Proposal for Discovering SNPs from Eight Commonly Used Inbred Rat Strains

Tim Aitman[1], Richard Gibbs[2], George Weinstock[2], Norbert Huebner[3], Michael Jensen-Seaman[4,] Daniel Maloney[5] and Howard J. Jacob[5]

[1]Physiological Genomics and Medicine Group, MRC Clinical Sciences Centre, Imperial College Faculty of Medicine, Hammersmith Hospital, Ducane Road, London W12 0NN, UK. [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [3]Max-Delbruck-Center for Molecular Medicine (MDC), Robert-Rossle-Str. 10, 13092 Berlin-Buch, Germany. [4]Dept. of Biological Sciences, Mellon Hall, Duquesne University, 600 Forbes Ave., Pittsburgh PA 15282. [5]Human and Molecular Genetics Center, Medical College of Wisconsin, 8710 Watertown Plank Road, Milwaukee WI 53236.

For Correspondence:
Howard Jacob, Ph. D.
Director, Human and Molecular Genetics Center and
Warren Knowles Professor
Department of Physiology
Medical College of Wisconsin
8710 Watertown Plank Road
Milwaukee WI 53236
jacob@mcw.edu
Ph. (414) 456-4887
Fax (414) 456-6516

Table of Contents

Importance of the organism: The importance of the laboratory rat in biomedical research is well established. Since 1966, there have been on average over 28,000 publications per year using rat (PubMed search, key word: rat); in the last eight years (1996-2003) there have been on average almost 37,000 publications annually. The initiation of the rat genome project has yielded a tremendous wealth of genomic resources including genetic maps; radiation hybrid (RH) cell lines and the associated RH maps (over 6,000 genetic markers and 16,000 genes and ESTs mapped); cDNA libraries generating more than 593,880 ESTs (with more being generated) clustered into over 63,000 UniGenes; over 10,033 genetic markers; and a published draft (~6.8 X) sequence of the genome based on the inbred BN (Brown Norway) strain. The physiology of the "sequenced" rat is being explored by placing each chromosome of the BN on to two genomic backgrounds (the SS and FHH strains) and then measuring over 200 phenotypes in each strain. The consomic rats and the data are available from the PhysGen Programs in Genomic Applications (PGA) project. The rat genome resources have also been expanded by three new BAC libraries (the F344, SS, and FHH) via the BAC resource White Paper proposal. Over 700 strains of rats, including over 200 transgenic lines, numerous congenics, advanced intercross lines, and recombinant inbred panels exist and are being actively used to advance the annotation of the human genome. This single nucleotide polymorphisms (SNP) project is needed to advance these studies. The rat is primarily known as a physiological model and there has been some question over the years about the likelihood that the rat genomic tools would be fully utilized. Figures 1 and 2 show the change in usage of the Rat Genome Database website (measured by number of pages accessed) and changes in the number of visits (defined as 30 minutes using the site without more than 5 minutes of inactivity), over the period 2000-2004. Recall that an advanced draft of the rat genomic sequence, together with an analysis and annotation, was published in April 2004 (Rat Genome Sequencing Project Consortium, 2004).

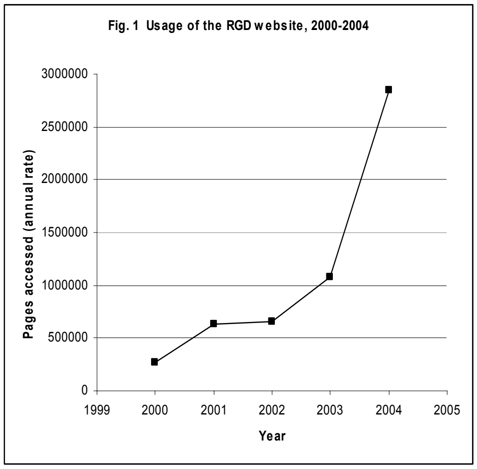
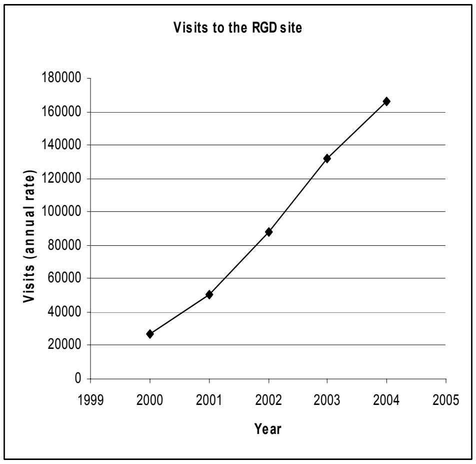


