nature genetics

SNP and haplotype mapping for genetic analysis in the rat

The STAR Consortium*

The laboratory rat is one of the most extensively studied model organisms. Inbred laboratory rat strains originated from limited *Rattus norvegicus* founder populations, and the inherited genetic variation provides an excellent resource for the correlation of genotype to phenotype. Here, we report a survey of genetic variation based on almost 3 million newly identified SNPs. We obtained accurate and complete genotypes for a subset of 20,238 SNPs across 167 distinct inbred rat strains, two rat recombinant inbred panels and an F_2 intercross. Using 81% of these SNPs, we constructed high-density genetic maps, creating a large dataset of fully characterized SNPs for disease gene mapping. Our data characterize the population structure and illustrate the degree of linkage disequilibrium. We provide a detailed SNP map and demonstrate its utility for mapping of quantitative trait loci. This community resource is openly available and augments the genetic tools for this workhorse of physiological studies.

The unique power of the laboratory rat resides in the extensive biological characterization of a wide range of inbred strains representing models for common human diseases (see URL 1 in URLs section of Methods below). Although the rat is primarily known as a physiological model, there has been a steady increase in the use of the rat in genetic and genomic studies over the last decade[1]. The genome of the Brown Norway rat (BN/NHsdMcwi) has been sequenced, but, as a sequence of a single inbred rat strain, it provided little insight into the genetic variation that is responsible for the wide range of disease phenotypes, drug resistance or variability in toxicology responses in the different rat strains. At present, genetic variability in the rat genome is usually assayed using a limited set of microsatellite markers[1–4]. A dense set of polymorphic markers, for which SNPs provide the most cost-effective solution, would transform the genetic toolkit available for rat biologists.

The breeding history of rat strains, in common with that of many other laboratory animals, is known to have a complex genesis[5], with a number of unknown relationships in the formation of the laboratory strains. In addition, strains often carry the same designation but substrains are not necessarily identical because in a number of cases breeding stocks were distributed before the line became inbred, with varying physiological consequences[6,7]. Thus it is important to have detailed marker information available on any specific substrain. The presence of a number of recombinant inbred lines (RI)—in particular the HXB-BXH sets, derived from the BN-Lx/Cub and SHR/Ola parental strains and the FXLE-LEXF sets, derived from the F344/Stm and the LE/Stm strains—and the presence of both congenics and consomics provides a rich set of renewable genetic resources available for rat biologists to examine the variation of phenotypes between different genotypes.

To study genome-wide genetic variation, we initiated the genetic dissection of the ancestral segments making up the most commonly used rat inbred lines, and we developed a comprehensive open resource of validated SNP markers for essentially any strain combination. We provide extensive maps of strain distribution patterns (SDPs) for the two largest rat recombinant inbred (RI) strains, estimations on linkage disequilibrium, and haplotype structure in the rat genome and evaluate the use of correlation between phenotype and ancestral sequence origin across many inbred strains required to facilitate the identification of underlying alleles. This study provides a set of permanent resources for rat genetics (SNPs, SDPs and genetic maps), immediately facilitates more statistically powerful analysis on the RI strains and provides insight in the genesis of the different rat strains available to researchers today.

RESULTS

Polymorphisms in the rat genome and generation of a SNP map

We generated a SNP map of the rat genome containing about 3 million distinct SNPs mapped to the draft genome sequence, at an average density of approximately one SNP per 800 bps. Three distinct sources of DNA sequencing reads were used for automated computational SNP discovery (**Table 1**): (i) shotgun sequence generated from the four strains SS/Jr, WKY/Bbb, GK/Ox and SHRSP/Bbb, all widely used in disease mapping experiments in crosses and congenic strains[8]; (ii) genome-wide shotgun sequence at ×1.5 coverage of the outbred Sprague-Dawley rat generated by Celera; and (iii) BAC end sequences from the F344/Stm rat, which is also widely used in genetic studies. For the Sprague-Dawley rat, approximately 14% of the SNPs were heterozygous on the basis of evidence from overlapping aligned sequencing reads. Because of the relatively low coverage, this is an underestimate of the heterozygosity in this strain. As expected by the relatively low sequencing coverage, we observed that homozygous SNPs were supported on average by fewer aligned reads than

*The complete lists of participants and affiliations appear at the end of the article. Correspondence should be addressed to N.H. (nhuebner@mdc-berlin.de).

Received 30 October 2007; accepted 27 February 2008; published online 28 April 2008; doi:10.1038/ng.124

Table 1 Number of sequencing reads and detected SNPs per strain

Source	Number of reads	Non-redundant SNPs
STAR shotgun sequence	249,525	128,976 (SS/Jr: 56,639)[a] (WKY/Bbb: 32,601)[a] (GK/Ox: 16,838)[a] (SHRSP/Bbb: 28,332)[a]
Celera Sprague-Dawley genome-wide shotgun sequence	7,990,225	2,650,525
F344 BAC end sequence	344,064	196,812
Total		2,976,313

[a]Number of SNPs before removal of redundancy.

heterozygous SNPs, which leads to a systematic underestimation of the true heterozygosity in this strain.

