Minireview

Insights from the rat genome sequence

Linda J Mullins and John J Mullins

Address: Molecular Physiology Laboratory, Wilkie Building, Teviot Place, University of Edinburgh Medical School, Edinburgh EH8 9AG, UK.

Correspondence: John J Mullins. E-mail: j.mullins@ed.ac.uk

Published: 30 April 2004

Genome ***Biology*** 2004, **5:**221

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/221

Abstract

The availability of the rat genome sequence, and detailed three-way comparison of the rat, mouse and human genomes, is revealing a great deal about mammalian genome evolution. Together with recent developments in cloning technologies, this heralds an important phase in rat research.

Historically, the rat has been the animal model of choice for research in key areas that inform human medicine, such as cardiovascular biology, neurobiology and nutrition. A huge body of knowledge has accumulated and over 230 disease models have been generated through selective breeding. The rat is also an indispensable tool in drug development, both for the assessment of therapeutic efficacy and for toxicity trials. The mouse has largely usurped the rat as the species of choice in biomedical research in general, however, because of its size, fecundity and ease of genetic manipulation - especially with the development of gene knock-out technology. After human, the mouse was the obvious next choice for whole-genome sequencing, and there was a naive belief that the rat genome sequence would prove to be redundant, given the morphological and evolutionary similarity between the rat and the mouse.

The recent publication in *Nature* of an initial rat genome-sequence analysis [1] has gone a long way to silence doubters and to inspire the rat research community. Using a combination of random whole-genome shotgun sequencing and a bacterial artificial chromosome (BAC) contig-building approach, a high-quality draft of the Brown Norway rat sequence, covering over 90% of the genome, has been achieved by the Rat Genome Sequencing Project Consortium - led by the Baylor College of Medicine (Houston, USA), and including Celera Genomics (Rockville, USA), Genome Therapeutics (Waltham, USA) and many academic centers worldwide. A three-way comparison of the rat sequence with the human and mouse genomes has revealed a great deal of new information about mammalian genome evolution. The rat genome (2.75 gigabases, Gb) is smaller than the human genome (2.9 Gb) but larger than that of the mouse (2.6 Gb). Global comparison of the three genomes reveals large chromosomal regions, referred to as orthologous chromosomal segments, which have been inherited with minimal rearrangement of gene order from the primate-rodent ancestor. These intact regions have become interspersed during large-scale chromosomal rearrangements since the separation of primate and murid ancestors approximately 75 million years ago, and since the split between rat and mouse 12-24 million years ago. Comparison of present day chromosomal configurations allows one tentatively to reconstruct the sequence and timing of the rearrangements, and confirms that the rate of rearrangements in murid rodents is much higher than in the primate lineage.

Large segmental repeats make up about 3% of the rat genome, a value intermediate between the mouse (1-2%) and human genomes (5-6%). These duplicated regions are enriched near telomeres and centromeric regions, and are associated with the recent expansion of major gene families. About 40% of the euchromatic rat genome aligns with both mouse and human sequences and thus represents the ancestral core; this core contains about 95% of the known coding exons and non-coding regulatory regions, both of which characteristically accumulate substitutions at a slower rate than 'neutral' DNA, indicating their critical role. Conservation

within three mammalian genomes has proved to be extremely useful for identifying non-coding regulatory elements, including transcription-factor binding sites and locus-control regions. Searching the human genome for 109 transcription-factor binding sites revealed over 186,000,000 potential sites. When conservation between the three genomes was a pre-requisite for a potential site, however, the number was reduced to 4,000,000, representing a 44-fold increase in specificity. Such analyses should aid in the location of enhancer sequences, boundary elements, and perhaps even matrix-attachment sites (at which DNA is thought to bind to chromosomal scaffolds). Given the long distances over which control elements act, apparent 'gene deserts' (gene-poor regions larger than 500 kilobases), which make up approximately 25% of the human genome, may prove to be a fertile source of important gene-regulatory elements [2].

The three mammalian genomes contain multiple copies of immobilized transposable elements, which constitute 40% of the mouse and rat genomes, and almost 50% of the human [1]. The long interspersed nucleotide element LINE-1 was active before the rodent-primate split, and over half a million copies, in various stages of decay, can still be recognized in the rat. Since the rat-mouse split, the L1 retrotransposon has remained active, and represents 12% of the rat genome and 10% of the mouse genome. Looking at rat euchromatin, 28% aligns only with mouse, and 40% of this consists of rodent-specific repeats, such as B2 SINEs (short interspersed nucleotide elements), which are still active, and the extinct B4 element. The Alu-like B1 element is still active in the mouse but probably became extinct in the rat soon after the mouse-rat split. On the other hand, the ID element, which is relatively minor in the mouse, is present in over 160,000 copies in the rat. The remainder of the euchromatic rat genome includes rat-specific repeats or rodent-specific repeats that have been lost from the mouse genome.

Rodent lineages have acquired more genomic changes than primates, including a three-fold higher rate of base substitution in neutral DNA. Interestingly, the rate of base substitution is 5-10% higher in the rat than the mouse branch, leading to a relative increase in GC content in the rat; and the rat has also accumulated microdeletions more rapidly than the mouse. Such biochemical changes may reflect increased recombination rates, and differences in repair and replication enzymes. One particular type of non-coding sequence, namely pseudogenes, is not subject to selective constraint, so pseudogenes accumulate sequence modifications neutrally. Approximately 20,000 pseudogenes were identified in the rat genome, a similar number to that found in human and mouse. The largest groups of pseudogenes have arisen from ribosomal-protein genes, olfactory receptors, glyceraldehyde 3-phosphate dehydrogenase, protein kinases and RNA-binding RNP-1 proteins. A large proportion of the pseudogenes (80%) is not found in human-rat syntenic regions and are probably retrotransposed and processed. In addition, when looking at coding sequences, analysis of in-frame changes to proteins suggests that trinucleotide repeats accumulated more often in secreted and nuclear proteins, transcription regulators and ligand-binding proteins, than in cytoplasmic and mitochondrial proteins. Transmembrane domains were found to be particularly refractory to trinucleotide accumulation (six-fold lower than would be expected if due to chance).

