Omics-Scale Bioinformatics Technology and Methods: from Data to Information

A Dissertation

Presented to

Faculty of the Department of Biology and Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Haosi Chen

December 2014

Omics-Scale Bioinformatics Technology and Methods: from Data to Information

----

Haosi Chen

APPROVED

Dr. Xiaolian Gao,

Chair of the Committee

Dr. James M. Briggs

Dr. Cecilia M. Williams

Dr. Fuli Yu

Dean, College of Natural

Sciences and Mathematics

Acknowledgements

I am using this opportunity to express my sincere appreciation to my advisor, Dr. Xiaolian Gao, for her inspiring guidance and for the tremendous learning opportunities she provided for my Ph.D. studies at the University of Houston. I also appreciate the rest of my committee, Drs. James Briggs, Cecilia Williams and Fuli Yu, for their invaluable constructive criticism and friendly advice during my graduate studies.

I am very grateful to all my collaborators: Ms. Ruijuan Zhu, she made great contributions for the miRFocus project; Ms. Chengping Sui, she collected the data for the miRNA-miRNA relationship database in the miRFocus project; Ms. Wen Wan, a student in University of Science and Technology of China, she performed the experiments of the Microchip-synthesized DNA project and generously shared the data with me; Dr. Qi Zhu, senior scientist at LC Sciences, he sequenced the nucleotides for the Microchip-synthetic DNA project; Dr. Bing Zhu, he performed the experiment of the Histone PepArrary chip project; Dr. Ailing Hong, senior chemist at LC Sciences, she made the PepArrary chips for the Histone PepArray project. And I also appreciate Dr. Wei Huang, Celise Robertson, Henry Vo for their valuable suggestions for revising my thesis.

I would like to thank my family and all my friends, for their company and support which extends beyond my Ph.D. student life.

Finally, I would like to specifically thank my wife. She is the most precious person to me in the world. Her invaluable help is essential for my completion of the thesis and graduate with Ph.D. degree.

Omics-Scale Bioinformatics Technology and Methods: from Data to Information

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Haosi Chen

December 2014

Abstract

Omics-scale bioinformatics is an emerging discipline of science that plays an essential role in analyzing and interpreting large scale biological data. In this thesis, I developed three different omics-scale bioinformatics methods to facilitate the studies of human miRNAs, synthetic DNA oligo library and histone post-translational modifications (PTMs), respectively.

MiRNAs, which are involved in various biological processes by regulating multiple genes, have been an area of research drawing intensive interest in the recent two decades. There exists an enormous amount of miRNA related information, and how to effectively mine the valuable information embedded in the large volume of literature has become an urgent problem. Because each of the existing online databases includes only partial information about human miRNAs, I created a comprehensive web-based resource ‘miRFocus’ for conveniently retrieving extensive and comprehensive human miRNA information and conducting pathway and Gene Ontology (GO) term enrichment analysis.

Current next-generation sequencing (NGS) technologies mainly focus on genome or transcriptome sequencing analysis and none of the existing NGS methods is suitable for high resolution nucleobase-specific analysis of libraries of synthetic oligonucleotides, which are used as materials for engineering long DNA fragments in synthetic biology applications. To meet such requirements, I developed an algorithm and software tool for analyzing synthetic oligo libraries. This approach is composed of two-step quality control and Bowtie2-based sequence alignments. It is proved that such a method successfully assessed the efficiency of etMICC-based error-removal method on synthetic oligos of different lengths and identified that etMICC columns has higher binding affinity with gap error structure than substitution error structure.

Epiproteomics examines diverse PTMs, such as histone methylation. However, traditional methods of studying histone PTMs are expensive in cost, labor and time. I developed a histone peptide array (hPepArray) for analyzing activities of cellular histone methyltransferases (HMTs). Lysine-containing peptides of hPepArray are directly generated from 10 histone proteins. In the hPepArray, two known methylation sites H3K122 and H4K59 are verified and one possible methylation site H2A-K74 is identified. The experimental results demonstrate that hPepArray and the method of analysis offer a high-throughput epiproteomic tool to assay activities of HMTs in nuclear lysates.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

bp	base pair
ddNTP	dideoxynucleotide triphosphate
DNA Array	DNA microarray
*Eco*MutS	*Escherichia coli* MutS
etMICC	etMutS immobilized cellulose column
etMutS	A mixture of *Eco*MutS and *Taq*MutS
GeneRIF	Gene Reference into Function database
GO	Gene Ontology
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
HMT	histone methyltransferase
hPepArray	histone peptide array
KEGG	Kyoto Encyclopedia of Genes and Genomes
KHMT	lysine-specific histone methyltransferase
lowess	locally weighted scatterplot smoothing
LWFP	long wave fluorescent protein
miRNA	microRNA
mRNA	message RNA
MutS	mismatch-binding protein
NCBI	national center for biotechnology information
NGS	next-generation sequencing

oligo	oligonucleotide
OMIM	Online Mendelian Inheritance in Man
PepArray	peptide microarray
PPI	protein-protein interaction
PTM	post-translational modification
RHMT	arginine-specific histone methyltransferase
RNA-Seq	high-throughput mRNA sequencing
SAM	S-Adenosyl methionine
SM	side mutation
SNP	single-nucleotide polymorphism
*Taq*MutS	*Thermus aquaticus* MutS
WT	wild type

Chapter 1

Introduction

1.1 Omics-Scale Life Sciences

1.1.1 Omics

When James Watson and Francis Crick first discovered, in 1953, the double helix structure of DNA [1], life sciences entered a new era of modern biology. In the past decades, more and more technologies have been invented, and more and more discoveries have led to the exponential knowledge growth in the fields of biochemistry, molecular biology, genetics and other life science disciplines. With the innovation of technologies and reduction of experimental costs, scientists are no longer satisfied with the study of individual molecules (such as gene, RNA and protein). They are beginning to be interested in the collection of many elements of the subject of their research. The format of scientific research is undergoing a transition from simplex to multiplex to system scale.

Omics refers to the study of collective elements that is used to explore the functions and relationships of the various types of molecules. Based on the types of biological molecules, omics can be divided into different fields, such as genomics, proteomics, glycomics, lipidomics, etc. Moreover, in light of the discrepancy of functions of the molecules, it can be further subdivided, for instance, genomics can be divided into cancer genomics, immunomics, epigenomics, and so on. Among the omics studies, the most well-known one is the Human Genome Project (HGP). HGP was an international scientific project, costing roughly $3-billion with the goal to benefit mankind. The project was founded in 1990 and announced completed in April 2003, having verified 99% of the human genome with 99.99% accuracy [2]. Although HGP was finished, as revealed by the project [3], more detailed analyses are still needed for genome functions.

1.1.2 Systems Biology

As the scope of research objects has been largely extended, scientists realize that it is not only difficult to examine the relationships between molecules, but also difficult to characterize clearly the function of a single molecule if only focused at the single molecule level. As a result, they turn to integrate the biological information in collaborative ways, which is the core idea of systems biology.

Systems biology, a biology-based inter-disciplinary study, aims at studying and interpreting biological phenomena at the macro-level by exploring the interactions of various molecules in the same system, for example, signaling network analysis, or by examining discrepancies of one type of biological molecules in the same biological system. The concept of systems biology has been widely used in various biosciences, e.g., HGP.

In the human genome, only 1.5% sequences of total 3.3 billion DNA base-pairs are coding DNA, which can be transcribed and translated into proteins, while more than 98% of the genome is consisted of noncoding DNAs, also called "Junk DNAs", without any particular known functions. Surprisingly, when researchers studied a part of the "Junk DNAs" at the systems level, they found that although the "Junk DNAs" themselves do not have any special function of conventional interpretation, some of it can regulate the expression and cellular activities of certain proteins through their particular structural properties or their transcribed RNAs [4]. Thus, "Junk DNAs" can actually be important regulatory molecules.

Among a variety of regulatory genes, microRNAs (miRNA) first discovered in 1990s is one of the well-known families [5]. MiRNAs are small non-coding RNAs, consisting of 16 ~ 28 nucleotides, and by prediction, they regulate about 60% of human genes [6]. Consequently, miRNAs have impact on all the pathways related to their target genes. Certainly, such a large impact won’t be noticed if research was only focused on one or a few molecules.

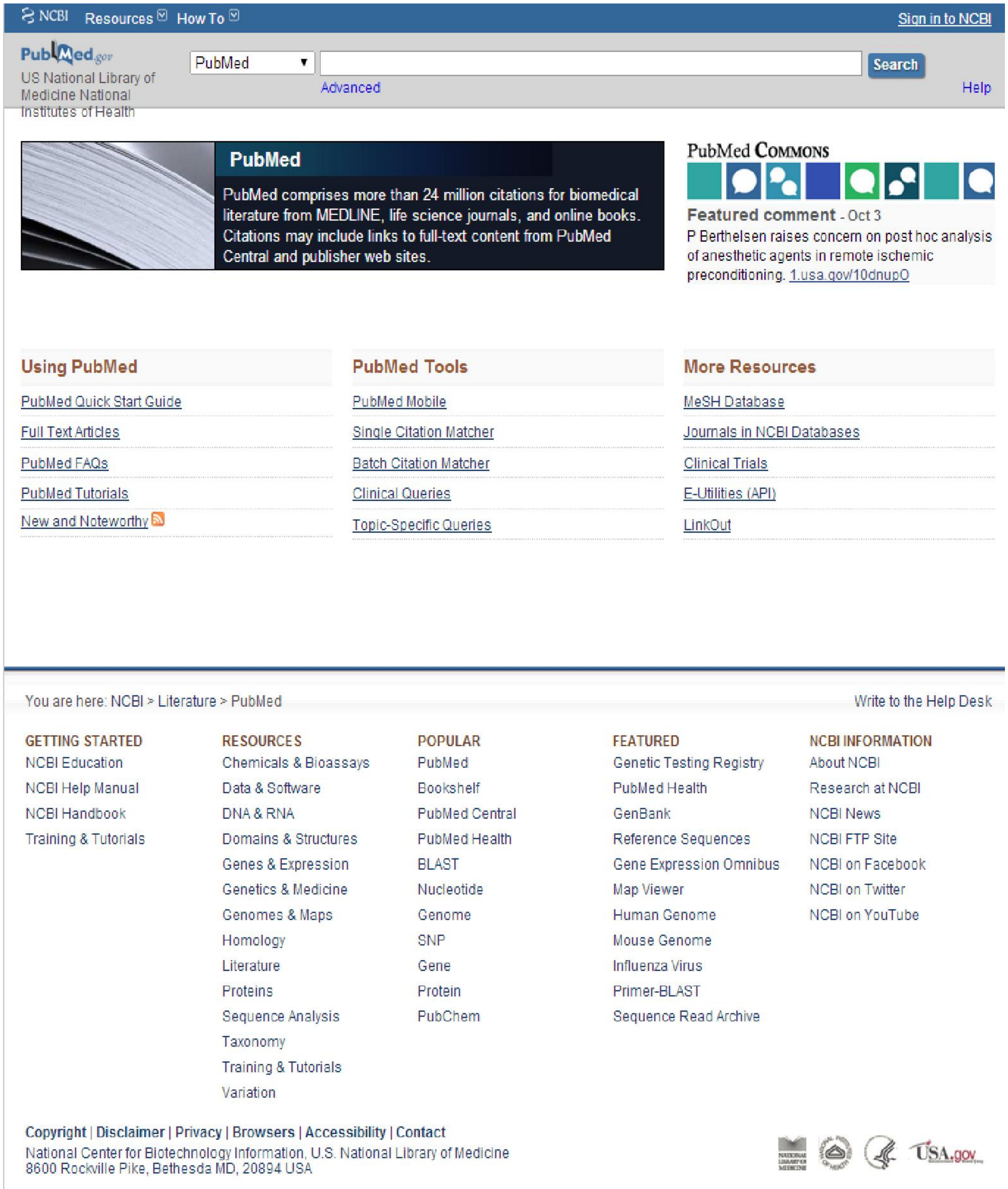
1.2 Bioinformatics

1.2.1 Background

With the dramatic increase in the quantity of experimental data, coupled with improvement of the depth and scope of study objects, it is unreasonable to store and analyze the data just with pen and paper. The prevalence of computers and internet communications provided a good opportunity to solve the problems caused by large data sets. It not only facilitates people's daily lives, but also offers great help to researchers' work. In order to dig out useful information from the mountains of biological data, knowledge of mathematics, databases, software tools, statistics and computer science are all integrated to develop the analytical methods. With the implementation and improvement of these methods, a new research field emerged. In 1970, Paulien Hogeweg and Ben Hesper used the term bioinformatics to refer to this new field [7]. Bioinformatics is an interdisciplinary field that processes the biological data with methods developed by combining mathematics, statistics and computer science. In the last few years, more and more investigations have entered bioinformatics because of its important role in biology, which highly promoted the development of bioinformatics and life sciences.

1.2.1 Databases

In the face of giant biological data, how to organize and store them effectively became an urgent problem. Databases are one of the most important means for biologists to store the acquired biological data, such as DNA or protein sequences and related information, with the assistance of computer scientists. It offers a user-convenient way to search and query. The most widely accessed literature database is PubMed, which is hosted by National Center for Biotechnology Information (NCBI), affiliated with the United States National Library of Medicine (NLM). The PubMed database is also called Medical Literature Analysis and Retrieval System Online (MEDLINE) Database. It is a database of article references and abstracts related to scientific subjects, including, chemistry, physics, mechanics, life sciences, medicine, health care and other topics [8]. Besides PubMed，NCBI also maintains other major databases, including GenBank for DNA sequences and GEO for the result of gene expression analysis. In addition to the databases authorized by NCBI, there are many other distinguished databases, including UniProt (A database of protein sequences and their functional information, developed by European Bioinformatics Institute (EBI), Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR)) [9], and GeneCards (A database of human genes and their related information, maintained by the Crown Human Genome Center at the Weizmann Institute of Science) [10].



NCBI Resources How To Sign in to NCBI

PubMed.gov
US National Library of Medicine National Institutes of Health

PubMed ▼ Search

Advanced Help

PubMed

PubMed comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubMed Commons

Featured comment - Oct 3

P Berthelsen raises concern on post hoc analysis of anesthetic agents in remote ischemic preconditioning. 1.usa.gov/10dnupO

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities (API)
- LinkOut

You are here: NCBI > Literature > PubMed

Write to the Help Desk

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials

RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Gene Expression Omnibus
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube

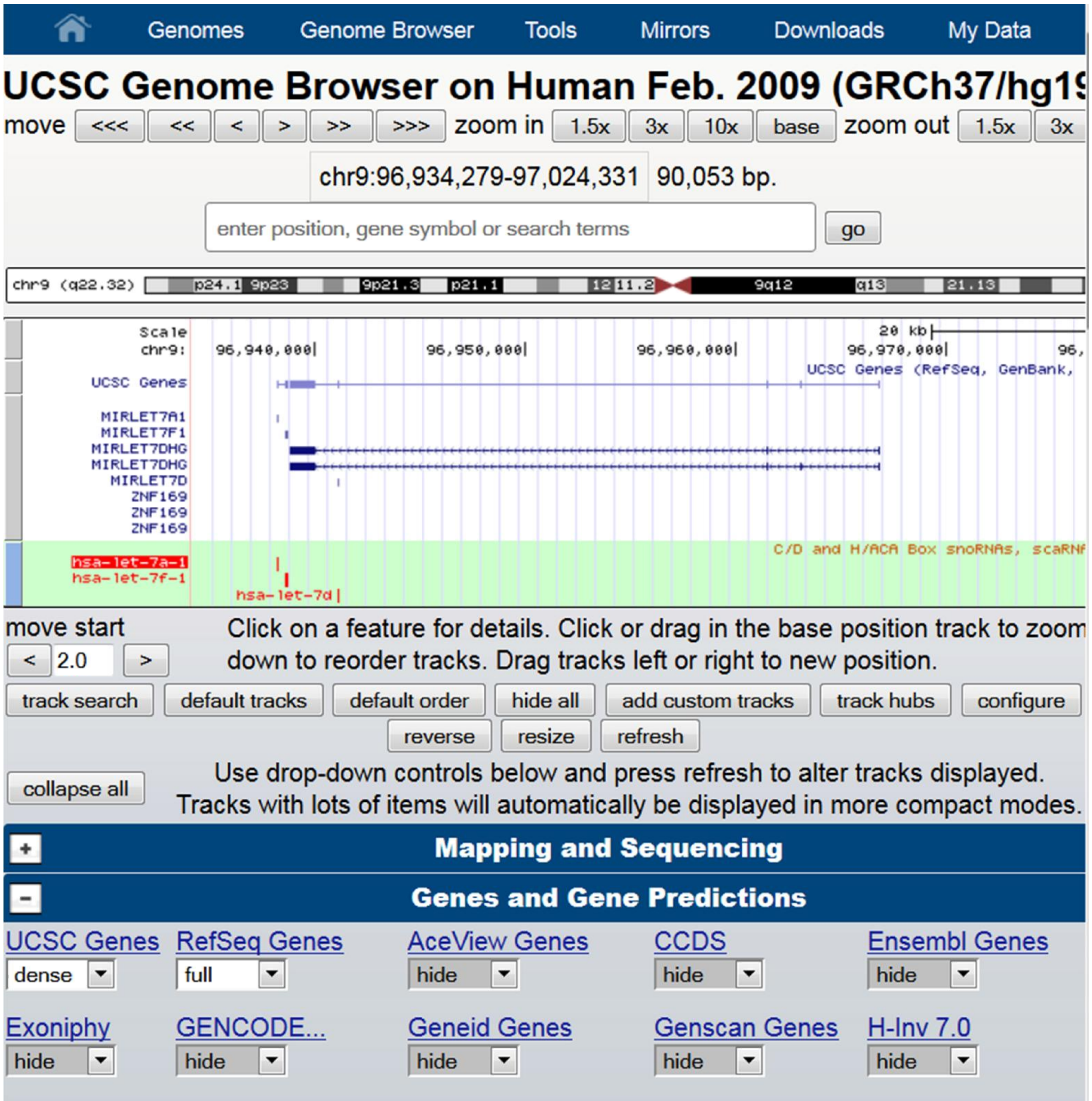
Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Figure 1-1. The main page of PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/).

1.2.3 Data Analysis and Visualization

With the large amount of biological data collected from many different systems at different levels, how can researchers link the related information together, and how can they extract logical and meaningful knowledge from a mass amount of information? Data mining is the answer. With the development and popularity of new technologies, a massive amount of data can be easily collected at one time from high-throughput experiments, such as NGS and microarray. With the large amount of data in hand, first, researchers have to analyze the data, extracting the meaningful information and removing the useless information; second, researchers wish to visualize the data, to convert the meaningful data into easy forms for reading and understanding. Nowadays several excellent databases with the visualization function are accessible freely. University of California Santa Cruz Genome Browser (UCSC) [11] and Cufflinks [12] are both such widely used tools. UCSC, hosted by the University of California, Santa Cruz, refers to annotating genes by aligning the existing biological information to genome sequences and displaying the results in a graphical viewer (Figure 1-2) [13]. Cufflinks aims at discovering validated and novel message RNA (mRNA), estimating the abundance of transcripts and profiling the transcriptome through analyzing high-throughput mRNA sequencing (RNA-Seq) data (Figure 1-3) [14].



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x

chr9:96,934,279-97,024,331 90,053 bp.

enter position, gene symbol or search terms

Click on a feature for details. Click or drag in the base position track to zoom
down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

UCSC Genes	RefSeq Genes	AceView Genes	CCDS	Ensembl Genes
dense	full	hide	hide	hide
Exoniphy	GENCODE...	Geneid Genes	Genscan Genes	H-Inv 7.0
hide	hide	hide	hide	hide

Figure 1-2. Screen image of UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr9%3A96934279-97024331). The genome browser includes a text input bar and several buttons for quick access to genome location search, different annotations as text tables and parts of select buttons for track-specific options.

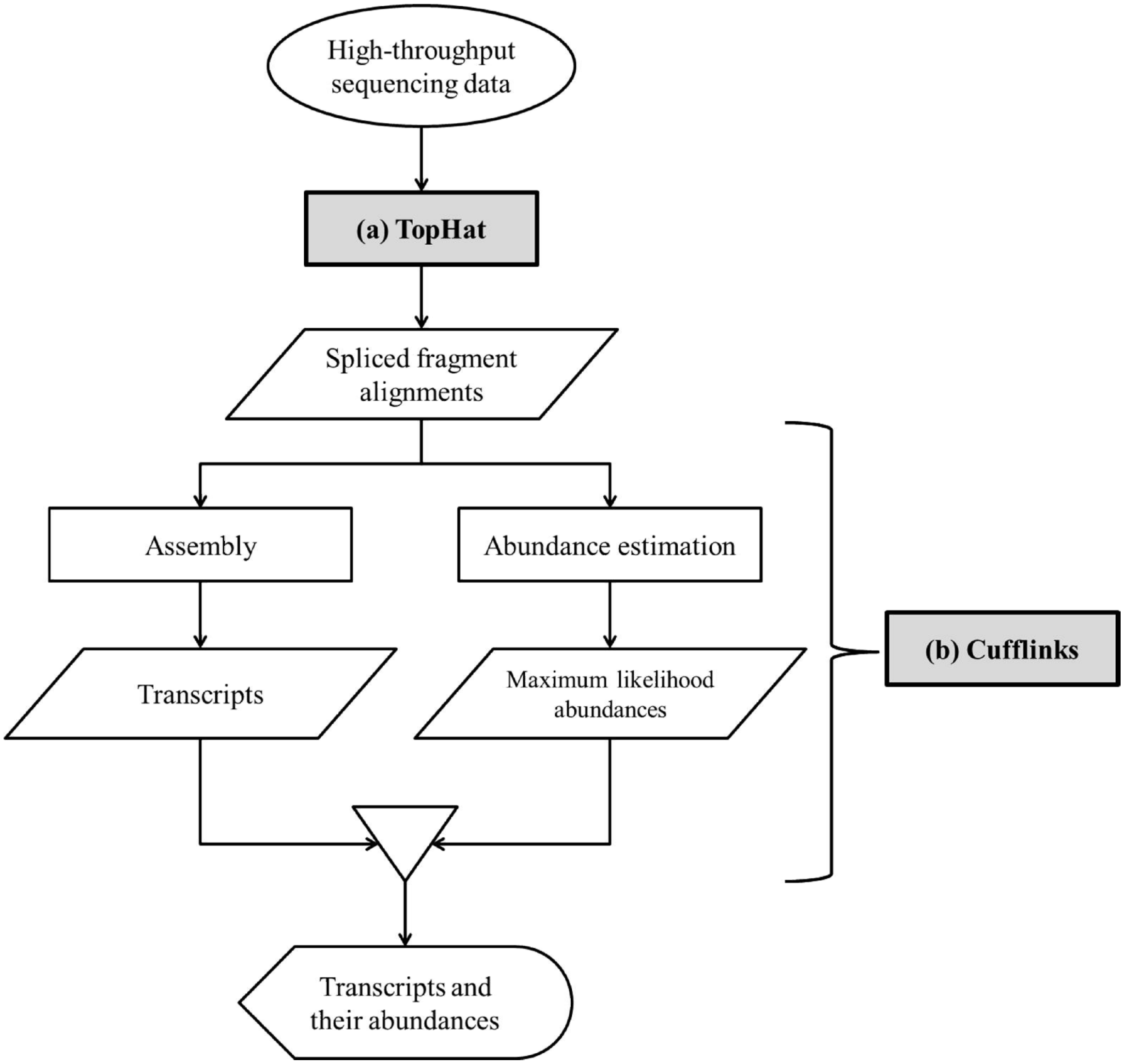

High-throughput sequencing data
(a) TopHat
Spliced fragment alignments
Assembly
Abundance estimation
Transcripts
Maximum likelihood abundances
(b) Cufflinks
Transcripts and their abundances

Figure 1-3. Flowchart of Cufflinks: (a) Align high-throughput sequencing data to reference genome by the alignment tool such as TopHat. (b) Assemble the mapping result to possible transcripts, and estimate these transcripts' maximum likelihood abundances.

1.2.4 Simulation and Prediction

Bioinformatics can not only analyze and process existing experimental data, but also in turn guide and design experiments. Through the years, the advancement of biotechnology has lowered the experimental costs, however, the large number of experiments are still time-consuming and expensive. To save time, energy and money, biological models, constructed by making use of the experimental data and effective information mined from it, are used to predict the processes and results of other related experiments. These methods are widely used in 3D structure simulation, assisting the drug discovery and drug design. Even without understanding of the mining information, researchers can obtain the results of future experiments only with the help of the tools based on statistics and informatics. In the biology field, this method is mainly used in the prediction tools related to sequence, including secondary structure prediction and miRNA-target interaction prediction. There are many such methods available. In particularly, one of the most well-known methods is machine learning, such as support vector machines (SVM) [15] and artificial neural networks (ANN) [16].

1.3 DNA Sequencing

1.3.1 Background

In 1953, the structure of DNA was first discovered by James Watson and Francis Crick [1]. More than ten years later, in the early 1970s, the first DNA sequence was acquired by an academic research group [17]. Almost two decades later, in 1990, HGP, an international scientific project, was formally launched [2]. Six years ago, in 2008, the 1000 Genomes Project, an international effort focusing on building the detailed catalogue of human genetic variations, was announced to start [18]. It is obvious that as time goes on, more giant projects will emerge, which are implemented not by individual groups or cooperative groups, but rather depending on the international collaborative efforts from multiple fields. From the development of DNA sequencing, we can see that life science research has advanced from simple to complex. For example, DNA sequencing, at first, could only determine one DNA sequence of 50 base pairs (bp). Nowadays, it can identify the DNA sequence up to 3 billion reads of 50-300 bp per run. The price of DNA sequencing has decreased dramatically from \$10 million to \$0.1 per million bases [19]. The application of DNA sequencing technology has also been largely extended. For example, at the very beginning, DNA sequencing technology could only be used in discovering and verifying DNA sequences, by this time, it has been made use of in transcriptome profiling, DNA-protein interactions analysis, single-nucleotide polymorphism (SNP) calling and others [20]. With the cost reduction and running speed acceleration of the sequencing technology, scientists are devoted to applying the DNA sequencing technology in medical research for the benefit of all mankind.

1.3.2 First-Generation Sequencing

Two major DNA sequencing methods were developed at the beginning. One is chemical degradation method developed by Allan Maxam and Walter Gilbert in 1977, also named Maxam-Gilbert sequencing. Maxam-Gilbert sequencing is a method to determine the location of each specific base by using gel electrophoresis to define the size of degraded DNA fragments, which are obtained by breaking down the template DNA sequences with radioactive labeling at the 5' end by different chemical reactions [21]. Another is chain-termination method proposed by Frederick Sanger *et al*. around the same period and therefore named Sanger sequencing. Sanger sequencing is a method to identify the location of each specific base by making use of gel electrophoresis separating and analyzing the size of the DNA fragments. Unlike Maxam-Gibert sequencing, the DNA fragments are generated when the process of DNA replication is randomly discontinued by dideoxynucleotide triphosphate (ddNTP) with radioactive labeling [22]. Sanger sequencing introduced error because of employing DNA polymerase. However, its advantages are obvious. The easier operation and less toxicity of Sanger sequencing make the wider range of application of Chain-termination method than of Chemical degradation method.

Due to the broad range of application of Sanger sequencing, scientists have made a number of improvements on it. For instance, by replacing radioisotopes with four different fluorescent dyes to label ddNTP, scientists not only improved the security of the sequencing technology, but also reduced the signal-to-noise ratio [23]. Furthermore, after substituting traditional gel electrophoresis with capillary gel electrophoresis, it not merely reduced the use of materials, but also improved the speed of reading fluorescence in the process of sequencing with the Sanger sequencing method [24]. Therefore, automated laser-fluorescence sequencing, which is based on Sanger sequencing, played a vital role in the HGP.

With the significant improvements, the Sanger sequencing method is largely applicable to analyze DNA fragments shorter than 1000 bps. To determine DNA sequences longer than 1000 bps, especially like chromosomes, scientists put forward shotgun sequencing. In shotgun sequencing, the principal change, compared with traditional DNA sequencing, is that it added at the beginning the step to prepare numerous DNA fragments, shorter than the defined limits, by breaking up randomly the targeted sequences using physical methods, chemical methods, or biological methods and constructing libraries. The following step is to sequence the DNA libraries through traditional DNA sequencing. The last step is to assemble the sequenced data using bioinformatics approaches [25].

1.3.2 Next-Generation Sequencing

Sequencing long DNA strands was implemented with shotgun sequencing. However, this is still time-consuming and expensive. Because of high demand for solutions of previous problems, NGS was developed.

NGS methods mainly include Roche 454 pyrosequencing, Illumina dye sequencing, SOLiD sequencing, and Ion Torrent semiconductor sequencing which was released in 2010. These methods, coupled with respective advantages and disadvantages, use different implementations, whereas they all aim to sequence high-throughput data by massive parallel sequencing. The workflow of Illumina sequencing is taken as an example here to introduce the basic process of NGS.

1.3.2.1 Library Preparation

In order to determine long DNA sequences, first, target DNA needs to be broken up into DNA fragments within the defined length range, in line with the idea of shotgun sequencing. Second, primers that are used for PCR amplification and adaptors are added to the DNA fragments. With this procedure, target DNA sequences can be either genomic DNA, cDNA library that is used for RNA-Seq, or even *de novo* synthesized oligonucleotides, which will be introduced in Chapter 3 (Figure 1-4 a1) [26].

1.3.2.2 PCR Amplification

PCR amplification aims to amplify the sequencing signal and reduce the background noise. It consists of three steps: attaching the single-strand DNA fragments to flow cell surface, PCR amplification to get DNA clusters, and denaturing the double-stranded DNA fragments. Unlike Illumina, in the PCR amplification, some of the other NGS technologies make use of beads instead of flow cell surface, and substitute bridge PCR with emulsion PCR for DNA fragments amplification (Figure 1-4 a2-4).

1.3.2.3 Sequencing to Data

As the core part of DNA sequencing, Illumina adopts sequencing-by-synthesis approach using reversible terminators to identify the sequences. Reversible terminators are the nucleotides with a base-unique fluorescent dye and 3' terminal group. Sequencing one base consists of three steps: 1) adding terminators and DNA polymerase into DNA clusters, thus linking terminators to DNA clusters, which is catalyzed by DNA polymerase; 2) removing uncombined terminators, then determining and storing the first base for each cluster by reading base-specific fluorescence emitted from the terminators; 3) trimming the blocking group and fluorescent label from binding terminators. Eventually, the whole-base sequence hybridizing with each DNA cluster would be determined by repeating steps from 1 to 3. Sequencing-by-synthesis approach primarily came from Sanger sequencing, one of the first-generation sequencing methods. This approach has been widely applied to some sequencing platforms, for example, Ion Torrent semiconductor sequencing. Ion Torrent semiconductor sequencing does not use the traditional fluorescence detecting technology, rather employs semiconductor which is used to detect the variation of H+ ion concentration emitted during the synthesis. On the other hand, sequencing-by-synthesis is not adopted by such other sequencing methods as SOLiD Sequencing, in which DNA ligase-mediated sequencing approach is used.

1.3.2.4 Further Analysis

The raw data directly from DNA sequencers need to be further analyzed for certain experimental purposes. For instance, to achieve the whole genomic DNA or cDNA sequences, analyses such as transcriptome profiling, DNA-protein interactions analysis, and SNP calling, are necessary with the aid of bioinformatics tools. The knowledge as to further analysis of DNA sequencing based on bioinformatics will be introduced in Section 1.3.3.

