Review
EASL JOURNAL OF HEPATOLOGY

Bioinformatics and database resources in hepatology

Andreas Teufel*

Department of Medicine I, University Medical Center, Regensburg, Germany

Summary

Lately, advances in high-throughput technologies in biomedical research have led to a dramatic increase in the accessibility of molecular insights at multiple biological levels in hepatology. Much of this information is available in publications, but an increasing number of large-scale analyses are currently being stored in databases. Scopes of these databases are very divergent and may range from large, general databases collecting information on almost every known disease, to very specialized databases covering only a specific liver disease or aspect of hepatology. Over recent years, these bioinformatics data repositories have rapidly evolved into an essential aid for molecular hepatology. However, although publicly available through the internet, many of these databases are only known to a few experts. To facilitate access to these resources, the publicly available databases supporting research on liver diseases are summarized in this review.

Introduction

Chronic liver disease is a major health burden worldwide. In developed countries, viral hepatitis B and C [1], but also NALFD/NASH [2,3], result in a steadily increasing number of patients with liver fibrosis, cirrhosis or hepatocellular carcinoma as common end-stage [4].

Recent advances in drug development for viral hepatitis will certainly be beneficial for many patients [5]. However, for patients with liver fibrosis, cirrhosis or hepatocellular carcinoma, only very few drugs, if any, are currently available [6,7]. Thus, investigating the molecular causes of chronic liver diseases, but also their common end-stages – fibrosis, cirrhosis and HCC – is still a major clinical need [4].

Over the past two decades, molecular targets and networks have increasingly gained attention [8,9], in particular due to the efforts of the Human Genome Project [10,11] and the introduction of large scale molecular analysis platforms [12]. As a result, human and many other genomes, transcriptomes, proteomes and resources from many other biological layers are publicly available. This vast amount of molecular data provides a rich source to better understand the molecular basis of chronic liver diseases, and to identify novel genomic targets for therapeutic intervention. However, in order to utilize this information, knowledge of the available database resources and bioinformatics tools is indispensable.

Data can be deposited in general or specific databases. In general databases such as the NIH-hosted PubMed [13], OMIM [13], or GAD [14], searching specific information on chronic liver diseases may still be time-consuming and result in a non-comprehensive overview. In addition, general databases, e.g. PubMed [13], lack specific, hepatology-centered search options and thus searching liver disease specific data remains elusive. In contrast, specialized databases with a focus on specific chronic liver diseases lack comprehensiveness or validity of curated data due to automated data curation, but offer the option of very detailed and specialized data retrieval, and thus are a valuable resource in hepatology research. Therefore, the currently available databases supporting research on liver diseases are summarized in this review. Furthermore, short summaries of their focus and applications in hepatology are provided.

General databases holding information on liver diseases

PubMed, PubMatrix, OMIM – mining biomedical literature

PubMed is a major resource of the available biomedical literature, maintained by the National Center for Biotechnology Information (NCBI); it currently comprises over 24 million citations for biomedical literature from MEDLINE, life science journals, and online books [13]. A useful tool to extract from PubMed organ- or disease-specific information, e.g. information on liver diseases, is Pubmatrix. Using available abstracts in PubMed, this web-based tool performs automated keyword searches for co-occurring terms such as specific gene names and diseases [15]. The Online Mendelian Inheritance in Man (OMIM) database, a compendium of human genes and genetic phenotypes, provides also genetics centered summaries of a large collection of diseases, including most chronic liver diseases [13].

Keywords: Bioinformatics; Computational biology; Database; Genetics; Proteomics; HBV; HCV; Viral hepatitis; NASH; NAFLD; HCC; Toxic liver disease.
Received 28 July 2014; received in revised form 10 October 2014; accepted 23 October 2014

* Address: University of Regensburg, Department of Internal Medicine I, Franz-Josef-Strauss Allee 11, 95053 Regensburg, Germany. Tel.: +49 941 944 7003; fax: +49 941 944 7004.
E-mail address: andreas.teufel@ukr.de.

Key Points

- Multiple databases serve as profound knowledge bases and provide significant support to molecular research in liver disease development and drug targeting
- These resources are not yet connected. However, a better connection of the available databases would certainly be desirable and holds potential prospects for further advances in understanding the molecular changes in liver disease
- Bioinformatics and data base resources for liver cancer are the best developed tools in hepatology
- Other areas in hepatology, such as viral hepatitis or liver fibrosis need significant support for the development of more advanced bioinformatics resources. The development in oncology may serve as an example of how to develop these resources
- Availability and easy access to pre-processed molecular high throughput data and the development of graphical outputs for hepatology-related data queries will further help to extend the use of publicly available data

Genetic Association Database (GAD)

The Genetic Association Database provides a collection of data on genetic association studies. Available data may be searched in various ways. With respect to liver diseases, the most relevant search option is the disease phenotypes mode. The listed, searchable diseases include many chronic liver diseases [14].

