
JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

# MEDICALGENOMICS.ORG - AN OPEN SOURCE DATABASE AND RETRIEVAL SYSTEM FOR BIOMEDICAL ANALYSIS

Dissertation
zur Erlangung des Grades
„Doktor der Naturwissenschaften“

am Fachbereich Biologie
der Johannes Gutenberg-Universität
in Mainz

vorgelegt von

Markus Krupp

geboren am 2. April 1980
in Bingen am Rhein

Mainz, den 12. Juni 2014

Mündliche Prüfung: 14. August 2014

**Erklärung**

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

*Mainz, 12. Juni 2014*

Markus Krupp

# ZUSAMMENFASSUNG

Die Molekularbiologie von Menschen ist ein hochkomplexes und vielfältiges Themengebiet, in dem in vielen Bereichen geforscht wird. Der Fokus liegt hier insbesondere auf den Bereichen der Genomik, Proteomik, Transkriptomik und Metabolomik, und Jahre der Forschung haben große Mengen an wertvollen Daten zusammengetragen. Diese Ansammlung wächst stetig und auch für die Zukunft ist keine Stagnation absehbar. Mittlerweile aber hat diese permanente Informationsflut wertvolles Wissen in unüberschaubaren, digitalen Datenbergen begraben und das Sammeln von forschungsspezifischen und zuverlässigen Informationen zu einer großen Herausforderung werden lassen.
Die in dieser Dissertation präsentierte Arbeit hat ein umfassendes Kompendium von humanen Geweben für biomedizinische Analysen generiert. Es trägt den Namen medicalgenomics.org und hat diverse biomedizinische Probleme auf der Suche nach spezifischem Wissen in zahlreichen Datenbanken gelöst. Das Kompendium ist das erste seiner Art und sein gewonnenes Wissen wird Wissenschaftlern helfen, einen besseren systematischen Überblick über spezifische Gene oder funktionaler Profile, mit Sicht auf Regulation sowie pathologische und physiologische Bedingungen, zu bekommen. Darüber hinaus ermöglichen verschiedene Abfragemethoden eine effiziente Analyse von signalgebenden Ereignissen, metabolischen Stoffwechselwegen sowie das Studieren der Gene auf der Expressionsebene. Die gesamte Vielfalt dieser Abfrageoptionen ermöglicht den Wissenschaftlern hoch spezialisierte, genetische Straßenkarten zu erstellen, mit deren Hilfe zukünftige Experimente genauer geplant werden können. Infolgedessen können wertvolle Ressourcen und Zeit eingespart werden, bei steigenden Erfolgsaussichten. Des Weiteren kann das umfassende Wissen des Kompendiums genutzt werden, um biomedizinische Hypothesen zu generieren und zu überprüfen.

## ABSTRACT

The molecular biology of humans is a highly complex and multifarious topic with research being carried out at different levels including genome, proteome, transcriptome as well as metabalome levels and years of research and development have created amounts of valuable data. This collection still continues and will always continue for the better development of human being. Meanwhile, however, this continuing increase in the scale of data being produced has buried treasures of knowledge in unmanageable amounts of digital data and it has become a great challenge to extract research specific and reliable information from this mass of data.

The work presented in this thesis has generated a comprehensive compendium of human tissues for biomedical analysis called medicalgenomics.org, and thus has solved several biomedical aspects of identifying specific knowledge out of the numerous online databases available in the Internet. This compendium is the first of its kind and its retrieved knowledge will aid researchers in getting a systematic overview about specific genes or functional profiles in view of regulation as well as pathological or physiological conditions. Moreover, several query options are integrated that enable researchers to allow the efficient analysis of signaling events and metabolic pathways as well as to enable gene studying on their specific expression levels. All those query opportunities will aid professional researchers to generate highly customized genetic road maps, which may compass future experiments so that researchers can orchestrate their experiments more precisely, consequently saving valuable resources and time, while also increasing success rate. Beyond, the comprehended knowledge of the compendium may be accessed to further examine biomedical assumptions as well as proving or generating novel biomedical hypotheses.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## Abbreviations

| | |
|---|---|
| ABL | Abl Tyrosine Kinase |
| AIH | Autoimmune Hepatitis |
| AKT1 | v-akt Murine Thymoma Viral Oncogene Homolog 1 |
| ANOVA | Analysis Of Variance |
| API | Application Programming Interface |
| BCL2 | B-cell CLL/Lymphoma 2 |
| BCR | Breakpoint Cluster Region |
| bp | Basepairs |
| BRCA2 | Breast Cancer 2, Early Onset |
| C9 | Complement Component 9 |
| CCC | Cholangiocellular Carcinoma |
| cDNA | Complementary Deoxyribonucleic Acid |
| ChIP | Chromatin Immunoprecipitation |
| CPS1 | Carbamoyl Phosphate Synthase 1 |
| cRNA | Complementary Ribonucleic Acid |
| DAVID | Database for Annotation, Visualization and Integrated Discovery |
| DNA | Deoxyribonucleic Acid |
| EBI | European Bioinformatics Institute |
| EGFR | Epidermal Growth Factor Receptor |
| ENCODE | Encyclopedia of DNA Elements |
| FPKM | Fragments Per Kilobase of exon model per Million mapped reads |
| GA | Genome Analyser |
| GEO | Gene Expression Omnibuss |
| HCC | Hepatocellular carcinoma |
| HPX | Hemopexin |
| HUGO | Human Genome Organization |
| INSDC | International Nucleotide Sequence Database Collaboration |
| IPA | Ingenuity Pathway Analysis |
| ITIH2 | Inter Alpha Trypsin Inhibitor Heavy Chain 2 |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |

| | |
|---|---|
| LoMA | Library of Molecular Associations |
| MAPK | Mitogen Activated Protein Kinase |
| Mb | Mega Basepairs |
| MEDLINE | Medical Literature Analysis and Retrieval System Online |
| MIAME | Minimum Information About a Microarray Experiment |
| miRNA | Micro Ribonucleic Acid |
| MM | Mismatch |
| mRNA | Messanger Ribonucleic Acid |
| MYC | v-myc Avian Myelocytomatosis Viral Oncogene Homolog |
| NASH | Nonalcoholic Steatohepatitis |
| NCBI | National Center for Biotechnology Information |
| NGS | Next Generation Sequencing |
| NIH | National Institutes of Health |
| NLM | National Library of Medicine |
| nt | Nucleotide |
| ORM1 | Orosomucoid 1 |
| PBC | Primary Biliary Cirrhosis |
| PCR | Polymerase Chain Reaction |
| PM | Perfect Match |
| RNA | Ribonucleic Acid |
| RPKM | Reads Per Kilobase of exon model per Million mapped reads |
| rRNA | Ribosomal Ribonucleic Acid |
| SAM | Sequence Alignment/Map |
| SBH | Sequencing By Hybridization |
| snoRNA | Small Nucleolar Ribonucleic Acid |
| SNP | Single Nucleotide Polymorphisms |
| SRA | Sequence Read Archive |
| TAT | Tyrosine Aminotransferase |
| TPM | Transcripts Per Million mapped reads |
| TP53 | Tumor Protein 53 |
| tRNA | Transfer Ribonucleic Acid |
| UCSC | University of California, Santa Cruz |

# 1 Introduction

## 1.1 Overview

Over the last two decades the field of biomedical science has undergone a drastic change. This change was particularly characterized by the enormous growth in knowledge for research. This ongoing growth can be attributed to the continual rise of novel technologies. One of the first milestones in modern biomedical science was the discovery of the polymerase chain reaction in 1971, which revolutionized research and diagnostics [1]. As a result of this technology, genetic linkage as well as mutation analysis has become a simple task and the identification of disease associated genes grew enormously. A further landmark, which promotes this identification, was the publication of a DNA sequencing method, presented by Frederick Sanger and colleagues in 1977 [2, 3]. Moreover during 1994, the advent of microarray technology has enabled the determination of disease associated genes on a large scale [4]. Since then, simultaneous measurement of the complex network of numerous genes has been possible. This progress has been followed by the decoding of the human genome by the Human Genome Project [5] and the competing Celera Genomics Project [6] in the beginning of the new millennium. These projects aim to create a physical map of the human genome to the best possible detail. With this fundamental basis, the identification of disease associated genes has again become considerably simpler, especially in diseases with polygenetic origins.
Nearly at the same time, the expansion and development of the Internet evolved from a local network composed of 334 users to a global universe with over billion users (Fig. 1.1). This novel communication network also affected the scientific community. Scientific discussions were no longer restricted to privileged places and attending scientists closely related to specific areas. Thereafter, they were also conducted in this novel communication network. This has addressed wider societies of scientists to follow and present their view to discussions.
The combination of accumulated scientific knowledge and a global communication network has contributed to the rise of new interdisciplinary research areas, such as biotechnology or bioinformatics. Moreover, innovative platforms were established to share the gathered scientific knowledge; most notably PubMed, the retrieval system for biomedical literature [7]. In addition, each new technology in biomedical science was accompanied with growing amounts of resulting data. Subsequently, huge databases were created to enable scientists to deposit and share supporting material to corresponding discoveries.
All these advances have led to a flood of freely available information and have opened the door to the post genomic era, in which in silico analyzes are essential to manage the abundance of data. However, despite all these advancements the elucidation of complex diseases, like diabetes and cancer, remains difficult. Thus, one of the biggest challenges in the post genomic era is to get a deeper understand-

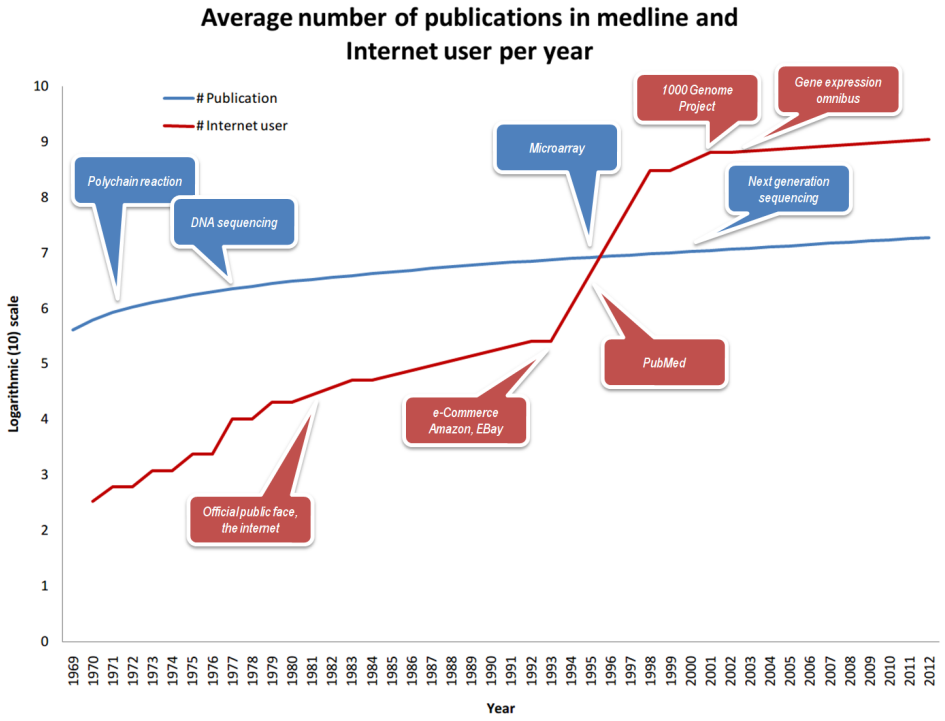


**Figure 1.1:** The graph reflect the growth of biomedical publications compared to the expansion of the Internet.

ing of the complex genome and its interacting mechanisms. A fundamental step in this process is to gain a better understanding of the valuable discoveries already produced and stored in various literature and databases.

## 1.2 Motivation

For the last 25 years, the biomedical community has gathered scientific knowledge on a large scale. This knowledge has improved biological research as well as clinical medicine, especially in diagnosing and understanding the behavior of various cells and organs in a human body as well as in maintaining and promoting health in humans in terms of basics of diseases and immunology. However, the huge increase in produced data in the post genomic era has buried treasures of knowledge. Thus, one of the greatest challenges facing the biomedical community today is to understand the wealth of data that has been produced by those innumerable projects. This process is being accompanied by an intensive filtering for specific knowledge related to the area of interest and comprehensive data analysis, if the data is present in raw format and is not complied. However, filtering for specific information can confront researchers with severe obstacles, which might become even invincible if the knowledge is hidden in raw format.

## 1.3 Objective

The work described in this thesis was aimed to offer professional researchers tools which support access to the wealth of biomedical knowledge with focus on human tissues, especially liver. Furthermore, specific tools, which are based on obtained knowledge and will help users to design their experiments individually that it will not only save valuable resources and time but also increase the success rate.

## 1.4 Synopsis of the thesis

This thesis is structured as followed:

- Chapter 2 introduces scientific literature databases, text-mining strategies to extract literature based knowledge, the developed workbench LoMA (Publication 1) and a systems biology analysis of LoMA (Publication 2).
- Chapter 3 reviews the microarray technology, analysis pipeline and databases. Further, the developed workbench CellMinerHCC and CellLineNavigator are outlined. Moreover, it introduces the discovery of a conserved expression profile in liver cancer.
- Chapter 4 reviews next generation technology, RNA-Seq analysis and databases. It also briefly describes the developed workbench RNA-Seq Atlas and its application to generate a healthy liver signature.
- Chapter 5 - Publication 1 - lists a copy of the original publication entitled: *Library of molecular associations: curating the complex molecular basis of liver diseases*. Published in BMC Genomics in 2010.
- Chapter 6 - Publication 2 - lists a copy of the original publication entitled: *A systems biology perspective on cholangiocellular carcinoma development: focus on MAPK-signaling and the extracellular environment*. Published in Journal of Hepatology in 2009.
- Chapter 7 - Publication 3 - lists a copy of the original publication entitled: *CellMinerHCC: a microarray-based expression database for hepatocellular carcinoma cell lines*. Published in Liver International in 2013.
- Chapter 8 - Publication 4 - lists a copy of the original publication entitled: *CellLineNavigator: a workbench for cancer cell line analysis*. Published in Nucleic Acids Research in 2013.
- Chapter 9 - Publication 5 - lists a copy of the original publication entitled: *RNA-Seq Atlas: a reference database for gene expression profiling in normal tissue by next generation sequencing*. Published in Bioinformatics in 2012.
- Chapter 10 discusses the achievement of the presented work and suggests possible areas of future work.

# BIBLIOGRAPHY


[1] K. Kleppe, E. Ohtsuka, R. Kleppe, I. Molineux, and H. G. Khorana. Studies on polynucleotides. XCVI. repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of molecular biology*, 56(2):341–361, **March 1971**. PMID: 4927950.

[2] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi x174 DNA. *Nature*, 265(5596):687–695, **February 1977**. PMID: 870828.

[3] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, **February 1977**. PMID: 265521.

[4] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, **October 1995**. PMID: 7569999.

[5] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L.

Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351, **February 2001**. PMID: 11181995.

[6] Complete sequence and gene map of a human major histocompatibility complex. the MHC sequencing consortium. *Nature*, 401(6756):921–923, **October 1999**. PMID: 10553908.

[7] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(Database issue):D8–D20, **January 2013**. PMID: 23193264.

# 2 The wealth of scientific knowledge - Treasures buried in millions of publications

## 2.1 Overview

On January 5th in 1665, the first scientific journal, the *Journal des scavans*, began publication in France. Three month later, the *Philosophical Transactions of the Royal Society* began publication in England, the first systematically publishing scientific journal. Already at the end of the 18th century more than a thousand research results were publically available, and the number has increased rapidly after that. With the advent of the computer area and the Internet, a dramatic shift in how research results enter the scientific community has taken place. Not only the way of submitting research article: from double keyboarding publications by hand, to scanning and using optical character recognition (OCR), to importing records supplied by publishers in eXtensible Markup Language (XML) format; has changed, but also the way of publication: from printed journals to make the publication on the Internet available (Fig. 2.1). Furthermore, the computer

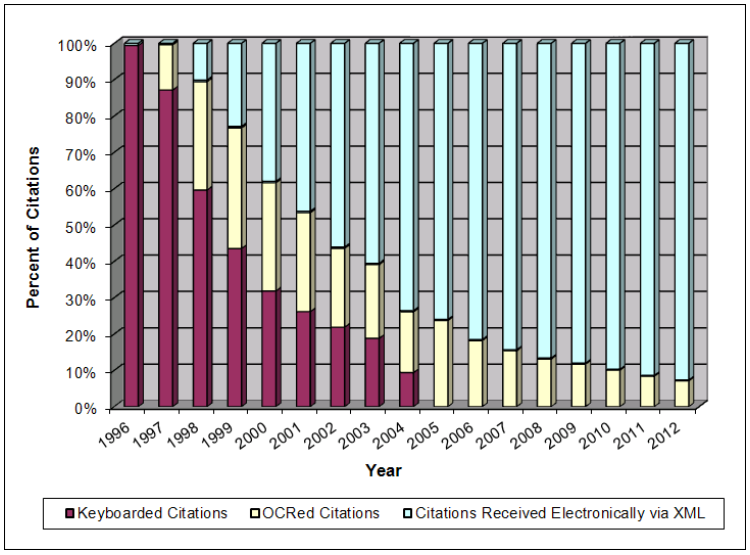


**Figure 2.1:** Beginning in the 1997, a dramatic shift in how research results enter the scientific community has taken place: from double keyboarding publications by hand; to scanning and using optical character recognition (OCR); to importing records supplied by publishers in eXtensible Markup Language (XML) format.

area was accompanied with advances of high-throughput technology and rapid

growth of research capacity in producing large-scale biomedical data leading to an exponential growth of biomedical literature. As a consequence of the large volume of scientific literature and its exponential growth, the acquisition of significant importance for researchers in making relevant discoveries had become increasingly difficult. In response, literature databases with powerful retrieval systems were developed to enable scientists to search for specific content in the wealth of biomedical knowledge distributed over thousands of journals. Among them PubMed, Thomson, Scientific, EMBASE, HighWire Press, Science Direct, Scopus and Cochrane Collaboration, which have become well established, scientific retrieval systems over the years (Table 2.1).

**Table 2.1:** Major scientific literature retrieval systems.

| Name | Hyperlink | Comment |
|---|---|---|
| PubMed | http://www.ncbi.nlm.nih.gov/pubmed | More than 7000 life science and biomedical journals |
| Thomson Scientific | http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=D | 8700 international scientific journals |
| EMBASE | http://www.elsevier.com/online-tools/embase | More than 7000 international biomedicine and pharmacology scientific journals |
| HighWire Press | http://highwire.stanford.edu/ | Covers more than 1000 scientific journals |
| Science Direct | http://www.sciencedirect.com/ | More than 2000 biomedical and other journals |
| Scopus | http://www.scopus.com/ | More than 12850 scientific, medical, social science, and technical academic journals |
| Cochrane Collaboration | http://www.cochrane.org/cochrane-reviews | Collection of databases featuring more than 2000 systematic reviews |

In conclusion, the availability of such a rich resource on knowledge can be a triumph and a tragedy at the same time!
A triumph in term of offering researchers a wide spectrum of journals broadening their horizon beyond the publications related to their fields, which has led to new interdisciplinary research areas, such as molecular biology, molecular genetics, computer science or bioinformatics. On the other hand a tragedy, because despite the continuously improvements of the database design and web services researchers can be confronted with unmanageable mass of information, not allowing them to find specific information. To this end, highly specified databases focusing on one particular issue might be of great value for researchers.

The following subchapter will address first: PubMed, the life science and biomedical journals retrieval system of the National Center for Biotechnology

Information (NCBI); second: typical text-mining strategies applied to scientific literature to further narrowing down database query results; third: the application of text-mining strategies to develop the Library of Molecular Associations (LoMA), the first available database providing a comprehensive view and analysis options for published molecular associations on multiple liver diseases; and finally, the performed systems biology analysis to LoMA that reveal essential functions and structures key to cholangiocellular carcinoma (CCC) progression. These data may provide a rationale for treatment of CCC with sorafenib.

A copy of the original publication of LoMA, published in the journal *BMC Genomics* 2010, is listed in Chapter 5; the publication of the systems biology analysis to LoMA, published in the *Journal of Hepatology* 2009, in Chapter 6 respectively.

## 2.2 PubMed: The life science and biomedical journals retrieval system of the National Center for Biotechnology Information

PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a retrieval system for biomedical literature from MEDLINE, life science journals, and online books. It encompasses publications, abstracts, indexing terms (refer to MeSH terms below) and outgoing links to publishers' web site. The free web service is developed and maintained by the National Center for Biotechnology Information (NCBI), one of the world's premier web sites for biomedical and bioinformatics research based within the National Library of Medicine (NLM) at the National Institutes of Health (NIH), USA. At current state PubMed contains over 22 million publications and with more than 19 million contributed publications, MEDLINE is the primary basis of PubMed [1, 2]. The remaining publications stored in PubMed come from some OLDMEDLINE publications that have not yet been updated with current vocabulary and converted MEDLINE status as well as publications that are out-of-scope from MEDLINE and some other life science journals.
MEDLINE is the NLM's premier bibliographic database. Its broad scope is biomedicine and health encompassing the areas of life science, behavioral science, chemical science, bioengineering and public health. Moreover, MEDLINE includes life science publications critical to biomedical researchers. Those publications cover aspects of plant, animal and environmental science, marine biology, chemistry as well as biology and biophysics. The information of approximately 5,600 international journals in 39 languages were listed in MEDLINE covering the years 1948 to the present, with some older material. Since 2005, each day Tuesday through Saturday between 2,000 to 4,000 complete references are added to MEDLINE; nearly 750,000 total added in 2012 (Fig. 2.2) [2]. A distinctive feature of MEDLINE is that the records are indexed with NLM Medical Subject Headings (MeSH) [3]. These indexes, also known as concepts or MeSH terms, link MEDLINE records to one another as *related publications* on the basis of computationally detected

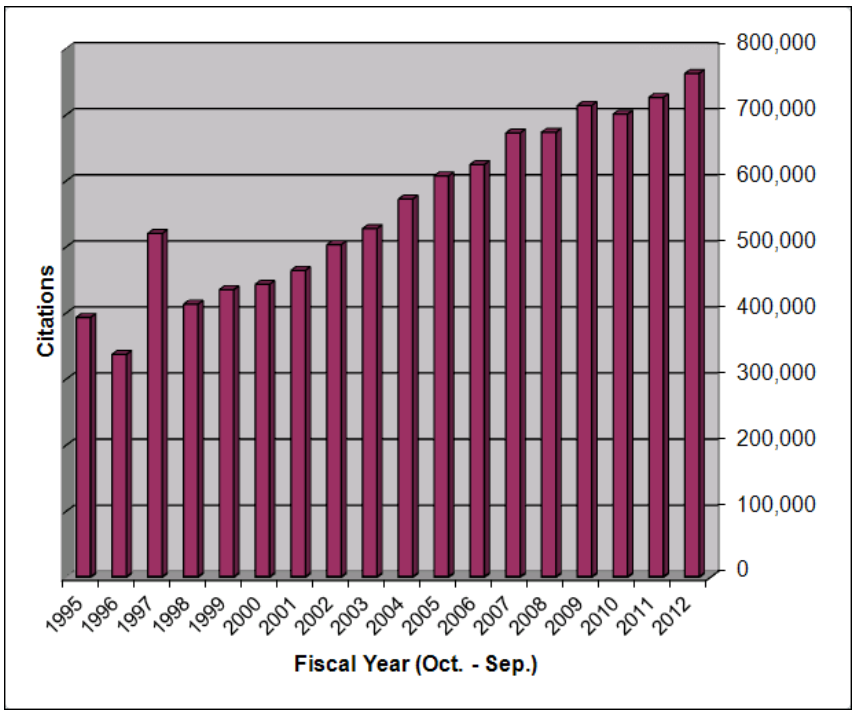


**Figure 2.2:** The graph reflect the number of publications added to MEDLINE during each fiscal year since 1995.

similarities. At current state, MEDLINE is organized into over 235,000 concepts, whereas an entry can be associated to several concepts [2]. This controlled vocabulary thesaurus is adopted to PubMed and can be accessed in the web interface using the [*MeSH*] at the end of a search query. For example, executing the search query *"Carcinoma, Hepatocellular"* [*MeSH*], PubMed will systematically return publications which are linked to hepatocellular carcinomas, Liver cancer, liver cell carcinomas, hepatocellular carcinoma, hepatoma, and 15 further related terms.

To conclude, PubMed is an open source retrieval system accessing over 22 million life science and biomedical publications, primarily accessed from the MEDLINE database. In addition to normal queries, PubMed can be queried by means of MeSH terms.

## 2.3 Text-mining strategies in life science

Twenty-two million papers in PubMed: a triumph or a tragedy? It has become a triumph with the modification of conventional or newly developed text-mining strategies to filter life science information on a large scale. After the adjustment of those underlying algorithms to users need, those strategies can be used to classify the unmanageable mass of information into helpful and unhelpful publications. Hereby, a fundamental distinction should be made between database retrieval systems, e.g.

PubMed [1, 2], and text-mining approaches. In both cases, a database system is the crucial basis. Whereas databases archive and index data to enable fast access to the information. In turn, database retrieval systems, exclusive designed for corresponding databases, can be used to querying and returning specific information. However, well defined queries, even with the help of logical expressions such as *AND*, *OR* or *NOT*, may result in outrageous volume of information and desired information merely coexists within the other valid pieces in the results. Because, database system will always return all publications that were linked to the queried terms, including publications were the terms are mentioned briefly, used as reference to other publications or just distributed within the information without any coherence. The great benefit of text-mining towards naive database queries is that text-mining can actively differ between rational relations between search terms (for example gene-disease relations) and undesirable by-products, such as information without any coherence, because key terms are randomly distributed. For example, researchers interested in the role of P53 in human liver cancer can use the retrieval system of PubMed to screen literature with the descriptive term *P53 AND "Carcinoma, Hepatocellular" [MeSH]* about what is already known. At current state, querying for this term result in 1,500 PubMed publications - an unmanageable amount of reading material. However, elaborated text-mining strategies, with the focus of analyzing the resulting list for truly gene-disease associations, narrowed down the resulting list of publications to less than 50.
Thus, text-mining studies in life science permit the discovery of genuinely new, previously unknown information, by automatically extracting knowledge from a usually large amount of publications. The process of text mining are implemented in programming languages like Perl (http://www.perl.org/) or Python (http://www.python.org/) and can be separated into following steps:

1. Text acquisition
2. Text preprocessing
3. Text transformation
4. Attribute selection
5. Pattern discovery
6. Interpretation / evaluation

**Text acquisition** occurs by a variety of ways and depends on researchers interests, e.g. in the subsequent analysis the text acquisition was based on PubMed and all publications linked to liver diseases. In the dependency of textual data volume and/or suit of application, document clustering is applied to the data [4]. Here, k-means and agglomerative hierarchical clustering methods are commonly used [5, 6].
In the **text preprocessing** step, textual data were normalized and afterwards single words were split into a set of tokens [5]. Normalization comprises removal of tab

stops, line breaks and figures as well as the reformatting of tables and formulas; and tokenization is needed to deal with issues like hyphens, apostrophes, special characters and terms (e.g. C++, G/T, A/C).
Step three, **text transformation**, can be subdivided into a) stop word removal b) stemming c) ranking of words and based thereupon d) feature selection. Stemming warrants that the dimensionality of a term is reduced. In other words, it identifies and convert the word by its roots (e.g. analyzing, analyzed, $\ldots \rightarrow$ analyze). Common algorithms applied for stemming are the KSTEM [7] and Porter's algorithm [8]. After executing a) and b) a ranking schema for each publication is applied. The most popular ranking schema normalized the word by its frequency $\mathtt{tfidf}$:

$$\mathtt{tfidf}(w) = \mathtt{tf}(w) * \log(\tfrac{\mathrm{N}}{\mathtt{df}(w)})$$

Whereas $\mathtt{tf}(w)$ equals to the number of word occurrences in a document, $\mathtt{df}(w)$ to the number of document containing the word, N to the number of all document in the data set and $\mathtt{tfidf}(w)$ the rank of the word $w$ in the document. In simple terms, the more often the word $w$ appears in the target publication and is seen in less other publications, the higher the rank of the word within the target publication. Finally, the thereupon feature selection for each publication is based on the highest ranking words within the corresponding publication [4].
The forth step of further reducing the dimensionality as well as removing irrelevant features in text mining, here described as **attribute selection**, is not necessary and depends on suit of application. Word statistics could be generated for example, examined and used for further filtering steps.
At this point the data are organized into a well structured text format. This structured text is the fundamental basis for the next step of pattern discovery. **Pattern discovery** and the subsequent **data interpretation or evaluation** is a purely application dependent stage. Therefore, depending on the research focus, highly specified algorithms must be individually designed. Based on their functionality those algorithms may generate genuinely new knowledge or may even formulate new hypotheses.
One of the first medical hypotheses generated by text-mining strategies was carried out by Swanson et al. [9] in 1998. Swanson et al. investigated the cause of migraine headaches by extracting evidence from various text fragments included in titles of publications; a hypothesis which did not exists in the literature before. These hypotheses suggest that magnesium deficiency may act as a key player in some migraine headaches. Some of the indications to generate the hypotheses can be paraphrased as followed:

- **Stress** ... is associated with migraines.
- **Stress** ... can lead to loss of magnesium.
- **Calcium channel blockers** prevent some migraines.
- ... magnesium is a natural **calcium channel blocker**.

- ... **Spreading Cortical Depression (SCD)** is implicated in some migraines.
- High levels of magnesium inhibit **SCD**.
- Migraine patients have high **platelet aggregability**.
- Magnesium can suppress **platelet aggregability**.

In 1989, subsequent analysis by Ramadan et al. [10] provides evidence for this text-mining generated magnesium-migraine hypothesis. A targeted application of text-mining to generate genuinely new knowledge on serious liver diseases is introduced in the next subchapter and the successfully, subsequent evaluation of gained knowledge was used to provide a rationale for treatment of cholangiocellular carcinoma with sorafenib in the next-but-one.

In conclusion, text-mining is a bunch of algorithms to reformat unstructured textual information into structured and thus computer interpretable content. The organized data is further interpreted by highly specified, application depended, algorithms to generate new knowledge or to establish novel hypotheses. However, during the whole process of text-mining, particular attention must be paid to ambiguous word or sentence text characteristics and multiple negated phrases, they can lead to inaccurate results.

## 2.4 Publication 1 - The workbench: Library of Molecular Associations - a targeted application of text-mining to generate a literature based workbench for liver disease

LoMA - the Library of Molecular Associations - is the first available database providing a comprehensive view and analysis option for published molecular associations on multiple liver diseases. This study was published in the journal *BMC Genomics* in 2010, a copy of the original publication can be found in Chapter 5. The fundamental database of this workbench was generated by initially querying PubMed for liver disease publications and subsequent application of developed text-mining algorithms to the resulting PubMed abstracts. All the filtered abstracts were further manually validated to confirm content and potential genetic associations and may therefore be highly trusted. All data were publicly available at http://medicalgenomics.org/loma, containing currently approximately 1,260 confirmed molecular associations to the chronic liver diseases autoimmune hepatitis (AIH), cholangiocellular carcinoma (CCC), fibrosis, hepatocellular carcinoma (HCC), nonalcoholic steatohepatitis (NASH) (fatty liver), primary biliary cirrhosis (PBC) and primary sclerosing cholangitis (PSC).

In order to establish this database, the complete PubMed [1] database, at this time containing over 17 million publications, has been searched for each liver disease or respective MeSH term [3]. For example, the MeSH search string to

identify disease associated abstracts for PBS was in detail: *"PBC" OR "primary biliary cirrhosis" [MeSH] OR "biliary cirrhosis, primary"*. Each abstract identified to be associated with the particular disease were processed by conventionally text-mining algorithms. As the result, the unstructured textual information hidden in the abstract was converted into structured, computer-accessible data. After that, highly sophisticated pattern discovery algorithms were developed in the programming language Perl (http://www.perl.org/), whose task was to screen for human, mouse and rat gene names and alias gene names as provided by the Human Genome Organization (HUGO, http://www.hugo-international.org.) in each previously selected abstract. For example, if the gene to be searched was p53, the abstract was searched for any combinations of signs starting with the letter p followed by the numbers 5 and 3 and all its alias gene names with the same approach. Furthermore, if the algorithms identified a sentence with the specific gene, the sentence was crosschecked for its association to the targeted disease. By this approach a total of 101,026 abstracts, potential holding information on generic applications to chronic liver diseases, were collected. In detail 917 abstracts suggesting genetic associations for AIH, 13,710 for CCC, 37,173 for liver fibrosis, 44,548 for HCC, 2,022 for NASH, 1,211 for PBC and 1,445 for PSC were identified. During the process of text-mining, the algorithms were designed to filter all abstracts associated to molecular associations in liver diseases. At this high amount of sensitivity, the algorithm lacks of specificity in the way of recognizing ambiguous word or sentence text characteristics and multiple negated phrases. But this lack of specificity is a general problem of text-mining algorithms. For example, the abstract may describe that *gene XY is related to disease Z, but not gene A*, which would link gene A to disease Z by the described text-mining strategy. In order to close the gap of sensitivity versus specificity all resulting abstract were manually reviewed for de facto molecular associations and if approved stored in the LoMA database. By the end of the text-mining and manual curation step, 310 molecular associations for CCC, 150 molecular associations for liver fibrosis, 574 molecular associations for HCC and 82 molecular associations for NASH were identified and stored in the medicalgenomics.org postgreSQL (http://www.postgresql.org) database. Only a few genes were identified to be related to the development of autoimmune liver disease: 29 abstracts describing molecular associations were found to be related to AIH, 56 to PBC, and 60 to PSC. Subsequently, this database was then made publicly accessible and searchable through a retrieval system (Fig. 5.1, 5.2, 5.3) implemented in PHP http://www.php.net, a download option is also supported. Since one of the major goals in implementing this database was to perform high throughput systems biology analyses, the LoMA genetic associations had to be linked to commonly used and established bioinformatics databases and knowledge repositories.

Overall, a total of 1,260 molecular associations for major chronic liver diseases were identified using semi-automated text-mining strategies and a powerful retrieval systems was implemented to allow a multitude of querying options to

the data. The so called LoMA workbench is the first available system providing a comprehensive view and analysis option for published molecular associations on multiple liver diseases. One of the versatile applications areas of LoMA is demonstrated in the subchapter that follows. It describes a systems biology analysis of the cholangiocellular carcinoma profile located in LoMA and the discovery of two significant enriched pathways which provide a rational for treatment of CCC with sorafenib.

## 2.5 Publication 2 - The discovery: System biology analysis of the LoMA cholangiocellular carcinoma profile reveals a rationale for treatment of CCC with sorafenib

The here described study was published in the *Journal of Hepatology* in 2009 a copy can be found in Chapter 6.

Cholangiocellular carcinoma (CCC) is a comparatively rare cancer arising from the bile ducts. However, rates of cholangiocellular carcinoma have been rising worldwide over the past several decades [11, 12] and therapeutic options for CCC currently remain very limited. Besides surgery, the use of chemotherapy is still a matter of intense debate with many arguing for best supportive care as the standard of treatment [13]. Multiple genes have been implicated in CCC development. However, this study of the CCC profile located in LoMA indicates that the overall neoplastic risk is associated with a much lower number of critical physiological pathways. Those results were furthermore validated in a subset analysis looking only at the microarray derived data, not subject to a possible selection bias by the scientific community. In particular, the MAPK pathway was consistently enriched in CCC. Comparing our data to genetic associations in HCC often successfully treated by a multityrosine kinase inhibitor, sorafenib, the study demonstrated a similar expression pattern of MAPK. These data may provide a rationale for treatment of CCC with sorafenib. Furthermore, genes coding for products in the extracellular environment were identified to be significantly enriched. To this end, CCC must be regarded as developing in the context of an altered extracellular environment.

This study suggests the liver microenvironment holds essential functions and structures key to CCC progression. Furthermore, the MAPK signaling pathway was identified to be consistently enriched, pointing towards a critical role in CCC development. These data may provide a rationale for treatment of CCC with sorafenib.

# BIBLIOGRAPHY


[1] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(Database issue):D8–D20, **January 2013**. PMID: 23193264.

[2] J. McEntyre and J. Ostell. The NCBI handbook, **2002**.

[3] W. SEWELL. MEDICAL SUBJECT HEADINGS IN MEDLARS. *Bulletin of the Medical Library Association*, 52:164–170, **January 1964**. PMID: 14119288.

[4] W. Fan, L. Wallace, S. Rich, and Z. Zhang. Tapping the power of text mining. *Commun. ACM*, 49(9):76–82, **September 2006**.

[5] M. P. Singh, editor. *Practical Handbook of Internet Computing*. Chapman Hall & CRC Press, Baton Rouge, **2004**.

[6] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, page 318–329, New York, NY, USA, **1992**. ACM.

[7] R. Krovetz. Viewing morphology as an inference process. *Artificial Intelligence*, 118(1–2):277 – 294, **2000**.

[8] M. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, **1980**.

[9] D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, **1988**. PMID: 3075738.

[10] N. M. Ramadan, H. Halvorson, A. Vande-Linde, S. R. Levine, J. A. Helpern, and K. M. Welch. Low brain magnesium in migraine. *Headache*, 29(9):590–593, **October 1989**. PMID: 2584000.

[11] S. A. Khan, H. C. Thomas, B. R. Davidson, and S. D. Taylor-Robinson. Cholangiocarcinoma. *The Lancet*, 366(9493):1303–1314, **August**.

[12] Y. H. Shaib, J. A. Davila, K. McGlynn, and H. B. El-Serag. Rising incidence of intrahepatic cholangiocarcinoma in the united states: a true increase? *Journal of Hepatology*, 40(3):472–477, **March 2004**.

[13] M. Shimoda and K. Kubota. Multi-disciplinary treatment for cholangiocellular carcinoma. *World journal of gastroenterology: WJG*, 13(10):1500–1504, **March 2007**. PMID: 17461440.

# 3 Microarray - A technology for gene expression analysis

## 3.1 Overview

In every living creature a complex network of thousands of genes and their products (e.g. mRNA or proteins) is functioning. The measurement of this complexity outreach traditional gene-gene methods, because these techniques were designed to deal with single genes and are thereof not powerful enough to create a global image of a living cell. Due to this a new technology was developed, the microarray technology. This technology is employed in screening for single nucleotide polymorphisms (SNPs), sequencing by hybridization (SBH), characterization of genomes and gene expression analysis [1, 2, 3, 4, 5]. Since the focus of the presented work in the following subchapters as well as Chapter 7 and 8 is restricted to gene expression analysis, the application of microarray to other areas was out of the scope. The microarray technology enables gene expression measurement of a whole genome on a single microarray within hours. In this process, the quantitative amount of a specific sequence is captured. This technology has developed from the detection of a bunch of selective sequences on nylon membranes to whole genome expression profiling on solid glass surfaces within a short period of time [6, 7]. Current microarrays have the potential to measure hundred of thousands of sequences simultaneously. The largest manufacturers for microarrays are Affymetrix (one-channel; http://www.affymetrix.com), Illumina (one- and two-channel; http://www.illumina.com/) and Agilent (one- and two-channel; http://www.agilent.com).

The following subchapter will briefly highlight the microarray technology with the focus on gene expression profiling, the associated data analysis and databases developed to store (or publish depending on point of view) microarray experiments. Moreover, the developed microarray workbenches CellminerHCC and CellLineNavigator were introduced as well as the discovery of a common expression profile on HCC.

A copy of the original publication of CellminerHCC, published in the journal *Liver International* 2013, is listed in Chapter 7; the publication of CellLineNavigator, published in the journal *Nucleic Acids Research* 2013, in Chapter 8 respectively.

## 3.2 Microarray technology - A scientific machinery for gene expression profiling

The genome contains all genetic information of an individual, which is consistent in all cells, but in dependency of cell type and its state (e.g. diseased or healthy) the genomic transcription and translation varies. The ability of microarrays to detect gene expression on a large scale enables studies of essential mechanisms and networks behind those processes.
A Microarray, also called gene chip, can carry hundred of thousands oligonucleotides, more than the total number of genes of any higher organism. Each individual positioning on the chip has a diameter of only 50 micrometer, the whole arrangement usually comprises a few centimeter. In the case of gene expression analysis, in situ synthesized oligonucleotides or cDNA fragments (one-channel, two-channel microarrays respectively, see below) were spotted onto the microarray, which represent mRNA of different genes. This study will conform with the nomenclature proposed by Daggan et al. [8], and sequences spotted to the solid substrate will be referred as probe. Most of the time, probes of the length of 25 to 80 basepairs (bp) are used. The elementary principle behind the microarray technology is basepair hybridization. The elements constituting this process include extraction of desired mRNAs from the sample (target), preprocessing the target to cDNAs by reverse transcriptase, labeling it with fluorescence material and applying the substance to a microarray to start the hybridization process (Fig. 3.1). Whereas

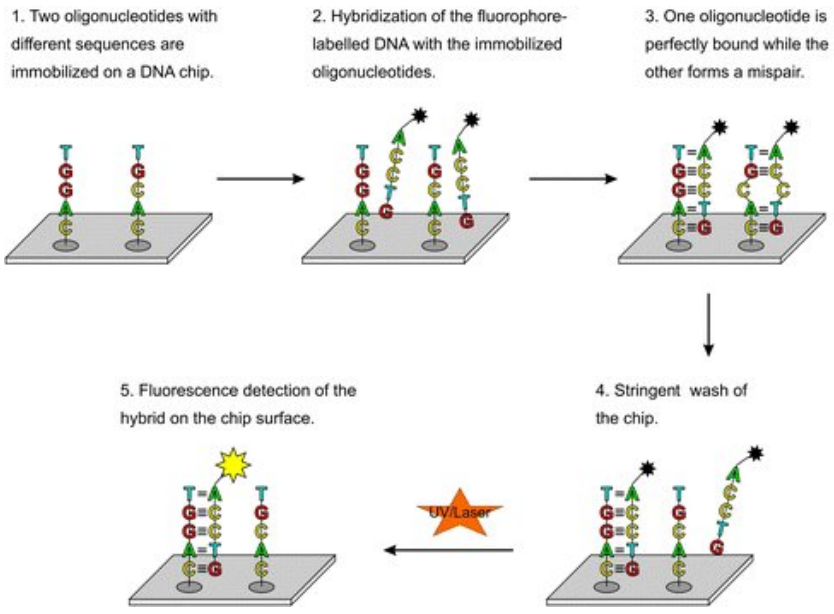


**Figure 3.1:** Principle diagram of oligonucleotide microarrays. T, G, A, C represent the four canonical building blocks of the DNA, the asterisk indicated the fluorescence label.

the sequence specificity ensures that a target binds to its specific position on the mi-

croarray. After the hybridization process, the unbound sequences are washed away and the quantified amount of hybridized target is detected by the use of laser technology. The resulting image files cover the emitting light intensity of each probe. By which the amount of fluorescence emitted by each spot will be proportional with the amount of mRNA isolated from a sample obtained under a particular biological state. Depending on whether one or two samples were spotted onto the chip, the experiment is defined as one-channel or two-channel microarray experiment. In the later case, two differently labeled samples competitively bind to the complementary probe on the microarray and two different lasers are used in two different steps to detect the brightness level of each sample individually. For example, sample A was labeled with a green and sample B with a red dye (Fig. 3.2). Focusing on one spot on the array, its expression intensity can be characterized as:

- No sequence has hybridized → No light is emitted.
- Sequences of sample A hybridized → Levels of green light is emitted.
- Sequences of sample B hybridized → Levels of red light is emitted.
- Sequences of sample A & B hybridized → Levels of yellow light is emitted.

The resulting image file would in some extent immediately reveal different expressed probes. Whereby the above example applied to one-channel experiments would result in two different image files and must be first superimposed before expression levels could be analyzed. Because in one-channel experiments each sample is distributed to different microarrays.
A judgment about the superiority of one-channel versus two channel and vice versa is difficult. The cDNA technology seems to be more flexible, allowing spotting of almost any PCR product whereas the in situ synthesized oligonucleotides technology seems more reliable and easier to use. The introduced analysis techniques and tools in the next but not one subchapter can be applied to any type of microarray.

To conclude, microarrays are versatile tools to measure thousand of gene expression values in parallel. It is distinguished between one-channel and two-channel microarrays; two-channel microarrays are typically hybridized with two samples labeled with two different fluorophores.

