Published in final edited form as:
Nat Genet. 2008 May ; 40(5): 560–566. doi:10.1038/ng.124.

SNP and haplotype mapping for genetic analysis in the Rat

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], **Paul Flicek**[4], **Mario Foglio**[6], **Asao Fujiyama**[7,8], **Ivo G. Gut**[6], **Dominique Gauguier**[3], **Roderic Guigo**[9], **Victor Guryev**[5], **Matthias Heinig**[1], **Oliver Hummel**[1], **Niels Jahn**[10], **Sven Klages**[2], **Vladimir Kren**[11], **Heiner Kuhl**[2], **Takashi Kuramoto**[12], **Yoko Kuroki**[7], **Doris Lechner**[6], **Young-Ae Lee**[1], **Nuria Lopez-Bigas**[9], **G. Mark Lathrop**[6], **Tomoji Mashimo**[12], **Michael Kube**[2], **Richard Mott**[3], **Giannino Patone**[1], **Jeanne-Antide Perrier-Cornet**[6], **Matthias Platzer**[10], **Michal Pravenec**[11], **Richard Reinhardt**[2], **Yoshiyuki Sakaki**[7], **Markus Schilhabel**[10], **Herbert Schulz**[1], **Tadao Serikawa**[12], **Medya Shikhagaie**[9], **Shouji Tatsumoto**[7], **Stefan Taudien**[10], **Atsushi Toyoda**[7], **Birger Voigt**[12], **Diana Zelenika**[6], **Heike Zimdahl**[1], and **Norbert Hubner**[1]

[1]Max-Delbrück-Center for Molecular Medicine (MDC), Berlin-Buch, Germany [2]Max Planck Institute for Molecular Genetics, Berlin, Germany [3]Wellcome Trust Centre for Human Genetics, University of Oxford, UK [4]European Bioinformatics Institute, Hinxton, UK [5]Hubrecht Institute, Utrecht, The Netherlands [6]CEA/Institut de Génomique, Centre National de Génotypage, Evry, France [7]RIKEN Genomic Sciences Center, Kanagawa 230-0045, Japan [8]National Institute of Informatics, Tokyo 101-8430, Japan [9]Centre de Regulacio Genomica, Barcelona, Spain [10]Leibniz-Institut für Altersforschung - Fritz-Lipmann-Institut, Jena, Germany [11]Institute of Physiology, Czech Academy of Sciences and 1st Medical Faculty, Charles University, Prague, Czech Republic [12]Institute of Laboratory Animals, Graduate School of Medicine, Kyoto University, Yoshidakonoe-cho, Sakyo-ku, Kyoto 606-8501, Japan

Abstract

The laboratory rat is one of the most extensively studied model organisms. Inbred laboratory rat strains have 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 novel SNPs. We obtained accurate and complete genotypes for a subset of 20,238 SNPs across 167 distinct inbred rat strains, two rat recombinant inbred (RI) 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 characterise the population structure and illustrate the degree of linkage disequilibrium. We provide a detailed SNP map and demonstrate its utility for QTL mapping studies. This community resource is openly available and augments the genetic tools for this workhorse of physiological studies.

Address correspondence to: Norbert Hubner, Max-Delbruck-Center for Molecular Medicine (MDC), Robert-Rossle-Str. 10, 13125 Berlin, Germany, nhuebner@mdc-berlin.de.

Accession numbers:
NCBI Nucleotide: AAXN01000001 - AAXN01072867, AAXL01000001 - AAXL01031928, AAXP01000001 - AAXP01073497, AAXM01000001 - AAXM01023012; DNA Databank of Japan: DH508174 - DH839445.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Introduction

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[URL1]. While 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 decade1. The genome of the BN/NHsdMcwi rat 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. Currently genetic variability in the rat genome is usually assayed using a limited set of microsatellite markers 1,2,3,4. A dense set of polymorphic markers, for which single nucleotide polymorphisms (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 many other laboratory animals, is known to have a complex genesis5, 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 consequences6,7. Thus it is important to have detailed marker information available on any specific sub-strain. The presence of a number of recombinant inbred lines (RI), in particular the HXB-BXH sets and the FXLE-LEXF sets, and the presence of both congenics and consomics provide 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 basically 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.

