

99. Comparative sequencing of vertebrate genomes

Matthew E. Portnoy and Eric D. Green

National Human Genome Research Institute, Bethesda, MD, US

1. Introduction

g204314 (P0001)

The past decade has brought astonishing growth in the generation of sequence data from eukaryotic genomes. This has largely been catalyzed by major technological and strategic advances in large-scale DNA sequencing (Green, 2001a; *see* article 99, **Shotgun sequencing strategies**, Volume 0), coupled with the intense effort to complete the Human Genome Project (*see* article 99, **The human genome project**, Volume 0) and, in particular, to finish the sequence of the first vertebrate genome – that of *Homo sapiens* (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001).

g204314 (P0002)

With a complete human genome sequence now available, attention has rapidly turned to understanding the functional information it encodes. Significant advances have been made in identifying the protein-coding portion of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001; *see* article 99, **Gene finding pipelines**, Volume 0); however, this portion only reflects an estimated 1–2% of the ~2.9-Gb human genome sequence. Importantly, an additional 3–4% of the human genome appears to be functional, but does not code for protein (Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004); these sequences include elements that provide temporal and spatial control of gene expression (Wasserman and Sandelin, 2004) as well as those involved in chromosome dynamics. It is now apparent that a comprehensive cataloging of all functional elements in the human genome, especially those that do not directly code for protein, will require a multifaceted approach, involving the generation of additional laboratory- and computational-based data and the development of new paradigms for assimilating and analyzing the resulting complex data sets.

g204314 (P0003)

One of the most powerful approaches for identifying functional genomic elements involves the comparison of genome sequences from species at distinct evolutionary positions (Miller *et al.*, 2004; Nobrega and Pennacchio, 2004; Boffelli

2 Model Organisms: Functional and Comparative Genomics

et al., 2004; Pennacchio, 2003; Hardison, 2003). The resulting information provides a working knowledge of the precise sequence-level similarities and differences among genomes, which in turn can be used to gain insight about genome function. For example, sequences found to be common (or conserved) among species separated by large evolutionary distances (e.g., >50–100 million years) can be considered candidates for serving a functional role; the process of identifying such conserved sequences has been termed *phylogenetic footprinting* (Duret and Bucher, 1997; Weitzman, 2003). In contrast, sequences found to be different among closely related species (e.g., primates) can be considered less likely to be functional; the process of “eliminating” such sequences from consideration (thereby leaving the remaining sequences as candidates for serving a functional role) has been termed *phylogenetic shadowing* (Boffelli *et al.*, 2003; Boffelli *et al.*, 2004). In short, strategies have emerged that involve the use of genome sequences from both closely and distantly related species to extract functional information by comparative sequence analysis.

g204314(P0004)

Two complementary approaches have been used in recent years to generate vertebrate genome sequences (Green, 2001a; *see* article 99, **Shotgun sequencing strategies**, Volume 0) en route to comparative analyses. In whole-genome sequencing projects, data are generated across an entire species’ genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001; Mouse Genome Sequencing Consortium, 2002; Aparicio *et al.*, 2002; Rat Genome Sequencing Project Consortium, 2004); while such efforts are comprehensive with respect to the individual genome, they are limited in terms of the total number of different genomes that can be compared (because of the costs associated with sequencing an entire vertebrate genome). A subset of the whole-genome sequencing projects performed to date has involved species that are used as experimental models (the so-called reference genomes, such as mouse, rat, and zebrafish; *see* article 99, **The mouse genome sequence**, Volume 0, article 99, **The rat as a model physiological system**, Volume 0, and article 99, **The Fugu and zebrafish genomes**, Volume 0, respectively). In targeted sequencing projects, data are generated for discrete genomic regions, typically from multiple species (Thomas *et al.*, 2003); while such efforts examine only a limited portion of the genome, they result in sequence comparisons that involve large collections of evolutionarily diverse species. The major vertebrate whole-genome and targeted sequencing efforts are listed at www.intlgenome.org.

