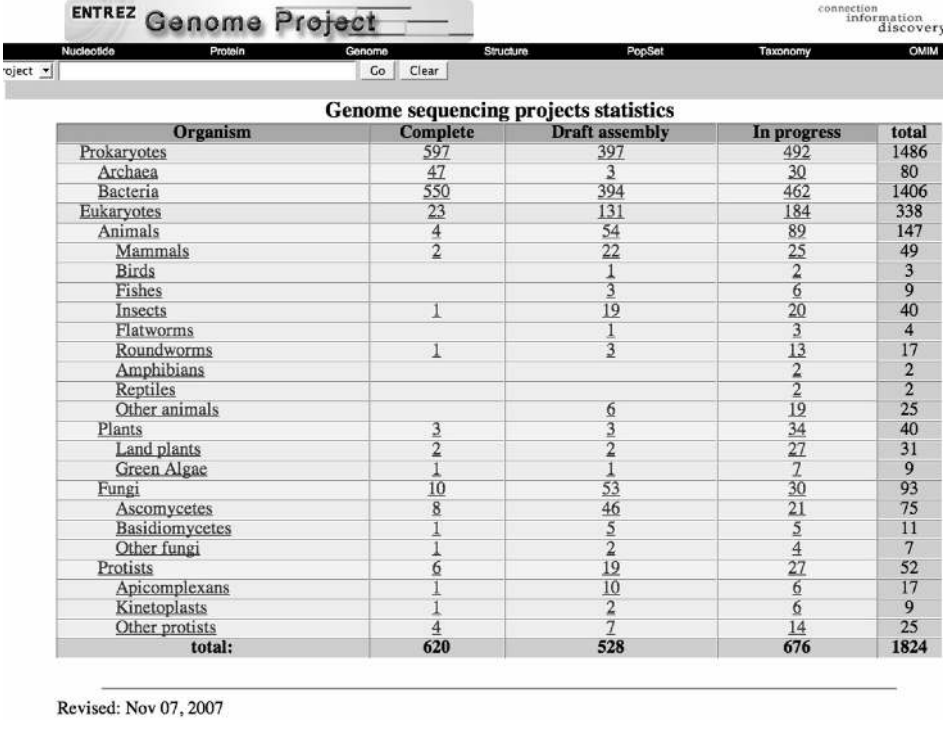
1

The Molecular Biology Data Explosion

The explosion of genome sequence data in the last decade has been so widely noted as to have almost become a cliché. The first microbial genome was only sequenced in 1995. However, by late 2007, web sites that track genome sequencing projects, such as NCBI's Entrez Genome Project site (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html) and the Genomes OnLine Database (GOLD) project (http://www.genomesonline.org) had cataloged approximately 1,000 complete microbial genome sequences. Similarly, the first complete genome of a multicellular organism (*C. elegans*) became available in 1998. Nine years later, there are complete or draft genome sequences available for more than 60 multicellular species, with low-coverage data or sequencing projects in progress for dozens of others. Figure 1.1 shows summary statistics for genomes that have been sequenced as of November 2007. Moreover, the rate at which genomes for new species and species variants are being sequenced continues to accelerate as novel sequencing technologies lower the cost of obtaining sequence data. For example, Figure 1.2 shows some of the 25 mammalian species whose genome sequencing is currently in progress. Meanwhile, along with increasing amounts of DNA sequence data, there has been a remarkable increase in the quantity of data describing how the information in the genome sequence is used to implement the functions of the organism.

With this explosion of data has come the opportunity – for those with the skill and ability to identify patterns and correlations among the data – to develop an ever more profound understanding of the way organisms function at their most fundamental levels. Genes are being identified that are involved in such basic human experiences as thinking, feeling, and communicating with spoken language. Genomes of hundreds of new microbial species are being sequenced, some of which may hold keys to humanity's most vexing problems in energy generation and environmental preservation. And genes and gene-regulatory mechanisms are rapidly being identified that underlie many of the most dread diseases, including cancer, heart disease, and the degenerative neurological diseases of aging.

Indeed, these are heady times in molecular biology. However, the same data explosion that is enabling all these advances is threatening to drown us in its very enormity.



Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	597	397	492	1486
Archaea	47	3	30	80
Bacteria	550	394	462	1406
Eukaryotes	23	131	184	338
Animals	4	54	89	147
Mammals	2	22	25	49
Birds		1	2	3
Fishes		3	6	9
Insects	1	19	20	40
Flatworms		1	3	4
Roundworms	1	3	13	17
Amphibians			2	2
Reptiles			2	2
Other animals		6	19	25
Plants	3	3	34	40
Land plants	2	2	27	31
Green Algae	1	1	7	9
Fungi	10	53	30	93
Ascomycetes	8	46	21	75
Basidiomycetes	1	5	5	11
Other fungi	1	2	4	7
Protists	6	19	27	52
Apicomplexans	1	10	6	17
Kinetoplasts	1	2	6	9
Other protists	4	7	14	25
total:	**620**	**528**	**676**	**1824**

Revised: Nov 07, 2007

Figure 1.1 Screenshot from NCBI's Entrez Genome Project (at http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html) showing the number of species whose genomes have been sequenced or whose genomes are in the process of being sequenced as of November 2007.

As the quantity of data increases, the task of discerning the critical interrelationships among this data has become increasingly difficult. Organizing biological information into dedicated databases of related data has been helpful. However, as the number of biological databases reaches into the thousands (the annual database review of the journal *Nucleic Acids Research* now regularly includes almost 1,000 new or significantly enhanced molecular biology databases each year), intelligently "mining" these data sources becomes ever more challenging.

To a large extent, the current difficulties of analyzing molecular biological data arise simply from the need to characterize such a large quantity of highly interrelated information. However, the biological research community has also brought some challenges of biological data integration and analysis onto itself by the way such data have historically been stored, transferred, and manipulated. Biology databases are located in many different locations. Many of these databases are only downloadable as flat files, as a result of which database searching may be awkward and slow, or else local relational databases may need to be set up. Varying data formats are used requiring the use of multiple data parsers for automated data analyses. As a result, integrating and comparing data from multiple biological databases is difficult and tedious.

Organism	Taxon ID	Genome Size (MB)	Number of Chromosomes	Method	Sequencing Center/Consortium
Lemur catta	9447			WGS	Broad Institute
Loxodonta africana	9785	3000	28	WGS	Broad Institute
Loxodonta africana	9785	3000	28	Clone-based	NISC - NIH Intramural Sequencing Center
Macaca fascicularis	9541				Washington University (Wash U)
Macropus eugenii	9315	3800	8	WGS & Clone-based	Baylor College of Medicine (more)
Manis pentadactyla	143292		20	WGS	Washington University (Wash U)
Nomascus leucogenys	61853				Washington University (Wash U)
Nomascus leucogenys	61853				Baylor College of Medicine
Ornithorhynchus anatinus	9258			WGS	Washington University (Wash U)
Ornithorhynchus anatinus	9258			Clone-based	NISC - NIH Intramural Sequencing Center
Oryctolagus cuniculus	9986	3500	22	Clone-based	NISC - NIH Intramural Sequencing Center
Otolemur garnettii	30611		31	Clone-based	NISC - NIH Intramural Sequencing Center
Pan troglodytes	9598	3100	24		Celera Genomics
Papio anubis	9555			Clone-based	NISC - NIH Intramural Sequencing Center
Papio anubis	9555			Clone-based	University of Oklahoma
Papio hamadryas	9557				Baylor College of Medicine
Pongo pygmaeus	9600			WGS	Baylor College of Medicine
Procavia capensis	9813		19		Baylor College of Medicine
Pteropus vampyrus	132908		19	WGS	Baylor College of Medicine
Pteropus vampyrus	132908		29	Clone-based	NISC - NIH Intramural Sequencing Center
Rhinolophus ferrumequinum	59479	1929		Clone-based	NISC - NIH Intramural Sequencing Center
Sorex araneus	42254	3000		Clone-based	NIH Intramural Sequencing Center (more)

Figure 1.2 Subset of mammalian genome sequencing projects in progress as of November 2007. Data taken from http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi.

