbalance in size between studies.

Since the number of animals in our studies varies from 1 to 220 (see **Figure 29**), and also considering the trade-off between power and type I error rates, we finally chose the original Egger test for our analysis. An asymmetry score is calculated as the ratio of intercept for the regression line to average value for the measurements in the group under analysis.

$$\text{Asy} = \frac{\text{intercept}}{\text{average\_value}}$$

[F344] Total:13; Asymmetry_(b/avg):-0.0064; Asymmetry_slope:0.0048

Funnel Plot for Body Weight (N vs value, Macaskill Method)

Funnel Plot for Body Weight (1/SD vs value, Egger Method)

Galbraith Plot for Body Weight (1/SD vs value/SD, Egger Method)
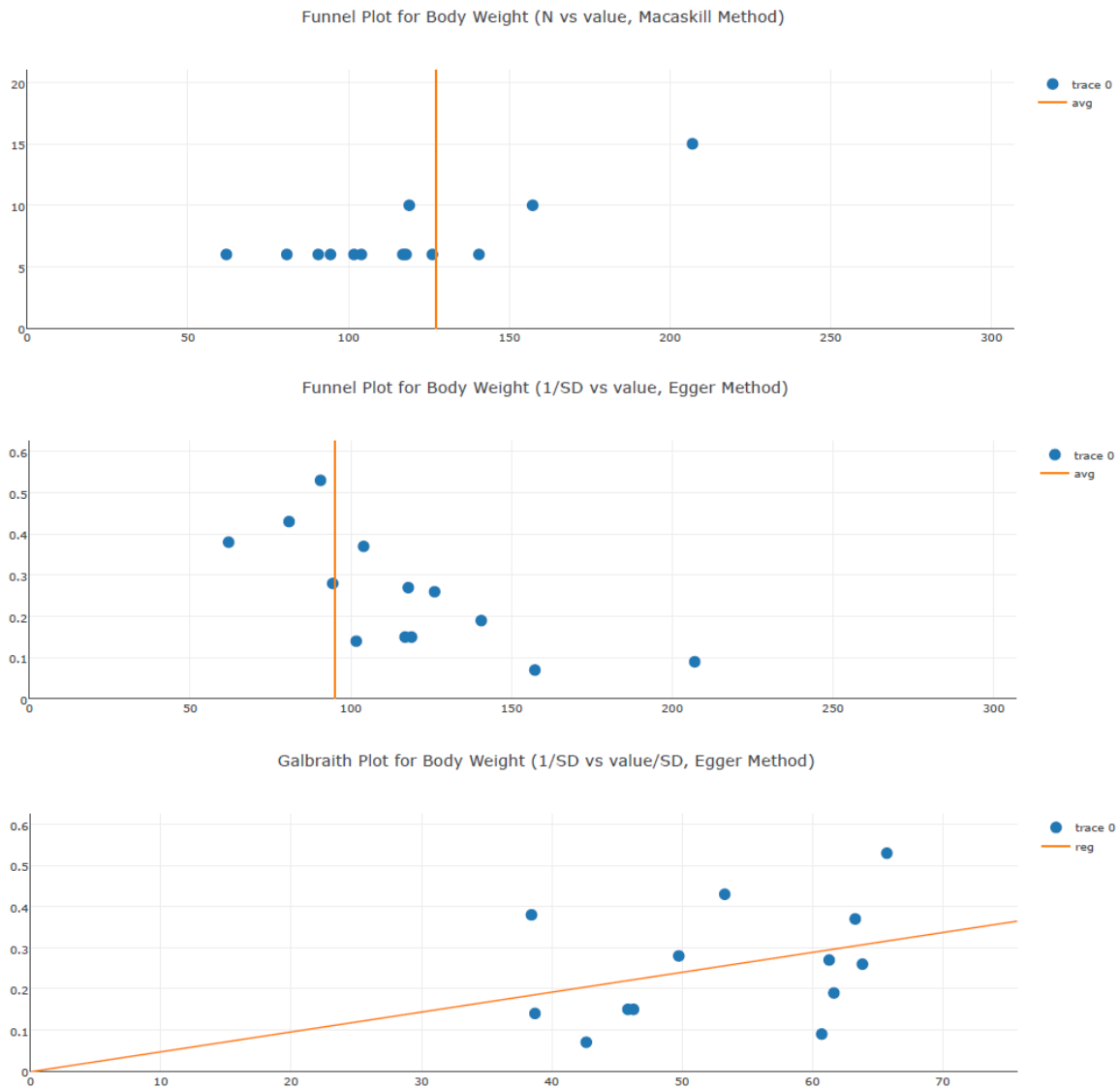
Figure 16 Publication Bias Examination Demonstrations for Body Weight

2. 4   Optimal Number of Experiments for Meta-Analysis Quality Control

To assure the quality of meta-analysis, we need to assign a confidence level to our meta-analysis model (here we used a binary parameter with value "confident" and "low_confidence") and results given a specific set of phenotype measurement data (in a single meta-analysis). While each meta-analysis is based on a unique phenotype-strain pair, for different phenotype-strain pairs, the total number of experiments can vary significantly.

Hence, we examined the relationship between Asymmetry score (Asy) from publication bias analysis and total number of experiments (in a single meta-analysis) to identify potential biases in our analysis. The example below used Body Weight data (**Figure 17**, **Figure 18**). The result shows that the Asymmetry score (Asy) was reduced significantly with *four* or more experiments in one meta-analysis.
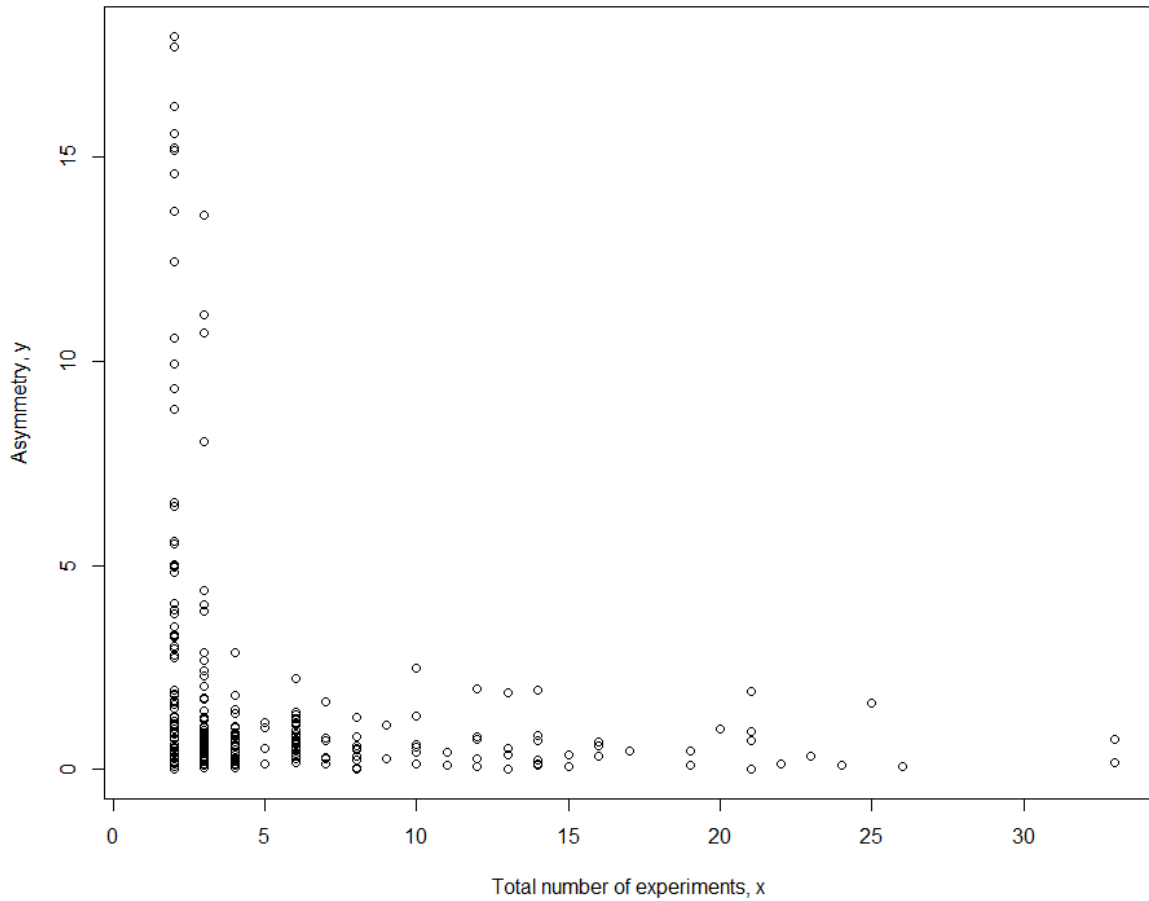
Figure 17 Relationship between Asymmetry Score and Total Number of Experiments for Body Weight
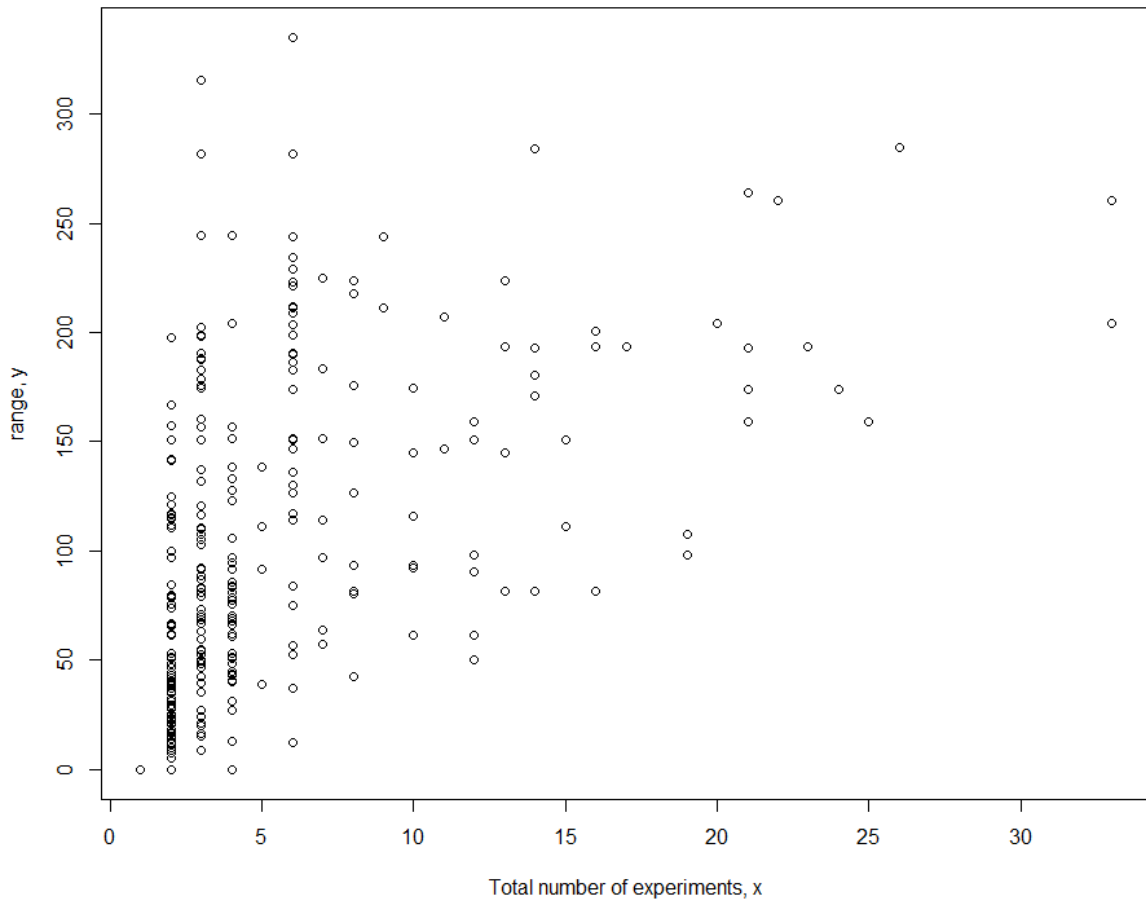
Figure 18 Relationship between Range and Total Number of Experiments for Body Weight

We also examined range distribution and its relationship with the total number of experiments (in a single meta-analysis). **Figure 17** shows that data range for different meta-analysis datasets also varies with total number of experiments for each meta-analysis. The range is defined as the difference between maximum value among all studies and minimum value among all studies. For a meta-analysis with fewer experiments, the range between studies can be falsely small. This indicates that meta-analysis without enough experiments might have false negative heterogeneity representations (heterogeneity not revealed) and thus the meta-analysis model choice might be wrong (fixed-effect or random-effect).

2. 5   Sensitivity Analyses with Forest Plot

Data with poor quality for non-systematic reasons is often an issue in meta-analysis so selection, inclusion and integration (or population stratification) of data is an important factor for consideration. Those decisions may affect major findings, so researchers usually carry out some sensitivity analysis prior to integration of data. The usual way of displaying data for sensitivity analysis is by a forest plot. This displays the findings from each individual study using a blob or square (53), the x-axis is the value of measurement or effect size. The size of the blob or square is proportional to the sample size. A horizontal line representing 95% confidence interval is drawn around each of the studies' squares to represent the uncertainty of the measurement. The meta-analysis result is displayed as a diamond. After exploring the forest plot of the study cohort, the main findings can be changed by varying the approach to integration (or population stratification). An effective sensitivity analysis will explore the effect of excluding various categories of studies, such as outlier data (outliers need to be excluded for justifiable reasons), data without specified sex information, or data from unpublished studies. It may also examine how consistent the results are across various subgroups (perhaps defined by subject population stratification, type of measurement method or condition).

A useful sensitivity analysis is a series of repeated meta-analyses, usually omitting one study at a time. A heterogeneity score is calculated and the meta-analysis model (fixed-effect or random-effect) is also chosen based on the heterogeneity score. Such an 'exclusion sensitivity plot' by Bax *et al.* (54) reveals any study/observation that has a particularly large influence (outlier) on the result of the meta-analysis. An interactive interface in our tool provides users with the ability to decide on inclusion/exclusion of any study/observation in the meta-analysis before proceeding with the next step in the analysis. Additionally, users of

the pipeline may identify subgroups in their data and may decide to adopt a modified

sensitivity analysis by excluding a group of studies/observations and creating new

stratification criteria. For example, users may find substrain A/X and A/Y to have different

measurements for phenotype T. Instead of analyzing phenotype T using all data for strain A,

users may want to further stratify the population using substrain characteristics. **Figure 19**

shows an example of sensitivity analysis that identified the study result (in red circle) to be

an outlier. The example is for the F344 strain Body Weight phenotype. After initial analysis,

we found that the large value of the outlier is due to sample age. 65 days might not be a

homogenous group member for young rat group in terms of Body Weight although it might

be acceptable for another phenotype.

I2 statistic: 99.2%, Q statistic: 1107.3149, Summary: 93.99(fixed-effect);94.89(random-effect)

I2 statistic: 99.3%, Q statistic: 1103.0054, Summary: 93.16(fixed-effect);93.91(random-effect)
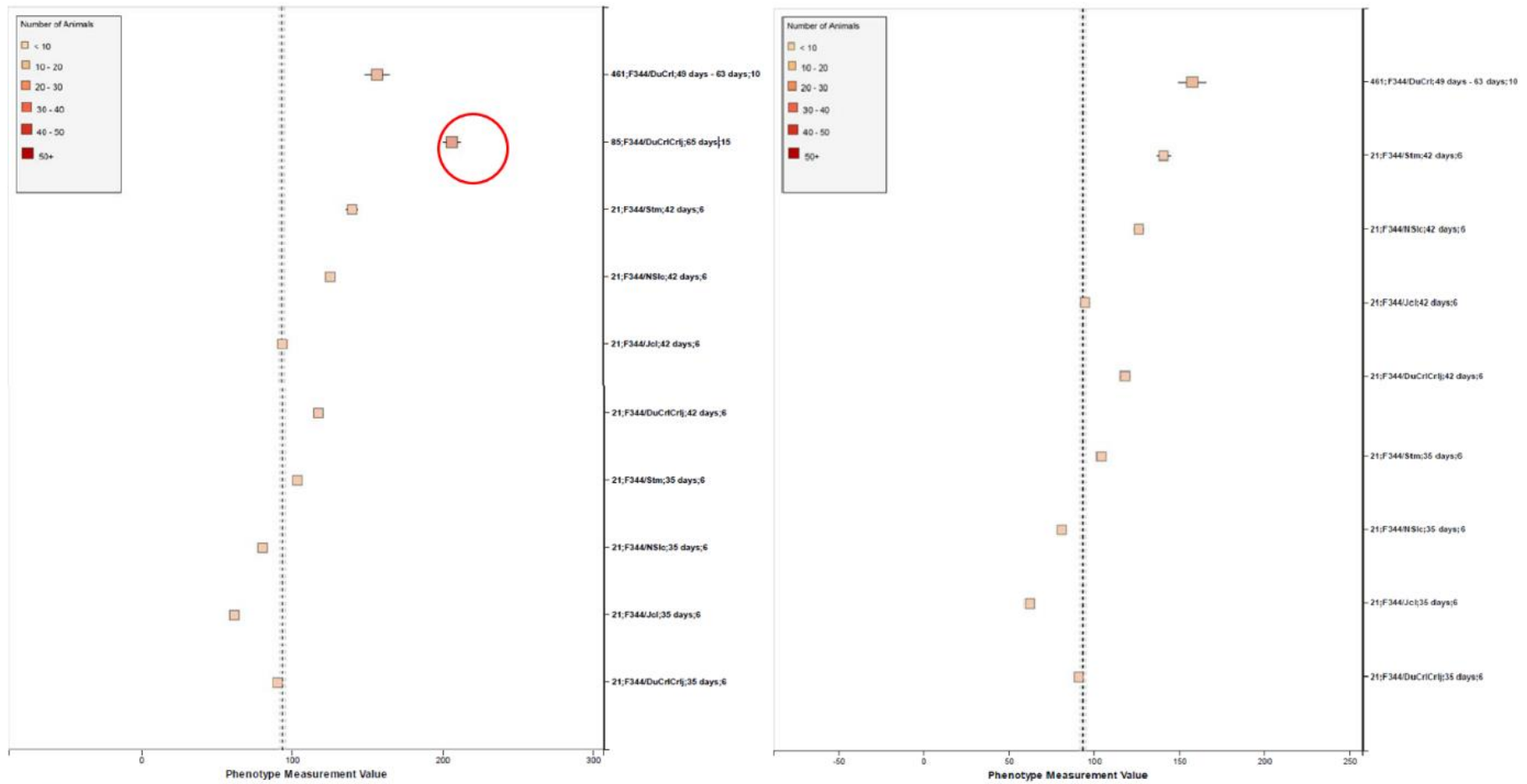
Figure 19 Sensitivity Analysis Example Before Outlier Removal (Left) and After Outlier Removal (Right)

2. 6   Statistical Meta-Analysis

1)   Heterogeneity and Meta-regression

In the meta-analysis, we needed to evaluate whether the results from different studies can be "combinable". This involved examining heterogeneity. Statistics commonly used for testing heterogeneity include Cochrane's Q ($Q = \sum (w \times ES^2) - \dfrac{\left[\sum (w \times ES)\right]^2}{\sum w}$), a statistic

based on the chi-squared test and the I^2 statistic ($I^2 = \begin{cases} \frac{Q-(k-1)}{Q} \times 100\% \; for \; Q > (k-1) \\ 0 \; for \; Q \leq (k-1) \end{cases}$)

(55, 56). The Weight for each study is $w_i = \frac{1}{s_i^2}$, where s_i is the standard deviation of that study. ES represents effect size, which is a key concept in meta-analysis when assessing heterogeneity. It aims to offer a standardized heterogeneity comparison among different measurements, similar to the idea behind the t-test. It can be calculated as $\overline{ES} = \dfrac{\overline{X}}{s}$ when the original

measurement value is used for meta-analysis. In our analysis, the Q statistic was unsatisfactory as it depends heavily on the scale of measurement and has no absolute interpretation for comparison. Thus the I^{2}statistic was more attractive because it scores heterogeneity between 0% and 100%, with 25% corresponding to low heterogeneity, 50% to moderate and 75% to high (57). It interprets the percent of the total variance that is due to between study heterogeneity. However, both methods may sometimes fail to detect heterogeneity when it is actually present (58).

If the study results for a sub-population are relatively homogenous, we can integrate the results using a general meta-analysis method (fixed-effect model). If heterogeneity exists, we can further stratify the current sub-population based on a conceptual stratification method (e. g. stratify based on age and sex) or use random-effect model meta-analysis. At

this point, it is important to investigate what may have caused the heterogeneity. Meta-regression is a technique which allows researchers to explore factors contributing to the heterogeneity (59, 60). The simplest type of meta-regression uses summary data from each entry, such as the average effect size, age, sex, experimental condition, and experimental method. The result from the meta-regression provides us with useful insights on how to further stratify sub-populations. This approach is valuable, but it has only limited ability to identify significant factors (61). If no factors were identified, that means there might be systematic differences between studies, leading to heterogeneity. Then we would need to define special statistical parameters (inter-study variance used in random-effect model) to interpret the systematic heterogeneity of the results.

