Gene Nomenclature by Default, or Blasting to Babel

*David R. Nelson**

The University of Tennessee Health Sciences Center, Memphis TN 38163, USA
* *Correspondence to*: Tel: +1 901 448 8303; Fax: +1 901 448-7360; E-mail: dnelson@utmem.edu

Date received (in revised form): 11th May 2005

Abstract

The current proliferation of mammalian genomes is creating a nomenclature issue caused by naming genes based on their best BLAST hit to a gene in another annotated genome. The rat genome is relying heavily on the mouse genome for nomenclature, but not all rat genes have direct orthologues in the mouse; often, there are paralogous groups of genes — due to expansions of that gene subfamily in one or the other genome. Many of these genes have already been assigned names in the rat, so that renaming them based on BLAST scores leads to duplicate sets of names. The supposed orthology created by name sharing across genomes is not always true. These inaccurate names are appearing in frequently used sites, such as the University of California Santa Cruz Genome Browser. The example of rat cytochrome P450 (*Cyp*) genes is presented here, but other gene families are also likely to be affected.

Keywords: *gene nomenclature, cytochrome P450, rat genome, mouse genome, orthologue*

The rat genome has been sequenced and assembled,[1] creating a need for rat gene nomenclature. The obvious source of gene nomenclature for the rat would seem to be the mouse genome. Ideally, orthologues should have the same name. This logic has led to an automated naming of rat genes — and this runs into problems of two kinds. First, the rat has long been an experimental animal. Genes from both rat and mouse were sequenced and named for nearly 20 years before the genomes were sequenced. In the example of cytochromes P450, the first mammalian sequences *Cyp2b1* and *Cyp2b2* were determined in the rat.[2] The mouse sequences began to appear two years later with *Cyp1a1* and *Cyp1a2*.[3,4] The established nomenclature for *CYP* genes has been in place since 1987,[5–7] and these names have been used in publications for several years. Because the names were assigned independently, mostly in chronological order, orthologues do not always carry the same name.

The second nomenclature problem has to do with divergence over time between species' genomes. Here, mouse and rat will be discussed, but the same applies to other species such as human and rhesus monkey. When similar genes appear in gene clusters, the one to one relationship of the genes between mouse and rat is often broken, meaning that the orthology is broken. Compared with the 57 *CYP* genes of the human, the mouse has greatly expanded its set of *Cyp* genes to 102 full-length genes;[8] the rat has been a little more conservative, with 87 *Cyp* genes.[9] The solo genes in a mammalian *CYP* subfamily, those that occur without related neighbours, are strict orthologues, and nomenclature by best reciprocal BLAST hit between mouse and rat is a viable strategy. This works for 31 mouse–rat gene pairs and one pseudogene. Eight-seven rat genes cannot match up to 102 mouse genes as orthologue pairs, however, and this nomenclature method can be seen to fail in the gene clusters.

Not all *Cyp* gene clusters are disordered between mouse and rat. For example, the *Cyp4f* gene cluster has nine genes in both species, and there is a clean 1:1 mapping between orthologous pairs (Table 1). In fact, there are 33 such pairs in the *Cyp* gene clusters (Table 1); two of these pairs involve matches to pseudogenes in the other species. After these 64 pairs are subtracted and a correction is made in the count for pseudogenes, there are still 40 mouse genes remaining to pair with 24 rat genes. These genes either have no orthologues (paired with an 'x' in Table 1) or they are in paralogous gene sets (shaded and boldened in Table 1).

The discrepancy is explained by extra duplications, mostly in the mouse, but in at least some cases, there is duplication in the rat that is not seen in the mouse. Figures 1 and 2 illustrate in detail two such gene clusters compared between rat and mouse. The *Cyp2d* cluster shows five genes in the rat and two pseudogenes. The mouse has nine genes and 17 pseudogenes. *Cyp2d5* and *Cyp2d1* in the rat are most similar by BLAST