## 3.3 Databases

The increasing influence of the microarray technology in the biological world has lead to the generation of a vast number of microarray experiments focusing a variety of different biological organisms, tissues or states or any combination out of them. Although many important conclusions have been drawn by these studies, the development of standards for presenting and exchanging such data has been neglected. To this end, the minimum information about a microarray experiment

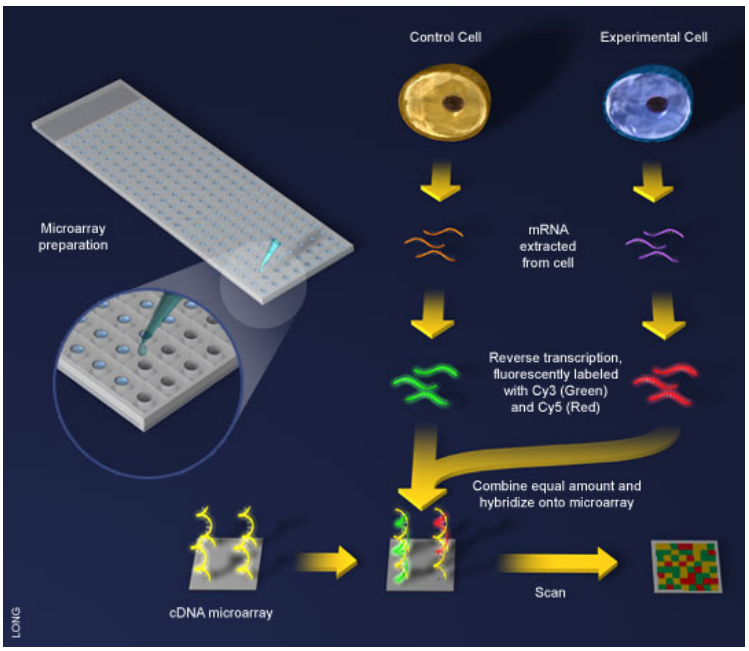


**Figure 3.2:** Workflow of a typical two-channel microarray experiment.

(MIAME) standards for microarray data was introduced in 2001 by Brazma et al. [9]. With this standard more and more microarray data were stored in public available database. Nowadays, three major international repositories exist: the Gene Expression Omnibus (GEO) [10], the ArrayExpress [11] and the Stanford Microarray Database [12] hosted at the NCBI, European Bioinformatics Institute (EBI) and Standford University respectively. Currently, the GEO database hosts more than 32,000 studies comprising over 800,000 samples, the ArrayExpress over 25,000 and more than 700,000. The Stanford Microarray Database stores data on 82,000 experiments, no information on samples size is supported.

## 3.4 In silico analysis - A computer-based challenge of analyzing hundred of thousands scientific experiments

The microarray technology has become a promising tool for various applications regarding the analysis of genetic material. It allows simultaneous nucleic acid hybridization for a large number of immobilized oligonucleotides on a small surface area. After the hybridization process with fluorescence labeled targets, the microarray is scanned and the resulting image file is analyzed such that the signal from each probe can be quantified into numerical values. The fundamental steps in analyzing resulting image files are:

1. Image processing
2. Quality control

3. Data preprocessing and normalization
4. Identification of differentially expressed genes
5. Functional analysis and biological interpretation of microarray data

**Image processing** encompasses the investigation of rectangular arrays of intensity values stored in digital images characterized by several resolutions and color depths. Each intensity value corresponds to several points in the image file and are defined as pixels and the total number of pixels is called resolution. The color depths is the number of bits used to save the intensity value of a each pixel. A successful analysis of microarray experiments highly depends on the resulting image file and due to this reason those files must satisfy minimum requirements of resolution and color depths. Minimum criteria for microarrays is a spot diameter of ten pixel and a color depth of 16 bits, which allow pixels to display $2^{16} = 65,536$ different intensity levels. To convert the visual information of mRNA levels present in the tested samples into numerical values several steps must be accomplished: microarray location, image segmentation, quantification, and spot quality assessment. Microarray location covers the spot finding procedure. Image segmentation is done either by spatial or intensity information, it is differentiated between the three algorithms: fixed circle, adaptive circle, and seeded region (Fig. 3.3). Quantification

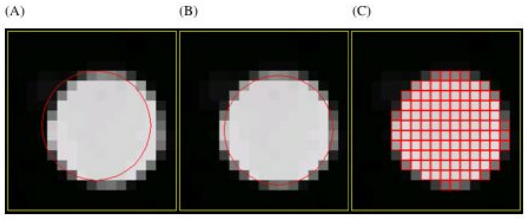


**Figure 3.3:** Microarray image segmentation: (A) fixed circle, (B) adaptive circle, and (C) seeded region. Yellow lines indicates the microarray location step. Red circles of squares indicate which pixel are considered signal for each method.

merges the values of various pixels to calculate a unique numerical value characterizing the expression level of the spot. This representative value is usually calculated using the median or mean of the corresponding signal intensities. Spot quality is usually evaluated by using the ratio between the signal area and the total spot area in combination with the shape regularity. Depending on the quality score genes are described as marginally present or absent.
The procedure of **quality control** should be considered as an extra step rather than a substitute for the quality control that is performed in the laboratory. The quality control assesses the condition of the microarray data obtained in a given experiment and ensures the further analysis and the correct interpretation of the data. In fact, only a single abnormal microarray can completely falsify the analysis of a large data set. Mostly, problems arise due to rapid mRNA degradation if the sample is not processed properly immediately after collection or with the quality of the

microarray itself. Various tools exist to indicate microarrays of poor quality, which should be removed from the data set. The most common ones are intensity distribution plots and box plots. Probe intensity images, quality control metrics and RNA degradation curves are further tools for quality assessment.
Having a data set of reliable **microarrays preprocessing and normalization** is initiated. Preprocessing extracts or increase relevant data characteristics, for example preprocessing is the logarithm transformation of the raw values or combining replicates. Normalization is a step to account for systematic difference across data sets by standardizing the data in such a way that they are independent of the particular experiment and technology used. In other words normalizations allow values corresponding to specific genes to be compared directly from one microarray to another. A plethora of refined normalization methods have been established over the time (e.g. background correction (mean, median), color normalization (curve fitting, LOWESS/LOESS), quantil) and should become attuned to experiment design [13, 14, 15, 16].
For the **identification of differentially expressed genes** a variety of methods is available, too. The calculation of fold change is widespread used:

$$\mathtt{fc}_g = \log_2(\frac{\mathtt{int}_{g,exp}}{\mathtt{int}_{g,ctrl}})$$

Whereas $\mathtt{fc}$ corresponds to the fold change, $\mathtt{int}_{g,exp}$ to the mean or median intensity values of the experiment, $\mathtt{int}_{g,ctrl}$ of the control, respectively, of gene $g$. In general, $\mathtt{fc}_g = 0$ implies that there is no difference between the gene expression in the experiment versus the sample. On the other hand, $\mathtt{fc}_g \geqslant 1$ ($\mathtt{fc}_g \leqslant -1$) indicate that twice (half) the amount of mRNA was found in the experiment compared to the control.
Although in widespread use, the method of defining a single gene as differentially expressed, not talking about gene set enrichment analysis, by just looking at the fold change and without a clear biological justification is just a blind guess. Because the fold change will constantly report few genes as regulated even if two identical tissues are compared to each other (false positive). Therefore, selecting differentially expressed genes must be accompanied by classical hypothesis testing approaches. Hereby, the Wilcoxon or t-test is a common univariate test statistics [17, 18] to search for significantly differentially expressed genes between two independent samples. If more than two samples (multivariate test statistics) have to be tested for gene regulation the ANalysis Of VAriance (ANOVA) should be used [19, 20, 21]. However, when testing for thousands of differentially expressed genes the correction for multiple comparisons is crucial, because some genes will appear as being significantly different just by chance [22, 23]. A standard method for adjusting p-values is the false discovery rate. Altogether, current statistical methods offer a reliable way to select significant regulated genes. Nevertheless, all such methods depend essentially on a precise experimental design and the existence of replicates.
Similar to text-mining, the step **functional analysis and biological interpretation of microarray data** is a purely application dependent state. According to research

topic further algorithms have to be implemented or already established functional analysis tools must be conducted.
The methods discussed above from quality control to identification of differentially expressed genes can be implemented in R [24], a open source programming language, development environment, as well as an integrated suite of software routines that allow efficient data manipulation, and graphical display. Moreover, the extension bioconductor [25], an open source and open development software project for the analysis and comprehension of genomic data, offers the possibility to load specific bioinformatics modules into the R environment. Depending on module, several analysis step are already implemented and can be applied to the microarray data.

In summary, gene expression microarrays contribute a snapshot of all the transcriptional activity in a biological sample. However, this technology certainly produces a huge amount of data, confronting the researcher to interpret it by exploiting modern computational and statistical methods.

## 3.5 Publication 3 - The workbench: CellMinerHCC - A microarray-based expression database for hepatocellular carcinoma cell lines

CellMinerHCC is a publicly available database providing a comprehensive view and analysis options for microarray data of the most commonly used HCC cell lines and may be of significant use for in vitro modeling of HCC. This study was published in the journal *Liver International* in 2013, a copy of the original publication can be found in Chapter 7. All data are publicly available at http://medicalgenomics.org/cellminerhcc.

In order to create CellMinerHCC, gene expression profiles of 18 HCC cell lines and normal liver samples were analyzed. The HCC cell lines are the following: 7703, Focus, Hep3B, Hep3B-TR, Hep40, HepG2, HLE, HLF, HUH-1, HUH-6, HUH-7, PLC/PRF/5, SK-Hep1, SNU-182, SNU-387, SNU-389, SNU-449 and SNU-475. As control for two-channel microarrays, a pool of total RNA from 19 normal liver samples was used [26]. Oligo microarrays were produced at the Advanced Technology Center at the National Cancer Institute, NIH, USA using 70-mer probes of 21,329 genes. After the isolation of mRNA from the cell cultures and normal liver samples fluorescently labeled cDNAs were synthesized (green cell cultures, red normal liver samples) followed by a competitively hybridization process on a microarray for each cell culture versus normal liver sample. Hybridized arrays were scanned at 10-µm resolution on a GenePix 4000A scanner (Axon Instruments, Foster City, CA) at variable photo-multiplier tube voltage to obtain maximal signal intensities with less than 1% probe saturation. Resulting images were analyzed via GenePix Pro v3.0 (Axon Instruments) as described in the manual. The application of a background model was used to identify well-measured spots and

spots not meeting the criteria were excluded from further analysis. After merging the intensity information of each spot using an average model, the expression ratios were $\log_2$ transformed and after normalization fold change values were calculated. Finally, the genetic profile of the 18 HCC cell lines were stored in the medicalgenomics.org postgreSQL (http://www.postgresql.org) database and made publically accessible and searchable through a retrieval system implemented in PHP (http://www.php.net) (Fig. 7.1, 7.2, 7.3, 7.4). A download option is also supported. Moreover, CellMinerHCC data were combined with commonly used and established bioinformatics databases and knowledge repositories. To enable further functional analysis, CellMinerHCC is linked to the publically available, web-based tool Database for Annotation, Visualization and Integrated Discovery (DAVID) [27].

Altogether, CellMinerHCC is the first database providing a comprehensive view and analysis options for microarray data of the most commonly used HCC cell lines and may be of significant use for in vitro modeling of HCC.

### 3.5.1 The discovery: A conserved expression profile of 195 genes in HCC - A promising starting point for revealing novel therapeutic options

Hepatocellular carcinoma is among the most common malignancies worldwide and its incidence is rising, especially in Asia and Sub-Saharan Africa, but also in Western countries. Simultaneously, the therapeutic options for this disease besides surgery still remain limited. Although both cellular changes that lead to HCC and the aetiological factors responsible for the majority of HCC cases have been recognized, the molecular pathogenesis of this disease lasts elusive [28]. The key to achieve further progress in the therapy of HCC will rely on a better understanding of the underlying biology of HCC progression and growth allowing the development of subsequent targeted therapies against essential molecular mechanisms. To this end, the gene expression profile of all HCC cell lines located in CellMinerHCC were evaluated for differentially expressed genes. With a cut-off of at least two fold-changes in gene expression, between 1,638 genes in HepG2 and 3,214 genes in SNU-398 were identified as regulated. Of these, 195 were identified as differentially expressed over all cell lines (Tab. 7.1). In detail, the 195 commonly regulated gene profile exhibit 163 genes consistently up- and 32 genes consistently down-regulated (Fig. 7.8). In order to obtain a comprehensive overview on the biological function of these genes, the list of commonly regulated genes was investigated using DAVID [27] and the Ingenuity Pathway Analysis (IPA) software [29]. In both methods, the majority of significantly enriched functions and pathways are highly specific for liver and a malignant phenotype, which proves the reliability of data and applied analysis methods (Tab. 7.1 and Fig. 7.6). In addition, other functional networks demonstrated a high association with the profile. Among them, inflammatory response was identified as significantly enriched. The link

between ongoing acute inflammation and HCC has become increasingly tight [30, 31]. Identification of these categories in this gene set is providing additional evidence for this association. Categories like immunological disease connect the inflammatory response with the immune system that is involved in HCC. Enrichment of additional categories like humoral immune response and immune cell trafficking point towards the important role the immune system is playing in development and progression of HCC [32]. Finally, the category lipid metabolism has been identified as third highest ranked biological function when analyzing the 195 commonly regulated HCC cell line genes. For aberrant lipid metabolism, an association with HCC has already been described [33] and may also play a role in obesity and chronic inflammation related development of HCC [34, 31].

Taken together, analysis of the commonly regulated genes among the 18 most often used HCC cell lines for enrichment of signaling pathways, proteins and interactions not only described a liver tumor phenotype, it also identified molecular associations and numerous categories currently under intense scientific development. To this end, further evaluation of HCC on the basis of the introduced conserved expression profile and its associated functional mechanisms may aid in revealing novel therapeutic options.

## 3.6 Publication 4 - The workbench: CellLineNavigator - A workbench for cancer cell line analysis

CellLineNavigator, a workbench for large scale comparisons of a massive collection of diverse cell lines, aims to support experimental design in the fields of genomics, systems biology and translational biomedical research. Currently, this compendium holds genome wide expression profiles of 317 different cancer cell lines, categorized into 57 different pathological states and 28 individual tissues. The CellLineNavigator workbench is freely available at http://www.medicalgenomics.org/celllinenavigator. This study was published in the journal *Nucleic Acids Research* in 2013, a copy of the original publication can be found in Chapter 8.

To establish CellLineNavigator genome-wide expression data of multiple cell lines were downloaded from ArrayExpress [11]. Briefly, the transcript abundance of 317 cancer cell lines was analyzed using a one-channel microarray, the Affymetrix Human Genome-U133 Plus2 GeneChip. This microarray covers the complete human genome for analysis of over 45,000 transcripts and more than 19,000 genes. All data were available in technical triplicates. Corresponding information on tissue site and disease state was supported for each cell line (Fig. 8.1). The differential expression was analyzed using the R-Project [24] bioconductor [25] suite with the following additional libraries: *affy* [35] , *hgu133plus2.db* [36] and *frma* [37, 38]. After quality control, two microarray experiments were neglected

for further analysis because of insufficient RNA level detection. All data were normalized using the *expresso* function of the *affy* package and following settings: background adjustment method: *mas*, normalization method: *quantiles*, Perfect-Match (PM) adjustment method: *mas* and the method used for the computation of expression values: *medianpolish*. Next, the expression median was calculated for each probe set for all cell lines. These values were subsequently used as control to calculate $\log_2$ transformed expression ratios (fold change), after the median expression was calculated for each cancer cell line. Fold change representing the expression levels of tissue sites and disease states were calculated accordingly. Gene expression barcodes were generated using the *frma* (frozen robust multiarray analysis) (default options) and *barcode* (output: Z-score) function implemented in the *frma* package. A frma Z-score $> 5$ suggested that a gene is expressed in a particular tissue. The frma Z-score was generated to allow comparison of the expression profiles with data already present at medicalgenomics.org and other microarray data sets processed with the frma method.
To ensure easy data access, a simple data and an intuitive querying interface were implemented (Fig. 8.2, 8.3, 8.4). It allows the user to explore and filter gene expression, focusing on pathological or physiological conditions. For a more complex search, the advanced query interface may be used to query for:

- Differentially expressed genes,
- pathological or physiological conditions, and
- gene names or functional attributes.

Finally, CellLineNavigator allows additional advanced analysis of differentially regulated genes by a direct link to DAVID [27].

In summary, CellLineNavigator is the first database providing comprehensive summary, display and analysis options for gene expression data of the most commonly used cancer cell lines. It provides access to large microarray data sets without advanced bioinformatics skills. Thus, CellLineNavigator may be of significant aid for in vitro modeling of cancer mechanisms and testing of novel therapeutic approaches.

# BIBLIOGRAPHY


[1] C. Debouck and P. N. Goodfellow. DNA microarrays in drug discovery and development. *Nature Genetics*, 21:48–50, **January 1999**.

[2] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, **January 1999**.

[3] E. M. Southern. DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. *Trends in Genetics*, 12(3):110–115, **March 1996**.

[4] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, **October 1995**. PMID: 7569999.

[5] A. D. Mirzabekov. DNA sequencing by hybridization — a megasequencing method and a diagnostic tool? *Trends in Biotechnology*, 12(1):27–32, **January 1994**.

[6] A. A. Ewis, Z. Zhelev, R. Bakalova, S. Fukuoka, Y. Shinohara, M. Ishikawa, and Y. Baba. A history of microarrays in biomedicine. *Expert Review of Molecular Diagnostics*, 5(3):315–328, **May 2005**. PMID: 15934810.

[7] J. K. Peeters and P. J. Van der Spek. Growing applications and advancements in microarray technology and analysis tools. *Cell Biochemistry and Biophysics*, 43(1):149–166, **2005**. PMID: 16043891.

[8] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. Expression profiling using cDNA microarrays. *Nature genetics*, 21(1 Suppl):10–14, **January 1999**. PMID: 9915494.

[9] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, 29(4):365–371, **December 2001**. PMID: 11726920.

[10] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, **November 2012**.

[11] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. Pedro Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Research*, 41(D1):D987–D990, **November 2012**.

[12] J. Hubble, J. Demeter, H. Jin, M. Mao, M. Nitzberg, T. B. K. Reddy, F. Wymore, Z. K. Zachariah, G. Sherlock, and C. A. Ball. Implementation of GenePattern within the stanford microarray database. *Nucleic acids research*, 37(Database issue):D898–901, **January 2009**. PMID: 18953035.

[13] C. M. L. S. Bouton and J. Pevsner. DRAGON and DRAGON view: information annotation and visualization tools for large-scale expression data. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 7:Unit 7.4, **August 2003**. PMID: 18428707.

[14] C. M. L. S. Bouton and J. Pevsner. DRAGON view: information visualization for annotated microarray data. *Bioinformatics (Oxford, England)*, 18(2):323–324, **February 2002**. PMID: 11847082.

[15] C. M. Bouton and J. Pevsner. DRAGON: database referencing of array genes online. *Bioinformatics (Oxford, England)*, 16(11):1038–1039, **November 2000**. PMID: 11159315.

[16] J. Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32 Suppl:496–501, **December 2002**. PMID: 12454644.

[17] N. Dean and A. E. Raftery. Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC bioinformatics*, 6:173, **2005**. PMID: 16011807.

[18] J. M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Human molecular genetics*, 6(10):1735–1744, **1997**. PMID: 9300666.

[19] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of computational biology: a journal of computational molecular cell biology*, 7(6):819–837, **2000**. PMID: 11382364.

[20] M. K. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical research*, 77(2):123–128, **April 2001**. PMID: 11355567.

[21] M. K. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical research*, 89(5-6):509–514, **December 2007**. PMID: 18976541.

[22] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, **1979**.

[23] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. Wiley, **2011**.

[24] R development core team (2008). r: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. available at http://www.R-project.org.

[25] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, **2004**. PMID: 15461798.

[26] J.-S. Lee, I.-S. Chu, J. Heo, D. F. Calvisi, Z. Sun, T. Roskams, A. Durnez, A. J. Demetris, and S. S. Thorgeirsson. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*, 40(3):667–676, **2004**.

[27] X. Jiao, B. T. Sherman, D. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806, **July 2012**. PMID: 22543366 PMCID: PMC3381967.

[28] T. Maass, I. Sfakianakis, F. Staib, M. Krupp, P. R. Galle, and A. Teufel. Microarray-based gene expression analysis of hepatocellular carcinoma. *Current Genomics*, 11(4):261–268, **June 2010**.

[29] Data were analyzed through the use of ingenuity pathways analysis (ingenuity® systems, www.ingenuity.com).

[30] G. Castello, S. Scala, G. Palmieri, S. A. Curley, and F. Izzo. HCV-related hepatocellular carcinoma: From chronic inflammation to cancer. *Clinical Immunology*, 134(3):237–250, **March 2010**.

[31] S. Toffanin, S. L. Friedman, and J. M. Llovet. Obesity, inflammatory signaling, and hepatocellular Carcinoma—An enlarging link. *Cancer Cell*, 17(2):115–117, **February 2010**.

[32] A. Budhu, M. Forgues, Q.-H. Ye, H.-L. Jia, P. He, K. A. Zanetti, U. S. Kammula, Y. Chen, L.-X. Qin, Z.-Y. Tang, and X. W. Wang. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell*, 10(2):99–111, **August 2006**.

[33] J.-M. Wu, N. J. Skill, and M. A. Maluccio. Evidence of aberrant lipid metabolism in hepatitis c and hepatocellular carcinoma. *HPB*, 12(9):625–636, **2010**.

[34] D. Becker, I. Sfakianakis, M. Krupp, F. Staib, A. Gerhold-Ay, A. Victor, H. Binder, M. Blettner, T. Maass, S. Thorgeirsson, P. R. Galle, and A. Teufel. Genetic signatures shared in embryonic liver development and liver cancer define prognostically relevant subgroups in HCC. *Molecular Cancer*, 11(1):55, **August 2012**. PMID: 22891627.

[35] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)*, 20(3):307–315, **February 2004**. PMID: 14960456.

[36] M. Carlson, S. Falcon, H. Pages, and N. Li. hgu133plus2.db: Affymetrix human genome u133 plus 2.0 array annotation data (chip hgu133plus2).

[37] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2):242–253, **April 2010**. PMID: 20097884.

[38] M. N. McCall, K. Uppal, H. A. Jaffee, M. J. Zilliox, and R. A. Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(Database issue):D1011–D1015, **January 2011**. PMID: 21177656 PMCID: PMC3013751.

# 4 Next Generation Sequencing - The biology tool towards Personalized Medicine

## 4.1 Overview

The Next Generation Sequencing (NGS) technology has revolutionized biomedical research. Its impact can be compared to the impact of the polymerase chain reaction (PCR) introduced in 1971 by Kleppe et al. [1]. NGS technology has set the important target to DNA sequencing, which is the determination of the orders of nucleotides in a DNA molecule. In 1977, the first method to sequence DNA was presented by Frederick Sanger and colleagues [2, 3]. He has therefore received the Nobel Prize for chemistry together with Walter Gilbert and Paul Berg in 1980. Since this time, DNA sequencing have become integral parts of biotechnology research and discovery, diagnostics, and forensic. Sanger's method utilizes dideoxunucleotides triphosphates to terminate DNA chain elongation followed by separation of the molecules by gel electrophoresis and detection of fluorescence labeled terminators. With this method, single sample sequences may be interrogated. A series of improvements were introduced to this method over time and lead to an increase in efficiency from 10Kb per 4h run in the late 1980 (slab gel sequencers) over 50Kb per 1h in 1990 (capillary sequencers, first generation NGS device), 20MB per 7h in 2005 (massive parallel pyrosequencing, second generation NGS device) and 1GB per 5d in 2007 (sequencing by synthesis, second generation NGS device) to 100GB per 5d in 2010 (single molecule sequencing, second generation NGS device). As a comparison, a single Illumina (http://www.illumina.com/) GA-II NGS device is as efficient as over 200 Applied Biosystems (http://www.appliedbiosystems.com) 3730xl traditional gel sequencing devices. This progress is acquired through the application of micro- and nanotechnology, which enables massively parallel sequencing reactions. Meanwhile, the second generation of NGS devices shape DNA sequencing, but third generation is being on the rise where the human genome can completely be sequenced within 15 minutes. Today, modern sequencing devices were applied in over 20 areas. The common areas are: DNA-Seq, ChIP-Seq and RNA-Seq. DNA-Seq is mainly used for de novo assembly of genomes; ChIP-Seq (chromatin immunoprecipitation) for analysis of DNA interactions with transcription factors, histone modifications, and chromatin binding proteins; and RNA-Seq for gene expression studies, miRNA analysis, non-coding RNA investigations, discovering splice variants, single nucleotide polymorphisms (SNPs), and RNA editing sites. The NGS analyses of this thesis are dedicated to application of RNA-Seq.

The following subchapter will give a brief overview about the NGS technolo-

gies, introduce NGS analysis with the focus on RNA-Seq and discusses databases for storing NGS data. Moreover, the developed workbench RNA-Seq Atlas is highlighted and the discovery of a healthy liver signature by using RNA-Seq Atlas is described.

A copy of the original publication of RNA-Seq Atlas, published in the journal *Bioinformatics* 2012, is listed in Chapter 9.

## 4.2 NGS technology - The evolution of transcriptomics

The NGS technology is challenging microarrays as the state-of-the-art method to measure the complex network of genes in order to create global images of living cells. Wwith focus on RNA-Seq analysis, the non-dependence of references in NGS enables the detection of novel splicing variances and novel transcripts. In addition, NGS has a higher resolution than whole genome microarrays and the same experimental protocol can be applied to various purposes, whereas specialized microarrays need to be designed. Moreover, the NGS technology has a high technical reproducibility. In comparison to the common principle of basepair hybridization being utilized over all microarrays by every producer, the NGS technology varies greatly between the manufactures (Tab. 4.1). Three different manufactures for massively parallel NGS production are in widespread use: Roche (http://454.com), Illumina (http://www.illumina.com/) and Applied Biosystem (http://www.lifetechnologies.com). Roche has become famous for its 454 GS FLX NGS platform, Illumina (previously Solexa) through its Genome Analyser (GA) and HiSeq 2000, and Applied Biosystem for its SOLiD.
A typical NGS gene expression profiling workflow is displayed in Fig. 4.1. The primary steps in the RNA-Seq workflow are platform independent but starts to differ during library construction. Similar to microarray analysis, the RNA-Seq protocol starts with the extraction of sample mRNA, followed by reverse transcription and PCR to amplify cDNA. Next, the cDNA is fragmented by DNase I treatment or sonication, which gives a more uniform coverage of each exon. This step is pursued by the library construction. A procedure in which manufacture specific adapters ligate to both ends of the fragments. Since this step the RNA-Seq protocol differ in dependence on manufactures. After the library construction, the clonal template amplification is applied. Illumina GA and HiSeq 2000 use bridge amplification, whereas Roche 454 GS FLX and Applied Biosystem SOLiD emulsion PCR and enrichment. Afterwards, the sequencing takes place. The Genome Analyser follows the approach of sequencing-by-synthesis, 454 GS FLX of pyrosequencing and SOLiD of ligation-based-sequencing. In dependence of fragments are sequenced from one direction only, the reads are defined as single end reads. In contrast, if fragments are sequenced from both, 5′ and 3′, ends the reads are called paired end reads.

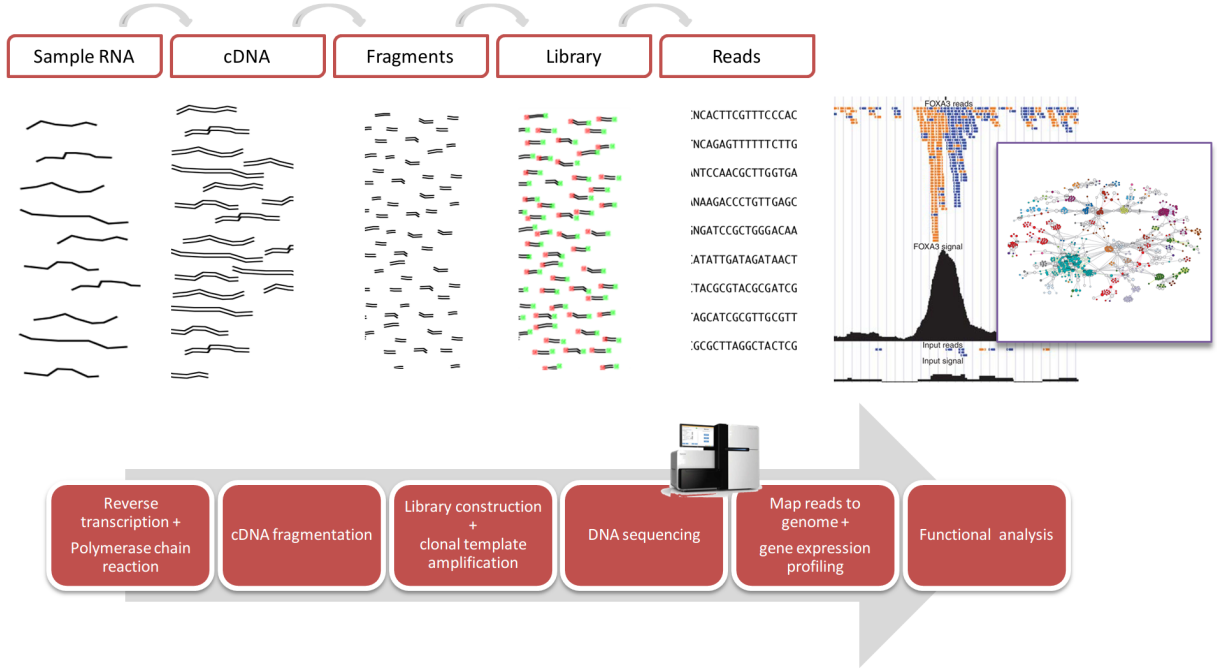


**Figure 4.1:** Principle diagram of the gene expression profiling by RNA-Seq technology.

The fundamental principles of **Illumina platforms**: DNA fragments are bound to the surface of a flow cell, DNA polymerase is initiated leading to the building of specific DNA fragment clusters. After the so called bridge amplification, complementary strands are removed and all four nucleotides are added to the flow cell simultaneously, along with DNA polymerase. The nucleotides carry a base-unique fluorescent label and the 3'-OH group is blocked such that each incorporation is a unique event. After each incorporation step an image is taken and the 3'-OH group is removed. This series of step continues up to a user defined setting (Fig. 4.2).
The basic idea of **Roche 454 GS FLX** is, that after binding DNA fragments to beads followed by emulsion PCR, light emitting nucleotides are bound to the fragments stepwise. The main difference in this system is the usage of a a specific nucleotide in each incorporation step instead of using a mixture of all nucleotides. This leads to the fact that if a homopolymer occur (stretch of same nucleotide) multiple complementary nucleotides bind to the fragment, emitting more light than a single base. For example an AAAAA fragment will bind TTTTT at a single step and the light intensity will be five times as high as a single T nucleotide. To get an idea about the average light intensity range of a single A, C, G or T, ACGT adaptor sequences are added at the start of each DNA fragment (library construction) to calibrate the analysis software for a single nucleotide. However, the saturation is achieved by $\leqslant 6$ nucleotides and thus homopolymers with a length $> 6$ cannot properly interpreted.

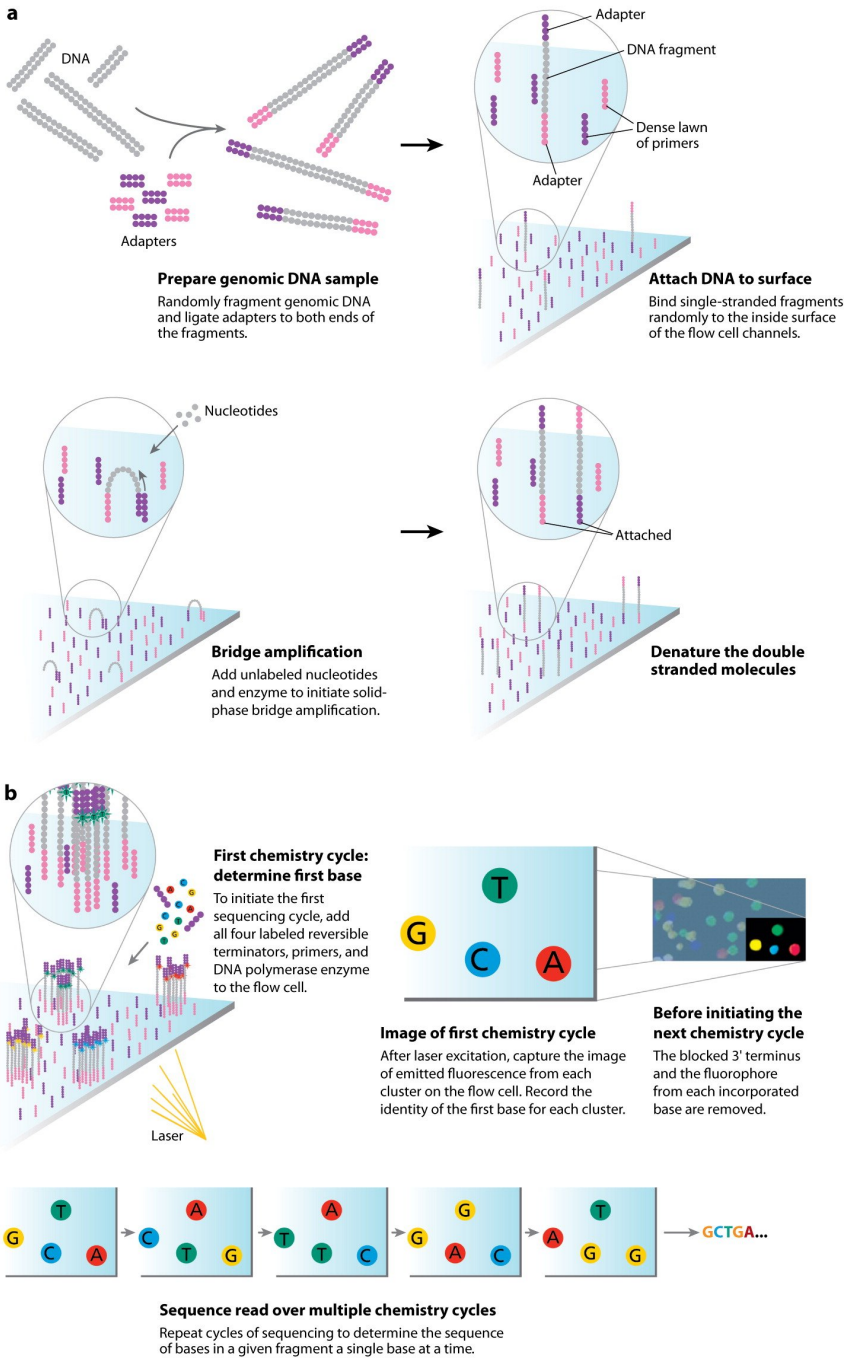


**Figure 4.2:** The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3′-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3′-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation.

Therefore the system is prone to base insertion and deletion errors, but very robust to base substitution errors (Fig. 4.3).
The **Applied Biosystem SOLiD** platform uses a similar bead technology as Roche 454 GS FLX to amplify the fragments for sequencing and uses a ligases-mediated step for sequencing. After the amplification process and assigning an adapter sequence to the 5'end (library construction), the beads were deposited on a flow cell slide. Unlike the other platforms, SOLiD uses shared adapter sequences on each amplified fragment and then DNA ligase is provided along with specific fluorescent-labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group. Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation (Fig. 4.4).
The resulting images, taken during each cycle and underlying each platform, consume several TByte of storage. In order to extract sequence information those images are decrypted by specific software supported by the manufactures.
Shaped by the differences in sequencing mechanisms each platform has its own throughput characteristics. The longest reads are produced by Roche 454 GS FLX with 700 bp. Whereas the best accuracy is achieved by the SOLiD systems through the application of shared adapter sequences. The highest throughput can be obtained with Illumina HiSeq 2000. These orders to 3 billion bases, which correspond to 600 GBytes data per run. A detailed comparison of the introduced NGS platforms and the traditional Sanger 3730xl sequencer is given in table 4.1, facts are acquired from Mardis et al. and Lia et al. [4, 5]:

Altogether, Roche 454, Illimina GA and HiSeq 2000 as well as Applied Biosystems SOLiD are the most common massively parallel sequencing platforms at current state. These technologies sequence hundreds of megabases to gigabases of nucleotide sequences reads on parallelized platforms in a single run. Hereby, the platforms differ in engineering configurations and sequencing chemistry, but share the mutual interest of massive parallel sequencing via spatially separated, clonally amplified DNA fragments.

## 4.3 RNA-Seq Databases

Next generation sequencing technologies have been rapidly applied in biomedical and biological research and numerous experiments have been made public available. However, not only does a single NGS run contains gigabytes of data (Tab. 4.1) but also no central database exits to deposit the data. Thus until now, most of the data are distributed over multiple places. One of the first major NGS projects enabling public access to their big data became known under the name 1000 Genome Project [6]. This consortium support sequence information of 697 individuals from seven populations and were initially analyzed to provide a deep characterization

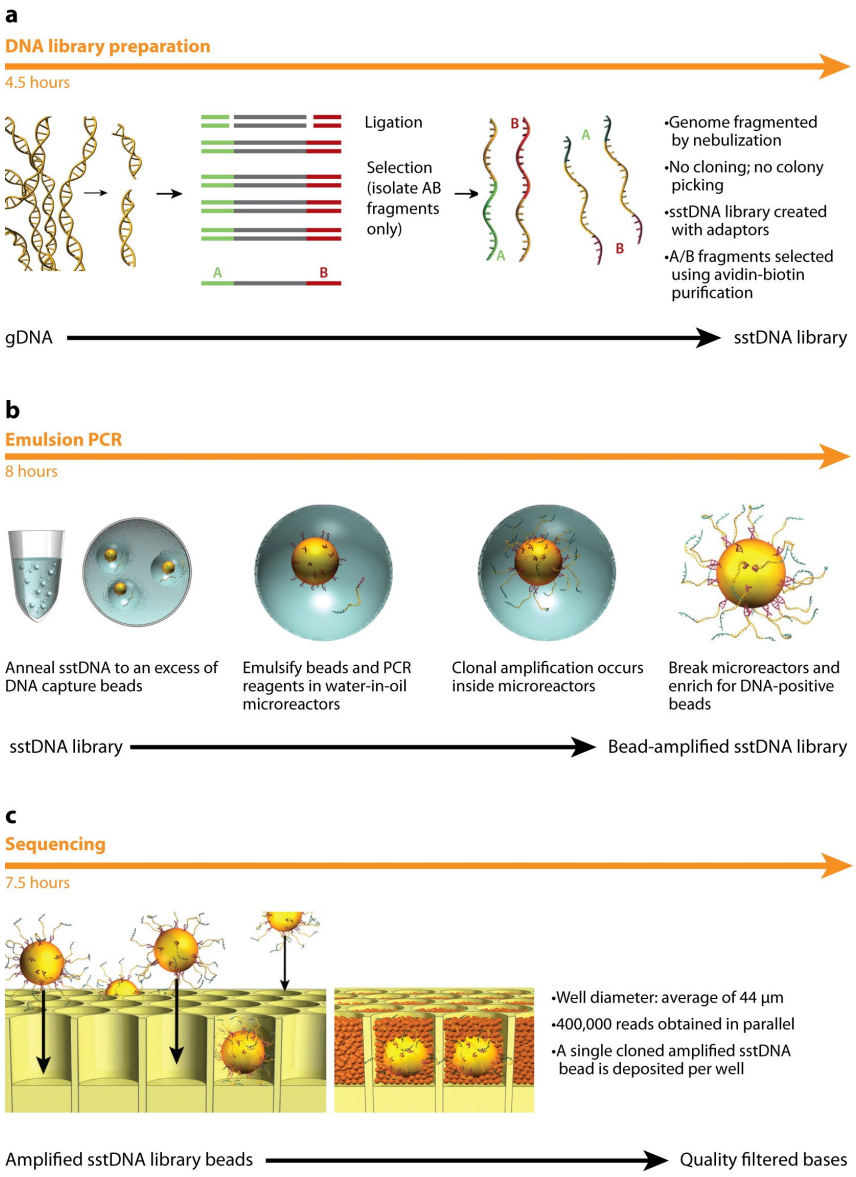


**Figure 4.3:** The method used by the Roche 454 GS FLX sequencer to amplify single-stranded DNA copies from a fragment library on agarose beads. A mixture of DNA fragments with agarose beads containing complementary oligonucleotides to the adapters at the fragment ends are mixed in an approximately 1:1 ratio. The mixture is encapsulated by vigorous vortexing into aqueous micelles that contain PCR reactants surrounded by oil, and pipetted into a 96-well microtiter plate for PCR amplification. The resulting beads are decorated with approximately 1 million copies of the original single-stranded fragment, which provides sufficient signal strength during the pyrosequencing reaction that follows to detect and record nucleotide incorporation events. The aberration sstDNA stands for single-stranded template DNA.

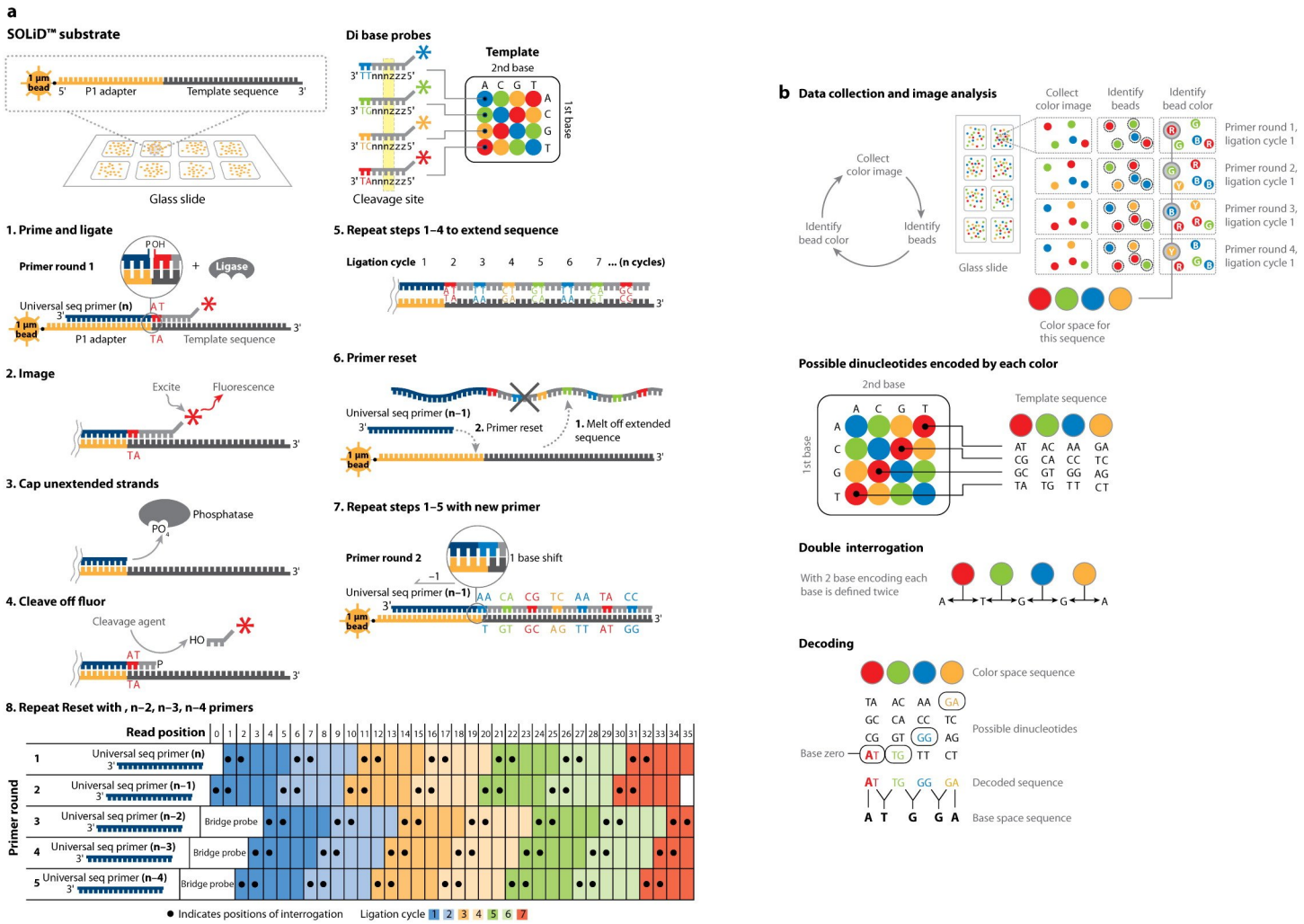


**Figure 4.4:** The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer. In a manner similar to Roche 454 GS FLX emulsion PCR amplification, DNA fragments for SOLiD sequencing are amplified on the surfaces of magnetic beads to provide sufficient signal during the sequencing reactions, and are then deposited onto a flow cell slide. Ligase-mediated sequencing begins by annealing a primer to the shared adapter sequences on each amplified fragment, and then DNA ligase is provided along with specific fluorescent-labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group. Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation. Principles of two-base encoding. Because each fluorescent group on a ligated 8mer identifies a two-base combination, the resulting sequence reads can be screened for base-calling errors versus true polymorphisms versus single base deletions by aligning the individual reads to a known high-quality reference sequence.

