Transcriptome analysis is the study of the transcriptome, the complete set of RNA transcripts that are produced under specific circumstances, using high-throughput methods. Transcription profiling, which follows total changes in the behavior of a cell, is used throughout diverse areas of biomedical research, including diagnosis of disease, biomarker discovery, risk assessment of new drugs or environmental chemicals, etc.

Transcriptome analysis is most commonly used to compare specific pairs of samples, for example, tumor tissue versus its healthy counterpart.

In this volume, Dr. Pyo Hong discusses the role of long RNA sequences in transcriptome analysis, Dr. Shinichi describes the next-generation single-cell sequencing technology developed by his team, Dr. Prasanta presents transcriptome analysis applied to rice under various environmental factors, Dr. Xiangyuan addresses the reproductive systems of flowering plants and Dr. Sadovsky compares codon usage in conifers.
Transcriptome Analysis

Edited by Miroslav Blumenberg

Published in London, United Kingdom
Supporting open minds since 2005
We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

4,400+ Open access books available
117,000+ International authors and editors
130M+ Downloads

151 Countries delivered to
Our authors are among the Top 1% most cited scientists
12.2% Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Meet the editor

Miroslav Blumenberg, PhD, was born in Subotica and received his BSc in Belgrade, Yugoslavia. He completed his PhD at MIT in Organic Chemistry; he followed up his PhD with two postdoctoral study periods at Stanford University. Since 1983, he has been a faculty member of the RO Perelman Department of Dermatology, NYU School of Medicine, where he is a codirector of a training grant in cutaneous biology. Dr. Blumenberg's research is focused on the epidermis, expression of keratin genes, transcription profiling, keratinocyte differentiation, inflammatory diseases and cancers, and most recently the effects of the microbiome on skin. He has published more than 100 peer-reviewed research articles and graduated numerous PhD and postdoctoral students. Dr. Blumenberg lives in New York, USA, with his wife and two children.
Preface

Section 1
Introduction

Chapter 1  Introductory Chapter: Transcriptome Analysis by Miroslav Blumenberg

Section 2
Tumor Transcriptome

Chapter 2  Single-Cell Transcriptome Analysis in Tumor Tissues by Sadahiro Iwabuchi and Shinichi Hashimoto

Section 3
Reference Transcriptomes

Chapter 3  Transcriptome Atlas by Long-Read RNA Sequencing: Contribution to a Reference Transcriptome by Dong Jin Lee and Chang Pyo Hong

Section 4
Transcriptome Analysis in Plants

Chapter 4  Plant Comparative Transcriptomics Reveals Functional Mechanisms and Gene Regulatory Networks Involved in Anther Development and Male Sterility by Xiangyuan Wan and Ziwen Li

Chapter 5  Transcriptome Analysis for Abiotic Stresses in Rice (Oryza sativa L.) by Ashutosh Kumar and Prasanta K. Dash

Chapter 6  Revealing the Symmetry of Conifer Transcriptomes through Triplet Statistics by Sadovsky Michael, Putintseva Yulia, Biryukov Vladislav and Senashova Maria
# Contents

**Preface**  
XIII

**Section 1**  
Introduction  
1

**Chapter 1**  
Introductory Chapter: Transcriptome Analysis  
*by Miroslav Blumenberg*  
3

**Section 2**  
Tumor Transcriptome  
9

**Chapter 2**  
Single-Cell Transcriptome Analysis in Tumor Tissues  
*by Sadahiro Iwabuchi and Shinichi Hashimoto*  
11

**Section 3**  
Reference Transcriptomes  
23

**Chapter 3**  
Transcriptome Atlas by Long-Read RNA Sequencing: Contribution to a Reference Transcriptome  
*by Dong Jin Lee and Chang Pyo Hong*  
25

**Section 4**  
Transcriptome Analysis in Plants  
37

**Chapter 4**  
Plant Comparative Transcriptomics Reveals Functional Mechanisms and Gene Regulatory Networks Involved in Anther Development and Male Sterility  
*by Xiangyuan Wan and Ziwen Li*  
39

**Chapter 5**  
Transcriptome Analysis for Abiotic Stresses in Rice (*Oryza sativa* L.)  
*by Ashutosh Kumar and Prasanta K. Dash*  
61

**Chapter 6**  
Revealing the Symmetry of Conifer Transcriptomes through Triplet Statistics  
*by Sadovsky Michael, Putintseva Yulia, Biryukov Vladislav and Senashova Maria*  
77
Transcriptome analysis is the study of the transcriptome, the complete set of RNA transcripts produced under specific circumstances, using high-throughput methods. Transcription profiling, which follows changes in the behavior of a cell, is used throughout diverse areas of biomedical research, including the diagnosis of disease, biomarker discovery, risk assessment of new drugs or environmental chemicals, etc. Transcription profiling can be applied to loss- and gain-of-function mutants to identify the changes associated with the mutant phenotype. Transcriptomics also allows the identification of pathways that respond to or ameliorate environmental stresses. RNA sequencing (RNA-Seq) detects all transcripts in a sample, including mRNAs as well as the regulatory siRNA and lncRNA transcripts. RNA-Seq can also identify disease-associated gene fusions, single nucleotide polymorphisms, and even allele-specific expression.