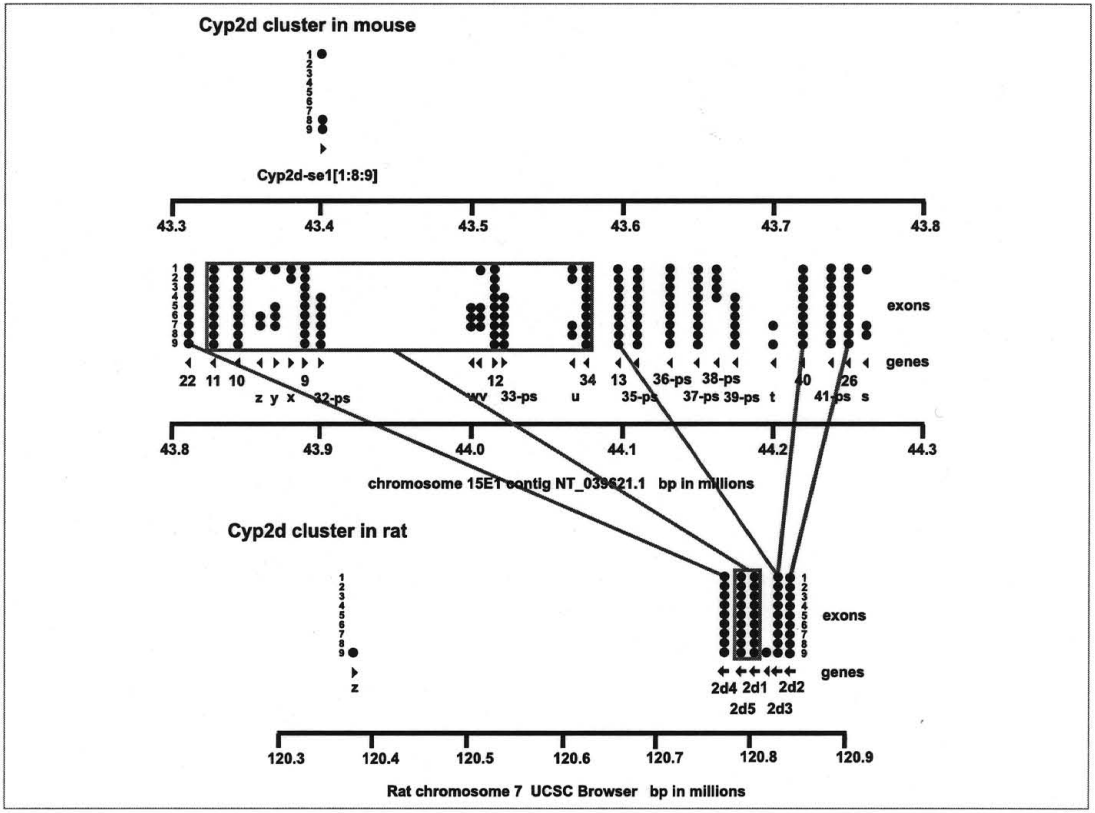
Table 1. Orthology between mouse and rat *Cyp* genes. Paralogues are boldened and yellow shaded.

Mouse	Rat	Mouse	Rat	Mouse	Rat
1a1	*1a1*	*2e1*	*2e1*	*4f13*	*4f6*
1a2	*1a2*	*2f2*	*2f4*	*4f14*	*4f1*
1b1	*1b1*	*2g1*	*2g1*	*4f15*	*4f4*
			4f16	*4f5*	
2a4	x	*2j5*	*2j5-ps*	*4f17*	*4f17*
2a5	*2a3*	*2j6*	*2j4*	*4f18*	*4f18*
2a12	*2a2*	***2j7***		*4f37*	*4f37*
2a22	*2a1*	***2j8***	***2j16***	*4f39*	*4f39*
		2j11		*4f40*	*4f40*
2b9	*2b3*	*2j9*	*2j3*		
2b10	*2b1*	*2j12*	*2j10*	*4v3*	*4v?*
2b13	*2b2*	*2j13*	*2j13*	*4x1*	*4x1*
	2b12				
2b19	***2b15***	*2r1*	*2r1*	*5a1*	*5a1*
	2b31	*2s1*	*2s1*	*7a1*	*7a1*
2b23	*2b21*	*2t4*	*2t1*	*7b1*	*7b1*
		2u1	*2u1*	*8a1*	*8a1*
2c37	x	*2w1*	*2w1*	*8b1*	*8b1*
2c29		*2ab1*	*2ab1*	*11a1*	*11a1*
2c38	***2c6***	*2ac1-ps*	*2ac1*	*11b1*	*11b1*
2c39	***2c7***			*11b2*	*11b2*
2c44	*2c23*	***3a11***		x	*11b3*
2c50	x	***3a16***	***3a1/3a23***	*17a1*	*17a1*
2c52-ps	*2c11*	***3a41***	***3a2***	*19a1*	*19a1*
2c54	x	***3a44***	***3a73***	*20a1*	*20a1*
2c55	***2c24***	*3a13*	*3a9*	*21a1*	*21a1*
	2c80	***3a25***		*24a1*	*24a1*
		3a57	***3a18***	*26a1*	*26a1*
2c65		***3a59***		*26b1*	*26b1*
2c66	***2c78***	x	*3a62*	*26c1*	*26c1*
2c70	*2c22*			*27a1*	*27a1*
2c40		***4a12a***		*27b1*	*27b1*
2c67	***2c12***	***4a12b***	***4a8***	*39a1*	*39a1*