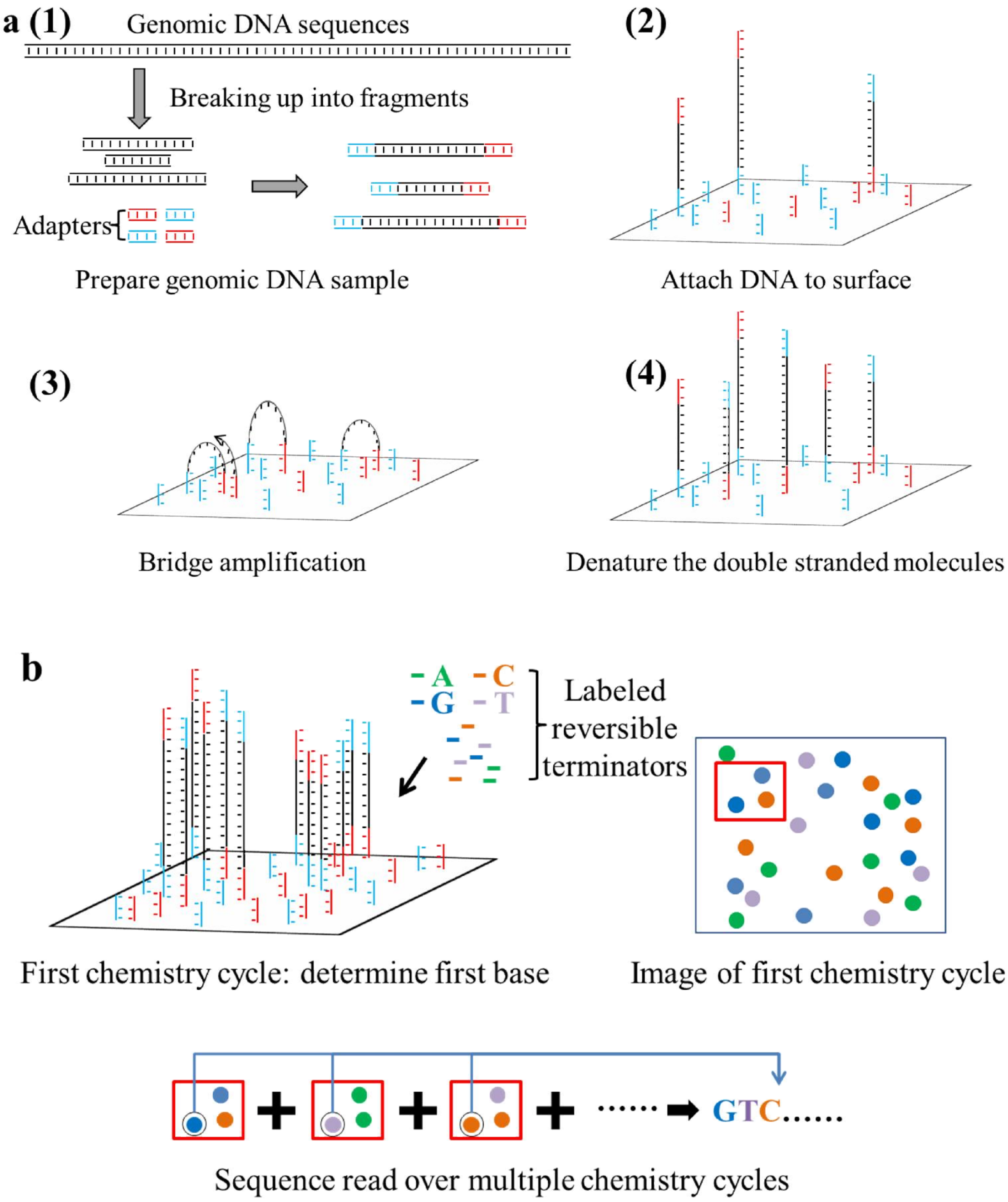

a (1)
Genomic DNA sequences
Breaking up into fragments
Adapters
Prepare genomic DNA sample
(2)
Attach DNA to surface
(3)
Bridge amplification
(4)
Denature the double stranded molecules
b
A
C
G
T
Labeled reversible terminators
First chemistry cycle: determine first base
Image of first chemistry cycle
+
+
+
......
GTC......
Sequence read over multiple chemistry cycles

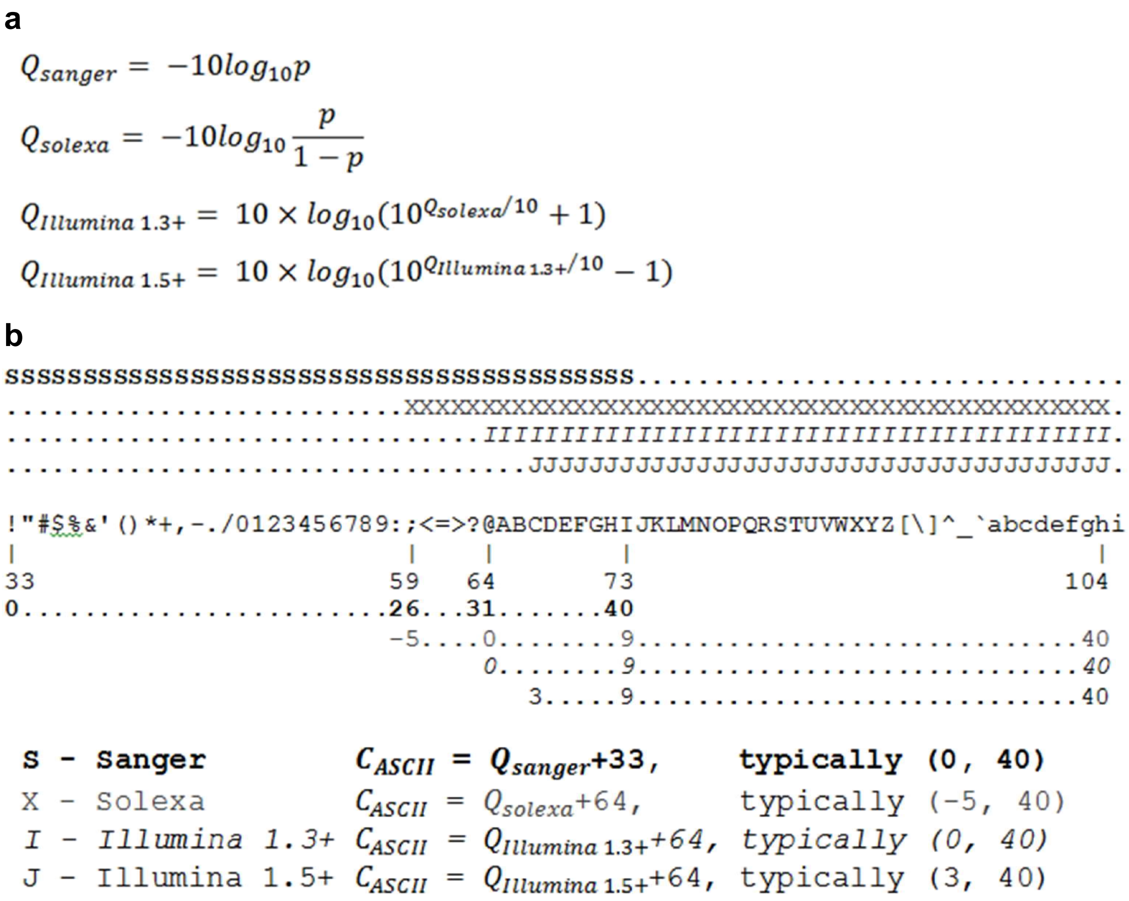
Figure 1-4. Illumina sequencing approach. (a) Library preparation and PCR amplification; (b) Sequencing to data.

1.3.3 Bioinformatics in DNA Sequencing

1.3.3.1 Raw Sequencing Output

Strictly speaking, the primary data from DNA sequencers based on fluorescence detection as Illumina sequencing is image data. Since the image data is in large quantity, thus difficult to process, DNA sequencers usually automatically, with the aid of the instant image processing software according to the electrical physical and chemical properties of the sequencers, transform the image data into digital data containing the sequences and their corresponding quality. Nevertheless, the file formats of different sequencers may not be exactly the same. For example, Illumina sequencing makes use of the FASTQ format based on Phred Scores, while SOLid employs CSFASTA format on the basis of color-space. Moreover, even if they are from the same platform in the same company, the file formats might vary with different versions of the sequencers. For instance, as to the Illumina sequencing platform, different versions of sequencers utilize different types of FASTQ format, such as Solexa/Illumina 1.0 FASTQ, Illumina 1.3+ FASTQ and Illumina 1.5+ FASTQ. Therefore, at the beginning of popularizing NGS, conversion among diverse file formats is one of the crucial processes. Fortunately, as a great amount of bioinformatics tools regarding NGS have been developed, scientists notice that it is very important to unify the file formats, and thus Sanger FASTQ has become the *de facto* standard file format [27]. So far, most of the sequencers provide the option to output the data in Sanger FASTQ format.

Sanger FASTQ developed by Wellcome Trust Sanger Institute is based on traditional sequence format, combining sequence and its corresponding quality score. In FASTQ format file, it generally uses four lines to demonstrate per sequence. The first line starts with character ‘@’, which is followed by the sequence identifier and optional sequence description. The second line is the sequence letter. The third line begins with character “+”, which is followed by discretionary sequence identifier the same as line 1 and any sequence description. The last line reveals the quality score of each corresponding base of the sequence in line 2. Each quality score represents the accuracy rate of each corresponding base, the larger the ASCII of the quality score the higher accuracy rate. The relationship of quality score and accuracy rate (or error rate) for each FASTQ format is represented in Figure 1-5.



a

$$Q_{sanger} = -10log_{10}p$$

$$Q_{solexa} = -10log_{10}\frac{p}{1-p}$$

$$Q_{Illumina\,1.3+} = 10 \times log_{10}(10^{Q_{solexa}/10} + 1)$$

$$Q_{Illumina\,1.5+} = 10 \times log_{10}(10^{Q_{Illumina\,1.3+}/10} - 1)$$

b

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.
..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghi
|                         |    |        |                              |
33                        59   64       73                             104
0.........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
```

S - Sanger $C_{ASCII} = Q_{sanger}$+33, typically (0, 40)

X - Solexa $C_{ASCII} = Q_{solexa}$+64, typically (-5, 40)

I - Illumina 1.3+ $C_{ASCII} = Q_{Illumina\,1.3+}$+64, typically (0, 40)

J - Illumina 1.5+ $C_{ASCII} = Q_{Illumina\,1.5+}$+64, typically (3, 40)

Figure 1-5. Phred quality score in different FASTQ formats. (a) Relation between Phred Quality Score (Q) and Probability of incorrect ball call (p) in Wellcome Trust Sanger and Illumina, where p means Probability of incorrect ball, Q_{sanger} means Phred Quality Score used by Wellcome Trust Sanger, Q_{solexa}, $Q_{Illumina\,1.3+}$, and $Q_{Illumina\,1.5+}$ mean Phred Quality Score used by Illumina. (b) Relation between Phred Quality Score (Q) and corresponding ASCII character (C_{ASCII}) in different FASTQ formats.

1.3.3.2 Read Trimming and Quality Control

The raw sequencing data files directly from the sequencer not only include the DNA sequences, but may also encompass adaptors unattached to the target sequences. All the sequences are called reads. Before analyzing the DNA sequences, it is necessary to remove the specific tag fragments and carry out the process of quality control.

Since adaptors linked to target DNA fragments will be collected while the size of target DNA fragments are smaller than that of defined read cycles, it is still necessary to get rid of these adaptors and their extending sequences prior to further analysis. Moreover, as the parallel sequencing method, NGS can sequence multiple samples at the same time. Thus, it will add disparate tag sequences to diverse samples for target sample sequence preparation. Furthermore, if it has to differentiate the direction of the DNA strand, two ends of the DNA fragment will be attached with different tag fragments. All of the tags will be removed after their assistance of dividing the reads into different groups.

After removing the specific tag sequences, assessing the quality of each read to determine whether it needs further analysis is also essential. As for different experimental purposes and objects, the requirements for quality control are not exactly the same. On one hand, with the given accuracy and efficiency of the experiment, it seems that the more rigid the requirement is, the better the result will be. On the other hand, if the requirement is too rigid, it will dramatically reduce the quantity of the data for further analysis, thus will lose useful information. Therefore, how to realize the maximum removal of the unqualified data as well as to achieve an adequate volume of data are important challenges for bioinformatics. There are two major methods for quality control. One is to, by directly assessing the base quality and content distribution, remove the whole non-confirmed reads or undesirable sequences attached to both ends of the reads. This method has been embedded in various quality control tools, such as FastQC [28], SolexaQA [29], NGS QC Toolkit [30], and so on. The other does the error correction for possible mismatched bases by sequence alignment. Software such as Reptile [31], ECHO [32], and RACER [33] are based on this method. Nowadays, a large number of software for data analysis exists, and they also provide a variety of parameters to satisfy the requirements of different researchers. With the wide spread of the application of NGS, however, new needs for quality control are emerging, which makes the development of the newer software more and more necessary.

1.3.3.3 Sequence Analysis

Reads that pass the quality control can be used for further analysis. In accordance with the different experimental purposes, there are diverse analytical methods. However, mapping and assembling are the basic steps for any analytical method.

Except for the result of *de novo* sequencing, the sequences of other sequencing methods all have their reference sequences, of which the most commonly used one is chromosome sequence. The aim of mapping is to map the reads to their reference sequences through sequence alignment. Sequence alignment is an important branch of bioinformatics, which encompasses not only DNA sequence alignment, but also protein-protein alignment, protein-nucleotide alignment, and so on. At the beginning of the first-generation sequencing, the first computer algorithm for sequence analysis had been published [34]. As for the research of sequence alignment, the most popular tool is BLAST [35], which is a pair-wise local alignment. Like BLAST, BLAT developed by

UCSC is a similar tool [36]. Since such alignment methods are not developed specifically for NGS, they face a big challenge on alignment efficiency and time-consuming if one tries to process an enormous number of reads. Focusing on large quantity and short reads of the output of NGS, new sequence alignment tools, such as Bowtie [37] and BWA [38], were developed. Bowtie can make the alignment at the rate of over 25 million per hour when mapping 25-bp reads to reference sequences. Based on Bowtie, Langmead *et al*. developed Bowtie2 with more functions. Compared with Bowtie, Bowtie2 starts to support gapped and local alignment modes [39].

In order to sequence a long sequence, it needs to break up the original sequence into short fragments. Thus, after obtaining the sequenced reads, sequence assembly is used to reconstruct the original sequence. Sequence assembly, in fact, is based on sequencing alignment. It joins the fragments together by overlaps of different reads, using pairwise alignment or multiple alignments. For reads with reference sequences, sequences can be assembled by mapping the reads to their references. Just like sequence mapping, sequence assembly meet similar problems, such as high-throughput data and short fragments. Nowadays, a number of sequence assembly tools specific for NGS are available, of which PGA [40] is a more widely used tool.

To meet different experimental purposes, there are many other diverse bioinformatics tools that are based on sequence mapping and sequence assembly. For example, Cufflink, an RNA-Seq analytical tool for transcriptome profiling [12], and MACS, a CHIP-Seq analytical tool which aims to discover the DNA-protein interactions [41], and SNPtools, a SNP calling tool which is used to identify sequence differences

[42]. It is believed that, with the increasingly wide application of NGS, more efficient bioinformatics tools will be developed.

1.3.4 Future DNA Sequencing

During the time when the NGS technologies were booming and widely used, a bunch of new DNA sequencing technologies emerged and constantly developed, such as Heliscope True Single Molecule Sequencing (tSMS) of Helicos Biosciences, Single Molecule Real Time Sequencing (SMRT), and Nanopore Sequencing of Oxford Technology. The accuracy and cost of these new DNA sequencing technologies are not better than those of the improved NGS. However, compared with NGS, the new DNA sequencing technologies skipped PCR amplification, the step before sequencing, thus eliminating the error which could have been introduced in this process. Therefore, the new technology was named Third-generation Sequencing or Next-next-generation sequencing.

In fact, the development of DNA sequencing technology involves not only the development of biology, but also the development and integration of physics, chemistry and engineering. Furthermore, with the assistance of these fields, of the DNA sequencing technology, the cost will be lower, the development will get faster, and the operation will be easier. Certainly, DNA sequencing in the future will not only facilitate more research areas, but also go to the market and become an important medical aid for personalized medicine.

1.4 Microarray

1.4.1 Background

In 1983, Tse Wen Chang first illustrated the concept of microarray in the study of using antibody matrix to determine specific cell surface antigens [43]. After that, Dr. Roger Ekin and colleagues made a great contribution to developing the theoretical background for ligand-binding assays based on protein microarray [44]. Twelve years later, in 1995, Patrick Brown's laboratory at Stanford University published the first article in Science describing the use of miniaturized microarrays for gene expression analysis [45]. Microarray technology is developed as a high-throughput technology for parallel analysis of multiple molecular targets on a miniaturized surface [46]. Simply defined, microarray is a collection of microscopic molecular features commonly known as probes orderly arranged on a planar solid substrate, which is usually made of glass or silicon. During the past years, the technology has advanced rapidly. For example, primarily due to miniaturization of the spots, the number of probes immobilized per cm^2 of solid surface has increased from less than 100 in 1995 to millions today [47]. And microfluidic technology has been applied to microarray manufacturing, which makes reactions in microarrays controllable spatially and temporally (Figure 1-6).

Based on the kind of the probes immobilized on the support substrate, microarrays can be categorized into DNA microarrays, protein microarrays, peptide microarrays, antibody microarrays, tissue microarrays, cellular microarrays, chemical compound microarrays, and so on. For purposes of this thesis, I focus on the introduction of DNA microarray (DNA Array) and peptide microarray (PepArray).

A kind of DNA Array, also known as oligonucleotide microarray or gene chip, is the most developed and the most widely used type of microarray, for example, gene-expression microarray, which is based on Watson-Crick complementary base pairing (i.e., guanine pairs with cytosine and adenine pairs with thymine or uracil), assays the gene expression level by using oligonucleotide probes to pair with mRNAs in the sample [48]. In addition, DNA Array technology also plays an essential role in various fields, such as biomarker determination, discovery of correlation between gene expression and diseases, and drug discovery.

It is known that peptides retain partial functions of proteins, and are much more stable than proteins immobilized on support substrate. Due to the insufficiency of DNA Array in proteomic research, as well as the instability of protein structure and activity, it is not surprising that PepArray plays a vital role in studying proteomics [49]. PepArray, also known as peptide chip, is a kind of device which is used for high-throughput analysis of protein samples via peptide probes orderly immobilized on the solid substrate. Over the last two decades, PepArray has been increasingly used as a high-throughput tool in diverse research fields, such as epitope mapping, drug discovery, biomarker discovery, disease diagnosis, and so on. One important application of pepArray is that the pepArray directly measures variations in levels of proteins and thus allows a more direct association of the array measurements to protein signaling pathway network activities..

As microarray technology has so many advantages, several large microarray corporates, such as Affymetrix, Agilent, emerged within only 20 years. The booming companies, coupled with the requirements of various research fields, also promoted the popularization of microarray technology.

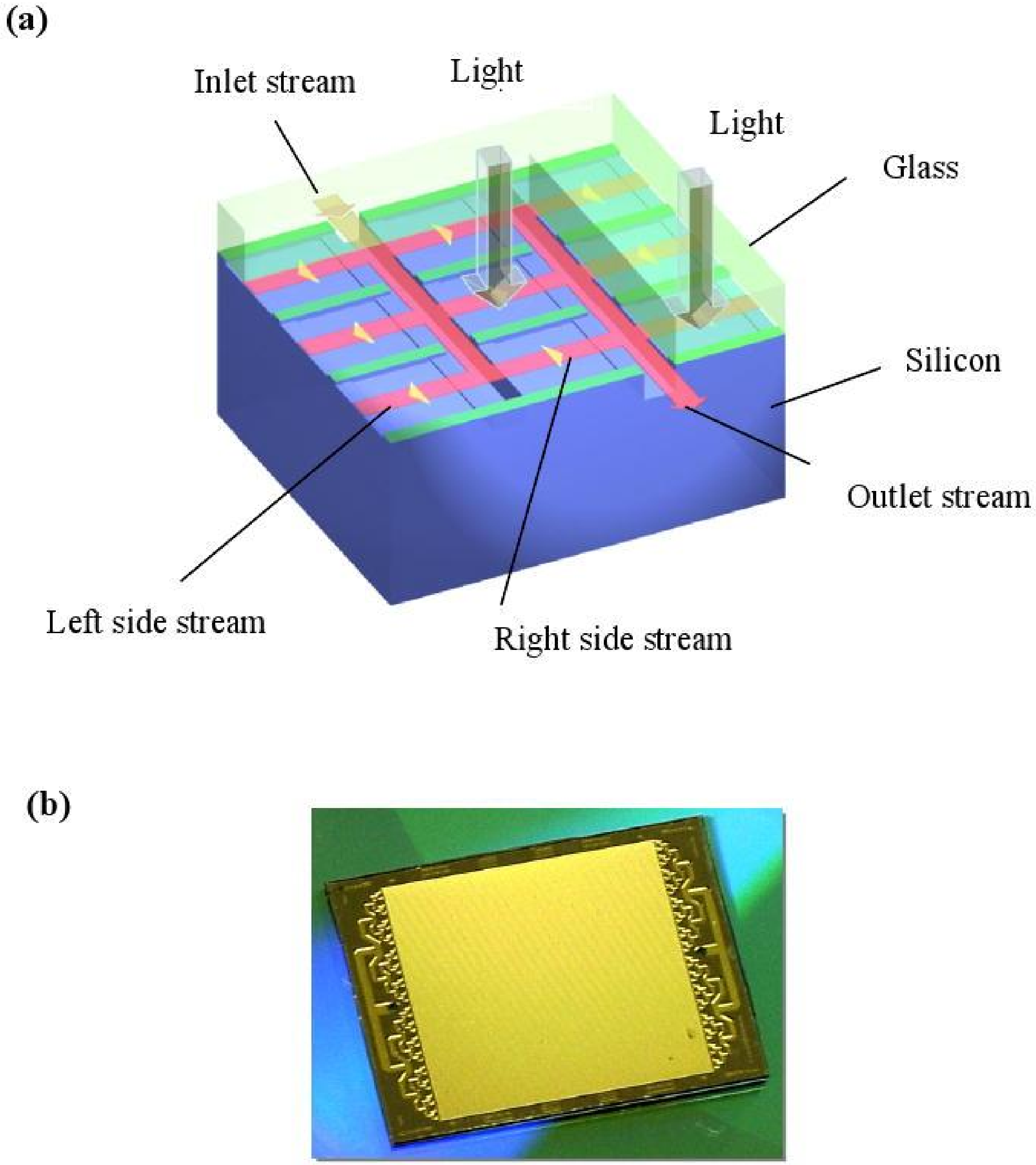


Figure 1-6. Microfluidic reactor array device. (a) Microfluidic technology has been applied to the structure of microarray. (b) Physical picture of microfluidic microarray. Reprinted from [50], Copyright 2009, with permission from Elsevier.

1.4.2 Manufacturing of Microarrays

Manufacturing of microarray plays an essential role in producing adequate high-quality microarrays, and it is a technology-demanding process. Microarray fabrication requires the expertise of biology, bioinformatics, physics, chemistry and engineering.

Fabricating microarray can be divided into three principal sections, including solid substrate selection, probe preparation and probe immobilization. In general, glass and silicon are usually adopted as substances for solid substrate. Glass is selected not only because it is economically affordable and readily accessible, but also due to its properties such as inertness, excellent flatness, as well as low fluorescence [51]. Glass surface offers silicon dioxide bonds which can react with nucleophile groups of nucleic acids or proteins to form covalent bond and immobilize these molecules on surface.

In microarrays, probes orderly deposited on the planar solid substrate are used to monitor or detect the target sequences in samples of interest. Therefore, it becomes crucial to winnow out the probes which are not only specific for their targets, but also react with their targets under similar condition.

It is well known that instable, unspecific immobilization of probes to supporting substrate will result in weak binding accuracy and thus produce inferior microarray experimental results. Therefore, the process of probe immobilization plays a central role in manufacturing of microarrays [52]. The probes can be deposited onto a supporting surface by disparate methods, such as spotted method vs *in situ* synthesis. As for spotting a method, it deposits pre-synthesized probes in predefined positions on a suitable surface with the help of robotic instruments. On the one hand, this method has some advantages, such as flexibility in array design due to its specific synthesis mechanism, high-quality synthesis, as well as low cost. On the other hand, the prerequisite of pre-synthesized probes makes it inflexible for array design [53]. Unlike spotted method, *in situ* synthesis can directly synthesize numerous probes in accurate positions on a suitable supporting surface. As for *in situ* synthesis method, appropriate chemical groups used for further synthesis will be capped on an array surface, then a variety of probes will be simultaneously synthesized in different locations via electrochemical technology or photolithography technology, which lets light go through physical masks or dynamic micromirror devices for producing specific probes with photolabile nucleoside monomers or photolabile amino acid monomers (Figure 1-7). *In situ* synthesis based on combinatorial chemistry is a major advancement for microarray development, which makes it possible to achieve microarrays containing a very large number of probes in short time and with low cost. Such microarray manufacturing process has distinct advantages, such as standardized automatic processing, array miniaturization, high probe quality, flexible synthesis, and so on. Compared with spotted method, *in situ* synthesis results in reduced flexibility owing to the hybridization and detection equipment. However, due to its parallel synthesis, *in situ* synthesis is efficient and economic. Therefore, *in situ* synthesis microarrays have also been used to provide DNA materials for *de novo* synthesis of long DNA sequences or genes for protein engineering of synthetic biology applications [54].

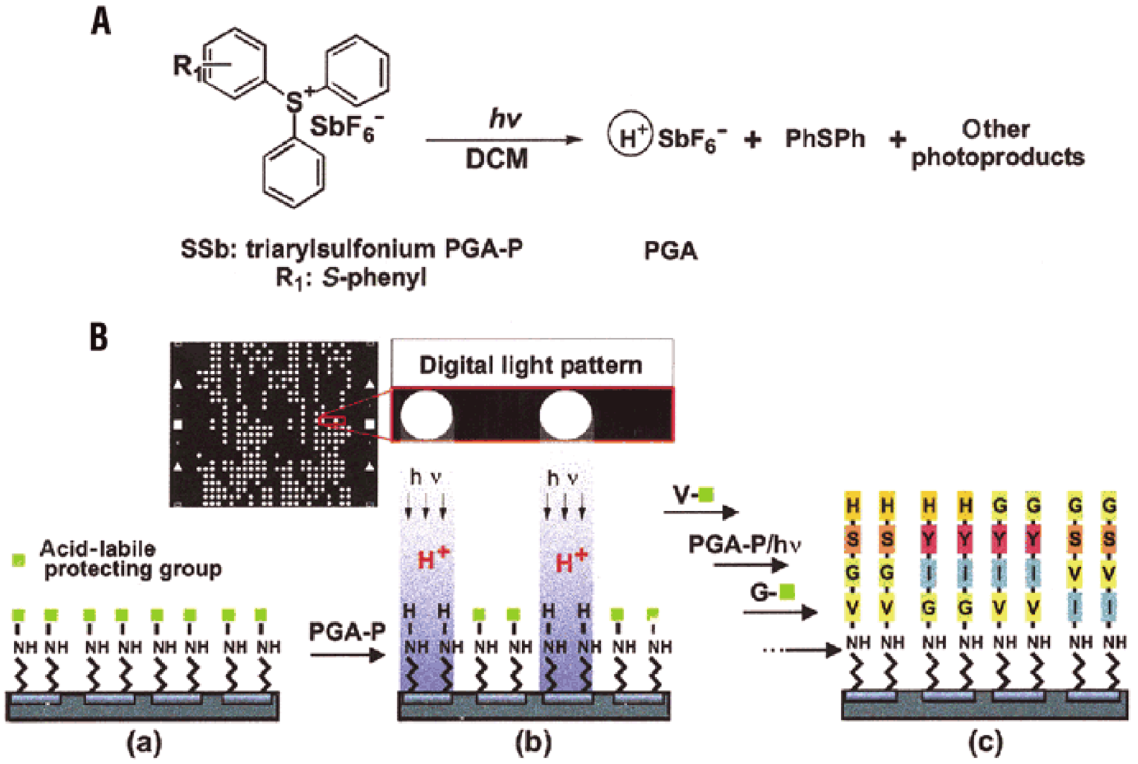

A
R1
S+
SbF6-
hv
DCM
H+ SbF6- + PhSPh + Other photoproducts
SSb: triarylsulfonium PGA-P
R1: S-phenyl
PGA
B
Digital light pattern
Acid-labile protecting group
NH
PGA-P
h v
H+
H
V-
PGA-P/hv
G-
(a)
(b)
(c)

Figure 1-7. *In situ* synthesis with photogenerated acid (PGA) reaction. (A) PGA is formed from deprotection in light irradiation. (B) In *in situ* peptide parallel synthesis, PGA deprotection is used as the gating step to synthesize peptide by Boc-protected amino acid. The reactions are under the control of predetermined digital light patterns. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [55], copyright 2002.

1.4.3 Bioinformatics of Microarray

With the improvement of microarray technology, over two million oligonucleotide probes can be deposited in only one microarray [56]. Therefore, microarray technologies make it possible for scientists to conduct experiments at the omic level, such as genome-wide expression profiling and proteomic profiling [57,58]. Thus, enormous amount of microarray data are continuously generated, which promote the rapid development of bioinformatics in this field.

Bioinformatics plays a vital role in analyzing the information embedded in a large amount of microarray data, thus interpreting the biological meanings. Over the past two decades, numerous bioinformatics tools supporting different microarray platforms have been developed, for example, microarray software suite TM4 [59]. However, the corresponding bioinformatics solutions to certain novel questions proposed by the newest microarray technologies are developing or remain to be found. In this thesis, I will take μParaflo microfluidic array as an example to illustrate how to analyze the microarray data with the assistance of bioinformatics.

1.4.3.1 Probe Design

As discussed above, microarrays would require probe design based on the target of interest. For example, if the goal is to study activities of HMTs in the nucleus of cells with peptide microarrays, it is important to make sure the wild type (WT) probes coming from histone proteins contain possible methylation sites, such as lysine (K) or arginine (R).

Diverse probes should be selected or designed for different experimental purposes. For example, in order to capture proteins with a specific function domain using PepArray, it is necessary to select the peptides, known as test probes, which can specifically bind to the target proteins under the defined experimental environment. In addition, to ensure high-quality experiments and raise the accuracy rate of data analysis, it is indispensable to design negative control probes corresponding to each test probe and quality control probes for array quality analysis, as well as enough replicated probes which are applied to reduce the effect of marginal data. Otherwise, it is impossible to achieve the goal.

1.4.3.2 Image digitization

Like direct output of NGS, the original data from optical detector are recorded in image files. Since they are hard to store and analyze, it is necessary to convert them to digital files prior to further analysis. Unlike those from NGS, digital files of microarrays do not include sequences, rather coordinates and the corresponding signals. In general, a microarray is composed of a set of reactive sites uniformly distributed on a planar surface, and each site is recorded by row and column. Probes of a site may unequally interact with their target, which results in that the signal of an individual site cannot exactly reveal the extent of reaction. To solve such problem, μParaflo takes over the existing functions of Array-Pro Analyzer which is microarray analysis software. For instance, the software will exploit parameters, including mean, median, and standard deviation values, calculated based on all the pixels of a suitable area, to demonstrate the quantity and quality of each site.

1.4.3.3 Background Subtraction

For microarrays, fluorescent signal of each probe includes not only the binding signal, but also background signal. For example, both non-specific binding and autofluoresence of probe and supporting surface contribute much to background signal. In order to winnow out the binding signal for further analysis, the background signal should be subtracted. Here, we take μParaflo as an example to introduce a method to remove the background signal. Due to obeying normal distribution, the characteristic of background can be extrapolated from the whole microarray signals. First, it generates a ranked raw signal profile graph, indexed from 1 to total probe as x axis, and sorts signal from low to high as y axis. Then the estimated average of background signal can be sought out at the first inflection.

1.4.3.4 Normalization

Normalization aims to make the probes within the same microarray or of different microarrays comparable by adjusting effects deriving from the technological variations rather than from biological differences [60]. The goal of a large number of microarray studies is to seek patterns under different environments by comparing probes in diverse samples. However, such factors as different ambient environments when experiments were processed print quality and disparate instruments used for synthesizing or scanning may result in variations between microarrays. Therefore, before conducting appropriate biological comparison, a great number of effects including the factors mentioned above must be eliminated.

With the development of microarrays for the past two decades, many normalization methods, such as total intensity normalization, log centering, and rank-invariant methods [61], have been created. Unlike those mentioned above, there is another widely used method, known as locally weighted scatterplot smoothing (lowess), which takes account of systematic biases appearing in most microarray experiments. Here, we provide a brief introduction of the main concept of lowess with an example normalizing two microarrays, M and N. First, the $log_2(M_i/N_i)$ ratio is plotted as a function of $log_{10}(M_i * N_i)$ product, where i denotes the index of each probe on the microarray. From the ratio-product plot, systematic deviations are detected, thus lowess corrects them via implementing a local weighted linear regression as a function of $log_{10}(M_i * N_i)$, as well as deducting the best-fit average $log_2(M_i/N_i)$ from the obtained ratio for each probe. Upon the correction, lowess adopts a weight function, in which the contributions of the probes far from the others will be unvalued. After the appropriate correction, two microarrays will reach the comparable level.

In addition, since following normal distribution is the prerequisite of most widespread normalization methods, it is necessary to develop other approaches for microarray data which does not obey that distribution. One option is to seek out various groups of specific probes which can be considered as background of different microarray experimental data which is not normally distributed. Conceivably, with novel problems related to normalization coming up, the current approaches will be generally improved, and original methods will be continually developed by diligent researchers.

1.4.3.5 In-depth Data Analysis

Most of the microarray experiments aim to seek out patterns by directly assaying the biological binding levels of tens of thousands of probes on comparable microarrays. Thus, the identified patterns play vital roles in a lot of aspects, such as functional annotation of novel probes based on known probes within the same pattern, diseases diagnostics based on the patterns between one or multiple pairs of comparable samples, biomarker discovery, and so on. In the early time of microarray technology, a cut-off method, for example, two-fold changes, was adopted to identify significant patterns. Following that, such parameters as mean, median, standard deviation, logarithm value, and others are included in slightly complex methods. Up to now, many sophisticated statistics approaches have been developed, such as ANOVA used to analyze differences between two or more samples developed by R.A. Fisher [62], hierarchical clustering which seek to create a hierarchy of clusters of probes or different samples, relevance networks useful in obtaining similarity of probes by comparing comprehensive pairwise features, principal-components analysis which can be used to find much coherent variables by omitting the less significant features, support vector machines which analyze data and identify patterns with learning algorithms in order to conduct classification and regression analysis, and so on. Furthermore, a large number of multifunctional tools for microarray analysis have been created, which definitely facilitate the research in many fields. For example, Multiexperiment Viewer (Mev) developed by the John Quackenbush laboratory is widely used versatile microarray data analysis software [63] which incorporates quite a few statistic methods, such as Hierarchical cluster (HCL) [64], K-means clustering (KMC) [65], relevance networks (RN) [66], template matching (PTM)

[67], significance analysis of microarrays (SAM) [68], one-way analysis of variance (ANOVA), T-Test and SVM [69].

Indeed a great number of microarray analytical tools are accessible today, however, a number of problems still exist. Given more or less minor deficiency of most existing microarray analytical approaches, as well as emerging challenges, researchers in this field tempt to fix them via modifying the available methods, and developing novel analytical functions and software, respectively. It is conceivable that the development of microarray technology will get more and more perfect, and in turn it will give more help to investigators in the field.

Chapter 2

miRFocus: Open source of Web-tool

for Human miRNA Annotation,

Target Gene and Pathway Analysis

2.1 Introduction

MiRNAs are a class of small single-stranded non-coding RNAs of approximately 22 nucleotides in length, which are excised from hairpin-shaped pre-miRNAs [70]. It is becoming clear that they play essential regulatory roles in diverse organisms via imperfect binding with target mRNAs, thus resulting in cleavage or translational regression (Figure 2-1) [71,72]. In 1993, the first miRNA from *C. elegans* was described in an article published by Lee, *et al* [5]. However, it was not until 2001 that their existence in vertebrates, as well as their important impact on gene regulation, was recognized, and the term miRNA was introduced [73,74].

With the help of computer sciences, researchers found that almost 60% of human genes are targeted by miRNAs, and each miRNA is capable to regulate numerous target genes, likewise every target gene is probably controlled by multiple miRNAs [75]. Through complicated regulations, miRNAs take an important part in many key biological processes, including cell development, tissue differentiation, cell proliferation and apoptosis. Therefore, either dysfunction of miRNAs or their inaccurate regulatory pathways will probably lead to diverse diseases, such as diabetes [76], obesity [77], cardiovascular disease [78], renal function disorders [79], as well as different types of cancers. For example, Nassirpour and colleagues illustrated that overexpressed miR-221 is linked to triple-negative breast cancer, and that tumor growth can be inhibited by knocking down miR-221 [80].

Due to the importance of the miRNAs in various aspects, studies on miRNAs have become a hotspot. It is worth mentioning that the number of publications as to miRNAs increases dramatically these years. To date, almost 14 years have passed since the term of miRNA was coined in 2001. Searching with key word either “miRNA” or “microRNA” in PubMed, the number of achieved articles has increased up to 34,310 in the past 7 years compared with mere 2,403 in the previous 7 years, of which 23,157 is relevant to humans (Figure 2-2a). In addition, the quantity of published human pre-miRNAs and mature miRNAs in miRBase has reached to 1,881 and 2,588 in 2014 (version 21) from 56 and 44 in 2003 (version 1), respectively (Figure 2-2b). MiRBase is a database which provides published miRNAs and their annotations. It is evident that high-throughput technologies, such as microarray and NGS, play a vital role in paving the way for diverse research in this field [81].

With rapid development of miRNA studies, enormous jumbled information is generated in many aspects. Therefore, how to systematically organize the unorganized information becomes an urgent problem. To satisfy the requirement, a number of databases concerning miRNAs are developed. Based on main functions, databases can be briefly divided into several categories, such as annotation, target-miRNA, miRNA-disease, pathway-miRNA, and so forth. MiRecords is a reservoir of miRNAs and target genes, which includes experimentally validated and predicted miRNA-target interactions [82]. Besides miRecords, miRTarBase [83] and TarBase [84] both focus on validated miRNA-target interactions collection. Moreover, as for target prediction web tools, TargetScan and miRanda adopt algorithms based on seed pairing [85]; PicTar predicts miRNA-target interactions on the basis of binding probability [86]; RNA22 determines possible miRNA-target in line with binding site similarity; and PITA identifies the interaction between miRNAs and target genes by considering not only the region mapping, but also the free energy used for structure construction [87]. Compared with resources for annotation and miRNA-target, there are only a handful of databases offering approaches for miRNA-pathways, such as DIANA-miRPath [88] and miRSystem [89]. In addition, miR2Disease is a manually curated database focusing on miRNA regulation in various diseases [90].

It is of no doubt that researchers in this field have received great help from available databases. For example, they can retrieve miRNA sequences from miRBase [91], diseases affected by miRNAs from miR2Disease, and predict observe target genes from miRecords or TargetScan. However, it is quite time-consuming to collect as much information as possible about one interesting miRNA. One of the reasons is that it takes significant time to identify the related databases. Although the databases are accessible, there are still several problems. For example, because diverse programs may require different versions of miRNA names which include 21 nomenclature versions in total, users have to spend time on name conversation. In addition, most databases allow users to query only one item each time, which is a big limitation when searching many miRNAs. In the face of these challenges, miRFocus was developed.

MiRFocus is an integrated resource and web tool focusing on human miRNA. Compared with existing databases and software, miRFocus incorporates more comprehensive information, including miRNA sequences and genomic clusters, co-expression miRNAs, experimentally validated and predicted target genes, miRNA-disease regulation, enrichment pathway analysis, as well as miRNA annotation for all Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. In particular, in the view of current problems mentioned above, miRFocus not only provides an approach to convert miRNA IDs from different versions into the latest version automatically, but also organizes a wealth of information in a reasonable way. In addition, miRFocus coupled with a user-friendly web site can make and facilitate significant contributions in accessing the information already available for the sequence and biology information of human miRNAs, so that the implication and rich information embedded in the data sets of miRNAs measurements can become evident and valuable biological clues extracted. The web server is freely accessible at (http://mirfocus.org).

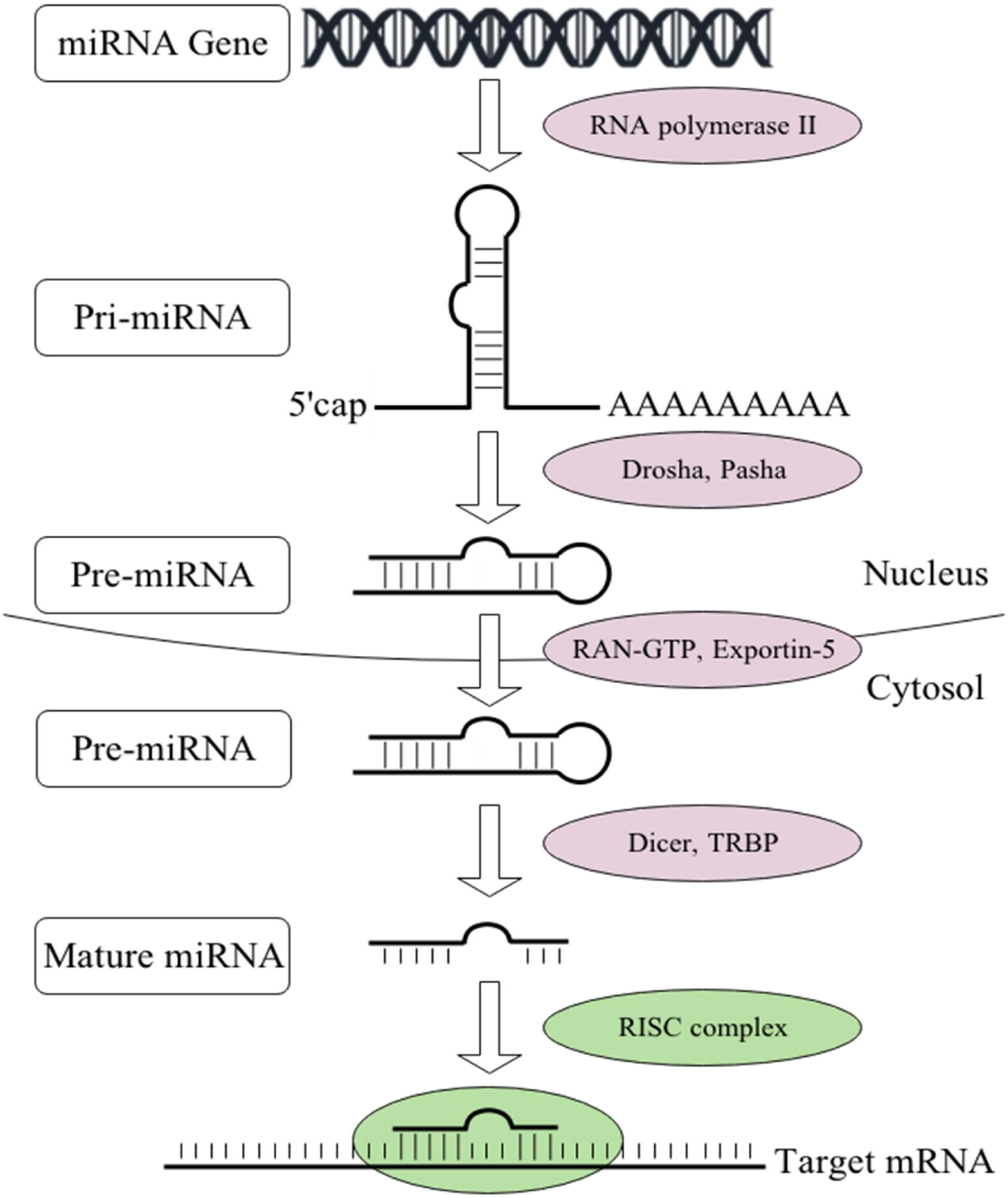


Figure 2-1. The biogenesis of miRNAs. In this figure, it shows the process of human miRNA formation, which starts from miRNA gene, through pri-miRNA, pre-miRNA, to mature miRNA. Subsequently, mature miRNAs can induce gene expression regression or mRNA cleavage by binding to the complementary site in target mRNA.

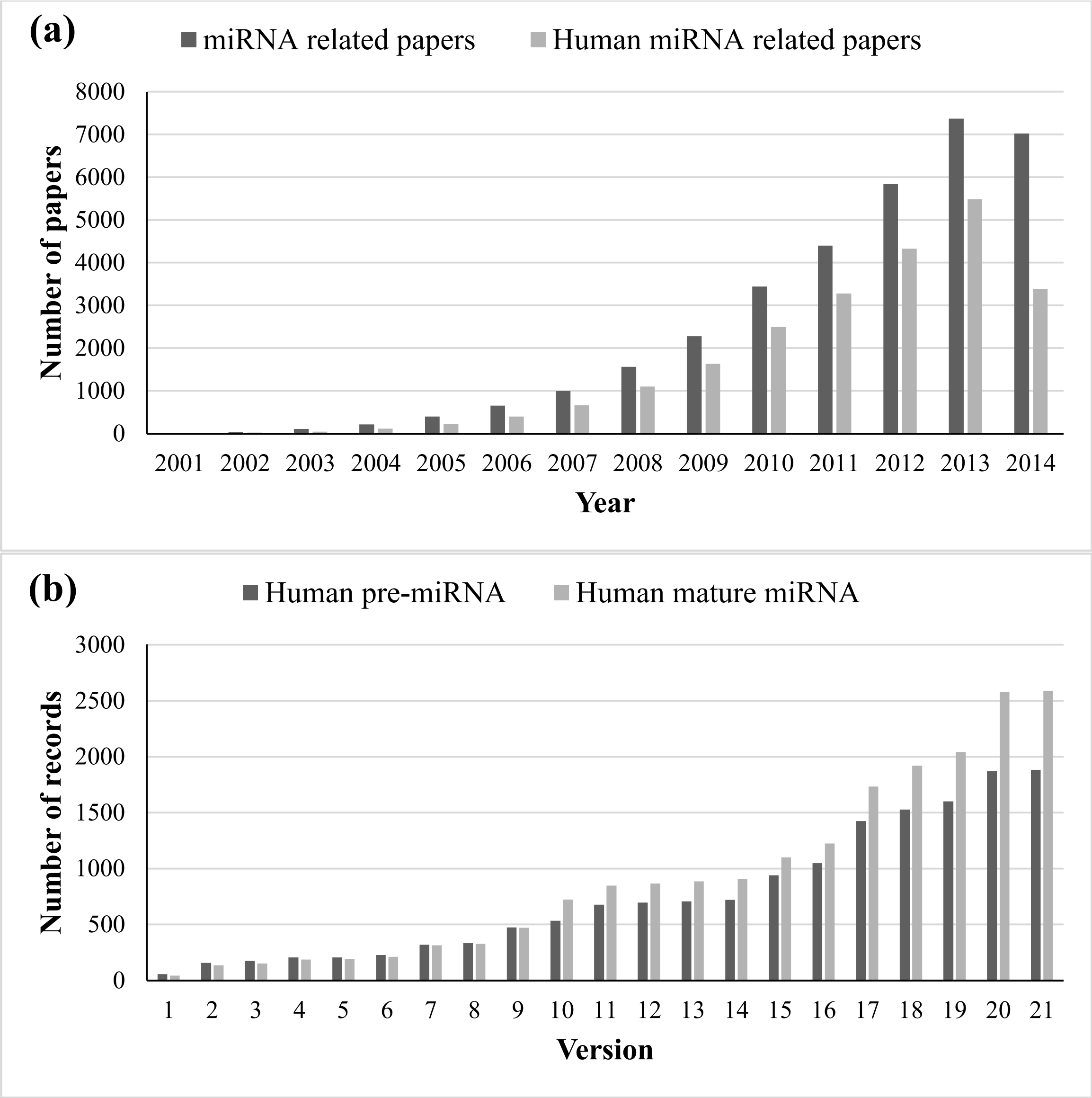


Figure 2-2. Histograms for human miRNA. (a) Published papers relevant to miRNA from 2001 to 2014 in PubMed. Data are collected from PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/); (b) Available pre-miRNAs and mature miRNAs from different versions of miRBase. Data are collected from miRBase database (http://mirbase.org/).

2.2 Methods

2.2.1 Overview

MiRFocus is a web-based resource including comprehensive information relevant to human miRNAs, such as sequences, target genes, diseases, as well as pathways. All data are saved in MySQL database, and the web interface is implemented with PHP5 and JavaScript language. Both of them are integrated in an Apache HTTP Server with all major browsers' support. MiRFocus is composed of three major modules: miRNA annotation, pathway analysis and miRpathway. An overview of miRFocus function is illustrated in Figure 2-3.

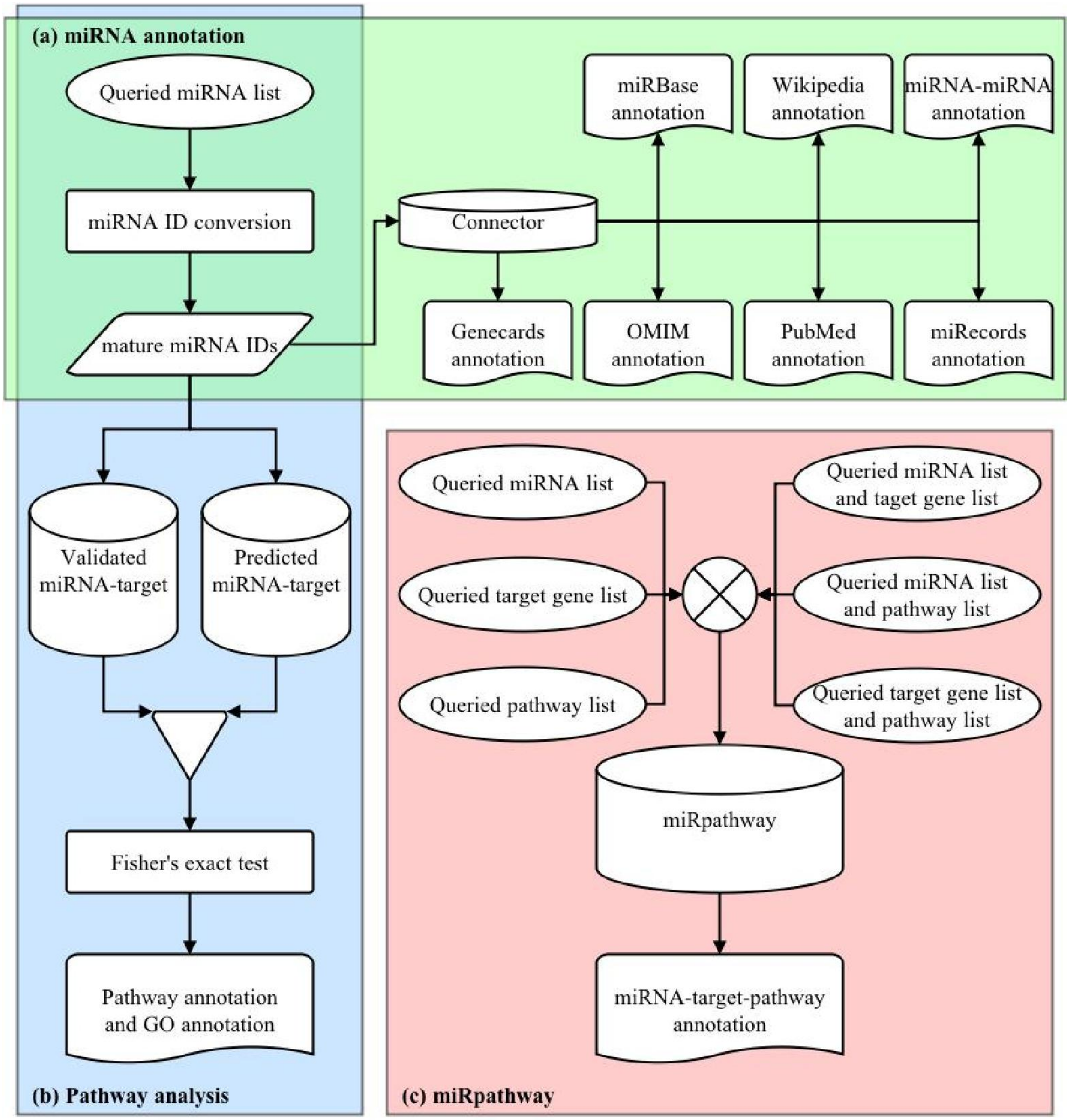
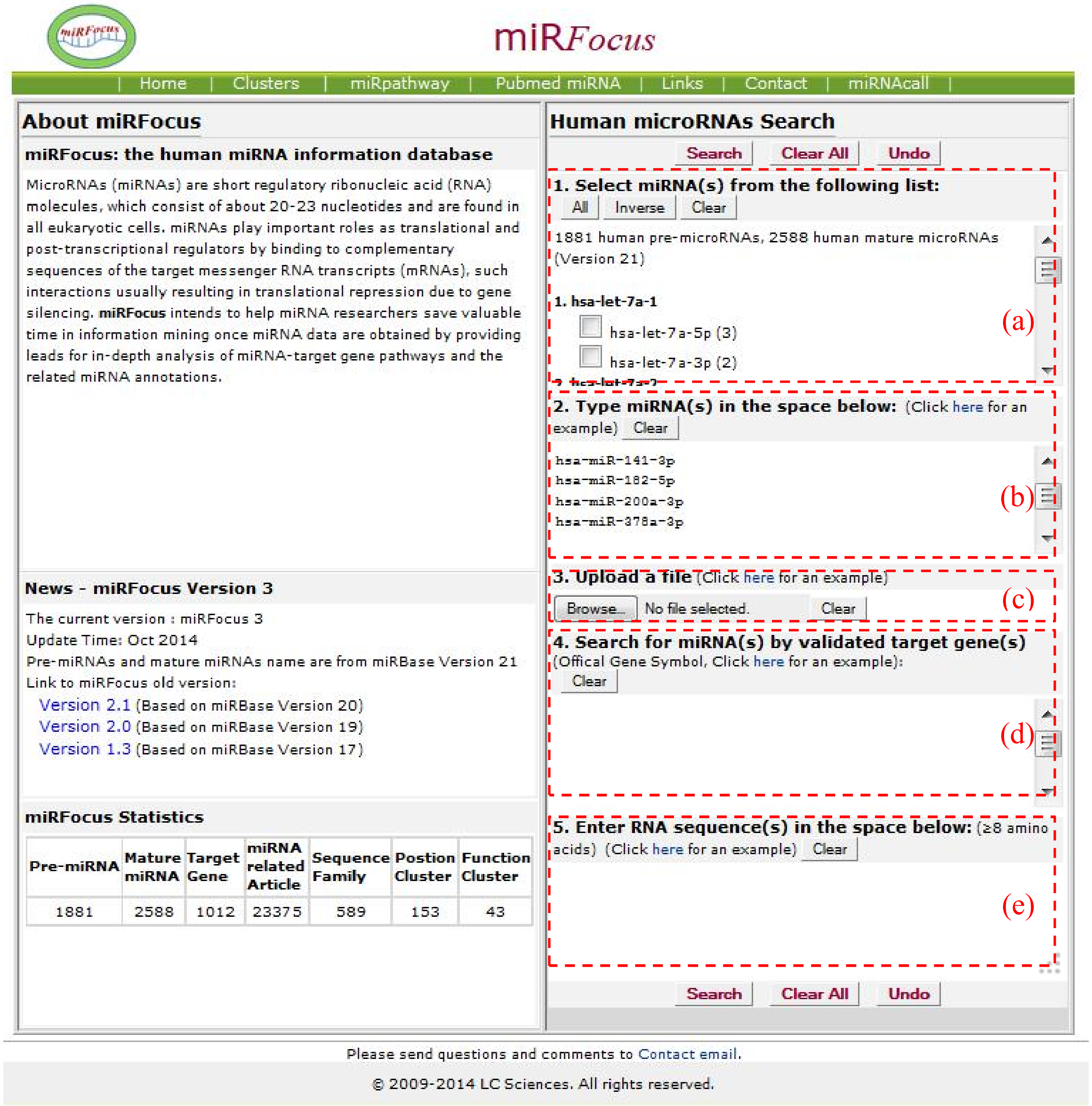


Figure 2-3. Major functions of miRFocus. (a) Green box denotes module “miRNA annotation”, which includes seven kinds of annotation of queried miRNAs. (b) Blue box denotes module “pathway analysis”, which aims at achieving the pathway annotation and GO annotation relevant to queried miRNAs. (c) Pink box denotes module “miRpathway”, which incorporates interrelations among miRNAs, target genes, and pathways.



About miRFocus

miRFocus: the human miRNA information database

MicroRNAs (miRNAs) are short regulatory ribonucleic acid (RNA) molecules, which consist of about 20-23 nucleotides and are found in all eukaryotic cells. miRNAs play important roles as translational and post-transcriptional regulators by binding to complementary sequences of the target messenger RNA transcripts (mRNAs), such interactions usually resulting in translational repression due to gene silencing. **miRFocus** intends to help miRNA researchers save valuable time in information mining once miRNA data are obtained by providing leads for in-depth analysis of miRNA-target gene pathways and the related miRNA annotations.

News - miRFocus Version 3

The current version : miRFocus 3
Update Time: Oct 2014
Pre-miRNAs and mature miRNAs name are from miRBase Version 21
Link to miRFocus old version:
Version 2.1 (Based on miRBase Version 20)
Version 2.0 (Based on miRBase Version 19)
Version 1.3 (Based on miRBase Version 17)

miRFocus Statistics

Pre-miRNA	Mature miRNA	Target Gene	miRNA related Article	Sequence Family	Postion Cluster	Function Cluster
1881	2588	1012	23375	589	153	43

Human microRNAs Search

1. Select miRNA(s) from the following list:

1881 human pre-microRNAs, 2588 human mature microRNAs (Version 21)

1. hsa-let-7a-1

□ hsa-let-7a-5p (3)

□ hsa-let-7a-3p (2)

2. Type miRNA(s) in the space below: (Click here for an example) Clear