Transcriptome analysis is most commonly used to compare specific pairs of samples. The differences may be due to different external environmental conditions, for example, hormonal effects or toxins. More commonly, healthy and disease states are compared. In general, transcriptome analysis is a very powerful hypothesis-generating tool rather than a theory-proving one. Transcriptome analyses have become indispensable in basic research and translational and clinical studies.

In this volume, Dr. Pyo Hong discusses the role of long RNA sequences in transcriptome analysis. It should be noted that early RNA-Seq methods generated rather short reads, 35 to a few hundred nucleotides, and relied on massive redundancy to achieve required accuracy. Newer methods, which provide longer reads, have significant advantages, for example, in the analysis of previously not sequenced genomes. Such approaches need tailored software methods.

Dr. Shinichi describes next-generation single-cell sequencing technology developed by his team. It can be used for single-cell transcriptome analysis in tumor tissues. This is an extremely important area nowadays because it is clear that most tumors are heterogeneous. Identifying the transcriptome of tumor stem cells may lead to specific targeting of these cells. Alternatively, single-cell transcriptome analysis can help in defining the tumor-infiltrating immune cells, a critical component of immunotherapies. Dr. Shinichi and his team developed a microwell device that can be easily transported and is relatively cheaper than most other RNA-Seq methods, which will be essential for the widespread use of transcription analysis, especially in the developing world.

Dr. Prasanta presents transcriptome analysis applied to rice, one of the world’s most essential staple foods. Rice production and yield are critically affected by environmental factors, including drought, flooding, high salinity, extreme temperatures, nutrient and mineral availability, toxins and pollutants, etc. Because of the complexity of influences on crop yield, it is essential to define the intricate regulatory gene networks and their signaling pathways involved in stress responses. High-throughput RNA-Seq data have provided an abundance of transcriptome data on rice. RNA-Seq provides data regarding not only coding mRNAs but also...
Transcriptome analysis is the study of the transcriptome, of the complete set of RNA transcripts that are produced under specific circumstances, using high-throughput methods. Transcription profiling, which follows total changes in the behavior of a cell, is used throughout diverse areas of biomedical research, including diagnosis of disease, biomarker discovery, risk assessment of new drugs or environmental chemicals, etc. Transcription profiling can be applied to loss-and gain-of-function mutants to identify the changes associated with the mutant phenotype. Transcriptomics also allows the identification of pathways that respond to or ameliorate environmental stresses. RNA sequencing (RNA-Seq) detects all transcripts in a sample, including mRNAs as well as the regulatory siRNA and lncRNA transcripts. RNA-Seq can also identify disease-associated gene fusions, single nucleotide polymorphisms, and even allele-specific expression.

Transcriptome analysis is most commonly used to compare specific pairs of samples. The differences may be due to different external environmental conditions, for example, hormonal effects or toxins. More commonly, healthy and disease states are compared. In general, transcriptome analysis is a very powerful hypothesis-generating tool rather than a theory-proving one. Transcriptome analyses have become indispensable in basic research and translational and clinical studies.

In this volume, Dr. Pyo Hong discusses the role of long RNA sequences in transcriptome analysis. It should be noted that early RNA-Seq methods generated rather short reads, 35 to a few hundred nucleotides, and relied on massive redundancy to achieve required accuracy. Newer methods, which provide longer reads, have significant advantages, for example, in the analysis of previously not sequenced genomes. Such approaches need tailored software methods.

Dr. Shinichi describes next-generation single-cell sequencing technology developed by his team. It can be used for single-cell transcriptome analysis in tumor tissues. This is an extremely important area nowadays because it is clear that most tumors are heterogeneous. Identifying the transcriptome of tumor stem cells may lead to specific targeting of these cells. Alternatively, single-cell transcriptome analysis can help in defining the tumor-infiltrating immune cells, a critical component of immunotherapies. Dr. Shinichi and his team developed a microwell device that can be easily transported and is relatively cheaper than most other RNA-Seq methods, which will be essential for the widespread use of transcription analysis, especially in the developing world.

Dr. Prasanta presents transcriptome analysis applied to rice, one of world’s most essential staple foods. Rice production and yield are critically affected by environmental factors, including drought, flooding, high salinity, extreme temperatures, nutrient and mineral availability, toxins and pollutants, etc. Because of the complexity of influences on crop yield, it is essential to define the intricate regulatory gene networks and their signaling pathways involved in stress responses. High-throughput RNA-Seq data have provided an abundance of transcriptome data on rice. RNA-Seq provides data regarding not only coding mRNAs but also
noncoding RNAs, components of regulatory gene networks involved in the stress response. These results may enable more optimal cultivating conditions and help to develop new tolerant varieties of rice.

Dr. Xiangyuan focused his studies on the reproductive systems of flowering plants, specifically the gene regulatory networks in anthers, the parts of the stamen that produce and contain pollen.

