A Genome-Wide Association Study of Inbred Rat Strains

Douglas Greiman[1], Atul Butte[2]

[1]Stanford Center for Professional Development, Stanford University,
[2]School of Medicine, Stanford University,
Stanford, CA 94305 USA
Email: Douglas Greiman - duggelz@gmail.com; Atul Butte - abutte@stanford.edu

This paper describes a genome-wide association study conducted on various inbred strains of Brown Norway rat. The study used preexisting, publicly available data. Phenotype data collected by NBRP-Rat, Kyoto, was merged with genotype data collected by the European STAR consortium. Applying a stringent Bonferroni correction, no statistically significant results were found. Applying a more lenient criterion based on false discovery rate led to several hundred possibly significant correlations between SNPs and phenotypes. These findings were combined with Gene Ontology annotations from the Rat Genome Database to associate particular phenotypes with particular Gene Ontology terms. The associations suggest biological pathways and mechanisms that may give rise to the phenotypic variations observed between various strains of rats.
Keywords: Genome-Wide Association Study; Rattus Norvegicus; STAR Consortium; National BioResource Project For the Rat; Rat Genome Database

1. Introduction

In the past two years, genome-wide association studies (GWAS) have been successful in finding novel genes related to Parkinson's Disease[1], Type II Diabetes[2], human height variation[3] and more. These studies share a number of common characteristics. The studies are done using human subjects. Each individual is measured for both genotype and phenotype. The phenotypes may be quantitative measurements such as height or blood pressure, or binary classifications into disease and control groups. Each study typically only addresses a single phenotype. The genotypes measured are pairs of alleles at single nucleotide polymorphisms (SNPs). The set of SNPs used is large and spans the entire genome, although it is not the complete set of all SNPs known. Often, data from the International HapMap project is used to infer values for unmeasured SNPs based on the measured SNPs near them.

The type of data gathered by a GWAS is very personal, private, and highly revealing. Thus, this data is not made publicly available. For a student researcher, this human data is hard to get. Access to the data requires going though an application process, and requires approval from a medical ethics board or similar body. Use of the data is restricted to the informed consent gathered from study participants. Uses outside of those already approved requires obtaining new consent from every subject. The application process is generally geared toward established investigators with degrees, laboratories, advisors, grants, and some idea of what they are doing. Furthermore, the approval process takes time. Finally, simply tracking down who has what phenotype data is not simple. For example, the Wellcome Trust Case Control Consortium aggregates data from many other sources. However, it only has very limited phenotype data. The complete data is held by the original investigators who must be contacted directly to get access to the data.

These difficulties stem from the fact that the study subjects are human. Given this, I decided to look for

data elsewhere. I decided to look to other species. I was partially successful. I did not find any publicly available data for the genotype and phenotype of individual animals. The process of genotyping an organism is cheaper than it has ever been, but it's still far from free, and no one has yet mass-genotyped animals and put that data on the Internet. However, a number of model organisms are maintained as inbred strains. These are animals of a particular species in which brother-sister pairs of animals have been bred for twenty or more generations to produce lines of animals which are almost genetically identical. It is possible to genotype a representative from such an inbred strain and expect that any other animal from that strain will have virtually the same genotype. Then, other animals from that strain can be measured for any desired phenotype. For this project, genotype data for inbred rat strains gathered by the European STAR consortium was combined with phenotype data for inbred rat strains gathered by the Kyoto Rat Phenotype Database. The Rat Genome Database was also used to obtain Gene Ontology (GO) annotations as described later.

The European STAR consortium[4] was formed in 2005 to genotype various strains of rats used in the laboratory. Organized as a consortium of laboratories, to date the project has obtained genotypes for a subset of 20,238 SNPs across 167 distinct inbred rat strains, two rat recombinant panels, and an F2 intercross. The data is publicly available and freely downloadable. For each SNP in their panel, the measured allele is given for each of 466 strains and substrains. Recall that as inbred strains, the organisms will be homozygous at most alleles. Each SNP is identified by a SNP identifier, as well as a chromosome number and chromosome base pair location relative to the RGSC v3.4 assembly. Of the 20,238 SNPs, 10,871 were measured on an Illumina platform, and the remaining 9,412 were measured on an Affymetrix platform. Each strain is identified using the standardized MGI nomenclature for rat strains.