Fig. 1 Usage of the RGD website, 2000-2004


Visits to the RGD site

The data show an over 10-fold increase in usage and a concomitant increase in traffic, with an increase in the number of pages used per visit from 10 to 17. Thus, not only is usage of the site increasing impressively, but site visitors are making increasing use of the analysis tools available. Data for visits to our PGA website (http://pga.mcw.edu), average 4,000 per month, also document an increase of over 50% in both page accesses and visits over the period 2002 – 2004. Collectively these data strongly suggest that the rat community is not only coming and using the data—but that the community is rapidly growing!

One reason the rat is a dominant model in nutrition, pharmacology and physiology is the large numbers of strains (728) (http://rgd.mcw.edu) most of which were developed as models for complex, common diseases (Mary Shimoyama, RGD, personal communication). For example, strains are used in studies of transplantation and immunogenetics, cancer, cardiovascular diseases, behavior, growth and reproduction, various metabolic disorders, neurological and neuromuscular disorders, diseases of the skin and hair, aging, sleep apnea, pulmonary disease, nutrition, endocrinology, and toxicology. As described in more detail in the strain selection section, 708 quantitative trait loci (QTLs) have been identified in rat (there are 1729 QTLs in mouse) that contain alleles for many common complex diseases. Identification of these regions has been possible primarily because of the large number of genetic markers characterized in the 48 most commonly used strains. In addition, a dense radiation hybrid map consisting of ~25,000 mapped elements has enabled the construction of detailed comparative maps for the rat, mouse and human (http://rgd.mcw.edu/VCMAP) thereby enabling the QTL maps to link rat physiology to mouse and human genetics.

While these successes in mapping QTLs provide valuable genomic information, they also indicate a general need for improving tools to be used in mapping genes affecting complex traits. Typically, QTLs can currently be localized to specific genetic intervals of about 2 - 10 cM but further characterization is difficult. A critical need is therefore the development of SNPs and haplotype maps in order to reduce the size of these regions. The SNP discovery in this proposal is the first step towards a SNP-based haplotype map that will improve gene hunting via two general avenues. One is the advancement of correlations between phenotypic data and ancestral sequence origin across many existing inbred strains. This will immediately identify very short genomic regions most likely to harbor responsible genes. The other is the identification of segments that would be shared by rat strains used for simple intercross/backcross experiments.

The second area that represents a general need for improvement in genomic tools is the advancement of re-sequencing technologies to allow rapid gene targeted analyses. These efforts are ongoing outside the current proposal, and will complement the emerging generic marker resources.

Research community: In contrast to the mouse community, the rat community is distributed across a large number of disease areas, with the biology, not the genetics, being the common binding theme. A rat genomics community has developed over the last 6 – 7 years and hosts an annual international meeting: 'even' years are held in Europe or Asia and 'odd' years at Cold Spring Harbor Laboratories. Concurrently, as rat genomic and genetic tools have been developed, there has been an increase in the application of these tools in rat research. As mentioned above, the use of the RGD website has greatly increased in the last two years. Since 1991, when the first QTL was mapped in rat, there have been 434 papers (702 in mouse) published relating to quantitative genetic studies in the rat (PubMed search: keywords quantitative trait and rat), with nearly half being published in the last 3 years. These papers include investigations of the genetic basis of arthritis (where a QTL gene has been cloned by position), copper metabolism, pituitary tumor growth, aerobic capacity, blood pressure and hypertension, diabetes (where a QTL gene has been cloned by position), cardiovascular disease (where a QTL gene has been cloned by position), ethanol tolerance, behavioral conditioning, anxiety, fat accumulation and chemical carcinogenesis as examples (PubMed abstracts).