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: (1) shotgun sequence generated from the four strains SS/Jr,

URL List:
URL 1 http://rgd.mcw.edu/strains/;

Europe PMC Funders Author Manuscripts

WKY/Bbb, GK/Ox, SHRSP/Bbb, all widely used in disease mapping experiments in crosses and congenic strains8; (2) whole genome shotgun sequence at 1.5x coverage of the outbred Sprague-Dawley (SD) rat generated by Celera; and (3) BAC end sequences from the F344/Stm rat, which is also widely used in genetic studies. The numbers of SNPs discovered and reads used from each resource are listed in Table 1. For the SD rat, approximately 14% of the SNPs were heterozygous based on evidence from overlapping aligned sequencing reads. Due to 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 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, 2 RI panels of 31 and 33 strains, and tested a subset (n=9,691) in 89 F2 animals. The allele-frequencies of SNPs across the inbred strains are approximately evenly distributed between 3% and 50%. We assayed seven 1,536-plex assays that can be run on the Illumina BeadLab station (10,752 SNPs). In parallel, we have developed a 10K rat Targeted Genotyping panel to be run on an Affymetrix platform (9,691 SNPs) (see supplementary Methods). 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. Both 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 is very stable even for those nodes that are not supported by a high bootstrap value (supplementary Figure 1).

Annotation of putative functional SNPs from inbred strains: funcSTAR database

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 SD 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 due to incomplete inbreeding of the colony. We estimated the selective pressure on amino acid replacement mutation for all residues with non-synonymous coding SNPs (n=1,160) by calculating the Omega value or non-synonymous/synonymous (dN/dS) substitution rate9. 56 SNPs with Omega value lower than 0.1 were identified (supplementary Table 1), which is indicative of a likely effect in protein function of which 7 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

Europe PMC Funders Author Manuscripts

site (AT). Finally, we assessed the potential effect of the SNPs on transcriptional and post-transcriptional regulation. 1,019 SNPs in promoter regions map into conserved transcription factor binding sites and 568 SNPs map into DNA triplexes10. 132 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 strains11. When considering the many rounds of inbreeding and interbreeding that most likely took place a complex evolutionary history can be expected. Therefore, we chose to visualise the inter-strain relationship and genetic proximity in a phylogenetic network12 rather than a tree (Figure 1). The observed strain relationships agree very well with the known history of rat strains and support all significant clusters of strains previously recognized in analysis of microsatellite markers5. The reticulation around the center of the network may reflect extensive genetic heterogeneity of the ancestral rat population. The network reveals 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 inbred6,7. The Brown Norway (BN) rats are 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 data5,13 and a study that used different SNP panels14. The inclusion of wild and outbred rats into the phylogenetic analysis results in their branching from the reticulate center and not outgrouping BN or non-BN rats. Within non-BN strains we define 10 clusters of strains that evidently shared breeding history, these clusters are highly supported as monophyletic groups on a traditional phylogenetic tree (supplementary Figure 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 was shown that BN and DA inbred substrains obtained from different locations harbor genetic difference14. Our dataset includes multiple substrain samples and confirms the presence of variation between substrains. This effect is 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 less 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

Athough the genotyped panel of markers is not dense enough to support conclusive evaluation of 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 of historical recombination15. Using

Haploview16, 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 2 SNPs, but these blocks are relatively small and in total cover less than 2 Mb. In contrast, the average size of 514 blocks that contain 3 or more SNPs is 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 Figure 2). 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[URL2]. 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 is larger in laboratory mice where haplotype blocks cover a larger fraction of the genome (35% compared to 12% in the rat), contain a higher proportion of informative markers (56% vs 21%) and have a greater average size (648 kb vs 388 kb). The direct comparison of LD decay profiles in rat and mouse (supplementary Figure 3) further substantiates this notion. On the other hand, linkage disequilibrium in the rat is larger than for cow, where haplotype blocks cover only 2.2% of autosomes17. Although LD in rats is less pronounced when compared to mice it still extends over hundreds of kilobases, unlike LD in humans or across dog breeds where r^2 drops below 0.1 at 100 kb18–20. 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 subspecies17,18,19, while there is no such evidence in the rat. At the same time, the phylogenetic relationships among groups of rats are hard to deduce (supplementary figure 1) reflecting more divergent genetic background of a rat founder population. Comparison of LD between rat and mouse shows that the size of haplotype blocks in syntenic regions exhibit small, but significant correlation (r^2=0.18), consistent with the previously observed correlation between murine recombination rate and fine LD structure21,22. Olfactory genes are the only gene class that is overrepresented in rat LD blocks (P<10-6). 21 distinct gene clusters harboring about one third (n=325) of all olfactory genes are 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.