```
hsa-miR-141-3p
hsa-miR-182-5p
hsa-miR-200a-3p
hsa-miR-378a-3p
```

3. Upload a file (Click here for an example)

4. Search for miRNA(s) by validated target gene(s) (Offical Gene Symbol, Click here for an example): Clear

5. Enter RNA sequence(s) in the space below: (≥8 amino acids) (Click here for an example) Clear

Please send questions and comments to Contact email.

© 2009-2014 LC Sciences. All rights reserved.

Figure 2-4. The home page of miRFocus. (a) Input by "Select miRNAs from the following list"; (b) Input by "Type miRNAs in the space below"; (c) Input by "Upload a file"; (d) Input by "Search for miRNAs by validated target genes"; (e) Input by "Enter RNA sequences in the space below".

2.2.2 miRNA Annotation

MiRNA annotation module is one of the most important sections of miRFocus. In this module, it displays in a well-organized way seven different sources of the queried miRNAs, including sequences and corresponding pre-miRNAs of miRNAs from miRBase, miRNA clusters from Wikipedia, miRNA-miRNA correlation in diseases manually curated from papers, annotation from GeneCards and OMIM, validated target genes from miRecords, literatures related to input miRNAs from PubMed, as well as pathway analysis.

In the home page, five different approaches are available to input miRNAs, which consist of “Select miRNAs from the following list”, “Type miRNAs in the space below”, “Upload a file”, “Search for miRNAs by validated target genes”, as well as “Enter RNA sequences in the space below” (Figure 2-4). Every method allows inputting of one or multiple miRNAs, which can really help users save time, compared with the methods permitting one item each time.

In particular, miRNA IDs used in different versions of miRBase releases will be converted into the latest version automatically through a miRNA ID conversion function. Mature miRNA IDs after miRNA ID conversion can be directly applied to pathway analysis, but which have to pass miRNA Gene ID conversion before achieving the information of miRNA annotation. The sources of miRNA annotation reported by miRFocus are shown in Table 2-1.

Table 2-1. Sources of miRNA annotation.

Source	Information
miRBase	miRNA/pre-miRNA basic information
Wikipedia	miRNA precursor family
miRNA-miRNA	miRNA-miRNA relationship
GeneCards	Gene aliases, genomic location and external IDs
OMIM	Genetic Disease
PubMed	Bibliography
miRecords	Validated target gene annotation

2.2.2.1 Construction of miRNA ID Conversion Function

MiRNA nomenclatures are gradually changing coupled with the development of miRNA studies, that is either because novel miRNAs are constantly discovered, or because names of some existing miRNAs may be removed or alternated (Figure 2-2b). It is common to see that the same miRNA appears in different databases or different literatures with distinct identities, which causes big troubles for researchers in this field. In order to solve this problem, miRNA ID conversion function is created, which connects miRNA identities from different versions by unique sequence accession number in miRBase. This function aims at converting miRNA IDs from all previous versions into the latest version (Version 21). Comparison of miRNA ID changes between version 21 and all previous versions is displayed in Table 2-2.

Table 2-2. ID tracking of miRNA ID in version 21.

State	miRNA ID count
Unchanged	1,886
Changed	689
New	13
Deleted	45

2.2.2.2 Construction of miRNA Gene ID Conversion Function

In module miRNA annotation, massive amount of information collected from numerous different sources has resulted in a big challenge: how to link them to queried miRNAs. Via miRNA IDs, some information can be readily extracted, but pre-miRNA annotation from GeneCards, disease annotation from OMIM and literatures from PubMed are hardly found.

To fix such problems, HUGO Gene Nomenclature Committee (HGNC) ID is utilized as the connector to link miRNA IDs with information from GeneCards, OMIM and PubMed, respectively. Corresponding HGNC IDs of input miRNAs can be found in miRBase, thus HGNC IDs are converted into official gene symbol used in GeneCards, OMIM ID in OMIM, and Entrez gene ID appearing in PubMed in accordance with the gene relation table downloaded from HGNC. Diverse IDs relevant to miRNA in miRFocus is shown in Table 2-3.

Table 2-3. Diverse IDs relevant to miRNA in miRFocus.

ID Type	Count
miRNA ID	2,588
Pre-miRNA ID	1,886
HGNC ID	1,861
Official Gene Symbol	1,860
Entrez Gene ID	1,860
OMIM ID	230

2.2.2.3 miRNA-Related Data Collection

The comprehensive information regarding human miRNAs is assembled from seven different sources, which will be introduced sequentially.

MiRBase annotation comes from miRBase (http://miRBase.org) [92], which consists of sequences and identities of mature miRNAs and their corresponding pre-miRNAs, as well as family clusters and position clusters of the pre-miRNAs. In particular, cluster can be used to define possible functions of novel miRNAs in line with known functions of miRNAs in the same cluster. In this part, it includes total 1,886 mature miRNAs, 2,588 pre-miRNAs, 589 sequence families and 153 position clusters.

MiRNA precursor families are collected from Wikipedia, which is a free online encyclopedia. This section contains total 43 miRNA precursor families and which cover 127 miRNAs (Version 2012).

Annotation of miRNA-miRNA correlation and diseases is curated manually from 693 publications, which includes 1,612 miRNA-miRNA correlations based on 533 miRNAs.

GeneCards is a well-known database which provides a wide range of information regarding human genes, including genomic and functional information [93]. Annotation from GeneCards encompasses a wide range of information indirectly relevant to miRNAs, which includes a variety of aliases, genomic location, and KEGG pathways. In this section, it contains total 1,860 miRNA genes which associate with 2,583 mature miRNAs.

Online Mendelian Inheritance in Man (OMIM) is comprehensive knowledge base of human genes and diseases [94]. In this section, it displays diverse useful information of each miRNA gene, including descriptions, functions, mapping results on chromosomes, as well as reliable references, which is collected from OMIM. This section covers total 230 miRNA genes which are relevant to 399 miRNAs.

In PubMed annotation section, references of the queried miRNAs are assembled from a variety of databases, including miRBase, Entrez Gene, Gene Reference into Function database (GeneRIF), Wikipedia, OMIM and miRNA-miRNA. Detailed information of references is supplied, such as title, author, journal, affiliation and abstract, and literatures relevant to queried miRNAs are displayed chronologically, from latest to earliest date. In addition, each title can be linked to PubMed, which facilitates users to find the full-text articles. Number of publications from different sources is shown in Table 2-4.

Experimentally validated target genes in the miRecords annotation section are from miRecords (Version 3). Besides target genes, miRecords annotation also provides GO and pathway annotation which company with each target gene. This section includes a total 1,448 interactions between 185 miRNAs and 1,045 regulated genes.

Table 2-4. Number of publications from different sources.

Source	Count
miRBase	110
Entrez Gene	5,005
GeneRIF	4,912
Wikipedia	74
miRNA-miRNA	693
OMIM	319
Total	5,321

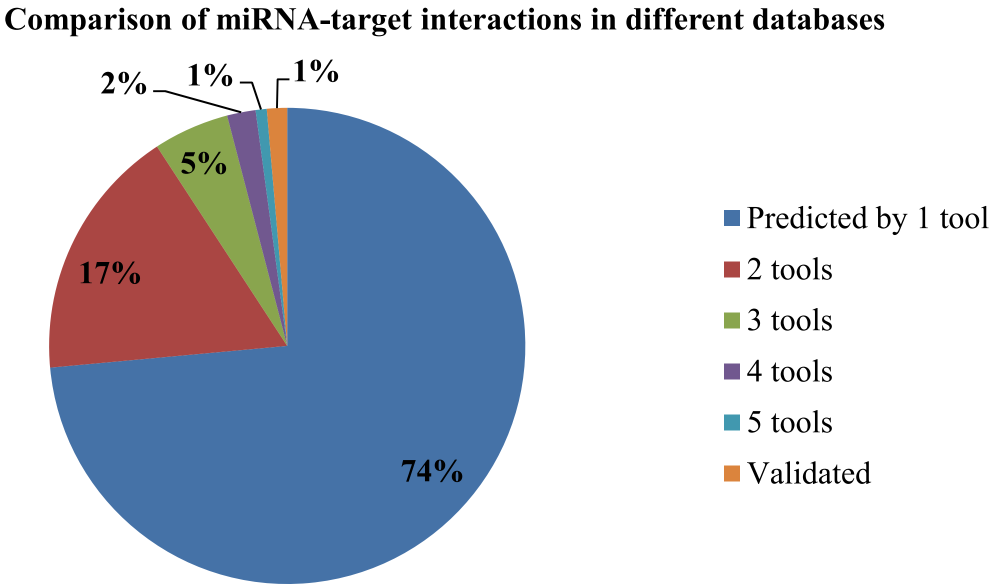


Figure 2-5. Distribution of miRNA-target interactions supported by different numbers of web tools.

2.2.3 Pathway Analysis

Pathway analysis is another essential part of miRFocus, which aims at achieving enriched KEGG pathways, BioCarta pathways and GO terms of queried miRNAs based on a statistical significance test – Fisher' exact test.

Since miRNAs function by base-pair binding with their target genes which are indispensable factors of some pathways, it is possible to further interpret miRNAs via their relevant pathways. Because a limited number of experimentally validated miRNA-target gene interactions are available, we incorporate a set of predicted miRNA-target gene interactions from popular target prediction databases.

MiRFocus includes both experimentally validated miRNA-target interactions collected from four databases: miR2Disease, miRecords, miRTarBase and TarBase, and predicted miRNA-target interactions which are based on five target prediction web tools including microT, MiRanda, MirTargets, PicTar and TargetScan. In order to integrate miRNA-target interactions of diverse databases, miRNA IDs of various versions are converted into corresponding IDs of the latest version via miRNA ID conversion function, and different gene IDs are substituted with Entrez Gene ID either through HGNC, biomart, or manually. A summary of miRNA-target interactions of different databases is shown in Table 2-5.

Because it is difficult to evaluate these target prediction databases due to limited experimentally validated miRNA-target interactions, we generate a pie graph (Figure 2-5) to represent miRNA-target interactions among five databases, which can be a guide for users to select databases according to their needs. From Figure 2-2 we can see that up to 74% miRNA-target interactions are predicted in only one database, therefore, it will be possible to construct pathway analysis on the basis of miRNA-target interactions predicted by at least three prediction databases and experimentally validated, which account for about 9% of the total interactions.

Table 2-5. Summary of miRNA-target interactions of different databases.

Type	Source	miRNA	Target gene	Interaction
Validated	miR2Disease	161	379	647
	miRecords v4.0	207	1,056	1,637
	miRTarBase v4.5	569	12,099	37,381
	Tarbase v5.0	88	852	1,025
Predicted	microT v3.0	553	17,446	1,441,631
	MiRanda	249	19,284	737,379
	MirTarget2 v4.0	1,911	16.663	691,805
	PicTar	1,142	12,816	361,961
	TargetScan v6.2	1,541	15,023	523,235
Total		1,975	20,713	2,755,481

2.2.4 miRpathway

MiRpathway module provides the useful correlations between miRNAs, target genes and KEGG pathways, which is established on basis of the experimentally validated miRNA-target interactions and pathway-gene relations from KEGG. A summary of components in miRpathway is shown in Table 2-6. In this module, it displays KEGG pathways in two levels of hierarchy. The first level consists of seven categories – Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Process, Organismal Systems, Human Diseases, and Drug Development, and each category includes numbers of subcategories. MiRNAs and target genes relevant to each pathway, subcategory and category are represented, respectively.

Table 2-6. Summary of components in miRpathway.

Elements	Count
miRNA	507
Target gene	4,608
KEGG pathway	281

In addition, this module also supplies useful functions for retrieving miRNAs, genes, or pathways on the basis of their correlation. For example, relevant pathways and genes will be achieved after inputting a miRNA list in “miRNA Query”. Compared with pathway analysis, in miRpathway, we only focus on KEGG pathways due to its greater popularity, and miRNA-target interactions are all experimentally validated in order to improve the credibility, and a statistic test is not applied in this module because of the small number of verified miRNA-target interactions. Furthermore, multiple functions in miRpathway facilitate users to obtain genes and pathways of queried miRNAs, or inputting specific miRNAs and pathways to achieve their correlated genes.

2.3 Results and Discussions

MiRFocus as a comprehensive human miRNA web-based resource provides a user-friendly interface. I will utilize an example to help users to become familiar with miRFocus.

It is published that miRNAs are able to access human blood with the help of exosomes in milk [95]. Recently, 602 unique miRNAs extracted from human breast milk exosomes using deep sequencing technology were published, and the top ten highest expressed miRNAs cover 62.3% of total counts (Table 2-7) [96]. Ten miRNAs will be selected as input data to introduce the input and result pages of miRFocus.

Table 2-7. Top ten highest expression miRNAs in breast milk exosomes.

Index	miRNA ID
1	hsa-miR-148a-3p
2	hsa-miR-30b-5p
3	hsa-let-7f-5p
4	hsa-miR-146b-5p
5	hsa-miR-29a-3p
6	hsa-let-7a-5p
7	hsa-miR-141-3p
8	hsa-miR-182-5p
9	hsa-miR-200a-3p
10	hsa-miR-378a-3p

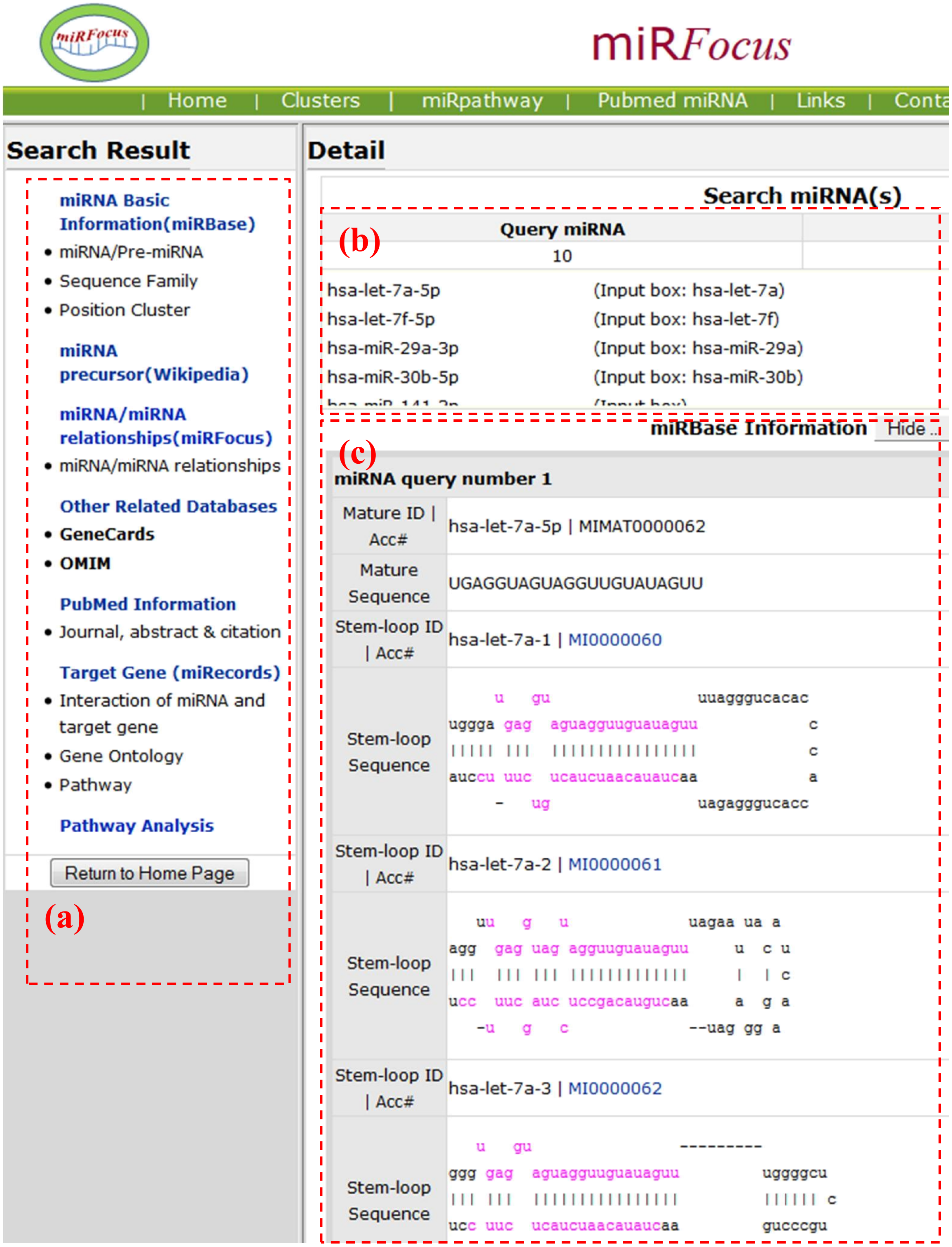
2.3.1 miRNA Annotation

In the home page, miRFocus provides five different methods to input miRNAs. I pasted ten miRNAs into input field (Figure 2-4 c), and clicked “Search” button, which processed to the result page.

The result page mainly consists of “Search Result” a navigation panel (Figure 2-6 a) and “Detail” which is used to display details relevant to input miRNAs. “Search Result” contains seven major sources, such as “miRNA Basic Information (miRBase)”, “miRNA precursor (Wikipedia)”, “miRNA-miRNA relationships (miRFocus)”, “Other Related Databases”, “PubMed Information”, “Target Gene (miRecords)” and “Pathway Analysis”.

In the navigation panel, each major source title colored with blue has the functions to either exhibit or hide corresponding details. For instance, the area of “miRBase Information” (Figure 2-6 c) will be concealed by clicking “miRNA Basic Information (miRBase)”, and the same area will be unveiled again if you click the same source title once again. At the top of “Detail”, queried miRNAs and their corresponding IDs of miRBase latest version are revealed (Figure 2-6 b). Details of “miRNAs Basic Information (miRBase)” are displayed in Figure 2-7, which supplies users with identities and sequences of mature miRNAs and pre-miRNAs, family clusters, position clusters, as well as references regarding input miRNAs which are collected from miRBase. Moreover, detailed information of both precursor functional families which facilitates to learn functions of miRNAs in the same family, and correlations between miRNAs and diseases is illustrated in Figure 2-8. In addition, Figure 2-9 demonstrates the a wide-range of information from GeneCards and OMIM, including diverse accession numbers of a variety databases which also link to source webpages, functions, genomic locations, and so forth. From Figure 2-10, it is readily to observe the output format for both references of miRNAs and validated target genes.

To display a particularly impressive range of information in one page is a characteristic of miRFocus, which helps users save lots of time on gleaning information from diverse databases. Meanwhile, users can really achieve comprehensive useful information of their interested miRNAs. For example, it can be seen from “PubMed Information” shown in Figure 2-10 (a) that up to 464 articles relevant to ten input miRNAs are listed, which means these miRNAs have been studied at certain levels; “miRNA-miRNA Information” in figure 2-8 (b) demonstrates that hsa-let-7a-5p, hsa-let-7f-5p, hsa-miR-29a-3p, hsa-miR-141-3p, hsa-miR-146b-5p, hsa-miR-182-5p, and hsa-miR-200a-3p correlate with human ovarian cancer, as well as hsa-miR-141-3p and hsa-miR-200a-3p are considered as an vital player in breast cancer.


miRFocus
| Home | Clusters | miRpathway | Pubmed miRNA | Links | Conta
(a)
(b)
(c)
Return to Home Page
Hide ..

miRFocus

Search Result

miRNA Basic Information(miRBase)

- miRNA/Pre-miRNA
- Sequence Family
- Position Cluster

miRNA precursor(Wikipedia)

miRNA/miRNA relationships(miRFocus)

- miRNA/miRNA relationships

Other Related Databases

- **GeneCards**
- **OMIM**

PubMed Information

- Journal, abstract & citation

Target Gene (miRecords)

- Interaction of miRNA and target gene
- Gene Ontology
- Pathway

Pathway Analysis

Detail

Search miRNA(s)

Query miRNA
10

hsa-let-7a-5p	(Input box: hsa-let-7a)
hsa-let-7f-5p	(Input box: hsa-let-7f)
hsa-miR-29a-3p	(Input box: hsa-miR-29a)
hsa-miR-30b-5p	(Input box: hsa-miR-30b)
hsa-miR-141-3p	(Input box)

miRBase Information

miRNA query number 1

Mature ID | Acc#	hsa-let-7a-5p | MIMAT0000062
Mature Sequence	UGAGGUAGUAGGUUGUAUAGUU
Stem-loop ID | Acc#	hsa-let-7a-1 | MI0000060
Stem-loop Sequence	see below