Prof. Sadovsky analyzed the coding sequences of few conifers, comparing the usage of triplet codons in cold-adjusted plants.

We can anticipate a greatly expanded usage of transcriptome analysis, especially when translated to the bedside, to provide better understanding and more specific diagnoses, enabling physicians to establish diagnoses quickly and reliably.

Miroslav Blumenberg
NYU School of Medicine,
USA
Section 1

Introduction
Chapter 1
Introductory Chapter: Transcriptome Analysis
Miroslav Blumenberg

The central dogma of molecular biology describes the flow of genetic information from genes to functions of the cells and organisms. This comprises a two-step process: first, DNA, the permanent, heritable, genetic information repository, is transcribed by the RNA polymerase enzymes into RNA, a short-lasting information carrier; second, a subset of RNA, the messenger RNAs, mRNAs, are translated into protein. The transcriptome, then, is the complete set of all RNA molecules in a cell, a population of cells or in an organism.

Importantly, not all RNAs are translated into proteins, some serve a structural function, for example, rRNAs in the assembly of ribosomes, others are transporters, e.g., tRNAs, yet others serve regulatory functions, for example, the siRNAs, short interfering RNA, or lncRNAs, long non-coding RNAs; these are not translated into proteins [1]. However, these non-coding RNAs can and often do play roles in human diseases such as cancer, cardiovascular, and neurological disorders. While transcriptomics is most commonly applied to the mRNAs, the coding transcripts, transcriptomics also provides important data regarding content of the cell noncoding RNAs, including rRNA, tRNA, lncRNA, siRNA, and others. Specific approaches address the analysis of splice variant of the same gene in different tissues.

1. Transcriptome analysis

Transcriptome Analysis is the study of the transcriptome, of the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell, using high-throughput methods. Transcription profiling, which follows changes in behavior of a cell in toto, not of a single gene or just a few genes, is used throughout diverse areas of biomedical research, including disease diagnosis, biomarker discovery, risk assessment of new drugs or environmental chemicals etc. Transcription profiling can be applied to loss- and gain-of-function mutants to identify the changes associated with the mutant phenotype. The transcriptomic techniques have been particularly useful in identifying the functions of genes. Transcriptomics also allows identification of pathways that respond to or ameliorate environmental stresses. RNA-Seq can also identify disease-associated gene fusions, single nucleotide polymorphisms and even allele-specific expression.

2. Uses of transcriptome analysis

Transcriptome Analysis is most commonly used to compare specific pairs of samples. The differences may be due to different external environmental conditions, e.g., hormonal effects or toxins. More commonly, healthy and disease states
Chapter 1

Introductory Chapter: Transcriptome Analysis

Miroslav Blumenberg

The central dogma of molecular biology describes the flow of genetic information from genes to functions of the cells and organisms. This comprises a two-step process: first, DNA, the permanent, heritable, genetic information repository, is transcribed by the RNA polymerase enzymes into RNA, a short-lasting information carrier; second, a subset of RNA, the messenger RNAs, mRNAs, are translated into protein. The transcriptome, then, is the complete set of all RNA molecules in a cell, a population of cells or in an organism.

Importantly, not all RNAs are translated into proteins, some serve a structural function, for example, rRNAs in the assembly of ribosomes, others are transporters, e.g., tRNAs, yet others serve regulatory functions, for example, the siRNAs, short interfering RNA, or lncRNAs, long non-coding RNAs; these are not translated into proteins [1]. However, these non-coding RNAs can and often do play roles in human diseases such as cancer, cardiovascular, and neurological disorders. While transcriptomics is most commonly applied to the mRNAs, the coding transcripts, transcriptomics also provides important data regarding content of the cell noncoding RNAs, including rRNA, tRNA, lncRNA, siRNA, and others. Specific approaches address the analysis of splice variant of the same gene in different tissues.

1. Transcriptome analysis

Transcriptome Analysis is the study of the transcriptome, of the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell, using high-throughput methods. Transcription profiling, which follows changes in behavior of a cell in toto, not of a single gene or just a few genes, is used throughout diverse areas of biomedical research, including disease diagnosis, biomarker discovery, risk assessment of new drugs or environmental chemicals etc. Transcription profiling can be applied to loss- and gain-of-function mutants to identify the changes associated with the mutant phenotype. The transcriptomic techniques have been particularly useful in identifying the functions of genes. Transcriptomics also allows identification of pathways that respond to or ameliorate environmental stresses. RNA-Seq can also identify disease-associated gene fusions, single nucleotide polymorphisms and even allele-specific expression.

2. Uses of transcriptome analysis

Transcriptome Analysis is most commonly used to compare specific pairs of samples. The differences may be due to different external environmental conditions, e.g., hormonal effects or toxins. More commonly, healthy and disease states
are compared. For example, in cancer, transcriptomics analyses address classification, the mechanisms of pathogenesis and even outcome prediction. Transcriptome studies can classify cancer beyond anatomical location and histopathology. Outcome predictions can establish gene-based benchmarks to predict tumor prognosis and therapy response. These approaches are already in use for personalized medicine, individualized cancer patient therapies.