Interestingly, we identified 939 inter-chromosomal SNP pairs that are in full linkage disequilibrium. These SNPs are heavily shifted towards low minor allele frequencies with only 38 and 4 of them having a minor allele frequency larger than 0.1 and 0.15, respectively. More detailed inspection reveals that, besides being rare, the vast majority of minor frequency alleles are private to branches of the phylogenetic tree (e.g. many of them are restricted to the SHR + WKY + GK cluster). Thus perfectly correlating SNPs on different

URL 2 http://www.broad.mit.edu/~claire/MouseHapMap;

chromosomes are unlikely to result from epistatic effects or genome assembly errors, but are more likely to represent a shared physical genomic structure (*i.e.* genetic background). It should be mentioned that imperfect but significant pairwise correlation (r^2>=0.5) is observed among about 0.2% of the inter-chromosomal SNP pairs. The highly correlated subset disappears almost completely when SNP alleles are randomized and could thus reflect epistatic interactions as well as ancestral relationships.

Prospects for whole genome association mapping using inbred Rat strains

One hundred of the most diverse inbred rat strains were evaluated by simulation for their potential for whole genome association (WGA) mapping of quantitative trait loci (QTL). The method originated in the mouse genetics community23, where it has generated much discussion24. 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 is no genetic component to the phenotypic variance is $\log P_{max}$ = 4.1, close to the Bonferroni estimate of 4.3. In contrast, for an infinitesimal model where many small-effect QTL, in total accounting for 50% of the total variance, are distributed uniformly across the genome, the median $\log P_{max}$=21.5, *i.e.* much higher. For a single major QTL explaining 50% of the variance, the genome-wide maximum coincides with the true position in 31% of simulations, and there is a local maximum exceeding the genome-wide null threshold of significance at the true QTL in 51% of simulations, and within 1Mb of the true position in 91% of cases. However, the median $\log P_{max}$ = 14.3, which lies between these thresholds and on average 72 putative QTL exceed the null threshold of Finally, for a realistic complex trait scenario of ten 5% QTL the median $\log P_{max}$ = 9.96, and the median number of putative QTL exceeding 4.1 is 1,412. Thus there is a very large number of false positive QTL, and consequently each true QTL is close to a putative QTL. Supplementary Table 2 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 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 F2 cross and recombinant inbred lines

A total of 20,283 SNPs were typed 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 9,691 SNPs were typed in 89 progeny of a F2 cross between BN/Par and GK/Ox rats (GKxBN). These populations have been used for mapping complex phenotypes for metabolic syndrome25, expression QTL26 and metabonomic traits27, and extensive phenotype characterizations as part of the Japanese phenome project[URL3],28. In addition, genotype analysis in F2 rats enabled assessment of the reliability of heterozygous genotype calls.

URL 3 http://www.anim.med.kyoto-u.ac.jp/nbr;

Genetic map construction was initially carried out in the GKxBN F2 cross with the JoinMap program as previously described2. This approach has the advantage that prior knowledge of markers' physical order is not required for calculating genetic distances. Over 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). Following genotype verifications, we confirmed the existence of strong distortion of segregation previously reported in chromosomes 3, 4, 9 and 132. Alignment of genetic and physical maps showed the general agreement of marker order and distance (supplementary Figure 4). However, we identified regions (from 1 SNP to 8Mb) where SNP-based genetic maps are inconsistent with the current rat genome assembly draft

Genetic mapping in this cross and both panels of RI strains was then repeated using the R and R/QTL software packages29,30 integrating SNP genotype and physical map data resulting in 16,543 SNPs mapped. Data were initially filtered to remove markers containing genotyping errors (e.g. absence of segregation in the cohort despite apparent allele variation in the parental strains or over 10% of heterozygous genotypes in the RI 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 and[URL4], and the resulting maps in supplementary Figure 5. Strong evidence of discrepancies between the genetic map and the draft genome assembly were found (Figure 2). In particular, genetic mapping in all three panels identified a p11-centromeric segment of chromosome 1, which has been wrongly assembled in the p14-telomeric region of chromosome 17. Genetic mapping data suggest further additional intra- and inter-chromosomal relocations in regions of chromosomes 2, 4, 11, 12, 14, 17. Known conflicts between rat genome assemblies, provided by BCM and Celera[URL5], indicate the relocation in the p14 region of chromosome 17 supporting the Celera assembly, and one conflict on chromosome 9 is resolved favouring the BCM assembly (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 occur on contig boundaries (supplementary Table 5). We rearranged the fragments of the chromosome resulting from splitting the contigs that were not linked (LOD < 2) in the order that generates the smallest average recombination fraction in the three populations (supplementary Figure 7a). Using the resulting marker positions we constructed three genetic maps (supplementary Figure 7b) summarized in supplementary Table 4.

