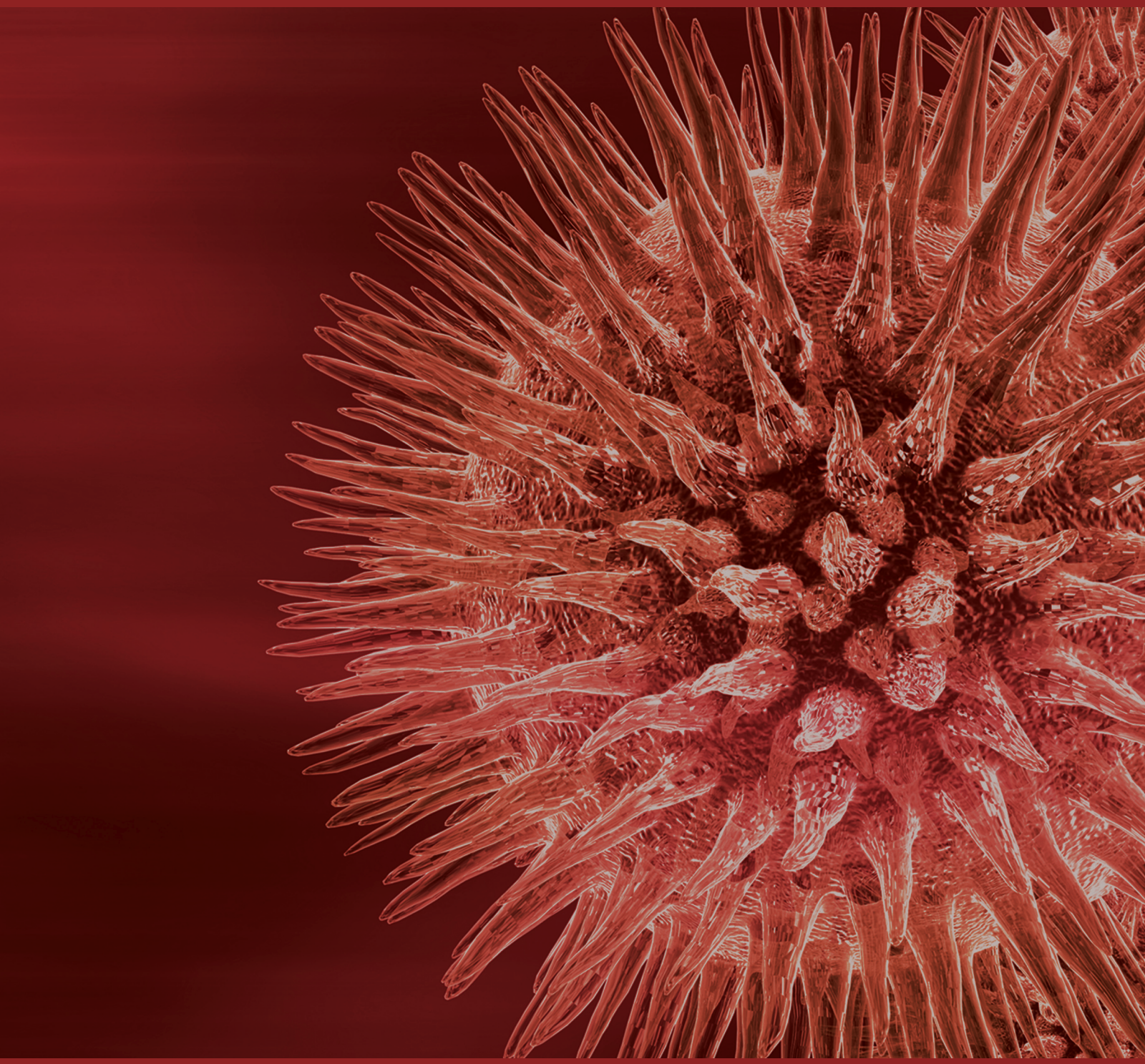
Journal of Biomedicine and Biotechnology

Next-Generation Sequencing

Guest Editors: Momiao Xiong, Zhongming Zhao, Jonathan Arnold, and Fuli Yu



Next-Generation Sequencing

Journal of Biomedicine and Biotechnology

Next-Generation Sequencing

Guest Editors: Momiao Xiong, Zhongming Zhao, Jonathan Arnold, and Fuli Yu

Editorial Board

The editorial board of the journal is organized into sections that correspond to the subject areas covered by the journal.

Agricultural Biotechnology

Guihua H. Bai, USA
Christopher P. Chanway, Canada
Ravindra N. Chibbar, Canada
Ian Godwin, Australia
Hari B. Krishnan, USA
Carol A. Mallory-Smith, USA
Dennis P. Murr, Canada
Rodomiro Ortiz, Mexico
B. C. Saha, USA
Mariam B. Sticklen, USA
Chiu-Chung Young, Taiwan

Animal Biotechnology

E. S. Chang, USA
Hans H. Cheng, USA
Bhanu P. Chowdhary, USA
Noelle E. Cockett, USA
Peter Dovc, Slovenia
Scott C. Fahrenkrug, USA
Dorian J. Garrick, USA
Thomas A. Hoagland, USA
Tosso Leeb, Switzerland
James D. Murray, USA
Anita M. Oberbauer, USA
Jorge A. Piedrahita, USA
Daniel Pomp, USA
Kent M. Reed, USA
Lawrence Reynolds, USA
Lawrence B. Schook, USA
Mari A. Smits, The Netherlands
Leon Spicer, USA
J. Verstegen, USA
Matthew B. Wheeler, USA
Kenneth L. White, USA

Biochemistry

Robert Blumenthal, USA
David Ronald Brown, UK
Saulius Butenas, USA
Vittorio Calabrese, Italy
F. Castellino, USA
Roberta Chiaraluce, Italy
D. M. Clarke, Canada
Francesca Cutruzzolà, Italy
Paul W. Doetsch, USA
Hicham Fenniri, Canada
Nick V. Grishin, USA
J. Guy Guillemette, Canada
Paul W. Huber, USA
Chen-Hsiung Hung, Taiwan
Michael Kalafatis, USA
B. E. Kemp, Australia
Phillip E. Klebba, USA
Wen-Hwa Lee, USA
Richard D. Ludescher, USA
George Makhatadze, USA
Leonid Medved, USA
Susan A. Rotenberg, USA
Jason Shearer, USA
Andrei Surguchov, USA
John B. Vincent, USA
Yujun George Zheng, USA

Bioinformatics

T. Akutsu, Japan
Miguel A. Andrade, Germany
Mark Y. Borodovsky, USA
Rita Casadio, Italy
Artem Cherkasov, Canada
David Corne, UK
Sorin Draghici, USA
Stavros J. Hamodrakas, Greece
Paul Harrison, USA
George Karypis, USA
Jack A. Leunissen, The Netherlands
Guohui Lin, Canada
Satoru Miyano, Japan
Zoran Obradovic, USA
Florencio Pazos, Spain
Zhirong Sun, China
Ying Xu, USA
A. Zelikovsky, USA
Albert Zomaya, Australia

Biophysics

Miguel Castanho, Portugal
P. Bryant Chase, USA
Kuo-Chen Chou, USA
Rizwan Khan, India
Ali A. Khraibi, Saudi Arabia
Rumiana Koynova, USA
Serdar Kuyucak, Australia
Jianjie Ma, USA
S. B. Petersen, Denmark
Peter Schuck, USA
Claudio M. Soares, Portugal

Cell Biology

Omar Benzakour, France
Sanford I. Bernstein, USA
Phillip I. Bird, Australia
Eric Bouhassira, USA
Mohamed Boutjdir, USA
Chung-Liang Chien, Taiwan
Richard Gomer, USA
Paul J. Higgins, USA
Pavel Hozak, Czech Republic
Xudong Huang, USA
Anton M. Jetten, USA
Seamus J. Martin, Ireland
Manuela Martins-Green, USA
Shoichiro Ono, USA
George Perry, USA
M. Piacentini, Italy
George E. Plopper, USA
Lawrence Rothblum, USA
Michael Sheetz, USA
James L. Sherley, USA
G. S. Stein, USA
Richard Tucker, USA
Thomas van Groen, USA
Andre Van Wijnen, USA
Steve Winder, UK
Chuanyue Wu, USA
Bin-Xian Zhang, USA

Genetics

Adewale Adeyinka, USA
Claude Bagnis, France
J. Birchler, USA
Susan Blanton, USA
Barry J. Byrne, USA
R. Chakraborty, USA
Domenico Coviello, Italy
Sarah H. Elsea, USA
Celina Janion, Poland
J. Spencer Johnston, USA
M. Ilyas Kamboh, USA
Feige Kaplan, Canada
Manfred Kayser, The Netherlands
Brynn Levy, USA
Xiao Jiang Li, USA
Thomas Liehr, Germany
James M. Mason, USA
Mohammed Rachidi, France
Raj S. Ramesar, South Africa
Elliot D. Rosen, USA
Dharambir K. Sanghera, USA
Michael Schmid, Germany
Markus Schuelke, Germany
Wolfgang Arthur Schulz, Germany
Jorge Sequeiros, Portugal
Mouldy Sioud, Norway
Rongjia Zhou, China

Genomics

Vladimir Bajic, Saudi Arabia
Margit Burmeister, USA
Settara Chandrasekharappa, USA
Yataro Daigo, Japan
J. Spencer Johnston, USA
Vladimir Larionov, USA
Thomas Lufkin, Singapore
Joakim Lundeberg, Sweden
John L. McGregor, France
John V. Moran, USA
Yasushi Okazaki, Japan
Gopi K. Podila, USA
Momiao Xiong, USA

Immunology

Hassan Alizadeh, USA
Peter Bretscher, Canada
Robert E. Cone, USA
Terry L. Delovitch, Canada
Anthony L. DeVico, USA
Nick Di Girolamo, Australia
Don Mark Estes, USA
Soldano Ferrone, USA
Jeffrey A. Frelinger, USA
John Robert Gordon, Canada
James D. Gorham, USA
Silvia Gregori, Italy
Thomas Griffith, USA
Young S. Hahn, USA
Dorothy E. Lewis, USA
Bradley W. McIntyre, USA
R. Mosley, USA
Marija Mostarica-Stojković, Serbia
Hans Konrad Muller, Australia
Ali Ouaissi, France
Kanury V. S. Rao, India
Yair Reisner, Israel
Harry W. Schroeder, USA
Wilhelm Schwaeble, UK
Nilabh Shastri, USA
Yufang Shi, China
Piet Stinissen, Belgium
Hannes Stockinger, Austria
J. W. Tervaert, The Netherlands
Graham R. Wallace, UK

Microbial Biotechnology

Jozef Anné, Belgium
Yoav Bashan, Mexico
Marco Bazzicalupo, Italy
Nico Boon, Belgium
Luca Simone Cocolin, Italy
Peter Coloe, Australia
Daniele Daffonchio, Italy
Han de Winde, The Netherlands
Yanhe Ma, China
Bernd H. A. Rehm, New Zealand
Angela Sessitsch, Austria
Effie Tsakalidou, Greece
J. Wiegel, USA

Microbiology

D. Beighton, UK
Steven R. Blanke, USA
Stanley Brul, The Netherlands
Isaac K. O. Cann, USA
Peter Dimroth, Switzerland
Stephen K. Farrand, USA
Alain Filloux, UK
Gad Frankel, UK
Roy Gross, Germany
Hans-Peter Klenk, Germany
Tanya Parish, UK
Gopi K. Podila, USA
Frederick D. Quinn, USA
Didier A. Raoult, France
Isabel Sá-Correia, Portugal
P. L. C. Small, USA
Lori Snyder, UK
Michael Thomm, Germany
H. C. van der Mei, The Netherlands
Schwan William, USA

Molecular Biology

Rudi Beyaert, Belgium
Michael Bustin, USA
Douglas Cyr, USA
K. Iatrou, Greece
Lokesh Joshi, Ireland
David W. Litchfield, Canada
Noel F. Lowndes, Ireland
Wuyuan Lu, USA
Patrick Matthias, Switzerland
John L. McGregor, France
S. L. Mowbray, Sweden
Elena Orlova, UK
Yeon-Kyun Shin, USA
William S. Trimble, Canada
Lisa Wiesmuller, Germany
Masamitsu Yamaguchi, Japan

Oncology

Colin Cooper, UK
F. M. J. Debruyne, The Netherlands
Nathan Ames Ellis, USA
Dominic Fan, USA
Gary E. Gallick, USA
Daila S. Gridley, USA
Xin-yuan Guan, Hong Kong
Anne Hamburger, USA
Manoor Prakash Hande, Singapore
Beric Henderson, Australia
Steve B. Jiang, USA
Daehee Kang, Republic of Korea
Abdul R. Khokhar, USA
Rakesh Kumar, USA
Macus Tien Kuo, USA
Eric W. Lam, UK
Sue-Hwa Lin, USA
Kapil Mehta, USA
Orhan Nalcioglu, USA
Vincent C. O. Njar, USA
P. J. Oefner, Germany
Allal Ouhtit, USA
Frank Pajonk, USA
Waldemar Priebe, USA
F. C. Schmitt, Portugal
Sonshin Takao, Japan
Ana Maria Tari, USA
Henk G. Van Der Poel, The Netherlands
Haodong Xu, USA
David J. Yang, USA

Pharmacology

Abdel A. Abdel-Rahman, USA
M. Badr, USA
Stelvio M. Bandiera, Canada
Ronald E. Baynes, USA
R. Keith Campbell, USA
Hak-Kim Chan, Australia
Michael D. Coleman, UK
J. Descotes, France
Dobromir Dobrev, Germany
Ayman El-Kadi, Canada
Jeffrey Hughes, USA
Kazim Husain, USA
Farhad Kamali, UK
Michael Kassiou, Australia
Joseph J. McArdle, USA
Mark J. McKeage, New Zealand
Daniel T. Monaghan, USA
T. Narahashi, USA
Kennerly S. Patrick, USA
Vickram Ramkumar, USA
Michael J. Spinella, USA
Quadiri Timour, France
Todd W. Vanderah, USA
Val J. Watts, USA
David J. Waxman, USA

Plant Biotechnology

P. L. Bhalla, Australia
J. R. Botella, Australia
Elvira Gonzalez De Mejia, USA
H. M. Häggman, Finland
Liwen Jiang, Hong Kong
Pulugurtha Bharadwaja Kirti, India
Yong Pyo Lim, Republic of Korea
Gopi K. Podila, USA
Ralf Reski, Germany
Sudhir Kumar Sopory, India

Toxicology

M. Aschner, USA
Michael L. Cunningham, USA
Laurence D. Fechter, USA
Hartmut Jaeschke, USA
Youmin James Kang, USA
M. Firoze Khan, USA
Pascal Kintz, France
R. S. Tjeerdema, USA
Kenneth Turteltaub, USA
Brad Upham, USA

Virology

Contents

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 370710, 2 pages
doi:10.1155/2010/370710

Editorial

Next-Generation Sequencing

Momiao Xiong,[1] Zhongming Zhao,[2] Jonathan Arnold,[3] and Fuli Yu[4]

[1] *Human Genetics Center, The University of Texas School of Public Health, Houston, TX 77025, USA*
[2] *Departments of Biomedical Informatics and Psychiatry and Vanderbilt-Ingram Cancer Center (VICC), Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[3] *Department of Genetics, The University of Georgia, Life Sciences Building, Athens, GA 30602, USA*
[4] *Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA*

Correspondence should be addressed to Zhongming Zhao, momiao.xiong@uth.tmc.edu

Received 31 December 2010; Accepted 31 December 2010

It has been widely appreciated that the genome sequence is shaping the future biomedical research. The genome sequence provides a general framework for assembling fragmentary DNA information into landscape of biological structure and function [1]. The rapid advances in DNA sequencing technology are revolutionizing biomedical research.

Starting in 2005, a variety of massively parallel sequencing instruments such as the Roche/454, the Life Technologies SOLiD, and the Illumina platforms which were largely different from the Sanger-based capillary sequencing were used to sequence the human and model organism genomes. Although each instrument has its own attributes, all massively parallel sequences machines share some common remarkable features [2]. First, the initial preparatory steps are reduced and simplified. Second, amplification of the library fragments is needed for all platforms. Third, sequencing reactions are performed and detected automatically. In the past decade, the amount of sequence output per run has been dramatically increased, the per-base cost of DNA sequencing has plummeted by ~100,000-fold, and base-calling accuracy has been largely improved. The current second-generation sequencing machines can read ~250 billion bases in a week.

When sequencing becomes simple and inexpensive, it is being routinely applied to biomedical research. To create comprehensive catalogues of genomic variants, the next-generation sequencing technologies have been used to produce sequence data in the 1000 Genomes Project. It plans to sequence more than 2000 individuals to find essentially all single-nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants with frequency >1% across the genome and >0.1% in protein-coding regions. After the project is completed by 2012, the full spectrum of human genomic variation in large, diverse sample sets will have been identified. The further reduction of cost and improvement of base-calling accuracy will uncover the genetic architectures of complex diseases and make clinical use of genome sequencing a routine practice and create great opportunities for genomic medicine.

Many layers of epigenomic information are being mapped by next-generation sequencing. Chromatic modification and protein binding can be mapped by chromatin immunoprecipitation sequencing (ChIP-Seq). The genome-wide single-base resolution of DNA methylation map has been performed by bisulfate sequencing, in which the methylated cytosines have been chemically modified. Massively parallel sequencing have also been applied to microRNA and mRNA profiling (RNA-Seq) to more accurately measure expressions of microRNA and mRNA, identify variability in microRNA sequence and mRNA sequence, and detect splice form of mRNA expressions.

Massively parallel sequencing platforms have significantly increased our ability to study the human genome and provided powerful new tools for genomic medicine. However, these technologies have also required profound changes to the data analysis. The major obstacle in genomic research is no longer data production. The major challenge in genome sequencing is the methods for data storages,

transfer, and data analysis. The classical statistical methods and computational algorithms are inadequate for analyzing the unprecedented amount of genomic sequence data. Novel analytic strategies for exploring new features of sequencing data, integrating various genomic and epigenomic data, unraveling the structure, organization, and function of the human genome, understanding fundamental principles of genomic biology, and discovering genetic and nongenetic bases of diseases are urgently needed.

This special issue includes six high-quality papers, which were selected after undergoing rigorous peer review. We briefly describe the papers in the following.

The first paper (V. Costa et al., 2010) provides a comprehensive survey of the RNA sequencing methodology. RNA sequencing is a major platform in the next-generation sequencing (NGS), aiming to accurately determine expression levels of specific genes, differential splicing, and allele-specific expression of transcripts at the transcriptome level. So far, RNA sequencing remains the most complex (NGS) application. The authors focus on the challenges that RNA sequencing presents both from a biological and a bioinformatics point of view.

In the second paper (Y. Qi et al., 2010), the authors apply high-throughput sequencing of microRNAs in adenovirus type 3 (AD3) infected human laryngeal epithelial (Hep2) cells. Using the SOLiD sequencing technology, analysis of microRNAs profiles identified 492 precursor microRNAs in the AD3 infected Hep2 cells and 540 precursor microRNAs in the control. Among them, 44 and 36 microRNAs showed high and lower expression in the AD3 infected cells than the control, respectively. The study demonstrates that NGS is efficient and powerful for microRNA profiling in the virus-infected cell lines.

L. Cui et al. (2010) also apply SOLiD sequencing to profile microRNAs involved in the host response to enterovirus 71 (EV71) infection. They found 64 microRNAs whose expression levels changed from more than 2-fold in response to EV71 infection in Hep2 cells. Functional analysis like Gene Ontology enrichment test revealed that many of these microRNAs might be involved in neurological process, immune response, and cell death pathways, which have known to be associated with the extreme virulence of EV71. As authors stated, this is the first paper on host microRNAs expression alteration in response to EV71 infection.

W. Wang et al. (2010) use another NGS technology, ChIP-Seq, to find the targeting microRNA genes of a transcription factor, EGR1, in human erythroleukemia cell line K562. They found EGR1 binding sites near the promoters of 124 distinct microRNA genes, accounting for about 42% of the miRNAs which have high-confidence predicted promoters (294). They also found that EGR1 binds to another 63 pre-miRNAs. This study provides the first global binding profile between the transcription factor EGR1 and its targeting miRNA genes in PMA-treated K562 cells.

S. Hasson et al. (2010) report the cloning of cDNA sequences encoding four groups or isoforms of the haemostasis-disruptive Serine protease proteins (SPs) from the venom glands of *Echis ocellatus*, whose bite is the leading cause of death and morbidity in Africa. Based on their observation of the extraordinary level of interspecific and intergeneric sequence conservation exhibited by the *Echis ocellatus* EoSPs and analogous serine proteases from other viper species, the authors speculate that antibodies to representative molecules should neutralise the biological function of this important group of venom toxins in vipers that are distributed throughout Africa, the Middle East, and the Indian subcontinent.

The last paper (Z. Zhao and C. Jiang 2010) conducts a comparative genome-wide polymorphism-fixation analysis of human codons, as previously investigators often analyze either interspecies fixed substitutions or intraspecies nucleotide polymorphisms, but not both data types simultaneously. The authors report many features in the recent codon evolution. They conclude that fixation process could effectively and quickly correct the volatile changes introduced by polymorphisms so that codon changes could be gradual and directional and that codon composition could be kept relatively stable during evolution. As numerous mutation data have been identified by sequencing and many more will be identified by NGS in the near future, such analysis may help us understand mutational process in the recent genome evolution.

Acknowledgment

We are especially grateful to the anonymous reviewers who helped improve the quality of the papers in this special issue.

Momiao Xiong
Zhongming Zhao
Jonathan Arnold
Fuli Yu

References

[1] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, no. 7333, pp. 187–197, 2011.

[2] E. R. Mardis, "A dacade's perspective on DNA sequencing technology," *Nature*, vol. 470, no. 7333, pp. 198–203, 2011.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 853916, 19 pages
doi:10.1155/2010/853916

Review Article

Uncovering the Complexity of Transcriptomes with RNA-Seq

Valerio Costa,[1] Claudia Angelini,[2] Italia De Feis,[2] and Alfredo Ciccodicola[1]

[1] *Institute of Genetics and Biophysics "A. Buzzati-Traverso", IGB-CNR, 80131 Naples, Italy*
[2] *Istituto per le Applicazioni del Calcolo "Mauro Picone", IAC-CNR, 80131 Naples, Italy*

Correspondence should be addressed to Valerio Costa, costav@igb.cnr.it

Received 22 February 2010; Accepted 7 April 2010

Academic Editor: Momiao Xiong

In recent years, the introduction of massively parallel sequencing platforms for Next Generation Sequencing (NGS) protocols, able to simultaneously sequence hundred thousand DNA fragments, dramatically changed the landscape of the genetics studies. RNA-Seq for transcriptome studies, Chip-Seq for DNA-proteins interaction, CNV-Seq for large genome nucleotide variations are only some of the intriguing new applications supported by these innovative platforms. Among them RNA-Seq is perhaps the most complex NGS application. Expression levels of specific genes, differential splicing, allele-specific expression of transcripts can be accurately determined by RNA-Seq experiments to address many biological-related issues. All these attributes are not readily achievable from previously widespread hybridization-based or tag sequence-based approaches. However, the unprecedented level of sensitivity and the large amount of available data produced by NGS platforms provide clear advantages as well as new challenges and issues. This technology brings the great power to make several new biological observations and discoveries, it also requires a considerable effort in the development of new bioinformatics tools to deal with these massive data files. The paper aims to give a survey of the RNA-Seq methodology, particularly focusing on the challenges that this application presents both from a biological and a bioinformatics point of view.

1. Introduction

It is commonly known that the genetic information is conveyed from DNA to proteins via the messenger RNA (mRNA) through a finely regulated process. To achieve such a regulation, the concerted action of multiple cis-acting proteins that bind to gene flanking regions—"core" and "auxiliary" regions—is necessary [1]. In particular, core elements, located at the exons' boundaries, are strictly required for initiating the pre-mRNA processing events, whereas auxiliary elements, variable in number and location, are crucial for their ability to enhance or inhibit the basal splicing activity of a gene.

Until recently—less than 10 years ago—the central dogma of genetics indicated with the term "gene" a DNA portion whose corresponding mRNA encodes a protein. According to this view, RNA was considered a "bridge" in the transfer of biological information between DNA and proteins, whereas the identity of each expressed gene, and of its transcriptional levels, were commonly indicated as "transcriptome" [2]. It was considered to mainly consist of ribosomal RNA (80–90%, rRNA), transfer RNA (5–15%, tRNA), mRNA (2–4%) and a small fraction of intragenic (i.e., intronic) and intergenic noncoding RNA (1%, ncRNA) with undefined regulatory functions [3]. Particularly, both intragenic and intergenic sequences, enriched in repetitive elements, have long been considered genetically inert, mainly composed of "junk" or "selfish" DNA [4]. More recently it has been shown that the amount of noncoding DNA (ncDNA) increases with organism complexity, ranging from 0.25% of prokaryotes' genome to 98.8% of humans [5]. These observations have strengthened the evidence that ncDNA, rather than being junk DNA, is likely to represent the main driving force accounting for diversity and biological complexity of living organisms.

Since the dawn of genetics, the relationship between DNA content and biological complexity of living organisms has been a fruitful field of speculation and debate [6]. To date, several studies, including recent analyses performed during the ENCODE project, have shown the pervasive nature of eukaryotic transcription with almost the full length of nonrepeat regions of the genome being transcribed [7].

The unexpected level of complexity emerging with the discovery of endogenous small interfering RNA (siRNA) and microRNA (miRNA) was only the tip of the iceberg [8]. Long interspersed noncoding RNA (lincRNA), promoter- and terminator-associated small RNA (PASR and TASR, resp.), transcription start site-associated RNA (TSSa-RNA), transcription initiation RNA (tiRNA) and many others [8] represent part of the interspersed and crosslinking pieces of a complicated transcription puzzle. Moreover, to cause further difficulties, there is the evidence that most of the pervasive transcripts identified thus far, have been found only in specific cell lines (in most of cases in mutant cell lines) with particular growth conditions, and/or particular tissues. In light of this, discovering and interpreting the complexity of a transcriptome represents a crucial aim for understanding the functional elements of such a genome. Revealing the complexity of the genetic code of living organisms by analyzing the molecular constituents of cells and tissues, will drive towards a more complete knowledge of many biological issues such as the onset of disease and progression.

The main goal of the whole transcriptome analyses is to identify, characterize and catalogue all the transcripts expressed within a specific cell/tissue—at a particular stage—with the great potential to determine the correct splicing patterns and the structure of genes, and to quantify the differential expression of transcripts in both physio- and pathological conditions [9].

In the last 15 years, the development of the hybridization technology, together with the tag sequence-based approaches, allowed to get a first deep insight into this field, but, beyond a shadow of doubt, the arrival on the marketplace of the NGS platforms, with all their "*Seq*" applications, has completely revolutionized the way of thinking the molecular biology.

The aim of this paper is to give an overview of the RNA-Seq methodology, trying to highlight all the challenges that this application presents from both the biological and bioinformatics point of view.

2. Next Generation Sequencing Technologies

Since the first complete nucleotide sequence of a gene, published in 1964 by Holley [10] and the initial developments of Maxam and Gilbert [11] and Sanger et al. [12] in the 1970s (see Figure 1), the world of nucleic acid sequencing was a RNA world and the history of nucleic acid sequencing technology was largely contained within the history of RNA sequencing.

In the last 30 years, molecular biology has undergone great advances and 2004 will be remembered as the year that revolutionized the field; thanks to the introduction of massively parallel sequencing platforms, the *Next Generation Sequencing*-era, [13–15], started. Pioneer of these instruments was the Roche (454) Genome Sequencer (GS) in 2004 (http://www.454.com/), able to simultaneously sequence several hundred thousand DNA fragments, with a read length greater than 100 base pairs (bp). The current GS FLX Titanium produces greater than 1 million reads in excess of 400 bp. It was followed in 2006 by the Illumina Genome Analyzer (GA) (http://www.illumina.com/) capable to generate tens of millions of 32-bp reads. Today, the Illumina GAIIx produces 200 million 75–100 bp reads. The last to arrive in the marketplace was the Applied Biosystems platform based on Sequencing by Oligo Ligation and Detection (SOLiD) (http://www3.appliedbiosystems.com/AB_Home/index.htm), capable of producing 400 million 50-bp reads, and the Helicos BioScience HeliScope (http://www.helicosbio.com/), the first single-molecule sequencer that produces 400 millions 25–35 bp reads.

While the individual approaches considerably vary in their technical details, the essence of these systems is the miniaturization of individual sequencing reactions. Each of these miniaturized reactions is seeded with DNA molecules, at limiting dilutions, such that there is a single DNA molecule in each, which is first amplified and then sequenced. To be more precise, the genomic DNA is randomly broken into smaller sizes from which either fragment templates or mate-pair templates are created. A common theme among NGS technologies is that the template is attached to a solid surface or support (immobilization by primer or template) or indirectly immobilized (by linking a polymerase to the support). The immobilization of spatially separated templates allows simultaneous thousands to billions of sequencing reactions. The physical design of these instruments allows for an optimal spatial arrangement of each reaction, enabling an efficient readout by laser scanning (or other methods) for millions of individual sequencing reactions onto a standard glass slide. While the immense volume of data generated is attractive, it is arguable that the elimination of the cloning step for the DNA fragments to sequence is the greatest benefit of these new technologies. All current methods allow the direct use of small DNA/RNA fragments not requiring their insertion into a plasmid or other vector, thereby removing a costly and time-consuming step of traditional Sanger sequencing.

It is beyond a shadow of doubt that the arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. The broadest application of NGS may be the resequencing of different genomes and in particular, human genomes to enhance our understanding of how genetic differences affect health and disease. Indeed, these platforms have been quickly applied to many genomic contexts giving rise to the following "Seq" protocols: RNA-Seq for transcriptomics, Chip-Seq for DNA-protein interaction, DNase-Seq for the identification of most active regulatory regions, CNV-Seq for copy number variation, and methyl-Seq for genome wide profiling of epigenetic marks.

3. RNA-Seq

RNA-Seq is perhaps one of the most complex next-generation applications. Expression levels, differential splicing, allele-specific expression, RNA editing and fusion transcripts constitute important information when comparing samples for disease-related studies. These attributes, not

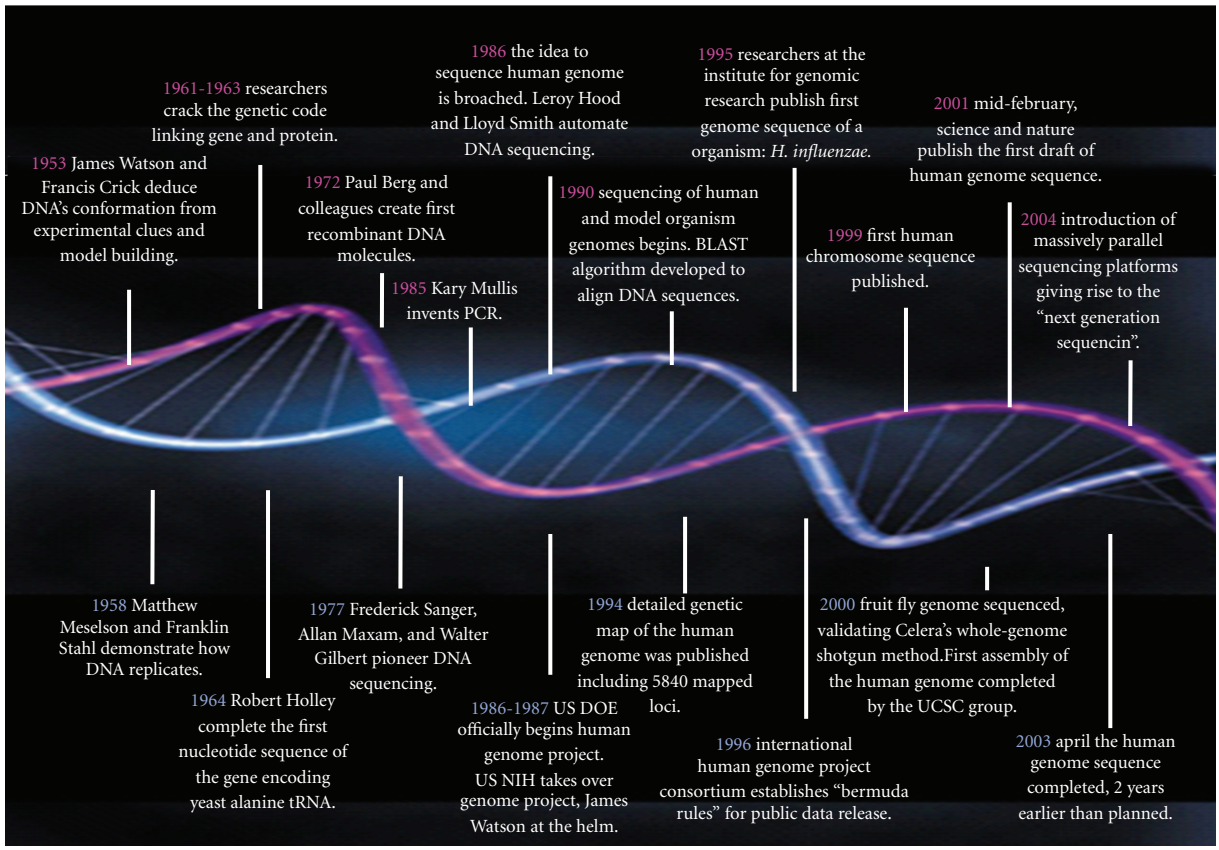


Figure 1: Evolution of DNA revolution.

readily available by hybridization-based or tag sequence-based approaches, can now be far more easily and precisely obtained if sufficient sequence coverage is achieved. However, many other essential subtleties in the RNA-Seq data remain to be faced and understood.

Hybridization-based approaches typically refer to the microarray platforms. Until recently, these platforms have offered to the scientific community a very useful tool to simultaneously investigate thousands of features within a single experiment, providing a reliable, rapid, and cost-effective technology to analyze the gene expression patterns. Due to their nature, they suffer from background and cross-hybridization issues and allow researchers to only measure the relative abundance of RNA transcripts included in the array design [16]. This technology, which measures gene expression by simply quantifying—via an indirect method—the hybridized and labeled cDNA, does not allow the detection of RNA transcripts from repeated sequences, offering a limited dynamic range, unable to detect very subtle changes in gene expression levels, critical in understanding any biological response to exogenous stimuli and/or environmental changes [9, 17, 18].

Other methods such as Serial, Cap Analysis of Gene Expression (SAGE and CAGE, resp.) and Polony Multiplex Analysis of Gene Expression (PMAGE), tag-based sequencing methods, measure the absolute abundance of transcripts in a cell/tissue/organ and do not require prior knowledge of any gene sequence as occurs for microarrays [19]. These analyses consist in the generation of sequence tags from fragmented cDNA and their following concatenation prior to cloning and sequencing [20]. SAGE is a powerful technique that can therefore be viewed as an unbiased digital microarray assay. However, although SAGE sequencing has been successfully used to explore the transcriptional landscape of various genetic disorders, such as diabetes [21, 22], cardiovascular diseases [23], and Downs syndrome [24, 25], it is quite laborious for the cloning and sequencing steps that have thus far limited its use.

In contrast, RNA-Seq on NGS platforms has clear advantages over the existing approaches [9, 26]. First, unlike hybridization-based technologies, RNA-Seq is not limited to the detection of known transcripts, thus allowing the identification, characterization and quantification of new splice isoforms. In addition, it allows researchers to determine the correct gene annotation, also defining—at single nucleotide resolution—the transcriptional boundaries of genes and the expressed Single Nucleotide Polymorphisms (SNPs). Other advantages of RNA-Seq compared to microarrays are the low "background signal," the absence of an upper limit for quantification and consequently, the larger dynamic range of expression levels over which transcripts can be detected. RNA-Seq data also show high levels of reproducibility for both technical and biological replicates.

Table 1: Selection of papers on mammalian RNA-Seq.

Reference	Organism	Cell type/tissue	NGS platform
Bainbridge et al., 2006 [27]	*Homo sapiens*	Prostate cancer cell line	Roche
Cloonan et al., 2008 [30]	*Mus musculus*	ES cells and Embryoid bodies	ABI
Core et al., 2008 [31]	*Homo sapiens*	Lung fibroblasts	IIlumina
Hashimoto et al., 2008 [32]	*Homo sapiens*	HT29 cell line	ABI
Li et al., 2008 [33]	*Homo sapiens*	Prostate cancer cell line	IIlumina
Marioni et al., 2008 [34]	*Homo sapiens*	Liver and kidney samples	IIlumina
Morin et al., 2008 [35]	*Homo sapiens*	ES cells and Embryoid bodies	IIlumina
Morin et al., 2008 [36]	Homo sapiens	HeLa S3 cell line	IIlumina
Mortazavi et al., 2008 [37]	*Mus musculus*	Brain, liver and skeletal muscle	IIlumina
Rosenkran et al., 2008 [38]	*Mus musculus*	ES cells	IIlumina
Sugarbaker et al., 2008 [39]	*Homo sapiens*	Malignant pleural mesothelioma, adenocarcinoma and normal lung	Roche
Sultan et al., 2008 [40]	*Homo sapiens*	Human embryonic kidney and B cell line	IIlumina
Asmann et al., 2009 [41]	*Homo sapiens*	Universal and brain human reference RNAs	IIlumina
Chepelev et al., 2009 [42]	*Homo sapiens*	Jurkat and $GD4^+$ T cells	IIlumina
Levin et al., 2009 [43]	*Homo sapiens*	K562	IIlumina
Maher et al., 2009 [44]	*Homo sapiens*	Prostate cancer cell lines	Roche IIlumina
Parkhomchuk et al., 2009 [45]	*Mus musculus*	Brain	IIlumina
Reddy et al., 2009 [46]	*Homo sapiens*	A549 cell line	IIlumina
Tang et al., 2009 [47]	*Mus musculus*	Blastomere and oocyte	ABI
Blekhman et al., 2010 [48]	*Homo sapiens,* *Pan troglodytes,* *Rhesus macaca.*	Liver	IIlumina
Heap et al., 2010 [49]	*Homo sapiens*	Primary $GD4^+$ T cells	IIlumina
Raha et al., 2010 [50]	*Homo sapiens*	K562 cell line	IIlumina

Recent studies have clearly demonstrated the advantages of using RNA-Seq [27–50]. Table 1 provides a short description of recent and more relevant papers on RNA-Seq in mammals.

Many research groups have been able to precisely quantify known transcripts, to discover new transcribed regions within intronic or intergenic regions, to characterize the antisense transcription, to identify alternative splicing with new combinations of known exon sequences or new transcribed exons, to evaluate the expression of repeat elements and to analyze a wide number of known and possible new candidate expressed SNPs, as well as to identify fusion transcripts and other new RNA categories.

3.1. Sample Isolation and Library Preparation. The first step in RNA-Seq experiments is the isolation of RNA samples; further RNA processing strictly depends on the kind of analysis to perform. Indeed, as "transcriptome" is defined as the complete collection of transcribed elements in a genome (see [2]), it consists of a wide variety of transcripts, both mRNA and non-mRNA, and a large amount (90–95%) of rRNA species. To perform a whole transcriptome analysis, not limited to annotated mRNAs, the selective depletion of abundant rRNA molecules (5S, 5.8S, 18S and 28S) is a key step. Hybridization with rRNA sequence-specific 5′-biotin labeled oligonucleotide probes, and the following removal with streptavidin-coated magnetic beads, is the main procedure to selectively deplete large rRNA molecules from total isolated RNA. Moreover, since rRNA—but not capped mRNAs—is characterized by the presence of 5′ phosphate, an useful approach for selective ribo-depletion is based on the use of an exonuclease able to specifically degrade RNA molecules bearing a 5′ phosphate (mRNA-ONLY kit, Epicentre). Compared to the polyadenylated (polyA+) mRNA fraction, the ribo-depleted RNA is enriched in non-polyA mRNA, preprocessed RNA, tRNA, regulatory molecules such as miRNA, siRNA, small ncRNA, and other RNA transcripts of yet unknown function (see review [8]).

How closely the RNA sequencing reflects the original RNA populations is mainly determined in the library preparation step, crucial in the whole transcriptome protocols. Although NGS protocols were first developed for the analysis of genomic DNA, these technical procedures have been rapidly and effectively adapted to the sequencing of double-strand (ds) cDNA for transcriptome studies [51].

A double-stranded cDNA library can be usually prepared by using: (1) fragmented double-stranded (ds) cDNA and (2) hydrolyzed or fragmented RNA.

The goal of the first approach is to generate high-quality, full-length cDNAs from RNA samples of interest to be fragmented and then ligated to an adapter for further amplification and sequencing. By the way, since the primer adaptor is ligated to a fragmented ds cDNA, any information on the transcriptional direction would completely be lost. Preserving the strandedness is fundamental for data analysis; it allows to determine the directionality of transcription and gene orientation and facilitates detection of opposing and overlapping transcripts. To take into account and thus to avoid this biologically relevant issue, many approaches, such as pretreating the RNA with sodium bisulphite to convert cytidine into uridine [52], have been so far developed. Other alternative protocols, differing in how the adaptors are inserted into ds cDNA, have been recently published: direct ligation of RNA adaptors to the RNA sample before or during reverse transcription [30, 31, 53], or incorporation of dUTP during second strand synthesis and digestion with uracil-Nglycosylase enzyme [45]. For instance, SOLiD Whole Transcriptome Kit contains two different sets of oligonucleotides with a single-stranded degenerate sequence at one end, and a defined sequence required for sequencing at the other end, constraining the orientation of RNA in the ligation reaction. The generation of ds cDNA from RNA involves a number of steps. First, RNA is converted into first-strand cDNA using reverse transcriptase with either random hexamers or oligo(dT) as primers. The resulting first-strand cDNA is then converted into double-stranded cDNA, further fragmented with DNAse I and then ligated to adapters for amplification and sequencing [54]. The advantage of using oligo dT is that the majority of cDNA produced should be polyadenylated mRNA, and hence more of the sequence obtained should be informative (nonribosomal). The significant disadvantage is that the reverse transcriptase enzyme will fall off of the template at a characteristic rate, resulting in a bias towards the 3′ end of transcripts. For long mRNAs this bias can be pronounced, resulting in an under representation (or worse in the absence) of the 5′ end of the transcript in the data. The use of random primers would therefore be the preferred method to avoid this problem and to allow a better representation of the 5′ end of long ORFs. However, when oligo dT primers are used for priming, the slope which is formed by the diminishing frequency of reads towards the 5′ end of the ORF can, in some cases, be useful for determining the strand of origin for new transcripts if strand information has not been retained [28, 37].