2) I^2 Statistics Cut-off Threshold

In the previous step, we decided to exclude or take a lower confidence in meta-analysis when the total number of experiments was below four. In this step, we needed to decide the cut-off threshold for the I^2 statistic to decide the model choice for each meta-analysis. In the example in **Figure 17**, we can see that I^2=0. 85 is an optimal cut-off threshold to separate high and low heterogeneity datasets. The four quadrants in **Figure 17** represent different characteristics of datasets for each meta-analysis task. Quadrant one represents a high asymmetry score and high heterogeneity, which may be caused by publication bias or true heterogeneity (e. g. extreme outliers). Quadrant two represents a high asymmetry score and low heterogeneity, which may indicate true publication bias. Quadrant three represents a low asymmetry score and low heterogeneity, which indicates that we should choose the fixed-effect model. Quadrant four represents a low asymmetry score and high heterogeneity, which indicates that we should choose the random-effect model. From this

summary plot, we can determine model choice for data in Quadrant three and four. For

data in Quadrant one and two, we need to further confirm the existence of publication bias

before any conclusions can be made.
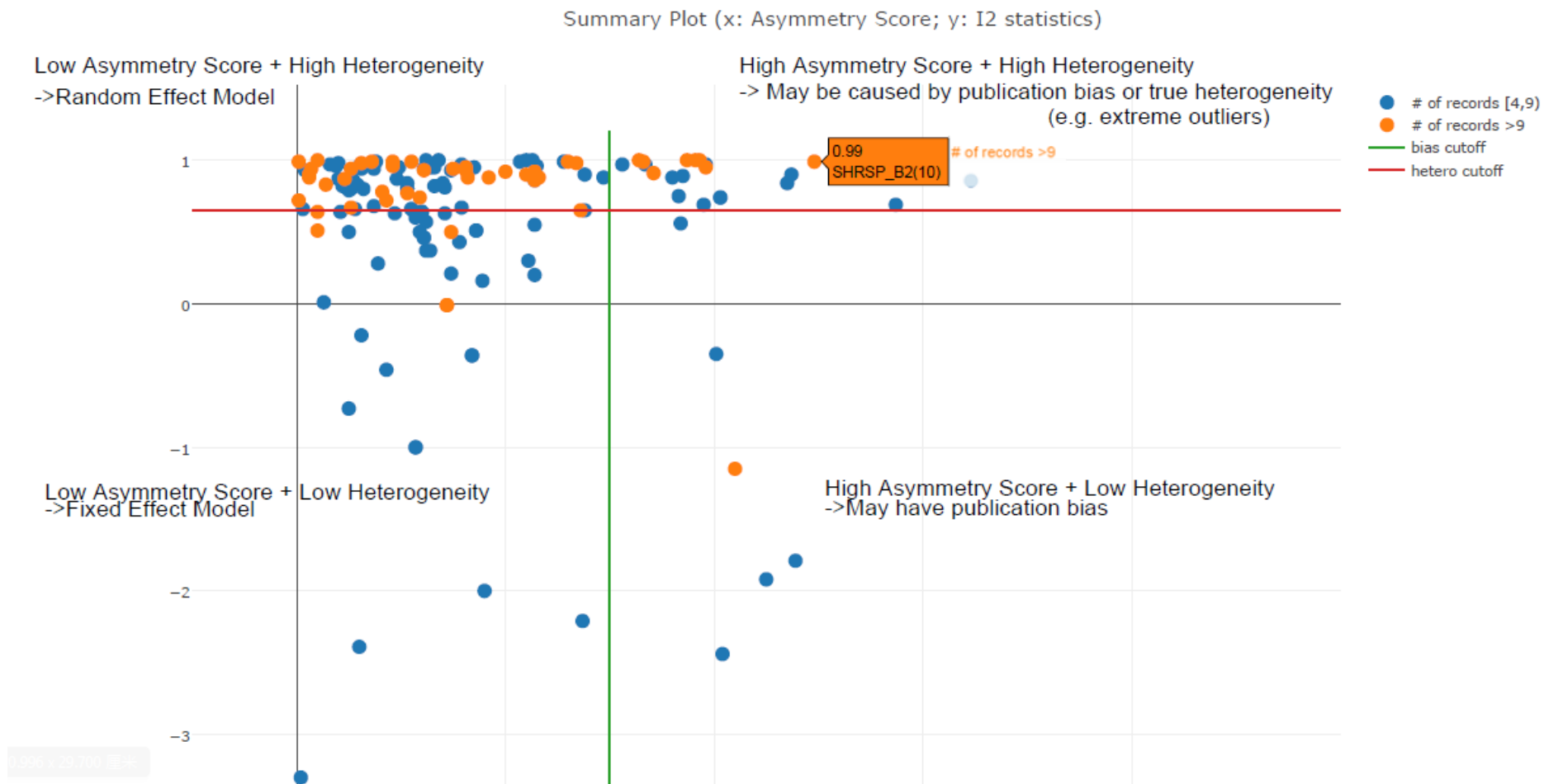
Figure 20 Scatter Plot of Asymmetry Score and I2 statistics for Different Meta-Analysis Results for Body Weight Data (x: Asymmetry Score; y: I2 statistics)

3) Fixed-Effect Model vs Random-Effect Model

The presence or absence of heterogeneity influences the subsequent method of analysis. If heterogeneity is absent, then the analysis employs a fixed-effect model. This assumes the size of the system effect is fixed across all studies and the variation seen between studies is completely random. However, as studies generally vary in size and variance, each study is considered to have different precision. In meta-analysis, the key concept is to assign a weight to each study while synthesizing results. Generally believed by statisticians, a study based on 100 subjects is assumed to provide a more "precise" estimate than a study based on 10 subjects. Therefore, larger studies should carry more "weight" in the analyses than smaller studies. This sample size based approach is a simple one. A better approach is to assign weight by the inverse variance ($w_i = \frac{1}{s_i^2}$). Thus, the meta-analysis mean is $m_w = \left.\sum_i w_i y_i \middle/ \sum_i w_i\right.$ and variance $var(m_w) = \left. 1 \middle/ \sqrt{\sum_i w_i} \right.$.

Ideally, population stratification should be sufficient for removing heterogeneity, and we should be able to construct our meta-analysis model with the fixed-effect model only. This should hold true in many cases as long as k (number of studies) and the within study sample sizes are large. In practice, k and sample size are usually not very large, so the estimate of variance may be too small, because when calculating variance $var(m_w)$, we ignore the variability in the weights, assuming weights to be fixed and known. However, in some cases, there may be latent factors contributing to the variance between studies. Thus, heterogeneity cannot be removed through further stratification. Several alternative approaches have been proposed to address this issue, including the weighted least squares approach (62) and the robust variance estimator (63). Both estimators perform better than

the original model, with the robust variance estimator a little better overall (63, 64). The robust variance estimator estimates the variance as $\text{var}(m_w) = \frac{\sum w_i(y_i - m_w)^2}{\sum w_i(k-1)}$.

Another more commonly used way is to adopt a random-effect model. Random-effect models assume that the treatment effect varies between studies. The observed measurement y_i, from the i-th study is made up of two additive components: the true measurement, θ_i, and the sampling error, e_i. That is, $y_i = \theta_i + e_i$ for i=1, …, k. The variance of e_i, can be estimated by s_i^2. Additionally, inter-study variance has to be considered in the formula. The first and most widely adopted random-effects models for meta-analysis was proposed by DerSimonian and Laird in 1986 (65). This method is now considered the "standard approach" for meta-analysis in medical and clinical research. In their model, they also adopted inverse variance weight. The total variance, however, is the sum of within-study variance (s_i^2) and inter-study variance (τ^2), leading to weight as $w_i = \frac{1}{\tau^2 + s_i^2}$. τ^2 which can be obtained using three methods: maximum likelihood (ML), restricted maximum likelihood (REML), and method-of-moments (MOM). The author compared the three methods and concluded that ML was biased downward, and that there was little difference between REML and MOM (66). They used MOM because it is non-iterative and easy to implement (ML and REML requires iteration to estimate τ^2). A general method-of-moments estimate t^2 for τ^2 is:

$$
t^2 = \frac{[\sum_i a_i (y_i - y_w)^2] - \left[\sum_i a_i s_i^2 - \frac{\sum_i a_i^2 s_i^2}{\sum_i a_i}\right]}{\left[\sum_i a_i - \frac{\sum_i a_i^2}{\sum_i a_i}\right]}
$$

where $a_1, \ldots, a_k$ are positive constants reflecting weights assigned to the k studies. Each set of $a_1, \ldots, a_k$ values will yield an alternative estimate for τ^2. In DerSimonian and Laird's (65) non-iterative estimate, $t^2(DL)$, $a_i = \frac{1}{s_i^2}$.

Further studies have compared DerSimonian and Laird's method with several more computer intensive methods (67-69). These estimators include Cochran's (70) analysis-of-variance (ANOVA) non-iterative estimate, $t^2(CA)$, where $a_i = \frac{1}{k}$; Paule and Mandel's (71) iterative estimate, $t^2(PM)$, where $a_i = \frac{1}{t^2(PM) + s_i^2}$; two-step estimate starting with Cochran's estimate of τ^2, $t^2(CA2)$, where $a_i = \frac{1}{t^2(CA2) + s_i^2}$; and two-step estimate starting with the DerSimonian and Laird estimate of τ^2, $t^2(DL2)$, where $a_i = \frac{1}{t^2(DL2) + s_i^2}$. They found that DerSimonian and Laird's method is sufficient in most cases. Actually several more sophisticated estimators for both iterative and non-iterative methods can be considered as special cases of a general DerSimonian and Laird's method with a slightly different formula to calculate weights assigned to the individual studies (67). To achieve both efficiency and performance, in this study, we used the original non-iterative DerSimonian and Laird estimator for τ^2.

2. 7   Meta-Analysis Workflow and Example

The previous sections described the analysis methods we used to determine parameters in the meta-analysis workflow. A decision tree of the workflow is shown in **Figure 21**.
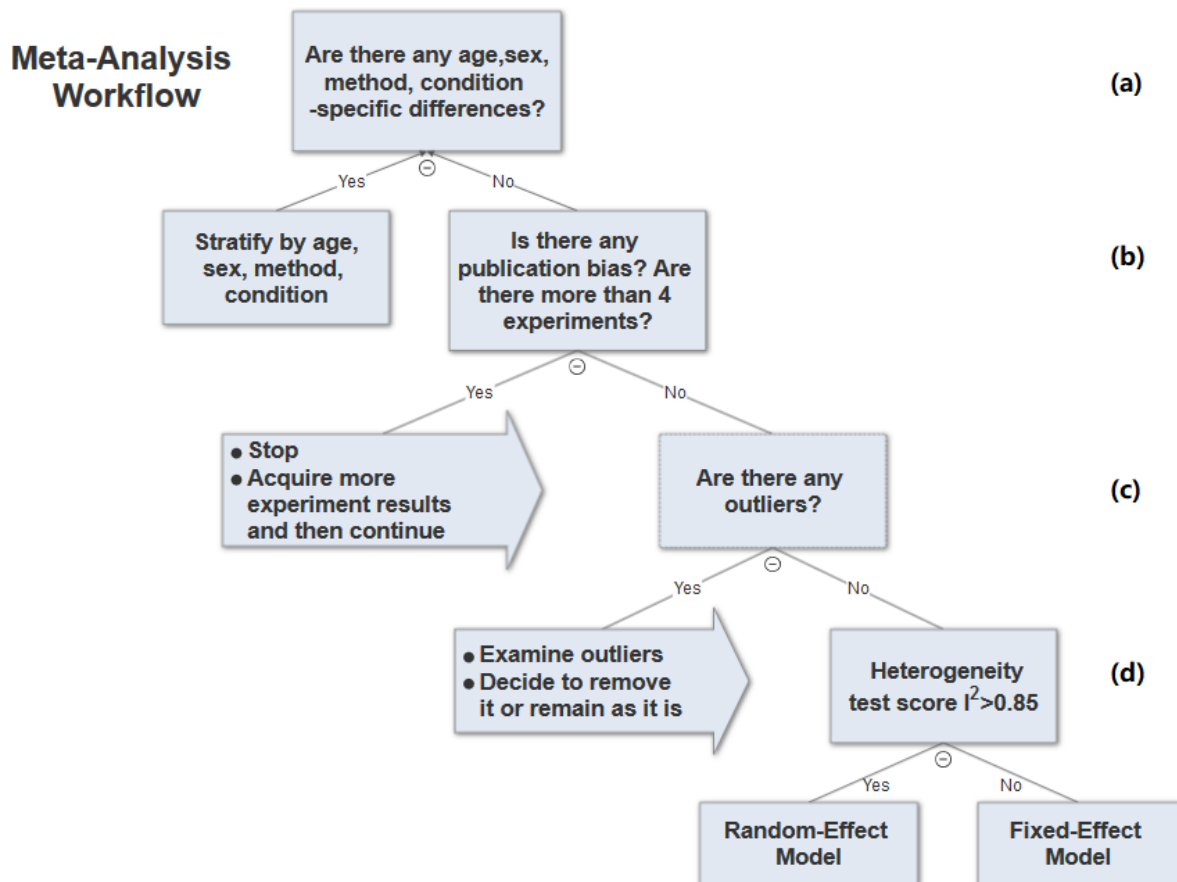
Figure 21 Meta-Analysis Workflow

In summary, our workflow is composed of the following key steps:

1. Perform exploratory analysis (**Figure 21 (b)**) to examine publication bias and total

    number of experiments. For examination of publication bias, we used the original Egger

    test considering the trade-off between power and type I error rates. |Asymmetry

    score| > 1. 5 or total number of experiments <4 is a sign of a potentially biased sample.

    Thus, conclusions from the meta-analysis might not be trustworthy. We will need to

    acquire more data in order to proceed with the analysis. In **Figure 22**, we showed an

    example of publication bias. The |Asymmetry score| is > 1. 5 from Egger Galbraith plot.

    In the Egger funnel plot, there is a clear bias towards publication of heavier rats.

[BN] Total:37; Asymmetry_socre:-1.8498; Asymmetry_slope:0.0071

Funnel Plot for Body Weight (N vs value, Macaskill Method)

Funnel Plot for Body Weight (1/SD vs value, Egger Method)

Galbraith Plot for Body Weight (1/SD vs value/SD, Egger Method)

Figure 22 Publication Bias Exist for BN Body Weight Measurement

2. Perform exploratory analysis (**Figure 21 (c)**) to determine inclusion/exclusion of individual study/observation in the meta-analysis. An example is provided in **Figure 19**.

3. Examination of heterogeneity (**Figure 21 (d)**) using Cochrane's Q

$(Q = \sum (w \times ES^2) - \dfrac{[\sum (w \times ES)]^2}{\sum w})$, a statistic based on the chi-squared test and the I^2

statistic ($I^2 = \begin{cases} \frac{Q-(k-1)}{Q} \times 100\% \ for \ Q > (k-1) \\ 0 \ for \ Q \leq (k-1) \end{cases}$). For each set of experiments

qualified for meta-analysis, we calculated Q and I^2, which were used to determine

model selection in the next step.

4. Choosing meta-analysis model (fixed-effect or random-effect). The fixed-effect and

   random-effect model choice threshold is set to I^2=0. 85, which is considered the

   optimal threshold to distinguish heterogeneity caused by a limited number of records

   or true inter-study variance.

For the fixed-effect model, meta-analysis mean is $m_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$ and variance

$var(m_w) = \frac{1}{\sqrt{\sum_i w_i}}$ $(w_i = \frac{1}{s_i^2})$.

For the random-effect model, since results show that DerSimonian and Laird's method

is sufficient in most cases, in this study, we used the original non-iterative DerSimonian

and Laird estimator for τ^2. Thus $w_i = \frac{1}{\tau^2 + s_i^2}$ and $\tau^2 = \frac{[\sum_i a_i(y_i - y_w)^2] - \left[\sum_i a_i s_i^2 - \frac{\sum_i a_i^2 s_i^2}{\sum_i a_i}\right]}{\left[\sum_i a_i - \frac{\sum_i a_i^2}{\sum_i a_i}\right]}$.

Based on the original model estimator, the meta-analysis mean is $m_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$

and variance $var(m_w) = \frac{1}{\sqrt{\sum_i w_i}}$.

An example of random effect model is shown in **Figure 23**. It shows that the I^2=0. 871>0.

85. Using fixed-effect model, the meta-analysis value will be 121. 17. However, using

the random-effect model, the meta-analysis value will be 128. 57.

[BN] I2 statistic: 87.1%, Summary: 121.17(fixed-effect);128.57(meta)
[] ci_start:126.63; ci_end:130.50; min: 113; max: 143; range: 30; number:10
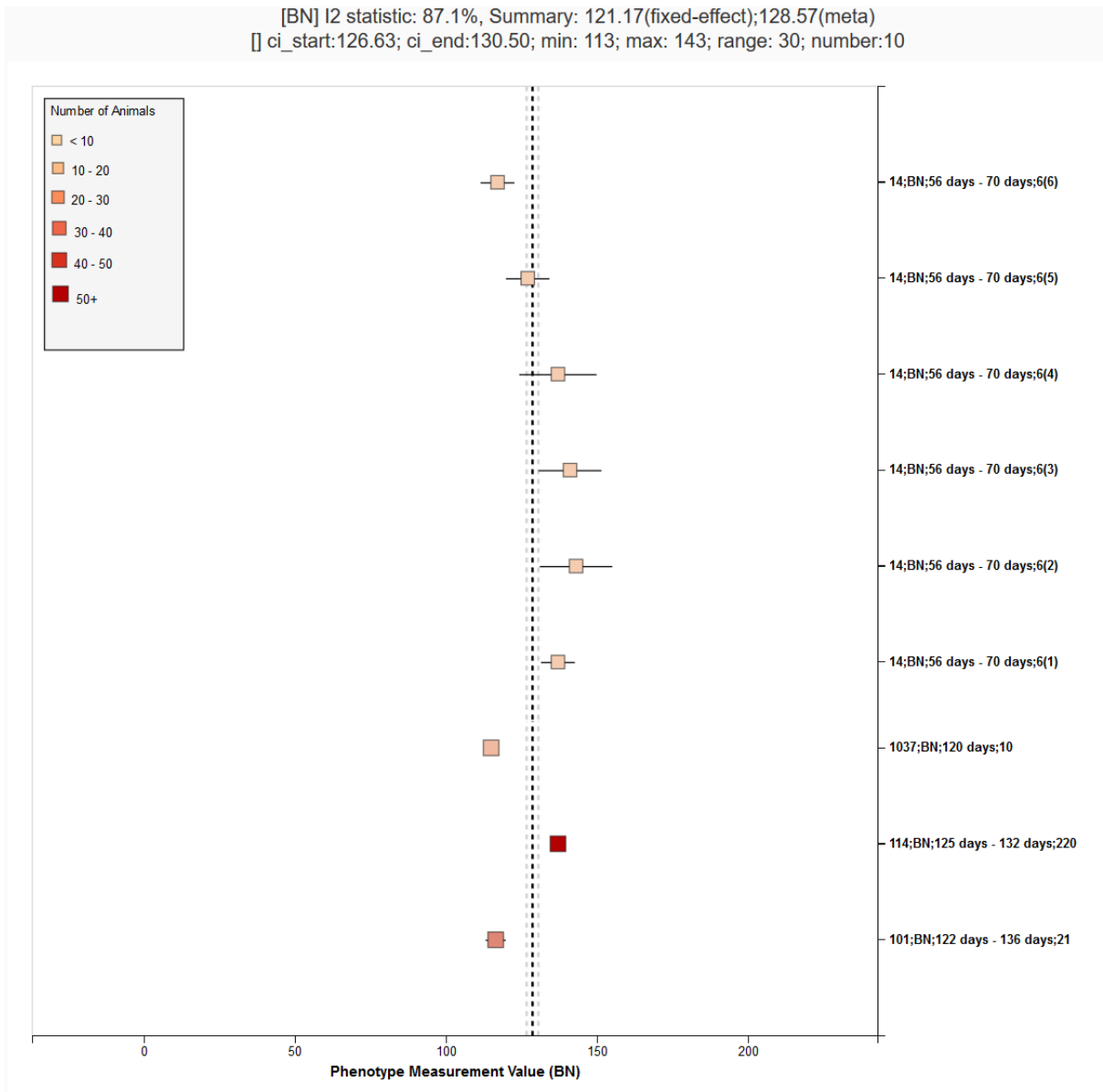
Figure 23 Random-Effect Model for BN "Systolic Blood Pressure"

