Upgrading the DNA Sequence of the Rat Genome

Richard Gibbs and George Weinstock
Baylor College of Medicine Human Genome Sequencing Center

The Brown Norway or laboratory rat (*Rattus norvegicus*) genome was sequenced in a project jointly supported by the NHGRI and the NHLBI (12). This was the third mammalian project undertaken by the NIH but the first that would only be taken to the draft stage. The project was a complex collaboration led by the BCM-HGSC (BAC skims, wgs, assembly, annotation, overall coordination) and including Celera Genomics Inc. (wgs reads), Genome Therapeutics Inc. (wgs reads, Y chromosome), University of Utah (wgs reads), TIGR (BAC end sequences), British Columbia Genome Sciences Centre (BAC fpc map), NISC-NHGRI (BAC sequencing), Washington University Genome Sequencing Center (BAC fpc map), and Children's Hospital Oakland Research Institute (BAC libraries). Annotation and analysis was performed by an international consortium of over 70 analysts.

Because the product was a draft sequence, there was no opportunity to correct errors during a finishing phase. Thus considerable effort was expended to increase the quality of the draft product. A new combined BAC skim - Whole Genome Shotgun strategy was developed to improve on the quality of the draft products of pure WGS approaches, and a new genome assembly program, Atlas (5), was developed to marry these two types of data into the final sequence.

The initial draft genome was examined in microscopic detail by an analysis group comprising a broad selection of disciplines focusing on genomic and proteomic aspects, from the genome-wide level to individual genes. This initial assembly was found adequate for this analysis and the goals of the project were reached. There has only been one significant upgrade to the sequence since then, which involved replacement of draft with finished sequence when it was available from BACs (e.g. from ENCODE regions). One disappointing aspect of this project was the lack of support for selective finishing of important or difficult-to-assemble regions. Thus there were no finished regions produced.

We also sought resources to improve the availability of full length cDNA sequences and of SNPs for the rat. Neither were forthcoming at the time, however the MGC project has since targeted several thousand rat cDNAs and more recently the BCM-HGSC has been 'approved' to target 3 million DNA sequence reads for an additional 8 rat strains in order to discover variation. In addition Applied Biosystems has now released approximately 1 X coverage of the genome in reads from an additional strain. This information will greatly benefit the use of rats in genetic studies.

In the year and a half since its publication the rat research community has used the sequence extensively. There is little need to detail the significance of the rat as an experimental system for human medical research, or the impact of the rat genome in the unfolding analysis of mammalian evolution (4, 8-10). In general the draft genome receives high marks not only providing hitherto unknown sequences but also correcting existing sequences (2). Despite the overall success of the project, there are a number of areas where improvements can have more than incremental impact, and we propose to address those next. The result will be an even more complete genome, with the high standard of accuracy extended into important regions that are currently problematic.

The various activities we propose are:

(1) reassembly of the genome with the latest version of the Atlas assembler
(2) targeted finishing of problematic regions
(3) draft sequence of the Y chromosome

Currently we work with the Rat Genome Database and provide an annual incremental update of the sequence based on information collected over the previous year. This is an inefficient and tedious approach to upgrading the genome sequence. It would be preferable to do the majority of this work in a single short-term project as proposed here.

The BCM-HGSC has also been given approval to generate a collection of SNPs from eight rat strains. This information will benefit rat genetic methodology and the upgraded sequence will further enhance its value.

We envisage that the long term aim of all the groups producing whole genome assemblies is to develop an objective set of standards and language by which any project can be assessed and described. We believe that the upgrading of the rat genome as described here will facilitate that by providing the highest possible quality data for a portion of the genome, so it can be compared at draft, pre-finished and finished stages. When knowledge from rat and other upgraded genomes are added to that from the human and mouse genomes, we may reach these goals.

The current rat assembly (RNOR 3.4)

Statistics of the current rat assembly are shown in the two tables below. The draft genome is 419 ultrabactigs (regions defined by a path of BACs) comprising 137,000 sequence contigs. The sequence contigs amount to 92% of the regions spanned by the ultrabactigs. About 96% of this sequence is mapped to chromosomes, the majority of the remainder being unmapped. Beyond these coarse statistics, the assembly includes 2.9% of its bases in segmental duplications (>5kb, >90% identity) (12, 13).

