C.C. Englbrecht, M. Han, M.T. Mader, A. Osanger, K.F.X. Mayer

MIPS, Institute for Bioinformatics
GSF – National Research Center for Environment and Health
Neuherberg, Germany

Review

Curated databases and their role in clinical bioinformatics

The "*–ics*" sciences and their role in molecular medicine

Within the last decade modern biology and medicine underwent a paradigm shift. The large scale genome data available now complements long established research routes that apply epidemiological and molecular studies on individual genes. The challenge to analyze causes and consequences of human-pathogen interactions and the molecular basis of human diseases and plagues on a genomic scale bears so far unknown opportunities for the understanding of molecular mechanisms and the development of effective therapy and drugs. However, the sheer amount of data is overwhelming. Therefore the successful usage of genome information depends on the comprehensive analysis of genome data, the storage of genome and genome associated data, tools for inter-genome comparisons and knowledge transfer, and the iterative enrichment of information resources with the most current research results.

Within this chapter we give an overview of the broad variety of genome and genome associated ("*-ics*") resources that are important for clinical research. An emphasis is put on the discussion of restrictions, challenges and opportunities of the various analyses and on the challenges and necessities in structuring and organizing the enormous amount of intrinsically heterogeneous data. Clearly it is mandatory to further develop database standards. We need them not only for the handling and organization of the data, but as essential tools and prerequisites in order to carry out any genome based research. Structuring and provision of large scale genomic data is a demanding task. Nevertheless, only with the fulfillment of this task, the wide range of opportunities offered by the data and their comparative as well as combinatorial potential can successfully be used and exploited for medical applications.

Genomes: of Mice and Men and More

History of Genomics

The genome projects of the past 10 years considerably increased the amount of data available for biomedical research (see Fig. 1 and Fig. 2a). Due to the immense development and rapid acceptance of the internet, molecular data spread quickly through web accessible databases. Genome sequences provide information on the composition and organization of particular chromosomes and genomes, on complete sets of genes and their location on the chromosomes. More in depth analyses can also address complex questions about relationships within one genome or among different species through comparative genomic means. As a consequence of the rapid increase of large scale sequence data, the number of databases increased dramatically. These databases exhibit both numerous interface varieties and an enormous heterogeneity with respect to data content, object description and the format of the data (see Table 1).

The generation of expressed sequence tags (ESTs) is a hallmark for the beginning of the genomic age (see Fig. 1). ESTs are transcribed sequences which are being sequenced partially and are of comparably low quality [1]. ESTs give insights into large portions of the transcribed genome and allow for first approximations of the particular genomes. Today, enormous amounts of ESTs from a wide variety of organisms exist. To eliminate redundancy and give comprehensive insights into the particular transcriptomes, computational strategies to collapse ESTs into clusters and assemblies have been developed [2-4].

The advent of a new type of mass sequence data, whole genome sequences, dates back to the mid-1990s (see Fig. 1). The start was made with the bacterium *Haemophilus influenza*

C.C. Englbrecht, M. Han, M.T. Mader, A. Osanger, K.F.X. Mayer

MIPS, Institute for Bioinformatics
GSF – National Research Center for Environment and Health
Neuherberg, Germany

Review

Curated databases and their role in clinical bioinformatics

The "*–ics*" sciences and their role in molecular medicine

Within the last decade modern biology and medicine underwent a paradigm shift. The large scale genome data available now complements long established research routes that apply epidemiological and molecular studies on individual genes. The challenge to analyze causes and consequences of human-pathogen interactions and the molecular basis of human diseases and plagues on a genomic scale bears so far unknown opportunities for the understanding of molecular mechanisms and the development of effective therapy and drugs. However, the sheer amount of data is overwhelming. Therefore the successful usage of genome information depends on the comprehensive analysis of genome data, the storage of genome and genome associated data, tools for inter-genome comparisons and knowledge transfer, and the iterative enrichment of information resources with the most current research results.

Within this chapter we give an overview of the broad variety of genome and genome associated ("*-ics*") resources that are important for clinical research. An emphasis is put on the discussion of restrictions, challenges and opportunities of the various analyses and on the challenges and necessities in structuring and organizing the enormous amount of intrinsically heterogeneous data. Clearly it is mandatory to further develop database standards. We need them not only for the handling and organization of the data, but as essential tools and prerequisites in order to carry out any genome based research. Structuring and provision of large scale genomic data is a demanding task. Nevertheless, only with the fulfillment of this task, the wide range of opportunities offered by the data and their comparative as well as combinatorial potential can successfully be used and exploited for medical applications.

Genomes: of Mice and Men and More

History of Genomics

The genome projects of the past 10 years considerably increased the amount of data available for biomedical research (see Fig. 1 and Fig. 2a). Due to the immense development and rapid acceptance of the internet, molecular data spread quickly through web accessible databases. Genome sequences provide information on the composition and organization of particular chromosomes and genomes, on complete sets of genes and their location on the chromosomes. More in depth analyses can also address complex questions about relationships within one genome or among different species through comparative genomic means. As a consequence of the rapid increase of large scale sequence data, the number of databases increased dramatically. These databases exhibit both numerous interface varieties and an enormous heterogeneity with respect to data content, object description and the format of the data (see Table 1).

The generation of expressed sequence tags (ESTs) is a hallmark for the beginning of the genomic age (see Fig. 1). ESTs are transcribed sequences which are being sequenced partially and are of comparably low quality [1]. ESTs give insights into large portions of the transcribed genome and allow for first approximations of the particular genomes. Today, enormous amounts of ESTs from a wide variety of organisms exist. To eliminate redundancy and give comprehensive insights into the particular transcriptomes, computational strategies to collapse ESTs into clusters and assemblies have been developed [2-4].

The advent of a new type of mass sequence data, whole genome sequences, dates back to the mid-1990s (see Fig. 1). The start was made with the bacterium *Haemophilus influenza*

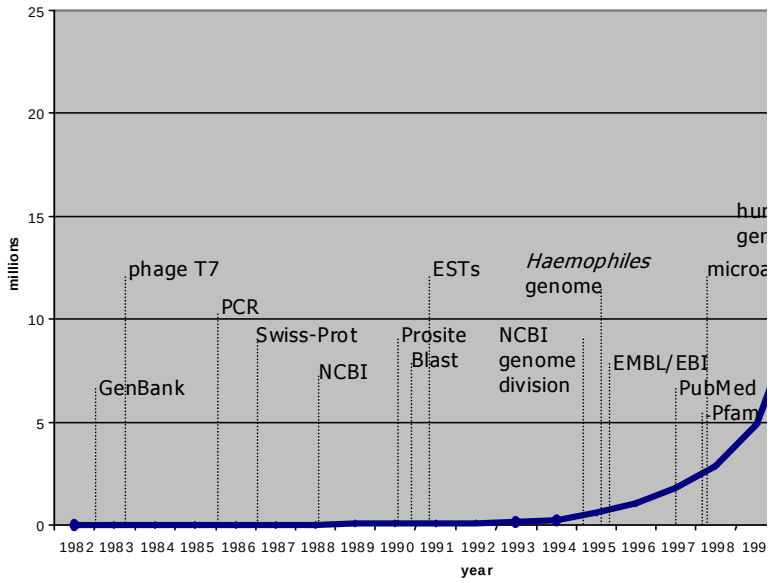

Fig. 1. Timeline of several essential developments in molecular biology, genomics and bioinformatics from 1982 till 2002, charted against the accumulation of increasing DNA sequences in GenBank. Cumulative sequences (in million basepairs) are shown in blue. (Information on the growth of Genbank is available at: http://www.ncbi.nlm.nih.gov/GenBank/genbankstats.html).

[5], soon followed by the yeast *Saccharomyces cerevisiae* [6]. Only two years later the genome sequence of a more-complex multicellular organism, the roundworm *Caenorhabditis elegans*, has been reported [7]. In 2000, the genome sequence of another important invertebrate model, the fruit fly *Drosophila melanogaster*, was published [8], and as a culmination the draft sequences of man [9, 10] and mouse [11] were reported. Today, the complete genomes of more than 100 organisms from all kingdoms of life are available (see Fig. 2a and Fig. 2b), and huge efforts are made to constantly update and improve the sequence data of the human genome [12].

Strategies for Genome Sequencing

The most common approach for the generation of sequences of large, complex genomes involves the establishment of an ordered subset of large-insert genomic clones from which a physical map of the respective genome is generated. Each genomic clone is sequenced with high accuracy and finally the sequences of individual clones are re-assembled into a total genome sequence. This classical strategy is named clone-by-clone shotgun (CBCS) sequencing [13] and is best exemplified by the efforts to sequence yeast [14], roundworm [7] and human [9] genomes.