(continued)

Table 1. *Continued*

Mouse	Rat	Mouse	Rat	Mouse	Rat
2c68	2c13			46a1	46a1
2c69		4a29	x	51a1	51a1
		4a30b	x		
2d9					
2d10		4a14	4a2		
2d11	2d1		4a3		
2d12	2d5				
2d34		4a10			
2d22	2d4	4a31	4a1		
2d26	2d2	4a32			
2d13					
2d40	2d3	4b1	4b1		



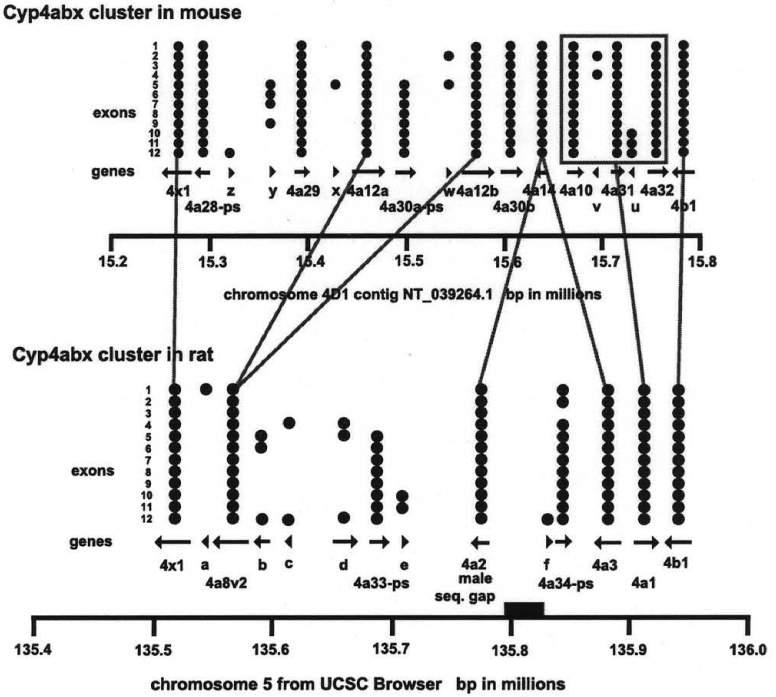


Figure 2. The *Cyp4abx* locus in the rat and the mouse. There has been differential expansion in both species. Strict orthology does not exist, except on the outer edges of the cluster. This is often a feature of genes in gene clusters: the edges of the cluster are more likely to be conserved.

searches to the five mouse genes — *Cyp2d11, Cyp2d10, Cyp2d9, Cyp2d12* and *Cyp2d34* — that are boxed in the mouse cluster; these represent paralogous sets of genes. Mouse *Cyp2d13* and *Cyp2d40* are about equally similar to *Cyp2d3* in the rat; in between these genes there are six pseudogenes. This whole cluster of genes and pseudogenes may have been derived from a *Cyp2d3*-like ancestor that expanded in the mouse. Of course, more complicated scenarios are also possible.

