Rats in the genomic era

K. C. Worley, G. M. Weinstock, and R. A. Gibbs

Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

Submitted 10 September 2007; accepted in final form 15 November 2007

Rats in the genomic era

Worley KC, Weinstock GM, Gibbs RA. Rats in the genomic era. Physiol Genomics 32: 273–282, 2008. First published November 20, 2007: doi:10.1152/physiolgenomics.00208.2007.—The rat genome project and the resources that it has generated are transforming the translation of rat biology to human medicine. The rat genome was sequenced to a high quality “draft,” the structure and location of the genes were predicted, and a global assessment was published (Gibbs RA et al., Nature 428: 493–521, 2004). Since that time, researchers have made use of the genome sequence and annotations and related resources. We take this opportunity to review the currently available rat genome resources and to discuss the progress and future plans for the rat genome.

Rattus norvegicus; genome sequence; single nucleotide polymorphisms; haplotype map; finished sequence

The laboratory rat is important for biomedical research, with more than 28,000 publications annually since 1966 and almost 37,000 annually since 1996 (3). The rat genome project has generated numerous resources (Table 1), including the draft genome sequence, genetic maps, cDNA sequences, consomic rat lines, and additional bacterial artificial chromosome (BAC) libraries. Over 700 strains (53), including transgenic, congenic, intercross, and recombinant inbred lines of rats are being used to understand human biology. The rat has traditionally been thought of as a model for physiological systems, but with the genome sequence and related resources genetic studies are becoming more common.

Although a researcher may think of genome sequences as standardized and defined entities, genome sequences are not all of the same quality and consistency. Understanding the limitations of the data is important to make the best use of them. Fig. 1 shows the steps used to subclone, sequence, and assemble genome sequences; presents some of the limitations of the assembly process; and describes the difference between a “draft” genome sequence and a “finished” genome sequence.

The initial publication of the draft human genome sequence in 2001 (32) was followed by additional draft genome publications (7–9, 30, 34, 35, 35a, 41a, 41b, 42, 48), along with the publication of the finished human reference genome (10, 12–15, 20, 21, 23, 25, 26, 28, 36–39, 43, 45, 46, 51, 56, 57). At the time of publication, the rat was only the second draft mammalian genome beyond human, and the triangulation possible with the three genomes (human, mouse, and rat) was powerful for informing the genome analysis of the three species.

RESULTS

The Brown Norway rat (Rattus norvegicus) genome was sequenced in a project supported by the National Human Genome Research Institute (NHGRI) and the National Heart, Lung, and Blood Institute (NHLBI) (41a) where the intended product was a draft sequence. This project produced a genome assembly using a variety of sequence and map resources [BAC skim sequence, whole genome shotgun (WGS) sequence, BAC end sequences, BAC fpc map (31)] and a new BAC skim-WGS combined strategy and a new genome assembly program, Atlas (24), that combined the two types of data. Care was taken to increase the quality of the draft since there would not be an opportunity to correct errors during a finishing phase.

The genome assembly was examined in great detail by an analysis group formed by experts in computational biology and rat genetics. Overall, the data were found to be of adequate quality to support the desired analyses and the resulting annotations (41a). Table 2 lists the number of gene predictions reported in the draft rat, mouse, and human genomes and the finished human genome. In addition to the quality of the genome sequence, the statistics depend upon the gene prediction method, the alignment criteria, and the amount of expressed sequence evidence, which is much more abundant for human and mouse than for rat (expressed sequence tags in GenBank in millions: 8.13 human, 4.85 mouse, and only 0.83 rat). Most of the human orthologs are represented, with ~90% of the rat genes possessing a single ortholog in the human genome, and careful analysis of rat proteases showed that 93% have 1:1 orthologs in mouse (41a). Reciprocal best hits between proteins from the three gene predictions sets identified 12,440 1:1:1 orthologs with a mean amino acid identity of 88.0% and mean nucleotide identity is 85.1% (41a). Although the number of exons per gene and average exon length are similar in the three species (mouse, human, rat), the intron lengths vary with the longest average intron in human, followed by rat and then mouse (41a).

The genome sequence has been upgraded once since the initial publication by replacing the draft with the available finished sequence from BACs (~55 Mb, e.g., from the ENCODE regions, marked in blue in Supplementary Table S1). The lack of support for producing finished regions in the initial genome project limited these resources.