```
      u  gu                 uuagggucacac
uggga gag  aguagguuguauaguu             c
||||| |||  ||||||||||||||||             c
auccu uuc  ucaucuaacauaucaa             a
      -  ug                 uagagggucacc
```

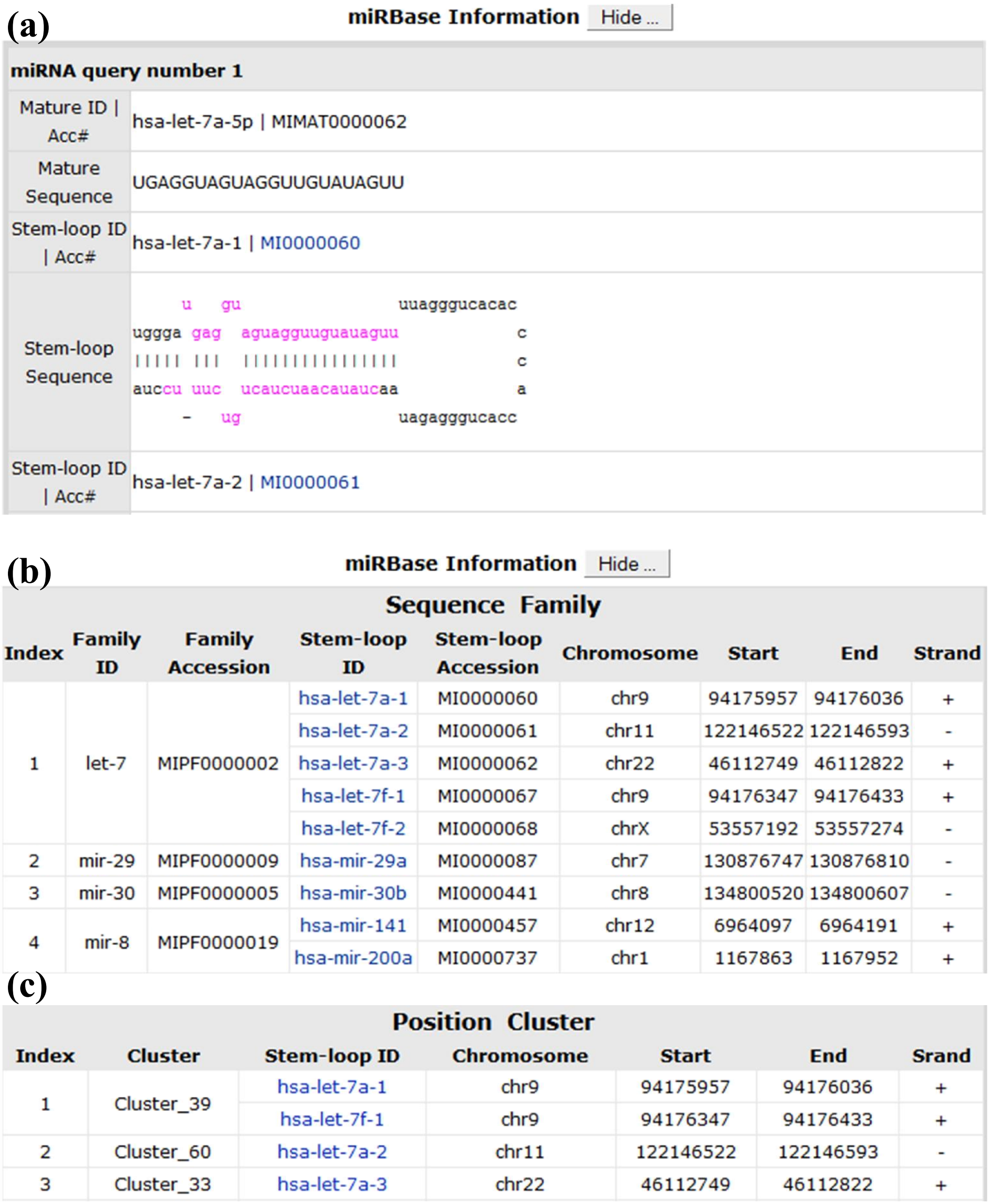
Stem-loop ID | Acc#	hsa-let-7a-2 | MI0000061
Stem-loop Sequence	see below

```
   uu   g   u             uagaa ua a
agg  gag uag agguuguauaguu     u  c u
|||  ||| ||| |||||||||||||     |  | c
ucc  uuc auc uccgacaugucaa     a  g a
   -u   g   c             --uag gg a
```

Stem-loop ID | Acc#	hsa-let-7a-3 | MI0000062
Stem-loop Sequence	see below

```
   u  gu                ---------
ggg gag  aguagguuguauaguu         uggggcu
||| |||  ||||||||||||||||         |||||| c
ucc uuc  ucaucuaacauaucaa         gucccgu
```

Figure 2-6. The result of miRNA annotation. (a) Annotation navigation panel. (b) ID conversion tracking panel. (c) Annotation revealing panel.



miRNA query number 1

Mature ID | Acc#	hsa-let-7a-5p | MIMAT0000062
Mature Sequence	UGAGGUAGUAGGUUGUAUAGUU
Stem-loop ID | Acc#	hsa-let-7a-1 | MI0000060

Stem-loop Sequence