To assess the utility of the SNPs for rat genetics we genotyped a subset ($n = 20{,}283$) in 167 inbred strains and two RI panels of 31 and 33 strains and tested a subset ($n = 9{,}691$) in 89 F_2 animals. The allele frequencies of SNPs across the inbred strains were approximately evenly distributed between 3% and 50%. We assayed seven 1,536-plex assays that could be run on the Illumina BeadLab station (10,752 SNPs). In parallel, we developed a 9,691-SNP rat targeted genotyping panel (called the '10K panel') to be run on an Affymetrix platform (see **Supplementary Methods** online). The rationale behind this separation was to evaluate the appropriate technology for efficient SNP genotyping in the rat. These genotyping tools now are commercially available from Illumina and Affymetrix, respectively. The two panels together yielded 20,283 validated SNPs. We genotyped 1,057 SNPs with both platforms in 231 rats as a control, with a concordance of 99.8%. Furthermore, we constructed separate phylogenetic trees using data from each genotyping platform and found that the clustering was very stable, even for those nodes that are not supported by a high bootstrap value (**Supplementary Fig. 1** online).

Annotation of putative functional SNPs from inbred strains

We predicted the functional effects of 325,788 SNPs. This analysis was restricted to SNPs derived from the five inbred strains used in the SNP discovery (SS/Jr, GK/Ox, SHRSP/Bbb, WKY/Bbb and F344/Stm). We excluded the large set of SNPs identified from the outbred Sprague-Dawley rat. This ensures that all predicted functional consequences can be tested experimentally in stable inbred strains, eliminating the uncertainty that a particular animal may not carry the described allele because of incomplete inbreeding of the colony. We estimated the selective pressure on amino acid replacement mutation for all residues with nonsynonymous coding SNPs ($n = 1{,}160$) by calculating the omega value, or ratio of nonsynonymous/synonymous substitution rates[9]. We identified 56 SNPs with omega values lower than 0.1 (**Supplementary Table 1** online), which is indicative of a likely effect in protein function; seven of these lie in rat orthologs of human disease genes involved in hereditary diseases or cancer. These include the gene *ALDH2*, involved in acute alcohol intolerance, the gene *PCCB*, involved in propionic acidemia, the cancer gene *AFF4* (AF4/FMR2 family member 4) and the proto-oncogene tyrosine kinase receptor ret precursor (*RET*).

Further, 324 SNPs are predicted to disrupt the normal pattern of splicing, and 57 SNPs and 63 SNPs, respectively, create a potential donor splice site (GT) or potential acceptor splice site (AT). Finally, we assessed the potential effect of the SNPs on transcriptional and post-transcriptional regulation. One thousand nineteen SNPs in promoter regions map into conserved transcription factor binding sites and 568 SNPs map into DNA triplexes[10]. One hundred thirty-two SNPs in 3′ UTR regions affect microRNA targets (**Supplementary Table 1**).

Phylogenetic relationships among rat strains

It is unclear from documented information how ancestral subspecies, strains and individual rats have contributed to shaping the genomes of the modern laboratory rat strains[11]. Considering the many rounds of inbreeding and interbreeding that most likely took place, a complex evolutionary history can be expected. Therefore, we chose to visualize the interstrain relationship and genetic proximity in a phylogenetic network[12] rather than a tree (**Fig. 1**). The observed strain relationships agreed very well with the known history of rat strains and supported all significant clusters of strains previously recognized in analysis of microsatellite markers[5]. The reticulation around the center of the network may reflect extensive genetic heterogeneity of the ancestral rat population. The network revealed a complex breeding history of WKY-related rat strains, corroborating the notion that breeding stocks of several strains (WKY, SHR, SHRSP) were distributed before the line became inbred[6,7]. The Brown Norway (BN) rats were placed as the most diverged strain, which might be explained by the SNP ascertainment bias, because the BN genome sequence was used as a reference sequence for SNP discovery. However, the separation of BN is also supported by microsatellite data[5,13] and a study that used different SNP panels[14]. The inclusion of wild and outbred rats into the phylogenetic

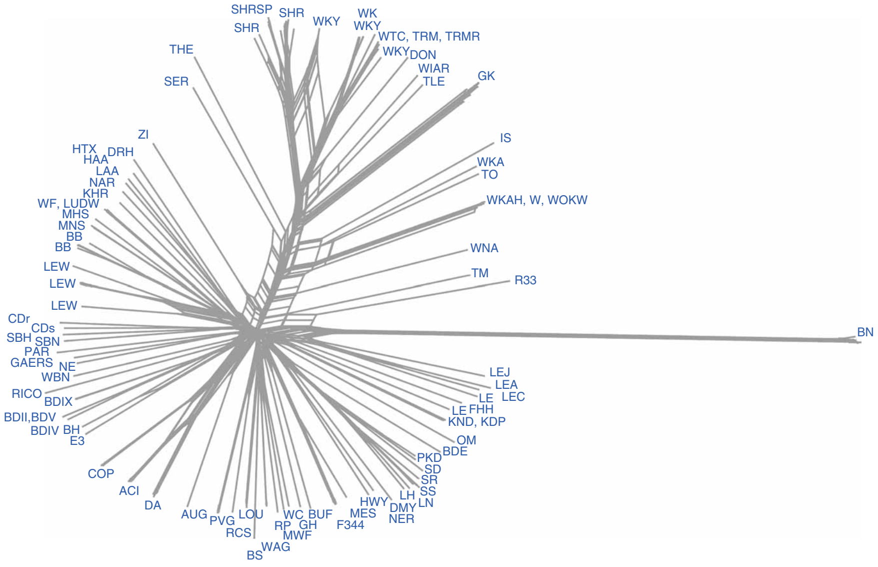

Figure 1 Phylogenetic neighbor-net network constructed from 20,283 polymorphic positions genotyped in 167 laboratory rats. Reticulation in the center of the network likely reflects genetic heterogeneity of the ancestral rat population. The group of WKY-related strains shows a complex pattern of relationships, due probably to incomplete inbreeding of stocks before they were disseminated to various laboratories and subsequently inbred to completion. The network also shows presence of residual inter-isolate variation within SHR, LEW, BB, WKY, LE, GK and BN inbred strains.

analysis resulted in their branching from the reticulate center and not outgrouping of BN or non-BN rats. Within non-BN strains, we defined ten clusters of strains that evidently shared breeding history; these clusters were highly supported as monophyletic groups on a traditional phylogenetic tree (**Supplementary Fig. 1**).

