9

Phylogenomic Resources at the UCSC Genome Browser

Kate Rosenbloom, James Taylor, Stephen Schaeffer, Jim Kent, David Haussler, and Webb Miller

Summary

The UC Santa Cruz Genome Browser provides a number of resources that can be used for phylogenomic studies, including (1) whole-genome sequence data from a number of vertebrate species, (2) pairwise alignments of the human genome sequence to a number of other vertebrate genome, (3) a simultaneous alignment of 17 vertebrate genomes (most of them incompletely sequenced) that covers all of the human sequence, (4) several independent sets of multiple alignments covering 1% of the human genome (ENCODE regions), (5) extensive sequence annotation for interpreting those sequences and alignments, and (6) sequence, alignments, and annotations from certain other species, including an alignment of nine insect genomes. We illustrate the use of these resources in the context of assigning rare genomic changes to the branch of the phylogenetic tree where they appear to have occurred, or of looking for evidence supporting a particular possible tree topology. Sample source code for performing such studies is available.

Key Words: Evolutionary event; phylogenetic tree; interspersed repeat; chromosomal break.

1. Introduction

Rare genomic changes (RGCs), such as retroposon integrations, indels (insertions/deletions) in protein-coding regions, and inversions, provide useful alternatives to the use of nucleotide and amino-acid substitutions for determining the evolutionary relationships among living organisms ***(1)***. If one knows where in the genome to look, the UCSC Genome Browser ***(2,3)*** and the associated database ***(4)*** make it easy to see evidence that an RGC occurred at a particular point in evolutionary history.

From: *Methods in Molecular Biology: Phylogenomics*
Edited by: W. J. Murphy © Humana Press Inc., Totowa, NJ

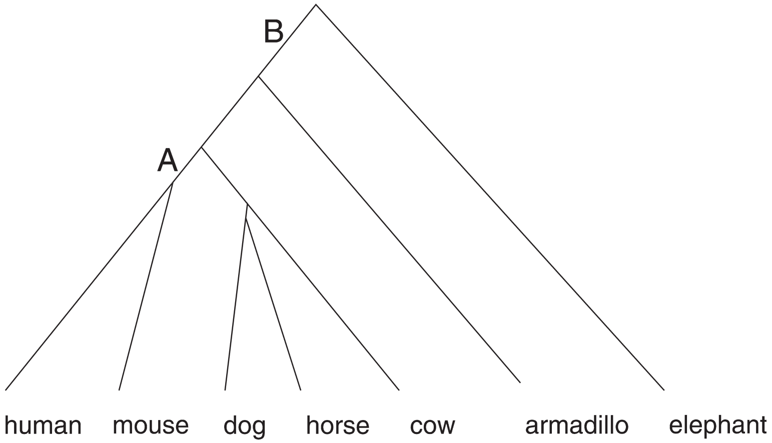


Fig. 1. A proposed evolutionary tree for several mammals, as discussed in this manuscript. Only the tree topology is sketched; branch lengths have no meaning.

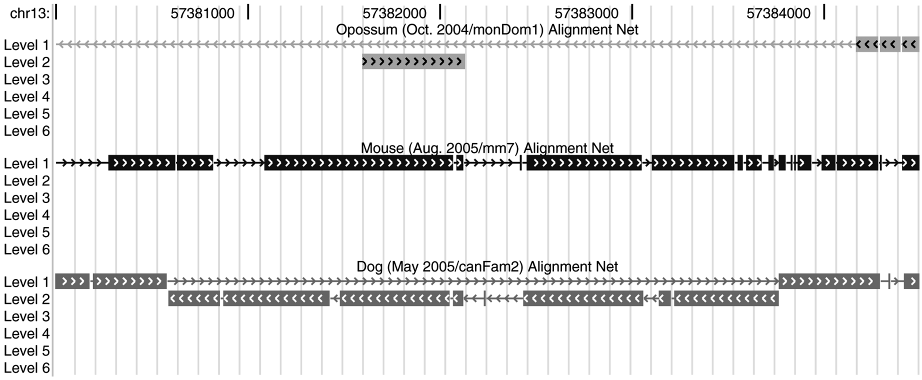


Fig. 2. View in the UCSC Human Genome Browser of alignments of opossum, mouse and dog to a region of human chromosome 13, showing a reversal that appears to have occurred along branch A in the tree of **Fig. 1**.