2. Vertebrate whole-genome sequences

g204314(P0005)

The central goals of the Human Genome Project included the generation of foundational information about the human genome and that of a handful of carefully selected other species, in particular, commonly used experimental models (Green, 2001b) (*see* article 99, **The human genome project**, Volume 0). Initially, only human and mouse were included among the vertebrates, but more recently that list has grown substantially (Table 1).

g204314(P0006)

A finished human genome sequence was generated using a clone-based shotgun-sequencing strategy (International Human Genome Sequencing Consortium, 2001;

g204314 (T01)

Table 1 Vertebrate whole-genome sequences

Common name	Species	Sequencing strategy ^a	Approximate coverage ^b	URL
Human	<i>Homo sapiens</i>	HSS	99%	www.ncbi.nlm.gov/genome/guide/human genome.ucsc.edu www.ensembl.org/Homo_sapiens genome.wustl.edu www.broad.mit.edu www.hgsc.bcm.tmc.edu
Chimpanzee	<i>Pan troglodytes</i>	WGS	4X	www.ensembl.org/Homo_sapiens genome.wustl.edu www.broad.mit.edu www.hgsc.bcm.tmc.edu
Rhesus macaque	<i>Macaca mulatta</i>	WGS/HSS	In progress	www.hgsc.bcm.tmc.edu www.broad.mit.edu www.hgsc.bcm.tmc.edu www.ncbi.nlm.gov/genome/guide/mouse genome.ucsc.edu www.ensembl.org/Mus_musculus www.broad.mit.edu
Cow	<i>Bos taurus</i>	WGS/HSS	In progress	www.hgsc.bcm.tmc.edu www.broad.mit.edu
Dog	<i>Canis familiaris</i>	WGS	In progress	www.broad.mit.edu www.hgsc.bcm.tmc.edu
Rat	<i>Rattus norvegicus</i>	WGS/HSS	>90%	www.hgsc.bcm.tmc.edu www.ncbi.nlm.gov/genome/guide/mouse genome.ucsc.edu www.ensembl.org/Mus_musculus www.broad.mit.edu
Mouse	<i>Mus musculus</i>	WGS/HSS	90–96%	www.ncbi.nlm.gov/genome/guide/mouse genome.ucsc.edu www.ensembl.org/Mus_musculus www.broad.mit.edu
Laboratory opossum	<i>Monodelphis domestica</i>	WGS	In progress	www.broad.mit.edu genome.wustl.edu www.jgi.doe.gov www.sanger.ac.uk www.jgi.doe.gov www.genoscope.cns.fr
Chicken	<i>Gallus gallus</i>	WGS	6.6X	genome.wustl.edu www.jgi.doe.gov
Frog	<i>Xenopus tropicalis</i>	WGS	In progress	www.jgi.doe.gov www.sanger.ac.uk www.jgi.doe.gov www.genoscope.cns.fr
Zebrafish	<i>Danio rerio</i>	WGS/HSS	5.7X	www.sanger.ac.uk www.jgi.doe.gov www.genoscope.cns.fr
Fugu	<i>Takifugu rubripes</i>	WGS	5.7X	www.jgi.doe.gov www.genoscope.cns.fr
Pufferfish	<i>Tetraodon nigroviridis</i>	WGS	6X	www.genoscope.cns.fr

^aHSS = clone-based hierarchical shotgun sequencing (*see* article 99, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 0); WGS = whole-genome shotgun sequencing.

^bSequence coverage indicated as redundant coverage (e.g., 4X) or percent of total covered by assembled sequence (e.g., 99%).

see article 99, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 0 and article 99, **Shotgun sequencing strategies**, Volume 0), whereby minimal overlapping sets of large-insert clones (mostly bacterial-artificial chromosome (BAC) clones (*see* article 99, **BAC and Fosmid libraries**, Volume 0)) were individually subjected to random shotgun sequencing, followed by directed finishing (Green, 2001a). The prior construction of a BAC-based physical map of the human genome (McPherson *et al.*, 2001) provided a key organizing framework for the long-range assembly of the whole-genome sequence.

g204314 (P0007)

Sequencing the genome of the most widely used experimental mammal – the laboratory mouse – began prior to completion of the Human Genome Project (Table 1; *see* article 99, **The mouse genome sequence**, Volume 0). The initial phase of this effort involved whole-genome shotgun sequencing and initial assembly (Mouse Genome Sequencing Consortium, 2002), with BAC-based finishing now ongoing. The mouse's critical role in biomedical research has prompted the finishing of its genome sequence to roughly the same quality as the completed human genome sequence. Interestingly, initial comparisons reveal that roughly 40% of the mouse genome sequence aligns with the corresponding (or orthologous)