**Table 4.1:** Comparison of next generation sequencing technologies.

| | Roche 454 GS FLX | Illumina HiSeq 2000 | Applied Biosystem SOLiD | Sanger 3730xl |
|---|---|---|---|---|
| Sequencing chemistry | Sequencing-by-synthesis | Sequencing-by-synthesis | Sequencing-by-ligation | Dideoxy chain termination |
| DNA support | 25-30 μm bead | Flow cell surface | 1 μm bead | |
| Amplification approach | Emulsion PCR | Cluster amplification | Emulsion PCR | Polymerase chain reaction |
| Sequencing surface | High density well plate | 8-channel flow cell | Single slide imaged in panel | Polyacrylamide-urea gel |
| Read length | 250-700 bp | 35 to 100 bp | 35-50 bp | 400-900bp |
| Accuracy | 99.9% | 98% | 99.94% | 99.99% |
| Output data/run | 0.7Gb | 600Gb | 120Gb | 1.9-84Kb |
| Time/run | 24 hours | 3-10 days | 7-14 days | 20 mins- 3 hours |
| Advantages | Read length, fast | High throughput | Accuracy | High quality, long read length |
| Disadvantages | Error rate, high costs | Short read assembly | Short read assembly | High costs, low throughput |

of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Soon afterwards a second large project on NGS data was launched by the UCSC, the Encyclopedia of DNA Elements (ENCODE). Meanwhile, this encyclopedia supports over 2,000 individual experiments on human [7] and mouse [8]. The most versatile database for storing NGS data is currently the Sequence Read Archive (SRA) of the International Nucleotide Sequence Database Collaboration (INSDC) [9]. The awareness level of this archive has progressed enormously since journals impose authors to make their data publically available by storing them in SRA. SRA contains over 850 terabasepairs of biological sequence data, adding more than a terabase daily [10]. However, this database is just for depositing files and it does not come with a retrieval system. Therefore, the collaborators integrated the data into their own retrieval systems. The best known of which is the GEO [11]. But there is still the need for improvement, because actually NGS data can just be effectively identified within GEO by using their assigned platform identifier (e.g. querying GEO for with the platform identifier GPL9115, which is associated to Illumina Genome Analyzer II (Homo sapiens), will result in 3466 samples; GPL11154, Illumina, HiSeq 2000 (Homo sapiens), with 1695 samples; or GPL9186, Roche 454 GS FLX (Homo sapiens), with 32 samples).
As concluding mark to this section, the SRA is on the right path to become a central database for depositing NGS experiments, but the enormous amount of data produced by NGS experiments run put SRA recently in its place. Simply because they

do not have enough data storage. In addition many published NGS studies are still not being made available anyway, but should be.

## 4.4 In silico RNA-Seq analysis - Exploring terabytes of scientific data

Multiple steps in RNA-Seq analyses benefit from previously developed approaches applied in gene expression profiling with microarrays. These well considered approaches could be partially transferred directly into the analysis pipeline, some with slight modifications. Thus, RNA-Seq analysis can also be characterized in:

- Image processing
- Quality control
- Data preprocessing and normalization
- Identification of differentially expressed genes
- Functional analysis and biological interpretation

The **image processing** steps differ only fundamentally from the microarray step since a) intensity values of four signals are measured and b) instead of one image a huge variety of images are sequentially analyzed and superimposed to study sequence information. The resulting sequence can be examined in a chromatogram trace image (Fig. 4.5). Subsequent **quality control** assigns a quality score to each

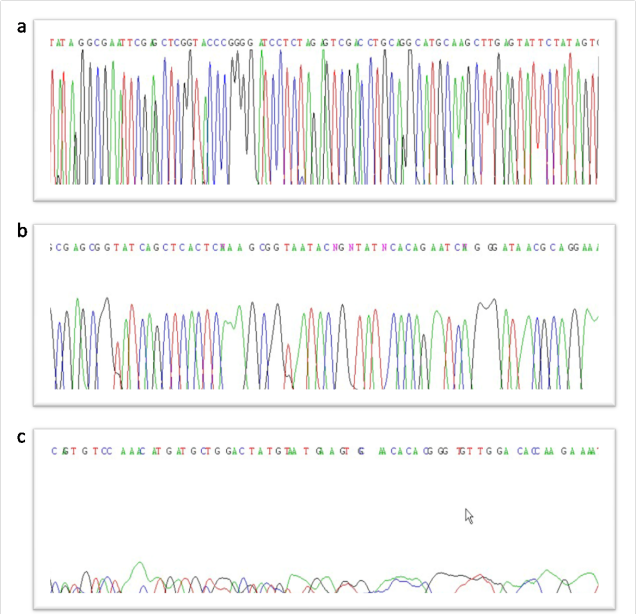


**Figure 4.5:** Next generation sequencing tracing: a) High quality region, no ambiguities, b) medium quality region, some ambiguities and c) poor quality region, low confidence.

base position and any read. This rating uses a modified phred algorithm [12, 13], which returns the probability of an incorrectly called base. Phred is based on Fourier analysis (decomposing the data into a series of sine waves) to examine chromatogram trace data (Fig. 4.6). The resulting phred scores are logarithmically related to the probability of an error. Whereas a score of 20 is generally considered the minimum acceptable score and represents an error rate of 1 in 100, with a corresponding call accuracy of 99%. Sequence and corresponding quality information

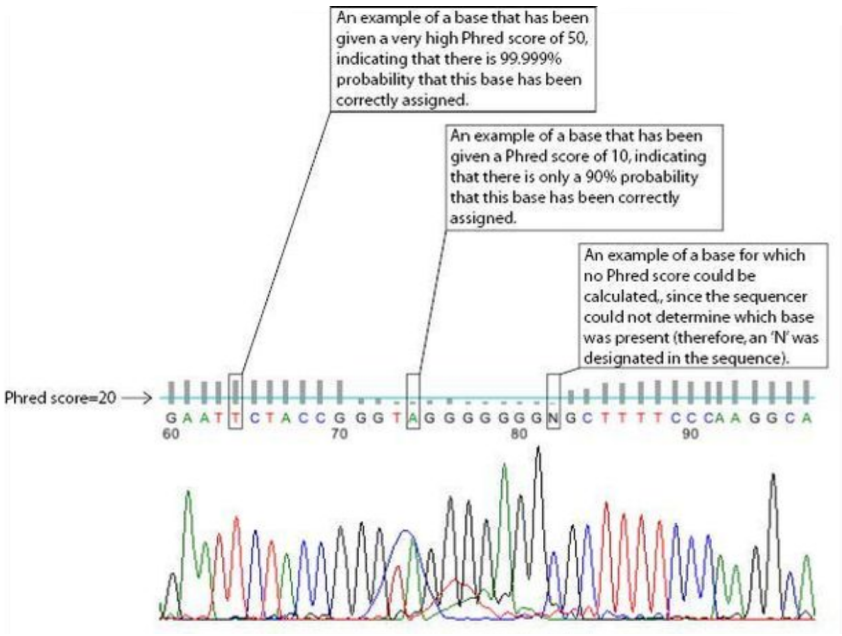


**Figure 4.6:** An example of DNA sequencing tracing and the phred score (grey bars) corresponding to each color peak. The colored peaks on the trace belongs to each DNA base. The light-blue horizontal line placed across the grey bars represents a phred score of 20 which is considered an acceptable level of accuracy.

are stored in a text-based file in the imposed FASTQ format. Whereas the quality score is encrypted in ASCII charters, character *!* represents the lowest quality while ~ is the highest. A FASTQ file normally consumes several GBytes storage and uses four lines per sequence:

- Line 1: Begins with character @ followed by a sequence identifier.
- Line 2: Contains sequence letters.
- Line 3: Begins with character + followed by sequence identifier and comments (optional).
- Line 4: Same length as sequence, each character encodes the quality of the base

A FASTQ example entry:

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3 +
4 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

In some cases, to save storage space, just the sequence information is stored. Those files are called FASTA file and contain only identifier and sequence information (line 1 and 2).
FASTQ or FASTA files are the starting point for **data preprocessing and normalization**. During preprocessing, the reads are mapped to the genome or transcriptome of the specific organisms from which samples are taken. This involves finding the place in the reference genome that each read matches to. The mechanism of mapping is also called alignment. Many novel algorithms were developed to solve this problem. Two of the most popular are BWA [14] and Bowtie [15]. To deal with the huge amount of input data, i.e. a FASTQ file and a reference genome, the Burrow-Wheeler Transform [16] is used in both methods. Typically, due to the high sequence similarity within members of the same sequence, around 80% of the reads can be aligned to the reference genome by the most aligners. However, it depends to a large extend on the configuration of the alignment algorithms. In general, alignment software can be configured to tolerate a specific number of alignment mismatches and to account for sequencing quality scores. A typical criterion is 1 to 2 mismatches for 36 bp reads with a quality score of at least 20. Of course, sequencing errors, multiple matches as well as deviations from the reference genome (SNPs, insertions, etc.) and problems with the aligner can also raise substantial dilemmas. Usually, the percentage amount of mapped reads is a good measure of data quality. The alignment is stored in the flexible Sequence Alignment/Map (SAM) format [17], a binary version (BAM) is also available allowing more efficient storage. Those files store the exact position and characteristic of each read within the reference genome. Among other things, these characteristics include the exact start position, the number and the position of the mismatches, and the information about how often a read could be mapped. Normalization in RNA-Seq analysis concerns a) different numbers of reads of a sample, and b) various sequence lengths of genes. This step merges together with the identification of differentially expressed genes.
Current state-of-the-art method regarding the **identification of differentially expressed genes** is analysis of those reads, which can be mapped uniquely to the genome. Other methods include multireads, i. e. a read that could be assigned to multiple positions on the reference genome. Those methods deal with multireads either by selecting just the read with the highest mapping score or by reporting all multireads. But in addition, new methods are established that use probabilistic models to assign fractions of reads to multiple positions or estimate the position with the highest alignment probability [18, 19]. Subsequent to the alignment, the abundance is estimated by counting reads. In more detail, (a) estimate the probabil-

ity of reads being encoded from a given transcript by counting the number of reads that align to that region and (b) normalize for transcript length:

$$(a)\ \theta_i = \frac{c_i}{N}$$

$$(b)\ \tau_i \propto \frac{\theta_i}{l_i'}$$

Whereas, $\theta_i$ is the probability of transcript $i$ to be transcribed, $c_i$ the number of reads mapped to transcript i and N the total number of mappable reads. $\tau_i$ refers to the expression value of transcript $i$ normalized to $l_i'$, the length of transcript $i$. Next these values are summarized to RPKM values for each transcript, which is designated to reads per kilobase of exon model per million mapped reads [20]:

$$RPKM_i = 1 * 10^9 * \frac{c_i}{N l_i'}$$

RPKM is slightly analogous to FPKM (fragments per kilobase of exon model per million mapped reads) [21] and is calculated using single end reads and the later is using paired end reads. A novel method to estimate expression levels is TPM (transcripts per million mapped reads) [22], which includes a normalization factor $Z$ for all expressed transcripts:

$$TPM_i = 1 * 10^6 * Z * \frac{c_i}{N l_i'}$$

with

$$Z = \sum_i \tau_i l_i'$$

The resulting normalized data is reported as reads (or transcripts) per million and should be preferred over RPKM/FPKM because of the normalization by factor $Z$ [22]. After estimating the expression values for each sample, statistically significant expressed genes can be calculated by applying univariate or multivariate tests (e.g. t-test or ANOVA). These tests as well as the follow-up adjustment for p-values are already discussed in microarray analysis; please refer to section 3.4 for more information.

Already mentioned above, **functional analysis and biological interpretation** are purely application dependent stage. According to a specific research question further algorithms have to be applied.

Despite the novelty of RNA-Seq analysis, some of the methods discussed above are already integrated in the software suite bioconductor [23] and can directly be applied to the data. However, many methods need to be improved or even have to be newly developed, depending on the outstanding issue that has to be solved. In NGS, the most challenging ones are 1) the mapping of short reads that makes mapping challenging, 2) GC and amplification bias, 3) sequencing errors, not all sequences are equally likely to be sequenced. But the biggest problem arises from 4) repetitive regions and much effort need to be invested to solve this dilemma.

In conclusion, several steps in RNA-Seq analysis were already solved during microarray studies and can be applied to those data. For the other steps sophisticated methods have been developed. However, there is still a need for improvements, in particular, the smart way of handling repetitive regions.

## 4.5 Publication 5 - The workbench: RNA-Seq Atlas - A reference database for gene expression profiling in normal tissue by next generation sequencing

Over the next years, the availability of next generation sequencing data will offer an entirely new perspective for clinical research and will speed up personalized medicine. So far, several databases offer storage space or downloads of NGS data [24, 25]. However, not only does a single NGS run contains gigabytes of data but the data analysis is also not feasible for researchers who are not familiar with computer science. Therefore, the data must be seen as not accessible for those researchers. To this end, RNA-Seq Atlas was implemented (http://medicalgenomics.org/rna_seq_atlas), which is an easily accessible database and retrieval system, offering access to evaluated NGS gene expression profiles. This study was published in the journal *Bioinformatics* in 2012, a copy of the original publication can be found in Chapter 9.

RNA-Seq Atlas originates from RNA-Seq data on eleven, healthy, human tissue samples pooled from multiple donors spanning 32,384 specific transcripts corresponding to 21,399 unique genes. The tissues include adipose, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes. Total RNA from the reference tissues were purchased from Ambion (Austin, USA) and represent a pool of RNA from multiple donors. Libraries were prepared as described in Armour et al. in 2009 [26], including both poly[A]+ and poly[A]- fractions. Sequencing was performed on the Illumina GA-II sequencer. An average of 50 million reads per tissue was generated, with sequence reads of 36 nucleotides (nt) or 50 nt depending on the tissue. After trimming reads to a common length of 28 nt to avoid aligning sequences of amplified primers, the obtained reads were aligned to the human hg18 genome assembly using BWA [14]. For mRNAs RefSeq transcript coordinates and associated gene symbols were downloaded from the University of California, Santa Cruz (UCSC) genome browser. Only the reads mapping to a single gene were used. Next, reads overlapping each transcript in the correct genomic orientation were determined. The expression levels were estimated by mapping and counting reads to single gene sequences derived from the UCSC genome browser followed by normalization to RPKM values. For this analysis, ncRNAs, pseudogenes, miRNAs, tRNAs, and rRNAs were removed. In addition, RNA-Seq data were linked to several microarray gene profiles, including BioGPS (http://biogps.org/) normal tissue profiles and NCI60 (http://discover.nci.nih.gov/cellminer/home.do) cancer cell line expression data

to enable an integrative detailed comparison between RNA-Seq and microarray expression profiles. Further, the RNA-Seq Atlas was linked to commonly used and established bioinformatics databases and knowledge repositories. Enabling access to deeper transcriptional information was achieved by linking the RNA-Seq Atlas data to the NCBI Nucleotide database [27]. Also, information on corresponding gene symbol, aliases, description, chromosomal location, Entrez ID as well as EnsEMBL ID were assembled from the NCBI Entrez and EnsEMBL databases [27, 28]. Additional outgoing links to HGNC [29], HPRD [30] , OMIM [31], BioGPS [32], Nextbio [33] and GENT [34] were supported. Moreover, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [35] was accessed to identify gene signaling as well as molecular pathway affiliations; and data on cellular component, biological process and molecular function were collected from the Gene Ontology database [36]. Finally, a retrieval system was implemented (Fig. 9.1, 9.2, 9.3, 9.4) and the RNA-Seq Atlas data were cross-linked to the data already present at medicalgenomics.org.

To conclude, RNA-Seq Atlas is the first database providing data mining tools and open access to large scale RNA-Seq expression profiles. Its applications are versatile, as it will be beneficial in identifying tissue specific genes and expression profiles, comparison of gene expression profiles among diverse tissues, but also systems biology approaches linking tissue function to gene expression changes.

## 4.6 The discovery: Identification of liver specific genes using next generation sequencing technology

The next generation sequencing technology offers an entirely new perspective for clinical research and will speed up personalized medicine. In this study, this new technology was used to identify a gene profile highly specific for healthy liver tissues, which could deal as a diagnostic tool to distinguish between a healthy or diseased liver. This study was introduced at the annual convention of *International Society for Computational Biology*, *European Association for the Study of Liver Diseases*, *American Association for the Study of Liver Diseases* and *German Association for the Study of Liver Diseases* in 2011 and 2012; it was not published in any scientific journal.

Initially, next generation sequencing data were accessed from RNA-Seq Atlas. This genome-wide expression compendium originates from multiple, healthy, human tissue samples pooled from multiple donors. The expression levels were estimated by mapping and counting reads to single gene sequences derived from the UCSC genome browser. To make the expression levels comparable across tissues, the RPKM method was applied.
Identification of liver specific genes was achieved by the help of a developed computer program implemented in R [37]. In the first place, the R program subdivided the RNA-Seq data into liver and non liver data (reference set), and cutoff values

were defined for the RPKM values of the transcripts (liver $\geqslant$ 10 RPKM; reference set $\leqslant$ 2 RPKM). Afterwards, the program successively and consistently tested each transcript for the redefined criteria and if a transcript passed the criteria the transcript was flagged as liver specific. However, special attention was dedicated to gene-transcripts affiliations, because a gene may be transcribed by several transcripts. Therefore, an additional filter step was applied to the set of liver specific transcript to identify genes exclusively regulated in the liver. The first step in the filtering process was the assignment of gene names to the transcripts. In the second step, each gene-transcript relation was evaluated. The evaluation stated a gene as liver specific, if all transcripts of a specific gene were allocated in the subset of liver specific transcripts. Detailed insights into the program are provided in Figure 4.7.

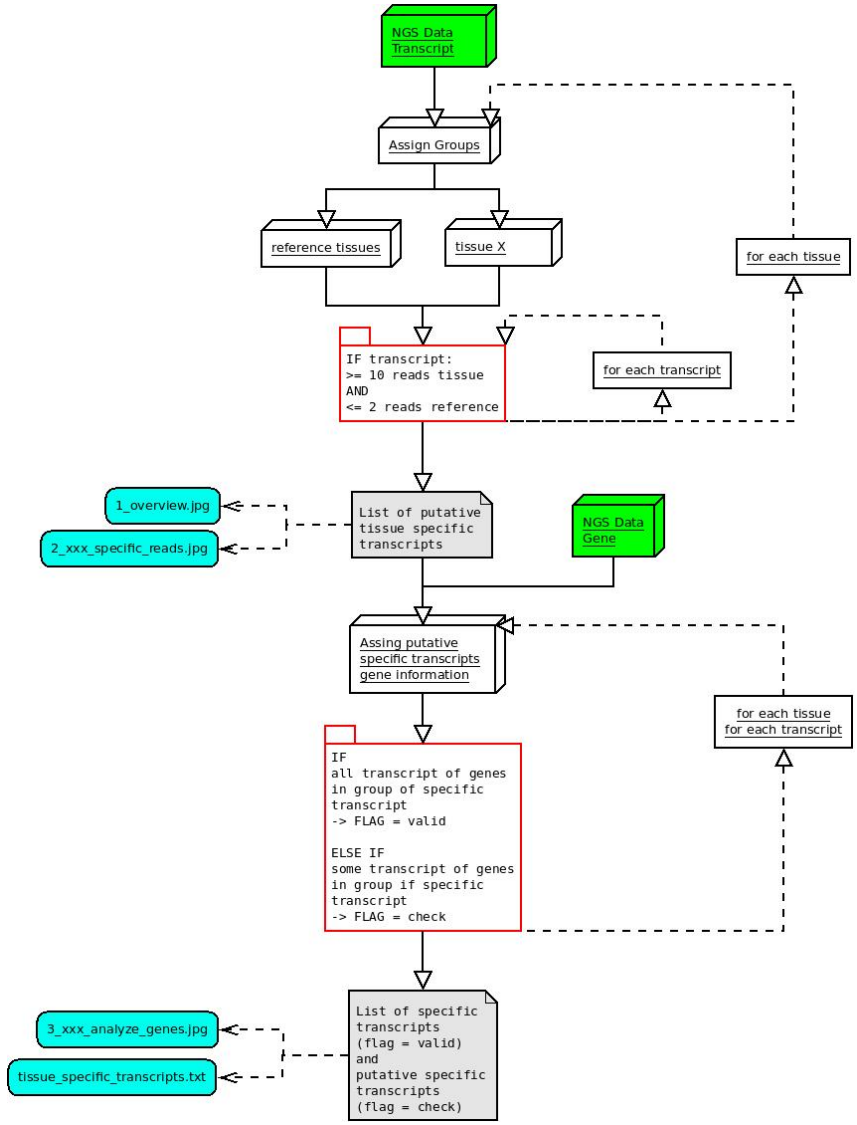


**Figure 4.7:** Principle diagram of the identification of liver specific genes.

The evaluation results in 3,164 liver transcripts that have a RPKM of at least 10; a subset of 127 out of them shows less or equal 2 RPKM within the reference set (Fig. 4.8). As a consequence of the second filtering step to the subset, a profile of 98 genes exclusively expressed by the liver were identified. With 518 RPKM, the gene complement component 9 (C9) ranks on top of this profile. Followed by orosomucoid 1 (ORM1), tyrosine aminotransferase (TAT), hemopexin (HPX), carbamoyl-phosphate synthase 1 (CPS1) and inter-alpha-trypsin inhibitor heavy chain 2 (ITIH2) with 395, 205, 249, 311 and 171 RPKM, respectively (Fig. 4.9). For the highest ranking gene C9 it is known that it encodes the final component of the complement system. Changes and defective regulation of this system was recently reported to be involved in the progression of liver fibrosis in children with chronic hepatitis C [38] and also described to be associated with metabolic disorders in the liver [39]. ORM1 transcribes an acute phase plasma protein and serves as biomarker for obesity [40]. Moreover, ORM1 is also an important player in the protein interaction network of the human liver [41]. TAT catalyzes the conversion of L-tyrosine into p-hydroxyphenylpyruvate and mutations in this gene are stated to contribute to the pathogenesis of hepatocellular carcinoma [42, 43]. Dysregulation in HPX, CPS1 and ITIH2 are also associated to liver disease and serious liver dysfunctions [44, 45, 46]. Further reaching investigation on pathway and functional level was accomplished with IPA [47]. Those analyses revealed significant associations to lipid metabolism, molecular transport, coagulation systems as well as acute phase response signaling (Fig. 4.10, 4.11). In the literature lipid metabolism is stated to be essential for controlling basic liver activities and disorders in this metabolisms is often reported with occurrence and progression of diabetes, cancer, obesity and hepatic steatosis [48]. Moreover, dysregulation in molecular transport, coagulation systems and acute phase response is also reported in a variety of liver diseases and critical liver dysfunctions [49, 50, 51].

To conclude, comparative analysis of multiple tissues by RNA-Seq data has revealed a signature of 98 genes exclusively expressed in normal liver tissues. For this profile it was shown that changes and defective regulation of its genes as well as their corresponding biological functions are closely connected to liver diseases and critical liver dysfunctions. To this end, the predicted liver specific profile may function as a diagnostic tool to distinguish between a healthy or diseased liver. However, to confirm the prediction further wet-lab analysis must be carried out.

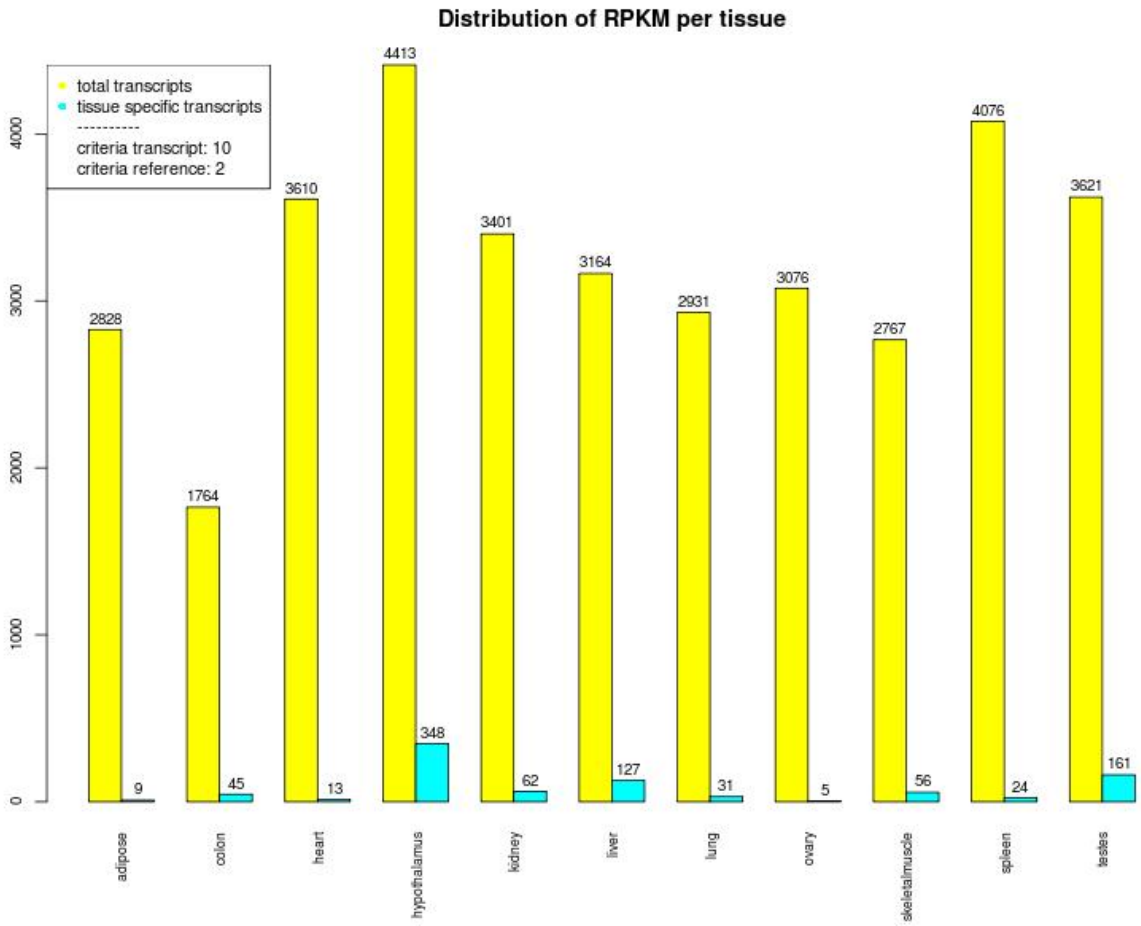


**Figure 4.8:** Distribution of RPKM values per tissues. Yellow bars indicate the total number of transcripts of a specific tissue, blue bars the exclusively expressed subset of the transcripts.

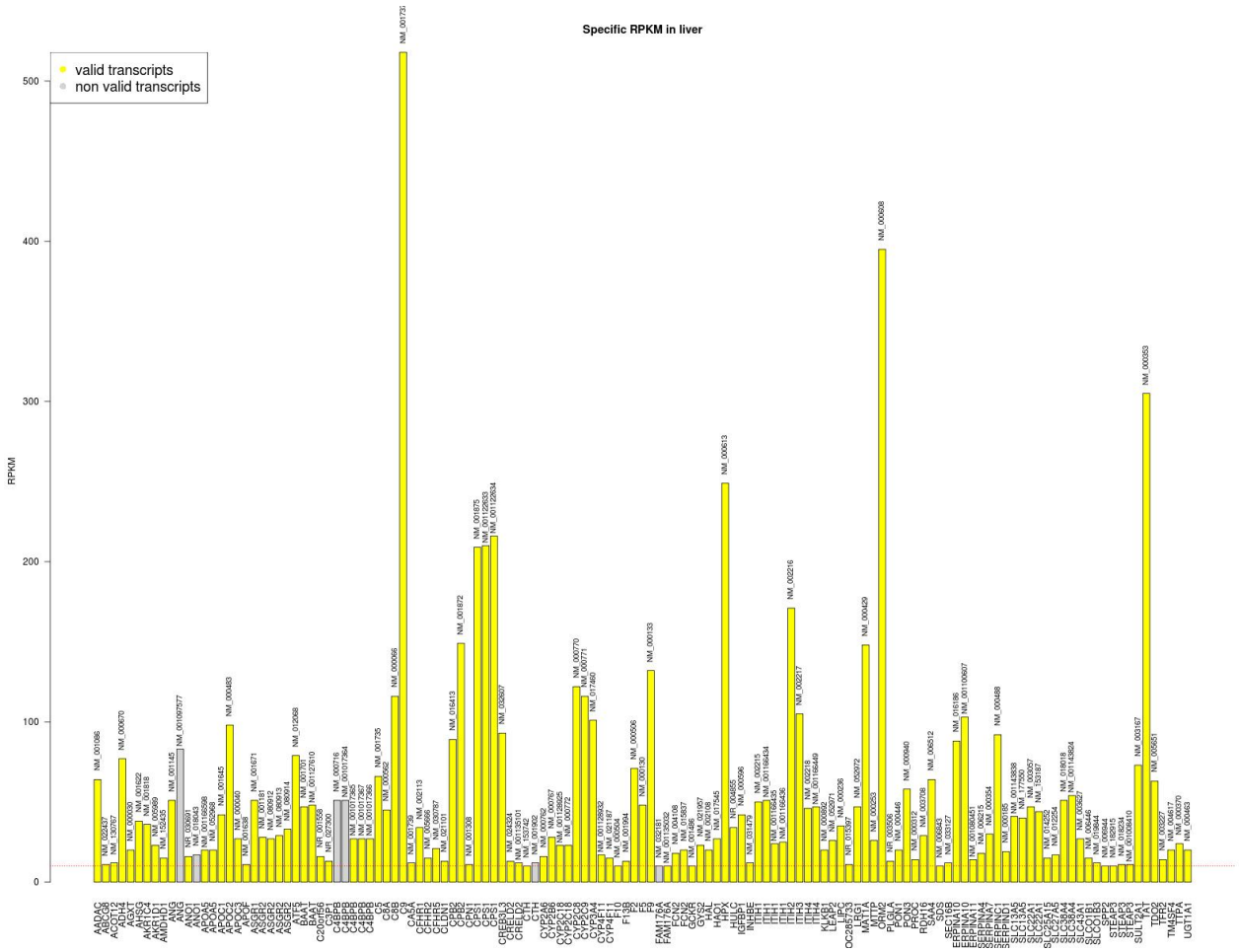


**Figure 4.9:** RPKM values of liver specific transcripts.

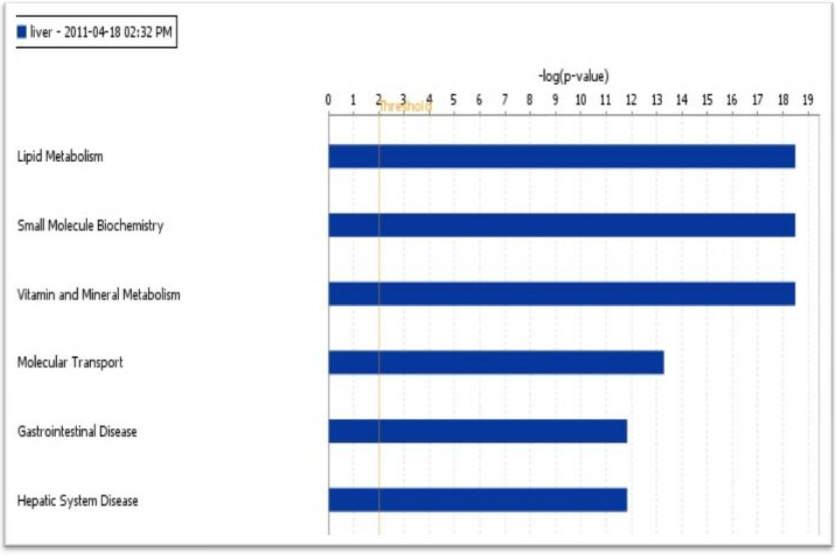


**Figure 4.10:** IPA investigation of the liver specific gene profile on functional level.

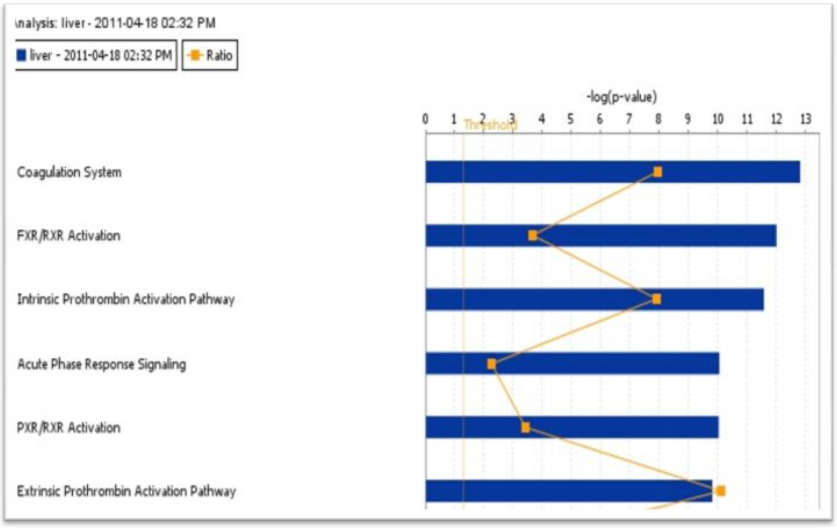


**Figure 4.11:** IPA investigation of the liver specific gene profile on pathway level.

# BIBLIOGRAPHY


[1] K. Kleppe, E. Ohtsuka, R. Kleppe, I. Molineux, and H. G. Khorana. Studies on polynucleotides. XCVI. repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of molecular biology*, 56(2):341–361, **March 1971**. PMID: 4927950.

[2] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi x174 DNA. *Nature*, 265(5596):687–695, **February 1977**. PMID: 870828.

[3] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, **February 1977**. PMID: 265521.

[4] E. R. Mardis. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9:387–402, **2008**. PMID: 18576944.

[5] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012, **July 2012**.

[6] 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, **October 2010**. PMID: 20981092.

[7] ENCODE Project Consortium, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, **September 2012**. PMID: 22955616.

[8] Mouse ENCODE Consortium, J. A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, J. Dekker, G. E. Crawford, C. A. Keller, W. Wu, C. Morrissey, S. A. Kumar, T. Mishra, D. Jain, M. Byrska-Bishop, D. Blankenberg, B. R. Lajoie1, G. Jain, A. Sanyal, K.-B. Chen, O. Denas, J. Taylor, G. A. Blobel, M. J. Weiss, M. Pimkin, W. Deng, G. K. Marinov, B. A. Williams, K. I. Fisher-Aylor, G. Desalvo, A. Kiralusha, D. Trout, H. Amrhein, A. Mortazavi, L. Edsall, D. McCleary, S. Kuan, Y. Shen, F. Yue, Z. Ye, C. A. Davis, C. Zaleski, S. Jha, C. Xue, A. Dobin, W. Lin, M. Fastuca, H.

Wang, R. Guigo, S. Djebali, J. Lagarde, T. Ryba, T. Sasaki, V. S. Malladi, M. S. Cline, V. M. Kirkup, K. Learned, K. R. Rosenbloom, W. J. Kent, E. A. Feingold, P. J. Good, M. Pazin, R. F. Lowdon, and L. B. Adams. An encyclopedia of mouse DNA elements (mouse ENCODE). *Genome biology*, 13(8):418, **August 2012**. PMID: 22889292.

[9] R. Leinonen, H. Sugawara, and M. Shumway. The sequence read archive. *Nucleic Acids Research*, 39(Database issue):D19–D21, **January 2011**. PMID: 21062823 PMCID: PMC3013647.

[10] Y. Kodama, M. Shumway, R. Leinonen, and on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1):D54–D56, **October 2011**.

[11] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, **November 2012**.

[12] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome research*, 8(3):175–185, **March 1998**. PMID: 9521921.

[13] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome research*, 8(3):186–194, **March 1998**. PMID: 9521922.

[14] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, **July 2009**. PMID: 19451168.

[15] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, **March 2009**. PMID: 19261174.

[16] B. M and W. DJ. A block sorting lossless data compression algorithm. *Technical Report 124. Palo Alto, CA: Digital Equipment Corporation;*, **1994**.

[17] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, **August 2009**. PMID: 19505943.

[18] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26(4):493–500, **February 2010**. PMID: 20022975.

[19] S.-K. Lou, J.-W. Li, H. Qin, A. K. Yim, L.-Y. Lo, B. Ni, K.-S. Leung, S. K. Tsui, and T.-F. Chan. Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length. *BMC Bioinformatics*, 12(Suppl 5):S2, **July 2011**. PMID: 21988959.

[20] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, **July 2008**.

[21] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, **May 2010**. PMID: 20436464.

[22] G. P. Wagner, K. Kin, and V. J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, **December 2012**.

[23] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, **2004**. PMID: 15461798.

[24] D. Altshuler. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, **October 2010**. PMID: 20981092.

[25] M. Shumway, G. Cochrane, and H. Sugawara. Archiving next generation sequencing data. 38(Database issue):D870–D871, **January 2010**. PMID: 19965774 PMCID: 2808927.

[26] C. D. Armour, J. C. Castle, R. Chen, T. Babak, P. Loerch, S. Jackson, J. K. Shah, J. Dey, C. A. Rohl, J. M. Johnson, and C. K. Raymond. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods*, 6(9):647–649, **September 2009**. PMID: 19668204.

[27] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(Database issue):D38–51, **January 2011**. PMID: 21097890.

[28] G. Spudich, X. M. Fernández-Suárez, and E. Birney. Genome browsing with ensembl: a practical overview. *Briefings in Functional Genomics & Proteomics*, 6(3):202–219, **September 2007**. PMID: 17967807.

[29] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, and E. A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, 39(Database issue):D514–519, **January 2011**. PMID: 20929869.

[30] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–772, **January 2009**. PMID: 18988627.

[31] Online mendelian inheritance in man, OMIM®. McKusick-Nathans institute of genetic medicine, johns hopkins university (baltimore, MD).

[32] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, r. Huss, Jon W, and A. I. Su. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10(11):R130, **2009**. PMID: 19919682.

[33] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. 5(9). PMID: 20927376 PMCID: 2947508.

[34] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim. GENT: gene expression database of normal and tumor tissues. *Cancer Informatics*, 10:149–157, **2011**. PMID: 21695066.

[35] K. F. Aoki and M. Kanehisa. Using the KEGG database resource. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, Chapter 1:Unit 1.12, **October 2005**. PMID: 18428742.

[36] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, **May 2000**. PMID: 10802651.

[37] R development core team (2008). r: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. available at http://www.R-project.org.

[38] B. E. Behairy, G. M. El-Mashad, R. S. Abd-Elghany, E. M. Ghoneim, and M. M. Sira. Serum complement c4a and its relation to liver fibrosis in children with chronic hepatitis c. *World journal of hepatology*, 5(8):445–451, **August 2013**. PMID: 24023984.

[39] J. Phieler, R. Garcia-Martin, J. D. Lambris, and T. Chavakis. The role of the complement system in metabolic organs and metabolic diseases. *Seminars in immunology*, 25(1):47–53, **February 2013**. PMID: 23684628.

[40] H. Rangé, C. Poitou, A. Boillot, C. Ciangura, S. Katsahian, J.-M. Lacorte, S. Czernichow, O. Meilhac, P. Bouchard, and C. Chaussain. Orosomucoid, a new biomarker in the association between obesity and periodontitis. *PloS one*, 8(3):e57645, **2013**. PMID: 23526947.

[41] J. Wang, K. Huo, L. Ma, L. Tang, D. Li, X. Huang, Y. Yuan, C. Li, W. Wang, W. Guan, H. Chen, C. Jin, J. Wei, W. Zhang, Y. Yang, Q. Liu, Y. Zhou, C. Zhang, Z. Wu, W. Xu, Y. Zhang, T. Liu, D. Yu, Y. Zhang, L. Chen, D. Zhu, X. Zhong, L. Kang, X. Gan, X. Yu, Q. Ma, J. Yan, L. Zhou, Z. Liu, Y. Zhu, T. Zhou, F. He, and X. Yang. Toward an understanding of the protein interaction network of the human liver. *Molecular systems biology*, 7:536, **2011**. PMID: 21988832.

[42] L. Fu, S.-S. Dong, Y.-W. Xie, L.-S. Tai, L. Chen, K. L. Kong, K. Man, D. Xie, Y. Li, Y. Cheng, Q. Tao, and X.-Y. Guan. Down-regulation of tyrosine aminotransferase at a frequently deleted region 16q22 contributes to the pathogenesis of hepatocellular carcinoma. *Hepatology (Baltimore, Md.)*, 51(5):1624–1634, **May 2010**. PMID: 20209601.

[43] K. K. Rehman, Q. Ayesha, A. A. Khan, N. Ahmed, and C. M. Habibullah. Tyrosine aminotransferase and gamma-glutamyl transferase activity in human fetal hepatocyte primary cultures under proliferative conditions. *Cell biochemistry and function*, 22(2):89–96, **April 2004**. PMID: 15027097.

[44] X. R. Xu, J. Huang, Z. G. Xu, B. Z. Qian, Z. D. Zhu, Q. Yan, T. Cai, X. Zhang, H. S. Xiao, J. Qu, F. Liu, Q. H. Huang, Z. H. Cheng, N. G. Li, J. J. Du, W. Hu, K. T. Shen, G. Lu, G. Fu, M. Zhong, S. H. Xu, W. Y. Gu, W. Huang, X. T. Zhao, G. X. Hu, J. R. Gu, Z. Chen, and Z. G. Han. Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15089–15094, **December 2001**. PMID: 11752456.

[45] A.-G. Wang, S. Y. Yoon, J.-H. Oh, Y.-J. Jeon, M. Kim, J.-M. Kim, S.-S. Byun, J. O. Yang, J. H. Kim, D.-G. Kim, Y.-I. Yeom, H.-S. Yoo, Y. S. Kim, and N.-S. Kim. Identification of intrahepatic cholangiocarcinoma related genes by comparison with normal liver tissues using expressed sequence tags. *Biochemical and biophysical research communications*, 345(3):1022–1032, **July 2006**. PMID: 16712791.

[46] H. Liu, H. Dong, K. Robertson, and C. Liu. DNA methylation suppresses expression of the urea cycle enzyme carbamoyl phosphate synthetase 1 (CPS1) in human hepatocellular carcinoma. *The American journal of pathology*, 178(2):652–661, **February 2011**. PMID: 21281797.

[47] Data were analyzed through the use of ingenuity pathways analysis (ingenuity® systems, www.ingenuity.com).

[48] P. G. Kopelman. Obesity as a medical problem. *Nature*, 404(6778):635–643, **April 2000**. PMID: 10766250.

[49] C. Duarte-Rey, D. Bogdanos, C.-Y. Yang, K. Roberts, P. S. C. Leung, J.-M. Anaya, H. J. Worman, and M. E. Gershwin. Primary biliary cirrhosis and the nuclear pore complex. *Autoimmunity reviews*, 11(12):898–902, **October 2012**. PMID: 22487189.

[50] E. Schaden, F. H. Saner, and K. Goerlinger. Coagulation pattern in critical liver dysfunction. *Current opinion in critical care*, 19(2):142–148, **April 2013**. PMID: 23400090.

[51] D. L. Thiele. Tumor necrosis factor, the acute phase response and the pathogenesis of alcoholic liver disease. *Hepatology (Baltimore, Md.)*, 9(3):497–499, **March 1989**. PMID: 2493417.

# 5 Publication 1 - Library of molecular associations: curating the complex molecular basis of liver diseases

Published in BMC Genomics in 2010

*BMC Genomics 11, 189 (2010)*

Stefan Buchkremer
Jasmin Hendel
Markus Krupp
Arndt Weinmann
Kai Schlamp
Thorsten Maass
Frank Staib
Peter R. Galle
Andreas Teufel

| Authors contribution: | |
|---|---|
| Draft writing | 20% |
| Review process | 30% |
| Data processing | 80% |
| Data analysis | 40% |
| Database design | 90% |
| GUI implementation | 90% |

## Abstract

**Background:** Systems biology approaches offer novel insights into the development of chronic liver diseases. Current genomic databases supporting systems biology analyses are mostly based on microarray data. Although these data often cover genome wide expression, the validity of single microarray experiments remains questionable. However, for systems biology approaches addressing the interactions of molecular networks comprehensive but also highly validated data are necessary.

**Results:** We have therefore generated the first comprehensive database for published molecular associations in human liver diseases. It is based on PubMed published abstracts and aimed to close the gap between genome wide coverage of low validity from microarray data and individual highly validated data from PubMed. After an initial text mining process, the extracted abstracts were all manually validated to confirm content and potential genetic associations and may therefore be highly trusted. All data were stored in a publicly available database, Library of Molecular Associations http://www.medicalgenomics.org/databases/loma/news, currently holding approximately 1260 confirmed molecular associations for chronic liver diseases such as HCC, CCC, liver fibrosis, NASH/fatty liver disease, AIH, PBC, and PSC. We furthermore transformed these data into a powerful resource for molecular liver research by connecting them to multiple biomedical information resources.