For instance, it is widely held ***(5)*** that the human lineage diverged from dog (and other so-called laurasiatheres) before it diverged from mouse. Thus, we can expect to observe evidence for genomic changes that occurred on the branch labeled A in **Fig. 1. Figure 2** shows the black-and-white version of a view in the Browser of a 4500-bp interval of human chromosome 13 (positions relative to the May 2004 human assembly). Dog shows a local inversion relative to human and mouse, and the outgroup opossum indicates that dog is in the ancestral orientation. The most parsimonious explanation is that a single inversion event occurred on branch A.

On the other hand, if one is not told where in the genome to look, a search for RGCs that occurred along a hypothesized short or ancient branch cannot be performed efficiently by eye; too many evolutionary events are recorded to make it feasible to start looking through them one at a time. For a systematic attack on the problem, it is often necessary to download large amounts of data and process them with custom-built software. Fortunately, a graduate student or talented undergraduate with a strong programming background can frequently write such programs, provided they get over a few initial hurdles. This chapter contains much of the necessary information.

A wealth of genomic data is available from the UCSC Browser, as well as alternative data providers, such as Ensembl ***(6)***, NCBI ***(7)***, and Galaxy ***(8)***. Sequence data, alignments, and annotations can be downloaded from the Table Browser ***(9)*** or the UCSC "downloads" page, then searched in arbitrary ways. The aim of this chapter is to illustrate how this can be done through three examples. The general problem treated by the examples is assigning RGCs to branches of the phylogenetic tree. Such an assignment is a natural part of an attempt to reconstruct evolutionary histories, as has been proposed for eutherian mammals ***(10)***. The first example focuses on a search for informative retroposon integrations in the 1% of the human genome encompassed by the 44 regions chosen for extensive study by the ENCODE project ***(11)***. For the second example, we look across the entire human genome for indels in protein-coding regions that provide evidence concerning the earliest divergences among eutherian mammals, e.g., perhaps occurring in the human lineage after divergence from elephants (afrotheres) but before divergence from armadillo (xenarthrans) (branch B in **Fig. 1**). In the third example, we look for particular kinds of chromosomal breaks in an alignment of nine insect genomes. To help the reader get started writing such programs, we make the source code used in those examples freely available at http://bio.cse.psu.edu/miller_lab/ under the title "Phylogenomic Tools." Although many biologists may hesitate to venture into such a project, we encourage them to consider hiring someone with solid programming skills to perform whole-genome searches, helped by the clues we provide here.

2. Methods

2.1. Interspersed Repeats in ENCODE Regions

As illustrated by Chapters 14 and 15 in this volume, interspersed repeat elements can sometimes be used to answer phylogenomic questions. In particular, insertion events that occurred very early in the mammalian radiation can provide evidence