```
     u   gu                uuagggucacac
uggga gag  aguagguuguauaguu            c
||||| |||  ||||||||||||||||            c
auccu uuc  ucaucuaacauaucaa            a
     -   ug                uagagggucacc
```

Stem-loop ID | Acc#	hsa-let-7a-2 | MI0000061

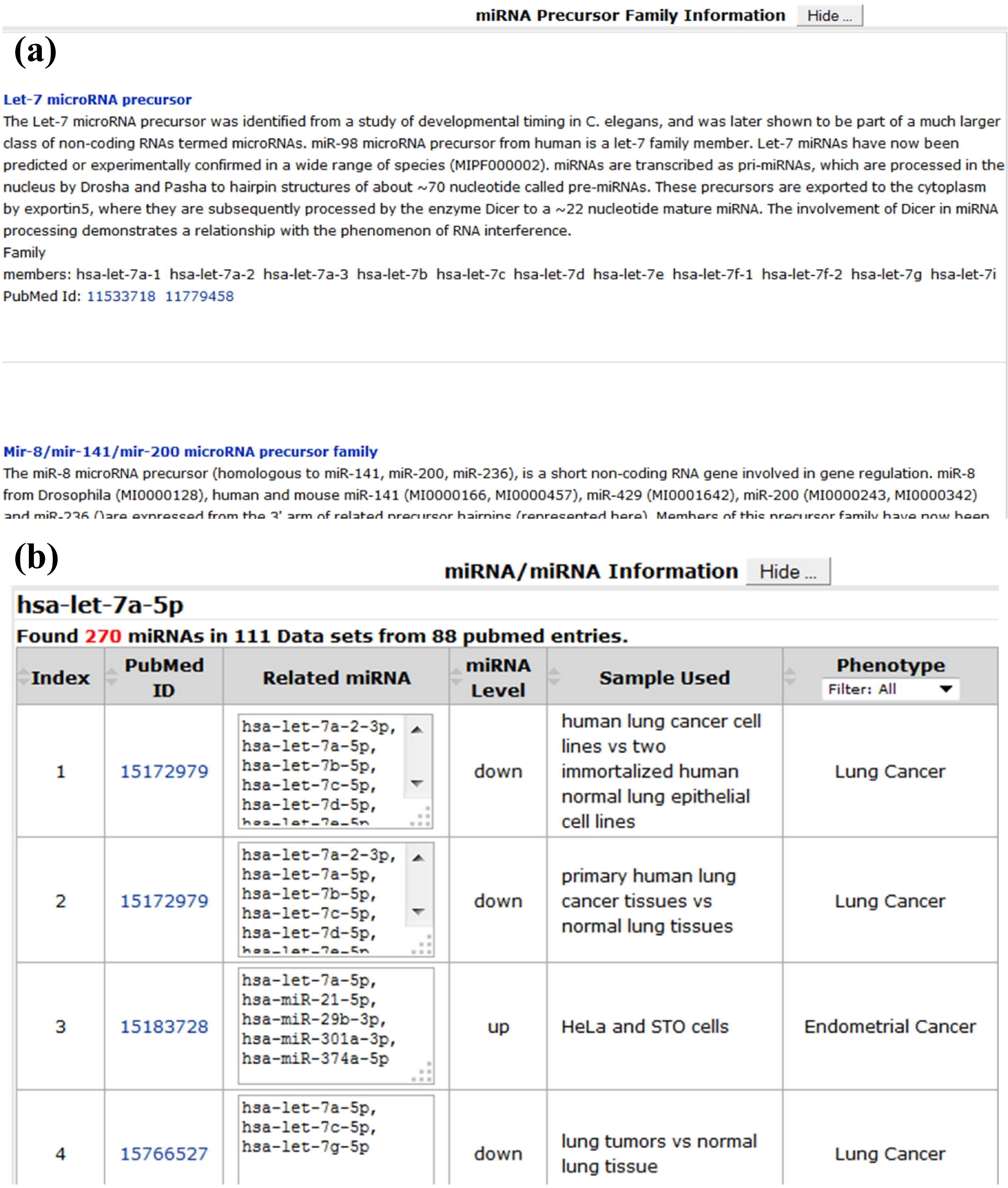
Sequence Family

Index	Family ID	Family Accession	Stem-loop ID	Stem-loop Accession	Chromosome	Start	End	Strand
1	let-7	MIPF0000002	hsa-let-7a-1	MI0000060	chr9	94175957	94176036	+
			hsa-let-7a-2	MI0000061	chr11	122146522	122146593	-
			hsa-let-7a-3	MI0000062	chr22	46112749	46112822	+
			hsa-let-7f-1	MI0000067	chr9	94176347	94176433	+
			hsa-let-7f-2	MI0000068	chrX	53557192	53557274	-
2	mir-29	MIPF0000009	hsa-mir-29a	MI0000087	chr7	130876747	130876810	-
3	mir-30	MIPF0000005	hsa-mir-30b	MI0000441	chr8	134800520	134800607	-
4	mir-8	MIPF0000019	hsa-mir-141	MI0000457	chr12	6964097	6964191	+
			hsa-mir-200a	MI0000737	chr1	1167863	1167952	+

Position Cluster

Index	Cluster	Stem-loop ID	Chromosome	Start	End	Srand
1	Cluster_39	hsa-let-7a-1	chr9	94175957	94176036	+
		hsa-let-7f-1	chr9	94176347	94176433	+
2	Cluster_60	hsa-let-7a-2	chr11	122146522	122146593	-
3	Cluster_33	hsa-let-7a-3	chr22	46112749	46112822	+

Figure 2-7. Details of "miRNA basic information (miRBase)": (a) Identities and sequences of queried miRNAs; (b) Family clusters relevant to queried miRNAs. (c) Position clusters relevant to queried miRNAs.



miRNA Precursor Family Information Hide ...

Let-7 microRNA precursor

The Let-7 microRNA precursor was identified from a study of developmental timing in C. elegans, and was later shown to be part of a much larger class of non-coding RNAs termed microRNAs. miR-98 microRNA precursor from human is a let-7 family member. Let-7 miRNAs have now been predicted or experimentally confirmed in a wide range of species (MIPF000002). miRNAs are transcribed as pri-miRNAs, which are processed in the nucleus by Drosha and Pasha to hairpin structures of about ~70 nucleotide called pre-miRNAs. These precursors are exported to the cytoplasm by exportin5, where they are subsequently processed by the enzyme Dicer to a ~22 nucleotide mature miRNA. The involvement of Dicer in miRNA processing demonstrates a relationship with the phenomenon of RNA interference.

Family members: hsa-let-7a-1 hsa-let-7a-2 hsa-let-7a-3 hsa-let-7b hsa-let-7c hsa-let-7d hsa-let-7e hsa-let-7f-1 hsa-let-7f-2 hsa-let-7g hsa-let-7i

PubMed Id: 11533718 11779458

Mir-8/mir-141/mir-200 microRNA precursor family

The miR-8 microRNA precursor (homologous to miR-141, miR-200, miR-236), is a short non-coding RNA gene involved in gene regulation. miR-8 from Drosophila (MI0000128), human and mouse miR-141 (MI0000166, MI0000457), miR-429 (MI0001642), miR-200 (MI0000243, MI0000342) and miR-236 ()are expressed from the 3' arm of related precursor hairpins (represented here). Members of this precursor family have now been

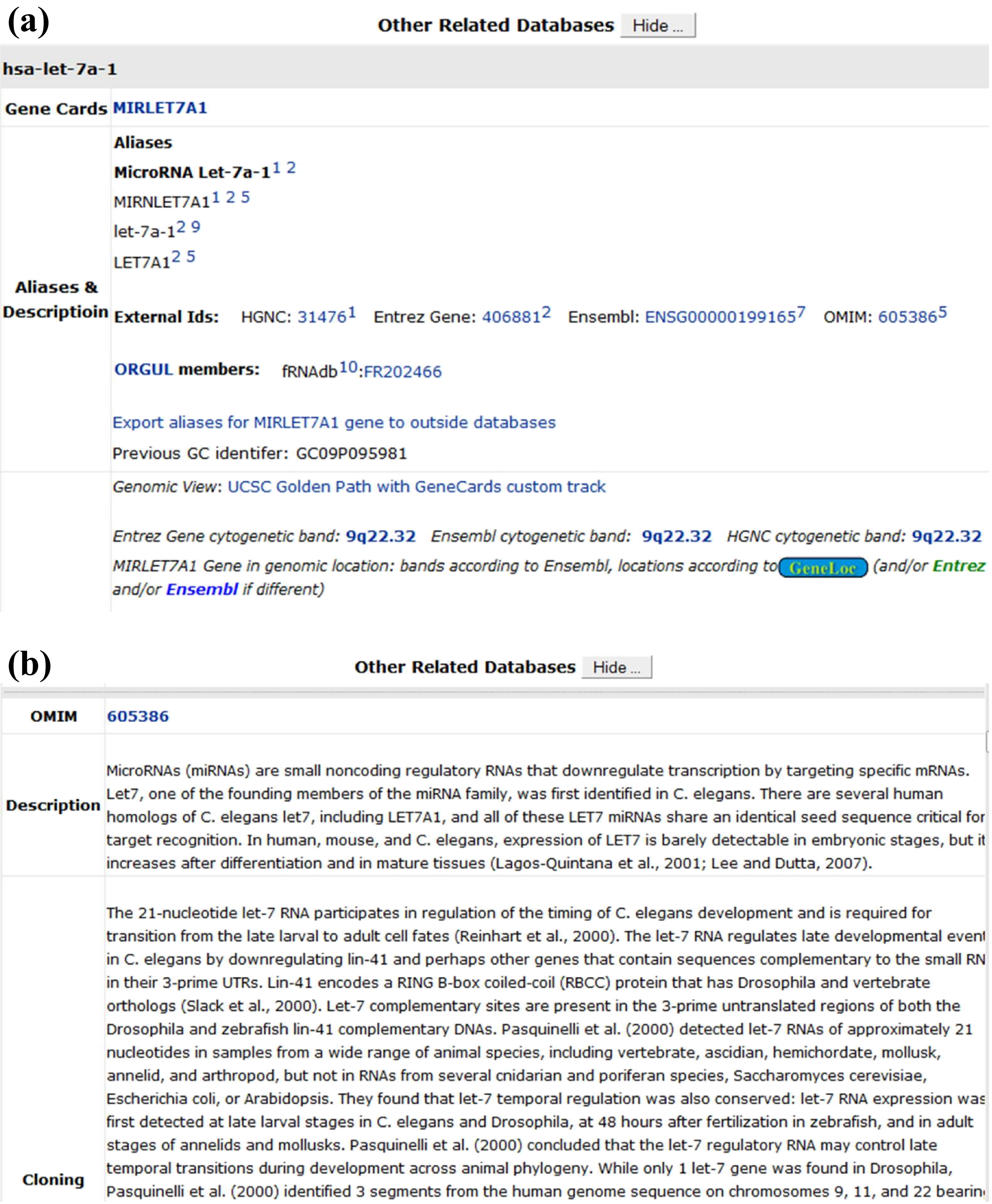
miRNA/miRNA Information Hide ...

hsa-let-7a-5p

Found 270 miRNAs in 111 Data sets from 88 pubmed entries.

Index	PubMed ID	Related miRNA	miRNA Level	Sample Used	Phenotype Filter: All
1	15172979	hsa-let-7a-2-3p, hsa-let-7a-5p, hsa-let-7b-5p, hsa-let-7c-5p, hsa-let-7d-5p, hsa-let-7e-5p	down	human lung cancer cell lines vs two immortalized human normal lung epithelial cell lines	Lung Cancer
2	15172979	hsa-let-7a-2-3p, hsa-let-7a-5p, hsa-let-7b-5p, hsa-let-7c-5p, hsa-let-7d-5p, hsa-let-7e-5p	down	primary human lung cancer tissues vs normal lung tissues	Lung Cancer
3	15183728	hsa-let-7a-5p, hsa-miR-21-5p, hsa-miR-29b-3p, hsa-miR-301a-3p, hsa-miR-374a-5p	up	HeLa and STO cells	Endometrial Cancer
4	15766527	hsa-let-7a-5p, hsa-let-7c-5p, hsa-let-7g-5p	down	lung tumors vs normal lung tissue	Lung Cancer

Figure 2-8. Detailed information of both "miRNA precursor (Wikipedia)" and "miRNA/miRNA relationships (miRNA-miRNA)": (a) precursor families concerning queried miRNAs; (b) miRNA-miRNA correlations in diverse diseases.



(a)

Other Related Databases Hide ...

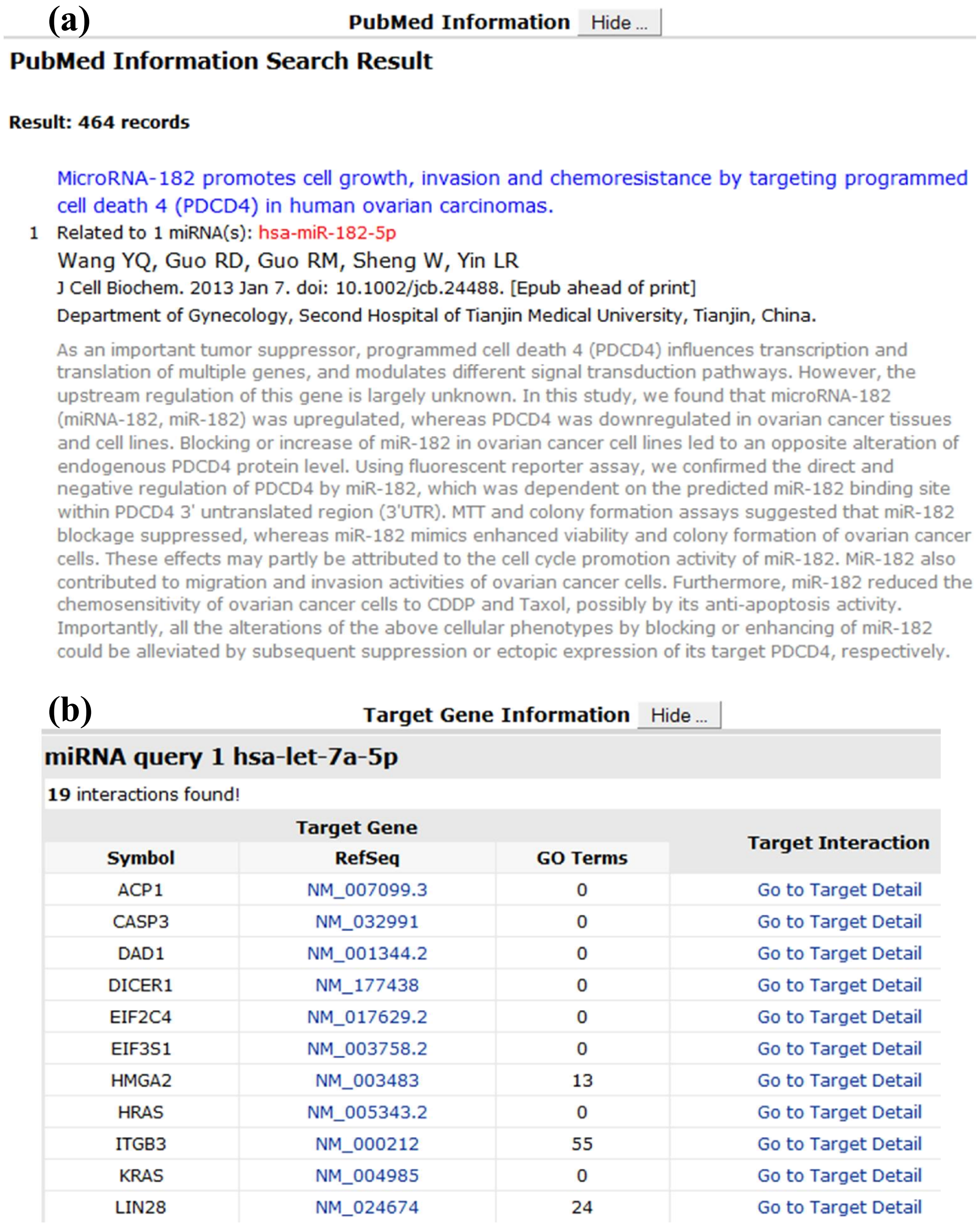
hsa-let-7a-1	
Gene Cards	**MIRLET7A1**
Aliases & Descriptioin	**Aliases** **MicroRNA Let-7a-1**[1 2] MIRNLET7A1[1 2 5] let-7a-1[2 9] LET7A1[2 5] **External Ids:** HGNC: 31476[1] Entrez Gene: 406881[2] Ensembl: ENSG00000199165[7] OMIM: 605386[5] **ORGUL members:** fRNAdb[10]:FR202466 Export aliases for MIRLET7A1 gene to outside databases Previous GC identifer: GC09P095981
	Genomic View: UCSC Golden Path with GeneCards custom track *Entrez Gene cytogenetic band:* **9q22.32** *Ensembl cytogenetic band:* **9q22.32** *HGNC cytogenetic band:* **9q22.32** *MIRLET7A1 Gene in genomic location: bands according to Ensembl, locations according to* GeneLoc *(and/or* ***Entrez*** *and/or* ***Ensembl*** *if different)*

(b)

Other Related Databases Hide ...

OMIM	**605386**
Description	MicroRNAs (miRNAs) are small noncoding regulatory RNAs that downregulate transcription by targeting specific mRNAs. Let7, one of the founding members of the miRNA family, was first identified in C. elegans. There are several human homologs of C. elegans let7, including LET7A1, and all of these LET7 miRNAs share an identical seed sequence critical for target recognition. In human, mouse, and C. elegans, expression of LET7 is barely detectable in embryonic stages, but it increases after differentiation and in mature tissues (Lagos-Quintana et al., 2001; Lee and Dutta, 2007).
Cloning	The 21-nucleotide let-7 RNA participates in regulation of the timing of C. elegans development and is required for transition from the late larval to adult cell fates (Reinhart et al., 2000). The let-7 RNA regulates late developmental event in C. elegans by downregulating lin-41 and perhaps other genes that contain sequences complementary to the small RN in their 3-prime UTRs. Lin-41 encodes a RING B-box coiled-coil (RBCC) protein that has Drosophila and vertebrate orthologs (Slack et al., 2000). Let-7 complementary sites are present in the 3-prime untranslated regions of both the Drosophila and zebrafish lin-41 complementary DNAs. Pasquinelli et al. (2000) detected let-7 RNAs of approximately 21 nucleotides in samples from a wide range of animal species, including vertebrate, ascidian, hemichordate, mollusk, annelid, and arthropod, but not in RNAs from several cnidarian and poriferan species, Saccharomyces cerevisiae, Escherichia coli, or Arabidopsis. They found that let-7 temporal regulation was also conserved: let-7 RNA expression was first detected at late larval stages in C. elegans and Drosophila, at 48 hours after fertilization in zebrafish, and in adult stages of annelids and mollusks. Pasquinelli et al. (2000) concluded that the let-7 regulatory RNA may control late temporal transitions during development across animal phylogeny. While only 1 let-7 gene was found in Drosophila, Pasquinelli et al. (2000) identified 3 segments from the human genome sequence on chromosomes 9, 11, and 22 bearin

Figure 2-9. Details of "Other Related Databases": (a) a variety of IDs and genomic description from GeneCards; (b) Functional descriptions and corresponding references of miRNAs from OMIM.



PubMed Information Search Result

Result: 464 records

MicroRNA-182 promotes cell growth, invasion and chemoresistance by targeting programmed cell death 4 (PDCD4) in human ovarian carcinomas.

1 Related to 1 miRNA(s): hsa-miR-182-5p

Wang YQ, Guo RD, Guo RM, Sheng W, Yin LR

J Cell Biochem. 2013 Jan 7. doi: 10.1002/jcb.24488. [Epub ahead of print]

Department of Gynecology, Second Hospital of Tianjin Medical University, Tianjin, China.

As an important tumor suppressor, programmed cell death 4 (PDCD4) influences transcription and translation of multiple genes, and modulates different signal transduction pathways. However, the upstream regulation of this gene is largely unknown. In this study, we found that microRNA-182 (miRNA-182, miR-182) was upregulated, whereas PDCD4 was downregulated in ovarian cancer tissues and cell lines. Blocking or increase of miR-182 in ovarian cancer cell lines led to an opposite alteration of endogenous PDCD4 protein level. Using fluorescent reporter assay, we confirmed the direct and negative regulation of PDCD4 by miR-182, which was dependent on the predicted miR-182 binding site within PDCD4 3' untranslated region (3'UTR). MTT and colony formation assays suggested that miR-182 blockage suppressed, whereas miR-182 mimics enhanced viability and colony formation of ovarian cancer cells. These effects may partly be attributed to the cell cycle promotion activity of miR-182. MiR-182 also contributed to migration and invasion activities of ovarian cancer cells. Furthermore, miR-182 reduced the chemosensitivity of ovarian cancer cells to CDDP and Taxol, possibly by its anti-apoptosis activity. Importantly, all the alterations of the above cellular phenotypes by blocking or enhancing of miR-182 could be alleviated by subsequent suppression or ectopic expression of its target PDCD4, respectively.

miRNA query 1 hsa-let-7a-5p

19 interactions found!

Target Gene			Target Interaction
Symbol	RefSeq	GO Terms	
ACP1	NM_007099.3	0	Go to Target Detail
CASP3	NM_032991	0	Go to Target Detail
DAD1	NM_001344.2	0	Go to Target Detail
DICER1	NM_177438	0	Go to Target Detail
EIF2C4	NM_017629.2	0	Go to Target Detail
EIF3S1	NM_003758.2	0	Go to Target Detail
HMGA2	NM_003483	13	Go to Target Detail
HRAS	NM_005343.2	0	Go to Target Detail
ITGB3	NM_000212	55	Go to Target Detail
KRAS	NM_004985	0	Go to Target Detail
LIN28	NM_024674	24	Go to Target Detail

Figure 2-10. Detailed information of both "PubMed Information" and "Target Gene (miRecords)": (a) Publications relevant to input miRNAs; (b) Experimentally validated target genes of input miRNAs.

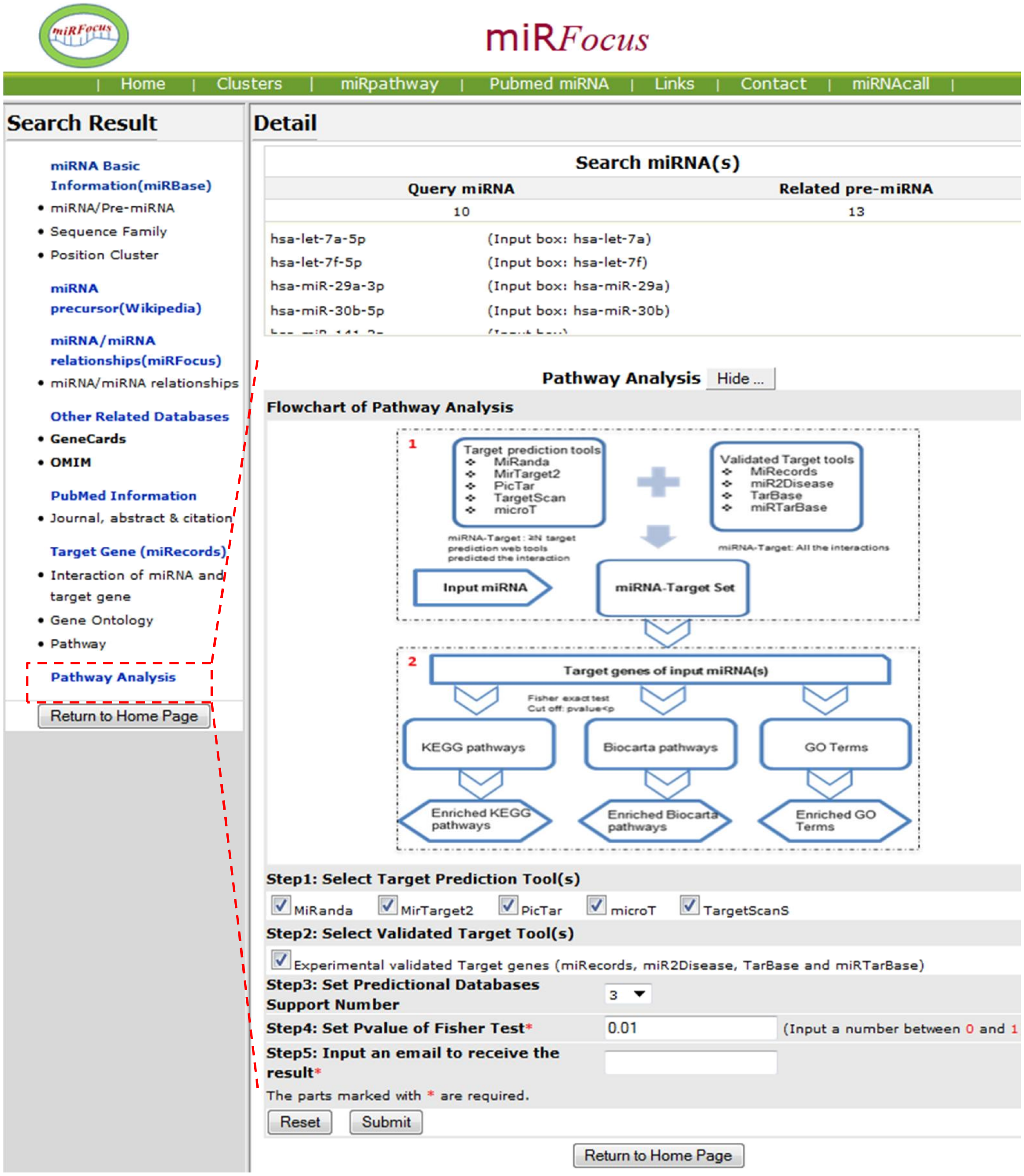
2.3.2 Pathway Analysis

"Pathway analysis" is important web tool for achieving enriched pathways and GO Terms based on the input miRNAs. The input area for "pathway Analysis" (Figure 2-11) will show up by clicking "Pathway Analysis" in the navigation panel which is on the left of the whole page. In the input page of pathway analysis module, a flowchart on the top gives a big picture of data analysis process; step 1 to 5 provides flexible options for users to choose the parameters in line with their needs; in the end the results will be sent to the email provided by users at step 5. In addition, we suggest set "3" in step 3, "Set Prediction Databases Support Number", which can remove excessive false results resulting from a large number of false positive target genes, and obtain enough useful enriched pathways and GO terms.

After receiving an email from mirfocus@mirfocus.org, the result page of pathway analysis will be opened up by clicking the link. From Figure 2-12 we can see that the result page of pathway analysis is composed of a navigation panel on the left and a details exhibition area on the right. The result of pathway analysis is divided into five modules, including "Summary", "KEGG Pathway", "GO Terms", "BioCarta", as well as "Download". The "Download" module is used to save analysis results on local computers. Detailed information of "Summary" reveals input miRNAs with target genes (Figure 2-12 a), and miRNA-target interaction supported by which databases (Figure 2-12 b). Since the outputs of "KEGG Pathway", "GO Terms" and "BioCarta" are displayed in the same format, the following will utilize "KEGG Pathway" as an example to introduce the results. Clicking "KEGG Pathway" in navigation panel will link to "KEGG Pathway Detail" (Figure 2-13 a) which includes KEGG pathway IDs that can link to

KEGG, -Log10 of P-value the probability of obtaining insignificant results, pathway description and “Show Detail”. Details of pathways containing input miRNAs and corresponding target genes will exhibited by clicking “Show Detail”.

It is evident that ten input miRNAs are highly related to Foxo signaling pathway, pathways in cancer, pathway related to chronic myeloid leukemia (CML) and so forth. And the miRNAs are involved in these pathways by regulating multiple target genes.



Search Result

miRNA Basic Information(miRBase)
- miRNA/Pre-miRNA
- Sequence Family
- Position Cluster

miRNA precursor(Wikipedia)

miRNA/miRNA relationships(miRFocus)
- miRNA/miRNA relationships

Other Related Databases
- GeneCards
- OMIM

PubMed Information
- Journal, abstract & citation

Target Gene (miRecords)
- Interaction of miRNA and target gene
- Gene Ontology
- Pathway

Pathway Analysis

Detail

Search miRNA(s)

Query miRNA	Related pre-miRNA
10	13

hsa-let-7a-5p (Input box: hsa-let-7a)
hsa-let-7f-5p (Input box: hsa-let-7f)
hsa-miR-29a-3p (Input box: hsa-miR-29a)
hsa-miR-30b-5p (Input box: hsa-miR-30b)

Pathway Analysis

Flowchart of Pathway Analysis

Step1: Select Target Prediction Tool(s)

☑ MiRanda ☑ MirTarget2 ☑ PicTar ☑ microT ☑ TargetScanS

Step2: Select Validated Target Tool(s)

☑ Experimental validated Target genes (miRecords, miR2Disease, TarBase and miRTarBase)

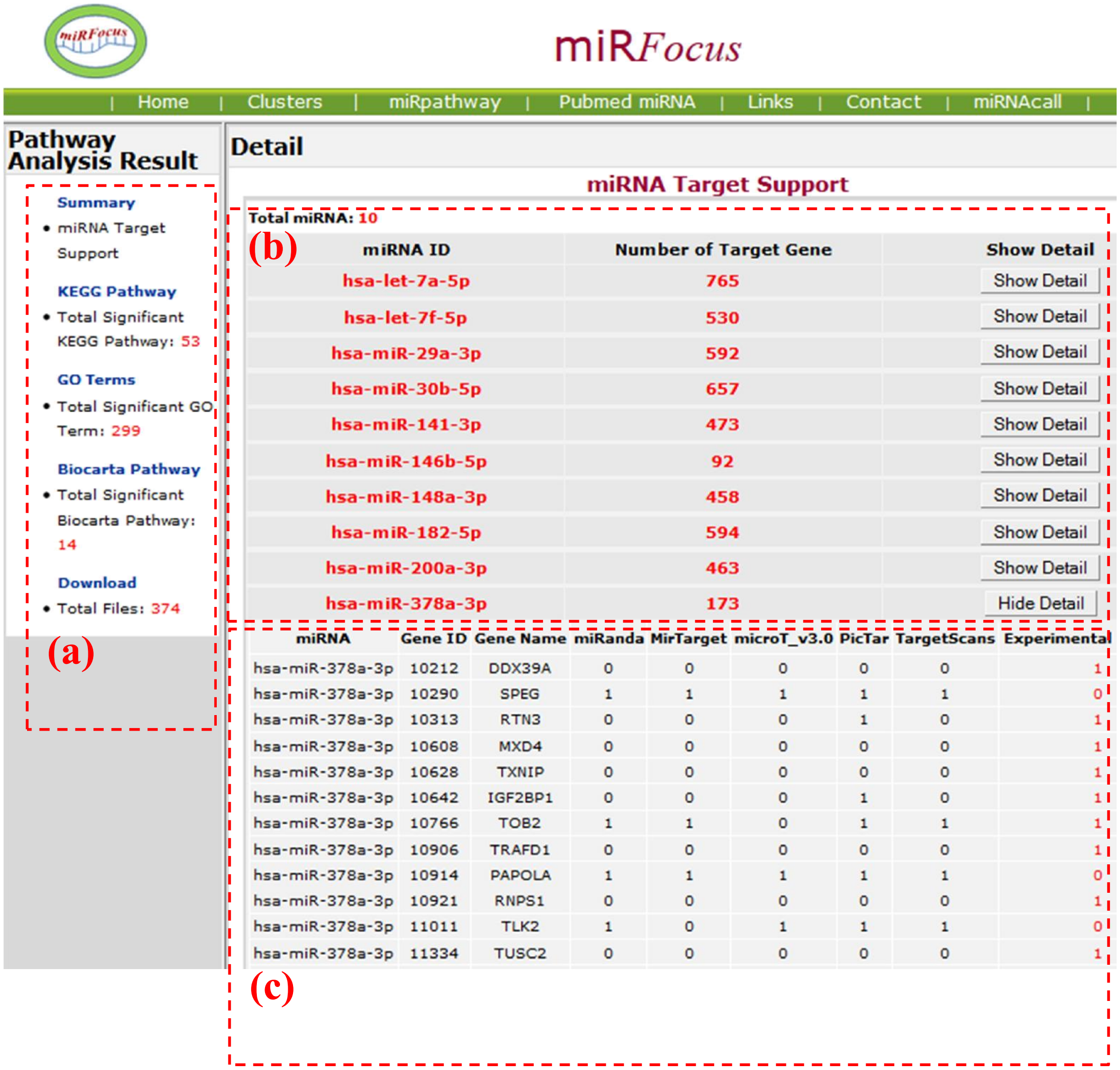
Step3: Set Predictional Databases Support Number 3

Step4: Set Pvalue of Fisher Test* 0.01 (Input a number between 0 and 1

Step5: Input an email to receive the result*

The parts marked with * are required.

Figure 2-11. The input page of pathway analysis module. Click Link "Pathway Analysis" to show this page.


miRFocus
miRFocus
| Home | Clusters | miRpathway | Pubmed miRNA | Links | Contact | miRNAcall |
(a)
(b)
(c)
Show Detail
Hide Detail

Pathway Analysis Result

Summary
- miRNA Target Support

KEGG Pathway
- Total Significant KEGG Pathway: 53

GO Terms
- Total Significant GO Term: 299

Biocarta Pathway
- Total Significant Biocarta Pathway: 14

Download
- Total Files: 374

Detail

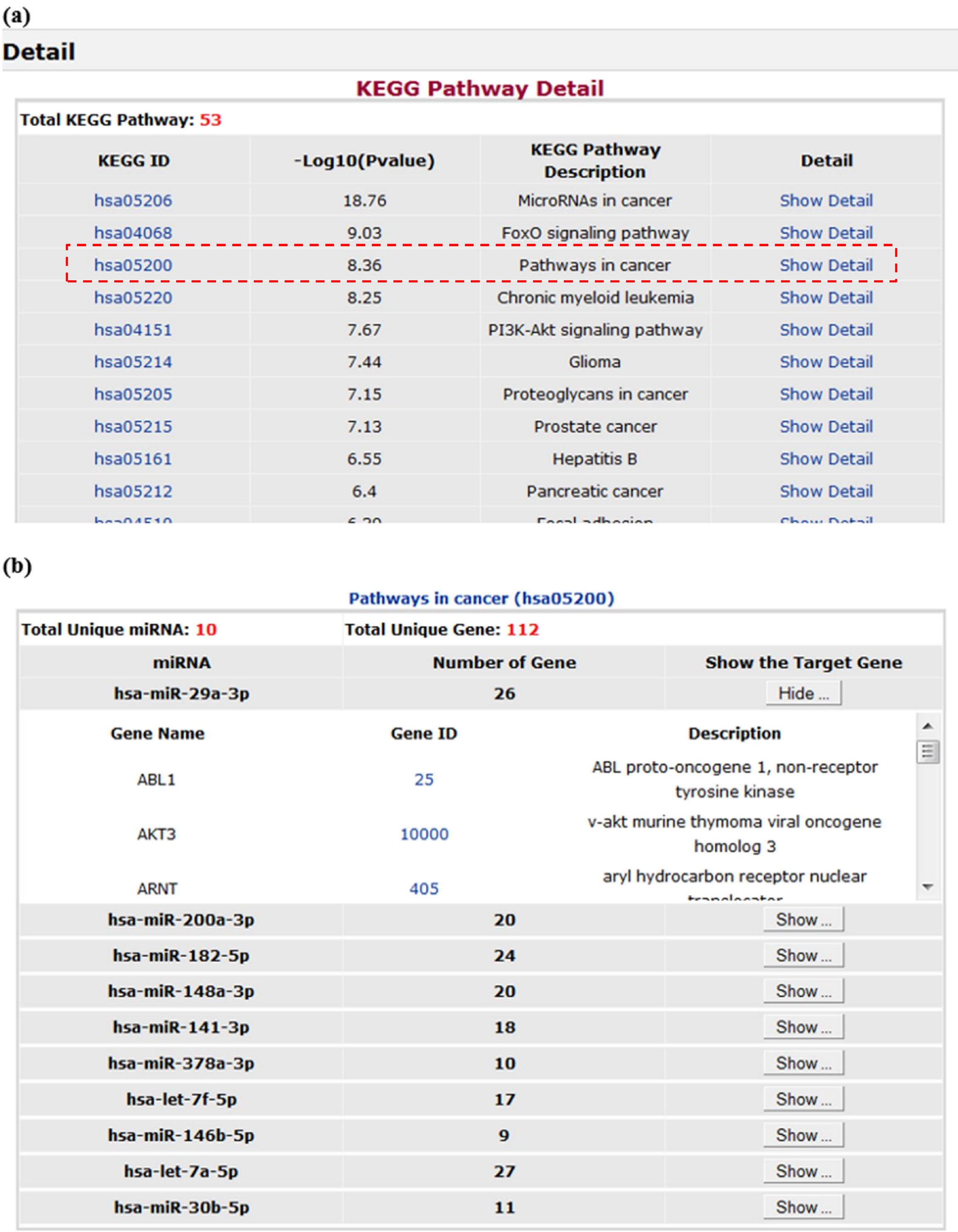
miRNA Target Support

Total miRNA: 10

miRNA ID	Number of Target Gene	Show Detail
hsa-let-7a-5p	765	Show Detail
hsa-let-7f-5p	530	Show Detail
hsa-miR-29a-3p	592	Show Detail
hsa-miR-30b-5p	657	Show Detail
hsa-miR-141-3p	473	Show Detail
hsa-miR-146b-5p	92	Show Detail
hsa-miR-148a-3p	458	Show Detail
hsa-miR-182-5p	594	Show Detail
hsa-miR-200a-3p	463	Show Detail
hsa-miR-378a-3p	173	Hide Detail

miRNA	Gene ID	Gene Name	miRanda	MirTarget	microT_v3.0	PicTar	TargetScans	Experimental
hsa-miR-378a-3p	10212	DDX39A	0	0	0	0	0	1
hsa-miR-378a-3p	10290	SPEG	1	1	1	1	1	0
hsa-miR-378a-3p	10313	RTN3	0	0	0	1	0	1
hsa-miR-378a-3p	10608	MXD4	0	0	0	0	0	1
hsa-miR-378a-3p	10628	TXNIP	0	0	0	0	0	1
hsa-miR-378a-3p	10642	IGF2BP1	0	0	0	1	0	1
hsa-miR-378a-3p	10766	TOB2	1	1	0	1	1	1
hsa-miR-378a-3p	10906	TRAFD1	0	0	0	0	0	1
hsa-miR-378a-3p	10914	PAPOLA	1	1	1	1	1	0
hsa-miR-378a-3p	10921	RNPS1	0	0	0	0	0	1
hsa-miR-378a-3p	11011	TLK2	1	0	1	1	1	0
hsa-miR-378a-3p	11334	TUSC2	0	0	0	0	0	1

Figure 2-12. The result page of pathway analysis. (a) Navigation panel. (b) Input miRNAs with target genes. (c) Target genes of each miRNA.



(a)

Detail

KEGG Pathway Detail

Total KEGG Pathway: 53

KEGG ID	-Log10(Pvalue)	KEGG Pathway Description	Detail
hsa05206	18.76	MicroRNAs in cancer	Show Detail
hsa04068	9.03	FoxO signaling pathway	Show Detail
hsa05200	8.36	Pathways in cancer	Show Detail
hsa05220	8.25	Chronic myeloid leukemia	Show Detail
hsa04151	7.67	PI3K-Akt signaling pathway	Show Detail
hsa05214	7.44	Glioma	Show Detail
hsa05205	7.15	Proteoglycans in cancer	Show Detail
hsa05215	7.13	Prostate cancer	Show Detail
hsa05161	6.55	Hepatitis B	Show Detail
hsa05212	6.4	Pancreatic cancer	Show Detail
hsa04510	6.2[illegible]	Focal adhesion	Show Detail

(b)

Pathways in cancer (hsa05200)

Total Unique miRNA: 10 **Total Unique Gene: 112**

miRNA	Number of Gene	Show the Target Gene
hsa-miR-29a-3p	26	Hide ...

Gene Name	Gene ID	Description
ABL1	25	ABL proto-oncogene 1, non-receptor tyrosine kinase
AKT3	10000	v-akt murine thymoma viral oncogene homolog 3
ARNT	405	aryl hydrocarbon receptor nuclear translocator

miRNA	Number of Gene	Show the Target Gene
hsa-miR-200a-3p	20	Show ...
hsa-miR-182-5p	24	Show ...
hsa-miR-148a-3p	20	Show ...
hsa-miR-141-3p	18	Show ...
hsa-miR-378a-3p	10	Show ...
hsa-let-7f-5p	17	Show ...
hsa-miR-146b-5p	9	Show ...
hsa-let-7a-5p	27	Show ...
hsa-miR-30b-5p	11	Show ...

Figure 2-13. The result of enriched KEGG pathways. (a) Details of enriched KEGG pathways. (b) Input miRNAs and their target genes of pathway in caner.

2.3.3 miRpathway

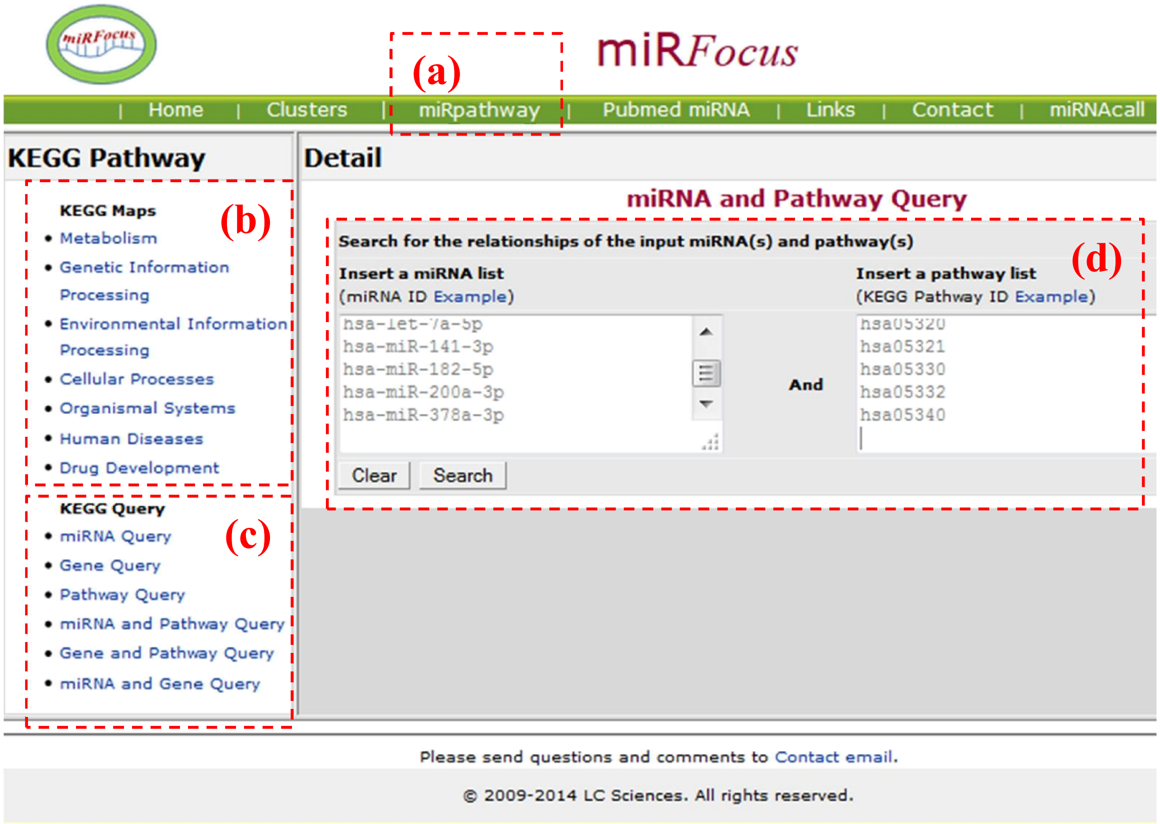
Based on the outputs of pathway analysis, it is easy to achieve the important relations between queried miRNAs and cancer related pathways. It is reported that miRNAs in breast cancer play a vital role in immune system, but which does not show up in the result of pathway analysis [97]. Therefore, we turn to miRpathway to study the correlations between top ten highest expression miRNAs in breast milk exosomes and pathways related to immune diseases.

The query page of miRpathway module will show up by clicking “miRpathway” (Figure 2-14a) on the navigation bar which is on the top of the whole page. In the miRpathway statistic page, it shows that there are 8 pathways related to immune diseases, which cover 320 unique genes and 201 unique miRNA in miRpathway database (Figure 2-15). The detailed information of these miRNAs and genes can be observed by click “Immune diseases” (Figure 2-15a).

To directly focus on relations between top ten highest expression miRNAs in breast milk exosomes and immune pathways, we can choose “miRNA and Pathway Query” (Figure 2-14c) from KEGG Query panel, and input ten miRNA IDs and eight KEGG pathway IDs into input panel (Figure 2-14d).

The query result (Figure 2-16) shows that almost all the pathways except for Primary immunodeficiency pathway are related to one or multiple of the input miRNAs. Figure 2-16b denotes correlations between miRNAs, genes and pathways, for example, hsa-miR-148a-3p takes part in autoimmune thyroid disease pathway，Allograft rejection pathway and Graft-versus-host disease pathway through regulating HLA-G, while hsa-let-7f-5p can interact with IL13 to affect asthma pathway. Therefore, it is of no doubt that miRNAs in breast milk exosomes have much to do with immune system.

From the results of miRpathway, it is discovered that miRNAs have effects on immune diseases related pathways by regulating only one or two genes, while they can bind to multiple genes to affect cancer related pathways. That is why immune diseases related pathways do not pass the cutoff of pathway analysis. It is true that we cannot ignore the effects of miRNAs on the pathways, even if they are just able to control the sole gene of the pathways. Studying correlations between miRNA, target gene and pathway beyond complicated interactions is one advantage of miRpathway.



miRNA and Pathway Query

Search for the relationships of the input miRNA(s) and pathway(s)

Insert a miRNA list
(miRNA ID Example)

```
hsa-let-7a-5p
hsa-miR-141-3p
hsa-miR-182-5p
hsa-miR-200a-3p
hsa-miR-378a-3p
```

And

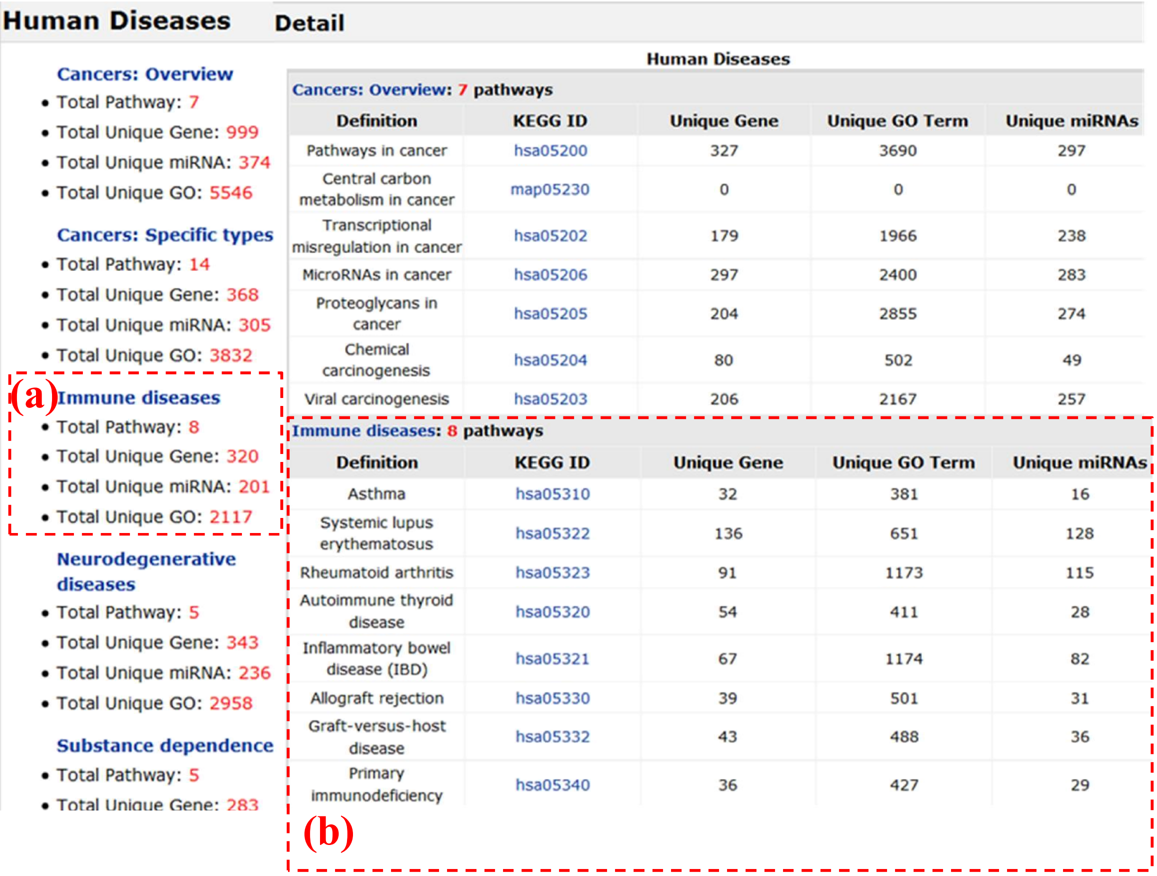
Insert a pathway list
(KEGG Pathway ID Example)