	Number	Average_Size	N50_Size	Total_Size
UltraBactigs	419	6.54 MB	19.0 MB	2.80 GB
Contigs	137,000	18.7 KB	37.2 KB	2.57 GB

	Sum_of_Contigs	Contigs+Gaps
Unmapped	62.2 MB	82.7 MB
Random_on_Chromosomes	27.6 MB	32.5 MB
Mapped_on_Chromosomes	2.48 GB	2.72 GB
Total sequence	2.57 GB	2.72 GB

As measured by comparison with finished sequence, contigs are mainly at the finished quality standard (12). Thus, what remains to be addressed are regions that pose special problems to genome assembly. These include regions with unusual repeat structures, with polymorphisms, or with problems whose origin are not known but could result from rearrangements in BACs, low coverage due to cloning biases, etc. By informal communication with the rat research community we have learned of some of these cases, described below, and they help to define how to upgrade the current assembly.

The Atlas assembly program has itself undergone considerable evolution since the initial version that was used to assemble the rat genome. For example, in the rat project when inclusion of sequences in the final assembly was tracked, about 4% of 32-mer sequences were not included, and as much as 18% of 32-mers that were present at copy numbers of 2-10 in the genome were not included. The algorithm for dealing with repeats has been improved since the initial version (the ‘overlapper’ and ‘binner’ components of the software suite), a result of adapting the program to do pure WGS assemblies, so the program is much more efficient in dealing with low copy repeats. In addition, new modules have been implemented in Atlas to address heterozygosity in highly polymorphic genomes (e.g. sea urchin) and more efficient handling of BAC skims (for our approach to BAC skims using clone pools). These and other improvements to Atlas are expected to more effectively assemble problem areas.

Examples of problem regions

Regions with unusual repeat structures are difficult to assemble and often require finishing to resolve their correct sequence. One example of such a region in the rat is the locus for the activatory Fc receptor FcgRIII, Fcgr3, implicated in autoimmune nephritis in a rat model and humans. Tim Aitman has evidence of duplication in this region (genomic and BAC Southerns, clonotype analysis) that was not represented in the genome assembly. The duplication appears in the rat but not the mouse genome. Fcgr3 has undergone at least two duplications since divergence of the rat and mouse lineages and the rat has at least 3 expressed genes. The region of duplication and deletion extends over at least 30 Kb, and more likely 100-150 Kb of genomic sequence around Fcgr3. Since this is a recent (since rat-mouse divergence) duplication, the genes are nearly identical in sequence.

According to Dr. Aitman, the syntenic region in humans also has undergone duplication deletion but this has not yet been resolved on the human genome assembly. In both rat and human, copy number is associated with and is probably a cause of a disease phenotype. Clearly, resolving the differences between human, mouse and rat is important for both disease studies and understanding mammalian evolution.

Another example is the region of duplication around rat Cd36 (involved in fatty acid transport with a number of possible disease phenotypes from cardiovascular to diabetes). Once again there is a recent duplication in rats since the mouse divergence (1), with one expressed gene and two pseudogenes. There is a single Cd36 gene in mouse and humans. Comparison of finished sequence for rat, mouse and human would be important for understanding complex diseases and their models. These genomic events could serve as paradigms for the effects of genome plasticity on evolution of genetically complex mammalian phenotypes.

There is concern that the rat titin locus is not properly assembled. It is the biggest protein known so far, spanning a genomic region of more than 2 MB and showing elaborate alternative splicing. This gene is under intense investigation for its role in several heart disorders and also as a signaling molecule.

A final example is a 5 MB region on chromosome 1 that is of interest to a many investigators, yet has need of repeated sequence resolution. In depth sequencing of this region would help analysis of a range of phenotypes including hypertension and stroke. The region encompasses a number of disorders and risk factors including stroke (11), hypertension (3), and metabolic syndrome (6) with candidate genes including P2ry2 (purinergic receptor P2Y, G-protein coupled), Pde2a (phosphodiesterase 2A, cGMP-stimulated), P2ry6 (pyrimidinergic receptor P2Y, G-protein coupled, 6), and Slco2b1 (solute carrier organic anion transporter family, member 2b1).