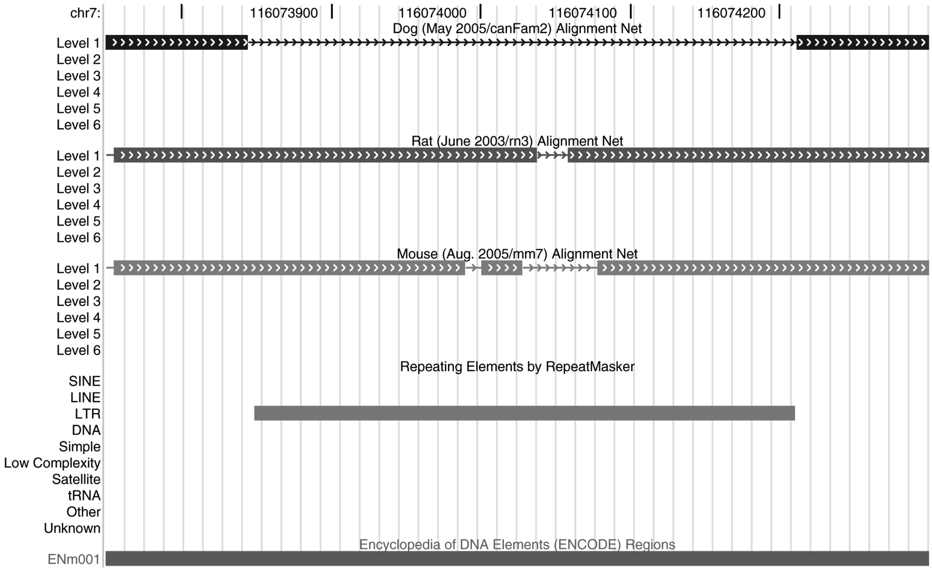


Fig. 3. Browser view of an interspersed repeat element in ENCODE region ENm001 that appears to have been inserted along branch A in **Fig. 1**.

to support certain hypotheses about phylogenetic relationships, such as that the human lineage diverged from dogs and cows before it diverged from mice ***(12)***, and that horses are more closely related to dogs than to cows ***(13)***.

A repetitive element that may have inserted on the branch labeled “A” in **Fig. 1** is shown in **Fig. 3.** Two important characteristics are that (i) a large fraction of the element aligns with mouse (for this to be possible, the element cannot be completely masked e.g., by replacing each of its nucleotides by “N”), and (ii) none of it aligns with dog. Those properties also hold of segments that were deleted in the dog lineage after it diverged from human and mouse, but here we have additional information. Namely, we recognize the region as an insertion element; for this to be a deletion in dog, the deleted interval would need to correspond to within a few nucleotides of the inserted element. (This can be seen by inspecting the alignment in detail, noting that the positions that align just to either side of the repetitive element are only a few nucleotides apart in dog.) Moreover, the element belongs to the family MLT1A0, which is known to have been transpositionally active early in the mammalian radiation. (For a list of such families, see Table 6 of the mouse genome paper in **ref. *14*.**) Finally, the percent identity with the consensus sequence for the class, 20.8%, is consistent with what one expects for an insertion that occurred early just before human–mouse divergence.

For further evidence, one can hope to identify the flanking direct repeats at the borders of the human element, and find a single copy of that sequence at the appropriate place in dog.

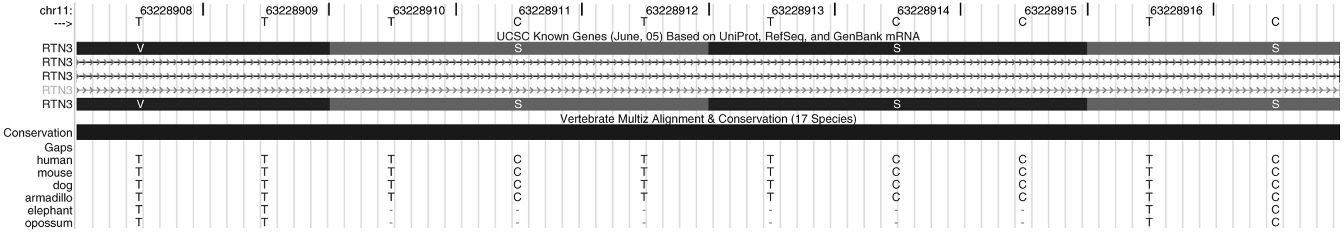
We wrote some simple computer programs to help search through large genomic regions for retroposons that are likely candidates for having been inserted at a particular branch of the phylogenetic tree. The program is essentially the same as what we had used earlier to find examples ***(12,13)***. These programs are used on "soft-masked" sequence, i.e., where repetitive DNA is given with lower-case letters ("a", "c", "g", and "t"), and the ".out" files produced by RepeatMasker (a table of information about the identified repeat elements). Both kinds of data can be downloaded from UCSC. The sequences are aligned with the blastz program ***(15)***, which can be freely obtained from http://www.bx.psu.edu/ miller_lab/

Some small programs in the "Phylogenomic Tools" package (from the same website as blastz) read the RepeatMasker ".out" files and the blastz alignments, then report repeat elements with the desired property. When we downloaded sequences and ".out" files for the ENCODE region ENm001, the tools identified the repeat element shown in **Fig. 3**. The Phylogenomic Tools package contains detailed instructions for how to proceed.