```
hsa05320
hsa05321
hsa05330
hsa05332
hsa05340
```

Please send questions and comments to Contact email.

© 2009-2014 LC Sciences. All rights reserved.

Figure 2-14. The home page of miRpathway. (a) Navigation button of miRpathway. (b) "KEGG Maps" panel for searching miRpathway statistic information of KEGG pathway catalogs. (c) "KEGG Query" panel for different types of query. (d) Input panel for query.



Human Diseases

Cancers: Overview

- Total Pathway: 7
- Total Unique Gene: 999
- Total Unique miRNA: 374
- Total Unique GO: 5546

Cancers: Specific types

- Total Pathway: 14
- Total Unique Gene: 368
- Total Unique miRNA: 305
- Total Unique GO: 3832

Immune diseases

- Total Pathway: 8
- Total Unique Gene: 320
- Total Unique miRNA: 201
- Total Unique GO: 2117

Neurodegenerative diseases

- Total Pathway: 5
- Total Unique Gene: 343
- Total Unique miRNA: 236
- Total Unique GO: 2958

Substance dependence

- Total Pathway: 5
- Total Unique Gene: 283

Detail

Human Diseases

Cancers: Overview: 7 pathways

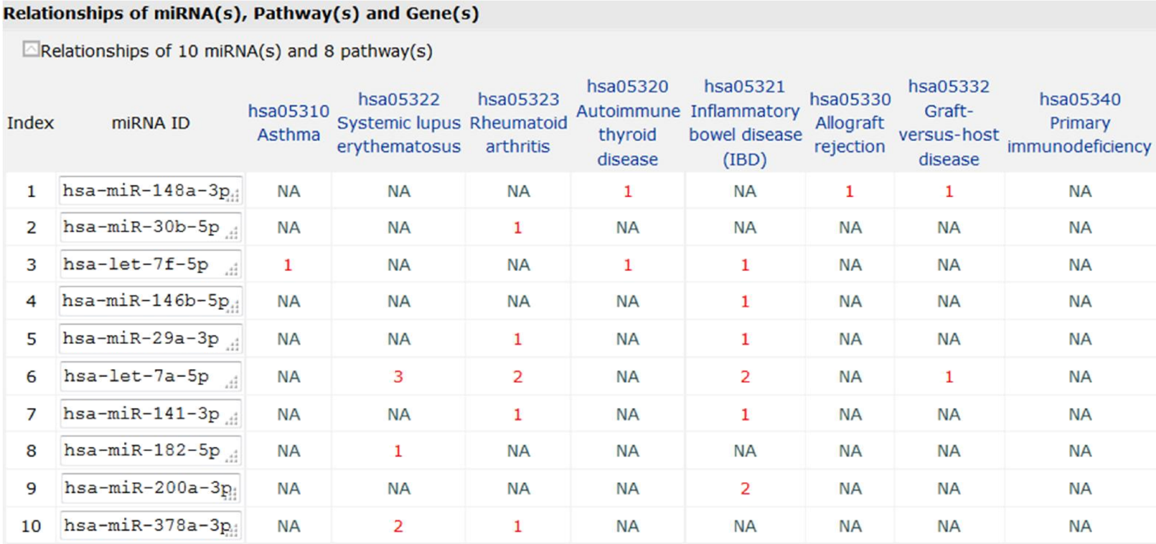
Definition	KEGG ID	Unique Gene	Unique GO Term	Unique miRNAs
Pathways in cancer	hsa05200	327	3690	297
Central carbon metabolism in cancer	map05230	0	0	0
Transcriptional misregulation in cancer	hsa05202	179	1966	238
MicroRNAs in cancer	hsa05206	297	2400	283
Proteoglycans in cancer	hsa05205	204	2855	274
Chemical carcinogenesis	hsa05204	80	502	49
Viral carcinogenesis	hsa05203	206	2167	257

Immune diseases: 8 pathways

Definition	KEGG ID	Unique Gene	Unique GO Term	Unique miRNAs
Asthma	hsa05310	32	381	16
Systemic lupus erythematosus	hsa05322	136	651	128
Rheumatoid arthritis	hsa05323	91	1173	115
Autoimmune thyroid disease	hsa05320	54	411	28
Inflammatory bowel disease (IBD)	hsa05321	67	1174	82
Allograft rejection	hsa05330	39	501	31
Graft-versus-host disease	hsa05332	43	488	36
Primary immunodeficiency	hsa05340	36	427	29

Figure 2-15. The statistics of "Human Diseases" catalog in miRpathway. (a) Statistic information of all immune diseases related pathways. (b) Detailed statistical information of each pathway related to immune diseases.

(a)

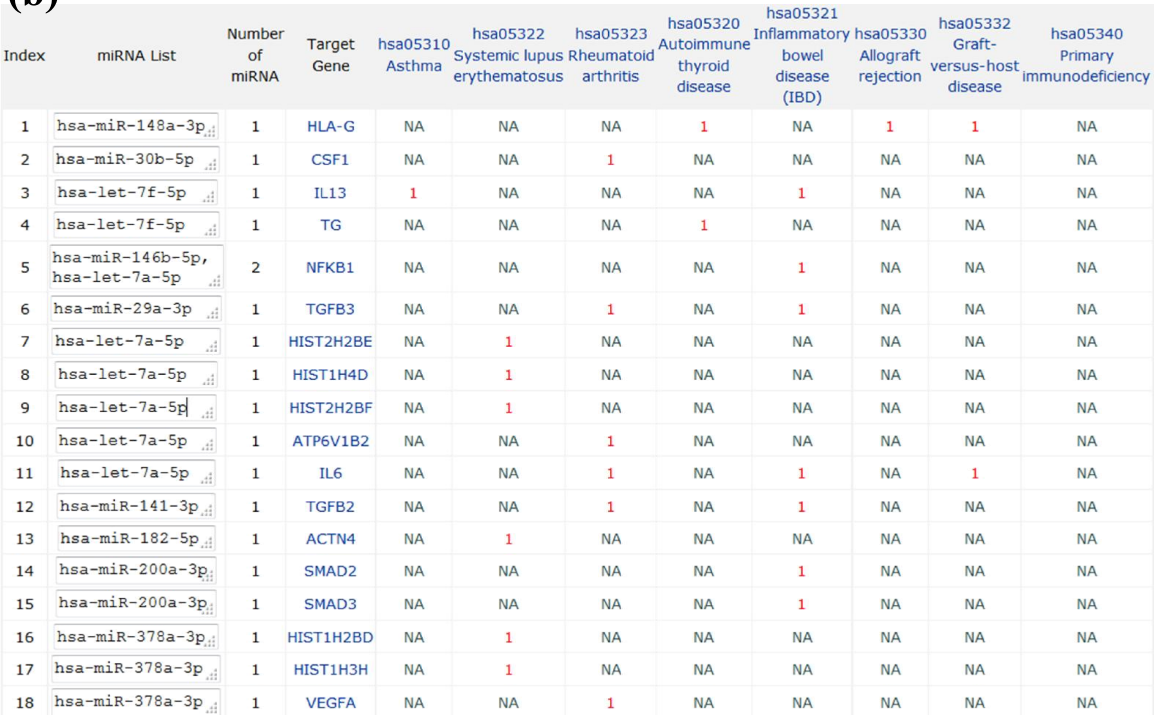


Relationships of miRNA(s), Pathway(s) and Gene(s)

□Relationships of 10 miRNA(s) and 8 pathway(s)

Index	miRNA ID	hsa05310 Asthma	hsa05322 Systemic lupus erythematosus	hsa05323 Rheumatoid arthritis	hsa05320 Autoimmune thyroid disease	hsa05321 Inflammatory bowel disease (IBD)	hsa05330 Allograft rejection	hsa05332 Graft-versus-host disease	hsa05340 Primary immunodeficiency
1	hsa-miR-148a-3p	NA	NA	NA	1	NA	1	1	NA
2	hsa-miR-30b-5p	NA	NA	1	NA	NA	NA	NA	NA
3	hsa-let-7f-5p	1	NA	NA	1	1	NA	NA	NA
4	hsa-miR-146b-5p	NA	NA	NA	NA	1	NA	NA	NA
5	hsa-miR-29a-3p	NA	NA	1	NA	1	NA	NA	NA
6	hsa-let-7a-5p	NA	3	2	NA	2	NA	1	NA
7	hsa-miR-141-3p	NA	NA	1	NA	1	NA	NA	NA
8	hsa-miR-182-5p	NA	1	NA	NA	NA	NA	NA	NA
9	hsa-miR-200a-3p	NA	NA	NA	NA	2	NA	NA	NA
10	hsa-miR-378a-3p	NA	2	1	NA	NA	NA	NA	NA

(b)



Index	miRNA List	Number of miRNA	Target Gene	hsa05310 Asthma	hsa05322 Systemic lupus erythematosus	hsa05323 Rheumatoid arthritis	hsa05320 Autoimmune thyroid disease	hsa05321 Inflammatory bowel disease (IBD)	hsa05330 Allograft rejection	hsa05332 Graft-versus-host disease	hsa05340 Primary immunodeficiency
1	hsa-miR-148a-3p	1	HLA-G	NA	NA	NA	1	NA	1	1	NA
2	hsa-miR-30b-5p	1	CSF1	NA	NA	1	NA	NA	NA	NA	NA
3	hsa-let-7f-5p	1	IL13	1	NA	NA	NA	1	NA	NA	NA
4	hsa-let-7f-5p	1	TG	NA	NA	NA	1	NA	NA	NA	NA
5	hsa-miR-146b-5p, hsa-let-7a-5p	2	NFKB1	NA	NA	NA	NA	1	NA	NA	NA
6	hsa-miR-29a-3p	1	TGFB3	NA	NA	1	NA	1	NA	NA	NA
7	hsa-let-7a-5p	1	HIST2H2BE	NA	1	NA	NA	NA	NA	NA	NA
8	hsa-let-7a-5p	1	HIST1H4D	NA	1	NA	NA	NA	NA	NA	NA
9	hsa-let-7a-5p	1	HIST2H2BF	NA	1	NA	NA	NA	NA	NA	NA
10	hsa-let-7a-5p	1	ATP6V1B2	NA	NA	1	NA	NA	NA	NA	NA
11	hsa-let-7a-5p	1	IL6	NA	NA	1	NA	1	NA	1	NA
12	hsa-miR-141-3p	1	TGFB2	NA	NA	1	NA	1	NA	NA	NA
13	hsa-miR-182-5p	1	ACTN4	NA	1	NA	NA	NA	NA	NA	NA
14	hsa-miR-200a-3p	1	SMAD2	NA	NA	NA	NA	1	NA	NA	NA
15	hsa-miR-200a-3p	1	SMAD3	NA	NA	NA	NA	1	NA	NA	NA
16	hsa-miR-378a-3p	1	HIST1H2BD	NA	1	NA	NA	NA	NA	NA	NA
17	hsa-miR-378a-3p	1	HIST1H3H	NA	1	NA	NA	NA	NA	NA	NA
18	hsa-miR-378a-3p	1	VEGFA	NA	NA	1	NA	NA	NA	NA	NA

Figure 2-16. The results of miRpathway by using query method "miRNA and pathway query". (a) The relations between queried miRNAs and KEGG pathways. (b) The relations between queried miRNAs and KEGG pathways and target genes.

2.4 Conclusions

In this chapter, we present an overview of three major modules of miRFocus, and demonstrate the useful and reasonable results with an example. It is easier to conclude that miRFocus is a comprehensive web-based resource for human miRNAs information retrieval and further analysis. Based on the results of miRNA annotation and pathway analysis, it is discovered that top ten highest expression miRNAs in breast milk exosomes have all been well studied, and these miRNAs are able to take part in cancer related pathways by regulating numerous target genes. Moreover, on the basis of analysis results of miRpathway module, we ensured that important connections between miRNAs of breast milk exosomes and immune diseases related pathways are really reliable, and also observed the candidate connectors – target genes.

MiRFocus includes three major modules: miRNA annotation, pathway analysis and miRpathway. MiRNA annotation contains the most comprehensive known annotation relevant to human miRNAs, including miRBase annotation, Wikipedia annotation, miRNA-miRNA annotation, GeneCards annotation, OMIM annotation and miRecords annotation. Prior to reaching the result page of miRNA annotation, different input miRNA IDs will be automatically converted into new miRNA IDs of the latest version (miRBase version 21) by miRNA ID conversion function, thus with the help of miRNA Gene ID conversion function new miRNA IDs will be transferred into Gene related ID. Pathway analysis module implements a statistic method to determine whether each pathway or GO term is significantly affected by input miRNAs or not. Since both pathway analysis and GO terms analysis require the interactions between miRNA and genes, an integrates miRNA-target set is created by collecting interactions from four experimentally validated target databases and five predicted target web tools. To facilitate users to achieve correlations between miRNAs, target genes, and KEGG pathways, miRpathway module not only summarizes miRNAs and experimentally validated target genes of each human KEGG pathways, but also creates multiple query methods.

Chapter 3

NGS Analysis for Error Removal in DNA Microarray-Based Synthetic Oligonucleotide Libraries

3.1 Introduction

As the technology of DNA sequencing develops, our understanding of DNA sequences and related biological knowledge is also gradually increasing. Scientists are not just satisfied by the existing natural DNA sequences in biological systems, but rather would like to re-design specialized DNA sequences based on learned biological knowledge to realize certain biological functions. Synthetic Biology is a discipline that studies how to design and synthesize new DNA sequences.

In the development of Synthetic Biology, how to *de novo* synthesize target DNA sequences of any length in a highly-efficient and inexpensive manner has always been a principal challenge. There are two major types of *de novo* synthesis: enzymatic synthesis and chemical synthesis, the latter of which has been used more commonly because it is relatively easier and cheaper. Chemical synthesis produces long-chain molecules through chemical reactions of simple chemicals, such as nucleotides or small precursor molecules. Currently, commonly used chemical approaches include phosphoramidite method and modified phosphoramidite-based methods [98]. Theoretically, it is possible to synthesize nucleotides of any length through chemical approaches, but practically, the usual synthesized length is just 150 bases because of the presence of secondary reactions. The product would be in very low amount if it exceeds such a length [99]. Therefore, to synthesize long-chain DNA, it usually starts from the synthesis of short oligonucleotides (oligos), then assembles them into long target DNA sequences through DNA enzymes, such as DNA ligase or polymerase, based on the overlapping regions among oligos, either *in vitro* or *in vivo* [100]. Using this method， Gibson and collaborators synthesized, assembled and cloned a new *Mycoplasma genitalium* genome, named *M. genitalium* JCVI-1.0, with 582,970 bps [101].

Similar as DNA sequencing technology, DNA synthesis technology has also begun to evolve toward the direction of high-throughput parallel experiments to promote the synthesis production and reduce the cost. Because of the advancement in the manufacturing of microarrays, scientists started to perform oligos synthesis on Microarray [102].

Since it is impossible for chemical reactions to reach 100% efficiency during chemical synthesis, errors are likely to be introduced in the oligos, such as substitution, insertion, and deletion of bases. For a single oligo, such a problem does not have significant influence. However, for the DNA sequence assembled from multiple oligos, the impact of error synthesis grows exponentially as the number of oligos increase. Compared to the error rate of 10^{-7} to 10^{-8} in the DNA replication in biological systems [103], the error rates of current chemical synthesis method are between 10^{-2} and 10^{-3} [104]. In the high-throughput microarray-based DNA synthesis, the error rates are usually even higher than that of classical chemical methods due to the large number of oligos. For the presence of non-target DNA sequences which include synthesis errors, we usually select target DNA sequences using cloning and sequencing method, the cost of which accounts for a significant part of the overall expense of DNA synthesis. Therefore, to increase the efficiency of DNA synthesis and lower the cost, it is a very important step to eliminate error-containing oligos in DNA synthesis technology.

There are different methods to eliminate error-containing oligos, such as High Performance Liquid Chromatography (HPLC) [105], polyacrylamide gel electrophoresis

(PAGE) [106]. Among these various methods, one eliminates erroneous oligos by taking advantage of a mismatch-binding proteins (MutS) that is able to recognize and bind to several types of DNA mismatching. During the DNA annealing process, error-containing oligos will complement with error-free oligos to form mismatch-containing double-stranded DNAs. MutS will bind to these mismatch-containing double-stranded DNAs and thus capture and eliminate them (Figure 3-1). Compared to other error correction methods, such screening method with MutS has wider application range, more reasonable price, and higher efficiency.

Our collaborators established a MutS-based, simple, inexpensive, and high-throughput error-correction method. They solidified fusion expression of Cellulose binding module (CBM) and etMutS, a mixture of *Escherichia coli* MutS (*Eco*MutS) and *Thermus aquaticus* MutS protein (*Taq*MutS), onto a cellulose gel column to construct an etMutS immobilized cellulose column (etMICC), which can eliminate synthetic oligos including errors that are introduced in the process of assembling and microarray-based synthesis [107].

To evaluate the error-removal efficiency of etMICC, our collaborator randomly selected single clone for sequencing, which is much expensive, time consuming and very random. In this chapter, a high-throughput sequencing method was adopted to assess the oligos pre- and post-etMICC. The result demonstrates that etMICC can reduce error rate from 1.69% to 0.28%. Moreover, by analyzing the data using NGS analysis, it is discovered that etMICC has higher efficiency in removing gap errors caused by insertion and deletion than eliminating mismatch errors resulting from substitution.

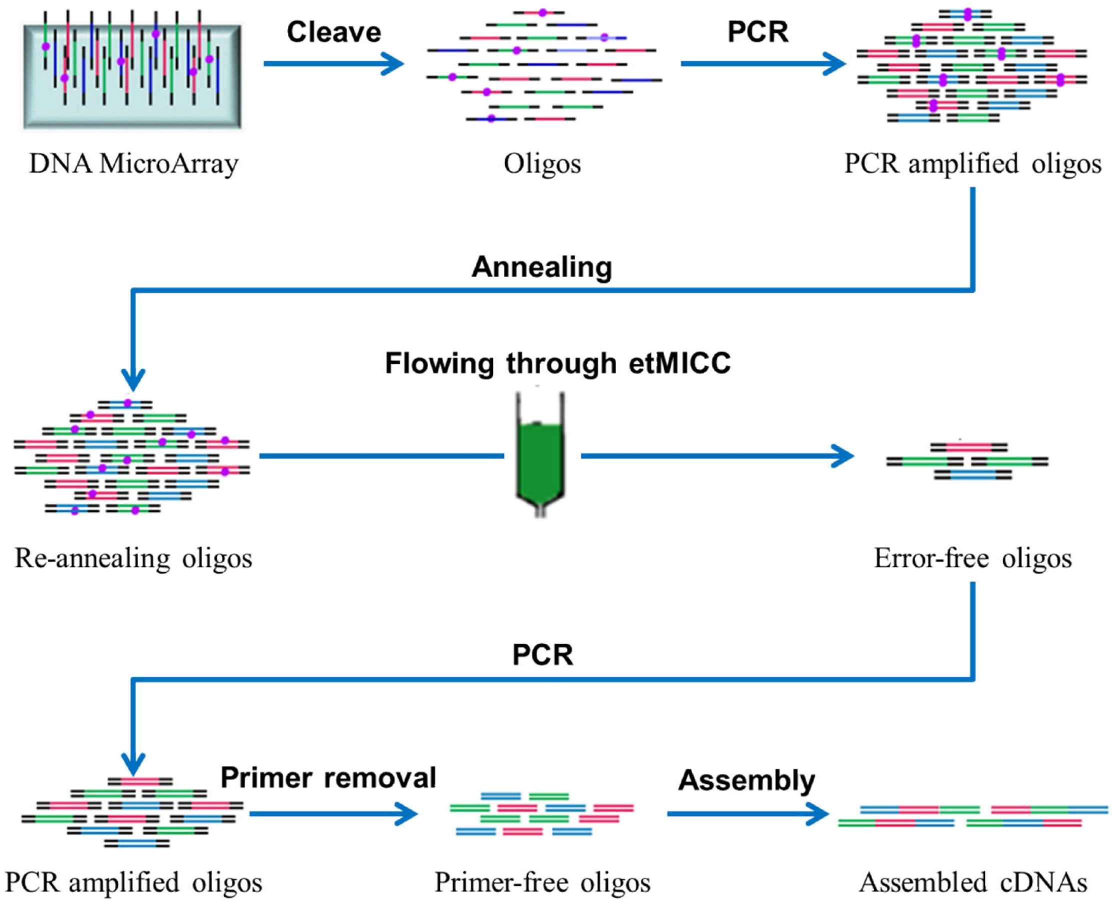


Figure 3-1. The schematic representation of error-containing oligos removal by using etMICC.

3.2 Materials and Methods

3.2.1 Materials

To study the property of etMICC-based error-correction method in eliminating error-containing oligos from DNA Microarray-based synthesis oligo library, we select the oligos before and after-process of etMICC as the sequencing sample.

3.2.1.1 LWFP Oligo Library Design

Genes for long wave fluorescent protein (LWFP) refer to red and far-red fluorescent protein (RFP and FRFP) genes. RFP can be achieved through expressing synthesized LWFP DNA sequences. Expression clones with RFP will be easier detected with naked eyes by introducing into red fluorescence. For LWFP oligo library design, 21 selected LWFP template genes consisting of 666 ~ 714 bases are truncated at homologous regions, which are found by alignment, to produce short oligos with length between 21 ~ 90, and total 399 unique oligos are obtained. The steps for cutting off gene sequences are shown in Figure 3-2.

3.2.1.2 Microarray-based Synthesis

LWFP oligo library was synthesized by using μParaflo microfluidic microarray technology in LC Sciences.

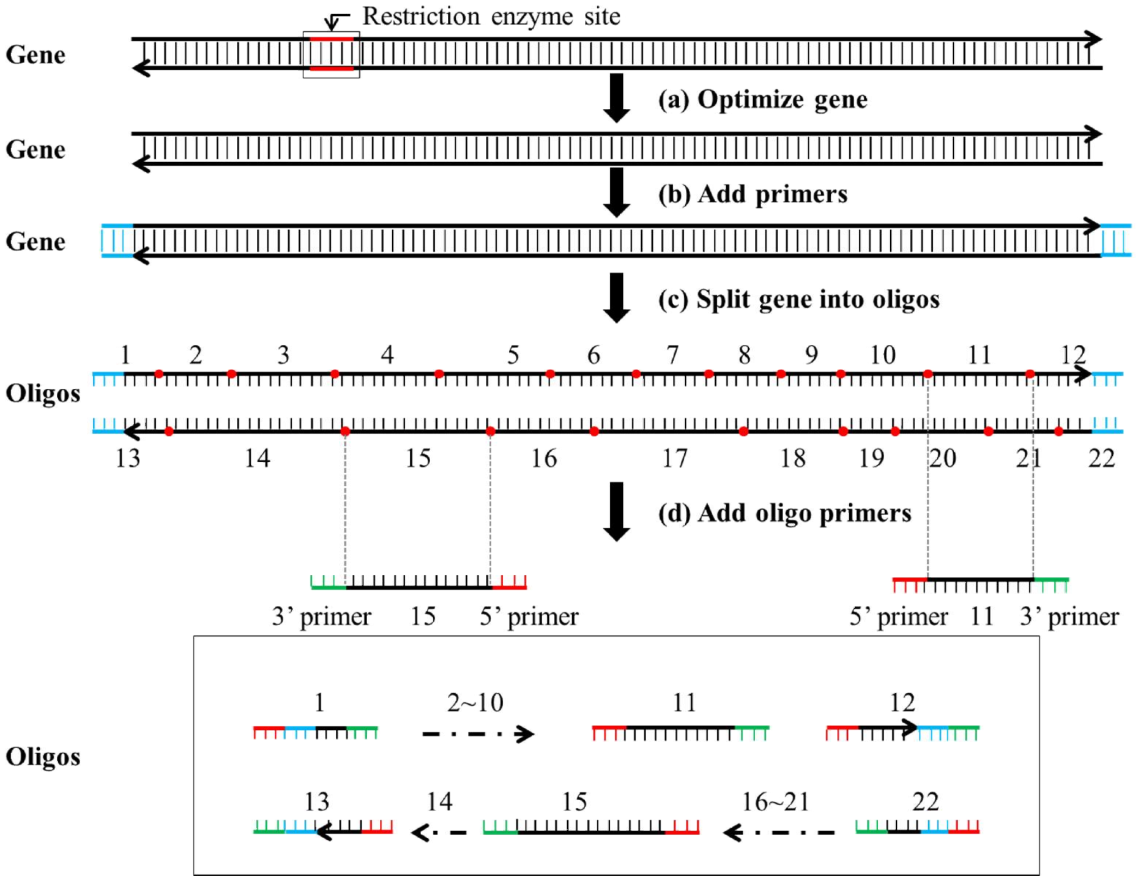


Figure 3-2. Flowchart of cutting off LWFP gene sequences to oligos. (a) Optimize gene sequences for exclusion of restriction enzymes sites; (b) Add primers for full length gene sequence amplication after assembly; (c) Split gene into oligos in homologous regions; (d) Add oligo primers for oligo amplications.

3.2.1.3 PCR Amplification

Microarray-eluted oligos are amplified in 30 PCR cycles thus purified with UNIQ-10 Oligonucleotide Cleanup Kit.

3.2.1.4 Error Removal Using etMICC

Re-anneal oligo sample to expose errors and thus load the oligo sample into etMICC to collect 18 filtrates.

3.2.1.5 PAGE Gel Detection of the Filtrates

Since the concentration of the oligos was too low to detect, they were re-amplified by PCR then the filtrates were detected with PAGE Gel. After the detection, filtrate 9 (F-9) and filtrate 10 (F-10) showed the brightest expected band in the detection, therefore, both F-9, F-10 and untreated sample were selected for further high-throughput sequencing.

3.2.1.6 High-throughput Sequencing

Before sequencing the sample with Illumina Sequencer in LC Sciences, four sequential steps were conducted: (1) Applying one more PCR amplification; (2) Removing primers from oligos by MlyI digestion; (3) Attaching a single adenine base (A) to oligos 3' end for adaptor ligation with the help of Taq DNA polymerase; (4) Adding "GATCGGAAGAGCACACGTCT", a genome DNA adaptor from Illumina, to the adenine base which is added in step three.

3.2.2 Getting Raw Sequencing Data

Three raw data files consisting of numerous 56-base reads in Sanger FASTQ format were obtained by using Illumina's Genome AnalyzerIIX System.

3.2.3 Read Filter

"Read Filter" refers to removing unqualified data prior to alignment, which includes "read trimming" and "quality filter".

3.2.3.1 Read Trimming

Because a single adenine and 3' end "GATCGGAAGAGCACACGTCT" adaptor (A-adaptor) were introduced into oligos during sample preparation, it is necessary to get rid of them before sequencing alignment. The process of eliminating A-adaptors includes two steps: determining the beginning of A-adaptors by traversing all reads in raw data files with seed sequence "AGATCGGA" which is the first eight bases of an A-adaptor, thus removing A-adaptor and sequences after A-adaptor from each read. Since each short referent oligo with 21 bases in length allows 5 errors, the new reads with the length less than 16 will be removed.

3.2.3.2 Quality Control

To study the correction capability of etMICC-based error-correction tool on different synthetic errors, it is very important to ensure that mismatches largely come from synthesis, rather than from sequencing. Therefore, it seems reasonable to make each base of the reads have a quality score higher than 30, which means the probability of incorrect ball call is smaller than 0.001. However, only about 50% reads passed the threshold 30 among the raw data. To achieve enough data for further analysis on the premise of each read with high quality score, a two-step disqualified sequence removal method is adopted: mapping the reads with average quality score no less than 30 to the referent sequence thus removing mapped reads which include at least one mismatch site with quality score less than 30.

3.2.4 Sequence Alignment

Unlike genome sequencing, oligo samples do not have to be broken up into fragments before sequencing since the sequencing capability of Illumina Sequencer covers the length range of referent oligos (21~90). Because oligo samples starts from 5' end, the mapping should begin from the 5' end of referent oligos. It is known that current high-throughput sequencing alignment tools usually map short reads to long referent sequences, and none of them can align short reads to short referent sequences. In the face of such problem, a Bowtie2-based analytical method for sequence alignment and data statistic was proposed in our study. In this method, Bowtie2 is used to search for all possible mapping results, which is followed by evaluating the mapping results through counting the number of error bases, partial of which are defined based on mapping positions.

Bowtie2 is a sequence alignment tool, which is good at fast mapping high-throughput short fragments to long DNA sequence, in particular the whole genomic sequence, and also supports gapped alignment. To improve the operating efficiency of Bowtie2, referent oligos and their reverse complementary sequences (RC-Seqs) are combined into a long DNA sequence by poly "N" sequences. During the process of combination, three points should be followed: (1) The length of "N" should be longer than 10; (2) Different poly "N" sequences do not have to be the same length; (3) Poly "N" sequences should be added before the first and after the last referent sequences, respectively. After combination is completed, Bowtie2 will align short oligos to the long DNA reference including poly "N" to get all possible mapping results using very sensitive parameters. The parameters used are shown in Table 3-1.

Table 3-1. Parameters used in Bowtie2 alignment.

Parameter	Meaning	Reason
-a	Report all alignments	For next-step filter
-norc	Do not align RC sequences	RC-Seqs are already added
--end-to-end	Global alignment	Entire read must align
-D 6	Give up extending after 6 failed	The max error allowed is 5
-R 3	Try 3 sets of seeds	More possible mapping
-N 1	Allow 1 error in seed alignment	More possible mapping
-L 16	Length of seed is 16	The minimum length is 21
-i S,1,0.50	Very-sensitive	More possible mapping

After the process of mapping, it is to select best mapping results from all the possible mapping results by counting the number of substitutions, insertions and deletions. During the process of counting, several points should be noted: (1) Because primers are added to pre- and post-oligos for using etMICC to remove error-containing sequences, the insertions and deletions before and after oligos will result in gaps, thus, such insertions and deletions should be counted. (2) Insertions which result from the remaining A-adaptor (the first eight base of “A” and adaptor) will be trimmed. (3) If the reads contain such a substitution site whose quality sore is lower than 30, such reads will be removed. (4) If one read mapped to n sites on referent sequence, it will be compared to 1/n read mapped to n referent oligos.

3.2.5 Further Data Analysis

Besides the methods of sequencing alignment, mapping result collection, as well as error counting, there are several other approaches for further analyzing previous results, which will be introduced in Section 3.3.

3.3 Results and Discussions

3.3.1 NGS Raw Data

The statistics of the raw sequencing data are show in Table 3-2.

Table 3-2. Statistics of raw sequencing data.

	Reads	Bases	Quality Score
Untreated	30,315,468	2,303,975,568	31.4
F-9	25,353,391	1,926,857,716	31.1
F-10	15,638,422	875,751,632	33.8

3.3.2 Read Filter

The statistics of the data after read filter are shown in Table 3-3. The comparison between filtered and raw data is shown in Table 3-4.

Table 3-4 reveals that the read filter not only ensured high quality score for sequences, but also guaranteed more than 80% sequences retained. In addition, approximatively 50% bases are removed, which means “read trimming” is extremely necessary.

Table 3-3. Statistics of filtered data.

	Reads	Bases	Average Score
Untreated	26,970,025	1,241,075,151	37.4
F-9	23,329,865	900,660,792	37.5
F-10	13,993,448	500,386,622	37.8

Table 3-4. Comparison between filtered data and raw data (Filtered/Raw).

Sample	Percentage of Reads	Percentage of Bases	Fold of Score
Untreated	88.9%	53.9%	1.19
F-9	92.0%	46.7%	1.20
F-10	89.5%	57.1%	1.12

3.3.3 Sequence Alignment

Sequence alignment consists of two sequential steps: mapping oligos to referent sequence and selecting best results by counting number of errors in each mapped result. After the second step, there were 14.9%, 9.0% and 12.9% of reads were removed from untreated, F-9 and F-10 samples, respectively (Table 3-5).

Table 3-5. The statistic of unqualified read filtered by sequence alignment.

Sample	Unqualified Reads	Percentage in Raw Data	Total Unqualified Reads
Untreated	1,181,664	3.9%	14.9%
F-9	263,493	1.0%	9.0%
F-10	368,848	2.4%	12.9%

After sequence alignment, the percentages of mappable reads are shown in Table 3-6, and the distribution of mapping reads in filtered data is revealed in (Figure 3-3). Figure 3-3 demonstrates that the percentage of perfect match reads after etMICC treatment is remarkably higher than that of untreated reads, and the percentage of perfect match reads in F-10 even reaches up to 91.1%.

Table 3-6. The statistic of mappable read.

Sample	Mappable Reads	Percentage in Filtered Data
Untreated	18,258,469	70.8%
F-9	15,654,864	67.9%
F-10	13,182,640	96.8%

Table 3-7 reveals that at the cost of losing partial error-free oligos, etMICC effectively removed error-containing oligos.

Table 3-7. The statistic of mapped referent oligos.

Sample	Mapped Referent Oligos	Covered Percentage
Untreated	399	100.0%
F-9	342	85.7%
F-10	362	90.7%

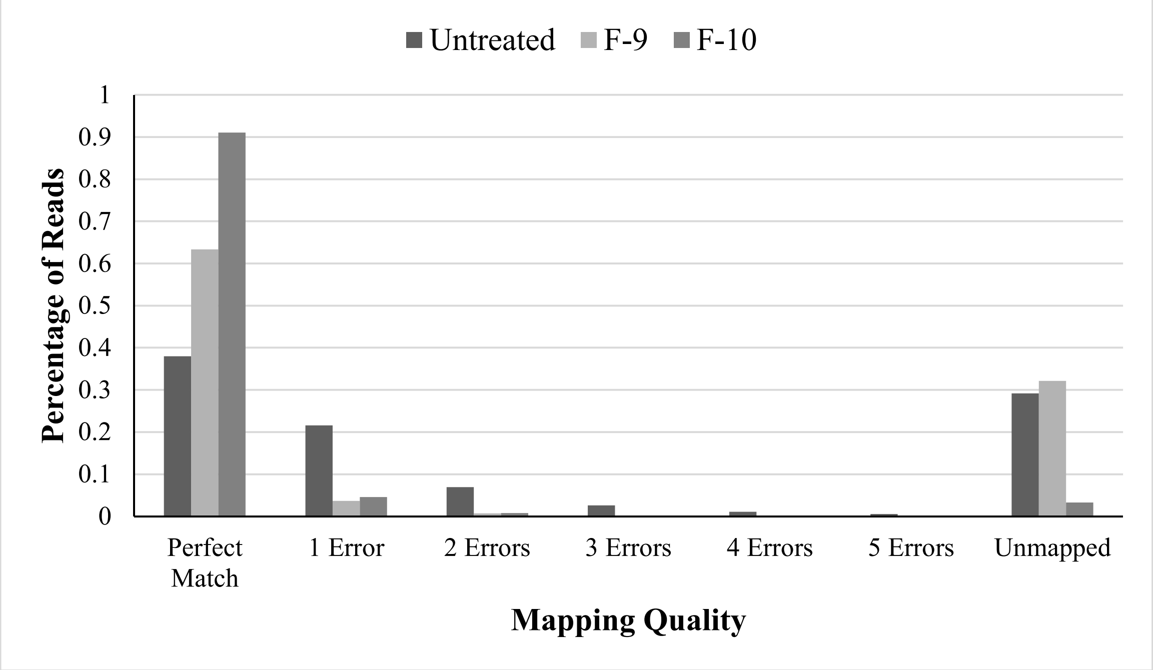


Figure 3-3. The statistics of perfect match and different error types of untreated sample, F-9 and F-10.

3.3.4 Unmapped Referent Oligos Analysis

To further analyze the reasons of the loss of partial referent oligos, mapped oligos and unmapped oligos in F-9 and F-10 are compared by using unpaired one-tailed Student's T-Test, respectively. The result of comparison shows that the average length of unmapped oligos is significantly longer than that of mapped oligos. The results of T-Test are displayed in Table 3-8.

Table 3-8. The results of Student's T-Test for length of mapped and unmapped oligos.

Sample	Length(Mapped)	Length(Unmapped)	T-Test P-Value	-log10 of P-Value
F-9	54.4 (342)	71.6 (57)	2.9×10^{-23}	22.5
F-10	55.4 (362)	70.9 (37)	4.4×10^{-12}	12.4

Furthermore, F9 has 20 more unique unmapped referent oligos than F10. One possible reason leading to the difference is that oligos with different lengths pass etMICC at different speeds. Thus, longer oligos likely need more time to get through the etMICC. Therefore, it is probably reasonable to improve the quality of error-free oligos by optimizing the time for samples passing through etMICC and the quantity of collected filtrate.

3.3.5 Error Ratio Analysis

Based on the sequences in mapping result, the average quantity of different errors in each sample and the folds of Untreated/F-9 and Untreated/F-10 are calculated, respectively, which are shown in Table 3-9. Table 3-9 shows that etMICC reduces the total error rate from 1.7% to 0.2% (7.8 fold) by comparing Untreated with F-10.

Table 3-9. Different errors in each sample and fold of Untreated/F-9 and Untreated/F-10.

Sample	Total	Insertion	Deletion	Substitution
Untreated	1.7%	0.2%	0.5%	0.9%
F-9	0.3%	0.0%	0.1%	0.2%
F-10	0.2%	0.0%	0.1%	0.1%
Untreated/F-9	6.0	9.4	5.6	5.6
Untreated/F-10	7.8	11.0	6.5	8.2

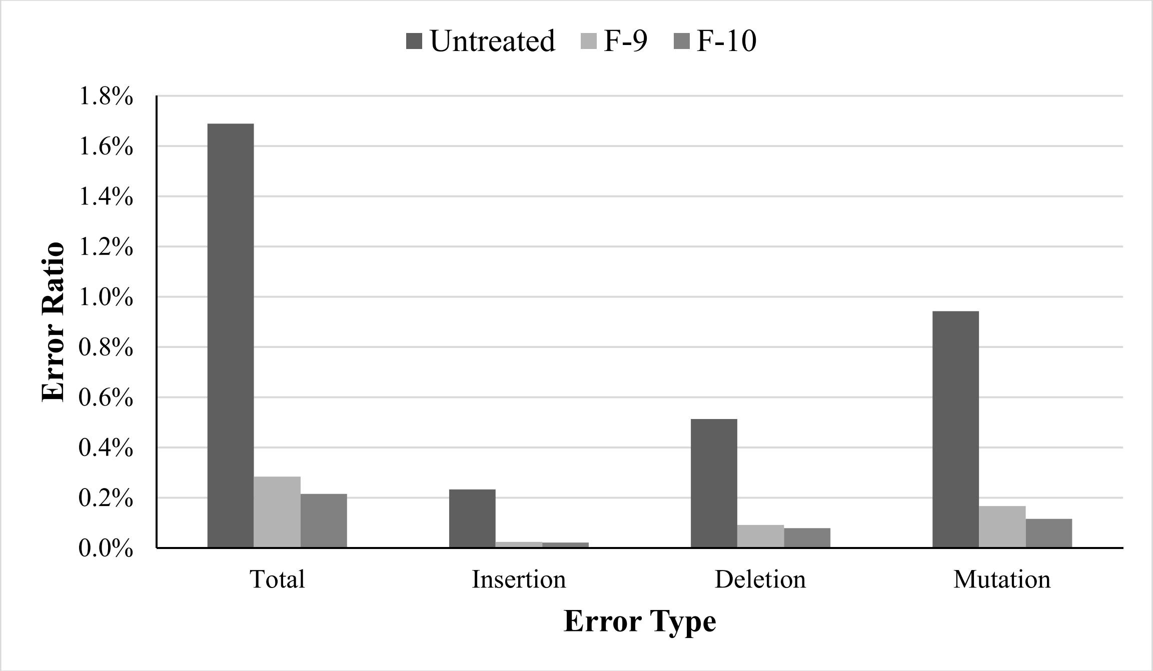


Figure 3-4. The histogram of ratio for different error types in each sample.

3.3.5 etMICC Function Analysis

Comparing the number of error-containing oligos before and after etMICC is one of the feasible methods to assess the binding affinity between different error types and etMICC. Error counting is based on the reads only mapping to referent oligos which include up to 56 (max length of read) bases. In addition, each error-containing read may include multiple error types and etMICC may not react with all of the error sites when removing such oligos, therefore, to simplify the experimental conditions, I only focus on one-error-containing reads.

Table 3-10. The statistics of one-error-containing reads in different samples.

Sample	Total	Insertion	Deletion	Substitution
Untreated	4,380,193	485,974	874,622	3,019,597
F-9	838,128	23,332	128,437	838,128
F-10	609,677	20,558	127,716	534,296

Table 3-11. The statistics of fold Untreated/F9 and Untreated/F10 based on one-error-containing reads in different samples.

Sample	Insertion	Deletion	Substitution
Untreated/F9	18.6	6.1	3.2
Untreated/F10	12.5	3.6	3.0

Summary of one-error-containing reads in three samples is shown in Table 3-10 and Table 3-12. To further analyze diverse substitutions, insertions and deletions in different samples, three different error types are divided into eight substitution errors and four gap errors. The gap errors integrate both insertion and deletion, because insertion and deletion cannot be differentiated by etMICC. Result of subdivided errors is displayed in Table 3-12, and the folds of Untreated/F9 and Untreated/F10 on the basis of the information in Table 3-11 is shown in Figure 3-5. From Table 3-12, it is clear that etMICC is more efficient in filtering insertion than others, and Table 3-14 reveals that subdivision errors of substitution A-G, as well as G Gap and A Gap have higher probability to be recognized by etMICC.