4 Model Organisms: Functional and Comparative Genomics

regions of the human genome sequence (Mouse Genome Sequencing Consortium, 2002), but only a small minority of that aligning sequence (totaling roughly 5% of the mouse or human genome) is actively conserved and presumed to be functionally important.

g204314 (P0008)

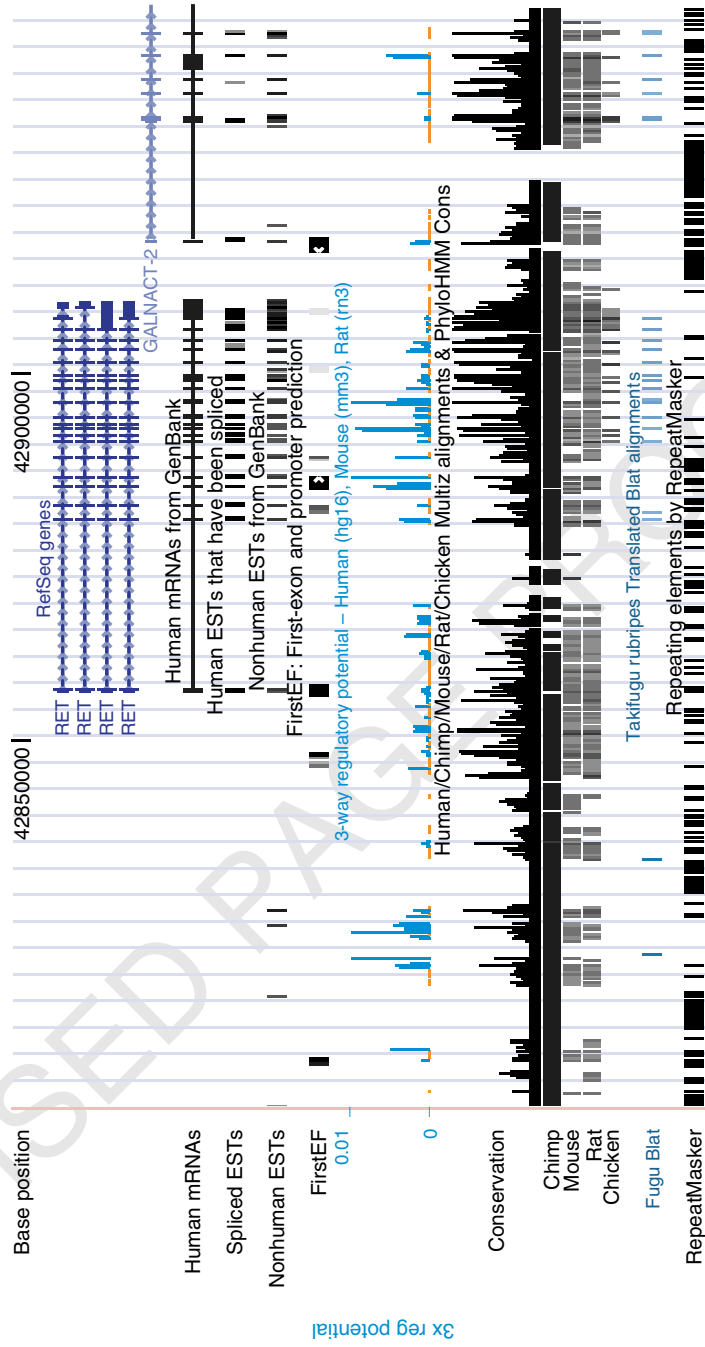
The third mammalian whole-genome sequence to be generated was that of the rat (Rat Genome Sequencing Project Consortium, 2004) (Table 1). This project utilized an integrated whole-genome and BAC-based shotgun-sequencing strategy, which yielded a very high-quality draft sequence. However, at present, there are no plans to finish the rat genome sequence to the same quality standards as were used for the human or mouse genome sequences. Indeed, until the costs of sequence finishing (*see* article 99, **Finishing strategies, alternative technologies, problems, unsequenceable regions, PCR versus primer walking off shotgun or large insert clones**, Volume 0) decrease substantially, it is unclear which other vertebrate genomes (if any) will be sequenced as accurately and completely as the human genome.

g204314 (P0009)