URL 4 http://www.snp-star.eu;
URL 5 http://rgd.mcw.edu/gbreport/gbrowser_error_conflicts.shtml;

Europe PMC Funders Author Manuscripts

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 strains31 and their parental progenitors F344/Stm and LE/Stm as part of the Japanese Rat Phenome Project28. 28.5 % (5,778) of the 20,283 SNPs tested were polymorphic between the parental strains with 1,033 distinct SDP across the 33 RI strains. In total, we identified 250 significant QTL (FDR<0.05) for 74 phenotypic parameters (supplementary Table 6)[URL6]. While we detected loci previously reported for a number of traits (e.g. cholesterol levels; supplementary Figure 6) most of the significant linkages are novel since the majority 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 increases more than 3-fold as compared to the existing microsatellite-based map13 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 behavioural traits and risk factors for complex diseases.

Accessibility of the data

The complete set of novel SNPs reported and the entire set of genotypes across all rat strains are publicly accessible, via Ensembl[URL7] and other web sites, which are summarized in the supplementary Note. The BioMart tool provides a particularly flexible interface to this data, where both genomic position queries and gene list queries can be used to select a set of SNPs which are polymorphic between two strain combinations. The novel 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[URL8], and visualization of polymorphisms along chromosomes is available[URL9]. To facilitate access to the data in the context of additional information the data represented here has been integrated in other databases, namely RGD[URL1] and GeneNetworks[URL10]. Finally, data presented for functional assessment of the SNP consequence type is comprehensively available[URL11].

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 approx. 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

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/gsBleadingEdge/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/;

Europe PMC Funders Author Manuscripts

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 expression QTL.

We have genotyped 20,283 selected SNPs that are 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 shows that there are 10 clusters of strains that share breeding history and that the Brown Norway strain separates phylogenetically from all other strains. Our first-generation haplotype map of the laboratory rat shows the genomic history of genomic segments and may allow for the imputation of genotypes in additional strains with sparser genotype and sequence data. Interestingly, our study shows different extent of LD in populations of laboratory mice and rats. Moreover, the phylogenetic relationships inferred from the genotype data suggest a more complex origin and relationships between rat strains if compared with mouse. 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 show that whole genome 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 on 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 F2 cross show 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 by lowering the cost and other barriers for the application of genetic tools to this organism provides numerous synergies between the vast array of existing working assays on this organism to be combined with a powerful genetic toolkit. We expect that this resource will lead in the future to the re-sequencing 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.

Materials and Methods

Animals

We used 167 inbred rat strains that cover 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[URL4]. Also, the majority of strains are listed in RGD[URL1]. For most of these strains, QTL data are available. In our analysis we captured strains that encompass about 90% of the rat QTL reported according the Rat Genome Database[URL1]. The sets of recombinant inbred strains and the F2 cross are described in the supplementary Methods.

Genomic shotgun fragment sequencing

Shotgun libraries of a single male rat of strains SS/Jr, WKY/Bbb, GK/Ox, and SHRSP/Bbb, respectively, were constructed by sonication of 15 μg of genomic DNA each. 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). Additional information on basecalling methods is available in the supplementary Methods.

BAC library construction and end sequencing

The RNB1 rat BAC library was produced by cloning of partially digested *Sac I* genomic DNA isolated from peripheral lymphocytes of a male rat of strain F344/Stm into pKS145 vector, which was developed by Fujiyama et al. (2002) 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 (DDBJ). BAC clones are available from the RIKEN BioResource Center DNA bank[URL12].

SNP calling

SNP discovery used the SSAHAsnp algorithm33. 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 (NQS), 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 BeadLab34. Samples were processed in 96-well plates. For Affymetrix Targeted Genotyping, genotyping was carried out using the GeneChip® Scanner 3000

URL 12 http://www.brc.riken.jp/lab/dna/en/index.html;

Europe PMC Funders Author Manuscripts

Targeted Genotyping System protocol from Affymetrix, originally described as MIP technology35,36.

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 where removed from the analysis in the final data set. Also, SNPs with a call rate below 90% were dropped. Our conclusive data set 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 130kb. Additional information about the genotyping design is given in the supplementary Methods.