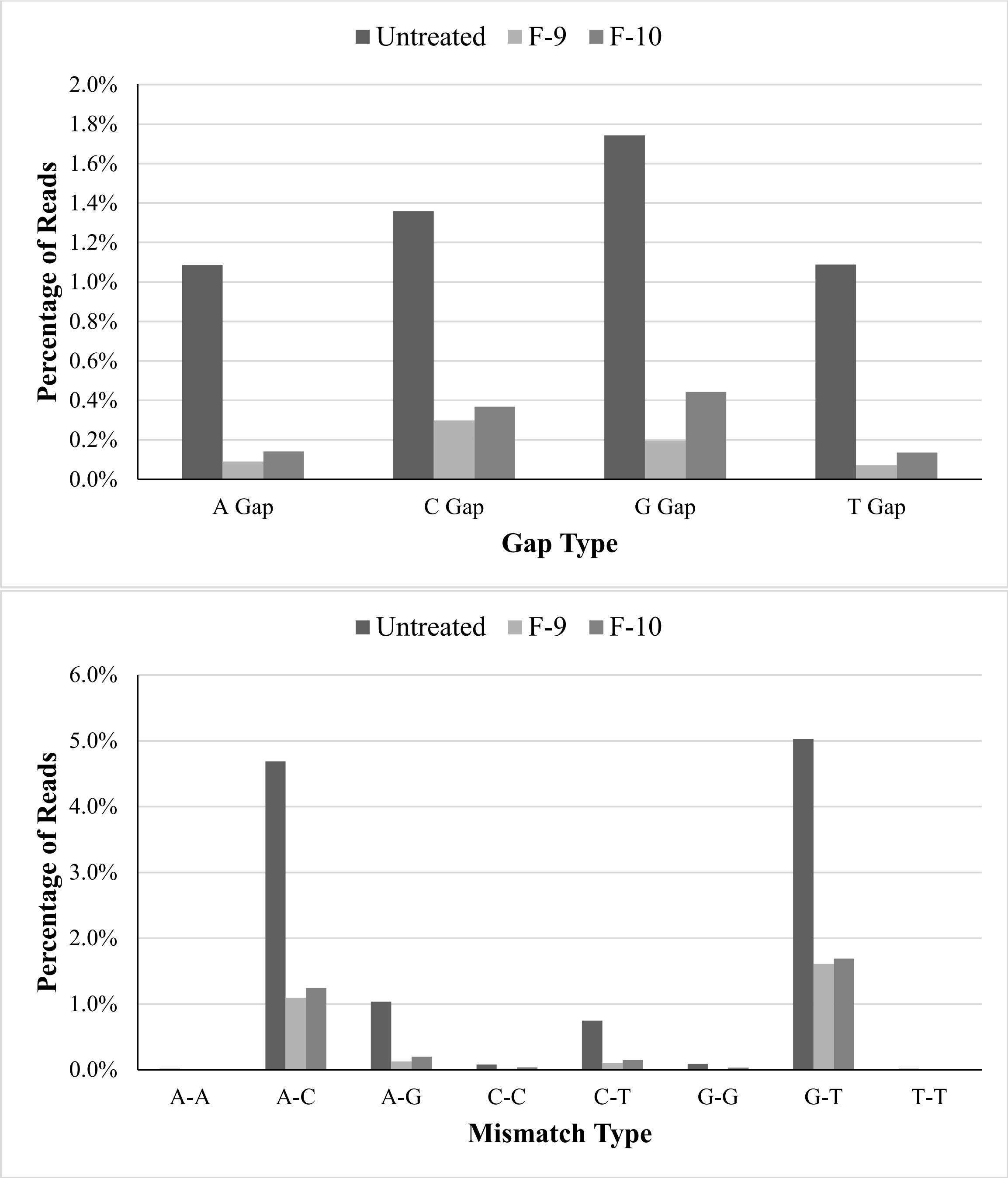


Figure 3-5. The statistics of gap errors and mismatch errors of untreated sample, F-9 and F-10. (a) Percentage of reads for different gap types. (b) Percentage of reads for different mismatch types.

Table 3-12. Comparison of error removal between filtrates and untreated sample.

Type	Error Type	Untreated/F-9	Untreated/F-10
A Gap	ins A or del T *	12.0	7.7
C Gap	ins C or del G *	4.5	3.7
T Gap	ins T or del A *	8.9	3.9
G Gap	ins G or del C *	15.0	8.0
All Gap		8.0	4.9
A-A	T->A	3.2	1.2
A-C	G->A or T->C	4.3	3.8
A-G	C->A or T->G	8.1	5.2
C-C	G->C	5.2	2.1
C-T	A->C or G->T	7.1	5.0
G-G	C->G	7.8	2.7
G-T	A->G or C->T	3.1	3,0
T-T	A->T	3.6	1.2
All Mismatch		3.9	3.5
Total		4.7	3.8

3.4 Conclusions

In this chapter, the property of etMICC-based error-containing method in correcting microarray-based synthetic oligos is analyzed by comparing read sequences before and after etMICC. To obtain enough high-quality reads for analysis, a two-step quality control method is adopted, which includes the step of eliminating oligos with average quality score lower than 30 and the following step of removing oligos in which at least one substitution site with quality score lower than 30. After the process of quality control, more than 80% high-quality reads are collected. Moreover, since referent oligos are 21-91 in length and none existing alignment method is efficient at mapping high-throughput sequencing data of NGS to short referent oligos, a Bowtie2-based sequence alignment approach is developed, and owing to such method about 50% reads successfully mapped to target oligos.

In addition, by comparing high-quality oligos before and after etMICC, it reveals that at the expense of losing partial long sequences, etMICC significantly increases the percentage of perfect match oligos from 38.0% to 91.1%, and decreases ratio of error-containing oligos from 1.7% to 0.3%. Moreover, through studying binding affinity of etMICC and one-error-containing oligos, it is demonstrated that etMICC has higher error-removal efficiency for gap error structure resulting from insertion and deletion than for substitution error structure which are caused by substitution. Furthermore, it is evident that the NGS analysis method is appropriate to help select best filtrates for following assembly based on quality evaluation of different filtrates, and is also able to assess the efficiencies and properties of other error-correction methods.

Chapter 4

Development of Proteomic Tools for Investigation of

Cellular Protein Functions

and Post-translational Modifications (PTMs)

4.1 Introduction

Histones are a series of alkaline proteins existing in eukaryotic cell nuclei, which include four core histones H2A, H2B, H3, H4 and two linker histones H1 and H5 [108]. An octamer containing two copies of each H2A, H2B, H3, and H4 wrapped by a segment of DNA forms a nucleosome, which is the basic unit of chromatin. The alternative structure and function of chromatin usually result from diverse post-translational modifications (PTMs) of histones. Among a myriad of PTMs, histone methylation catalyzed by specific enzymes has been increasingly recognized as players responsible for a major signaling mechanism in eukaryotic cells.

Histone methylation is a process in which histone methyltransferases (HMTs) transfer methyl groups (CH3) from S-Adenosyl methionine (SAM) onto lysine or arginine residues of histones [109]. Surprisingly, histone lysine methylation is much more common than histone arginine methylation. It is studied that lysine is capable to be mono-, bi-, or tri-methylated via substituting hydrogen (H) of NH3 group with methyl groups (Figure 4-1) [110], while arginine can only be mono- or bi-methylated [111]. For example, H3K37 can be methylated to H3K37me1, H3K37me2 and H3K37me3 while H3R17 can be only methylated to H3R17me1 and H3R17me2 [112].

HMT is a kind of enzyme catalyzing the histone methylation by transferring methyl groups to residues of histone proteins. HMTs consist of two major types: lysine-specific and arginine-specific. Lysine-specific HMTs (KHMTs) can be divided into SET (Suppressor of variegation, Enhancer of Zeste, Trithorax) domain proteins and no-SET proteins, for example MLL4 for H3K4 contains SET domain and DOT1L for H3K79 does not [113].

Several lines of evidence ensure that histone methylation plays essential roles in epigenetic alternations, which can either activate or repress gene expression. For example, H3K4me2/3 activates gene transcription while H3K9me2/3 inactivates gene expression [114]. With the first discovery of histone demethylases (HDMs), the fact that histone methylation with revisable nature starts to be accepted [115]. The regulation will become complicated if both HMT and HDM appear in the common complex. It is known that mis-regulation of HMTs is associated with diverse diseases. For example, EZH2 a transcriptional repressor has been perceived as a biomarker for prostate cancer [116]. Therefore, as a member of the suite of epigenetic modifiers HMTs are a new and promising class of therapeutic targets.

In cancer, there is a growing body of evidence suggesting that changes in the activity of HMTs contributing to gene expression activation and uncontrolled cell proliferation are hallmarks of these devastating diseases. To identify the activity of HMTs *in vivo*, the traditional method is: (1) extract histone fractions from cell nucleus by trichloroacetic acid (TCA) extraction or high-salt extraction; (2) make use of specific anti-PTM histone antibodies, for example, anti-H3K36me1, to analyze particular methylated sites on histone proteins [117,118]. For the conventional method, it is feasible to investigate a particular PTM of a single site with an appropriate antibody, however, a variety of anti-PTM histone antibodies are required if the goal is to assay activities of all HMTs in the whole nuclear lysate, which results in not only more expensive, but also more difficult detection of PTMs in each site due to different impacts of distinct anti-PTM histone antibodies [119]. Therefore, we designed a histone peptide microarray (hPepArray) platform to analyze activities of nuclear KHMTs *in vitro* by using anti-pan methyl lysine antibodies.

In this study, we designed and synthesized a serial of histone peptides on a microfluidic chip, which contains 160 unique peptide groups covering five histone proteins, such as H1, H2A, H2B, H3 and H4. Each unique peptide group includes one WT peptide directly extracted from histone protein and several mutant peptides and control peptides were used to make sure the methylation was caused by the target lysine residue or not. Moreover, we obtained nuclear extract from a breast cancer cell line – T47D, and applied the protein lysates to hPepArray. After incubation hPepArray was loaded with the nuclear extract and the general methylated lysine antibody, significant signals were detected at the sites represented by peptides corresponding to H2AK74, H3K122, and H4K59, while null signal was found at corresponding mutant peptides and control peptides.

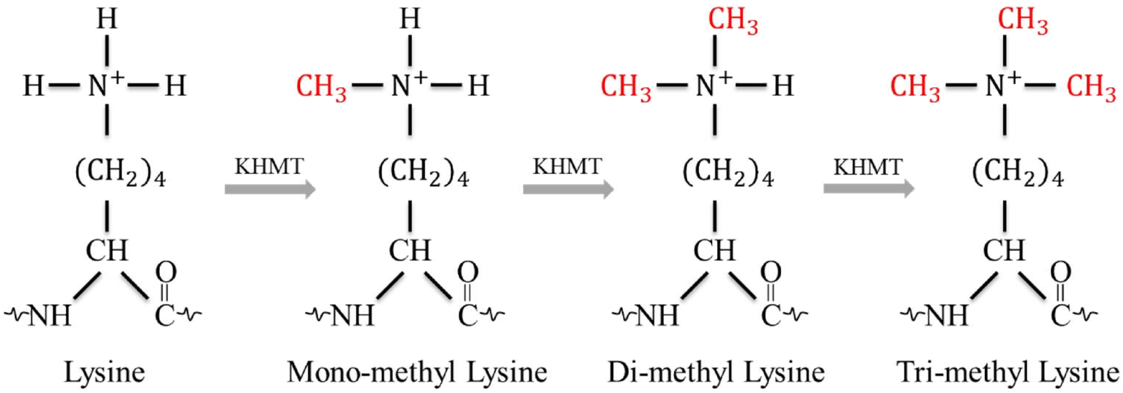


Figure 4-1. Histone lysine methylation states catalyzed by lysine histone methyltransferases (KHMT).

4.2 Materials and Methods

4.2.1 Histone PepArray chip probe design

The studies of sequence specificity of the substrates of HMTs with spotted peptide arrays [120,121] demonstrated that histone methylations are sequence dependent. We assume sequences with nine residues can be recognized by specific HMTs. To simplify the experiment, only lysine-specific probes are designed to focus on activities of lysine-specific HMTs. Moreover, to assay as many KHMTs as possible, the probes were extensively derived from ten histone protein sequences, including core histones H2A, H2B, H3 and H4 from NCBI RefSeq database, and liner protein H1 [122]. Each nine-residue peptide consists of a lysine residue in the center and four residues on each side. Counting all lysine residues in ten histone proteins, 160 unique peptides in total were generated (Table 4-1).

Total 160 unique peptides directly extracted from histones are named as WT peptides, in which the central lysine residues are called core lysine. Each WT peptide has one corresponding negative control, named as A control, in which the core lysine (K) is mutated into alanine (A). Because nuclear lysate employed in hPepArray contains not only KHMTs, but also other PTM enzymes which may impede methylation by interacting with peptides primarily, it is necessary to remove the impact of other PTM enzymes. To simplify the experiment, for each WT peptide, we eliminated effects of phosphorylation enzymes and histone acetyltransferases (HATs) by mutating serine (S) and threonine (T) and tyrosine (Y) (three phosphorylation sites) into alanine and substituting lysine and arginine residues on both sides of the core lysine which could be either acetylated or methylated with similar size residue glutamine, respectively. These newly designed peptides with mutant residues on both sides of the core lysine are called Side Mutation (SM) peptides. However, certain WT peptides, such as H1.1-K16 and H3.3-K27 and H3.3C-K27, without residues like S, T, Y, K and R on each side of the core lysine have no SM peptides. In addition, negative control peptide for each SM peptide, named SM A control, is designed by replacing the core lysine with alanine. Entire 634 unique peptides including WT peptides, WT A control peptides, SM peptides and SM A control peptides are designed. Further, 634 unique peptides will be synthesized on an hPepArray in which each peptide has one or two repeats.

Table 4-1. Statistics of reference peptides.

Source	Number of reference peptides
H1.1	57
H1t	40
H2A	14
H2B	20
H3	13
H3.1	13
H3.1t	13
H3.3	13
H3.3C	11
H4	11
Total unique peptides	160

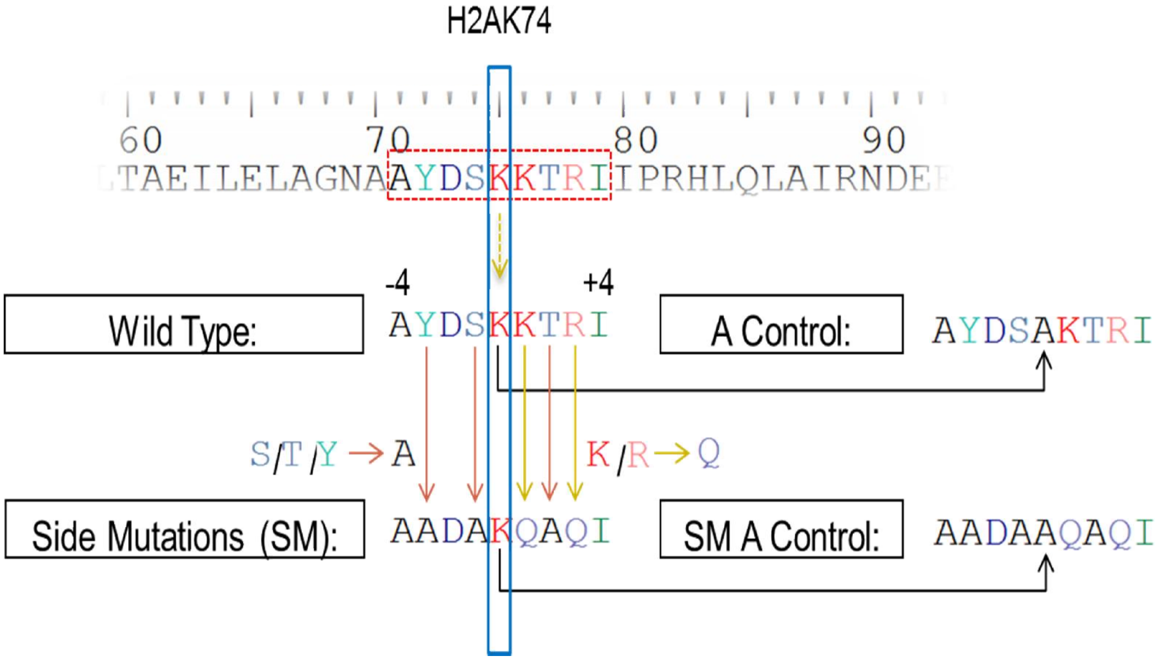


Figure 4-2. Strategy of peptide design. Blue frame: the core lysine; Red frame: the origin nine-residue peptide located in histone sequence named as Wild Type (WT); A Control: the negative control peptide for the corresponding WT peptide by mutating the core lysine (K) into alanine (A); Side Mutation (SM): for each WT peptide, the residues serine (S), threonine (T), tyrosine (Y) on both sides of the core lysine are replaced with alanine (A) and another two residues such as lysine (K), arginine (R) on each side of the core lysine are replaced with glutamine (Q); SM A Control: the negative control peptide for each SM peptide by substituting the core lysine (K) with alanine (A).

4.2.2 Experiments

4.2.2.1 Microarray-based Synthesis

hPepArray was generated by using µParaflo microfluidic microarray technology in LC Sciences, Houston, TX.

4.2.2.2 Nuclear Lysate Extraction

Nuclear lysate is extracted from a breast cancer cell line – T47D by using Abcam nuclear extraction kit according to vendor protocol.

4.2.2.3 Methylation and Immunoblotting

Nuclear lysate and SAM were loaded on the hPepArray and incubated at 37º C for 3 hours. Thus hPepArray was stained with anti-pan methylated lysine antibody from Abcam according to vendor protocol. After previous processes, the chip was scanned with Axon GenePix 4400A.

4.2.3 Data Analysis

4.2.3.1 Image Digitization

Digital density data was transferred from image data by using ArrayPro Analyzer.

4.2.3.2 Background Subtraction and Signal Significance Analysis

Reducing background and selecting detectable peptides whose signals are significantly higher than that of the background are processed according to the protocol of LC Sciences, Houston, TX [123].

4.3 Results and Discussions

4.3.1 Methylation detected on WT peptides

By comparing with A control peptides, three WT peptides corresponding to H2AK74, H3K122, and H4K59 with significantly higher signals have been identified (Figure 4-2). It is known that H3K122 can be either acetylated or methylated [124,125] and methylation of H4K59 plays an essential role in transcriptional silencing [126], while methylation of H2AK74 has not been reported. Although hPepArray is an *in vitro* experimental study that may not exactly demonstrate the activities *in vivo*, it is likely that H2A-K74 is a methylation site because partial histone lysine methylations catalyzed by specific HMTs are sequence dependent [111,112].

4.3.2 Methylation detected on SM peptides

By comparing with corresponding A Control peptides, six SM peptides in total with significantly higher signals have been determined. These SM peptides considered as potential methylation probes for further optimization of hPepArray are shown in Table 4-2.

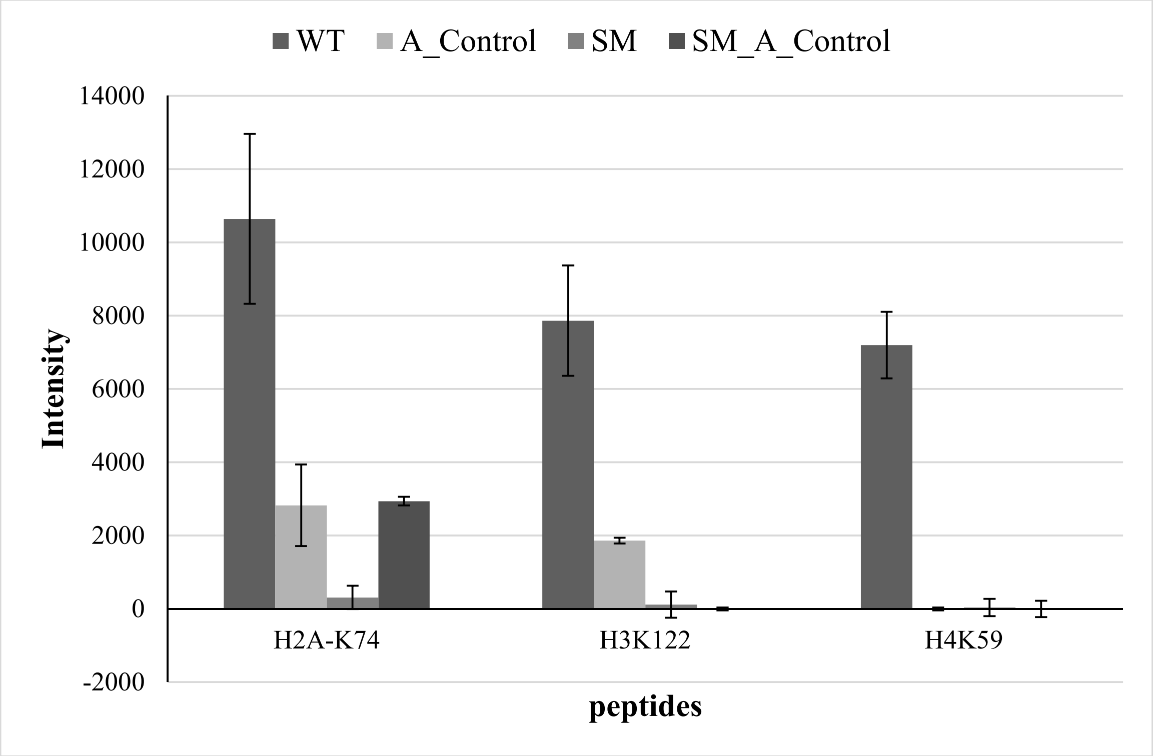


Figure 4-3. Intensity of peptides in three peptide groups corresponding to H2A-K74, H3K122 and H4K59. The higher of the intensity means the higher level of methylation.

Table 4-2. Six potential methylation probes.

Peptide ID	Sequence	Density (SM)	Density (SM A Control)
H1.1-K140_SM	QQLQKAAGA	8,091.2	0.0
H1.1-K178_SM	VQPQKVAQA	15,574.8	305.5
H1.1-K192_SM	AVQPKAAQA	13,068.4	87.6
H1t-K147_SM	PQAAKANQQ	15,657.9	515.8
H2B-K23_SM	AQAQKQDAQ	23,293.8	24.0
H4K5_SM	AGQGKGGQG	39,047.3	0.0

4.4 Conclusions

By comparing the signals of lysine peptides extracted from hPepArray, interestingly, three lysine positions – H2A-K74, H3K122, and H4K59 – have extremely higher signals than others. It is demonstrate that histone peptide microarrays are able to reveal the activity profiles of KHMTs at specific histone sites in cellular system.

It is studied that H3K122 and H4K59 can be methylated, while H2A-K74 is a novel methylation site *in vitro*. Therefore, further experiments are required to verify the methylation of H2A-K74 *in vivo*.

Moreover, it is published that H3K36 is methylated in breast cancer cell line [127], while methylation of H3K36 was not revealed in hPepArray. Thus, it is necessary to ensure that hPepArray is able to fully reflect the activity profiles of KHMTs and demonstrate the properties of the substrates of KHMTs by improving peptide design.

Currently, we only completed the study of histone methylation in T47D. Future studies will embrace histone methylation in different cell lines to identify the differentiations of KHMTs activity profiles between diverse cell lines.

Although the study of histone methylation described in this chapter is the first step of hPepArray development, the results indicate that it is feasible to learn the activity profiles of KHMTs in cellular system with hPepArray. In contrast to traditional methods for studying histone modifications, hPepArray not only allows high-throughput experiments, but also replaces site-specific anti-methylated lysine antibody with general anti-methylated lysine antibody, which highly reduced the experimental complexities and lowered the experimental costs. In addition, besides analyzing activity profiles of KHMTs, hPepArray can also be used to assay activity profiles of arginine-specific histone methyltransferases (RHMTs). Furthermore, aside from methylation, hPepArray is able to study the properties of other types of histone PTMs by employing specific antibodies and designing different peptides as is required.

While mass spectrometry-based proteomic approaches are only applied to proteomics quantitative profiling [128], hPepArray focuses on proteomics activity profiling which reflects real impacts on histone PTMs of certain proteins in nuclear lysate. In consequence, hPepArray can be adopted to select inhibitors of specific PTMs as targets for therapeutic development.

Chapter 5

Overall Conclusions and Future Directions

In the development of biological science and bio-technology, scientists start the research on large-scale biology 'omics' and systems biology with the help of advancement in high-throughput technologies, such as NGS and microarrays, which result in tremendous amount of data. Thus, bioinformatics for bio-data analysis is becoming an indispensable player in designing experiments and interpreting biological meanings embedded in the enormous amount of data. In this thesis, I developed such applications of bioinformatics in three different fields.

5.1 Conclusion and Future Development of miRFocus

In Chapter 2, I developed a comprehensive web-based resource 'miRFocus' for effectively retrieving and analyzing human miRNAs by combining knowledge of computer sciences with information in biology. By submitting a list of miRNAs obtained from miRNA expression experiments to the webpage of miRFocus, a wide range of information associated with query miRNAs, such as sequences, genomic clusters, co-expression miRNAs, diseases, publications, experimentally validated and predicted target genes, as well as enriched pathways and GO terms will be quickly provided in the result page. Moreover, miRFocus allows multiple miRNAs input through one or more different input methods, thus miRNA nomenclatures of diverse versions will be converted into the names of the latest version (miRBase version 21) automatically. Furthermore, miRFocus also implemented the function to query either target genes or KEGG pathways for searching miRNAs.

Although miRFocus offers detailed information relevant to interrelations between miRNAs and target genes and KEGG pathways, it does not demonstrate how certain miRNA affects the function of a KEGG pathway. For example, based on the information in miRFocus, it reveals that IL13, a gene of Asthma pathway, is the target gene of miRNA hsa-let-7f-5p, but it is not clear that whether hsa-let-7f-5p affects Asthma pathway by down-regulating or up-regulating IL13. Therefore, to incorporate such useful knowledge into miRFocus, in future, we have to collect much more data regarding this aspect from existing databases and published literatures. It seems that to curate such information manually is not reasonable; however, the emergence of text-mining technology which is used for deriving high-quality information from text makes it possible [129]. Moreover, since current text-mining technologies focus on the searching by key words, it is necessary to develop advanced algorithms for getting more detailed regulation information between miRNA and genes.

As for predicted target genes and enriched pathways, it is not easy to assess their reliability due to the lack of enough experimentally validated target genes. In this thesis, I choose the target genes predicted by at least three largely-used prediction web tools as the more reliable candidate targets, which are based on the analysis of databases embedded in miRFocus. To observe the more reliable enrichment pathways, some scientists imported some factors such as an observed to expected (O/E) miRNA expression ratio and tissue-specific expression signatures into the process of pathway analysis [89,130]. Even though they are meaningful attempts, there are still multiple questions without appropriate answers. For example, as for tissue-specific expression signatures, it assumes the low expression genes in specific tissues will be ignored in the process of pathway analysis, even though they are the targets of queried miRNAs. The question is that removal of the low expression genes may results in the loss of valuable information, since miRNAs are able to activate or inactivate certain pathways by down-regulating such target genes [131]. Although effective solutions to such problems have not come out, it is believed more precise methods for miRNA regulation pathway analysis will be developed as more experimentally validated target genes are identified. Therefore, in the absence of adequate experimentally validated target genes of known miRNAs for pathway analysis, it is necessary to develop advanced methods which are able to achieve enriched pathways by utilizing broader interactions of miRNAs and genes. For example, in the database of miRpathway, only 507 human miRNAs, approximately one-fifth of total human miRNAs, are included. To cover more extensive interrelations of miRNAs and genes, besides the experimentally validated interactions of miRNAs and target genes, I would like to incorporate predicted target genes of high reliability. To include much more interrelations of genes and pathways, I would like to take into account more genes which are associated with the genes included in current pathways by protein-protein interactions (PPI) from existing PPI databases such as UniProtKB [132] and HomoMINT [133]. The final goal of miRFocus is collecting comprehensive information of human miRNAs and genes and pathways, and to establish a bioinformatics tool similar as GPS, which can be used to search and illustrate/navigate the functions and interaction networks of each biological molecule in biological systems.

5.2 Conclusion and Future Development of Synthetic Oligo NGS Analysis

In Chapter 3, I developed an efficient NGS analysis for analyzing synthetic oligo library, which consists of millions of short DNA sequences with less than 100 nucleotides in length. Through two steps of quality control and subsequent Bowtie2-based alignment, this method successfully evaluated the error removal efficiency of etMICC-based error-removal method on microarray-based synthetic oligos of various lengths, and analyzed binding affinities between etMICC columns and different types error-containing oligo sequences. Moreover, since current NGS technologies are mainly applied to large scale genomic or transcriptomic sequencing and none of existing NGS methods is suitable for synthetic oligo library analysis, the approach I established is to fill the gap, and it is much useful for scientists who are interested in such field to assay synthetic oligos.

The newly designed method for synthetic oligo NGS analysis in my thesis can be improved in several aspects. Firstly, it is necessary to speed up the running rate of the analysis method, which is because in most cases it usually takes about six hours to process a library of 30 million reads (6GB data size) by a computer with 2.4 GHz and 8GB ram. Secondly, to improve the newly designed NGS analysis method, it has to reduce the size of temporary files generated during the process of alignment, since up to 25GB temporary files were created when aligning 6GB raw data to referent sequences by using sensitive parameters of Bowtie2. Developing a novel sequencing alignment method for synthetic oligo library analysis is one feasible solution to improve running efficiency and save operating spaces. In the novel alignment method, first, I would like to index the short length referent oligo sequences based on existing index methods, such as Burrow-Wheeler Transform (BWT), instead of combining them into a long referent sequence, which will reduce the operating steps and running time; second, I would like to integrate the two-step alignment employed in current method into one-step alignment to reduce the size of temporary files and operating time.

In the development of NGS technologies, it is believed that sequencing technologies will become faster, much less expensive and more accurate for obtaining DNA sequence information, which in turn will lead to extensive usage of NGS technologies and thus enormous volume of sequencing results is going to be generated. Therefore how to properly transmit and analyze such large data set emerged as a tough problem. It is known that the ongoing 1000 Genomes project aims to build the most detailed catalogue of human genetic variation, which has cumulatively yielded approximately 233 TB of sequencing data (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data) [134]. Since this project incorporates the efforts of different research laboratories around the world, it requires considerable network resources for data transmission and analysis. To implement large amount of resource sharing, cloud computing emerged [135]. Cloud computing refers to computing which integrates large number of remote servers for data storage and allows users to access and analyze data stored in various resources through a remote registry service. Due to cloud computing-based data sharing, storage space is largely saved, cost is efficiently reduced, and high operation capability is achieved. However, most of existing NGS analysis methods have not been optimized in terms of cloud computing. Therefore, the future work is to improve the capability of portability and parallel computing of NGS analysis method by making use of existing cloud computing infrastructure, such as CloudBioLinux [136].

5.3 Conclusion and Future Development of hPepArray

In Chapter 4, to study activities of HMTs, I designed an hPepArray including 160 unique peptide groups that are generated based on lysine residues on ten histone protein sequences of four core histones H2A, H2B, H3, H4 and one linker histone H1. From the hPepArray, HMT activities in nuclear lysate of T47D (breast cancer cell line) are successfully detected. The experimental results show that two known histone methylated peptides corresponding to H3K122 and H4K59, respectively, and one potential methylation site H2A-K74 are identified, which demonstrated that hPepArray has the ability to reveal activity profiles of HMTs in a cellular system.