Figure 2 shows the *Cyp4abx* clusters. Notice how the rat *Cyp4a1* gene has given rise to three *Cyp4a* genes in the mouse. By contrast, mouse *Cyp4a14* has duplicated to make *Cyp4a2* and *Cyp4a3* in the rat, based on BLAST similarities. The mouse cluster is further complicated by an approximately 100 kilobase duplication involving the *Cyp4a12* and *Cyp4a30* genes. This did not happen in the rat and there does not seem to be a *Cyp4a30* equivalent in the rat — unless it might be the rat *Cyp4a33-ps* pseudogene. There are seven *Cyp* gene clusters in the rat, and some are even more complex than that described for the *Cyp2d* and *Cyp4abx* clusters.

The mouse versus rat *Cyp* genes is the example that has been chosen in this paper, but they are by no means the only gene set that will have this problem. In the 5th December, 2002 issue of *Nature*, in which the mouse genome is reported,[10] Table 11 (p. 542) shows the top 50 InterPro domain families in mouse compared with that in human, fish, worm and fly. Cytochrome P450 is

Q1 **Figure 1.** (See page 3) The rat and mouse *Cyp2d* locus. Expansion in the mouse leads to non-orthologous relationships between these genes. These rat genes have been named for more than 15 years; in fact, they were the first five genes in the *Cyp2d* subfamily to be identified. For more details on rat P450 nomenclature, see http://drnelson.utmem.edu/cytochromeP450.html. Abbreviation: bp, base pairs.

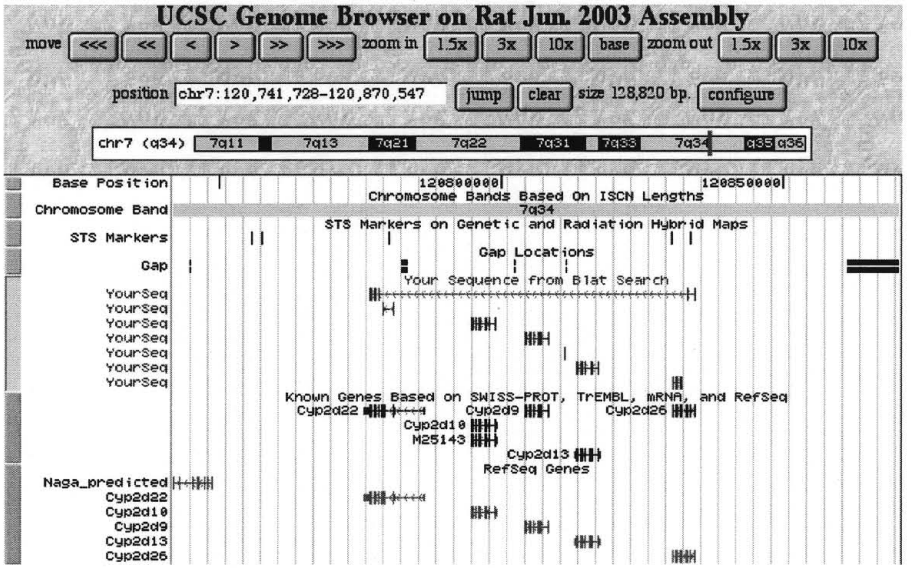


Figure 3. The rat *Cyp2d* locus, as shown at the University of California Santa Cruz Genome Browser. Note that the gene nomen clature being used follows the existing mouse gene nomenclature, and this is incorrect. Mouse *Cyp2d22* is actually the orthologue of rat *Cyp2d4*. Mouse *Cyp2d26* is actually the orthologue of rat *Cyp2d2*.

ranked 46th in the mouse and 52nd in human. The 45 other families that are more abundant than *Cyp* in the mouse will potentially have similar nomenclature issues. Fortunately, some of these groups (eg the homeobox genes) have a firmly established nomenclature, and they will not be renamed. It is not so clear what confusion will descend on the ATPases, kinases, zinc finger proteins and the many other gene families.