Regions that overlap but were not joined owing to apparent polymorphism were observed anecdotally in the assembly despite the inbred strain used. Such overlapping haplotypes create artificial duplications that can result in misleading interpretations of gene families. As noted above, a new module of Atlas was designed specifically to assemble these regions.

Other possible regions of misassembly are identified by comparison of the genetic and sequence maps (7, 14). There does not appear to be more than a few such regions per chromosome, and it is not clear if the inconsistency is due to the genetic map or sequence assembly. Nevertheless, detailed validation of the assembly in these regions would contribute to the overall quality and utility of both the sequence and genetic map.

Y chromosome

The original approved plan for sequencing the rat genome included a draft assembly of the Y chromosome, but this was not achieved. The WGS component of the rat project was performed with a female DNA source to maintain even stoichiometry between chromosomes. Next it was planned to purify the Y chromosome by flow cytometry in a collaboration between Los Alamos National Laboratory and Genome Therapeutics, who would then provide the WGS reads. Preliminary sorts by the Los Alamos laboratory however revealed no chromosome of the predicted size – instead a new peak 1.8 x larger than expected was seen in male cell lines with dye-binding characteristics similar to the characterized Y peaks. PCR analysis with Y-specific probes indicated this was likely the Brown Norway rat Y chromosome. Subsequent investigation indicated that the BN Y chromosome was the largest among 17 rat strains. Consistent with an enlarged Y, Monte Turner found the Sry locus from the BN rat contains 6 copies of the Sry gene, as contrasted to one in human.

Subsequently, the Y was sorted from the other chromosomes at Los Alamos. However the amount of material was only sufficient for low coverage and the project had since moved from Genome Therapeutics to Agencourt. About 2x WGS coverage was produced.

Given that there appear to be no technical limitations on the sequencing the Y, and this had been approved for the original project, we propose to produce a draft sequence as part of the upgrading of the genome.

Proposed activities

(1) Assembly of the rat genome with the current version of Atlas. As mentioned above, the current version of the Atlas assembly suite is a significantly more mature program than was used for the rat project. The current version has been used successfully for the genomes of *Drosophila pseudoobscura*, the honey bee, the sea urchin, the cow, and the macaque and each of these projects have contributed to the Atlas refinement. Before embarking on the upgrading activities described below*, it will be advantageous to start with the best draft assembly that can be produced*. We believe this will be readily accomplished with the current version of Atlas.

The data for this process is available in the NCBI trace archive. Part of the process will include comparing the RNOR 3.4 assembly versus the new assembly in both the problem regions that have been described to us (detailed above), in the same finished regions that were used to benchmark the original assembly (12), as well as in genome-wide comparisons. We are currently involved in comparing multiple assemblies of the macaque genome from different centers, so by the time the rat upgrade is performed we should have established paths to follow in these comparisons.

We also note that Celera recently released their data for wgs sequencing of a different strain of rat. We will investigate whether addition of these reads to the assembly improves the quality or if strain differences reduce the overall utility of the mixed data.

(2) Resolution of problem regions. The most direct approach to resolve the problem areas described above will be to finish them. This will follow our standard finishing pipeline, used for human and mouse and rat ENCODE regions. As targets for finishing we will solicit input from the rat research community, but also identify possible problem areas by map comparison, by identifying sequences have not been mapped (e.g. that may map to multiple loci possibly due to misassembly), by identifying genes that are incomplete (with respect to mouse and human), by identifying low coverage regions of the genome, and by other methods suggested by the community. Analysis of regions in the rat genome that are breakpoints in chromosome evolution will also be considered. We anticipate that not more than 100 MB (3% of the genome) would need to be finished.

(3) Sequence of the Y chromosome. This will be performed by generating WGS reads from flow sorted chromosomes, as well as by BAC sequence skimming using a male rat BAC library available from Pieter DeJong. Flow sorting will either be performed at Los Alamos or at a new facility being developed at the NCI (Frederick) by William Modi. The project will require about 600,000 reads. Following sequencing and assembly, an annotation group will be convened for analysis.