Whole-genome draft sequences have also been generated for a trio of fish species: zebrafish (*Danio rerio*) and two types of pufferfish, *Fugu rubripes* (Aparicio *et al.*, 2002) and *Tetraodon nigroviridis* (Table 1) (*see* article 99, **The Fugu and zebrafish genomes**, Volume 0). The zebrafish genome sequence will be critical for the rapidly growing research community using these fish as experimental models, whereas the two pufferfish species were selected for their notably compact genomes (among the smallest of all vertebrates). Among the primates, a whole-genome draft sequence has been generated for the chimpanzee, our closest relative in evolution, while that of the rhesus macaque is actively being produced. The current list of available or planned vertebrate whole-genome sequences includes that of two additional nonprimate eutherian mammals (dog and cow), a marsupial (the laboratory opossum *Monodelphis domestica*), a bird (the chicken), and an amphibian (the frog *Xenopus tropicalis*) (Table 1). As the costs of large-scale DNA sequencing decline, the list of generated vertebrate whole-genome sequences (such as in Table 1) will inevitably grow, with the strategies for generating and analyzing that sequence likely changing as well (*see* below).

g204314 (P0010)

The generation of an ever-growing set of vertebrate genome sequences, coupled with initial efforts to compare them in a rigorous fashion, has yielded spectacular amounts of data. To ensure that this information is readily accessible and comprehensible, especially to the general biomedical research community, several groups have developed “genome browser” systems. These consist of convenient navigational tools for accessing and utilizing the genomic sequence of different vertebrates as well as organized frameworks for assimilating relevant annotations, including those emanating from comparative analyses. The three most widely used genome browsers are the UCSC Genome Browser (genome.ucsc.edu), Ensembl (www.ensembl.org), and NCBI Map Viewer (www.ncbi.nlm.nih.gov/mapview). As a representative example, Figure 1 depicts a ~150-kb segment of the human genome, as displayed by the UCSC Genome Browser. Note the ability to observe simultaneously the structure of known genes in the region (*RET* and *GALNACT-2*) as well as various other types of information (e.g., promoters, repetitive elements, and the results of comparative analyses with a handful of other species). Indeed, this figure illustrates how genomes will likely be viewed in the coming years, as



g204314 (tbl) **Figure 1** View of a roughly 150-kb segment of the human genome, as displayed on the UCSC Genome Browser. A segment of human chromosome 10 encompassing the *RET* gene is shown (July 2003 build of the human genome sequence, UCSC version hg16/NCBI build 34; coordinates chr10:42,800,001-42,950,000). Various annotations are displayed; in particular, note the tracks depicting the results of comparative sequence analyses (see genome.ucsc.edu for details)

6 Model Organisms: Functional and Comparative Genomics

increasingly detailed information about genes, other functional elements, regions of sequence conserved among various species, and other genome features of interest are assimilated layer by layer.

3. Multivertebrate sequences of targeted genomic regions

g204314 (P0011)

While the above whole-genome sequencing efforts are providing valuable data for comparative analyses, they are limited in the total number of vertebrates being studied (see Table 1). To complement these projects, the sequence of smaller, targeted genomic regions can be generated from a greater number of species, resulting in comparative sequence analyses with larger, more evolutionarily diverse collections of vertebrates (Thomas and Touchman, 2002). For example, the NISC Comparative Sequencing Program (see www.nisc.nih.gov) is currently sequencing more than 150 targeted regions of the human genome in multiple vertebrates, in some cases generating orthologous sequence data from over 30 species (Thomas *et al.*, 2003).

g204314 (P0012)

These studies have already yielded some interesting findings. First, the resulting data have reflected the first-available genomic sequence for a number of vertebrates, providing new insights about the genetic blueprints of these species. These have included information about gene density, the relative degree of genome compression/expansion, the amounts and types of repetitive sequences, the extent and types of mutational events that have uniquely sculpted each genome, and the general patterns of conservation seen upon comparison with other species' sequences (Thomas *et al.*, 2003). Second, the generation of orthologous sequences from large sets of vertebrates has catalyzed the development of computational methods for multispecies comparative sequence analyses. For example, new approaches have been developed for identifying sequences that are highly conserved across multiple species (called *Multispecies Conserved Sequences* or *MCSs*) (Margulies *et al.*, 2003; Margulies *et al.*, 2004). Interestingly, vertebrates differ with respect to the effectiveness of their sequence for detecting MCSs in the human genome (Margulies *et al.*, 2003; Margulies *et al.*, 2004). Finally, targeted sequencing projects are particularly well suited for genome-evolution studies because they can readily yield sequence data from carefully selected species of interest (Thomas *et al.*, 2003). Such studies can include in-depth surveying of multiple species at a particular phylogenetic node (e.g., primates), which, at least for the foreseeable future, is not possible with whole-genome sequence data sets.