GWAS central

Genetic changes associated with liver diseases may be obtained from GWAS central, a centralized repository of genetic association studies. Searching data by means of phenotypes offers multiple options for liver disease-related searches. Top ranked genetic changes from each displayed analysis can be downloaded for further bioinformatics or wet-lab analysis [16].

Gene Expression Omnibus (GEO) profiles, Oncomine, ArrayExpress – mining gene expression microarray data

Gene expression microarray data are available from multiple public and commercial databases. The most widely used array repositories are Gene Expression Omnibus (GEO) Profiles, Oncomine, and ArrayExpress. The GEO Profiles database provides graphic gene expression profiles derived from microarray experiments stored at NCBI's GEO microarray resource [17]. These profiles are displayed as bar charts and searchable by means of full text search. Therefore, by entering liver diseases as search terms, the database returns graphic results of differential gene expression in multiple gene expression experiments.

The commercial microarray data repository Oncomine holds gene expression profiles of thousands of cancer patient genomes. The database offers free differential expression searches for single genes across different experiments, displaying expression levels in all samples within a given dataset. Searching multiple genes and gene expression profiles is limited to the commercial license of the database. With respect to hepatology, the data repository contains multiple microarray experiments on hepatocellular and cholangio carcinoma.

The European microarray resource Array Express and its graphical phenotype-based expression profiler Array Atlas, currently does not contain a considerable amount of data on liver diseases that can be extracted easily. Thus, the large collection on microarray data with a focus on liver diseases, is for the most part accessible only to researchers with advanced bioinformatics knowledge [18].

Tissue-specific Gene Expression and Regulation (TiGER)

The TiGER database contains tissue-specific gene expression profiles or expressed sequence tag (EST), cis-regulatory module (CRM), and combinatorial gene regulation data. Among the multiple search options, the database may be searched for genes preferentially expressed in liver, CRM detections in liver, or pairs of transcription factors co-regulated in liver tissue [19].

RNA-Seq Atlas

RNA-Seq Atlas is a web-based repository of RNA-Seq gene expression profiles and query tools. The database may be used to compare tissues and find genes with specific expression patterns. It offers liver specific gene search options. All data are linked to common functional and genetic databases, offering in particular information on genes, signaling pathway analysis and evaluation of biological functions by means of gene ontologies. Additionally, data are linked to several microarray gene profiles, including BioGPS normal tissue profiles and NCI60 cancer cell line expression data [20].

General liver physiology

The Liver Specific Gene Promoter Database (LSPD)

The Liver Specific Gene Promoter Database (LSPD) provides a collection of liver specific genes and their regulatory elements. Information on these regulatory elements is primarily sequence-based. It may be searched in multiple ways, including by querying for known regulatory elements, and allow retrieving promoter sequences. It also offers query options to discover novel motifs or build promoter models. Finally, the database provides insights into interactions among known transcriptions factors expressed in liver.

Liverbase

Based on a large-scale analysis effort of the Chinese human liver proteome project (CNHLPP), Liverbase contains extensive knowledge on the human liver proteome. It provides key protein information such as protein function, abundance, and subcellular localization. This information is linked to disease specific information. Through an intuitive web-interface, the database may be searched by disease, but also offers queries centered on pathway or gene ontology. In addition, the database holds information on the liver transcriptome from Chromatin Immunoprecipitation DNA Sequencing (CHIP-Seq) and Massively Parallel Signature Sequencing (MPSS) experiments [21].

Table 1. Summary of the currently available bioinformatics and database resources in hepatology.