The Japanese National BioResource Project (NBRP) is a national project that aims to collect, preserve, and provide bioresources for life sciences research. The NBRP Project for the Rat, Institute of Laboratory Animals, Kyoto University, Japan, maintains the Kyoto Rat Phenotype Database. The institute has measured 74 different phenotypes for 204 strains of rat. The data is made publicly available through an interactive web interface spread over a several different web pages. There are eight major categories of phenotypes: Body Weight at Various Ages, Spontaneous Locomotor Activity, Passive Avoidance Test, Blood Pressure and Body Temperature, Blood Biochemistry, Hematology, Urine Volume and Urine Electrolyte Values, and Organ Weights. For each phenotype, six male rats from each strain were measured and the results averaged to produce a single strain value. The rats were of the same age and from the same laboratory environment. Measurements of female rats are also available for most strains. For this study, only male rat measurements are used. The specifics of how each phenotype was measured is described on their web site. Strains are identified using the same nomenclature that the STAR consortium uses. The NBRP-Rat project also maintains genotype information on the strains in its database. However, the genotype information is limited to 400 SSLPs for each strain. The SSLP data was not used for this project.

2. Data Acquisition

The STAR data is available as a single plain text file in tab-separated-values (TSV) format, one click away from the home page of the STAR consortium. This data was downloaded, and no further processing was required.

The NBRP-Rat data is harder to access in bulk. The data is made available through a searchable, sortable, database-backed website implemented on ASP.NET. In order to obtain data, I wrote a screen-scraping tool in Python using the Beautiful Soup HTML parsing library. This tool extracted data fields from 24 web pages, parsed them, and merged them all into a single table of 74 rows and 204 columns,

saved as a local CSV file.

The final stage of this project used data from the Rat Genome Database (RGD). The RGD offers programmatic access to its data via a number of methods. There is a web form-based interface, a Perl library, a REST-like API using XML formatted requests, and direct download of the database files. I wrote a Python program to query the Rat Genome Database via the REST-like API.

3. Methods

In a typical GWAS, hundreds, thousands, or even tens of thousands of individuals are used. Only a single phenotype is measured or studied at a time. Hundred of thousands of SNPs are directly measured, and if HapMap data is used, millions of SNPs are inferred.

By contrast, this study uses only 20,000 SNPs. However, the STAR consortium picked these SNPs from the entire span of the rat genome. Thus, this data is less dense than typical human data, but still covers all the chromosomes of the rat. The number of results found will clearly be smaller, but the analysis is otherwise the similar. Now look at phenotypes. We examine 74 different phenotypes, which is a many more than the single phenotype measured in many GWAS studies. This also doesn't materially affect the analysis, it simply means that more results will be found. In both cases, a correction for multiple hypothesis testing will be required, as for any GWAS. Now look at the samples. Instead of samples from thousands of human individuals, we have samples from 51 strains. This small sample size will reduce the confidence level of all our results, and cause us to miss results that a study with larger sample sizes would find. Finally, instead of measure genotypes and phenotypes from individuals where the two are directly comparable, in this study the genotypes for each strain are assumed to be identical for all individuals in that strain, and the phenotypes are the averaged measurements from six males of that strain reared in identical environments. If the inbred rat strains truly are virtually genetically identical, then this simply add an additional noise factor into the analysis proportional to the degree to which individuals aren't actually genetically identical. However, an additional complication is that the various strains are correlated to each other, in both genotype and phenotype. The strains are not produced by random mating, but are in fact, commonly closely related to each other. This is a hard problem, and this study doesn't try to address the issue.

In the best case, the STAR consortium and NBRP-Rat would measure the same strains. In the worst case there would be no overlap. Reality is somewhere in the middle. Between the STAR consortium's 466 strains, and the NBRP-Rat's 204 strains, there are 84 strains in common. However, of these, 33 are part of a set of rat recombinant inbred strains between the LE/Stm and F344 strains. The tight coupling between the genomes and phenomes of these strains made them unsuitable for inclusion in this project's analysis, and so they were excluded. This left 51 strains in common between STAR and NBRP-Rat.