Genome databases offer solutions to these problems. By aggregating data from scores of primary databases and integrating data in a uniform and standardized manner, they enable researchers to formulate complex biological queries involving data that were originally located in multiple databases. Learning how to effectively query such interrelated biological data is the primary focus of this book. However, before we can begin this task, we need to spend a little time describing what a genome database is, what the main types of data that it includes are, and how such a database is designed and constructed.

1.1 What is a genome database?

By a genome database, we will mean a data repository (generally implemented via multiple relational databases) that includes all or most of the genomic DNA sequence data of one or more organisms. Generally, a genome database will also include additional data (usually referred to as "annotations") that either describe features of the DNA sequence itself or other biological properties of the species. A genome database typically also includes a web-based user interface – referred to as a "genome browser" – that offers the ability to visualize disparate annotations of genes and other genomic locations in ways that were not possible previously.

Early genome databases and browsers focused on integrating data from a single species, generally one of the biological research community's "model organisms." There was WormBase, for the nematode worm, *Caenorhabditis elegans*; FlyBase, for the fruit fly, *Drosophila melanogaster*; the Saccharomyces Genome Database (SGD) for budding yeast; the Mouse Genome Database (MGD); and so on. Since the completion of the sequencing of the human genome, three additional databases have been developed – EBI's Ensembl Database, the NCBI MapViewer Database, and the UCSC Genome Database – that contain not only integrated human genomic data but also data from many other species as well. This latter feature is important as it becomes increasingly apparent that to interpret the genome of a single species, we need to compare it with its evolutionary relatives. As we will see in detail later in this book, the NCBI, Ensembl, and UCSC Genome Database projects each provide somewhat different, largely complementary resources. Collectively, these projects provide tools and data for genomic analysis that have become indispensable for modern biological research, as evidenced by the fact that UCSC, Ensembl, and MapViewer papers have been referenced more than 3,000 times to date in the scientific literature.

1.2 What classes of annotations are found in the genome databases?

Annotations in the genome databases can be roughly separated into two different classes. The first class includes what might be called "local chromosomal" annotations, as they are associated with a specific region along a chromosome. Examples of such localized annotations include (definitions of unfamiliar terms can be found in the glossary):

- Locations of genes
- Gene-structure annotations indicating a gene's exon-intron boundaries
- Locations of known and putative gene regulatory regions such as promoters, transcriptional enhancers, CpG islands, splicing enhancers and silencers, DNase hypersensitive sites, nucleosome sites, and so on
- Transcript alignments indicating the genomic sources of observed proteins, mRNAs/cDNAs, and expressed sequence tags (ESTs)
- Alignments of protein, mRNA, and EST sequences from related species

- General chromosomal features such as repetitive sequences, recombination "hotspots," and variations in local CG%
- Alignments of genomic DNA from other species, which can provide clues regarding sequence conservation and chromosomal evolution
- Annotations of regions that vary within a population of individuals, including single nucleotide polymorphisms (SNPs), short indels, large structural or copy-number variations, and correlations among sequence variations, such as those that have been identified by the haplotype mapping projects (e.g., HapMap)
- Genome-wide RNA expression data from tiling-array and related projects
- Sequence features that are used in the process of assembling the genome, such as sequence tagged sites (STSs) from genetic and radiation hybrid maps

The other class of annotations includes those that are not directly associated with a genomic region, such as:

- Protein structure data
- Evolutionary data, including evolutionary relationships among individual genes as well as among chromosomal regions and entire genomes
- Annotations describing phenotype variations
- Metabolic- and signaling-pathway data
- Protein-interaction data, such as data from yeast two-hybrid system experiments and data derived from protein-chip expression analysis

To be sure, this distinction between annotations associated with a genomic region and other data is not rigid. However, it can be useful to consider to what extent any given annotation describes a local feature because of the powerful ability provided by the genome databases to address queries involving multiple annotations associated with the same region.