An important issue for studies that involve inbred rats is their degree of inbreeding and the variation between substrains that are maintained at different laboratories. Previously it has been shown that BN and DA inbred substrains obtained from different locations harbor genetic differences[14]. Our dataset included multiple substrain samples and confirmed the presence of variation between substrains. This effect was most pronounced in LE (29% of genotyped variable sites were different in the pairwise substrain comparison), WKY (up to 19%), LEW (13%), SHR (11%), BB (10%), PKD (5%) and, to a lesser degree, in GK (1%) and BN (0.6%) inbred strains. These observations indicate that, at least for some inbred strains, the use of different substrains may have a significant effect on the outcome and reproducibility of experimental results.

Estimations on linkage disequilibrium and haplotype structure in the rat genome

Although the genotyped panel of markers is not dense enough to support conclusive evaluation of linkage disequilibrium (LD) structure and a complete haplotype map, it can be used for obtaining the estimates on haplotype length and its comparison with that of other organisms. Using 15,901 SNPs with minor allele frequency of $>5\%$ among all genotyped samples, we defined the haplotype blocks as adjacent SNPs lacking historical recombination[15]. Using Haploview[16], a total of 837 blocks were detected, encompassing 19% of SNPs and covering about 12% of the rat genome, with an average block size of 411 kb. These included 323 blocks of only two SNPs, but these blocks were relatively small and in total covered less than 2 Mb. In contrast, the average size of the 514 blocks that contained three or more SNPs was 665 kb. The observed block structure completely disappeared upon permutation of SNP positions and was relatively stable when the number of substrains or density of markers were randomly reduced by 10% (**Supplementary Fig. 2** online). If we extrapolate from the current data, we estimate that it will require another 50,000 to 75,000 SNPs to define the remaining haplotype structure comprehensively.

We compared rat haplotype structure with unpublished mouse haplotype data from the Broad Institute (see URL 2 below). We balanced the mouse and rat sets to contain the same number of strains ($n = 38$), SNP density (6.4 kb^{-1}) and inter-SNP distance distribution. Under the same criteria for haplotype block partitioning, the extent of LD was larger in laboratory mice, where haplotype blocks covered a larger fraction of the genome (35% compared to 12% in the rat), contained a higher proportion of informative markers (56% versus 21%) and had a greater average size (648 kb versus 388 kb). The direct comparison of LD decay profiles in rat and mouse (**Supplementary Fig. 3** online) further substantiated this notion. On the other hand, linkage disequilibrium in the rat was larger than that in cow, for which haplotype blocks cover only 2.2% of autosomes[17]. Although LD in rats was less pronounced than that in mice, it still extended over hundreds of kilobases, unlike LD in humans or across dog breeds, for which the correlation coefficient r^2 drops below 0.1 at 100 kb[18–20]. These results suggest that the breeding histories of laboratory rats and mice are qualitatively different. However, there are also considerable differences in extent of LD and complexity of phylogenetic relationships when rat and mouse laboratory populations are compared. In the mouse, large LD blocks can be recognized that reflect ancestral contributions from different subspecies[17–19], whereas there is no such evidence in the rat. At the same time, the phylogenetic relationships among groups of rats are hard to deduce (**Supplementary Fig. 1**), reflecting more divergent genetic background of a rat founder population. Comparison of LD between rat and mouse showed that the size of haplotype blocks in syntenic regions showed small, but significant, correlation ($r^2 = 0.18$), consistent with the previously observed correlation between murine (mouse and rat) recombination rate and fine LD structure[21,22]. Olfactory genes were the only gene class overrepresented in rat LD blocks ($P < 10^{-6}$). Twenty-one distinct gene clusters harboring about one-third ($n = 325$) of all olfactory genes were located in LD blocks longer than 500 kb. The same phenomenon was observed in mice (130 genes, $P < 10^{-7}$ for blocks exceeding 2 Mb), suggesting an increased selective pressure on rodent genes involved in sensory perception of smell.

Notably, we identified 939 interchromosomal SNP pairs in full linkage disequilibrium. These SNPs were heavily shifted toward low minor-allele frequencies, with only 38 and 4 of them having minor-allele frequencies larger than 0.1 and 0.15, respectively. More detailed inspection revealed that, besides being rare, the vast majority of these variants were private to branches of the phylogenetic tree (for example, many of them were restricted to the SHR + WKY + GK cluster). Thus, perfectly correlated SNPs on different chromosomes are unlikely to result from epistatic effects or genome assembly errors, but are more likely to represent a shared physical genomic structure (that is, genetic background). It should be mentioned that imperfect but significant pairwise correlation ($r^2 \geq 0.5$) was observed among about 0.2% of the interchromosomal SNP pairs. The highly correlated subset disappeared almost completely when SNP alleles were randomized and could thus reflect epistatic interactions as well as ancestral relationships.