Organisms and tissues at various stages of development can be molecularly characterized. The transcriptomes of stem cells help to understand the processes of cellular differentiation or embryonic development. Because of its very broad approach transcriptome analysis is a great source for identifying targets for treatment.

2.1 Methodologies

The early approach to study whole transcriptomes used microarrays, a set of defined sequences arranged on a solid substrate [2]. Microarrays almost exclusively represented mRNAs, that is, genes that are translated into proteins.

Nowadays the microarray approach is supplanted by high-throughput RNA sequencing, RNA-Seq, which detects all transcripts in a sample, including the regulatory siRNA and IncRNA transcripts [3]. In this methodology, the bulk RNA is extracted from the sample and copied into stable double-stranded copy DNA, ds-cDNA, which is then sequenced using various sequencing methods [4]. The sequences obtained are aligned to reference genome sequences, available in data banks, to identify which genes are transcribed. Quantitatively, the results provide the expression levels for the transcribed genes. Compared to microarrays, RNA-Seq can measure both the low-abundance and high-abundance RNAs over a five orders of magnitude range and, importantly, RNA-Seq requires much less starting material (nanograms vs. micrograms and even as little as 50 pg) [5]. This made possible analyses of transcriptomes in a single cell, a great advance over bulk tissue RNA analyses [6]. RNA-seq can be used to identify alternative splicing, novel transcripts, and fusion genes (Table 1).

In principle, the assembly of RNA-Seq reads is not dependent on reference genomes and can be used for gene expression studies of poorly characterized species with limited genomic resources. It can also be used to identify novel protein coding regions in sequenced genomes. RNA-seq can be performed using many next-generation sequencing platforms, however, each platform has its own requirements of sample preparation and the instrument design.

![Image of a table comparing RNA-seq methodologies](image)

**Table 1.** Comparison of RNA-seq methodologies.
Transcriptome Analysis

are compared. For example, in cancer, transcriptomics analyses address classification, the mechanisms of pathogenesis and even outcome prediction. Transcriptome studies can classify cancer beyond anatomical location and histopathology. Outcome predictions can establish gene-based benchmarks to predict tumor prognosis and therapy response. These approaches are already in use for personalized medicine, individualized cancer patient therapies.

Organisms and tissues at various stages of development can be molecularly characterized. The transcriptomes of stem cells help to understand the processes of cellular differentiation or embryonic development. Because of its very broad approach transcriptome analysis is a great source for identifying targets for treatment.

2.1 Methodologies

The early approach to study whole transcriptomes used microarrays, a set of defined sequences arranged on a solid substrate [2]. Microarrays almost exclusively represented mRNAs, that is, genes that are translated into proteins.

Nowadays the microarray approach is supplanted by high-throughput RNA sequencing, RNA-Seq, which detects all transcripts in a sample, including the regulatory siRNA and lncRNA transcripts [3]. In this methodology, the bulk RNA is extracted from the sample and copied into stable double-stranded copy DNA, ds-cDNA, which is then sequenced using various sequencing methods [4]. The sequences obtained are aligned to reference genome sequences, available in data banks, to identify which genes are transcribed. Quantitatively, the results provide the expression levels for the transcribed genes. Compared to microarrays, RNA-Seq can measure both the low-abundance and high-abundance RNAs over a five orders of magnitude range and, importantly, RNA-Seq requires much less starting material (nanograms vs. micrograms and even as little as 50 pg) [5]. This made possible analyses of transcriptomes in a single cell, a great advance over bulk tissue RNA analyses [6]. RNA-seq can be used to identify alternative splicing, novel transcripts, and fusion genes.

Table 1. Comparison of RNA-seq methodologies.

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>High-throughput, able to detect all transcripts, including regulatory RNA, no need for reference genome, can be used for novel gene discovery</td>
</tr>
<tr>
<td>Microarrays</td>
<td>Relatively low throughput, limited to detecting mRNAs, requires reference genome, no discovery of novel genes</td>
</tr>
</tbody>
</table>

2.2 Data analysis, repositories and presentation

Improved sequencing technologies necessitated improved data analysis methods to deal with the increased volume of data produced by each transcriptome experiment. Importantly, the results are deposited into transcriptome databases, essential tools for transcriptome analysis. For example Gene Expression Omnibus, www.ncbi.nlm.nih.gov, contains millions of transcription profiling experiments. Such data have potential applications beyond the original aims of an experiment. Typical outputs include quantitative tables of the transcript levels. This requires specific analysis algorithms, often specific to the methodology used. There are software packages to bridge data from disparate methodologies, to identify groups of similar expressed genes, or differentially expressed functionally significant regulatory or metabolic pathways.