Computing functional predictions

Selective pressure has been estimated by calculating the Omega value as previously described9. Transcription Factor Binding Sites (TFBS) in the upstream region of the genes were identified by scanning the promoter region with MatScan and JASPARS 37collection of matrices. Next, we identified the TFBS conserved across species (human and rat) using meta-alignments38. Finally, we have identified the SNPs that map into the conserved TFBS (1,019). These SNPs are considered to have a putative effect in the expression of the gene.

Phylogenetic structure predictions

The genotype information encompassing 20,283 genome positions from 167 inbred strains was employed to determine the phylogenetic relationships among the strains. We used NeighborNet method with uncorrected p-distances implemented in Splitstree4.8 software12 for building split network. We also produced more traditional tree-like structure calculated by MEGA4 package39 using Neighbor-Joining method with uncorrected p-distances and bootstrap test with 1,000 replicates.

Linkage disequilibrium and haplotype structure

We used Haploview 3.32 software16 to estimate haplotype block structure in rat and mouse laboratory strains. Custom Perl scripts were written to allow selection of SNP, most divergent rat strains, to facilitate SNPs/strains randomisation 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-server40. Additional information on haplotype analysis is given in the supplementary Methods.

Genetic map constructions and QTL Analysis

Genetic mapping in the GKxBN F2 cross was carried out with JoinMap as previously described2. We then used an automatic construction procedure for the genetic map of the HXB-BXH and FXLE-LEXF RI populations and the GKxBN cross from SNP markers and the physical map positions of the SNPs. We propose to use the empirical observation of larger number of recombinations between markers with increasing physical distance for an automated reconstruction of the map. The procedure is systematically evaluating the removal of markers that generate suspiciously large distances in the map. The criterion to call an interval suspicious is defined by a linear model. The model is defined by a user-specified

intercept, which is the minimal genetic distance at which distances are considered for removal and a slope that is computed chromosome-wise from the data. We set this threshold to 3cM. In order to determine the slope, an initial genetic map is estimated for all markers using the order defined by the physical map. Then all map distances greater than the 95% quantile are removed and the slope is defined as the sum of the remaining genetic distances over the sum of physical distances between markers. The algorithm performs 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 Ver. 2.5[URL13]. Composite Interval Mapping (CIM) was used as QTL mapping strategy. A detailed description of the QTL mapping strategy is given in the supplementary Methods. Our analysis on simulations on genome wide association is given in the supplementary Methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by EU grant LSGH-2004-005235 and by grant 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 RIKEN GSC 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. holds a Wellcome Trust Senior Fellowship in Basic Biomedical Science (057733). M.T.B. and D.G. acknowledge support from the Wellcome Cardiovascular Functional Genomics Initiative (CFG, 066780/Z/01/Z). M.Pr. is an international research scholar of the Howard Hughes Medical Institute and is supported by the Grant Agency of the Czech Republic. M. Pr. and V.K are supported by grants from the Ministry of Education of the Czech Republic.

Reference List

1. Jacob HJ, Kwitek AE. Rat genetics: attaching physiology and pharmacology to the genome. Nat Rev Genet. 2002; 3:33–42. [PubMed: 11823789]
2. Bihoreau MT, et al. A linkage map of the rat genome derived from three F2 crosses. Genome Res. 1997; 7:434–440. [PubMed: 9149940]
3. Guryev V, Berezikov E, Malik R, Plasterk RH, Cuppen E. Single nucleotide polymorphisms associated with rat expressed sequences. Genome Res. 2004; 14:1438–1443. [PubMed: 15231757]
4. Zimdahl H, et al. A SNP map of the rat genome generated from cDNA sequences. Science. 2004; 303:807. [PubMed: 14764869]
5. Thomas MA, Chen CF, Jensen-Seaman MI, Tonellato PJ, Twigger SN. Phylogenetics of rat inbred strains. Mamm Genome. 2003; 14:61–64. [PubMed: 12532268]
6. Kurtz TW, Morris RC Jr. Biological variability in Wistar-Kyoto rats. Implications for research with the spontaneously hypertensive rat. Hypertension. 1987; 10:127–131. [PubMed: 3596765]

URL 13 http://www.statgen.ncsu.edu/qtlcart/WQTLCart.htm.