Similarly, a CRISP search on NIH funding using the key words “rat” and “QTL” identifies 32 NIH-funded grants (compared with 92 in the mouse) on topics including alcohol, hypertension, breast and pituitary cancer and autoimmune disease. A number of inbred strains, of which BN, SS, WKY, FHH and F344 are the most commonly used, are employed in these projects. Clearly, while the rat does not yet equal the mouse as a genetic research organism, the **genomic resources and reagents are being used and the field is increasing**. The combination of the genomic sequence and a preliminary SNP database will further enhance the tool kit for investigators using rats to study human disease.

Finally, this Whitepaper has been written by several investigators, who collectively represent a larger research community and who will use the SNPs. We will also be the developers of the haplotype maps, and the distributors of the data via RGD and EMBL and others. At the annual “Rat Meeting” held in Copenhagen, in September 2004, we held meetings with the major users in the community, and there was a broad consensus that development of SNPs and haplotype maps was the most critical need at this time.

Current SNP resources and discovery efforts. dbSNP contains 43,229 rat RefSNPs, of which only 729 are validated by actual functional assays (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). It is worth pointing out, however, that Dr. Huebner from Germany, a Co-author on this whitepaper, has been funded by the EU to develop 150,000 SNPs from cDNAs from four strains SHR, WKY, GK and SS. His team will then genotype these SNPs in 50 additional strains. He will submit an additional proposal to support the typing of SNPs generated in the project proposed here in the same strains. The genotyping of these strains will enable the SNPs generated here and in Europe to be validated as well as to enable the development of the haplotype map. Dr. Hubner’s group generated a rat SNP map based on cDNA from the SHRSP, BN, WKY and SD strains identifying 12,395 interstrain comparisons with a polymorphic site (Zimdahl et al. 2004): based on comparisons between the BN and other strains, the authors estimate the discovery rate at 1 SNP per 1100 base pairs, even with this limited comparison. This level is consistent with the sequencing work done in the Jacob lab, where they sequenced 1Kb, every 25Kb over 5Mb in 10 inbred strains and identified a SNP every 840bp.

Smits et al (2004) screened 55 genes in 96 strains, finding a total of 103 SNPs. Considering only intronic SNPs and grouping closely linked polymorphisms as a single SNP, they calculate a frequency of 1 SNP per 367 base pairs, reflecting the large number of strains examined. Extracting the genotype data for the eight strains proposed in this white paper, we calculated that we would miss about half of the SNPs if using only our eight strains rather than all 96, and that we could expect to find one SNP every 719 bp in intronic DNA. This is probably a conservative estimate for the expectation in intergenic noncoding DNA since this intronic data deliberately included intronic sequence immediately adjacent to exon sequence, which should be more conserved than typical noncoding DNA. The same logic applied to the UTR data from Smits et al. (2004) gives an expectation of one SNP, on average, every 527bp in UTRs in our eight strains.

Guryev et al. (2004) looked at a much larger data set, comparing public WGS, EST, and mRNA sequence data. They also resequenced in a handful of strains, including quite a few outbred Sprague-Dawleys. While it is difficult to extrapolate from the Gurvey et al data to an average SNP frequency, their data are useful in several respects. In particular, they found that most of the confirmed polymorphisms were "common" variants (two-thirds had a minor allele frequency of >20%), consistent with our reanalysis of the Smits et al data showing that you can identify most (or half) of the polymorphisms with a small subset of strains. Second, they found a high frequency of confirmed polymorphisms, with one SNP every 252bp in lab strains. This is a greater frequency than with the Smits et al. data because they included several outbred SDs.