Figure 1.
Graphic representations of transcriptome analysis data. (A) Heat map with clustering tree. (B) Venn diagrams of regulated genes.
Transcriptome Analysis

The results of transcriptomic analyses are graphically often presented as heat maps, a system of color-coding that represents different levels of expression of given genes in different samples (Figure 1A). Such presentations also frequently display a clustering of samples, this helps to identify samples with similar gene expression. Another common graphical presentation uses Venn diagrams, which count the transcripts which are equivalently regulated in multiple samples (Figure 1B).

Transcriptome analyses have become indispensable in basic research, translational, and clinical studies. In general, transcriptome analysis is a very powerful hypothesis-generating tool, more than a theory proving one.

3. Specific example: transcriptome analysis applied to human skin

Easily accessible, skin was among the first targets analyzed using ‘omics’ and dermatology embraced the approaches very early [7]. A classic example of coordinated transcriptional regulation was observed in cultured fibroblasts after serum stimulation [2]. Serum addition causes not only rapid recommencement of the cell cycle but, characteristically a wound-healing response, a physiological role of fibroblasts in wound healing [8]. Transcriptional responses of epidermal keratinocytes to UV light, hormones, vitamins, infections, inflammatory and immunomodulating cytokines, toxins and allergens have been characterized, as were the changes associated with epidermal differentiation [9, 10].

The expression signatures that define the various cell types in human skin, were used to define 20 specific gene signatures, including those for keratinocytes, melanocytes, endothelia, adipocytes, immune cells, hair follicles, sebaceous, sweat, and apocrine glands. This resource provided a resource named SkinSig, which was then used to analyze 18 skin conditions, providing in-context interpretation of, for example, influx in immune cells in inflammation or differentiation changes in disorders of cornification [11].

In the future we can anticipate a greatly expanded usage of transcriptome analysis. Translated to the bedside, it can provide better understanding and more specific diagnoses of diseases. This, of course, requires additional advances in the technology, both in the lab-bench components reducing the costs and guaranteeing reproducibility and accuracy, as well as in the computer-based components, algorithms that enable physicians to establish diagnosis quickly and reliably. In a generation, this approach will become routine.

Author details

Miroslav Blumenberg
NYU School of Medicine, USA

*Address all correspondence to: miroslav.blumenberg@nyulangone.org

IntechOpen
© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Introductory Chapter: Transcriptome Analysis
DOI: http://dx.doi.org/10.5772/intechopen.85980

References


Section 2

Tumor Transcriptome
Chapter 2
Single-Cell Transcriptome Analysis in Tumor Tissues
Sadahiro Iwabuchi and Shinichi Hashimoto

Abstract
The tumor microenvironment is comprised of cancer cells and their surroundings, including various normal cells and non-cellular components, and each tumor tissue has a distinctive microenvironment. Cancer progression is affected by different microenvironmental states, such as the heterogeneity of infiltrating immune cells. Therefore, it is necessary to understand the complex cell-to-cell interactions associated with tumor developmental stages in different tissues. Recent revolution of single-cell RNA sequencing technology can uncover the tumor microenvironment diversity. We have developed a novel strategy of single-cell transcriptome analysis: next generation 1-cell sequencing (Nx1-seq) technology, and it allows for profiling of thousands of single cells from tumor tissue. Our microwell with cell bar-code beads device can detect genes with high sensitivity, and it is easily transported anywhere without any other dedicated devices. Further, the developmental cost is relatively cheaper than other single-cell RNA sequencing methods. In this study, we introduce representative application of the single-cell RNA sequencing technique in gynecological cancers, and we show the result of Nx1-seq application in human endometrioid adenocarcinoma tissue.

Keywords: tumor microenvironment, single-cell transcriptome analysis, Nx1-seq

1. Introduction
Tumor tissues are aggregates of various cell populations, and each single cell or cell population plays an important role for cancer progression and regression. The representative cell populations of the tumor microenvironment are cancer cells, surrounding normal cells, and infiltrated immune cells of all types. Anticancer agents and immune checkpoint blockers, such as programmed death receptor-1 (PD-1) and its ligand, have been widely used in patients, and the curative effect is great. However, for many patients, these treatments are ineffective because the minor cell populations escape the immune system. Therefore, a deeper understanding of the tumor microenvironment immunology will be critical for immunotherapy to become a standard therapy. In addition, it is important to clarify patient and tumor-dependent cell phenotypes by gene expression analysis because the composition and functions of the tumor microenvironment are heterogeneous between cancers and patients.

Previous gene expression measurements have been performed on bulk samples. Conventional bulk-based RNA sequencing or microarrays alone or in combination with flow cytometry can provide a full view of all gene expression, and it is useful to
Chapter 2

Single-Cell Transcriptome Analysis in Tumor Tissues

Sadahiro Iwabuchi and Shinichi Hashimoto

Abstract

The tumor microenvironment is comprised of cancer cells and their surroundings, including various normal cells and non-cellular components, and each tumor tissue has a distinctive microenvironment. Cancer progression is affected by different microenvironmental states, such as the heterogeneity of infiltrating immune cells. Therefore, it is necessary to understand the complex cell-to-cell interactions associated with tumor developmental stages in different tissues. Recent revolution of single-cell RNA sequencing technology can uncover the tumor microenvironment diversity. We have developed a novel strategy of single-cell transcriptome analysis: next generation 1-cell sequencing (Nx1-seq) technology, and it allows for profiling of thousands of single cells from tumor tissue. Our microwell with cell bar-code beads device can detect genes with high sensitivity, and it is easily transported anywhere without any other dedicated devices. Further, the developmental cost is relatively cheaper than other single-cell RNA sequencing methods. In this study, we introduce representative application of the single-cell RNA sequencing technique in gynecological cancers, and we show the result of Nx1-seq application in human endometrioid adenocarcinoma tissue.