Fragmenting RNA, rather than DNA, has the clear advantage of reducing possible secondary structures, particularly for tRNA and miRNA, resulting in a major heterogeneity in coverage and can also lead to a more comprehensive transcriptome analysis (Figure 2). In this case, the RNA sample is first fragmented by using controlled temperature or chemical/enzymatic hydrolysis, ligated to adapters and retro-transcribed by complementary primers. Different protocols have been so far developed. Indeed, the adaptor sequences may be directly ligated to the previously fragmented RNA molecules by using T4 RNA ligase, and the resulting library can be reverse transcribed with primer pairs specifically suited on the adaptor sequences, and then sequenced. Another approach, recently described in [55], consists in the *in vitro* polyadenilation of RNA fragments in order to have a template for the next step of reverse transcription using poly(dT) primers containing both adaptor sequences (linkers), separated back-to-back by an endonuclease site. The resulting cDNAs are circularized and then cleaved at endonuclease site in the adaptors, thus leaving ss cDNA with the adaptors at both ends [55]. A third protocol described by [33], named double random priming method, uses biotinylated random primers (a sequencing primer P1 at the 5′ end, and a random octamer at the 3′ end). After a first random priming reaction, the products are isolated by using streptavidin beads and a second random priming reaction is performed on a solid phase with a random octamer carrying the sequencing primer P2. Afterwards, second random priming products are released from streptavidin beads by heat, PCR-amplified, gel-purified, and finally subjected to sequencing process from the P1 primer. Moreover, as already mentioned, in [45] the authors used dUTP—a surrogate for dTTP—during the second-strand synthesis to allow a selective degradation of second cDNA strand after adaptor ligation using a uracil-N-glycosylase. The use of engineered DNA adaptors, combined to the dUTP protocol, ensures that only the cDNA strand corresponding to the "real" transcript is used for library amplification and sequencing, reserving the strandedness of gene transcription [45].

However, independently on the library construction procedure, particular care should be taken to avoid complete degradation during RNA fragmentation.

The next step of the sequencing protocols is the clonally amplification of the cDNA fragments.

Illumina, 454 and SOLiD use clonally amplified templates. In particular, the last two platforms use an innovative procedure, emulsion PCR (emPCR), to prepare sequencing templates in a cell-free system. cDNA fragments from a fragment or paired-end library are separated into single strands and captured onto beads under conditions that favour one DNA molecule per bead. After the emPCR and beads enrichment, millions of them are chemically crosslinked to an amino-coated glass surface (SOLiD) or deposited into individual PicoTiterPlate (PTP) wells (454) in which the NGS chemistry can be performed. Solid-phase amplification (Illumina) can also be used to produce randomly distributed, clonally amplified clusters from fragment or mate-pair templates on a glass slide. High-density forward and reverse primers are covalently attached to the slide, and the ratio of the primers to the template defines the surface density. This procedure can produce up to 200 million spatially separated template clusters, providing ends for primer hybridization, needed to initiate the NGS reaction. A different approach is the use of single molecules templates (Helicos BioScience) usually immobilized on solid supports, in which PCR amplification is no more required, thus avoiding the insertion of possible confounding mutations in the templates. Furthermore, AT- and GC-rich sequences present amplification issues, with over- or under-representation bias

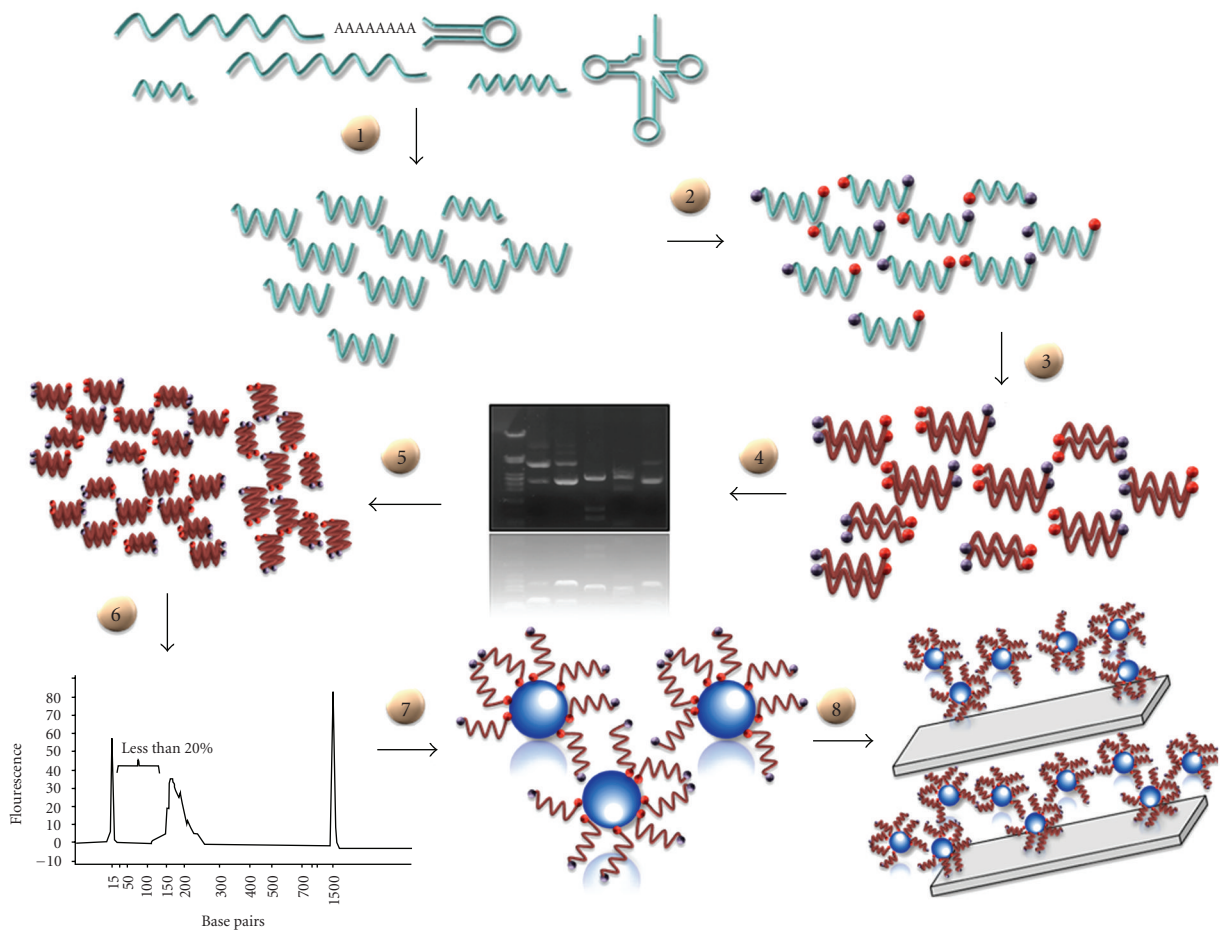


Figure 2: *Library preparation and clonal amplification.* Schematic representation of a workflow for library preparation in RNA-Seq experiments on the SOLiD platform. In the figure is depicted a total RNA sample after depletion of rRNA, containing both polyA and non-polyA mRNA, tRNAs, miRNAs and small noncoding RNAs. Ribo-depleted total RNA is fragmented (1), then ligated to specific adaptor sequences (2) and retro-transcribed (3). The resulting cDNA is size selected by gel electrophoresis (4), and cDNAs are PCR amplified (5). Then size distribution is evaluated (6). Emulsion PCR, with one cDNA fragment per bead, is used for the clonal amplification of cDNA libraries (7). Purified and enriched beads are finally deposited onto glass slides (8), ready to be sequenced by ligation.

in genome alignments and assemblies. Specific adaptors are bound to the fragmented templates, then hybridized to spatially distributed primers covalently attached to the solid support [56].

3.2. Sequencing and Imaging. NGS platforms use different sequencing chemistry and methodological procedures.

Illumina and HeliScope use the Cyclic Reversible Termination (CRT), which implies the use of reversible terminators (modified nucleotide) in a cyclic method. A DNA polymerase, bound to the primed template, adds one fluorescently modified nucleotide per cycle; then the remaining unincorporated nucleotides are washed away and imaging capture is performed. A cleavage step precedes the next incorporation cycle to remove the terminating/inhibiting group and the fluorescent dye, followed by an additional washing. Although these two platforms use the same methodology, Illumina employs the four-colour CRT method, simultaneously incorporating all 4 nucleotides with different dyes; HeliScope uses the one-colour (Cy5 dye) CRT method.

Substitutions are the most common error type, with a higher portion of errors occurring when the previous incorporated nucleotide is a G base [57]. Under representation of AT-rich and GC-rich regions, probably due to amplification bias during template preparation [57–59], is a common drawback.

In contrast, SOLiD system uses the Sequencing by Ligation (SBL) with 1, 2-nucleotide probes, based on colour space, which is an unique feature of SOLiD. It has the main advantage to improve accuracy in colour and single nucleotide variations (SNV) calling, the latter of which requires an adjacent valid colour change. In particular, a universal primer is hybridized to the template beads, and a library of 1, 2-nucleotide probes is added. Following four-colour imaging, the ligated probes are chemically cleaved to generate a 5′-phosphate group. Probe hybridization and ligation, imaging, and probe cleavage is repeated ten times to yield ten colour calls spaced in five-base intervals. The extended primer is then stripped from the solid-phase-bound templates. A second ligation round is performed with

a $n - 1$ primer, which resets the interrogation bases and the corresponding ten colour calls one position to the left. Ten ligation cycles ensue, followed by three rounds of ligation cycles. Colour calls from the five-ligation rounds are then ordered into a linear sequence (the csfasta colour space) and aligned to a reference genome to decode the sequence. The most common error type observed by using this platform are substitutions, and, similar to Illumina, SOLiD data have also revealed an under representation of AT- and GC-rich regions [58].

Another approach is pyrosequencing (on 454), a non-electrophoretic bioluminescence method, that unlike the above-mentioned sequencing approaches is able to measure the release of pyrophosphate by proportionally converting it into visible light after enzymatic reactions. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle. The enzymatic cascade generates a light recorded as a flowgram with a series of picks corresponding to a particular DNA sequence. Insertions and deletions are the most common error types.

An excellent and detailed review about the biotechnological aspects of NGS platforms can be found in [15].

3.3. From Biology to Bioinformatics. The unprecedented level of sensitivity in the data produced by NGS platforms brings with it the power to make many new biological observations, at the cost of a considerable effort in the development of new bioinformatics tools to deal with these massive data files.

First of all, the raw image files from one run of some next generation sequencers can require terabytes of storage, meaning that simply moving the data off the machine can represent a technical challenge for the computer networks of many research centers. Moreover, even when the data are transferred from the machine for subsequent processing, common desktop computer will be hopelessly outmatched by the volume of data from a single run. As a result, the use of a small cluster of computers is extremely beneficial to reduce computational bottleneck.

Another issue is the availability of software required to perform downstream analysis. Indeed after image and signal processing the output of a RNA-Seq experiment consists of 10–400 millions of short reads (together with their base-call quality values), typically of 30–400 bp, depending on the DNA sequencing technology used, its version and the total cost of the experiments.

NGS data analysis heavily relies on proper mapping of sequencing reads to corresponding reference genomes or on their efficient *de novo* assembly. Mapping NGS reads with high efficiency and reliability currently faces several challenges. As noticed by [60], differences between the sequencing platforms in samples preparation, chemistry, type and volume of raw data, and data formats are very large, implying that each platform produces data affected by characteristic error profiles. For example the 454 system can produce reads with insertion or deletion errors during homopolymer runs and generate fewer, but longer, sequences in fasta like format allowing to adapt classical alignment algorithms; the Illumina has an increased likelihood to accumulate sequence errors toward the end of the read and produce fasta reads, but they are shorter, hence requiring specific alignment algorithms; the SOLiD also tends to accumulate bias at the end of the reads, but uses di-base encoding strategy and each sequence output is encoded in a colour space csfasta format. Hence, some sequence errors are correctable, providing better discrimination between sequencing error and polymorphism, at the cost of requiring analysis tools explicitly built for handling this aspect of the data. It is not surprising that there are no "box standard" software available for end-users, hence the implementation of individualized data processing pipelines, combining third part packages and new computational methods, is the only advisable approach. While some existing packages are already enabling to solve general aspects of RNA-Seq analysis, they also require a time consuming effort due to the lack of clear documentation in most of the algorithms and the variety of the formats. Indeed, a much clear documentation of the algorithms is needed to ensure a full understanding of the processed data. Community adoption of input/output data formats for reference alignments, assemblies and detected variants is also essential for ease the data management problem. Solving these issues may simply shift the software gap from sequence processing (base-calling, alignment or assembly, positional counting and variant detection) to sequence analysis (annotation and functional impact).

3.4. Genome Alignment and Reads Assembly. The first step of any NGS data analysis consists of mapping the sequence reads to a reference genome (and/or to known annotated transcribed sequences) if available, or *de novo* assembling to produce a genome-scale transcriptional map. (see Figure 3 for an illustration of a classical RNA-Seq computational pipeline). The decision to use one of strategies is mainly based on the specific application. However, independently on the followed approach, there is a preliminary step that can be useful to perform which involves the application of a quality filtering to remove poor quality reads and to reduce the computational time and the effort for further analysis.

Analyzing the transcriptome of organisms without a specific reference genome requires *de novo* assembling (or a guided assembly with the help of closely related organisms) of expressed sequence tags (ESTs) using short-read assembly programs such as [61, 62]. A reasonable strategy for improving the quality of the assembly is to increase the read coverage and to mix different reads types. However RNA-Seq experiments without a reference genome propose specific features and challenges that are out of the scope of the present paper; we refer the readers to [63, 64] for further details.

In most cases, the reference genome is available and the mapping can be carried out using either the whole genome or known transcribed sequences (see, e.g., [28–30, 32, 34, 37, 40, 46, 47]). In both cases, this preliminary but crucial step is the most computationally intensive of the entire process and strongly depends on the type of available sequences (read-length, error profile, amount of data and data format). It is

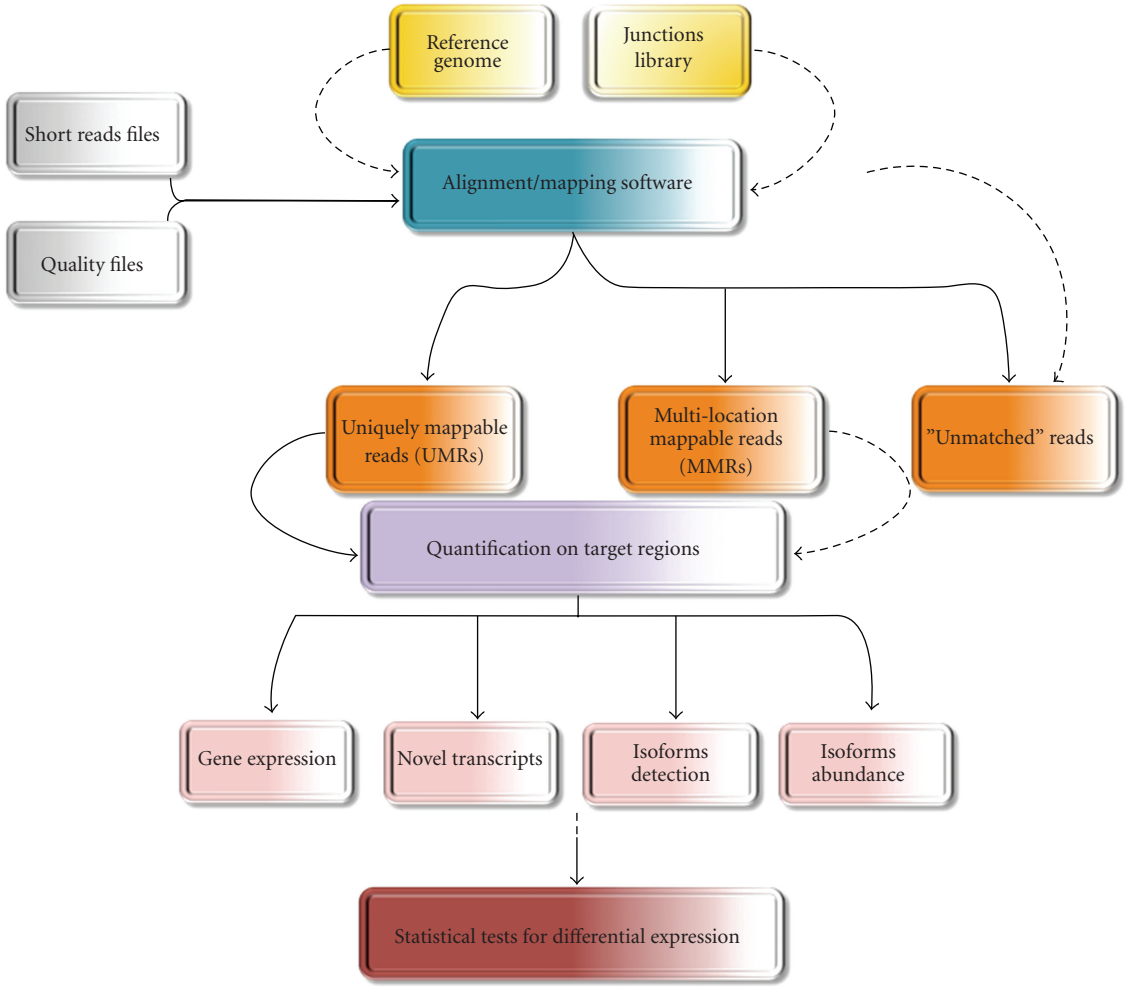


Figure 3: RNA-Seq computational pipeline.

not surprising that such nodal point still constitutes a very prominent area of research (see, e.g., [65–67] for a review) and has produced a great number of different algorithms in the last couple of years (e.g., [68–78]). Clearly, not all of them completely support the available platforms or are scalable for all amount of throughput or genome size. Nevertheless, the sequencing technologies are still in a developing phase with a very fast pace of increase in throughput, reads length and data formats after few months. Consequently, the already available mapping/assembly software are continuously under evolution in order to adapt themselves to the new data formats, to scale with the amount of data and to reduce their computational demand. New softwares are also continuously complementing the panorama. Moreover, the alignment phase of reads from RNA-Seq experiments presents many other subtleties to be considered; standard mapping algorithms are not able to fully exploit the complexity of the transcriptome, requiring to be modified or adapted in order to account for splicing events in eucaryotes.

The easiest way to handle such difficulty is to map the reads directly on known transcribed sequences, with the obvious drawback of missing new transcripts. Alternatively, the reads can be mapped continuously to the genome, but with the added opportunity of mapping reads that cross splice junctions. In this case, the algorithms differ from whether they require or not junctions's model. Algorithms such as Erange [37] or RNA-mate [79] require library of junctions constructed using known splice junctions extracted from data-bases and also supplemented with any set of putative splice junctions obtained, for instance, using a combinatorial approach on genes' model or ESTs sequences. Clearly, such approaches do not allow to map junctions not previously assembled in the junctions' library. On the other hand, algorithms like the WT [69], QPALMA [80], TopHat [81], G.Mo.R-Se [63], and PASS [78] potentially allow to detect new splice isoforms, since they use a more sophisticated mapping strategy. For instance, WT [69] splits the reads in left and right pieces, aligns each part to the genome, then attempts to extend each alignment on the other side to detect the junction. Whereas TopHat [81] first maps the reads against the whole reference genome using [77], second aggregates the mapped reads in islands of candidate exons on which compute a consensus measure, then generates potential donor/acceptor splice sites using

neighboring exons, and finally tries to align the reads, unmapped to the genome, to these splice junction sequences.

Most of the RNA-Seq packages are built on top of optimized short read *core* mappers [68, 69, 72, 77] and the mapping strategy is carried out by performing multiple runs or cycles. At the end of each cycle the unmatched reads are trimmed from one extreme and another step of alignment is attempted (see, e.g., [79]). Specific tolerances can be set for each alignment in order to increase the amount of mappable data. Obviously the simplest *core* approach is to map the sequence reads across the genome allowing the user to specify only the number of tolerated mismatches, although other methods allow to use also gapped alignment. Such flexibility can be beneficial for the rest of the analysis since both sequencing errors, that usually increase with the length of the sequence, and SNPs may cause substitutions and insertion/deletion of nucleotides in the reads. On the other hand, increasing the mapping flexibility also introduces a higher level of noise in the data. The compromise between the number of mapped reads and the quality of the resulting mapping is a very time consuming process without an optimal solution.

At the end of the mapping algorithm one can distinguish between three types of reads: reads that map uniquely to the genome or to the splice junctions (Uniquely Mappable Reads, UMR), reads with multiple (equally or similarly likely) locations either to the genome or to the splice junctions (Multilocation Mappable Reads, MMR) and reads without a specific mapping location. MMRs arise predominantly from conserved domains of paralogous gene families and from repeats. The fraction of mappable reads that are MMRs depends on the length of the read, the genome under investigation, and the expression in the individual sample; however it is typically between 10–40% for mammalian derived libraries [30, 37]. Most of the studies [28, 34] usually discarded MMRs from further analysis, limiting the attention only to UMRs. Clearly, this omission introduces experimental bias, decreases the coverage and reduces the possibility of investigating expressed regions such as active retrotransposons and gene families. An alternative strategy for the removal of the MMRs is to probabilistically assign them to each genomic location they map to. The simplest assignment considers equal probabilities. However, far better results have been obtained using a guilt-by-association strategy that calculates the probability of a MMRs originating from a particular locus. In [82], the authors proposed to proportionally assign MMRs to each of their mapping locations based on unique coincidences with either UMRs and other MMRs. Such a technique was later adopted in [79]. By contrast, in [83], the authors computed the probability as the ratio between the number of UMRs occurring in a nominal window surrounding each locus occupied by the considered MMR and the total number of UMRs proximal to all loci associated with that MMR. Similarly, in [37] the MMRs were fractionally assigned to their different possible locations considering the expression levels of their respective gene models. All these rescue strategies lead to substantially higher transcriptome coverage and give expression estimates in better agreement with microarrays than those using only UMRs (see, [37, 83]). Very recently, a more sophisticated approach was proposed in [84]. The authors introduced latent random variables representing the true mappings, with the parameters of the graphical model corresponding to isoform expression levels, read distributions across transcripts, and sequencing error. They allocated MMRs by maximizing the likelihood of the expression levels using an Expectation-Maximization (EM) algorithm. Additionally, they also showed that previous rescue methods introduced in [37, 82] are roughly equivalent to one iteration of EM. Independently on the specific proposal, we observe that all the above mentioned techniques work much better with data that preserve RNA strandedness. Alternatively, the use of paired-end protocols should help to alleviate the MMRs problem. Indeed, when one of the paired reads maps to a highly repetitive element in the genome but the second does not, it allows both reads to be unambiguously mapped to the reference genome. This is accomplished by first matching the first nonrepeat read uniquely to a genomic position and then looking within a size window, based on the known size range of the library fragments, for a match for the second read. The usefulness of this approach was demonstrated to improve read matching from 85% (single reads) to 93% (paired reads) [70], allowing a significant improvement in genome coverage, particularly in repeat regions. Currently, all of the next generation sequencing technologies are capable for generating data from paired-end reads, but unfortunately, till now only few RNA-Seq software support the use of paired-end reads in conjunction with the splice junctions mapping.

One of the possible reasons for reads not mapping to the genome and splice junctions is the presence of higher sequencing errors in the sequence. Other reasons can be identified in higher polymorphisms, insertion/deletion, complex exon-exon junctions, miRNA and small ncRNA: such situations could potentially be recovered by more sophisticated or combined alignment strategy.

Once mapping is completed, the user can display and explore the alignment on a genome browser (see Figure 4 for a screen-shot example) such as UCSC Genome Browser [85] (http://genome.ucsc.edu/) or the Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv), or on specifically devoted browsers such as EagleView [86], MapView [87] or Tablet [88], that can provide some highly informative views of the results at different levels of aggregations. Such tools allow to incorporate the obtained alignment with database annotations and other source of information, to observe specific polymorphism against sequence error, to identify well documented artifacts due to the DNA amplifications, as well as to detect other source of problems such as the not uniformity of the reads coverage across the transcript. Unfortunately, in many cases the direct visualization of the data is hampered by the lack of a common format for the alignment algorithm, causing a tremendous amount of extra work in format conversion for visualization purposes, feature extraction and other downstream analysis. Only recently, the SAM (Sequencing Alignment/Map) format [89] has been proposed as a possible standard for storing read alignment against reference sequences.

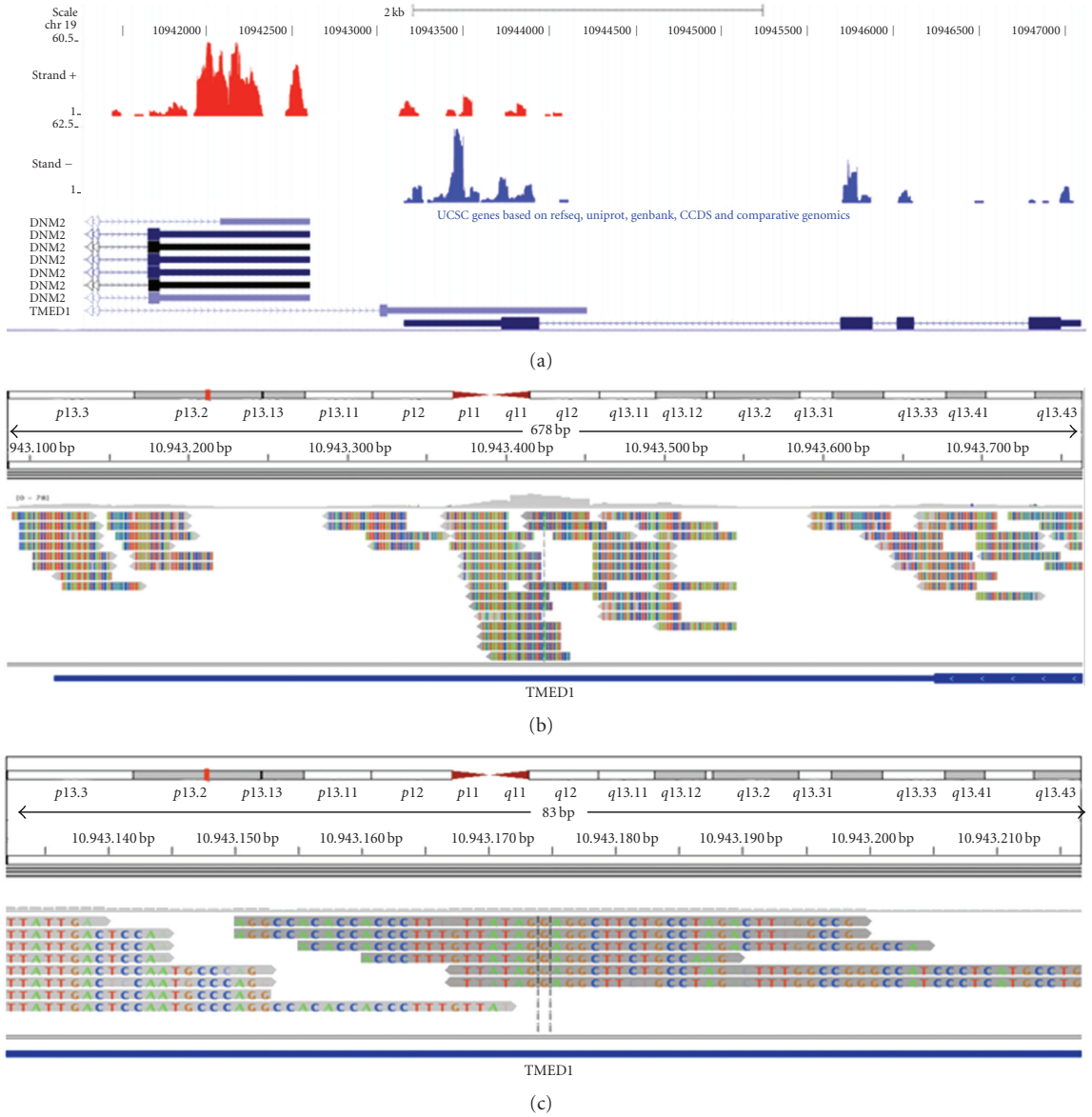


Figure 4: *Strand-Specific Read Distribution in UCSC Genome Browser and IGV.* (a) UCSC Genome Browser showing an example of stranded sequences generated by RNA-Seq experiment on NGS platform. In particular, the screenshot—of a characteristic "tail to tail" orientation of two human genes—clearly shows the specific expression in both strands where these two genes overlap, indicating that the strandedness of reads is preserved. (b) The same genomic location in the IGV browser, showing the reads (coloured blocks) distribution along TMED1 gene. The grey arrows indicate the sense of transcription. The specific expression in both strands where the genes overlap, indicates that the strandedness of reads is preserved. In (c) a greater magnification of the reads mapping to the same region at nucleotide level, useful to SNP analysis. The chromosome positions are shown at the top and genomic loci of the genes are shown at the bottom of each panel.

3.5. Quantifying Gene Expression and Isoforms' Abundance. Browser-driven analyses are very important for visualizing the quality of the data and to interpret specific events on the basis of the available annotations and mapped reads. However they only provide a qualitative picture of the phenomenon under investigation and the enormous amount of data does not allow to easily focus on the most relevant details. Hence, the second phase of most of the RNA-Seq pipeline consists of the automatic quantification of the transcriptional events across the entire genome (see Figure 4). From this point of view the interest is both quantifying known elements (i.e., genes or exons already annotated) and detecting new transcribed regions, defined as transcribed segments of DNA not yet annotated as exons in databases. The ability to detect these unannotated regions, even though biologically relevant, is one of the main advantages of the RNA-Seq over microarray technology. Usually, the quantification step is preliminary to any differential expression approach, see Figure 5.

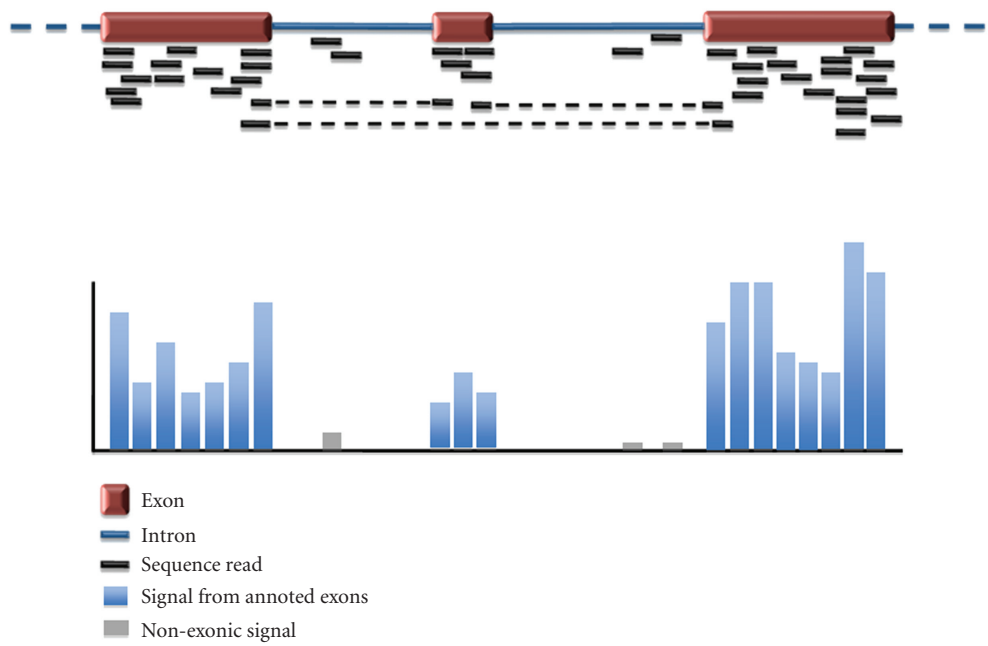


Figure 5: *Mapping and quantification of the signal.* RNA-seq experiments produce short reads sequenced from processed mRNAs. When a reference genome is available the reads can be mapped on it using efficient alignment software. Classical alignment tools will accurately map reads that fall within an exon, but they will fail to map spliced reads. To handle such problem suitable mappers, based either on junctions library or on more sophisticated approaches, need to be considered. After the mapping step annotated features can be quantified.

In order to derive a quantitative expression for annotated elements (such as exons or genes) within a genome, the simplest approach is to provide the expression as the total number of reads mapping to the coordinates of each annotated element. In the classical form, such method weights all the reads equally, even though they map the genome with different stringency. Alternatively, gene expression can be calculated as the sum of the number of reads covering each base position of the annotated element; in this way the expression is provided in terms of base coverage. In both cases, the results depend on the accuracy of the used gene models and the quantitative measures are a function of the number of mapped reads, the length of the region of interest and the molar concentration of the specific transcript. A straightforward solution to account for the sample size effect is to normalize the observed counts for the length of the element and the number of mapped reads. In [37], the authors proposed the *Reads Per Kilobase per Million* of mapped reads (RPKM) as a quantitative normalized measure for comparing both different genes within the same sample and differences of expression across biological conditions. In [84], the authors considered two alternative measures of relative expression: the fraction of transcripts and the fraction of nucleotides of the transcriptome made up by a given gene or isoform.

Although apparently easy to obtain, RPKM values can have several differences between software packages, hidden at first sight, due to the lack of a clear documentation of the analysis algorithms used. For example ERANGE [37] uses a union of known and new exon models to aggregate reads and determines a value for each region that includes spliced reads and assigned multireads too, whereas [30, 40, 81, 90] are restricted to known or prespecified exons/gene models. However, as noticed in [91], several experimental issues influence the RPKM quantification, including the integrity of the input RNA, the extent of ribosomal RNA remaining in the sample, the size selection steps and the accuracy of the gene models used.

In principle, RPKMs should reflect the true RNA concentration; this is true when samples have relatively uniform sequence coverage across the entire gene model. The problem is that all protocols currently fall short of providing the desired uniformity, see for example [37], where the Kolmogorov-Smirnov statistics is used to compare the observed reads distribution on each selected exon model with the theoretical uniform one. Similar conclusions are also illustrated in [57, 58], among others.

Additionally, it should be noted that RPKM measure should not be considered as the panacea for all RNA-Seq experiments. Despite the importance of the issue, the expression quantification did not receive the necessary attention from the community and in most of the cases the choice has been done regardless of the fact that the main question is the detection of differentially expressed elements. Regarding this point in [92] it is illustrated the inherent bias in transcript length that affect RNA-Seq experiments. In fact the total number of reads for a given transcript is roughly proportional to both the expression level and the length of the transcript. In other words, a long transcript will have more reads mapping to it compared to a short gene of similar expression. Since the power of an experiment is proportional to the sampling size, there will be more statistical power

to detect differential expression for longer genes. Therefore, short transcripts will always be at a statistical disadvantage relative to long transcripts in the same sample. RPKM-type measures provide an expression level normalized by the length of the gene and this only apparently solves the problem; it gives an unbiased measure of the expression level, but also changes the variance of the data in a length dependent manner, resulting in the same bias to differential expression estimation. In order to account for such an inherent bias, in [92] the authors proposed to use a fixed length window approach, with a window size smaller than the smallest gene. This method can calculate aggregated tag counts for each window and consequently assess them for differential expression. However, since the analysis is performed at the window level some proportion of the data will be discarded; moreover such an approach suffers for a reduced power and highly expressed genes are more likely to be detected due to the fact that the sample variance decreases with the expression level. Indeed, it should be noticed that the sample variance depends on both the transcript length and the expression level.

Finally, we observe that annotation files are often inaccurate; boundaries are not always mapped precisely, ambiguities and overlaps among transcripts often occur and are not yet completely solved. Concerning this issue in [93] the authors proposed a method based on the definition of "union-intersection genes" to define the genomic region of interest and normalized absolute and relative expression measures within. Also, in this case we observe that all strategies work much better with data that preserve RNA strandedness, which is an extremely valuable information for transcriptome annotation, especially for regions with overlapping transcription from opposite directions.

The quantification methods described above do not account for new transcribed region. Although several studies have already demonstrated that RNA-Seq experiments, with their high resolution and sensitivity have great potentiality in revealing many new transcribed regions, unidentifiable by microarrays, the detection of new transcribed regions is mainly obtained by means of a sliding window and heuristic approaches. In [94] stretches of contiguous expression in intergenic regions are identified after removing all UTRs from the intergenic search space by using a combination of information arising from tiling-chip and sequence data and visual inspection and manual curation. The procedure is quite complex and is mainly due to the lack of strandedness information in their experiment. On the contrary, the hybridization data are less affected by these issues because they distinguish transcriptional direction and do not show any 5′ bias (see [94] for further details). Then, new transcribed regions are required to have a length of at least 70 bp and an average sequence coverage of 5 reads per bp. A similar approach, with different choices of the threshold and the window, was proposed in [40], where the authors investigated either intergenic and intronic regions. The choices of the parameters are assessed by estimating noise levels by means of a Poisson model of the noncoding part of the genome. In [45] the whole genome is split into 50 bp windows (non-overlapping). A genomic region is defined as a new transcribed region if it results from the union of two consecutive windows, with at least two sequence reads mapped per window. Additionally, the gap between each new transcribed regions should be at least 50 bp, and the gap between a new transcribed region and an annotated gene (with the same strand) at least 100 bp. A slightly more sophisticated approach is used in ERANGE [37]. Reads that do not fall within known exons are aggregated into candidate exons by requiring regions with at least 15 reads, whose starts are not separated by more than 30 bp. Most of the candidate exons are assigned to neighboring gene models when they are within a specifiable distance of the model.

These studies, among others, reveal many of these new transcribed regions. Unfortunately, most of them do not seem to encode any protein, and hence their functions remain often to be determined. In any case, these new transcribed regions, combined with many undiscovered new splicing variants, suggest that there is considerably more transcript complexity than previously appreciated. Consequently further RNA-Seq experiments and more sophisticated analysis methods can disclose it.

The complexity of mammalian transcriptomes is also compounded by alternative splicing which allows one gene to produce multiple transcript isoforms. Alternative splicing includes events such as exon skipping, alternative 5′ or 3′ splicing, mutually exclusive exons, intron retention, and "cryptic" splice sites (see Figure 6). The frequency of occurrence of alternative splicing events is still underestimated. However it is well known that multiple transcript isoforms produced from a single gene can lead to protein isoforms with distinct functions, and that alternative splicing is widely involved in different physiological and pathological processes. One of the most important advantages of the RNA-Seq experiments is the possibility of understanding and comparing the transcriptome at the isoform level (see [95, 96]). In this context, two computational problems need to be solved: the detection of different isoforms and their quantification in terms of transcript abundance.

Initial proposals for solving these problems were essentially based on a gene-by-gene manual inspection usually focusing the attention to the detection of the presence of alternative splicing forms rather than to their quantification. For example, the knowledge of exon-exon junction reads and of junctions that fall into some isoform-specific regions can provide useful information for identifying different isoforms. The reliability of a splicing junction is usually assessed by counting features like the number of reads mapping to the junction, the number of mismatches on each mapped read, the mapping position on the junction and the mismatches location in a sort of heuristic approach. Unfortunately, these techniques cannot be scaled to the genome level and they are affected by a high false positive and false negative rate.

Following the above mentioned ideas, in [40] the authors detected junctions by computing the probability of a random hits for a read of length R on the splice junctions of length J with at most a certain number of mismatches. In [95], the authors used several information similar to those described above to train classifiers based on logistic regression for splicing junction detection. In [97], the authors introduced

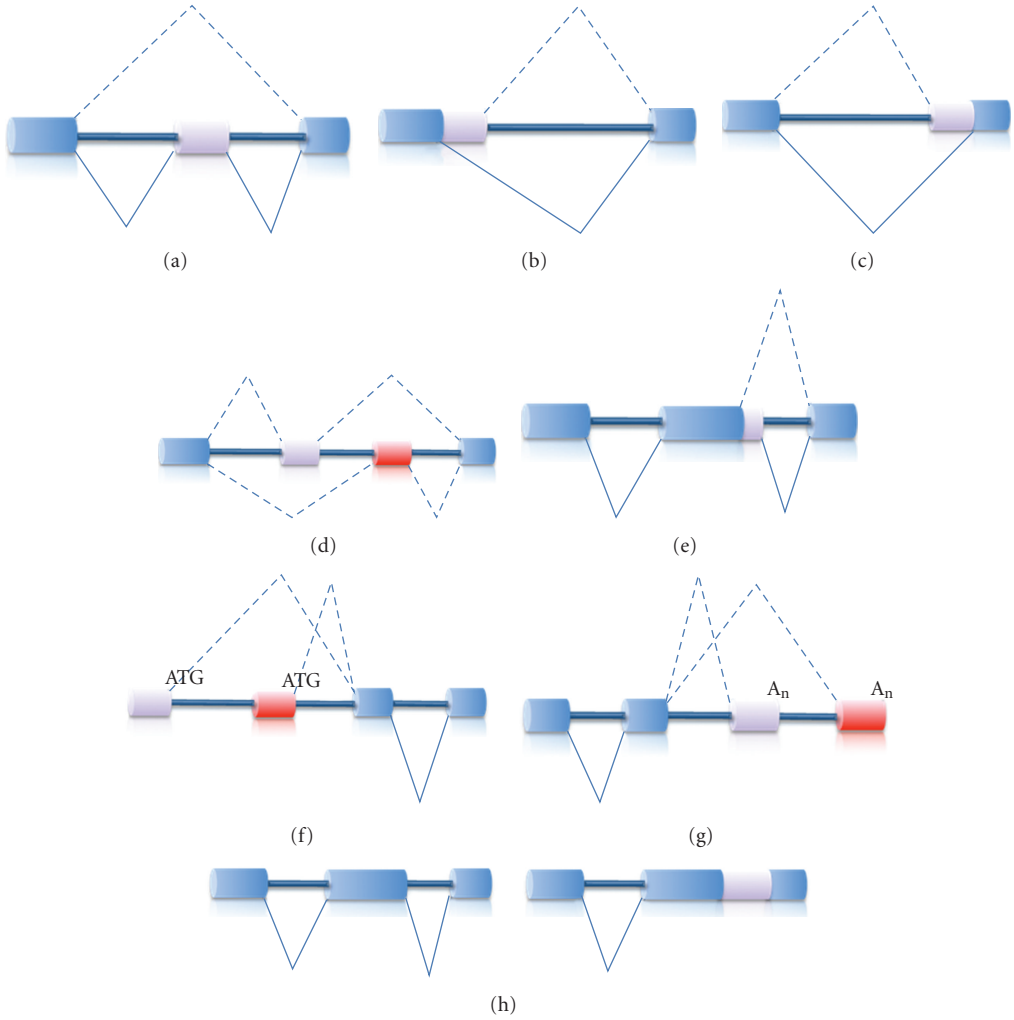


Figure 6: *Alternative splicing*. Schematic representation of the possible patterns of alternative splicing of a gene. Boxes are discrete exons that can be independently included or excluded from the mRNA transcript. Light blue boxes represent constitutive exons, violet and red boxes are alternatively spliced exons. Dashed lines represent alternative splicing events. (a) Canonical exon skipping; (b) 5′ or (c) 3′ alternative splicing; (d) Mutually exclusive splicing event involving the selection of only one from two or more exon variants; (e) Intra-exonic "cryptic" splice site causing the exclusion of a portion of the exon from the transcript; (f) Usage of new alternative 5′ or (g) 3′ exons; (h) Intron retention.

a new metric to measure the quality of each junction read. Then they estimated the distribution of such metric either with respect to known exon splice junctions and random splice junctions, and implemented an empirical statistical model to detect exon junctions evaluating the probability that an observed alignment distribution comes from a true junction.