g204314 (P0013)

Multivertebrate sequencing of targeted genomic regions is playing an important role in the recently launched ENCODE (Encyclopedia of DNA Elements) project, which aims to identify all of the functional elements in the human genome (genome.gov/ENCODE). ENCODE's initial goal is to catalog comprehensively the functional elements in a selected 1% (~30 Mb) of the human genome, using a diverse set of experimental and computational approaches. These targeted ~30 Mb, which are distributed across 44 different genomic regions, are being sequenced in multiple vertebrates. The resulting sequences will be subjected to myriad comparative analyses, with the results in turn compared to various other types of data (computational and experimental) generated for the same ~30 Mb. Eventually,

this process should provide important insights into the utility of multispecies sequence comparisons for unraveling the complexities of genome function.

4. Deducing genome function through sequence comparisons

g204314 (P0014) Comparative analyses of whole-genome and targeted-genome sequences from various vertebrates have been shown to be valuable for the study of genome function. Simple alignments of orthologous sequences from two or more genomes can be used to identify the presence and structure of genes (Batzoglou *et al.*, 2000; Miller *et al.*, 2004). More refined approaches have been developed for gene prediction that involve sequence comparisons; for example, using the human and mouse genome sequences, TWINSKAN (Korf *et al.*, 2001) and SLAM (Alexandersson *et al.*, 2003) have been used to produce a conservative estimate of 25,622 genes in the human genome (Flicek *et al.*, 2003) and to detect roughly 80% of the predicted human exons in the NCBI RefSeq gene collection (see www.ncbi.nlm.nih.gov/RefSeq) (Alexandersson *et al.*, 2003). In a more targeted fashion, human–mouse sequence comparisons directly led to the discovery of the apolipoprotein A5 gene (*APOA5*) (Pennacchio *et al.*, 2001); subsequent functional studies showed the importance of *APOA5* in regulating triglyceride levels (Pennacchio, 2003).

g204314 (P0015) Comparative sequence analyses are also proving critical for detecting conserved sequences outside of coding regions, which are candidates for functional noncoding elements (e.g., such as those regulating gene expression). For example, phylogenetic footprinting of sequences from a set of diverse mammals identified several regulatory elements upstream of the ϵ -globin gene (Gumucio *et al.*, 1993). Similarly, human–mouse sequence comparisons of the interleukin gene cluster identified several conserved noncoding sequences, the longest of which was shown to be a cis coactivator of several nearby interleukin genes (Hardison, 2000; Loots *et al.*, 2000). More global methods have now been developed for identifying sequences that are most highly conserved across multiple species (Margulies *et al.*, 2003; Dermitzakis *et al.*, 2003), many of which are likely to be functionally important. Interestingly, in the human genome, there are more such highly conserved sequences within noncoding regions compared to coding regions (Margulies *et al.*, 2003; Margulies *et al.*, 2004). By a different strategy (specifically, phylogenetic-shadowing methods), sequence comparisons involving closely related species have been used to identify potential regulatory elements (Boffelli *et al.*, 2003; Boffelli *et al.*, 2004).

5. Future prospects

g204314 (P0016) The landscape of comparative vertebrate sequencing is changing rapidly. The major commitments to date for whole-genome sequencing mostly involve vertebrates associated with large research communities that will directly exploit the resulting sequence data. As such, the genome sequences of species such as human (International Human Genome Sequencing Consortium, 2001; Venter *et al.*,

8 Model Organisms: Functional and Comparative Genomics

2001), mouse (Mouse Genome Sequencing Consortium, 2002), rat (Rat Genome Sequencing Project Consortium, 2004), chicken, zebrafish, *Xenopus*, dog, and cow provide reference information of great value. In addition, each whole-genome sequence provides secondary value by contributing to an ever-expanding repertoire of comparative sequence analyses, which more broadly advance our knowledge of complex genomes. However, only a few remaining vertebrates can be regarded as true reference species, and thus a primary rationale for most future genome-sequencing projects will be the acquisition of data for comparative studies.

g204314 (P0017)