We ran our programs on all ENCODE regions, looking for repeats that inserted on branch A. We ignored Alu repeats (because they inserted after human–mouse divergence), as well as MIR and L2 repeats (because they are thought to have died out before the eutherian radiation). Four MLTA0 repeats and two L1s were identified. As a check for specificity (i.e., lack of false positives), we reversed the roles of mouse and dog; no repeats were predicted to have inserted after the human lineage diverged from mouse but before it diverged from dog.

2.2. Coding Indels

Indels (insertions/deletions) in protein-coding regions are another class of RGC that has been utilized for phylogenetic analysis ***(16),*** including deletions that apparently happened on branch A of **Fig. 1** ***(17)***. We have written several programs to automate most of the job of searching UCSC-generated alignments for informative coding indels and tested them in a search for events on branch B of **Fig. 1**. The first program reads a file of gene locations downloaded from UCSC and produces a simplified list of coding exons (including the removal of redundant copies in the case of splice isoforms). The second program reads the list of exons and three pairwise alignments of human, to, say, armadillo, elephant, and opossum, and looks for human exons that align without a gap to armadillo, but align to elephant and opossum with gaps (lengths divisible by 3) in precisely the same places. *See* **Fig. 4** for one of the exons that it found.


UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
RTN3
RTN3
RTN3
RTN3
RTN3
V
S
S
S
V
S
S
S
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
Gaps

```
chr11:     63228908  63228909  63228910  63228911  63228912  63228913  63228914  63228915  63228916
--->       T         T         T         C         T         T         C         C         T         C
human      T         T         T         C         T         T         C         C         T         C
mouse      T         T         T         C         T         T         C         C         T         C
dog        T         T         T         C         T         T         C         C         T         C
armadillo  T         T         T         C         T         T         C         C         T         C
elephant   T         T         -         -         -         -         -         -         T         C
opossum    T         T         -         -         -         -         -         -         T         C
```

Fig. 4. Browser view of an alternatively spliced exon of the RTN3 gene, showing two codons in the human, mouse, dog, and armadillo genes that are missing in elephant and opossum. This is a candidate for an insertion on branch B of **Fig. 1.**

The most parsimonious explanation is that two codons were inserted in a single event on branch B.

2.3. Chromosomal Breaks in Drosophila

One of the earliest uses of genetic data to reconstruct phylogenies was by Dobzhansky and Sturtevant when they examined polymorphisms in the polytene chromosomes isolated from *Drosophila* salivary gland cells (***18***). The polytene chromosomes have reproducible banding and puffing patterns that reflect the order of genes on chromosomes. Dobzhansky and Sturtevant showed that *Drosophila pseudoobscura* had a wealth of gene order differences in natural populations and inferred that the paracentric inversion mutations could relate the different chromosomes to each other in an unrooted network (***19***). Remarkably, all of the intermediate chromosomes in the network of gene arrangements have been collected from nature over the years with one exception, the Hypothetical chromosome (***20***). It is now well documented that chromosomal inversions have been important in the evolution of *Drosophila* because many of the species are polymorphic for gene order based on polytene chromosomal data *(**21**)*. The banding and puffing patterns of polytene chromosomes of different *Drosophila* species are quite different and prevent direct comparison of gene order. With the advent of complete genomic sequencing, we can now infer ancestral rearrangements between species by comparing inferred gene orders. Comparison of two species of *Drosophila* has shown that at least 10 inversions occur per million years (***22***) and that in some cases repetitive elements may be responsible for the process of rearrangement. These analyses also show that some regions tend to resist rearrangement because the size of conserved linkage blocks is too large given the number of inversions that have occurred over esvolutionary time, suggesting a functional conservation of gene order to maintain coordinate regulation (***23***).

Our software toolkit contains programs to read multiway alignments downloaded from UC Santa Cruz and to look for places where one group of species shows conservation of order and orientation of genomic features, but other species have a chromosomal break. *See* **Fig. 5** for an example.