1.3 Building and maintaining a genome database

Building a genome database is a complex multiphase task. Although some of these tasks vary with the specific annotations included within the particular database and with the way the database is designed, certain basic tasks are necessary for the construction of essentially any genome database. These fundamental tasks include:

- Sequencing the genomic DNA
- Assembling the fragments of DNA sequence data into continuous pieces spanning all or most of the length of the organism's chromosomes
- Aligning transcript data to the genomic sequence
- Identifying the locations of the genes within the genome sequence
- Designing and implementing the data-storage architecture to house the data
- Maintaining and updating the database as additional data become available

In many cases, responsibility for the completion of each of these tasks belongs to a different project team. For example, genome sequencing is generally the responsibility of one of the major sequencing centers such as the Broad Institute, the Wellcome-Trust Sanger Center, Washington University, Baylor University, or the Joint Genome Institute. In contrast, sequence assembly is performed by other groups; for example, the human genome assembly was carried out initially by UCSC, and independently by Celera Genomics, and is now performed by the NCBI. Sequence annotation, particularly transcript alignment and gene prediction, have been carried out by yet other groups, for example, Ensembl and NCBI for the human genome. Finally, construction of the genome databases themselves is the responsibility of the groups that will actually provide the genome browser interfaces and maintain the databases, for example, Ensembl, NCBI, and UCSC.

In the following sections, we will introduce each of these tasks briefly. We will return to some of them in more detail later in the context of examining how they impact the information that is available from the genome databases. In addition, these topics are quite broad, and entire books could be (and in some cases have been) written about them. References to the literature are included for those readers who would like to learn more about these important topics.

1.3.1 Sequencing and assembly

To date, nearly all genomic sequencing has been carried out using the conventional Sanger sequencing protocols. With Sanger sequencing, the genome is first randomly cut (e.g., using mechanical shearing) into pieces of between 10 kilobases and 1 megabase, depending on the specific protocol. These pieces are then amplified and subsequently sequenced through a multistep process that involves fluorescent labeling, sequence priming, sequence extension using chain-terminator nucleotides, and electrophoresis (e.g., see chapter 7 of Primrose and Twyman, 2006, for a detailed description). It is worth noting that novel technologies are emerging that show promise for supplanting conventional Sanger sequencing, at least for some applications. These new methods significantly lower costs and increase sequencing output compared to conventional methods. We will describe the potential impact of these emerging technologies in the final chapter.

Because of the random nature with which the original genome is cut, sequencing protocols require that far more bases be sequenced than the number of bases in the entire genome. This is necessary to increase the likelihood that each base will occur and be sequenced from at least one clone. The average number of times any base in the genome has been sequenced in a sequencing project is referred to as the *coverage* of the genome (e.g., a sequencing project of a 100-megabase chromosome with five-fold (5x) coverage involves the sequencing of 500 megabases). For example, so-called draft sequences have 4x to 5x coverage, whereas a finished sequence typically has 8x to 9x coverage. In some cases, for economic reasons, only "low coverage" – that is, 1x to 2x coverage – sequencing is performed (for an interesting discussion of the trade-offs involved in low coverage sequencing, see Green, 2007).

Once a genome has been sequenced, or even partly sequenced, the sequence data needs to go through a process called *sequence assembly*. This is because current Sanger sequencing technology is limited to sequencing no more than approximately 1,000 base pairs in a single data acquisition or "read." (Note: The newer sequencing technologies, though faster and cheaper, have even shorter reads.) In contrast, chromosomes may be 100 million base pairs or more in length. Consequently, chromosome sequence assembly is a complex process in which a large number of overlapping reads are stitched together into longer contiguous regions called "contigs." Subsequently, contigs separated by distances of approximately known length are linked together into "scaffolds." Depending on the sequencing and assembly technology, an additional intermediate step may be necessary to determine the sequence of the individual clones, that is, the pieces of DNA into which the chromosomes were sheared in the initial phase of the sequencing process.

Although this assembly process is straightforward in principle, problems arise in regions where the sequence is highly repetitive or in regions where there are gaps between individual reads. To address these problems, two general strategies for genome sequence assembly have been developed – clone mapping and whole genome shotgun assembly (WGSA). In clone mapping, one first builds a genomic "map" of each chromosome, which includes a list of genetic features or landmarks (e.g., sequence tagged sites) with their relative positions along the chromosome. Using these landmarks, clone and contig sequences can be "anchored" to regions of the chromosome, making it possible to distinguish sequences that are duplicated in other parts of the genome.