5. All the summary values for the phenotype under analysis are displayed in a summary forest plot (shown in **Figure 24**). The center of the box represents the meta-analysis mean and the range determined by one standard deviation above and below the meta-analysis value. The color of the boxes showed the total number of experiments that made up of the meta-analysis range. It is a sign of confidence for the resulting phenotype range. On the right side, the legend shows the strain and sex of which the

range is representing. In the bracket, we also noted the confidence level of analysis considering the total number of experiments (<4 low_confidence, >=4 confident).
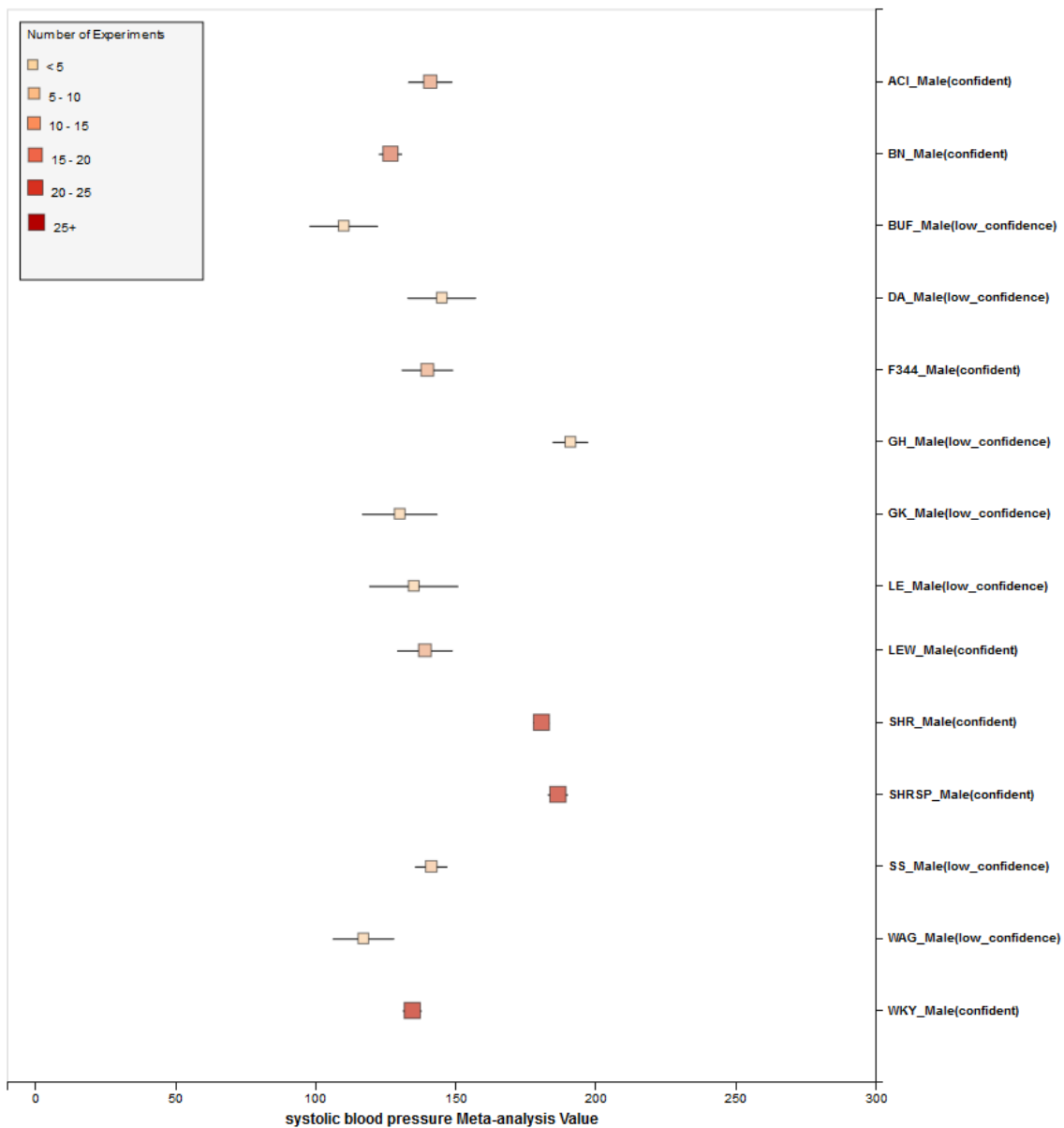
Figure 24 Meta-Analysis Final Summary for "Systolic Blood Pressure" (Male)

2. 8   Applications of Meta-Analysis in the Biomedical Field

Since the publication of DerSimonian and Laird's paper (65), the paper has been cited 19, 145 times according to Google Scholar with more than 50% of those citations occurring in the last few years. Their approach to integrate the findings across related clinical trials has

become increasingly popular in medical research. According to a recent survey done by DerSimonian (66), the top five research topics constituting about half of the total 7, 342 citations during 2010–2014 are in oncology (17%), internal medicine (12%), science & technology (8%), cardiology (7%) and surgery (7%). For example, Ruff *et al.* compared the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation using DerSimonian and Laird's method (72). Beelen *et al.* examined the effects of long-term exposure to air pollution on natural-cause mortality by performing meta-analysis on 22 European cohorts within the multicentre ESCAPE project (73). Recently, meta-analysis was even used to combine multiple different genome-wide association (GWAS) studies in a single integrated analysis in order to identify associations with very small effect sizes (74). The list of meta-analysis applications in the biomedical field is exhaustive showing that applying the meta-analysis method to generate standard phenotype measurement ranges is justified and may provide significant insights from rat phenotype data.

2. 9   Limitations of Meta-Analysis

Meta-analysis is an effective method to integrate data from heterogeneous data sources, with different sample sizes or even different sub-populations/substrains. Such cases include clinical treatment effectiveness test, drug safety evaluation, and effect of diet and exercise on physiological representations, etc. However, it has some limitations:

1. If publication bias is identified, in a preliminary stage, there currently are no statistical methods to overcome this bias and its affects. The existence of publication bias means that any conclusions made might be biased so that meta-analysis results might not reach sufficient levels of confidence. Therefore, researchers who would like to conduct meta-

analysis with a biased dataset need to be aware of and fully disclose the existing publication bias when they make conclusions.

2. Meta-analysis quality and confidence are fundamentally limited by the quality of the underlying studies, so it is necessary for the researchers to evaluate the quality of data by examining the sample size of each experiment, the institutes' credibility, the protocols they used to perform analysis, and other details before meta-analysis can be conducted.

3. Random-effects models tend to increase the variance of the summary measure, making it more difficult to obtain significant results. When the amount of heterogeneity is large, it may even be inappropriate to calculate an overall summary measure of effect size.

4. The results of several meta-analysis studies were later compared to large-scale, well-conducted, randomized controlled trials (so-called 'mega trials'). Some showed good agreement between meta-analysis results and mega trial results while others showed discrepancies (75, 76).

In conclusion, it is extremely important to be diligent at every step while conducting meta-analysis, to examine data patterns to make sure no publication bias exists, to carefully examine the cause of heterogeneity if there is any, and finally to choose a model based on data patterns.

2. 7   Data Source Overview

RGD implemented the PhenoMiner project (29) to integrate quantitative phenotype data from the PhysGen Program for Genomic Application (24), the PhysGen Knockout Project (30) and the Rat Phenome Project (20) under National Bio Resource Project (NBRP) (77) in Japan as well as data from publications or submitted by researchers. This combined data includes

56,710 active entries in the PhenoMiner tool (28). PhenoMiner includes data for 1,218

strains (including substrains), and 678 phenotype parameters (248 under naïve control

conditions) across neurobehavioral, cardiovascular, biochemical, hematological and

morphological categories (78). Entries are grouped by experiment within a study. For some

studies, such as NBRP or PhysGen, numerous experiments make up the study while in

others, such as those from a publication, a study may only be comprised of two or three

experiments. **Figure 25** provides some insight on the wide spread use of some strains, such

as SHR, SS, FHH, as disease models and WKY as a control model. **Figure 26** shows that a

majority of entries in PhenoMiner comes from Study 41 (NBRP Organ Weight). **Figure 27**

further illustrates the large number of quantitative phenotype records available for strains

used as cardiovascular disease. Organ weight records are also numerous in PhenoMiner and

are mostly from Study 41 (NBRP Organ Weight) and Study 21 (NBRP Body Weight) in which

morphological measurements were a primary focus.
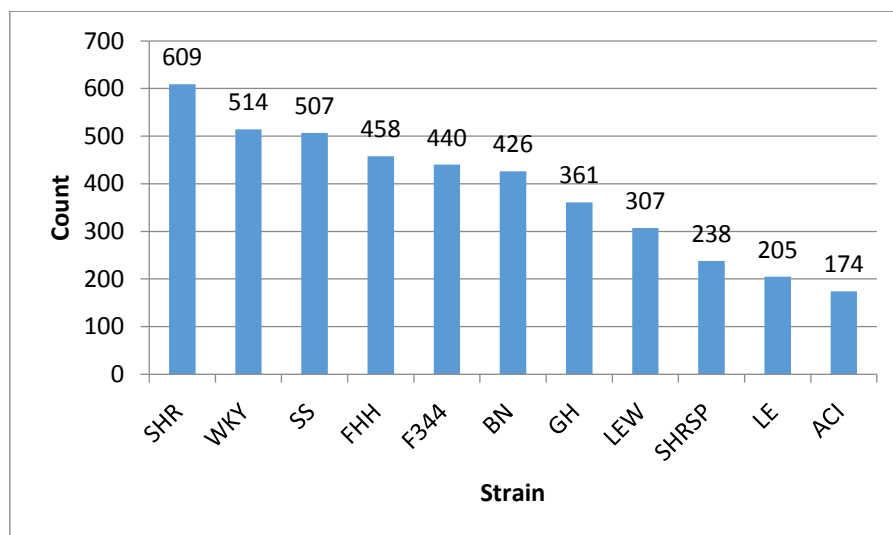
Figure 25 Number of Entries per Strain Group in PhenoMiner Database
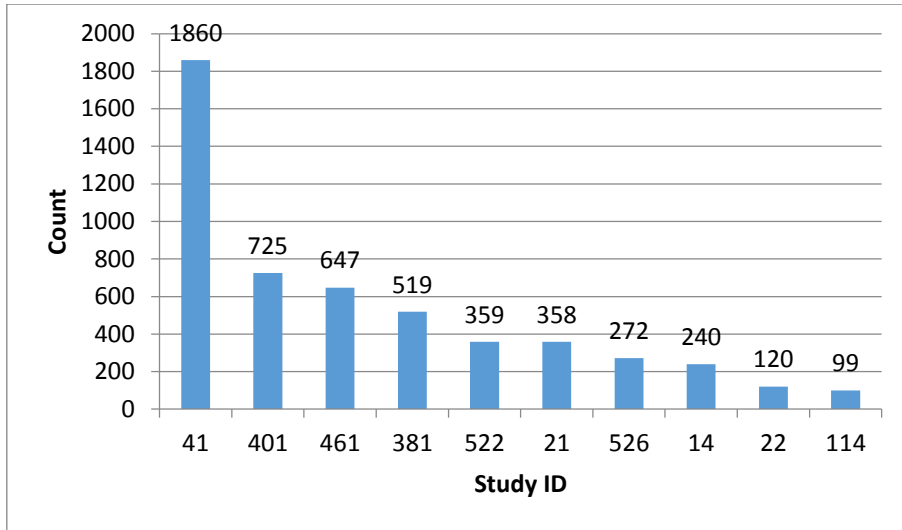
Figure 26 Number of Entries per Study in PhenoMiner Database
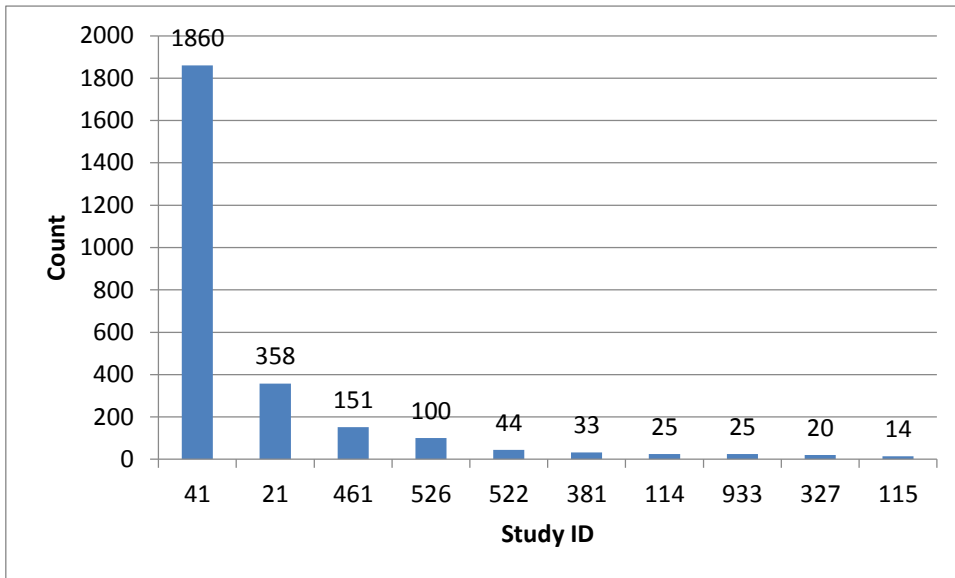
Figure 27 Number of Entries about Organ Weight by Study in PhenoMiner Database

From **Figure 28** we can see that the majority of phenotype records in PhenoMiner are for

organ weights and cardiovascular phenotypes. This is in accordance with the fact that rats

are widely used in cardiovascular research.
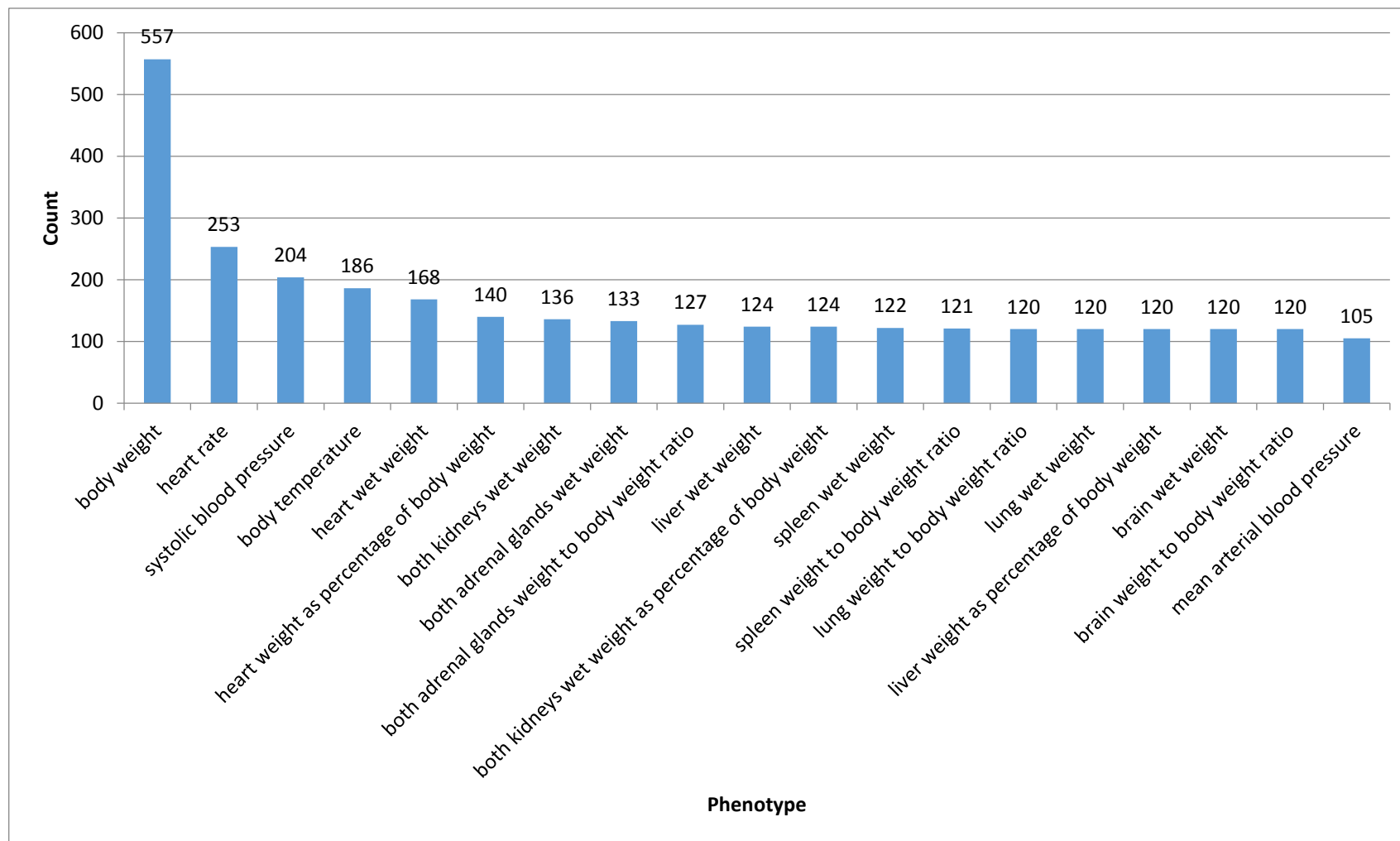
Figure 28 Number of Entries per Phenotype in PhenoMiner Database

Figure 29 shows the number of experiments utilizing particular sample sizes. As might be expected, a sample size of 6 was the most numerous since every experiment in NBRP has used 6 animals. As we mentioned previously, the number of animals in each study range from 1 to 220. Overall, there are more small-scale studies than large ones.
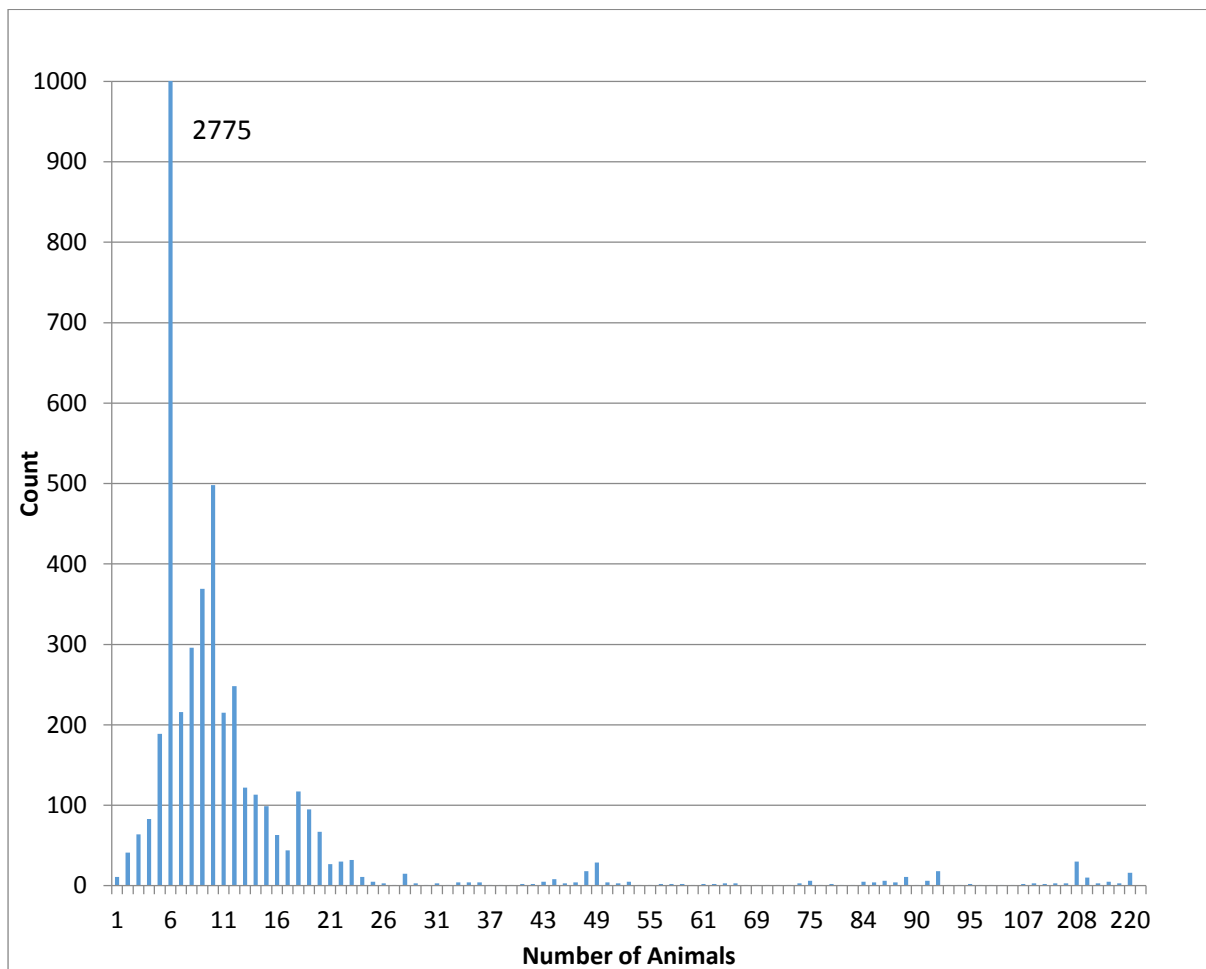
Figure 29 Number of Experiments with Particular Sample Sizes in PhenoMiner

We also separately analyzed sequenced strains and non-sequenced strains (**Table 2** shows a sub-section of our table). From the analysis we found that on average, sequenced strains have more than double the phenotype measurements than non-sequenced strains (**Figure 30**). This is in line with the fact the community promoted particular strains to be sequenced

based on their wide usage as disease models or controls. The availability of the sequence would provide additional resources to link phenotypes to genotypes. The table also provides us with insight on what phenotypes to acquire from research labs in the future to fill in gaps and provide sufficient data for analysis and what strains to sequence given the abundance of phenotype data that already exists. Such efforts would expand the resources available for researchers to investigate the impact of genotype on phenotype. Our analysis shows that the strains ACI, BUF, DA, and SHRSP are lacking some data on blood chemistries related to cardiovascular physiology which could supplement their extensive data for heart and blood pressure characterization including data on serum triglyceride level, plasma total cholesterol level, plasma triglyceride level, and serum free fatty acids level which are important metabolites in cardiovascular physiology (**Table 2**). Compared to systolic blood pressure and mean arterial blood pressure, diastolic blood pressure data were less frequently available for the PhenoMiner project. Such findings could help the PhenoMiner project group to focus on acquiring data from targeted publications or individual researchers in order to fill out comprehensive phenotype profiles for strains. Among non-sequenced strains, we found MWF has a considerable amount of data related to renal function and blood chemistry phenotypes. Findings such as these could also help to focus on acquisition of data from other phenotype areas and to promote strains such as MWF as a candidate worthy of being sequenced.