**Conclusion:** Together, this database is the first available database providing a comprehensive view and analysis options for published molecular associations on multiple liver diseases.

## Background

The completely sequenced human genome has made it possible for modern medicine to step into an era rich in genetic information and high-throughput genomic analysis [1]. Large gene expression databases [2, 3] and advancing technologies in proteomics [4] provide rich sources for systemic evaluations of the development of chronic liver diseases.
These novel and readily available genetic resources and analytical tools may be the key to unravel the molecular basis of diverse chronic liver diseases as many of these must be regarded to be complex multigenic diseases. Moreover, since an efficient treatment for many of these conditions and diseases is lacking, further understanding of the genetic background of chronic liver disease will be crucial in order to develop new therapies aimed at selected targets [5, 6, 7, 8, 9, 10].
At present, large genetic association studies for liver diseases are mostly based on microarray data or SAGE [11, 12]. Some of these data have recently lead to the identification of prognosticly relevant subgroups in HCC suggesting that a large quantity of microarray data may aid in the identification of biologically relevant biochemical mechanisms [13, 14, 15]. However, most publicly available microarray data on chronic liver disease covers only a few samples [3]. Although these microarrays face several limitations, the data cover large expression profiles. Arguably, the biggest disadvantage is the need of confirming single microarray data by means of molecular biology, e.g. Northern Blot or RT-PCR. Single microarray experiments have been demonstrated to lack reliability with respect to validity of individual single gene expression profiles [16]. Thus, more recent microarray experiments of single probe experiments include confirmation of the proposed hypothesis by means of molecular biology. However, these experiments can be time consuming and costly. To overcome these limitations for systems biology approaches to chronic liver disease, we created a novel resource for systems biology analysis of chronic liver diseases by using PubMed published molecular associations. As multiple molecular factor genes have already been investigated in association these published studies provide a rich source of known molecular associations.

## Implementation

## Data Acquisition

In order to establish this database, the complete PubMed database, currently containing more than 17 million publications, has initially been searched by means of MeSH terms and text mining semi-automated searches [17].
Initially, for each individual disease all abstracts were searched for the disease name or respective MeSH terms providing alternative names or abbreviations which may also be used in the literature to describe the respective disease. In detail the used

MeSH search strings in PubMed used for searching disease associated abstracts were:

1. "Hepatocellular" [MeSH] OR "hepatocellular carcinoma" OR "HCC" OR "hepatoma" OR "liver cancer" OR "primary liver cancer" OR "liver tumor" OR "liver carcinoma" OR "primary liver cancer" OR "hepatic tumor" for HCC
2. "biliary tract cancer" OR "gallbladder cancer" OR "cholangiocellular carcinoma" OR cholangiocarcinoma for CCC
3. "fibrosis" OR "fibroses" for liver fibrosis
4. "NASH" OR "NAFLD" OR "nonalcoholic steatohepatitis" OR "non-alcoholic steatohepatitis" OR "nonalcoholic fatty liver disease" [MeSH] for NASH
5. "AIH" OR "hepatitis, autoimmune" [MeSH] OR "autoimmune hepatitis" for AIH
6. "PBC" OR "primary biliary cirrhosis" [MeSH] OR "biliary cirrhosis, primary"
7. "PSC" OR "sclerosing cholangitis" OR "cholangitis, sclerosing" [MeSH] OR "primary sclerosing cholangitis" for PSC

The abstracts identified to be associated with the particular diseases were then searched for human, mouse, and rat gene names and alias gene names as provided by the Human Genome Organization (HUGO, http://www.hugo-international.org. Making use of the pattern matching capabilities of the Perl programming language http://www.perl.org, we used a pattern matching approach to identify gene names in the previously selected abstracts. E.g. if the gene to be searched was p53, the abstract was searched for any combinations of signs starting with the letter p followed by the numbers 5 and 3. This approach ensured a most flexible search strategy.
Mouse and rat gene names were also searched as not all authors of published abstracts went conform with the HUGO nomenclature and some of them did use murine gene names in (comparative) human studies.
By this approach we gathered a total of 101026 abstracts, potentially holding information on genetic associations to chronic liver disease. In detail we identified 44548 abstracts suggesting genetic associations for HCC, 13710 for CCC, 917 for AIH, 37173 for liver fibrosis, 2022 for NASH, 1211 for PBC and 1445 for PSC.
This strategy revealed all abstracts containing both the disease name and a gene name. However, also this semi-automated search provided a first approximation to genetic associations to liver diseases, as in multiple abstracts this genetic association could not be confirmed by reading the full abstract. E.g. the abstract may read that the gene XY is not related to disease Z, which would have also been detected by the described search strategy. Thus these automatically, by means of text mining identified abstracts, were then all individually read to confirm the suggested genetic association with the particular disease. We thereby obtained a large number of manually confirmed genetic associations to liver diseases.

Thereby, we finally identified 574 molecular associations for HCC, 150 molecular associations for liver fibrosis, 310 molecular associations for CCC, and 82 molecular associations for NASH. Only a few genes were identified to be related to the development of autoimmune liver disease: 29 abstracts describing molecular associations were found to be related to AIH, 56 to PBC, and 60 to PSC. Overall, we were able to identify a total of 1260 molecular associations for major chronic liver diseases. As all these molecular associations were manually confirmed by reading the individual full published abstract, and thus these molecular associations can be trusted to be highly reliable.

## Data organization, Webinterface

The above described strategy of identifying potential genetic associations with chronic liver diseases identified 1260 genetic associations for several diverse chronic liver diseases. Initially the retrieved genetic associations were stored locally in a postgreSQL database http://www.postgresql.org. Subsequently, this database was then made publicly accessible and searchable through a webinterface (Fig. 5.1) implemented in PHP http://de.php.net. It may also be downloaded as a single text file.

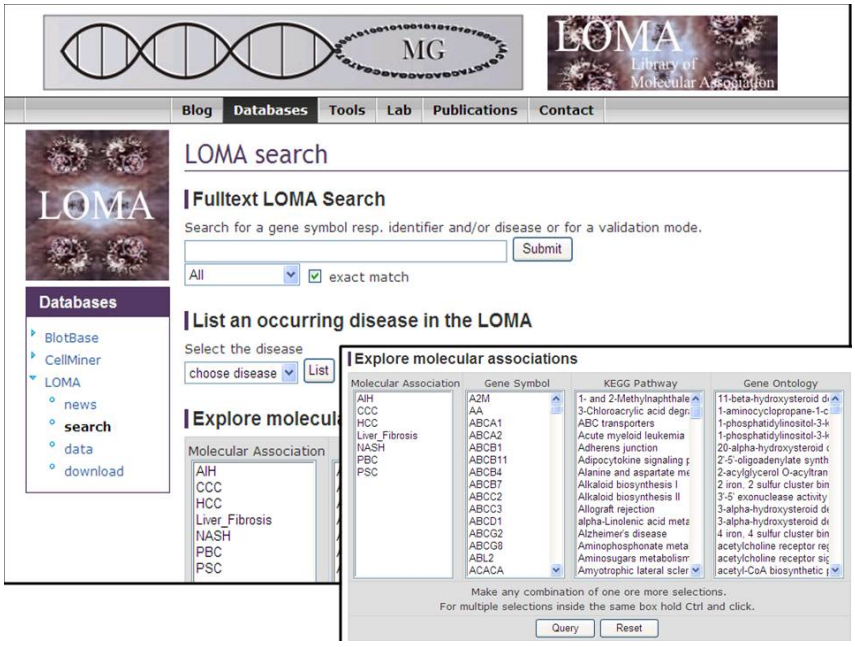


**Figure 5.1:** LOMA data search interface. LOMA offers multiple search options. Searches may be performed by means of individual gene names, NCBI Gene IDs, Ensembl Gene IDs, or disease names. Also more complex searches may be performed by selecting disease, gene symbol, a genetic pathway from KEGG, or a gene ontology from the "explore genetic association" panel.

### Linkage to structural and functional bioinformatics information repositories

Since one of the major goals in implementing this database was to perform high throughput systems biology analyses, the LOMA genetic associations had to be linked to commonly used and established bioinformatics databases and knowledge repositories.
Gene descriptions were assembled from the NCBI Entrez database [17], chromosomal location and Ensembl ID information [18]. Furthermore, data on gene signaling and molecular pathway affiliation were collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG, [19]). Finally, the Gene Ontology database was accessed to identify cellular component, biological process and molecular function information for each gene.

## Results and Discussion

### Database design and rationale

A wide variety of human diseases have been demonstrated to be genetic (inherited). Genetic mutations and a variable genetic background have been demonstrated to significantly influence the development and course of multiple diseases as well as the efficiency of treatment with diverse drugs.
Over the past decades molecular mechanisms and individual factors have been shown to be involved in the development of liver diseases and it has become clear that most liver diseases such as liver cancer, cholangiocellular carcinoma, liver fibrosis, NASH or autoimmune liver diseases are complex systemic diseases. Thus they must not only be investigated focusing on individual, potentially key regulatory genes but also with respect to underlying genetic clusters and networks [7, 20, 21, 22]. However, to investigate these complex molecular interactions, data resources providing a comprehensive collection of all genes involved in the development of the diseases are urgently needed. Microarray and SAGE databases hold a vast amount of gene expression profiles [2, 3]. However, the validity of individual microarray data remains low compared to data generated by means of RT-PCR, Northern-Blot, Western-Blot, RFLP, or even DNA Sequencing. As the later molecular techniques may have a higher validity they have mostly been published in individual publications, currently not available for high throughput analysis. Furthermore extracting and analyzing information on genetic associations in liver diseases already published is extremely time consuming as the respective databases may only be searched for individual publications. However, in total, these data provide a rich source of genetic information.
To overcome these obstacles, we designed a publicly available database for genetic associations with human (liver) diseases, Library of Molecular Associations (LOMA). Currently, this database holds 1260 molecular associations for a total of

seven liver diseases, HCC, CCC, liver fibrosis, AIH, PBC and PSC. Most molecular associations were identified to be associated with HCC, 595, followed by CCC and liver fibrosis, 310 and 150 database entries, respectively. 82 entries were associated with the development of NASH. As expected and in concordance with a currently missing clear association of genetic networks with autoimmune liver diseases, only few genes were reported to be associated with AIH, PBC, or PSC. However, as some of these diseases, especially PBC, have been demonstrated an increased relative risk of the disease in twins and first grade relatives, a genetic basis of the disease must be suspected. Thus further research into the genetic basis of the disease is warranted to identify targets for therapeutic treatment of the disease.
In contrast to other available genetic association databases such as the Genetic Association Database [23], our database contains all published genetic associations with each specific diseases as our semi-automated search was designed to completely capture all associations.

### Database usage

The LOMA database provides multiple search options to support complex genetic analyses. Firstly, LOMA offers the option to search for individual genes and their association with different liver diseases. This search may be performed by means of a search for individual gene names, NCBI Gene IDs [17], Ensembl Gene IDs [18], or disease names. Also more complex searches may be performed by selecting disease, gene symbol, a genetic pathway from KEGG [19], or a gene ontology from the "explore genetic association" panel, providing a highly detailed search option (Fig. 5.1).
After executing a search, the result page for these searches offers the genetic associations to individual diseases if present. Furthermore, the results page gives a summary on gene name, associated disease, NCBI Gene ID [17], Ensembl Gene ID [18], information on the species in which the gene's association to the disease was published (if the respective gene was found to be associated with the disease in human, this category was set to "human" as default). The validation mode column gives a rough estimate, whether a genetic association was only published in a single article (low) or if the genetic association was documented in two or more articles (Fig. 5.2). Finally, more details on the specific gene such as gene alias names, chromosomal location, the association documenting reference(s), gene ontology information, and associated genetic pathways were provided in the "details" section (Fig. 5.3). For example, if one wants to know all molecular associations with the Wnt signaling pathway that have been published to play a role in HCC development, this is now easily possible with our database. On the search site under Explore molecular associations one would select "HCC" from the "Molecular Associations" column and "Wnt Signaling Pathway" from the KEGG column. The executed search will then return a number of Wnt signaling associated genes and target genes, APC, AXIN1, CTNNB1, MMP7, PRKCA, SMAD4, TP53.

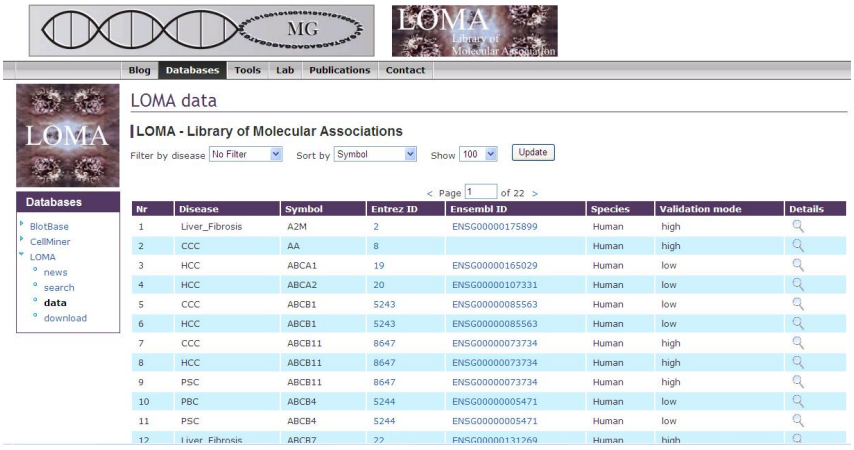


**Figure 5.2:** LOMA results page. The result page provides information on disease and individual information as well as summaries on NCBI Gene ID, Ensembl Gene ID, the species in which the molecular association to the disease was published, and number of publications reporting the molecular association ("high" stands for two or more publications). The details link provides linkage to a rich source of individual molecular information as shown in Figure 5.3.

For these molecular associations further information is linked especially in the details section of each gene. With this information one could for example evaluate the enrichment of the Wnt signaling pathway among all CCC related molecular factors.

### Linkage to common bioinformatics databases

A key issue in developing this database was to provide the hepatologic community with a powerful but simultaneously highly reliable and comprehensive database to perform systems biology based high-throughput searches and comparison of gene expression, our database was linked to multiple other sources of genomic or genetic information and gene expression information in particular. This rich embedding of our database into the current scenery of bioinformatics repositories provides valuable connections which may support advanced search and evaluation strategies.
In detail, LOMA has been linked to the most commonly used bioinformatics databases, such as PubMed [17], the European Bioinformatics Institute Website Ensembl [18], the bioinformatics resource of the National Center of Biotechnology Information Entrez Gene [17], the Mouse Genome Informatics Website (MGI, [24]), and the Gene Ontology database, holding functional information on genes and proteins [25]. These links were selected as they may in addition support automated correlation with additional genomic information such as multiple sequence information, microarray expression data, conserved domains, as well as information on a gene's function.

| Detail information to gene: ABCB1 | |
|---|---|
| Gene symbol | ABCB1 |
| Gene description | ATP-binding cassette, sub-family B (MDR/TAP), member 1 |
| Gene aliases | ABC20<br>CD243<br>CLCS<br>GP170<br>MDR1<br>MGC163296<br>P-gp<br>PGY1 |
| Species | Human |
| Chromosomal location | 7q21.1 |
| Entrez ID | 5243 |
| Ensembl ID | ENSG00000085563 |
| Molecular association | HCC |
| Literature | PubMed: 1346405 |
| Further molecular associations | CCC |
| GeneOntology | nucleotide binding<br>transporter activity<br>protein binding<br>ATP...,<br>membrane fraction<br>transport<br>xenobiotic-transporting ATPase activity<br>cell surface membrane<br>integral to membrane<br>hydrolase activity<br>ATPase activity<br>response to drug |
| Associated Pathways | ABC transporters |

**Figure 5.3:** LOMA results page. The "Details" section of the results page provides extensive additional information and linkage to gene alias names, chromosomal location, the association documenting reference(s), gene ontology information, and associated genetic pathways.

## Comparison to other genetic association databases

Our database has been evaluated against other public databases such as Genetic Association database, HuGENavigator, or OMIM. This evaluation was performed using the molecular associations to CCC development. Comparing our text mining strategy to a manually searched sample set of 1000 randomly selected CCC associated abstracts, we documented a sensitivity of our approach of 98% and a false negative rate for abstracts not selected by or text mining approach but containing molecular associations to CCC of 2%.
For CCC development our database contained all associations also listed by other databases with two exceptions, MRP2/ABCC2 which was published only recently and the miRNA370 which was missed by our search strategy [26]. In contrast however, we provide a significantly larger list of genetic associations to CCC development of 310 molecular associations compared to 6, 19, and 39 in Genetic Association database, HuGENavigator, or OMIM, respectively.

## Conclusion

The Library of Molecular Associations (LOMA) was designed as a comprehensive database of highly reliable molecular associations conceived to close the gap between high-throughput molecular data for automated analysis and individual reliable experimental data by molecular biology. Currently this database supports information on molecular associations for several liver diseases, HCC, CCC, liver fibrosis, NASH/fatty liver disease, AIH, PBC and PSC. In addition, the database was extensively embedded into the currently available genomics repositories supporting advanced searches and cross analyses with other databases.
Together, this database is the first available database providing a comprehensive view and analysis options for published molecular associations on multiple liver diseases.

# BIBLIOGRAPHY


[1] J. F. Hocquette. Where are we in genomics? *Journal of physiology and pharmacology: an official journal of the Polish Physiological Society*, 56 Suppl 3:37–70, **June 2005**. PMID: 16077195.

[2] R. J. Marinelli, K. Montgomery, C. L. Liu, N. H. Shah, W. Prapong, M. Nitzberg, Z. K. Zachariah, G. J. Sherlock, Y. Natkunam, R. B. West, M. van de Rijn, P. O. Brown, and C. A. Ball. The stanford tissue microarray database. *Nucleic acids research*, 36(Database issue):D871–877, **January 2008**. PMID: 17989087.

[3] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–872, **January 2009**. PMID: 19015125.

[4] P. Conrotto and S. Souchelnytskyi. Proteomic approaches in biological and medical sciences: principles and applications. *Experimental oncology*, 30(3):171–180, **September 2008**. PMID: 18806738.

[5] C. Nugent and Z. M. Younossi. Evaluation and management of obesity-related nonalcoholic fatty liver disease. *Nature clinical practice. Gastroenterology & hepatology*, 4(8):432–441, **August 2007**. PMID: 17667992.

[6] D. C. Rockey. Antifibrotic therapy in chronic liver disease. *Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association*, 3(2):95–107, **February 2005**. PMID: 15704042.

[7] A. Teufel, F. Staib, S. Kanzler, A. Weinmann, H. Schulze-Bergkamen, and P.-R. Galle. Genetics of hepatocellular carcinoma. *World journal of gastroenterology: WJG*, 13(16):2271–2282, **April 2007**. PMID: 17511024.

[8] T. Yau, P. Chan, R. Epstein, and R.-T. Poon. Evolution of systemic therapy of advanced hepatocellular carcinoma. *World journal of gastroenterology: WJG*, 14(42):6437–6441, **November 2008**. PMID: 19030192.

[9] A. Teufel, P. R. Galle, and S. Kanzler. Update on autoimmune hepatitis. *World journal of gastroenterology: WJG*, 15(9):1035–1041, **March 2009**. PMID: 19266594.

[10] P. Muratori, A. Granito, G. Pappas, L. Muratori, M. Lenzi, and F. B. Bianchi. Autoimmune liver disease 2007. *Molecular aspects of medicine*, 29(1-2):96–102, **April 2008**. PMID: 18067956.

[11] S. Kaneko and K. Kobayashi. Clinical application of a DNA chip in the field of liver diseases. *Journal of gastroenterology*, 38 Suppl 15:85–88, **March 2003**. PMID: 12698878.

[12] K. S. Swanson. Using genomic biology to study liver metabolism. *Journal of animal physiology and animal nutrition*, 92(3):246–252, **June 2008**. PMID: 18477304.

[13] J.-S. Lee and S. S. Thorgeirsson. Comparative and integrative functional genomics of HCC. *Oncogene*, 25(27):3801–3809, **June 2006**. PMID: 16799621.

[14] S. S. Thorgeirsson, J.-S. Lee, and J. W. Grisham. Functional genomics of hepatocellular carcinoma. *Hepatology (Baltimore, Md.)*, 43(2 Suppl 1):S145–150, **February 2006**. PMID: 16447291.

[15] J.-S. Lee, J. Heo, L. Libbrecht, I.-S. Chu, P. Kaposi-Novak, D. F. Calvisi, A. Mikaelyan, L. R. Roberts, A. J. Demetris, Z. Sun, F. Nevens, T. Roskams, and S. S. Thorgeirsson. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nature medicine*, 12(4):410–416, **April 2006**. PMID: 16532004.

[16] M. Tuomela, I. Stanescu, and K. Krohn. Validation overview of bio-analytical methods. *Gene therapy*, 12 Suppl 1:S131–138, **October 2005**. PMID: 16231045.

[17] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 37(suppl 1):D5–D15, **January 2009**. PMID: 18940862.

[18] G. Spudich, X. M. Fernández-Suárez, and E. Birney. Genome browsing with ensembl: a practical overview. *Briefings in functional genomics & proteomics*, 6(3):202–219, **September 2007**. PMID: 17967807.

[19] K. F. Aoki and M. Kanehisa. Using the KEGG database resource. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 1:Unit 1.12, **October 2005**. PMID: 18428742.

[20] C. Wang, T. Maass, M. Krupp, F. Thieringer, S. Strand, M. A. Wörns, A.-P. Barreiros, P. R. Galle, and A. Teufel. A systems biology perspective on cholangiocellular carcinoma development: focus on MAPK-signaling and the extracellular environment. *Journal of hepatology*, 50(6):1122–1131, **June 2009**. PMID: 19395114.

[21] C. H. Osterreicher, F. Stickel, and D. A. Brenner. Genomics of liver fibrosis and cirrhosis. *Seminars in liver disease*, 27(1):28–43, **February 2007**. PMID: 17295175.

[22] S. Weber, O. A. Gressner, R. Hall, F. Grünhage, and F. Lammert. Genetic determinants in hepatic fibrosis: from experimental models to fibrogenic gene signatures in humans. *Clinics in liver disease*, 12(4):747–757, vii, **November 2008**. PMID: 18984464.

[23] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang. The genetic association database. *Nature genetics*, 36(5):431–432, **May 2004**. PMID: 15118671.

[24] J. A. Blake, J. E. Richardson, M. T. Davisson, and J. T. Eppig. The mouse genome database (MGD). a comprehensive public resource of genetic, phenotypic and genomic data. the mouse genome informatics group. *Nucleic acids research*, 25(1):85–91, **January 1997**. PMID: 9045213.

[25] Gene ontology; http://www.geneontology.org.

[26] F. Meng, H. Wehbe-Janek, R. Henson, H. Smith, and T. Patel. Epigenetic regulation of microRNA-370 by interleukin-6 in malignant human cholangiocytes. *Oncogene*, 27(3):378–386, **January 2008**. PMID: 17621267.

# 6 Publication 2 - A systems biology perspective on cholangiocellular carcinoma development: focus on MAPK-signaling and the extracellular environment

Published in Journal of Hepatology in 2009

*J. Hepatol 50, 1122-1131 (2009)*

Chunxia Wang
Thorsten Maass
Markus Krupp
Florian Thieringer
Susanne Strand
Marcus A. Wörns
Ana-Paula Barreiros
Peter R. Galle
Andreas Teufel

| Authors contribution: | |
|---|---|
| Draft writing | 10% |
| Review process | 10% |
| Data processing | 90% |
| Data analysis | 30% |
| Database design | not supported |
| GUI implementation | not supported |

## Abstract

**Background/Aims:** Multiple genes have been implicated in cholangiocellular carcinoma (CCC) development. However, the overall neoplastic risk is likely associated with a much lower number of critical physiological pathways.

**Methods:** To investigate this hypothesis, we extracted all published genetic associations for the development of CCC from PubMed (genetic association studies, but also studies associating genes and CCC in general, i.e. functional studies in cell lines, genetic studies in humans, knockout mice etc.) and integrated CCC microarray data.

**Results:** We demonstrated the MAPK pathway was consistently enriched in CCC. Comparing our data to genetic associations in HCC often successfully treated by a multityrosine kinase inhibitor, sorafenib, we demonstrated a similar overrepresentation of MAPK. In contrast, most cancer-related genetic studies focusing on genes related to transcription and cell cycle control, we consistently found genes coding for products in the extracellular environment to be significantly enriched. Thus, CCC must be regarded as developing in the context of an altered extracellular environment.

**Conclusions:** Our study suggests the liver microenvironment holds essential functions and structures key to CCC progression. Furthermore, we identified the MAPK signaling pathway consistently enriched, pointing towards a critical role in CCC development. These data may provide a rationale for treatment of CCC with sorafenib.

## Introduction

Cholangiocellular carcinoma (CCC) is a comparatively rare cancer arising from the bile ducts. However, rates of cholangiocellular carcinoma have been rising worldwide over the past several decades, particularly rates of intrahepatic CCC [1, 2].
Known risk factors account for only a few cases of CCC. Highest incidences of CCC were documented among patients with primary sclerosing cholangitis (PSC) [3]. Furthermore, particularly in Asia, infection with parasitic liver flukes causes a significant number of CCC [4].
Therapeutic options for CCC currently remain very limited. Besides surgery, the use of chemotherapy is still a matter of intense debate with many arguing for best supportive care as the standard of treatment [5]. Recently, targeted therapies have become novel therapeutic options in a variety of diseases. For example the multi-tyrosine kinase inhibitor sorafenib was the first drug to significantly improve overall survival in hepatocellular carcinoma (HCC), previously regarded as resistant to conventional chemotherapeutic strategies [6].
Numerous individual genes have been studied with respect to their level of expression in liver tissue. However, the overall picture is still undefined and general rules or factors regulating gene expression in the liver have not yet been established [7]. A vast number of genes have been reported to be involved in cancer development [8]. With this sheer number of genes the overall neoplastic risk has been suggested to be influenced rather by a much lower number of essential (physiological) pathways [9, 10]. Thus, we performed a genome-wide analysis of genetic factors involved in CCC development.

## Material and methods

### Data aquisition and accessibility

In order to establish a comprehensive dataset on genetic associations of CCC, the complete PubMed database, currently containing approximately 17 million publications, has initially been searched by means of MeSH terms for abstracts containing the terms “CCC”, “biliary tract cancer”, “gallbladder cancer”, “cholangiocellular carcinoma” and “cholangiocarcinoma”. Subsequently, the extracted abstracts were searched for the official gene names of the Human Genome Organiszation (HUGO, http://www.hugo-international.org/) including all alias gene names obtained from the NCBI Gene Website as well as the corresponding murine gene names also from NCBI [11]. This strategy revealed 13710 abstracts assumed to potentially describe genetic associations in CCC development. Subsequently, these publications were individually and manually validated for genetic associations, leaving 236 confirmed potential genetic associations for CCC development. Publications included not only genetic association studies (case-control studies), but studies associating genes and cholangiocarcinoma in general, i.e. functional stud-

ies in cell lines, genetic studies in humans, knockout mice etc. These genes as well as the genes regulated in the microarray experiment by Obama et al. [12] were designated "CCC associated" in this paper. Finally, validated information was stored in our publicly available database, Library of Genetic Associations (http://www.medicalgenomics.org/databases/loga/news, Fig. 6.1).

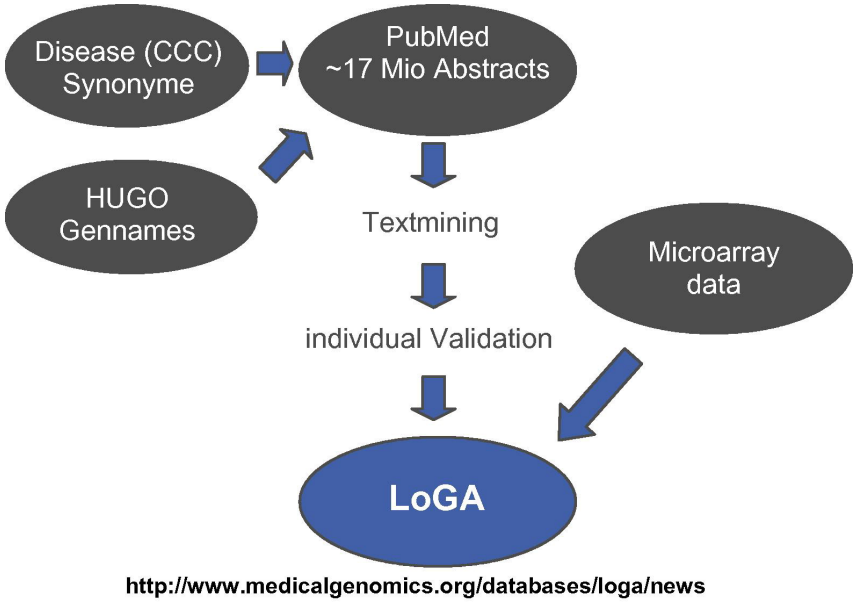


**Figure 6.1:** Schematic drawing of the data acquisition process. Identified potential genetic associations to CCC development were made publicly available through the LoGA data base (http://www.medicalgenomics.org/databases/loga/news)

## Microarray data

Microarray data were obtained from the microarray experiment by Obama et al. as published (http://www.interscience.wiley.com/jpages/0270-9139/suppmat/index.html, [12]). Of these, 51 upregulated and 324 downregulated genes could be assigned an individual gene ID and thus were then available for automated analysis. Genes/EST not being assigned an Entrez Gene ID were mostly EST of unknown functional relevance.

## Automated analysis

As for many genes multiple alias names exist, we had to address each gene with the gene-ID provided by the NCBI Entrez website [11]. For some genes, generally uncharacterized genes, unique IDs have not yet been assigned. Thus, after matching individual Entrez-Gene-IDs to the individual genes we obtained a gene set of 601

genes for further automated analysis, 592 of which were available for WebGestalt (see [13]) analysis as they had an individual gene name assigned and the associaton with CCC was found in human tissue. As for those genes not having an Entrez-Gene-ID, no functional information was available, we did not expect the removal of a few genes from the gene set to be critical for our further functional analysis.

### Gene set analysis

Gene set analysis of the PubMed isolated gene sets and microarray gene sets as well as the complete data set were performed using the WebGestalt Toolkit [13]. For comparison to reference collectives, we selected the human reference gene set "WEBGESTALT_HUMAN". For statistical testing choice was left on "no preference". Due to the strictly human reference gene set, we discarded genes that had only been demonstrated in mouse or rat to be associated with CCC from this analysis. As these were only four genes confirmed in rat 3 in mouse, and 1 in hamster we did not expect this to be critical to the analysis.

### Analysis of genetic association with hepatocellular carcinoma (HCC)

In order to be able to investigate genetic associations for hepatocellular carcinoma, we repeated the same selection strategy as for CCC. MeSH terms were "Hepatocellular" (MeSH) OR "hepatocellular carcinoma" OR "HCC" OR "hepatoma" OR "liver cancer" OR "primary liver cancer" OR "liver tumor" OR "liver carcinoma" OR "primary liver cancer" OR "hepatic tumor". We identified 608 confirmed human potential genetic associations for HCC. The validated associations were stored in our publicly available database (http://www.medicalgenomics.org/databases/loga/news).
Microarray data were provided by Lee et al. and had previously been published [14] and [15].

## Result

### Chromosomal distribution

To identify genetic hotspots we analyzed the chromosomal distribution of the genes associated with CCC. Looking at the complete list of genes, these were found to be evenly distributed. An exception to this was the Y-chromosome which contained only two of these genes. Also at Chromosome 17 a regional accumulation of CCC associated genes was observed. However, none of these genes were neighbouring one another. This general distribution pattern was observed in all three subgroups of the complete list, the PubMed extracted genes and the array of up/downregulated genes (Fig. 6.2).

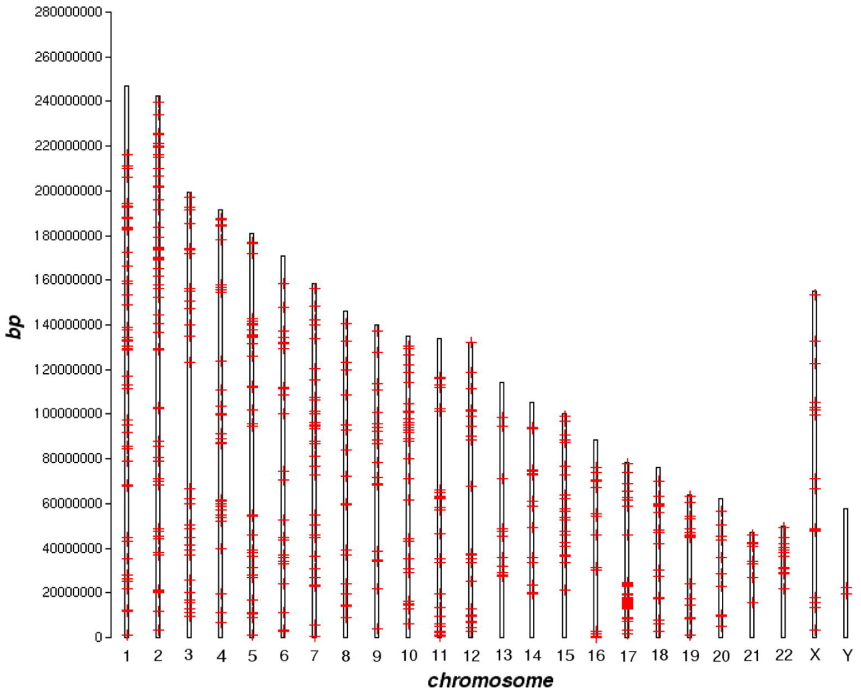


**Figure 6.2:** Chromosomal distribution of genes associated with cholangiocellular carcinoma development.

## Pathway and cluster analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG)

The Kyoto Encyclopedia of Genes and Genomes (KEGG) currently holds 344 reference pathways. Our complete gene list was evaluated for enriched pathways. Overall, a genetic cluster summarizing the functions of focal adhesion was found to have the highest significance with respect to enrichment ($p = 8.16e-23$). Looking at genetic pathways as listed in KEGG involved in cancer development, we found pathways such as the MAPK ($p = 3.26e-12$), insulin signaling ($p = 2.84e-7$), GnRH ($p = 3.87e-7$), mTOR ($p = 5.53e-7$), VEGF ($p = 9.45e-6$), Fc epsilon RI ($p = 1.58e-5$), calcium signaling ($p = 1.90e-5$), Jak-STAT signaling ($p = 2.09e-5$), complement and coagulation cascades ($p = 4.32e-5$), toll-like receptor ($p = 4.67e-5$), adipocytokine signaling ($p = 6.36e-5$), T cell receptor signaling ($p = 3.45e-4$), and Wnt signaling pathway ($p = 1.55e-3$), and to be significantly enriched. Besides classical pathways, several clusters of genes associated with major cancer entities have been demonstrated to be overrepresented pointing towards a common oncogenic basis: colorectal cancer ($p = 5.51e-18$), Glioma ($p = 1.15e-16$), pancreatic cancer ($p = 1.56e-15$), chronic myeloid leukemia ($p = 3.11e-12$) (Table 6.1A).

Analyzing the subset of genes found to be regulated in the microarray study by Obama et al. [12] genes associated with the MAPK signaling pathway ($p = 7.91e-5$) and toll-like receptor signaling pathway ($p = 5.99e-4$), were still demonstrated

**Table 6.1:** Enrichment of common genetic pathways in our gene set of genes associated with CCC development. Enrichment was analyzed in KEGG (A) and Biocarta (B) annotated pathways as the two major resources of genetic pathway annotation. For KEGG the most overrepresentated genetic pathways/clusters as well as the commonly known signaling pathways of the KEGG "Environmental Information Processing" and "Cellular Processses" sections were displayed. For Biocarta the most highly overrepresented signaling pathways were summarized. Pathways were sorted strictly by p-value.

(A)

| KEGG pathway | Gene number | Enrichment |
|---|---|---|
| Focal adhesion | 35 | p = 8.16e-23 |
| Colorectal cancer | 22 | p = 5.51e-18 |
| Glioma | 19 | p = 1.15e-16 |
| Pancreatic cancer | 19 | p = 1.56e-15 |
| **MAPK signaling pathway** | **27** | **p = 3.26e-12** |
| Metabolism of xenobiotics by cytochrome P450 | 16 | p = 1.45e-13 |
| Chronic myeloid leukemia | 16 | p = 3.11e-12 |
| Regulation of actin cytoskeleton | 23 | p = 1.02e-11 |
| Cytokine-cytokine receptor interaction | 25 | p = 1.47e-11 |
| Apoptosis | 16 | p = 1.47e-11 |
| Insulin signaling pathway | 14 | p = 2.84e-7 |
| GnRH signaling pathway | 12 | p = 3.87e-7 |
| mTOR signaling pathway | 9 | p = 5.53e-7 |
| VEGF signaling pathway | 9 | p = 9.45e-6 |
| Fc epsilon RI signaling pathway | 9 | p = 1.58e-5 |
| Calcium signaling pathway | 13 | p = 1.90e-5 |
| Jak-STAT signaling pathway | 12 | p = 2.09e-5 |
| Complement and coagulation cascades | 8 | p = 4.32e-5 |
| Toll-like receptor signaling pathway | 9 | p = 4.67e-5 |
| Adipocytokine signaling pathway | 8 | p = 6.36e-5 |
| T cell receptor signaling pathway | 8 | p = 3.45e-4 |
| Wnt signaling pathway | 9 | p = 1.55e-3 |

(B)

| Biocarta pathway | Gene number | Enrichment |
|---|---|---|
| p53 Signaling pathway | 10 | p = 6.81e-12 |
| Role of ERBB2 in signal transduction and oncology | 10 | p = 8.41e-11 |
| CXCR4 signaling pathway | 10 | p = 1.20e-10 |
| IL-2 receptor beta chain in T cell activation | 11 | p = 2.07e-10 |
| EGF signaling pathway | 10 | p = 4.45e-10 |
| Regulation of BAD phosphorylation | 9 | p = 5.39e-10 |
| Antiapoptotic path. from IGF-1R lead to BAD phosphorylation | 9 | p = 5.39e-10 |
| Signaling pathway from G-protein families | 9 | p = 1.57e-9 |
| Trefoil factors initiate mucosal healing | 9 | p = 2.18e-9 |
| Influence of Ras and Rho proteins on G1 to S transition | 9 | p = 2.98e-9 |
| Keratinocyte differentiation | 10 | p = 6.10e-9 |
| NFAT and hypertrophy of the heart (transcription in the broken heart) | 10 | p = 6.10e-9 |
| Telomeres, telomerase, cellular aging and immortality | 8 | p = 1.04e-8 |
| **Erk1/Erk2 Mapk signaling pathway** | **9** | **p = 1.22e-8** |
| **MAPKinase signaling pathway** | **9** | **p = 3.04e-5** |

to be significantly enriched in this gene set whereas other pathways such as the Jak-STAT ($p = 2.58e-2$), GnRH ($9.54e-2$), VEGF ($p = 5.46e-1$), or Wnt signaling pathway ($p = 2.25e-1$) could not be confirmed.
However, the enrichment of a genetic cluster attributed to focal adhesion ($p = 3.33e-6$) was also identified in CCC by means of microarray studies.

### Pathway and cluster analysis, Bicoarta pathway database

Similar to KEGG the Biocarta database currently holds 353 reference pathways. A large number of pathways commonly known to be involved in cancer development was found to be enriched in our CCC gene list (Table 1B). The top five highly overrepresented pathways were the p53 signaling pathway ($p = 6.81e-12$), role of ERBB2 in signal transduction and oncology ($p = 8.41e-11$), CXCR4 signaling pathway ($p = 1.20e-10$), IL-2 receptor beta chain in T cell activation ($p = 2.07e-10$), and EGF signaling pathway ($p = 4.45e-10$). Besides, the (Erk1) MAPK signaling pathway was also demonstrated to be among the enriched signaling pathways ($p = 1.22e-8$ and $p = 3.04e-5$, respectively).

### Comparison of MAPK overrepresentation to HCC

Applying the same algorithms to HCC development, we identified 559 genes to be involved in the disease's development and available for automated analysis. Analyzing these genes for overrepresented pathways, the MAPK signaling pathway was highly significantly overrepresented in KEGG ($p = 4.07e-24$).
Again, we performed the same analysis with the extensive microarray data. We analyzed the previously published data of Lee [14, 15] and again were able to also confirm a highly overrepresented MAPK signaling pathway ($p = 9.65e-4$).

### Clustering localizations of genes

Consistently, the analysis of the complete list of CCC associated genes as well as all individual subsets demonstrated a strong significant overrepresentation of genes with the attributed localization "extracellular region" ($p = 1.70e-12$). This localization is defined as "the space external to the outermost structure of a cell... or the space outside of the plasma membrane". Similarly, for "extracellular matrix", defined as "a structure lying external to one or more cells, which provides structural support for cells or tissues..." a significant overrepresentation of attributed genes was identified in all subsets as well as the complete gene list ($p = 3.33e-5$ for the complete list).

### Evolutionary conservation

To evaluate the identified genes associated with CCC for evolutionary conservation, we were able to assign 566 HUGO individual symbols to our data set of human genes associated with CCC development. Highest conservation was found in mouse with 443 orthologues genes present in the mouse genome (78%). Slightly lower conservation was demonstrated to be present in Rattus norvegicus (413 genes conserved, 73%) and Xenopus tropicalis (301 genes conserved, 53%). In Danio rerio still 218 (39%) of the human CCC associated genes were conserved. In non-vertebrates, Drosophila melanogaster, only 50 (9%) of all genes were found to be conserved orthologues genes. Finally, 19 orthologous genes were identified for yeast, Saccharomyces cerevisiae (3%).

### Analysis of biological functions of the extracellular region genes

The extracellular region and microenvironment were the only compartments being significantly overrepresented within our dataset of genetic associations. Looking at the GO terms [16] associated with these genes, the biological functions overrepresented were "response to stimulus" ($p = 2.73e-13$), "development" ($p = 7.65e-6$), "interaction between organisms" ($p = 4.28e-4$), "developmental maturation" ($p = 2.75e-4$), and "growth" ($p = 8.27e-3$).
With respect to the molecular function "enzyme regulatory activity" ($p = 1.72e-4$) and "binding" ($p = 1.70e-3$) were significantly overrepresented. However, besides these classical functional entities, 13 extracellular structural proteins were also highly overrepresented ($p = 6.00e-4$), not commonly reported to be involved in cancer development (see Table 6.2 and Table 6.4).

### Discussion

Over the past decades multiple genes have been detected which may be involved in cholangiocellular development. We summarized 601 genes being essential to cholangiocellular carcinoma development [17]. Through this analysis, it has become clear that cancer is a systemic disease and must be investigated not only by focusing on individual potentially key regulatory genes, but also with respect to underlying genetic clusters and networks.
Analyzing this large amount of data obviously requires a high degree of automatization and thus comprehensive databases containing genetic associations of the respective cancer entities to be investigated. For hepatocellular carcinoma, we and others have established those databases in the past (Medicalgenomics [17]; EHCO [18]). However, for CCC such a data repository had not been established so far. We therefore extracted all published potential genetic associations with CCC from currently more than 17 million publications of the NCBI PubMed database [11] and

**Table 6.2:** Genes associated with CCC development and attributed GO term extracellular region.

| Gene ID | Symbol | Gene ID | Symbol | Gene ID | Symbol |
|---|---|---|---|---|---|
| 4583 | MUC2 | 5730 | PTGDS | 4060 | LUM |
| 80144 | FRAS1 | 3491 | CYR61 | 59277 | NTN4 |
| 3514 | IGKC | 2064 | ERBB2 | 6366 | CCL21 |
| 7276 | TTR | 4069 | LYZ | 6363 | CCL19 |
| 3486 | IGFBP3 | 4313 | MMP2 | 7980 | TFPI2 |
| 3455 | IFNAR2 | 2638 | GC | 4586 | MUC5AC |
| 5444 | PON1 | 345 | APOC3 | 1294 | COL7A1 |
| 7057 | THBS1 | 9510 | ADAMTS1 | 7076 | TIMP1 |
| 3908 | LAMA2 | 4314 | MMP3 | 3240 | HP |
| 8743 | TNFSF10 | 7078 | TIMP3 | 7033 | TFF3 |
| 4239 | MFAP4 | 462 | SERPINC1 | 3479 | IGF1 |
| 7045 | TGFBI | 3570 | IL6R | 183 | AGT |
| 7373 | COL14A1 | 710 | SERPING1 | 7422 | VEGFA |
| 9068 | ANGPTL1 | 7018 | TF | 6387 | CXCL12 |
| 1440 | CSF3 | 5741 | PTH | 4192 | MDK |
| 9353 | SLIT2 | 5678 | PSG9 | 3569 | IL6 |
| 375790 | AGRIN | 9235 | IL32 | 6906 | SERPINA7 |
| 1956 | EGFR | 4316 | MMP7 | 2056 | EPO |
| 1066 | CES1 | 1191 | CLU | 1490 | CTGF |
| 54097 | FAM3B | 5950 | RBP4 | 2243 | FGA |
| 338 | APOB | 1292 | COL6A2 | 3339 | HSPG2 |
| 3026 | HABP2 | 2246 | FGF1 | 6696 | SPP1 |
| 6948 | TCN2 | 2335 | FN1 | 5549 | PRELP |
| 55959 | SULF2 | 3484 | IGFBP1 | 2719 | GPC3 |
| 213 | ALB | 5327 | PLAT | 2266 | FGG |
| 1890 | ECGF1 | 4582 | MUC1 | 2778 | GNAS |
| 7448 | VTN | 259 | AMBP | 6278 | S100A7 |
| 3481 | IGF2 | 2153 | F5 | 7123 | CLEC3B |
| 1950 | EGF | 10395 | DLC1 | 1401 | CRP |
| 3512 | IGJ | 2200 | FBN1 | 350 | APOH |

**Table 6.3:** Genes associated with CCC development and attributed GO term extracellular region coding for structural proteins.

| Gene ID | Symbol |
|---|---|
| 1292 | COL6A2 |
| 1294 | COL7A1 |
| 2200 | FBN1 |
| 2335 | FN1 |
| 3908 | LAMA2 |
| 4060 | LUM |
| 4583 | MUC2 |
| 4586 | MUC5AC |
| 5549 | PRELP |
| 7057 | THBS1 |
| 7373 | COL14A1 |
| 7980 | TFPI2 |
| 375790 | AGRIN |

furthermore included potential genetic associations identified by microarray experiments [12].