An alternative strategy for genome sequencing is to apply a whole-genome shotgun (WGS) sequencing strategy [13]. Here unordered, highly redundant shotgun sequence libraries of the entire genome are generated. Subsequent massive application of bioinformatic and computer assisted analysis aim to assemble the millions of short nucleotide sequence reads into a complete genome sequence. In principle, this shortcut bypasses the need for a clone-based physical map.

Advantages and disadvantages of the WGS strategy have been discussed controversially [15,16]. Nevertheless, the two strategies are not mutually exclusive. There has been a remarkable convergence in the use of these sequencing approaches, resulting in the advent of hybrid strategies that incorporate elements of both. Prominent examples for this mixed approach are the sequencing projects of mouse, rat and zebrafish [13].

Status of the Human Genome

The sequences of human and other organisms represent fundamental information for biology and biomedicine. It became clear that the structure of the human genome is extraordinarily complex and the elucidation of the function of the genome in its whole complexity is far from being understood. Only 1–2% of its bases encode proteins [9] and up to now the full set of protein-coding sequences has not been determined with high reliability. Initially, an approximate number of about 25.000-35.000 human genes have been reported. In addition, a large number of thus far uncharacterized, non-coding sequences is under selective pressure, suggesting functional importance of these regions [11,17]. One significant class of genes, often missing from contemporary genome annotations, is the group of non-protein-coding RNAs (ncRNAs). ncRNAs constitute a major functional output of the genome and play a major role in protein synthesis, genomic imprinting [18], and the control of genetic networks [19]. Even less is known about the role of roughly half of the genome which consists of highly repetitive sequences. Furthermore, the relatively small number of new genes detected in the human genome has led to a renewed focus on the role of the precise regulation of gene transcription as well as alternative splicing in mediating the complexity of mammals.

Models as a Means to Study Human Genes

With the availability of a "finished" sequence of the human genome, the primary focus is to identify the complete set of both protein-coding and non-protein-coding mammalian genes. Although certainly being a primary goal, a comprehensive data resource that contains the complete description of a mammalian transcriptome, thus far has only been partially realized. With the shortcomings in computational

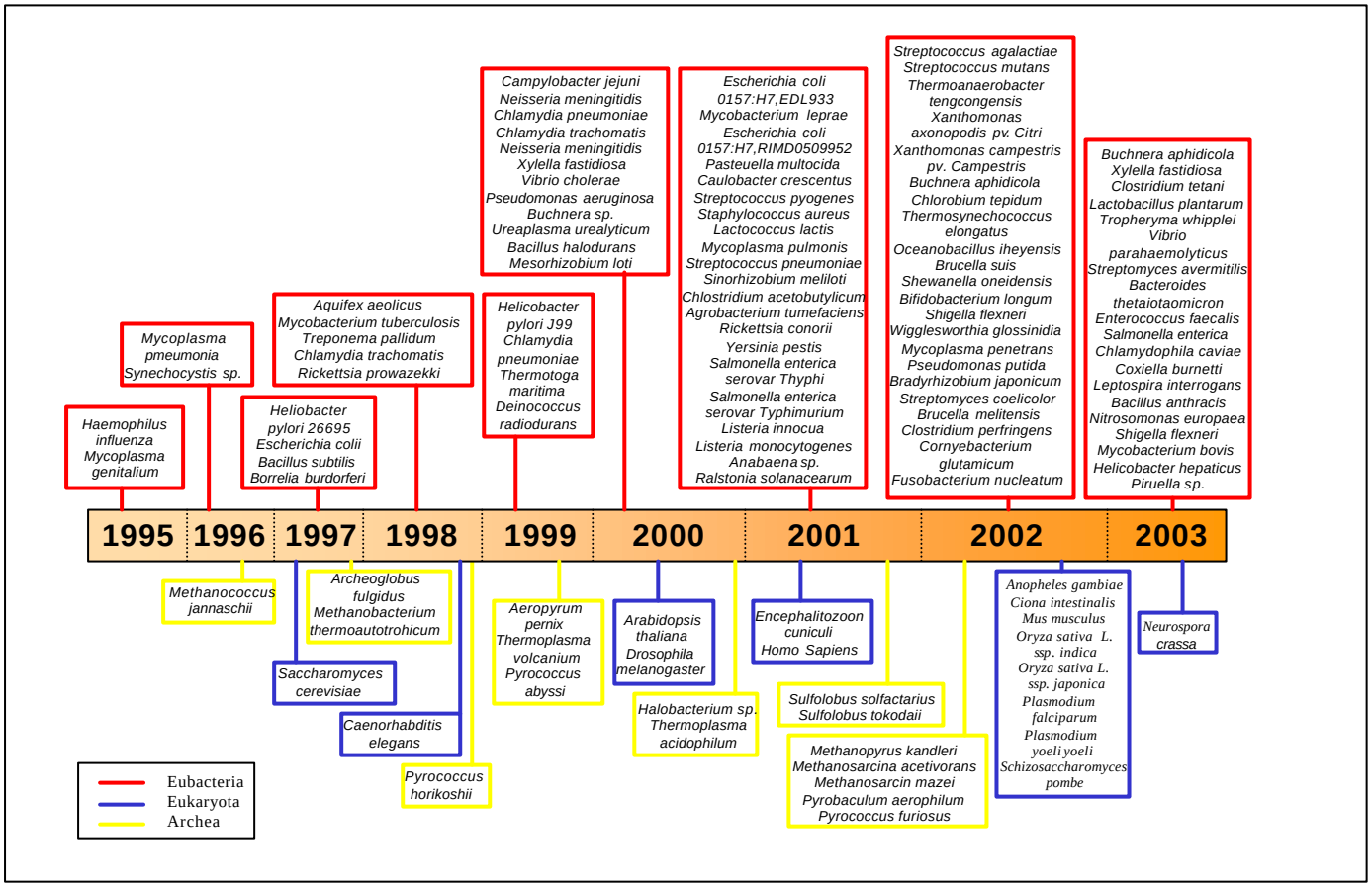

Fig. 2a. Genomes of living organisms (Archea, Eubacteria, and Eukaryota) sequenced between 1995 and 2003 (data taken from: Complete Genome Tracking Database: http://maine.ebi.ac.uk:8000/services/cogent).

detection and definition of genes, large scale experimental identification of transcribed units is a highly useful resource for defining genetic features on the genomic backbone. One approach is the systematic isolation and characterization of full-length cDNA sequences [20-23]. The generation of a set of cDNAs that contains the complete and uninterrupted protein coding regions of all human and/or mouse genes provides a valuable means for the accurate identification of genes, products of alternative splicing and the systematic and comprehensive analysis of protein [24].

Major challenges inherent in further programs for the discovery of genes are the experimental identification and validation of alternate splice forms and mRNAs that are expressed in a very restricted

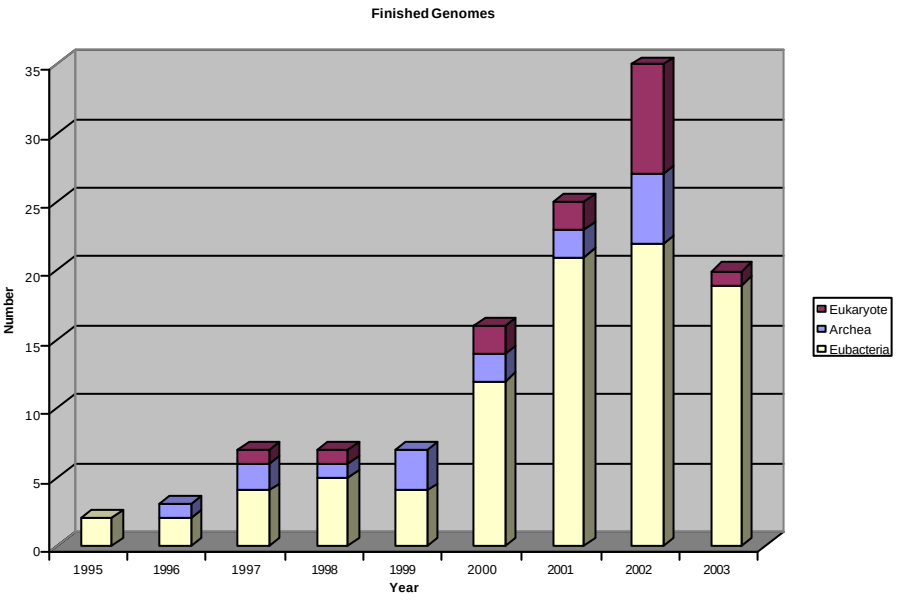

Fig. 2b. The growth of finished genomes per year, from 1995 until 2003. Stacked columns indicate the cumulative number of finished organisms. Eukaryota shown in purple, archea in blue, and eubacteria in yellow (data taken from: Complete Genome Tracking Database: http://maine.ebi.ac.uk:8000/services/cogent).

way. However, additional information of protein-protein interaction, cDNA microarray expression profiling, protein localization and structural genomics experiments have to be integrated with existing data.