Database	URL	Ref.
General databases holding information on liver diseases		
Pubmed	http://www.ncbi.nlm.nih.gov/pubmed	[13]
Pubmatrix	http://pubmatrix.grc.nia.nih.gov	[15]
OMIM	http://www.ncbi.nlm.nih.gov/omim	[13]
Genetic Association database (GAD)	http://geneticassociationdb.nih.gov	[14]
GWAS central	http://www.gwascentral.org	[16]
Gene Expression Omnibus (GEO) profiles	http://www.ncbi.nlm.nih.gov/geoprofiles	[17]
Oncomine	https://www.oncomine.org	
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	[18]
Tissue-specific Gene Expression and Regulation (TiGER)	http://bioinfo.wilmer.jhu.edu/tiger	[19]
RNA-Seq Atlas	http://medicalgenomics.org/rna_seq_atlas	[20]
General liver physiology		
The Liver Specific Gene Promoter Database (LSPD)	http://rulai.cshl.edu/LSPD	
Liverbase	http://liverbase.hupo.org.cn	[21]
Mouse liver protein database	http://proteome.biochem.mpg.de/liver	[23]
Liver disease specific databases		
Library of Molecular Associations (LOMA)	http://medicalgenomics.org/loma	[24]
JAX-Mice database-Liver defects	http://jaxmice.jax.org/list/ra102.html	
Toxic liver disease		
LiverTox	http://livertox.nih.gov	
Liver Toxicity Knowledge Base (LTKB)	http://www.fda.gov/ScienceResearch/BioinformaticsTools/LiverToxicityKnowledgeBase	[25]
Viral hepatitis-General		
The Hepatitis Virus Database	http://s2as02.genes.nig.ac.jp	[26]
VirusMINT	http://mint.bio.uniroma2.it/virusmint	[27]
Virus hepatitis B		
HepSEQ	http://www.hepseq.org	[28]
HBVRegDB	http://lancelot.otago.ac.nz	[29]
Oxford HBV Automated Subtyping Tool	http://www.bioafrica.net/rega-genotype/html/subtypinghbv.html	[30]
HBV STAR	http://www.vgb.ucl.ac.uk/starn.shtml	[31]
jpHMM	http://jphmm.gobics.de/submission_hbv	[32]
HBVseq	http://hivdb.stanford.edu/HBV/HBVseq/development/HBVseq.html	[33]
SeqHepB	http://www.seqhepb.com	[34]
Viral hepatitis C		
The HCV database at the Los Alamos National Laboratory	http://hcv.lanl.gov	[35,36]
Oxford HCV Subtyping Tool	http://www.bioafrica.net/rega-genotype/html/subtypinghcv.html	[30]
Virus Pathogen Database and Analysis Resource (ViPR)	http://www.viprbrc.org/brc/home.spg?decorator=flavi_hcv	[37]
euHCVdb,European HCV database	http://euhcvdb.ibcp.fr	[38]
HCVpro, Hepatitis C virus protein interaction database	http://cbrc.kaust.edu.sa/hcvpro	[39]
Hep-Druginteractions.org	http://hep-druginteractions.org	
Hepatocellular carcinoma		
OncoDB.HCC	http://oncodb.hcc.ibms.sinica.edu.tw	[40]
Encyclopedia of Hepatocellular Carcinoma Online gene (EHCO)	http://ehco.iis.sinica.edu.tw	[41]
CellMinerHCC	http://medicalgenomics.org/cellminerhcc	[42]
Integrated Clinical Omics Database (iCOD)	http://omics.tmd.ac.jp/icod_pub_eng/portal/top.do	[43]
The Cancer Genome Atlas (TCGA)	http://cancergenome.nih.gov	[44]
International Cancer Genome Consortium (ICGC)	https://icgc.org	[45]

Mouse liver protein database

The mouse liver proteome database is based on mass spectrometry analysis of mouse liver tissue. It lists proteins with at least two peptides identified in mass spectrometry. This information was combined with the previously reported organelle map [22]. Through a public web-interface, the database may easily be searched by means of protein name, peptide sequence, IPI accession, but also using BLAST to compare novel protein sequences to protein entries present the database [23].

Liver disease specific databases

Library of Molecular Associations (LoMA)

The Library of Molecular Associations (LoMA) is based on PubMed-published abstracts and aims at closing the gap between genome-wide coverage of low validity from microarray data and individual highly-validated data from PubMed. After an initial automated text mining process, the extracted abstracts were all manually validated. The database holds confirmed molecular associations for chronic liver diseases such as HCC, CCC, liver fibrosis, NASH/fatty liver disease, AIH, PBC, and PSC. The database may be searched not only by means of disease or gene names, but also using pathway or gene ontology information provided by the LoMA database [24].

JAX-Mice database – liver defects

The Jackson Laboratory, harboring one of the largest collections of gene modified mouse strains, provides a detailed overview of mouse strains related to liver specific phenotypes. For each mouse model, a short summary on the specific gene function and resulting phenotype are given.

Toxic liver disease

LiverTox

Database resources for toxic liver disease remain limited. However, the National Institute of Diabetes and Digestive and Kidney Diseases in collaboration with the National Library of Medicine and the Drug-Induced Liver Injury Network study group have established LiverTox, a comprehensive clinical information website on drug-induced liver injury. The website provides an overview of multiple drugs (its chemical nature, indications, recommended doses and regimens, and frequency of use), followed by a detailed summary of the pattern and course of the associated liver injury, and related PubMed references.