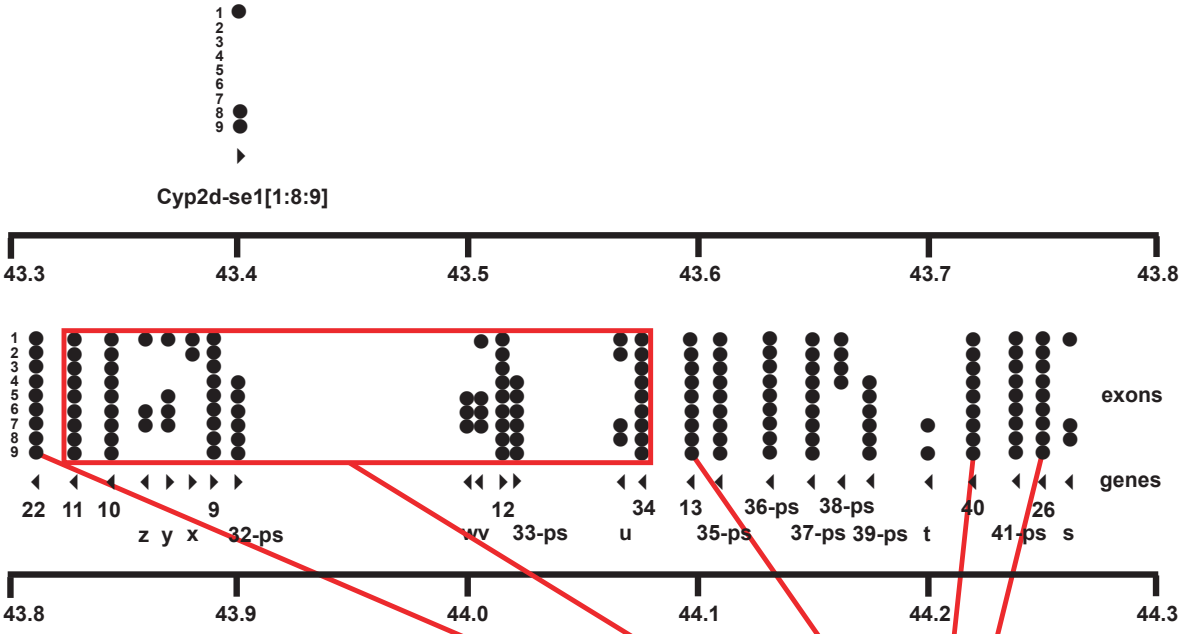
The point made here by these figures and tables is that: naming genes cannot be an automatic process, unless one wishes to create confusion. Best reciprocal BLAST hits can be used in assigning names, but they should not be used indiscriminately — if they are, the result found in Figure 3 might occur. In fact, it has occurred. Figure 3 is a screenshot of the University of California Santa Cruz (UCSC) browser showing the rat *Cyp2d* cluster, with its five genes. Note that these genes are named *Cyp2d22, Cyp2d10, Cyp2d9, Cyp2d13* and *Cyp2d26*. From Figure 2, it can be seen that these rat genes had already been named *Cyp2d4, Cyp2d5, Cyp2d1, Cyp2d3* and *Cyp2d2*. These names were assigned between 1987 and 1989 by the Committee on Standardized Cytochrome P450 Nomenclature, and they are official names used in many dozens or hundreds of publications. The two outside 'rat genes *Cyp2d22* and *Cyp2d26*' (Figure 3) are in fact orthologues of rat *Cyp2d4* and *Cyp2d2* (Table 1), but the other three rat genes in between *Cyp2d22* and *Cyp2d26* in Figure 3 — *Cyp2d10, Cyp2d9* and *Cyp2d13* — are not orthologous pairs. Thus, rat *Cyp* genes that already have official names have been renamed to match seemingly orthologous mouse *Cyp* genes. On other views in the UCSC browser, rat *Cyp* genes that already have official names have been renamed for human *CYP* genes that are not their orthologues. These names are wrong, and they appear in the Genbank database, where they will probably be used by companies making microarrays and by genome browsers like UCSC and ENSEMBL. This is a very unfortunate practice that may require considerable effort to correct.

Gene nomenclature committees have been established to impose order on gene families and in whole genomes, to prevent duplication of names and multiple uses of the same root symbol. Gene nomenclature committees have been established to provide an authority that can be trusted. Ignoring the existence of naming systems in order to assign hundreds, or thousands, of names quickly to the rat genes to match genes in other genomes, will come with a price, and the price will be in failed communication and widespread confusion. These problems are not that different from what must occur when a carefully constructed language is corrupted.