Quality of Sequences and Models

While unraveling large genomes, technical complications and limitations in data quality and related annotation have been reported. A number of reports discussed sequence errors and chromosomal mis-assignments for gene sequences and entire contigs of the first human genome draft [25,26]. However, for cases where the sequence of a closely related genome has already been finished (e.g. human vs. chimpanzee), the implementation of the WGS strategy has the potential to provide a considerable acceleration of the assembly procedure. In such cases conserved syntenic regions can be used as additional information for the assembly [27].

In simpler eukaryotic organisms such as yeast, the majority of the genome encodes for proteins, and individual genes generally have a well-defined start and stop and a single mRNA transcript. The mammalian genome organization is considerably more complex and so is the challenge to detect individual elements. Only a small portion of the genome encodes mRNAs and the detection and modelling of genes is very complex [28,29]. In the fruit fly *Drosophila melanogaster*, *ab initio* gene prediction methods correctly predicted about 79% of individual exons [30], in contrast to about 70% in human [31]. In part, this is caused by small exons which can be separated by long introns, or the usage of rare and unusual splice sites and alternative genes products. Thus, further experimental evidence, e.g. from cDNAs and ESTs, is a highly valuable resource for the detection and accurate modelling of genes within complex genomes [32].

Why Genome Sequences of Related Mammals?

The increasing number of finalized and draft metazoan genome sequences provides new opportunities for biomedicine and genetics. Comparative genomics approaches probably represent the most powerful strategies [33].

With the availability of the assembled mouse and human genome the alignment and comparison of two large vertebrate genomes has now become feasible [11]. The completion of the sequencing of the mouse genome enables to delineate human genes with greater accuracy. While current *ab initio* gene prediction programs are remarkably sensitive (i.e., they predict at least a fragment of most genes), their specificity is often low and they predict a large number of (probably) false-positive genes in the human genome. Human-mouse sequence conservation at the protein level helps to eliminate some of those.

For the study of diversification, the comparison of closely related genomes is a highly valuable instrument. For this purpose, the alignment of the human genome with those of apes and monkeys are of importance. The comparison of cDNA sequences of the cynomolgus monkey (*Macaca fascicularis*) with human genome sequences already proved the usefulness of this approach [34] and genome sequencing of the chimpanzee (*Pan troglodytes*) is underway. As a complementary approach, the analysis of high throughput cDNA from orang utan (*Pongo pygmaeus*) has already been initiated (S. Wiemann, personal communication). Without doubt, comparative genomic analysis between human and ape genomes will lead to unprecedented insights into human evolution [27].

Data mining of all available genomic databases from completely or partially sequenced organisms enables the detection of orthologous genes. Interspecies genomic comparison is a powerful tool to infer the function of genes as it allows to project knowledge gained within one particular organism to a related organism [33]. Comparative genomics databases facilitate the identification of evolutionarily conserved genomic sequences, genes and gene families, and thus to enrich the annotation of the human genome. These analyses have the potential to identify new exons and highly conserved non-coding regulatory elements by the comparison of the upstream regions from orthologous genes. However, with the apparent high sequence redundancy within mammalian genomes, orthology assignment based on pure sequence homology often leads to ambiguous data. Thus beside homology based assignment, orthology assignment by syntenious localization of the respective gene pairs is of importance [33,35].

Presentation and Visualization

A first and immediate outcome of any genome analysis project is the deposition of unordered and fragmented genomic sequences in public sequence databases. While this information is a valuable resource for researchers interested in particular genes, many biomedical scientists need to gain knowledge on additional contextual information. For example, in positional cloning projects it is preferable to know the order and relative orientation of genes, markers and repeats within a given interval. This information can only be derived from assembled consensus sequences. Therefore efforts are made to assemble overlapping genomic fragments into contigs and anchor them to individual regions of the respective genomes. Although this is carried out for a variety of species and their respective genomes, a special emphasis is of course put on the human genome. Among others the "Golden Path" assemblies at the University of California, Santa Cruz (UCSC, http://genome.ucsc.edu) and the contig assemblies from the National Center

of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) have been representative major resources. Various websites now offer the opportunity to browse annotations of the human genome as well as draft sequences from other species. Among the most prominent and exhaustive data collections are the UCSC Human Genome Browser [36], the NCBI Map Viewer [37] and the Ensembl database [38] (see also Table1).

Comparative analysis and visualization tools have been developed that allow the exploitation of genome sequences of related organisms in order to search for orthologous genes. Among the most commonly used comparative genomic tools are VISTA (Visualization Tool for Alignment) and PipMaker (Percent Identity Plot maker) [35,39,40] (see Table1).

Pharmacogenetics: from SNPs to Therapy

Pharmacogenetics is the combination of pharmaceutical knowledge and genetic information. The most frequently discussed pharmacogenetic application is the analysis of the genetic influence on an individuals response to drugs. The goal is the development of personalized medication based on the knowledge of an individual's genetic profile and of the genetic background of the disease, which could reduce side effects and increase the effectiveness of the treatment. An important strategy is the analysis of genetic variations, such as single nucleotide polymorphisms (SNPs) and microsatellites in order to detect important genotypes of potential drug target genes and regulatory elements, e.g. drug-metabolizing enzymes and receptors [41-43].

The functional classification of genes, the association of functional categories with pathway information, and the association of data on drug metabolism with the function of known and unknown genes is essential in order to be able to select genomic regions that are of potential interest for further experimental analyses. This is of outstanding importance as quantitative trait loci (QTLs) - genetic loci that influence quantitative traits like weight or immunological parameters, in contrast to mendelian loci, which show dichotomous phenotypes - might affect phenotypes in less obvious ways. Complex diseases (e.g. asthma), which show phenotypes determined by multiple factors, are most likely influenced by multiple loci [44] and require the analysis of multiple candidate genes/regions. These analyses can only be performed with the help of integrated databases. Such databases are hence invaluable for pharmacogenetic studies, as only the combination of genome based sequence data, experimental evidence, annotation and clinical results can lead to the identification and characterization of the complex genetic background of a given phenotype.

Currently, most databases that provide special information on pharmacogenetics are commercial. However, publicly available databases are increasing in number (Table 1). An example is PharmGKB. It contains the annotation of genes, including pharmacogenetically relevant publications, associated diseases and genetic variations.

SNP Based Studies

The use of SNPs is increasing in research and diagnostics world wide. It became very popular to analyze the genetic makeup of populations and individuals [45].

Most studies dedicated to the characterization of pharmacogenetically interesting genotypes are based on linkage and association studies [46]. While conducting such studies, the genetic association of markers with the attribute of interest is analyzed. Besides microsatellite markers, SNP based markers are widely applied. SNP based markers offer additional opportunities to other marker types. They are available at comparably high densities and in ideal cases a particular SNP already tags the particular gene and genome position (e.g. the amino acid codon) that causes the observed effect. With the development of new techniques the costs for genotyping will be further reduced, which will make large-scale whole genome SNP screens economically feasible in the near future [47].

One approach to reduce the numbers of SNP markers used for studies and therefore the costs involved in a whole genome analysis, is the development of haplotype maps [48]. These maps represent sets of SNPs that always co-occur due to a high linkage disequilibrium in the respective genomic areas. Hence, it is possible to represent a complete haplotype block by typing a single SNP only. The amount of SNPs that have to be analyzed can be reduced 20 to 40 fold from 10.000.000-20.000.000 to 500.000. This in turn drastically reduces the costs of SNP based genotyping required for standard analysis [49].

Many SNPs published in public SNP collections lack validation and manual curation. Especially the increase in accuracy of SNP detection from EST assemblies [50] and also the efforts of publicly funded revalidation projects will lead to an improved quality of the data. In addition, the availability of information on population specific SNPs allows to focus on epidemiological genetic features. This has been demonstrated by the company deCODE genetics, which is using comprehensive genetic, disease and genealogical information of Iceland's founder population to detect disease linked genes [51].

Candidate-gene approaches, e.g. attempts to isolate particular genes closely associated to or causative for a

specific phenotype, are the most frequently used analyzes carried out. Often these kinds of analyzes are supported by genome wide linkage studies. For a successful and time efficient completion of such projects detailed knowledge and information of the genetic background is a necessity, because it allows for the rapid detection of appropriate markers within the region of interest. In addition, the screening for candidate genes in the respective region has the potential to speed up the analysis process dramatically. Screens for candidate genes are merely based on the molecular characteristics of the genes located within the regions. Thus, to understand the impact of potential candidate genes located inside the regions of interest, knowledge of their function and position in functional networks (metabolism, signaling, etc.) is essential. Hence the availability of high quality, manually curated information on metabolites, associated pathways, diseases and effects on drugs is a prerequisite for the characterization of genetic influences on diseases and provides the link for associating epidemiology to molecular biology (Table 1).