Liver Toxicity Knowledge Base (LTKB)

Since liver toxicity is the most common cause for the discontinuation of clinical trials and an approved drug's withdrawal from sales, Liver Toxicity Knowledge Base collects and summarizes mechanisms of toxicity, drug metabolism, histopathology, therapeutic use, targets, side effects, etc. This information is linked to information on individual drugs, and the use of systems biology analysis to assess and predict drug-induced liver injury. Importantly, both conventional and high-throughput molecular biomarker assays will be conducted for selected drugs to develop novel biomarkers based on knowledge acquired from the project [25].

Viral hepatitis – general

Viral mutations may accumulate during chronic infection or in response to external pressure. Resulting drug resistance or vaccine escape mutants have become a major concern for the treatment of individual patients with viral hepatitis, but also in general for public health. Therefore, bioinformatics algorithms and databases focusing on viral genetic mutation or recombination are a major aid to identify those genetic changes and to successfully target the resulting liver disease.

The Hepatitis Virus Database

The Japanese Hepatitis Virus Database holds information on diverse hepatotropic viruses. This repository mainly provides resources as a phylogenetic analysis of publicly HBV, HCV, and HEV sequences, both nucleic acid and protein sequences [26].

VirusMINT

A detailed understanding of the molecular interactions between host physiology and viral infection will certainly be a major aid in targeting viral hepatitis. The VirusMINT database provides a comprehensive summary of protein interactions between viral and human proteins reported in the literature. Among the different viral strains provided are information on viral hepatitis B and C. Virus-specific search pages offer easy query options, and results provided by the database can be displayed with a graphical viewer [27].

Viral hepatitis B

For HBV research, several genomic databases have been established to investigate and monitor the genetic variability of HBV sequences and viral resistance to treatment. These databases may aid in the development of new diagnostic reagents as well as in the monitoring of polymerase and envelope protein mutations selected under different antiviral treatments.

HepSEQ

HepSEQ is a collection of sequence, clinical and epidemiological data related to hepatitis B virus (HBV) infection. The data were collected from patients of 40 different nationalities and diverse ethnic backgrounds, and are quality-controlled. Upon registration, they may be accessed through a web front-end allowing search and manipulation of the stored data, but also extraction and visualization of information on epidemiological, virological, clinical, nucleotide sequence and mutational aspects. Additional bioinformatics tools provide further information on HBV genotype, identify mutations with known clinical significance (e.g. vaccine escape, precore and antiviral-resistant mutations), and carry out sequence homology searches against other deposited strains [28].

HBVRegDB

HBVRegDB combines multiples NCBI reference sequence annotations of the HBV genome with comparative analysis tools such as blastn, a standard sequence alignment tool. Major strength of the database is its graphic presentation of the results, which uses the generic genome browser (GBrowse) adapted for the analysis of viral genomes [29].

Oxford HBV automated subtyping tool

More recently, genotyping algorithms that implement a sliding window to generate multiple overlapping segments of a query sequence and its reference dataset, have been developed for HBV. The resulting shorter sequence pieces are individually analyzed and afterwards assembled to obtain the full sequence analysis. This procedure increases the accuracy and reliability of the results, especially when analyzing recombinant viruses. The algorithm allows analysis of up to 1000 sequences simultaneously; it is available on several webservers, among which the Bioafrica server [30].

HBV STAR

Comparable to the above-mentioned algorithm, HBV STAR performs sequence analysis and genotyping by using a sliding window of 150 bp with a step interval of one base. The underlying algorithm generates position-specific scoring matrices, which allow the analysis of subgenomic sequences as the basis of sequence comparison and the identification of potential recombinants (if sequences are divergent more than 1% from ascribed genotypes). A web interface to HBV STAR is available [31].

jpHMM

Similar to the above-described algorithms, Schultz *et al.* implemented a genotyping algorithm that may indentify and describe intersubtype recombinations using a probabilistic approach, the so-called jumping profile Hidden Markov Model (jpHMM). This algorithm models the most likely subtype at each nucleotide, thereby being able to identify diverse subtypes and recombinations in one sequence. The jpHMM algorithm is also available through a web interface [32].

HBVseq

In addition to subtype analyses of a sequence query, the HBVseq uses information from a local HBV drug resistance database (HBVrt DB) to retrieve the prevalence of each mutation according to genotype and treatment [33].

SeqHepB

Similar to HBVseq, SeqHepB is a commercial database that determines the HBV genotype and aims at identifying key viral mutations associated with antiviral resistance [34].

Viral hepatitis C

Six genotypes of the hepatitis C virus (HCV) have been established that differ in treatment response. However, on the basis of nucleotide sequences many more additional subtypes may be characterized. The high sequence variability may very well be involved in escape and resistance mutations of the virus. Over the past decade, as an accompanying effort of the development of increasingly effective drugs targeting the virus, several HCV specific databases have been established with the aim of collecting HCV sequences as well as structural and functional analyses of the virus.