Table 2 Summary of Cardiovascular Phenotypes for Sequenced Strains

Phenotype	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR	SHRSP	SS
heart rate	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR	SHRSP	SS
systolic blood pressure	ACI	BN	WKY	BUF	DA	F344		GH	LE	LEW	SHR	SHRSP	SS
heart wet weight	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR	SHRSP	SS
heart weight as percentage of body weight	ACI	BN	WKY	BUF	DA	F344	FHH		LE	LEW	SHR	SHRSP	SS
mean arterial blood pressure		BN	WKY			F344	FHH	GH	LE	LEW	SHR	SHRSP	SS
diastolic blood pressure		BN	WKY					GH		LEW	SHR	SHRSP	
heart right ventricle weight to left ventricle weight ratio		BN	WKY			F344	FHH	GH	LE	LEW	SHR		SS
heart weight to body weight ratio	ACI	BN	WKY	BUF	DA	F344					SHR	SHRSP	
heart left ventricle wet weight		BN	WKY		DA	F344		GH		LEW	SHR		
heart left ventricle weight to body weight ratio			WKY		DA	F344				LEW	SHR		
heart right ventricle wet weight			WKY			F344							
hematocrit	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR	SHRSP	SS
serum total cholesterol level		BN	WKY		DA	F344	FHH	GH	LE	LEW	SHR	SHRSP	SS
blood hemoglobin level	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR		SS
red blood cell count	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR		SS
mean corpuscular volume	ACI	BN	WKY	BUF	DA	F344	FHH	GH	LE	LEW	SHR		SS
serum calcium level		BN	WKY			F344	FHH	GH	LE	LEW	SHR		SS
serum aspartate aminotransferase activity level		BN	WKY			F344	FHH	GH	LE	LEW	SHR		SS
serum potassium level		BN	WKY			F344	FHH	GH	LE	LEW	SHR		SS
serum chloride level		BN	WKY			F344	FHH	GH	LE	LEW	SHR		SS
serum triglyceride level		BN			DA	F344					SHR		
plasma total cholesterol level										LEW			
plasma triglyceride level										LEW			
serum free fatty acids level						F344							
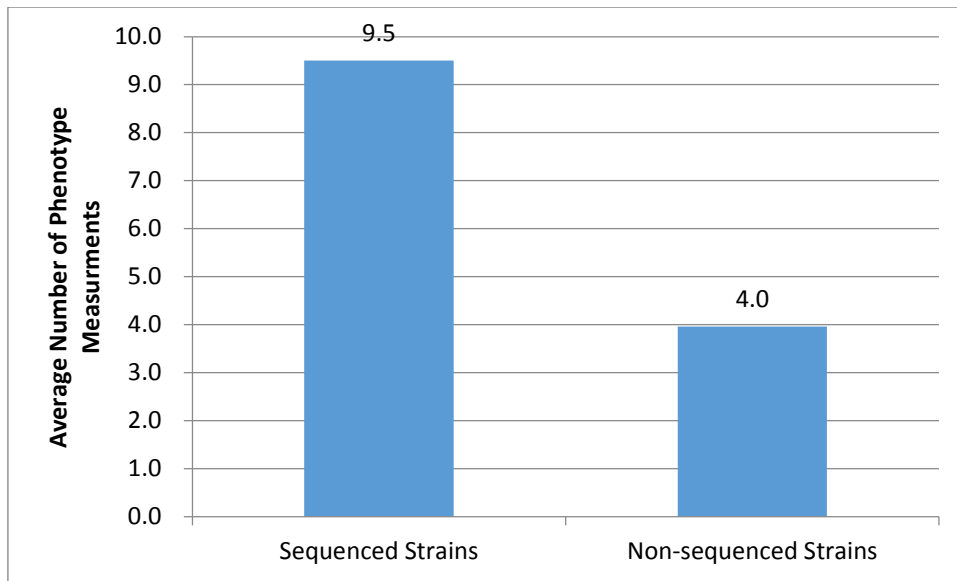
Figure 30 Average Number of Phenotype Measurements for Sequenced Strains and Non-Sequenced Strains

To evaluate potential heterogeneity in phenotype measurements, we also investigated the methods used to measure each phenotype. We found 191 phenotypes with only one measurement method and 52 phenotypes with records divided among two or more measurement methods (**Figure 31**). Phenotypes such as systolic blood pressure were measured using seven different methods (shown in **Figure 32**). As PhenoMiner developed, it was evident that measurement method had an impact on values collected, something that was commonly known among researchers but had not been quantified. We found that measurement methods used also correlated with ages of rats in experiments at times. For example, it is hard to measure younger rats with the intra-vascular method for blood pressure; hence younger rats are usually measured using tail cuff methods. Because of the impact of measurement method on the value of the measurement, it is one of the parameters used to group and stratify records within phenotype areas.
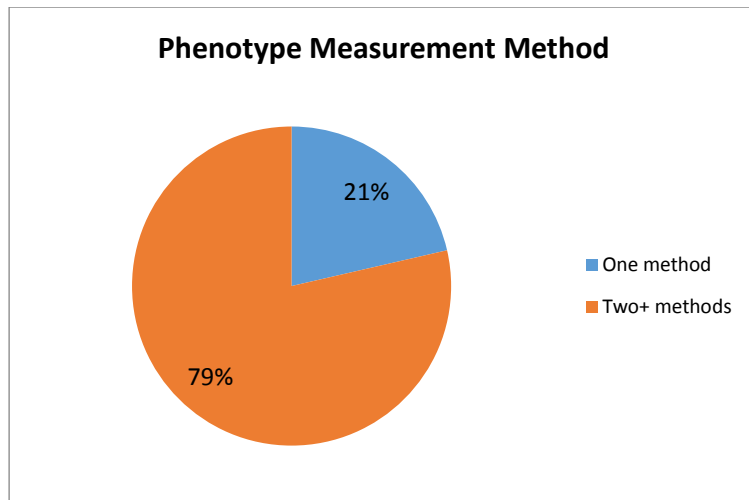
Figure 31 Number of Phenotypes Measured by One or More Than One Method
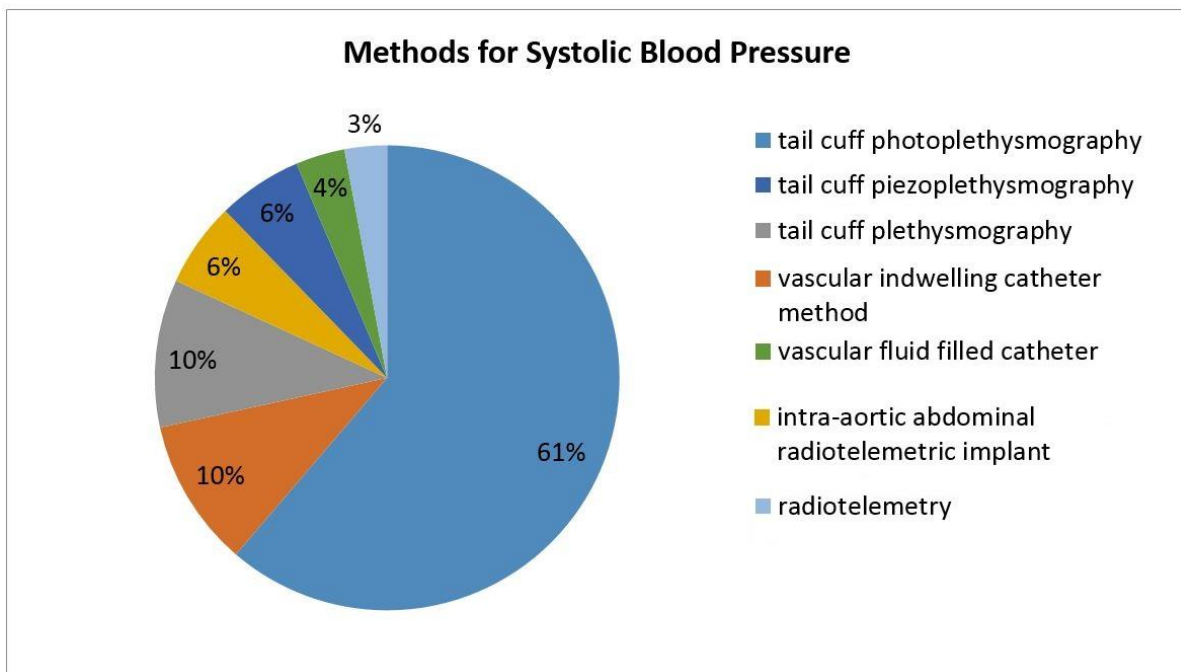
Figure 32 Percentage of Phenotype Records for Systolic Blood Pressure by Measurement Type

2. 8   Meta-Analysis Result and Evaluation

Our meta-analysis analysis provided expected ranges for 24 cardiovascular related

phenotypes. We analyzed all the available sequenced strains for each phenotype. For non-

sequenced strains, we focused on MWF since most non-sequenced strains had limited data

available for these phenotype areas.

In the first step, our meta-analysis used only experiment records under control conditions for inbred stains. In this way, the meta-analysis range can be regarded as an "Expected Range" for a specific strain or strain group. One type of strain group was created by grouping all substrains under a certain parent strain name according to the Rat Strain Ontology. In accordance with official standards, a strain is given a substrain designation when it is bred for 20 generations or more in a different facility or laboratory. For example, strain group "ACI" include all substrains "ACI/Eur", "ACI/Kun", "ACI/N", "ACI/SegHsd", "ACI/Ti", "ACI/Ztm". Control strain groups were also created based on their widespread use as control animals and acceptance as exhibiting "normal" measurements. For example, the "Normal Systolic Blood Pressure Strain Group" consists of strains that are considered commonly used as controls in blood pressure experiments and to exhibit "normal systolic blood pressure". The "Normal Systolic Blood Pressure Strain Group" was created in an iterative process using a domain expert with extensive experience in large scale phenotyping projects: 1. Strains commonly used as controls were identified based on experience and prior knowledge and designated as "founder control strains" e. g. BN and WKY which have long been used as control models. 2. An initial "Expected Range" was constructed based on phenotype ranges of those "founder" strains using the highest and lowest values of the previously determined expected ranges for each strain. 3. The overlap of previously determined expected ranges of other strains with this initial "Control Expected Range" was examined to determine whether additional strains could be included in the normal phenotype strain group. 4. An updated "Control Expected Range" was constructed using all strains added to the normal phenotype strain group.

In addition to constructing a general "Expected Range", we stratified our analysis by age, sex, and measurement method when data was available for different ages, sexes, or

methods. We then constructed age, sex, method-specific "Expected Ranges" using the same workflow. The results of constructing phenotype "Expected Ranges" are discussed in the first section below.

After the initial "Expected Ranges" for 24 phenotypes using inbred strains were constructed, we also performed meta-analysis for outbred and mutant strains, when sufficient data existed. We also evaluated the applicability of the analysis under non-control conditions using inbred strains in which a measured salt diet was the experimental condition. Creating such expected ranges for particular experimental conditions will further assist researchers in choosing model strains for particular experiments and provide data for developing tools and statistical processes that would allow them to analyze their own data.

1) Phenotype Expected Ranges

Phenotype data was available for inbred, outbred, consomic, congenic, mutant and transgenic strains. However, for this study, initial development of the algorithms and workflows and expected ranges of phenotypes were established using only inbred strains. In addition, phenotype records in which the experimental conditions naïve control, were used and those involving experimental diets, exercise, application of drugs or chemicals or other manipulated conditions were not initially used in this study. **Figure 33** shows an example of a forest plot summary produced for "systolic blood pressure" for different age groups. We were then able to identify strains with expected ranges within or have overlap with previously constructed "Control Expected Range" (ACI, BN, BUF, DA, F344, GK, LE, LEW, LN, M520, MNS, MR, MWF, and WKY), and strains with expected ranges outside of the constructed "Control Expected Range"(GH, LH, MHS, SHR, and SHRSP), strains for naïve control (BN, WKY) from the graph.
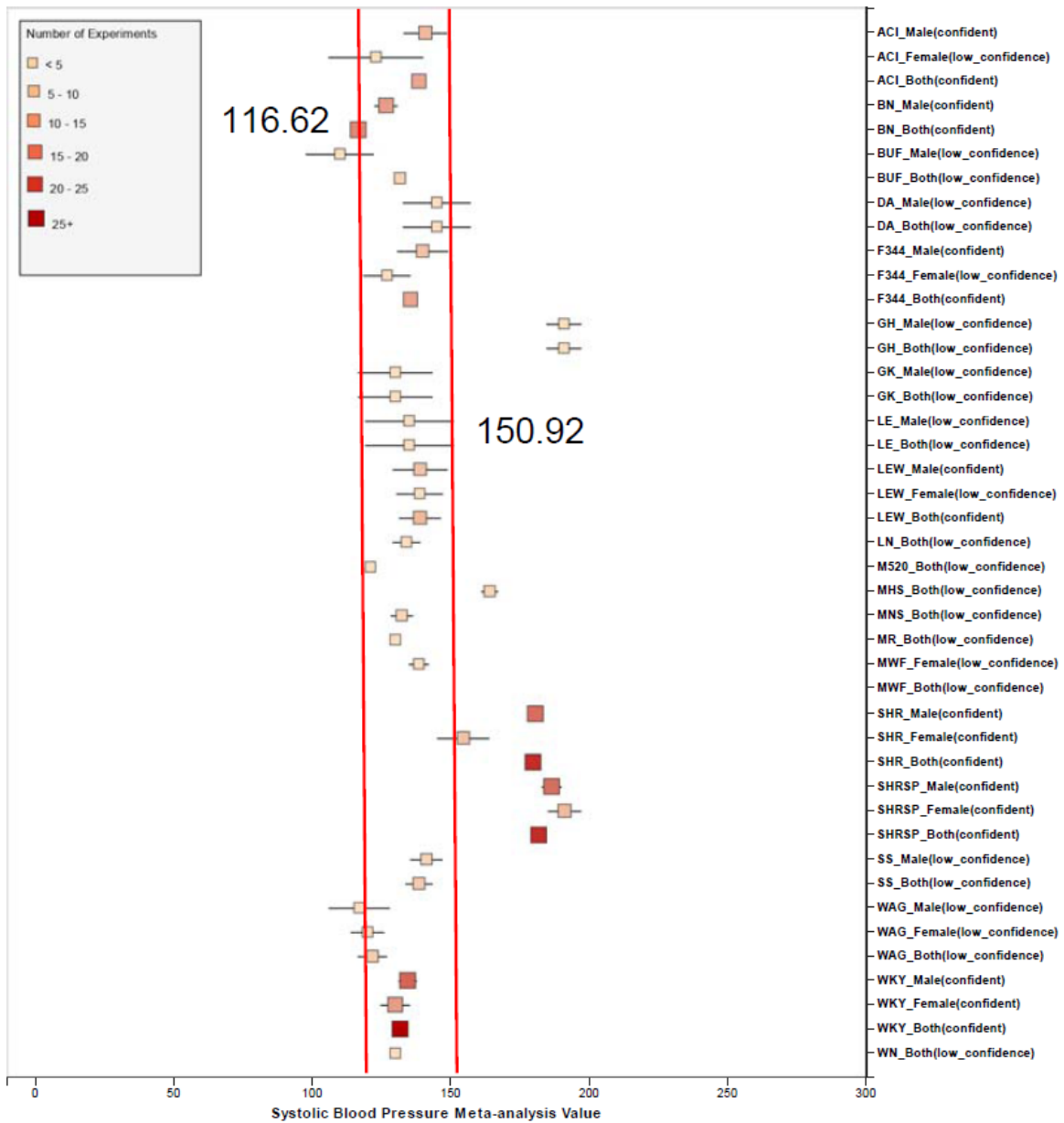
Figure 33 Forest Plot of Meta-Analysis Summary for "Systolic Blood Pressure"

APPENDIX A includes all the meta-analysis summaries for 24 cardiovascular phenotypes

(blood hemoglobin level, diastolic blood pressure, heart left ventricle weight to body weight

ratio, heart left ventricle wet weight, heart rate, heart right ventricle weight to left ventricle

weight ratio, heart right ventricle wet weight, heart weight as percentage of body weight,

heart weight to body weight ratio, heart wet weight, hematocrit, mean arterial blood

pressure, mean corpuscular volume, plasma total cholesterol level, plasma triglyceride level,

red blood cell count, serum aspartate aminotransferase activity level, serum calcium level, serum chloride level, serum free fatty acids level, serum potassium level, serum total cholesterol level, serum triglyceride level, systolic blood pressure). Information about abnormal, normal, naïve control strains for 24 phenotypes is available in **APPENDIX B**.

We also created separate summaries for different age groups (age1: <70 days, age2: 70-99 days, age3:100+ days) and gender groups if significant age or gender difference was observed in the all age analysis. For example, in **Figure 33**, systolic blood pressure between male and female differs from each other and shows an obvious pattern (usually females has lower blood pressure). After analysis, we decide that we should produce a separate summary for both gender and for each age group as well. Same situation applies to phenotypes such as diastolic blood pressure, heart weight as percentage of body weight, heart weight to body weight ratio, heart rate, heart wet weight, and mean arterial blood pressure. First, those phenotypes have enough data for age and gender stratification while still produces meaningful meta-analysis results. Phenotypes related to ventricle weight and blood metabolite measurement suffer from lack of data for stratification (heart right ventricle weight to left ventricle weight ratio, heart right ventricle wet weight, plasma total cholesterol level, plasma triglyceride level, serum aspartate aminotransferase activity level, serum calcium level, serum chloride level, serum free fatty acids level, serum potassium level only have data for one age group).

Results from the meta-analysis were reviewed by a domain expert who previously had classified strains potentially within and outside of predicted reference ranges of "normal" and outside of "normal" based on wide spread use and characterization as control and non-control strains. Disagreements between the meta-analysis results and the domain expert's

classification were marked out on the graph with a red rectangle. The percentage of cases in which there was consistent agreement were calculated at 98% (**APPENDIX C**).

We also found that for some strains considered to exhibit "normal" phenotypes, although the meta-analysis mean value was within the overall expected range, the upper bound or the lower bound of its individual range was beyond the overall expected range for "normal", perhaps indicating 1) more data is needed from more institutes for more confident conclusions about individual expected ranges, or 2) variability for those strain groups may be due to potential genetic drift so that the genotypes of substrains have become more diverse or certain strains could be more susceptible to outside influences such as housing, handlers and other environmental factors on phenotypes. A good example would be ACI_Female, BUF_Male and WAG_Female/Male/Both in **Figure 33**. We also found that the obvious pattern between male and female (usually females has lower blood pressure) sometimes didn't hold true. For example, in **Figure 34**, SHRSP the sex specific pattern is opposite of the common pattern where female rats have lower blood pressure than male rats. On the other hand, SHR exhibits the common pattern.
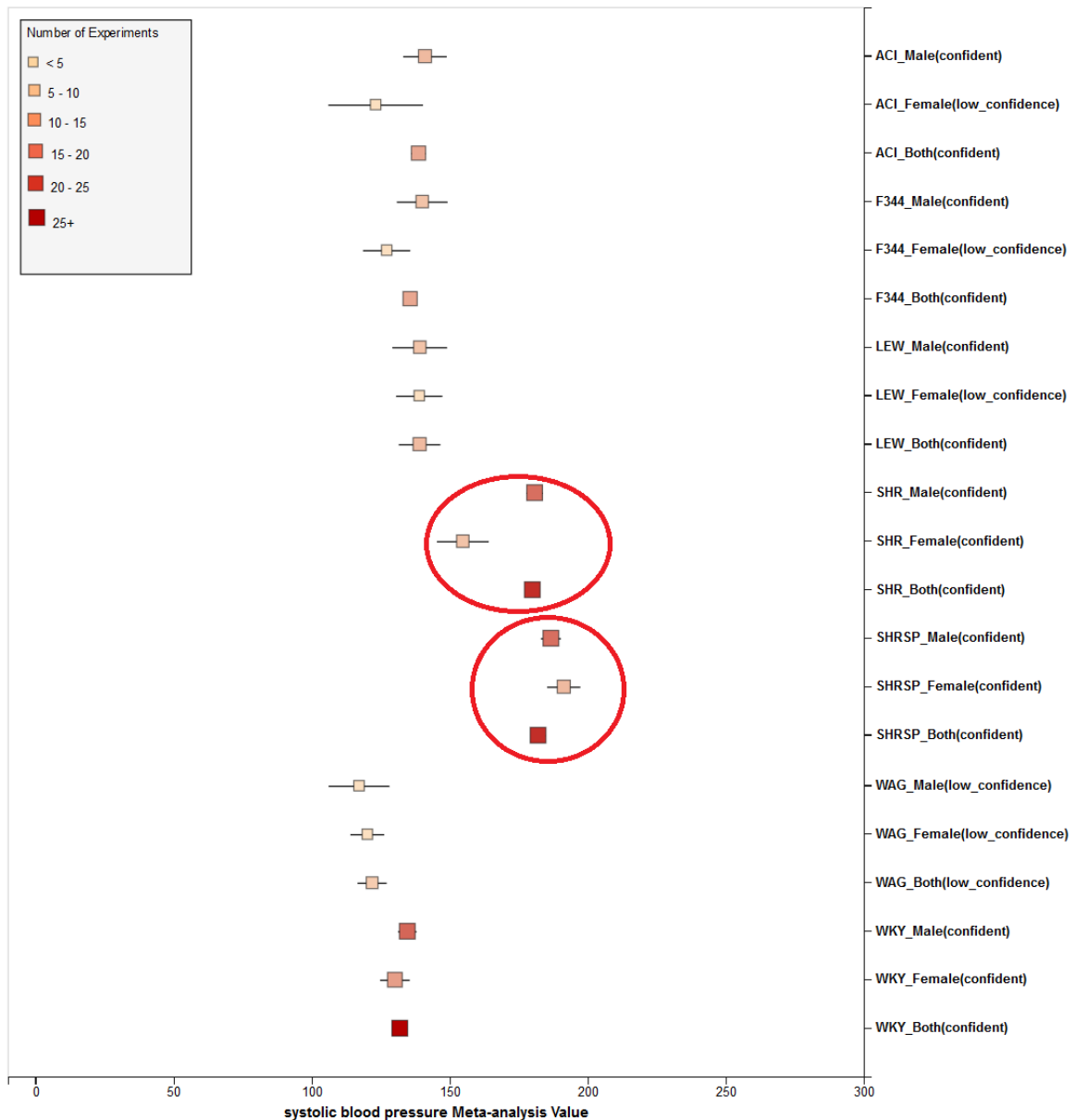
Figure 34 Summary for "Systolic Blood Pressure" Shows Sex Pattern Does Not Always Hold True

Summary data which stretched beyond the overall expected ranges or exhibited odd patterns generally came from meta-analyses with a limited number of experiments (usually <5). This is evidence of low confidence for the meta-analysis result, which indicates more experimental data are needed to establish a trustworthy range. This is also evidence that for meta-analysis, the number of studies included is vital to eliminate random experimental error and generate trustworthy results. The number of studies in general was more

important than the number of animals in each study. Our analysis method demonstrated its potential to be used: 1) to provide expected ranges for rat phenotypes; 2) to facilitate research planning by visualizing current gaps and suggesting potential research directions to fill in the gaps.