Alternative Transcripts

It has been estimated that between 35% - 59% [52] of all human genes encode for more than one transcripts. The impact of these alternative products on function and regulation has become a major focus in recent years [53]. In addition the identification of splicing-regulatory elements like exonic splice enhancers [54] and the search for conserved splice variants between species led to the establishment of various databases harboring information on alternatively spliced transcripts. As the cause and effect of alternative splicing in a genome, transcriptome and proteome context seem to vary widely, most public databases of alternative splicing focus on the alignment of sequences of alternative products in the context of a genomic reference or full length cDNA sequence. Differential expression, gain/loss of function, and the influence of point mutations on the occurrence of splice patterns are only rarely included. Therefore the current focus is mainly directed towards the development of high-throughput methods to structure and analyze the enormous amount of ESTs from public repositories rather than on the manual curation of detected splice patterns.

An important and complementary focus of research conducted on alternative splicing is the identification of transcript specific expression patterns [55]. The relevance of alternative splice products in diseases seems to be high [56,57]. Therefore the development of strategies that allow to distinguish between the expression of alternatively spliced transcripts will lead to new insights and will also have a significant implication in the field of expression analysis.

Infection: Viruses, Bacteria and Other Pests

One of the biggest burdens for humanity is infectious disease. Bacteria, viruses and protozoan parasites are still major causes of death despite the steadily increasing understanding of the mechanisms of pathogenicity and the constant effort to develop novel drugs. According to the National Institute of Allergy and Infectious Diseases (http://www.niaid.nih.gov/) about 13 million people die annually due to infectious diseases which accounts for one quarter of deaths around the world. Approximately 1.7 million people die every year from tuberculosis, 1.1 million from malaria and 2.9 million from human immunodeficiency virus (HIV) infections. However, the biggest threat comes from infections of the lower respiratory tract with 3.9 million deaths per year.

Conventional drugs that used to be quite effective, such as common antibiotics or chloroquine, against *Plasmodium falciparum*, the malaria parasite, are losing their efficiency due to rapidly spreading resistance. The most recent alarming case was the emergence of a vancomycin resistant strain of *Staphylococcus aureus* [58]. Diseases that were thought to be under control are returning. For example tuberculosis has reemerged due to decreasing standards of hygiene and as a consequence of the spread of HIV. There is an urgent need to improve existing drugs and vaccines and to develop new ones against known microorganisms and newly emerging ones, e.g. the SARS-coronavirus (SARS-CoV). The development of diagnostic tools that allow a rapid test is also mandatory in order to fight infectious diseases efficiently.

With the advent of the genomic age, it became evident that information on the genomes of pathogens and their relatives could be of tremendous help to fight infectious diseases. The first genome of a free living organism ever sequenced was that of *Haemophilus influenzae* [5]. Since then, over one hundred bacterial genomes and more than 1000 viral genomes have been completed, the SARS-CoV being the latest one as of the writing of this article. The sequencing of major protozoan parasites is also under way and finished for some genomes like *Plasmodium falciparum*, the causal agent of malaria. Major centers for sequencing of bacterial genomes are the Sanger Center in the UK (http://www.sanger.ac.uk/Projects/Microbes/) and the Institute of Genomic Research (TIGR, http://www.tigr.org/tdb/) in the US, but there are also other institutions like ACGT (Univ. Oklahoma, http://www.genome.ou.edu/) or Univ. Wisconsin-Madison Genome Sequencing (UWisc, http://www.genome.wisc.edu/) to name a few (see also Table1).

Mechanisms of Pathogenicity

The fight against pathogens focuses on mechanisms of pathogenicity and the immune response of the host. Main virulence factors of bacteria are capsules, cell wall components, toxins and adhesins. The first and crucial step in infection is the invasion of the host. There are extracellular pathogens like *Staphylococcus aureus* that break down host barriers without the invasion of cells. Intracellular pathogens penetrate the cell membrane and persist inside the cells either in the cytosol (e.g. *Lysteria monocytogenes*) or in vacuoles (e.g. *Mycobacterium tuberculosis*).

The human host has several defenses against invaders in critical places such as epithelial cells of the skin and the respiratory tract, lysozyme in tears and low pH in the stomach. After successful invasion, the immune system of the host reacts to more general pathogen associated molecular patterns or specific antigens.

Variability is known for the human immunoglobulin genes, the major histocompatibility complex and virulence genes, as in the case of the cag island in *Helicobacter* [59]. Reciprocal polymorphisms in genes involved in host-pathogen interaction are interesting candidates for co-evolution, the adaptive genetic changes between interacting species [60]. It is known that a great deal of bacterial evolution is mediated through genome rearrangements (collectively referred to as horizontal transfer) and plays an important role in the molecular evolution of novel pathogens. In the process of horizontal transfer, genomic islands from a donor organism are incorporated in the genome of the recipient organism. These islands can contain large blocks of virulence genes (pathogenicity islands). Plasmids often carry genes that confer antibiotic resistance and bacteriophages also contribute to the horizontal transfer of virulence genes.

Pathogen Functional Genomics

The wealth of genome sequences available for pathogens and also their non-pathogenic relatives has opened up possibilities to understand pathogenesis and find candidates for virulence genes. The understanding of pathogenesis and the identification of virulence factors provides a basis for the development of anti-microbial drugs, vaccines and diagnostic tools. One key methodology is the *in silico* comparison of complete genomes, though the analysis of single genomes also allows to find potential virulence factors. With more and more genomes being finished, a plethora of meaningful and promising comparisons is possible. Once genomes are available that allow for meaningful comparisons, several aspects can be studied that in combination help to reach the goal of an effective fight against pathogens.

The intent of genome studies and comparisons varies. Depending on the hypothesis, either closely related or distantly related genomes, whole genomes, open reading frames (ORFs) only or regulatory regions are of interest. As described above, many virulence factors are already known. Some of them combine features that can be searched for *in silico* even when only the sequence of a single genome is available. Cell-wall or secreted proteins are of particular interest. Pizza *et al*. could identify 570 putative cell-wall or secreted proteins in *Neisseria meningitidis* [61]. Further immunization screens in mice led to the extraction of two conserved vaccine candidates. Another promising approach is the comparison of strains or species that are related but have differing virulence (e.g. *Neisseria meningitidis* serotype A and B), infect different tissues in the human host (e.g. *Bacillus anthracis* and *Bacillus cereus*), or infect hosts other than human (e.g. *Mycobacterium tuberculosis* and *Mycobacterium bovis*).

Horizontal Gene Transfer

Not too long ago only a limited amount of information from fairly distantly related genomes was available. Thus, comparisons were mainly focused on phylogenetic studies that allowed for some insight into the relatedness among bacterial groups and horizontal transfer among bacteria which can substantially blur phylogenetic signals. The understanding of the evolution of bacteria, i. e. the history and mechanism of the exchange of genetic material, will help to elucidate the virulence mechanisms of emerging and reemerging infectious diseases and changes in virulence associated with these infections.

Horizontally transferred DNA plays an important role in the exchange of virulence genes and the acquirement of resistance. Recently, the analysis of the complete sequences of the vancomycin resistant *Enterococcus faecalis* (an opportunistic pathogen that is the major cause of urinary tract infections, bacteremia and infective endocarditis) showed that more than one quarter of the genome consists of probable mobile or foreign DNA [62]. The authors argue that the propensity for the incorporation of mobile elements probably contributed to the rapid acquisition and dissemination of drug resistance.

Buchrieser *et al*. showed that horizontal transfer accounts for differences in *Lysteria monocytogenes* (a food-borne pathogen) and *Lysteria innocua* (non-pathogenic) [63]. Perna *et al*. compared *Eschericha coli* K12 with *Eschericha coli* 0157:H7, an entero-haemorrhagic relative [64]. They found that lateral gene transfer is extensive and offered a wealth of candidate genes that may be involved in virulence.

Pathogenicity in Different Tissues and Hosts

Recently, attention has focused on *Bacillus anthracis*, because it became notorious as a bioweapon. It is an endospore-forming bacterium that

causes inhalational anthrax. The comparison of the genomes of *Bacillus anthracis* and *Bacillus cereus*, an opportunistic pathogen causing food poisoning [65,66] facilitates the identification of candidate genes responsible for pathogenesis. The comparative genome sequencing of *Bacillus anthracis* by Read *et al*. revealed markers in the highly monomorphic species that divide the anthrax isolates into distinct families [67]. Members of the *Mycobacterium* group are responsible for many millions of deaths worldwide. Starting with the genome sequence of the laboratory strain of *Mycobacterium tuberculosis* H37Rv, the causal agent of tuberculosis, in 1998 [68], by and by the genome sequences of several relatives have been published: the one of the leprosy bacillus *Mycobacterium leprae* [69], of *Mycobacterium tuberculosis* CDC155, a clinical isolate [70] and of *Mycobacterium bovis* [71] which is a pathogen primarily in cattle but also in humans. The comparison of these species and strains with differing host preferences and pathogenicity will provide insight into the evolution of this species complex and the virulence factors involved in pathology.