Since hPepArray is processed *in vitro*, it is important to verify methylations of H3K122, H4K59 and H2A-K74 in fact also occur *in vivo*. This can be done through traditional histone methylation profiling [114] or mass spectrometry analysis. In addition, H3K4, H3K27 and H4K20 are absent in hPepArray experiment, although it has been published all of them are relevant to breast cancer [137]. Therefore, it seems that a part of histone lysine methylations are not identified by hPepArray, which requires verification by *in vivo* tests. If methylations of such lysine sites are testified *in vivo*, we will make efforts to find out the reasons. As for the possible failed detections, there are two probable reasons. One probable reason is that some of the designed peptides cannot reflect the property of specific methylations, due to which we can alter the length of peptides or change the position of core lysine residue to improve peptide design. The other possible reason of the absence of several published histone methylation sites in hPepArray is that some lysine residues participate in multiple PTMs, and some PTMs will influence the detection of methylation. For example, the loss of methylation of H3K4 in hPepArray is probably because acetylation occupied the lysine residue. Therefore, it is feasible to substantiate the hypothesis by using specific antibodies.

The results of completed hPepArray experiment primarily proved that hPepArray can reveal activities of HMTs in nuclear lysate. Therefore, it is believed that by designing different peptides corresponding for specific PTMs, hPepArray can be used to assay activities of diverse enzymes which catalyze different PTMs, such as acetylation, phosphorylation and SUMOylation. Furthermore, besides breast cancer cell lines, in future, we will apply this technology to different cell lines such as prostate cancer cell lines, lung cancer cell lines and gastric adenocarcinoma cell lines which have been demonstrated relevant to histone methylations [137]. It is the final goal to find several biomarkers for each cell line by comparing peptides among different cell lines.

In addition, there are three main directions for development of omics-scale bioinformatics technology. Firstly, it is to integrate existing knowledge in various levels， including genomes, transcriptomes and proteomes, through data mining to form a whole knowledge system which aims to facilitate the process of studying diverse biological phenomena. For example, miRFocus a comprehensive web resource (Chapter 2) aims to combine as much human miRNA relevant information as possible and thus benefit the studies in this field. Secondly, developing analytical approaches based on cloud computing is another major direction. In the development and popularization of high-throughput technologies, such as microarray and NGS, a large amount of data has been generated. Gene Expression Omnibus (GEO) repository was established to store and share the high-throughput data [138]. In order to store and analyze the enormous amount of data effectively, developing methods in terms of cloud computing is one of the feasible solutions. Thirdly, practical application like medical diagnosis is one of important final goals of omics-scale bioinformatics technology. Through omics-scale bioinformatics technology and methods, it is convenient to design and utilize different methods in terms of diverse of requirements to examine biological information of patients. For example, high-throughput technology can be used to process overall evaluation of patients through assaying their genome, transcriptome and proteome, and as for a specific disease, omics-scale bioinformatics methods are utilized to diagnose underlying conditions by testing particular SNPs and relevant proteomic activities. It is believed bioinformatics technologies of omics studies will make great contributions to individual health care in future.

Chapter 6

References

1. Watson JD, Crick FH, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature. 1953 Apr 25; **171**(4356): p.737-8.

2. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salazar A, Tsai M, Myers RM, *Quality assessment of the human genome sequence.* Nature. 2004 May 27; **429**(6990): p.365-8.

3. International Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome.* Nature. 2004 Oct 21; **431**(7011): p.931-45.

4. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP, *The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.* Science. 2005 Dec 16; **310**(5755): p.1817-21.

5. Lee RC, Feinbaum RL, Ambros V, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell. 1993 Dec 3; **75**(5): p.843-54.

6. Sayed D, Abdellatif M, *MicroRNAs in development and disease.* Physiol Rev. 2011 Jul; **91**(3): p.827-87.

7. Hogeweg P, *The roots of bioinformatics in theoretical biology.* PLoS Comput Biol. 2011 Mar; **7**(3): e1002021.

8. Lindberg DA, *Internet access to the National Library of Medicine.* Eff Clin Pract. 2000 Sep-Oct; **3**(5): p.256-60.

9. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS, *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res. 2004 Jan 1; **32**(Database issue): p.115-9.

10. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D, *GeneCards: integrating information about genes, proteins and diseases.* Trends Genet. 1997 Apr; **13**(4): p.163.

11. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, *The human genome browser at UCSC.* Genome Res. 2002 Jun; **12**(6): p.996-1006.

12. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L, *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol. 2010 May; **28**(5): p.511-5.

13. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, *The human genome browser at UCSC.* Genome Res. 2002 Jun; **12**(6): p.996-1006.

14. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L, *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol. 2010 May; **28**(5): p.511-5.

15. Cortes C, Vapnik V, *Support-vector networks.* Machine Learning **20**(3): p.273-9.

16. Judith E, Dayhoff Ph.D, James M. DeLeo, *Artificial neural networks.* Cancer. 2001 April 15; **91**(8): p.1615-35.

17. Min Jou W, Haegeman G, Ysebaert M, Fiers W, *Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein.* Nature. 1972 May 12; **237**(5350): p.82-8.

18. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA, *A map of human genome variation from population-scale sequencing.* Nature. 2010 Oct 28; **467**(7319): p.1061-73.

19. Pettersson E, Lundeberg J, Ahmadian A, *Generations of sequencing technologies.* Genomics. 2009 Feb; **93**(2): p.105-11.

20. de Magalhães JP, Finch CE, Janssens G, *Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions.* Ageing Res Rev. 2010 Jul; **9**(3): p.315-23.

21. Maxam AM, Gilbert W, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A. 1977 Feb; **74**(2): p.560-4.

22. Sanger F, Nicklen S, Coulson AR, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A. 1977 Dec; **74**(12): p.5463-7.

23. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE, *Fluorescence detection in automated DNA sequence analysis.* Nature. 1986 Jun 12-18; **321**(6071): p.674-9.

24. Swerdlow H, Gesteland R, *Capillary gel electrophoresis for rapid, high resolution DNA sequencing.* Nucleic Acids Res. 1990 Mar 25; **18**(6): p.1415-9.

25. Staden R, *A strategy of DNA sequencing employing computer programs.* Nucleic Acids Res. 1979 Jun 11; **6**(7): p.2601-10.

26. Mardis ER, *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet. 2008; **9**: p.387-402.

27. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.* Nucleic Acids Res. 2010 Apr; **38**(6): p.1767-71.

28. Ramirez-Gonzalez RH, Leggett RM, Waite D, Thanki A, Drou N, Caccamo M, Davey R, *StatsDB: platform-agnostic storage and understanding of next generation sequencing run metrics.* F1000Res. 2013 Nov 15; **2**: p.248.

29. Cox MP, Peterson DA, Biggs PJ, *SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data.* BMC Bioinformatics. 2010 Sep 27; **11**: p.485.

30. Patel RK1, Jain M, *NGS QC Toolkit: a toolkit for quality control of next generation sequencing data.* PLoS One. 2012; **7**(2): e30619.

31. Yang X, Dorman KS, Aluru S, *Reptile: representative tiling for short read error* correction. Bioinformatics. 2010 Oct 15; **26**(20): p.2526-33.

32. Kao WC, Chan AH, Song YS, *ECHO: a reference-free short-read error correction algorithm.* Genome Res. 2011 Jul; **21**(7): p.1181-92.

33. Ilie L, Molnar M, *RACER: Rapid and accurate correction of errors in reads.* Bioinformatics. 2013 Oct 1; **29**(19): p.2490-3.

34. Needleman SB, Wunsch CD, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol. 1970 Mar; **48**(3): p.443-53.

35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, *Basic local alignment search tool.* J Mol Biol. 1990 Oct 5; **215**(3): p.403-10.

36. Kent WJ, *BLAT--the BLAST-like alignment tool.* Genome Res. 2002 Apr; **12**(4): p.656-64.

37. Langmead B, Trapnell C, Pop M, Salzberg SL, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol. 2009; **10**(3): R25.

38. Li H, Durbin R, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics. 2009 Jul 15; **25**(14): p.1754-60.

39. Langmead B, Salzberg SL, *Fast gapped-read alignment with Bowtie 2. Nat Methods.* 2012 Mar 4; **9**(4): p.357-9.

40. Zhao F, Zhao F, Li T, Bryant DA, *A new pheromone trail-based genetic algorithm for comparative genome assembly. Nucleic Acids Res.* 2008 Jun; **36**(10): p.3455-62.

41. Feng J, Liu T, Qin B, Zhang Y, Liu XS, *Identifying ChIP-seq enrichment using MACS.* Nat Protoc. 2012 Sep; **7**(9): p.1728-40.

42. Wang Y, Lu J, Yu J, Gibbs RA, Yu F, *An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data.* Genome Res. 2013 May; **23**(5): p.833-42.

43. Chang TW, *Binding of cells to matrixes of distinct antibodies coated on solid surface.* J Immunol Methods. 1983 Dec 16; **65**(1-2): p.217-23.

44. Ekins RP, *Multi-analyte immunoassay.* J Pharm Biomed Anal. 1989; **7**(2): p.155-68.

45. Schena M1, Shalon D, Davis RW, Brown PO, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science. 1995 Oct 20; **270**(5235): p.467-70.

46. Haab BB, Dunham MJ, Brown PO, *Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions*. Genome Biol. 2001; **2**(2): research0004.

47. Dufva M, *Fabrication of high quality microarrays.* Biomol Eng. 2005 Dec; **22**(5-6): p.173-84.

48. Watson A, Mazumder A, Stewart M, Balasubramanian S, *Technology for microarray analysis of gene expression.* Curr Opin Biotechnol. 1998 Dec; **9**(6): p.609-14.

49. Tapia VE, Ay B, Volkmer R, *Exploring and profiling protein function with peptide arrays.* Methods Mol Biol. 2009; **570**: p.3-17.

50. Srivannavit O, Gulari M, Hua Z, Gao X, Zhou X, Hong A, Zhou T, Gulari E, *Microfluidic Reactor Array Device for Massively Parallel In-situ Synthesis of Oligonucleotides.* Sens Actuators B Chem. 2009 Jul; **140**(2): p.473-81.

51. Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM, *Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports.* Nucleic Acids Res. 1994 Dec 11; **22**(24): p.5456-65.

52. Sun H1, Chen GY, Yao SQ, *Recent advances in microarray technologies for proteomics.* Chem Biol. 2013 May 23; **20**(5): p.685-99.

53. Gao X, Gulari E, Zhou X, *In situ synthesis of oligonucleotide microarrays. Biopolymers.* 2004 Apr 5; **73**(5): p.579-96.

54. Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G, *Accurate multiplex gene synthesis from programmable DNA microchips.* Nature. 2004 Dec 23; **432**(7020): p.1050-4.

55. Pellois JP, Zhou X, Srivannavit O, Zhou T, Gulari E, Gao X, *Individually addressable parallel peptide synthesis on microchips.* Nat Biotechnol. 2002 Sep; **20**(9): p.922-6.

56. Blow N, *Genomics: catch me if you can.* Nat Methods. 2009 Jul; **6**(7): p.539-544.

57. Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, Doi H, Wood WH 3rd, Becker KG, Ko MS, *Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray.* Proc Natl Acad Sci U S A. 2000 Aug 1; **97**(16): p.9127-32.

58. Schröder C, Jacob A, Tonack S, Radon TP, Sill M, Zucknick M, Rüffer S, Costello E, Neoptolemos JP, Crnogorac-Jurcevic T, Bauer A,Fellenberg K, Hoheisel JD, *Dual-color proteomic profiling of complex samples with a microarray of 810 cancer-related antibodies.* Mol Cell Proteomics. 2010 Jun; **9**(6): p.1271-80.

59. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J, *TM4: a free, open-source system for microarray data management and analysis.* Biotechniques. 2003 Feb; **34**(2): p.374-8.

60. Smyth GK, Speed T, *Normalization of cDNA microarray data. Methods.* 2003 Dec; **31**(4): p.265-73.

61. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH, *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.* Nucleic Acids Res. 2001 Jun 15; **29**(12): p.2549-57.

62. Ronald A. Fisher, *The Correlation Between Relatives on the Supposition of Mendelian Inheritance.* Philosophical Transactions of the Royal Society of Edinburgh. 1918; **52**: p.399–433.

63. Howe EA, Sinha R, Schlauch D, Quackenbush J, *RNA-Seq analysis in MeV.* Bioinformatics. 2011 Nov 15; **27**(22): p.3209-10.

64. Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns.* Proc. Natl. Acad. Sci. USA. 1998; **95**: p.14863-8.

65. Soukas, A., P. Cohen, N.D. Socci, and J.M. Friedman, *Leptin-specific patterns of gene expression in white adipose tissue.* Genes Dev. 2000; **14**: p.963-80.

66. Butte, A.J., P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane. *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.* Proc. Natl. Acad. Sci. USA. 2000; **97**: p.12182-6.

67. Pavlidis, P., and W.S. Noble, *Analysis of strain and regional variation in gene expression in mouse brain.* Genome Biology. 2001; **2**: research0042.1-15.

68. Tusher, V.G., R. Tibshirani and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences USA; 2001; **98**: p.5116-21.

69. Brown, M.P., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler, *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc. Natl. Acad. Sci. USA; 2000; **97**: p.262-7.

70. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T, *A uniform system for microRNA annotation.* RNA. 2003 Mar; **9**(3): p.277-9.

71. Bartel DP, *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell. 2004 Jan 23; **116**(2): p.281-97.

72. Esquela-Kerscher A, Slack FJ, *Oncomirs - microRNAs with a role in cancer.* Nat Rev Cancer. 2006 Apr; **6**(4): p.259-69.

73. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T, *Identification of novel genes coding for small expressed RNAs.* Science. 2001 Oct 26; **294**(5543): p.853-8.

74. Lee RC, Ambros V, *An extensive class of small RNAs in Caenorhabditis elegans.* Science. 2001 Oct 26; **294**(5543): p.862-4.

75. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A, *Rfam: annotating non-coding RNAs in complete genomes.* Nucleic Acids Res. 2005 Jan 1; **33**(Database issue): p.121-4.

76. Shantikumar S, Caporali A, Emanueli C, *Role of microRNAs in diabetes and its cardiovascular complications.* Cardiovasc Res. 2012 Mar 15; **93**(4): p.583-93.

77. Ortega FJ, Mercader JM, Catalán V, Moreno-Navarrete JM, Pueyo N, Sabater M, Gómez-Ambrosi J, Anglada R, Fernández-Formoso JA, Ricart W, Frühbeck G, Fernández-Real JM, *Targeting the circulating microRNA signature of obesity.* Clin Chem. 2013 May; **59**(5): p.781-92.

78. van Rooij E1, Olson EN, *MicroRNA therapeutics for cardiovascular disease: opportunities and obstacles.* Nat Rev Drug Discov. 2012 Nov; **11**(11): p.860-72.

79. Ha TY, *MicroRNAs in Human Diseases: From Lung, Liver and Kidney Diseases to Infectious Disease, Sickle Cell Disease and Endometrium Disease.* Immune Netw. 2011 Dec; **11**(6): p.309-23.

80. Nassirpour R, Mehta PP, Baxi SM, Yin MJ, *miR-221 promotes tumorigenesis in human triple negative breast cancer cells.* PLoS One. 2013 Apr 24; **8**(4): e62170.

81. Baskerville S, Bartel DP, *Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.* RNA. 2005 Mar; **11**(3): p.241-7.

82. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T, *miRecords: an integrated resource for microRNA-target interactions.* Nucleic Acids Res. 2009 Jan; **37**(Database issue): p105-10.

83. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD, *miRTarBase: a database curates experimentally validated microRNA-target interactions.* Nucleic Acids Res. 2011 Jan; **39**(Database issue): D163-9.

84. Sethupathy P, Corda B, Hatzigeorgiou AG, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* RNA. 2006 Feb; **12**(2): p.192-7.

85. Lewis BP, Burge CB, Bartel DP, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell. 2005 Jan 14; **120**(1): p.15-20.

86. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N, *Combinatorial microRNA target predictions.* Nat Genet. 2005 May; **37**(5): p.495-500.

87. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E, *The role of site accessibility in microRNA target recognition.* Nat Genet. 2007 Oct; **39**(10): p.1278-84.

88. Vlachos IS, Kostoulas N, Vergoulis T, Georgakilas G, Reczko M, Maragkakis M, Paraskevopoulou MD, Prionidis K, Dalamagas T, Hatzigeorgiou AG, *DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways.* Nucleic Acids Res. 2012 Jul; **40**(Web Server issue): W498-504.

89. Lu TP, Lee CY, Tsai MH, Chiu YC, Hsiao CK, Lai LC, Chuang EY, *miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets.* PLoS One. 2012; **7**(8): e42390.

90. Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu, *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids Res. Jan 2009; **37**(Database issue): D98-104.

91. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ, *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res. 2006 Jan 1; **34**(Database issue): D140-4.

92. Kozomara A, Griffiths-Jones S, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Res. 2014 Jan; **42**(Database issue): D68-73.

93. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D, *GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.* Bioinformatics. 1998; **14**(8): p.656-64.

94. Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick, *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res. Jan 1, 2005; **33**(Database issue): D514-7.

95. Melnik BC, John SM, Schmitz G, *Milk: an exosomal microRNA transmitter promoting thymic regulatory T cell maturation preventing the development of atopy?* J Transl Med. 2014 Feb 12; **12**: p.43.

96. Zhou Q, Li M, Wang X, Li Q, Wang T, Zhu Q, Zhou X, Wang X, Gao X, Li X, *Immune-related microRNAs are abundant in breast milk exosomes.* Int J Biol Sci. 2012; **8**(1): p.118-23.

97. Kosaka N, Izumi H, Sekine K, Ochiya T, *microRNA as a new immune-regulatory agent in breast milk.* Silence. 2010 Mar 1; **1**(1): p.7.

98. Sierzchala AB, Dellinger DJ, Betley JR, Wyrzykiewicz TK, Yamada CM, Caruthers MH, *Solid-phase oligodeoxynucleotide synthesis: a two-step cycle using peroxy anion deprotection.* J Am Chem Soc. 2003 Nov 5; **125**(44): p.13427-41.

99. LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH, *Synthesis of high-quality libraries of long (150mer) oligonucleotides by a*

novel depurination controlled process. Nucleic Acids Res. 2010 May; **38**(8): p.2522-40.

100. Czar MJ, Anderson JC, Bader JS, Peccoud J, *Gene synthesis demystified.* Trends Biotechnol. 2009 Feb; **27**(2): p.63-72.

101. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, Merryman C, Young L, Noskov VN, Glass JI, Venter JC, Hutchison CA 3rd, Smith HO, *Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome.* Science. 2008 Feb 29; **319**(5867): p.1215-20.

102. Tian J, Ma K, Saaem I, *Advancing high-throughput gene synthesis technology.* Mol Biosyst. 2009 Jul; **5**(7): p.714-22.

103. Schofield MJ, Hsieh P, *DNA mismatch repair: molecular mechanisms and biological function.* Annu Rev Microbiol. 2003; **57**: p.579-608.

104. Xiong AS, Yao QH, Peng RH, Li X, Fan HQ, Cheng ZM, Li Y. *A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences.* Nucleic Acids Res. 2004 Jul 7; **32**(12): e98.

105. Andrus A, Kuimelis RG, *Analysis and purification of synthetic nucleic acids using HPLC.* Curr Protoc Nucleic Acid Chem. 2001 May; **10**: Unit 10.5.

106. Ellington A, Pollard JD Jr, *Introduction to the synthesis and purification of oligonucleotides.* Curr Protoc Nucleic Acid Chem. 2001 May; **3**: Appendix 3C.

107. Wan W, Li L, Xu Q, Wang Z, Yao Y, Wang R, Zhang J, Liu H, Gao X, Hong J, *Error removal in microchip-synthesized DNA using immobilized MutS.* Nucleic Acids Res. 2014; **42**(12): e102.

108. Baake M, Bäuerle M, Doenecke D, Albig W, *Core histones and linker histones are imported into the nucleus by different pathways.* Eur J Cell Biol. 2001 Nov; **80**(11): p.669-77.

109. Wang Y, Jia S, *Degrees make all the difference: the multifunctionality of histone H4 lysine 20 methylation.* Epigenetics. 2009 Jul 1; **4**(5): p.273-6.

110. Copeland RA, Moyer MP, Richon VM, *Targeting genetic alterations in protein methyltransferases for personalized cancer therapeutics.* Oncogene. 2013 Feb 21; **32**(8): p.939-46.

111. Bannister AJ, Kouzarides T, *Histone methylation: recognizing the methyl mark.* Methods Enzymol. 2004; **376**: p.269-88.

112. Khare SP, Habib F, Sharma R, Gadewal N, Gupta S, Galande S, *HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes.* Nucleic Acids Res. 2012 Jan; **40**(Database issue): D337-42.

113. Albert M, Helin K, *Histone methyltransferases in cancer.* Semin Cell Dev Biol. 2010 Apr; **21**(2): p.209-20.

114. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K, *High-resolution profiling of histone methylations in the human genome.* Cell. 2007 May 18; **129**(4): p.823-37.

115. Agger K, Christensen J, Cloos PA, Helin K, *The emerging functions of histone demethylases.* Curr Opin Genet Dev. 2008 Apr; **18**(2): p.159-68.

116. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RG, Otte AP, Rubin MA, Chinnaiyan AM, *The polycomb group protein EZH2 is involved in progression of prostate cancer.* Nature. 2002 Oct 10; **419**(6907): p.624-9.

117. Judith Erkmann, *Histone Modification Research Methods.* Mater Methods 2011; **1**: p.92.

118. Shechter D, Dormann HL, Allis CD, Hake SB, *Extraction, purification and analysis of histones. Nat Protoc.* 2007; **2**(6): p.1445-57.

119. Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, Cheung MS, Day DS, Gadel S, Gorchakov AA, Gu T, Kharchenko PV, Kuan S, Latorre I, Linder-Basso D, Luu Y, Ngo Q, Perry M, Rechtsteiner A, Riddle NC, Schwartz YB, Shanower GA, Vielle A, Ahringer J, Elgin SC, Kuroda MI, Pirrotta V, Ren B, Strome S, Park PJ, Karpen GH, Hawkins RD, Lieb JD, *An assessment of histone-modification antibody quality.* Nat Struct Mol Biol. 2011 Jan; **18**(1): p.91-3.

120. Bock I, Dhayalan A, Kudithipudi S, Brandt O, Rathert P, Jeltsch A, *Detailed specificity analysis of antibodies binding to modified histone tails with peptide arrays.* Epigenetics. 2011 Feb; **6**(2): p.256-63.

121. Bock I, Kudithipudi S, Tamas R, Kungulovski G, Dhayalan A, Jeltsch A, *Application of Celluspots peptide arrays for the analysis of the binding specificity*

of epigenetic reading domains to modified histone tails. BMC Biochem. 2011 Aug 31; **12**: p.48.

122. Weiss T, Hergeth S, Zeissler U, Izzo A, Tropberger P, Zee BM, Dundr M, Garcia BA, Daujat S, Schneider R, *Histone H1 variant-specific lysine methylation by G9a/KMT1C and Glp1/KMT1D.* Epigenetics Chromatin. 2010 Mar 24; **3**(1): p.7.

123. Zhou X, Zhu Q, Eicken C, Sheng N, Zhang X, Yang L, Gao X, *MicroRNA profiling using µParaflo microfluidic array technology.* Methods Mol Biol. 2012; **822**: p.153-82.

124. Tropberger P, Pott S, Keller C, Kamieniarz-Gdula K, Caron M, Richter F, Li G, Mittler G, Liu ET, Bühler M,Margueron R, Schneider R, *Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer.* Cell. 2013 Feb 14; **152**(4): p.859-72.

125. Hainer SJ, Martens JA, *Identification of histone mutants that are defective for transcription-coupled nucleosome occupancy.* Mol Cell Biol. 2011 Sep; **31**(17): p.3557-68.

126. Zhang L, Eugeni EE, Parthun MR, Freitas MA, *Identification of novel histone post-translational modifications by peptide mass fingerprinting.* Chromosoma. 2003 Aug; **112**(2): p.77-86.

127. Feng Q, Zhang Z, Shea MJ, Creighton CJ, Coarfa C, Hilsenbeck SG, Lanz R, He B, Wang L, Fu X, Nardone A, Song Y, Bradner J, Mitsiades N, Mitsiades CS, Osborne CK, Schiff R, O'Malley BW, *An epigenomic approach to therapy for tamoxifen-resistant breast cancer.* Cell Res. 2014 Jul; **24**(7): p.809-19.

128. Guerrero C, Tagwerker C, Kaiser P, Huang L, *An integrated mass spectrometry-based proteomic approach: quantitative analysis of tandem affinity-purified in vivo cross-linked protein complexes (QTAX) to decipher the 26 S proteasome-interacting network.* Mol Cell Proteomics. 2006 Feb; **5**(2): p.366-78.

129. Wei CH, Kao HY, Lu Z. *PubTator: a web-based text mining tool for assisting biocuration.* Nucleic Acids Res. 2013 Jul; **41**: W518-22.

130. Kowarsch A, Preusse M, Marr C, Theis FJ, *miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs.* RNA. 2011 May; **17**(5): p.809-19.

131. Palacios F, Abreu C, Prieto D, Morande P, Ruiz S, Fernández-Calero T, Naya H, Libisch G, Robello C, Landoni AI, Gabus R,Dighiero G, Oppezzo P, *Activation of the PI3K/AKT pathway by microRNA-22 results in CLL B-cell proliferation.* Leukemia. 2014 June 10.

132. Magrane M, Consortium U, *UniProt Knowledgebase: a hub of integrated protein data.* Database (Oxford). 2011 Mar 29; **2011**: bar009.

133. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G, *HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.* BMC Bioinformatics. 2005 Dec 1; **6**(Suppl 4): S21.

134. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA, *An*

integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1; **491**(7422): p.56-65.

135. Baker M, *Next-generation sequencing: adjusting to data overload.* Nat Methods. 2010 Jul; **7**(7): p.495-9.

136. Afgan E, Chapman B, Jadan M, Franke V, Taylor J, *Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy.* Curr Protoc Bioinformatics. 2012 Jun; **Chapter 11**: Unit11.9.

137. Greer EL, Shi Y, *Histone methylation: a dynamic mark in health, disease and inheritance.* Nat Rev Genet. 2012 Apr 3; **13**(5): p.343-57.

138. Edgar R, Domrachev M, Lash AE, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res. 2002 Jan 1; **30**(1): p.207-10.