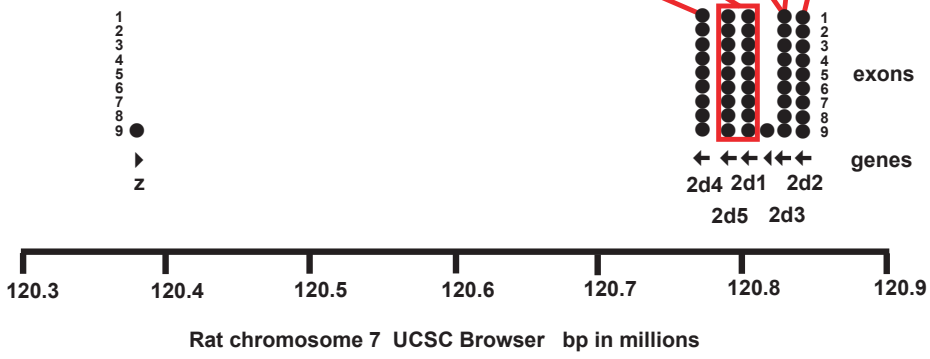
References

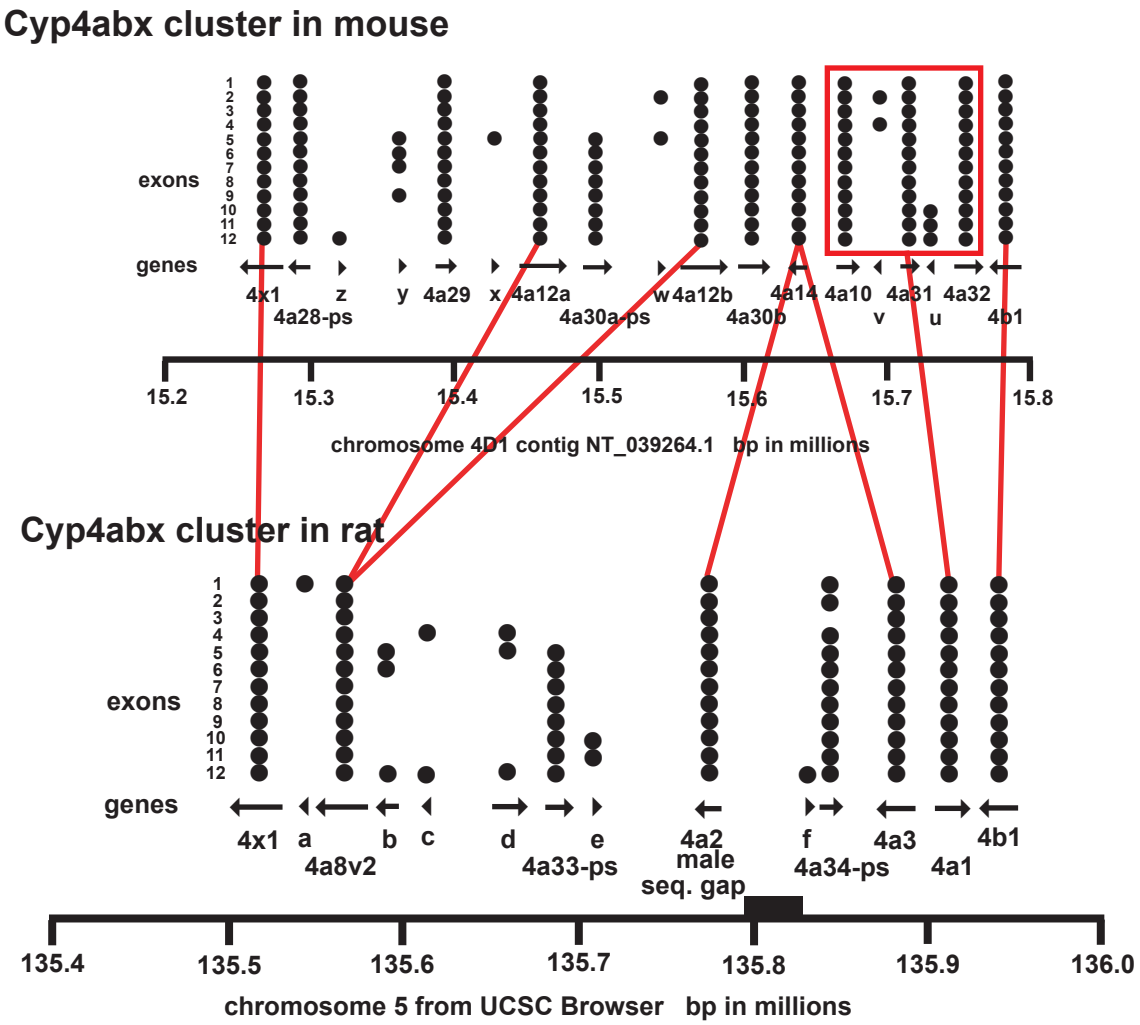
1. Rat Genome Sequencing Project Consortium. (2004), 'Genome sequence of the Brown Norway rat yields insights into mammalian evolution', *Nature* Vol. 428, pp. 493–521.
2. Fujii-Kuriyama, Y., Mizukami, Y., Kawajiri, K. *et al.* (1982), 'Primary structure of a cytochrome P-450: Coding nucleotide sequence of phenobarbital-inducible cytochrome P-450 cDNA from rat liver', *Proc. Natl. Acad. Sci. USA* Vol. 79, pp. 2793–2797.
3. Kimura, S., Gonzalez, F.J. and Nebert, D.W. (1984), 'The murine *Ah* locus Comparison of the complete cytochrome P_1-450 and P_3-450 cDNA nucleotide and amino acid sequences', *J. Biol. Chem.* Vol. 259, pp. 10705–10713.
4. Kimura, S., Gonzalez, F.J. and Nebert, D.W. (1984), 'Mouse cytochrome P3-450: Complete cDNA and amino acid sequence', *Nucleic Acids Res.* Vol. 12, pp. 2917–2928.
5. Nebert, D.W., Adesnik, M., Coon, M.J. *et al.* (1987), 'The P450 gene superfamily: Recommended nomenclature', *DNA* Vol. 6, pp. 1–11.
6. Nelson, D.R., Kamataki, T., Waxman, D.J. *et al.* (1993), 'The P450 superfamily: Update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature', *DNA Cell Biol.* Vol. 12, pp. 1–51.
7. Nelson, D.R., Koymans, L., Kamataki, T. *et al.* (1996), 'P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature', *Pharmacogenetics* Vol. 6, pp. 1–42.
8. Nelson, D.R., Zeldin, D., Hoffman, S. *et al.* (2004), 'Comparison of cytochrome P450 (*CYP*) genes from the mouse and human genomes including nomenclature recommendations for genes, pseudogenes, and alternative-splice variants', *Pharmacogenetics* Vol. 14, pp. 1–18.
9. http://drnelson.utmem.edu/cytochromeP450.html. 'The Cytochrome P450 homepage' (last accessed 30th March, 2005).
10. Mouse Genome Sequencing Consortium. (2002), 'Initial sequencing and comparative analysis of the mouse genome', *Nature* Vol. 420, pp. 520–562.