7. Kurtz TW, Montano M, Chan L, Kabra P. Molecular evidence of genetic heterogeneity in Wistar-Kyoto rats: implications for research with the spontaneously hypertensive rat. Hypertension. 1989; 13:188–192. [PubMed: 2914738]

8. Gauguier, D. The rat as a model physiological system. Encyclopedia of Genetics. Jorde, LB.Little, P.Dunn, M., Subramaniam, S., editors. Vol. 3. Wiley; London, U.K.: 2006. p. 1154-1171.

9. Arbiza L, et al. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol. 2006; 358:1390–1404. [PubMed: 16584746]

10. Goni JR, de l C X, Orozco M. Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res. 2004; 32:354–360. [PubMed: 14726484]

11. Hedrich, HJ., editor. Genetic Monitoring of Inbred Strains of Rat. Stuttgart, New York: Gustav Fischer Verlag; 1990.

12. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006; 23:254–267. [PubMed: 16221896]

13. Mashimo T, et al. A set of highly informative rat simple sequence length polymorphism (SSLP) markers and genetically defined rat strains. BMC Genet. 2006; 7:19. [PubMed: 16584579]

14. Smits BM, et al. Efficient single nucleotide polymorphism discovery in laboratory rat strains using wild rat-derived SNP candidates. BMC Genomics. 2005; 6:170. [PubMed: 16316463]

15. Gabriel SB, et al. The structure of haplotype blocks in the human genome. Science. 2002; 296:2225–2229. [PubMed: 12029063]

16. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]

17. Wade CM, et al. The mosaic structure of variation in the laboratory mouse genome. Nature. 2002; 420:574–578. [PubMed: 12466852]

18. Frazer KA, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 2007; 448:1050–1053. [PubMed: 17660834]

19. Yang H, Bell TA, Churchill GA, Pardo-Manuel d V. On the subspecific origin of the laboratory mouse. Nat Genet. 2007; 39:1100–1107. [PubMed: 17660819]

20. Lindblad-Toh K, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005; 438:803–819. [PubMed: 16341006]

21. Guryev V, et al. Haplotype block structure is conserved across mammals. PLoS Genet. 2006; 2:e121. [PubMed: 16895449]

22. Jensen-Seaman MI, et al. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 2004; 14:528–538. [PubMed: 15059993]

23. Grupe A, et al. In silico mapping of complex disease-related traits in mice. Science. 2001; 292:1915–1918. [PubMed: 11397946]

24. Payseur BA, Place M. Prospects for association mapping in classical inbred mouse strains. Genetics. 2007; 175:1999–2008. [PubMed: 17277361]

25. Gauguier D, et al. Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat. Nat Genet. 1996; 12:38–43. [PubMed: 8528248]

26. Hubner N, et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet. 2005; 37:243–253. [PubMed: 15711544]

27. Dumas ME, et al. Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. Nat Genet. 2007; 39:666–672. [PubMed: 17435758]

28. Mashimo T, Voigt B, Kuramoto T, Serikawa T. Rat Phenome Project: the untapped potential of existing rat strains. J Appl Physiol. 2005; 98:371–379. [PubMed: 15591307]

29. Ihaka R, Gentleman RR. A Language for Data Analysis and Graphics. J Comput Graph Statist. 1996; 5:299–314.

30. Broman KW. The genomes of recombinant inbred lines. Genetics. 2005; 169:1133–1146. [PubMed: 15545647]

31. Shisa H, et al. The LEXF: a new set of rat recombinant inbred strains between LE/Stm and F344. Mamm Genome. 1997; 8:324–327. [PubMed: 9107675]

32. Fujiyama A, et al. Construction and analysis of a human-chimpanzee comparative clone map. Science. 2002; 295:131–134. [PubMed: 11778049]

33. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001; 11:1725–1729. [PubMed: 11591649]
34. Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques. 2002; (Suppl):56–1.
35. Hardenbol P, et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol. 2003; 21:673–678. [PubMed: 12730666]
36. Hardenbol P, et al. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res. 2005; 15:269–275. [PubMed: 15687290]
37. Vlieghe D, et al. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res. 2006; 34:D95–D97. [PubMed: 16381983]
38. Blanco E, Messeguer X, Smith TF, Guigo R. Transcription factor map alignment of promoter regions. PLoS Comput Biol. 2006; 2:e49. [PubMed: 16733547]
39. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007; 24:1596–1599. [PubMed: 17488738]
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. 2007; 35:W193–W200. [PubMed: 17478515]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

