Poster B-22

Online dynamic visualization and statistical analyses of rat phenome data

Authors:
Anne E Kwitek (*Medical College of Wisconsin*)
Zhanchi Wang (*Medical College of Wisconsin*)
Allen W. Cowley Jr. (*Medical College of Wisconsin*)
Andrew S. Greene (*Medical College of Wisconsin*)
Richard J. Roman (*Medical College of Wisconsin*)
Howard J. Jacob (*Medical College of Wisconsin*)

Short Abstract: PhysGen has generated and characterized over 200 phenotypes in two panels of rat chromosome substitution strains. All data is released to the public via our website (http://pga.mcw.edu). To facilitate more streamlined mining of the PhysGen data, we have developed flexible and dynamic visualization tools, with linked online statistical analysis.

Long Abstract:

The post-genome era has provided resources to link disease phenotypes to the genomic sequence, i.e. creating a disease 'phenome'. The rat, with >300 models that capture key aspects of human complex is ideal for a phenome project. The availability of the genome sequence of the BN/SsNHsd/Mcwi (BN) rat ushers in a new era for investigators using the rat in their research programs. Our detailed characterization of the sequenced BN rat strain (BN/NHsdMcwi) provides the first concerted effort in creating a direct link between a sequenced genome and its resulting biology. As a major step in generating a comprehensive cardiovascular and pulmonary disease phenome, we measured 281 traits related to diseases of the heart, lung, kidney, vasculature, and blood (http://pga.mcw.edu) in the sequenced BN and 10 other strains (8 inbred and 2 outbred), enabling us to determine "normal" and "abnormal" physiological values for the BN. In addition, to map these traits at the chromosomal level in the rat, we have transfered each of the 22 chromosomes (20 autosomes and the X and Y chromosomes) from the BN onto two different genome backgrounds: the SS/JrHsd/Mcwi (SS) and FHH/EurMcwi (FHH) hypertensive rat strains. Each of the 44 chromosome substitution (consomic) strains of rat is characterized for the same phenotypes as those measured in the other inbred and outbred strains. We found that the BN genome contains protective genes for heart, lung and renal disease on some chromosomes, and susceptibility genes for these same traits on other chromosomes. Collectively, we have determined over 8,000 physiological measures per strain and investigated the impact of genome background on the "sequenced" BN strain, providing a powerful and unique tool kit for functional genomics and systems biology in the rat.

All the PhysGen data, including phenotypic measures and genotypes of all characterized strains, are housed in Oracle relational databases. PhysGen maintains three functional databases, all running on the same Oracle 9i platform; they include a Sun 450 for our development server, Sun 480 for production site, and a Sun 880 for a mirror site. The public database has a website interface that allows a user to visualize, analyze and download ALL the phenotypic (raw or mean data) and genotypic data. New data are released to the public

on a quarterly basis through our public website (http://pga.mcw.edu). A user can obtain all raw phenotypic data through a download option. Currently, with characterization of 54 of the 55 parental and consomic strains, there are 429,603 mean data values in the database.

With such a large volume of data, we have developed visualization and statistical tools on the website, enabling an investigator/user to evaluate the data in advance of downloading the data. We have created a visualization tool to dynamically display strain distribution patterns for each trait (http://pga.mcw.edu/pga/data_status.html), partitioned by protocol. To visualize a trait, an investigator can select 'Data;' on the PhysGen home page, then choose 'Phenotype Data Download and Analysis'. On this page, one can choose the protocol of interest, e.g. Biochemistry, and select 'Visualization and Statistics'. The following web page offers a variety of options to dynamically view strain phenotype distribution patterns. For instance, an investigator could choose to visualize the 'SSBN consomic strains', under 'Hypoxic (12% O2) preconditioning, and 'Males', for the phenotype 'plasma cholesterol' and 'Do Analysis'. The result is presented as a histogram showing the mean value for each strain, automatically sorted in ascending order. Below the histogram is a table listing additional detail (e.g. mean + SEM, N, environment, diet) as well as options for dynamic statistical analysis (described below). Each bar has mouse-over capability that displays the strain name, the environmental and diet conditions, gender, and the mean value. Clicking on the histogram bar returns a pop-up window displaying the distribution of the raw trait values.

Several statistical tests exist on the website. These dynamic analysis tools were used for the strain distribution patterns, and can be used for any phenotypic data selected for visualization. Each test is automatically performed on a hierarchical basis depending on a series of decision trees. The first test is for homoscedasticity of the data using Levene's test. If Levene's test passed, a parametric ANOVA or T-test is used, depending on the number of strains selected for statistical analysis. If a T-test is selected, a parametric test is automatically selected based upon the results of Levene's test; if an ANOVA is chosen, results of a conventional ANOVA are provided. For the ANOVA, two post-hoc analysis methods can be selected, a pairwise analysis using Tukey's test, or comparison to a control strain using Dunnett's test. Correction for multiple testing is accounted for in both tests. If the Levine's test for homoscedasticity fails, a non-parametric analysis is used. T-test analyses are performed using a Mann-Whitney test; non-parametric ANOVA analyses are performed using the Kruskal-Wallis test. Post-hoc analyses for the non-parametric ANOVA include Dunn's test for all pairwise comparisons and a non-parametric Dunnet's test for pairwise comparisons to a selected control strain. Again, each test includes a correction for multiple testing errors.

In addition to visualizing the distribution of mean trait values across all parental strains using the dynamic online visualization and analysis tools, we also generated a strain report that enables an investigator to identify traits that are significantly different from the "population" mean in the BN strain, as well as the other 10 parental strains, thus enabling the determination of additional correlated cardiovascular or pulmonary traits. For each strain, a list of significantly different traits can be obtained online in a 'Strain Profile' at http://pga.mcw.edu/pga-bin/strain_profile.cgi, with links to graphical visualization and statistical analyses for each trait. These data can also be directly downloaded as mean and raw data values.