Prospects for genome-wide association mapping using inbred rat strains

One hundred of the most diverse inbred rat strains were evaluated by simulation for their potential for genome-wide association mapping of quantitative trait loci (QTL). The method originated in the mouse genetics community[23], where it has generated much discussion[24]. From our simulations, we found that the threshold for genome-wide significance varied depending on the extent and nature of the genetic component of the phenotypic variance. For example, the genome-wide threshold for significance when there was no genetic component to the phenotypic variance was $\log P_{\max} = 4.1$, close to the Bonferroni estimate of 4.3. In contrast, for an infinitesimal model in which many small-effect QTL, in total accounting for 50% of the total variance, were distributed uniformly across the genome, there was a median $\log P_{\max} = 21.5$; that is, much higher. For a single major QTL explaining 50% of the variance, the genome-wide maximum coincided with the true position in 31% of simulations, and there was a local maximum exceeding the genome-wide null threshold of significance at the true QTL in 51% of simulations and within 1 Mb of the true position in 91% of cases. However, there was a median $\log P_{\max} = 14.3$, which lies between the two thresholds above, and, on average, 72 putative QTL exceeded the null threshold of 4.1. Finally, for a realistic complex trait scenario of ten 5% QTL, there was a median $\log P_{\max} = 9.96$, and the median number of putative QTL exceeding 4.1 was 1,412. Thus, there were a very large number of false positive QTL, and consequently each true QTL was close to a putative QTL. **Supplementary Table 2** online gives the numbers of putative QTL identified at different thresholds and illustrates the problem of balancing true and false positive rates. For example, a threshold of 8 limits the number of putative QTL to only twice the number of true QTL, but the majority of putative QTL locations do not coincide with the true locations at this threshold.

Construction of a rat genetic map using an F_2 cross and recombinant inbred lines

We typed a total of 20,283 SNPs in two independent panels of recombinant inbred (RI) strains derived from SHR and BN-Lx rats (HXB-BXH) ($n = 31$) and from F344/Stm and LE/Stm rats (FXLE-LEXF) ($n = 33$), and we typed 9,691 SNPs in 89 progeny of an F_2 cross between BN/Par and GK/Ox rats (GK × BN). These populations have been used for mapping complex phenotypes for metabolic syndrome[25], expression QTL (eQTL)[26] and metabonomic traits[27], and extensive phenotype characterizations as part of the Japanese phenome project (see URL 3 below and ref. 28). In addition, genotype analysis in F_2 rats enabled assessment of the reliability of heterozygous genotype calls.

Genetic map construction was initially carried out in the GK × BN F_2 cross with the JoinMap program as previously described[2]. This approach has the advantage that prior knowledge of markers' physical order is not required for calculating genetic distances. More than 8,400 microsatellite and SNP markers have now been mapped in this cross and SNP typing has significantly improved the resolution of the genetic maps (**Supplementary Table 3** online). Following genotype verifications, we confirmed the existence of strong distortion of segregation previously reported in chromosomes 3, 4, 9 and 13 (ref. 2). Alignment of genetic and physical maps showed the general agreement of marker order and distance (**Supplementary Fig. 4** online). However, we identified regions (from 1 SNP to 8 Mb) where SNP-based genetic maps were inconsistent with the current rat genome assembly draft.

We then repeated genetic mapping in this cross and both panels of RI strains using the R and R/QTL software packages[29,30] integrating SNP genotype and physical map data, resulting in 16,543 SNPs mapped. Data were initially filtered to remove markers containing genotyping errors (for example, absence of segregation in the cohort despite apparent allele variation in the parental strains) and blocks of adjacent SNPs with identical segregation patterns were collapsed into SDPs. Markers that generated suspiciously large map distances were removed, using criteria derived from the approximately linear relationship of genetic and physical distances. Details of the typed markers and mapped positions are given in **Supplementary Table 4** online and URL 4 below and the resulting maps in **Supplementary Figure 5** online**.** We found strong evidence of discrepancies between the genetic map and the draft genome assembly (**Fig. 2**). In particular, genetic mapping in all three panels identified a p11-centromeric segment of chromosome 1 that has been wrongly assembled in the p14-telomeric region of chromosome 17. Genetic mapping data suggested further intra- and interchromosomal relocations in regions of chromosomes 2, 4, 11, 12, 14, 17. Known conflicts between rat genome assemblies, provided by BCM and Celera (see URL 5 below), indicated the relocation in the p14 region of chromosome 17, supporting the Celera assembly, and one conflict on chromosome 9 was resolved favoring the BCM assembly (data not shown). The other conflicting mapping results require further independent verifications.

When we set out to construct a genetic map for the X chromosome based on the physical order of markers, we detected several unlinked markers, which rendered the mapping impossible. In-depth investigation of these linkage breaks revealed that they occurred on contig boundaries (**Supplementary Table 5** online). We rearranged the fragments of the chromosome resulting from splitting the contigs that were not linked (lod score < 2) in the order that generated the smallest average recombination fraction in the three populations (**Supplementary Fig. 6a** online). Using the resulting marker positions, we constructed three genetic maps (**Supplementary Fig. 6b**), summarized in **Supplementary Table 4**.