In contrast, with WGSA the initial step of building a genomic map is skipped. Instead, the WGSA process includes the cloning of longer (20–50 KB) sequence fragments. One KB of both ends of these clones are then sequenced in individual sequence reads. Using these "paired-end reads," it is then possible to build a scaffold assembly that jumps over ambiguous, duplicated genomic regions without requiring a map of genetic landmarks. Initially, it was unclear whether WGSA would be capable of assembling large genome sequences. However, the effectiveness of WGSA was demonstrated in the assembly of the fly and human genomes, and WGSA has become the primary method of genome-sequence assembly.

Because of the ambiguities in determining the precise location of sequence fragments during genome assembly – no matter which assembly strategy is used – a feature, such as a gene, may be located precisely within a clone or a contig but its location within the entire chromosome might be much less well established. It is for this reason that feature locations are sometimes given in contig or clone coordinates as well as, or instead of, in chromosomal coordinates. Even so-called finished assemblies, such as the current assemblies of the human and mouse genomes, still have gaps. These sequence gaps – for example, those in the centromeric regions – can be quite large. In fact, "finished" assemblies are not really complete at all. Rather, they are simply assemblies that are considered to be as complete as possible within the limits of current technology.

For low coverage (1x–2x) sequences, assembly is much more difficult. In fact, low coverage sequences can generally only be assembled if the genome assembly of a closely related species is available to use as a reference scaffold for ordering the sequence fragments. In addition, with low coverage sequencing, identified genes are often missing exons or are otherwise incomplete. On the other hand, because the costs of sequencing a genome are roughly proportional to coverage, one can sequence approximately four times as many genomes at 2x coverage than one could sequence at 8x. Moreover, since for many comparative genomics applications it is more important to have data from many related species than to have complete-gene sequence data, low coverage sequencing is used in many sequencing projects (see Pontius et al., 2007, for an example of how low coverage sequencing data can be used in the analysis of mammalian genomes). For more details on sequencing and assembly methods, the reader is referred to any modern molecular biology or genomics text, such as Primrose and Twyman (2006).

1.3.2 Transcript alignment and gene prediction

Once the genome has been at least partially assembled, the next step is to locate important biological features – and particularly genes and their exon-intron boundaries – on the assembled sequence. This is not an easy task, and several different strategies, each with its own advantages and disadvantages, have been developed for this purpose. In general, these strategies can be divided into those that are based on transcript alignments, those generated by purely *ab initio* computational predictions, and those that include a combination of alignment and computational approaches. Transcript-based alignments include alignments of proteins, cDNA/mRNAs, and ESTs, both from the genome of the species being sequenced as well as from homologs from related species. In addition, the transcript alignments may be performed completely automatically by computer or may involve manual curation of computer-generated alignments.

In general, gene annotation methods involving manual curation yield fewer false positives – that is, pseudogenes that are annotated as functional genes – than purely computational approaches. However, manually curated approaches are much more labor intensive and tend to generate more false negatives, that is, missed genes. Consequently, depending on the requirements of the specific application (e.g., whether it is more important that one has high confidence that all annotations are correct than that no true genes are being missed), one approach may be preferred over the other.

1.3.2.1 Manually curated gene annotation

The two main projects for manual curation of transcript-based mammalian gene annotations are the Reference Sequence (RefSeq) Project of the NCBI and the Vertebrate Genome Annotation (VEGA) Project of the Sanger Institute. Although the specifics of the RefSeq and VEGA annotation algorithms vary considerably (see Ashurst

et al., 2005, and Pruitt et al., 2007, for details), they both are based on manually curated alignments of transcripts to the genome. Consequently, the RefSeq and VEGA datasets often agree, particularly in gene detection and in distinguishing functional genes from pseudogenes. However, RefSeq and VEGA annotations do not always agree, especially in terms of their predicted exon-intron boundary locations.

To address the fact that RefSeq and VEGA annotations sometimes differ, yet another manual curation project, the Consensus Coding Sequence (CCDS) Project, has been started (http://www.ncbi.nlm.nih.gov/CCDS). The goal of CCDS is to identify highly reliable gene annotations, namely those for which there is 100% agreement between the RefSeq and VEGA annotations and that meet other quality tests developed by the CCDS Project, for example, tests to confirm that the predicted gene is neither a processed pseudogene nor produces a transcript that would be subject to nonsense mediated decay. Currently, the CCDS dataset is restricted to human gene annotations; however, expansion to other mammalian species (e.g., mouse) is planned.