### Evolutionary conservation

Analyzing the evolutionary conservation of the complete list of 592 genes available for automated evaluation, we were able to demonstrate a high conservation of these genes in mammals (78% and 73%), in contrast to a much lower conservation of these genes in non-vertebrates (9% in Drosophila). In the lowest eukaryotic organism, yeast (Saccharomyces cervisiae) we observed only a few genes to be conserved, only 3%. This was comparable to the reported conservation of genes associated with the development of HCC [12].

### Representation of genetic signaling pathways

At present two large repositories were established holding genetic pathway information, KEGG [19] and Biocarta [20]. Comparing our data set to the KEGG data repository, focal adhesion was found to have the most significant enrichment of individual genes of the pathway. This is of significant interest as the extracellular microenvironment was found to be significantly enriched. Key features of this pathway are the interaction of the extracellular matrix and integrin receptors transmitting signals from the extracellular microenvironment into the cell. This strong enrichment of the pathway including genes relevant to focal adhesion was further-

more validated in a subset analysis looking only at the microarray derived data, not subject to a possible selection bias by the scientific community.

### Enrichment of genes associated with MAPK signaling – a rationale for sorafenib treatment of CCC ?

Several genetic pathways previously demonstrated to be involved in cancer development in general were also identified to be essential for CCC development. Among these were the MAPK [21], Jak-STAT GnRH [22], VEGF [23], the mTOR [24], or the toll-like receptor [25], T cell receptor signaling [26], and the Wnt signaling pathway [27]. Furthermore, the MAPK signaling pathway was also demonstrated to be enriched by an additional investigated pathway repository, the Biocarta database (Table 6.1).
However, these data collected from PubMed may underlie a selection bias, since pathways successfully linked to CCC development may be subject to greater attention and more extensive investigation by the scientific community. In order to prevent such a selection bias we furthermore investigated microarray data [12]. Analyzing all regulated genes in this dataset, the MAPK signaling pathway was also highly enriched. Thus, the MAPK signaling enrichment was confirmed by the microarray dataset and demonstrated not to be subject to a selection bias in PubMed (Table 6.1).
Over recent years, several studies have demonstrated a pivotal role of mitogen activated protein kinase (MAPK) signaling pathways in CCC. Traditionally, the classical p44/p42 MAPK pathway has been connected to regulation of cell growth and differentiation and the Ras oncogene transmits extracellular growth signals through this cascade [28]. K-Ras mutations were seen in early carcinogenesis of intrahepatic cholangiocellular carcinomas [29] and associated with an aggressive tumour occurrence [30], correlating to poor prognosis [31]. Finally, a disruption of the MAPK pathway was detected in approximately 62% of CCC [32].
p38 MAPK and JNK pathways are both strongly activated in response to environmental stresses and inflammatory cytokines [28]. Bile acids promote cholangiocellular carcinogenesis by inducing EGFR phosphorylation, thereby enhancing different MAPK pathways.
It has to be critically acknowledged though that most of the genes associated with MAPK signaling in the Obama microarray data set were downregulated as this data set identified most genes regulated to be downregulated in expression. Some of these downregulated genes were suppressors of the classical MAPK signaling pathway, e.g. DUSP1 [33]. Others were not directly linked to the classical EGF-MAPK-pathway and thus not necessarily interacting or inhibiting the activation of the classical EGF-MAPK-pathway, e.g. NUSS7. The only downregulated gene not fitting into the concept of an activated classical MAPK signaling pathway was SOS2. SOS2 was demonstrated to contribute to receptor-mediated Ras activation as it interacts with GRB2 [34] and thus a downregulation of SOS2 would result

in a inhibition of classical MAPK signalling. However, with 12 genes of the classical MAPK signaling pathway individually and independently being demonstrated to be activated in CCC development, it seems questionable whether SOS2 is commonly downregulated in CCC development. Even more, as of the 421 genes being demonstrated to be downregulated in CCC, SOS2 was only number 369 with respect to a comparably low 2.86-fold negative regulation of expression of SOS2, which was on the lower end of regulated genes.
Together, these findings of a significant role of MAPK signaling in CCC development may provide additional support for the evaluation of novel targeted treatment approaches. Sorafenib is a multikinase inhibitor blocking tumour cell proliferation and angiogenesis [35]. Sorafenib treatment dose-dependently blocked growth-factor-induced activation of the MAPK and inhibited the proliferation of EGI-1 and TFK-1 CC cells in a time- and dose-dependent manner [36]. Since sorafenib has been proven to be safe and efficient in HCC therapy these results may advocate therapeutic trials with sorafenib in CCC.

## Comparison of MAPK overrepresentation to HCC

If MAPK overrepresentation was postulated to be a rationale for sorafenib treatment of CCC one would argue that MAPK should also be overrepresented in data sets of genetic associations with HCC development, were the drug has already clinically proven benefit. Those data have recently been published by Hsu et al. [18] and were reproduced by us. Analyzing these genes for overrepresented pathways, we demonstrated the MAPK signaling pathway to be highly significantly overrepresented in both the gene set obtained by means of PubMed text mining ($p = 4.07e-24$) as well as an independent microarray data set ($p = 9.65e-4$, [14] and [15]). Thus, by applying our approach to HCC we were able to provide even more support for a significant role of MAPK signaling in CCC. These results further argue in favour of a trial of sorafenib and MAPK targeting in CCC.

## Gene ontology evaluation: focus on the extracellular microenvironment

In order to evaluate the extracted gene sets for their biological function and relevance with respect to tumour development, we made use of gene ontologies (GO, [37]), a standardized vocabulary to describe gene and gene product attributes.
However, analyzing the cellular compartment, one would expect to have the nuclear or cytosplasmatic compartments to be overrepresented as the majority of cancer related genetic studies have focussed on genes involved in transcription and cell cycle control. This is reflected by searching for abstracts containing cancer and cell cycle against cancer and extracellular matrix or microenvironment a ratio of 10:1 in number of articles opened up (92.306 vs. 9.220 articles). In contrast to these expectations, the only cellular compartment associated with CCC development in the complete set of identified associations but also in both individual subsets, the

NCBI PubMed extracted genes and the microarray data was the extracellular matrix and extracellular region in general. Those results were somewhat surprising, as it shifts the center of interest away from intracellular genetic pathways and argues for a much more complex picture of CCC development (Table 6.2 and Table 6.4).
The extracellular microenvironment and matrix has been thought for decades to be simply the "glue" between cells, responsible for shape and coherence of tissues and organs as well as a reservoir of body fluids. Rather, the liver microenvironment seems to be an important piece in a complex puzzle so cancers must be regarded as heterogeneous multicellular entities containing cells of multiple lineages whose interactions with one another, the extracellular matrix (ECM), and soluble molecules in their vicinity are dynamic. Furthermore these interactions favour cancer cell proliferation, movement, differentiation, and ECM metabolism, while simultaneously restricting cell death, stationary polarized growth, and ECM stability [38]. However, the microenvironment is also essential to key processes such as angiogenesis. Also important is the appreciation that cell migration and matrix and tissue remodeling are not unique properties of cancerous growths but instead are regulated programs normally utilized during development and in adult tissues responding to acute tissue stress [38]. Reciprocal interactions between these responding "normal" cells, their mediators, structural components of the ECM, and genetically altered neoplastic cells regulate all aspects of tumourigenicity. An additional role was suggested by harbouring and complex interaction of the extra cellular tumour stroma with (mesenchymal) stem cells being involved in cancer growth [39].
Interestingly, genes associated with CCC development and associated with the extracellular microenvironment not only encoded growths factors or related proteins such as receptors or known regulatory proteins. Also 12 genes coding for structural proteins were highly enriched. This points towards key roles of the texture and structure of the extracellular microenvironment [40].

## Conclusion

Analysis of the underlying genetic networks in CCC development revealed a significant enrichment of the MAPK signaling pathway, supporting evidence for further evaluation of sorafenib in treatment of CCC. Furthermore, the liver microenvironment was found to be the only consistently enriched biological compartment assigned to genes within the CCC dataset. Thus the cellular matrix and microenvironment must be assumed to hold essential functions and structures key to CCC development. Thus future research on CCC development must extend the focus of molecular research to the liver microenvironment as a compartment absolutely essential to tumour development.

# BIBLIOGRAPHY


[1] S. A. Khan, H. C. Thomas, B. R. Davidson, and S. D. Taylor-Robinson. Cholangiocarcinoma. *The Lancet*, 366(9493):1303–1314, **August**.

[2] Y. H. Shaib, J. A. Davila, K. McGlynn, and H. B. El-Serag. Rising incidence of intrahepatic cholangiocarcinoma in the united states: a true increase? *Journal of Hepatology*, 40(3):472–477, **March 2004**.

[3] Y. Shaib and H. B. El-Serag. The epidemiology of cholangiocarcinoma. *Seminars in liver disease*, 24(2):115–125, **May 2004**. PMID: 15192785.

[4] T. Ben-Menachem. Risk factors for cholangiocarcinoma. *European Journal of Gastroenterology & Hepatology*, 19(8):615–617, **August 2007**.

[5] M. Shimoda and K. Kubota. Multi-disciplinary treatment for cholangiocellular carcinoma. *World journal of gastroenterology: WJG*, 13(10):1500–1504, **March 2007**. PMID: 17461440.

[6] A. X. Zhu. Development of sorafenib and other molecularly targeted agents in hepatocellular carcinoma. *Cancer*, 112(2):250–259, **2008**.

[7] A. Teufel, A. Weinmann, M. Krupp, M. Budinger, and P. R. Galle. Genome-wide analysis of factors regulating gene expression in liver. *Gene*, 389(2):114–121, **March 2007**.

[8] T. Sakatani and P. Onyango. Oncogenomics: prospects for the future. *Expert Review of Anticancer Therapy*, 3(6):891–901, **December 2003**.

[9] A. Teufel, M. Krupp, A. Weinmann, and P. R. Galle. Current bioinformatics tools in genomic biomedical research (review). *International journal of molecular medicine*, 17(6):967–973, **June 2006**. PMID: 16685403.

[10] P. Hernández, J. Huerta-Cepas, D. Montaner, F. Al-Shahrour, J. Valls, L. Gómez, G. Capellá, J. Dopazo, and M. A. Pujana. Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics*, 8(1):185, **June 2007**. PMID: 17584915.

[11] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(Database issue):D5–12, **January 2007**. PMID: 17170002.

[12] K. Obama, K. Ura, M. Li, T. Katagiri, T. Tsunoda, A. Nomura, S. Satoh, Y. Nakamura, and Y. Furukawa. Genome-wide analysis of gene expression in human intrahepatic cholangiocarcinoma. *Hepatology*, 41(6):1339–1348, **2005**.

[13] B. Zhang, S. Kirov, and J. Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33(suppl 2):W741–W748, **July 2005**. PMID: 15980575.

[14] J.-S. Lee, J. Heo, L. Libbrecht, I.-S. Chu, P. Kaposi-Novak, D. F. Calvisi, A. Mikaelyan, L. R. Roberts, A. J. Demetris, Z. Sun, F. Nevens, T. Roskams, and S. S. Thorgeirsson. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nature Medicine*, 12(4):410–416, **April 2006**.

[15] J.-S. Lee and S. S. Thorgeirsson. Comparative and integrative functional genomics of HCC. *Oncogene*, 25(27):3801–3809, **2006**.

[16] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25–29, **May 2000**. PMID: 10802651.

[17] http://www.medicalgenomics.org/databases/loga/news.

[18] C.-N. Hsu, J.-M. Lai, C.-H. Liu, H.-H. Tseng, C.-Y. Lin, K.-T. Lin, H.-H. Yeh, T.-Y. Sung, W.-L. Hsu, L.-J. Su, S.-A. Lee, C.-H. Chen, G.-C. Lee, D. T. Lee, Y.-L. Shiue, C.-W. Yeh, C.-H. Chang, C.-Y. Kao, and C.-Y. F. Huang. Detection of the inferred interaction network in hepatocellular carcinoma from EHCO (encyclopedia of hepatocellular carcinoma genes online). *BMC Bioinformatics*, 8(1):66, **February 2007**. PMID: 17326819.

[19] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, **January 1999**. PMID: 9847135.

[20] Biocarta. http://www.biocarta.com.

[21] G. Pimienta and J. Pascual. Canonical and alternative MAPK signaling. *Cell cycle (Georgetown, Tex.)*, 6(21):2628–2632, **November 2007**. PMID: 17957138.

[22] A. Ferrajoli, S. Faderl, F. Ravandi, and Z. Estrov. The JAK-STAT pathway: A therapeutic target in hematological malignancies. *Current Cancer Drug Targets*, 6(8):671–679, **December 2006**.

[23] M. Kowanetz and N. Ferrara. Vascular endothelial growth factor signaling pathways: Therapeutic perspective. *Clinical Cancer Research*, 12(17):5018–5022, **September 2006**. PMID: 16951216.

[24] Y. Mamane, E. Petroulakis, O. LeBacquer, and N. Sonenberg. mTOR, translation initiation and cancer. *Oncogene*, 25(48):6416–6422, **2006**.

[25] J. Krishnan, K. Selvarajoo, M. Tsuchiya, G. Lee, and S. Choi. Toll-like receptor signal transduction. *Experimental & Molecular Medicine*, 39(4):421–438, **2007**.

[26] C. Geisler. TCR trafficking in resting and stimulated t cells. *Critical Reviews in Immunology*, 24(1):67–86, **2004**.

[27] N. Gavert and A. Ben-Ze'ev. Beta-catenin signaling in biological control and cancer. *Journal of Cellular Biochemistry*, 102(4):820–828, **2007**.

[28] P. P. Roux and J. Blenis. ERK and p38 MAPK-Activated protein kinases: a family of protein kinases with diverse biological functions. *Microbiology and Molecular Biology Reviews*, 68(2):320–344, **June 2004**. PMID: 15187187.

[29] K. Ohashi, Y. Nakajima, H. Kanehiro, M. Tsutsumi, J. Taki, Y. Aomatsu, A. Yoshimura, S. Ko, T. Kin, and K. Yagura. Ki-ras mutations and p53 protein expressions in intrahepatic cholangiocarcinomas: relation to gross tumor morphology. *Gastroenterology*, 109(5):1612–1617, **November 1995**. PMID: 7557145.

[30] T. Isa, S. Tomita, A. Nakachi, H. Miyazato, H. Shimoji, T. Kusano, Y. Muto, and M. Furukawa. Analysis of microsatellite instability, k-ras gene mutation and p53 protein overexpression in intrahepatic cholangiocarcinoma. *Hepatogastroenterology*, 49(45):604–608, **June 2002**. PMID: 12063950.

[31] A. Rashid, T. Ueki, Y.-T. Gao, P. S. Houlihan, C. Wallace, B.-S. Wang, M.-C. Shen, J. Deng, and A. W. Hsing. K-ras mutation, p53 overexpression, and microsatellite instability in biliary tract cancers: a population-based study in china. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 8(10):3156–3163, **October 2002**. PMID: 12374683.

[32] A. Tannapfel, F. Sommerer, M. Benicke, A. Katalinic, D. Uhlmann, H. Witzigmann, J. Hauss, and C. Wittekind. Mutations of the BRAF gene in cholangiocarcinoma but not in hepatocellular carcinoma. *Gut*, 52(5):706–712, **May 2003**. PMID: 12692057.

[33] S. M. Abraham and A. R. Clark. Dual-specificity phosphatase 1: a critical regulator of innate immune responses. *Biochemical Society transactions*, 34(Pt 6):1018–1023, **December 2006**. PMID: 17073741.

[34] S. S. Yang, L. Van Aelst, and D. Bar-Sagi. Differential interactions of human sos1 and sos2 with grb2. *The Journal of biological chemistry*, 270(31):18212–18215, **August 1995**. PMID: 7629138.

[35] S. M. Wilhelm, C. Carter, L. Tang, D. Wilkie, A. McNabola, H. Rong, C. Chen, X. Zhang, P. Vincent, M. McHugh, Y. Cao, J. Shujath, S. Gawlak, D. Eveleigh, B. Rowley, L. Liu, L. Adnane, M. Lynch, D. Auclair, I. Taylor, R. Gedrich, A.

Voznesensky, B. Riedl, L. E. Post, G. Bollag, and P. A. Trail. BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. *Cancer Research*, 64(19):7099–7109, **October 2004**. PMID: 15466206.

[36] A. Huether, M. Höpfner, V. Baradari, D. Schuppan, and H. Scherübl. Sorafenib alone or as combination therapy for growth control of cholangiocarcinoma. *Biochemical Pharmacology*, 73(9):1308–1317, **May 2007**.

[37] J. Lomax. Get ready to GO! a biologist's guide to the gene ontology. *Briefings in Bioinformatics*, 6(3):298–304, **September 2005**. PMID: 16212777.

[38] T. D. Tlsty and L. M. Coussens. Tumor stroma and regulation of cancer development. *Annual Review of Pathology: Mechanisms of Disease*, 1(1):119–150, **2006**. PMID: 18039110.

[39] A. E. Karnoub, A. B. Dash, A. P. Vo, A. Sullivan, M. W. Brooks, G. W. Bell, A. L. Richardson, K. Polyak, R. Tubo, and R. A. Weinberg. Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature*, 449(7162):557–563, **October 2007**.

[40] M. E. Lukashev and Z. Werb. ECM signalling: orchestrating cell behaviour and misbehaviour. *Trends in Cell Biology*, 8(11):437–441, **November 1998**.

## Supplementary information

| Gene ID | Symbol | PMID |
|---|---|---|
| 8 | AA | 12143044, 16966336 |
| 9 | NAT1 | 15901993 |
| 58 | ACTA1 | 8835746 |
| 259 | AMBP | 16712791 |
| 301 | ANXA1 | 16712791 |
| 302 | ANXA2 | 16712791 |
| 320 | APBA1 | 12213730 |
| 324 | APC | 10212000, 11260864, 11798885, 11866974, 12607585 |
| | | , 15467712 |
| 331 | XIAP | 15578516 |
| 355 | FAS | 10393851, 10482689 |
| 367 | AR | 16292515 |
| 399 | RHOH | 16627262 |
| 427 | ASAH1 | 16292515 |
| 444 | ASPH | 10728685 |
| 462 | SERPINC1 | 16712791 |
| 581 | BAX | 12668978 |
| 596 | BCL2 | 10617275, 9328309, 12201864, 15484294 |
| 598 | BCL2L1 | 11729212, 12235080 |
| 638 | BIK | 16865775 |
| 673 | BRAF | 12692057 |
| 761 | CA3 | 12927591 |
| 864 | RUNX3 | 15471559, 17470130 |
| 887 | CCKBR | 11281558 |
| 902 | CCNH | 12819026, 15017593 |
| 924 | CD7 | 16758167 |
| 934 | CD24 | 16125303, 17436128 |
| 947 | CD34 | 15484322, 16201082, 16465407 |
| 960 | CD44 | 9062875, 9537436, 14598145 |
| 968 | CD68 | 9869396, 16201082 |
| 1001 | CDH3 | 15880566 |
| 1017 | CDK2 | 12483273 |
| 1026 | CDKN1A | 11124819, 12829999 |
| 1027 | CDKN1B | 12717390 |
| 1029 | CDKN2A | 7796400, 10334895, 10718212, 10769677, 11034592 |
| | | , 11100316, 11340376, 12107841, 12210082, 12213730 |
| | | , 12360471, 12378511, 12668978, 12738733, 14647920 |
| | | , 15014024, 15467712, 15471559, 15619210, 15915369 |
| | | , 16373701 |
| 1045 | CDX2 | 15048136, 16116640, 16758167, 16794828, 17295772 |
| 1154 | CISH | 17241887 |
| 1163 | CKS1B | 17384652 |
| 1326 | MAP3K8 | 17621267 |

| | | |
|---|---|---|
| 1401 | CRP | 17006987 |
| 1440 | CSF3 | 16850133 |
| 1485 | CTAG1B | 15466353 |
| 1499 | CTNNB1 | 15288479 |
| 1544 | CYP1A2 | 15901993 |
| 1555 | CYP2B6 | 8896890 |
| 1604 | CD55 | 11131459 |
| 1611 | DAP | 12213730, 15471559 |
| 1612 | DAPK1 | 15467712, 15915369, 17690039 |
| 1630 | DCC | 10719364, 11004673, 11260864, 11798885, 11866974 |
| 1636 | ACE | 7975626, 15853991 |
| 1647 | GADD45A | 12829999 |
| 1806 | DPYD | 15884115, 16820886 |
| 1950 | EGF | 2839749 |
| 1956 | EGFR | 11261824, 12143054, 12360423, 12438243, 15213623 |
| | | , 15665568, 15833824, 15855163, 15921858, 16032426 |
| | | , 16116640, 16551849, 16790433, 16984600, 17079474 |
| 1977 | EIF4E | 12829998, 16982703 |
| 2056 | EPO | 16790433 |
| 2064 | ERBB2 | 11478488, 12044528 |
| 2065 | ERBB3 | 11261824 |
| 2150 | F2RL1 | 15492786 |
| 2195 | FAT | 16311342 |
| 2246 | FGF1 | 16201082 |
| 2272 | FHIT | 12668978, 12867802, 15625784, 16343073 |
| 2335 | FN1 | 8617420 |
| 2668 | GDNF | 11815973 |
| 2670 | GFAP | 16271085 |
| 2719 | GPC3 | 16162153 |
| 2744 | GLS | 12066193, 16204727 |
| 2778 | GNAS | 17356712 |
| 2950 | GSTP1 | 12213730, 12805482 |
| 3043 | HBB | 9444033 |
| 3082 | HGF | 11748452, 12143044, 15683434, 15921858, 16950403 |
| 3091 | HIF1A | 16094703 |
| 3159 | HMGA1 | 14714251 |
| 3198 | HOXA1 | 17550320 |
| 3265 | HRAS | 1739910, 2153451, 8957064, 10212000, 10469215 |
| | | , 10735605, 11034592, 11580146, 12063950, 15529594 |
| 3339 | HSPG2 | 2477943, 16292515 |
| 3429 | IFI27 | 17217824 |
| 3479 | IGF1 | 16773710 |
| 3480 | IGF1R | 16936263, 17266941 |
| 3481 | IGF2 | 17550320 |
| 3484 | IGFBP1 | 17006947 |
| 3486 | IGFBP3 | 17006947 |

| | | |
|---|---|---|
| 3569 | IL6 | 7827297, 9527063, 9762547, 10534331, 10869285 |
| | | , 14598145, 15917303, 15921858, 15940637, 16336976 |
| | | , 16469407, 17072955, 17079474, 17241887, 17621267 |
| 3659 | IRF1 | 9618305 |
| 3845 | KRAS | 10564951, 12692057, 17294242 |
| 3855 | KRT7 | 10617275, 10843291, 12747467, 15048994, 16680226 |
| | | , 16758167, 17342308, 17527078 |
| 3856 | KRT8 | 10617275, 11182037, 15048994, 16680226 |
| 3861 | KRT14 | 17405896 |
| 3880 | KRT19 | 9822920, 10617275, 12667323, 15048994, 15362741 |
| | | , 16627262, 16680226 |
| 3930 | LBR | 2477943 |
| 3956 | LGALS1 | 11274640 |
| 4072 | TACSTD1 | 10398165 |
| 4089 | SMAD4 | 12607585, 15573254, 16767220 |
| 4163 | MCC | 11798885, 11866974 |
| 4170 | MCL1 | 15940637, 16317687 |
| 4193 | MDM2 | 10718212, 10963376 |
| 4248 | MGAT3 | 15238704 |
| 4255 | MGMT | 12213730, 15467712, 15915369, 17550320 |
| 4288 | MKI67 | 10711445, 10895069, 10963376, 11029528, 11340376 |
| | | , 11490816, 11495045, 12028409, 17523309 |
| 4292 | MLH1 | 12175538, 12402306, 14506736, 15617839, 15915369 |
| | | , 16773692 |
| 4311 | MME | 16260277 |
| 4313 | MMP2 | 12439941, 15567754 |
| 4314 | MMP3 | 8675149 |
| 4316 | MMP7 | 11900228 |
| 4436 | MSH2 | 17051350 |
| 4507 | MTAP | 16373701 |
| 4524 | MTHFR | 17201138 |
| 4582 | MUC1 | 8903378, 11710692, 11920540, 12495310, 14752841 |
| | | , 15213623, 15690474, 16094706, 16124042, 17380013 |
| 4583 | MUC2 | 8903378, 16842244 |
| 4585 | MUC4 | 11680592, 14752841, 15690474 |
| 4586 | MUC5AC | 9495202, 9503463, 16116640, 16124042, 16679351 |
| | | , 16842244, 17397518 |
| 4588 | MUC6 | 9495202, 16842244 |
| 4589 | MUC7 | 9495202 |
| 4680 | CEACAM6 | 16868542 |
| 4684 | NCAM1 | 12031086, 17553067 |
| 4762 | NEUROG1 | 17550320 |
| 4771 | NF2 | 12668978, 16865775 |
| 4790 | NFKB1 | 15655831, 17524042 |
| 4843 | NOS2A | 11208728, 16094703 |
| 4968 | OGG1 | 11260864, 11798885, 11866974 |

| | | |
|---|---|---|
| 5111 | PCNA | 8097902, 9403719, 10711445, 11144921 |
| 5228 | PGF | 17006947 |
| 5243 | ABCB1 | 15884115 |
| 5268 | SERPINB5 | 15608662 |
| 5327 | PLAT | 1332997, 2411964, 2852388, 3028121 |
| 5465 | PPARA | 16927143 |
| 5578 | PRKCA | 11281558 |
| 5595 | MAPK3 | 16794828, 17294242, 17461449 |
| 5604 | MAP2K1 | 16950403, 17461449 |
| 5610 | EIF2AK2 | 11169059 |
| 5629 | PROX1 | 17069925 |
| 5704 | PSMC4 | 12819026 |
| 5708 | PSMD2 | 8174460, 11819329 |
| 5728 | PTEN | 12668978, 16767220 |
| 5731 | PTGER1 | 15855163, 17551669 |
| 5741 | PTH | 16850133 |
| 5743 | PTGS2 | 11870367, 14999691, 15921858, 16361272 |
| 5747 | PTK2 | 16270396, 16818654, 17687194 |
| 5894 | RAF1 | 12029618 |
| 5920 | RARRES3 | 15742394 |
| 5925 | RB1 | 11866974, 11798885 |
| 5967 | REG1A | 11343228 |
| 6046 | BRD2 | 15901993 |
| 6275 | S100A4 | 11029520 |
| 6277 | S100A6 | 12191675 |
| 6382 | SDC1 | 9771483 |
| 6383 | SDC2 | 11350606 |
| 6387 | CXCL12 | 16565491, 16792547 |
| 6446 | SGK1 | 15917303 |
| 6502 | SKP2 | 15596046 |
| 6508 | SLC4A3 | 11353065 |
| 6513 | SLC2A1 | 15362741 |
| 6521 | SLC4A1 | 2449824, 7539881, 10935654, 11353065, 11494016 |
| | | , 15836844 |
| 6696 | SPP1 | 15101999 |
| 6752 | SSTR2 | 7768398, 12483273 |
| 6774 | STAT3 | 16317687 |
| 6794 | STK11 | 12668978 |
| 6822 | SULT2A1 | 11228044 |
| 7012 | TERC | 10498642 |
| 7015 | TERT | 16168519 |
| 7031 | TFF1 | 14563942, 16830362, 17397518 |
| 7032 | TFF2 | 16830362 |
| 7033 | TFF3 | 16830362 |
| 7045 | TGFBI | 17006947 |
| 7057 | THBS1 | 9828214, 11927969, 12213730, 16465407 |

| | | |
|---|---|---|
| 7076 | TIMP1 | 8675149, 8707287 |
| 7078 | TIMP3 | 12213730 |
| 7080 | NKX2-1 | 16627262 |
| 7133 | TNFRSF1B | 15723721 |
| 7157 | TP53 | 7557145, 8280380, 10212000, 10362118, 10383693 |
| | | , 10469215, 10564951, 10711445, 10718212, 10735605 |
| | | , 10963376, 11144921, 11260864, 11340376, 11580146 |
| | | , 11798885, 11866974, 12063950, 12175538, 12378511 |
| | | , 15192787, 15529594,15794670, 15998419, 16596244 |
| | | , 16937443 |
| 7422 | VEGFA | 9828214, 12615726, 16465407, 16601431 |
| 7424 | VEGFC | 16601431, 16688800 |
| 7852 | CXCR4 | 16565491, 17461449 |
| 7869 | SEMA3B | 15704097 |
| 7905 | REEP5 | 12481012, 14506736 |
| 7979 | SHFM1 | 11906618 |
| 8312 | AXIN1 | 12668978 |
| 8647 | ABCB11 | 17452236 |
| 8714 | ABCC3 | 15884115 |
| 8743 | TNFSF10 | 10960444, 16166346 |
| 8795 | TNFRSF10B | 15150106 |
| 8797 | TNFRSF10A | 11495087, 15150106 |
| 8837 | CFLAR | 10573518, 14612959, 17071604 |
| 8851 | CDK5R1 | 14647920 |
| 8941 | CDK5R2 | 8916147 |
| 9021 | SOCS3 | 17241887 |
| 9166 | EBAG9 | 12044529, 12971966, 16048560 |
| 9235 | IL32 | 9881707 |
| 9338 | TCEAL1 | 2574140, 2839749, 3039284, 9618305, 10469215 |
| | | , 11340376, 12829999 |
| 9360 | PPIG | 8896890 |
| 9377 | COX5A | 11115827, 11870367, 14973068, 14999691, 15086604 |
| | | , 15855163, 15921858, 16966336, 17165088 |
| 9446 | GSTO1 | 15992993 |
| 10164 | CHST4 | 17006947 |
| 10197 | PSME3 | 1336666, 7557145, 7830335, 8280380, 8600693 |
| | | , 8826860, 10398901, 10735605, 10765128, 10895069 |
| | | , 11340376, 11798885, 11866974, 15966189, 16696030 |
| 10233 | LRRC23 | 12819026, 16673309 |
| 10561 | IFI44 | 10534331, 11124819, 12717393, 15014024 |
| 10573 | MRPL28 | 12107841 |
| 10660 | LBX1 | 17069925 |
| 10682 | EBP | 7805444, 8174474 |
| 10766 | TOB2 | 16865775 |
| 11162 | NUDT6 | 7513265, 9242465, 11478488 |
| 11186 | RASSF1 | 12399230, 15467712, 15704097, 16343073 |

| | | |
|---|---|---|
| 22915 | MMRN1 | 14559985 |
| 23643 | LY96 | 14615419 |
| 23645 | PPP1R15A | 16138120 |
| 26503 | SLC17A5 | 14614295, 16077978 |
| 27352 | SGSM3 | 16469407 |
| 29108 | PYCARD | 16911948 |
| 50818 | NULL | 11281558, 14604861 |
| 51364 | ZMYND10 | 15704097 |
| 53353 | LRP1B | 12668978 |
| 54474 | KRT20 | 9822920, 10843291, 16260277, 16679351 |
| 54575 | UGT1A10 | 9230212, 17006947 |
| 54657 | UGT1A4 | 9230212 |
| 55743 | CHFR | 17550320 |
| 57534 | MIB1 | 10435558, 10711445 |
| 79811 | SLTM | 11357901, 11895493, 14691921, 15690474, 15892172 |
| | | , 15921858, 16950403 |
| 80142 | PTGES2 | 15855163, 16966336 |
| 83478 | ARHGAP24 | 10362118, 11555575, 15467712 |
| 84676 | TRIM63 | 9618305 |
| 84909 | C9orf3 | 10393851, 10482689 |
| 94025 | MUC16 | 1332997, 1666951, 15948244 |
| 114049 | WBSCR22 | 17621267 |
| 119391 | GSTO2 | 15992993 |
| 133482 | SLCO6A1 | 1663474, 15992993 |
| 221002 | RASGEF1A | 17121879 |
| 285440 | CYP4V2 | 10765128 |
| 375790 | AGRN | 17640714 |
| 406991 | MIRN21 | 16762633 |

**Table 6.4:** Genes associated with CCC development and attributed GO term extracellular region coding for structural proteins.

# 7 Publication 3 - CellMiner-HCC: A microarray-based expression database for hepatocellular carcinoma cell lines

Published in Liver International in 2013

*Liver Int. 2013 Jul 31. doi: 10.1111/liv.12292*

Markus Krupp
Frank Staib
Thorsten Maass
Timo Itzel
Arndt Weinmann
Ju-Seog Lee
Bertil Schmidt
Martina Müller
Snorri S. Thorgeirsson
Peter R. Galle
Andreas Teufel

| Authors contribution: | |
|---|---|
| Draft writing | 50% |
| Review process | 50% |
| Data processing | 90% |
| Data analysis | 50% |
| Database design | 95% |
| GUI implementation | 95% |

## Abstract

**Background and aims:** Therapeutic options for hepatocellular carcinoma (HCC) still remain limited. Development of gene targeted therapies is a promising option. A better understanding of the underlying molecular biology is gained in in vitro experiments. However, even with targeted manipulation of gene expression varying treatment responses were observed in diverse HCC cell lines. Therefore, information on gene expression profiles of various HCC cell lines may be crucial to experimental designs. To generate a publicly available database containing microarray expression profiles of diverse HCC cell lines.
**Methods:** Microarray data were analyzed using an individually scripted R program package. Data were stored in a PostgreSQL database with a PHP written web interface. Evaluation and comparison of individual cell line expression profiles are supported via public web interface.
**Results:** This database allows evaluation of gene expression profiles of 18 HCC cell lines and comparison of differential gene expression between multiple cell lines. Analysis of commonly regulated genes for signaling pathway enrichment and interactions demonstrates a liver tumor phenotype with enrichment of major cancer related KEGG signatures like 'cancer' and 'inflammatory response'. Further molecular associations of strong scientific interest, e.g. 'lipid metabolism', were also identified.
**Conclusion:** We have generated CellMinerHCC (http://www.medicalgenomics.org/cellminerhcc), a publicly available database containing gene expression data of 18 HCC cell lines. This database will aid in the design of in vitro experiments in HCC research, because the genetic specificities of various HCC cell lines will be considered.

## Introduction

Hepatocellular carcinoma (HCC) is among the most common malignancies worldwide and its incidence is rising, especially in Asia and Sub-Saharan Africa, but also in Western countries. Simultaneously, the therapeutic options for this disease besides surgery still remain limited. Although both cellular changes that lead to HCC and the aetiological factors responsible for the majority of HCC cases have been recognized, the molecular pathogenesis of this disease still remains elusive [1]. However, key to achieve further progress in the therapy of HCC will rely on a better understanding of the underlying biology of HCC development and growth allowing the development of subsequent targeted therapies against essential molecular mechanisms.

In vitro cell culture experiments provide the opportunity of modelling the complex mechanisms of HCC development. Thus, in vitro experiments may be used to easily test gene expression and corresponding biological behaviour of HCC cells in order to identify genetic signatures that are related to tumour growth and proliferation and may therefore be potential therapeutic targets. Furthermore, these cells are used in xenograft models of HCC development in mice in order to study tumour development and biological behaviour in vivo.

Although these in vitro and in vivo approaches using immortalized tumour cell lines have provided some insights in the molecular mechanisms underlying HCC development, it was simultaneously and repeatedly noted that diverse cell lines show diverse biological behaviour in response to drug treatment or genetic manipulation [2].

These diverse forms of behaviour and underlying genetic signatures made it necessary to take these differences in gene expression signatures into account, especially while planning targeted strategies to influence the biological behaviour. For example, a potential target of interest may be expressed very low in a particular cell line and therefore not be suitable as a therapeutic target whereas other cell lines may exhibit a strong expression, which may very well be blocked by therapeutic strategies in order to slow down tumour growth or migration.

Being aware of these differences may be crucial to the outcome of an experiment and therefore should be of importance in experimental design. Thus, information on the diverse gene expression profiles of the HCC cell lines may be crucial to experimental designs of modelling HCC in vitro.

The completely assembled human genome provides the opportunity to approach this issue with a large amount of genetic information and high-throughput genomic data [3] . However, analysis of these data is not feasible for biologists or physicians not familiar with bioinformatics or at least microarray analysis. Even with a profound experience in microarray analysis, analysis of such data may not be done easily and would take a considerable amount of time. In conclusion, the available gene expression profiles of HCC cell lines must therefore be considered as not accessible for the very most of the hepatological community. As a result, gene expression

profiles of common HCC cell lines are still not considered in most publications on in vitro models studying molecular mechanisms of HCC development.
We have, therefore, established the CellMinerHCC database providing easy access to the microarray (raw) data on the 18 most commonly used HCC cell lines. Also, within this publicly accessible database, these data were made easily comparable in order to assist in future designing of in vitro HCC modelling.

## Material and Methods

### Data sources

Gene expression data were kindly provided by Lee and Thorgeirsson, Laboratory of Experimental Carcinogenesis, National Cancer Institute, National Institutes of Health, USA. The data include 18 HCC cell lines and normal liver samples. The HCC cell lines are the following: 7703, Focus, Hep3B, Hep3B-TR, Hep40, HepG2, HLE, HLF, HUH-1, HUH-6, HUH-7, PLC/PRF/5, SK-Hep1, SNU-182, SNU-387, SNU-389, SNU-449 and SNU-475. The HCC cell lines were handled as previously described [4] . As control for dual channel microarrays, a pool of total RNA from 19 normal liver samples was used [5] . Oligo microarrays were produced at the Advanced Technology Center at the National Cancer Institute, NIH, USA using 70-mer probes of 21 329 genes. Microarray experiments were performed as previously described [4], with the exception of dye swap experiments that were not included in this data set.

### Data organization, webinterface

The genetic profile of the 18 HCC cell lines were stored in a PostgreSQL database (http://www.postgresql.org) and made publically accessible and searchable through a web interface implemented in PHP (http://de.php.net) (Fig. 7.1) , as previously described [6] . The data may also be downloaded as text files. Furthermore, CellMinerHCC data were linked to multiple well established databases such as NCBI Entrez database [7] , HGNC [8] , HPRD [9] , OMIM [10] , BioGPS [11] , Nextbio [12] and Gent [13] . Furthermore, the data were connected to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14] and the Gene Ontology (GO) database [15] to enable an integrative comparison and querying for functional, molecular and signalling events. Moreover, CellMinerHCC is linked to the publically available, web-based tool DAVID ('Database for Annotation, Visualization and Integrated Discovery'; http://david.abcc.ncifcrf.gov/home.jsp), enabling further functional analysis [16, 17]. Finally, the CellMinerHCC was cross-linked to our expression profiling database on normal tissues by next-generation sequencing database RNA-Seq Atlas [18] as well as our liver-specific databases on molecular associations LoMA [6].

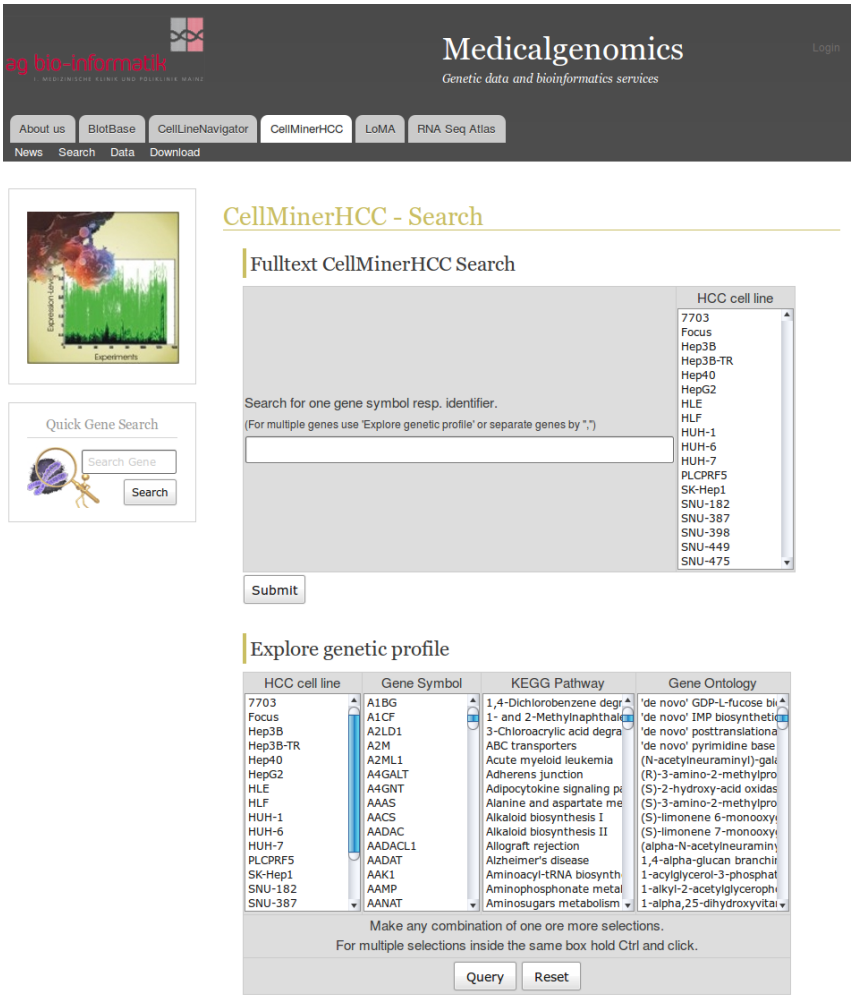


**Figure 7.1:** CellMinerHCC offers multiple search options. Searches may be performed by means of individual gene names, NCBI Gene IDs, Ensembl Gene IDs, or disease names. Also, more complex searches may be performed by selecting disease, gene symbol, a genetic pathway from KEGG or a gene ontology from the 'explore genetic association' panel.

### Functional analysis

For annotation of commonly regulated genes across all 18 HCC cell lines with GO-terms, KEGG pathways, and SP-PIR Keywords, as defined by the SwissProt/Uniprot and PIR groups, DAVID has been used with default settings [16, 17].
Ingenuity Pathway Analysis (http://www.ingenuity.com/) was performed using 195 commonly regulated genes (M-value >1, respectively <-1) across all 18 HCC cell lines applying default settings.