2) Rat Strain Model Comparison

Initially, expected ranges for individual phenotypes were established for inbred strains under the naïve control condition. We then applied the analysis process calculate expected ranges for phenotype measurements for outbred, mutant, congenic and transgenic strains. Next, we calculated expected ranges for inbred, outbred, mutant, congenic and transgenic strains on measurements acquired under different diet conditions. As we developed the analysis pipeline, we created a user-friendly interface, so researchers could compare expected ranges for different rat models under both control and salt diet conditions. (inbred, outbred, mutant, congenic and transgenic; naïve control vs salt diet).

Outbred and Mutant Strains

Figure 35 shows expected ranges for "systolic blood pressure" under naïve control conditions for outbred and mutant rat strains. The dotted lines represent the expected range for the "Control Strain Group" based on 14 inbred strains used as controls for this phenotype. This presentation clearly indicates which strains have expected ranges within the "normal" control range and which do not.
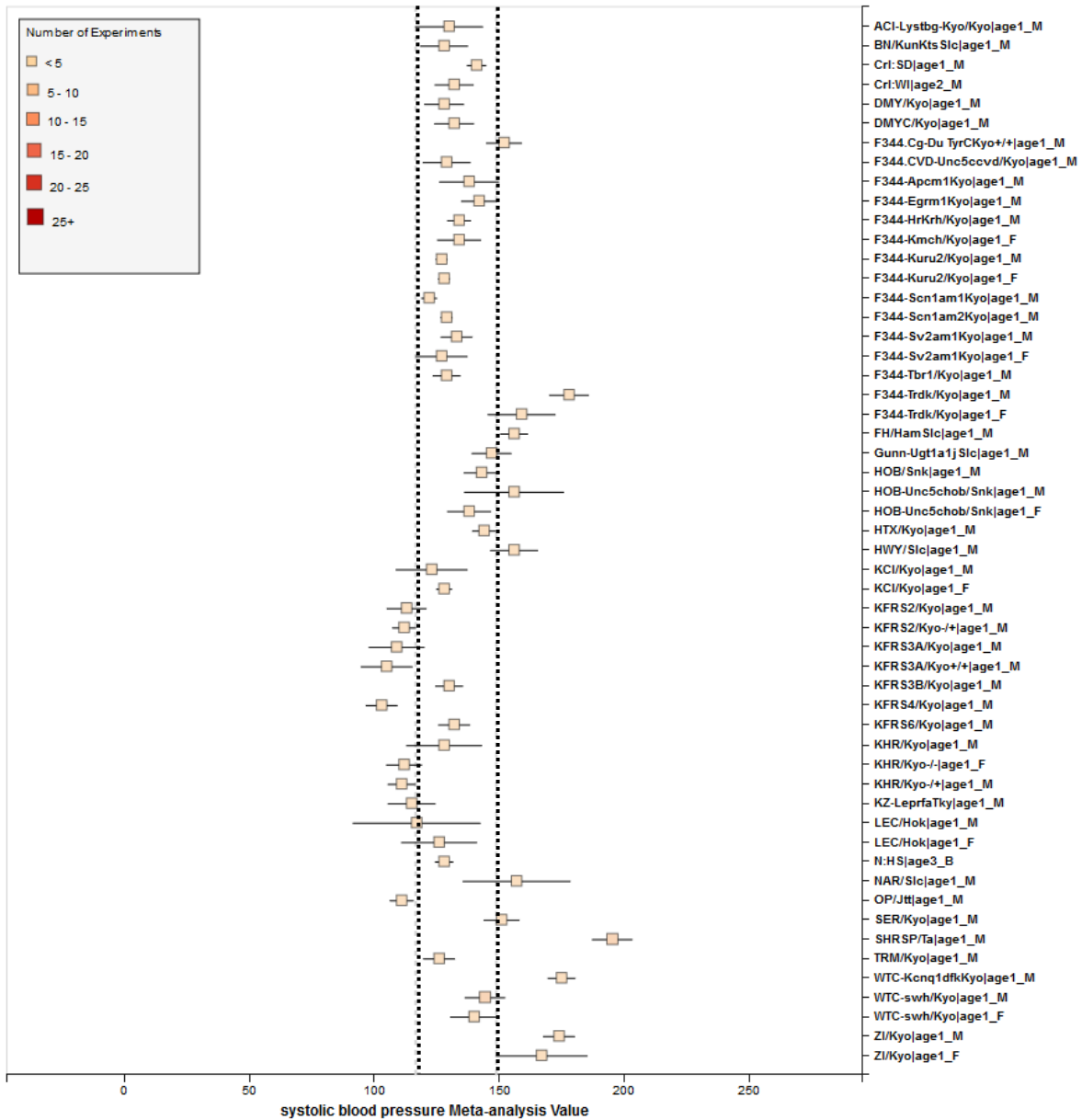
Figure 35 RGD Phenotype Expected Ranges for "Systolic Blood Pressure" for Outbred and Mutant Strains

2. 9   Tools to Mine and Present Expected Ranges

While the determination of expected ranges for individual phenotypes within specific strains

and across strains is a valuable resource, data mining and visualization tools increase the

value of this resource for investigators. Modifications to existing database structures and

the development of query and visualization tools make this possible.

1) Database Structure

The existing PhenoMiner database structure (**Figure 36**) was modified to incorporate the expected range values for each strain group. The tables with alternating shaded cells are newly added TABLEs. EXPECTED_RANGE TABLE was created to store data for phenotype expected ranges. EXPECTED_RANGE_EXPERIMENT_REC TABLE links the experiment records that make up the phenotype expected ranges to the expected ranges. STRAIN_GROUP TABLE was created to link the strains that make up the phenotype expected ranges which may be several substrains or substrains and parent strain, or that the designated control strains to create the "normal" or control expected range for that phenotype.
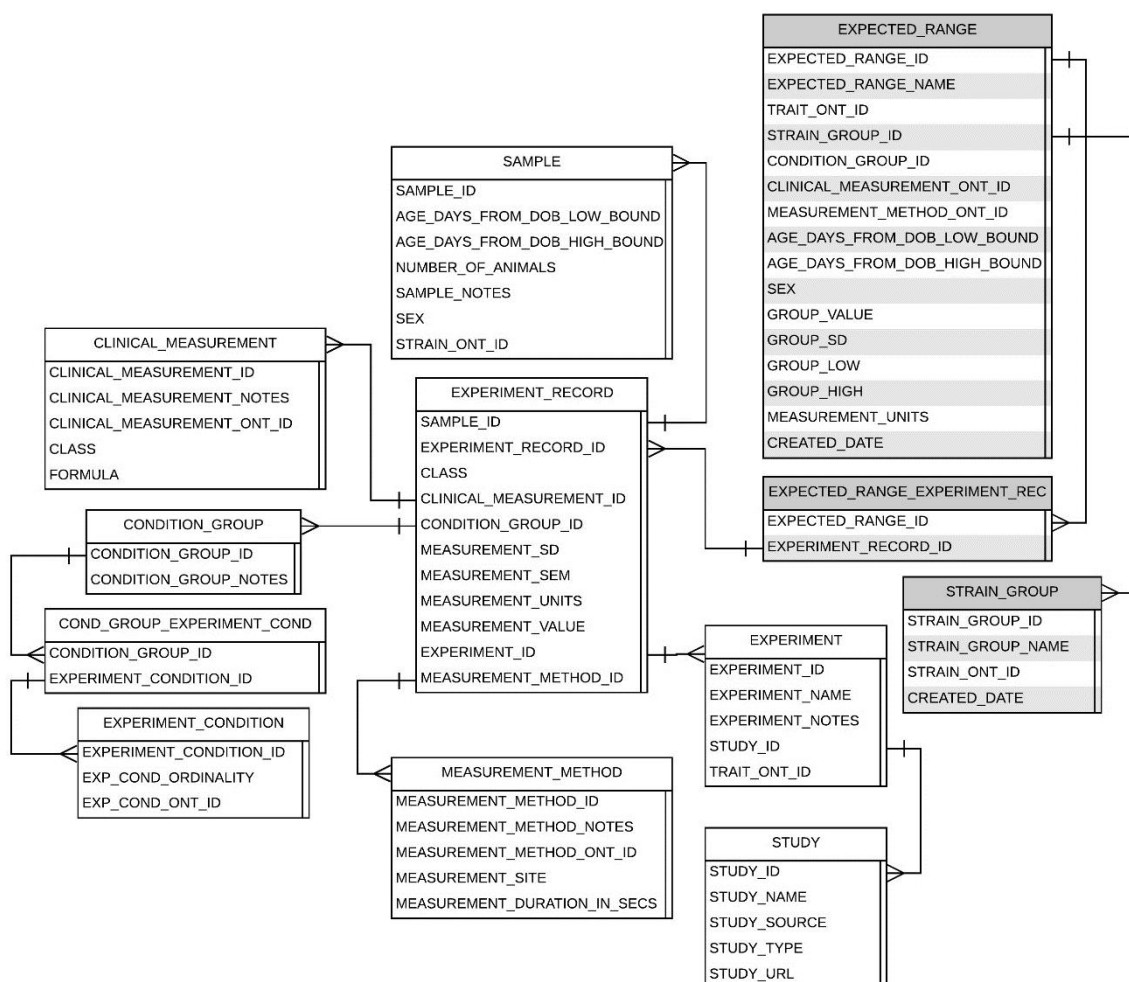

Figure 36 Data Structure for Storing Expected Range Data

2) Home Page and Query Tool

To highlight the results of the meta-analysis and provide tools to mine and display data, a component for the Rat Genome Database has been created. The query begins with a search by trait area in a pulldown menu at the top right corner of the home page (**Figure 37**). This results in a listing of the quantitative phenotype measurements commonly used to assess aspects of that trait area. In the example provided, circulatory system trait, the phenotypes presented include systolic, diastolic and mean arterial blood pressure, various weight measurements for the heart and heart rate. For each phenotype, the normal range that was calculated from control strains is presented in a table. The table also includes 1) number of strains with available expected ranges; 2) number of strains with sex-specific expected ranges (at least two expected ranges for mixed, female, or male, so that expected ranges can be compared between sex groups); 3) number of strains with age-specific expected ranges (at least two expected ranges for 0-79 days, 80-99 days, or 100+ days, so that expected ranges can be compared between age groups).

Figure 37 Summary of Number of Strains with Expected Ranges (Circulatory System Trait)

The user can click on the phenotype term or the number of strains value to retrieve the associated data. For example, clicking on the term Diastolic Blood pressure or Circle 1 retrieves all expected ranges of available strains for diastolic blood pressure; clicking on Circle 2 retrieves expected ranges for the 7 strains that for which there are separate ranges for each sex for diastolic blood pressure; and clicking on Circle 3 Retrieves the data for the 9 strains for which there are separate expected ranges for multiple age groups.

We also provided a listing of number of expected ranges of (circulatory system) phenotypes available for each strain (**Figure 38**).

Figure 38 Summary of Number of Phenotypes with Expected Ranges (Circulatory System Trait)

3) Data Summaries and Visualization

Retrieved data are presented in an interface that includes three major components: a selection panel, a visualization panel and a data panel (**Figure 39**).

The selection panel built with a bootstrap JavaScript library (https://getbootstrap. com/) provides the user with the information on the strains, measurement methods, experimental conditions, sex and age groups associated with the returned data, This allows the user to further customize the returned data by narrowing the results to smaller strain sets, specific measurement methods, ages or sex by clicking off the options not desired. As an initial selection is made, options for further filtering are automatically updated to present those available based on the initial selection. As **Figure 40** shows, when the user chooses Age Group 0-79 Days, the panel and visualization change to show that data is only available for the SHR, SHRSP and WKY strain groups. **Figure 41** presents the visualization when the user limits the results to males only in the first step and **Figure 42** when the user selects specific strains in the first step.

diastolic blood pressure - Expected Ranges

Select Phenotype:
diastolic blood pressure

Options

Strains
BN/Elh
BN/OrlIcoCrlf
GH/Omr
HTG
LEW
LH/Mav
LH/MavRrrc
LN/Mav
LN/MavRrrc
MHS/Gib

☑ Strain_Group:
- ☑ BN   ☑ LH   ☑ SHR
- ☑ GH   ☑ LN   ☑ SHRSP
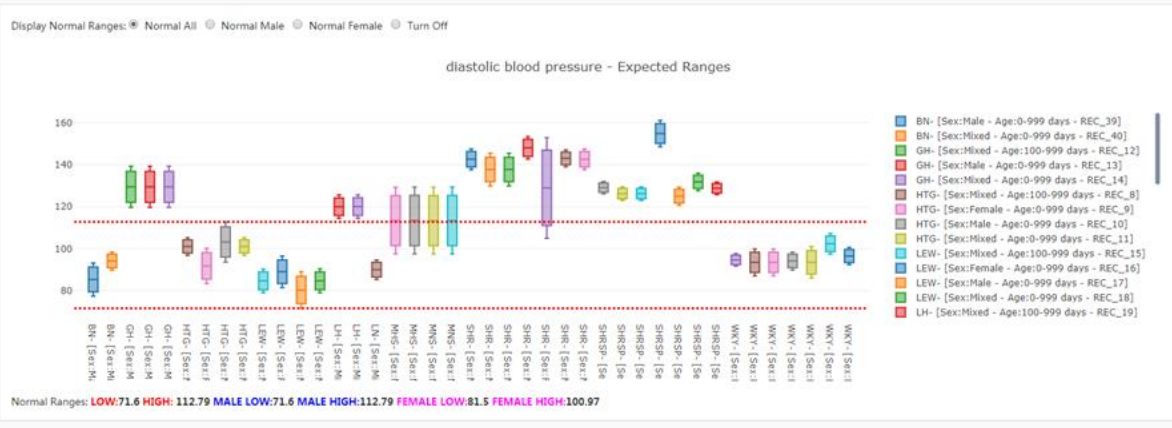- ☑ HTG ☑ MHS ☑ WKY
- ☑ LEW ☑ MNS

☑ Conditions
- ☑ control condition
- ☑ naive control condition

☑ Methods
- ☑ intra-aortic abdominal radiotelemetry
- ☑ radiotelemetry
- ☑ vascular fluid filled catheter
- ☑ vascular indwelling catheter method

☑ Age
- ☑ 0 - 79 days
- ☑ 80 - 99 days
- ☑ 100 - 999 days
- ☑ 0 - 999 days

☑ Sex
- ☑ Male
- ☑ Mixed
- ☑ Female

Submit

Display Normal Ranges: ◉ Normal All ○ Normal Male ○ Normal Female ○ Turn Off

diastolic blood pressure - Expected Ranges

Legend:
- BN- [Sex:Male - Age:0-999 days - REC_39]
- BN- [Sex:Mixed - Age:0-999 days - REC_40]
- GH- [Sex:Mixed - Age:100-999 days - REC_12]
- GH- [Sex:Male - Age:0-999 days - REC_13]
- GH- [Sex:Mixed - Age:0-999 days - REC_14]
- HTG- [Sex:Mixed - Age:100-999 days - REC_8]
- HTG- [Sex:Female - Age:0-999 days - REC_9]
- HTG- [Sex:Male - Age:0-999 days - REC_10]
- HTG- [Sex:Mixed - Age:0-999 days - REC_11]
- LEW- [Sex:Mixed - Age:100-999 days - REC_15]
- LEW- [Sex:Female - Age:0-999 days - REC_16]
- LEW- [Sex:Male - Age:0-999 days - REC_17]
- LEW- [Sex:Mixed - Age:0-999 days - REC_18]
- LH- [Sex:Mixed - Age:100-999 days - REC_19]

Normal Ranges: LOW:71.6 HIGH: 112.79 MALE LOW:71.6 MALE HIGH:112.79 FEMALE LOW:81.5 FEMALE HIGH:100.97

REC_ID	Strain Group	Measurement	Method	Sex	Age	Range Mean	Range SD	Range Low	Range High
39	BN	diastolic blood pressure		Male	0 - 999 days	85.3	7.86	77.44	93.16
40	BN	diastolic blood pressure		Mixed	0 - 999 days	94.16	4.29	89.87	98.44
12	GH	diastolic blood pressure		Mixed	100 - 999 days	129.5	9.82	119.68	139.32
13	GH	diastolic blood pressure		Male	0 - 999 days	129.5	9.82	119.68	139.32
14	GH	diastolic blood pressure		Mixed	0 - 999 days	129.5	9.82	119.68	139.32
8	HTG	diastolic blood pressure		Mixed	100 - 999 days	101.06	4.18	96.88	105.23
9	HTG	diastolic blood pressure		Female	0 - 999 days	91.8	8.4	83.4	100.2
10	HTG	diastolic blood pressure		Male	0 - 999 days	103.19	9.6	93.59	112.79
11	HTG	diastolic blood pressure		Mixed	0 - 999 days	101.06	4.18	96.88	105.23
15	LEW	diastolic blood pressure		Mixed	100 - 999 days	84.67	5.68	78.99	90.35
16	LEW	diastolic blood pressure		Female	0 - 999 days	89.0	7.5	81.5	96.5
17	LEW	diastolic blood pressure		Male	0 - 999 days	80.3	8.7	71.6	89.0
18	LEW	diastolic blood pressure		Mixed	0 - 999 days	84.68	5.69	78.99	90.36
19	LH	diastolic blood pressure		Mixed	100 - 999 days	120.0	5.6	114.4	125.6
20	LH	diastolic blood pressure		Mixed	0 - 999 days	120.05	5.65	114.4	125.7
21	LN	diastolic blood pressure		Mixed	0 - 999 days	90.01	4.66	85.35	94.66
22	MHS	diastolic blood pressure		Mixed	100 - 999 days	113.4	15.9	97.5	129.3
23	MHS	diastolic blood pressure		Mixed	0 - 999 days	113.45	15.95	97.5	129.4
24	MNS	diastolic blood pressure		Mixed	100 - 999 days	113.4	15.9	97.5	129.3
25	MNS	diastolic blood pressure		Mixed	0 - 999 days	113.45	15.95	97.5	129.4
26	SHR	diastolic blood pressure		Mixed	0 - 999 days	142.62	4.91	137.71	147.52
27	SHR	diastolic blood pressure		Mixed	0 - 79 days	137.74	7.77	129.97	145.5
28	SHR	diastolic blood pressure		Mixed	80 - 99 days	137.74	7.77	129.97	145.5
29	SHR	diastolic blood pressure		Mixed	100 - 999 days	148.08	5.38	142.7	153.45
30	SHR	diastolic blood pressure		Female	0 - 999 days	129.0	24.0	105.0	153.0
31	SHR	diastolic blood pressure		Male	0 - 999 days	143.06	4.03	139.03	147.09
32	SHR	diastolic blood pressure		Mixed	0 - 999 days	142.62	4.91	137.71	147.52
1	SHRSP	diastolic blood pressure		Mixed	0 - 999 days	129.11	2.93	126.17	132.04
2	SHRSP	diastolic blood pressure		Mixed	0 - 79 days	126.25	3.25	123.0	129.5
3	SHRSP	diastolic blood pressure		Mixed	80 - 99 days	126.25	3.25	123.0	129.5
4	SHRSP	diastolic blood pressure		Mixed	100 - 999 days	154.88	6.28	148.6	161.16
5	SHRSP	diastolic blood pressure		Female	0 - 999 days	125.01	4.3	120.71	129.31
6	SHRSP	diastolic blood pressure		Male	0 - 999 days	131.85	4.19	127.66	136.04
7	SHRSP	diastolic blood pressure		Mixed	0 - 999 days	128.79	3.0	125.79	131.79