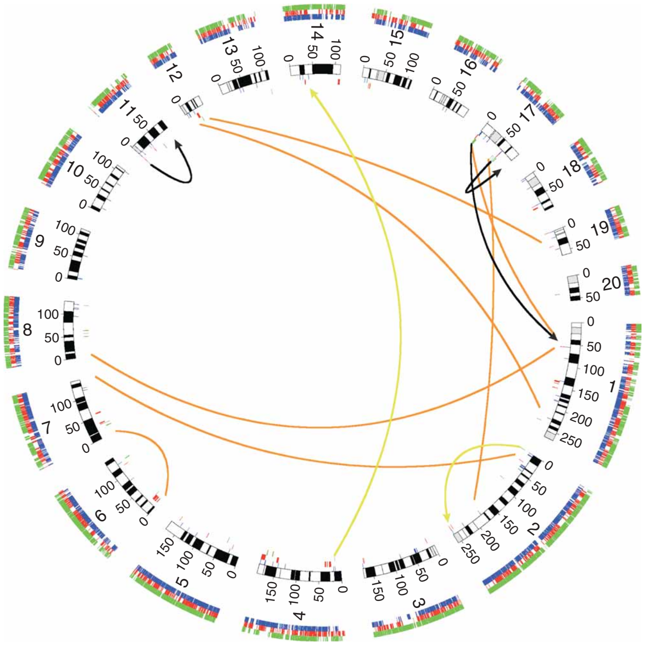

Figure 2 Identified discrepancies between rat genome assembly and genetic maps. Rearrangement of the physical map according to genetic mapping information. Data from each cohort is color coded (red, FXLE-LEXF; green, HXB-BXH; blue, GK × BN). Outer circle, positions of informative SNPs for each cohort. Bars in the inner circle, conflicts in the genetic map. Arrows, relocation, according to minimal recombination fraction, of SNP markers that had extreme genetic distances compared to their physical distance from adjacent markers. For black lines, all crosses support the rearrangement; lime green, HXB-BXH and F_2 cross support the rearrangement; orange, unresolved genomic conflicts.

SDPs for mapping quantitative traits in rat recombinant inbred strains

We carried out quantitative trait mapping for a subset of 74 traits that were measured in the FXLE-LEXF panel of 33 RI strains[31] and their parental progenitors F344/Stm and LE/Stm as part of the Japanese Rat Phenome Project[28]. Of the 20,283 SNPs tested, 28.5% (5,778) were polymorphic between the parental strains, with 1,033 distinct SDPs across the 33 RI strains. In total, we identified 250 significant QTL (false discovery rate < 0.05) for 74 phenotypic parameters (see **Supplementary Table 6** online and URL 2 below). Although we detected loci previously reported for a number of traits (for example, cholesterol levels; **Supplementary Fig. 7** online), most of the significant linkages were new, as most of the phenotypes assessed here have not been mapped in the rat previously (**Supplementary Table 6**). The number of SDP identified by the current SNP map increased more than threefold over the existing microsatellite-based map[13] and determined 3,766 recombination events for the 33 RI strains. This resulted in a marked improvement of genome wide coverage and greater QTL mapping resolution. Our results demonstrate that this RI resource, historically generated to study genes involved in tumorigenesis, is applicable to the detection of physiological and behavioral traits and risk factors for complex diseases.

Accessibility of the data

The complete set of newly reported SNPs and the entire set of genotypes across all rat strains are publicly accessible, through Ensembl (see URL 7 below) and other web sites (**Supplementary**

Note online). The BioMart tool provides a particularly flexible interface to this data, whereby both genomic position queries and gene list queries can be used to select a set of SNPs that are polymorphic between two strain combinations. The new SNPs are fully integrated with SNPs from other sources on Ensembl overview displays such as ContigView. A tool to select subsets of SNPs is available (see URL 8 below), as is visualization of polymorphisms along chromosomes (see URL 9). To facilitate access to the data in the context of further information, the data represented here has been integrated in other databases, namely the Rat Genome Database (RGD; see URL 1) and GeneNetworks (see URL 10). Finally, data presented for functional assessment of the SNP consequence type is comprehensively available (see URL 11).

DISCUSSION

We present a comprehensive study of genetic variation in the laboratory rat in order to accelerate its use as a model of human complex diseases. To this end, we have identified approximately 3 million SNPs and predicted the functional effects of 325,788 of them. This brings a far richer genetic toolkit to this common toxicology and physiology model mammal, and the presence of pre-typed renewable genetic resources, such as RI lines, provides a resource in which any phenotypic assay available on rat can be augmented by a genotype scan, provided the assay can be performed on the RI panel. The functional analysis of SNPs is in its infancy, but already provides a useful priority list of potential functional variants for further testing, in particular when combined with other data, such as eQTL.

We genotyped 20,283 selected SNPs that were distributed evenly across the genome in 167 inbred strains and 64 recombinant inbred lines, resulting in a community resource of validated polymorphic markers for any strain combination. These strains represent founders of crosses with more than 90% of the rat QTL reported in the literature, and thus this resource will serve as a valuable tool for functional genomics and facilitate positional cloning of QTL and the identification of causal variants.

Our analysis of the evolutionary history of the rat based on these data showed that there were ten clusters of strains sharing breeding history and that the Brown Norway strain separated phylogenetically from all other strains. Our first-generation haplotype map of the laboratory rat showed the genomic history of genomic segments and may allow for the imputation of genotypes in other strains with sparser genotype and sequence data. Notably, our study showed different extents of LD in populations of laboratory mice and rats. Moreover, the phylogenetic relationships inferred from the genotype data suggested a more complex origin and relationships between rat strains than between mouse strains. Theoretically, applying correlation between phenotype and ancestral sequence origin across many inbred strains could enable the identification of genomic regions that are likely to contain the responsible genes. However, our simulations showed that genome-wide association mapping using 100 inbred rat strains is only practicable for single large-effect QTL, and even in these contexts it is not guaranteed to identify the QTL location. Nevertheless, the method does show promise for single large-effect eQTL. Moreover, knowledge of phylogenetic relationships between strains may help in the selection of informative strains for further phenotypic characterization.