Each combination of SNP and phenotype was examined in turn. For each combination of SNP and phenotype, there were 51 measurements of phenotype, one per strain. There were also 51 measurements of allele class, one per strain. Each SNP has previously been assigned a major allele, based on the genotype of the benchmark BN strain, and a minor allele, based on the most commonly occurring nucleotide other than the major allele. Only SNPs with two allelic forms were measured by the STAR consortium. This gives three allele classes that each strain can fall into: homozygous for the major allele, homozygous for the minor allele, and heterozygous. Although inbred strains are homozygous at most loci, a small fraction are heterozygous. This might happen, for example, if the minor allele encodes for a lethal recessive trait.

These 51 samples for each SNP and phenotype combination, divided into three allele classes, gives

fairly small sample sizes. At most each class will have 26 members. Considering this, I didn't want to assume that the phenotype measurements in each class are normally distributed. So, I used a non-parametric statistical test. I also used pair-wise comparisons between the three classes of alleles, rather than comparing all three distributions in a single test. That is, I compared the distributions of the phenotype measures of homozygous major vs homozygous minor, then homozygous major vs heterozygous, and finally heterozygous vs homozygous minor. Each two distributions were compared with a null hypothesis that the distributions were the same. To compare distributions, I used the Mann-Whitney-Wilcoxon (MWW) rank test. The MWW is very similar to performing an ordinary parametric two-sample t-test on the data after assigning a ranking to each measurement over the combined set of samples. Unlike other GWAS studies, I did not try to estimate the magnitude of the effect of the SNP on the phenotype measures, as it seemed unlikely to give reliable results from the sample size used.

When setting significance thresholds, correcting for the presence of multiple hypothesis testing is necessary. Starting with a target p-value cutoff of 0.05, and applying the simple but conservative Bonferroni correction gives up an adjusted p-value cutoff of 10^{-8}. At this cutoff value, I found zero significant results. This isn't too surprising given the small sample size. So, I decided to use a more lenient approach. I wanted to estimate the false discovery rate (FDR) of my method. So I ran the analysis twice. The second time, the phenotype measurements were randomly shuffled between the 51 strains while the allele classes remained the same. This retained the size and structure of the allele classes for each SNP. Setting a target q-value of around 0.02, a p-value cutoff of 10^{-5} gave 341 significant results from the first analysis and 5 false discoveries from the second analysis. Note however, that this shuffling also obscures the effects of the correlations in genotype and phenotype between the various strains, which might bias the results. No corrected for this possible bias was attempted.

4. Putting the Results in Context

After performing the genome-wide association, I was left with a list of 341 SNPs that are significantly statistically associated with variations in 42 phenotypes. This doesn't immediately seem extremely useful. Instead of focusing on the SNPs, I decided to focus on the phenotypes. What could make an organism have a bigger liver?

In an attempt to discern more meaning from this list, I turned to the Rat Genome Database, and the Gene Ontology annotations contained within. For each phenotype, I have a list of SNPs. For each of these SNPs, I can use the RGD to identify genes that occur near the location of the SNP. The RGD can be queried for all genes within a certain range of locations in the genome. For this study, I searched a range 10,000 basepairs upstream and downstream of each SNP. For comparison, the SNPs measured by the STAR consortium average 150,000 base pairs apart.

The genes in RGD have been given Gene Ontology (GO) annotations from a variety of sources. The query process above produced a list of genes, and a list of GO terms from those genes. By modeling the terms by the hypergeometric distribution, I found those GO terms most significantly enriched for each phenotype. The top terms are listed in Table 1, along with the SNPs associated with phenotypes. GO terms needed to occur at least twice, and have a p-value of less than 0.01 to be considered significant.

As a concrete example, one of the phenotypes is size of the testes in grams. One of the significant GO terms associated with this phenotype is GO:0007131 (reciprocal meiotic recombination) via SNPs near the gene Mre11a. Clearly a gene involved in meiosis might be relevant to the development of the testes. In fact, a recent study found that mRNA from this gene is expressed at high levels in testes tissue[7].

5. Conclusion

To summarize, I performed a GWAS on inbred rat strains. I merged phenotype data from one source with genotype data from another source, and was able to find statistically significant correlations between phenotypes and genotypes. I used these results to associate phenotypes with Gene Ontology terms, and to suggest possible biological mechanisms for complex and high-level phenotypes such as organ weight.