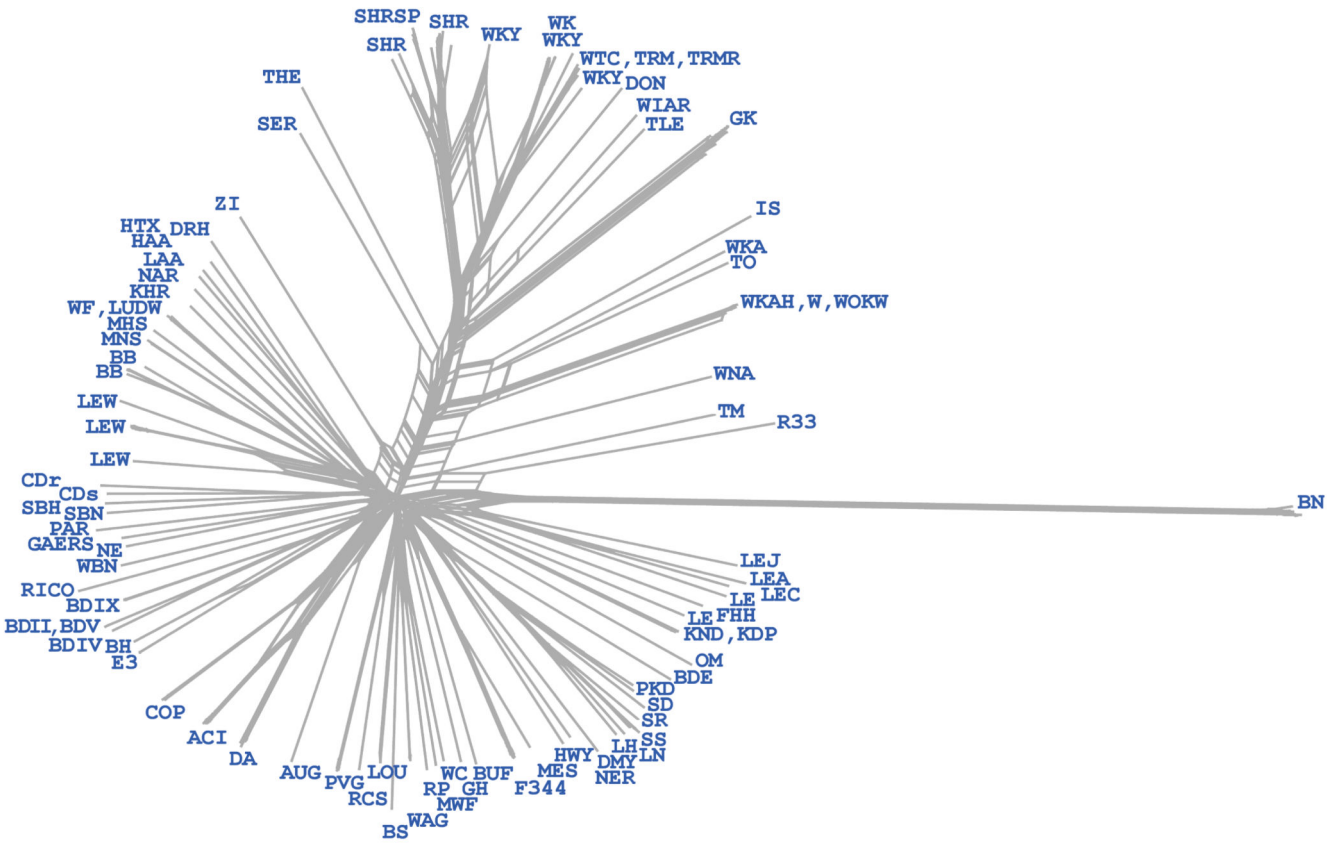

Figure 1.

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 ancestral rat population. The group of WKY-related strains exhibits complex pattern of relationships probably due 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.