The simple detection of specific isoforms does not provide useful information about their quantitative abundance. In principle, the quantification methods described above are equally applicable to quantify isoform expression. In practice, however, it is difficult to compute isoform-specific expression because most reads that are mapped to the genes are shared by more than one isoform and then it becomes difficult to assign each read only to a specific isoform. As a consequence, the assignment should rely on inferential methods that consider all data mapping to a certain region.

Several proposed methods for inferring isoforms' abundance are based on the preliminary knowledge of precise isoforms' annotation, on the assumption of uniform distribution of the reads across the transcript, on Poisson model for the reads' counts and equal weight for each read, regardless the quality of the match. The methods are often limited to handle only the cases where there is a relative small number of isoforms without confounding effects due to the overlap between genes. In particular in [98], the authors showed that the complexity of some isoform sets may still render the estimation problem nonidentifiable based on current RNA-Seq protocols and derived a mathematical characterization of identifiable isoform set. The main reason for such an effect is that current protocols with short single-end reads RNA-Seq are only able to asses local properties of a transcript. It is possible that the combination of short-read data with longer reads or paired-end reads will be able to go further in addressing such challenges.

Recently, in [90] the authors proposed a statistical method where, similar to [34], the count of reads falling into an annotated gene with multiple isoforms is modeled as a Poisson variable. They inferred the expression of each individual isoform using maximum likelihood approach, whose solution has been obtained by solving a convex optimization problem. In order to quantify the degree of uncertainty of the estimates, they carried out statistical inferences about the parameters from the posterior distribution by importance sampling. Interestingly, they showed that their method can be viewed as an extension of the RPKM concept and reduces to the RPKM index when there is only one isoform. An attempt to relax the assumption of uniform reads sampling is proposed in [84]. In this paper, the authors unified the notions of reads that map to multiple locations, that is, that could be potentially assigned to several genes, with those of reads that map to multiple isoforms through the introduction of latent random variables representing the true mappings. Then, they estimated the isoforms' abundance as the maximum likelihood expression levels using the EM algorithm. The Poisson distribution is also the main assumption in [99], where a comprehensive approach to the problem of alternative isoforms prediction is presented. In particular, the presence of alternative splicing event within the same sample is assessed by using Pearson's chi-square test on the parameter of a multinomial distribution and the EM algorithm is used to estimate the abundance of each isoform.

3.6. Differential Expression. The final goal in the majority of transcriptome studies is to quantify differences in expression across multiple samples in order to capture differential gene expression, to identify sample-specific alternative splicing isoforms and their differential abundance.

Mimicking the methods used for microarray analysis, researchers started to approach such crucial question using statistical hypothesis' tests combined with multiple comparisons error procedures on the observed counts (or on the RPKM values) at the gene, isoform or exon level. Indeed, in [30] the authors applied the empirical Bayes moderated t-test proposed in [100] to the normalized RPKM. However in microarray experiments, the abundance of a particular transcript is measured as a fluorescence intensity, that can be effectively modeled as a continuous response, whereas for RNA-Seq data the abundance is usually a count. Therefore, procedures that are successful for microarrays do not seem to be appropriate for dealing with such type of data.

One of the pioneering works to handle such difference is [34], where the authors modeled the aggregated reads count for each gene using Poisson distribution. One can prove that the number of reads observed from a gene (or transcript isoform) follows a binomial distribution that can be approximated by a Poisson distribution, under the assumption that RNA-Seq reads follow a random sampling process, in which each read is sampled independently and uniformly from every possible nucleotide in the sample. In this set-up, in [34] the authors used a likelihood ratio test to test for significant differences between the two conditions. The Poisson model was also employed by [40], where the authors used the method proposed in [101] to determine the significance of differential expression. On the contrary, in [83], the authors simply estimated the difference in expression of a gene between two conditions through the difference of the count proportions p_1 and p_2 computed using a classical Z-test statistics. In [18], the authors employed the Fishers exact test to better weigh the genes with relatively small counts. Similarly in [99] the authors used Poisson model and Fishers exact test to detect alternative exon usage between conditions.

Recently, more sophisticated approaches have been proposed in [102, 103]. In [102], the authors proposed an empirical Bayesian approach, based on the negative binomial distribution; it results very flexible and reduces to the Poisson model for a particular choice of the hyperparameter. They carried out differential expression testing using a moderated Bayes approach similar in the spirit to the one described in [100], but adapted for data that are counts. We observed that the method is designed for finding changes between two or more groups when at least one of the groups has replicated measurements. In [103], the observed counts of reads mapped to a specific gene obtained from a certain sample was modeled using Binomial distribution. Under such assumption, it can be proved that the log ratio between the two samples conditioned to the intensity signal (i.e., the average of the two logs counts) follows an approximate normal distribution, that is used for assessing the significance of the test. All the above-mentioned methods assume that the quantification of the features of interest under the experimental conditions has been already done and each read has been assigned to only one elements, hence the methods are directly applicable to detect genes or exons differences provided that overlapping elements are properly filtered out. By contrast the above described methods are not directly suited for detecting isoforms' differences unless the quantification of the isoform abundance has been carried out using specific approaches. To handle such difficulties, in [104], the authors proposed a hierarchical Bayesian model to directly infer the differential expression level of each transcript isoform in response to two conditions. The difference in expression of each isoform is modeled by means of an inverse gamma model and a latent variable is introduced for guiding the isoform's selection. The model can handle the heteroskedasticity of the sequence read coverage and inference is carried out using Gibbs sampler.

It should be noticed that although these techniques already provide interesting biological insights, they have not been sufficiently validated on several real data-sets where different type of replicates are available, neither sufficiently compared each others in terms of advantages and disadvantages. As with any new biotechnology it is important to carefully study the different sources of variation that can affect measure of the biological effects of interest and to statistically asses the reproducibility of the biological findings in a rigorous way, and to date this has been often omitted. Indeed, it should be considered that there are a variety of experimental effects that could possibly increase the variability, the bias, or be confounded with sequencing-based measures, causing miss-understanding of the results.

Unfortunately, such problems have received little of attention until now. In order to fill this gap, in [93] the authors presented a statistical inference framework for transcriptome analysis using RNA-Seq mapped read data. In particular, they proposed a new statistical method based on log-linear regression for investigating relationships between read counts and biological and experimental variables describing input samples as well as genomic regions of interest. The main advantage of the log-linear regression approach is that it allows to account both for biological effect and a variety of experimental effects. Their paper represents one of the few attempts of looking at the analysis of RNA-Seq data from a general point of view.

4. Challenges and Perspective for NGS

From the development of the Sanger method to the completion of the HGP, genetics has made significant advances towards the understanding of gene content and function. Even though significant achievements were reached by Human Genome, HapMap and ENCODE Projects [7, 105, 106], we are far from an exhaustive comprehension of the genomic diversity among humans and across the species, and from understanding gene expression variations and its regulation in both physio and pathological conditions. Since the appearance of first NGS platforms in the 2004, it was clear that understanding this diversity at a cost of around $5–10 million per genome sequence [107], placed it outside the real possibilities of most research laboratories, and very far from single individual economical potential. To date, we are in the "$1,000 genome" era, and, although this important barrier has not yet been broken, its a current assumption that this target is going to be reached within the end of 2010. It is likely that the rapid evolution of DNA sequencing technology, able to provide researchers with the ability to generate data about genetic variation and patterns of gene expression at an unprecedented scale, will become a routine tool for researchers and clinicians within just a few years.

As we can see, the number of applications and the great amount of biological questions that can be addressed by "Seq" experiments on NGS platforms is leading a revolution in the landscape of molecular biology, but the imbalance between the pace at which technology innovations are introduced in the platforms and the biological discoveries derivable from them is growing up. The risk is the creation of a glut of "under-used" information that in few months becomes of no use because the new one is produced. It is necessary to invest in an equivalent development of new computational strategies and expertise to deal with the volumes of data created by the current generation of new sequencing instruments, to maximize their potential benefit.

These platforms are creating a new world to explore, not only in the definition of experimental/technical procedures of large-scale analyses, but also in the downstream computational analysis and in the bioinformatics infrastructures support required for high-quality data generation and for their correct biological interpretation. In practice, they have shifted the bottleneck from the generation of experimental data to their management and to their statistical and computational analysis. There are few key points to consider. The first one is the data management: downstream computational analysis becomes difficult without appropriate Information Technology (IT) infrastructure. The terabytes of data produced by each sequencing run requires conspicuous storage and backup capacity, which increases considerably the experimental costs. The second one regards the protocols used for the production of raw data: each platform has its peculiarity in both sample preparation and type and volume of raw data produced, hence they require individualized laboratory expertise and data processing pipelines. Third, beside vendor specific and commercial software, several other open-source analysis tools are continuously appearing. Unfortunately, there is often an incomplete documentation and it is easy to spend more time in evaluating software suites than in analyzing the output data. Whichever software is used, the most important question is to understand its limitations and assumptions. Community adoption of input/output data standards is also essential to efficiently handle the data management problem. Till now the effort has been mainly devoted to the technological development rather than to the methodological counterpart. The choice of a careful experimental design has been also not always adequately considered.

As regards the RNA-Seq, we have still to face several critical issues either from a biological and computational point of view. RNA-seq protocols are extremely sensitive and need a very careful quality control for each wet laboratory step. For instance, the contamination of reagents with RNAse and the degradation of RNA, even partial, must be avoided during all the technical procedures. The quality of total isolated RNA is the first, and probably the most crucial point for an RNA-Seq experiment. Poor yield of polyA enrichment or low efficiency of total RNA ribodepletion are also critical issues for preparing high-quality RNA towards the library construction. It is clear that, independently on the library construction procedure, particular care should be taken to avoid complete degradation of RNA during the controlled RNA fragmentation step. Furthermore, in order to correctly determine the directionality of gene transcription and to facilitate the detection of opposing and overlapping transcripts within gene-dense genomic regions, particular care should be taken to preserve the strandedness of RNA fragments during the library preparation. In addition, to provide a more uniform coverage throughout the transcript length, random priming for reverse transcription protocols, rather than oligo dT priming (with the bias of low coverage at the 5′ ends), should be done after removal of rRNA. Finally, it should be considered that for the platforms based on CRT and SBL, substitutions and under representation of AT-rich and GC-rich regions, probably due to amplification bias during template preparation, are the most common error type. In contrast, for pyrosequencing platforms, insertions and deletions represent a common drawback.

For what concern the data analysis, to the above-mentioned points, we should note that most of the available software for read alignment are designed for genomic

mapping hence they are not fully capable to discover exon junctions. The classical extension for handling RNA-Seq data involves the preconstruction of junction libraries reducing the possibility of discovering new junctions. It would be desirable to develop new methods that allow either new junction detection and also the use of paired-end reads, that are particularly promising for more accurate study. Additionally further developments are required to assess the significance of new transcribed regions, the construction of new putative genes and the precise quantification of each isoform, for which there is still a lack of statistical methodologies. For what concerns the detection of differential expression, existing techniques were not sufficiently validated on biological data and compared in terms of specificity and sensitivity. Moreover, of potentially great impact, is the lack of biological replicates which precludes gauging the magnitude of individual effects in relation to technical effects. Biological replicates is essential in a RNA-Seq experiment to draw generalized conclusions about the "real" differences observed between two or more biological groups.

Facing such multidisciplinary challenges will be the key point for a fruitful transfer from laboratory studies to clinical applications. Indeed, the availability of low-cost, efficient and accurate technologies for gene expression and genome sequencing will be useful in providing pathological gene expression profiles in a wide number of common genetic disorders including type II diabetes, cardiovascular disease, Parkinson disease and Downs syndrome. Moreover, the application of NGS to the emerging disciplines of pharmacogenomics and nutrigenomics will allow to understand drug response and nutrient-gene interactions on the basis of individual patient's genetic make-up, leading in turn to the development of targeted therapies for many human diseases or tailored nutrient supplementation [108].

Acknowledgment

We are grateful to the anonymous referees whose valuable comments helped to substantially improve the paper. This work was supported by the CNR-Bioinformatics Project.

References

[1] D. D. Licatalosi and R. B. Darnell, "RNA processing and its regulation: global insights into biological networks," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 75–87, 2010.

[2] V. E. Velculescu, L. Zhang, W. Zhou, et al., "Characterization of the yeast transcriptome," *Cell*, vol. 88, no. 2, pp. 243–251, 1997.

[3] J. Lindberg and J. Lundeberg, "The plasticity of the mammalian transcriptome," *Genomics*, vol. 95, no. 1, pp. 1–6, 2010.

[4] W. F. Doolittle and C. Sapienza, "Selfish genes, the phenotype paradigm and genome evolution," *Nature*, vol. 284, no. 5757, pp. 601–603, 1980.

[5] R. J. Taft, M. Pheasant, and J. S. Mattick, "The relationship between non-protein-coding DNA and eukaryotic complexity," *BioEssays*, vol. 29, no. 3, pp. 288–299, 2007.

[6] T. Cavalier-Smith, "Cell volume and the evolution of eukaryote genome size," in *The Evolution of Genome Size*, T. Cavalier-Smith, Ed., pp. 105–184, John Wiley & Sons, Chichester, UK, 1985.

[7] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, no. 7146, pp. 799–816, 2007.

[8] A. Jacquier, "The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs," *Nature Reviews Genetics*, vol. 10, no. 12, pp. 833–844, 2009.

[9] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[10] R. W. Holley, "Alanine transfer RNA," in *Nobel Lectures in Molecular Biology 1933–1975*, pp. 285–300, Elsevier North Holland, New York, NY, USA, 1977.

[11] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 2, pp. 560–564, 1977.

[12] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.

[13] E. R. Mardis, "Next-generation DNA sequencing methods," *Annual Review of Genomics and Human Genetics*, vol. 9, pp. 387–402, 2008.

[14] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[15] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[16] R. A. Irizarry, D. Warren, F. Spencer, et al., "Multiple-laboratory comparison of microarray platforms," *Nature Methods*, vol. 2, no. 5, pp. 345–349, 2005.

[17] P. A. C. 't Hoen, Y. Ariyurek, H. H. Thygesen, et al., "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms," *Nucleic Acids Research*, vol. 36, no. 21, article e141, 2008.

[18] J. S. Bloom, Z. Khan, L. Kruglyak, M. Singh, and A. A. Caudy, "Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays," *BMC Genomics*, vol. 10, article 221, 2009.

[19] M. Harbers and P. Carninci, "Tag-based approaches for transcriptome research and genome annotation," *Nature Methods*, vol. 2, no. 7, pp. 495–502, 2005.

[20] M. P. Horan, "Application of serial analysis of gene expression to the study of human genetic disease," *Human Genetics*, vol. 126, no. 5, pp. 605–614, 2009.

[21] H. Misu, T. Takamura, N. Matsuzawa, et al., "Genes involved in oxidative phosphorylation are coordinately upregulated with fasting hyperglycaemia in livers of patients with type 2 diabetes," *Diabetologia*, vol. 50, no. 2, pp. 268–277, 2007.

[22] T. Takamura, H. Misu, T. Yamashita, and S. Kaneko, "SAGE application in the study of diabetes," *Current Pharmaceutical Biotechnology*, vol. 9, no. 5, pp. 392–399, 2008.

[23] D. V. Gnatenko, J. J. Dunn, S. R. McCorkle, D. Weissmann, P. L. Perrotta, and W. F. Bahou, "Transcript profiling of human platelets using microarray and serial analysis of gene expression," *Blood*, vol. 101, no. 6, pp. 2285–2293, 2003.

[24] C. A. Sommer, E. C. Pavarino-Bertelli, E. M. Goloni-Bertollo, and F. Henrique-Silva, "Identification of dysregulated genes in lymphocytes from children with Down syndrome," *Genome*, vol. 51, no. 1, pp. 19–29, 2008.

[25] W. Malagó Jr., C. A. Sommer, C. Del Cistia Andrade, et al., "Gene expression profile of human Down syndrome leukocytes," *Croatian Medical Journal*, vol. 46, no. 4, pp. 647–656, 2005.

[26] B. T. Wilhelm and J.-R. Landry, "RNA-Seq-quantitative measurement of expression through massively parallel RNA-Sequencing," *Methods*, vol. 48, no. 3, pp. 249–257, 2009.

[27] M. N. Bainbridge, R. L. Warren, M. Hirst, et al., "Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach," *BMC Genomics*, vol. 7, article 246, 2006.

[28] U. Nagalakshmi, Z. Wang, K. Waern, et al., "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, 2008.

[29] T. T. Torres, M. Metta, B. Ottenwälder, and C. Schlötterer, "Gene expression profiling by massively parallel sequencing," *Genome Research*, vol. 18, no. 1, pp. 172–177, 2008.

[30] N. Cloonan, A. R. R. Forrest, G. Kolle, et al., "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.

[31] L. J. Core, J. J. Waterfall, and J. T. Lis, "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters," *Science*, vol. 322, no. 5909, pp. 1845–1848, 2008.

[32] S.-I. Hashimoto, W. Qu, B. Ahsan, et al., "High-resolution analysis of the 5′-end transcriptome using a next generation DNA sequencer," *PLoS ONE*, vol. 4, no. 1, article e4108, 2009.

[33] H. Li, M. T. Lovci, Y.-S. Kwon, M. G. Rosenfeld, X.-D. Fu, and G. W. Yeo, "Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20179–20184, 2008.

[34] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[35] R. D. Morin, M. D. O'Connor, M. Griffith, et al., "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, no. 4, pp. 610–621, 2008.

[36] R. D. Morin, M. Bainbridge, A. Fejes, et al., "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing," *BioTechniques*, vol. 45, no. 1, pp. 81–94, 2008.

[37] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[38] R. Rosenkranz, T. Borodina, H. Lehrach, and H. Himmelbauer, "Characterizing the mouse ES cell transcriptome with Illumina sequencing," *Genomics*, vol. 92, no. 4, pp. 187–194, 2008.

[39] D. J. Sugarbaker, W. G. Richards, G. J. Gordon, et al., "Transcriptome sequencing of malignant pleural mesothelioma tumors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 9, pp. 3521–3526, 2008.

[40] M. Sultan, M. H. Schulz, H. Richard, et al., "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008.

[41] Y. W. Asmann, E. W. Klee, E. A. Thompson, et al., "3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer," *BMC Genomics*, vol. 10, article 531, 2009.

[42] I. Chepelev, G. Wei, Q. Tang, and K. Zhao, "Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq," *Nucleic Acids Research*, vol. 37, no. 16, article e106, 2009.

[43] J. Z. Levin, M. F. Berger, X. Adiconis, et al., "Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts," *Genome Biology*, vol. 10, no. 10, article R115, 2009.

[44] C. A. Maher, N. Palanisamy, J. C. Brenner, et al., "Chimeric transcript discovery by paired-end transcriptome sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 30, pp. 12353–12358, 2009.

[45] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, et al., "Transcriptome analysis by strand-specific sequencing of complementary DNA," *Nucleic Acids Research*, vol. 37, no. 18, article e123, 2009.

[46] T. E. Reddy, F. Pauli, R. O. Sprouse, et al., "Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation," *Genome Research*, vol. 19, no. 12, pp. 2163–2171, 2009.

[47] F. Tang, C. Barbacioru, Y. Wang, et al., "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[48] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad, "Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, vol. 20, no. 2, pp. 180–189, 2010.

[49] G. A. Heap, J. H. M. Yang, K. Downes, et al., "Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing," *Human Molecular Genetics*, vol. 19, no. 1, pp. 122–134, 2010.

[50] D. Raha, Z. Wang, Z. Moqtaderi, et al., "Close association of RNA polymerase II and many transcription factors with Pol III genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 8, pp. 3639–3644, 2010.

[51] S. Marguerat and J. Bahler, "RNA-Seq: from technology to biology," *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, 2010.

[52] Y. He, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, and K. W. Kinzler, "The antisense transcriptomes of human cells," *Science*, vol. 322, no. 5909, pp. 1855–1857, 2008.

[53] R. Lister, R. C. O'Malley, J. Tonti-Filippini, et al., "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," *Cell*, vol. 133, no. 3, pp. 523–536, 2008.

[54] B. T. Wilhelm, S. Marguerat, I. Goodhead, and J. Bahler, "Defining transcribed regions using RNA-Seq," *Nature Protocols*, vol. 5, no. 2, pp. 255–266, 2010.

[55] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling," *Science*, vol. 324, no. 5924, pp. 218–223, 2009.

[56] T. D. Harris, P. R. Buzby, H. Babcock, et al., "Single-molecule DNA sequencing of a viral genome," *Science*, vol. 320, no. 5872, pp. 106–109, 2008.

[57] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, vol. 36, no. 16, article e105, 2008.

[58] O. Harismendy, P. C. Ng, R. L. Strausberg, et al., "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome Biology*, vol. 10, no. 3, article R32, 2009.

[59] L. W. Hillier, G. T. Marth, A. R. Quinlan, et al., "Whole-genome sequencing and variant discovery in *C. elegans*," *Nature Methods*, vol. 5, no. 2, pp. 183–188, 2008.

[60] J. D. McPherson, "Next-generation gap," *Nature Methods*, vol. 6, no. 11S, pp. S2–S5, 2009.

[61] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821–829, 2008.

[62] I. Birol, S. D. Jackman, C. B. Nielsen, et al., "De novo transcriptome assembly with ABySS," *Bioinformatics*, vol. 25, no. 21, pp. 2872–2877, 2009.

[63] F. Denoeud, J.-M. Aury, C. Da Silva, et al., "Annotating genomes with massive-scale RNA sequencing," *Genome Biology*, vol. 9, no. 12, article R175, 2008.

[64] M. Yassoura, T. Kaplana, H. B. Fraser, et al., "Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 9, pp. 3264–3269, 2009.

[65] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes," *Nature Biotechnology*, vol. 27, no. 5, pp. 455–457, 2009.

[66] P. Flicek and E. Birney, "Sense from sequence reads: methods for alignment and assembly," *Nature Methods*, vol. 6, supplement 11, pp. S6–S12, 2009.

[67] D. S. Horner, G. Pavesi, T. Castrignanò, et al., "Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing," *Briefings in Bioinformatics*, vol. 11, no. 2, pp. 181–197, 2009.

[68] A. Cox, "ELAND: efficient local alignment of nucleotide data," unpublished, http://bioit.dbi.udel.edu/howto/eland.

[69] "Applied Biosystems mappread and whole transcriptome software tools," http://www.solidsoftwaretools.com/.

[70] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[71] A. D. Smith, Z. Xuan, and M. Q. Zhang, "Using quality scores and longer reads improves accuracy of Solexa read mapping," *BMC Bioinformatics*, vol. 9, article 128, 2008.

[72] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[73] R. Li, C. Yu, Y. Li, et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[74] B. D. Ondov, A. Varadarajan, K. D. Passalacqua, and N. H. Bergman, "Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications," *Bioinformatics*, vol. 24, no. 23, pp. 2776–2777, 2008.

[75] H. Jiang and W. H. Wong, "SeqMap: mapping massive amount of oligonucleotides to the genome," *Bioinformatics*, vol. 24, no. 20, pp. 2395–2396, 2008.

[76] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li, "ZOOM! Zillions of oligos mapped," *Bioinformatics*, vol. 24, no. 21, pp. 2431–2437, 2008.

[77] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[78] D. Campagna, A. Albiero, A. Bilardi, et al., "PASS: a program to align short sequences," *Bioinformatics*, vol. 25, no. 7, pp. 967–968, 2009.

[79] N. Cloonan, Q. Xu, G. J. Faulkner, et al., "RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data," *Bioinformatics*, vol. 25, no. 19, pp. 2615–2616, 2009.

[80] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," *Bioinformatics*, vol. 24, no. 16, pp. i174–i180, 2008.

[81] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[82] G. J. Faulkner, A. R. R. Forrest, A. M. Chalk, et al., "A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE," *Genomics*, vol. 91, no. 3, pp. 281–288, 2008.

[83] T. Hashimoto, M. J. L. de Hoon, S. M. Grimmond, C. O. Daub, Y. Hayashizaki, and G. J. Faulkner, "Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite," *Bioinformatics*, vol. 25, no. 19, pp. 2613–2614, 2009.

[84] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, no. 4, pp. 493–500, 2009.

[85] W. J. Kent, C. W. Sugnet, T. S. Furey, et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[86] W. Huang and G. Marth, "EagleView: a genome assembly viewer for next-generation sequencing technologies," *Genome Research*, vol. 18, no. 9, pp. 1538–1543, 2008.

[87] H. Bao, H. Guo, J. Wang, R. Zhou, X. Lu, and S. Shi, "MapView: visualization of short reads alignment on a desktop computer," *Bioinformatics*, vol. 25, no. 12, pp. 1554–1555, 2009.

[88] I. Milne, M. Bayer, L. Cardle, et al., "Tablet-next generation sequence assembly visualization," *Bioinformatics*, vol. 26, no. 3, pp. 401–402, 2010.

[89] H. Li, B. Handsaker, A. Wysoker, et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[90] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009.

[91] S. Pepke, B. Wold, and A. Mortazavi, "Computation for ChIP-Seq and RNA-Seq studies," *Nature Methods*, vol. 6, no. 11S, pp. S22–S32, 2009.

[92] A. Oshlack and M. J. Wakefield, "Transcript length bias in RNA-Seq data confounds systems biology," *Biology Direct*, vol. 4, article 14, 2009.

[93] J. H. Bullard, E. A. Purdom, K. D. Hansen, S. Durinck, and S. Dudoit, "Statistical inference in mRNA-Seq: exploratory data analysis and differential expression," Tech. Rep. 247/2009, University of California, Berkeley, 2009.
[94] B. T. Wilhelm, S. Marguerat, S. Watt, et al., "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution," *Nature*, vol. 453, no. 7199, pp. 1239–1243, 2008.
[95] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.
[96] E. T. Wang, R. Sandberg, S. Luo, et al., "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.
[97] L. Wang, Y. Xi, J. Yu, L. Dong, L. Yen, and W. Li, "A statistical method for the detection of alternative splicing using RNA-Seq," *PLoS ONE*, vol. 5, no. 1, article e8529, 2010.
[98] D. Hiller, H. Jiang, W. Xu, and W. H. Wong, "Identifiability of isoform deconvolution from junction arrays and RNA-Seq," *Bioinformatics*, vol. 25, no. 23, pp. 3056–3059, 2009.
[99] H. Richard, M. H. Schulz, M. Sultan, et al., "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments," *Nucleic Acids Research*, vol. 38, no. 10, p. e112, 2010.
[100] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.
[101] S. Audic and J.-M. Claverie, "The significance of digital gene expression profiles," *Genome Research*, vol. 7, no. 10, pp. 986–995, 1997.
[102] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
[103] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data," *Bioinformatics*, vol. 26, no. 1, pp. 136–138, 2009.
[104] S. Zheng and L. Chen, "A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level," *Nucleic Acids Research*, vol. 37, no. 10, article e75, 2009.
[105] F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.
[106] International Human Genome Sequencing Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005.
[107] E. R. Mardis, "Anticipating the 1,000 dollar genome," *Genome Biology*, vol. 7, no. 7, article 112, 2006.
[108] V. Costa, A. Casamassimi, and A. Ciccodicola, "Nutritional genomics era: opportunities toward a genome-tailored nutritional regimen," *The Journal of Nutritional Biochemistry*, vol. 21, no. 6, pp. 457–467, 2010.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 915980, 8 pages
doi:10.1155/2010/915980

Research Article

High-Throughput Sequencing of MicroRNAs in Adenovirus Type 3 Infected Human Laryngeal Epithelial Cells

Yuhua Qi,[1,2] Jing Tu,[1] Lunbiao Cui,[2] Xiling Guo,[2] Zhiyang Shi,[2] Shuchun Li,[1] Wenting Shi,[1] Yunfeng Shan,[2] Yiyue Ge,[2] Jun Shan,[2] Hua Wang,[2] and Zuhong Lu[1]

[1] *State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, SiPaiLou2#, Nanjing 210096, China*
[2] *Key Laboratory of Enteric Pathogenic Microbiology, Ministry of Health, Microbiological Laboratory, Jiangsu Center for Disease Prevention and Control (CDC), 172 Jiangsu Rd, Nanjing 210009, China*

Correspondence should be addressed to Zuhong Lu, zhlu@seu.edu.cn

Received 11 February 2010; Revised 27 March 2010; Accepted 31 March 2010

Academic Editor: Momiao Xiong

Adenovirus infection can cause various illnesses depending on the infecting serotype, such as gastroenteritis, conjunctivitis, cystitis, and rash illness, but the infection mechanism is still unknown. MicroRNAs (miRNA) have been reported to play essential roles in cell proliferation, cell differentiation, and pathogenesis of human diseases including viral infections. We analyzed the miRNA expression profiles from adenovirus type 3 (AD3) infected Human laryngeal epithelial (Hep2) cells using a SOLiD deep sequencing. 492 precursor miRNAs were identified in the AD3 infected Hep2 cells, and 540 precursor miRNAs were identified in the control. A total of 44 miRNAs demonstrated high expression and 36 miRNAs showed lower expression in the AD3 infected cells than control. The biogenesis of miRNAs has been analyzed, and some of the SOLiD results were confirmed by Quantitative PCR analysis. The present studies may provide a useful clue for the biological function research into AD3 infection.

1. Introduction

Adenovirus (AD) belongs to a family of nonenveloped icosahedral viruses containing a double-stranded DNA genome, *Adenoviridae*, and genus *Mastadenovirus.* There are 53 human AD serotypes divided into seven subgroups (A to G) based on immunologic, biologic, and biochemical characteristics [1]. Among these, viruses in groups B1, C, and E cause respiratory infections, group B2 viruses infect the kidney and urinary tract; group F viruses cause gastroenteritis; and several group D serotypes are associated with epidemic keratoconjunctivitis [1]. Most people recover from AD infections by themselves, but people with immunodeficiency and small children are at higher risk of fatal adenovirus infection. Studies showed that AD can evade host immune responses, such as inhibition of interferon functions by RNA and E1A, and inhibition of intrinsic cellular apoptosis in infected cells [2]. The interaction between the virus and its host is still not clear. Recently, several studies evaluated human adenoviruses which play an important role in the study of tumorigenesis, including oncogenic transformation of cells, immune evasion, angiogenesis, and metastasis [3, 4]. There are no antiviral drugs to treat adenoviral infections, so treatment is largely directed at the symptoms. The discovery of the AD yielded further evidence that viruses can sometimes interfere with each other during the growth and replication within a host animal. Other researchers found that this interference can be mediated by interferon produced by the host animal. So interferon has become prominent in the treatment of a variety of cancers and infectious diseases, such as hepatitis C [5]. Recently, several studies demonstrated that RNA interference (RNAi) can also be used for treatment with viral infection in vitro, such as hepatitis B virus, coxsackievirus B3, and Coxsackievirus A16 [6–8]. These findings indicate that those specific siRNAs targeted viruses are not as effective as expected.

RNAi can also be triggered by microRNAs (miRNAs), which are short ~22 nucleotide RNA sequences that bind complementary sequences in the 3′ UTR of multiple target

mRNAs, playing an important role in the development, proliferation, differentiation and apoptosis of organisms. Growing evidences have indicated a strong association between microRNAs and various viral infections. For example, miR-125b, mirR-150, miR-382, and miR-223 have been reported to contribute to the maintenance of human immunodeficiency virus-1 (HIV-1) latency in resting $CD4^+$ T cells [9]. Epstein-Barr virus and Kaposi sarcoma herpes viruses (KSHVs) have been reported to encode miRNAs [10], but the functions of most of them are not known. miRNAs have been reported to be used as a tool for gene specific therapeutics against viral infections [11]. The potential use of miRNAs as novel noninvasive biomarkers for diagnosis has been suggested in other studies [12–14].

In this study, we used a strategy of initial screening by Applied Biosystems' next generation sequencing system which is Sequencing by Oligonucleotide Ligation and Detection (SOLiD) and validation by quantitative RT-PCR (qRT-PCR), to analyze the different miRNA expression profiling in AD3 infected Hep2 cells.

2. Materials and Methods

2.1. Cell Culture and Virus Infection. Human laryngeal epithelial (Hep2) cells were cultivated in complete RPMI medium 1640 (Invitrogen) supplemented with 10% fetal calf serum (FCS) at 37°C in an atmosphere of 5% CO_2. When the Hep2 cells were grown in 25-cm^2 flasks reached to 70% confluence, they were infected with AD3 at a multiplicity of infection (m.o.i.) of 0.4 under 50% tissue culture infectious doses ($TCID_{50}$), and maintained after infection at 37°C in RPMI 1640 medium with 2% FCS.

2.2. RNA Isolation. Hep2 cells were infected with AD3 as described above. At 6 hours, 24 hours, 48 hours, and 72 hours post infection (p.i.), RNA samples were prepared from AD3 infected Hep2 cells and controlled Hep2 cells using a mirVana miRNA Isolation Kit (Ambion): 10^2–10^7 cultured cells were washed in 1% PBS for three times, washed cells were mixed with 600 μl Lysis/Binding Solution and 1/10 volume of miRNA Homogenate Additive thoroughly by vortexing and incubated on ice for 10 minutes An equal volume of acid/phenol/chloroform (Ambion) was then added to the each aliquot. The solution was centrifuged for 5 minutes at 10,000 g at room temperature. For miRNA isolation, the extraction was mixed thoroughly with 1/3 volume of 100% ethanol and passed through a column (Ambion). The flow-through was contained the small RNAs. The filtrate was then mixed thoroughly with 2/3 volume 100% ethanol and passed through the second column (Ambion). The column was washed following the protocol, and miRNAs were eluted in 100 μl of elution buffer (95°C). For total RNA isolation, the extraction was mixed thoroughly with 1.25 volume of 100% ethanol and passed through a column (Ambion). The column was washed following the protocol and total RNA were eluted in 100 μl of elution buffer (95°C). A total of 1 μl of the eluate was quantified by Nanodrop.

2.3. SOLiD Sequencing and Sequence Analysis. MiRNA samples (100 ng) isolated from AD3 infected Hep2 cells and control Hep2 cells were processed into sequencing libraries using the Small RNA Expression Kit (Applied Biosystems). Briefly, RNA was ligated overnight with the adapters from the kit, reverse transcribed, RNAse H-treated, and PCR-amplified before size selection on Agarose gels to contain 16–61 nt of inserted sequences. Libraries were amplified onto beads using emulsion PCR, deposited on slides, and sequenced using the SOLiD v 2 sequencing system (Applied Biosystems) [15] at the State Key Lab of Bioelectronics Laboratory, Southeast University of China. Finally, SOLiD data was first analyzed by SOLiD System Small RNA Analysis Pipeline Tool (RNA2MAP). Firstly, the parameters set for alignments were the maximum length (18 nt) and the tolerance of 3 mismatches when generating initial seeds locations. In the extension step, less than 6 mismatches were allowed in full length mapping. The miRBase sequences (Sanger) of human being was download from miRBase (http://www.mirbase.org/). Acceptable sequences were compared to the miRBase database (release 14.0). To increase signal above noise, we conservatively selected only those alignments corresponding to beads sampled a minimum of 10 times in any of the libraries.

2.4. Quantitative RT-PCR (qRT-PCR). To confirm the expression of miRNAs by deep sequencing approach, stem-loop quantitative RT-PCR (qRT-PCR) was performed. Briefly, cDNA was synthesized from total RNA by using AMV reverse transcriptase (TaKaRa). The 20 μL reactions were incubated for 15 minutes at 16°C, 30 minutes at 42°C, and 5 minutes at 85°C, and then held at 4°C. Subsequently, the 20 μL PCR reactions included 1 μl RT-PCR product, 10 μl Premix Ex Taq (TaKaRa), and 1 μl SYBR green (Invitrogen). The reactions were incubated in a 96-well optical plate at 95°C for 10 minutes, followed by 40 cycles of 95°C for 15 seconds, and 60°C for 1 minute. Q-PCR assay was performed in triplicate using an ABI 7500 machine (Applied Biosystems). The expression levels of candidate miRNAs were measured in terms of threshold cycle value (C_T) and normalized to Human small nuclear RNA U6. The ratio of two groups of miRNAs was calculated by using the equation $2^{-\Delta\Delta CT}$ [16], in which ΔCT_1(infected cells) = $C_{T1} - C_{TU6}$, ΔCT_2(control cells) = $C_{T2} - C_{TU6}$ and $\Delta\Delta CT = \Delta CT_1 - \Delta CT_2$.

2.5. Statistical Analysis. We firstly used the Z-test to determine the statistical significance of the differences between the two libraries. This approach is to look at the number of copies of a specic miRNA per cell as a fraction or proportion of the total number of miRNA molecules in that cell. The same proportion (p) of specic tags should be present in the miRNA library of all sequenced tags [17]

$$p = \frac{n_{\text{specific miRNA}}/\text{cell}}{N_{\text{total miRNA}}/\text{cell}} = \frac{n_{\text{specific miRNA}}}{N_{\text{total miRNA}}}. \quad (1)$$

So the Z-value for a specific miRNA could be calculated as following:

$$Z\text{-value} = \frac{p_{\text{test}} - p_{\text{control}}}{\sqrt{p_0(1-p_0)/N_{\text{test}} + p_0(1-p_0)/N_{\text{control}}}}. \quad (2)$$

The proportion p_0, the expected proportion when the null hypothesis is true, is calculated as $p_0 = (n_{\text{test}} + n_{\text{control}})/(N_{\text{test}} + N_{\text{control}})$. The Z statistic is approximately normally distributed and can be compared with critical Z value for the two-sided significance level α [17].

A Z-test was used to deal with the raw data by considering both the copies and fold change of miRNAs. Candidate miRNAs should be full three criteria: (1) mean fold change > 2, (2) having at least 10 copies by SOLiD sequencing, and (3) Z score > 1.96 or < −1.96 for $P < .05$.

3. Results

3.1. Replication of AD3 in Hep2 Cells. Cytopathic effect (CPE) was first observed at 24 hours post infection and progressed to moderate and severe CPE at 72 hours and 96 hours, respectively. miRNA extracted from cells after 72 hours of infection was used for SOLiD deep sequencing.

3.2. SOLiD Sequencing of miRNAs from Hep2 Cells Infected by AD3 and Control Hep2 Cells. To search for the miRNAs expression in AD3 infected Hep2 cells and the control, we prepared these two miRNA samples for sequencing by the SOLiD system. 358,297 effective reads were obtained in AD3 infected cells and 479,414 effective reads in control cells. After filtrating reads contaminated by rRNAetc (rRNA, tRNA, snRNA, and snoRNA), repeat fractions, and degraded fractions, the remaining effective reads were mapped to the human precursor libraries. 37,800 and 78,143 reads were obtained from infected and control cells, respectively. In these reads, the most abundant size class was 22 nt in both of AD3 infected Hep2 cells and the control. The percentage was 29.3% and 34.6% respectively. Following was 23 nt (27.7% in AD3 infected Hep2 cells and 26.3% in the control) (Figure 1). Compared to the miRBase (14.0), 492 precursor, miRNA out of 734 known precursor miRNAs were identified among the AD3 infected Hep2 cells, and 540 precursor miRNAs were identified among the control Hep2 cells. All of the known miRNAs have been characterized in the Sanger miRBase sequence database. Figure 2 showed the distribution of genomic loci in the chromosome. All of the human chromosome were located in the precursor miRNAs. Chromosomes 19 and X contain more precursor miRNAs (12%, and 11%, resp.). 314 miRNAs (51%) were located in introns of protein-coding genes, 231 miRNAs (38%) were located in intergenic regions, while only 43 miRNAs (7%) were located in exons of noncoding RNAs or UTR of protein coding genes. Interestedly, 23 miRNAs (4%) were located in either an exon or an intron depending on alternative splicing of the host transcript (Table 1).

3.3. Expression Profiles of miRNAs in AD3 Infected Cells. To study the differential expression profile of known miRNAs between AD3 infected Hep2 cells and the control, both the Z-test and fold-change analysis were made based on the sequencing results. There are 80 precursor miRNAs to meet the three criteria described above, and could be selected as candidates. Among these candidates, 44 precursor miRNAs were shown upregulated expression in AD3 infected cells (Table 2). Z-test value was from 1.99 to 49.33. 19 out of 44

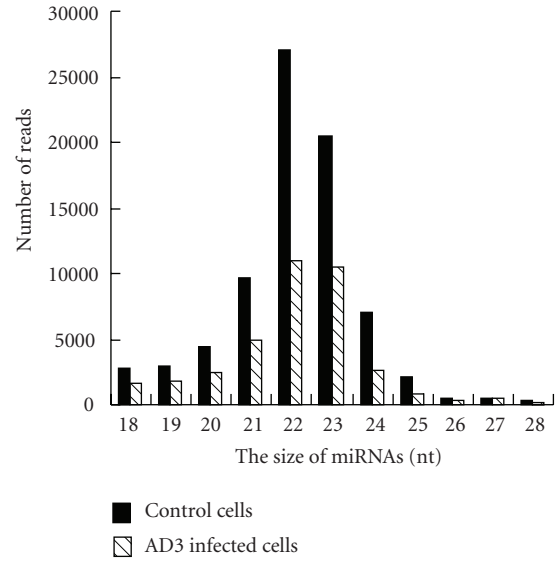


Figure 1: Size distribution of sequenced miRNAs.

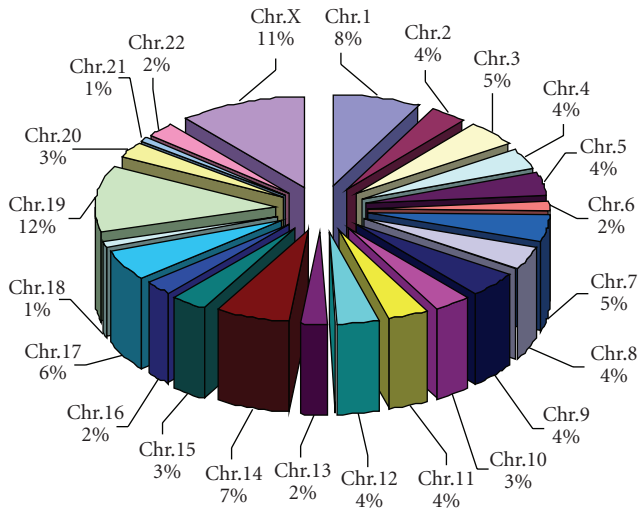


Figure 2: Genomic loci of the sequenced miRNAs (Chr: chromosome).

Table 1: Transcript loci of sequenced miRNAs.