This study could be expanded in many ways. First and foremost, more samples would lead to higher confidence results. Only 51 rat strains were examined, but many more strains exist and could be measured. To validate the results of this study, note that rats are not the only species maintained as inbred strains. Mice would be an excellent candidate for a similar study. The degree of overlap between the two studies would provide a clear indicator of the level of confidence that should be ascribed to the results of each. Finally, promising SNPs could be followed up on with focused genotyping and research into their biological properties.

References

1. Hon-Chung Fung, *et al.* Genome-wide Genotyping in Parkinson's Disease and Neurologically Normal Controls: First Stage Analysis and Public Release of Data. *Lancet Neurology* **5**, 911-916 (2006).
2. Rampersaud, E. *et al.* Identification of Novel Candidate Genes for Type 2 Diabetes From a Genome-Wide Association Scan in the Old Order Amish. *Diabetes.* **56(12),** 3053-62 (2007).
3. Visscher, P.M. Sizing up human height variation. *Nat. Genet.* **40**, 489-490 (2008).
4. STAR Consortium. SNP and haplotype mapping for genetic analysis in the rat. *Nat. Genet.* **40**, 560–566 (2008).
5. Mashimo, T. *et al.* Rat Phenome Project: The untapped potential of existing rat strains. *J. Appl. Physiol.* **98,** 371-37 (2004).
6. Twigger, S. *et al.* Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res.* **30**, 125–128 (2002).
7. Lanson, N.A. *et al.* The MRE11-NBS1-RAD50 pathway is perturbed in SV40 large T antigen-immortalized AT-1, AT-2 and HL-1 cardiomyocytes. *Nucleic Acids Res.* **15(28)**, 2882-92 (2000).

Table 1: SNPs and GO Terms by Phenotype

Phenotype	3 Most Significant SNPs	Top 3 GO Terms
A/G	rat013_005_o19.p1ca_200 WKYOa35c04_s1_110 rat104_044_e22.q1ca_462	GO:0005859(muscle myosin complex) GO:0006939(smooth muscle contraction) GO:0003774(motor activity)
ALP (IU/L)	gko-14h12_rp2_b1_747 gko-77c18_fp2_b1_98 WKY-G-i-18b06_f1_831	GO:0030097(hemopoiesis) GO:0004361(glutaryl-CoA dehydrogenase activity) GO:0014731(spectrin-associated cytoskeleton)
Adrenals (mg)	rdahl-52a14_rp2_b1_409	
BW (g)	J476656 Cpn_10043983403 J599023	
Body Temperature (C)	J656718 J692319 gnl|ti|896517011_19866867024693_213	GO:0030141(secretory granule)
Body Weight 5 weeks (g)	WKYc48f05_r1_76	
Body Weight 6 weeks (g)	WKYc48f05_r1_76 J551516 Cpn_1159013781	GO:0005515(protein binding) GO:0005783(endoplasmic reticulum)
Brain (g%)	J476656 J551516 DahlSb01f09_r1_445	GO:0005783(endoplasmic reticulum) GO:0006468(protein amino acid phosphorylation) GO:0005515(protein binding)
Brain (g)	J669843 J650031 J659861	
CRE (mg/dL)	gko-48k22_rp2_b1_373	
Ca (mg/dL)	rat108_007_d14.q1ca_214 J562674 rat101_001_o15.q1ca_656	GO:0030018(Z disc) GO:0051017(actin filament bundle formation) GO:0051015(actin filament binding)
Cl/Body Weight(µEq/100g/6hrs)	WKYc92c01_s1_786 WKYc68c09_r1_206 J1282979	GO:0030169(low-density lipoprotein binding) GO:0016021(integral to membrane)
HDL-C (mg/dL)	J505540	GO:0004984(olfactory receptor activity) GO:0007186(G-protein coupled receptor protein signaling pathway)