Considering all these data, we can estimate the expected frequency of SNPs that we will find in our study at about 1 SNP in 500 – 600 bp. This level of variation is more than sufficient to help maximize the efficiency of a SNP discovery project via shot-gun sequencing across these eight strains.

A SNP HapMap. The most important question in this project is precisely how many SNPs are required to enable the aim of discovering important alleles. As with the human HapMap project at its inception, the data on SNP density and LD block size in rat are not sufficient to predict accurately how many SNPs we need to identify in order to obtain complete assurance of marking most LD blocks. Some guidance may be obtained from contrasting the human (outbred) and mouse (inbred) studies, however. Gabriel et al (2002) scanned autosomal regions totaling 13 Mb in 275 humans using 3,738 SNPs. They identified LD blocks ranging from 1 – 93 kb (average 9 kb) in Africans and African-Americans and from 2 – 186 kb (average 18 kb) in all others, suggesting about 175,000 (LD at 18kb) haplotype blocks in human. In the mouse, the regions are much longer. Frazer et al (2004) screened 4.6 Mb in 15 mouse strains, identifying 18,366 SNPs that defined a set of 50 LD blocks ranging from 12 to 608 kb in length (average 92 kb) or about 25,000 LD blocks in the laboratory mouse genome. Conservatively, and based on preliminary sequencing results from the Jacob lab, we expect the LD blocks in the rat to be intermediate in size between human and mouse (and probably closer to the mouse): ~70,000 blocks with an average size of 40 kb. Therefore, identifying minimally 280,000 rat SNPs (4 SNPs per block) should provide sufficient information to mark most LD blocks in any given cross.

Required number of SNPs. Under the assumption that we will have to cover 70,000 haplotype blocks, and that we will want 4 SNPs per haplotype block to provide maximum coverage and flexibility for the users, we require 280,000 SNPs be developed. We estimate that the use of the 8 strains will result in an average one SNP per 640 bp, or about 1 per sequencing read. Therefore, there will need to be a total of 280,000 reads per strain for a total of 2,240,000 reads for this project. The strains we have selected are designed to be maximally informative, as well as being most likely to facilitate positional cloning. Finally, it is worth noting that the EU has funded the STAR program (Dr. Huebner, PI), which has the goal of generating an additional 150,000 SNPs, primarily from cDNA. Cumulatively, this project and the EU project plus existing SNPs would yield a total of 485,624 SNPs. We anticipate this level of coverage would meet the genetic mapping needs. To facilitate the positional cloning projects in rats and to help maximize annotating the genome with function we have selected eight strains that cover the "evolutionary" range of the known genetic variation in the lab rat, which in addition to maximizing diversity, also cover the strains where large numbers of QTLs have been mapped and other genetic tools such as congenics, consomics, and recombinant inbred strains exist.

Strain Selection. We selected the eight strains based on three criteria. First, the strains cover the maximal degree of known genetic variation. Second, the strains selected carry significant resources and data to warrant having SNP discovery within them, as these strains will directly benefit from this program. Third, the data were not only complimentary but could be integrated with the EU project. The rat genome project from its inception has carefully selected strains based on providing maximal utility to the global community. For example, the BN was selected as the strain to sequence as it was the most genetically distinct, thereby maximizing the chances of finding SNPs and genetic markers when compared to the other strains (Figure 3). The combination of the large set of inbred disease models with the genomic resources described has encouraged researchers to attack the analysis of multigenic diseases by identifying QTLs when many gene variants appear to contribute to the final disease state.

Strain	No. QTLs
BN	216
SHR/SHRSP	210
SS	155
SD	136
F344	112
WKY	94
BB	87
DA	74
LEW	76
OLETF	58
ACI	20
GK	15
BUF	4
FHH	4
PVG	1

Table 1. Shows the QTLs per strain for strains considered for this project. The strains considered were based on their positions in figure 3. Note QTLs are mapped using two strains so the numbers do not add up to 708 QTLs.