Keywords: tumor microenvironment, single-cell transcriptome analysis, Nx1-seq

1. Introduction

Tumor tissues are aggregates of various cell populations, and each single cell or cell population plays an important role for cancer progression and regression. The representative cell populations of the tumor microenvironment are cancer cells, surrounding normal cells, and infiltrated immune cells of all types. Anticancer agents and immune checkpoint blockers, such as programmed death receptor-1 (PD-1) and its ligand, have been widely used in patients, and the curative effect is great. However, for many patients, these treatments are ineffective because the minor cell populations escape the immune system. Therefore, a deeper understanding of the tumor microenvironment immunology will be critical for immunotherapy to become a standard therapy. In addition, it is important to clarify patient and tumor-dependent cell phenotypes by gene expression analysis because the composition and functions of the tumor microenvironment are heterogeneous between cancers and patients.

Previous gene expression measurements have been performed on bulk samples. Conventional bulk-based RNA sequencing or microarrays alone or in combination with flow cytometry can provide a full view of all gene expression, and it is useful to
investigate the tumor microenvironment. However, a blended gene expression analysis might mask the minor cell population, which may be the origin of tumor progression. To overcome this problem, RNA sequencing methods that can analyze mRNA expression at the single-cell level from thousands of individual cells are required. The fundamentally necessary approaches of single-cell RNA sequencing are: (1) single-cell isolation with high survival rate, (2) cell lysis to obtain mRNA, (3) conversion of mRNA into cDNA, (4) specific amplification of cDNA, (5) cDNA fragmentation process, and (6) creation of high-quality sequencing libraries. After single-cell isolation, there are some innovative single-cell transcriptome analysis methods (e.g., CEL-seq [1], Quartz-seq [2], Quartz-seq2 [3], Smart-seq [4], Drop-seq [5], iDrop RNA sequencing [6], Cyto-Seq [7], automated microwell-based RNA sequencing [8], and our next generation 1-cell sequencing; Nx1-seq [9]), and every method uses oligo-dT primers containing cell-specific bar-codes, which tag cDNA from single cells.

Although cell number, tissue volume analyzed, analysis sensitivity, and overall cost for creating libraries are completely different, any methods with an efficient data analysis procedure would be particularly useful to understand cellular heterogeneity and to identify rare cell populations. For example, six prominent single-cell RNA sequencing methods: CEL-seq2, Drop-seq, MARS-seq, SCRB-seq, Smart-seq, and Smart-seq2 have been compared in mouse embryonic stem cells [10]. If single-cell transcriptome analysis were performed in a limited number of cells or small tissue volume, SCREB-seq and MARS-seq will have better sensitivity. Yet, Smart-seq2 may detect the highest number of genes per cell with amplification noise. Drop-seq is a preferable and more cost-effective method for large numbers of cells with low sequencing depth. In terms of the number of reads per cell and genes, we also compared our Nx1-seq and Drop-seq, and it revealed similar sensitivity [9].

Recently, another microwell-based RNA sequencing has been developed [11], and this is a simple, high-throughput, and low-cost device. The principle of their device is similar to Cyto-Seq and Nx1-seq, but the differences are the beads material and the loading order of single cells and beads to the microwell. They have attempted to construct a “mouse cell atlas” by using over 50 mouse tissues, organs, and cell cultures. One of the reasons they can analyze a large sample amount is the low-cost device without any expensive, exclusive apparatus and kits for capturing mRNA from a single cell. Previously, a detailed description of each method was thoroughly reviewed [12]; yet, innovate new technologies for single-cell RNA sequencing are still to be developed. We also continue improving our Nx1-seq device progressively.