Europe PMC Funders Author Manuscripts

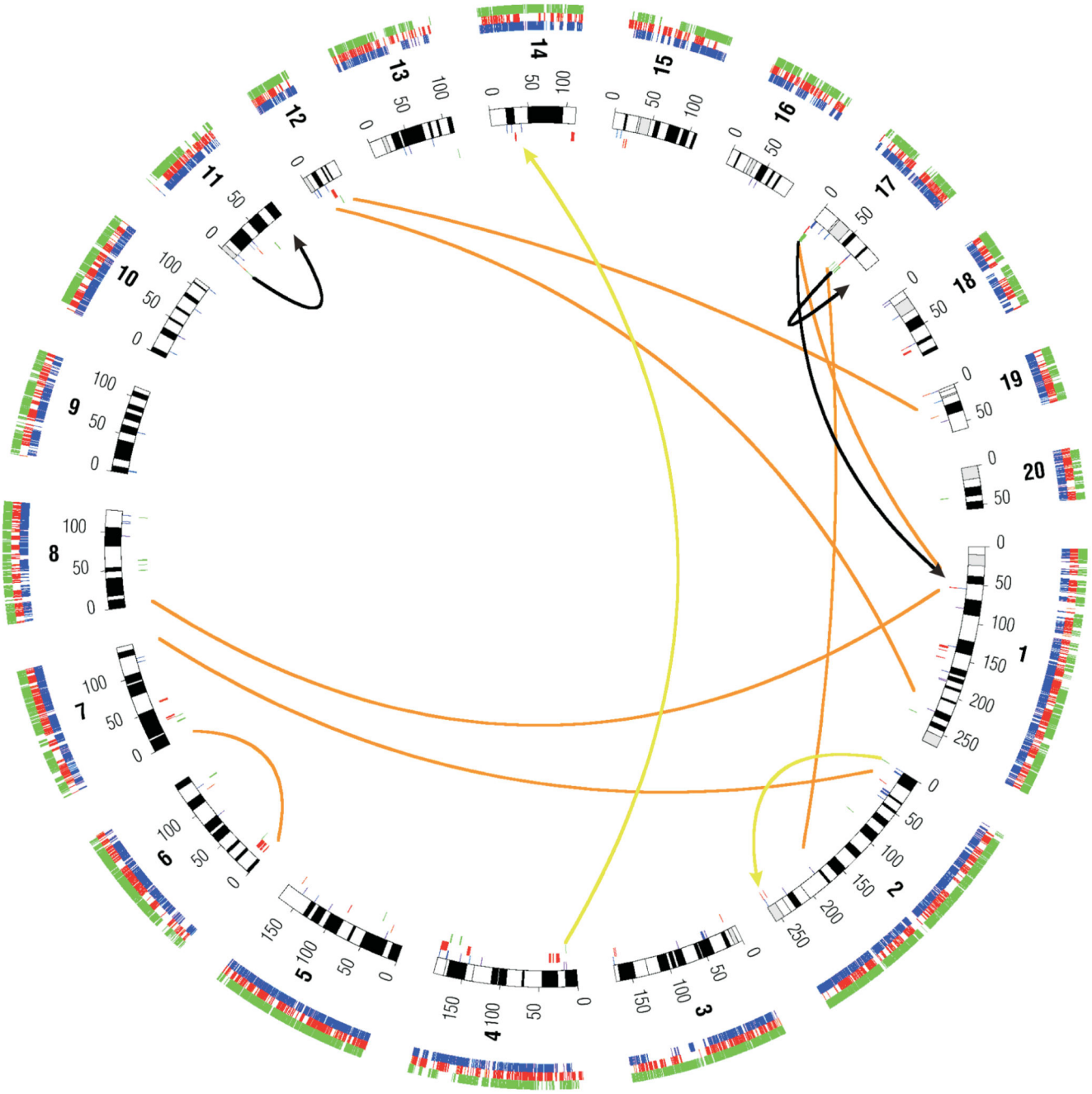

Figure 2.

Identified discrepancies between rat genome assembly and genetic maps. Rearrangement of the physical map according to genetic mapping information. Data from each cohort are color coded (red: FXLE-LEXF, green: HXB-BXH, blue: GKxBN). Black lines: all crosses support this rearrangement; lime green: HXB-BXH and F2 cross support this rearrangement. Orange lines indicate unresolved genomic conflicts. The outer circle marks positions of informative SNPs for each cohort. Arrows indicate the relocation of SNP markers that had extreme genetic distances compared to their physical distance from adjacent markers. Markers were

relocated according to minimal recombination fraction. Conflicts in the genetic map are marked by bars in the inner circle.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1
Number of sequencing reads and detected SNPs per strain.

Source	Number of reads	Non-redundant SNPs
(1) STAR shotgun sequence	249,525	128,976
		(SS/Jr: 56,639) *
		(WKY/Bbb: 32,601) *
		(GK/Ox: 16,838) *
		(SHRSP/Bbb: 28,332) *
(2) Celera SD whole genome shotgun sequence	7,990,225	2,650,525
(3) F344 BAC end sequence	344,064	196,812
total		2,976,313

* number of SNPs before removal of redundancy