Heart (g%)	rat106_017_k22.q1ca_520 J584831 J507345	GO:0004138(deoxyguanosine kinase activity) GO:0004421(hydroxymethylglutaryl-CoA synthase activity) GO:0008465(glycerate dehydrogenase activity) GO:0016618(hydroxypyruvate reductase activity)
Heart (g)	Cpn_1156642138 gko-107n14_rp2_b1_139 J502162	GO:0004984(olfactory receptor activity) GO:0007186(G-protein coupled receptor protein signaling pathway) GO:0008017(microtubule binding)
Kidneys (g%)	J1260586 J482894 rat106_017_k22.q1ca_520	GO:0004138(deoxyguanosine kinase activity) GO:0046070(dGTP metabolic process) GO:0001533(cornified envelope)
Kidneys (g)	Cpn_10043983403	
Liver (g%)	J567264	
Liver (g)	Cpn_10043983403 J476656 J551516	GO:0005515(protein binding) GO:0005783(endoplasmic reticulum)
Locomotor Activity (0-10 min)	gnl|ti|896879677_19866868593860_352 gnl|ti|896485062_19866867270516_284 gnl|ti|897024745_19866867480173_298	
Locomotor Activity (10-20 min)	J504572 J587168 rat102_005_m08.p1ca_127	GO:0042903(tubulin deacetylase activity) GO:0042826(histone deacetylase binding) GO:0006476(protein amino acid deacetylation)
Lung (g%)	WKYc70b07_s1_107 J700079 rat106_025_d22.q1ca_454	GO:0003883(CTP synthase activity) GO:0005776(autophagic vacuole) GO:0006606(protein import into nucleus)
Lung (g)	SHRSPc66c03_r1_275	
MCH (pg)	J562720 rat101_012_f21.p1ca_136	
MCV (fL)	rdahl-97l17_rp2_b1_86	
Na (mEq/L)	WKYc86a10_s1_493	
Na/Body Weight(µEq/100g/6hrs)	J1282979 WKYc92c01_s1_786 WKYc68c09_r1_206	GO:0016021(integral to membrane)
PT (seconds)	SHRSPc35a07_r1_685 J571830	
Plasma Cl (mEq/L)	SHRSPa27b04_s1_351	GO:0005622(intracellular) GO:0008270(zinc ion binding)

Platelets $x10^{-4}/\mu L$	J548847 WKYc08c02_s1_570 WKYc49g10_r1_749	GO:0000247(C-8 sterol isomerase activity) GO:0015385(sodium) GO:0045745(positive regulation of G-protein coupled receptor protein signaling pathway)
RBC $x10^{-4}/\mu L$	J700746	
Spleen (g)	Cpn_10043983403 J551516 J515738	GO:0030198(extracellular matrix organization) GO:0005737(cytoplasm) GO:0001501(skeletal system development)
T-BIL (mg/dL)	J505223 J563928 J474896	GO:0006306(DNA methylation) GO:0003886(DNA (cytosine-5-)-methyltransferase activity) GO:0006346(methylation-dependent chromatin silencing)
T-CHO (mg/dL)	rat101_005_o14.q1ca_181	
TP (g/dL)	rat104_044_e22.q1ca_462 J667481	
Testes (g%)	Cpn_10043983403 J661664 WKYc42e11_s1_216	GO:0006302(double-strand break repair) GO:0000075(cell cycle checkpoint) GO:0007131(reciprocal meiotic recombination) GO:0030870(Mre11 complex)
Total Locomotor Activity (0-30 min)	DS-g-a-13a01_f1_592 gko-106d3_fp2_b1_266 gnl|ti|896485062_19866867270516_284	GO:0042903(tubulin deacetylase activity) GO:0031032(actomyosin structure organization) GO:0042826(histone deacetylase binding)
UN (mg/dL)	rat102_030_m04.p1ca_307 gnl|ti|896761734_19866868058125_285 gko-78p9_rp2_b1_713	
Urine Volume (mL/6 hrs)	J667481	
WBC Lym. %	rdahl-64j5_fp2_b1_256 gnl|ti|896665172_19866866986781_292 J874898	
WBC Seg. %	SHRSPa68d11_r1_317 J645635 rat104_069_m20.q1ca_347	GO:0005681(spliceosome) GO:0043234(protein complex)

WBC $x10^{-2}/\mu L$	rdahl-74l5_fp2_b1_435 J683015 rdahl-86a10_rp2_b1_26	GO:0003858(3-hydroxybutyrate dehydrogenase activity) GO:0002752(cell surface pattern recognition receptor signaling pathway) GO:0042834(peptidoglycan binding) GO:0042892(chloramphenicol transport) GO:0051076(Gram-positive bacterial binding)