2. Single-cell transcriptome analysis for cancer tissues

To find new molecular targets for a cancer prognosis prediction method, it requires an understanding of the single-cell level transcriptome heterogeneity in tumor tissues and their microenvironment. Bulk-based RNA sequencing may also contribute to development of new minimally invasive monitoring of circulating tumor cells or cancer gene-transferred macrophages and lymphoid cells. If the targeted cancer antigen and/or cell surface protein were held in small cell populations, the intensity signal of the gene expression would be weak. In this case, single cell transcriptome analysis is a useful tool to identify the small cell population and obtain all of the gene information in this population. In the next chapter, we describe our Nx1-seq methods in detail and show a representative Nx1-seq application in human endometrioid adenocarcinoma tissue. At this time, there are no reports about single-cell transcriptome analysis for endometrioid adenocarcinoma, except our research [9]. Here, we briefly summarize recent applications of single-cell RNA sequencing in one of the major gynecological cancers, breast cancer.
Chung et al. conducted single-cell transcriptome analysis for 11 primary tumors and 2 metastatic lymph nodes from 11 patients, representing 4 breast cancer subtypes [13]. It clearly displayed the carcinoma and tumor-infiltrating immune cells population using the 10–17 μm integrated fluidic circuit mRNA sequencing chip in the C1™ Single-Cell Auto Prep System of Fluidigm®. The C1™ integrate fluidic circuit is an integrated microfluidic system that can automatically isolate individual cells from suspended cells. Subsequently, cell lysis buffer is automatically applied to individual cells to capture mRNA. It takes 5 h to make sequence-ready libraries from isolated cells, and the operation is simple [14]. The authors demonstrated that many T cells with high cytokine and chemokine expression were observed in three triple negative breast cancers (TNBC), and their phenotypes were regulatory T cells (two out of three patients) and another one was exhaustion and cytotoxicity signatures [13]. This result indicates that immune checkpoint blockers may be effective in the patient.

Recently, single-cell transcriptome analysis using 10X Genomics Chromium was reported in breast cancer [15, 16]. Cazet et al. investigated the anti-tumor inhibitor effects in a mouse tumor model, in terms of changes in the gene expression profiles of each cell population [15]. Tumor development and progression were associated with stiffness of the extracellular matrix, and collagen density in the tumor-stromal interface was reduced by small molecule inhibitor of smoothened (SMO) treatment. They also showed that the chemotherapy significantly slowed tumor growth and reduced the frequency of metastatic disease in xenograft models of human TNBC. In another article, an infiltrating T cell population in breast cancer was classified from 123 patients, and it demonstrated the importance of qualitative identification of CD8+ T cell subtypes [16]. CD8+ CD103+ T cells contained features of read tissue-resident memory, including high granzyme B, PD-1, and cytotoxic T lymphocyte (associated) antigen 4 (CTLA-4), rather than CD8+ CD103- T cells, meaning that these are target cells for immune checkpoint brokers.

The above representative reports using single-cell read transcriptome analysis were well analyzed, but we speculate that the cost for creating a sequence library per sample using commercially available device-dependent kits may be expensive. Many samples should be analyzed in a clinical study because the observed microenvironment heterogeneity is patient-, malignant-, or organ-dependent. In addition, if characterization of tumor gene expression profiling was recognized according to the individual’s region or country, it should be performed locally because fresh samples, not frozen ones, are better to analyze for RNA sequencing. From this standpoint, a device with low-cost, in high sensitivity, and easy performance is recommended.

3. Nx1-seq

The major component of Nx1-seq (next generation 1-cell sequencing) consists of bar-code beads and a specifically processed microwell. In this chapter, we describe these devices in further detail.

3.1 Bar-code beads

Oligonucleotides on beads have the following sequence: (1) “root array” is used as a priming site for subsequent PCR; (2) “cell bar-code” allocates 12 bp of oligonucleotide to identify cells, and the bar-code has $4^{12} = 16,777,216$ various patterns; (3) “UMI” (a unique molecular identifier) has 8 bp of oligonucleotide to
eliminate gene duplication bias and improve signal/noise ratio by PCR, meaning that 1-cell bar-code has 1 UMI; and (4) “poly-dT” array consists of 25 bp oligo dT sequences for capturing polyadenylated mRNA. The bar-code beads (“root”-“cell bar-code”-“UMI”-“poly-dT”) were made by following a modified instruction manual for the GS Junior Titanium emulsion PCR Kit (Lib-L) from Roche® Applied Science or synthesized by ChemGenes Corporation (Wilmington, MA, USA) with additional annealing and ligation of the poly-dT array in our laboratory. The detailed method for generating bar-code beads using the emulsion PCR kit is described in our previous report [9]. We could get randomly synthesized various “cell bar-code” inserted bar-code beads, and the beads were washed with Low TE buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA pH 8.0) and stored at −20°C until use.

3.2 Microwell slide

The microwell plate was prepared using polydimethylsiloxane (PDMS) and was cut 2 × 2 × 2 cm using cutting dies (Noda Co. Ltd., Osaka, Japan) which contained 1.3–1.6 × 10⁵ microwells. The size of one microwell was 25 ± 3 μm diameter, 40 ± 8 μm height, 20 ± 9 μL capacity (column-shape), and the distance between microwells was 5 μm (YODAKA CO., Ltd., Kanagawa, Japan). If the size of the target cell was not between 15 and 25 μm (YODAKA CO., Ltd., Kanagawa, Japan). If the size of the target cell was not between 15 and 25 μm, the diameter and height sizes were easily adjusted. The PDMS microwell plate was placed in an oxygen plasma chamber for hydrophilic processing because PDMS is a hydrophobic material. The microwell plate was quickly set into the Nunc™ Lab-Tek™ Chamber slide system (Thermo Fisher Scientific, Waltham, MA, USA), and bar-code beads were applied to the microwell plate. If the expected number of cells obtained from the tumor tissue was <1 × 10⁵ cells, the PDMS microwell was cut ~1/4 or 1/2 of its size and set into the appropriate Nunc™ Lab-Tek™ Chamber slide system (Figure 1). The PDMS microwell plate was kept at 4°C, meaning that the Nx1-seq device can be stored until use.