The genetic maps that were generated from RI panels and an F_2 cross showed that the draft genome sequence is largely correct, but did also reveal several regions that need further investigation. In addition, we provided a high-resolution map of the contribution of ancestral genomic segments for every individual strain in two rat recombinant inbred panels. The utility of such information was illustrated by mapping QTL for 74 phenotypic parameters in one of these RI panels (FXLE-LEXF).

The availability of robustly assayed SNPs and renewable genetic resources provided here constitutes the next step for the genetic toolkit for the rat. The rat is extensively used in many biological assays, and lowering the cost and other barriers for the application of genetic tools to this organism provides many synergies between the vast range of existing working assays on this organism and a powerful genetic toolkit. We expect that this resource will lead in the future to the resequencing of key strains, the discovery of more genetic associations, and their final resolution to a molecular variant, leading to a new avenue to research human disease.

METHODS

Animals. We used 167 inbred rat strains that covered the diversity of the most commonly used strains in research. Tissue was provided by researchers from the rat community and DNA extraction was performed at the MDC. The list of strains with designation and ILAR code can be found at the web page of the STAR consortium (see URL 4 below). Also, most of the strains are listed in RGD (see URL 1). For most of these strains, QTL data are available. In our analysis we captured strains that encompass about 90% of the rat QTL reported in the RGD (see URL 1). The sets of recombinant inbred strains and the F_2 cross are described in the **Supplementary Methods**.

Genomic shotgun fragment sequencing. Shotgun libraries of a single male rat for each of strains SS/Jr, WKY/Bbb, GK/Ox and SHRSP/Bbb were constructed by sonication of 15 μg of genomic DNA. Fragments between 800–2,000 bp in length were subcloned into pUC18 and clones were sequenced from both ends using BigDye terminator chemistry (v3.1) and ABI3730 sequencers (Applied Biosystems). Further information on base calling methods is available in the **Supplementary Methods**.

BAC library construction and end sequencing. The RNB1 rat BAC library was produced by cloning partially SacI-digested genomic DNA isolated from peripheral lymphocytes of a male rat of strain F344/Stm into the pKS145 vector[32]. The BAC library, consisting of 172,800 clones, was used for BAC-end sequencing. BAC DNA extractions were performed using the PI-1100 plasmid isolator (Kurabo), and BAC clones were sequenced using BigDye terminator (v3.1) sequencing kits and ABI 3730 sequencers (Applied Biosystems). Raw sequence data were base-called by KB Basecaller. All sequences were submitted to the DNA Databank of Japan. BAC clones are available from the RIKEN BioResource Center DNA bank (see URL 12 below).

SNP calling. SNP discovery used the SSAHAsnp algorithm[33]. Briefly, this procedure aligned the sequencing reads above to version 3.4 of the rat genome assembly. We apply several filters on alignment quality and neighborhood quality standard, which is defined by the PHRED score of the variant base and surrounding bases.

SNP selection and genotyping. For Illumina GoldenGate, genotyping was carried out using the GoldenGate protocol in a fully automated BeadLab[34]. Samples were processed in 96-well plates. For Affymetrix Targeted Genotyping, genotyping was carried out using the GeneChip Scanner 3000 Targeted Genotyping System protocol from Affymetrix, originally described as MIP technology[35,36].

Data was subjected to stringent quality control procedures eliminating samples and SNPs that did not reach sufficiently high call rates. All SNPs with more than 10% heterozygous genotypes in the inbred strains were removed from the analysis in the final dataset. Also, SNPs with a call rate below 90% were dropped. Our conclusive dataset of 20,283 SNPs comprised 99.2% of all SNPs genotyped, with an overall success rate of 98.7%, covering the genome with an average distance of 130 kb. More information about the genotyping design is given in the **Supplementary Methods**.

Computing functional predictions. We estimated selective pressure by calculating the omega value as previously described[9]. Transcription factor binding sites (TFBSs) in the upstream region of the genes were identified by scanning the promoter region with the MatScan and JASPARS[37] collection of matrices. Next, we identified the TFBSs conserved between species (human and rat) using meta-alignments[38]. Finally, we identified the SNPs that mapped into the conserved TFBSs (1,019). These SNPs are considered to have a putative effect in the expression of the gene.

Phylogenetic structure predictions. We used the genotype information, encompassing 20,283 genome positions from 167 inbred strains, to determine the phylogenetic relationships among the strains. For building a split network, we used the NeighborNet method with uncorrected p-distances implemented in Splitstree4.8 software[12]. We also produced more traditional tree-like structure calculated by MEGA4 package[39] using the neighbor-joining method with uncorrected p-distances and bootstrap test with 1,000 replicates.

Linkage disequilibrium and haplotype structure. We used Haploview 3.32 software[16] to estimate haplotype block structure in rat and mouse laboratory strains. Custom Perl scripts were written to allow selection of SNPs and identification of the most genetically divergent rat strains, to facilitate SNPs/strains randomization or random removal and to calculate LD decay profiles. These scripts are available from authors upon request. Functional analysis of overrepresentation of gene ontology terms for genes located in high LD regions was done with gProfiler web-server[40]. Further information on haplotype analysis is given in the **Supplementary Methods**.