The current plans for vertebrate genome sequencing largely reflect these changing priorities. Ongoing or soon-to-be-initiated sequencing projects include a wider sampling of vertebrates across the phylogenetic tree and exploration of a larger set of primates. Significantly, recent findings indicate that the identification of highly conserved sequences from evolutionarily diverse species can be accomplished with lower-quality draft sequence. Specifically, comparative analyses using low-redundancy genomic sequences (e.g., providing one- to twofold coverage) from a larger number of species appear to be more effective in identifying the most highly conserved genomic elements than those using high-redundancy sequences (e.g., providing two- to fourfold coverage) from a smaller number of species (unpublished data). These findings are prompting efforts to acquire low-redundancy genomic sequence from a large, diverse group of vertebrates. Such an endeavor would not be for the purpose of generating an assembled sequence of each genome but rather to amass a large data set that can be used predominantly for comparative analyses. This approach reflects a strategic shift from the sequencing projects performed under the auspices of the Human Genome Project, but one that resonates with the high-priority efforts to interpret the human genome sequence in a comprehensive fashion.

g204314 (P0018)

More futuristic views of comparative vertebrate sequencing largely depend on the knowledge of the relative costs of large-scale DNA sequencing. Should those costs continue to decline substantially, then the genomes of much larger collections of vertebrates (and indeed invertebrates as well) would inevitably be sequenced, with the resulting massive data sets greatly empowering comparative studies. Regardless, the lessons learned to date clearly indicate that the sequence of each species' genome contains a treasure trove of information about evolutionary history and that comparisons of those histories are critical for understanding the complexities of genome structure and function.

Acknowledgments

g204314 (P0019)

We thank Elliott Margulies, Bob Blakesley, Nancy Hansen, and Monica Janossy for the critical reading of this chapter.

References

g204314 (P0020)

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**, 1301–1310.

- Alexandersson M, Cawley S and Pachter L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, **13**, 496–502.
- Batzoglou S, Pachter L, Mesirov JP, Berger B and Lander ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, **10**, 950–958.
- Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L and Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
- Boffelli D, Nobrega MA and Rubin EM (2004) Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics*, **5**, 456–465.
- Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C and Antonarakis SE (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, **302**, 1033–1035.
- Duret L and Bucher P (1997) Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology*, **7**, 399–406.
- Flicek P, Keibler E, Hu P, Korf I and Brent MR (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Research*, **13**, 46–54.
- Green ED (2001a) Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics*, **2**, 573–583.
- Green ED (2001b) The human genome project and its impact on the study of human disease. In *The Metabolic and Molecular Bases of Inherited Disease*, Eighth Edition, Scriver CR, Beaudet AL, Sly WS, Valle D, Childs B, Kinzler KW and Vogelstein B (Eds.), McGraw-Hill: New York, NY, pp. 259–298.
- Gumucio DL, Shelton DA, Bailey WJ, Slightom JL and Goodman M (1993) Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 6018–6022.
- Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics*, **16**, 369–372.
- Hardison RC (2003) Comparative Genomics. *Public Library of Science Biology*, **1**, 156–160.
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Korf I, Flicek P, Duan D and Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.
- Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM and Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
- McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al.* (2001) A physical map of the human genome. *Nature*, **409**, 934–941.
- Margulies EH, Blanchette M, Haussler D and Green ED (2003) Identification and characterization of multi-species conserved sequences. *Genome Research*, **13**, 2507–2518.
- Margulies EH, NISC Comparative Sequencing Program and Green ED (2004) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harbor Symposia on Quantitative Biology*, Vol 68: *The Genome of Homo Sapiens*, CSHL Press: Woodbury, NY, pp. 255–263.
- Miller W, Makova KD, Nekrutenko A and Hardison RC (2004) Comparative genomics. *Annual Review of Genomics*, **5**, 15–56.
- Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
- Nobrega MA and Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. *Journal of Physiology*, **554**, 31–39.
- Pennacchio LA (2003) Insights from human/mouse genome comparisons. *Mammalian Genome*, **14**, 429–436.

10 Model Organisms: Functional and Comparative Genomics

- Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM and Rubin EM (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*, **294**, 169–173.
- Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
- Thomas JW and Touchman JW (2002) Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends in Genetics*, **18**, 104–108.
- Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
- Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, **5**, 276–287.
- Weitzman JB (2003) Tracking evolution's footprints in the genome. *Journal of Biology*, **2**, 9.

REVISED PAGE PROOFS