The HCV database at the Los Alamos National Laboratory

The HCV database at the Los Alamos National Laboratory (USA) contains manually annotated sequence information (HCV sequence database) and information on immunological epitopes (HCV immunology database). The HCV sequence database allows diverse search options for HCV sequences, in particular for sequence length or specific regions of interest. Furthermore, user's sequence results may be incorporated and subject to sequence comparisons (e.g. for similarity, principal coordinate analysis, and others), or phylogenetic analyses. Search results may be aligned using implemented alignment tools. Noteworthy, a unique geographic tool provides pie charts of the numbers of sequences of each genotype on geographical maps [35,36].

Oxford HCV subtyping tool

Comparable to the HBV subtyping tool described above, an HCV analysis algorithm was also implemented by the same group. Again, a sliding window generates multiple overlapping segments of a query sequence and its reference dataset. These shorter pieces of sequence are individually analyzed and in a second step assembled to obtain the full sequence analysis. The algorithm is available through a public webserver [30].

Virus Pathogen Database and Analysis Resource (ViPR)

The Virus Pathogen database and Analysis Resource (ViPR) contains major information and bioinformatics tools for HCV research. It provides a large collection of sequence genetics and protein sequence information, but also data on three-dimensional protein structures, immune epitopes, or host factors. Information retrieved from the comprehensive database may then directly be subject to further analysis via advanced bioinformatics algorithms such as alignment tools, BLAST searches, phylogenetic comparisons or three-dimensional visualizations of protein structures [37].

euHCVdb, European HCV database

The European HCV database offers access to a computer-annotated set of sequences and molecular models of HCV proteins, and focuses on protein sequence, structure and function analysis. It is composed of two parts, static and dynamic part. The static part provides access to multiple pre-computed reference sequences and alignments, but also three-dimensional protein structures. Bioinformatics interfaces such as Jmol or alignment tools allow comparison and visualization of these sequence and molecules. The dynamic part of the website allows submitting specific queries using selections from a pre-defined query interface, containing multiple parameters for each specific query. Finally, three-dimensional protein models for variants not yet available in the database can be built with the Geno3D Web server [38].

HCVpro, hepatitis C virus protein interaction database

The HCVpro database provides manually verified hepatitis C virus–virus and virus–human protein interactions curated from literature and databases. The available data were extracted from different sources among them euHCVdb, HCVdb, or VirusMint, but also other protein interaction resources such as BIND and other relevant biology repositories. Besides pure protein interaction, proteins curated in HCVpro have also been extensively cross-referenced to other essential bioinformatics annotations, in particular gene ontologies, canonical pathways, or Online Mendelian Inheritance in Man (OMIM). Finally, HCVpro holds summaries on structure and functions of HCV proteins and current state of drug and vaccine development, and links references [39].

Hep-Druginteractions.org

Differently from databases summarizing molecular knowledge on disease development, Hep-Druginteractions.org summarizes recent developments on novel antiviral substances and their obvious drug interactions to multiple know substances. Hep-Druginteractions.org illustrates these interactions by providing a printable chart of the currently known drug interactions as well as an Internet-based search tool. The interactions may be searched as an alphabetical list of drugs or by drug classes. As an output, the Internet-based knowledge base returns information on drugs that should not be co-administered or may have potential interactions.

Hepatocellular carcinoma

OncoDB.HCC

OncoDB.HCC provides detailed bioinformatics resources for HCC research by integrating data from a wide variety of genomic aberration studies. These data include chromosomal aberration studies from loss of heterozygosity (LOH) and comparative genome hybridization (CGH) analysis as well as altered gene expression data from microarray and proteomic studies. Data were collected from public databases such as PubMed, journals of the NCBI literature database and the Stanford Microarray Database. The intuitive web interface offers several comprehensive search options. The genomic aberration search displays a graphic overview on human and rodent QTLs, or genetic loci associated with HCC development. In expression view, a total of 9785 genes may be searched for a potential involvement in HCC development. In addition, the HCC significant search option offers a summary of 614 genes and their association with several lines of evidence such as microarray data, RT-PCR, QPCR, IHC, Western blot, and others [40].

Encyclopedia of Hepatocellular Carcinoma Online gene (EHCO)

Encyclopedia of Hepatocellular Carcinoma Online gene (EHCO) is an integrative platform collecting, organizing and comparing systematically unsorted HCC-related studies by using natural language processing as well as individual and manual validation. Currently, two versions of EHCO are available, a 2007 version and a 2008 version. The 2008 version (EHCO II) contains 13 gene sets related to HCC from several diverse technological platforms such as PubMed, SAGE, microarray analysis as well as proteomics data. Data from these sources were extracted in an automated fashion using intelligent software robots. The 2008 version contains 4020 genes in total; however, most genes are only included once (65%), suggesting that tremendous efforts need to be exerted to further validate these automatically obtained genes and to characterize the relationship between HCC and these genes. Genes associated to HCC may be searched by multiple identifiers such as Gene Symbol, UniGene ID, Entrez Gene ID, or Ensembl ID. In addition, the website offers a full list of genes, a reference to the dataset in which a gene is found differentially expressed, and provide counts of the number of datasets where an up- or down-regulation of the specific gene is reported [41].