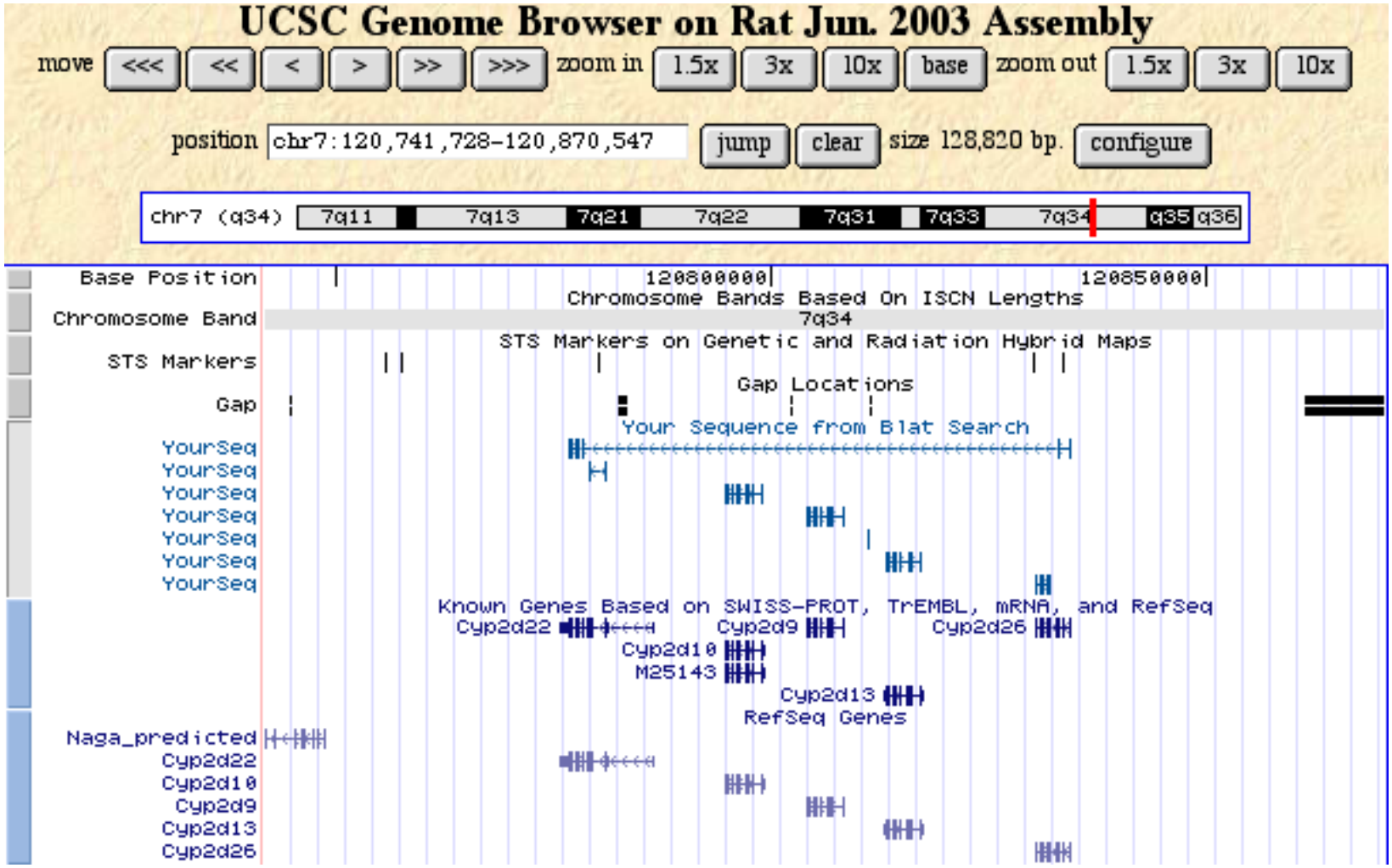
Cyp2d cluster in mouse



Cyp2d cluster in rat




Cyp4abx cluster in mouse
exons
1 2 3 4 5 6 7 8 9 10 11 12
genes
4x1
4a28-ps
z
y
4a29
x
4a12a
4a30a-ps
w
4a12b
4a30b
4a14
4a10
v
4a31
u
4a32
4b1
15.2
15.3
15.4
15.5
15.6
15.7
15.8
chromosome 4D1 contig NT_039264.1 bp in millions
Cyp4abx cluster in rat
exons
1 2 3 4 5 6 7 8 9 10 11 12
genes
4x1
a
4a8v2
b
c
d
e
4a33-ps
4a2
male
seq. gap
f
4a34-ps
4a3
4a1
4b1
135.4
135.5
135.6
135.7
135.8
135.9
136.0
chromosome 5 from UCSC Browser bp in millions


UCSC Genome Browser on Rat Jun. 2003 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position chr7:120,741,728-120,870,547 jump clear size 128,820 bp. configure
chr7 (q34) 7q11 7q13 7q21 7q22 7q31 7q33 7q34 q35 q36
Base Position
120800000
120850000
Chromosome Bands Based On ISCN Lengths
Chromosome Band
7q34
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
Gap Locations
Gap
Your Sequence from Blat Search
YourSeq
YourSeq
YourSeq
YourSeq
YourSeq
YourSeq
YourSeq
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Cyp2d22
Cyp2d9
Cyp2d26
Cyp2d10
M25143
Cyp2d13
RefSeq Genes
Naga_predicted
Cyp2d22
Cyp2d10
Cyp2d9
Cyp2d13
Cyp2d26