## Results

### CellMinerHCC database system

Diverse biological behaviour of HCC cell lines may result from diverse underlying genetic profiles and expression signatures, which may differ significantly among immortalized HCC tumour cell lines. The awareness of these differences made it necessary to take the diverse gene expression signatures into account, especially while planning targeted strategies to influence the biological behaviour.
However, the analysis of these data does not seem to be feasible for biologists or physicians not familiar with bioinformatics or at least microarray analysis. Even with profound experience in microarray analysis, analysis of such data is a complex and time-consuming task. We have, therefore, established the CellMinerHCC database providing easy access to the microarray (raw) data on the 18 most commonly used HCC cell lines. Currently, this database holds genome wide expression profiles of the 18 most commonly used HCC cell lines namely 7703, Focus, Hep3B, Hep3B-TR, Hep40, HepG2, HLE, HLF, HUH-1, HUH-6, HUH-7, PLC/PRF/5, SK-Hep1, SNU-182, SNU-387, SNU-389, SNU-449 and SNU-475. For all cell lines, our database contains the expression data of 21 329 genes. As some genes were represented by more than only one probe, expression data of all probes representing the same gene were averaged. In order to provide a rapid overview over the expression of individual genes in multiple cell lines, we furthermore colour coded the gene expression profiles. Mouse over the individual spots allows then to retrieve the detailed expression values for each gene and cell line.
The CellMinerHCC database provides multiple search options to support complex genetic analyses. Firstly, CellMinerHCC offers the option to search for individual genes and their genetic profile. This search may be performed by means of a search for individual gene names or cell lines or even a combination of both from the search page of the CellMinerHCC drop down menu. Also, more complex searches may be performed by a genetic pathway from KEGG [14], or a gene ontology from the 'explore genetic association' panel, providing a highly detailed search option. These search options together offer valuable complex analysis options (Fig. 7.1) . For example, one can now, for the first time, easily select all genes associated with the Wnt signalling pathway and display their expression profile as an overview in

the 18 HCC cell lines. Furthermore, one could search all gene products located to the nucleus by means of gene ontologies in order to identify transcription factors associated with the development of HCC and also display their expression profiles as a consolidated overview in all 18 HCC cell lines. In addition, the data search site offers the option of a free text search in order to provide a maximum flexibility in designing database queries (Fig. 7.2). After executing a search, the result page for these searches offers the genetic associations to individual diseases if present. Mainly, the results page provides a summary on gene name, corresponding NCBI Gene ID [7], Ensembl Gene ID [19], and most importantly the expression profiles of the gene in selected cell lines (Fig. 7.3). Additionally, the DAVID [16, 17] integration enables the user to further analyse their resulting gene set. Besides, more details on the specific gene such as gene alias names, chromosomal location, information on biological function, participation in biological processes and subcellular localization by supplying gene ontology information, and associated genetic pathways are accessible through the details button and respective site for each gene (Fig. 7.4). This information may be of significant value in designing complex and highly selective queries to the database.
Since local availability of the data in this repository may significantly speed up high-throughput searches for interested users, we also provide data in flat files for a complete download.
A key issue in developing this database was to provide the hepatological community with a powerful but simultaneously highly reliable and comprehensive database to perform systems biology-based high-throughput searches and comparison of gene expression. Our database was linked to multiple other bioinformatics resources, providing valuable connections and supporting advanced search and evaluation strategies. These links offer a strong backbone for bioinformatics research on chronic liver disease. In detail, CellMinerHCC has been linked to the most commonly used bioinformatics databases, such as PubMed [7], the European Bioinformatics Institute Website Ensembl [19], the Bioinformatics Resource of the National Center of Biotechnology Information Entrez Gene [7], and the Gene Ontology database, holding such as multiple sequence information, microarray expression data, conserved domains, as well as information on a gene's function [20].

## Differential gene expression in HCC cell lines

Evaluating all 18 HCC cell lines for differential gene expression with a cut-off of at least two-fold changes in gene expression, we identified between 1.638 genes in HepG2 and 3.214 genes in SNU-398 as being differentially regulated. A box-plot providing the distribution of differentially regulated genes with differentiation into up- and downregulated genes is provided in figure 7.8 in the supplement. The pattern of genes being upregulated vs. genes being downregulated was mostly evenly distributed. Solely in SNU-182 cells, significantly more genes were found to be downregulated as compared to genes upregulated. Distribution of M-values as a

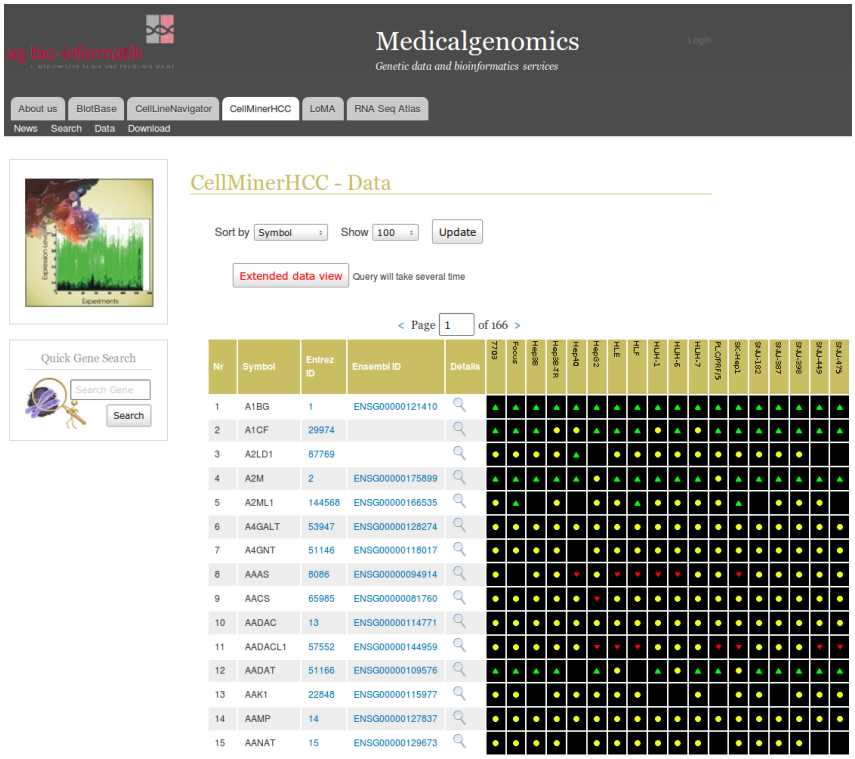


**Figure 7.2:** The CellMinerHCC data overview provides a colour-coded rapid overview of gene expression of diverse genes in HCC cell lines. Furthermore, it summarizes information on the HUGO gene symbol, Entrez Gene ID and Ensembl Gene ID. Mouse over any of the colour-coded gene expression data points provided the exact M-value of that particular gene as shown for the gene 'A2M' in the cell line 'Hep40'. Finally, to enable further analysis we provided a separate link for each gene forwarding to a details page providing a comprehensive source of individual molecular information.

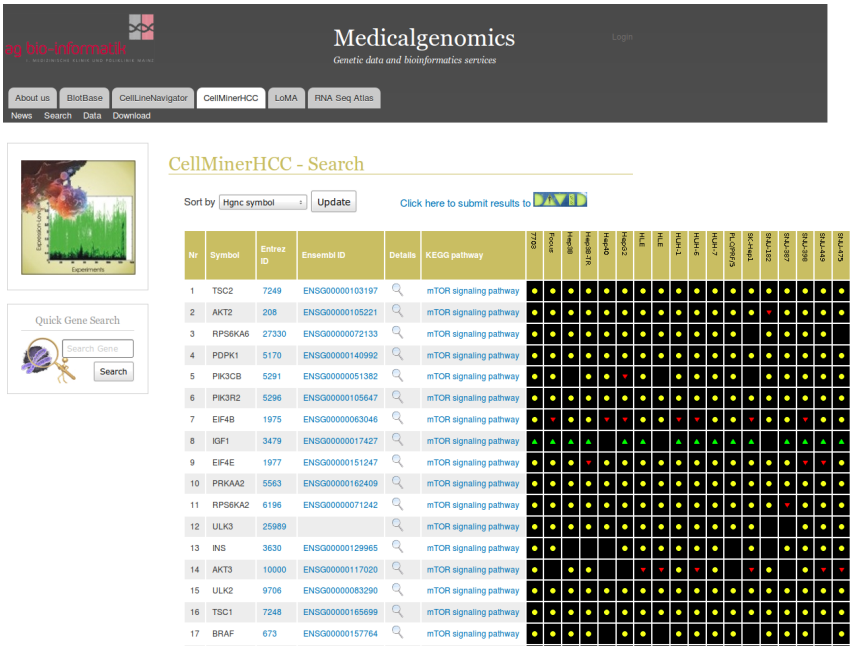


**Figure 7.3:** The results of a CellMinerHCC search for the KEGG pathway 'mTOR signalling' provide additional information on any individual gene by Entrez Gene IDs, Ensembl Gene IDs, associated KEGG pathways and Gene Ontologies. Furthermore, for each individual gene a separate link is provided forwarding to a details page providing comprehensive individual molecular information.

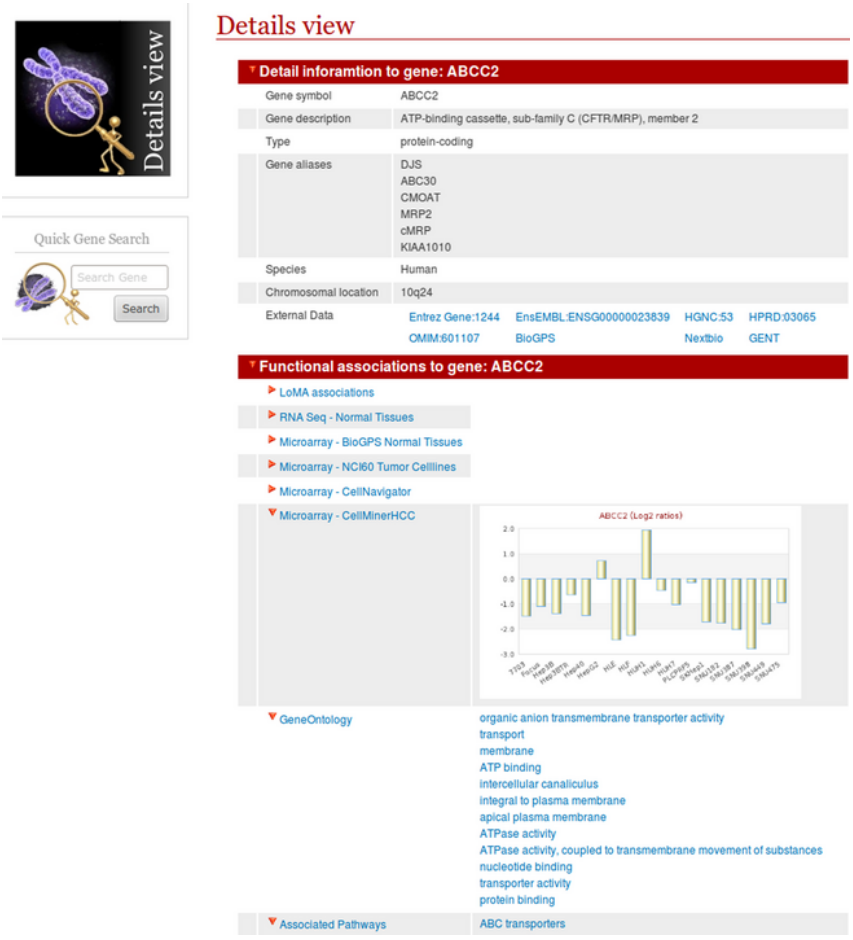


**Figure 7.4:** The 'Details' section of search results of the data view provides extensive additional information and linkage to gene description, alias names, chromosomal location and functional associations to CellMinerHCC data across all cell lines of this particular gene, associations found in our LOMA databank, gene ontology information as well as associated KEGG pathways.

measure of change in gene expression followed a 'normal distribution curve' in all cell lines ( Fig. 7.8 in supplement ). Of these, 195 genes were identified as consistently either up- or downregulated with 163 genes being downregulated and 32 genes being upregulated ( Tab. 7.1 in supplement). ).

### Biological functions of commonly regulated genes in HCC cell lines

Having established such a rich data resource, we were curious about the biological functions of these 195 most commonly regulated genes among these 18 cell lines as they may hold key functions related to liver cancer development. In order to obtain a comprehensive overview on the biological functions of these genes, we performed several advanced bioinformatics analyses. The list of 195 commonly regulated genes was imported into DAVID [16, 17]. The majority of significantly enriched GO-terms, KEGG pathways and SP-PIR Keywords are highly specific for liver and a malignant phenotype. Among the top 20 enriched categories were 'liver' (P = 6.5 × 10E-14), 'plasma' (P = 2.0 × 10E-25), 'complement and coagulation cascades' (P = 1.2 × 10E-14), 'metalloprotein' (P = 1.5 × 10E-7), 'extracellular space' (P = 1.1 × 10E-9) and 'disease mutation' (P = 8.3 × 10E-8) (Fig. 7.5). Furthermore, all

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| | SP_PIR_KEYWORDS | plasma | RT | | 25 | 13.4 | 5.1E-28 | 2.0E-25 |
| | KEGG_PATHWAY | Complement and coagulation cascades | RT | | 19 | 10.2 | 6.1E-17 | 1.2E-14 |
| | SP_PIR_KEYWORDS | liver | RT | | 15 | 8.0 | 3.1E-16 | 6.5E-14 |
| | SP_PIR_KEYWORDS | Secreted | RT | | 52 | 27.8 | 7.1E-14 | 9.2E-12 |
| | SP_PIR_KEYWORDS | acute phase | RT | | 11 | 5.9 | 9.1E-14 | 8.9E-12 |
| | GOTERM_BP_FAT | acute inflammatory response | RT | | 16 | 8.6 | 4.2E-13 | 6.3E-10 |
| | GOTERM_CC_FAT | extracellular space | RT | | 35 | 18.7 | 4.4E-12 | 1.1E-9 |
| | SP_PIR_KEYWORDS | signal | RT | | 67 | 35.8 | 5.0E-10 | 3.9E-8 |
| | UP_SEQ_FEATURE | signal peptide | RT | | 67 | 35.8 | 6.5E-10 | 4.3E-7 |
| | SP_PIR_KEYWORDS | disease mutation | RT | | 43 | 23.0 | 1.3E-9 | 8.3E-8 |
| | GOTERM_CC_FAT | extracellular region | RT | | 57 | 30.5 | 2.4E-9 | 3.0E-7 |
| | GOTERM_BP_FAT | response to wounding | RT | | 26 | 13.9 | 2.4E-9 | 1.8E-6 |
| | SP_PIR_KEYWORDS | metalloprotein | RT | | 13 | 7.0 | 2.7E-9 | 1.5E-7 |
| | GOTERM_BP_FAT | protein maturation by peptide bond cleavage | RT | | 12 | 6.4 | 4.5E-9 | 2.3E-6 |
| | UP_SEQ_FEATURE | domain:Peptidase S1 | RT | | 12 | 6.4 | 6.5E-9 | 2.1E-6 |
| | SP_PIR_KEYWORDS | blood coagulation | RT | | 9 | 4.8 | 8.3E-9 | 4.1E-7 |
| | UP_SEQ_FEATURE | domain:Sushi 2 | RT | | 9 | 4.8 | 8.7E-9 | 1.9E-6 |
| | UP_SEQ_FEATURE | domain:Sushi 1 | RT | | 9 | 4.8 | 8.7E-9 | 1.9E-6 |
| | GOTERM_CC_FAT | extracellular region part | RT | | 36 | 19.3 | 9.0E-9 | 7.7E-7 |
| | SMART | Tryp_SPc | RT | | 12 | 6.4 | 1.0E-8 | 8.4E-7 |

**Figure 7.5:** Top 20 DAVID annotation chart for the 195 commonly regulated genes across all 18 HCC cell lines sorted by their degree of significance. Terms are identified by gene enrichment and functional annotation analysis. The categories on their left side provide the original databases and resources where the enriched terms originate from. The number of genes involved in each particular term out of the 195 gene list is provided graphically, in total numbers, and in percentage to the list analysed. Finally, the P-value of the gene enrichment in relation to the human genome and the Benjamini correction for multiple testing are provided on the right.

commonly regulated genes were analysed by means of Ingenuity Pathway Anal-

ysis (Ingenuity® Systems, www.ingenuity.com) (Fig. 7.6). As a proof of principle, the highest ranking disease terms associated with these commonly regulated genes in HCC cell lines were 'cancer' (76 out of 195 genes; P = 3.8 × 10E-10 - 1.2 × 10E-2) followed by 'gastrointestinal disease' (54 out of 195 genes; P = 3.8 × 10E-10 - 1.2 × 10E-2), 'hepatic system disease' (31 out of 195 genes; P = 3.8 × 10E-10 - 1.2 × 10E-2) and 'inflammatory response' (23 out of 195 genes; P = 9.2 × 10E-7 - 1.2 × 10E-2). In addition, other networks also demonstrated a high association with HCC cell lines as well. In particular, the lipid metabolism ranked as third most significantly enriched category (36 out of 195 genes; P = 6.2 × 10E-9 - 1.2 × 10E-2). In concordance with the DAVID analysis pointing out a significant role of many genes in metabolic processes, these data point towards an important role of gene/protein members of the lipid metabolism functional pathway [21]. Furthermore, top biofunctions contained several categories related to immune functions. Among the top 20 biological functions, 'immunological disease' (32 out of 195 genes; P = 3.0 × 10E-7 - 1.2 × 10E-2), and 'immune cell trafficking' (19 out of 195 genes; P = 9.2 × 10E-7 - 1.2 × 10E-2) were identified. As further, 'diseases and disorders' term 'inflammatory response' (23 out of 195 genes; P = 9.2 × 10E-7 - 1.2 × 10E-2) was highly significant enriched. Thus, among the 20 highest ranked biological functions were three categories related to immune system and inflammatory responses indicating that immune system and inflammation may play an important role in this disease.
We were also interested in expression of hepatocellular cancer-like stem cell genes among those commonly regulated genes in all 18 HCC cell lines. Lee et al. previously described a hepatoblast-like HCC subgroup, which included the expression of hepatic oval cell marker genes like KRT7, KRT19 am VIM [22]. These genes were not found among the 195 commonly regulated genes. However, in a CellMinerHCC search, these genes have easily been identified as differentially regulated across the 18 HCC cell lines. All three marker genes are upregulated in SK-Hep1 and SNU-182, while they are downregulated in SNU-387. A recent study by Woo et al. described an eight gene stem cell-like signature associated with TP53 mutations and poor prognosis in patients with HCC. Again, these genes were not listed among the 195 commonly regulated genes. A CellMinerHCC search identified six out of these eight genes as minor regulated (within the two-fold change) except for ES1 (corresponds with HES1 in CellMinerHCC), which is significantly downregulated in most of the cell lines.
Altogether, these data indicate that these 195 commonly regulated genes across all 18 HCC cell lines not only reflect typical liver biofunctions and a liver cancer phenotype, there are also multiple additional enriched functions that may play an important role in HCC.

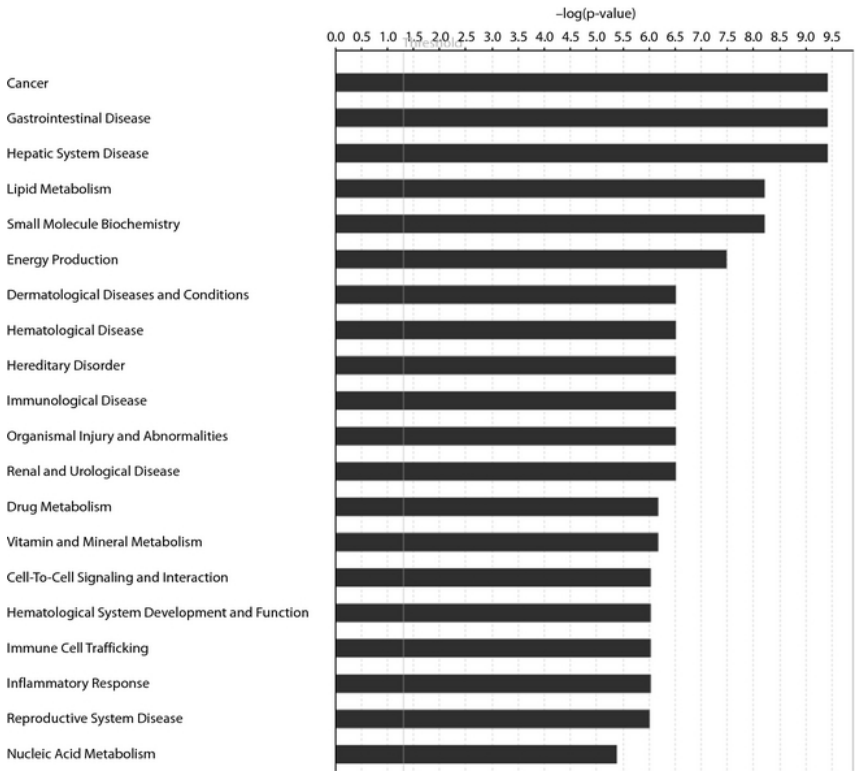


**Figure 7.6:** Top 20 significantly enriched biological functions identified by Ingenuity Pathway Analysis using the 195 commonly regulated genes across all 18 HCC cell lines. The bars represent the degree of significance (-log of the P-value) on the x-axis. The threshold line indicates the level above which the degree of enrichment has become statistically significant. The significant categories are provided along the y-axis.

## Discussion

Genetic mutations and a variable genetic background have been demonstrated to significantly influence the development and course of HCC as well as the efficiency of its treatment with diverse drugs.
Over the past decades, multiple molecular mechanisms and individual factors have been shown to be involved in the development of HCC and it has become clear that development and course of this disease underlay complex genetic interactions. We have, in the past, published a large genetic database summarizing genes known to be involved in HCC development [6]. However, these data do not represent a genome-wide screen as they were collected through a text mining approach of the known literature.
However, to investigate these complex molecular interactions, data resources providing a comprehensive collection of differential gene expression data involved in the development of HCC are urgently needed. We therefore present a novel database resource targeted to be of significant aid in the complex modelling of HCC development and evaluation of treatment options in vitro.
In vitro cell culture experiments provide the opportunity of modelling the complex mechanisms of HCC development. Thus, in vitro experiments may be used to easily test gene expression and corresponding biological behaviour of HCC cells in order to identify genetic signatures that are related to tumour growth and proliferation and may therefore be potential therapeutic targets. However, the diverse genetic backgrounds and differential gene expression profiles of these cell lines were not publicly available. We therefore analysed the 18 most commonly used liver cancer cell lines by means of microarray analysis in order to provide the basis for a more specific selection of liver cancer cell lines for future design of experiments in molecular hepatology research. To our knowledge, this is one of the first databases of its kind summarizing genome-wide gene expression profiles of cell lines of a specific tumour type. Thus, comparative transcriptomics analyses to similar databases were not possible. However, we thought it would be of definite interest to evaluate the genes being consistently over- or underexpressed in all cell lines as they may be crucial to the biological mechanisms of liver cancer development.
The analysis by DAVID revealed several GO-terms, KEGG pathways and SP-PIR keywords describing liver tissue and many of its physiological functions, which may at least partially be involved in HCC [16, 21] . Among these are 'complement and coagulation', a KEGG pathway that has been previously described to be possibly involved in HCC [23, 24], an 'acute inflammatory response' may be involved in carcinogenesis if an imbalance and prolongation occur [25], 'disease mutation' plays multiple roles in carcinogenesis in general, and 'metalloprotein' involved in the generation of the extracellular matrix may also be involved in HCC, when unequal distributions occur [26] and even more when their decomposing metalloproteinases are disturbed [27]. These findings are complemented and further differentiated by Ingenuity Pathway Analysis. Here, the disease terms describe a significant enrichment and association of 76 out of 195 commonly regulated genes in all

18 cell lines with 'cancer'. With 54 and 31 out of 195 commonly regulated genes, the terms 'gastrointestinal disease' and a 'hepatic system disease' provide a more general heading for HCC. Thus, many of these 195 genes have already been associated with cancerous diseases in the gastrointestinal tract. The disease term 'inflammatory response' is complementary to the DAVID analysis revealing the GO term 'acute inflammatory response'. The link between ongoing acute inflammation and HCC has become increasingly tight [28, 29]. Identification of these categories in this gene set is providing additional evidence for this association. Categories like 'immunological disease' connect the inflammatory response with the immune system that is involved in HCC. Enrichment of additional categories like 'humoral immune response' and 'immune cell trafficking' point towards the important role the immune system is playing in development and progression of HCC [30]. Finally, the category 'lipid metabolism' has been identified as third highest ranked biological function when analysing the 195 commonly regulated HCC cell line genes. For aberrant lipid metabolism, an association with HCC has already been described [31] and may also play a role in obesity and chronic inflammation related development of HCC [21, 29]. Taken together, analysis of the commonly regulated genes among the 18 most often used HCC cell lines for enrichment of signalling pathways, proteins and interactions not only described a liver tumour phenotype, it also identified molecular associations and numerous categories currently under intense scientific development.

Over the past decade, it has become increasingly clear that stem cells are not only beneficial and that tumours of various origins also contain stem cells that help them to proliferate but also to escape conventional chemotherapy [32, 33]. Such cancer stem cells were identified in multiple tumours. The existence of cancer stem cells in hepatocellular carcinoma still remains elusive. However, there is accumulating evidence for the concept of cancer stem cells in HCC similar to other solid tumours, where cancer stem cells have already been conclusively identified [34, 35]. We therefore analysed our differentially regulated genes for the appearance of stem cell markers. However, among the 195 genes consistently regulated in all 18 HCC cell lines, we did not find any of the conventional stem cell markers. In addition, searching the 195 commonly expressed genes in all 18 HCC cell lines for expression of a set of 3 classical markers for hepatic oval cells (KRT7, KRT19 and VIM) that have previously been identified among a group of genes defining a hepatoblast HCC subgroup associated with poor prognosis did not reveal any hit [36]. However, running a CellMinerHCC search for these three target genes revealed a differential expression pattern across all cell lines. The graphical output of CellMinerHCC made identification of those cell lines with up- or downregulation of all three marker genes easy indicating its usefulness and easy-to-use properties.

Since our database is the first of its kind and such a novel tool in HCC research, we set a high value on a user friendly but at the same time usability for advanced bioinformatics analyses. To guarantee easy data access and connectivity, we transformed this database into a powerful web application. Since multiple systems biology queries and analyses require complex connections between information from

diverse databases, the provided links offer a strong backbone for bioinformatics research on chronic liver disease.

We furthermore used our novel genetic resource in HCC research to identify key issues in transformed HCC cell lines such as the enrichment of genetic signalling pathways or biological functions within these complex transcriptomics queries. We therefore realized a rich embedding of our database into the current scenery of bioinformatics repositories providing valuable connections, which may support advanced search and evaluation strategies. The provided links to further bioinformatics repositories were selected as they may in addition support automated correlation with additional genomic information such as multiple sequence information, microarray expression data, conserved domains, as well as information on a gene's function. Since many users of a preliminary version of our database were interested in genetic pathway analysis, we made this information instantly available for advanced queries through a drop down menu at the front page of the data search option.

Altogether, CellMinerHCC is the first database providing a comprehensive view and analysis options for microarray data of the most commonly used HCC cell lines and may be of significant use for in vitro modelling of HCC. CellMinerHCC is freely accessible at http://www.medicalgenomics.org/cellminerhcc.

# BIBLIOGRAPHY


[1] T. Maass, I. Sfakianakis, F. Staib, M. Krupp, P. R. Galle, and A. Teufel. Microarray-based gene expression analysis of hepatocellular carcinoma. *Current Genomics*, 11(4):261–268, **June 2010**.

[2] K. Herzer, T. G. Hofmann, A. Teufel, C. C. Schimanski, M. Moehler, S. Kanzler, H. Schulze-Bergkamen, and P. R. Galle. IFN-alpha-induced apoptosis in hepatocellular carcinoma involves promyelocytic leukemia protein and TRAIL independently of p53. *Cancer research*, 69(3):855–862, **February 2009**. PMID: 19141642.

[3] J. F. Hocquette. Where are we in genomics? *Journal of physiology and pharmacology: an official journal of the Polish Physiological Society*, 56 Suppl 3:37–70, **June 2005**. PMID: 16077195.

[4] J.-S. Lee and S. S. Thorgeirsson. Functional and genomic implications of global gene expression profiles in cell lines from human hepatocellular cancer. *Hepatology*, 35(5):1134–1143, **2002**.

[5] J.-S. Lee, I.-S. Chu, J. Heo, D. F. Calvisi, Z. Sun, T. Roskams, A. Durnez, A. J. Demetris, and S. S. Thorgeirsson. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*, 40(3):667–676, **2004**.

[6] S. Buchkremer, J. Hendel, M. Krupp, A. Weinmann, K. Schlamp, T. Maass, F. Staib, P. R. Galle, and A. Teufel. Library of molecular associations: curating the complex molecular basis of liver diseases. *BMC Genomics*, 11(1):189, **March 2010**. PMID: 20302666.

[7] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, **January 2011**. PMID: 21097890.

[8] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, and E. A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, 39(suppl 1):D514–D519, **January 2011**. PMID: 20929869.

[9] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database—2009 update. *Nucleic Acids Research*, 37(suppl 1):D767–D772, **January 2009**. PMID: 18988627.

[10] V. A. McKusick. Mendelian inheritance in man and its online version, OMIM. *The American Journal of Human Genetics*, 80(4):588–604, **April 2007**.

[11] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, J. W. Huss, and A. I. Su. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10(11):R130, **November 2009**. PMID: 19919682.

[12] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS ONE*, 5(9):e13066, **September 2010**.

[13] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim. GENT: gene expression database of normal and tumor tissues. *Cancer informatics*, 10:149–157, **2011**. PMID: 21695066.

[14] K. F. Aoki and M. Kanehisa. Using the KEGG database resource. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 1:Unit 1.12, **October 2005**. PMID: 18428742.

[15] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25–29, **May 2000**. PMID: 10802651.

[16] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, **January 2009**. PMID: 19033363.

[17] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, **December 2008**.

[18] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel. RNA-Seq atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, 28(8):1184–1185, **April 2012**. PMID: 22345621.

[19] G. Spudich, X. M. Fernández-Suárez, and E. Birney. Genome browsing with ensembl: a practical overview. *Briefings in Functional Genomics & Proteomics*, 6(3):202–219, **September 2007**. PMID: 17967807.

[20] T. G. O. Consortium. Creating the gene ontology resource: Design and implementation. *Genome Research*, 11(8):1425–1433, **August 2001**. PMID: 11483584.

[21] D. Becker, I. Sfakianakis, M. Krupp, F. Staib, A. Gerhold-Ay, A. Victor, H. Binder, M. Blettner, T. Maass, S. Thorgeirsson, P. R. Galle, and A. Teufel. Genetic signatures shared in embryonic liver development and liver cancer define prognostically relevant subgroups in HCC. *Molecular Cancer*, 11(1):55, **August 2012**. PMID: 22891627.

[22] J.-S. Lee, J. Heo, L. Libbrecht, I.-S. Chu, P. Kaposi-Novak, D. F. Calvisi, A. Mikaelyan, L. R. Roberts, A. J. Demetris, Z. Sun, F. Nevens, T. Roskams, and S. S. Thorgeirsson. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nature Medicine*, 12(4):410–416, **April 2006**.

[23] T. Li, B. Wan, J. Huang, and X. Zhang. Comparison of gene expression in hepatocellular carcinoma, liver development, and liver regeneration. *Molecular Genetics and Genomics*, 283(5):485–492, **May 2010**.

[24] E. Wurmbach, Y.-b. Chen, G. Khitrov, W. Zhang, S. Roayaie, M. Schwartz, I. Fiel, S. Thung, V. Mazzaferro, J. Bruix, E. Bottinger, S. Friedman, S. Waxman, and J. M. Llovet. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology*, 45(4):938–947, **2007**.

[25] A. J. Schetter, N. H. H. Heegaard, and C. C. Harris. Inflammation and cancer: interweaving microRNA, free radical, cytokine and p53 pathways. *Carcinogenesis*, 31(1):37–49, **January 2010**. PMID: 19955394.

[26] Y. Liu, L. Li, Y. Gao, C. Chen, B. Li, W. He, Y. Huang, and Z. Chai. Distribution of metalloproteins in hepatocellular carcinoma and surrounding tissues. *Hepato-gastroenterology*, 54(80):2291–2296, **December 2007**. PMID: 18265650.

[27] P. Vihinen and V.-M. Kähäri. Matrix metalloproteinases in cancer: Prognostic markers and therapeutic targets. *International Journal of Cancer*, 99(2):157–166, **2002**.

[28] G. Castello, S. Scala, G. Palmieri, S. A. Curley, and F. Izzo. HCV-related hepatocellular carcinoma: From chronic inflammation to cancer. *Clinical Immunology*, 134(3):237–250, **March 2010**.

[29] S. Toffanin, S. L. Friedman, and J. M. Llovet. Obesity, inflammatory signaling, and hepatocellular Carcinoma—An enlarging link. *Cancer Cell*, 17(2):115–117, **February 2010**.

[30] A. Budhu, M. Forgues, Q.-H. Ye, H.-L. Jia, P. He, K. A. Zanetti, U. S. Kammula, Y. Chen, L.-X. Qin, Z.-Y. Tang, and X. W. Wang. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell*, 10(2):99–111, **August 2006**.

[31] J.-M. Wu, N. J. Skill, and M. A. Maluccio. Evidence of aberrant lipid metabolism in hepatitis c and hepatocellular carcinoma. *HPB*, 12(9):625–636, **2010**.

[32] T. K. W. Lee, A. Castilho, S. Ma, and I. O. L. Ng. Liver cancer stem cells: implications for a new therapeutic target. *Liver International*, 29(7):955–965, **2009**.

[33] A. Teufel and P. R. Galle. Collecting evidence for a stem cell hypothesis in HCC. *Gut*, 59(7):870–871, **July 2010**. PMID: 20581232.

[34] F. Iovino, S. Meraviglia, M. Spina, V. Orlando, V. Saladino, F. Dieli, G. Stassi, and M. Todaro. Immunotherapy targeting colon cancer stem cells. *Immunotherapy*, 3(1):97–106, **January 2011**.

[35] J. Papailiou, K. J. Bramis, M. Gazouli, and G. Theodoropoulos. Stem cells in colon cancer. a new era in cancer theory begins. *International Journal of Colorectal Disease*, 26(1):1–11, **January 2011**.

[36] H. G. Woo, X. W. Wang, A. Budhu, Y. H. Kim, S. M. Kwon, Z. Tang, Z. Sun, C. C. Harris, and S. S. Thorgeirsson. Association of TP53 mutations with stem cell-like gene expression and survival of patients with hepatocellular carcinoma. *Gastroenterology*, 140(3):1063–1070.e8, **March 2011**.

## Supplementary information

| Symbol | Entrez ID | Entrez Gene Name | Fold Change |
|---|---|---|---|
| A1BG | 1 | alpha-1-B glycoprotein | -4.76 |
| ACAA1 | 30 | acetyl-CoAacyltransferase 1 | -2.46 |
| ACAA2 | 10449 | acetyl-CoAacyltransferase 2 | -1.9 |
| ACP2 | 53 | acidphosphatase 2, lysosomal | -2.31 |
| ACSL1 | 2180 | acyl-CoA synthetase long-chain family member 1 | -4.24 |
| ACSM2A | 233799 | acyl-CoA synthetase medium-chain family member 2A | -4.03 |
| ACTL6A | 86 | actin-like 6A | 1.9 |
| ADH4 | 127 | alcohol dehydrogenase 4 (class II), pi polypeptide | -4.13 |
| AGL | 178 | amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase | -3.08 |
| ALDH2 | 217 | aldehydedehydrogenase 2 family (mitochondrial) | -2.45 |
| ALDH6A1 | 4329 | aldehyde dehydrogenase 6 family, member A1 | -3.14 |
| ALDOA | 226 | aldolase A, fructose-bisphosphate | 1.94 |
| ALDOB | 229 | aldolase B, fructose-bisphosphate | -4.98 |
| ANGPTL3 | 27329 | angiopoietin-like 3 | -4.06 |
| AP3B2 | 8120 | adaptor-related protein complex 3, beta 2 subunit | -1.94 |
| APCS | 325 | amyloid P component, serum | -4.7 |
| APOC2 | 344 | apolipoprotein C-II | -4.02 |
| APOC3 | 345 | apolipoprotein C-III | -3.94 |
| ASPRV1 | 151516 | asparticpeptidase, retroviral-like 1 | -1.82 |
| BAG3 | 9531 | BCL2-associated athanogene 3 | -3.88 |
| BHLHE40 | 8553 | basic helix-loop-helix family, member e40 | -2.13 |
| BHMT | 635 | betaine–homocysteine S-methyltransferase | -4.45 |
| BTD | 686 | biotinidase | -1.83 |
| C1orf183 | 55924 | chromosome 1 open readingframe 183 | -2.73 |
| C1QTNF4 | 114900 | C1q and tumor necrosis factor related protein 4 | -1.98 |
| C1R | 715 | complementcomponent 1, r subcomponent | -4.7 |
| C1S | 716 | complementcomponent 1, s subcomponent | -3.47 |
| C20orf160 | 140706 | chromosome 20 open readingframe 160 | -4.54 |
| C20orf70 | 140683 | chromosome 20 open readingframe 70 | -1.45 |
| C4orf34 | 201895 | chromosome 4 open readingframe 34 | -1.86 |
| C6 | 729 | complementcomponent 6 | -4.15 |
| C9 | 735 | complementcomponent 9 | -4.75 |
| CAT | 847 | catalase | -1.85 |
| CCDC103 | 388389 | coiled-coildomaincontaining 103 | -3.11 |
| CCNA2 | 890 | cyclin A2 | 2.28 |
| CCNB1 | 891 | cyclin B1 | 2.11 |
| CCT7 | 10574 | chaperonin containing TCP1, subunit 7 (eta) | 1.75 |
| CES2 | 8824 | carboxylesterase 2 | -3 |
| CFB | 629 | complementfactor B | -4.14 |
| CFH | 3075 | complementfactor H | -5.15 |
| CFHR2 | 3080 | complementfactor H-related 2 | -6.04 |
| CFHR3 | 10878 | complementfactor H-related 3 | -4.03 |

| | | | |
|---|---|---|---|
| CFI | 3426 | complementfactor I | -3.55 |
| CLIC1 | 1192 | chlorideintracellularchannel 1 | 2.08 |
| CLRN1 | 7401 | clarin 1 | -4.42 |
| COX19 | 90639 | COX19 cytochrome c oxidase assembly homolog (S. cerevisiae) | -2 |
| CP | 1356 | ceruloplasmin (ferroxidase) | -4.34 |
| CPA5 | 93979 | carboxypeptidase A5 | -2.37 |
| CPB2 | 1361 | carboxypeptidase B2 (plasma) | -4.6 |
| CS | 1431 | citratesynthase | 1.82 |
| CSAG2 | 9598 | CSAG family, member 2 | -2.54 |
| CSMD3 | 114788 | CUB and Sushi multiple domains 3 | -2.91 |
| CTH | 1491 | cystathionase (cystathionine gamma-lyase) | -1.78 |
| CYB5A | 1528 | cytochrome b5 type A (microsomal) | -3.07 |
| CYP2C9 | 1559 | cytochrome P450, family 2, subfamily C, polypeptide 9 | -3.87 |
| CYP2D6 | 1565 | cytochrome P450, family 2, subfamily D, polypeptide 6 | -3.9 |
| CYP2E1 | 1571 | cytochrome P450, family 2, subfamily E, polypeptide 1 | -3.87 |
| CYP3A4 | 1576 | cytochrome P450, family 3, subfamily A, polypeptide 4 | -4.92 |
| CYP3A7 | 1551 | cytochrome P450, family 3, subfamily A, polypeptide 7 | -5.2 |
| CYP4F2 | 8529 | cytochrome P450, family 4, subfamily F, polypeptide 2 | -3.24 |
| CYTH3 | 9265 | cytohesin 3 | -2.31 |
| DECR1 | 1666 | 2,4-dienoyl CoAreductase 1, mitochondrial | -1.99 |
| DLGAP3 | 58512 | discs, large (Drosophila) homolog-associated protein 3 | -2.14 |
| DMGDH | 29958 | dimethylglycinedehydrogenase | -3.23 |
| DPEP1 | 1800 | dipeptidase 1 (renal) | -2.47 |
| DPYS | 1807 | dihydropyrimidinase | -2.74 |
| ENG | 2022 | endoglin | -2.1 |
| F12 | 2161 | coagulation factor XII (Hageman factor) | -3.9 |
| F9 | 2158 | coagulationfactor IX | -3.33 |
| FAM189A1 | 23359 | family with sequence similarity 189, member A1 | -2.43 |
| FBL | 2091 | fibrillarin | 2.25 |
| FGA | 2243 | fibrinogenalphachain | -3.98 |
| FGG | 2266 | fibrinogengammachain | -3.94 |
| FLJ30679 | 146512 | hypotheticalprotein FLJ30679 | -3.06 |
| GAD2 | 2572 | glutamate decarboxylase 2 (pancreatic islets and brain, 65kDa) | -3.56 |
| GADD45B | 4616 | growth arrest and DNA-damage-inducible, beta | -2.66 |
| GATM | 2628 | glycine amidinotransferase (L-arginine:glycineamidinotransferase) | -3.79 |
| GC | 2638 | group-specific component (vitamin D binding protein) | -3.84 |
| GCC2 | 9648 | GRIP and coiled-coil domain containing 2 | -2.79 |
| GHR | 2690 | growthhormonereceptor | -3.61 |
| GLYATL1 | 92292 | glycine-N-acyltransferase-like 1 | -3.46 |
| GNE | 10020 | glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase | -2.14 |
| GNMT | 27232 | glycine N-methyltransferase | -2.54 |
| GRB14 | 2888 | growth factor receptor-bound protein 14 | -4 |
| GSTZ1 | 2954 | glutathionetransferasezeta 1 | -1.5 |

| | | | |
|---|---|---|---|
| GZMB | 3002 | granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1) | -2.28 |
| HABP2 | 3026 | hyaluronanbindingprotein 2 | -3.45 |
| HBB | 3043 | hemoglobin, beta | -3.89 |
| HCLS1 | 3059 | hematopoietic cell-specific Lyn substrate 1 | -2.58 |
| HGFAC | 3083 | HGF activator | -3.05 |
| HIST4H4 | 8364 | histonecluster 1, H4c | 2.15 |
| HLA-DRA | 3122 | major histocompatibility complex, class II, DR alpha | -4.86 |
| HMGCS2 | 3158 | 3-hydroxy-3-methylglutaryl-CoA synthase 2 (mitochondrial) | -3.01 |
| HNRNPA1 | 3178 | heterogeneousnuclearribonucleoprotein A1 | 1.83 |
| HPR | 3250 | haptoglobin-relatedprotein | -3.24 |
| HRG | 3273 | histidine-richglycoprotein | -3.63 |
| HSD17B6 | 8630 | hydroxysteroid (17-beta) dehydrogenase 6 homolog (mouse) | -3 |
| IGHG1 | 3500 | immunoglobulin heavy constantgamma 1 (G1m marker) | -3.35 |
| IL3 | 3562 | interleukin 3 (colony-stimulating factor, multiple) | -4.01 |
| INE1 | 8552 | inactivation escape 1 (non-protein coding) | -2.39 |
| IQCB1 | 9657 | IQ motifcontaining B1 | -3.53 |
| ITIH1 | 3697 | inter-alpha (globulin) inhibitor H1 | -3.15 |
| ITIH4 | 3700 | inter-alpha (globulin) inhibitor H4 (plasmaKallikrein-sensitive glycoprotein) | -3.38 |
| KCTD4 | 386618 | potassium channel tetramerisation domain containing 4 | -4.61 |
| KHDRBS1 | 10657 | KH domain containing, RNA binding, signal transduction associated 1 | 1.59 |
| KNG1 | 3827 | kininogen 1 | -4.28 |
| LBP | 3929 | lipopolysaccharidebindingprotein | -3.94 |
| LOC150371 | 150371 | hypothetical LOC150371 | -1.79 |
| LRFN5 | 145581 | leucine rich repeat and fibronectin type III domain containing 5 | -2.47 |
| MBL2 | 4153 | mannose-binding lectin (protein C) 2, soluble (opsonic defect) | -2.86 |
| MCM2 | 4171 | minichromosomemaintenancecomplexcomponent 2 | 1.62 |
| MCM3 | 4172 | minichromosomemaintenancecomplexcomponent 3 | 2.11 |
| MFGE8 | 4240 | milk fat globule-EGF factor 8 protein | -2.49 |
| MGC12982 | 84793 | hypotheticalprotein MGC12982 | -4.27 |
| MRM1 | 79922 | mitochondrial rRNAmethyltransferase 1 homolog (S. cerevisiae) | -1.89 |
| MSH2 | 4436 | mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) | 2.3 |
| MTERFD1 | 51001 | MTERF domaincontaining 1 | -3.97 |
| MUT | 4594 | methylmalonylCoAmutase | -2.26 |
| NAP1L1 | 4673 | nucleosomeassemblyprotein 1-like 1 | 2.49 |
| NCRNA00176 | 284739 | non-protein coding RNA 176 | -3.87 |
| NDRG2 | 57447 | NDRG familymember 2 | -3.09 |
| NGEF | 25791 | neuronal guanine nucleotide exchange factor | -4.08 |
| NME2 | 4831 | non-metastatic cells 2, protein (NM23B) expressed in | 1.75 |
| NMRAL1 | 57407 | NmrA-like family domain containing 1 | -2.28 |

| | | | |
|---|---|---|---|
| NNMT | 4837 | nicotinamide N-methyltransferase | -3.31 |
| NPM1 | 4869 | nucleophosmin (nucleolarphosphoprotein B23, numatrin) | 2.25 |
| NR1H3 | 10062 | nuclear receptor subfamily 1, group H, member 3 | -2.18 |
| NUP205 | 23165 | nucleoporin 205kDa | 1.98 |
| ORM1/ORM2 | 5005 | orosomucoid 1 | -2.55 |
| OTC | 5009 | ornithinecarbamoyltransferase | -2.89 |
| PBLD | 64081 | phenazine biosynthesis-like protein domain containing | -3.38 |
| PCSK6 | 5046 | proproteinconvertasesubtilisin/kexin type 6 | -2.75 |
| PDK2 | 5164 | pyruvatedehydrogenasekinase, isozyme 2 | -2.38 |
| PFDN4 | 5203 | prefoldinsubunit 4 | 2.35 |
| PFN2 | 5217 | profilin 2 | 2.46 |
| PKM2 | 5315 | pyruvatekinase, muscle | 2.63 |
| PLA2G2A | 5320 | phospholipase A2, group IIA (platelets, synovial fluid) | -3.03 |
| PLG | 5340 | plasminogen | -4.67 |
| PLGLB1/PLGLB2 | 5342 | plasminogen-like B2 | -4.49 |
| PON3 | 5446 | paraoxonase 3 | -3.75 |
| PPP1CC | 5501 | protein phosphatase 1, catalytic subunit, gamma isozyme | 1.77 |
| PPP1R13L | 10848 | protein phosphatase 1, regulatory (inhibitor) subunit 13 like | -2.91 |
| PPP1R1A | 5502 | protein phosphatase 1, regulatory (inhibitor) subunit 1A | -2.89 |
| PROS1 | 5627 | protein S (alpha) | -2.94 |
| PRSS22 | 64063 | protease, serine, 22 | -2.76 |
| PTGES3 | 10728 | prostaglandin E synthase 3 (cytosolic) | 1.64 |
| PTPRCAP | 5790 | protein tyrosine phosphatase, receptor type, C-associated protein | -1.74 |
| RAN | 5901 | RAN, member RAS oncogene family | 2.13 |
| RCC2 | 55920 | regulatorofchromosomecondensation 2 | 2.02 |
| RDH5 | 5959 | retinoldehydrogenase 5 (11-cis/9-cis) | -1.77 |
| RELB | 5971 | v-relreticuloendotheliosis viral oncogene homolog B | -4.13 |
| RGS9 | 8787 | regulator of G-protein signaling 9 | -4.41 |
| RRAD | 6236 | Ras-related associated with diabetes | -2.48 |
| RRH | 10692 | retinal pigment epithelium-derived rhodopsin homolog | -3.28 |
| RTBDN | 83546 | retbindin | -1.78 |
| SAE1 | 10055 | SUMO1 activatingenzymesubunit 1 | 1.68 |
| SCGB1A1 | 7356 | secretoglobin, family 1A, member 1 (uteroglobin) | -2.09 |
| SDS | 10993 | serinedehydratase | -2.67 |
| SEC14L2 | 23541 | SEC14-like 2 (S. cerevisiae) | -1.62 |
| SERPINA3 | 12 | serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 | -3.63 |
| SERPINA6 | 866 | serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6 | -2.87 |
| SERPINC1 | 462 | serpin peptidase inhibitor, clade C (antithrombin), member 1 | -5.02 |
| SERPINF2 | 5345 | serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 | -3.43 |

| | | | |
|---|---|---|---|
| SERPING1 | 710 | serpin peptidase inhibitor, clade G (C1 inhibitor), member 1 | -3.87 |
| SLC17A7 | 57030 | solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 7 | -3.05 |
| SLC1A5 | 6510 | solute carrier family 1 (neutral amino acid transporter), member 5 | 2.07 |
| SLC22A1 | 6580 | solute carrier family 22 (organic cation transporter), member 1 | -4.07 |
| SLC27A5 | 10998 | solute carrier family 27 (fatty acid transporter), member 5 | -2.37 |
| SLC35C1 | 55343 | solute carrier family 35, member C1 | -2.05 |
| SLC38A3 | 10991 | solutecarrierfamily 38, member 3 | -1.91 |
| SLITRK6 | 84189 | SLIT and NTRK-like family, member 6 | -3.91 |
| SNCG | 6623 | synuclein, gamma (breast cancer-specific protein 1) | -2.76 |
| STAG3 | 10734 | stromalantigen 3 | -2.37 |
| SULT2A1 | 6822 | sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1 | -3.58 |
| TAT | 6898 | tyrosineaminotransferase | -3.51 |
| TBC1D29 | 26083 | TBC1 domainfamily, member 29 | -2.84 |
| TBX4 | 9496 | T-box 4 | -3.78 |
| TCTE3 | 6991 | t-complex-associated-testis-expressed 3 | -3 |
| TDO2 | 6999 | tryptophan 2,3-dioxygenase | -4.59 |
| TFR2 | 7036 | transferrinreceptor 2 | -3.55 |
| THOC4 | 10189 | THO complex 4 | 1.73 |
| THRSP | 7069 | thyroidhormoneresponsive | -4.22 |
| TOP2A | 7153 | topoisomerase (DNA) II alpha 170kDa | 2.31 |
| TREM2 | 54209 | triggering receptor expressed on myeloid cells 2 | -2.83 |
| TTR | 7276 | transthyretin | -3.38 |
| TUBB | 203068 | tubulin, beta | 2.08 |
| TUBB4Q | 56604 | tubulin, beta polypeptide 4, member Q | 1.84 |
| TUBG1 | 7283 | tubulin, gamma 1 | 2.04 |
| TYROBP | 7305 | TYRO proteintyrosinekinasebindingprotein | -1.79 |
| UGT2B4 | 7363 | UDP glucuronosyltransferase 2 family, polypeptide B4 | -4.95 |
| UGT2B7 | 7364 | UDP glucuronosyltransferase 2 family, polypeptide B7 | -4.33 |
| UNKL | 64718 | unkempt homolog (Drosophila)-like | -3.44 |
| UPRT | 139596 | uracil phosphoribosyltransferase (FUR1) homolog (S. cerevisiae) | -3.39 |
| VAMP5 | 10791 | vesicle-associated membrane protein 5 (myobrevin) | -2.82 |
| ZNF821 | 55565 | zincfingerprotein 821 | -4.21 |

**Table 7.1:** List of 195 genes commonly regulated across all 18 HCC cell lines. The gene symbol, Entrez Gene ID and Name, as well as its mean fold change are given.