The Rat Genome Database contains records of some 708 QTLs reported in the literature (Mary Shimoyama, RGD, personal communication). Table 1 shows how many of these loci were identified in crosses involving the most commonly used strains (and their derivatives). Clearly, the five most commonly used strains account for the majority of the QTLs.

Based on Figure 3 (reproduced from Thomas et al (2003)) we selected the strains in Table 1. We then evaluated each of these strains to determine what they have to offer in addition to their position in the phylogenetic tree and how they compared to the BN, which is the strain that was sequenced, in part because it is the outlier strain. In principle, cataloguing SNPs present in these strains will identify a significant portion of the known genetic variation in rat. Here we justify each strain from the top of Figure 3 down.

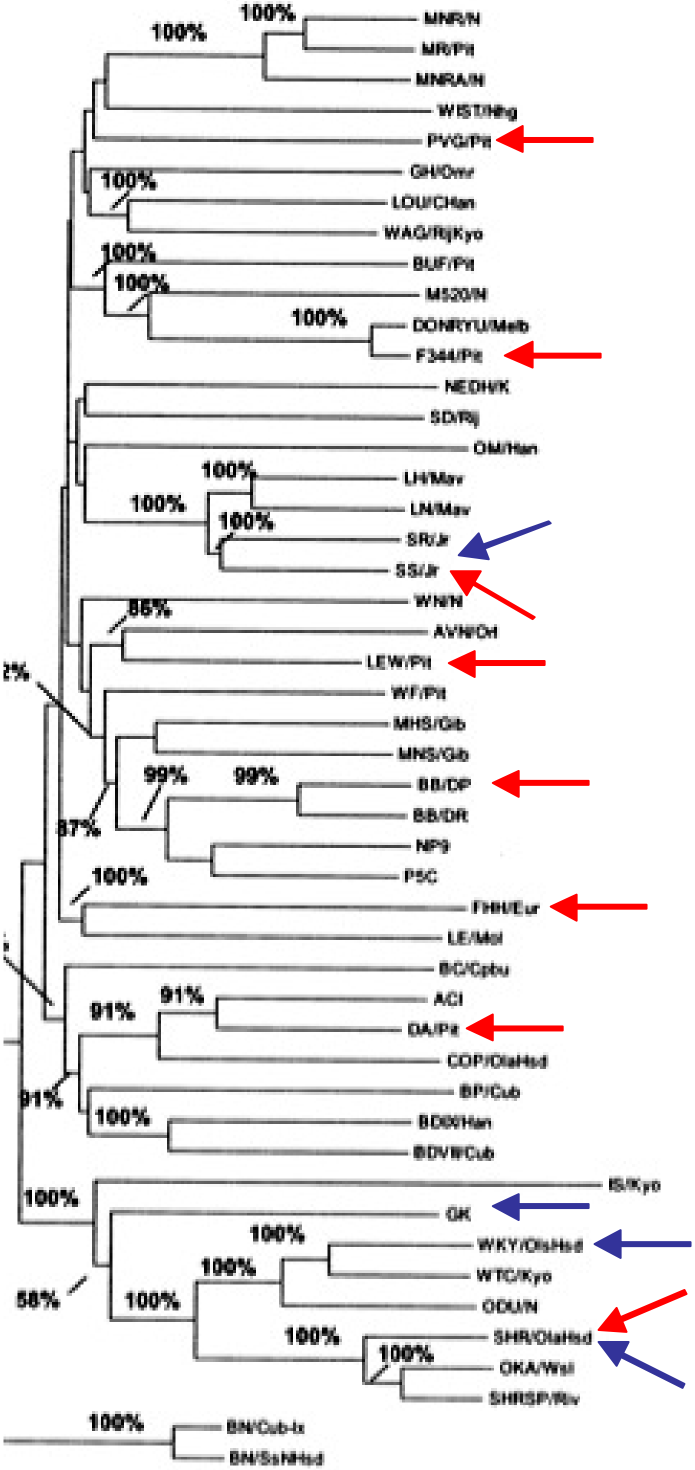


Figure 3. Phylogenetic tree of inbred rat strains based on data from 4256 microsatellite loci, reproduced from Thomas et al. 2003. Red arrows point to the eight strains proposed for the SNP study herein. Blue arrows point the four strains to be used bv the EU