2.4. Obtaining Sequences, Alignments, and Annotation from UCSC

The UCSC Genome Browser site provides two main facilities for obtaining genomic sequence, alignments, and annotations:

- a download server that supports bulk retrieval via FTP or HTTP access
- the Table Browser data retrieval tool *(8)*, useful for selective querying by region

Because this chapter is oriented toward high-throughput, whole-genome analysis, this section focuses on the download server, although we will briefly describe how to use the Table Browser to obtain alignments in selected regions.

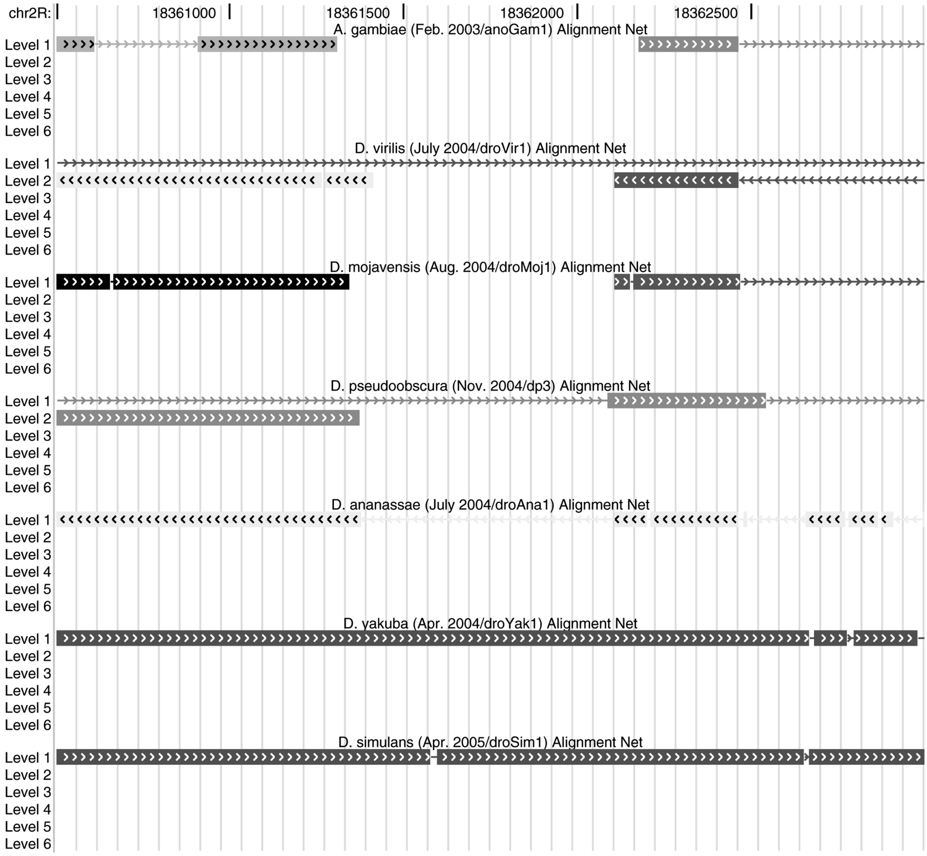

chr2R:
18361000
18361500
18362000
18362500
A. gambiae (Feb. 2003/anoGam1) Alignment Net
D. virilis (July 2004/droVir1) Alignment Net
D. mojavensis (Aug. 2004/droMoj1) Alignment Net
D. pseudoobscura (Nov. 2004/dp3) Alignment Net
D. ananassae (July 2004/droAna1) Alignment Net
D. yakuba (Apr. 2004/droYak1) Alignment Net
D. simulans (Apr. 2005/droSim1) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6

Fig. 5. Browser view of a region on *Drosophila melanogaster* chromosome 2R where *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. ananassae* show colinear alignment, but where *D. pseudoobscura*, *D. mojanavensis*, *D. virilis*, and *Anopheles gambiae* have a chromosomal break.

2.4.1. UCSC Download Server

The download server can be accessed from the “Downloads” menu link (http://hgdownload.cse.ucsc.edu) on the UCSC Browser main page or by anonymous FTP (http://hgdownload.cse.ucsc.edu/goldenPath). The Downloads web page provides a directory of links to the data files available, grouped by species and assembly.