Genetic map constructions and QTL analysis. Genetic mapping in the GK × BN F_2 cross was carried out with JoinMap as previously described[2]. We then used an automatic construction procedure for the genetic map of the HXB-BXH and FXLE-LEXF RI populations and the GK × BN cross from SNP markers and the physical map positions of the SNPs. Next we used the empirical observation of larger number of recombinations between markers with increasing physical distance for an automated reconstruction of the map. The procedure involved systematically evaluating the removal of markers that generate suspiciously large distances in the map. The criterion to call an interval suspicious was defined by a linear model. The model was defined by a user-specified intercept, which is the minimal genetic distance at which distances are considered for removal, and a slope that was computed chromosome-wise from the data. We set this threshold to 3 cM. In order to determine the slope, an initial genetic map was estimated for all markers using the order defined by the physical map. Then all map distances greater than the 95% quantile were removed and the slope was defined as the sum of the remaining genetic distances over the sum of physical distances between markers. The algorithm performed these steps for each chromosome: (i) compute the initial map based on the physical order of markers; (ii) estimate the linear model; (iii) while the size of the genetic map is reduced, evaluate the size of the genetic map when removing candidate markers and select the marker leading to the minimal map size.

For QTL mapping in LEXF-FXLE RI strains, calculations were performed with WinQTL Cart version 2.5 (see URL 13 below). Composite interval mapping was used as QTL mapping strategy. A detailed description of the QTL mapping strategy is given in the **Supplementary Methods**, as is our analysis of simulations on genome-wide association.

URLs. URL 1, http://rgd.mcw.edu/strains/; URL 2, http://www.broad.mit.edu/~claire/MouseHapMap; URL 3, http://www.anim.med.kyoto-u.ac.jp/nbr; URL 4, http://www.snp-star.eu; URL 5, http://rgd.mcw.edu/gbreport/gbrowser_error_conflicts.shtml; URL 6, http://www.anim.med.kyoto-u.ac.jp/nbr/RI_SNPs.html; URL 7, http://www.ensembl.org/; URL 8, http://gscan.well.ox.ac.uk/rats/rat.snp.selector.cgi; URL 9, http://www.well.ox.ac.uk/rat_mapping_resources/SNPbased_maps.html; URL 10, http://www.genenetwork.org/; URL 11, http://bg.upf.edu/funcSTAR/; URL 12, http://www.brc.riken.jp/lab/dna/en/index.html; URL 13, http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

Accession codes. GenBank nucleotide: AAXN01000001–AAXN01072867, AAXL01000001–AAXL01031928, AAXP01000001–AAXP01073497, AAXM01000001–AAXM01023012. DNA Databank of Japan: DH508174–DH839445. dbSNP Submitter Method: RAT_STRAIN-READS_SNPS_200712, RAT_STRAIN-GENOTYPES_200712.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
This work was supported by European Union grants LSGH-2004-005235 and LSHG-CT-2005-019015. We acknowledge funding from the National Genome Research Network of the German Ministry of Science and Education. We thank all of the technical staff of the Sequencing Technology Team at the RIKEN Genomic Sciences Center for their assistance. Part of this work was supported by the National BioResource Project of the Ministry of Education, Culture, Sports, Science and Technology of Japan. D.G. is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (057733/Z/99/A). M.-T.B. and D.G. acknowledge support from the Wellcome Cardiovascular Functional Genomics Initiative (066780/Z/01/Z). M. Pravenec is supported by the Howard Hughes Medical Institute and by the Grant Agency of the Czech Republic. M. Pravenec and V.K. are supported by grants from the Ministry of Education of the Czech Republic.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Jacob, H.J. & Kwitek, A.E. Rat genetics: attaching physiology and pharmacology to the genome. *Nat. Rev. Genet.* **3**, 33–42 (2002).
2. Bihoreau, M.T. *et al.* A linkage map of the rat genome derived from three F_2 crosses. *Genome Res.* **7**, 434–440 (1997).
3. Guryev, V., Berezikov, E., Malik, R., Plasterk, R.H. & Cuppen, E. Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res.* **14**, 1438–1443 (2004).
4. Zimdahl, H. *et al.* A SNP map of the rat genome generated from cDNA sequences. *Science* **303**, 807 (2004).
5. Thomas, M.A., Chen, C.F., Jensen-Seaman, M.I., Tonellato, P.J. & Twigger, S.N. Phylogenetics of rat inbred strains. *Mamm. Genome* **14**, 61–64 (2003).
6. Kurtz, T.W. & Morris, R.C. Jr. Biological variability in Wistar-Kyoto rats. Implications for research with the spontaneously hypertensive rat. *Hypertension* **10**, 127–131 (1987).
7. Kurtz, T.W., Montano, M., Chan, L. & Kabra, P. Molecular evidence of genetic heterogeneity in Wistar-Kyoto rats: implications for research with the spontaneously hypertensive rat. *Hypertension* **13**, 188–192 (1989).
8. Gauguier, D. The rat as a model physiological system. In *Encyclopedia of Genetics* vol. 3 (eds. Jorde, L.B., Little, P., Dunn, M. & Subramaniam, S.) 1154–1171 (Wiley, London, 2006).
9. Arbiza, L. *et al.* Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J. Mol. Biol.* **358**, 1390–1404 (2006).
10. Goñi, J.R., de la Cruz, X. & Orozco, M. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.* **32**, 354–360 (2004).
11. Hedrich, H.J. (ed.) *Genetic Monitoring of Inbred Strains of Rat* (Gustav Fischer, Stuttgart, New York, 1990).
12. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
13. Mashimo, T. *et al.* A set of highly informative rat simple sequence length polymorphism (SSLP) markers and genetically defined rat strains. *BMC Genet.* **7**, 19 (2006).
14. Smits, B.M. *et al.* Efficient single nucleotide polymorphism discovery in laboratory rat strains using wild rat-derived SNP candidates. *BMC Genomics* **6**, 170 (2005).
15. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
16. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
17. Wade, C.M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**, 574–578 (2002).
18. Frazer, K.A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).
19. Yang, H., Bell, T.A., Churchill, G.A. & Pardo-Manuel de Villena, F. On the subspecific origin of the laboratory mouse. *Nat. Genet.* **39**, 1100–1107 (2007).
20. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
21. Guryev, V. *et al.* Haplotype block structure is conserved across mammals. *PLoS Genet.* **2**, e121 (2006).
22. Jensen-Seaman, M.I. *et al.* Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**, 528–538 (2004).
23. Grupe, A. *et al.* In silico mapping of complex disease-related traits in mice. *Science* **292**, 1915–1918 (2001).
24. Payseur, B.A. & Place, M. Prospects for association mapping in classical inbred mouse strains. *Genetics* **175**, 1999–2008 (2007).
25. Gauguier, D. *et al.* Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat. *Nat. Genet.* **12**, 38–43 (1996).
26. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**, 243–253 (2005).
27. Dumas, M.E. *et al.* Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nat. Genet.* **39**, 666–672 (2007).