The PVG (1 QTL, 669 papers) was developed in the 1940s and inbred at Glaxo. Under normal conditions it is perfectly healthy and has been used as a control in studies of alloreactivity and transplantation, induced arthritis and glomerulosclerosis. The rats have shown increased expression of anxiety related genes when exposed to cats. It appears only distantly related to another strain in its clade, the MNR.

The F344 rat (112 QTLs, 23,944 papers) is the most widely used rat in toxicological and pharmacological studies. The F344 was developed in 1920 by Curtiss and Dunning. Studies of the F344 and its substrains involve traits as disparate as inhalation toxicology, induced carcinoma, numerous autoimmune diseases, pain, maternal behavior, locomotor behavior and dietary effects. There is also a BAC library for this strain.

The SS rat (155 QTLs, 1,200 papers) has enormous genetic infrastructure with the development of the consomic rat strains, where one chromosome at a time from the sequenced BN strain has been introgressed onto the SS genome background. The consomics have been extensively phenotyped with more than 200 traits in each of the 22 consomic lines (http://physgen.mcw.edu). In addition there is a BAC library and this strain is also to be used by the EU project, enabling us to have some quality control measures between the EU and US project. Traits studied include the cardiovascular traits NO synthase, hypertension, albuminuria, infarct size and hyperlipidemia, congestive heart failure, as well as alcohol consumption and insulin resistance. The strain was develop in the late 1960s by Dahl and was then inbred by John Rapp in the 1970s.

The LEW rat (76 QTLs, 16,603 publications) is also a common model system used in the pharmaceutical industry as well as academia. The large number of QTLs and publications warrants its inclusion in the group of eight rats. It was developed by Lewis from a Wistar stock in the 1940s and 1950s. Lewis rats are susceptible to allergic encephalomyelitis and inflammatory disease due to deficient counter-regulation of immune responsiveness. It is commonly used in transplant studies and studies of immune responsiveness. The LEW rat appears to be a model of human Lyme disease. It has been used as the inbred parent for a number of MHC congenic lines.

The BB rat (87 QTLs in BB, BB/DR and BB/DP, 2,792 publications) was developed in the early 1970s by inbreeding a colony of Wistar rats. It has been the subject of studies on diabetes, peripheral neuropathy, collagen-induced arthritis and lymphopenia including many studies of islet transplantation and insulin treatment. The BB represents a large clade of laboratory rat strains in this proposal and fills it in with the LEW at the other end.

The FHH (4 QTLs, 53 publications) while not used for too many QTL studies to date, already possess a significant infrastructural investment including the consomic panels described above and a large, publicly available dataset of physiological measurements related to cardiovascular function, a BAC library and a large scale ENU mutagenesis program underway in this strain. Historically it derives from the FH outbred rat in the 1980s, but the FH rat itself comes from 3 different strains in Europe that are within this clade. As we show in the table above, numerous QTLs have been mapped in crosses between the FHH, SS and BN strains so the SNP data could immediately be applied to QTL fine mapping and gene discovery.

The DA strain (74 QTLs, 193 articles) was developed in the 1960s from an unknown outbred strain. The rats develop arthritis, urinary bladder tumors, induced tongue carcinomas and hormone-dependent endometrial adenocarcinomas at high frequency. They have been used in studies of drug metabolism, anxiety-related behavior, autoimmune encephalitis and transplantation. It is the only strain in its clade to appear in this request.