On the download server, each genome assembly is represented by a directory labeled with the UCSC assembly release name. For example, files related to the May 2004 human genome assembly are stored in the “hg17” directory. Assembly releases and versions are described on the FAQ page (http://genome.ucsc.edu/FAQ/FAQreleases).

The exact items vary by assembly, but the following data are typically available:

Link	Directory	Description
Full data set	bigZips	Misc assembly and annotation
Data set by chromosome	chromosomes	Repeatmasked sequence
Annotation database	database	Tab-delimited nightly database dump
Multiple alignment	(eight multiz8way species)	Multiz alignments (MAF format)
Conservation scores	phastCons	Per-base scores for multiple alignment
Mouse pairwise alignments	vsMm7	Blastz/chain/net alignments (AXT format)
Dog pairwise alignments	vsCanFam2	Blastz/chain/net alignments (AXT format)
LiftOver files	liftOver	Conversion chains to other assemblies

The section below details where to download the comparative genomics resources used in this chapter. *Note:* all referenced human genome data are from the May 2004 (UCSC release hg17) human genome assembly.

1. Human genome multiple alignments
   Link: Multiple alignments of 16 vertebrate genomes to human
   Directory: hg17/multiz17way

2. Human alignments to opossum, armadillo, and elephant
   Opossum Genome: Oct. 2004 assembly (UCSC release monDom1)
   Link: Opossum/Human (hg17) alignments
   Directory: monDom1/vsHg17
   Armadillo Genome: May 2005 assembly (not released in UCSC browser)
   Directory: dasNov1/vsHg17
   Elephant Genome: May 2005 assembly (not released in UCSC browser)
   Directory: loxAfr1/vsHg17

3. Human gene locations
   Link: Annotation database
   Directory: database
   Tables: refFlat (RefSeq genes)
   knownGene (UCSC Known genes)
   encodeGencodeGeneKnown (ENCODE Gencode genes)
   Files: *.txt.gz (data), *.sql (table schema)

4. Insect genome multiple alignment

Assembly: *Drosophila melanogaster* Genome, April 2004 assembly
(UCSC release dm2)
Link: Multiple alignment of eight insects with *D. melanogaster*
Directory: dm2/multiz9way

The ENCODE project has a project-specific downloads page, accessed from the "Downloads" menu link on the UCSC ENCODE portal page (http://genome.ucsc.edu/ENCODE). The following comparative genomics resources in the ENCODE regions can be found on the May 2004 human genome assembly (currently the ENCODE Reference Assembly):

1. ENCODE region sequence and RepeatMasker output for human genome
   Link: Nucleotide sequences
   Directory: hg17/encode/regions
2. ENCODE orthologous region sequence and RepeatMasker output for other species, used to produce multiple alignments
   Link: Multiple sequence alignments
   Directory: hg17/encode/alignments/<freeze-data>/sequences
3. ENCODE region multiple alignments
   Link: Multiple sequence alignments
   Directory: hg17/encode/alignments/<freeze-date>/alignments

The alignment file formats are described in the following help files:
MAF: http://genome.ucsc.edu/goldenPath/help/maf.html
AXT: http://genome.ucsc.edu/goldenPath/help/axt.html
Chain: http://genome.ucsc.edu/goldenPath/help/chain.html

2.4.2. UCSC Table Browser

The Table Browser is accessed via a menu link on the UCSC Genome Browser home page. Pull-down menus on the Table Browser allow selection of an organism, assembly, genomic region, data type, and output format; options are provided for filtering, intersecting, and correlating tables.

To obtain multiple alignments in a genomic region via the Table Browser, select the following menu options:

Group: Comparative Genomics
Track: Conservation
Table: multiz*
Output format: MAF

For ENCODE alignments, use:

Group: ENCODE Comparative Genomics
Tracks: TBA, MAVID, or MLAGAN Alignment
Tables: encodeTbaAlign, encodeMavidAlign, encodeMlaganAlign
Output format: MAF

For compactness, select the gzip compression output option.

2.4.3. Access Guidelines

Data in the UCSC Genome database are generally freely available for public use; any limitations are described in the Conditions of Use section of the UCSC Browser home page. When downloading multiple files, we recommend using the FTP site rather than the web access.