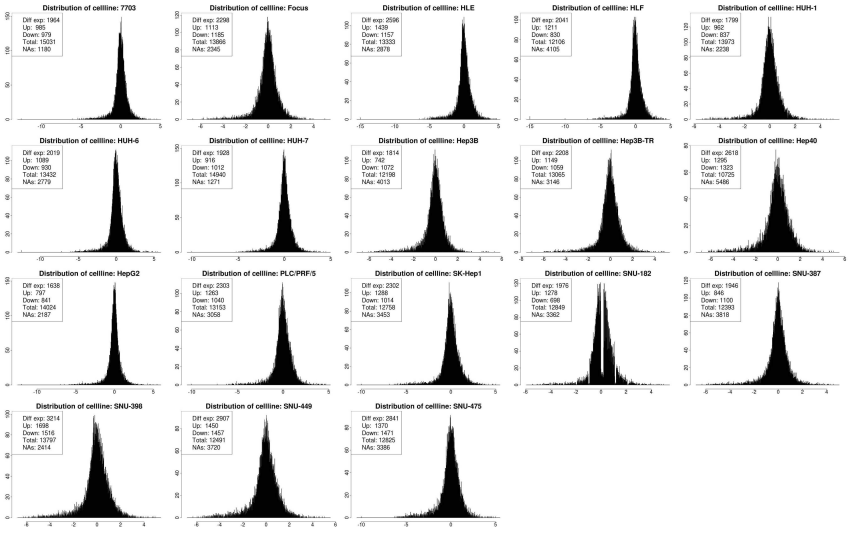


**Figure 7.7:** Box plot summary of all 195 genes commonly, differentially expressed in all 18 HCC cell lines investigated. The gene names are provided on the y-axis (a list of these genes is provided in the Supplement). The bar in the middle of the box of each gene represents average fold changes (log2 transformed). Expression values <-1 summarize genes with a more than two-fold reduction in expression. Values >1 represent genes with an expression change of more than two-fold. The x-axis provides fold changes (log2-transformed).

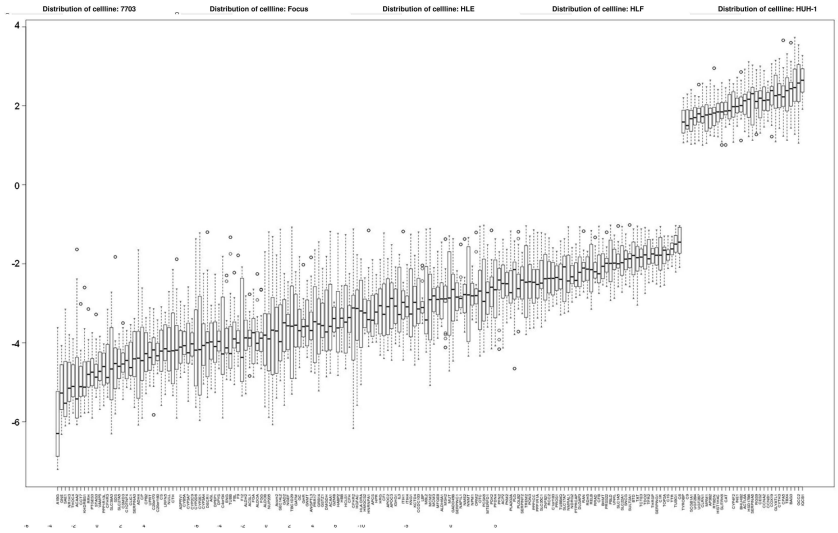


**Figure 7.8:** Distribution plots of the gene expression data for each cell line. The frequency of M-values is shown on the x-axis, while the M-values are given on the y-axis. The legend provides the number of differentially regulated genes (M-values >1 and <-1), their differentiation into up- and downregulated genes, the total number of genes used for generation of these plots (Total) and the number of genes for which no data were available (NAs).

# 8 Publication 4 - CellLineNavigator: a workbench for cancer cell line analysis

Published in Nucleic Acids Research in 2013

*Nucleic Acids Res. (2013). doi:10.1093/nar/gks1012*

Markus Krupp
Timo Itzel
Thorsten Maass
Andreas Hildebrandt
Peter R. Galle
Andreas Teufel

| Authors contribution: | |
|---|---|
| Draft writing | 80% |
| Review process | 80% |
| Data processing | 90% |
| Data analysis | 90% |
| Database design | 75% |
| GUI implementation | 75% |

## Abstract

The CellLineNavigator database, freely available at http://www.medicalgenomics.org/celllinenavigator, is a web-based workbench for large scale comparisons of a large collection of diverse cell lines. It aims to support experimental design in the fields of genomics, systems biology and translational biomedical research. Currently, this compendium holds genome wide expression profiles of 317 different cancer cell lines, categorized into 57 different pathological states and 28 individual tissues. To enlarge the scope of CellLineNavigator, the database was furthermore closely linked to commonly used bioinformatics databases and knowledge repositories. To ensure easy data access and search ability, a simple data and an intuitive querying interface were implemented. It allows the user to explore and filter gene expression, focusing on pathological or physiological conditions. For a more complex search, the advanced query interface may be used to query for (i) differentially expressed genes; (ii) pathological or physiological conditions; or (iii) gene names or functional attributes, such as Kyoto Encyclopaedia of Genes and Genomes pathway maps. These queries may also be combined. Finally, CellLineNavigator allows additional advanced analysis of differentially regulated genes by a direct link to the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources.

## Introduction

In vitro cancer cell culture experiments provide the opportunity of analysing and modelling the complex mechanisms of tumour biology through facile experimental manipulations, global as well as detailed mechanistic studies. They are, therefore, of significant aid in molecular biomedical research. A crucial role of cancer cell lines for medical, scientific and pharmaceutical institutions was elucidated by systematic analysis on lung cancer cell lines [1, 2] . They revealed not only the amazingly complex role of the cancer genome but also identified and characterized driver mutations in those cell lines. Further studies on cancer cell lines lead to the characterization of tumor protein 53 (TP53) and the understanding of multiple genetic mutations, mutant allele-specific imbalances and copy number losses in cancer [3, 4, 5] . Moreover, the ability to translate these findings to clinical applications had led to rational therapeutic drug selection [6] . For example, activating mutations in the epidermal growth factor receptor (EGFR) kinase domain have major clinical implications in lung cancer, and it was shown in cell line experiments that tumours with this mutation are sensitive to tyrosine kinase inhibitors [7] . However, repeatedly a varying response to treatment or targeted manipulation of gene expression was observed in diverse cancer cell lines. This was attributed to a diverse genetic background and, subsequently, a diverse gene expression. Thus, information on these diverse gene expression profiles in cancer cell lines may be crucial to experimental designs of modelling cancer in vitro and testing for novel therapeutic approaches.
We have, therefore, generated CellLineNavigator, a workbench for the biomedical community, which allows querying the transcriptom of a great variety of cancer cell lines to screen for the most suitable cell line for upcoming experiments. To enlarge the scope of this database, the data were linked to common functional and genetic databases, enabling querying for a more systematic view on cell line expression profiles.
In summary, we have generated a comprehensive database containing expression profiles of 317 cancer cell lines representing 57 different pathological states and 28 individual tissues. This database will aid the design of in vitro experiments in cancer research, as it will allow taking the genetic background of these cell lines into consideration. The CellLineNavigator database is publicly available at http://www.medicalgenomics.org/celllinenavigator/.

## Material and Methods

### Data source, data processing

Genome-wide expression data of multiple cell lines, freely available at ArrayExpress (database ID: E-MTAB-37 [8] ), were publicly provided by Greshock et al. (Laboratory of Cancer Metabolism Drug Discovery, GlaxoSmithKline, Collegeville, PA, USA). The cell lines were handled as previously described [9] . Briefly, the tran-

script abundance of 317 cancer cell lines was analysed using the Affymetrix Human Genome-U133 Plus2 GeneChip technology. This chip covers the complete human genome for analysis of >45 000 transcripts and >19 000 genes. All data were available in technical triplicates. Corresponding information on tissue site and disease state was supported for each cell line (Fig. 8.1). The differential expression was analysed using the R-Project [10] /bioconductor [11] suite with the following additional libraries: 'affy' [12] , 'hgu133plus2.db' [13] and 'frma' [14, 15] . After quality control, two microarray experiments (cell line SNU398—Replicate 1 and cell line SNU423—Replicate 2) were neglected for further analysis because of insufficient RNA level detection. All data were normalized using the 'expresso' function of the 'affy' package and following settings: background adjustment method: 'mas', normalization method: 'quantiles', PerfectMatch (PM) adjustment method: 'mas' and the method used for the computation of expression values: 'medianpolish'. Next, we calculated the expression median for each probe set for all cell lines. These values were subsequently used as control to calculate log2 transformed expression ratios (M-values), after the median expression was calculated for each cancer cell line. M-values representing the expression levels of tissue sites and disease states were calculated accordingly. Gene expression barcodes were generated using the 'frma' (frozen robust multiarray analysis) (default options) and 'barcode' (output: Z-score) function implemented in the 'frma' package. A frma Z-score of >5 suggested that a gene is expressed in a particular tissue. The frma Z-score was generated to allow comparison of the expression profiles with data already present at medicalgenomics.org [16, 17] and other microarray data sets processed with the frma method. Official gene symbols and National Center for Biotechnology Information (NCBI) Entrez GeneIDs were assigned to the data using the 'hgu133plus2.db' package.
To enable an integrative comparison and querying between gene expression and biological function information, all data were linked to commonly used and established bioinformatics databases and knowledge repositories, such as NCBI Entrez database [18] , HUGO Gene Nomenclature Committe (HGNC) [19] , Human Protein Reference Database (HPRD) [20] , Online Mendelian Inheritance in Man (OMIM) [19] , BioGPS [21] , Nextbio [22] and Gent [23] . Moreover, the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [24] was connected to identify gene signalling and molecular pathway associations. Data on cellular component, biological process and molecular function were collected from the Gene Ontology database [25] . Finally, the CellLineNavigator was cross-linked to our RNA-Seq expression profiling database on normal tissues, RNA-Seq Atlas [16] , and our liver-specific Library of Molecular Associations (LoMA [17] ).

## Data organization and Webinterface

The backbone of CellLineNavigator is a Linux-Postgre SQL-Apache-PHP stack implemented in a content management system (Drupal: http://drupal.org/). The

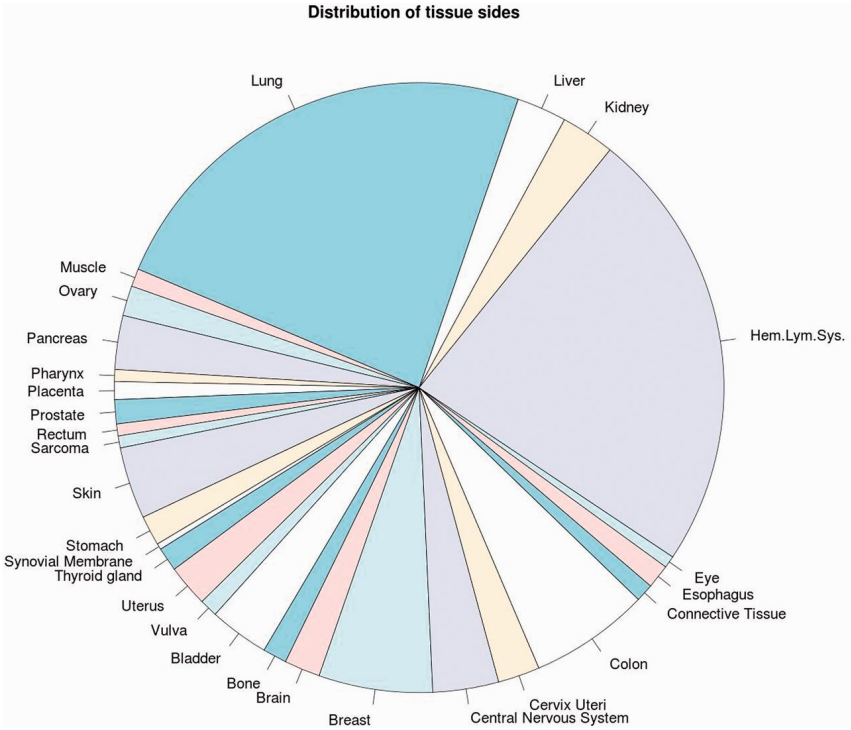


**Figure 8.1:** Distribution of tissues within CellLineNavigator.

database organization is founded on a menu, allowing to directly accessing the following sections: news, data, search, download and help section.
Information on current statistics and recent changes were posted in the news section to keep the users up-to-date, whereas the download section provides the possibility to download the complete CellLineNavigator database in tab separated text file format.
CellLineNavigator may easily be accessed through a simple (data section) or advanced querying interface (search section).
The data section offers the possibility to explore tissues (default) or disease states (Fig. 8.2). Filtering options for expression levels within individual tissues or disease states are supported. The default filter is set to list all genes with a different expression level of at least 2-fold in comparison with the respective control. Six additional levels of expression filtering are supported (from 1.5- to 5-fold). However, the user may also set the filter criteria to none (no filter) or no regulation (list all genes whose M-values are in the range of -1 to +1).
To allow users a high degree of flexibility to access CellLineNavigator, we implemented an advanced search section, offering the user 'Fulltext search' or 'Explore profile' options (Fig. 8.3). The 'Fulltext search' may be used to query for individual genes provided by the user to query for expression levels within specific cell lines, tissues or disease states (or any combination of all). Using the 'Explore profile' query option, the user may query for specific expression levels classified in the fields of (i) genes (ii) KEGG pathway maps (iii) gene ontologies (iv) cell lines (v)

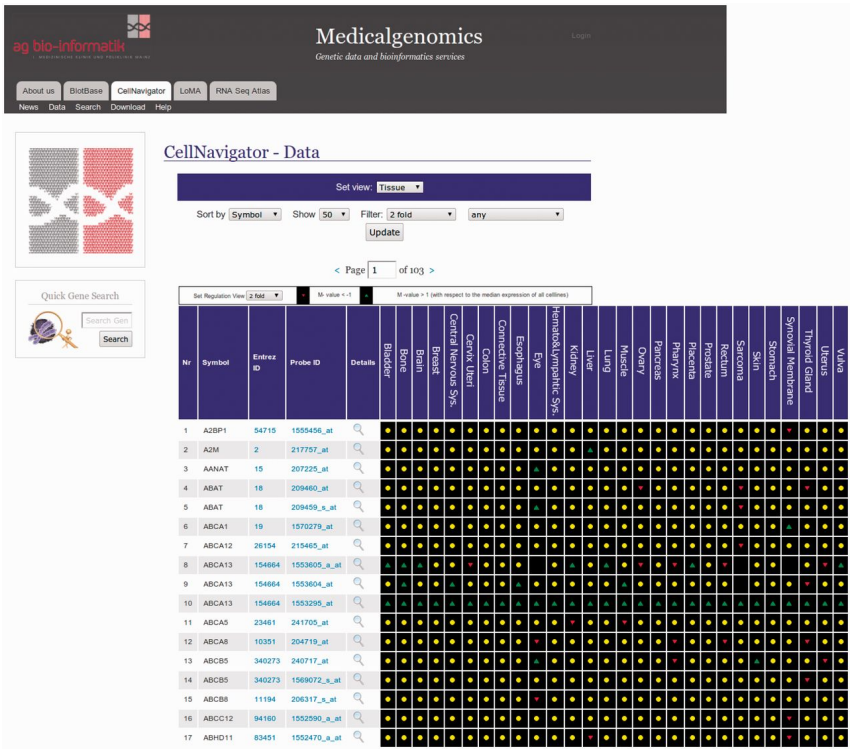


**Figure 8.2:** The data section offers the possibility to explore tissues or disease states. In default, the user will get a list of all genes that have shown a different expression level of at least 2fold in comparison with the respective control within the specific tissues. The set view option allows the user to switch between tissues and disease state views. Additional six filter options for expression levels are supported (default 2fold). Further, the user can set the expression filter criteria to none (no filter) or no regulation (list all genes whose M-values are in the range of $-1$ and 1). To allow users a more customizable way in displaying the data, the user may change the cut-off criteria for differentially expressed genes (default: 2fold).

tissues or (vi) disease states. Again, a combination of all query types is possible. Moreover, the user may also define cut-off criteria to filter for specific expression levels. The resulting gene list is shown in an interface providing the same features mentioned in the data section with one exception, the filter criteria is adjusted to the preceding query. These features may again be used for further filtering the resulting gene list.
To allow users a more customizable way of displaying the expression results, an extra option for setting the regulation view is supported (default: 2-fold).
Moreover, a powerful resource within our database extending the full impact of the individual gene–cell line relations is provided in the details section (Fig. 8.4). This section may be accessed by either clicking on the detail link or the specific expression icon in the results tables. It provides additional information on gene symbol, description, aliases, chromosomal location, Entrez ID, Ensembl ID, Gene Ontology, KEGG pathway and expression profiles. Although the expression profiles were individualized to the previous user query, for example, did the user click on the expression icon of tissue side 'bladder', the details view will show a barchar with an overview of the expression within all tissues and, more importantly, with a barchar representing the specific expression values of the cell lines corresponding to the tissue of interest. For further comparison with already available data at medicalgenomics.org, such as RNA-Seq Atlas, the details view may be switched from M-value to frma Z-score representation.
Finally, a major strength of the database is its direct connection to the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources [26] . Gene lists generated in CellLineNavigator may automatically be transferred to the DAVID analysis tools.

## Discussion

The current scope of biomedical studies on tumorigenesis requires a tremendous amount of human tumour material. Stringent restrictions on the international exchange of biological reagents and increasing requirements from institutes, ethics committees and government are limiting the availability of those human tumour materials. Thus, in vitro cell culture is highly useful for modelling the complex mechanisms of cancer development to identify molecular mechanisms related to tumour development and potential therapeutic targets.
Cell lines are capable of infinite replication and, therefore, offer an unlimited source of biomedical material that can be distributed to laboratories worldwide and thus, allow direct comparison of research results if originating from identical material. As a matter of fact, these cell lines are widely used in biomedical research. However, the detailed knowledge about their genetic profiles is still limited and has not been summarized in a large comparative database.
Diverse biological behaviour of cancer cell lines may result from diverse underlying genetic profiles and expression signatures, which may differ significantly

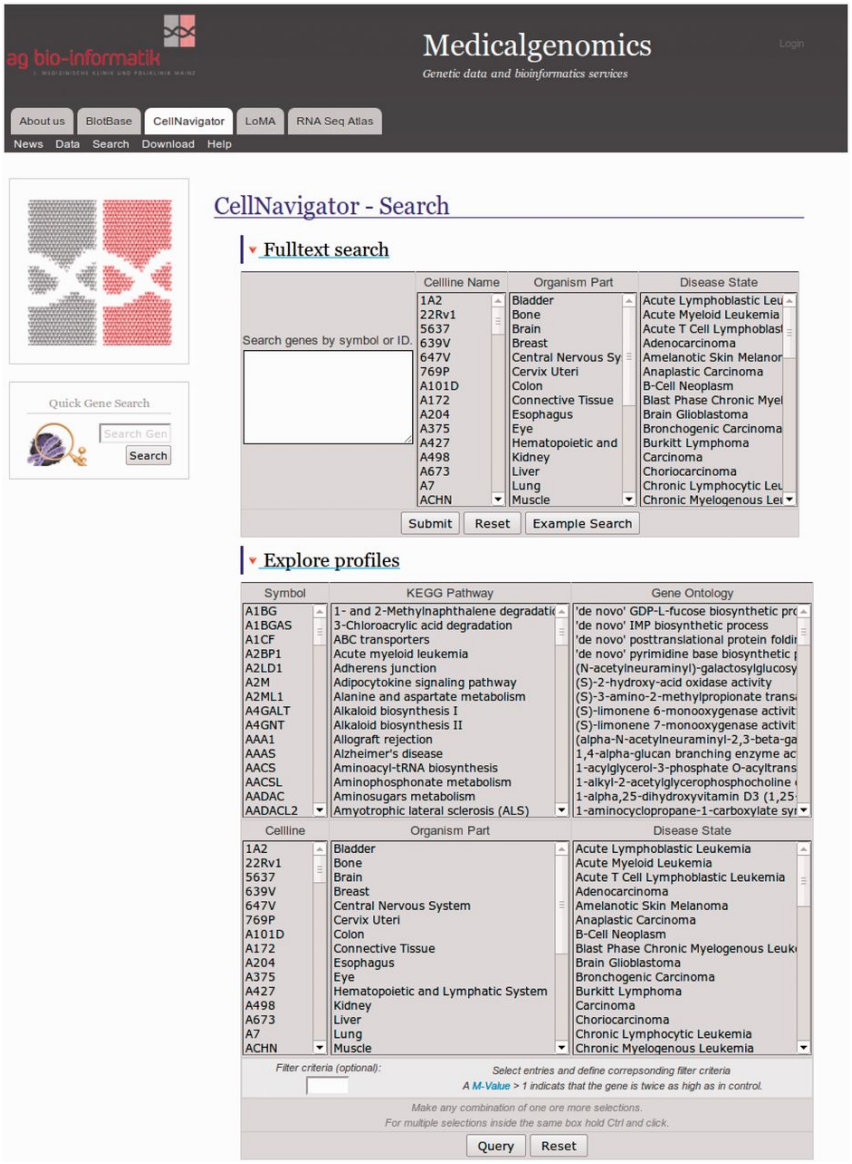


**Figure 8.3:** The search section allows users to choose between the 'Fulltext search' or 'Explore profile' option. In the 'Fulltext search', the user can provide a gene list that can be queried for expression levels within specific cell lines, tissue sides or disease states (or any combination of all). The 'Explore Profile' allows the user to query for specific expression levels classified in the fields (i) gene, (ii) KEGG pathway maps, (iii) Gene Ontology, (iv) cell line, (v) tissue side or (vi) disease state. A combination of all query types is possible. Additionally, the user may defin a cut-off criteria to filter for specific expression levels.

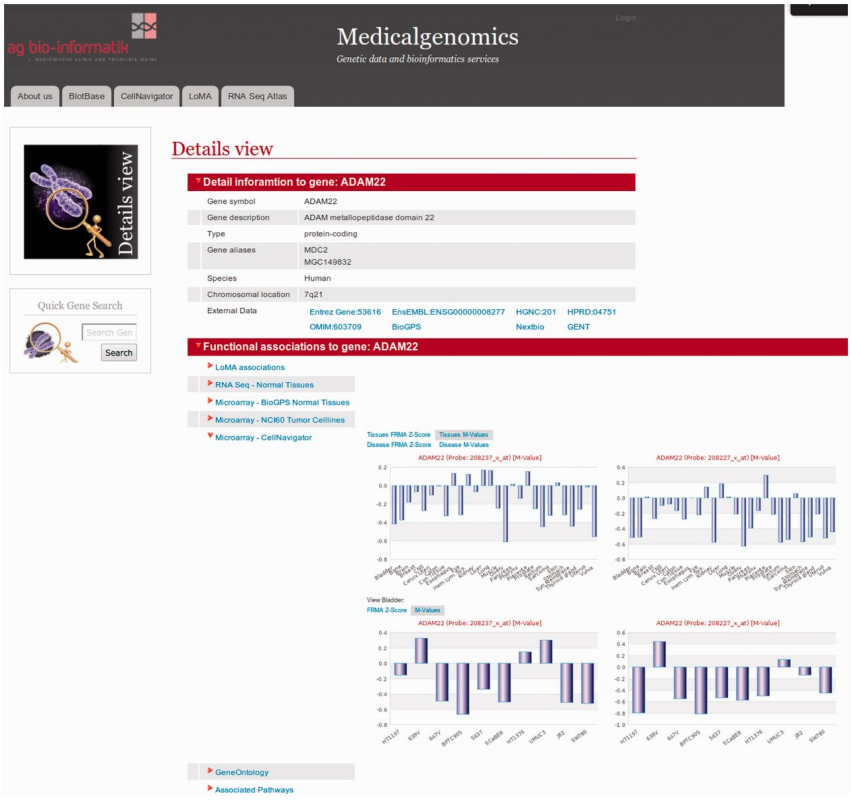


**Figure 8.4:** The details section offers a powerful option to access the full impact of the individual gene–cell line relation. The expression profiles are individualized to the previous user query, for example, if the user is interested in the expression of the tissue 'bladder', the details view display a barchar showing an overview of the expression within all tissues and, more importantly, with a barchar representing the specific expression values of the cell line(s) corresponding to the to the tissue 'bladder'. Further, additional information on gene symbol, description, aliases, chromosomal location, Entrez ID, Ensembl ID, Gene Ontology and KEGG pathway are supported.

among immortalized tumour cell lines. The awareness of these differences makes it useful/necessary to take the diverse gene expression signatures into account, especially while planning targeted strategies to influence the biological behaviour. Large scale microarray experiments to unravel the genetic profile of these cell lines are available through public databases, such as ArrayExpress [8] of Gene Expression Omnibus [27].

However, the analysis of these data is hardly feasible for biologists or physicians without substantial bioinformatics skills or at least knowledge on microarray analysis. Even with profound experience in microarray technology, analysis of such data is complex and time consuming task. So far to our knowledge, the Gene Expression Atlas [8] is the only database that provides access to these cell line expression profiles. However, the main focus of this database is not on cancer cell lines, and thus, it just contains the expression profile of ~90 cell lines from various species. Moreover, not only the classification into specific phenotypes but also data collected from multiple laboratories are incomplete and, therefore, exhibit multiple experimental conditions, making a comparison between the multiple expression profiles extremely difficult.

The database, CellLineNavigator, presented here contains gene expression profiles of >300 human cancer cell lines. These expression profiles were generated in the same laboratory under nearly the same experimental conditions and thus, guarantee a highest degree on comparability. Further, depending on phenotypic information, these cell lines were classified into corresponding tissues of origin and disease states. The main focus of CellLineNavigator is not simply on summarizing these data but rather on an easy and user friendly availability as well as the linkage to advanced bioinformatics analyses tools. To guarantee easy data access and connectivity, we implemented a mostly self-explaining Web application as a user friendly front end to the data base. This Web application allows users to query for (i) differentially expressed genes; (ii) pathological (e.g. melanoma) or physiological (e.g. lung) conditions or (iii) gene names or functional attributes, such as KEGG pathway maps. A combination of all query types is possible.

Comparative analysis of differential gene expression between cell lines or diseases of interest will initially often result in (large) lists of genes being differentially regulated. To further characterize the differences between the respective samples and thus a major advance in the usability of this database, these collections of genes need to be further characterized with respect to functional or structural similarities. We, therefore, chose to link and provide an automated data transfer of gene lists of interest to DAVID, a large bioinformatics suite providing functional and structural analyses, such as pathway enrichment, gene ontology enrichment or analysis of functional domains. This automated linkage to DAVID brings our database resource and analysis tool to the next level of not only comparing genetic changes but also functionally and structurally characterizing the differences by means of advanced bioinformatics. In summary, CellLineNavigator is the first database providing comprehensive summary, display and analysis options for gene expression data of the most commonly used cancer cell lines. It provides access to large mi-

croarray data sets without advanced bioinformatics skills. Thus, CellLineNavigator may be of significant aid for in vitro modelling of cancer mechanisms and testing of novel therapeutic approaches.

## Funding

Boehringer Ingeheim Funds and Roche (to AT). Funding for open access charge: Department of Medicine I of the Johannes Gutenberg University Mainz.

# BIBLIOGRAPHY


[1] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, **December 2009**.

[2] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, **April 2009**.

[3] J. Gandhi, J. Zhang, Y. Xie, J. Soh, H. Shigematsu, W. Zhang, H. Yamamoto, M. Peyton, L. Girard, W. W. Lockwood, W. L. Lam, M. Varella-Garcia, J. D. Minna, and A. F. Gazdar. Alterations in genes of the EGFR signaling pathway and their relationship to EGFR tyrosine kinase inhibitor sensitivity in lung cancer cell lines. *PLoS ONE*, 4(2):e4576, **February 2009**.

[4] A. Singh, P. Greninger, D. Rhodes, L. Koopman, S. Violette, N. Bardeesy, and J. Settleman. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. *Cancer cell*, 15(6):489–500, **June 2009**. PMID: 19477428.

[5] J. Soh, N. Okumura, W. W. Lockwood, H. Yamamoto, H. Shigematsu, W. Zhang, R. Chari, D. S. Shames, X. Tang, C. MacAulay, M. Varella-Garcia, T. Vooder, I. I. Wistuba, S. Lam, R. Brekken, S. Toyooka, J. D. Minna, W. L. Lam, and A. F. Gazdar. Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells. *PLoS ONE*, 4(10):e7464, **October 2009**.

[6] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, **January 2000**. PMID: 10647931.

[7] J. G. Paez, P. A. Jänne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon, K. Naoki, H. Sasaki, Y. Fujii, M. J. Eck, W. R. Sellers, B. E. Johnson, and M. Meyerson. EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500, **June 2004**.

[8] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–872, **January 2009**. PMID: 19015125.

[9] J. Greshock, K. E. Bachman, Y. Y. Degenhardt, J. Jing, Y. H. Wen, S. Eastman, E. McNeil, C. Moy, R. Wegrzyn, K. Auger, M. A. Hardwicke, and R. Wooster. Molecular target class is predictive of in vitro response profile. *Cancer research*, 70(9):3677–3686, **May 2010**. PMID: 20406975.

[10] R development core team (2011). r: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. isbn 3-900051-07-0, url http://www.r-project.org.

[11] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, **2004**. PMID: 15461798.

[12] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)*, 20(3):307–315, **February 2004**. PMID: 14960456.

[13] M. Carlson, S. Falcon, H. Pages, and N. Li. hgu133plus2.db: Affymetrix human genome u133 plus 2.0 array annotation data (chip hgu133plus2).

[14] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2):242–253, **April 2010**. PMID: 20097884.

[15] M. N. McCall, K. Uppal, H. A. Jaffee, M. J. Zilliox, and R. A. Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(Database issue):D1011–D1015, **January 2011**. PMID: 21177656 PMCID: PMC3013751.

[16] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel. RNA-Seq atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics (Oxford, England)*, 28(8):1184–1185, **April 2012**. PMID: 22345621.

[17] S. Buchkremer, J. Hendel, M. Krupp, A. Weinmann, K. Schlamp, T. Maass, F. Staib, P. R. Galle, and A. Teufel. Library of molecular associations: curating the complex molecular basis of liver diseases. *BMC genomics*, 11:189, **2010**. PMID: 20302666.

[18] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, **January 2011**. PMID: 21097890.

[19] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, and E. A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic acids research*, 39(Database issue):D514–519, **January 2011**. PMID: 20929869.

[20] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic acids research*, 37(Database issue):D767–772, **January 2009**. PMID: 18988627.

[21] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. Hodge, J. Haase, J. Janes, J. Huss, and A. Su. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10(11):R130, **November 2009**.

[22] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS ONE*, 5(9):e13066, **September 2010**.

[23] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim. GENT: gene expression database of normal and tumor tissues. *Cancer informatics*, 10:149–157, **2011**. PMID: 21695066.

[24] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–114, **January 2012**. PMID: 22080510.

[25] The gene ontology: enhancements for 2011. *Nucleic acids research*, 40(Database issue):D559–564, **January 2012**. PMID: 22102568.

[26] X. Jiao, B. T. Sherman, D. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806, **July 2012**. PMID: 22543366 PMCID: PMC3381967.

[27] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research*, 39(Database issue):D1005–1010, **January 2011**. PMID: 21097893.

# 9 Publication 5 - RNA-Seq Atlas: A reference database for gene expression profiling in normal tissue by next generation sequencing

Published in Bioinformatics 2012

*Bioinformatics (Oxford, England) (2012).doi:10.1093/bioinformatics/bts084*

Markus Krupp
Jens U. Marquardt
Ugor Sahin
Peter R. Galle
John Castle
Andreas Teufel

| Authors contribution: | |
|---|---|
| Draft writing | 80% |
| Review process | 80% |
| Data processing | 80% |
| Data analysis | 90% |
| Database design | 75% |
| GUI implementation | 75% |

## Abstract

**Motivation:** Next-generation sequencing technology enables an entirely new perspective for clinical research and will speed up personalized medicine. In contrast to microarray-based approaches, RNA-Seq analysis provides a much more comprehensive and unbiased view of gene expression. Although the perspective is clear and the long-term success of this new technology obvious, bioinformatics resources making these data easily available especially to the biomedical research community are still evolving.

**Results:** We have generated RNA-Seq Atlas, a web-based repository of RNA-Seq gene expression profiles and query tools. The website offers open and easy access to RNA-Seq gene expression profiles and tools to both compare tissues and find genes with specific expression patterns. To enlarge the scope of the RNA-Seq Atlas, the data were linked to common functional and genetic databases, in particular offering information on the respective gene, signaling pathway analysis and evaluation of biological functions by means of gene ontologies. Additionally, data were linked to several microarray gene profiles, including BioGPS normal tissue profiles and NCI60 cancer cell line expression data. Our data search interface allows an integrative detailed comparison between our RNA-Seq data and the microarray information. This is the first database providing data mining tools and open access to large scale RNA-Seq expression profiles. Its applications will be versatile, as it will be beneficial in identifying tissue specific genes and expression profiles, comparison of gene expression profiles among diverse tissues, but also systems biology approaches linking tissue function to gene expression changes.

**Availability and implementation:** http://medicalgenomics.org/rna_seq_atlas

Supplementary information: Supplementary data are available at Bioinformatics online.

## Introduction

Over the next years, the availability of next-generation sequencing (NGS) data will offer an entirely new perspective for clinical research and help usher in personalized medicine. So far, several databases offer storage or download of NGS data [1, 2] . However, to access the valuable information of this promising new technique, the user has to manually download the data and be familiar with their analysis to extract the valuable information. This is currently not the case for most biomedical researchers. We therefore created the RNA-Seq Atlas, a database and user interface (UI) providing easy access to NGS data. Currently, RNA-Seq Atlas holds gene expression profiles on eleven human, healthy tissues and can be accessed over an intuitive web interface. To further increase the utility of the RNA-Seq Atlas, the data were linked to multiple microarray gene profiles representing normal and pathological states. Furthermore, various query tools were designed to offer a great variability of individual analysis.

## Database organization and access

### Data sources

The provided genome-wide expression compendium originates from eleven healthy, human tissue samples pooled from multiple donors spanning 32 384 specific transcripts corresponding to 21 399 unique genes [3] (ENA ERP000257; ArrayExpress E-MTAB-305). The tissues include adipose, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes. Sequencing was performed on an Illumina GA-II sequencer, generating an average of 50 million reads per tissue, with sequence reads of 36 or 50 nt depending on tissue. The expression level were estimated by mapping and counting reads to single gene sequences derived from the UCSC genome browser, followed by normalization to Reads Per Kilobase of exon model per Million mapped reads values [4] .
Moreover, to enable an integrative comparison between RNA-Seq and microarray expression profiles we integrated a panel of 84 microarrays from BioGPS [5, 6] , Normal Tissue Gene Expression Study [7] as well as [8, 9] NCI60 into the RNA-Seq Atlas. These gene expression profiles correspond to the equivalent tissues included in the RNA-Seq Atlas and involve >39 000 transcripts from pathological (i.e. cancer) and normal tissues states. Detailed information about data processing and integration can be found in the Supplementary Materials S1 and S2.
Further, the RNA-Seq Atlas was linked to commonly used and established bioinformatics databases and knowledge repositories. Enabling access to deeper transcriptional information was achieved by linking the RNA-Seq Atlas data to the NCBI Nucleotide database [10] . Also, information on corresponding gene symbol, aliases, description, chromosomal location, Entrez ID as well as EnsEMBL ID were assembled from the NCBI Entrez and EnsEMBL databases [10, 11] . Additional out-

going links to HGNC [12] , HPRD [13] , OMIM [14] , BioGPS [6] , Nextbio [15] and GENT [16] were supported. Finally, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] was accessed to identify gene signaling as well as molecular pathway affiliations; and data on cellular component, biological process and molecular function were collected from the Gene Ontology database [18] . Finally, the RNA-Seq Atlas was cross-linked to our liver-specific databases LoMA [19].

## Data organization and web interface

RNA-Seq Atlas is implemented within a Drupal content management system environment over a Linux-PostgreSQL-Apache-PHP stack. The database organization is founded upon a menu which allows access to the news, data, search and download sections.
The news section keeps users up to date about recent changes and current statistics, whereas the download section give the possibility to download the RNA-Seq Atlas in tab separated text file format. RNA-Seq atlas can be accessed through simple (data section) or advanced (search section) query forms. The advanced query offers four detailed options:

1. Full text search.
2. Comparison of specific tissues profiles; also allowing for comparative analysis not only between normal tissue information but also to NCI60 data and thus between normal and tumor tissues.
3. Explore common (and diverse) gene expression profiles between tissues.
4. Explore pathway profile; e.g. selecting one or multiple KEGG pathway resulting in a list of involved genes.

Finally, a 'details' link provides additional information including: gene symbol, description, aliases, chromosomal location, Entrez ID, Ensembl ID, Gene Ontology, KEGG pathway as well as the expression profile within the normal human tissues and cancer cell lines.

## Future directions

Future directions include an incorporation of more data from healthy and cancer tissue to provide a richer source of comparative transcriptomics and implementation of a Gene Set Enrichment Analysis (GSEA) analysis engine within the RNA-Seq Atlas.

## Discussion

In this work, we present RNA-Seq Atlas, an easily accessible database and UI, offering access to NGS gene expression profiles. Furthermore, to enhance the bioin-

formatics integration, the data is linked to a wide variety of commonly used and established databases and knowledge repositories. To further enlarge the very broad scope of RNA-Seq Atlas and to facilitate the analysis of gene expression profiles of several pathological conditions, the data were linked to cancer cell line expression profiles. Finally, the implementation of a wide variety of querying tools allows the user to start individual analysis, enabling for both bioinformaticians and experimental researchers.

*Funding:* This study was supported by a research grant of the Boehringer Ingelheim Foundation and funding of the core facility bioinformatics of the University Hosiptal of the Johannes Gutenberg University Mainz, Germany.

# BIBLIOGRAPHY


[1] D. Altshuler. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, **October 2010**. PMID: 20981092.

[2] M. Shumway, G. Cochrane, and H. Sugawara. Archiving next generation sequencing data. 38(Database issue):D870–D871, **January 2010**. PMID: 19965774 PMCID: 2808927.

[3] J. C. Castle, C. D. Armour, M. Löwer, D. Haynor, M. Biery, H. Bouzek, R. Chen, S. Jackson, J. M. Johnson, C. A. Rohl, and C. K. Raymond. Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. *PloS One*, 5(7):e11779, **2010**. PMID: 20668672.

[4] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, **July 2008**. PMID: 18516045.

[5] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, **April 2004**. PMID: 15075390.

[6] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, r. Huss, Jon W, and A. I. Su. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10(11):R130, **2009**. PMID: 19919682.

[7] X. Ge, S. Yamamoto, S. Tsutsumi, Y. Midorikawa, S. Ihara, S. M. Wang, and H. Aburatani. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, 86(2):127–141, **August 2005**. PMID: 15950434.

[8] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3):227–235, **March 2000**. PMID: 10700174.