CellMinerHCC

CellMiner*HCC* is a publicly available database containing microarray expression profiles of diverse HCC cell lines. At present, this database holds genome wide microarray profiles of 18 HCC cell lines. The database allows the evaluation of expression profiles of individual HCC cell lines and the comparison of differential gene expression between multiple cell lines. Evaluation of cell line expression profiles and their differences is supported via a public web interface that provides easy access to microarray data [42].

Integrated Clinical Omics Database (iCOD)

The iCOD database is an Integrated Clinical Omics Database (iCOD) containing molecular omics data such as CGH (comparative genomic hybridization) and gene expression profiles, and also comprehensive clinical information such as clinical manifestations, medical images, laboratory tests, or drug histories from 140 patients with HCC. The iCOD database may therefore be a major aid in linking molecular omics data and disease pathways to clinico-pathological findings. A major strength of the database is certainly the strong link to numerous clinical data. On the downside, results from the database are hard to extract for further analyses [43].

The Cancer Genome Atlas (TCGA)

Under the lead of the National Cancer Institute and the National Human Genome Research Institute, The Cancer Genome Atlas (TCGA) database has developed into a massive and powerful data resource. It contains data on a large number of patients and for many of them data on multiple biological levels such as copy number variation, gene expression, miRNA expression, or methylation status. For most of the samples, control tissue/normal tissue data are available to compare the cancerous tissue with. The TCGA Data Portal allows to filter the available data for a cancer of interest and to select manually the data files of interest. Major advantage of the database is the availability of comprehensive data on multiple biological layers. However, to obtain and analyze the data, advanced bioinformatics skills are necessary as most data provided are unprocessed raw data from the original experiments [44].

International Cancer Genome Consortium (ICGC)

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumors from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Hepatocellular carcinoma is found among the tumors studied in this consortium. The consortium's data hub currently hosts five large-scale experiments and analyses on genomics, transcriptomics, and epigenetics of hepatocellular carcinoma. The underlying diseases vary from alcoholic liver disease to viral hepatitis based liver cirrhosis. Patients were collected in China, Japan, France, and USA. Strength of the database is its comprehensive data on multiple biological

Review

layers. However in order to obtain and analyse the data advanced bioinformatics skills are necessary as most data provided are unprocessed raw data from the original experiments [45].

Discussion

Promoted by the decoding of the human genome as part of the human genome project, bioinformatics algorithms and database resources have become a major aid for the analysis of molecular changes leading to chronic liver disease, cirrhosis and cancer. With the aim of expanding the current knowledge underlying disease mechanisms and treatment options, but also to identify and characterize biomarkers, the creation of genetic fingerprints for individualized diagnosis, prognosis and treatment of patients has shifted to the center of interest in translational hepatology. Furthermore, recently developed high-throughput analysis technologies rely essentially on sufficient bioinformatics databases and build expectations for more effective, less prone to side effects, and economically reasonable therapies. The challenges of the next few years relate to the broadening of the knowledge base, the establishment of reliable and standardized technologies, and the development of intelligent bioinformatics strategies for data analysis and data integration. In contrast to other fields in medicine, e.g. cancer biology, the availability of specialized bioinformatics resources and databases remains limited. However, the potential of powerful bioinformatics resources can nicely be illustrated with the example of hepatocellular carcinoma, which is efficiently covered by multiple efforts in the context of cancer biology [8].

If for example one wants to estimate the role of a given gene in HCC biology, multiple database resources may provide substantial information to assemble the full picture of transcriptional regulation. First LoMA [29] holds published molecular associations of liver cancer and genetic regulations. If not listed in LoMA, the gene may further be analyzed by means of Pubmatrix [15] matching the gene name against key biological terms such as "cancer", "liver carcinogenesis", or "HCC". Transcriptomic changes may then be obtained from GEO profiles using "HCC" or "liver cancer" as search terms; GEO profiles would provide graphical illustrations of differential regulations found in microarray experiments stored in the GEO microarray database [17]. Finally, differential gene expression may be linked to clinical course of disease using the iCOD database [43]. Entering a gene name, the database provides information on gene expression from 60 patients and links the gene expression to clinical parameters such as tumor size, TNM stage, AFP level, survival and others. A Student's *t*-test is performed to estimate the statistical relevance. The most comprehensive data are provided by the ICGC [44] and TCGA [45] projects. These projects offer clinical data combined with comprehensive molecular data, not only on gene expression, but also genetic and epigenetic information. If capable of performing the bioinformatics analysis, these resources offer to assemble a complete picture of the influence of genomics, transcriptomics or epigenetics changes for the development and clinical course of hepatocellular carcinoma.