Loci	Number of miRNAs	Percentage (%)
intron	314	51
intergenic	231	38
Exon and Utr	42	7
Exon, intron (mixed)	23	4

Table 2: Upexpressed miRNAs in infected cells compared with noninfected cells.

miRNA Name	Reads in infected cells	Reads in non-infected cells	Z score	Ratio of infected/non-infected (Normalized)
hsa-miR-1975	1937	548	49.33	7.47
hsa-miR-1302-2	788	228	31.13	7.31
hsa-miR-191	868	566	23.16	3.24
hsa-miR-17	1238	1105	21.65	2.37
hsa-miR-92a-2	1166	1084	20.17	2.27
hsa-miR-1979	444	201	20.02	4.67
hsa-miR-92a-1	1184	1146	19.48	2.18
hsa-miR-1302-3	197	58	15.45	7.18
hsa-miR-1302-8	191	54	15.38	7.48
hsa-miR-1302-6	191	57	15.16	7.09
hsa-miR-1302-7	191	57	15.16	7.09
hsa-miR-1302-1	180	59	14.32	6.45
hsa-miR-1302-5	165	50	14.03	6.98
hsa-miR-1302-4	161	50	13.76	6.81
hsa-miR-7-3	111	18	13.13	13.04
hsa-miR-7-1	150	50	13.00	6.34
hsa-miR-7-2	109	21	12.64	10.97
hsa-miR-1974	262	188	11.88	2.95
hsa-miR-1246	186	121	10.70	3.25
hsa-miR-378	220	199	8.95	2.34
hsa-miR-18a	185	153	8.92	2.56
hsa-miR-619	79	41	7.92	4.07
hsa-miR-25	178	172	7.52	2.19
hsa-miR-421	57	38	5.82	3.17
hsa-miR-1273	54	37	5.57	3.09
hsa-miR-18b	62	51	5.18	2.57
hsa-miR-345	50	37	5.07	2.86
hsa-miR-342	83	84	4.87	2.09
hsa-miR-744	28	19	4.03	3.12
hsa-miR-505	22	13	3.90	3.58
hsa-miR-1247	17	10	3.43	3.59
hsa-miR-942	12	6	3.14	4.23
hsa-miR-635	11	5	3.14	4.65
hsa-miR-627	12	7	2.90	3.62
hsa-miR-150	10	5	2.87	4.23
hsa-miR-566	10	5	2.87	4.23
hsa-miR-324	20	18	2.71	2.35
hsa-miR-146b	10	6	2.60	3.52
hsa-miR-454	10	6	2.60	3.52
hsa-miR-1972	14	12	2.37	2.47
hsa-miR-320c-2	14	13	2.20	2.28
hsa-miR-589	16	16	2.17	2.11

upregulated precursor miRNAs showed higher values (over 10). Among this, miR-1975 and miR-1302-2 have the highest Z values with 49.33 and 31.13, respectively. Table 2 showed the ratio of miRNAs expression level of AD3 infected cells versus control cells. Among this, miR-1975 and miR-1302-2 have the higher ratio (7.47-fold and 7.31-fold). 36 precursor miRNAs were shown downregulated expression in AD3 infected cells (Table 3). Among this there are 11 miRNAs that have the higher Z values (over 10). Table 3 showed the ratio of AD3 infected cells/control cells (expression level

Table 3: Downexpressed miRNAs in infected cells compared with noninfected cells.

miRNA Name	Reads in infected cells	Reads in non-infected cells	Z score	Ratio of infected/non-infected (Normalized)
hsa-miR-1184	3	21	−2.06	0.30
hsa-miR-1180	10	43	−2.06	0.49
hsa-miR-219-1	8	38	−2.14	0.45
hsa-miR-519a-1	6	33	−2.24	0.38
hsa-miR-433	6	35	−2.40	0.36
hsa-miR-650	6	35	−2.40	0.36
hsa-miR-148b	12	55	−2.49	0.46
hsa-miR-148a	13	60	−2.62	0.46
hsa-miR-196b	4	31	−2.62	0.27
hsa-miR-1259	6	46	−3.18	0.28
hsa-miR-362	30	128	−3.53	0.50
hsa-miR-181b-1	7	60	−3.80	0.25
hsa-miR-374b	8	65	−3.87	0.26
hsa-miR-655	26	126	−3.96	0.44
hsa-miR-33a	16	95	−3.99	0.36
hsa-miR-101-1	44	191	−4.40	0.49
hsa-miR-101-2	44	192	−4.43	0.48
hsa-miR-452	29	158	−4.87	0.39
hsa-miR-181b-2	8	91	−5.12	0.19
hsa-miR-26b	31	172	−5.14	0.38
hsa-let-7e	33	199	−5.84	0.35
hsa-miR-1274b	15	168	−6.93	0.19
hsa-miR-27b	139	677	−9.25	0.43
hsa-miR-34a	206	904	−9.70	0.48
hsa-miR-582	82	531	−9.95	0.33
hsa-miR-338	113	905	−14.40	0.26
hsa-miR-27a	248	1382	−14.68	0.38
hsa-miR-125b-1	399	2082	−17.23	0.41
hsa-miR-125b-2	443	2220	−17.25	0.42
hsa-miR-30b	219	1572	−18.12	0.29
hsa-miR-30c-2	233	1812	−20.19	0.27
hsa-miR-30c-1	232	1811	−20.22	0.27
hsa-miR-210	207	1738	−20.40	0.25
hsa-miR-19b-1	538	2952	−21.36	0.39
hsa-miR-19b-2	536	2946	−21.36	0.38
hsa-miR-30a	260	2083	−21.95	0.26

for precursor miRNAs). Among this miR-181b-2 and miR-1274b showed lowest change in AD3 infected cells (0.19 fold) compared with control cells.

3.4. Q-PCR Detection for miRNAs Profiles in AD3 Infected Cells. To confirm the differential expression of miRNAs in AD3 infected Hep2 cells, we performed some miRNAs tests with stem-loop Q-PCR assays in total RNA isolated from AD3 infected Hep2 cells and the control, respectively. The expression levels of candidate miRNAs were measured in terms of threshold cycle value (C_T) and normalized to Human small nuclear RNA U6. The ratio of miRNAs in two groups was calculated by using the equation $2^{-\Delta\Delta CT}$. The results showed that there was general consistency between Q-PCR assay and SOLiD sequencing analysis. The seven upregulated miRNAs in SOLiD results showed increased ratios measured in Q-PCR assays (Figure 3). For example, miR-1974, miR-1975, and miR-7 showed increasing values of 2.51, 4.87 and 2.44 respectively measured in Q-PCR assay, and increasing values of 2.95, 7.47 and 6.34 were obtained in SOLiD sequencing (Table 2). The seven downregulated

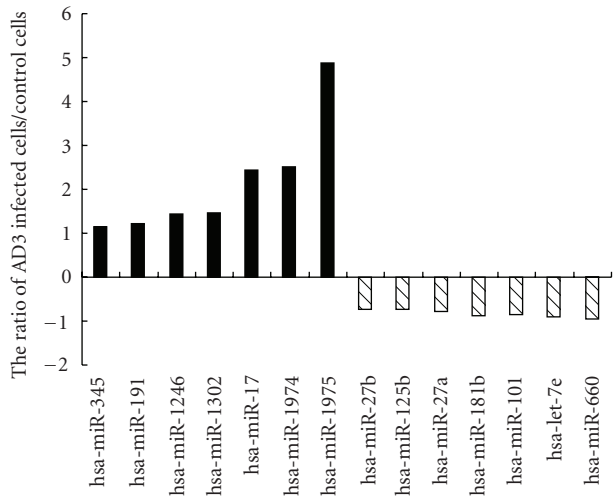


Figure 3: Quantitation of miRNA expression levels in the AD3 infected Hep2 cells and controls. Y values are the ratio of two groups of miRNAs calculated by using the equation $2^{-\Delta\Delta CT}$. The black column is for upregulated miRNAs, and the gray for downregulated miRNAs.

miRNAs in SOLiD results showed decrease ratios measured in Q-PCR assays (Figure 3). For example, the results showed decreased expression levels in miR-27b, miR-125b, and miR-27a in AD3 infected Hep2 cells with fold change of 0.74, 0.75, and 0.77 respectively (Figure 3). While in SOLiD sequencing analysis, these miRNAs have shown decreased expression with fold change 0.43, 0.41 and 0.38, respectively (Table 3). To further understand the miRNAs expression in AD3 infected Hep2 cells at 6 hours, 24 hours, 48 hours, and 72 hours post infection, we performed Q-PCR experiments on miRNAs chosen from Figure 3. The ratio of miRNAs in AD3 infected cells versus controls was calculated by using the equation $2^{-\Delta\Delta CT}$. Surprisingly, all of the miRNAs shown in Table 4 exhibited increased expression in AD3 infected cells at 6 hours post infection. However, at 48 hours p.i. and 72 hours p.i., miR-27b, miR-125b, and miR-181b showed downregulated expression in AD3 infected cells comparing controls (Table 4). At 48 hours p.i. and 72 hours p.i., miR-1974 showed increased expression in AD3 infected cells with fold change 1.36 and 1.26 respectively. However, at 24 hours p.i., miR-1974 showed decreased expression in AD3 infected cells. Table 4 showed that the miR-17 expression was upregulated at 6, 24, and 72 hours p.i. in AD3 infected cells. But at 48 hours p.i., miR-17 showed downregulated expression in AD3 infected cells.

Table 4: The ratio of ADV-3 infected cells/control cells at different time course after the infection.

miRNAs	6 hours p.i.	24 hours p.i.	48 hours p.i.	72 hours p.i.
hsa-miR-27b	2.77	1.25	0.32	0.66
hsa-miR-125b	2.39	0.47	0.56	0.84
hsa-miR-181b	1.73	0.62	0.52	0.62
hsa-miR-1974	1.62	0.4	1.36	1.26
hsa-miR-17	2.81	1.6	0.44	1.33

p.i. refers to post infection. The ratio of miRNAs in two groups was calculated by using the equation $2^{-\Delta\Delta CT}$.

4. Discussion

As second-generation sequencing platforms have matured, all three major high-throughput sequencing systems—Illumina's Genome Analyzer (Solexa), Life Technologies' ABI SOLiD, and Roche's 454 GS FLX have been used to study miRNA expression profiles associated with disease such as cancer, viral and metabolic diseases. Compared with the microarray technology, deep sequencing can generate millions of small RNA sequence reads from a given sample, giving possible to discover the novel miRNAs or study the function of virus encoded miRNAs by using bioinformatics and molecular biology tools. miRNAs have recently been recognized as major regulators of gene expression in various viral infections. Virus-encoded miRNAs seem to evolve rapidly and regulate both the viral life cycle and the interaction between viruses and their hosts. In the present study, we have screened out some miRNA candidates that have obvious differential expression from AD3 infected Hep2 cells by SOLiD deep sequencing system.

This study identified 492 precursor miRNAs in the AD3 infected cells and more precursor miRNAs in control cells by mapping the human miRBase database. Out of these almost 38% were located in intergenic regions, which usually contain their own miRNA gene promoter and regulatory units [18]. However, as much as 51% miRNAs were located in introns, and only 7% were located in exons of noncoding RNAs or UTR of protein coding genes. These usually show a concurrent transcription and regulation expression profile originating from a common promoter with their host genes [19–21]. Some pre-miRNAs existing in drosphila and *c elegants* have been reported to be spliced from the introns in which they reside without having to undergo the microprocessor machinery. These miRNAs are called mirtrons [22, 23]. Since our study identified 51% miRNAs were located in introns, we think more mirtons could be discovered through bioinformatics and molecular biology tools.

The interaction between the AD and its host is still unknown. One of the mechanisms is that the virus must interfere with the host cell antiviral defense mechanisms to maximize escape and spread of the progeny virus during a virus infection [24]. Previous studies determine the changes in the host cell gene expression profiles upon infection with AD using DNA microarray technique [24, 25]. For example, many transcriptions factor E2F-dependent host cell genes showed different expression during AD infection. 45% of the upregulated genes and 25% of the downregulated genes contained potential E2F binding sites [24]. It is showen that E2F-dependent host cell genes are subjected to a selective regulation. However, the mechanism of AD infection after transcription is still unknown. miRNA can regulate gene expression through translational repression or mRNA cleavage. Previous research has reported that the

expression of E2F1 is negatively regulated by two c-myc-regulated miRNAs, miR-17-5p, and miR-20a [26]. In our study, upregulated miR-17 may target with transcription factor E2F1 [26], playing important role in cell differentiation, proliferation, and apoptosis. Out of these, miR-17 has been reported to be overexpression in B-cell lymphoma samples [27]. The high miR-17 cluster miRNAs suppress cell death [27]. This suggests that during a virus infection, cellular apoptosis must delay long enough and the virus needs to establish optimal conditions for replication to ensure efficient production of progeny virus. Our study may provide a useful clue for the biological function research into AD3 infected host cells. Further investigation needs to clarify the roles of identified miRNAs in the pathogenesis of AD3 infected host cells.

5. Conclusion

SOLiD sequencing provides a useful method for identification of the miRNAs profiles in AD3 infected Hep2 cells. 492 precursor miRNAs were identified in the AD3 infected Hep2 cells, and 540 precursor miRNAs were identified in the control. A total of 44 miRNAs demonstrated high expression and 36 miRNAs showed lower expression in the AD3 infected cells than the control. Further studies are required to identify the miRNA target genes and the functions of the miRNAs in the complex molecular network regulation during the virus infection host cells using bioinformatics tools.

Acknowledgments

The paper was supported by the National Natural Science Foundation of China (30901285), the Natural Science Foundation of Jiangsu Province of China (SBK200922783), and the Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University of China. Both Yuhua Qi and Jin Tu contributed equally to this work.

References

[1] M. Echavarría, "Adenoviruses in immunocompromised hosts," *Clinical Microbiology Reviews*, vol. 21, no. 4, pp. 704–715, 2008.

[2] J. A. Mahr and L. R. Gooding, "Immune evasion by adenoviruses," *Immunological Reviews*, vol. 168, pp. 121–130, 1999.

[3] H. Guan, J. F. Williams, and R. P. Ricciardi, "Induction of neuronal and tumor-related genes by adenovirus type 12 E1A," *Journal of Virology*, vol. 83, no. 2, pp. 651–661, 2009.

[4] G. Chinnadurai, "Modulation of oncogenic transformation by the human adenovirus E1A C-terminal region," *Current Topics in Microbiology and Immunology*, vol. 273, pp. 139–161, 2004.

[5] K. A. Sharieff, D. Duncan, and Z. Younossi, "Advances in treatment of chronic hepatitis C: 'pegylated' interferons," *Cleveland Clinic Journal of Medicine*, vol. 69, no. 2, pp. 155–159, 2002.

[6] Z. Wu, Y. Gao, L. Sun, P. Tien, and Q. Jin, "Quick identification of effective small interfering RNAs that inhibit the replication of coxsackievirus A16," *Antiviral Research*, vol. 80, no. 3, pp. 295–301, 2008.

[7] K.-L. Wu, X. Zhang, J. Zhang, et al., "Inhibition of hepatitis B virus gene expression by single and dual small interfering RNA treatment," *Virus Research*, vol. 112, no. 1-2, pp. 100–107, 2005.

[8] J. Yuan, P. K. M. Cheung, H. M. Zhang, D. Chau, and D. Yang, "Inhibition of coxsackievirus B3 replication by small interfering RNAs requires perfect sequence match in the central region of the viral positive strand," *Journal of Virology*, vol. 79, no. 4, pp. 2151–2159, 2005.

[9] J. Huang, F. Wang, E. Argyris, et al., "Cellular microRNAs contribute to HIV-1 latency in resting primary $CD4^+$ T lymphocytes," *Nature Medicine*, vol. 13, no. 10, pp. 1241–1247, 2007.

[10] V. Nair and M. Zavolan, "Virus-encoded microRNAs: novel regulators of gene expression," *Trends in Microbiology*, vol. 14, no. 4, pp. 169–175, 2006.

[11] S.-Y. Ying and S.-L. Lin, "Current perspectives in intronic micro RNAs (miRNAs)," *Journal of Biomedical Science*, vol. 13, no. 1, pp. 5–15, 2006.

[12] J. Skog, T. Wurdinger, S. van Rijn, et al., "Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers," *Nature Cell Biology*, vol. 10, no. 12, pp. 1470–1476, 2008.

[13] K. E. Resnick, H. Alder, J. P. Hagan, D. L. Richardson, C. M. Croce, and D. E. Cohn, "The detection of differentially expressed microRNAs from the serum of ovarian cancer patients using a novel real-time PCR platform," *Gynecologic Oncology*, vol. 112, no. 1, pp. 55–59, 2009.

[14] S. Gilad, E. Meiri, Y. Yogev, et al., "Serum microRNAs are promising novel biomarkers," *PLoS ONE*, vol. 3, no. 9, article e3148, 2008.

[15] L. A. Goff, J. Davila, M. R. Swerdel, et al., "Ago2 immunoprecipitation identifies predicted MicroRNAs in human embryonic stem cells and neural precursors," *PLoS ONE*, vol. 4, no. 9, article e7192, 2009.

[16] T. D. Schmittgen and K. J. Livak, "Analyzing real-time PCR data by the comparative CT method," *Nature Protocols*, vol. 3, no. 6, pp. 1101–1108, 2008.

[17] A. J. Kal, A. J. van Zonneveld, V. Benes, et al., "Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources," *Molecular Biology of the Cell*, vol. 10, no. 6, pp. 1859–1872, 1999.

[18] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel, "An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*," *Science*, vol. 294, no. 5543, pp. 858–862, 2001.

[19] Y.-K. Kim and V. N. Kim, "Processing of intronic microRNAs," *The EMBO Journal*, vol. 26, no. 3, pp. 775–783, 2007.

[20] S. Baskerville and D. P. Bartel, "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes," *RNA*, vol. 11, no. 3, pp. 241–247, 2005.

[21] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley, "Identification of mammalian microRNA host genes and transcription units," *Genome Research*, vol. 14, no. 10, pp. 1902–1910, 2004.

[22] K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai, "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila," *Cell*, vol. 130, no. 1, pp. 89–100, 2007.

[23] J. G. Ruby, C. H. Jan, and D. P. Bartel, "Intronic microRNA precursors that bypass Drosha processing," *Nature*, vol. 448, no. 7149, pp. 83–86, 2007.

[24] H. Zhao, F. Granberg, L. Elfineh, U. Pettersson, and C. Svensson, "Strategic attack on host cell gene expression during adenovirus infection," *Journal of Virology*, vol. 77, no. 20, pp. 11006–11015, 2003.

[25] A. Dorn, H. Zhao, F. Granberg, et al., "Identification of specific cellular genes up-regulated late in adenovirus type 12 infection," *Journal of Virology*, vol. 79, no. 4, pp. 2404–2412, 2005.

[26] K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell, "c-Myc-regulated microRNAs modulate E2F1 expression," *Nature*, vol. 435, no. 7043, pp. 839–843, 2005.

[27] L. He, J. M. Thomson, M. T. Hemann, et al., "A microRNA polycistron as a potential human oncogene," *Nature*, vol. 435, no. 7043, pp. 828–833, 2005.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 425939, 8 pages
doi:10.1155/2010/425939

Research Article

Identification of microRNAs Involved in the Host Response to Enterovirus 71 Infection by a Deep Sequencing Approach

Lunbiao Cui,[1] Xiling Guo,[1] Yuhua Qi,[1,2] Xian Qi,[1] Yiyue Ge,[1] Zhiyang Shi,[1] Tao Wu,[1] Jun Shan,[1] Yunfeng Shan,[1] Zheng Zhu,[1] and Hua Wang[1]

[1] *Institute of Pathogenic Microbiology, Jiangsu Provincial Center for Diseases Prevention and Control, 172 Jiangsu Road, Nanjing 210009, China*
[2] *State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Lunbiao Cui, lbcui@njmu.edu.cn

Received 2 February 2010; Accepted 7 April 2010

Academic Editor: Zhongming Zhao

Role of microRNA (miRNA) has been highlighted in pathogen-host interactions recently. To identify cellular miRNAs involved in the host response to enterovirus 71 (EV71) infection, we performed a comprehensive miRNA profiling in EV71-infected Hep2 cells through deep sequencing. 64 miRNAs were found whose expression levels changed for more than 2-fold in response to EV71 infection. Gene ontology analysis revealed that many of these mRNAs play roles in neurological process, immune response, and cell death pathways, which are known to be associated with the extreme virulence of EV71. To our knowledge, this is the first study on host miRNAs expression alteration response to EV71 infection. Our findings supported the hypothesis that certain miRNAs might be essential in the host-pathogen interactions.

1. Introduction

Hand, foot, and mouth disease (HFMD), a common febrile illness in children, is usually caused by human enteroviruses. Enterovirus 71 (EV71) and coxsackievirus A16 (CA16) are two major causative agents of HFMD. EV71 and CA16 infections manifesting as lesions on the skin and oral mucosa are clinically similar, but EV71 infection is more frequently associated with serious neurological diseases such as aseptic meningitis, encephalitis, and acute flaccid paralysis and fatalities [1–4] while the CA16-associated HFMD has a milder outcome [5]. More than 500,000 HFMD cases caused by EV71 were reported nationwide, including 176 fatal cases in China since March 2008 [6]. Elucidating the cellular events following EV71 infection will facilitate the development of strategies to prevent and treat this virus. However, the molecular mechanisms of the host response to EV71 infection are not completely understood.

microRNAs (miRNAs) have emerged as key regulators in many biological processes, from development to defense, at almost all organismal levels through mRNA degradation or translational repression of their targets [7, 8]. Recently, their role has been highlighted in pathogen-host interactions. Not only the miRNA encoded by viral genomes but host encoded miRNAs have been found participating in host-virus interactions. They can stimulate as well as suppress viral infections. For example, liver-specic miR-122 is an indispensable factor in supporting hepatitis C virus (HCV) replication [9], whereas, miR-125b and miR-223 directly target human immunodeficiency virus-1 (HIV-1) mRNA, thereby attenuating viral gene expression in resting $CD4^+$ T cells [10]. It has been showed that Epstein-Barr virus (EBV) encodes five miRNAs in its large DNA genome. These miRNAs can potentially regulate several genes encoding proteins involved in apoptosis, cell proliferation, signal transduction, transcription regulation, and immune response [11]. In contrast, the global changes, miRNAs expression during EV71 infection have not yet to be extensively elucidated. To determine which cellular microRNAs play a role in the host response to enterovirus infection, in this study, we performed a comprehensive miRNA proling in EV71-infected Hep2 cells through deep sequencing.

2. Materials and Methods

2.1. Cell Culture and Virus Infection. Human epidermoid carcinoma (Hep2) cells were grown in RPMI 1640 medium (Invitrogen) supplemented with 10% fetal calf serum (FCS). When the Hep2 cells were grown to 70% confluence in 25 cm^2 flasks, they were infected with EV71 at a multiplicity of infection (m.o.i.) of 0.03 50% tissue culture infectious doses ($TCID_{50}$) and maintained after infection at 37°C in RPMI 1640 medium with 2% FCS.

2.2. Isolation of RNA and miRNA. Hep2 cells were infected with EV71 as described above. At 6 h, 24 h, 48 h, 72 h, and 96 h post infection, total RNA or miRNA was extracted from cells using the mirVana kit according to the manufacturer's protocol (Ambion). RNA was also extracted from noninfected control cells at the same times. The quality and the concentration of the RNA samples were monitored by gel electrophoresis and absorbance at A260/280 ratio.

2.3. Library Construction, SOLiD Sequencing and Analysis. The miRNA sequencing library construction followed standard procedure of SOLiD small RNA expression kit (Applied Biosystem). All SOLiD run parameters followed standard Applied Biosystems protocols. Different barcodes were introduced to two samples in the polymerase chain reaction of library construction, and all the samples were sequenced in a single sequencing run.

2.4. Data Analysis. SOLiD data were first analyzed by SOLiD System Small RNA Analysis Pipeline Tool (RNA2MAP). The miRBase sequences (Sanger) of human being were downloaded from miRBase (http://www.mirbase.org/).The number of bases to use when generating initial seeds locations was 18 with a tolerance of 3 mismatches. After extension step, at most 6 mismatches were allowed in full length mapping.

Potential conserved target genes of differentially expressed miRNAs were firstly predicted by targetscan (http://www.targetscan.org/) [12–14]. Target genes of some miRNAs, such as mir-1972, mir-1974, mir-1975, mir-1979, and mir-764, could not be predicted in targetscan database, we further predicted those miRNAs targets using DIANA-microT v3.0 [15, 16]. In brief, each differentially expressed miRNA was submitted to targetscan individually and all of its targets predicted in targetscan or microT v3.0 were used for the following gene ontology (GO) analysis (http://www.babelomics.bioinfo.cipf.es/). All targets of induced and repressed miRNAs were submitted to FatiGO program [17]. Functional category enrichment based on the GO terms was evaluated on the targets of these differentially expressed miRNAs.

2.5. Confirmation of Differentially Expressed miRNAs by Real-time Quantitative RT-PCR. To confirm the expression of miRNAs by deep sequencing approach, stem-loop quantitative RT-PCR (qRT-PCR) was performed. In brief, cDNA was synthesized from total RNA by using AMV reverse transcriptase (TaKaRa). The 20 μl reactions were incubated for 15 min at 16°C, 30 min at 42°C and 5 min at 85°C, and then held at 4°C. Subsequently, real-time quantification was performed using an Applied Biosystems 7500 Sequence Detection system (Applied Biosystems). The 20 μl PCR reactions included 1 μl RT-PCR product, 10 μl Premix Ex Taq (TaKaRa), and 1 μl SYBR green (Invitrogen). The reactions were incubated in a 96-well optical plate at 95°C for 10 min, followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. All reactions were run in triplicate. After reaction, the threshold cycle value (CT) data were determined using default threshold settings, and the mean CT was determined from the duplicate PCRs. Human small nuclear RNA U6 was used for normalization. The expression levels of miRNAs were measured in terms of CT and normalized to U6 using $2^{-\Delta\Delta CT}$ [18].

Table 1: Number of reads of miRNAs from EV71 infected and non-infected Hep2 cells.

	Infected cells	Control cells
High quality/both adapter	35,272	78,143
Exact match to known human miRNAs	2,731	8,457
Loose match to known human miRNAs	20,389	47,652

2.6. Statistical Analysis. We firstly used the Z test to determine the statistical significance of the differences between the two libraries [19]. This approach is to look at the number of copies of a specic miRNA per cell as a fraction or proportion of the total number of miRNA molecules in that cell. The same proportion of specic tags should be present in the miRNA library of all sequenced tags. In this test a false discovery rate less than 5% was selected. miRNAs were considered significantly altered only when they fullled three criteria: (1) mean fold change >2 or <0.5, (2) having at least 10 copies by SOLiD sequencing, and (3) Z score > 1.96 or < -1.96.

3. Results

3.1. Replication of EV71 in Hep2 Cells. Cytopathic effect (CPE) was first observed at 24 h post infection and progressed to moderate and severe CPE at 72 and 96 h, respectively. miRNA extracted from cells after 72 h infection was used for SOLiD deep sequencing.

3.2. Deep Sequencing. A total of 411,419 and 479,414 filtered high quality reads were obtained from infected and control cells by deep sequencing, respectively. After filtrating reads contaminated by rRNA, tRNA, snRNA, and snoRNA, 35,272 and 78,143 reads were obtained from infected and non-infected cells, respectively. Out of these reads, 11,188 of these high quality reads were exact matches while other 68,041 reads were loose matches to known human miRNAs (Table 1). Loose matches were defined by sequence reads that aligned with human miRNA consensus sequence with 1–4 mismatches. These may represent sequencing errors (when

Table 2: Upexpressed miRNAs in infected cells compared with non-infected cells.

miRNA	Reads in infected	Reads in non-infected	Z score	Ratio infected/non-infected (Normalized)
hsa-mir-636	11	0	5.05958439	#
hsa-mir-619	305	41	23.59443534	17.31099765
hsa-mir-1302-8	387	54	26.46428291	16.67720702
hsa-mir-1302-2	1560	228	53.05155216	15.92193571
hsa-mir-1302-6,7	387	57	26.28151611	15.79945928
hsa-mir-1302-3	390	58	26.34977656	15.64741958
hsa-mir-1302-1	366	59	25.24116716	14.4356116
hsa-mir-1302-5	305	50	22.98061978	14.19501807
hsa-mir-1302-4	300	50	22.73564953	13.96231285
hsa-mir-1273	162	37	15.80907527	10.18871479
hsa-mir-1290	34	8	7.194848151	9.889971606
hsa-mir-1268	17	4	5.08718048	9.889971606
hsa-mir-1178	11	3	3.959263353	8.532524522
hsa-mir-1246	382	121	22.48247616	7.346561309
hsa-mir-1972	34	12	6.488332484	6.593314404
hsa-mir-1285-1	96	35	10.79491649	6.382771591
hsa-mir-566	13	5	3.901900612	6.05033557
hsa-mir-1226	10	4	3.376339014	5.817630356
hsa-mir-635	12	5	3.645008994	5.584925142
hsa-mir-627	16	7	4.132798028	5.318976326
hsa-mir-518c	13	6	3.647443806	5.041946309
hsa-mir-1185-2	20	10	4.373804587	4.654104285
hsa-mir-1289-1	14	7	3.659280977	4.654104285
hsa-mir-539	11	6	3.116008646	4.266262261
hsa-mir-1247	18	10	3.95066374	4.188693856
hsa-mir-764	12	7	3.147056335	3.989232244
hsa-mir-1979	330	201	16.15562506	3.820533368
hsa-mir-1272	11	7	2.873705427	3.656796224
hsa-mir-324	28	18	4.558616136	3.619858888
hsa-mir-1975	829	548	24.5078521	3.520303332
hsa-mir-744	28	19	4.414239693	3.429339999
hsa-mir-541	19	16	3.126630669	2.763374419
hsa-mir-1254	13	11	2.576283222	2.750152532
hsa-mir-421	42	38	4.379270931	2.572005
hsa-mir-1237	11	10	2.231381474	2.559757357
hsa-mir-320c-1	74	77	5.081196578	2.236387773
hsa-mir-615	16	17	2.309101549	2.190166722
hsa-mir-720	14	15	2.140061026	2.171915333
hsa-mir-140	184	208	7.299247353	2.058546126
hsa-mir-451	188	214	7.31642594	2.044326181
hsa-mir-21	3445	3992	31.35678948	2.008190038

occurring in low copy numbers), mutations, and/or RNA editing events.

The size distribution of sequence reads showed that the majority of miRNAs was 18–25 for both libraries (>90%), with 22 nt small RNA being the most abundant (Figure 1), which is within the typical size range of human miRNAs.

After reads were compared with an miRBase database (release 14.0), 569 miRNAs were detected in EV71-infected

Table 3: Down-expressed miRNAs in infected cells compared with non-infected cells.

miRNA	Reads in infected	Reads in non-infected	Z score	Ratio infected/non-infected (Normalized)
hsa-mir-584	0	10	−2.07306	0
hsa-mir-221	4	70	−4.6261	0.132974
hsa-let-7e	23	199	−6.40515	0.268956
hsa-mir-1180	5	43	−2.96815	0.270587
hsa-mir-1259	6	46	−2.91293	0.303529
hsa-mir-338	128	905	−12.4248	0.32913
hsa-mir-30a	304	2083	−18.6008	0.339618
hsa-mir-19b-1	487	2952	−20.5604	0.383901
hsa-mir-19b-2	491	2946	−20.3931	0.387842
hsa-mir-545	7	41	−2.33845	0.397302
hsa-mir-433	6	35	−2.15407	0.398923
hsa-mir-582	93	531	−8.27217	0.407563
hsa-mir-26a-1	129	735	−9.72425	0.408421
hsa-mir-26a-2	132	746	−9.73602	0.411757
hsa-mir-27b	123	677	−9.08034	0.422788
hsa-mir-452	31	158	−4.09666	0.456574
hsa-mir-1974	38	188	−4.34508	0.470362
hsa-mir-27a	280	1382	−11.8087	0.471472
hsa-mir-26b	35	172	−4.12863	0.473528
hsa-mir-30e	86	419	−6.39521	0.477629
hsa-mir-222	50	236	−4.64215	0.49302
hsa-mir-660	48	225	−4.4992	0.496438

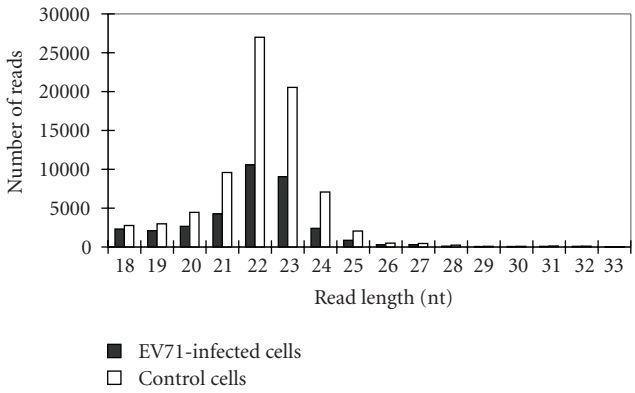


Figure 1: Size distribution of sequenced short RNAs.

cells while 540 miRNAs were detected in non-infected control cells.

3.3. Aberrant miRNAs Expression in EV71-Infected Cells. On the basis of differentially expressed miRNA, we found 64 that miRNAs were differentially expressed between infected and non-infected cells. More miRNAs (42 out of 64 miRNAs) were upregulated than down-regulated during EV71 infection in Hep2 cells (Tables 2 and 3).

3.4. Confirmation of Differentially Expressed miRNAs. Quantitative RT-PCR assays were used to confirm the expression pattern of differentially expressed miRNAs in Hep2 cells. There was general consistency between quantitative RT-PCR assay and deep sequence analysis in four miRNAs (miR-1246, miR-1237, miR-30a, and miR-222) in terms of directions of regulation and significance. Specifically, there was a 1.92-fold upregulation (7.35-fold in deep sequencing analysis) in miR-1246, 1.58-fold upregulation (2.58-fold in deep sequencing analysis) in mir-1237, ($P < .05$), 2.45-fold down-regulation (2.94-fold in deep sequencing analysis) in miR-30a, and 1.42-fold down-regulation (2.02-fold in deep sequencing analysis) in miR-222.

3.5. Gene Ontology Analysis. Targets were predicted for all identified differentially expressed miRNA families. In total 5765 unique target genes were predicted for 64 of the differentially expressed miRNAs (see Table S1 in Supplementray Material available online at doi:115/2010/425939). On the basis of the biological functions described by FatiGO program (http://www.babelomics.bioinfo.cipf.es/), these target genes can be grouped into 72 categories (S2). The top 30 Gene Ontology terms are shown in Figure 2. The majority of targets fall into the category of metabolic process, regulation of biological process, and cell communication indicating intense biological change in Hep2 cells after EV71

Table 4: Reverse regulatory association of up-regulated miRNAs and their predicted mRNA targets.

Up-regulated miRNA	Target genes	Functions	Reference
hsa-mir-1289-1	Interferon regulatory factor 2	IFN response	[20]
hsa-mir-140,hsa-mir-320c	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	Oncogene	[20]
hsa-mir-421	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog		[21]
hsa-mir-1226	Interleukin 10 receptor, beta	Immune response	[20]
hsa-mir-1290	Nuclear receptor coactivator 2	Transcriptional regulators	[20]
hsa-mir-320c,hsa-mir-1290	Transcription factor AP-2 beta		[20]
hsa-mir-320c	pre-B-cell leukemia homeobox 3		[21]
hsa-mir-636,hsa-miR-1979,hsa-mir-1302,hsa-mir-518c,hsa-mir-1226,hsa-mir-1290,hsa-mir-421,hsa-mir-21,hsa-mir-140	ribosomal protein S6 kinase	Kinases and phosphatases	[20]
hsa-mir-1289-1,hsa-mir-1272,hsa-mir-421	Splicing factor, arginine/serine-rich 1	RNA synthesis and modification	[20]
hsa-mir-1272,hsa-mir-140,hsa-mir-636	Neuro-oncological ventral antigen 1	Neuron specific	[20]
hsa-mir-539	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5		[21]
hsa-mir-1254,hsa-mir-627,hsa-mir-320c,hsa-mir-421,hsa-mir-21,hsa-mir-140	cyclin-dependent kinase 6	cell cycle	[21]
hsa-mir-619	protocadherin 9	cell adhesion	[21]
hsa-mir-627	RER1 retention in endoplasmic reticulum 1 homolog (S. cerevisiae)	trafficking and targeting proteins	[21]
hsa-mir-140	sorting nexin 2		[21]
hsa-mir-1289-1	xenotropic and polytropic retrovirus receptor	Cellular receptors	[21]
hsa-mir-21	interleukin 12A (natural killer cell stimulatory factor 1, cytotoxic lymphocyte maturation factor 1, p35)	cell signalling, extracellular communication	[21]
hsa-mir-1178,hsa-mir-140	transforming growth factor, alpha		[21]
hsa-mir-1178	dual specificity phosphatase 6	intracellular trasducers/effector/modulators	[21]

infection. Several other groups contain genes regulating death (255 target genes including 239 apoptosis-related target genes), neurological process (211 target genes), and immune response (120 target genes).

Hundreds of altered transcripts in response to EV71 infection were found in two transcriptomic studies [20, 21]. To determine whether miRNAs might be modulators of mRNAs that were differentially expressed, we investigated whether those differentially expressed mRNAs are enriched for predicted targets and interrogated their inversely correlated targets for functional associations. To our surprise, targets for differentially expressed miRNAs or miRNA families in the present study account for 12.1% (19 out of 157) of the differentially expressed mRNAs in one study [20] whereas they account for 22.4% (13 out of 58) of the transcripts that were significantly altered in another study (Tables 4 and 5) [21].

4. Discussion

Many technologies have been developed for miRNA profiling, including real-time quantitative RT-PCR [22], northern blotting [23, 24], and microarray analyses based on either direct hybridization or hybridization coupled with enzymatic extension [25, 26]. These methods have been used successfully in a variety of studies. But they still have some technical limitations. For example, first two methods were not high-throughput while microarray method needs large amounts of starting materials. Recent progress in high-throughput sequencing technologies allows deep sequencing of large libraries of short RNAs [27–30]. The longest reads are obtained by the 454 technology, which currently gives reads of about 400 base pairs (bp). However, this technology yields much less reads than other techniques (about 400 000 per sample) [27]. The Solexa (Illumina) and SOLiD

Table 5: Reverse regulatory association of down-regulated miRNAs and their predicted mRNA targets.

Down-regulated miRNA	Target genes	Functions	Reference
hsa-mir-27	Cytochrome P450,superfamily I polypeptide 1	Mitochondrial function	[20]
hsa-mir-26a,hsa-mir-27	BCL2-antagonist/killer 1	Apoptosis	[20]
hsa-mir-26a	Programmed cell death 10		[20]
hsa-mir-582	Amyloid beta (A4) precursor protein		[20]
hsa-mir-19	BCL2-associated athanogene 5		[20]
hsa-mir-1180,hsa-mir-19,hsa-mir-582,hsa-mir-26a,hsa-mir-30a,hsa-let-7e,hsa-mir-545	calcium channel, voltage-dependent, L type, alpha 1C subunit	Membrane transporters	[20]
hsa-mir-433,hsa-let-7e	Cathepsin C	Protein degradation	[20]
hsa-mir-26a,hsa-mir-545	Muscle RAS oncogene homolog	Oncogene	[20]
hsa-mir-19	RAB5B, member RAS oncogene family		[20]
hsa-mir-19	neurotrophic tyrosine kinase, receptor, type 2	Kinases and phosphatases	[20]
hsa-mir-19	Fas-activated serine/threonine kinase		[20]
hsa-mir-582,hsa-let-7e	adaptor-related protein complex 1, sigma 1 subunit	trafficking and targeting proteins	[21]
hsa-mir-27	actin, alpha 2, smooth muscle, aorta	cytoskeleton/motility	[21]

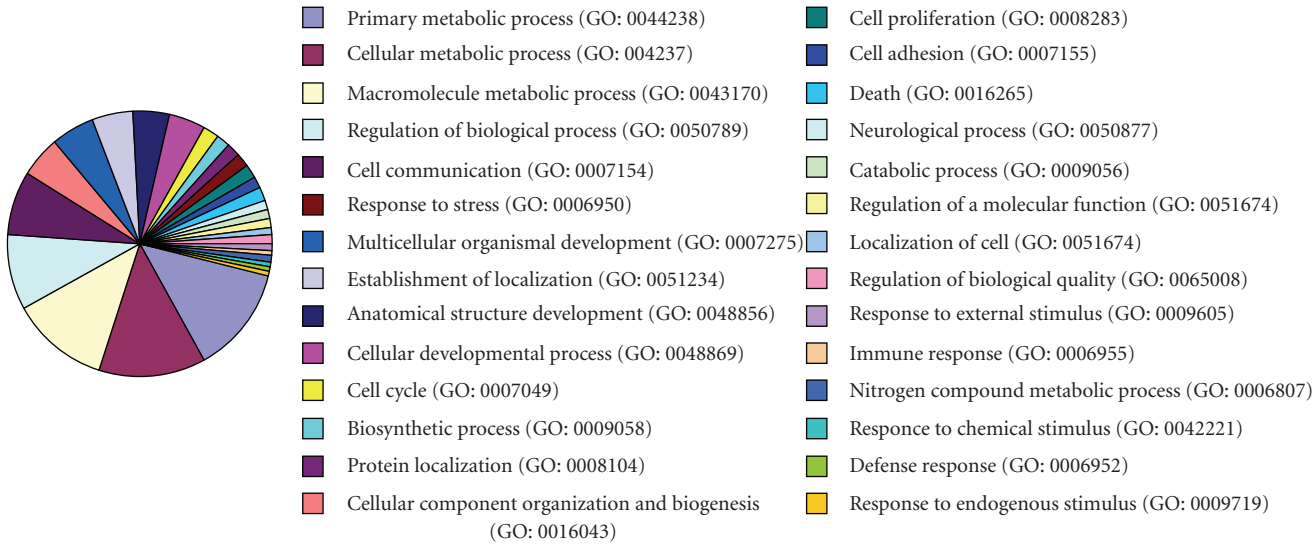


Figure 2: Predicted targets gene ontology terms in the biological process category.

platform (Applied Biosystems) generates shorter reads (up to 35 bp) but yields 1–3 million reads per sample [28, 29]. The other high-throughput technique, such as massively parallel sequencing (MPSS), gives more reads than Solexa but the reads are even shorter, only 17 bp [30]. In the present study, we used SOLiD platform to explore differential miRNA proling of Hep2 cells response to EV71 infection. We obtained about 0.4 million reads in both samples, well below the capacity of SOLiD. One possibility is that we used cell line as infection model but not tissue. Numbers of small RNAs were not expressed in Hep2 cells. Even now, we found that there was general consistency between quantitative RT-PCR assay and deep sequence analysis. So, SOLiD deep sequencing successfully revealed miRNA proling in EV71-infected and control Hep2 cells.

Based on a comprehensive examination of miRNA expression from EV71-infected and non-infected control Hep2 cells, we identified 64 miRNAs that were differentially expressed, with most of them (65.6%) upregulated in EV71-infected cells. The altered patterns of cellular miRNAs we

observed for EV71-infected cells are similar in some respects to changes seen for cells infected with Hepatitis C virus (HCV) [31] but contrast with those with Epstein-Barr virus (EBV) [32] and human cytomegalovirus (HCMV) in which more miRNAs were dow-regulated in response to virus infection [33]. Most of the miRNAs affected by EV71 are different from those affected by HCV, EBV, and HCMV. We observed miR-636 and miR-584 expressed in only infected or non-infected cells. Though their functions during virus infection have not been explored, their predicted target genes have been identified to be involved in virus entry, replication and propagation. For example, reticulon 3 (RTN3), one predicted target gene of miR-636, can bind the 2C protein of enterovirus 71 and is required for viral replication, [34]. Poliovirus receptor-related 1 (herpesvirus entry mediator C) (PVRL1, also known as nectin-1), another target gene of miR-636, can serve as receptor for herpes simplex virus and pseudorabies virus entry [35]. Abhydrolase domain containing 2 (ABHD2), one predicted target gene of miR-584, is essential for Hepatitis B virus propagation [36]. Thus, it will be important to investigate the mechanisms of regulation of miRNA levels during virus infection, which could be at the stages of transcription, maturation, and/or degradation.