[9] U. T. Shankavaram, S. Varma, D. Kane, M. Sunshine, K. K. Chary, W. C. Reinhold, Y. Pommier, and J. N. Weinstein. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*, 10:277, **2009**. PMID: 19549304.

[10] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(Database issue):D38–51, **January 2011**. PMID: 21097890.

[11] G. Spudich, X. M. Fernández-Suárez, and E. Birney. Genome browsing with ensembl: a practical overview. *Briefings in Functional Genomics & Proteomics*, 6(3):202–219, **September 2007**. PMID: 17967807.

[12] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, and E. A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, 39(Database issue):D514–519, **January 2011**. PMID: 20929869.

[13] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–772, **January 2009**. PMID: 18988627.

[14] Online mendelian inheritance in man, OMIM®. McKusick-Nathans institute of genetic medicine, johns hopkins university (baltimore, MD).

[15] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. 5(9). PMID: 20927376 PMCID: 2947508.

[16] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim. GENT: gene expression database of normal and tumor tissues. *Cancer Informatics*, 10:149–157, **2011**. PMID: 21695066.

[17] K. F. Aoki and M. Kanehisa. Using the KEGG database resource. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, Chapter 1:Unit 1.12, **October 2005**. PMID: 18428742.

[18] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald,

G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, **May 2000**. PMID: 10802651.

[19] S. Buchkremer, J. Hendel, M. Krupp, A. Weinmann, K. Schlamp, T. Maass, F. Staib, P. R. Galle, and A. Teufel. Library of molecular associations: curating the complex molecular basis of liver diseases. *BMC Genomics*, 11(1):189, **March 2010**. PMID: 20302666.

[20] C. D. Armour, J. C. Castle, R. Chen, T. Babak, P. Loerch, S. Jackson, J. K. Shah, J. Dey, C. A. Rohl, J. M. Johnson, and C. K. Raymond. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods*, 6(9):647–649, **September 2009**. PMID: 19668204.

[21] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, **July 2009**. PMID: 19451168.

[22] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2):242–253, **April 2010**. PMID: 20097884.

[23] M. N. McCall, K. Uppal, H. A. Jaffee, M. J. Zilliox, and R. A. Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(Database issue):D1011–1015, **January 2011**. PMID: 21177656.

## Supplementary information

## Material / Data Sources

**RNA-Seq Data:**
The provided genome-wide expression compendium originates from eleven, healthy, human tissues samples pooled from multiple donors spanning 32384 specific transcripts corresponding to 21399 unique genes. The tissues include adipose, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes. Total RNA from the reference tissues were purchased from Ambion (Austin, USA) and represent a pool of RNA from multiple donors. Libraries were prepared as described in Armour et al, 2009 [20] , including both poly[A]+ and poly[A]- fractions.
Sequencing was performed on the Illumina GA-II sequencer. An average of 50 million reads per tissue was generated, with sequence reads of 36 nt or 50 nt depending on the tissue. The unprocessed data was deposited at EMBL (ENA ERP000257; ArrayExpress E-MTAB-305). After trimming reads to a common length of 28 nt to avoid aligning sequences of amplified primers, the obtained reads were aligned to the human hg18 genome assembly using BWA [21] . For mRNAs RefSeq transcript coordinates and associated gene symbols were downloaded from the UCSC genome browser. Only the reads mapping to a single gene were used. Next, we determined the reads overlapping each transcript in the correct genomic orientation. The expression levels were estimated by mapping and counting reads to single gene sequences derived from the UCSC genome browser followed by normalization to RPKM values. For ncRNAs, the corresponding genomic coordinates were downloaded from the two tracks in the UCSC genome browser, assembly hg18, tracks RNA Genes and sno/miRNAs. For this analysis, pseudogenes, miRNAs, tRNAs, and rRNAs were removed. Genes labeled as "related" were combined, such as 7SK and 7SK-related, into a single cluster while preserving all genomic locations [3].

**Microarray Data:**
Multiple Microarray experiments, representing normal and disease states (i.e. cancer), were included into RNA-Seq Atlas to enable an integrative detailed comparison between RNA-Seq and microarray expression profiles. The BioGPS (http://biogps.org/) and 'Normal tissue gene expression study' (http://www.genome.rcast.u-tokyo.ac.jp/normal/) gene profiles were used as the basis for normal and the NCI60 (http://discover.nci.nih.gov/cellminer/home.do) gene profiles for pathological states.
To ensure homogenous and comparable gene expression profiles only microarrays from the Affymetrix Human Genome U133A Array (http://www.affymetrix.com) platform were integrated into RNA-Seq Atlas. The Human Genome U133 (HG-U133) Set contains almost 45,000 probe sets representing more than 39,000 transcripts derived from approximately 33,000 well-substantiated human genes. The

data were downloaded in CEL file format from the Gene Expression Omnibus web-interface of the NCBI (http://www.ncbi.nlm.nih.gov/geo/). Furthermore, normal tissue arrays were limited to tissues represented in the RNA-Seq Atlas. Thus, 16 gene expression profiles of the BioGPS project were included, 9 of the 'Normal tissue gene expression study' and 59 (LC:NCI_H23 microarray experiment failed) of the NCI60 (supplement table1) . The included tissues representing the normal states are as followed: colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes; Cancer tissues: breast, CNS, colon, kidney, leukemia, lung, melanoma, ovary and prostate.

## Methods

The FRMA Z-Score transformation of the microarray data was implemented in R-Project (http://www.r-project.org) using the bioconductor (http://www.bioconductor.org) libraries affy, hgu133a.db and fRMA [22, 23].

**Processing Affymetrix HG U133A:**
After loading gene expression profiles into an AffyBatch the phenotype data were assigned. Next, background correction, normalization and summarization were performed by applying the frma function [22] of the fRMA package to the AffyBatch with default options.

**Z-Score transformation:**
The Z-Score was calculated using the barcode function [23] of the fRMA package to standardize the gene values of the microarray data. Barcode options were set to platform='GPL96' and output='z-score'. The Z-Score for a specific gene in barcode is defined as:

$$Z = \frac{X - mu}{tau} \tag{1}$$

The Z represents the Z-Score, X the expression value of the specific transcript, mu (tau) the precomputed, standardized mean (standard deviation) of the unexpressed distribution of the specific transcript supported by the barcode function for the platform 'GPL96' [23] . Next, the Z-Scores were summarized via mean for each tissue and each pathological state (normal, cancer), whereas a Z-Score greater than 5 might suggest that the gene is expressed in that tissue. Finally, the transformed expression values were stored in the PostgreSQL database.

## Use cases

**Use Case 1: Biomarker revaluation:**
Provided that a researcher is interested in the evaluation of an identified biomarker (e.g. cancer). A keyword search at the RNA-Seq. Atlas enables a quick and integra-

tive overview of the tissue specific expression of the gene of interest in cancer and normal tissue.
This query will provide important information on different layers of translational biology:

1. By comparing the expression of the marker in the tissue of interest the researcher will get immediate information on the potential of the biomarker to assess pathological states.

2. By comparing the expression of the biomarker in different tissue the researcher can estimate tissue specificity.

*Query pipeline:*

1. *Open the 'Fulltext Search' of RNA-Seq Atlas and enter BioMarker of interest (e.g. EPCAM) (Fig. 9.1).*

2. *Execute query and use the preview advantage of the RNA-Seq Atlas result table by 'mouseover' over the expression charts to get a quick overview about the expression (Fig. 9.2).*

3. *Proceed to the details section by clicking on the magnifier glass and compare the expression of the biomarker in different tissue (Fig. 9.3).*

Thus, the RNA-Seq. Atlas reveals important information of both the diagnostic and therapeutic potential of the identified marker as a basis for subsequent and more detailed investigations.

**Use Case 2: Identification of liver specific genes:**
In case that the researcher is interested in a transcriptional profile highly specific for a healthy tissue which could deal as a diagnostic tool to distinguish between a healthy or diseased tissue, he can query the RNA-Seq Atlas.
This query will provide important information to tissue specific genes representing normal cell behavior.

1. By comparing the expression of the tissue of interested in comparison to the expression of the reference tissues the researcher will get immediate information on genes exclusively expressed within the tissue of interested.

*Query pipeline:*

1. *Open the 'Compare specific tissue profile' search of RNA-Seq Atlas and define appropriate cutoff values (e.g. in case for a highly specific liver transcriptional profile liver >= 10 and reference tissue <= 2) (Fig. 9.4).*

2. *Execute query*

Thus, the RNA-seq. Atlas provides integrative information of genes physiologically and specifically expressed within a given tissue. This enables a context specific view on genes that can be used to study disease states. Furthermore, investigations on pathway and functional levels can be revealed by navigating to the details section.

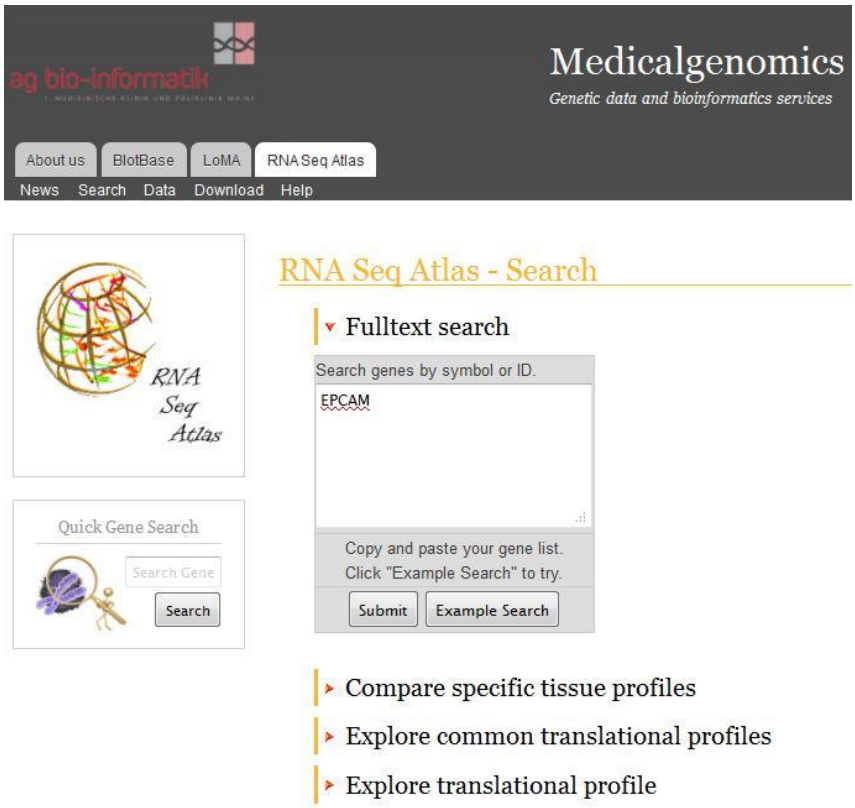


**Figure 9.1:** 'Fulltext Search' of RNA-Seq Atlas

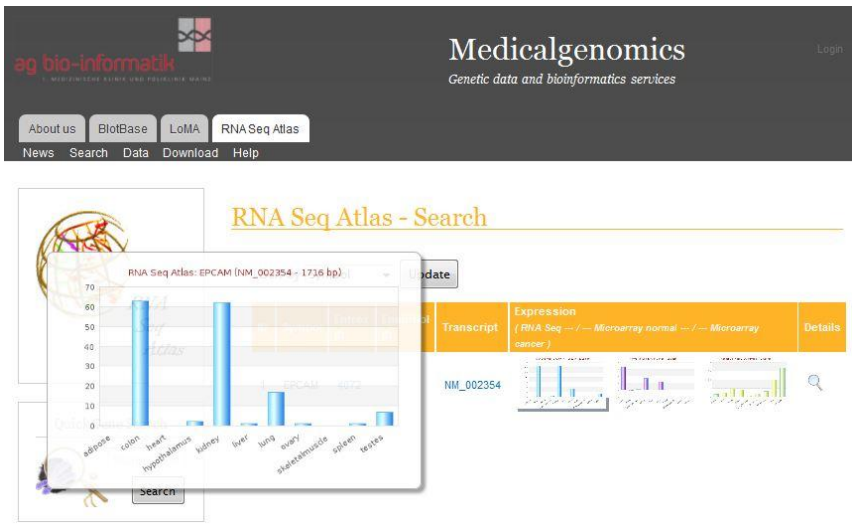


**Figure 9.2:** Result preview of RNA-Seq Atlas

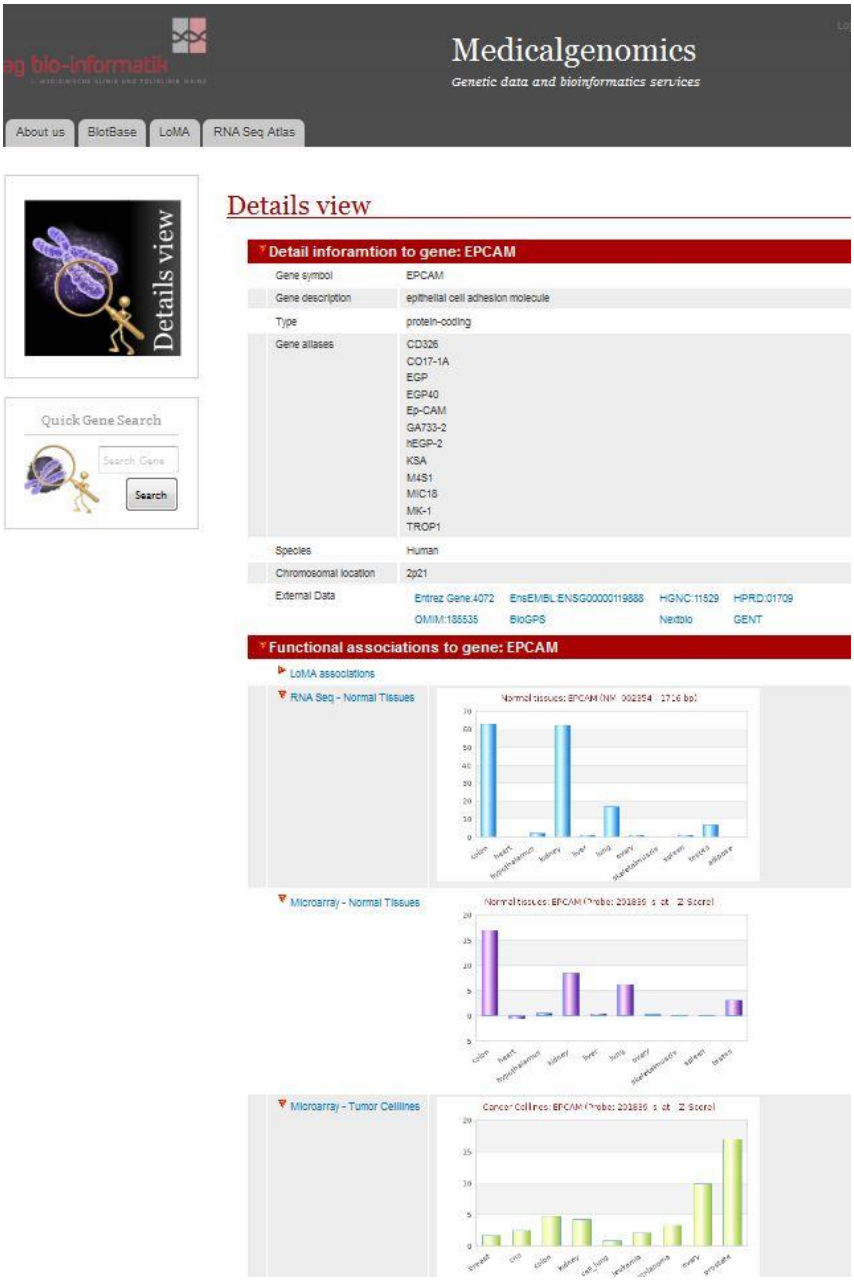


**Figure 9.3:** Details section of RNA-Seq Atlas

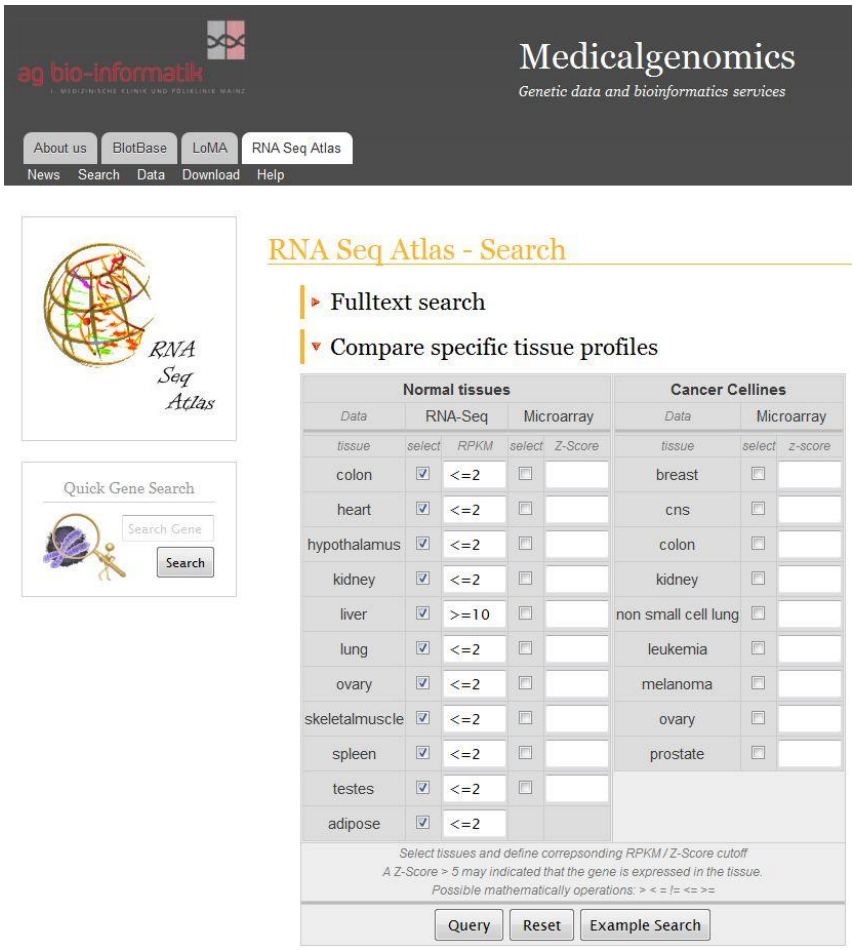


**Figure 9.4:** 'Compare specific tissue profile' search of RNA-Seq Atlas

# 10 Discussion and Outlook

## 10.1 Discussion

The work presented in the thesis has generated a comprehensive compendium to human tissues and has contributed massively to medicalgenomics.org, and therefore has solved several biomedical aspects of the identification of specific knowledge out of unmanageable amounts of digital data distributed over multiple places in the Internet. This compendium is the first of its kind and its salvaged knowledge will aid researchers in getting a systematic overview about specific genes or functional profiles in view of regulation as well as pathological or physiological conditions. Moreover, it enables researchers to design their experiments more individually, which will not only save valuable resources and time but also increase the success rate.

Up until January 2014, the compendium has been cited over 60 times. In most of the publications it was accessed to investigate on specific gene expression levels to prove hypotheses. For example, it was recently used to validate that the disruption of the gene Sec24d results in early embryonic lethality in the mouse [1]. Moreover, it was also used to support ongoing investigations. In one recent study, the compendium emphasises the importance of the CTCF-binding sites potential to function as insulators [2]. The compendium has also motivated other people to create a similar compendium to the domestic pig (http://www.macrophages.com/pig-atlas) [3], and used it to generate a interaction network [4].

The compendium is implemented within a Drupal (https://drupal.org/) content management system environment over a Linux-PostgreSQL-Apache-PHP stack. It subsists on four workbenches: LoMA [5], CellMinerHCC [6], CellLineNavigator [7] and RNA-Seq Atlas [8]. For each workbench, individual data mining techniques were developed in dependence of data foundation and evaluated to support highly valid gene disease associations and gene expression information. To ensure easy data access and searchability, a powerful but user-friendly application programming interface (API) was developed. In orientation on data foundation, the API further adapts and authorizes researchers with more specific queries. In addition, the compendium is connected to the most commonly applied biomedical databases and a comprehensive functional analysis tool.
Each of the introduced workbenches are the first of its kind and according to the workbench the compendium grants researchers the potential to examine biomedical assumptions as well as proving or generating novel biomedical hypotheses in the context of:

- Investigation on pathological conditions
- Investigation on physiological conditions
- Investigation on biological networks
- Investigation on single genes
- Investigation on gene expression levels
  - Filter expression levels
  - Compare expression levels
- Any combination of the above

One of the greatest functions of the compendium lies in the characterization of genes into pathological and physiological conditions. Thus, plenty of highly specialized gene signatures can be generated in dependence to the researchers need. Due to different data foundations and corresponding evaluation methods gene signatures associated with pathological conditions provided in the workbench, LoMA can be considered as most valid, because those associations were evaluated by text-mining algorithms followed by manual validation [5]. However, the text-mining of each specific disease relies on its own vocabulary and its creation is a complex process. Therefore, pathological gene signatures in LoMA are currently restricted to severe liver diseases. The enormous capability of such gene-disease signatures is demonstrated in Chapter 6. This chapter deals with the application of LoMA to generate a gene signature associated with cholangiocellular carcinoma. This signature was essential in the demonstration of the consistently enriched pathway mitogen activated protein kinase in CCC, which under deeper investigation has provide a rational for treatment of CCC with sorafenib [9]. In addition, the compendium supports a multitude of querying options to filter or compare gene expression profiles in order to create research-specific gene profiles. Those gene expression based investigations are reserved for the workbenches CellMinerHCC, CellLineNavigator and RNA-Seq Atlas, because their fundamental basis is generated from highly reliable raw expression data [6, 7, 8]. The potential of those query mechanisms are outlined in Chapter 4.6 and 7. In those chapters the compendium was queried to generate two conserved gene signatures reflecting the tissue expression state of a tumorous and healthy liver. Such signatures can be used to reveal essential functional differences between healthy and diseases tissue states and hence, help to find novel therapeutics options [10, 11]. In addition, those expression profiles are valuable in the discovery of potential biomarkers and may further deal as diagnostic tools to distinguish between healthy or diseased tissue states [12, 13, 14]. Many other studies on gene profiles were successfully carried out [15, 16]. In addition, to seven literature based disease profiles in LoMA, ten healthy and ten tumorous tissue signatures are already parsed in the compendium and integrated into the RNA-Seq Atlas and CellMinerHCC. Those profiles can be easily accessed by using the extended compendium GUI. Options to tune conserved expression profiles

exists, thus enabling a high amount of flexibility to fit researchers' needs. Moreover, the extended GUI supports several options for comparing two or more tissue types and/or states by the mean of expression values. This enables researchers to generate finely tuned signatures to prove or generate hypotheses in the field of comparative genomics. Beyond LoMA, RNA-Seq Atlas and CellMinerHCC, the workbench CellLineNavigator may also be applied to generate further expression profiles to over 80 different tissue states. However, the main potential of CellLineNavigator lies in the large scale comparison of a vast amount of diverse cell lines to support experimental design in the fields of genomics, systems biology and translational biomedical research. Those cell lines are capable of infinite replication and therefore, offer an unlimited source of biomedical material that can be distributed to laboratories worldwide and thus, allow direct comparison of research results if originating from identical material. As a matter of fact, these cell lines are widely used in biomedical research. But the detailed knowledge about their genetic profiles is still limited. CellLineNavigator closed this gap by offering access to over 300 different human cancer cell lines. Its versatile GUI provides a comprehensive summary, display and analysis options for gene expression data and thus, CellLineNavigator is of significant aid for in vitro modeling of cancer mechanisms and testing of novel therapeutic approaches.

Another strength of the compendium yields in the possibility to carry out investigations on functional genomic levels. In order to enable this option, metabolic reaction and signaling event information were collected from the Kyoto Encyclopedia of Genes and Genomes [17] and Gene Ontology [18] databases. These databases are categorized into several functional concepts and after parsing, those concepts were assimilated into the compendium. After the adjustment of the GUI accordingly, functional expression profiles to over 200 KEGG pathway maps and more than 6000 gene ontology categories can be easily generated. This powerful option in combination with pathological or physiological conditions enables researchers to create individual and research-targeted expression profiles (e.g. to cell communication in hepatocellular carcinoma). The resulting profiles empower scientists not only to get a comprehensive overview but also to identify key regulatory genes in an efficient way. The knowledge of such key regulatory genes might be of prime importance for a scientist in successfully future designing her/his research.

Finally, the utilization of medicalgenomics.org is also a good choice to further validate biomarkers. A biomarker is a molecule or gene, which can be used as indicator for normal or pathological biological processes or pharmacological reaction to therapeutic interventions. In modern cancer therapy, several gene biomarkers have meanwhile become established for diagnosis and therapy. In cancer, those biomarkers are associated with translocations (e.g. BCR-ABL t(9;22)) [19, 20], transmembrane receptors (e.g. EGFR) [20, 21, 22] , DNA repair system (e.g. BRCA2) [23, 24], tumor suppressors (e.g. TP53) [25, 26], anti apoptosis (e.g. BCL-2) [27, 28], transcriptions (e.g. MYC) [29, 30] or kinases (e.g. AKT1) [22, 31]. Figure 10.1 shows the expression profile of the transcription factor MYC in the compendium, Figure 10.2 of kinase AKT1, respectively. Further information, like the association of MYC to


**Figure 10.1:** The expression profile of MYC, generated with medicalgenomics.org. The upper bars exhibit the expression values in normal tissues, the lower bars in tumorous tissues. The red line indicates the Z-Score cut-off of 5. No data to RNA-Seq expression were found in the compendium.


**Figure 10.2:** The expression profile of AKT1, generated with medicalgenomics.org. The upper bars exhibit the expression values in normal tissues, the lower bars in tumorous tissues. The red line indicates the Z-Score cut-off of 5.

HCC, the connection to Gene Ontology terms like *regulation of transcription from RNA polymerase II promoter* or *transcription factor regulation*, or associated KEGG pathways like *ErbB signaling pathway* or *MAPK signaling pathway* are not displayed. The same applies for AKT1, which reveals among other things in the compendium associations to CCC, *protein modification process* or *negative regulation of apoptosis* in the Gene Ontology, or to *apoptosis* or *mTOR signialing pathway* in KEGG. The resulting profiles clearly exhibit the enriched expression levels of MYC and AKT1 in tumorous tissues compared to healthy tissues.
Additional analysis options with the compendium are conceivable and the described application areas on the thesis are limited to the author's scientific view. As an example, the salvaged knowledge of medicalgenomics.org about liver tissues was accessed to generate an interaction network of the human liver [4]. Those interaction networks provide substantial new insights into systems biology, disease research and drug discovery.

Because this compendium is the first of its kind, the direct comparison to other systems is not feasible. However, the comparison to the broadest sense related systems with focus on gene expression investigations in biomedical research reveals considerable advantages for using the presented compendium over the other systems.
The Gene Expression Atlas [32] is the only other database that provides access to cell line expression profiles. However, the main focus of this database is not dedicated to biomedical analysis and thus, it only contains the expression profile of around 90 cell lines from various species. In contrast, more than 300 different human cancer cell lines are stored in the workbench. In addition, the data foundation of CellLineNavigator was generated in the same laboratory under nearly the same experimental conditions and therefore, guarantees the highest degree on comparability. Moreover, each cell line is organized into pathological and physiological conditions. In comparison, Gene Expression Atlas data were collected from multiple laboratories and also classification into specific phenotypes is incomplete and therefore, exhibits multiple experimental conditions, making a comparison between the multiple expression profiles extremely difficult.
Compared to the GEO web application GEO2R [33] for identifying differentially expressed genes, the GUI of medicalgenomics.org is superior. GEO2R is just an extended frontend to R that allows an easy data import. Therefore, after the import the data are still in raw format and itself meaningless before analysis. Consequently, if the user has no substantial bioinformatics skills and experience in programming R the gene expression values remain coded. Within the compendium, all expression profile can be easily accessed and studied without prior knowledge of programming and bioinformatics. Because all expression profiles were analysed beforehand for gene expression by incisive developed computer algorithms and afterwards stored in the database in order to support a comprehensive knowledge base that can be easily accessed by researchers.

Advantages over medicalgenomics.org may be found in BioGPS [34]. This system supports more data than medicalgenomics.org. However, as it has been already stated above, the compendium was, among other things, designed in order to support comparative genomics analysis. Therefore, the data foundation was based on highly reliable data. Similar to Gene Expression Atlas, BioGPS data were collected from multiple laboratories and therefore, may be biased for comparability. In addition, BioGPS does not examine data on expression levels, but especially the knowledge of misleading gene regulation is essential for biomedical analysis.
A further advantage of the compendium compared to the other systems is the clearly arranged data representation. After querying the compendium, the results are displayed in an understandable graphical representation, which can be adjusted to the researchers needs to guarantee an accurate overview of the resulting data. This overview allows researchers the efficient identification of important genes in comparison to the non important ones. The overview can be united to form a detailed data picture by using the supported details view for each gene. In this view, the gene is moved into the spotlight so that all available information supported by the compendium for the selected gene can be accessed. Primary, this would include an accurate representation of the gene with respect to the executed query and its workbench (e.g. gene disease expression profile). But the researcher can also acquire the full knowledge of the gene existing in the compendium, which might broaden the horizon of the researcher and provide novel insights into not considered possibilities (e.g. gene disease expression profile plus associated normal expression profile and signaling events). Furthermore, the details view supports information on corresponding gene symbol, aliases, description, chromosomal location, Entrez gene ID as well as EnsEMBL gene ID, assembled from the NCBI Entrez and EnsEMBL databases [35, 36]. Additional outgoing links to HGNC [37], HPRD [38], OMIM [39], BioGPS [34], Nextbio [40], GENT [41], Kyoto Encyclopedia of Genes and Genomes [17] and Gene Ontology [18] databases are supported. Thus, enabling the researcher to acquire further knowledge of the selected gene beyond the scope of medicalgenomics.org. Information on gene symbol, Entrez gene ID as well as EnsEMBL gene ID are also supported in the overview. In addition, this view provides an export function as well as additional functional analysis by reporting the resulting genes directly to the Database for Annotation, Visualization and Integrated Discovery [42]. Therefore, the researchers have the possibility to investigate their generated gene lists in view on superior systematic levels. These additional analysis features and the assorted outgoing links to various databases are not supported by any similar system.

## 10.2 Outlook

In the near future, the RNA-Seq Atlas workbench in the compendium will be enlarged with more expression data to underline its pioneering position. In this close cooperation with the Translational Oncology Institute of the Johannes Gutenberg

University the interface and visualization will be improved firstly. Second, the consisting infrastructure will be updated with over 300 publicly available and locally processed RNA-Seq tissues, including normal adult, cell type, and cancer tissues. Third, sample reports will be created that include immune-relevant mutations. Fourth, the integrated query tools will be adjusted to allow effective sample analysis, such as *identify and return a breast tumor cell line expressing EGFR at over 10 RPKM with a BRAF V600E mutation* and *identify and return the genes expressed at over 10 RPKM in any melanoma sample, below 1 RPKM in all normal tissues, and having a kinase domain*. Finally, a broadcast feature should be implemented to interconnect the resultant gene lists to gene set enrichment pathway tools.
These extensions will enable the researchers to analyse and compare a large number of new pathological and physiological conditions. In addition, the interconnection to enrichment tools will allow the researchers to identify essential biological mechanisms in their analysed gene profiles.

# BIBLIOGRAPHY


[1] A. C. Baines, E. J. Adams, B. Zhang, and D. Ginsburg. Disruption of the sec24d gene results in early embryonic lethality in the mouse. *PloS one*, 8(4):e61114, **2013**. PMID: 23596517.

[2] J. D. Ziebarth, A. Bhattacharya, and Y. Cui. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic acids research*, 41(Database issue):D188–194, **January 2013**. PMID: 23193294.

[3] T. C. Freeman, A. Ivens, J. K. Baillie, D. Beraldi, M. W. Barnett, D. Dorward, A. Downing, L. Fairbairn, R. Kapetanovic, S. Raza, A. Tomoiu, R. Alberio, C. Wu, A. I. Su, K. M. Summers, C. K. Tuggle, A. L. Archibald, and D. A. Hume. A gene expression atlas of the domestic pig. *BMC biology*, 10:90, **2012**. PMID: 23153189.

[4] J. Wang, K. Huo, L. Ma, L. Tang, D. Li, X. Huang, Y. Yuan, C. Li, W. Wang, W. Guan, H. Chen, C. Jin, J. Wei, W. Zhang, Y. Yang, Q. Liu, Y. Zhou, C. Zhang, Z. Wu, W. Xu, Y. Zhang, T. Liu, D. Yu, Y. Zhang, L. Chen, D. Zhu, X. Zhong, L. Kang, X. Gan, X. Yu, Q. Ma, J. Yan, L. Zhou, Z. Liu, Y. Zhu, T. Zhou, F. He, and X. Yang. Toward an understanding of the protein interaction network of the human liver. *Molecular systems biology*, 7:536, **2011**. PMID: 21988832.

[5] S. Buchkremer, J. Hendel, M. Krupp, A. Weinmann, K. Schlamp, T. Maass, F. Staib, P. R. Galle, and A. Teufel. Library of molecular associations: curating the complex molecular basis of liver diseases. *BMC Genomics*, 11(1):189, **March 2010**. PMID: 20302666.

[6] F. Staib, M. Krupp, T. Maass, T. Itzel, A. Weinmann, J.-S. Lee, B. Schmidt, M. Müller, S. S. Thorgeirsson, P. R. Galle, and A. Teufel. CellMinerHCC: a microarray based expression database for hepatocellular carcinoma cell lines. *Liver International*, page n/a–n/a, **2013**.

[7] M. Krupp, T. Itzel, T. Maass, A. Hildebrandt, P. R. Galle, and A. Teufel. CellLineNavigator: a workbench for cancer cell line analysis. *Nucleic acids research*, **October 2012**. PMID: 23118487.

[8] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel. RNA-Seq atlas - a reference database for gene expression profiling in normal tissue by next generation sequencing. *Bioinformatics (Oxford, England)*, **February 2012**. PMID: 22345621.

[9] C. Wang, T. Maass, M. Krupp, F. Thieringer, S. Strand, M. A. Wörns, A.-P. Barreiros, P. R. Galle, and A. Teufel. A systems biology perspective on cholangio-

cellular carcinoma development: focus on MAPK-signaling and the extracellular environment. *Journal of Hepatology*, 50(6):1122–1131, **June 2009**. PMID: 19395114.

[10] J. Hou, M. Lambers, B. den Hamer, M. A. den Bakker, H. C. Hoogsteden, F. Grosveld, J. Hegmans, J. Aerts, and S. Philipsen. Expression profiling-based subtyping identifies novel non-small cell lung cancer subgroups and implicates putative resistance to pemetrexed therapy. *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, 7(1):105–114, **January 2012**. PMID: 22134068.

[11] M. Koti, R. J. Gooding, P. Nuin, A. Haslehurst, C. Crane, J. Weberpals, T. Childs, P. Bryson, M. Dharsee, K. Evans, H. E. Feilotter, P. C. Park, and J. A. Squire. Identification of the IGF1/PI3K/NFkB/ERK gene signalling networks associated with chemotherapy resistance and treatment response in high-grade serous epithelial ovarian cancer. *BMC cancer*, 13(1):549, **November 2013**. PMID: 24237932.

[12] A. Teufel, D. Becker, S. N. Weber, S. Dooley, K. Breitkopf-Heinlein, T. Maass, K. Hochrath, M. Krupp, J. U. Marquardt, M. Kolb, B. Korn, C. Niehrs, T. Zimmermann, P. Godoy, P. R. Galle, and F. Lammert. Identification of RARRES1 as a core regulator in liver fibrosis. *Journal of Molecular Medicine (Berlin, Germany)*, **June 2012**. PMID: 22669512.

[13] Y.-L. Chen, T.-H. Wang, H.-C. Hsu, R.-H. Yuan, and Y.-M. Jeng. Overexpression of CTHRC1 in hepatocellular carcinoma promotes tumor invasion and predicts poor prognosis. *PloS one*, 8(7):e70324, **2013**. PMID: 23922981.

[14] J.-S. Lee, J. Heo, L. Libbrecht, I.-S. Chu, P. Kaposi-Novak, D. F. Calvisi, A. Mikaelyan, L. R. Roberts, A. J. Demetris, Z. Sun, F. Nevens, T. Roskams, and S. S. Thorgeirsson. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nature medicine*, 12(4):410–416, **April 2006**. PMID: 16532004.

[15] M. Koch and M. Wiese. Gene expression signatures of angiocidin and darapladib treatment connect to therapy options in cervical cancer. *Journal of cancer research and clinical oncology*, 139(2):259–267, **February 2013**. PMID: 23052694.

[16] S. Nawrocki, T. Skacel, and T. Brodowicz. From microarrays to new therapeutic approaches in bladder cancer. *Pharmacogenomics*, 4(2):179–189, **March 2003**. PMID: 12605552.

[17] K. F. Aoki and M. Kanehisa. Using the KEGG database resource. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, Chapter 1:Unit 1.12, **October 2005**. PMID: 18428742.

[18] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, **May 2000**. PMID: 10802651.

[19] R. D. Press, C. Galderisi, R. Yang, C. Rempfer, S. G. Willis, M. J. Mauro, B. J. Druker, and M. W. N. Deininger. A half-log increase in BCR-ABL RNA predicts a higher risk of relapse in patients with chronic myeloid leukemia with an imatinib-induced complete cytogenetic response. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 13(20):6136–6143, **October 2007**. PMID: 17947479.

[20] P. La Rose, S. Holm-Eriksen, H. Konig, N. Hartel, T. Ernst, J. Debatin, M. C. Mueller, P. Erben, A. Binckebanck, L. Wunderle, Y. Shou, M. Dugan, R. Hehlmann, O. G. Ottmann, and A. Hochhaus. Phospho-CRKL monitoring for the assessment of BCR-ABL activity in imatinib-resistant chronic myeloid leukemia or ph+ acute lymphoblastic leukemia patients treated with nilotinib. *Haematologica*, 93(5):765–769, **May 2008**. PMID: 18367481.

[21] F. Baty, S. Rothschild, M. FrÃ¼h, D. Betticher, C. Droge, R. Cathomas, D. Rauch, O. Gautschi, L. Bubendorf, S. Crowe, F. Zappa, M. Pless, M. Brutsche, and Swiss Group for Clinical Cancer Research. EGFR exon-level biomarkers of the response to Bevacizumab/Erlotinib in non-small cell lung cancer. *PloS one*, 8(9):e72966, **2013**. PMID: 24039832.

[22] D. Dionysopoulos, K. Pavlakis, V. Kotoula, E. Fountzilas, K. Markou, I. Karasmanis, N. Angouridakis, A. Nikolaou, K. T. Kalogeras, and G. Fountzilas. Cyclin d1, EGFR, and Akt/mTOR pathway. potential prognostic markers in localized laryngeal squamous cell carcinoma. *Strahlentherapie und Onkologie: Organ der Deutschen Rontgengesellschaft ... [et al]*, 189(3):202–214, **March 2013**. PMID: 23400686.

[23] J. R. Diamond, V. F. Borges, S. G. Eckhardt, and A. Jimeno. BRCA in breast cancer: from risk assessment to therapeutic prediction. *Drug news & perspectives*, 22(10):603–608, **December 2009**. PMID: 20140280.

[24] I. Locke, Z. Kote-Jarai, E. Bancroft, S. Bullock, S. Jugurnauth, P. Osin, A. Nerurkar, L. Izatt, G. Pichert, G. P. H. Gui, and R. A. Eeles. Loss of heterozygosity at the BRCA1 and BRCA2 loci detected in ductal lavage fluid from BRCA gene mutation carriers and controls. *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 15(7):1399–1402, **July 2006**. PMID: 16835343.

[25] H.-L. Fu, L. Shao, Q. Wang, T. Jia, M. Li, and D.-P. Yang. A systematic review of p53 as a biomarker of survival in patients with osteosarcoma. *Tumour biology: the journal of the International Society for Oncodevelopmental Biology and Medicine*, **September 2013**. PMID: 24014053.

[26] L. Li, M. Fukumoto, and D. Liu. Prognostic significance of p53 immunoexpression in the survival of oral squamous cell carcinoma patients treated with surgery and neoadjuvant chemotherapy. *Oncology letters*, 6(6):1611–1615, **December 2013**. PMID: 24260053.

[27] A. Stone, M. J. Cowley, F. Valdes-Mora, R. A. McCloy, C. M. Sergio, D. Gallego-Ortega, C. E. Caldon, C. J. Ormandy, A. V. Biankin, J. M. W. Gee, R. I. Nicholson, C. G. Print, S. J. Clark, and E. A. Musgrove. BCL-2 hypermethylation is a potential biomarker of sensitivity to antimitotic chemotherapy in endocrine-resistant breast cancer. *Molecular cancer therapeutics*, 12(9):1874–1885, **September 2013**. PMID: 23861345.

[28] Q. Gao, S. Yang, and M.-Q. Kang. Influence of survivin and bcl-2 expression on the biological behavior of non-small cell lung cancer. *Molecular medicine reports*, 5(6):1409–1414, **June 2012**. PMID: 22446832.

[29] F. Pedica, A. Ruzzenente, F. Bagante, P. Capelli, I. Cataldo, S. Pedron, C. Iacono, M. Chilosi, A. Scarpa, M. Brunelli, A. Tomezzoli, G. Martignoni, and A. Guglielmi. A re-emerging marker for prognosis in hepatocellular carcinoma: the add-value of fishing c-myc gene for early relapse. *PloS one*, 8(7):e68203, **2013**. PMID: 23874541.

[30] M. D. Planas-Silva, R. D. Bruggeman, R. T. Grenko, and J. S. Smith. Overexpression of c-myc and bcl-2 during progression and distant metastasis of hormone-treated breast cancer. *Experimental and molecular pathology*, 82(1):85–90, **February 2007**. PMID: 17046747.

[31] S. Matsuda, A. Nakanishi, Y. Wada, and Y. Kitagishi. Roles of PI3K/AKT/PTEN pathway as a target for pharmaceutical therapy. *The open medicinal chemistry journal*, 7:23–29, **2013**. PMID: 24222802.

[32] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–872, **January 2009**. PMID: 19015125.

[33] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, **January 2013**. PMID: 23193258 PMCID: PMC3531084.

[34] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. Hodge, J. Haase, J. Janes, J. Huss, and A. Su. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10(11):R130, **November 2009**.

[35] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(Database issue):D38–51, **January 2011**. PMID: 21097890.

[36] G. Spudich, X. M. Fernández-Suárez, and E. Birney. Genome browsing with ensembl: a practical overview. *Briefings in Functional Genomics & Proteomics*, 6(3):202–219, **September 2007**. PMID: 17967807.

[37] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, and E. A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, 39(Database issue):D514–519, **January 2011**. PMID: 20929869.

[38] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–772, **January 2009**. PMID: 18988627.

[39] Online mendelian inheritance in man, OMIM®. McKusick-Nathans institute of genetic medicine, johns hopkins university (baltimore, MD).

[40] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. 5(9). PMID: 20927376 PMCID: 2947508.

[41] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim. GENT: gene expression database of normal and tumor tissues. *Cancer Informatics*, 10:149–157, **2011**. PMID: 21695066.

[42] X. Jiao, B. T. Sherman, D. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806, **July 2012**. PMID: 22543366 PMCID: PMC3381967.

# A Copyrights

Some figures in this thesis are protected by copyright, and thus license numbers for previously published content have been acquired through the Copyright Clearance Center. In the following list, we give the original sources and the license number where appropriate.

- Figure 2.1 is adopted from the National Library of Medicine, "link: http://www.nlm.nih.gov/bsd/stats/data_entry_bar.gif ". NLM Web sites is in the public domain.
- Figure 2.2 is adopted from the National Library of Medicine, "link: http://www.nlm.nih.gov/bsd/stats/cit_added_FY.gif ". NLM Web sites is in the public domain.
- Figure 3.1 is used with permission from *Molecular BioSystems*: F. Seela and S. Budow. Mismatch formation in solution and on DNA microarrays: how modified nucleosides can overcome shortcomings of imperfect hybridization caused by oligonucleotide composition and base pairing.,4:232–245, **2008**. License number: 3281950293892
- Figure 3.2 is adapted from the Wikimedia Commons file "File:Microarray-schema.jpg". The copyright holder release it into the public domain.
- Figure 4.3 is used from *Annu.Rev,Genomics Hum Genet.* with no need to obtain permission: E. R. Mardis. Next-Generation DNA Sequencing Methods,9:387-402, **2008**.
- Figure 4.4 is used from *Annu.Rev,Genomics Hum Genet.* with no need to obtain permission: E. R. Mardis. Next-Generation DNA Sequencing Methods,9:387-402, **2008**.
- Figure 4.2 is used from *Annu.Rev,Genomics Hum Genet.* with no need to obtain permission: E. R. Mardis. Next-Generation DNA Sequencing Methods,9:387-402, **2008**.