Having illustrated the *in silico* options in liver cancer biology, the shortcomings in other aspects of chronic liver disease become obvious. For example, searching for a gene's role in viral hepatitis biology remains difficult, if not impossible. The currently available bioinformatics resources focus on identifying viral subtypes or genetic changes within the viral genome, but offer no insights on the role of particular genes in host biology. Similarly, the options in investigating the role of a gene in liver fibrosis remain limited. Besides few microarray experiments available from the GEO [17] or ArrayExpress [18] databases and mostly performed in mice, there is currently no option to estimate a gene function in fibrosis development.

Overall, multiple database resources have been established over the past years (Table 1). However, their development is far behind other areas of clinical medicine such as oncology. Further development of bioinformatics and database resources in hepatology will certainly be essential to progress further in unraveling the molecular changes in liver disease.

In order to develop bioinformatics and database resources in hepatology, the recent developments in oncology may serve as an example. The first step to take will be the establishment of a comprehensive knowledgebase unifying the multiple available bioinformatics tools. Next, the integration of available microarray data in a hepatology specific repository and the graphical illustration of the data, e.g. for differential gene expression, will certainly help make these data available to the hepatologic community without requiring advanced bioinformatics knowledge. Finally, large-scale and multi-institutional efforts will be needed to collect samples and data for comprehensive databases in fields of hepatology other than liver cancer, in particular liver fibrosis and viral hepatitis.

Conclusions

Multiple databases offer a wide range of specific and mostly comprehensive resources for hepatology research. They may serve as profound knowledge bases, but also provide significant support to molecular research in liver disease development and drug targeting. However, it would a certainly be desirable to further develop the available databases and to better connect them, since this may hold potential prospects for further advances in understanding the molecular changes in liver disease.

Conflict of interest

The author declared that he does not have anything to disclose regarding funding or conflict of interest with respect to this manuscript.

References

[1] Hahne SJ, Veldhuijzen IK, Wiessing L, Lim TA, Salminen M, Laar M. Infection with hepatitis B and C virus in Europe: a systematic review of prevalence and cost-effectiveness of screening. BMC Infect Dis 2013;13:181.

[2] Vernon G, Baranova A, Younossi ZM. Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Aliment Pharmacol Ther 2011;34:274–285.

[3] Armstrong MJ, Houlihan DD, Bentham L, Shaw JC, Cramb R, Olliff S, et al. Presence and severity of non-alcoholic fatty liver disease in a large prospective primary care cohort. J Hepatol 2012;56:234–240.

[4] Liou IW. Management of end-stage liver disease. Med Clin North Am 2014;98:119–152.

[5] Manns MP, von Hahn T. Novel therapies for hepatitis C – one pill fits all? Nat Rev Drug Discov 2013;12:595–610.

[6] Schuppan D, Kim YO. Evolving therapies for liver fibrosis. J Clin Invest 2013;123:1887–1901.

[7] Worns MA, Galle PR. HCC therapies – lessons learned. Nat Rev Gastroenterol Hepatol 2014;11:447–452.

[8] Marquardt JU, Galle PR, Teufel A. Molecular diagnosis and therapy of hepatocellular carcinoma (HCC): an emerging field for advanced technologies. J Hepatol 2012;56:267–275.

[9] Teufel A, Marquardt JU, Dooley S, Lammert F, Galle PR. Personalised hepatology – current concepts, developments and expectations in the post-genome era. Z Gastroenterol 2012;50:41–46.

[10] Collins FS, McKusick VA. Implications of the Human Genome Project for medical science. JAMA 2001;285:540–544.

[11] Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. Transl Res 2009;154:277–287.

[12] Merrick BA, London RE, Bushel PR, Grissom SF, Paules RS. Platforms for biomarker analysis using high-throughput approaches in genomics, transcriptomics, proteomics, metabolomics, and bioinformatics. IARC Sci Publ 2011:121–142.

[13] Database resources of the National Center for Biotechnology. Nucleic Acids Res 2014;42:D7–D17.

[14] Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet 2004;36:431–432.

[15] Becker KG, Hosack DA, Dennis Jr G, Lempicki RA, Bright TJ, Cheadle C, et al. PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics 2003;4:61.

[16] Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur J Hum Genet 2014;22:949–952.

[17] Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 2006;411: 352–369.