The host response to viral infection represents complex orchestration of divergent pathways deigned to eliminate the virus and protect the host. Viruses impact on many aspects of the host cell's biology and function [20]. As seen from gene ontology analysis, the top 3 gene ontology terms related to metabolic process which indicated EV71 infection have enormous effect on Hep2 cells metabolism. Intriguingly, neurological process, apoptosis, and immune response related GO terms in biological process were also enriched from the predicted targets. Considering the fact that EV71 can cause lethal encephalitis or myocarditis, both apoptosis and immune response contribute to EV71 pathogenesis. Myocarditis, for example, represents an intricate interplay between virus and patient responses, in which both direct viral injury and immunopathologic damage caused by innocent-bystander phenomena affect the disease course [37]. Further study of the functions of those underlying miRNAs related to neurological process, apoptosis, and immune response will help to elucidate the molecular mechanisms of EV71 pathogenesis.

Our study has some limitations that will need to be addressed in future studies. We did not assess the roles in infection of miRNAs whose expression were altered after infection. As microRNAs predominately function as repressors of target gene expression, we indeed found numbers of targets in an another transcriptomic research whose expression was inversely correlated with the expression of dysregulated miRNAs in the present study. Although all assays were executed when cells demonstrated similar CPE, the virus may interact with miRNA regulatory pathways differently in different cell types in which human neural SF268 cells and rhabdomyosarcoma cells were used in those studies, respectively. Even then, these findings suggest that expression modification of host miRNAs during EV71 infection could be related to a number of cellular physiological processes that eventually control the cell fate.

In summary, in this study, we identified the miRNAs involved in the host response to EV71 infection using deep sequencing technology. Our findings provide a deeper understanding of the mechanisms underlying EV71 infection. Once the role of these miRNAs in the regulation of host-EV71 interaction has been determined, it will improve the protection and treatment strategies in enterovirus infection.

Acknowledgments

The first two authors contributed equally to this study. The study was supported by the grant from Important National Science & Technology Specific Projects during the eleventh five-year plan period (2008ZX10004-002), the National Natural Science Foundation of China (30901285), the Natural Science Foundation of Jiangsu Province (SBK200922783), and the Open Research Fund of State Key Laboratory of Bioelectronics Southeast University.

References

[1] P. McMinn, I. Stratov, L. Nagarajan, and S. Davis, "Neurological manifestations of enterovirus 71 infection in children during an outbreak of hand, foot, and mouth disease in Western Australia," *Clinical Infectious Diseases*, vol. 32, no. 2, pp. 236–242, 2001.

[2] M. Portolani, A. M. Bartoletti, P. Pietrosemoli, et al., "Non-specific febrile illness by Coxsackievirus A16 in a 6-day-old newbornMalattia febbrile aspecifica da Coxackievirus A16 in un neonato di 6 giorni," *Minerva Pediatrica*, vol. 56, no. 3, pp. 341–347, 2004.

[3] L.-Y. Chang, L.-M. Huang, S. S.-F. Gau, et al., "Neurodevelopment and cognition in children after enterovirus 71 infection," *The New England Journal of Medicine*, vol. 356, no. 12, pp. 1226–1234, 2007.

[4] C.-C. Huang, C.-C. Liu, Y.-C. Chang, C.-Y. Chen, S.-T. Wang, and T.-F. Yeh, "Neurologic complications in children with enterovirus 71 infection," *The New England Journal of Medicine*, vol. 341, no. 13, pp. 936–942, 1999.

[5] L.-Y. Chang, T.-Y. Lin, Y.-C. Huang, et al., "Comparison of enterovirus 71 and coxsackievirus A16 clinical illnesses during the Taiwan enterovirus epidemic, 1998," *Pediatric Infectious Disease Journal*, vol. 18, no. 12, pp. 1092–1096, 1999.

[6] N.-Z. Ding, X.-M. Wang, S.-W. Sun, Q. Song, S.-N. Li, and C.-Q. He, "Appearance of mosaic enterovirus 71 in the 2008 outbreak of China," *Virus Research*, vol. 145, no. 1, pp. 157–161, 2009.

[7] J. C. Carrington and V. Ambros, "Role of microRNAs in plant and animal development," *Science*, vol. 301, no. 5631, pp. 336–338, 2003.

[8] A. Kurzyńska-Kokorniak, P. Jackowiak, M. Figlerowicz, and M. Figlerowicz, "Human- and virus-encoded microRNAs as potential targets of antiviral therapy," *Mini-Reviews in Medicinal Chemistry*, vol. 9, no. 8, pp. 927–937, 2009.

[9] C. L. Jopling, M. Yi, A. M. Lancaster, S. M. Lemon, and P. Sarnow, "Molecular biology: modulation of hepatitis C virus RNA abundance by a liver-specific microRNA," *Science*, vol. 309, no. 5740, pp. 1577–1581, 2005.

[10] J. Huang, F. Wang, E. Argyris, et al., "Cellular microRNAs contribute to HIV-1 latency in resting primary $CD4^+$ T lymphocytes," *Nature Medicine*, vol. 13, no. 10, pp. 1241–1247, 2007.

[11] S. Pfeffer, M. Zavolan, F. A. Grässer, et al., "Identification of virus-encoded microRNAs," *Science*, vol. 304, no. 5671, pp. 734–736, 2004.

[12] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[13] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel, "MicroRNA targeting specificity in mammals: determinants beyond seed pairing," *Molecular Cell*, vol. 27, no. 1, pp. 91–105, 2007.

[14] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," *Genome Research*, vol. 19, no. 1, pp. 92–105, 2009.

[15] M. Maragkakis, P. Alexiou, G. L. Papadopoulos, et al., "Accurate microRNA target prediction correlates with protein repression levels," *BMC Bioinformatics*, vol. 10, article 295, 2009.

[16] M. Maragkakis, M. Reczko, V. A. Simossis, et al., "DIANA-microT web server: elucidating microRNA functions through target prediction," *Nucleic Acids Research*, vol. 37, web server issue, pp. W273–W276, 2009.

[17] M. J. Lodes, M. Caraballo, D. Suciu, S. Munro, A. Kumar, and B. Anderson, "Detection of cancer with serum miRNAs on an oligonucleotide microarray," *PLoS ONE*, vol. 4, no. 7, article e6229, 2009.

[18] T. D. Schmittgen and K. J. Livak, "Analyzing real-time PCR data by the comparative C method," *Nature Protocols*, vol. 3, no. 6, pp. 1101–1108, 2008.

[19] A. J. Kal, A. J. van Zonneveld, V. Benes, et al., "Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources," *Molecular Biology of the Cell*, vol. 10, no. 6, pp. 1859–1872, 1999.

[20] S.-R. Shih, V. Stollar, J.-Y. Lin, S.-C. Chang, G.-W. Chen, and M.-L. Li, "Identification of genes involved in the host response to enterovirus 71 infection," *Journal of NeuroVirology*, vol. 10, no. 5, pp. 293–304, 2004.

[21] W. F. Leong and V. T. K. Chow, "Transcriptomic and proteomic analyses of rhabdomyosarcoma cells reveal differential cellular gene expression in response to enterovirus 71 infection," *Cellular Microbiology*, vol. 8, no. 4, pp. 565–580, 2006.

[22] C. Chen, D. A. Ridzon, A. J. Broomer, et al., "Real-time quantification of microRNAs by stem-loop RT-PCR," *Nucleic Acids Research*, vol. 33, no. 20, article e179, 2005.

[23] L. F. Sempere, S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky, and V. Ambros, "Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation," *Genome Biology*, vol. 5, no. 3, article R13, 2004.

[24] J. M. Cummins, Y. He, R. J. Leary, et al., "The colorectal microRNAome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 10, pp. 3687–3692, 2006.

[25] C.-G. Liu, G. A. Calin, B. Meloon, et al., "An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9740–9744, 2004.

[26] P. T. Nelson, D. A. Baldwin, L. M. Scearce, J. C. Oberholtzer, J. W. Tobias, and Z. Mourelatos, "Microarray-based, high-throughput gene expression profiling of microRNAs," *Nat Methods*, vol. 1, no. 2, pp. 155–161, 2004.

[27] S. Moxon, R. Jing, G. Szittya, et al., "Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening," *Genome Research*, vol. 18, no. 10, pp. 1602–1609, 2008.

[28] E. A. Glazov, P. A. Cottee, W. C. Barris, R. J. Moore, B. P. Dalrymple, and M. L. Tizard, "A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach," *Genome Research*, vol. 18, no. 6, pp. 957–964, 2008.

[29] L. A. Goff, J. Davila, M. R. Swerdel, et al., "Ago2 immunoprecipitation identifies predicted MicroRNAs in human embryonic stem cells and neural precursors," *PLoS ONE*, vol. 4, no. 9, article e7192, 2009.

[30] C. Lu, K. Kulkarni, F. F. Souret, et al., "MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant," *Genome Research*, vol. 16, no. 10, pp. 1276–1288, 2006.

[31] X. Liu, T. Wang, T. Wakita, and W. Yang, "Systematic identification of microRNA and messenger RNA profiles in hepatitis C virus-infected human hepatoma cells," *Virology*, vol. 398, no. 1, pp. 57–67, 2010.

[32] S. E. Godshalk, S. Bhaduri-McIntosh, and F. J. Slack, "Epstein-Barr virus-mediated dysregulation of human microRNA expression," *Cell Cycle*, vol. 7, no. 22, pp. 3595–3600, 2008.

[33] F.-Z. Wang, F. Weber, C. Croce, C.-G. Liu, X. Liao, and P. E. Pellett, "Human cytomegalovirus infection alters the expression of cellular MicroRNA species that affect its replication," *Journal of Virology*, vol. 82, no. 18, pp. 9065–9074, 2008.

[34] W.-F. Tang, S.-Y. Yang, B.-W. Wu, et al., "Reticulon 3 binds the 2C protein of enterovirus 71 and is required for viral replication," *Journal of Biological Chemistry*, vol. 282, no. 8, pp. 5888–5898, 2007.

[35] M. Yoon and P. G. Spear, "Disruption of adherens junctions liberates nectin-1 to serve as receptor for herpes simplex virus and pseudorabies virus entry," *Journal of Virology*, vol. 76, no. 14, pp. 7203–7208, 2002.

[36] X. R. Ding, J. Yang, D. C. Sun, S. K. Lou, and S. Q. Wang, "Whole genome expression profiling of hepatitis B virus-transfected cell line reveals the potential targets of anti-HBV drugs," *Pharmacogenomics Journal*, vol. 8, no. 1, pp. 61–70, 2008.

[37] H. A. Rotbart and F. G. Hayden, "Picornavirus infections: a primer for the practitioner," *Archives of Family Medicine*, vol. 9, no. 9, pp. 913–920, 2000.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 867517, 11 pages
doi:10.1155/2010/867517

Research Article

Global Egr1-miRNAs Binding Analysis in PMA-Induced K562 Cells Using ChIP-Seq

Wei Wang,[1] Dequang Zhou,[2] Xiaolong Shi,[1] Chao Tang,[3] Xueying Xie,[4] Jing Tu,[1] Qinyu Ge,[4] and Zuhong Lu[1,4]

[1] *State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China*
[2] *School of Life Sciences, Lanzhou University, Lanzhou 730000, China*
[3] *School of Basic Medical Sciences, Southeast University, Nanjing 210096, China*
[4] *Research Center for Learning Science, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Zuhong Lu, zhlu@seu.edu.cn

Received 11 February 2010; Revised 17 May 2010; Accepted 28 June 2010

Academic Editor: Fuli Yu

Although much is known about microRNAs' regulation in gene expression and their contributions in cell fate, to date, globally lineage-(cell-) specific identification of the binding events between a transcription factor and its targeting microRNA genes is still waiting for elucidation. In this paper, we performed a ChIP-Seq experiment to find the targeting microRNA genes of a transcription factor, Egr1, in human erythroleukemia cell line K562. We found Egr1 binding sites near the promoters of 124 distinct microRNA genes, accounting for about 42% of the miRNAs which have high-confidence predicted promoters (294). We also found EGR1 bind to another 63 pre-miRNAs. We chose 12 of the 187 microRNAs with Egr1 binding sites to perform ChIP-PCR assays and the positive binding signal from ChIP-PCR confirmed the ChIP-Seq results. Our experiments provide the first global binding profile between Egr1 and its targeting microRNA genes in PMA-treated K562 cells, which may facilitate the understanding of pathways controlling microRNA biology in this specific cell line.

1. Introduction

MicroRNAs (miRNAs) are a family of ~22-nucleotide small noncoding RNAs in eukaryotes and mainly involve in regulation at posttranscriptional level by translational repression or degradation their target mRNAs [1, 2]. More than 700 human miRNAs have been identified up till now [3] and they are estimated to control about one-third of human known genes [4]. miRNAs have been reported to regulate hematopoietic lineage differentiation, angiogenesis, cell adhesion and so on [1, 5]. K562 is a cell line deriving from chronic myeloid leukemia, which is a common progenitor of megakaryocytic and erythroid lineages of the hematopoietic stem cell differentiation. It can be induced to differentiate into erythrocytes or megakaryocytes (MK) by hemin and Phorbol-12-myristate 13-acetate (PMA), respectively, [6–9]. Recently, miRNAs have been found to play a key role in K562 differentiation. miR-27a, miR-34a, miR-223 were up-regulated when K562 was induced to MK status while miR-27a, miR-223, miR-103, miR-130a, miR-210, and miR-18b were downregulated when K562 was induced to erythroid differentiation [10–13]. The changes of miRNAs expression level gave clues for their functions in hematopoietic lineage differentiation but a more detailed regulation pathway is anticipated to be understood: by which targets miRNAs realize their functions in hematopoietic differentiation and miRNAs are subjected to which factors controlling their transcription? Garzon et al. confirmed that miR-130 targets the transcription factor MAFB and participates in MK differentiation by up-regulating its expression level in $CD34^+$ hematopoietic progenitors [14]. Navarro et al. found that independently of p53, miR-34a directly regulates expression

of MYB facilitating megakaryocytic differentiation of K562 cells and of CDK4 and CDK6, to inhibit the G1/S transition [11]. Lu et al. found that miR-150 regulates megakaryocyte-erythrocyte progenitors (MEPs) differentiation and is preferentially expressed in megakaryocytic lineage. Besides, Lu et al. identified that transcription factor MYB is also a critical target of miR-150 in this regulation [15]. A feedback loop between Runx1 and miR-27a was found by Ben-Ami et al. that miR-27a plays a regulatory role in megakaryocytic differentiation by attenuating Runx1 expression, and during megakaryopoiesis, Runx1 exerts positive regulation of miR-27a expression [10]. While the researches, which deepen our understanding of the underlying mechanism of cellular differentiation involving miRNAs, are based on individual specific miRNA, currently the interactions between a TF and its target miRNAs in this cellular process on a large scale have been sparsely investigated.

Egr1 is an immediate-early response protein which is rapidly and transiently induced by various stimuli, such as different growth factors, cytokines, mechanical injury, shear stress [16]. As a transcription factor, many of its known transcription target genes are protein coding genes. Among the known transcriptional targets of Egr1, a part of genes were implicated in the pathogenesis of vascular disease, including PDGF-A, PDGF-B, FGF-2, SOD1, p53, CD44 (see commentary [17]). Besides, Egr1 may regulate cell interaction with the extracellular matrix by coordinated induction of TGF-β1, FN, and PAI-1 in human glioblastoma cells [18]. Moreover, the Egr1 gene is functionally implicated in cell proliferation and in the regulation of apoptosis and is considered as a potential target for prostate cancer therapy [19]. Few studies in the past, on the other hand, has been focused on the Egr1's regulation role on noncoding target genes except a finding of Egr1's role in hsa-miR-106a transcription. Through hsa-miR-106a, Egr1 indirectly regulates the IL-10 expression [20]. Previous studies showed that PMA-induced activation of Egr1 expression is involved in megakaryocytic differentiation of K562 cells but the regulation pathway has not been investigated. As an early response factor, Egr1 may regulate both downstream protein-coding and noncoding genes. ChIP-Seq, a technique based on next-generation sequencing, is able to capture the genome-wide snapshot of protein-DNA interactions in vivo [21]. The interactions between a TF and its target genes can be identified by ChIP-Seq and miRNA genes are also the targets of TFs. Here we identified the transcription factor Egr1 binding sites near miRNA promoters and premiRNAs in PMA-induced K562 cells using ChIP-Seq. Combining with the Navarro's miRNA expression profile [11] before and after PMA or hemin treatment in K562 cells, we analyzed the potential role of Egr1 in differentiation fate of K562 cells through targeting miRNA genes.

2. Materials and Methods

2.1. Cell Culture. K562 cell was cultured in RMPI 1640 medium containing 10% fetal bovine serum (GIBCO) at 37°C with 5% CO_2 incubator. K562 cells were treated by PMA (10 ng/ml) for 2 hours.

2.2. Chromatin Immunoprecipitation (ChIP). ChIP experiment was carried out as described previously [22]. Briefly, PMA treated K562 cells were crosslinked with 1% formaldehyde and incubated at room temperature (RT) for 12 minutes, prior to the addition of glycine to 0.125 M followed by incubation at RT for 5 minutes. Chromatin were sonicated and was capture with the beads coated with EGR1 antibody (588) (from Santa Cruz Biotechnology) or no IgG. The beads were then precipitated and washed. Bead complexes were eluted with elution buffer. Samples were incubated overnight at 65°C for reversing crosslink. Samples were then incubated with Proteinase K and RNase A for 2 hours at 50°C. DNA was purified by two rounds of phenol-chloroform extraction and ethanol precipitation and resuspended in 50 μl dH_2O.

2.3. Sequencing. Briefly, EGR1 ChIP DNA was sheared into 90–150 bp and then was repaired by DNA end-repair kit (Epicentre). SOLiD adaptors were then added and DNA was subjected to 19 cycles of PCR (2 × enhancer (invitrogen) was added to increase the amplification for GC-rich sequences). PCR product of 180–220 bp size was incised from gel and then was subjected to emulsion PCR. Template positive beads were collected and extended in the presence of terminal transferase and Bead Linker (SOLiD reagents, Applied Biosystems). Extended beads were then deposited and covalently attached onto slides at >20,000 beads per panel (SOLiD reagent, Applied Biosystems). Template bead slides were then loaded onto a SOLiD Analyzer, and cycled ligation sequencing using the SOLiD Sequencing System was performed.

2.4. Mapping of the SOLiD Sequencing Data. Sequence reads from ChIP fragments were aligned to human reference genome using standard SOLiD mapping pipeline tool (Corona Lite, [23]), allowing up to 3 mismatches out of 35 bp. All mapping was performed in the SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA.

2.5. Peak Detection and Motif Analysis. The uniquely mapped reads were retained for subsequent analysis. An integrated software CisGenome was used to detect peaks which are enriched from background reads [24]. The algorithm applied the sliding window method to count the reads and the window size was set 500 bp, a tradeoff size which had best coverage for the known Egr1 targets and was consistent with length of DNA fragments from ChIP experiments. The position with highest density in a window should contain at least nine reads otherwise the regions will be discarded. We used the strand filtering function to make sure each strand meet the minimum three reads requirement.

All the motif analysis, including known motif mapping and new motif discovery, in this study were completed by CisGenome. The known motif sequence (degenerate pattern: 5′-G[T]C[A]GG[T]GGGCG[A,T]-3′) for EGR1 was curated from literature [17, 25].

2.6. miRNA Promoters and PremiRNAs Annotation of Egr1 Binding Sites. We directly applied 294 predicted high-confidence miRNA promoters [26] with minor revisions to annotate our Egr1 ChIP-Seq dataset. Firstly, the coordinates of the predicted promoter regions were converted to the current version of the human genome (hg18) from hg17 using the liftover utility provided by UCSC Genome Browser [27]. Then, the miRNA gene names were revised based on miRBase v13 [3] because they are supposed to be criterion to link the miRNA expression profile in the subsequent analysis.

For sequence-specific factors such as Egr1, previous studies used a more relaxed region of 8 kb surrounding the promoters [26, 28]. As the length of those predicted promoters varies, here we define that the enriched regions which are located within 5 kb regions of predicted promoters and premiRNAs are candidate Egr1 binding sites.

2.7. Gene Ontology Analysis of the Targets of miRNA. The experimentally supported miRNA targets were downloaded from TarBase v5.0 [29]. We chose the Batch-Genes tools provided by GOEAST to do Gene Ontology(GO) enrichment analysis [30]. Given a set of genes, Batch-genes can find statistically significantly enriched GO terms among them by using a species-specific random background. Hypergeometric test was chose and the *P*-value was set to .01 (Yekutieli FDR adjusted) to call an enriched GO term.

2.8. ChIP-PCR Analysis. We randomly selected seven miRNAs together with another five up-regulated miRNAs after PMA treatment (hsa-miR-135b, hsa-miR-141, hsa-miR-152, hsa-miR-199b, hsa-let-7g) [11] among Egr1 target miRNAs to perform ChIP-PCR analysis and their corresponding ChIP DNA fragments resulted from two different cell conditions: before PMA treatment and after PMA treatment of K562 cells. The designed PCR primers are listed in supplementary Table 1at supplementary material available online at doi:10.1155/2010/867517.

3. Results

We performed a ChIP-Seq experiment on next-generation sequencer SOLiD system for the transcription factor Egr1 in PMA-induced K562 cells. Relying on the hypothesis that chromatin signatures (H3K4me3) can be used to locate transcription start sites of most genes in genome, Marson et al. [26] predicted 294 distinct high-confidence miRNA promoters utilizing the H3K4me3 dataset from ChIP-chip and ChIP-Seq experiment. We adopted these promoters to annotate our Egr1 binding sites and anticipated to find Egr1-miRNA gene interactions. We also surveyed the premiRNA with Egr1 binding signal. ChIP-PCR was conducted to assess the reliability of part of those binding events. Besides, a recent published miRNAs expression profile in the similar cell condition facilitated to analyze the expression changes of those miRNAs which are bound by Egr1 in PMA-induced MK differentiation of K562 cells [11].

3.1. Bioinformatics Evaluation of the Quality of ChIP-Seq Experiment. By setting the cutoff 9 during the peak detection, we got 14,636 enriched regions. Although the negative binominal model adopted by CisGenome is able to estimate a stringent false discovery date (FDR) and the FDR was less than 10% when the peak height is equal to and higher than 17 at the window size of 500 bp, we chose to combine the FDR estimation with further experimentally validations to determine the threshold and the cutoff 9 is supposed to compromise the coverage and specificity of the results.

Here we evaluate our ChIP-Seq experiments by motif analysis. The enriched regions were scanned by the canonical EGR1 motif (degenerate pattern: 5′-G[T]C[A]GG[T]GGGCG[A,T]-3′ [17, 25]) within different peak height rank. Among the peaks with height ≥ 17, 60.8% possessed the canonical EGR1 motifs. We observed that the percentages of the canonical EGR1 binding motifs decreased with peak rank. 43.2% of the full list of peaks can be found canonical EGR1 binding motifs. Besides, the sequences extracted from 2042 peaks with high quality (peak height more than 20) were used to do de novo motif finding analysis. We found two novel motifs, one (5′-GCGG[T]GGGC[T]GG-3′) resembling the known consensus of EGR1 binding site [31], and one is the other (5′-GGGGC[T]GGGG-3′) resembling the consensus of EGR1 binding sites which has been identified by Kubosaki using ChIP-chip [32]. These information are presumed to give an overview of the technical quality of this ChIP-Seq experiments in a bioinformatics way.

We next examined the EGR1 binding sites distribution relative to protein-gene structure. We found the majority of the EGR1 binding loci (40.7% of the 14,636 sites) are within the range from 10 kb upstream of the transcription start site (TSS) to 10 kb downstream of the transcription end site. About 12.2% of the peaks are located within 1 kb upstream of the TSS of the known genes, suggesting the transcriptional relevance of EGR1. A more detailed discussion of protein-coding gene targets of EGR1 using this ChIP-Seq experiment can be found in [33]. Due to the unbiased identification of sequencing method, it is unsurprising to find that many EGR1 binding sites are located in exons, introns, and intergenic regions.

3.2. Identification of the Occupancy of miRNA Promoters and PremiRNAs by Egr1 in K562 Cells. That 37.3% and 50.2% of the 14,636 enriched regions are located in intron regions and intergenic regions, respectively, elicited our curiosity to check if EGR1 is able to bind to noncoding genes like miRNAs, as 5,091 (91%, 5,604 miRNAs' context is known in miRBase v13 [3]) miRNAs' context is intron. With the emergence of the predicted miRNA promoters, we firstly defined those regions, which are located within 5 kb regions upstream and downstream of the predicted promoters, are candidate binding sites by Egr1. The promoters of 124 different miRNAs were found with ChIP-Seq signals and they accounts for 42% of those miRNAs with known predicted promoters. hsa-mir-10a is one of the miRNA genes and its Egr1 binding signal is visualized in Figure 1. hsa-mir-10a has two Egr1 ChIP-Seq enriched binding signals (chr17:44008884-44011068, reads

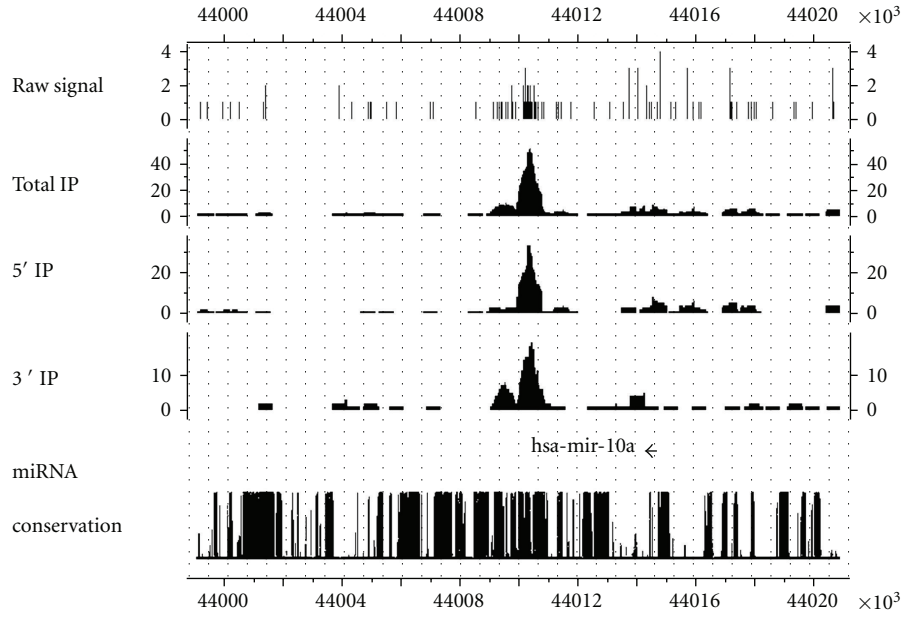


Figure 1: Visualization of Egr1 ChIP-Seq peaks enriched around the promoter of miRNA hsa-mir-10a. Two peaks (chr17:44008884-44011068, reads count:51; chr17:44013523-44015182, reads count:9) around the promoter of hsa-mir-10a can be detected. Raw signal represents the aligned tags. Total IP, 5′ IP and 3′ IP are the signal profile built by sliding windows. 5′ IP and 3′ IP means immunoprecipitationed DNA signal which are clustered by the forward and reverse aligned tags, respectively. The figure is drawed by CisGenome browser [24].

count:51; chr17:44013523-44015182, reads count:9) near its predicted promoters (chr17:44014310-44014709).

Although most miRNA genes (469) are still lacking high-confidence predicted promoters, to get a comprehensive view of the target miRNA genes, we also surveyed the EGR1 binding signal around all precursor miRNAs (miRBase v13 [3]). 125 precursor miRNAs were found EGR1 binding signal by the same distance standard for the miRNAs with promoters (also listed in supplementary Table 1). Among the 125 target miRNAs identified by precursor coordinates and 124 target miRNAs identified by promoter coordinates, 62 miRNA genes were found present in both lists (see Figure 2), thus, 187 miRNAs were found EGR1 binding signal in all. The reason for those 62 miRNA genes which can be identified by both coordinates is that the relative distance to mature miRNAs is one of elements to score the predicted miRNA promoters. Among the 12 miRNA genes whose ChIP-Seq EGR1 binding were confirmed by ChIP-PCR (see Figure 3), 9 (mir-10a, mir-130b, mir-95, mir-23b, mir-17, mir-152, mir-141, mir-135b, and mir-199b) were among the 62 miRNAs whose EGR1 binding signal can be identified by both the promoters and precursors of miRNAs and the left 3 ones were only found near the predicted promoters of miRNAs.

A region containing canonical consensus of a transcription factor has great potential to be the targets of this factor. To determine if the extra 63 miRNAs should be accepted as the targets of Egr1, the peaks associated with miRNAs were scanned by Egr1 canonical consensus sequence (degenerate pattern: 5′-G[T]C[A]GG[T]GGGCG[A,T]-3′ [17, 25]).

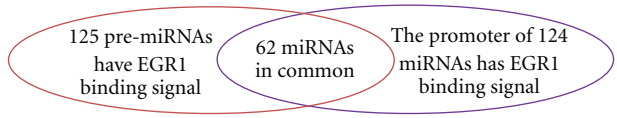


Figure 2: The overlap of miRNA list with EGR1 binding. 125 precursor miRNAs were found EGR1 binding signal in addition that the promoters of 124 different miRNAs were found with ChIP-Seq signals. 62 miRNA genes were found present in both lists.

61.2% of the peaks associated with miRNA genes (220 peaks in all) were found at least one Egr1 canonical motif (see Table 1), which is significantly more than the ratio of the full set peaks (Fisher's exact test, P-value $= 9.8 \times 10^{-8}$). However, the difference of the motif frequency between the peaks associated with the promoters of miRNAs and the peaks linking to premiRNAs is not significant (Fisher's exact test, P-value $= .1$), so we decided to view the extra 63 miRNAs as the putative targets of Egr1. All miRNAs with Egr1 binding sites enriched from ChIP-Seq signals are listed in supplementary Table 1. It is noteworthy that a miRNA gene may have two or more predicted promoters and a promoter may have more than one Egr1 binding sites.

Among the 14,636 peaks, most are located in intergenic and intron region as mentioned above, so we asked the genomic features of those peaks relating to miRNA. 220 unique peaks are associated with the 187 miRNA genes. Among them, 40.1% of peaks are located in intron region

Table 1: Genomic features (especially intergenic region and intron region) and motif properties of the ChIP-Seq enriched regions associated with miRNA genes.

	No. of peaks located in intergenic region (ratio)	No. of peaks located in intron region (ratio)	total peak no.	No. of peaks with canonical motif (ratio)
All peaks enriched	7340 (50.2%)	5460 (37.3%)	14,636	6318 (43.2%)
Peaks associated with miRNA genes (187)	95 (43.2%)	90 (40.1%)	220	135 (61.2%)
Peaks associated with the promoter of miRNA genes (124)	64 (46.0%)	67 (48.2%)	139	94 (67.6%)
Peaks associated with pre miRNAs (125)	60 (46.9%)	46 (35.9%)	128	74 (57.8%)

(see Table 1), which is a little more than the ratio of the same genomic peaks (37.3%) among the whole set of peaks (14,636). Higher ratio of peaks (48.2%) located in intron region was observed for peaks near the predicted promoter of miRNA genes. However, lower ratio of peaks (35.9%) located in intron region was observed for peaks near the premiRNAs. This is because most premiRNA genes are already located in intron region and the average length of human intron regions is 6,206 bp. In this case, peaks around the -5 kb $\sim$ $+5$ kb of the precursor of miRNA genes can easily be outside of the intron region. In general, no significant increase of the amount of peaks as putative Egr1-miRNA binding are located in introns (Fisher's exact test, P-value $= .29$).

3.3. Gene Ontology Enrichment Analysis of Egr1-Bound miRNAs' Targets. In order to elucidate the functions of the 187 Egr1-bound miRNAs, we examined gene ontologies (GO) using the web-based analysis toolkit GOEAST [30] for the targets of those miRNAs as miRNAs regulate cellular activities via pairing to mRNAs. Since our experimental results are nonmicroarray-based high throughput, we chose the Batch-Genes tools provided by GOEAST to do GO enrichment analysis. Given a set of genes, Batch-genes can find statistically significantly enriched GO terms among them by using a species-specific random background.

Although most mammalian mRNAs are conserved targets of miRNAs, a significant proportion of false positives are thought to be included in the predicted miRNA targets. Here, we chose to analyze those experimentally supported miRNA targets. Among the 187 Egr1-bound miRNAs, 38 miRNAs were found experimentally validated targets (247) according to TarBase v5.0 [29]. In the GOEAST analysis, the 247 genes were compared to 18,596 genes as background and the P-value of a GO term less than.01 (Yekutieli FDR adjusted) was thought to be enriched using hypergeometric test. We present the first 25 enriched biological process terms in Table 2 and the full enriched molecular function terms in Table 3. Interestingly, the statistically significantly overrepresented GO biological process terms are highly enriched for hemopoiesis, immune system development, cell differentiation and so on, which are the cases supposed to be occurring in this cell condition. Moreover, among enriched GO molecular function terms, the targets of miRNA genes with EGR1 binding signal include binding, protein binding, similar to the previous GO enrichment finding of gene targets of EGR1 using ChIP-chip [32]. Besides, molecular function GO terms are significant for transcription factor transcription, transcription factor activity, transcription activator binding, and transcription activator activity. This results is consistent with the discovery that transcription factors also prevail among human miRNA targets [34]. It also supports the notion that EGR1 acts as an early response protein in cell events and may trigger a series of transcription factor activity. Three graphical outputs showing the hierarchical relationships of enriched GO terms ($P \leq .01$, Yekutieli FDR adjusted) in each GO category (can be seen in supplementary Figures 1–3).

3.4. An Evaluation of the Predicted miRNA Promoters Using ChIP-Seq. All the predicted miRNA promoters were also scanned by Egr1 canonical consensus sequence (degenerate pattern: 5′-G[T]C[A]GG[T]GGGCG[A,T]-3′ [17, 25]) and 77 predicted promoters (refer to 74 miRNA genes) were found canonical Egr1 motifs. Among the 77 promoters with motif, 33 (42.9%, refer to 31 miRNA genes) have real Egr1 binding signals based on our ChIP-Seq experiment and they are listed in Table 4. According to the data, we found that hsa-mir-142 has 45 consensus sites dispersed among its promoters and correspondingly, 10 Egr1 binding regions identified by ChIP-Seq were enriched with reads count above nine. miR-142 is one of five miRNAs (miR-142, miR-144, miR-150, miR-155 and miR-223) which are highly specific for hematopoietic cells according expression comparison [35].

3.5. Egr1 Bound-miRNAs' Expression in PMA-Induced K562 Cells. The set of Egr1 bound-miRNA genes identified by ChIP-Seq do not necessarily mean that they are regulated by the transcription factor Egr1. However, if the expression level of Egr1 bound-miRNAs fluctuates concomitantly with the expression changes of Egr1 during MK differentiation induced by PMA, it may suggest the function relevance of their interactions. Recently, Navarro et al. compared 286 miRNAs expression at different time point (0 days, 2 days, and 4 days) after PMA treatment of K562 cells by miRNA microarray [11]. The expression level of most miRNA genes increased with PMA induction according to the microarray analysis. Since the K562 cells used in our

Table 2: Enrichment of Gene Ontology biological process terms in targets of miRNAs with EGR1 ChIP-Seq hits.

GO term IDs	Enriched GO terms	*P*-value
GO:0009987	Cellular process	$5.06E-08$
GO:0048513	Organ development	$9.35E-06$
GO:0006357	Regulation of transcription from RNA polymerase II promoter	$1.09E-05$
GO:0060056	Mammary gland involution	$1.29E-05$
GO:0044237	Cellular metabolic process	$3.30E-05$
GO:0060443	Mammary gland morphogenesis	$3.65E-05$
GO:0022612	Gland morphogenesis	$6.49E-05$
GO:0048731	System development	$1.26E-04$
GO:0030097	Hemopoiesis	$1.32E-04$
GO:0002520	Immune system development	$1.43E-04$
GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	$1.45E-04$
GO:0030154	Cell differentiation	$1.53E-04$
GO:0030155	Regulation of cell adhesion	$1.80E-04$
GO:0048856	Anatomical structure development	$1.85E-04$
GO:0042127	Regulation of cell proliferation	$2.03E-04$
GO:0009888	Tissue development	$3.87E-04$
GO:0048869	Cellular developmental process	$3.87E-04$
GO:0048534	Hemopoietic or lymphoid organ development	$3.87E-04$
GO:0010468	Regulation of gene expression	$4.29E-04$
GO:0008285	Negative regulation of cell proliferation	$5.44E-04$
GO:0006928	Cellular component movement	$6.05E-04$
GO:0007162	Negative regulation of cell adhesion	$6.10E-04$
GO:0042060	Wound healing	$6.14E-04$
GO:0014068	Positive regulation of phosphoinositide 3-kinase cascade	$8.61E-04$
GO:0032502	Developmental process	$9.71E-04$

Table 3: Enrichment of Gene Ontology molecular function terms in targets of miRNAs with EGR1 ChIP-Seq hits.

GO term IDs	Enriched GO terms	*P*-value
GO:0005515	Protein binding	$2.93E-05$
GO:0030528	Transcription regulator activity	$1.89E-04$
GO:0051879	Hsp90 protein binding	$2.64E-04$
GO:0003700	Transcription factor activity	$8.36E-04$
GO:0004722	Protein serine/threonine phosphatase activity	$9.73E-04$
GO:0004716	Receptor signaling protein tyrosine kinase activity	$1.53E-03$
GO:0030618	Transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity	$1.62E-03$
GO:0005488	Binding	$2.74E-03$
GO:0003702	RNA polymerase II transcription factor activity	$3.38E-03$
GO:0033613	Transcription activator binding	$6.08E-03$
GO:0016563	Transcription activator activity	$7.22E-03$
GO:0005072	Transforming growth factor beta receptor, cytoplasmic mediator activity	$8.44E-03$
GO:0034713	Type I transforming growth factor beta receptor binding	$8.44E-03$

ChIP experiment were also treated by PMA and the snapshot of interactions between Egr1 and its target DNAs were captured after 1.5 h PMA treatment, we directly applied this expression profile of miRNAs to investigate if the binding by Egr1 contributes to the regulation of miRNA expression. According to the microarray analysis of protein-coding genes (data not shown), we found that the expression of Egr1 in PMA-treated K562 cells increased to 7.5-fold of that in the untreated cells. We found that among those miRNAs with Egr1 binding sites, the expression level of 46 mature miRNAs had increased at least two fold after two days PMA treatment and 16 of them showed significant expression changes with at least five fold increase based on Navarro's miRNA microarray analysis [11]. These 16 mature miRNAs

Table 4: The list of miRNA promoters with both Egr1 motif and ChIP-Seq peak. mir_TSS_start and mir_TSS_end is the start and stop of the predicted promoters. score is the cumulative score for each predicted promoter. motif_num is the count of the Egr1 motifs in the promoter regions of miRNAs.

mir_name	mir_TSS_start	mir_TSS_end	score	strand	chr	reads_count	peak_summit	motif_num
hsa-mir-200b	1088265	1090140	10	+	chr1	11	1091461	2
hsa-mir-200b	1088265	1090140	10	+	chr1	10	1094077	2
hsa-mir-200b	1088265	1090140	10	+	chr1	9	1086834	2
hsa-mir-30e/30c-1 (cluster-hsa-mir-30a)	40929851	40930051	19	+	chr1	25	40930041	1
hsa-mir-92b	153429179	153434335	10	+	chr1	13	153430306	3
hsa-mir-92b	153429179	153434335	10	+	chr1	11	153428517	3
hsa-mir-146b	104184797	104186757	10	+	chr10	23	104185405	2
hsa-mir-132/212 (cluster-hsa-mir-132)	1899412	1901670	10	−	chr17	23	1900085	4
hsa-mir-132/212 (cluster-hsa-mir-132)	1899412	1901670	10	−	chr17	23	1903969	4
hsa-mir-10a	44014310	44014709	15	−	chr17	51	44010351	1
hsa-mir-10a	44014310	44014709	15	−	chr17	9	44014535	1
hsa-mir-142	53379405	53785839	10	−	chr17	14	53757151	45
hsa-mir-142	53379405	53785839	10	−	chr17	14	53516188	45
hsa-mir-142	53379405	53785839	10	−	chr17	13	53765628	45
hsa-mir-142	53379405	53785839	10	−	chr17	11	53760775	45
hsa-mir-142	53379405	53785839	10	−	chr17	11	53383253	45
hsa-mir-142	53379405	53785839	10	−	chr17	10	53740717	45
hsa-mir-142	53379405	53785839	10	−	chr17	10	53671689	45
hsa-mir-142	53379405	53785839	10	−	chr17	9	53739084	45
hsa-mir-142	53379405	53785839	10	−	chr17	9	53711312	45
hsa-mir-142	53379405	53785839	10	−	chr17	9	53747880	45
hsa-mir-24-2/27a/23a (cluster-has-mir-23a)	13818427	13819944	1	−	chr19	10	13814707	2
hsa-mir-642	50863241	50863441	20	+	chr19	18	50863778	1
hsa-mir-99b/let-7e/125a (cluster-has-mir-99b)	56883717	56886813	10	+	chr19	17	56886253	5
hsa-mir-99b/let-7e/125a (cluster-has-mir-99b)	56883717	56886813	10	+	chr19	10	56884005	5
hsa-mir-149	241040800	241041000	9	+	chr2	10	241042627	1
hsa-mir-149	241043770	241044582	10	+	chr2	10	241042627	1
hsa-mir-124-3	61276703	61277978	9	+	chr20	33	61280947	2
hsa-mir-124-3	61276703	61277978	9	+	chr20	10	61274186	2
hsa-mir-124-3	61278418	61280978	10	+	chr20	33	61280947	3
hsa-mir-124-3	61278418	61280978	10	+	chr20	10	61274186	3
hsa-mir-185	18388530	18388730	20	+	chr22	38	18388406	1
hsa-mir-185	18388530	18388730	20	+	chr22	36	18384025	1
hsa-mir-185	18388530	18388730	20	+	chr22	10	18392254	1
hsa-mir-301b/130b (cluster-hsa-mir-130b)	20335495	20337979	10	+	chr22	20	20338454	1
hsa-mir-301b/130b (cluster-hsa-mir-130b)	20335495	20337979	10	+	chr22	19	20341434	1
hsa-mir-301b/130b (cluster-hsa-mir-130b)	20335495	20337979	10	+	chr22	9	20332712	1
hsa-mir-301b/130b (cluster-hsa-mir-130b)	20335495	20337979	10	+	chr22	9	20330784	1
hsa-mir-425/191 (cluster-hsa-mir-191)	49029742	49035883	10	−	chr3	11	49033588	3

Table 4: Continued.

mir_name	mir_TSS_start	mir_TSS_end	score	strand	chr	reads_count	peak_summit	motif_num
hsa-mir-148a	25955148	25959801	10	–	chr7	12	25957595	2
hsa-mir-96/183 (cluster-hsa-mir-96)	129204548	129210298	10	–	chr7	23	129207821	7
hsa-mir-96/183 (cluster-hsa-mir-96)	129204548	129210298	10	–	chr7	9	129200022	7
hsa-mir-335	129919123	129919323	10	+	chr7	13	129919298	2
hsa-mir-596	1752393	1753192	10	+	chr8	13	1752487	1
hsa-mir-32	110921946	110922146	19	–	chr9	9	110922058	1
hsa-mir-455	115957951	115958151	20	+	chr9	14	115956609	1
hsa-mir-219-2	130194040	130195803	10	-	chr9	9	130196326	2

were hsa-miR-181a, hsa-miR-181b, hsa-miR-212, hsa-miR-132, hsa-miR-135b, hsa-miR-141, hsa-miR-152, hsa-miR-199b, hsa-let-7g, hsa-let-7b, hsa-miR-149, hsa-miR-153, hsa-miR-346, hsa-miR-375, hsa-miR-200a, and hsa-miR-200b. miR-181b and miR-181a are members of miR-181 cluster, which increased 133-fold and 19-fold, respectively. miR-132 and miR-212, in the miRNA miR-132/miR-212 cluster, share a promoter, whose region has two Egr1 binding sites. They were also highly up-regulated with the former increasing 10-fold and the latter increasing 6-fold. These data may suggest that due to PMA stimulating, coupling to Egr1 up-regulation during megakaryocytic differentiation, the expression of miRNA genes, which are bound and activated by the transcription factor Egr1, also increased.