Once potential virulence genes are found they can be used in vaccine screens. Virulence genes can be quite polymorphic. Therefore it is of interest to compare different strains of the same pathogenic species. For vaccine development, it is favorable to choose factors that are conserved within a species. This applies even more broadly for the development of antibiotic drugs. The pharmaceutical industry is mainly interested in the development of broad-spectrum antibiotics that act on many different species. The comparison of genomes of several strains and/or species can pinpoint the highly conserved genes that are potentially involved in virulence and exclude those that are too variable. There are exceptions as in the case of *Helicobacter pylori* which has an elevated intra-strain variation [72,73]. In the treatment of the chronic diseases caused by *Helicobacter pylori* (an organism that can cause ulcer and cancer) it can be advantageous to develop species specific antibiotics for long term treatment in order to avoid stress for the host.

The comparison of different strains is crucial in the study of the HIV. It elucidates the rapid evolution of some genes, e.g. *env* [74] and *gag* [75], that can render many therapy and vaccination approaches useless.

The causal agent of malaria is the protozoan parasite *Plasmodium falciparum*. It is transmitted to the human host by the mosquito *Anopheles gambiae*. The parasite needs these two hosts to complete its life-cycle. By the end of 2002, a wealth of genomic information was published that could considerably accelerate malaria research. In addition to the human sequence, genome sequences of *Anopheles gambiae* [76], *Plasmodium falciparum* [77] as well as *Plasmodium yoelii yoelii*, the causal agent of rodent malaria [78], were published. The comparative genomic analyzes of these genomes, including the well studied genome of *Drosophila melanogaster* will provide insight into host and parasite specific virulence and defense mechanisms, respectively [79-81].

In order to find novel drugs and vaccines, the study of the human host genome of course is equally important. Drugs and vaccines will fail if they inhibit essential functions in the host. They will also fail if they are self-antigens and hence poor immunogens or cause the production of auto-antibodies.

Prerequisites for Further Experimental Studies

With the help of software tools and high quality databases, clinical bioinformatics can find potential candidates for drug targets and vaccines and generate important clues about which genes to pursue for functional analysis. Hence bioinformatics approaches narrow down the number of potential targets and make "wet-lab" work more directed and efficient. The complete and correct prediction of ORFs and their functional annotation are also fundamental for the design and production of gene chips for gene expression studies. Gene expression studies that compare either the expression profiles of pathogens or hosts in different stages of infection can give important clues to finding new drug targets.

Transcriptome Analysis and Microarray Databases

The extraction of functional information from "one-dimensional" genome data is a major challenge of current biological and biomedical research. Functional information obtained by genome wide approaches needs to be related to genomic sequence data and the attached information. A widely applied and exceptionally powerful technique is the high-throughput expression analysis. The technology is well established and frequently employed in clinical research that is directed towards the elucidation of the molecular dynamics that underlie infection, inflammation and cancer, as well as in pharmacogenomically directed research and for diagnostic purposes.

Whole-genome and high-density setups have been implemented by means of dual-color cDNA microarrays, printed oligo arrays, and *in situ* synthesized oligo arrays (i.e. Affymetrix GeneChips®). In addition, several specialized low-density systems, mainly on the basis of cDNA or antibody arrays, gained influence for specific research topics and, especially, diagnosis. However, the underlying technologies are rapidly developing and new approaches like triple-color arrays, capillary 3D arrays or combined cDNA-antibody arrays are already emerging.

Applications in Medicine

The technology of expression analysis is applied in a broad range of medical research fields. Research topics as diverse as host-response after infection and immunology, cancer classification, chemotherapy administration, or embryogenesis and tissue differentiation are analyzed employing high-throughput expression analysis techniques [82-84]. However, microarray studies on cancer related issues are by far the most frequent and largest ones, followed by studies in inflammation and immunology.

Inflammation and Immunology

In the fight against infectious diseases, the genetic response of the human host to various microorganisms is of particular interest. Microarray studies have considerably contributed to the increasing knowledge about gene regulation in the host's defense mechanisms against pathogens [85,86] and the allergic reaction [87] as well as in macrophage activation [88]. This is mainly due to the fact that the high throughput of microarray technology allows to cluster series (e.g. with self-organizing maps) of measurements (e.g. time series, dose response series). Hence, genes with similar induction patterns can be identified and gene interaction networks can be proposed [89]. However, Ehrt *et al.* showed that tests for differential expression can be sufficient to unravel knowledge on signaling cascades if the experiment is well designed and replicated [90].

Cancer Research

Cancer type classification by expression profiling is by far the most comprehensively studied topic in the area of expression analysis [91]. Methods like LDA/QDA (linear and quadratic discriminant analysis), PLS (partial least squares), SVM (support vector machines), and ANN (artificial neural networks) have proven their capabilities in outperforming conventional histological analysis (e.g. tumor grading) [92]. Computational results have even been employed to refine histological grading techniques. However, there are still many cancer types (e.g. bronchioalveolar carcinomas) which remain hard to distinguish even when employing these techniques. The understanding of the molecular steps and the underlying molecular networks that play a role in the development of tumors and the distinction of different cancer types still remains a challenge in genomic based cancer research.

Another major topic in cancer expression analysis is the response to chemotherapy. This includes, amongst others, the question if a patient's expression signature can be used to predict the response to a chemotherapy. For bioinformatics, this represents a typical classification problem. As with the cancer type classification, many algorithms have been trained to distinguish responders and non-responders at high accuracy levels.

Besides, the increase of knowledge about gene function and regulatory networks in oncogenesis and chemotherapy response is remaining a demanding task in bioinformatics. Many studies try to discover new oncogenes and oncogenic pathways involved in specific tumorigenetic processes. Frequently used techniques are hierarchical clustering, principal component analysis (PCA), and self-organizing maps (SOM). However, with data on more tumor states and types becoming available, the dissection of the mechanisms of malignancy and metastasis will be challenging as well.

Transcriptomics Databases and Repositories

Database systems and repositories play a crucial role in transcriptomics [93] (see Table 1). Not only do they handle the vast amounts of data with ease; they also allow for user-transparency with respect to the heterogeneous vendor and hardware specific data and file formats.

Current web-based data management systems can be grouped into "database systems" and "repositories". Database systems merely support the upload and/or retrieval of expression values and annotation in a limited number of (predefined) formats (e.g. complete experiments or all data from one microarray/GeneChip). Many of them are purely public, i.e. users cannot specify private permissions for unpublished data.

Repositories, on the other hand, are web-based software suites comprising data storage, cross-linking with various other data sources (e.g. genomic, proteomic and functional databases), mining-tools (i.e. complex, user-driven dataset creation), statistics tools, and graphical interfaces. A typical repository supports data privacy via user accounts, user-defined groups and permission control. Especially collaborating laboratories frequently acknowledge these advantages compared to databases or LIMS (Laboratory Information Management Systems).

Databases and repositories are, in the long run, the only reasonable way to ensure comparability between transcriptomic data and motivate comparisons between laboratories and technologies. This will not only ensure confidentiality but also can save considerable time and money. Similar studies and sometimes even completely distinct studies can, for example, improve the classification accuracy due to an increase in the training data set's size.

A surplus of repositories is their mining capability. This gives researchers fast access to data in support or contradiction to special hypotheses. For instance, it becomes straightforward to extract all measurements of oncogenes from human gastrointestinal samples and to check for their differential expression between "normal" tissue and carcinogenic tissue (e.g. tumors).

Limitations of Array Data

However, beside the enormous advantages of high-density expression

analysis, the transcriptomics researchers are confronted with some drawbacks.

Hardly any commercial and only very few customer-based array productions include a thorough quality assessment step before shipping/hybridization. Some technologies tend to have a low reliablity of the spotted DNA sequences leading to wrong expression intensity assignments for some of the covered genes. Additionally, many experimental parameters, such as RNA quality, labeling reactions, hybridization, and scanner configuration influence the quality of the results. Since control experiments are frequently missing, algorithmic quality measures, commonly available in repositories, are essential for asserting quality standards.

The other main drawback is the heterogeneous, often very limited, amount of information about the experimental procedures and the biological samples ("biomaterial") supplied. This does not only hinder inter-laboratory comparability but also restricts the number of analytical procedures applicable (e.g. missing survival times in cancer research). Curated database systems and repositories are very useful in enforcing a minimum annotation level. Furthermore, they can engage researchers to include additional information and guide them to do so in a structured way. For example, the MIPS Expression repository requires that the user annotates the experiments and biomedical samples at a minimum level, engages MIAME (Minimum Information About Microarray Experiments) [94] compliant levels and guides further annotation by an intelligent combination of controlled and free text. Therefore, curated databases and repositories are the best means to enforce upcoming annotation standards like MIAME and assert a certain minimum quality level (i.e. implement a control for quality standards).