Figure 39 Entire Interface includes a selection panel, a visualization panel and a data panel
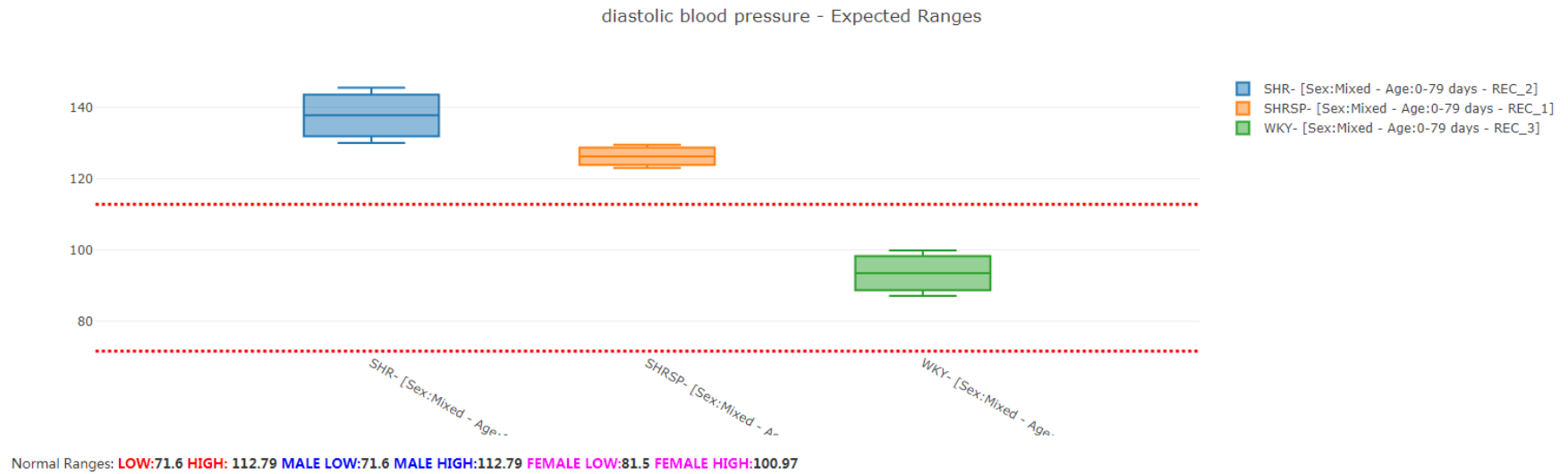
Figure 40 Visualization of All Rat Strains' Expected Range for Diastolic Blood Pressure (with selected age groups)
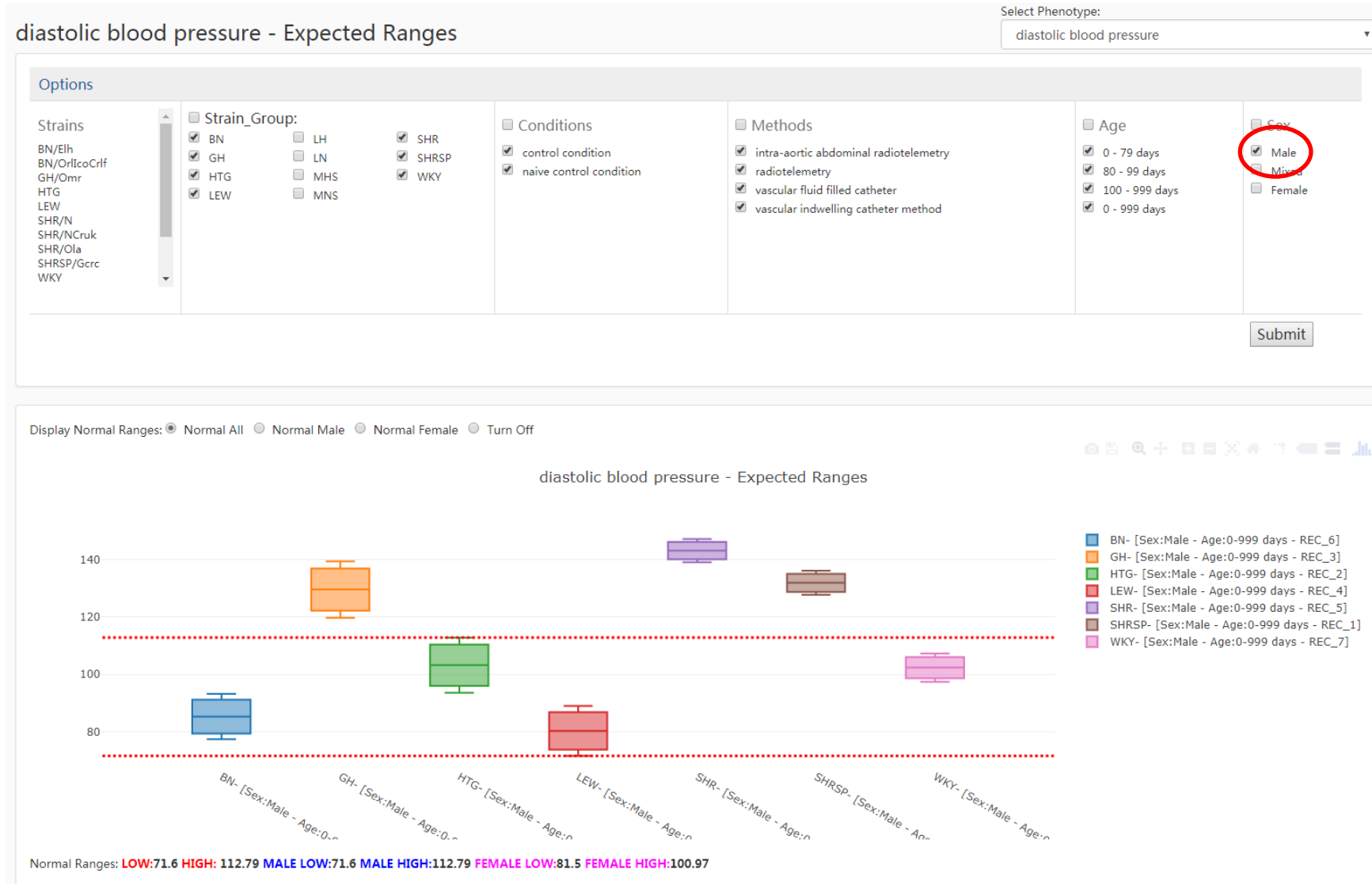
Figure 41 Visualization of All Rat Strains' Expected Range for Diastolic Blood Pressure (with selected sex groups)
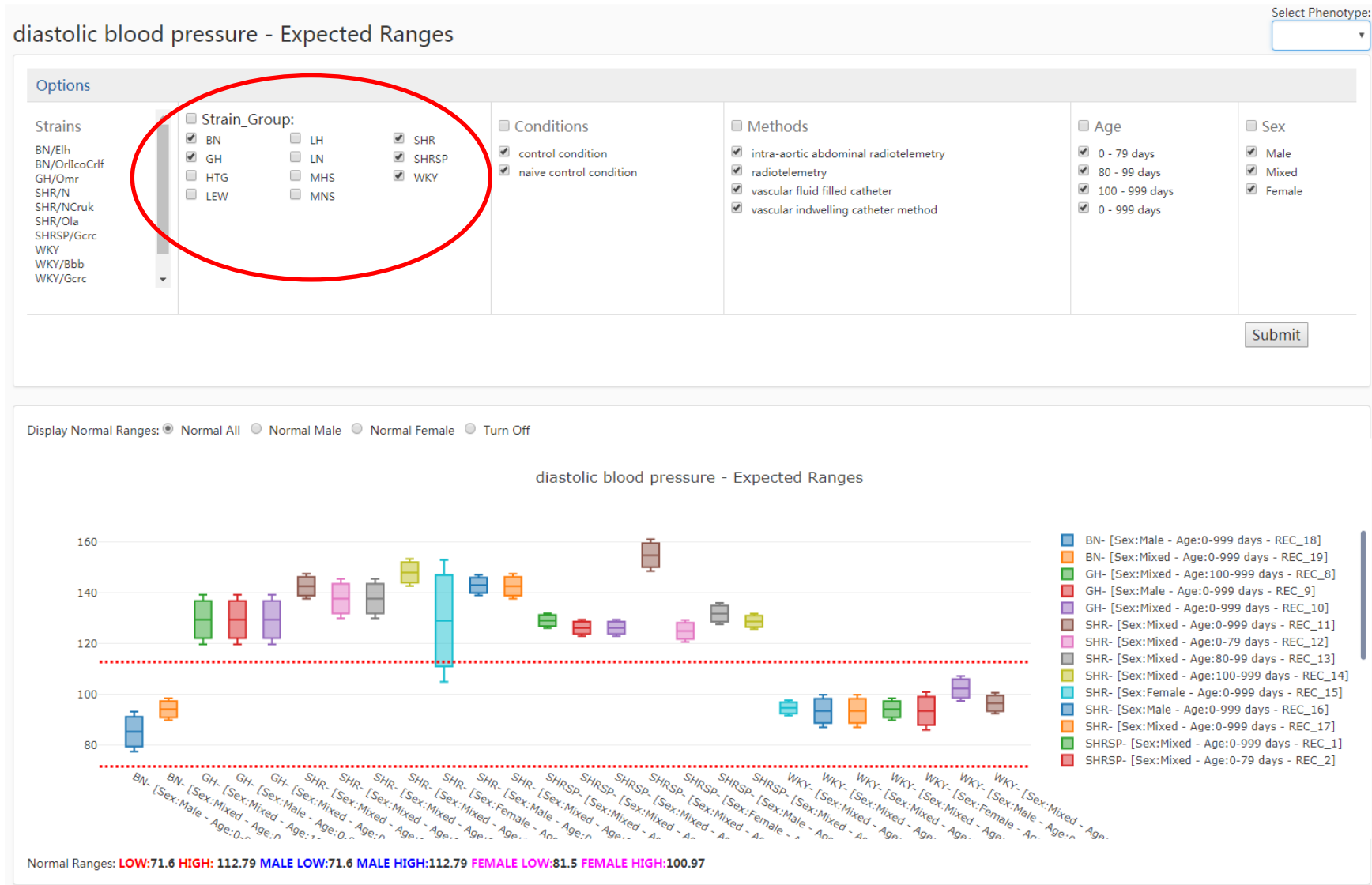
Figure 42 Visualization of All Rat Strains' Expected Range for Diastolic Blood Pressure (with selected strains)

The visualization panel was created using Plotly (https://plot. ly/) and D3js (https://d3js.

org/) libraries. Each box on the chart shows the expected range for a phenotype

measurement based the parameters that constitute a record - strain group, sex, age,

method and experimental condition. The red dotted line provides a reference showing the

expected range that is considered a normal range for control groups. The user can choose to

show the overall reference expected range or that for either sex. Currently only ranges

under naive control condition are available in the tool with those already determined for

salt diets to be added. This tool easily illustrates for researchers which strains have large or

small expected ranges, which expected ranges for commonly used disease models actually

overlap expected ranges for control groups. As shown in **Figure 39** and **Figure 43** it can

easily be seen that MHS, MNS and SHR_Females have very large expected ranges that

overlap to an extent with normal ranges while the SHRSP strain exhibits narrow ranges and

all SHRSP expected ranges are above the normal range. Such information is useful when

investigators are designing research projects.

The legend on the right side is also clickable. By clicking on individual boxes on the right-side

legend list, one can choose to view or hide specific expected range.

Furthermore, users can choose to compare expected ranges for different strains against a

selected reference. We provide reference expected ranges for female rats only, male rats

only, rats with age 0-79 days, rats with age 80-99 days, as well as rats with age 100-999 days.

Age-specific references are not available for all phenotypes. The reference values can be

viewed at the bottom of the graph.

Figure 43 Visualization of All Rat Strains' Expected Range for Diastolic Blood Pressure (showing reference of Female Rats)

Below the box plot graph is the data panel where we included a table with all the details (strains, measurement methods, sex, age, range mean, range SD, range low and range high) of the visualized information. Users can look up the details of a specific data entry (box) by searching its record ID in the table.

3   Rat Sub-Population Identification with Multiple Phenotype Profiles

In addition to providing access and tools to the determined expected ranges, additional analyses can be conducted on these results to further identify disease models and groups of strains appropriate for disease studies.

As noted rats are commonly used to study complex diseases such as cardiovascular, endocrine and neurological diseases. Most of the disease phenotypes related to polygenic (complex) diseases are continuous and are only considered abnormal when the measurement value for the phenotype exceeds or is below a certain threshold. Moreover, a number of phenotypes are affected in complex diseases, and the phenotypes that represent a complex disease are usually correlated. However, those correlated phenotypes related to a specific disease might have unique representation patterns in individuals. Heterogeneous representations of phenotypes may indicate the existence of disease subtypes with different biological mechanisms and may require different treatment plans. In addition, such differences could indicate the influence of environmental factors on the expression of the phenotype patterns.

Kim *et al*. showed that the 53 clinical phenotypes in the asthma dataset from the Severe Asthma Research Program (SARP) (86) can be represented as a network with several densely connected subnetworks (87). They leveraged a graph (**Figure 44**) of multiple asthma-related
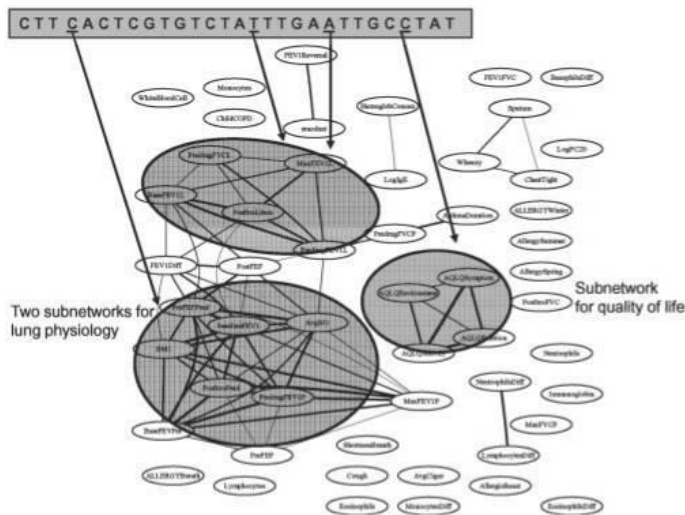
Figure 44 Illustration of Association Analysis using Phenotype Correlation Graph for Asthma Dataset (87)

quantitative traits as prior knowledge. From the assumption that the densely connected subgraphs over these correlated traits are more likely to be influenced by the same or heavily overlapping subset(s) of SNPs, regression analysis can be done to examine relationship between the quantitative traits in the subnetwork and causal SNPs subsets.

Rats are frequently used as cardiovascular disease models. Identifying complex disease sub-populations would further assist investigators in choosing appropriate models for their research. For example, subtypes of hypertension such as isolated systolic hypertension (ISH), isolated diastolic hypertension (IDH), and combined systo-diastolic hypertension (SDH) may represent different mechanisms and may indicate different risks for future cardiovascular events such as stroke and coronary heart disease (88). Furthermore, differences in organ damage and co-morbidities exist among different hypertension subtypes including hypertension accompanied by renal dysfunction, obesity, sleep apnea, or autonomic failure (89).

Very limited research in the identification of disease sub-populations has been done in model organism domain while there has been increasing interest in the clinical domain in patient sub-population identification and patient similarity-based predictive modeling for diagnosis and prognosis (90). Currently this interest is also extending to organisms used in pre-clinical research, such as the rat. Developing well-defined comprehensive phenotype

profiles for specific rat strains will enhance the ability of researchers to choose appropriate models that mimic the phenotypic profiles of humans with a particular disease. Currently clustering-based algorithms are the most commonly used methods to identify disease sub-populations. Cluster-based algorithms also use distance measurement to retrieve a group of similar subjects. Hence new subjects can be assigned to a predefined cluster based on their similarity/distance to each cluster. Various types of similarity metrics can be used to describe similarity between subjects, most of them are distance-based. Distances are measured to quantify similarity between subjects to retrieve a sub-population of similar subjects and find the closest class to a new subject. Distance measures include: Absolute distances (the absolute distances between 2 and 5 is |2-5|=3), Euclidean distance ("ordinary" straight-line distance between two points in Euclidean space; D_E $= \|\mathbf{q} - \mathbf{p}\| = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p} \cdot \mathbf{q}}$), Mahalanobis distance (a measure of the distance between a point P and a distribution D; $D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$, S is the covariance matrix). We employed a clustering-based method to calculate rat strain similarity with Euclidian distance to identify subpopulations of rat strains based on their phenotype profiles as determined by expected ranges. Using strain, experimental condition and multiple quantitative phenotypes as features, our cluster-based approach identified sub-populations with different profiles in a general disease category. The resulting model represents various rat population subtypes.

The algorithm was implemented using Weka (91). Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Here we used a standalone Weka java application for data pre-processing, clustering, and visualization.

3.1 Material and Method

1) Data Source

The data used for clustering included the meta-analysis expected range values for five cardiovascular related phenotypes: mean arterial blood pressure, diastolic blood pressure, systolic blood pressure, heart rate, and heart wet weight.

2) Model Construction

Hierarchical clustering was conducted using a method with Euclidian distance as the similarity measurement method. The algorithm was implemented using Weka (91).

3) Model Selection

We used subsets of data with different numbers of phenotype features to construct our clustering model. We then compared the "Distinguish Power" of each model to select the best clustering model. "Distinguish Power" is defined as the ratio of number of clusters to wrong cluster assignments (or small clusters with only one member). The Gold standard of cluster membership was provided by RGD curators. The number of cluster k was also selected based on the "Distinguish Power" metric. Small clusters were merged into big clusters so that k was considered the optimal number of clusters. The results of the models and respective scores are listed in **Table 3** (items marked with * represent wrong cluster assignment). Five phenotypes were numbered as mean arterial blood pressure (1), diastolic blood pressure (2), systolic blood pressure (3), heart rate (4), and heart wet weight (5).

Table 3 Clustering Results and Distinguish Power for Different Models

Model	Cluster	Distinguish Power
1+2	WKY_Both, WKY_Female, WKY_Male, BN_Both, BN_Male, LEW_Both*, LEW_Female*, LEW_Male*	3/4
	SHRSP_Both, SHRSP_Male, SHRSP_Female, SHR_Female*	
	SHR_Both, SHR_Male, GH_Both, GH_Male	

2+3	WKY_Both, WKY_Male, WKY_Female, MNS_Both, LN_Both, LEW_Both*, LEW_Female*, LEW_Male*, BN_Both, BN_Male	2/5
	SHRSP_Both, SHRSP_Female, SHRSP_Male, GH_Both, GH_Male, SHR_Both, SHR_Male	
	SHR_Female*, MHS_Both*	
1+2+3	WKY_Both, WKY_Female, WKY_Male, LEW_Both*, LEW_Female*, LEW_Male*, BN_Both, BN_Male	2/4
	SHRSP_Both, SHRSP_Male, SHRSP_Female, SHR_Both, SHR_Male, GH_Both, GH_Male	
	SHR_Female*	
1+4	WKY_Both, WKY_Female, WKY_Male, BN_Both, BN_Female, BN_Male, F344_Both, F344_Female, F344_Male, FHH_Both, FHH_Female, LE_Both, LE_Female, LE_Male, LEW_Both*, LEW_Male*, SS_Both, SS_Female, SS_Male, SHR_Female, SHRSP_Both, SHRSP_Male, SHRSP_Female, SHR_Both, SHR_Male, GH_Both, GH_Female, GH_Male	2/4
	FHH_Male*	
	LEW_Female*	
1+4+5	WKY_Both, WKY_Female, WKY_Male, BN_Both, BN_Female, BN_Male, F344_Both, F344_Female, F344_Male, LE_Both, LE_Male, LE_Female, LEW_Both*, LEW_Male*, FHH_Both, FHH_Female, FHH_Male, SS_Both, SS_Female, SS_Male, SHRSP_Both, SHRSP_Female, SHRSP_Male, SHR_Both*, SHR_Female*	3/4
	SHR_Male, GH_Both, GH_Female, GH_Male	
	LEW_Female	
1+2+3+ 4+5	WKY_Both, WKY_Male, WKY_Female, BN_Both, BN_Male	3/1(optimal) or 4/2
	SHR_Female*	
	SHRSP_Both, SHRSP_Male, SHRSP_Female, SHR_Both, SHR_Male, GH_Both, GH_Male	
	LEW_Both, LEW_Male, LEW_Female	
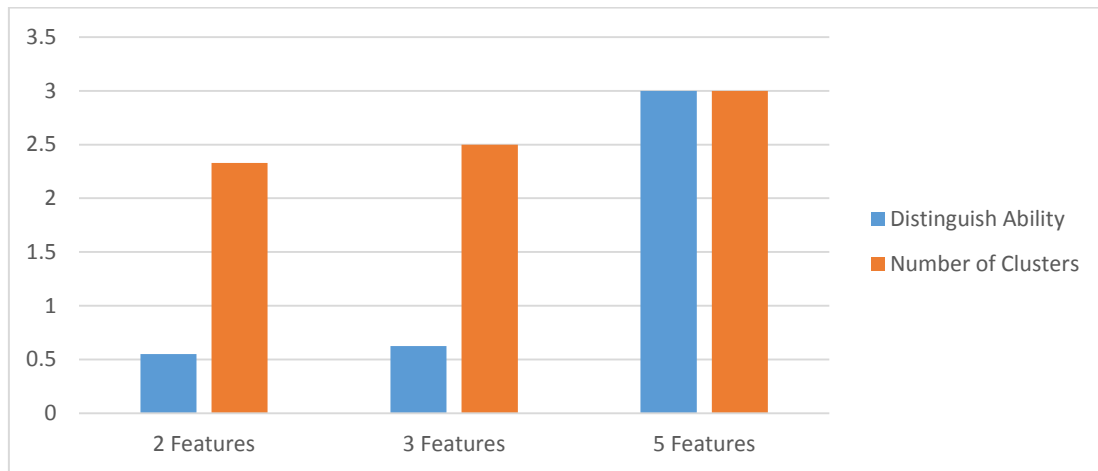
Figure 45 Distinguish Power and Number of Clusters for Different Models

3.2   Results and Conclusion

The Model (1+2+3+4+5) proved to be the best model and the optimal number of clusters k

was 3 after merging "SHR_Female" into cluster "SHRSP_Both, SHRSP_Male, SHRSP_Female,

SHR_Both, SHR_Male, GH_Both, GH_Male". K=3 was optimal because if we kept the

"SHR_Female" cluster and moved "SHR_Both, SHR_Male" into the small cluster, there were