The SHR (146 QTLs, 12,576 publications) was created in Japan from the outbred Wistar Kyoto strain and released in 1963. It is widely used in studies of hypertension. Extensive work has been done on organ and tissue specific changes during development of hypertension as well as on intervention strategies. The rat is used as a model for childhood hyperactivity and ADHD and is used in drug testing both for hypertension and ADHD as well as for a number of unrelated conditions.

Two other candidate strains the WKY and the GK were not selected for SNP discovery in this proposal, although both are important strains, but are too close to the SHR, as well as because they are included in the EU proposal. As a result of the EU proposal, and the genotyping of the SNPs discovered here being typed in these two strains, they will have more than adequate coverage across the genome.

Summary. This proposal seeks to generate 280,000 SNPs using eight carefully selected strains. This project was assembled by a group of international investigators, including the Baylor sequencing center. As with all other rat genome projects it has been designed to dovetail into existing infrastructure and to expand the field as a whole. In the case of this proposal we are fortunate that the EU will be making a significant contribution to the effort by providing an additional 150,000 SNPs and the genotypes that will convert this SNP discovery program into a SNP HapMap without additional NIH funds. Given the success the sequencing has had on the rat field and the need to rapidly annotate the human genome with complex disease biology, this proposal is timely and falls squarely within the mission of the NHGRI.

Literature cited

Aitman,T.J. *et al.* 1999. Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. Nat Genet 21:76-83.

Frazer KA, Wade CM, Hinds DA, Patil N, Cox DR, Daly MJ. 2004. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. Genome Res 14(8):1493-500.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. 2002. Science 296(5576):2225-9.

Guryev V, Berezikov E, Malik R, Plasterk RHA, Cuppen E. 2004. Single nucleotide polymorphisms associated with rat expressed sequences. Genome Research 14:1438-1443.

Marion E, Kaisaki PJ, Pouillon V, Gueydan C, Levy JC, Bodson A, Krzentowski G, Daubresse JC, Mockel J, Behrends J, Servais G, Szpirer C, Kruys V, Gauguier D, Schurmans S. 2002. The gene INPPL1, encoding the lipid phosphatase SHIP2, is a candidate for type 2 diabetes in rat and man. Diabetes 51:2012-2017.

Olofsson P, Holmberg J, Tordsson J, Lu S, Akerstrom B, Holmdahl R.2003. Positional identification of Ncf1 as a gene that regulates arthritis severity in rats. Nat Genet 33:25-32.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428(6982):493-521.

Seda O, Sedova L. 2003. New apolipoprotein A-V: comparative genomics meets metabolism. Physiol Res 52(2):141-6.

Smits BMG, van Zutphen BFM, Plasterk RHA, Cuppen E (2004) Genetic variation in coding regions between and within commonly used inbred rat strains. Genome Research 14:1285-1290.

Thomas MA, Chen C-F, Jensen-Seaman MI, Tonellato PJ, Twigger SN. 2003. Phylogenetics of rat inbred strains. Mamm Genome 14(1):61-4.

Wong AH, Macciardi F, Klempan T, Kawczynski W, Barr CL, Lakatoo S, Wong M, Buckle C, Trakalo J, Boffa E, Oak J, Azevedo MH, Dourado A, Coelho I, Macedo A, Vicente A, Valente J, Ferreira CP, Pato MT, Pato CN, Kennedy JL, Van Tol HH. 2003. Identification of candidate genes for psychosis in rat models, and possible association between schizophrenia and the 14-3-3eta gene. Mol Psychiatry. 8(2):156-66.

Zimdahl H, Nyakatura G, Brandt P, Schulz H, Hummel O, Fartmann B, Brett D , Droege M, Monti J, Lee Y-A, Sun Y, Zhao S,[4] Winter EE, Ponting CP , Chen Y,[6] Kasprzyk A , Birney E , Ganten D, Hubner N. A SNP Map of the Rat Genome Generated from cDNA Sequences. *Science* 303(5659), 807 , 6 February 2004.