3.6. Analyzing Egr1-miRNA Binding Events by ChIP-PCR. ChIP-PCR assay is a conventional and reliable measure of whether a locus is a true binding target of a protein of interest. To experimentally validate the enriched regions from ChIP-Seq, we performed ChIP-PCR assays for 12 miRNA genes. As showed in Figure 3, significant binding signals of the transcription factor Egr1 can be observed for 12 miRNA genes within the PMA-induced K562 cells, while for the input DNA, there is no signal can be observed (see Figures 3(a) and 3(b)). The results here can partially verified the ChIP-Seq experiment, however, it cannot be concluded that the sensitivity of this ChIP-Seq is as high as 100%. It can be easily envisioned if more ChIP-Seq found targets were subjected to ChIP-PCR, false positive rate may arise.

To further check that if the expression or the increase of expression levels of miRNAs can partially attribute to the binding of the transcription factor Egr1, we also performed ChIP-PCR analysis for 12 miRNA genes before PMA treatment (see Figure 3) and made a qualitative comparison for their Egr1-miRNA binding levels between these two cell conditions. Among the 12 miRNAs, 5 miRNA genes (the right column in Figure 3) showed at least 5-fold expression changes before and after PMA-treatment according to Navarro's microarray analysis for microRNA genes [11]. The ChIP-PCR results for two cell conditions showed that some Egr1-miRNA binding levels significantly changed after PMA treatment as well as those showed similar binding signals for different cell conditions. Among the 5 miRNAs whose expression levels significantly increased after PMA-treatment, miR-135b, miR-141, miR-152 showed no Egr1 binding signals before PMA treatment (see the right ChIP-PCR results in Figure 3), however, the PMA induced expression increase of transcription factor Egr1 subsequently turned on the Egr1 binding switch to these three miRNA genes. As analyzed above, miR-135b, miR-141, miR-152 were among the 16 mature miRNAs with at least five fold expression increase after PMA treatment, so before PMA induced, they still express at a relatively low level due to other factors although without the factor Egr1's transcription. We suspect that like general housekeeping genes which have low levels of genic transcription whereby gene products are constitutively produced at low levels [36], the low level expression of miRNA genes may also functionalize as housekeeping miRNAs. Once bound by the key transcription factors and transcriptionally stimulated, the explosion of their expression is likely to functionally correlate with the cell fate. For miRNA genes let-7g and miR-199b, we can observe an Egr1 ChIP binding signal enhancement before and after PMA treatment, which may mainly be caused by up-regulation of Egr1. As let-7g and miR-199b are also among the 16 mature miRNAs with significant fold changes, we can conclude that Egr1 may actively involve in these miRNA genes transcription after PMA induction.

For the left seven miRNA genes, based on Navrro' microarray analysis, miR-10a, miR-95, miR-326, miR-23b, miR-130b also showed two fold expression increase, despite not as high as that of the above five miRNA genes, and the expression level of miR-25 and miR-17 did not show any changes. The ChIP-PCR results of six of them within the nontreatment K562 cells showed positive binding signals. The similar binding signal before and after PMA treatment of K562 cell between the transcription factor Egr1 and these, or may be not limited to, six miRNA genes, suggests that the great expression increases of Egr1 may not be able to increase all the Egr1-miRNA bindings and therefore stimulate their transcriptions. It also suggests that only a part of miRNA genes, which form a regulatory pattern, contributes to this PMA-induced differentiation process. As the gene expression is such a complicated cellular process, different regulatory patterns may exist at distinct cellular processes.

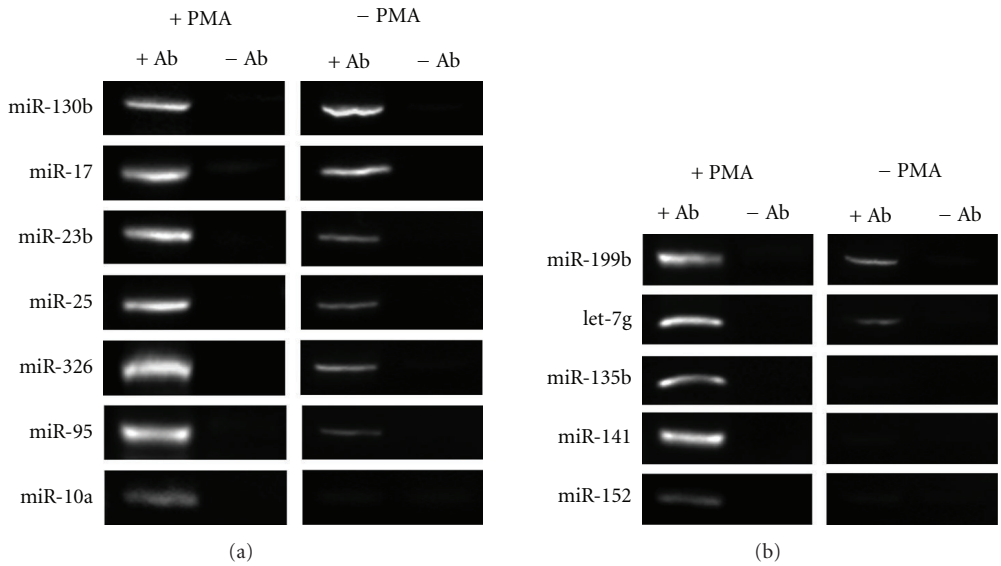


Figure 3: ChIP-PCR results of 12 miRNAs before (−) and after (+) PMA treatment. All the 12 miRNAs showed positive Egr1-binding signals after PMA treatment, which verified the ChIP-Seq results (+Ab).

4. Discussion

Cellular activities can be considered as a succession of hierarchically acting regulatory states [37]. Identification of miRNA's targets is necessarily of great significance, but it is also significant to detect what factors determine miRNAs expression. In this report, we have taken the approach of ChIP-Seq to identify all the miRNAs with known predicted promoters which are bound by the transcription factor Egr1. We found that the promoters of 124 distinct miRNAs have at least one Egr1 binding site. Besides, EGR1 binding signal was also found near extra 63 premiRNAs. Combing with the miRNA expression profile of PMA-induced K562 cells, we were able to find that the binding of the transcription factor Egr1 may be the reason for expression changes of parts of miRNAs. We believe that these will be instrumental in unraveling the mechanism of transcriptional regulation of miRNA genes.

Lineage specification is critical in developmental biology and evidences showed that miRNAs mediate control of cell fate in megakaryocy-erythrocyte progenitors [15]. As accumulative evidence showed that miRNAs expression differs between PMA- and hemin-induced K562 differentiation, the possible driven power is under revealing. EGR1 is a zinc finger family TF who is known as "immediate-early response factor" and has been implied in megakaryocytic differentiation of K562 cells. Evidences were also found that Egr1 activation is correlated with downregulation of erythroid-specific genes and up-regulation of the megakaryocyte-spcefic gene CD41a [8]. The identification of candidate EGR1-targeted miRNAs through this high-throughput approach significantly broadens the repertoire of biological effects attributable to the TF. As many of our annotated miRNAs are known to play crucial rules in hematopoietic differentiation (miR-27a, miR-10a, etc.), the role of Egr1 in hematopoietic differentiation is also implicated here.

During K562 cell erythroid/MK differentiation, there are miRNAs expression signatures specific to each cell lineage. For example, a subset of 12 highly up-regulated miRNAs (miR-34a, miR-181b, miR-299-5p, miR-134, miR-375, miR-181a, miR-139, miR-222*, miR-409-3p, miR-221, miR-212, miR-132) from Navarro's miRNA microarray in TPA-treated K562 cells was found to be specific to MK differentiations, while none of them was up-regulated during hemin-induced erythroid differentiation by Northern blot assay [11]. miR-223 was also up-regulated during PMA-induced megakaryocytic differentiation but downregulated during hemin-induced erythroid differentiation [12]. Induction of megakaryocytic differentiation in K562 cells by PMA markedly increased miR-27a expression; however, down-regulation of miR-27a expression occurs upon induction of K562 cells toward erythroid differentiation. Runx1 was the transcription factor which plays a role in the erythroid/megakaryocytic lineage determination through mediating regulation of miR-27a [10]. PMA-induced activation of Egr1 expression has been proved to involve in megakaryocytic differentiation of K562 cells but the regulation pathway has not been investigated. We therefore investigate several Egr1-bound miRNAs which behave differently in erythroid/megakaryocytic cell lineage to see if Egr1 plays a role in erythroid/megakaryocytic lineage determination of K562 cells. The promoters of miRNA cluster hsa-mir-181a (hsa-mir-181a/181b) and cluster hsa-mir-132 (hsa-mir-132/212) cluster were found Egr1 binding sites according to our ChIP-Seq experiment. These two miRNA clusters are possible candidates through which Egr1 acts as a lineage determinator. Our ChIP-Seq data also showed that near the two predicted promoters of miR-27a, each had an Egr1 binding sites, being enriched by 13 reads and 10

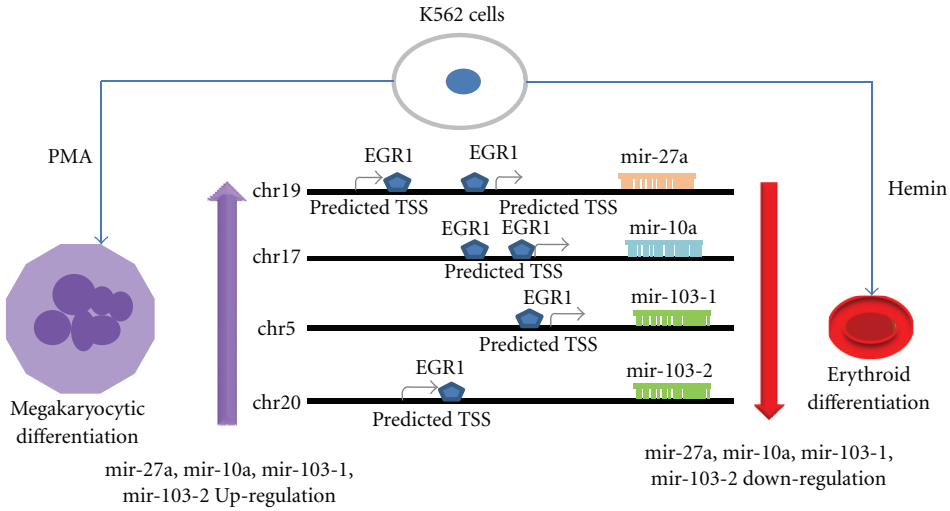


Figure 4: Schematic diagram of miRNA related Egr1's contribution to erythroid/megakaryocytic lineage determination of K562 cells. mir-27a, mir-10a, mir-103-1 and mir-103-2 were found Egr1 binding signals in this study. Based on literature data, when K562 cells were induced by PMA, the expression levels of these miRNAs increased. On the contrary, their expression level were found decreased when K562 cells were induced to erythroid differentiation.

reads, respectively. DNA sequence of the second ChIP-Seq enriched regions was found canonical Egr1 motif (degenerate pattern: 5′-G[T]C[A]GG[T]GGGCG[A,T]-3′) and we found three sites with motif, TAGGGGGCG, TCGTGGGCG, and TCGGGGGCA. So, we suspect that except that Runx1 exerts transcriptional stimulation on miR-27a expression, the binding of Egr1 may also regulate miR-27a expression level and therefore contribute to lineage determination of K562 cells.

Yang et al. characterized and validated that miRNAs of miR-103 and miR-10a exhibited downregulation after hemin induction [13], while Navarro's miRNA microarray analysis showed that miR-103 and miR-10a increased two fold after PMA induction [11]. These differences imply that like miR-27a, the two miRNAs may also participate in lineage determination of K562 cells. Meanwhile, our Egr1 ChIP-Seq results supported that Egr1 binds to the transcription initiation regions of hsa-mir-103-1 (reads count:14, canonical motif: TAGGGGGCG), hsa-mir-103-2 (reads count:25), hsa-mir-10a (two binding sites, reads count:51; canonical motifs: GCGGGGGCA, GCGTGGGCG, TCGGGGGCA, GCGTGGGCA; reads count:9; canonical motifs, GAGGGGGCG), so Egr1 would be the candidate transcription factor which regulates the expression changes of miR-103 and miR-10a. Our ChIP-PCR assay confirmed the Egr1-mir-10a binding. All these data provides key clues for establishment of regulation network of K562 cell differentiation. The relationship between Egr1 and its targeting miRNAs discussed above is summarized in Figure 4.

5. Conclusion

By ChIP-Seq, we provide a global binding profile between the transcription factor Egr1 and its targeting miRNA genes in PMA-treated K562 cells and found that EGR1 binds to promoters of 124 miRNAs and precusors of 125 miRNAs. Among the miRNAs which are bound by Egr1, miR-10a, miR-25, miR-23b, miR-135b, miR-130b, miR-326, miR-141, miR-152, miR-199b, miR-95, and let-7g are validated according to ChIP-PCR.

Acknowledgment

This paper is supported by Project no. 30871393 and no. 30973375 from the National Natural Science Foundation of China and the Project no. 2006AA020702 from the National High-Tech Research and Development Program of China.

References

[1] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 522–531, 2004.

[2] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[3] S. Griffiths-Jones, H. K. Saini, S. V. Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, pp. D154–D158, 2008.

[4] S. Sassen, E. A. Miska, and C. Caldas, "MicroRNA—implications for cancer," *Virchows Archiv*, vol. 452, no. 1, pp. 1–10, 2008.

[5] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.

[6] X.-F. Huo, J. Yu, H. Peng et al., "Differential expression changes in K562 cells during the hemin-induced erythroid differentiation and the phorbol myristate acetate (PMA)-induced megakaryocytic differentiation," *Molecular and Cellular Biochemistry*, vol. 292, no. 1-2, pp. 155–167, 2006.

[7] A. R. Green, E. DeLuca, and C. G. Begley, "Antisense SCL suppresses self-renewal and enhances spontaneous erythroid differentiation of the human leukaemic cell line K562," *The EMBO Journal*, vol. 10, no. 13, pp. 4153–4158, 1991.

[8] T. Cheng, Y. S. Wang, and W. Dai, "Transcription factor egr-1 is involved in phorbol 12-myristate 13-acetate- induced megakaryocytic differentiation of K562 cells," *The Journal of Biological Chemistry*, vol. 269, no. 49, pp. 30848–30853, 1994.

[9] A. R. Green, S. Rockman, E. DeLuca, and C. G. Begley, "Induced myeloid differentiation of K562 cells with down-regulation of erythroid and megakaryocytic transcription factors: a novel experimental model for hemopoietic lineage restriction," *Experimental Hematology*, vol. 21, no. 4, pp. 525–531, 1993.

[10] O. Ben-Ami, N. Pencovich, J. Lotem, D. Levanon, and Y. Groner, "A regulatory interplay between miR-27a and Runx1 during megakaryopoiesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 1, pp. 238–243, 2009.

[11] F. Navarro, D. Gutman, E. Meire et al., "miR-34a contributes to megakaryocytic differentiation of K562 cells independently of p53," *Blood*, vol. 114, no. 10, pp. 2181–2192, 2009.

[12] J.-Y. Yuan, F. Wang, J. Yu, G.-H. Yang, X.-L. Liu, and J.-W. Zhang, "MicroRNA-223 reversibly regulates erythroid and megakaryocytic differentiation of K562 cells," *Journal of Cellular and Molecular Medicine*, vol. 13, no. 11-12, pp. 4551–4559, 2009.

[13] G.-H. Yang, F. Wang, J. Yu, X.-S. Wang, J.-Y. Yuan, and J.-W. Zhang, "MicroRNAs are involved in erythroid differentiation control," *Journal of Cellular Biochemistry*, vol. 107, no. 3, pp. 548–556, 2009.

[14] R. Garzon, F. Pichiorri, T. Palumbo et al., "MicroRNA fingerprints during human megakaryocytopoiesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 13, pp. 5078–5083, 2006.

[15] J. Lu, S. Q. Guo, B. L. Ebert et al., "MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors," *Developmental Cell*, vol. 14, no. 6, pp. 843–853, 2008.

[16] A. Gashler and V. P. Sukhatme, "Early growth response protein 1 (Egr-1): prototype of a zinc-finger family of transcription factors," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 50, pp. 191–224, 1995.

[17] E. S. Silverman and T. Collins, "Pathways of Egr-1-mediated gene transcription in vascular biology," *American Journal of Pathology*, vol. 154, no. 3, pp. 665–670, 1999.

[18] C. T. Liu, J. Yao, D. Mercola, and E. Adamson, "The transcription factor EGR-1 directly transactivates the fibronectin gene and enhances attachment of human glioblastoma cell line U251," *The Journal of Biological Chemistry*, vol. 275, no. 27, pp. 20315–20323, 2000.

[19] D. Gitenay and V. T. Baron, "Is EGR1 a potential target for prostate cancer therapy?" *Future Oncology*, vol. 5, no. 7, pp. 993–1003, 2009.

[20] A. Sharma, M. Kumar, J. Aich et al., "Posttranscriptional regulation of interleukin-10 expression by hsa-miR-106a," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5761–5766, 2009.

[21] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[22] K. E. Boyd, J. Wells, J. Gutman, S. M. Bartley, and P. J. Farnham, "c-Myc target gene specificity is determined by a post-DNA-binding mechanism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 23, pp. 13887–13892, 1998.

[23] SOLiD Analysis Tools, http://solidsoftwaretools.com/gf/project/corona/.

[24] H. K. Ji, H. Jiang, W. X. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, vol. 26, no. 11, pp. 1293–1300, 2008.

[25] C. Liu, E. Adamson, and D. Mercola, "Transcription factor EGR-1 suppresses the growth and transformation of human HT-1080 fibrosarcoma cells by induction of transforming growth factor β1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 21, pp. 11831–11836, 1996.

[26] A. Marson, S. S. Levine, M. F. Cole et al., "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic dtem cells," *Cell*, vol. 134, no. 3, pp. 521–533, 2008.

[27] R. M. Kuhn, D. Karolchik, A. S. Zweig et al., "The UCSC genome browser database: update 2009," *Nucleic Acids Research*, vol. 37, pp. D755–D761, 2009.

[28] L. A. Boyer, T. I. Lee, M. F. Cole et al., "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, vol. 122, no. 6, pp. 947–956, 2005.

[29] G. L. Papadopoulos, M. Reczko, V. A. Simossis, P. Sethupathy, and A. G. Hatzigeorgiou, "The database of experimentally supported targets: a functional update of TarBase," *Nucleic Acids Research*, vol. 37, pp. D155–D158, 2009.

[30] Q. Zheng and X. J. Wang, "GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis," *Nucleic Acids Research*, vol. 36, pp. W358–W363, 2008.

[31] J. C. Bryne, E. Valen, M.-H. E. Tang et al., "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update," *Nucleic Acids Research*, vol. 36, pp. D102–D106, 2008.

[32] A. Kubosaki, Y. Tomaru, M. Tagami et al., "Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation," *Genome Biology*, vol. 10, no. 4, article R41, 2009.

[33] C. Tang, X. Shi, W. Wang, et al., "Global analysis of in vivo EGR1-binding sites in erythroleukemia cell using chromatin immunoprecipitation and massively parallel sequencing," *Electrophoresis*. In press.

[34] K. Tu, H. Yu, Y.-J. Hua et al., "Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms," *Nucleic Acids Research*, vol. 37, no. 18, pp. 5969–5980, 2009.

[35] P. Landgraf, M. Rusu, R. Sheridan et al., "A mammalian microRNA expression atlas based on small RNA library sequencing," *Cell*, vol. 129, no. 7, pp. 1401–1414, 2007.

[36] C. Z. Jiang and B. F. Pugh, "Nucleosome positioning and gene regulation: advances through genomics," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 161–172, 2009.

[37] O. Hobert, "Gene regulation by transcription factors and MicroRNAs," *Science*, vol. 319, no. 5871, pp. 1785–1786, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 134232, 12 pages
doi:10.1155/2010/134232

Research Article

Serine Protease Variants Encoded by *Echis ocellatus* Venom Gland cDNA: Cloning and Sequencing Analysis

S. S. Hasson,[1] R. A. Mothana,[2] T. A. Sallam,[3] M. S. Al-balushi,[1] M. T. Rahman,[1] and A. A. Al-Jabri[1]

[1] *Division of Immunology, Department of Microbiology and Immunology, College of Medicine and Health Sciences, Sultan Qaboos University, P.O. Box 35, Muscat, 123, Oman*
[2] *Department of Pharmacognosy, College of Pharmacy, King Saud University, P.O. Box 2457, Riyadh 11451, Saudi Arabia*
[3] *Department of Community Health, Faculty of Medical Sciences, Al-Baha University, Al-Baha, P.O. Box 2457, Al-Baha 11451, Saudi Arabia*

Correspondence should be addressed to S. S. Hasson, shyahasson@yahoo.co.uk

Received 28 May 2010; Accepted 20 July 2010

Academic Editor: Fuli Yu

Envenoming by *Echis* saw-scaled viper is the leading cause of death and morbidity in Africa due to snake bite. Despite its medical importance, there have been few investigations into the toxin composition of the venom of this viper. Here, we report the cloning of cDNA sequences encoding four groups or isoforms of the haemostasis-disruptive Serine protease proteins (SPs) from the venom glands of *Echis ocellatus*. All these SP sequences encoded the cysteine residues scaffold that form the 6-disulphide bonds responsible for the characteristic tertiary structure of venom serine proteases. All the *Echis ocellatus EoSP* groups showed varying degrees of sequence similarity to published viper venom SPs. However, these groups also showed marked intercluster sequence conservation across them which were significantly different from that of previously published viper SPs. Because viper venom SPs exhibit a high degree of sequence similarity and yet exert profoundly different effects on the mammalian haemostatic system, no attempt was made to assign functionality to the new *Echis ocellatus EoSPs* on the basis of sequence alone. The extraordinary level of interspecific and intergeneric sequence conservation exhibited by the *Echis ocellatus EoSPs* and analogous serine proteases from other viper species leads us to speculate that antibodies to representative molecules should neutralise (that we will exploit, by epidermal DNA immunization) the biological function of this important group of venom toxins in vipers that are distributed throughout Africa, the Middle East, and the Indian subcontinent.

1. Introduction

Envenoming resulting from snake bites is an important public health hazard in many regions, particularly in tropical and subtropical countries [1, 2]. The saw-scaled viper *Echis ocellatus* is the most abundant [3] and medically important viper species in West Africa [4]. Envenoming by saw-scaled viper (*Viperidae: Echis*) species is thought to be responsible for more snakebite deaths worldwide than any other snake genus [5]. In northern Nigeria, *E. ocellatus* is responsible for 95% of all envenoming by snakes [6], causing several hundred deaths annually [7]. The precise incidence of snakebite is difficult to determine and is often grossly underestimated, but in some areas of the Nigerian savannas, victims of *E. ocellatus* envenoming may occupy more than 10% of hospital beds [8]. In the Benue valley of Nigeria, for example, the estimated incidence is 497 per 100 000 population per year with 10%–20% untreated mortality [9]. Local effects of *Echis* viper envenoming include pain, swelling, blistering, and haemorrhage which, in severe cases, can lead to necrosis, permanent disfigurement, and even amputation of the affected limb [10]. Systemic effects include potentially lethal consumption coagulopathy, haemorrhage and hypovolaemic shock [10].

Snake venoms contain a great variety of toxic proteases [11, 12]. Many of these components are proteases, for example, metalloproteases [13], serine proteases [14], phospholipases A_2 [15] and C-type lectins [16] and mediate their toxicity by either stimulating or inhibiting the haemostatic system of human victims or experimental animals, resulting

in clinical complications of blood clotting or uncontrolled haemorrhage [12, 17–19]. Several of these proteinases cleave plasma proteins of the victims in a specific manner with varying degrees of substrate specificity. Thus, while some serine proteases have both fibrinogenolytic and fibrinolytic activities, others have only fibrinogenolytic activity and are called "thrombin-like" proteases [19–25]. Approximately 100 snake venom toxins have been identified as "thrombin-like" enzymes activating the blood coagulation factor [26]. These "thrombin-like" proteases hydrolyze fibrinogen specifically and release either fibrinopeptide A or B or both [27] resulting in the disruption of the blood coagulation system by producing abnormal fibrin clots composed of short polymers that are rapidly dispersed and no longer cross-linked by activated factor XIII [28].

Another group of serine proteases of *Batroxobin, Crotalase,* and Ancrod venoms affect other substrates, for example, plasminogen [27] by cleaving fibrinogen in manner distinct from that of thrombin. Other venom serine proteases function like mammalian kallikrein (or kininogenase) releasing bradykinin from kininogen [29–31] and are called "kallikrein-like" proteases [29], an example of this is halystase [32], a kallikrein-like serine protease isolated from *A. halys blomhoffii* venom, which cleaves the β chain at Arg^{42} and slowly degrades the α chain of fibrinogen to generate a product that is no longer converted to normal fibrin clots by thrombin; this results in both reduction of blood pressure as well as inhibiting fibrinogen clotting in the victims. Another kallikrein-like serine protease with potent biological activity but with different physicochemical properties from those of halystase has been isolated from the venoms of *A. caliginosus, C. atrox, C. viridis, and Trimeresurus mucrosquamatus* [29, 30, 32–34]. The latter showed both a strong β-fibrinogenolytic and kallikrein-like activities, cleaving β-chain of fibrinogen molecules specifically and releasing bradykinin from kininogen, respectively. Moreover, the purified enzymes indicated that they have specificities different from thrombin and thrombin-like proteases of snake venom reported previously by decreasing fibrinogen levels in plasma and prolonging bleeding without formation of fibrin clots. They also exhibit amidase activity against *N*-benzoyl-Pro-Phe-Arg-*p*-nitroanilide, which is a specific synthetic substrate for kallikrein-like proteases.

In addition, there have been a few reports on venom serine proteases with a unique activity, such as ACC-C, a protein C activator isolated from the *A. contortrix* venom [35] (which inhibits blood coagulation by inactivating the activated forms of factor V and VIII), a plasminogen activator such as TSV-PA isolated from the *T. stejnegeri* venom [36, 37], PA-BJ, a platelet aggregating enzyme isolated from the *B. jararaca* and *Trimeresurus mucrosquamatus* venoms [38], and RVV-V, a factor V-activating enzyme isolated from the *V. russelli* venom [39].

These data indicate that snake venom serine proteases comprise an enzyme superfamily with multifunctional activities that may have diverged or have undergone gene duplication resulting in alteration of their biological properties during the process of evolution thus acquiring special functions [40, 41]. Although a considerable amount of data is now available, no standardised grouping of these venom serine proteases has yet been documented. However, in 2001 Wang et al. [27] compared sequences of 40 serine proteinases isolated from different snake venoms, using a constructed phylogram in which such sequences were clustered into three groups designated as coagulating enzymes, kininogenases, and plasminogen activators.

No Serine proteinases have yet been purified from venom of the West African saw-scaled viper *Echis ocellatus,* in particular or for members of the *Echis* genus in general. However, the fact that the serine protease superfamily was important in the venom of the Viperidae suggested that such enzymes should be present in the venom of *E. ocellatus* and that serine protease-specific antibodies are likely to be an important factor in *E. ocellatus* envenoming. We therefore screened the *E. ocellatus* cDNA library in order to isolate and characterise different isoforms or variants of this enzyme superfamily.

2. Materials and Methods

2.1. Animals. Adult *E. ocellatus* (Nigeria) carpet viper used in this study was maintained in the herpetarium, Liverpool School of Tropical Medicine, Liverpool, UK.

2.2. Extraction of Total Venom Gland RNA and Construction of cDNA Libraries. Venom glands were dissected from three *Echis ocellatus* snakes. The vipers were sacrificed 3 days after venom extraction when toxin gene transcription rates are at a peak. Glands were homogenized under liquid Nitrogen and total RNA extracted using guanidinium thiocyanate-phenol-chloroform as described previously [15]. Lambda phage cDNA libraries for *E. ocellatus* were constructed by RT-PCR using the SMART cDNA library construction kit (Clontech, California, USA). The lambda phage of the *E. ocellatus* was packaged using Gigapack III Gold Packaging Extract (Stratagene) and boiled for 5 min prior to being used as targets of polymerase chain reaction (PCR) amplification.

2.3. Isolation and Analysis of cDNA Sequences. A PCR strategy [42] was used to isolate sequences encoding serine proteinases from the cDNA libraries. A sense primer (5V-**GGA-TCC**-ATG-GTG-CTG-ATC-AGA-GTG-CTA-**ATC**-GCA-3V) and an antisense primer (5V-**CTC-GAG**-TGG-GGG-GCA-AGT-CGC-AGT-TGT-ATT-TCC-3V) complimentary to highly conserved amino-terminal signal peptide (M-V-L-I-R-V) and to the less conserved carboxy-terminal (T-T-A-T-C-P-P) domains of published serine proteinases DNA sequences of related viper species were synthesized commercially (Sigma-Genosys, UK). A TAG stop codon was inserted in the 3′ primer and BamH1 and Xho1 restriction endonuclease sites (bold) were included in the 5′ and 3′ primers, respectively, to facilitate future subcloning into mammalian expression plasmids. PCR was performed using an initial denaturation (95°C—6 minutes) and annealing (55°C—1 minute) step, followed by 35 cycles (1 minute each) of extension (74°C), denaturation (94°C), and annealing (55°C), and a terminal extension step (7 minutes)

at 72°C in a thermal cycler (Gene Cycler, BioRad, Hercules, CA, USA). The inclusion of water-only controls with each PCR reaction allowed us to monitor and prevent cross-over contamination. The amplicons were subcloned into the TA cloning vector, pCR 2.1-TOPO, (Invitrogen, Groningen, The Netherlands) and used to transform chemically competent *E. coli* cells (TOP10F', Invitrogen) under ampicillin selection. Plasmid DNA was extracted (Mini-spin prep kit, Qiagen, Hilden, Germany) and digested with BamH1 and Xho1 at 37°C to select plasmids containing inserts of the predicted size for DNA sequencing. DNA sequencing was carried out by the dideoxy-nucleotide chain-termination method in a Beckman Coulter CEQk 2000 XL DNA Analysis System. To confirm that the cDNA sequences encoded CTLs, the predicted amino acid sequences were subjected to BLAST searches of the GenBank, PDB, SwissProt, PIR, and PRF databases. All the cDNAs exhibited significant sequence homology to Serine protienases of related vipers. The CLUSTALW program [43] with PAM 250 residue weight matrix was used to align deduced amino acid sequences representing each *E. ocellatus* Serineprotienases isoforms with analogues in venoms from related Viperidae species as illustrated in Table 1. Serine proteinase (CAB62591) from V. lebetina [44], Serine protease 1 (AAR24534) from B. gabonica [45], Thrombin-like enzyme pre. (AAK12273) from *D. acutus* [46], Venom serine protease 5 (AAN52350) from *T. stejnegeri* [47], Serine proteinase 3 pre. (O13063) from *gramineus* [48], Serine proteinase A Precursor (Q9PTU8) from *B. jararaca* [46, 49], Serine proteinase 2A pre. (O13060) from *T. gramineus* [45, 48], Serine protease (AAP42416) from *B. jararacussu* [50], KN-BJ2 (BAA20283) from *B. jararaca* [51], Serine proteinase 1 pre. (AAG10788) from *T. jerdonii* [52], Thrombin-like serine protease (AAL68708) from *G. ussuriensis* [53], and, finally, Serine protease catroxase I pre. (AAL77226) from *C. atrox* [54]. The phylogenetic trees constructed from the above alignments were generated by a neighbour-joining [55] algorithm in Lasergene software (DNASTAR, USA). The predicted antigenic profile [56] of the published and new *Echis ocellatus* serine protease (*EoSer*) isoforms analysed here was determined using Protean Software (DNASTAR).

3. Results

3.1. Isolation of cDNAs Encoding E. Ocellatus Serine Protease. PCR screening of the *Echis ocellatus* venom gland cDNA libraries resulted in a total of 14 *E. ocellatus* (*Eo*) cDNAs whose sequences matched (BLAST searches) those of published Serine proteases. The cDNAs consisted of 822 nucleotides (Figures 1 and 2(a)) and were predicted to encode an open reading frame proteins of 264 amino acids (28.5 kDa) (Figure 2(b)). Alignment of the predicted amino acid sequences of the 14 specific cDNAs encoding the *EoSP* proteins (Figure 2(b)) revealed sequence variations. The sequence similarity between the *EoSP* variants proteins was less than 60% for the mature protein-coding region but over 90% for regions coding both the signal peptide and the carboxyl-terminal end. Where two or more identical sequences were obtained from any one of these libraries, a single representative cDNA was used for subsequent analysis. Structural properties analysis (Emin algorithm-DNASTAR, USA) (Figure 3) was used to categorise the 14 Serine protease sequences into four distinct groups, based solely on sequence alignment.

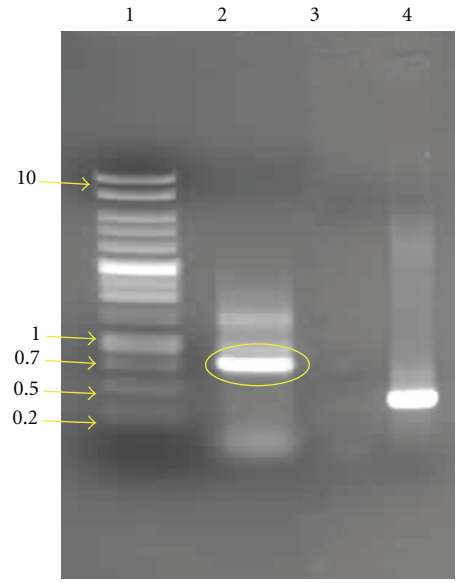
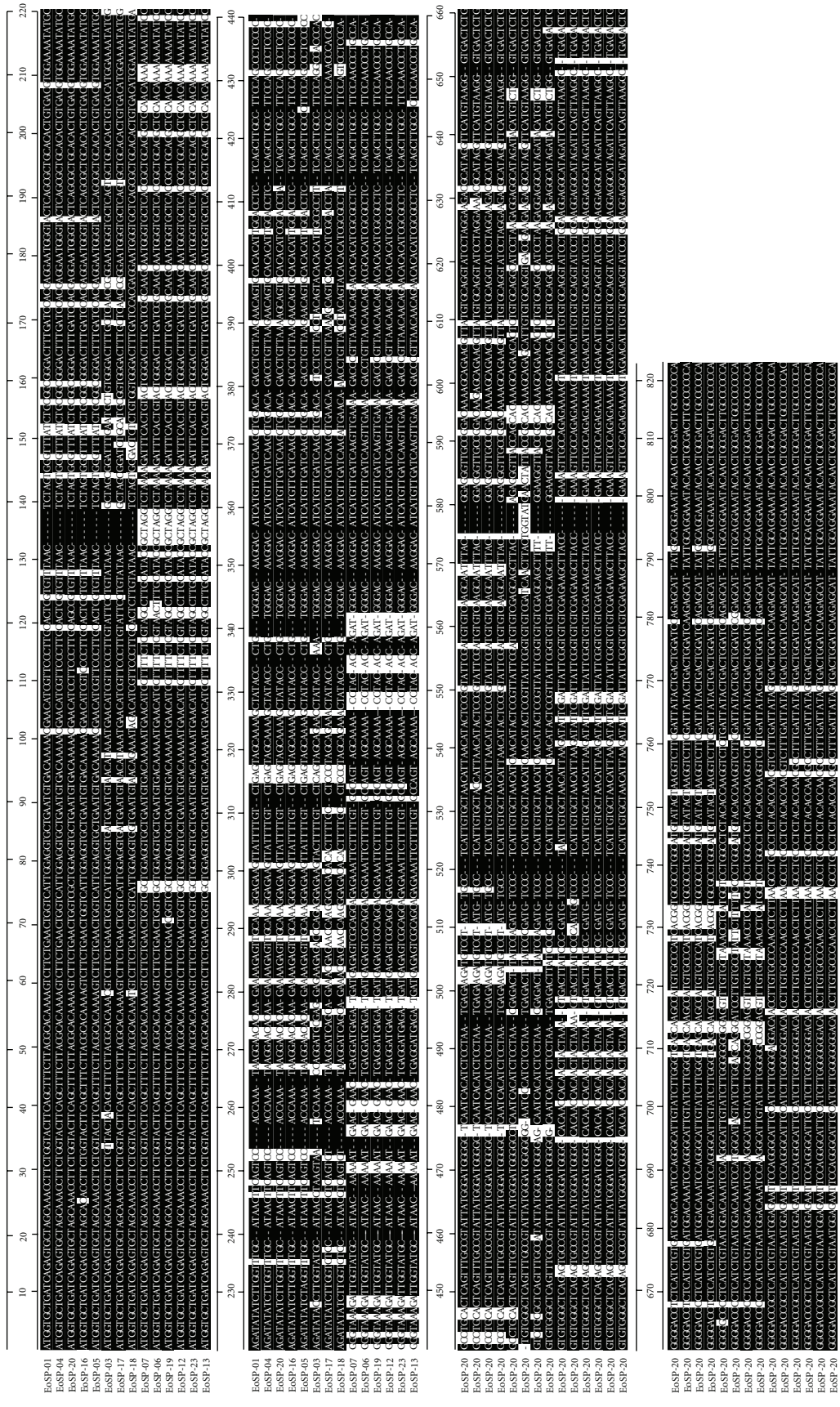


Figure 1: PCR product of the *E. ocellatus* serine proteases. Analysis of PCR amplification products by 0.7% agarose gel electrophoresis. Bands were visualised using the ultraviolet transillumination. Lane 2: represents the amplified PCR product (circled) of about 800 bp from *E. ocellatus* venom glands cDNA compared with Lane 1:1 kb ladder DNA-marker bands, of known molecular weight. Lanes 3 and 4 represent a H_2O negative control and a SOD positive control, respectively.

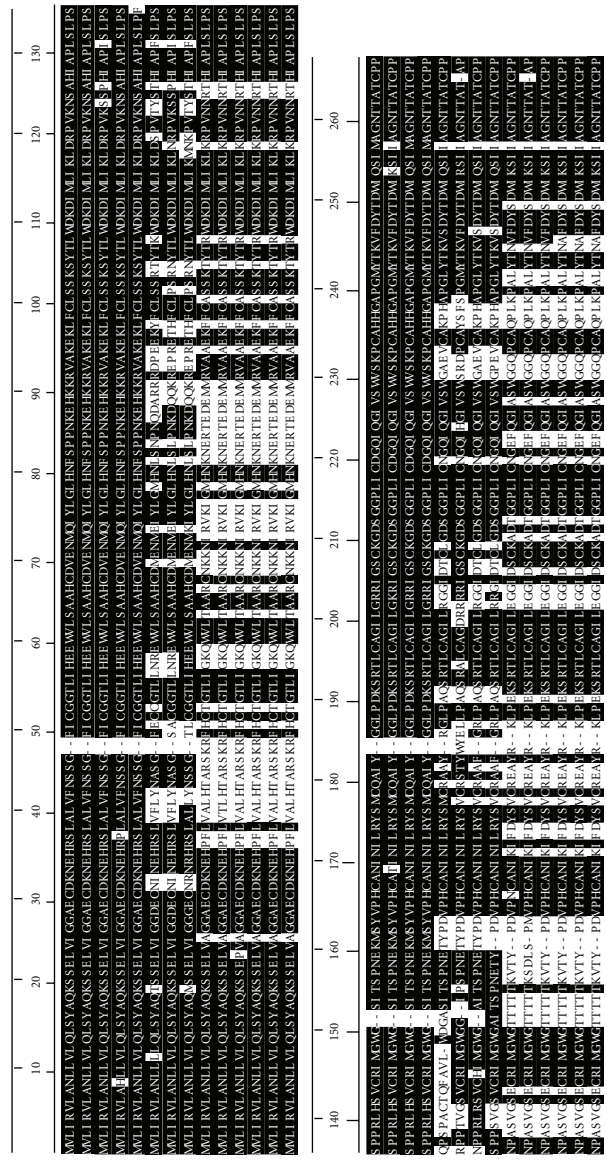
3.2. BLAST Search of the Predicted Amino Acid Sequence. Accession numbers assigned to the new *Echis ocellatus* Serine protease sequences are as follows: "group 1" *EoSer*-1 (GU562413), "group 2" EoSer-3 (GU592440), "group 3" *EoSer*-17 (GU592441), and "group 4" *EoSer*-7 (GU592439). The predicted amino acid sequences of the *EoSP-01, EoSP-03, EoSP-07,* and *EoSP-17* were submitted to BLAST searches of the genetic data bases and their similarity to published viper serine protease (Table 1) confirmed that the *EoSP* cDNAs encoded serine proteases.

3.3. Comparison of E. Ocellatus cDNAs with Analogous Serine Proteases from Other Viper Species. All the *EoSer*-variants contained the serine protease-consensus 24 amino acid signal peptide sequence (Figure 4, arrows), including the six-amino acids-activated motif. The signal peptide residues were followed by a protease domain of 236 residues. The deduced primary structures of all *EoSP* cDNA clones include the requisite, highly conserved, 12 cysteine residues that form the 6-disulphide bonds responsible for the characteristic tertiary structure of venom serine proteases. The complete amino acid sequences of the *EoSP* variants were aligned with those of other venom serine proteases (Figure 4). Viper



(a)

Figure 2: Continued.