2 wrong assignments and the "Distinguish Power" was reduced.
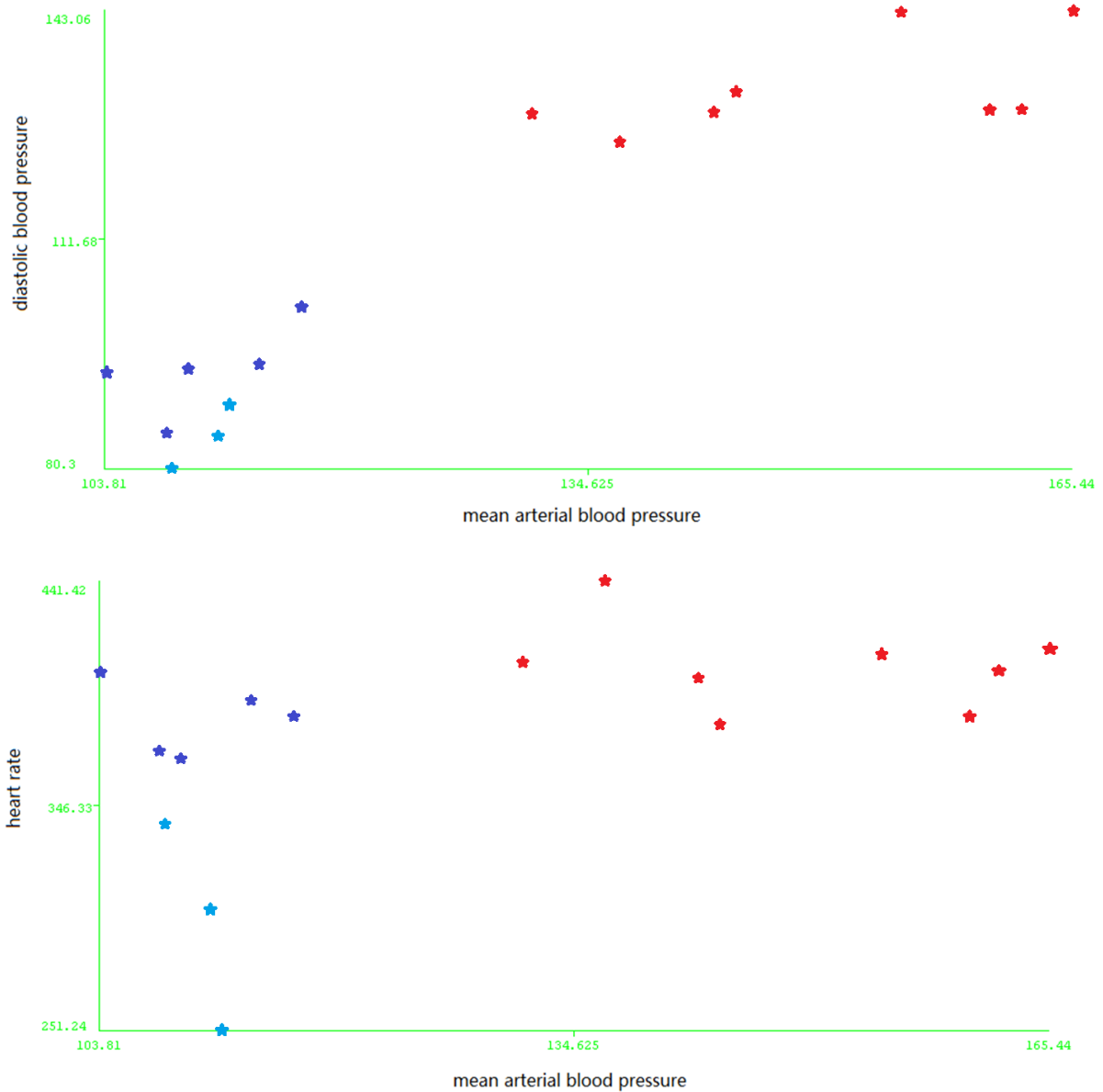
Figure 46 Scatter Plot for Model (1+2+3+4+5) with Different Features as XY Axis (Three Clusters)

Figure 46 shows the scatter plot for Model (1+2+3+4+5) with different features for the XY

axes. There are three clusters in the graph. We can see that the dark blue and light blue

clusters are actually very close to each other in the two dimensional scatter plot. However,

we were able to successfully distinguish the two clusters "WKY_Both, WKY_Male,

WKY_Female, BN_Both, BN_Male" and "LEW_Both, LEW_Male, LEW_Female" using five features.

We then included data from mutant strains and outbred strains into our analysis. We found that mutant strains and outbred strains only include Features 3, 4 and 5 (systolic blood pressure, heart rate, and heart wet weight). We were only able to identify two significant clusters (a small cluster with only one item was merged into the first cluster) (**Table 4**). **Figure 47** shows the scatter plot for Model (3+4+5) with different features on the XY axes. There are three clusters in the graph but we merged one cluster with only one item into the dark blue cluster (first cluster in **Table 4**).

Table 4 Clustering Results for Inbred, Outbred and Mutant Strains under Naïve Control Condition

Model	Cluster
3+4+5	ACI-Lystbg-Kyo/Kyo, BN/KunKtsSlc, BN_Both, BN_Male, Crl:SD, Crl:WI, DMY/Kyo, DMYC/Kyo, F344. Cg-Du TyrCKyo+/+, F344. CVD-Unc5ccvd/Kyo, F344_Both, F344_Female, F344_Male, F344-Apcm1Kyo, F344-Egrm1Kyo, F344-HrKrh/Kyo, F344-Kmch/Kyo, F344-Kuru2/Kyo, F344-Scn1am1Kyo, F344-Scn1am2Kyo, F344-Sv2am1Kyo, F344-Tbr1/Kyo, F344-Trdk/Kyo, FH/HamSlc, Gunn-Ugt1a1jSlc, HOB/Snk, HOB-Unc5chob/Snk, HTX/Kyo, HWY/Slc, KCI/Kyo, KFRS2/Kyo, KFRS2/Kyo-/+, KFRS3A/Kyo, KFRS3A/Kyo+/+, KFRS3B/Kyo, KFRS4/Kyo, KFRS6/Kyo, KHR/Kyo, KHR/Kyo-/-, KHR/Kyo-/+, KZ-LeprfaTky, LE_Both, LE_Male, LEC/Hok, LEW_Both, LEW_Female, LEW_Male, NAR/Slc, SER/Kyo, SHR_Female, SS_Both, SS_Male, TRM/Kyo, WKY_Both, WKY_Female, WKY_Male, WTC-swh/Kyo
	GH_Both, GH_Male, SHR_Both, SHR_Male, SHRSP/Ta, SHRSP_Both, SHRSP_Female, SHRSP_Male, WTC-Kcnq1dfkKyo, ZI/Kyo
	OP/Jtt

Figure 47 Scatter Plot for Model (3+4+5) with Different Features as XY Axis

The result agrees with the general classification or normal (control) and abnormal (hypertensive) strains but the ability to further classify normal strains into sub-clusters is lower than the full model with five features. This result indicates that more data for outbred and mutant strains on diastolic blood pressure and mean arterial blood pressure will be needed to further cluster strains into sub-populations.

As previously noted, researchers have often been constrained by availability of particular strains and familiarity or previous studies using those strains. Statistically clustering rat strains based on phenotypic profiles, such as what was done with the multiple cardiovascular phenotypes provides researchers with another tool to 1) identify the subpopulation of strains similar to the one in use to analyze data from these similar strains, 2) examine if they could use an alternative strain that is more readily available to them to substitute for those used by other researchers (given their closeness in the cluster); 3) examine any sex or age specific differences that might position some strains on the border of being "normal" or "abnormal" and thus not ideal to be used as a control strain.

4 Conclusions, Discussion and Future Directions

This project resulted in a platform consisting of 1) an analysis pipeline with user interface to generate expected ranges for phenotypes, and 2) a web-based data mining and visualization tools to provide easy access to this data for researchers. The pipeline and interface provides the means to 1) identify expected ranges for many types of quantitative phenotypes for different strains and different experimental conditions, 2) identify phenotypes without sufficient data to determine an expected range to prioritize these for acquisition through direct contact with researchers or extraction from published literature, 3) alert RGD staff of new phenotype data in PhenoMiner and potential changes in expected range so the pipeline can be run to determine expected ranges with this new additional data.

The data mining and visualization tools created in this project provide the ability to 1) query by phenotype or strain to retrieve expected ranges, 2) generate a profile of multiple phenotypes for a single strain, 3) compare expected ranges for phenotypes across strains, sexes and ages, 4) compare expected ranges for particular phenotypes and strains with the expected range for control animals. Such resources will allow researchers with information necessary to design studies and choose appropriate strains for their needs. A pilot cluster analysis using the expected ranges for multiple phenotypes demonstrated that such data can be used to group strains with similar phenotype profiles, providing additional resources for researchers interested in comparisons among similar strains or to assist in choosing strains with similar profiles.
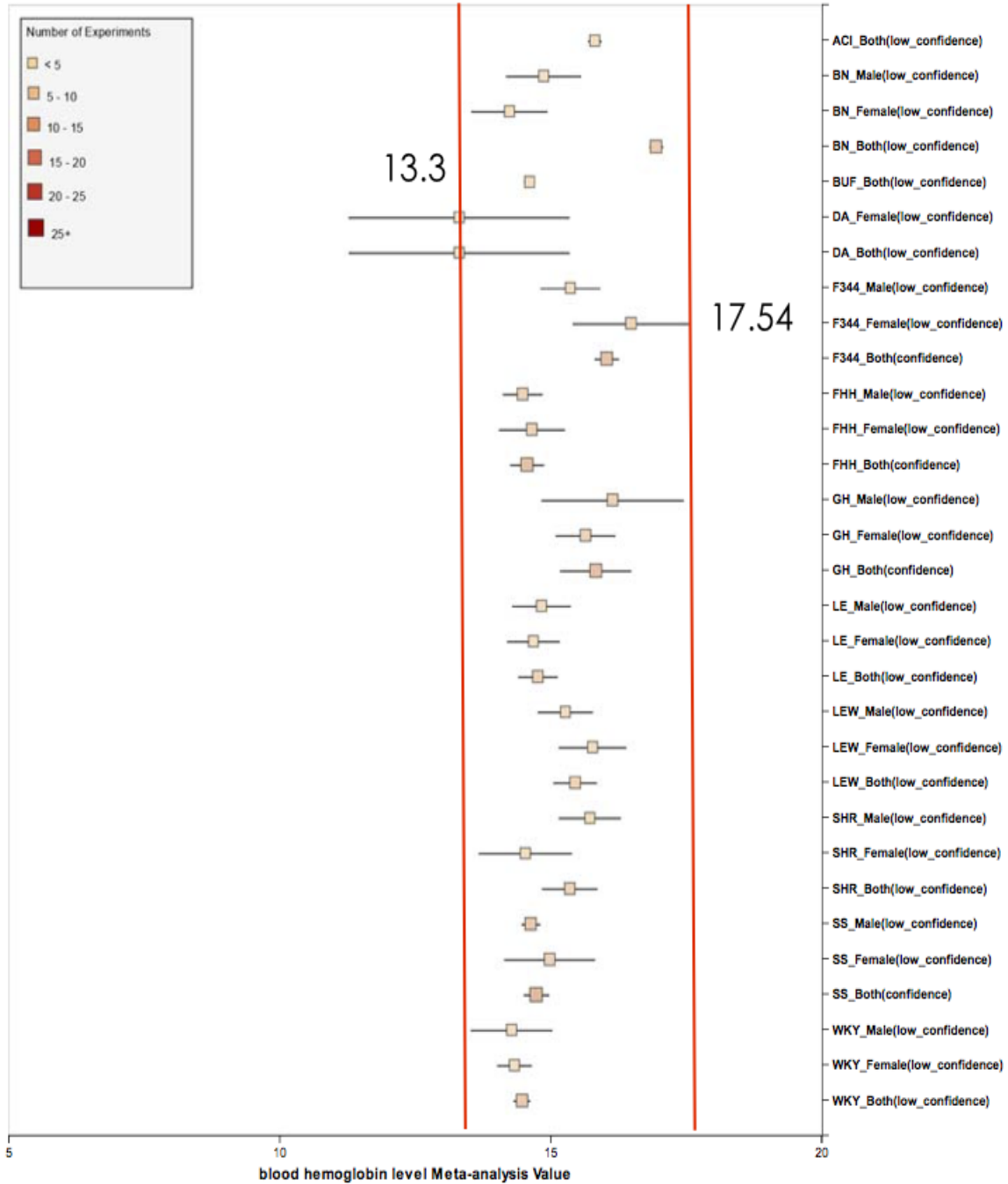
Based on the success of this project, the Rat Genome Database will develop a Precision Models Portal to present these data and link to others to provide a rich resource for investigators. The results of this study will be used to target phenotype areas for data
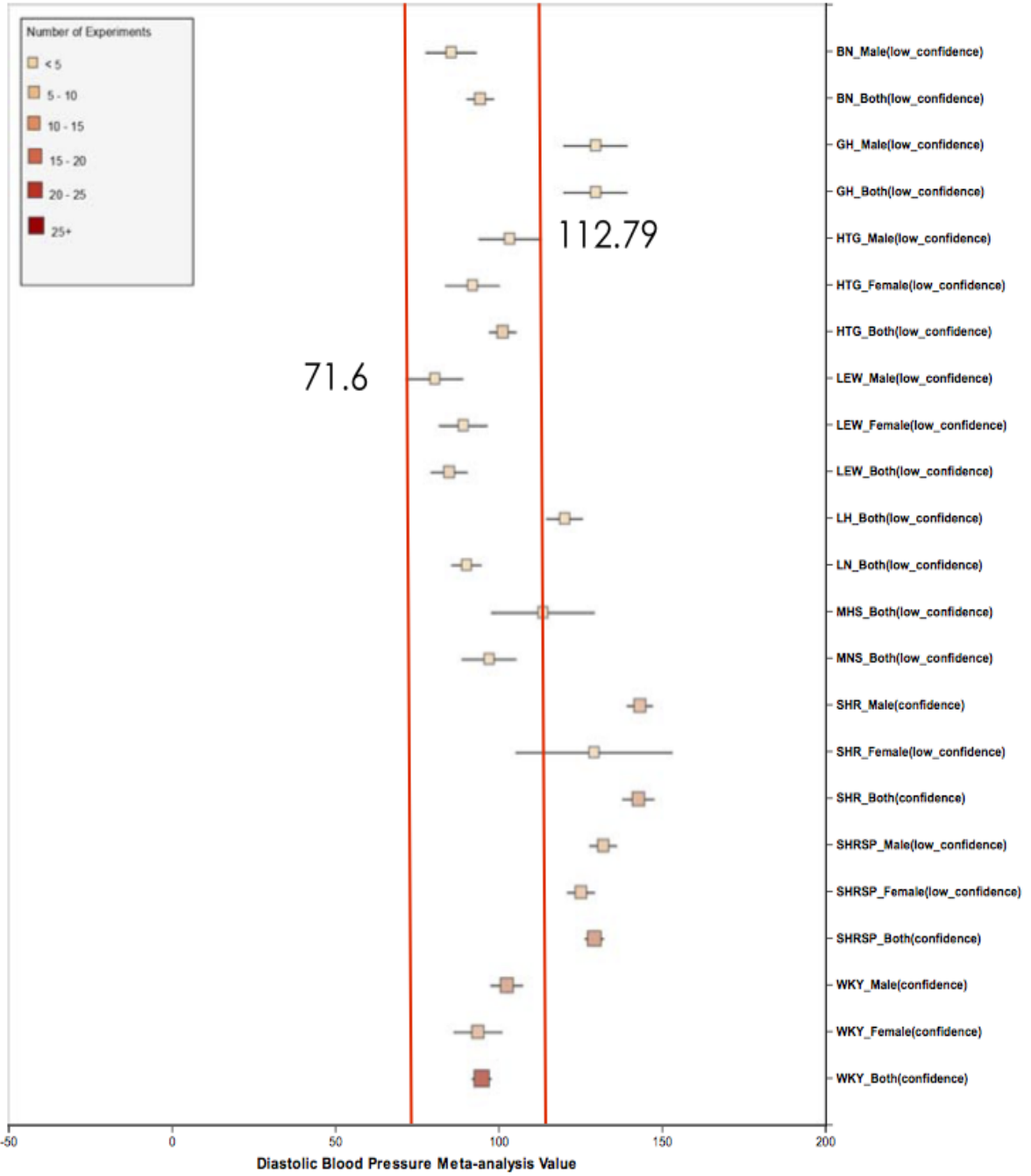
acquisition and analysis. The goal will be to provide comprehensive profiles based on the expected ranges for phenotypes across all major physiological systems. The availability of sequence and variant data for a number of the strains will offer the opportunity to provide a complementary genotype profile with the phenotype profile to enhance the ability of researchers to choose models based on both genotype and phenotype. Continued cluster analysis of the strains across multiple phenotype areas and as data expands, will provide additional information for researchers. Expanding the analysis method into a tool will also allow researchers to customize the profile of phenotypes for which they want to classify strains. Expansion of data mining and visualization tools could include allowing users to 1) upload their own data for comparison, 2) choose more than one phenotype to compare across strains in a single view, 3) integrate genotype in a visual manner.

Animal models used to study human diseases are developed to exhibit phenotypes similar to those of humans suffering from a particular disease (90-92). Researchers have a long history of developing such models, particularly in rat, and quantitatively measuring phenotypes of interest and statistically analyzing these against those measured in control animals. However, there has not been a comprehensive effort to statistically analyze quantitative phenotype measurements across multiple experiments to establish expected ranges for disease model rat strains and controls. The methods and statistical algorithms established in this project provide the platform for identifying and validating expected ranges for a variety of quantitative phenotypes. The development of interfaces and automation of the pipeline provide the ability to continuously add new expected ranges for particular phenotypes or strains and update existing ranges as new data becomes available. This data resource and the data mining and visualization tools will promote understanding of rat disease models, guide researchers to choose optimal strains for their research needs,

and encourage data sharing from different research hubs. Such resources help to promote research reproducibility. The value of the methods, platforms and data produced here go beyond the rat research community. The analysis pipeline can easily be adapted for multiple types of studies involving quantitative clinical or experimental results. Such methods have already been used to some extent with clinical trials but could be adapted for electronic phenotype analyses using data from Electronic Medical Records. The creation of a Precision Model Portal as an outgrowth of this project will provide a valuable resource for the Precision Medicine research efforts. It will increase the ability of those identifying variants that potentially impact a phenotype or set of phenotypes in a patient or human population to be able to identify the appropriate rat strain that mimics both the phenotype profile and genotype profile of that patient or population. Expanding on the analysis pipelines, data produced and interactive platforms created in this project will continue to provide a valuable resource for Translational Research.

APPENDIX A: META-ANALYSIS SUMMARIES FOR 24 CARDIOVASCULAR PHENOTYPES

Diastolic Blood Pressure Meta-analysis Value

heart left ventricle weight to body weight ratio Meta-analysis Value

heart left ventricle wet weight Meta-analysis Value

Heart Rate Meta-analysis Value

heart right ventricle weight to left ventricle weight ratio Meta-analysis Value

heart right ventricle wet weight Meta-analysis Value

heart weight as percentage of body weight Meta-analysis Value

heart weight to body weight ratio Meta-analysis Value

Heart Wt Meta-analysis Value

Arterial Blood Pressure Meta-analysis Value

mean corpuscular volume Meta-analysis Value

plasma total cholesterol level Meta-analysis Value

plasma triglyceride level Meta-analysis Value

red blood cell count Meta-analysis Value

serum aspartate aminotransferase activity level Meta-analysis Value

serum calcium level Meta-analysis Value

serum chloride level Meta-analysis Value

serum free fatty acids level Meta-analysis Value

serum potassium level Meta-analysis Value

serum total cholesterol level Meta-analysis Value

serum triglyceride level Meta-analysis Value

Systolic Blood Pressure Meta-analysis Value

APPENDIX B: SUMMARY OF ABNORMAL AND NORMAL STRAINS FOR 24 CARDIOVASCULAR PHENOTYPES

Phenotype	Abnormal	Normal	Naïve Control
blood hemoglobin level		all	
diastolic blood pressure	GH, LH, SHR, SHRSP	BN, HTG, LEW, MHS, WKY	BN, WKY
heart left ventricle weight to body weight ratio	SHR	COP, DA, DRY, F344, LEW, WKY	WKY
heart left ventricle wet weight	GH	BN, COP, DA, DRY, F344, LEW, SHR, WKY	BN, WKY
heart rate	ACI, DA, GK	BN, BUF, F344, FHH, GH, LE, LEW, SHR, SHRSP, SS, WAG, WKY	BN, WKY
heart right ventricle weight to left ventricle weight ratio	F344, FHH, GH	LE, LEW, SHR, SS, WKY	WKY
heart right ventricle wet weight		F344, WKY	WKY
heart weight as percentage of body weight	FHH, SHR, SHRSP	ACI, BN, BUF, COP, DA, F344, GK, LE, LEW, SS, WAG, WKY	BN, WKY
heart weight to body weight ratio	SHRSP	ACI, BN, BUF, COP, DA, F344, ISIAH, M520, MWF, SHR, WAG, WKY, WN	BN, WKY
heart wet weight	BUF, FHH, GH, SS	ACI, BN, COP, DA, F344, LE, LEW, MWK, WAG, WKY	BN, WKY
hematocrit		all	
mean arterial blood pressure	GH, SHR, SHRSP	BN, LE, LEW, SS, WKY	BN, WKY
mean corpuscular volume		all	
plasma total cholesterol level		HTG, LEW	
plasma triglyceride level		COP, HTG, LEW	
red blood cell count		all	
serum aspartate aminotransferase activity level	SHR	BN, F344, FHH, GH, LE, LEW, SS, WKY	BN, WKY
serum calcium level	F344, GH, LE	BN, FHH, LEW, SHR, SS, WKY	BN, WKY

serum chloride level		all	
serum free fatty acids level		F344	
serum potassium level		all	
serum total cholesterol level	GH, MWF, SHRSP	BN, DA, EHC, F344, FHH, GH, LEW, SHR, SS, WKY, WOKW	BN, WKY
serum triglyceride level	DA, WOKW	BN, EHC, F344, MWF, SHR	BN
systolic blood pressure	GH, LH, MHS, SHR, SHRSP	ACI, BN, BUF, DA, F344, GK, LE, LEW, LN, M520, MNS, MR, MWF, WKY	BN, WKY

APPENDIX C: AGREEMENTS BETWEEN META-ANALYSIS RESULTS AND DOMAIN EXPERT EMPIRICAL KNOWLEDGE

Phenotype	Agreement	Disagreement
blood hemoglobin level	31	0
diastolic blood pressure	23	0
heart left ventricle weight to body weight ratio	13	0
heart left ventricle wet weight	15	2
heart rate	41	1
heart right ventricle weight to left ventricle weight ratio	24	3
heart right ventricle wet weight	2	0
heart weight as percentage of body weight	36	3
heart weight to body weight ratio	20	0
heart wet weight	44	1
hematocrit	33	0
mean arterial blood pressure	30	0
mean corpuscular volume	31	0
plasma total cholesterol level	5	0
plasma triglyceride level	7	0
red blood cell count	27	0
serum aspartate aminotransferase activity level	27	0
serum calcium level	27	0
serum chloride level	27	0
serum free fatty acids level	2	0
serum potassium level	27	0
serum total cholesterol level	35	0
serum triglyceride level	12	0
systolic blood pressure	27	1

REFERENCES

1.	Jahoda CA, Kljuic A, O'Shaughnessy R, Crossley N, Whitehouse CJ, Robinson M, et al. The lanceolate hair rat phenotype results from a missense mutation in a calcium coordinating site of the desmoglein 4 gene. Genomics. 2004; 83(5):747-56.