Budget and timeframe

We anticipate the project can be completed in one year.

Costs:
Assembly and informatics for updating annotation and submission: $50,000
Finishing 100MB: $3M
Y chromosome: $0.5M

Total: $3,550,000

Literature Cited

1. **Aitman, T. J., A. M. Glazier, C. A. Wallace, L. D. Cooper, P. J. Norsworthy, F. N. Wahid, K. M. Al-Majali, P. M. Trembling, C. J. Mann, C. C. Shoulders, D. Graf, E. St Lezin, T. W. Kurtz, V. Kren, M. Pravenec, A. Ibrahimi, N. A. Abumrad, L. W. Stanton, and J. Scott.** 1999. Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. Nat Genet **21:**76-83.
2. **Daniels, D., A. Suzuki, E. Shapiro, L. Luo, D. K. Yee, and S. J. Fluharty.** 2005. Rattus norvegicus melanocortin 3 receptor: A corrected sequence. Peptides.
3. **Frantz, S., J. R. Clemitson, M. T. Bihoreau, D. Gauguier, and N. J. Samani.** 2001. Genetic dissection of region around the Sa gene on rat chromosome 1: evidence for multiple loci affecting blood pressure. Hypertension **38:**216-21.
4. **Hancock, J. M.** 2004. A bigger mouse? The rat genome unveiled. Bioessays **26:**1039-42.
5. **Havlak, P., R. Chen, K. J. Durbin, A. Egan, Y. Ren, X. Z. Song, G. M. Weinstock, and R. A. Gibbs.** 2004. The Atlas genome assembly system. Genome Res **14:**721-32.
6. **Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Muller, S. A. Cook, T. W. Kurtz, J. Whittaker, M. Pravenec, and T. J. Aitman.** 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet **37:**243-53.
7. **Krzywinski, M., J. Wallis, C. Gosele, I. Bosdet, R. Chiu, T. Graves, O. Hummel, D. Layman, C. Mathewson, N. Wye, B. Zhu, D. Albracht, J. Asano, S. Barber, M. Brown-John, S. Chan, S. Chand, A. Cloutier, J. Davito, C. Fjell, T. Gaige, D. Ganten, N. Girn, K. Guggenheimer, H. Himmelbauer, T. Kreitler, S. Leach, D. Lee, H. Lehrach, M. Mayo, K. Mead, T. Olson, P. Pandoh, A. L. Prabhu, H. Shin, S. Tanzer, J. Thompson, M. Tsai, J. Walker, G. Yang, M. Sekhon, L. Hillier, H. Zimdahl, A. Marziali, K. Osoegawa, S. Zhao, A. Siddiqui, P. J. de Jong, W. Warren, E. Mardis, J. D. McPherson, R. Wilson, N. Hubner, S. Jones, M. Marra, and J. Schein.** 2004. Integrated and sequence-ordered BAC- and YAC-based physical maps for the rat genome. Genome Res **14:**766-79.
8. **Lindblad-Toh, K.** 2004. Genome sequencing: three's company. Nature **428:**475-6.
9. **Pennisi, E.** 2004. Genomics. New sequence boosts rats' research appeal. Science **303:**455-8.
10. **Petit-Zeman, S.** 2004. Rat genome sequence reignites preclinical model debate. Nat Rev Drug Discov **3:**287-8.
11. **Rubattu, S., M. Volpe, R. Kreutz, U. Ganten, D. Ganten, and K. Lindpaintner.** 1996.

Chromosomal mapping of quantitative trait loci contributing to stroke in a rat model of complex human disease. Nat Genet **13:**429-34.

12. **The Rat Genome Sequencing Consortium.** 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428:**493-521.
13. **Tuzun, E., J. A. Bailey, and E. E. Eichler.** 2004. Recent segmental duplications in the working draft assembly of the brown Norway rat. Genome Res **14:**493-506.
14. **Wilder, S. P., M. T. Bihoreau, K. Argoud, T. K. Watanabe, M. Lathrop, and D. Gauguier.** 2004. Integration of the rat recombination and EST maps in the rat genomic sequence and comparative mapping analysis with the mouse genome. Genome Res **14:**758-65.