28. Mashimo, T., Voigt, B., Kuramoto, T. & Serikawa, T. Rat Phenome Project: the untapped potential of existing rat strains. *J. Appl. Physiol.* **98**, 371–379 (2005).
29. Ihaka, R. & Gentleman, R.R. A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5**, 299–314 (1996).
30. Broman, K.W. The genomes of recombinant inbred lines. *Genetics* **169**, 1133–1146 (2005).
31. Shisa, H. *et al.* The LEXF: a new set of rat recombinant inbred strains between LE/Stm and F344. *Mamm. Genome* **8**, 324–327 (1997).
32. Fujiyama, A. *et al.* Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295**, 131–134 (2002).
33. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
34. Oliphant, A., Barker, D.L., Stuelpnagel, J.R. & Chee, M.S. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **32** (suppl.), 56–58, 60–61 (2002).
35. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
36. Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269–275 (2005).
37. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–D97 (2006).
38. Blanco, E., Messeguer, X., Smith, T.F. & Guigo, R. Transcription factor map alignment of promoter regions. *PLOS Comput. Biol.* **2**, e49 (2006).
39. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
40. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).

The complete list of authors is as follows:

The STAR Consortium: Kathrin Saar[1], Alfred Beck[2], Marie-Thérèse Bihoreau[3], Ewan Birney[4], Denise Brocklebank[3], Yuan Chen[4], Edwin Cuppen[5], Stephanie Demonchy[6], Joaquin Dopazo[7], Paul Flicek[4], Mario Foglio[6], Asao Fujiyama[8,9], Ivo G Gut[6], Dominique Gauguier[3], Roderic Guigo[10,11], Victor Guryev[5], Matthias Heinig[1], Oliver Hummel[1], Niels Jahn[12], Sven Klages[2], Vladimir Kren[13,14], Michael Kube[2], Heiner Kuhl[2], Takashi Kuramoto[15], Yoko Kuroki[8], Doris Lechner[6], Young-Ae Lee[1,16], Nuria Lopez-Bigas[10,11], G Mark Lathrop[6], Tomoji Mashimo[15], Ignacio Medina[7], Richard Mott[3], Giannino Patone[1], Jeanne-Antide Perrier-Cornet[6], Matthias Platzer[12], Michal Pravenec[13,14], Richard Reinhardt[2], Yoshiyuki Sakaki[8], Markus Schilhabel[12], Herbert Schulz[1], Tadao Serikawa[15], Medya Shikhagaie[11], Shouji Tatsumoto[8], Stefan Taudien[12], Atsushi Toyoda[8], Birger Voigt[15], Diana Zelenika[6], Heike Zimdahl[1] & Norbert Hubner[1]

Affiliations for participants: [1]Max-Delbrück Center for Molecular Medicine, Robert-Rossle-Straße 10, 13125 Berlin, Germany. [2]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford, OX3 7BN, UK. [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [5]Hubrecht Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. [6]Commissariat à l'Énergie Atomique, Institut de Génomique, Centre National de Génotypage, 2 rue Gaston Crémieux CP 5721, 91 057 Evry Cedex, France. [7]Department of Bioinformatics, and Functional Genomics Node, Centro de Investigación Príncipe Felipe, Avenida Autopista del Saler 16, 46012 Valencia, Spain. [8]RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [9]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan. [10]Center for Genomic Regulation, C/Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. [11]Experimental and Health Science Department, Universitat Pompeu Fabra, C/Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. [12]Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute, Beutenbergstraße 11, 07745 Jena, Germany. [13]Institute of Physiology, Czech Academy of Sciences, Videnska 1083, 14220 Prague 4, Czech Republic. [14]Institute of Biology and Medical Genetics, First Medical Faculty, Charles University, Albertov 4, 12800 Prague 2, Czech Republic. [15]Institute of Laboratory Animals, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. [16]Pediatric Pneumology and Immunology, Charite, Campus Virchow Klinikum, Augustenburger Platz 1, 13353 Berlin, Germany.