Since the initial genome project, the Mammalian Gene Collection (MGC) project (34a, 34b, 50) has targeted the sequencing of 6,200 rat cDNAs [5,349 are available in GenBank (15a)], and the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) has produced ~3 million WGS sequence reads for an additional eight strains to discover sequence variation. In addition, Applied Biosystems

1 The online version of this article contains supplemental material.
has released ∼1× coverage of the genome in reads from the Sprague-Dawley (SD) strain. The description of the strains that have additional sequence data and the number of quantitative trait loci (QTLs) that have been described in each are given in Table 3, along with the number of sequence reads available and the number of single nucleotide polymorphisms (SNPs) derived from the sequence.

The rat research community has made extensive use of the genome sequence. In addition to its significance as an experimental model for human disease, the rat genome has had a major impact in the analysis of mammalian evolution (22, 33, 40, 41). The draft genome has received high marks for correcting existing sequences in addition to adding previously unknown sequences (11). A number of improvements were proposed and funded by the NHGRI to provide an even more complete genome with improved accuracy. This upgrade, the latest version of the Atlas assembler, targeted finishing of problematic regions, and a draft sequence of the Y chromosome, comprising 137,000 sequence contigs. The sequence contigs represent 92% of the regions spanned by the ultrabactigs. About 96% of this sequence is mapped on chromosomes; ½ of the remainder is of the regions spanned by the ultrabactigs. About 96% of this sequence is mapped on chromosomes; ½ of the remainder is unmapped, while ½ is mapped to a chromosome but unplaced.

The quality of the sequence is reflected in a number of measures. The assembly includes segmental duplications (defined as >5 kb and >90% identity) for 2.9% of its bases (41a, 52). Compared with finished sequence, individual contigs are by and large of finished quality (41a). Thus, the areas that remain to be addressed are regions that pose special problems to genome assembly. These include regions with unusual repeat structures, polymorphisms, possible BAC rearrangements, low sequence coverage (perhaps due to cloning bias), etc. A few specific cases have been pointed out by the rat research community and are being used to define the methods required to upgrade the assembly.

For example, Dr. Tim Aitman presented evidence of a duplication in the Fcgr3 locus implicated in autoimmune nephritis that was not represented in the rat genome assembly. The duplication appears in rat but not in the mouse genome. Fcgr3 has undergone at least two duplications since the divergence of the mouse and rat lineages, and the rat has at least three expressed genes. The syntenic region in humans also undergone copy number variation associated with disease, and the region is not well resolved in the human genome assembly. Resolving the differences between human, mouse, and rat is important for understanding both disease studies and mammalian evolution.

A second example region of a recent duplication in rats is the Cd36 region that is involved in fatty acid transport and linked to a number of cardiovascular and diabetes disease phenotypes (4). Comparison of finished sequence for human, mouse, and rat would be important for understanding complex diseases and their models. The genomic events of duplication and emergence of pseudogenes are paradigms for the effects of genome plasticity on evolution of complex phenotypes.

A third region of concern is the rat titan locus. Titin is the largest protein known to date, spanning >2 MB in the genome and showing elaborate alternative splicing.

A final example is a 5 MB region on chromosome 1 that is of interest for phenotypes of stroke (44), hypertension (16), and metabolic syndrome (27), with a number of candidate genes (P2ry2, P2Y, P2ry6, Pde2a, and Slco2b1). This region has repeat sequences that need to be resolved.

There were also anecdotal reports of regions in the assembly that were not merged due to apparent polymorphism. There were also a few possible regions of misassembly identified by...
RATS IN THE GENOMIC ERA

Female Genome Example
20 autosomes and X chromosome

1Mb
BACs
Fosmids
Plasmids

Sequence from each end, maintain pairing and insert size information

Overlaps - sequence similarity

Different clone insert sizes step across features of different sizes

Greater coverage improves consensus sequence assembly
Fig. 1.—Continued
comparison with the genetic maps (31, 54). Validation of the assembly in these regions would improve the quality and utility of both the sequence and the genetic map.

The later versions of the Atlas software have improved methods to deal with repeats and new modules for heterozygosity in highly polymorphic genomes (48) and more efficient handling of BAC sequences from clone pools. These methods are anticipated to improve the assembly of problematic regions.

Since the release of the v. 3.4 genome assembly, an additional 266 BACs spanning 100 Mb have been finished to human quality standards (47), and 598 BACs have been brought to an ENCODE ordered and oriented status (6) by the BCM-HGSC. A summary of the available finished sequence is given in Table 4, and details about each finished clone are given in Supplemental Table S1. These finished sequences are available for incorporation in an upgraded genome assembly.