Requirements of Information Resources

As illustrated above genome information and the organization of data within databases became an essential component for modern biological and biomedical research. Expression analysis, linkage analysis and the study of pathogens all heavily rely on the availability of large scale data. It is fundamental for the accomplishment and interpretation of such studies that the data generated are available in reliable quality and comprehensive annotation. The sequence data need to be stored in databases, have to be kept up to date and annotated as well as possible. There is a plethora of genome resources available on the internet, but the quality of annotation and curation are multifaceted. One major problem are inconsistencies when genome data are stored in several independent databases that are maintained at different levels of accuracy. Entries can be redundant and functional annotations vary. As illustrated above genome information and the organization of data within databases became an essential component for modern biological and biomedical research. So far the main focus has been directed towards the analysis of individual genomes and the comprehensive detection and study of the individual genetic elements.

Database Maintenance: How to Stay Up to Date

Inherent to genome data is that it solely represents a snapshot of what is known at a particular point in time. Although the causes are multifaceted, it is apparent that static data repositories are rapidly becoming obsolete. For example, assigned homologies can become outdated over night due to a newly characterized gene, or gene structure assignments need to undergo revision due to novel cDNA or EST data. In order to keep data resources up to date all kinds of information constantly need to be integrated. However, the amount of information is immense and the data are very heterogeneous (sequence data, expression data, mass spectometry (MS) data, printed publications, etc.). A widely followed route for the integration of annotations from different sources is the use of ontologies and controlled vocabulary like in the Functional Catalogue (FunCat, [95]) and Gene Ontology (GO, [96]). By using controlled vocabulary, genes can be assigned to and retrieved from pathways and functions. With the exponential increase in genomic data it has become more and more evident that traditional ways of curation and functional assignment, e.g. manual curation by expert groups, reach their limits. Moreover consistency problems, typically caused by differing biological interpretations, became increasingly evident. One way to tackle this problem, are concerted, ontology approaches that are based on model organisms and expert curated assignments [96]. An alternative or complementary approach to transfer biological knowledge derived from model organisms to new molecular data are automatic processes. These approaches use sequence and domain characteristics and their association with functional annotation from reference organisms to project the information on novel genomes. Although there are evident problems caused by difficulties in orthology assignment, pseudogenes, paralogs and tissue and development dependant function, these approaches have the potential to reach a high degree of selectivity and thus are a cost-effective way for first pass functional annotation. However, as biological knowledge on particular genes is not static but rapidly evolving, the problem arises of how to stay up to date with this knowledge and new publications. As ever repeating cycles of manual updates are unrealistic, auto-

mated approaches including automated literature mining are promising strategies. Thus, for the maintenance and further development of genome databases and their content, automatic tools gain more and more importance.

Database Access

One of the major endeavors is the consistent integration of heterogeneous information into resources that enable easy and intuitive access. Concepts and standards to achieve this are the focus of current research and development, and retrieval systems such as the Sequence Retrieval System (SRS) [97] already enable to browse numerous pre-selected databases with heterogeneous content and formats.

Scattered distribution, heterogeneous design and object models are characteristic to genome and functional genomic databases. Hence linkage and integration of different sources is difficult to realize. Whereas for DNA and protein sequences unambiguous linking (usually) is easy to achieve, the vocabulary used, gene and locus identification numbers, references to regulatory regions, etc. need to be tightly controlled to enable automatic and dynamic integration. In the public domain, distributed annotation systems and BioMOBY [98] object definition standards are currently being developed and will find broad application in the near future. This will help to overcome current limitations and inconveniences, such as repeated navigation between different web sites and databases and the collection of information fragments at each individual site that later need to be assembled in order to get an overall picture.

Combinatorial Genomics: Linking All Information

Only for a small percentage of genes encoded in the genomes detailed experimental data on the biochemical and functional role exist. Thus the challenge of postgenomic analyzes is to elucidate the functional role of huge amounts of the respective genomes. Besides the administration and structuring of already existing data, a role of bioinformatics is to analyze unknown genes by sensitive tools and thereby generate hypotheses for their function and role. Additionally, postgenomic functional genomics is building on genome scale assays and analyses (transcriptomics, metabolomics, proteomics...) and aims to enrich knowledge on functional and systemic properties of the respective organisms. Clearly the future challenge is to enable the connectivity of multiple data types to explore various combinatorial relationships and dependencies among different data types and experiments and thus to fully exploit combinatorial genomic opportunities. This demanding task can only be fulfilled with the help of excellent, curated databases.

Table 1. Selected Collection of curated and regularly updated Molecular Biological Databases.

Major DNA and Protein sequence respositories:

Database	URL	Knowledge	Refs.
DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp	All known nucleotide and protein sequences	[99]
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl.html	All known nucleotide and protein sequences	[97]
GeneBank	http://www.ncbi.nlm.nih.gov/	All known nucleotide and protein sequences	[100]
InterPro	http://www.ebi.ac.uk/interpro	Protein families and domains	[101]
Pfam	http://www.sanger.ac.uk/Software/Pfam	Multiple sequence alignments/HMM models of common protein domains	[102]
PROSITE	http://www.expasy.org/prosite	Biologically-significant protein patterns and profiles	[101]
Protein Information Resource	http://pir.georgetown.edu	Comprehensive, annotated, non-redundant protein sequence database	[103]
SWISS-PROT/TrEMBL	http://www.expasy.ch/sprot	Curated protein sequences	[104]
Unigene	http://www.ncbi.nlm.nih.gov/UniGene/	Non-redundant, gene orientated clusters	[37]

Major Eukaryotic Genome Databases:

MIPS	http://mips.gsf.de	Protein and genomic sequences	[106]
Mouse Genome Database	http://www.informatics.jax.org	Mouse genomics, alleles and phenotypes	[107]
PEDANT Genome Database	http://pedant.gsf.de	Automated analysis of genomic sequences	[95]
Rat Genome Database	http://www.rgd.mcw.edu	Rat genomic database	[108]
WormBase	http://www.wormbase.org	Genomic data on nematodes	[109]

Ressources for Pathogenic Genome, Protein and Annotation Information:

Advanced Center for Genome Technology	http://www.genome.ou.edu/	Microbial genome and annotation information	n.a.
AnoBase	http://www.anobase.org/AnoBase/index.html	Genome and information of *Amopheles gambiae*	n.a.
The Comprehensive Microbial Resources	http://www.tigr.org/tdb/ http://www.sanger.ac.uk/Projects/Microbes	Microbial sequence and annotation database	[110]
HIV Sequence Database	http://hiv-web.lanl.gov/content/hiv-db/mainpage.html	Information about HIV biology	n.a.
HOBACGEN	http://pbil.univ-lyon1.fr/databases/hobacgen.html	Database of all protein sequences of bacteria organized into families	[111]
PlasmoDB	http://plasmodb.org/	Genome and information of *Plasmodium falciparum*	[112]
Tuberculist	http://genolist.pasteur.fr/TubercuList/	Genome and information of *Mycobacterium tuberculosis*	n.a.
E.coli Genome Project	http://www.genome.wisc.edu/	Microbial genome and annotation information	n.a.

Essential Comparative Genomic Ressources and Visualisation Tools:

Clusters of Orthologous Groups	http://www.ncbi.nlm.nih.gov/COG	Phylogenetic classification of protein	[113]
CORG	http://corg.molgen.mpg.de	Conserved non-coding sequence blocks	[114]
PipMaker	http://bio.cse.psu.edu/pipmaker http://bio.cse.psu.edu/genome/hummus/	Alignment and visualisation of similar regions in DNA sequences	[40]
VISTA Genome Browser	http://pipeline.lbl.gov/	Comparison and visualisation of human and mouse genomes	[39]

Pharmacogenetics, Alternative Splicing and SNP Ressources:

ASAP	http://www.bioinformatics.ucla.edu/ASAP/	Alternative spliced isoforms	[115]
ASD	http://www.ebi.ac.uk/asd/	Three databases concerning alternative splicing mechanisms	n.a.
ASDB	http://hazelton.lbl.gov/~teplitski/alt/	Database of alternatively spliced genes	[116]
dbSNP	http://www.ncbi.nih.gov/SNP/	Search, batch and information on SNPS of several organisms	[37]
EASED	http://www.bioinf.mdc-berlin.de/splice/db/	Database for alternative splice forms of nine organisms	[117]
HASDB	http://www.bioinformatics.ucla.edu/~splice/HASDB/	Human alternative splicing database	[118]
HapMap	http://hapmap.cshl.org/	Datawarehouse for haplotypes, SNPs and allele frequencies	n.a.
HGVbase	http://hgvbase.cgb.ki.se/	Collection of all human genome sequence variations	[119]
jSNP	http://snp.ims.u-tokyo.ac.jp/	Database of Japanese SNPs	[120]
PALS db	http://palsdb.ym.edu.tw/	Putative alternative splicing database	n.a.
PharmGKP	http://www.pharmgkb.org	Relationship of variations in human genes and response to drugs	n.a.
ProSplicer	http://bioinfo.csie.ncu.edu.tw/ProSplicer/	Alternative splicing database	[121]
SpliceDB	http://www.softberry.com/spldb/SpliceDB.html	Canonical and non-canonical mammalian splice sites	[122]
SpliceNest	http://splicenest.molgen.mpg.de	Visualizing splicing of genes from EST data sets	[123]
SNP Consortium Database	http://www.snp.cshl.org	SNP Consortium data	[124]