The three mammalian genomes have been predicted to encode similar numbers of genes [1] and it is estimated that 90% of rat genes have orthologs in the mouse and human genomes that have persisted since they shared a common ancestor. The remaining genes are associated with gene-family expansions - a major source of genetic differences between the rat and the mouse - reflecting differences in chemosensation and aspects of reproduction. Detailed analysis of olfactory receptors, for example, indicates that the rat contains a potential olfactory repertoire of around 1,400 proteins. This is significantly more than the approximately 1,200 mouse olfactory receptors, although any functional implications for the animals' relative ability to discriminate odorants are not known at present. Another class of odorant-binding proteins, the α_{2u}-globulin pheromone-binding proteins have also undergone gene-family expansion. The orthologous human genomic region possesses a single homolog, probably mirroring the common rodent-primate ancestor, while the C57BL/6J mouse has four homologous genes (the major urinary proteins, MUPs) and seven pseudogenes. The rat genome contains 10 α_{2u}-globulin genes and 12 pseudogenes in one of several gene clusters, which have arisen by gene duplication since the rat-mouse split. Rapid evolution has also been observed in protease and protease-inhibitor genes, and also in the cytochrome P450 family of proteins. The latter are involved in the metabolism of both endogenous and toxic compounds. Given that rats are an important model for human drug metabolism and toxicity trials, it is essential to be aware of this species-specific variation in P450 subfamilies because it may have a significant bearing on such trials.

More than 1,000 human disorders that show Mendelian inheritance have been associated with specific gene loci, and these were compared with predicted rat genes. For over 75% of the disease genes, a 1:1 rat ortholog was predicted by Ensembl [3], and of the remaining 25%, the vast majority had likely orthologs among genomic, cDNA, expressed sequence tag (EST) and protein sequences. This suggests that, as a class, disease genes have been highly conserved since the rodent-primate split. When the genes were grouped by disease type, the neurological gene set exhibited fewer non-synonymous base substitutions than neutral DNA (suggesting the presence of selective constraints), whereas genes whose associated disease was classed as pulmonary, hematological or immunological manifested higher non-synonymous base substitution rates than neutral DNA, indicating positive selection or reduced constraints. These differences reflect

different evolutionary rates for the various disease systems. Multigenic disorders are investigated in humans using association studies and linkage analyses. With better definition of syntenic boundaries as a result of comparative genomic analysis, it may now be possible to narrow down the identity of candidate gene(s) and/or functional non-coding sequences within quantitative-trait loci. Recently developed consomic rat lines, in which an entire chromosome from one inbred strain is introgressed onto the background of a second inbred strain [4], congenic strains and recombinant inbred strains [5] can all be used to complement these studies, as can microarray technology.

Finally, one should consider the problem of assigning a function to all the genes identified by genome sequencing. One of the most effective means of determining gene function is by a targeted knock-out of the gene. Although this technology has proved elusive in the rat, random mutagenesis, by treatment of adult rats with ethylnitrosourea (ENU) [6], has successfully generated 'knock-out' rats. Mutations in target genes were identified in the F1 offspring of the ENU-treated adults using PCR combined with a yeast selection assay. Even more encouraging is the report that blastocyst-derived cells resembling rat embryonic stem (ES) cells have been maintained in culture for over 50 passages [7]. This is certainly long enough for targeted genetic modifications to be introduced, and given recent advances in nuclear transfer in the rat [8], the possibility of gene targeting by a combination of these techniques is tantalizingly close. In conclusion, the rat genome sequence is already proving its worth! The rat is not just a big mouse - it can now begin to take its rightful place in functional genomics and integrative physiology.

References

1. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, *et al.*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428:**493-521.
2. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers.** *Science* 2003, **302:**413.
3. **Ensembl Genome Browser** [http://www.ensembl.org/]
4. Cowley AW Jr, Roman RJ, Jacob HJ: **Application of chromosomal substitution techniques in gene-function discovery.** *J Physiol* 2004, **554:**46-55.
5. Printz MP, Jirout M, Jaworski R, Alemayehu A, Kren V: **Genetic models in applied physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics.** *J Appl Physiol* 2003, **94:**2510-2522.
6. Zan Y, Haag JD, Chen KS, Shepel LA, Wigington D, Wang YR, Hu R, Lopez-Guajardo CC, Brose HL, Porter KI, *et al.*: **Production of knockout rats using ENU mutagenesis and a yeast-based screening assay.** *Nat Biotechnol* 2003, **21:**645-651.
7. Buehr M, Nichols J, Stenhouse F, Mountford P, Greenhalgh CJ, Kantachuvesiri S, Brooker G, Mullins J, Smith AG: **Rapid loss of oct-4 and pluripotency in cultured rodent blastocysts and derivative cell lines.** *Biol Reprod* 2003, **68:**222-229.
8. Roh S, Guo J, Malakooti N, Morrison J, Trounson A, Du Z: **Birth of rats by nuclear transplantation using 2-cell stage embryo as donor nucleus and recipient cytoplasm.** *Theriogenology* 2003, **59:**283.