The current assembly has additional room for improvement in regions that are not captured by the BAC tiling path.

Versions of the Atlas assembly software have been developed and used for WGS-only assemblies of honey bee (8), sea urchin (48), bovine, and macaque (41b). The macaque genome

Table 2. Published gene predictions in human, mouse, and rat genomes

<table>
<thead>
<tr>
<th>Genome Assembly</th>
<th>Human Draft</th>
<th>Mouse Draft</th>
<th>Rat Draft</th>
<th>Human Finished</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size, Gb</td>
<td>2.9</td>
<td>2.5</td>
<td>2.75</td>
<td>2.85</td>
</tr>
<tr>
<td>Predicted transcripts</td>
<td>44,860</td>
<td>29,201</td>
<td>28,516</td>
<td>34,214</td>
</tr>
<tr>
<td>Predicted genes</td>
<td>31,778</td>
<td>22,011</td>
<td>20,973</td>
<td>22,287</td>
</tr>
<tr>
<td>Total predicted exons</td>
<td>170,211</td>
<td>213,562</td>
<td>205,623</td>
<td>231,667</td>
</tr>
</tbody>
</table>

Table 3. Strains data summary

<table>
<thead>
<tr>
<th>Strain</th>
<th>Traces in NCBI Trace Archive, n</th>
<th>QTLs, n</th>
<th>SNPs Compared With Genome, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>261,294 (BCM)</td>
<td>87</td>
<td>61,425 Ensembl 9,690 dbSNP</td>
</tr>
<tr>
<td>DA</td>
<td>332,682 (BCM)</td>
<td>74</td>
<td>61,405 Ensembl 9,903 dbSNP</td>
</tr>
<tr>
<td>F344</td>
<td>353,337 (BCM)</td>
<td>112</td>
<td>84,908 Ensembl 14,187 dbSNP</td>
</tr>
<tr>
<td>FHH</td>
<td>401,952 (BCM)</td>
<td>4</td>
<td>19,716 Ensembl 3,204 dbSNP</td>
</tr>
<tr>
<td>GK*</td>
<td>76,110 (MPIMG)</td>
<td>15</td>
<td>121,058 Ensembl 16,218 dbSNP</td>
</tr>
<tr>
<td>LEW</td>
<td>322,426 (BCM)</td>
<td>76</td>
<td>97,424 Ensembl 15,865 dbSNP</td>
</tr>
<tr>
<td>PVG</td>
<td>272,117 (BCM)</td>
<td>1</td>
<td>61,476 Ensembl 9,872 dbSNP</td>
</tr>
<tr>
<td>SD</td>
<td>7,990,225 (CRA)</td>
<td>136</td>
<td>2,645,606 Ensembl 32,759 dbSNP</td>
</tr>
<tr>
<td>SHR‡</td>
<td>292,536 (BCM)</td>
<td>210</td>
<td>75,274 Ensembl 4,653 dbSNP</td>
</tr>
<tr>
<td>SS§</td>
<td>376,992 (BCM)</td>
<td>155</td>
<td>92,678 Ensembl 6,539 dbSNP</td>
</tr>
<tr>
<td>WKY†</td>
<td>49,790 (IMBGA)</td>
<td>94</td>
<td>99,319 Ensembl 10,511 dbSNP</td>
</tr>
<tr>
<td>Wild rat</td>
<td>27,793 (MPIMG)</td>
<td></td>
<td>19,339 Ensembl 3,124 dbSNP</td>
</tr>
<tr>
<td>BN</td>
<td>16,851,990 (BCM)</td>
<td></td>
<td>84,224 Ensembl 15,041 dbSNP</td>
</tr>
</tbody>
</table>

Table 4. Finished BAC clones§

<table>
<thead>
<tr>
<th>BAC type</th>
<th>O &amp; O*</th>
<th>Finished†</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCODE</td>
<td>56</td>
<td>209</td>
<td>265</td>
</tr>
<tr>
<td>ENCODE and QTL</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Gene rich</td>
<td>382</td>
<td>49</td>
<td>431</td>
</tr>
<tr>
<td>QTL‡</td>
<td>158</td>
<td>77</td>
<td>235</td>
</tr>
<tr>
<td>Total</td>
<td>598</td>
<td>337</td>
<td>935</td>
</tr>
</tbody>
</table>

*Ordered and Oriented contigs, still contains gaps. †Finished to human genome standards. ‡Three QTL clones were from the RP32 library (2 O&O, 1 Finished), the remainder were from the CH230 library. §Details including clone name, BCM project identifier, NCBI phase, reason for finishing, and NCBI accession number are listed for all O&O and finished BAC clones in Supplementary Table S1.