1.3.2.2 Automated gene annotation

Compared to the curated RefSeq, VEGA, and CCDS datasets, the fully automated gene prediction systems provide larger numbers of gene annotations, and producing them is much less labor intensive. Such automated gene-prediction algorithms include both systems, such as the Ensembl Pipeline (Curwen et al., 2004; Potter et al., 2004), which are based largely on transcript alignments, and the *ab initio* computational gene-prediction programs. Furthermore, the *ab initio* programs can be partitioned into two major subclasses: single species gene-prediction programs, such as GENSCAN (Burge and Karlin, 1997), and newer programs that use multiple-species sequence alignments (Gross and Brent, 2006). Gene finders based on multiple sequence alignments rely on the fact that genes and gene structures are typically conserved in related species. Consequently, if a predicted gene splice junction has a consensus splice-site sequence that is conserved in other species, then the site is more likely to be genuine than if the splice-site sequence is not conserved. By using the additional information contained in sequence alignments, multispecies gene prediction programs usually have considerably lower false positive rates than single species programs (Brent, 2007).

Whether they are based on transcript alignment or on *ab initio* predictions, datasets produced by the automated pipelines generally have higher false positive rates or incorrect intron-exon boundaries than the manually curated datasets. However, despite their higher false positive rates, the non-curated datasets can be very useful. For example, many genuine genes are expressed only at low levels or in specific tissues or developmental stages. Consequently, transcripts of these genes may not have been experimentally detected to date, and such genes will generally not be included in the curated datasets. And of course, for non-model species for which there is little transcript data, non-transcript-based computational approaches are the only available tool. Nevertheless, it is important to remember that the higher false

positive rates for non-curated datasets means that these datasets need to be viewed with more caution.

1.3.2.3 Accuracy of gene prediction methods

With so many different ways of predicting genes, it is important to be able to assess the relative accuracies of the different approaches. Addressing this question for human gene annotation is one of the goals of the "Encyclopedia of DNA Elements," or ENCODE Project (Birney et al., 2007). The ENCODE Project ultimately seeks to annotate all functional DNA motifs in the human genome. In its pilot phase, the ENCODE Consortium has generated a large number of annotations for a small (around 30 MB) subset of the human genome.

One of ENCODE's initial objectives has been to generate a complete list of functional protein-coding transcripts in the 30-MB ENCODE regions, the so-called GENCODE gene set. This list of transcripts was generated by gene predictions from multiple computational and curated gene annotation methods, followed by experimental PCR-based validation, in over 20 different types of human tissue. The results (Harrow et al., 2006) indicate that current methods of gene annotation are quite good – but far from perfect. In addition, they confirm that current manual curation methods are generally more specific, but less sensitive, than fully automated approaches. For example, ENCODE determined that Ensembl's automated gene prediction pipeline detected 84.0% of the validated gene exons, whereas RefSeq's manually curated algorithm detected 80.0%. On the other hand, 98.3% of RefSeq's exon predictions could be experimentally verified, as compared to 91.5% for Ensembl.

Because there are so many different approaches for genomic gene identification, with different strengths and limitations, one often finds multiple different "gene" annotations in the genome browsers. We will consider this topic in more detail in later chapters. For now, suffice it to say that, for example, in the hg18 build of the UCSC Human Genome Database, there are more than a dozen different sets of protein-gene annotations. So if one is searching for data about a specific human gene, which annotation set should one use? Although there are no hard-and-fast rules, a useful guide would be to first check whether the gene is annotated in the CCDS dataset.[1] If not, one could check if it is included in VEGA or RefSeq. If the gene is not found in any of the manually curated sets, one could check an automated gene annotation dataset such as UCSC genes or Ensembl genes, or a modern *ab initio* gene prediction program such as N-SCAN. Finally, we note that this discussion has been limited largely to annotations of vertebrate genes; other curated gene annotation datasets are available for the non-vertebrate model species, such as the SGD Gene Set for yeast genes and the FlyBase Gene Set for *D. melanogaster* fly genes.

[1] If the region of interest is within the ENCODE regions, using the GENCODE gene set would also be a good choice.