[18] Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, et al. ArrayExpress update – trends in database growth and links to data analysis tools. Nucleic Acids Res 2013;41:D987–D990.

[19] Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics 2008;9:271.

[20] Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. RNA-Seq Atlas – a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics 2012;28:1184–1185.

[21] Sun A, Jiang Y, Wang X, Liu Q, Zhong F, He Q, et al. Liverbase: a comprehensive view of human liver biology. J Proteome Res 2010;9: 50–58.

[22] Foster LJ, de Hoog CL, Zhang Y, Xie X, Mootha VK, Mann M. A mammalian organelle map by protein correlation profiling. Cell 2006;125:187–199.

[23] Shi R, Kumar C, Zougman A, Zhang Y, Podtelejnikov A, Cox J, et al. Analysis of the mouse liver proteome using advanced mass spectrometry. J Proteome Res 2007;6:2963–2972.

[24] Buchkremer S, Hendel J, Krupp M, Weinmann A, Schlamp K, Maass T, et al. Library of molecular associations: curating the complex molecular basis of liver diseases. BMC Genomics 2010;11:189.

[25] Chen M, Zhang J, Wang Y, Liu Z, Kelly R, Zhou G, et al. The liver toxicity knowledge base: a systems approach to a complex end point. Clin Pharmacol Ther 2013;93:409–412.

[26] Shin IT, Tanaka Y, Tateno Y, Mizokami M. Development and public release of a comprehensive hepatitis virus database. Hepatol Res 2008;38:234–243.

[27] Chatr-aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S, Sacco F, et al. VirusMINT: a viral protein interaction database. Nucleic Acids Res 2009;37: D669–D673.

[28] Gnaneshan S, Ijaz S, Moran J, Ramsay M, Green J. HepSEQ: International Public Health Repository for Hepatitis B. Nucleic Acids Res 2007;35: D367–D370.

[29] Panjaworayan N, Roessner SK, Firth AE, Brown CM. HBVRegDB: annotation, comparison, detection and visualization of regulatory elements in hepatitis B virus sequences. Virol J 2007;4:136.

[30] Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res 2009;37:W634–W642.

[31] Myers R, Clark C, Khan A, Kellam P, Tedder R. Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. J Gen Virol 2006;87:1459–1464.

[32] Schultz AK, Bulla I, Abdou-Chekaraou M, Gordien E, Morgenstern B, Zoaulim F, et al. jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. Nucleic Acids Res 2012;40:W193–W198.

[33] Rhee SY, Margeridon-Thermet S, Nguyen MH, Liu TF, Kagan RM, Beggel B, et al. Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. Antiviral Res 2010;88:269–275.

[34] Yuen LK, Ayres A, Littlejohn M, Colledge D, Edgely A, Maskill WJ, et al. SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B. Antiviral Res 2007;75:64–74.

[35] Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos hepatitis C sequence database. Bioinformatics 2005;21:379–384.

[36] Yusim K, Richardson R, Tao N, Dalwani A, Agrawal A, Szinger J, et al. Los Alamos hepatitis C immunology database. Appl Bioinformatics 2005;4: 217–225.

[37] Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 2012;40:D593–D598.

[38] Combet C, Bettler E, Terreux R, Garnier N, Deleage G. The euHCVdb suite of in silico tools for investigating the structural impact of mutations in hepatitis C virus proteins. Infect Disord Drug Targets 2009;9:272–278.

[39] Kwofie SK, Schaefer U, Sundararajan VS, Bajic VB, Christoffels A. HCVpro: hepatitis C virus protein interaction database. Infect Genet Evol 2011;11: 1971–1977.

[40] Su WH, Chao CC, Yeh SH, Chen DS, Chen PJ, Jou YS. OncoDB.HCC: an integrated oncogenomic database of hepatocellular carcinoma revealed aberrant cancer target genes and loci. Nucleic Acids Res 2007;35: D727–D731.

[41] Hsu CN, Lai JM, Liu CH, Tseng HH, Lin CY, Lin KT, et al. Detection of the inferred interaction network in hepatocellular carcinoma from EHCO (Encyclopedia of Hepatocellular Carcinoma genes Online). BMC Bioinformatics 2007;8:66.

[42] Staib F, Krupp M, Maass T, Itzel T, Weinmann A, Lee JS, et al. Cell MinerHCC: a microarray-based expression database for hepatocellular carcinoma cell lines. Liver Int 2014;34:621–631.

[43] Shimokawa K, Mogushi K, Shoji S, Hiraishi A, Ido K, Mizushima H, et al. iCOD: an integrated clinical omics database based on the systems-pathology view of disease. BMC Genomics 2010;11:S19.

[44] Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature 2010;464: 993–998.

[45] <https://tcga-data.nci.nih.gov/tcga>, retrieved October 05, 2014.