NCBI, National Center for Biotechnology Information; QTL, quantitative trait locus; SNP, single nucleotide polymorphism; BCM, Baylor College of Medicine Human Genome Sequencing Center; MPIMG, Max Planck Institute of Molecular Genetics; CRA, Celera Genomics; IMBGA, Institute of Molecular Biotechnology Genome Analysis; ABC, Agencourt Bioscience Corporation. *GK is GK/OX in TA; †WKY is WKY/BBB; ‡SHR is SHRSP/BBB; §§ is SS/JRNHSD.
comparisons of multiple assemblies demonstrated the value of combining assemblies to gain benefits from different methods (41b). The phase 2 genome will combine unique sequence data from a WGS-only assembly with the data from the combined WGS-plus-BAC assembly for a more complete representation of the genome.

During the initial genome project, the Y chromosome was sequenced to about twofold WGS coverage. One of the reasons for the low coverage was the unusually large size of the BN Y chromosome. The phase 2 genome upgrade includes higher level of sequencing on the Y chromosome using a rat strain with a more representatively sized Y chromosome. The sequencing strategy will combine WGS sequence and low coverage sequence skims from male BAC library clones.

Traditional finishing strategies use a BAC subclone as the substrate and layer additional Sanger sequencing reads from subclones or directed PCR products to increase coverage. Newer sequencing technologies (see review Ref. 5), such as those from Roche’s 454 Life Sciences, Illumina’s Solexa, and Applied Biosystems’ Solid, are available that produce large volumes of sequence data in a short period of time. Combining these high-volume sequence production methods with other methods for finishing such as those that use pooled BAC substrates or regions selected by array hybridization will make it possible to finish many more regions of the rat genome for substrates or regions selected by array hybridization will make methods for finishing such as those that use pooled BAC substrates or regions selected by array hybridization will make it possible to finish many more regions of the rat genome for the same price. These developments will make it possible to address many more of the remaining challenging regions and perhaps to finish the entire rat genome.

There are a large number of rat strains, most of which were developed as models for common complex human diseases (3). A dense radiation hybrid map with ~25,000 elements and the detailed comparative maps for rat, mouse, and human link the rat physiology and QTLs to mouse and human genetics. Current limits on QTL mapping in the rat are 2–10 cM; additional SNPs and haplotype maps will be required to reduce the size of these regions.

Although the rat does not yet have as many well-characterized phenotypes associated with single genes as does the mouse, it leads the way in known quantitative genetic traits. A number of rat QTLs have been mapped (e.g., copper metabolism, pituitary tumor growth, aerobic capacity, blood pressure and hypertension, ethanol tolerance, behavioral conditioning, anxiety, fat accumulation, chemical carcinogenesis, arthritis, diabetes, and cardiovascular disease), and in some cases (e.g., the last three listed) the genes have been cloned by position. Current SNP resources. There are 43,798 rat RefSNPs in dbSNP (37a), with 1,605 validated using functional assays. Dr. Norbert Hübner, funded through the EU STAR program, has generated 2,707,538 SNPs from cDNAs from four strains (SHR, WKY, GK, and SS). These SNPs, as well as the SNPs from the US NHGRI-funded SNP discovery effort (3), will be genotyped in an additional 50 strains through the EU STAR program funding. This genotyping will enable the development of a haplotype map. The previous rat SNP map based on cDNA from four strains (SHRSP, BN, WKY, and SD) has 12,395 interstrain polymorphic sites (55). The US NHGRI-funded SNP discovery effort (3) has sequenced eight strains (PVG, F344, SS, LEW, BB, FHH, DA, and SHR). The available sequence data is summarized in Table 3. The estimated yield for these strains is 1 SNP per 500–600 bp.

Conclusions

The phase 2 upgrade to the genomic sequence will provide a more complete and accurate genome sequence that will, in combination with the additional markers from the SNP discovery project, improve gene hunting via improved correlations between phenotypic data and ancestral sequence origin across many inbred strains and the identification of shared segments in rat strains used for intercross/backcross experiments. In additional to these rat-specific resources, general improvement in genomic tools through the advanced resequencing technologies will allow rapid gene targeted analyses. All of these developments will continue to push the translation of rat models of disease to better understand human health.

GRANTS

The phase 2 genome project and the SNP discovery project are funded by NHGRI Grant 2 U54 HG-003273.

REFERENCES