Selected Gene Expression and Profiling Databases:

ArrayExpress	http://www.ebi.ac.uk/arrayexpress	Public collection of microarray gene expression data	[125]
GEO	http://www.ncbi.nlm.nih.gov/geo/	Gene expression and hybridization array data respository	[126]
maxdSQL	http://www.bioinf.man.ac.uk/microarray/maxd/	Data warehouse and visualisation environment for expression data	n.a.
MIPS Expression repository	http://mips.gsf.de/proj/mouseExpress/ME.html	Repository with focus on mammalian expression data	n.a.
READ	http://read.gsc.riken.go.jp/	RIKEN expression array database	[127]
Stanford Microarray Database	http://genome-www4.stanford.edu/MicroArray/SMD	Raw and normalized data from microarray experiments	[128]

Major Biochemical/Metabolic Pathways Databases:

EcoCyc	http://ecocyc.org	*E. coli* K-12 genome, metabolic pathways and gene regulation	[129]
Kyoto Encyclopedia of Genes and Genomes	http://www.genome.ad.jp/kegg	Metabolic and regulatory pathways	[130]
MetaCyc	http://www.metacyc.org/	Database for genes and biochemical pathways of over 150 different organisms	[129]
RegulonDB	http://kinich.cifn.unam.mx:8850/db/regulondb_intro.frameset	Database devoted to microbial regulation entities, operons and regulons	[131]

Varied Biomedical Content (Literature, Genome Browser, Controlled Vocabulary):

Ensembl	http://www.ensembl.org/	Annotated information on eukaryotic genomes	[38]
GeneCards	http://bioinfo.weizmann.ac.il/cards/	Integrated database of human genes, maps, proteins, diseases	[132]
Gene Ontology Consortium	http://geneontology.org/ http://www.ebi.ac.uk/GOA/	Established dynamic controlled vocabulary for annotation applied to a set of proteins	[96]
MIPS FunCat	http://mips.gsf.de	Hierarchical functional catalogue to address biological function(s)	[95]
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/	Genomic information by chromosomal location	[37]
PubMed	http://www.ncbi.nlm.nih.gov/entrez/	Access to MEDLINE citations and life science journals	[37]
Locus Link/RefSeq	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html	Reference sequence standards for genomes, genes, transcripts and proteins	[133]
UCSC Genome Browser	http://genome.ucsc.edu/	Genome assemblies and annotation	[36]

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 1991;252:1651-6.
2. Lee Y, Sultana R, Pertea G , Cho J, Karamycheva S, Tsai J, et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res 2002;12:493-502.
3. Flores-Morales A, Stahlberg N, Tollet-Egnell P, Lundeberg J , Malek RL, Quackenbush J, et al. Microarray analysis of the in vivo effects of hypophysectomy and growth hormone treatment on gene expression in the rat. Endocrinology 2001;142:3163-76.
4. Rudd S. Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci. 2003;8:321-9.
5. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995; 269:496-512.
6. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science 1996;274:546, 563-67
7. The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 1998;282:2012-8.
8. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science 2000;287:2185-95.
9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.

10. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304-51.
11. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002; 420:520-62.
12. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science 2003; 300:286-90.
13. Green ED. Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2001; 2:573-83.
14. Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, et al. The yeast genome directory. Nature 1997;387 Suppl:5-6.
15. Waterston RH, Lander ES, Sulston JE. On the sequencing of the human genome. Proc Natl Acad Sci U.S.A. 2002;99:3712-6.
16. Waterston RH, Lander ES, Sulston JE. More on the sequencing of the human genome. Proc Natl Acad Sci U.S.A. 2003;100:3022-4.
17. Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2001;2:919-29.
18. Reik W, Walter J. Genomic imprinting: parental influence on the genome. Nat Rev Genet 2001;2:21-32.
19. Mattick JS, Gagen MJ. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol 2002;18:1611-30.
20. Nomura N, Miyajima N, Sazuka T, Tanaka A, Kawarabayasi Y, Sato S, et al. Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1 (supplement). DNA Res 1994;1:47-56.
21. Strausberg RL, Feingold EA, Klausner RD, Collins FS. The mammalian gene collection. Science 1999;286:455-7.
22. Wiemann S, Weil B, Wellenreuther R, Gassenhuber J, Glassl S, Ansorge W, et al. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. Genome Res 2001;11:422-35.
23. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 2002;420:563-73.
24. Cyranoski D. Geneticists lay foundations for human transcriptome database. Nature 2002;419:3-4.
25. Quackenbush J. The power of public access: the human genome project and the scientific process. Nat Genet 2001;29:4-6.
26. Katsanis N, Worley KC, Lupski JR. An evaluation of the draft human genome sequence. Nat Genet 2001;29:88-91.
27. Olson MV, Varki A. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet 2003;4:20-8.
28. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 2002;3:698-709.
29. Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 2002;30:4103-17.
30. Reese MG, Kulp D, Tammana H, Haussler D. Genie—gene finding in Drosophila melanogaster. Genome Res 2000;10:529-38.
31. Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences. Genome Res 2000;10:1631-42.
32. Wolfsberg TG, Landsman D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res 1997;25,1626-32.
33. Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 2003;4:251-62.
34. Osada N, Hida M, Kusuda J, Tanuma R, Iseki K, Hirata M. et al. Assignment of 118 novel cDNAs of cynomolgus monkey brain to human chromosomes. Gene 2001; 275:31-7.
35. Pennacchio LA, Rubin EM. Comparative genomic tools and databases: providing insights into the human genome. J Clin Invest 2003; 111:1099-106.
36. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. Nucleic Acids Res 2003;31:51-4.
37. Wheeler DL, Church DM, Federhen S , Lash AE, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res 2003;31:28-33.
38. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y. et al. Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res 2003; 31:38-42.
39. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, et al. VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 2000;16:1046-7.
40. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, et al. PipMaker - a web server for aligning two genomic DNA sequences. Genome Res 2000;10:577-86.
41. Ma MK, Woo MH, McLeod HL. Genetic basis of drug metabolism. Am J Health Syst Pharm 2002; 59:2061-9.
42. Pirmohamed M, Park BK. Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci 2001;22:298-305.
43. Halapi E, Hakonarson H. Advances in the development of genetic markers for the diagnosis of disease and drug response. Expert Rev Mol Diagn 2002;2:411-21.
44. Immervoll T, Wjst, M. Current status of the Asthma and Allergy Database. Nucleic Acids Res 1999;27:213-4.
45. Melton L. Pharmacogenetics and genotyping: on the trail of SNPs. Nature 2003; 422:917-923.
46: Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 2001;273:1516-7.
47. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 2003;33:518-21.
48. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. Trends Genet 2003;19:135-40.
49. Couzin J. Human genome. HapMap launched with pledges of $100 million. Science 2002;298:941-2.
50. Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, et al. Snipping polymorphisms from large EST collection in barley (Hordeum vulgare L.). Journal of Molecular Genetics and Genomics. In press 2003.
51. Stefansson SE, Jonsson H, Ingvarsson T, Manolescu I, Jonsson HH, Olafsdottir G, et al. Genomewide scan for hand osteoarthritis: a novel mutation in matrilin-3. Am J Hum Genet 2003;72:1448-59.
52. Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet 2002; 30:13-9.
53. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, et al. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 2003;13:1290-300.
54. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science 2002;297:1007-13.
55. Yeakley JM, Fan JB, Doucet D, Luo L, Wickham E, Ye Z, et al. Profiling alternative splicing on fiber-optic arrays. Nat Biotechnol 2002;20:353-8.
56. Claes K, Poppe B, Machackova E, Coene I, Foretova L, De Paepe A, et al. Differentiating pathogenic mutations from polymorphic alterations in the splice sites