(b)

Figure 2: (a) The nucleotide sequence of the fourteen *E. ocellatus* venom gland cDNAs resulting from PCR amplification. (b) Deduced amino acid sequences of *E. ocellatus* venom gland serine protease cDNAs.

venom SP sequences in the genetic databases were compared with the *E. o* groups (Table 2 and Figure 4) by BLAST. Groups 1–4 represent novel, highly similar, SP isomers with less than 65% sequence similarity to analogues in related viper species. Group 4 showed the greatest sequence similarity (80% and 82%) to the Serine protease of the African *V. lebetina* and *B. gabonica* vipers, respectively. Of all the *EoSP* clusters seemed to represent a SP sequence which showed the highest sequence similarity range between 62% to 70% to the SP of the vipers. None of the clusters showed more than 72% sequence similarity to the partial peptide sequences for the Thrombin-like serine protease isolated from the venom of the *G. ussuriensis* viper [52]. Similarly, the Serine protease catroxase I pre. of *C. atrox* venom showed no greater than 65% sequence similarity to any of the *EoSP* sequences.

3.4. Predicted Antigenic Profile Analysis of E. ocellatus Serine Proteases with Analogous Molecules. Since the main focus of our research is to develop toxin neutralising antibodies by immunisation with DNA encoding specific toxins in venoms of the most medically important African vipers [15, 59, 60], we next compared the algorithm-predicted immunogenicity of the *E. ocellatus* serine protease cluster cDNA sequences with those of all the published SPs from vipers of African

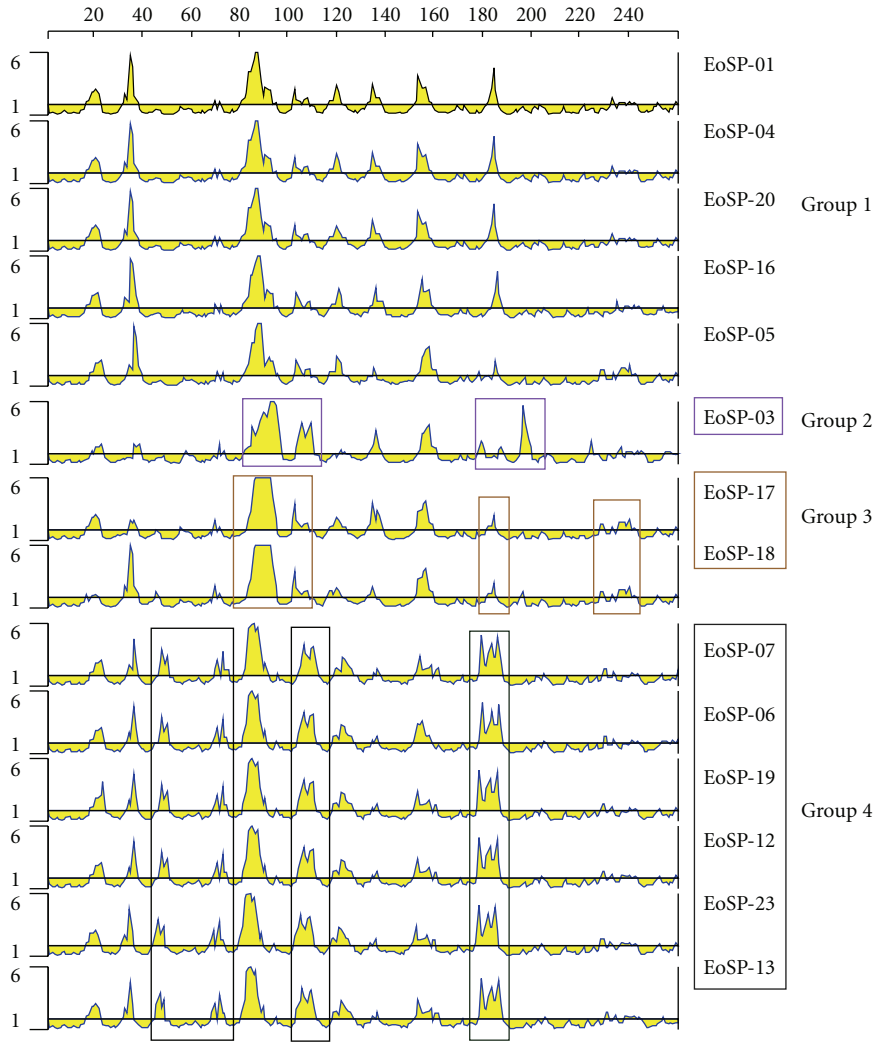


Figure 3: Differentiation of the fourteen cDNA-encoding *E. ocellatus* venom gland serine proteases. The predicted surface probabilities (Emin algorithm, DNASTAR, USA) of the 14 *E. ocellatus* serine protease cDNAs were aligned. The boxed areas indicate group specific structural motifs.

Table 1: Percent sequence similarity between *E. ocellatus* serine proteases and analogous molecules from related viper species.

Species	Accession no.	References	Serine protease	*EoSP*-1	*EoSP*-3	*EoSP*-7	*EoSP*-17
V. lebetina	CAB62591	Siigur et al. [44]	Serine proteinase	65	66	80	64
B. gabonica	AAR24534	Francischetti et al. [45]	Serine protease 1	62	63	82	65
D. acutus	AAK12273	Liang et al. [46]	Thrombin-like enzyme pre.	67	67	71	70
T. stejnegeri	AAN52350	Lee and Zhang [47]	Venom serine protease 5	66	66	65	69
T. gramineus	O13063	Deshimaru et al. [48]	Serine proteinase 3 pre.	71	71	61	76
B. jararaca	Q9PTU8	Murayama, [49]	Serine proteinase A pre.	66	68	66	74
T. gramineus	O13060	Deshimaru et al. [48]	Serine proteinase 2A pre.	65	65	73	70
B. jararacussu	AAP42416	Kashima et al. [50]	Serine protease	63	63	72	68
B. jararaca	BAA20283	Serrano et al. [51]	KN-BJ2	69	68	62	69
T. jerdonii	AAG10788	Lu et al. [52]	Serine proteinase 1 pre.	65	65	71	70
G. ussuriensis	AAL68708	Zhao et al. [53]	Thrombin-like serine protease	67	67	72	71
C. atrox	AAL77226	Tsai et al. [54]	Serine protease catroxase I pre.	66	65	63	65

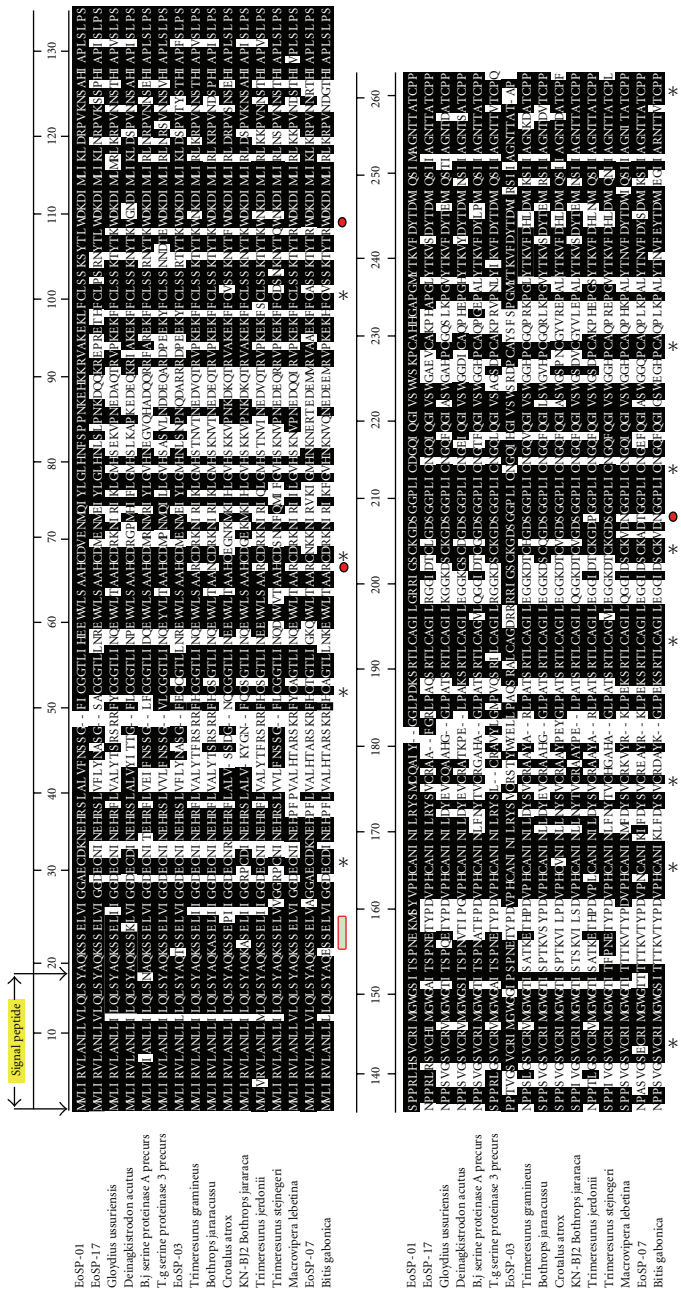


FIGURE 4: Amino acid sequence similarity between *EoSP* Variants and serine proteases from related vipers. The residues shaded in black correspond to residues that are identical to *EoSP-01*. The asteriks [*] represented the tweleve conserved cysteine residues. The catalytic traid **His**/**Arg** (67), **Asp** (110) and **Ser** (208) are represented in red circles. Activated peptide where the mature proteine cleaved is represented by green rectangle.

origin (Figure 5). The predicted antigenic profiles of the published and new *E. ocellatus* serine proteases were analysed as shown in Figure 5 using Protean Software (DNASTAR, USA) [53]. The deduced signal peptide domains of the *EoSP* variants are separated by a vertical dotted line, as these would normally be cleaved from the native proteins during posttranslational. The thin vertical boxes depict the residues comprising the catalytic traid, H/R/N, D/G/N, and S/P/N/T (67, 110, and 208), that show the greatest immunogenic domains conservation common to all the new and published African viper venom SPs sequences as demonstrated in Figure 5.

4. Discussion

Serine proteases are a major component of viper venoms and are thought to disrupt several distinct elements of the blood coagulation system of envenomed victims.

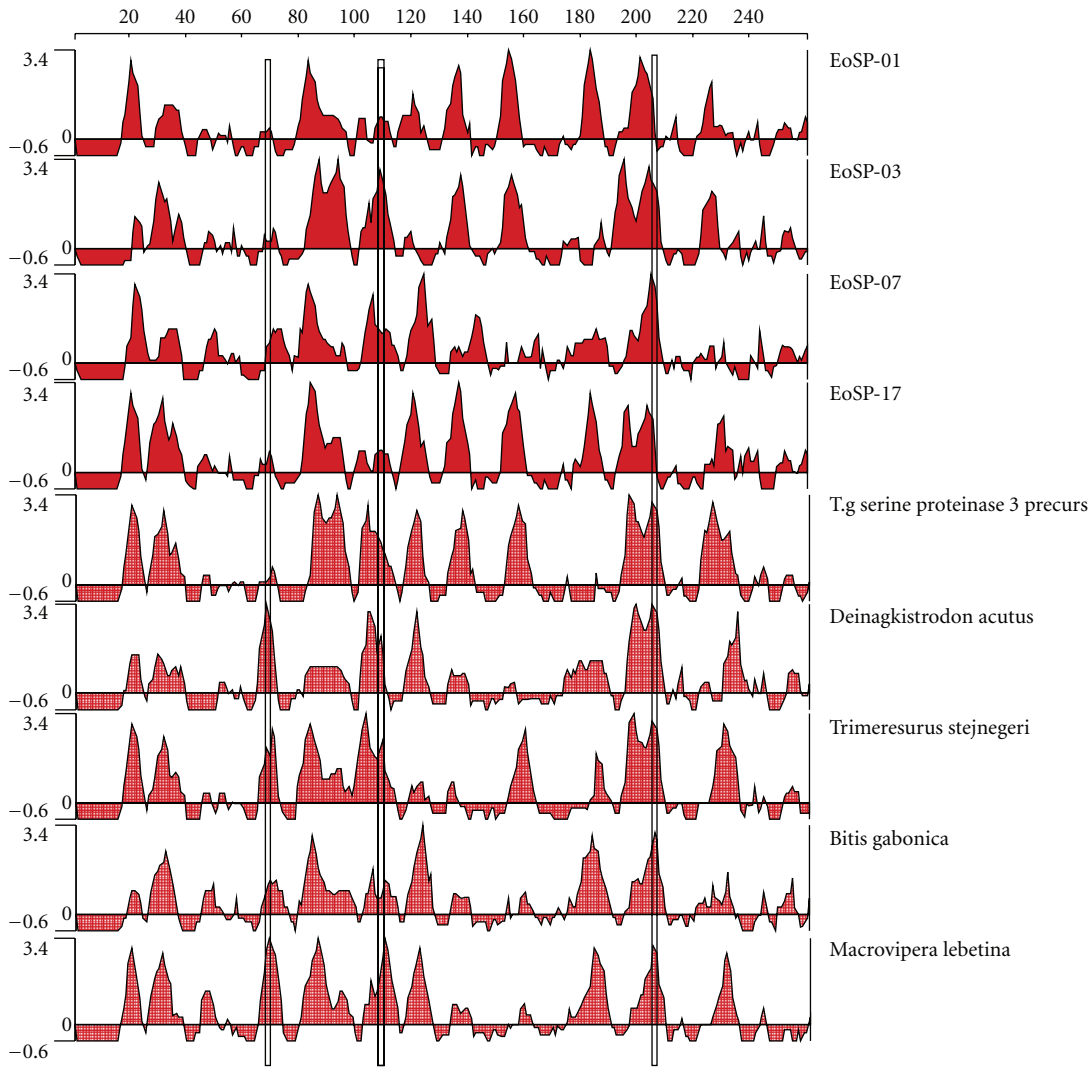


Figure 5: Comparison of antigenic profile of the *EoSer* variants with analogous serine proteases used in Figure 4. The top horizontal scale represents the number of amino acid residues. The conserved signal peptide is separated from the mature protein by a vertical dotted line. The three vertical boxes were drawn to indicate the conserved catalytic traid regions described in the text.

Table 2: Comparison of amino acid motifs which are responsible for the potent effects and characterisation of some published venom serine proteases with the four *EoSP* cDNAs.

Amino acid	TSV-PA	Batroxobin	Ancrod	*EoSP*-1	*EoSP*-17	*EoSP*-3	*EoSP*-7	References
H/R	H^{57}	$\mathbf{H}^{57}$	H^{57}	$\mathbf{H}^{67}$	H^{67}	$\mathbf{H}^{67}$	R^{67}	Braud et al. [57]
D	D^{102}	$\mathbf{D}^{102}$	D^{102}	$\mathbf{D}^{112}$	D^{112}	$\mathbf{D}^{112}$	D^{112}	
S	S^{195}	$\mathbf{S}^{195}$	S^{195}	$\mathbf{S}^{208}$	S^{208}	$\mathbf{S}^{208}$	T^{208}	
H	H^{192}	$\mathbf{G}^{192}$	N^{192}	$\mathbf{K}^{205}$	L^{205}	$\mathbf{K}^{205}$	K^{205}	
F	F^{193}	$\mathbf{G}^{193}$	S^{193}	$\mathbf{G}^{206}$	G^{206}	$\mathbf{G}^{206}$	A^{206}	Guinto et al. [58]
D	D^{189}	$\mathbf{D}^{189}$	D^{189}	$\mathbf{G}^{202}$	D^{202}	$\mathbf{G}^{202}$	D^{202}	
P	P^{225}	$\mathbf{P}^{225}$	P^{225}	$\mathbf{P}^{235}$	P^{235}	$\mathbf{P}^{235}$	P^{235}	
P	P^{219}	$\mathbf{P}^{219}$	P^{219}	$\mathbf{P}^{228}$	V^{228}	$\mathbf{P}^{228}$	P^{228}	Braud et al. [57]
D	D^{96}	$\mathbf{N}^{96}$	R^{96}	$\mathbf{Y}^{106}$	Y^{106}	$\mathbf{Y}^{106}$	Y^{106}	Lee and Zhang [47]
D	D^{97}	$\mathbf{V}^{97}$	T^{97}	$\mathbf{T}^{107}$	T^{107}	$\mathbf{T}^{107}$	T^{107}	
E	E^{98}	$\mathbf{I}^{98}$	S^{98}	$\mathbf{L}^{108}$	L^{108}	$\mathbf{K}^{108}$	R^{108}	

HDS: Catalytic Traid; H/F: substrate specificity; D & P: Architecture of water channel; P: Evolutionary region to kallikrein; DDE: substrate specificity to plasminogen.

A detailed understanding of the functions of these enzymes is important for both acquiring a full understanding of the pathology of envenoming and because these venom proteins have shown a vital role in treating blood coagulation disorders.

In general, serine proteinases including fibrinogenolytic enzymes are very abundant in *Viperidae* venoms in which they may account for 20% of their total protein content [61]. The unique specificity of snake venom proteinases makes them potentially useful in research of fibrinogen-depletion and limited proteolysis [62, 63]. This may be due to the existence of multiple forms of serine proteases in the venom of a single viper species which is likely to contribute to the diverse biological effects exerted by the whole venom. Therefore, screening the *E. ocellatus* cDNA library to isolate different isoforms or variants of serine proteases was the aim of this research work. The results obtained in this work provide the first molecular sequence data for *E. ocellatus* serine proteases they also reveal that the serine protease composition of *E. ocellatus* is as complex as that of the better characterised *Viperidae* species. The utilization of PCR amplification of *E. ocellatus* venom gland cDNA with the new viper serine protease-specific primers was successful and produced fourteen cDNAs sequences that were identified (BLAST) as belonging to the serine protease enzyme family. All *EoSP* cDNAs were of similar total length (approximately 0.80 kb, Figure 1) and encoded 260 amino acids (Figure 2(b)) with a predicted molecular weight of 28.5 kDa. To differentiate between the isolated *EoSP* clones a surface probability algorithm was used to assign the 14 *E. ocellatus* serine protease cDNAs into four main groups (Figure 3). A single representative clone from each group was chosen for further analyses as described earlier. The sequence similarity between the *EoSP* variants proteins was less than 60% for the mature protein-coding region but over 90% for regions coding both the signal peptide and the carboxyl-terminal end. Thus the latter two regions are highly conserved, which explains why the PCR experiment to amplify the cDNAs-encoding *EoSP* clones was successful.

The *EoSP* cDNA sequences were confirmed by BLAST searches as encoding serine proteases (Figure 4). The greatest sequence similarity was between *EoSer-7* and *B. gabonica* and *V. labetina* (80% and 85%) with the remaining *EoSP* cDNAs showing 60%–76% sequence similarity with other snake venom serine proteinases as illustrated in Table 1. From the proteins with known biological activity, sequence similarities of the *EoSP* variants (i.e., *EoSer-01*, *EoSer-03*, *EoSer-07* and *EoSer-17*) were 62%–69% with the kinin-releasing and fibrinogen-clotting serine protease (KN-BJ) from venom of *B. jararaca* [51] (Table 1). The putative 18 amino acid signal-peptide of the *EoSP* variants was as conserved (over 90% sequence similarity) as that in the serine proteases of other viper species (Figure 4, arrows). Following the signal peptide all the *EoSP* variants contained the predicted six-amino acid cleavage (activation) site **Q-K/T/M/E-S-S-E-L/P** (Figure 4 in green) as proposed for batroxobin [64]; thus cleavage generates a hydrophilic zymogen peptide, based on the processing site of pre-peptides of mammalian serine proteinases [65–67]. Comparison of the *EoSP* variants with analogous members of the serine protease family revealed that all *EoSP* variants encoded the presumed catalytic triad, which is common to venom serine proteases H67, D110 and S208 as shown in Figure 4. Such residues were highly conserved in groups 1–3, except proteins of group 4 (Figures 2(b) and 4) which contain R instead of H at the same position (Figure 4). Furthermore, comparison of the *EoSP* amino acid sequence alignment with analogous venom serine proteases (Figure 4) revealed a conserved consensus active site of L-T/S-A-A-H/R/N-C corresponding to position 63–68, as previously determined [68]. Most SVSPs are likely to be glycoproteins showing a variable number of N- or O-glycosylation sites in sequence positions that differ from one SVSP to the other [69]. Using the primary structure of *EoSP* variants (Figure 4) the putative N-linked glycosylation sites, Asn-X-Thr/Ser [45], were found and are located at two different positions. *EoSer-01, EoSer-03, and EoSer-17* [$\mathbf{N}^{44}$-$\mathbf{X}^{45}$-$\mathbf{S}^{46}$ and $\mathbf{N}^{257}$-$\mathbf{X}^{258}$-$\mathbf{T}^{259}$] and *EoSer*-07 [$\mathbf{N}^{124}$-$\mathbf{R}^{125}$-$\mathbf{T}^{126}$ and $\mathbf{N}^{257}$-$\mathbf{T}^{258}$-$\mathbf{T}^{258}$]. Although such motifs are thought to be needed for protein stabilization rather than for the catalytic function of the venom enzymes [30], confirmation of the roles of such motifs in venom proteases remain to be investigated. All serine proteases have a common pattern of 6-disulfide bridges [69, 70]. They contain twelve cysteine residues, ten of which form five disulfide bonds, based on the homology with trypsin [64]; the remaining two cysteines form a unique and conserved bridge among SVSPs, involving Cys245e (chymotrypsinogen numbering), found in the C-terminal extension [35].

From the results obtained this was found in all *EoSP* clones (Figure 2(b)) that encoded the common 12 cysteine residues in which are strongly conserved forming putative disulphide bridges which are located at $\mathbf{Cys}^{31}$ $\mathbf{Cys}^{52}$, $\mathbf{C}^{68}$, $\mathbf{C}^{100}$, $\mathbf{C}^{145}$, $\mathbf{C}^{165}$, $\mathbf{C}^{176}$, $\mathbf{C}^{204}$, $\mathbf{C}^{214}$, $\mathbf{C}^{229}$, and $\mathbf{C}^{260}$ (Figure 4). This suggests that the *EoSP* proteins possess a similar tertiary structure to that of other serine proteases which are well characterized.

Despite such sequence and structural conservation, viper venom serine proteases show very divergent effects on haemostasis as previously stated. In some cases certain amino acid sequences have been shown to be responsible for such effects as demonstrated in Table 2. Although such table gives a preliminary prediction of the functional characterization of the *EoSP* cDNAs in comparison with well-known characterized venom serine proteases, it cannot be considered as a functional confirmation or even a categorization strategy to differentiate between the four *EoSP* cDNAs. However, from Figure 4 and Table 2 it can be generally concluded that such comparison demonstrates that the enzymes encoded by the four *EoSP* cDNAs confer multiple haemostasis-disruptive activities to *E. ocellatus* venom. Furthermore, the sequence and predicted structural similarities of these four *EoSP* groups suggest that an antibody generated to one group may be capable of neutralizing the other group of *EoSP*s. To examine this permeability the sequences of *EoSP* groups were subjected to a more specific algorithm that predicted amino acid motifs of high immunogenicity. A protein structure-predicting algorithm [56] has been used (i) to identify domains of strong antigenic potential in the

toxin gene product and (ii) to determine whether these domains are conserved in analogous venom toxin gene products of related vipers. The signal peptide was separated from the mature protein by dotted line as would be cleaved posttranslationally. The peaks shown by the *EoSP*s profile indicate the numerous domains predicted to have a surface location and potential for antibody induction. Although the antigenic peaks of the catalytic traid of the *EoSP*s showed less similarity with that of the analogous venom SPs particularly those at residues 67 and 110, many antigenic residue similarities of *EoSP*s are shared with other SVSPs of related vipers. Therefore, it is likely that antibodies raised by *EoSP* DNA immunisation are likely to possess considerable cross-reactivity and might competitively inhibit the function of these domains in the similar venom toxins of related vipers. However, binding of antibodies specific to conserved antigenic domains without a known function are equally as likely to disrupt protein function by virtue of steric hindrance. The veracity of these speculations need to be confirmed experimentally and thus is a focus of our current research.

In conclusion, the predicted Jameson-Wolf antigenic profiles (DNASTAR, USA) of the *EoSP* variants aligned with very low identity to their (BLAST) analogous serine proteases. This observation strongly suggests that an antibody raised by immunisation with group one *EoSP* DNA is likely to be less effective against the gene products of groups 2, 3, or 4. Therefore additional antibodies generated against antigenic index that showed less conservation will be required.

Acknowledgments

Funding for this project was provided by the Wellcome Trust (RAH, Grant no. 061325), the University of Science and Technology, Yemen, and the Gunter Trust (S. Hasson). The authors would like to thank Dr. R. A. Harrison, Prof. R. D. G. Theakston, and Mr. Paul Rowley for their assistance during extraction of the venom glands from snakes and Dr. A. Nasidi, Federal Ministry of Health, Nigeria, for obtaining the snakes.

References

[1] J. M. Gutiérrez, R. D. G. Theakston, and D. A. Warrell, "Confronting the neglected problem of snake bite envenoming: the need for a global partnership," *PLoS Medicine*, vol. 3, no. 6, article e150, 2006.

[2] WHO, *Rabies and Envenomings: A Neglected Public Health Issue*, Report of a consultative meeting, WHO, Geneva, Switzerland, 2007.

[3] J. F. Trape, G. Pison, E. Guyavarch, and Y. Mane, "High mortality from snakebite in south-eastern Senegal," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 95, no. 4, pp. 420–423, 2001.

[4] S. C. Wagstaff and R. A. Harrison, "Venom gland EST analysis of the saw-scaled viper, *Echis ocellatus*, reveals novel $\alpha 9\beta 1$ integrin-binding motifs in venom metalloproteinases and a new group of putative toxins, renin-like aspartic proteases," *Gene*, vol. 377, no. 1-2, pp. 21–32, 2006.

[5] N. R. Casewell, R. A. Harrison, W. Wüster, and S. C. Wagstaff, "Comparative venom gland transcriptome surveys of the saw-scaled vipers (*Viperidae: Echis*) reveal substantial intra-family gene diversity and novel venom transcripts," *BMC Genomics*, vol. 10, article 564, 2009.

[6] S. S. Hasson, A. A. Al-Jabri, T. A. Sallam, M. S. Al-Balushi, and R. A. A. Mothana, "Antisnake venom activity of *Hibiscus aethiopicus* L. against *Echis ocellatus* and *Naja n. nigricollis*," *Journal of Toxicology*, vol. 2010, Article ID 837864, 8 pages, 2010.

[7] J. P. Chippaux, "The treatment of snake bites: analysis of requirements and assessment of therapeutic efficacy in tropical Africa," in *Perspectives in Molecular Toxinology*, A. Menez, Ed., pp. 457–472, Wiley, New York, NY, USA, 2002.

[8] P. Revault, "Ecology of *Echis ocellatus* and peri-urban bites in Ouagadougou," *Toxicon*, vol. 34, no. 2, p. 144, 1996.

[9] A. G. Habib, S. B. Abubakar, I. S. Abubakar et al., "Envenoming after carpet viper (*Echis ocellatus*) bite during pregnancy: timely use of effective antivenom improves maternal and foetal outcomes," *Tropical Medicine and International Health*, vol. 13, no. 9, pp. 1172–1175, 2008.

[10] D. A. Warrell, N. McD. Davidson, and B. M. Greenwood, "Poisoning by bites of the saw scaled or carpet viper (*Echis carinatus*) in Nigeria," *Quarterly Journal of Medicine*, vol. 46, no. 181, pp. 33–62, 1977.

[11] M. L. D. Weinberg, L. F. Felicori, C. A. Bello et al., "Biochemical properties of a bushmaster snake venom serine proteinase (LV-Ka), and its kinin releasing activity evaluated in rat mesenteric arterial rings," *Journal of Pharmacological Sciences*, vol. 96, no. 3, pp. 333–342, 2004.

[12] A. B. Sallau and M. A. Ibrahim, "Characterization of phospholipase A2 (PLA2) from *Echis ocellatus* venom," *African Journal of Biochemistry Research*, vol. 2, no. 4, pp. 98–101, 2008.

[13] J.-M. Howes, R. D. G. Theakston, and G. D. Laing, "Antigenic relationships and relative immunogenicities of isolated metalloproteinases from *Echis ocellatus* venom," *Toxicon*, vol. 45, no. 5, pp. 677–680, 2005.

[14] S. Liu, M.-Z. Sun, C. Sun, B. Zhao, F. T. Greenaway, and Q. Zheng, "A novel serine protease from the snake venom of *Agkistrodon blomhoffii ussurensis*," *Toxicon*, vol. 52, no. 7, pp. 760–768, 2008.

[15] K. Bharati, S. S. Hasson, J. Oliver, G. D. Laing, R. D. G. Theakston, and R. A. Harrison, "Molecular cloning of phospholipases A2 from venom glands of *Echis* carpet vipers," *Toxicon*, vol. 41, no. 8, pp. 941–947, 2003.

[16] Q. Lu, A. Navdaev, J. M. Clemetson, and K. J. Clemetson, "Snake venom C-type lectins interacting with platelet receptors. Structure-function relationships and effects on haemostasis," *Toxicon*, vol. 45, no. 8, pp. 1089–1098, 2005.

[17] K. Stocker, "Snake venom proteins affecting hemostasis and brinolysis," in *Medical Use of Snake Venom Proteins*, K. F. Stocker, Ed., pp. 97–160, CRC Press, Boca Raton, Fla, USA, 1990.

[18] T. Matsui, Y. Fujimura, and K. Titani, "Snake venom proteases affecting hemostasis and thrombosis," *Biochimica et Biophysica Acta*, vol. 1477, no. 1-2, pp. 146–156, 2000.

[19] F. S. Markland Jr., "Preface: snake venoms and hemostasis," *Toxin Reviews*, vol. 25, no. 4, pp. 319–321, 2006.

[20] F. S. Markland Jr., "Inventory of α- and β-fibrinogenases from snake venoms. For the Subcommittee on Nomenclature of Exogenous Hemostatic Factors of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis," *Thrombosis and Haemostasis*, vol. 65, no. 4, pp. 438–443, 1991.

[21] H. Pirkle, "Thrombin-like enzymes from snake venoms: an updated inventory," *Thrombosis and Haemostasis*, vol. 79, no. 3, pp. 675–683, 1998.

[22] H. Pirkle and K. Stocker, "Thrombin-like enzymes from snake venoms: an inventory. For the Subcommittee on Nomenclature of Exogenous Hemostatic Factors of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis," *Thrombosis and Haemostasis*, vol. 65, no. 4, pp. 444–450, 1991.

[23] H. Pirkle and I. Theodor, "Thrombin-like venom enzymes: structure and function," *Advances in Experimental Medicine and Biology*, vol. 281, pp. 165–176, 1991.

[24] A. V. Pérez, A. Rucavado, L. Sanz, J. J. Calvete, and J. M. Gutiérrez, "Isolation and characterization of a serine proteinase with thrombin-like activity from the venom of the snake Bothrops asper," *Brazilian Journal of Medical and Biological Research*, vol. 41, no. 1, pp. 12–17, 2008.

[25] H. C. Castro, R. B. Zingali, M. G. Albuquerque, M. Pujol-Luz, and C. R. Rodrigues, "Snake venom thrombin-like enzymes: from reptilase to now," *Cellular and Molecular Life Sciences*, vol. 61, no. 7-8, pp. 843–856, 2004.

[26] I. Panfoli, D. Calzia, S. Ravera, and A. Morelli, "Inhibition of hemorragic snake venom components: old and new approaches," *Toxins*, vol. 2, pp. 417–427, 2010.

[27] Y.-M. Wang, S.-R. Wang, and I.-H. Tsai, "Serine protease isoforms of *Deinagkistrodon acutus* venom: cloning, sequencing and phylogenetic analysis," *Biochemical Journal*, vol. 354, no. 1, pp. 161–168, 2001.

[28] R. A. Hutton and D. A. Warrell, "Action of snake venom components on the haemostatic system," *Blood Reviews*, vol. 7, no. 3, pp. 176–189, 1993.

[29] J. B. Bjarnason, A. Barish, and G. S. Direnzo, "Kallikrein-like enzymes from Crotalus atrox venom," *Journal of Biological Chemistry*, vol. 258, no. 20, pp. 12566–12573, 1983.

[30] Y. Komori, T. Nikai, and H. Sugihara, "Comparison of the lethal components in Vipera aspis aspis and Vipera aspis zinnikeri venom," *Journal of Natural Toxins*, vol. 7, no. 2, pp. 101–108, 1998.

[31] L. F. Felicori, C. T. Souza, D. T. Velarde et al., "Kallikrein-like proteinase from *bushmaster* snake venom," *Protein Expression and Purification*, vol. 30, no. 1, pp. 32–42, 2003.

[32] T. Matsui, Y. Sakurai, Y. Fujimura et al., "Purification and amino acid sequence of halystase from snake venom of Agkistrodon halys blomhoffii, a serine protease that cleaves specifically fibrinogen and kininogen," *European Journal of Biochemistry*, vol. 252, no. 3, pp. 569–575, 1998.

[33] S. Iwanaga, G. Oshima, and T. Suzuki, "Proteinases from the venom of Agkistrodon halys blomhoffi," *Methods in Enzymology*, vol. 45, pp. 459–468, 1976.

[34] C.-C. Hung and S.-H. Chiou, "Fibrinogenolytic proteases isolated from the snake venom of Taiwan Habu: serine proteases with kallikrein-like and angiotensin-degrading activities," *Biochemical and Biophysical Research Communications*, vol. 281, no. 4, pp. 1012–1018, 2001.

[35] M. A. A. Parry, U. Jacob, R. Huber, A. Wisner, C. Bon, and W. Bode, "The crystal structure of the novel snake venom plasminogen activator TSV-PA: a prototype structure for snake venom serine proteinases," *Structure*, vol. 6, no. 9, pp. 1195–1206, 1998.

[36] Y. Zhang, A. Wisner, Y. Xiong, and C. Bon, "A novel plasminogen activator from snake venom: purification, characterization, and molecular cloning," *Journal of Biological Chemistry*, vol. 270, no. 17, pp. 10246–10255, 1995.

[37] Y. Zhang, A. Wisner, R. C. Maroun, V. Choumet, Y. Xiong, and C. Bon, "*Trimeresurus stejnegeri* snake venom plasminogen activator: site-directed mutagenesis and molecular modeling," *Journal of Biological Chemistry*, vol. 272, no. 33, pp. 20531–20537, 1997.

[38] S. M. T. Serrano, R. Mentele, C. A. M. Sampaio, and E. Fink, "Purification, characterization, and amino acid sequence of a serine proteinase, PA-BJ, with platelet-aggregating activity from the venom of Bothrops jararaca," *Biochemistry*, vol. 34, no. 21, pp. 7186–7193, 1995.

[39] F. Tokunaga, F. Nagasawa, S. Tamura, T. Miyata, S. Iwanaga, and W. Kisiel, "The factor V-activating enzyme (RVV-V) from Russell's viper venom. Identification of isoproteins RVV-V(α), -Vβ, and -Vγ and their complete amino acid sequences," *Journal of Biological Chemistry*, vol. 263, no. 33, pp. 17471–17481, 1988.

[40] J. J. Calvete, C. Marcinkiewicz, D. Monleón et al., "Snake venom disintegrins: evolution of structure and function," *Toxicon*, vol. 45, no. 8, pp. 1063–1074, 2005.

[41] L. Sanz, A. Bazaa, N. Marrakchi et al., "Molecular cloning of disintegrins from *Cerastes vipera* and *Macrovipera lebetina transmediterranea* venom gland cDNA libraries: insight into the evolution of the snake venom integrin-inhibition system," *Biochemical Journal*, vol. 395, no. 2, pp. 385–392, 2006.

[42] D. Israel, "A PCR-based method for high stringency screening of DNA libraries," *Nucleic Acids Research*, vol. 21, no. 11, pp. 2627–2631, 1993.

[43] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[44] E. Siigur, A. Aaspõllu, and J. Siigur, "Sequence diversity of Vipera lebetina snake venom gland serine proteinase homologs—result of alternative-splicing or genome alteration," *Gene*, vol. 263, no. 1-2, pp. 199–203, 2001.

[45] I. M.B. Francischetti, V. My-Pham, J. Harrison, M. K. Garfield, and J. M.C. Ribeiro, "Bitis gabonica (Gaboon viper) snake venom gland: toward a catalog for the full-length transcripts (cDNA) and proteins," *Gene*, vol. 337, supplement, pp. 55–69, 2004.

[46] N. S. Liang, T. T. Liu, H. P. He, Y. A. Xie, Z. Q. Meng, and A. M. Liang, "Cloning and sequence analysis of the new cDNA of thrombin-like enzyme from *Deinagkistrodon acutus*," *Guangxi Yi Ke Da Xue Xue Bao*, vol. 19, no. 1, pp. 27–30, 2002.

[47] W. H. Lee and Y. Zhang, "Molecular cloning and sequence comparison of serine proteases from the venom of *Trimeresurus stejnegeri*," Unpublished, Direct Submission, BLAST Search, 2002.

[48] M. Deshimaru, T. Ogawa, K.-I. Nakashima et al., "Accelerated evolution of crotalinae snake venom gland serine proteases," *FEBS Letters*, vol. 397, no. 1, pp. 83–88, 1996.

[49] N. Murayama, "Thrombin-like snake venom serine protease," Direct Submission, Unpublished, BLAST Search, 1999.

[50] S. Kashima, P. G. Roberto, A. M. Soares et al., "Analysis of Bothrops jararacussu venomous gland transcriptome focusing on structural and functional aspects: I-gene expression profile of highly expressed phospholipases A_2," *Biochimie*, vol. 86, no. 3, pp. 211–219, 2004.

[51] S. M. T. Serrano, Y. Hagiwara, N. Murayama et al., "Purification and characterization of a kinin-releasing and fibrinogen-clotting serine proteinase (KN-BJ) from the venom of Bothrops jararaca, and molecular cloning and sequence analysis of

its cDNA," *European Journal of Biochemistry*, vol. 251, no. 3, pp. 845–853, 1998.

[52] Q.-M. Lu, Y. Jin, D.-S. Li, W.-Y. Wang, and Y.-L. Xiong, "Characterization of a thrombin-like enzyme from the venom of *Trimeresurus jerdonii*," *Toxicon*, vol. 38, no. 9, pp. 1225–1236, 2000.

[53] Y. Zhao, K. Fang, and K. Sun, "cDNA for thrombin-like serine protease from venom gland of Agkistrodon ussuriensis," Unpublished, Direct Submission, BLAST Search, 2001.

[54] I. H. Tsai, J. C. Hsu, and Y. M. Wang, "Catroxase I and II, the serine proteases of Crotalus Atrox venom: cloning, complete sequencing and functional characterization," Unpublished, Direct Submission, BLAST Search, 2002.

[55] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[56] B. A. Jameson and H. Wolf, "The antigenic index: a novel algorithm for predicting antigenic determinants," *Computer Applications in the Biosciences*, vol. 4, no. 1, pp. 181–186, 1988.

[57] S. Braud, M. A. Parry, R. Maroun, C. Bon, and A. Wisner, "The contribution of residues 192 and 193 to the specificity of snake venom serine proteinases," *The Journal of Biological Chemistry*, vol. 275, pp. 1823–1828, 2000.

[58] E. R. Cuinto, S. Caccia, T. Rose, K. Futterer, G. Waksman, and E. Di Cera, "Unexpected crucial rule of residue 225 in serine proteases," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 1852–1857, 1999.

[59] S. S. Hasson, R. D. G. Theakston, and R. A. Harrison, "Cloning of a prothrombin activator-like metalloproteinase from the West African saw-scaled viper, *Echis ocellatus*," *Toxicon*, vol. 42, no. 6, pp. 629–634, 2003.

[60] R. A. Harrison, J. Oliver, S. S. Hasson, K. Bharati, and R. D. G. Theakston, "Novel sequences encoding venom C-type lectins are conserved in phylogenetically and geographically distinct *Echis* and *Bitis* viper species," *Gene*, vol. 315, no. 1-2, pp. 95–102, 2003.

[61] A. Wisner, S. Braud, and C. Bon, "Snake venom proteinases as tools in hemostasis studies: structure-function relationship of a plasminogen activator purified from *Trimeresurus stejnegeri* venom," *Haemostasis*, vol. 31, no. 3–6, pp. 133–140, 2001.

[62] D. C. I. Koh, A. Armugam, and K. Jeyaseelan, "Snake venom components and their applications in biomedicine," *Cellular and Molecular Life Sciences*, vol. 63, no. 24, pp. 3030–3041, 2006.

[63] E. E. Gardiner and R. K. Andrews, "The cut of the clot(h): snake venom fibrinogenases as therapeutic agents," *Journal of Thrombosis and Haemostasis*, vol. 6, no. 8, pp. 1360–1362, 2008.

[64] N. Itoh, N. Tanaka, S. Mihashi, and I. Yamashina, "Molecular cloning and sequence analysis of cDNA for batroxobin, a thrombin-like snake venom enzyme," *Journal of Biological Chemistry*, vol. 262, no. 7, pp. 3132–3135, 1987.

[65] G. H. Swift, J. C. Dagorn, P. L. Ashley, S. W. Cummings, and R. J. MacDonald, "Rat pancreatic kallikrein mRNA: nucleotide sequence and amino acid sequence of the encoded preproenzyme," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 23, pp. 7263–7267, 1982.

[66] R. J. MacDonald, S. J. Stary, and G. H. Swift, "Two similar but nonallelic rat pancreatic trypsinogens. Nucleotide sequences of the cloned cDNAs," *Journal of Biological Chemistry*, vol. 257, no. 16, pp. 9724–9732, 1982.

[67] R. J. MacDonald, G. H. Swift, C. Quinto et al., "Primary structure of two distinct rat pancreatic preproelastases determined by sequence analysis of the complete cloned messenger ribonucleic acid sequences," *Biochemistry*, vol. 21, no. 6, pp. 1453–1463, 1982.

[68] S. Brenner, "The molecular evolution of genes and proteins: a tale of two serines," *Nature*, vol. 334, no. 6182, pp. 528–529, 1988.

[69] S. M. T. Serrano and R. C. Maroun, "Snake venom serine proteinases: sequence homology vs. substrate specificity, a paradox to be solved," *Toxicon*, vol. 45, no. 8, pp. 1115–1132, 2005.