Postscript

Since this chapter was first written, several groups have employed RGCs to predict early mammalian divergences. Kriegs et al. ***(24)*** and Nishihara et al. ***(25)*** used interspersed repeats, whereas Murphy et al. ***(26)*** employed insertions/deletions in protein-coding regions as well as interspersed repeats. The computer programs used by Murphy et al. to find informative coding indels will be made available along with the code described in this chapter.

Acknowledgments

Daryl Thomas and Elliott Margulies manage the ENCODE alignments and other resources, Brian Raney produced alignments to the low-redundancy genomes (including elephant and armadillo), and Angie Hinrichs generated the insect multiple alignments.

References

1. Rokas, A. and Holland, P. W. H. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15,** 454–459.
2. Kent, W. J., Sugnet, C. W., Furey, T. S., et al. (2002) The human genome browser at UCSC. *Genome Res.* **12,** 996–1006.
3. Hinrichs, A. S., Karolchik, D., Baertsch, R., et al. (2006) The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34** (Database issue)**,** D590–D598.
4. Karolchik, D., Baertsch, R., Diekhans, M., et al. (2003) The UCSC genome browser database. *Nucleic Acids Res.* **31,** 51–54.
5. Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature* **409,** 614–618.
6. Birney, E., Andrews, D., Caccamo, M., et al. (2006) Ensembl 2006. *Nucleic Acids Res.* **34** (Database issue)**,** D556–D561.
7. Wheeler, D. L., Church, D. M., Edgar, R., et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32,** D35–D40.
8. Giardine, B., Riemer, C., Hardison, R. C., et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15,** 1451–1455.
9. Karolchik, D., Hinrichs, A. S., Furey, T. S., et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**(Suppl 1)**,** D493–D496.

10. Blanchette, M., Green, E., Miller, W., and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome *in silico*. *Genome Res.* **14,** 2412–2423.
11. The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia of DNA Elements) project. *Science* **306,** 636–640.
12. Thomas, J. W., Touchman, J. W., Blakesley, R. W., et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793.
13. Schwartz, S., Elnitski, E., Li, M., et al. (2003) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31,** 3518–3524.
14. Waterston, R. H., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562.
15. Schwartz, S., Kent, W. J., Smit, A., et al. (2003) Human–mouse alignments with blastz. *Genome Res.* **13,** 103–107.
16. De Jong, W. W., van Dijk, M. A. M., Poux, C., Kappe, G., van Rheede, T., and Madsen, O. (2003) Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. *Mol. Phylogenet. Evol.* **28,** 328–340.
17. Poux, C., van Rheede, T., Madsen, O., and de Jong, W. W. (2002) Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19,** 2035–2037.
18. Dobzhansky, T. and Sturtevant, A. H. (1938) Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23,** 28–64.
19. Dobzhansky, T. (1944) Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis*. *Carnegie Inst. Washington Publ.* **554,** 47–144.
20. Anderson, W. W., Arnold, J., Baldwin, D. G., et al. (1991) Four decades of inversion polymorphism in *Drosophila pseudoobscura*. *Proc. Natl Acad. Sci. USA* **88,** 10,367–10,371.
21. Sperlich, D. and Pfriem, P. (1986) Chromosomal polymorphism in natural and experimental populations. In: *The Genetics and Biology of* Drosophila*, 3rd edition* (Ashburner, M., Carson, H. L., and Thomson, J. N., eds), pp. 257–309, Academic, New York.
22. Richards, S., Liu, Y., Bettencourt, B. R., et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene and *cis*-element evolution. *Genome Res.* **15,** 1–18.
23. Stolc, V., Gauhar, Z., Mason, C., et al. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306,** 655–660.
24. Kriegs, J. O., Churakow, G., Kiefmann, M., Jordan, U., Brosius, J., and Schmitz, J. (2006) Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol.* **4,** e91.
25. Nishihara, H., Hasegawa, M., and Okada, N. (2006) Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *Proc. Natl Acad. Sci. U S A* **103,** 9929–9934.
26. Murphy, W. J., Pringle, T. H., Crider, T., Springer, M. S., and Miller, W. (2006) Using genomic data to unravel the root of the placental mammal phylogeny. *Submitted.*