2.	Steiniger B, Stehling O, Scriba A, Grau V. Monocytes in the rat: phenotype and function during acute allograft rejection. Immunological reviews. 2001; 184(1):38-44.

3.	Rocha R, Rudolph AE, Frierdich GE, Nachowiak DA, Kekec BK, Blomme EA, et al. Aldosterone induces a vascular inflammatory phenotype in the rat heart. American Journal of Physiology-Heart and Circulatory Physiology. 2002; 283(5):H1802-H10.

4.	Monte MJ, Badia MD, Serrano MA, Sacristan MP, Cassio D, Marin JJ. Predominance of human versus rat phenotype in the metabolic pathways for bile acid synthesis by hybrid WIF-B9 cells. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 2001; 1534(1):45-55.

5.	Szyf M, Weaver IC, Champagne FA, Diorio J, Meaney MJ. Maternal programming of steroid receptor expression and phenotype through DNA methylation in the rat. Frontiers in neuroendocrinology. 2005; 26(3):139-62.

6.	Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, et al. Interleukin 7 receptor α chain (IL7R) shows allelic and functional association with multiple sclerosis. Nature genetics. 2007; 39(9):1083-91.

7.	Hajri T, Han XX, Bonen A, Abumrad NA. Defective fatty acid uptake modulates insulin responsiveness and metabolic responses to diet in CD36-null mice. The Journal of clinical investigation. 2002; 109(10):1381-9.

8.	Yokoi N, Komeda K, Wang H-Y, Yano H, Kitada K, Saitoh Y, et al. Cblb is a major susceptibility gene for rat type 1 diabetes mellitus. Nature genetics. 2002; 31(4):391-4.

9.	MacMurray AJ, Moralejo DH, Kwitek AE, Rutledge EA, Van Yserloo B, Gohlke P, et al. Lymphopenia in the BB rat model of type 1 diabetes is due to a mutation in a novel immune-associated nucleotide (Ian)-related gene. Genome research. 2002; 12(7):1029-39.

10.	Olofsson P, Holmberg J, Tordsson J, Lu S, Åkerström B, Holmdahl R. Positional identification of Ncf1 as a gene that regulates arthritis severity in rats. Nature genetics. 2003; 33(1):25-32.

11.	Jacob HJ. Functional genomics and rat models. Genome research. 1999; 9(11):1013-6.

12.	Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004; 428(6982):493-521.

13.	Petretto E, Sarwar R, Grieve I, Lu H, Kumaran MK, Muckett PJ, et al. Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass. Nature genetics. 2008; 40(5):546-52.

14.	Monti J, Fischer J, Paskas S, Heinig M, Schulz H, Gösele C, et al. Soluble epoxide hydrolase is a susceptibility factor for heart failure in a rat model of human disease. Nature genetics. 2008; 40(5):529-37.

15.	Samuelson DJ, Hesselson SE, Aperavich BA, Zan Y, Haag JD, Trentham-Dietz A, et al. Rat Mcs5a is a compound quantitative trait locus with orthologous human loci that associate with breast cancer risk. Proceedings of the National Academy of Sciences. 2007; 104(15):6299-304.

16.	Swanberg M, Lidman O, Padyukov L, Eriksson P, Åkesson E, Jagodic M, et al. MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. Nature genetics. 2005; 37(5):486-94.

17.	Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006; 439(7078):851-5.

18.	Behmoaras J, Bhangal G, Smith J, McDonald K, Mutch B, Lai PC, et al. Jund is a determinant of macrophage activation and is associated with glomerulonephritis susceptibility. Nature genetics. 2008; 40(5):553-9.

19.	Smith CL, Eppig JT. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mammalian genome. 2012; 23(9-10):653-68.
20.	Rat Phenome Project http://www. anim. med. kyoto-u. ac. jp/nbr/phenome. Available from: http://www. anim. med. kyoto-u. ac. jp/nbr/phenome.
21.	Cowley AW, Roman RJ, Jacob HJ. Application of chromosomal substitution techniques in gene‐function discovery. The Journal of physiology. 2004; 554(1):46-55.
22.	Cowley A, Liang M, Roman R, Greene A, Jacob H. Consomic rat model systems for physiological genomics. Acta Physiologica. 2004; 181(4):585-92.
23.	Smits BM, Mudde JB, van de Belt J, Verheul M, Olivier J, Homberg J, et al. Generation of gene knockouts and mutant models in the laboratory rat by ENU-driven target-selected mutagenesis. Pharmacogenetics and genomics. 2006; 16(3):159-69.
24.	PhysGen http://pga. mcw. edu/. Available from: http://pga. mcw. edu/.
25.	Dwinell MR. Online tools for understanding rat physiology. Briefings in bioinformatics. 2010; 11(4):431-9.
26.	Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic acids research. 2013; 42(D1):D802-D9.
27.	Shimoyama M, De Pons J, Hayman GT, Laulederkind SJ, Liu W, Nigam R, et al. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic acids research. 2014; 43(D1):D743-D50.
28.	Laulederkind SJ, Hayman GT, Wang S-J, Smith JR, Lowry TF, Nigam R, et al. The Rat Genome Database 2013—data, tools and users. Briefings in bioinformatics. 2013; 14(4):520-6.
29.	Wang S-J, Laulederkind SJ, Hayman GT, Petri V, Liu W, Smith JR, et al. PhenoMiner: a quantitative phenotype database for the laboratory rat, Rattus norvegicus. Application in hypertension and renal disease. Database. 2015; 2015:bau128.
30.	PhysGen Knockout Project http://www. rgd. mcw. edu/wg/physgenknockouts. Available from: http://www. rgd. mcw. edu/wg/physgenknockouts.
31.	Nigam R, Munzenmaier DH, Worthey EA, Dwinell MR, Shimoyama M, Jacob HJ. Rat Strain Ontology: structured controlled vocabulary designed to facilitate access to strain data at RGD. Journal of biomedical semantics. 2013; 4(1):1.
32.	Shimoyama M, Nigam R, McIntosh LS, Nagarajan R, Rice T, Rao D, et al. Three ontologies to define phenotype measurement data. 2012.
33.	Smith JR, Park CA, Nigam R, Laulederkind SJ, Hayman GT, Wang S-J, et al. The clinical measurement, measurement method and experimental condition ontologies: expansion, improvements and new applications. Journal of biomedical semantics. 2013; 4(1):1.
34.	Colborn T. Commentary: Setting aside tradition when dealing with endocrine disruptors. ILAR journal. 2004; 45(4):394-400.
35.	Spearow JL, Doemeny P, Sera R, Leffler R, Barkley M. Genetic variation in susceptibility to endocrine disruption by estrogen in mice. Science. 1999; 285(5431):1259-61.
36.	Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS biology. 2018; 16(2):e2003693.
37.	Bailar III JC. The promise and problems of meta-analysis. New England Journal of Medicine. 1997; 337(8):559-61.
38.	Mulrow CD. Rationale for systematic reviews. BMJ: British Medical Journal. 1994; 309(6954):597.
39.	Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. Journal of clinical epidemiology. 2000; 53(2):207-16.
40.	Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997; 315(7109):629-34.
41.	Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. Journal of clinical epidemiology. 2001; 54(10):1046-55.

42.   Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. Jama. 2006; 295(6):676.

43.   Egger M, Smith GD. Misleading meta-analysis. BMJ: British Medical Journal. 1995; 311(7007):753.

44.   Sterne JA, Egger M, Smith GD. Investigating and dealing with publication and other biases in meta-analysis. British Medical Journal. 2001; 323(7304):101.

45.   Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. Bmj. 2006; 333(7568):597-600.

46.   Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. Journal of Clinical Epidemiology. 2005; 58(9):894-901.

47.   Ioannidis JPA. Interpretation of tests of heterogeneity and bias in meta-analysis. Journal of Evaluation in Clinical Practice. 2008; 14(5):951–7.

48.   Sterne JAC, Gavaghan D, Egger M. Publication and related bias in meta-analysis : Power of statistical tests and prevalence in the literature. Journal of Clinical Epidemiology. 2000; 53(11):1119.

49.   Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. Bmj. 2006; 333(7568):597-600.

50.   Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Statistics in Medicine. 2001; 20(4):641.

51.   Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Statistics in Medicine. 1988; 7(8):889-94.

52.   Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Statistics in Medicine. 2006; 25(20):3443–57.

53.   Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. British Medical Journal. 2001; 322(7300):1479.

54.   Bax L, Yu LM, Ikeda N, Tsuruta H, Moons KG. Development and validation of MIX: comprehensive free software for meta-analysis of causal research data. BMC Medical Research Methodology, 6, 1(2006-10-13). 2006; 6(1):50.

55.   Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003; 327(7414):557-60.

56.   Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I² index? Psychological methods. 2006; 11(2):193.

57.   Higgins J, Thompson SG. Quantifying heterogeneity in a meta‐analysis. Statistics in medicine. 2002; 21(11):1539-58.

58.   Evangelou E, Ioannidis JP, Patsopoulos NA. Uncertainty in heterogeneity estimates in meta-analyses. BMJ: British Medical Journal. 2007; 335(7626):914-6.

59.   Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random‐effects regression model for meta‐analysis. Statistics in medicine. 1995; 14(4):395-411.

60.   Thompson SG, Higgins J. How should meta‐regression analyses be undertaken and interpreted? Statistics in medicine. 2002; 21(11):1559-73.

61.   Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. Journal of clinical epidemiology. 2004; 57(7):683-97.

62.   Hartung J, Knapp G. A refined method for the meta‐analysis of controlled clinical trials with binary outcome. Statistics in Medicine. 2001; 20(24):3875-89.

63.   Sidik K, Jonkman JN. A note on variance estimation in random effects meta-regression. Journal of Biopharmaceutical Statistics. 2005; 15(5):823-38.

64.   IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC medical research methodology. 2014; 14(1):1.

65.   DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986; 7(3):177-88.

66.   DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemporary clinical trials. 2015; 45:139-45.

67.   DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemporary clinical trials. 2007; 28(2):105-14.

68.   Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? Journal of Statistical Planning and Inference. 2010; 140(4):961-70.

69.   Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. Statistical methods in medical research. 2010:0962280210392008.

70.   Cochran WG. The combination of estimates from different experiments. Biometrics. 1954; 10(1):101-29.

71.   Paule RC, Mandel J. Consensus values and weighting factors. Journal of Research of the National Bureau of Standards. 1982; 87(5):377-85.

72.   Ruff CT, Giugliano RP, Braunwald E, Hoffman EB, Deenadayalu N, Ezekowitz MD, et al. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. The Lancet. 2014; 383(9921):955-62.

73.   Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, et al. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. The Lancet. 2014; 383(9919):785-95.

74.   Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. The American Journal of Human Genetics. 2011; 88(5):586-98.

75.   LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. New England Journal of Medicine. 1997; 337(8):536-42.

76.   Villar J, Carroli G, Belizan J. Predictive ability of meta-analyses of randomised controlled trials. The lancet. 1995; 345(8952):772-6.

77.   National Bio Resource Project http://www. nbrp. jp/. Available from: http://www. nbrp. jp/.

78.   Mashimo T, Voigt B, Kuramoto T, Serikawa T. Rat Phenome Project: the untapped potential of existing rat strains. Journal of Applied Physiology. 2005; 98(1):371-9.

79.   Aitman TJ, Critser JK, Cuppen E, Dominiczak A, Fernandez-Suarez XM, Flint J, et al. Progress and prospects in rat genetics: a community view. Nature genetics. 2008; 40(5):516.

80.   Mather K. Polygenic inheritance and natural selection. Biological Reviews. 1943; 18(1):32-64.

81.   Pociot F, Akolkar B, Concannon P, Erlich HA, Julier C, Morahan G, et al. Genetics of type 1 diabetes: what's next? Diabetes. 2010; 59(7):1561-71.

82.   Holmes C. Genotype and phenotype in Alzheimer's disease. The British Journal of Psychiatry. 2002; 180(2):131-4.

83.   Dominiczak AF, McBride MW. Genetics of common polygenic stroke. Nature genetics. 2003; 35(2):116-7.

84.   Moore WC, Bleecker ER, Curran-Everett D, Erzurum SC, Ameredes BT, Bacharier L, et al. Characterization of the severe asthma phenotype by the national heart, lung, and blood institute's severe asthma research program. Journal of Allergy and Clinical Immunology. 2007; 119(2):405-13.

85.   Kim S, Sohn K-A, Xing EP. A multivariate regression approach to association analysis of a quantitative trait network. Bioinformatics. 2009; 25(12):i204-i12.

86.   Li H, Kong F, Xu J, Zhang M, Wang A, Zhang Y. Hypertension subtypes and risk of cardiovascular diseases in a Mongolian population, inner Mongolia, China. Clinical and Experimental Hypertension. 2016; 38(1):39-44.

87.   Kawano Y, Horio T, Matayoshi T, Kamide K. Masked hypertension: subtypes and target organ damage. Clinical and experimental hypertension. 2008; 30(3-4):289-96.

88.   Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. Jmir Med Inform. 2017; 5(1):e7.

89. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009; 11(1):10-8.

90. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(21):8685-90.

91. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(11):4323-8.

92. Rual JF, Venkatesan K, Hao T, Hirozanekishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437(7062):1173-8.

93. Staff F. Ivacaftor (Kalydeco): A treatment of cystic fibrosis for patients with G551D mutation in the CFTR gene. Formulary. 2012.

94. Roychowdhury S, Chinnaiyan AM. Advancing precision medicine for prostate cancer through genomics. Journal of Clinical Oncology Official Journal of the American Society of Clinical Oncology. 2013; 31(15):1866.

95. Doroshow J. National Cancer Institute's Precision Medicine Initiatives for the New National Clinical Trials Network. 2014; 34:71.

96. De VS, Swanepoel A, Bester J, Pretorius E. Novel Diagnostic and Monitoring Tools in Stroke: an Individualized Patient-Centered Precision Medicine Approach. Journal of Atherosclerosis & Thrombosis. 2015; 23(5).

97. Grace A. Prophylactic Implantable Defibrillators for Hypertrophic Cardiomyopathy: Disarray in the Era of Precision Medicine. Circulation Arrhythmia & Electrophysiology. 2015; 8(4):763-6.

98. Cunningham D, Danley DE, Geoghegan KF, Griffor MC, Hawkins JL, Subashi TA, et al. Structural and biophysical studies of PCSK9 and its mutants linked to familial hypercholesterolemia. Nature Structural & Molecular Biology. 2007; 14(5):413.

99. Homer VM, Marais AD, Charlton F, Laurie AD, Hurndell N, Scott R, et al. Identification and characterization of two non-secreted PCSK9 mutants associated with familial hypercholesterolemia in cohorts from New Zealand and South Africa. Atherosclerosis. 2008; 196(2):659-66.

100. Raal FJ, Stein EA, Dufour R, Turner T, Civeira F, Burgess L, et al. PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD-2): a randomised, double-blind, placebo-controlled trial. Journal of Vascular Surgery. 2015; 62(5):1368-.

101. alirocumab, Inhibitor P. Alirocumab (Praluent) to Lower LDL-Cholesterol. Jama. 2015; 314(12):1284.

102. evolocumab, Inhibitor P. Evolocumab (Repatha)--a second PCSK9 inhibitor to lower LDL-Cholesterol. Jama. 2015; 314(21):2298-9.

103. Pfeufer A, Noord CV, Marciante KD, Dan EA, Larson MG, Smith AV, et al. Genome-wide association study of PR interval. Nature Genetics. 2010; 42(2):153-9.

104. Macrae CA, Vasan RS. The Future of Genetics and Genomics. Circulation. 2016; 133(25):2634-9.

105. Ho JE, Enserro D, Brouwers FP, Kizer JR, Shah SJ, Psaty BM, et al. Predicting Heart Failure With Preserved and Reduced Ejection Fraction: The International Collaboration on Heart Failure Subtypes. Circulation Heart Failure. 2016; 9(6):e003116.

106. Panahiazar M, Taslimitehrani V, Pereira NL, Pathak J. Using EHRs for heart failure therapy recommendation using multidimensional patient similarity analytics. Studies in health technology and informatics. 2015; 210:369.

CURRICULUM VITAE

Yiqing Zhao

EDUCATION

University of Wisconsin, Milwaukee
Ph.D., Medical Informatics, GPA 3.86/4   Expected graduate: May. 2018
University of California, Los Angeles
M.S., Department of Bioengineering, Minor: Biostatistics GPA 3.73/4   Sep. 2014
Fudan University
B.S., Dept. of Chemistry , GPA 3.53/4   Jun. 2012

PROFESSIONAL EXPERIENCE

Biostatistics Intern, Mayo Clinic   May. 2017 - Aug. 2017
Project: Mapping Rochester Epidemiology Project Research Elements to Common Data Model
- Using content based approach to create a standardized data model to represent Rochester Epidemiology Project cohort study dataset content.

Project: Rare Disease Knowledge Base
- Using Natural Language Processing to mine phenotype and disease associations from EMR.
- Using network analysis to examine and filter excess phenotype and disease associations and create a disease-disease association network.

Data Scientist, Department of Surgery, Medical College of Wisconsin   Jan. 2016 – now
Thesis: Rat Strain Stratification with Phenotype and Genotype Landscape
- Meta-analysis on rat phenotype data to establish a normal range for quantitative traits (such as systolic blood pressure, etc) considering inter-laboratory variations or other confounding factors.
- Construct phenotype network by mining OMIM database; Use Frequent Pattern Mining to cluster phenotypes; Bridge human and rat phenotype ontology by comparing human and rat phenotype subgraph features
- Using phenotype subgraph features to perform graph based LASSO to boost statistical power for association analysis

Research Scientist, The Center for Human Genetics, Marshfield Clinic Research Foundation
                                              May. 2016 - Aug.2016
Project: Development of a Big Data toolset for identifying functional genetic variants
- Developed a web-based variant annotation program called cloudANNOVAR on a Big Data (Cassandra) infrastructure

Research Assistant, Department of Healthcare Administration, UWM   Sep. 2015 - May. 2016
Project: Bridging the Gap of Medical Image and Text Representation
- NLP-based Semantic Searching using UMLS Ontology Lexicon (JAVA)
- Radiology Knowledge Database Establishment and Web Application Development (JSP, JavaScript, HTML5, CSS3)

Term Project: Computational Cloud Platform and Data Exploration of Clinical Trials
- Using Hadoop and Spark to build a cloud based tool for visualizing and analyzing (using K-means clustering) geographic patterns of global clinical trials using the published clinical trial data

PUBLICATION & PRESENTATION

Lab On a Chip , "Interfacial Organic Synthesis in a Simple Droplet-Based Microfluidic System"
Jan. 2012
Journal of Medical Informatics, "Big Data Application in Biomedical Informatics"   May 2015
Biomed Inform Insights, "Big Data Application in Biomedical Research and Health Care: A Literature Review"   Jan. 2016
PLOS One, "Pathogenic Mutations in Cancer-Predisposing Genes: A Survey of 300 Patients with Whole-Genome Sequencing and Lifetime Electronic Health Records"   Dec. 2016
AMIA Joint Summit Poster "Reference Mining to Improve Cohort Establishment Method Consistency**"**   Mar. 2018
Accepted BMC Clinical Decision Making, "Using Data-Driven Sublanguage Pattern Mining to Induce Knowledge Models: Application in Medical Image Reports Knowledge Representation"
Accepted Journal of Medical Systems "Semantic-Enhanced Query Expansion System for Retrieving Medical Imaging Notes" (#JOMS-D-18-00048R1)
Submitted "Mapping Rochester Epidemiology Project Research Elements to Common Data Model"

HONORS AND AWARDS

Chinese National First-Prize Fellowship; Rank1/ 107   Sep. 2009
Chinese National Fellowship; Rank2/107   Sep. 2010
Chinese Scholarship Council   Sep. 2012-June 2014
UWM Chancellor's award   Jan.2015