of BRCA1 and BRCA2. Genes Chromosomes Cancer 2003;37:314-20.
57. Tsunoda T, Inada H, Kalembeyi I, Imanaka-Yoshida K, Sakakibara M, Okada R, et al. Involvement of large tenascin-C splice variants in breast cancer progression. Am J Pathol 2003;162:1857-67.
58. Smith TL, Pearson ML, Wilcox KR, Cruz C, Lancaster MV, Robinson-Dunn B, et al. Emergence of vancomycin resistance in Staphylococcus aureus. Glycopeptide-Intermediate Staphylococcus aureus Working Group. N Engl J Med 1999;340:493-501.
59. Tomasini ML, Zanussi S, Sozzi M, Tedeschi R, Basaglia G, De Paoli P. Heterogeneity of cag genotypes in Helicobacter pylori isolates from human biopsy specimens. J Clin Microbiol 2003;41:976-80.
60. Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet 2002;32:569-77.
61. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000;287: 1816-20.
62. Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, Read TD, et al. Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. Science 2003;299:2071-4.
63. Buchrieser C, Rusniok C, Kunst F, Cossart P, Glaser P. Comparison of the genome sequences of Listeria monocytogenes and Listeria innocua: clues for evolution and pathogenicity. FEMS Immunol Med Microbiol 2003;35:207-13.
64. Perna NT, Plunkett G, III, Burland V, Mau B, Glasner, JD, Rose DJ, et al. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 2001;409:529-33.
65. Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B , Kapatral V,et al. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature 2003;423:87-91.
66. Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, et al. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. Nature 2003;423:81-6.
67. Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, et al. Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. Science 2002;296:2028-33.
68. Cole ST, Barrell BG. Analysis of the genome of Mycobacterium tuberculosis H37Rv. Novartis Found Symp 1998; 217:160-72.
69. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, et al. Massive gene decay in the leprosy bacillus. Nature 2001;409:1007-11.
70. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, et al. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J Bacteriol 2002; 184:5479-90.
71. Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, et al. The complete genome sequence of Myco - bacterium bovis. Proc Natl Acad Sci U.S.A. 2003;100:7877-82.
72. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature 1999;397: 176-80.
73. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. Proc Natl Acad Sci U.S.A. 2000;97:14668-73.
74. Balfe P, Simmonds P, Ludlam CA, Bishop JO, Brown AJ. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J Virol 1990;64:6221-33.
75. Yoshimura FK, DiemK, Learn GH, Jr., Riddell S, Corey L. Intrapatient sequence variation of the gag gene of human immunodeficiency virus type 1 plasma virions. J Virol 1996;70:8879-87.
76. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002;298:129-49.
77. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 2002;419:498-511.
78. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, et al. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 2002;419:512-9.
79. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, et al. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science 2002;298:149-59.
80. Ranson H, Claudianos C, Ortelli F, Abgrall C, Hemingway J, Sharakhova MV, et al. Evolution of supergene families associated with insecticide resistance. Science 2002;298:179-81.
81. Wu CH, Yamaguchi Y, Benjamin LR, Horvat-Gordon M, Washinsky J, Enerly E, et al. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. Genes Dev 2003;17: 1402-14.
82. Gu CC, Rao DC, Stormo G, Hicks C, Province MA. Role of gene expression microarray analysis in finding complex disease genes. Genet Epidemiol 2002;23:37-56.
83. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 2002;32 Suppl:502-8.
84. Valafar F. Pattern recognition techniques in microarray data analysis: a survey. Ann.N.Y.Acad.Sci 2002;980:41-64.
85. Glynne RJ, Watson SR. The immune system and gene expression microarrays—new answers to old questions. J Pathol 2001;195:20-30.
86. Granucci F, Vizzardelli C, Virzi E, Rescigno M, Ricciardi-Castagnoli P. Transcriptional reprogramming of dendritic cells by differentiation stimuli. Eur J Immunol 2001;31:2539-46.
87. Schmidt-Weber CB, Wohlfahrt JG, Blaser K. DNA arrays in allergy and immunology. Int Arch Allergy Immunol 2001;126:1-10.
88. Nau GJ, Richmond JF, Schlesinger A, Jennings EG, Lander ES, Young RA. Human macrophage activation programs induced by bacterial pathogens. Proc Natl Acad Sci U.S.A. 2002;99:1503-8.
89. Zhang Y, Luxon BA, Casola A, Garofalo RP, Jamaluddin M, Brasier AR. Expression of respiratory syncytial virus-induced chemokine gene networks in lower airway epithelial cells revealed by cDNA microarrays. J Virol 2001;75:9044-58.
90. Ehrt S, Schnappinger D, Bekiranov S, Drenkow J, Shi S, Gingeras TR, et al. Reprogramming of the macrophage transcriptome in response to interferon-gamma and Mycobacterium tuberculosis: signaling roles of nitric oxide synthase-2 and phagocyte oxidase. J Exp Med 2001;194:1123-40.
91. 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.
92. Ringner M, Peterson C, Khan J. Analyzing array data using supervised methods. Pharmacogenomics 2002;3:403-15.
93. Fayyad UP-SGS. From Data Mining to Knowledge Discovery in Databases. AI Magazine Fall 1996;37-54.
94. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet

2001;29:365-71.
95. Frishman D, Mokrejs M , Kosykh D, Kastenmuller G, Kolesov G, Zubrzycki I, et al. The PEDANT genome database. Nucleic Acids Res 2003;31,207-11.
96. Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res 2001;11:1425-33.
97. Stoesser G, Baker W, van den BA, Garcia-Pastor M, Kanz C, Kulikova T, et al. The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Res 2003;31:17-22.
98. Wilkinson MD, Links M. (2002) BioMOBY: an open source biological web services proposal. Brief Bioinform 2002;3:331-41.
99. Miyazaki S, Sugawara H, Gojobori T, Tateno Y. DNA Data Bank of Japan (DDBJ) in XML. Nucleic Acids Res 2003;31:13-6.
100. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2002; 30:17-20.
101. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, et al. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003;31:315-8.
102. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res 2003;31:371-3.
103. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J , Chen Y, et al. The Protein Information Resource. Nucleic Acids Res 2003;31:345-7.
104. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365-70.
105. FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res 2003;31:172-5.
106. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Res 2002;30:31-4.
107. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. (2003) MGD: the Mouse Genome Database. Nucleic Acids Res 2003;31:193-5.
108. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, et al. Rat Genome Database (RGD): mapping disease onto the genome. Nucleic Acids Res 2002;30:125-8.
109. Harris TW, Lee R, Schwarz E, Bradnam K, Lawson D, Chen W, et al. WormBase: a cross-species database for comparative genomics. Nucleic Acids Res 2003; 31:133-7.
110. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The Comprehensive Microbial Resource. Nucleic Acids Res 2001;29:123-5.
111. Perriere G, Duret L, Gouy M. HOBACGEN: database system for comparative genomics in bacteria. Genome Res 2000;10:379-85.
112. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, et al. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. Nucleic Acids Res 2003;31:212-5.
113. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22-8.
114. Dieterich C, Wang H , Rateitschak K, Luz H, Vingron M. CORG: a database for COmparative Regulatory Genomics. Nucleic Acids Res 2003;31:55-7.
115. Glasner JD, Liss P, Plunkett G, III, Darling A, Prasad T, Rusch M, et al. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res 2003;31:147-51.
116. Dralyuk I, Brudno M, Gelfand MS, Zorn M, Dubchak I. ASDB: database of alternatively spliced genes. Nucleic Acids Res 2000;28:296-7.
117. Pospisil H, Herrmann A, Pankow H, Reich JG. A database on alternative splice forms on the Integrated Genetic Map Service (IGMS). In Silico.Biol 2002;3:20.
118. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 2001;29:2850-9.
119. Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 2002;30:387-91.
120. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 2002;30:158-62.
121. Huang HD, Horng JT, Lee CC, Liu BJ. ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. Genome Biol 2003;4:R29.
122. Burset M, Seledtsov IA, Solovyev VV. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res 2001;29:255-9.
123. Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. Nucleic Acids Res., 30, 299-300.
124. Thorisson GA, Stein LD. The SNP Consortium website: past, present and future. Nucleic Acids Res 2003;31:124-7.
125. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31:68-71.
126. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002; 30:207-10.
127. Bono H, Kasukawa T, HayashizakiY, Okazaki Y. READ: RIKEN Expression Array Database. Nucleic Acids Res 2002;30:211-3.
128. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, et al. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res 2003;31:94-6.
129. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. The EcoCyc and MetaCyc databases. Nucleic Acids Res 2000;28:56-9.
130. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30:42-6.
131. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, et al. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. Nucleic Acids Res 2001;29:72-4.
132. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, et al. GeneCards 2002: towards a complete, object-oriented, human gene compendium. Bioinformatics 2002;18:1542-3.
133. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence project: update and current status. Nucleic Acids Res 2003;31:34-7.

Address of the authors:
Claudia C. Englbrecht, Michael Han,
Michael T. Mader, Andreas Osanger,
Klaus F. X. Mayer*
MIPS, Institute for Bioinformatics
GSF - National Research Center for
Environment and Health
85758 Neuherberg, Germany
E-mail: kmayer@gsf.de

*Corresponding author