[70] T. Nikai, A. Ohara, Y. Komori, J. W. Fox, and H. Sugihara, "Primary structure of a coagulant enzyme, bilineobin, from Agkistrodon bilineatus venom," *Archives of Biochemistry and Biophysics*, vol. 318, no. 1, pp. 89–96, 1995.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 202918, 9 pages
doi:10.1155/2010/202918

Research Article

Features of Recent Codon Evolution: A Comparative Polymorphism-Fixation Study

Zhongming Zhao[1, 2, 3] and Cizhong Jiang[4]

[1] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[2] *Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[3] *Bioinformatics Resource Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN 37203, USA*
[4] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA*

Correspondence should be addressed to Zhongming Zhao, zhongming.zhao@vanderbilt.edu

Received 14 March 2010; Accepted 31 March 2010

Academic Editor: Momiao Xiong

Features of amino-acid and codon changes can provide us important insights on protein evolution. So far, investigators have often examined mutation patterns at either interspecies fixed substitution or intraspecies nucleotide polymorphism level, but not both. Here, we performed a unique analysis of a combined set of intra-species polymorphisms and inter-species substitutions in human codons. Strong difference in mutational pattern was found at codon positions 1, 2, and 3 between the polymorphism and fixation data. Fixation had strong bias towards increasing the rarest codons but decreasing the most frequently used codons, suggesting that codon equilibrium has not been reached yet. We detected strong CpG effect on CG-containing codons and subsequent suppression by fixation. Finally, we detected the signature of purifying selection against A|U dinucleotides at synonymous dicodon boundaries. Overall, fixation process could effectively and quickly correct the volatile changes introduced by polymorphisms so that codon changes could be gradual and directional and that codon composition could be kept relatively stable during evolution.

1. Introduction

Proteins are typically constructed from 20 different amino acids that are encoded by triplets of adjacent nucleotides, namely, codons. Investigation of the features and trends of amino-acid or codon changes among species can provide us important insights into protein or genome evolution. The order of introducing amino acids into the genetic code was recently inferred based on the change in the frequency of amino acids between the last universal ancestor of all extant species and today [1]. More recently, Trifonov [2] reconstructed the chronological order of entering amino acids and codons into the genetic code by applying 60 different criteria. Both studies revealed that the early amino acids were among those synthesized in imitation experiments [3], which suggested their abundance in the prebiotic environment, but the late amino acids were nonexistent or rare in the prebiotic environment.

The reconstructed amino acid chronology provided a useful index for further studying the evolutionary trend of amino acid changes. Brooks and Fresco [4] found that the frequencies of Cys, Try, and Phe have increased since the last universal ancestor. Jordan et al. [5] systematically compared sets of orthologous proteins in 15 taxa and revealed a universal trend of amino acid gain and loss, that is, amino acids under declining are those entered the genetic code earliest while those under increasing are generally late coming. This universal trend was later confirmed by simulations; however, the underlining mechanism has been under debate [6, 7].

So far, most studies have relied on the comparison of amino acid frequencies among different genomes. Such comparison may not have sufficient resolution for revealing mutational mechanisms because many genetic factors such as insertions, deletions, and duplications might have had dramatic changes of sequences, and recurrent and back

mutations might occur at the same site in the course of evolution. Single nucleotide polymorphisms (SNPs) and their recently fixed substitutions provide us an alternative and unique model for reliably estimating the mutational trend in protein sequences. A systematic comparison of inter- and intraspecific mutational changes observed at codons rather than amino acids should have a fine resolution of tracing the recent trend of mutations in protein sequences. Such a systematic comparison at the codon level has rarely been performed.

For this purpose, we performed a unique comparison of the mutation patterns at the polymorphic and fixed sites in a lineage in order to identify features and trends of mutations that led to codon gain and loss. Our comparative analysis indicated a prominent strand bias in the complementary transition pairs in amino-acid coding sequences. Both polymorphism and fixation had strong bias toward G/C → A/T relative to A/T → G/C changes, leading to a decrease of GC content in coding regions in genome evolution. We examined the correlation between the frequency change of codons and their attributes such as physiochemical properties, chronology, and codon usage. Several interesting features were observed, implying for the trend of codon gain and loss in the recent evolutionary history. Our results indicated that spontaneous mutations have volatile effects on shaping both the codon and amino acid pool but the fixation process could effectively keep composition relatively stable and minimize the structural disruption of the encoded proteins.

2. Materials and Methods

2.1. SNP Data. The human SNPs were downloaded from the NCBI dbSNP database (ftp://ftp.ncbi.nih.gov/snp/, dbSNP build 126). We extracted those SNPs that were biallelic, validated, and uniquely mapped in the human genome and excluded those SNPs in the repetitive sequences using the SNP process pipeline in our previous study [8]. We retrieved the gene annotation information from the ENSEMBL database (ftp://ftp.ensembl.org/pub/, version 32.35e). Among the SNPs selected above, we found 18,368 mapped in the exonic regions.

An exonic site may be annotated in more than one transcript because of alternative splicing or multiple genes at the same location; therefore, we removed such a site when it had more than one position in codon or it was encoded in different strands. We separated the retained exonic SNPs into three groups based on their positions (position 1, 2, or 3) in the corresponding codons. Each site was counted once to calculate GC content. The GC content at position 1, 2, or 3 was denoted by GC_1, GC_2, and GC_3, respectively. When the alleles and flanking sequence of a SNP had different orientation from the corresponding cDNA sequence, the alleles and flanking sequence were reversely complemented. These SNPs were used to infer the mutation direction at the polymorphic sites.

2.2. Inference of Mutation Direction at the Polymorphic Sites. We used chimpanzee as the outgroup to infer the ancestral allele of each human SNP. The chimpanzee genome was downloaded from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/, build 1, version 1). We ran MegaBLAST [9] to map human SNP sequences to the chimpanzee genome and then inferred the ancestral allele of each SNP using the stringent criteria as in our recent study [8]. The mutation direction was inferred by the maximum parsimony principle. For example, if a SNP A/G matches nucleotide "A" in chimpanzee sequence, the mutation direction would be A → G.

2.3. Substitutions at the Fixed Sites. For consistency, we used the alignments of 7552 triplets of human-chimpanzee-mouse orthologous genes (coding regions only) prepared in Jordan et al. [5]. To estimate the nucleotide substitutions for the human lineage, we identified the sites where the mouse and the chimpanzee carried the same nucleotide (e.g., A) but the human carried a different nucleotide (e.g., G). Because the rate of nucleotide change in mammalian genomes is low ($\sim 10^{-10}$) per site per year [10]), according to the maximum parsimony principle, we could infer the direction of nucleotide substitution (A → G in this example). We excluded the sites where any insertion or deletion occurred in a lineage. Similar to the polymorphisms, these sites were separated into three groups according to their positions (position 1, 2, or 3) in the corresponding codons. The GC content at position 1, 2, or 3, denoted by GC_1, GC_2, and GC_3, was calculated correspondingly.

2.4. Nucleotide and Codon Changes. The frequency of each nucleotide change was initially calculated by

$$f_{i->j} = \frac{n_{i->j}}{\sum_i \sum_{j \neq i} n_{i->j}} \times 100\%, \quad (1)$$

where $n_{i->j}$ is the counts of nucleotide changes from the ith type to the jth type (i, j = A, C, G or T). Because of unequal nucleotide composition in sequences, the frequency of each nucleotide change was then normalized by

$$f_{i->j} = \frac{n_{i->j}/N_i}{\sum_i \sum_{j \neq i} \left(n_{i->j}/N_i\right)} \times 100\%, \quad (2)$$

where N_i is the total counts of nucleotide i in the sequences.

The normalized difference (gain and loss) of a codon was defined as the difference of the number of mutations removing the codon from that creating the codon and divided by the total number of mutations in the codon. The equation is

$$\text{Codon change} = \frac{n_+ - n_-}{n_+ + n_-}, \quad (3)$$

where $n_+(n_-)$ denotes the mutations creating (removing) the codon. The gain or loss of an amino acid was calculated similarly.

2.5. Codon Usage Data. Frequencies of codon usage in the human were obtained from the Codon Usage Database

[11], which contained 38,691,091 codons from human genes (build: NCBI-GenBank Flat File Release 156.0, http://www.kazusa.or.jp/codon/).

3. Results and Discussion

3.1. Mutational Features at the Polymorphic and Fixed Sites in Coding Regions. After removing uncertain or low quality SNPs, we obtained a total of 11,253 exonic SNPs that were biallelic and uniquely mapped in the human genome and whose ancestral alleles could be inferred from their alignment with chimpanzee genomic sequences. Separately, we identified 14,007 substitutions that occurred in the human lineage based on the alignments of human-chimpanzee-mouse orthologous gene sequences. The direction of nucleotide changes was used to examine the mutational features at the polymorphic and fixed sites as well as gain and loss of amino acids and their codons.

We first examined the polymorphisms at each position (1, 2, or 3) of a codon. We observed more than half of the SNPs located at position 3 (2803 at position 1, 2466 at position 2, and 5984 at position 3). The frequencies of transitions (changes between a purine and another purine or between a pyrimidine and another pyrimidine) were much higher than those of transversions (changes between a purine and a pyrimidine) (Figure 1). Specifically, the frequencies of polymorphisms G → A (C → T) were higher than those of A → G (T → C) at each codon position. Importantly, the frequency was notably different between A → G (e.g., 15.9% at position 1) and T → C (e.g., 11.8% at position 1) at each position; such difference could not be detected in the whole genome or in noncoding regions in our previous study [8]. This feature might reflect the strand bias in the coding regions. Indeed, we observed 26.8% of A and 17.3% of T at position 1 and similar composition bias at other two positions in the coding regions (see Supplemental Table 1 in supplementary material available online at doi:10.1155/2010/202918). The strand bias might be resulted from different functional constraints. For example, the two DNA strands of a transcribed gene are under different selection pressure due to transcriptional process.

We obtained a total of 14,007 fixed sites (2765 at position 1, 2199 at position 2, and 9043 at position 3) based on the alignments of triplets of human-chimpanzee-mouse orthologous gene sequences. Similar to the polymorphism data, the frequencies of transitions were much higher than transversions (Figure 1), and the frequencies were not symmetric for the complementary substitution pairs (e.g., G → A versus A → G).

Opposite to the similar frequency of each transversion type, we observed strong difference in the frequency of each transition type at each position, especially at position 3, when compared polymorphism and fixation data (Figure 1). At positions 1 and 2, mutation G → A occurred most frequently based at the polymorphic and fixed sites. At position 3, mutation C → T occurred most frequently at the polymorphic sties whereas T → C occurred most frequently at the fixed sties.

There was a significant excess of G/C → A/T mutations at both polymorphic and fixed sites when compared to A/T → G/C mutations (Table 1). Here, G/C → A/T denotes changes of G or C to A or T. For example, the number of G/C → A/T and A/T → G/C changes was 1443 and 991 at position 1 at the fixed sites, respectively. The difference was significant ($P = 5.1 \times 10^{-20}$) assuming that they had the same chance in a random mutation model. This observation suggested a declining GC content in coding regions, which has higher GC content than noncoding regions. These results were consistent with our previous finding of more G/C → A/T than A/T → G/C mutations at polymorphic sites in each categorized human genomic region [8]. However, another study reported a weak fixation bias favoring mutations that increase GC content in noncoding regions despite of no significant difference at the fixed sites between G/C → A/T and A/T → G/C changes [12]. Of note, the difference between the frequencies of G/C → A/T and A/T → G/C changes was much smaller at the fixed sites than the polymorphic sites (Table 1). For example, for all mutation data in codons, the frequencies of G/C → A/T and A/T → G/C changes were 47.8% and 42.8% at the fixed sites, compared to 57.6% and 30.2% at the polymorphic sites. This indicated that fixation allows less variation in GC content than polymorphism in the course of genome evolution.

3.2. Trends of Amino Acid Gain and Loss. We next examined the trend of mutation at the amino acid level. We measured gain and loss of an amino acid by the difference of the number of mutations removing the amino acid from creating the amino acid divided by the total number of mutations in the amino acid (see Materials and Methods). Supplementary Figure 1 displays the gain and loss trend for the 20 amino acids in the human lineage. There was a great variation towards gain or loss among amino acids; however, the extent of variation by the polymorphisms was remarkably stronger than that by the fixed substitutions. For example, the normalized difference in Gly was −0.13 by the polymorphisms, compared to −0.016 by the fixed substitutions. Here, a minus value indicates the trend of loss in the recent history. This suggests that many new neutral mutations would have been eliminated by genetic factors such as genetic drift and natural selection.

We examined the trend of gain and loss of amino acids by their chronology [2]. Overall, the early-coming amino acids (Gly, Ala, and Asp) tended to be lost; an opposite trend (i.e., gain), though much weaker, was observed for the late coming ones (e.g., Trp, Phe, Cys, and His) (Supplementary Figure 1). Interestingly, polymorphisms resulted in an accumulation of almost all the latest 10 amino acids except Tyr which had a small loss. This implies that the ancient amino acids have been under loss whereas the late coming amino acids have been under accumulation. This observation supports the previous studies using substitution data among more evolutionarily distant species [4, 5]. Of note, Met, a late amino acid, tended to lose by fixation. This may reflect the functional constraint on this amino acid because it is

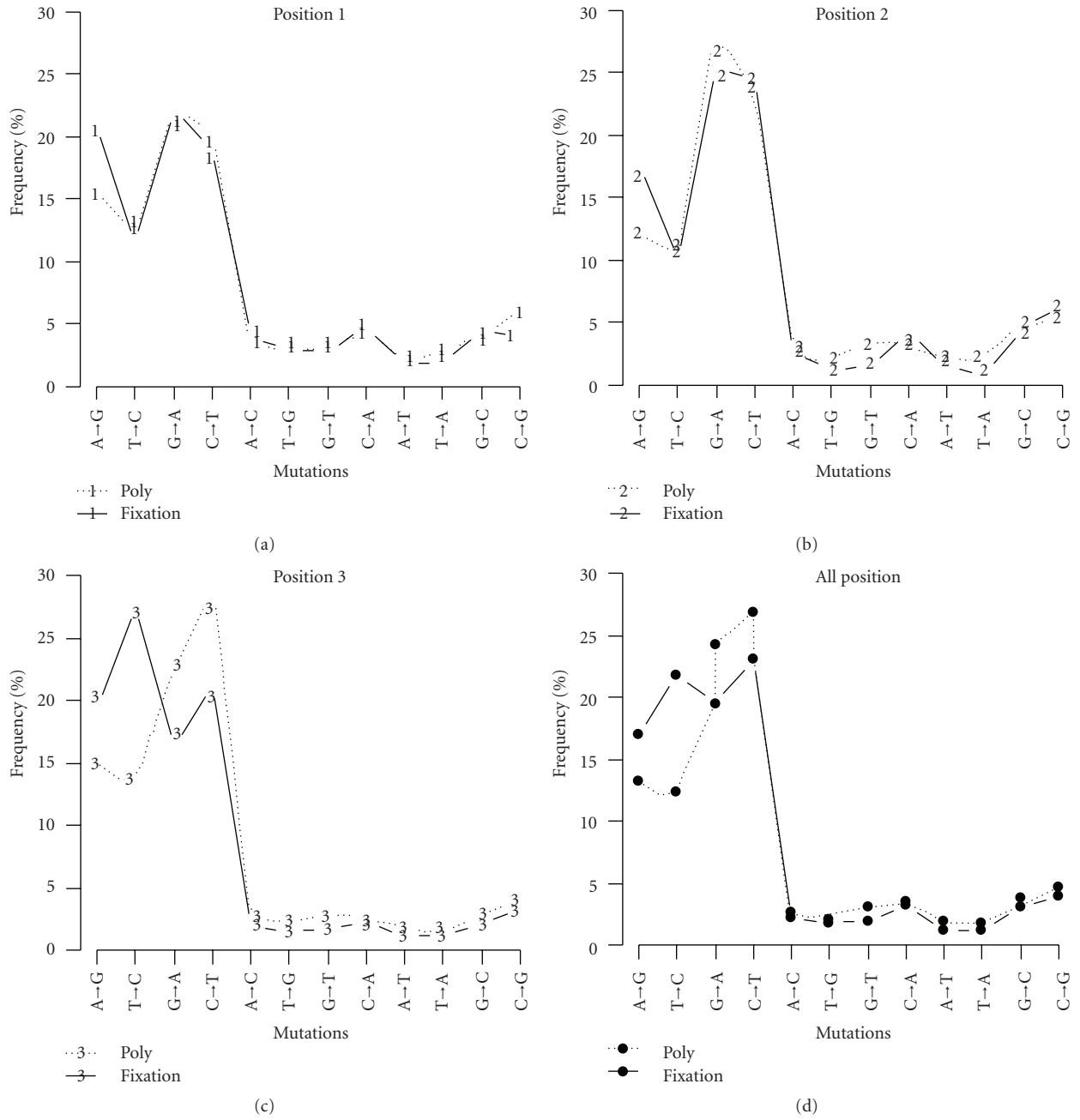


Figure 1: Normalized frequencies of nucleotide changes at the polymorphic and fixed sites in amino-acid coding regions. Figures 1(a)–1(c) show the frequencies of polymorphism (Poly) and fixation mutations at the first (a), second (b), and third (c) positions of a codon. (d) Frequency of all the mutation data. Dotted lines indicate the polymorphism and solid lines indicate the fixation.

Table 1: Comparison of mutations $G/C \rightarrow A/T$ and $A/T \rightarrow G/C$ at the polymorphic and fixed sites in coding regions.

	Position 1		Position 2		Position 3		Total	
	Poly (%)	Fixed (%)	Poly (%)	Fixed (%)	Poly (%)	Fixed (%)	Poly (%)	Fixed (%)
G/C → A/T	1547 (50.9)	1443 (48.7)	1324 (59.2)	1051 (55.4)	3791 (57.5)	4369 (41.7)	6662 (57.6)	6863 (47.8)
A/T → G/C	831 (33.7)	991 (39.3)	793 (26.6)	878 (31.4)	1593 (32.6)	3937 (50.6)	3217 (30.2)	5806 (42.8)
P-value	3.7×10^{-5}		2.0×10^{-7}		1.3×10^{-95}		1.8×10^{-90}	

The frequency of each mutation type was included in the parentheses. P-values were calculated by χ^2 test for 2×2 contingency tables at each position or for the total data. Poly: polymorphisms. Fixed: fixed substitutions.

the first amino acid in most eukaryotic proteins. Further investigation is warranted.

We performed linear regression analysis of amino acid changes with their physiochemical characters including charge, polarity, and polarity and volume using Zhang's classification [13]. No significant correlation was found, suggesting that the physiochemical attributes may not have a major effect on governing the recent amino acid composition changes.

3.3. Codon Gain or Loss with Chronology. Because most amino acids are encoded by more than one codon, a detailed examination of nucleotide substitution at the codon level should provide more insights on amino-acid compositional changes in protein evolution. We first examined the gain and loss of codons for the amino acids ordered by amino-acid chronology. There was no clear trend of loss of codons for early-coming amino acids or gain of codons for late coming amino acids; however, almost all amino acids that have multiple codons (synonymous codons) had a balancing effect by gaining some codons while losing the others (Supplemental Figure 2). For example, Gly has four codons. Both the polymorphisms and fixation resulted in two of them (GGC and GGG, early codons) to lose but the other two (GGA and GGT, late codons) to gain. This feature seems to be attributed to the different chronological orders of amino acids and codons entering the genetic code. Moreover, late coming amino acids might grab codon(s) encoding other amino acids through codon capture, which has been commonly proposed as a mechanism for adding amino acids into the code [14, 15]. For example, it was reported that AUA (Ile) also encodes methionine (Met) in yeast [16] and that AAA encodes asparagine (Asn) rather than lysine (Lys) in some animal mitochondria [17].

Therefore, we re-examined the gain and loss of codons in codon's temporal order constructed by Trifonov [2]. Remarkably, we observed an overall trend that the late coming codons were gainers whereas the early-coming codons were losers, with only a few exceptions (e.g., AAG and GTC) (Figure 2). Compared to the fixation, polymorphisms had a stronger trend. All the latest 10 codons were increasing while 10 of the earliest 11 codons were decreasing. This observation implies that, in the course of evolution, late coming codons were continuously added into the genetic code and encoded amino acids in the primitive proteins; meanwhile, the usage of early codons in proteins gradually declined. Finally, we did not find significant correlation between the codon changes and the physiochemical characters (charge, polarity, and polarity and volume) of amino acids encoded by the codons.

3.4. Codon Gain or Loss with Codon Usage. We further examined the relationship between codon usage and the trend of codon changes by the polymorphisms and fixation. We obtained the frequencies of codon usage in the human genome from the Codon Usage Database [11]. We defined a codon being rare when its frequency was ≤13 per 1,000 codons in the genome, as previously suggested in [18, 19]. Figure 3 shows the changes of codons ordered by the codon usage frequency. Interestingly, fixation had strong, but opposite, effect on rarely and frequently used codons (Figure 3). For rare codons, fixation increased 18 while decreased 6 codons. In contrast, for nonrare codons, fixation increased 13 but decreased 24 codons. The difference was statistically significant (2×2 contingency table, $P = .005$). Strikingly, fixation resulted in a strong increase of the frequencies (i.e., gain) of all the 10 rarest codons (Figure 3). The extent of increasing these codons by fixation was stronger than other codons. Conversely, for the 16 most frequent codons, we observed 12 under loss by fixation.

Polymorphisms did not show such a contrast pattern as observed in fixation. However, there was an interesting feature. There is a total of eight codons that contain CG dinucleotides; they could be termed NCG and CGN where "N" denotes any nucleotide (A, C, G, or T). When we examined the nucleotide attributes of rare codons, we found that all these CG containing codons belonged to rare codons. Interestingly, polymorphisms tended to decrease all these codons except CGT, which had a small increase (Figure 3). It is well known that the mutation rate of CG → TG/CA is much higher because of the hypermutability of methylated CpGs, that is, CpG effect [20, 21]. Our analysis of mutation direction using SNP data indicated much higher frequencies of CG → TG/CA than TG/CA → CG changes in the NCG and CGN codons (1,600 versus 560, 2.86 folds, Table 2), which confirmed the strong CpG effect on these codons. A more detailed examination revealed that this difference was largely attributed to the asymmetrical synonymous transition between NCG and NCA. For example, for CCG codon, we observed 206 synonymous transitions CCG → CCA compared to 80 CCA → CCG. Consequently, the frequencies of the NCG and CGN codons tended to decrease by polymorphisms. The only exception is CGT, a codon gainer. There were only 6 CGT → CGC compared to 53 CGC → CGT changes. This unique asymmetry of synonymous transition at the third position of CG containing codons resulted in an overall accumulation.

We further compared CG → TG/CA versus TG/CA → CG changes in the eight CG containing codons using the polymorphism and fixation data. In contrast to the strong CpG effect in the spontaneous nucleotide polymorphisms, we detected a strong suppression of CpG effect on NCG codons as well as a moderate suppression on CGN codons by fixation. For example, in polymorphisms, the ratios of CG → TG/CA over TG/CA → CG changes were at least 2; however, in fixation, most of CpG containing codons, especially the NCG codons, had the ratio not greater than 1 (Table 2). When we combined all eight codons, the ratio was 2.86 by polymorphisms but only 0.50 by fixation. Based on these results, we proposed that fixation process could fix the potential rapid loss of CpG containing codons caused by the CpG effect, therefore, avoid dramatic change of codons during short evolutionary period. It is worth noting that a similar suppression of polymorphisms at the CpG dinucleotides in CpG islands, which are often found in the promoter regions, was found as a mechanism to maintain high CpGs and GC content in the promoter regions [22, 23]. Purifying selection was suggested to play a role in

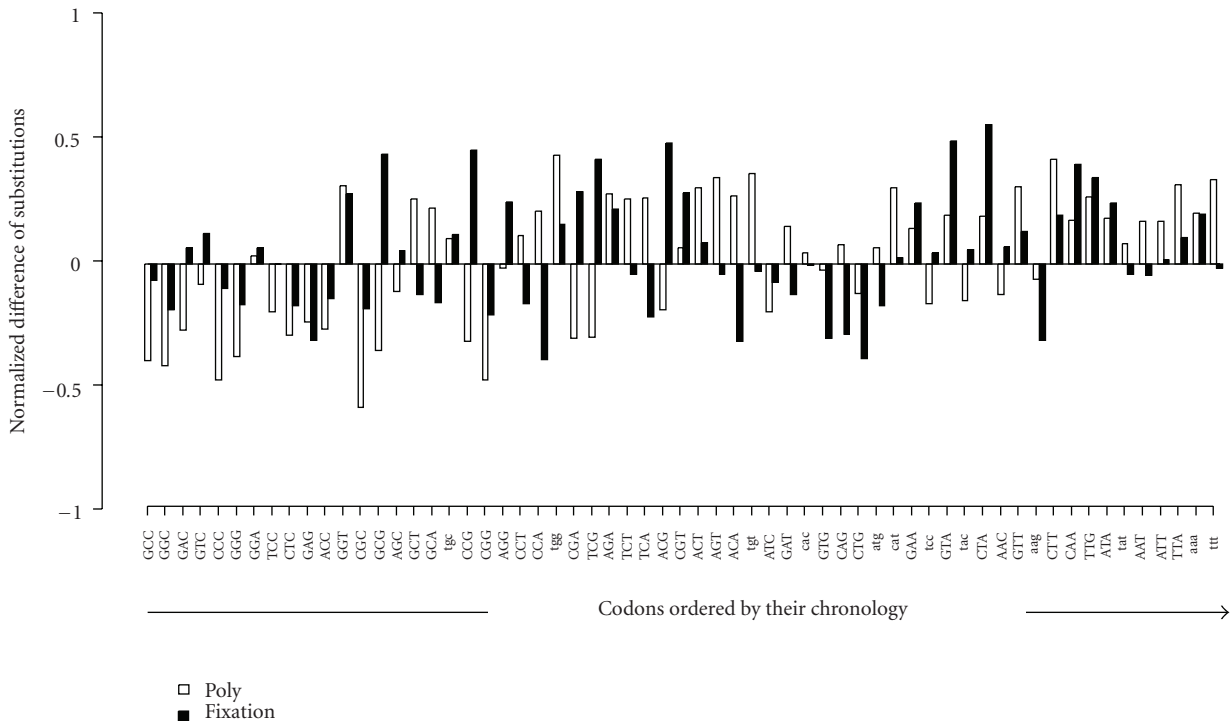


Figure 2: Normalized difference of codon changes (gain and loss) at the polymorphic and fixed sites in amino-acid coding regions. The normalized difference of a codon was defined as the difference of the number of mutations removing the codon from that creating the codon and divided by the total number of mutations in the codon. Codons are ordered (left: ancient; right: recent) in their temporal order in the genetic code according to Trifonov [2]. The codons associated with codon capture are in lower case.

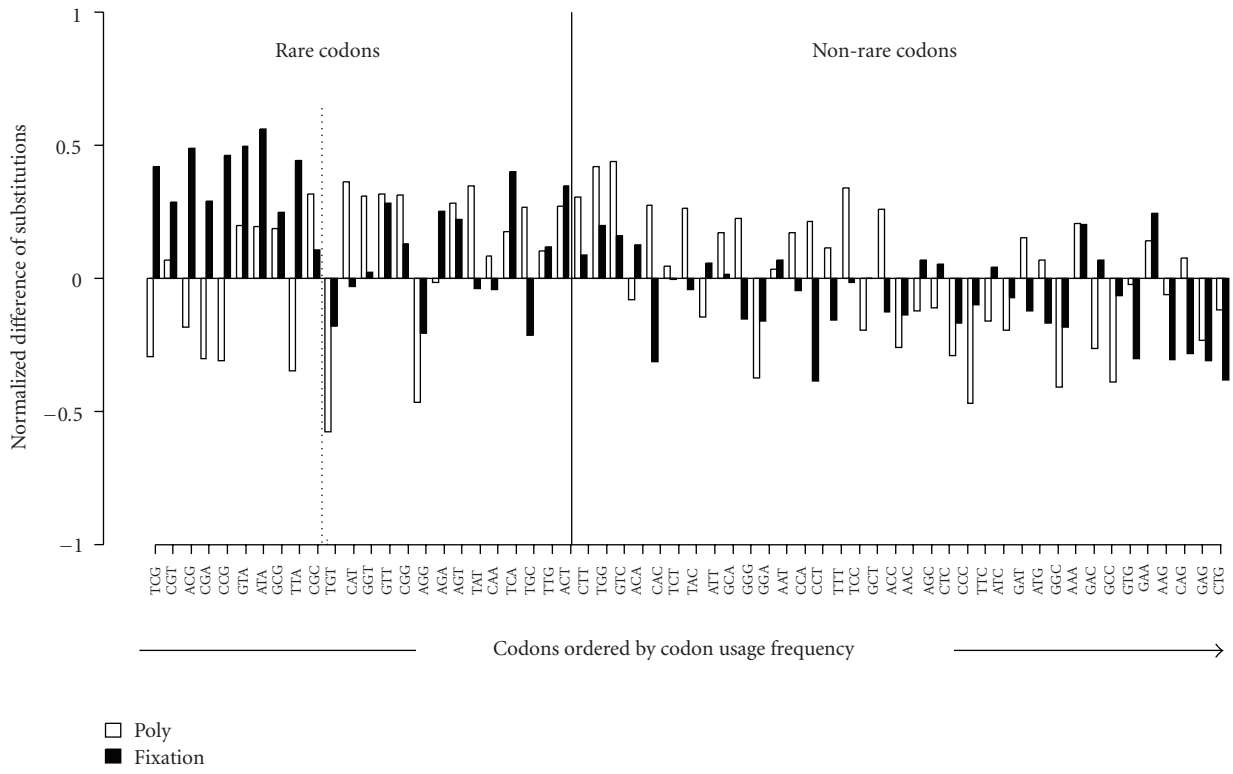


Figure 3: Normalized difference of codon changes (gain and loss) at polymorphic and fixes sites. Codons are ordered in the ascending order of frequencies in human codon usage based on the Codon Usage Database. The solid vertical line separates rare and nonrare codons. The 10 rarest codons are shown in the left side of the dotted line.

Table 2: Statistics of mutations *CG* ↔ *TG/CA* in eight CG-containing codons.

Codon	Polymorphism			Fixation		
	CG → TG/CA	TG/CA → CG	Ratio[a]	CG → TG/CA	TG/CA → CG	Ratio[a]
ACG	282	141	2.00	89	245	0.36
TCG	193	61	3.16	52	121	0.43
CCG	282	115	2.45	87	225	0.39
GCG	271	81	3.35	75	180	0.42
CGA	88	12	7.33	17	14	1.21
CGT	106	53	2.00	14	25	0.56
CGC	163	29	5.62	47	22	2.14
CGG	215	68	3.16	65	65	1.00
Total	1600	560	2.86	446	897	0.50

[a]Ratio of CG → TG/CA over TG/CA → CG changes.

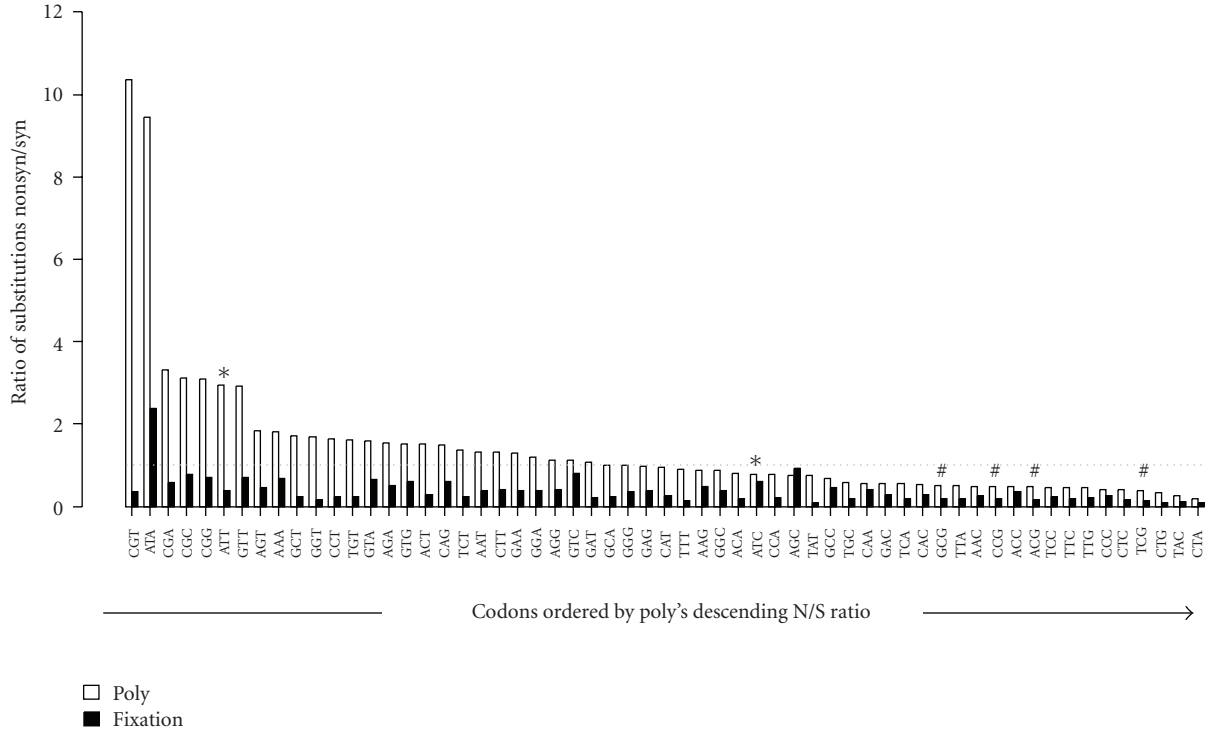


Figure 4: Nonsynonymous to synonymous substitutions (N/S) ratio calculated by the polymorphic and fixed nucleotide changes. Two codons (ATG and TGG) are not included because of their lack of synonymous substitutions. Symbol "#" indicates NCG codons. Codons are placed by the descending order of the polymorphism's N/S ratios.

the suppression of polymorphisms at the CpG dinucleotides in CpG islands [23]. Therefore, polymorphisms in the functional regions might have been under elimination due to functional constraints.

3.5. Selective Constraints on Codons. The different features of polymorphisms and fixation in codons, especially CpG containing codons, suggest that genetic factors such as selective constraints might have played important roles in codon evolution. We next calculated the ratio of nonsynonymous over synonymous substitution (N/S) for each codon, which is a typical method for detecting functional constraints on coding sequences [24]. Based on the fixation data, all codons except ATA had the N/S ratio smaller than 1, and most codons had the ratio much smaller than 1 (Figure 4). This suggests strong purifying selection on most codons during evolution. As expected, there were more synonymous mutations than nonsynonymous mutations because all codons, except ATG (coding Met), TGG (coding Trp), and ATA (coding Ile), could have transitions at their third position without changing amino acids and because the transition over transversion ratio has been observed to be approximate 2 in many vertebrate genomes [21, 25]. This type of synonymous transitional mutations was known to be "cheap" and occurred most frequently [26]. However, ATA lacks of such "cheap" mutations, which explains why the N/S ratio was greater than 1.

Table 3: Mutation pattern on $A|U$ dicodon boundaries.

	Polymorphism		Fixation	
	A|T	A|G	A|T	A|G
T/C/G → A	328	613	300	507
A → T/C/G	228	269	229	517
P-value[a]	4.49×10^{-5}		7.12×10^{-3}	

P-values were calculated by χ^2 test for 2×2 contingency tables. A|G dicodon boundaries were considered as control.
[a]Note the opposite pattern on A|U dicodon boundaries by polymorphism and fixation.

The N/S ratios using the polymorphism data were remarkably higher than the corresponding ones using the fixation data, with one exception (AGC, serine) (Figure 4). Nearly half of the codons had the N/S ratio greater than 1. Seven codons had ratio greater than 2, including four CGN codons. We found that spontaneous mutations favored transitions at the first and second positions (nonsynonymous) over transitions at the third position (mostly synonymous) in these seven codons, particularly in the CGN codons. For example, there were 106 CGT → TGT/CAT nonsynonymous changes but only 6 CGT → CGC synonymous changes. This resulted in a very high N/S ratio for codon CGT (Figure 4). However, these nonsynonymous changes were directly linked to the CpG effect, which conversely led to a high abundance of synonymous mutations at the second and third positions of NCG codons. Therefore, we observed small N/S ratios for NCG codons (Figure 4). As shown in Table 2, fixation had a much smaller ratio of CG → TG/CA over TG/CA → CG changes for NCG than for CGN codons. Taken together, CpG effect plays a dominant role in preferring mutations at the polymorphic sites. However, selective constraints during fixation process could largely correct it to balance the codon composition.

3.6. Selection Effect on Dicodon Boundary. An early study reported that AU dinucleotides was a cleavage site of RNase L, the 2′,5′-oligoadenylate-dependent ribonuclease [27]. Interestingly, a recent study extended this analysis in yeast and found that mRNAs containing A|U dinucleotides at synonymous dicodon boundaries had a short half-life due to more efficient 3′–5′ degradation by endonucleolytic cleavage [19]. If this was true in humans, we may detect the signature of purifying selection on A|U dinucleotides at synonymous dicodon boundaries. That is, mutations toward A at the third position of synonymous codon leading to A|U dicodon boundaries would not favor, but mutations from A at the third position of synonymous codon leading to non-A|U dicodon boundaries would favor. The mRNA half-life is irrelative of the use of synonymous A|G dicodon boundary [19]; therefore, we used synonymous A|G dicodon boundary as control. Our analysis of polymorphism data revealed a dramatic deficit of A|U dicodon boundary (Table 3). However, we observed an opposite effect of fixation on A|U dicodon boundary, that is, fixation would favor A|U dicodon boundaries. This opposite feature has not been reported in literature. The mechanism is unclear and needs further investigation.

4. Conclusions

Amino acid compositions and nucleotide substitutions have been extensively studied because they are fundamental in protein and genome evolution. So far, either interspecies nucleotide substitutions or intraspecies nucleotide polymorphisms, but not both, have been analyzed. In this study, we uniquely and systematically compared the inter- and intraspecies nucleotide changes in amino-acid coding sequences, especially at the codon level. Our results provided a detailed view on the spontaneous point mutations in codons and subsequent and rapid fixation in a lineage (e.g., human lineage) in recent genome evolution. We observed a trend of loss of the ancient codons while gain of the latest codons. Fixation had a strong bias towards increasing the rarest codons but decreasing the most frequent codons, suggesting that codon equilibrium has not been reached yet. Another major feature is that fixation could effectively suppress the strong CpG effect on the CG containing codons. Functional constraints such as purifying selection have likely played a major role in codon changes. Finally, we reported a unique and opposite mutation pattern on A|U dicodon boundaries. Although these findings are limited to the trends of codon gain and loss in the recent history, more specifically in the recent human history, they provide important insights on how nucleotide changes in the protein-coding regions have likely shaped the genome and protein sequence composition.

Acknowledgment

This project was partially supported by NIH Grants (LM009598 and AA017437) and the NARSAD young investigator award to Z. Zhao.

References

[1] D. J. Brooks, J. R. Fresco, A. M. Lesk, and M. Singh, "Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code," *Molecular Biology and Evolution*, vol. 19, no. 10, pp. 1645–1655, 2002.

[2] E. N. Trifonov, "The triplet code from first principles," *Journal of Biomolecular Structure and Dynamics*, vol. 22, no. 1, pp. 1–11, 2004.

[3] S. L. Miller, "A production of amino acids under possible primitive earth conditions," *Science*, vol. 117, no. 3046, pp. 528–529, 1953.

[4] D. J. Brooks and J. R. Fresco, "Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor," *Molecular & Cellular Proteomics*, vol. 1, no. 2, pp. 125–131, 2002.

[5] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, et al., "A universal trend of amino acid gain and loss in protein evolution," *Nature*, vol. 433, no. 7026, pp. 633–638, 2005.

[6] J. H. McDonald, "Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 240–244, 2006.

[7] L. D. Hurst, E. J. Feil, and E. P. C. Rocha, "Protein evolution: causes of trends in amino-acid gain and loss," *Nature*, vol. 442, no. 7105, pp. E11–E12, 2006.

[8] C. Jiang and Z. Zhao, "Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms," *Genomics*, vol. 88, no. 5, pp. 527–534, 2006.

[9] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 203–214, 2000.

[10] Z. Zhao, L. Jin, Y.-X. Fu, et al., "Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 21, pp. 11354–11358, 2000.

[11] Y. Nakamura, T. Gojobori, and T. Ikemura, "Codon usage tabulated from international DNA sequence databases: status for the year 2000," *Nucleic Acids Research*, vol. 28, no. 1, p. 292, 2000.

[12] M. T. Webster, N. G. C. Smith, and H. Ellegren, "Compositional evolution of noncoding DNA in the human and chimpanzee genomes," *Molecular Biology and Evolution*, vol. 20, no. 2, pp. 278–286, 2003.

[13] J. Zhang, "Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes," *Journal of Molecular Evolution*, vol. 50, no. 1, pp. 56–68, 2000.

[14] J. T. Wong, "A co-evolution theory of the genetic code," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, no. 5, pp. 1909–1912, 1975.

[15] F. H. C. Crick, "The origin of the genetic code," *Journal of Molecular Biology*, vol. 38, no. 3, pp. 367–379, 1968.

[16] B. G. Barrell, S. Anderson, and A. T. Bankier, "Different pattern of codon recognition by mammalian mitochondrial tRNAs," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 6, pp. 3164–3166, 1980.

[17] T. Ohama, S. Osawa, K. Watanabe, and T. H. Jukes, "Evolution of the mitochondrial genetic code IV. AAA as an asparagine codon in some animal mitochondria," *Journal of Molecular Evolution*, vol. 30, no. 4, pp. 329–332, 1990.

[18] G. Caponigro, D. Muhlrad, and R. Parker, "A small segment of the MATα 1 transcript promotes mRNA decay in Saccharomyces cerevisiae: a stimulatory role for rare codons," *Molecular and Cellular Biology*, vol. 13, no. 9, pp. 5141–5148, 1993.

[19] D. B. Carlini, "Context-dependent codon bias and messenger RNA longevity in the yeast transcriptome," *Molecular Biology and Evolution*, vol. 22, no. 6, pp. 1403–1411, 2005.

[20] B. K. Duncan and J. H. Miller, "Mutagenic deamination of cytosine residues in DNA," *Nature*, vol. 287, no. 5782, pp. 560–561, 1980.

[21] F. Zhang and Z. Zhao, "The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs," *Genomics*, vol. 84, no. 5, pp. 785–795, 2004.

[22] Z. Zhao and F. Zhang, "Sequence context analysis in the mouse genome: single nucleotide polymorphisms and CpG island sequences," *Genomics*, vol. 87, no. 1, pp. 68–74, 2006.

[23] Z. Zhao and F. Zhang, "Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome," *Gene*, vol. 366, no. 2, pp. 316–324, 2006.

[24] A. Nekrutenko, K. D. Makova, and W.-H. Li, "The K_A/ K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study," *Genome Research*, vol. 12, no. 1, pp. 198–202, 2002.

[25] Z. Zhao and E. Boerwinkle, "Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome," *Genome Research*, vol. 12, no. 11, pp. 1679–1686, 2002.

[26] A. D. Yoder, R. Vilgalys, and M. Ruvolo, "Molecular evolutionary dynamics of cytochrome b in strepsirrhine primates: the phylogenetic significance of third-position transversions," *Molecular Biology and Evolution*, vol. 13, no. 10, pp. 1339–1350, 1996.

[27] S. S. Carroll, E. Chen, T. Viscount, et al., "Cleavage of oligoribonucleotides by the 2′,5′-oligoadenylate- dependent ribonuclease L," *The Journal of Biological Chemistry*, vol. 271, no. 9, pp. 4988–4992, 1996.