Figure 1.
Schematic drawings of Nx1-seq. Cell bar-code beads (see the structure of cell bar-code bead) were filled into microwells, and an adequate number of single cells was applied. Cells were dissolved in lysis buffer, and mRNA from the cell was captured by cell bar-code beads in each microwell. After cellular lysis, all beads were collected into a single tube. Images of the microwell plate show that our device had some variations for differences of the number of applied cells.
3.3 Lysis of cells

After single cell isolation, ~1–2 × 10^5 cells mixed with 3.7 mL of cold PBS were applied to PDMS microwell plate (2 × 2 × 2 cm) and put the cover without entering a bubble. The microwell plate was put on ice for 10–15 min, which let the cells settle into the microwell by gravity. About 5% of whole microwells were filled with single cells according to Poisson distribution. The solution was removed from the microwell plate, and 1 mL of fresh cold PBS was gently applied. The washing process was repeated by 3–4 times. The reagent composition of 1 mL of cell lysis buffer was; 2 mg of N-Lauroylsarcosine sodium salt, 200 μL of 1 M Tris-HCl pH 7.5, 40 μL of 0.5 M EDTA pH 8.0, 750 μL of deionized water, 50 μL of 1 M dithiothreitol solution. The microwell plate was put on a microscopy, and we found the microwell which contains only cell without bar-code bead, then PBS was removed and 1 mL of cell lysis buffer was gently applied from the corner of the microwell. Most cells were getting to dissolve within 1–3 min, but it kept for 8 min. The cell lysis buffer was removed carefully and washing buffer (200 mM Tris-HCl pH 7.5, 20 mM EDTA, 50 mM DTT, 0.2% N-Lauroylsarcosine sodium salt, 2% Ficoll) was added. Conversion of mRNA into cDNA was done by SuperScript™ II or IV Reverse Transcriptase (Thermo Fisher Scientific).

4. Nx1-seq application to human endometrioid adenocarcinoma tissues

Previously, we reported the application of Nx1-seq to human endometrioid adenocarcinoma (EA) tissues [9]. Here, we summarize the result shortly. EA tissues were removed from the myometrial infiltration side (M-side) and endometrial side (E-side). Myometrial invasion is an independent prognostic parameter of EA, and invasion is correlated with the risk of metastasis to the lymph nodes. Single-cell analysis in each side revealed that EA had six cancer (cluster #0, 1, 2, 3, 5, 6), two macrophage (#4, 8), and one T cell population (#7) (Figure 2A). To analyze the sequencing data, we used Seurat software (http://satijalab.org/seurat/), which is an open tool for analyzing single-cell genomics in R (http://www.R-project.org/). As shown in Figure 2B, the distribution of cancer cells on the E-side and M-side differed, and the majority of the macrophage cluster (#4) was on the M-side. The number of infiltrating macrophages was not different between sides (Figure 2C), but macrophage specificity was more cytotoxic T lymphocytes (CTL)-like on the M-side. Macrophages on the M-side had higher expression of inflammatory chemokines, C-X-C motif chemokine ligand 3 and 8 (CXCL3 and CXCL8) and NF-κ-B inhibitorα (NFKBIA) (Figure 2D). The proportion of macrophages expressing the inflammatory factors CCL5, IL10 and IL6 did not differ among the two sides (data not shown). It has been widely believed that many cells expressing some malignancy-related genes exist on the M-side; however, our previous result showed that cancer cells on the E-side were highly malignant when compared to those on the M-side.

In addition, a cancer stem-like cell population was also higher on the E-side (e.g., the ratio of SOX2+ cells on E-side vs. M-side was 17 vs. 6%, respectively) [9]. These data reveal that cells with high malignant potential (HMP) are present at the same site of cancer tissue (E-side) in EA. To confirm our hypothesis, we focused on the ubiquitin C-terminal hydrolase L1 (UCHL1) gene. Protein ubiquitination or de-ubiquitination regulates cell growth, differentiation, transcription, and tumor prognosis. The function of UCHL1 in neurodegenerative disorders, particularly in Alzheimer's disease and Parkinson's disease has been reported, and decreased
Hydrolase activity and UCHL1 ligase activity may affect the neurodegeneration [17, 18]. In our EA tissue, the relative intensity of UCHL1 expression was higher on the E-side (Figure 2E). The functional role of UCHL1 in human tumor malignancy is still unresolved, but this gene has been reported to be cancer-related in endometrial cancer patients [19]. Goto et al. demonstrated that activation of UCHL1 via hypoxia inducible factor-1 (HIF-1) is the key regulator for underling mechanism of tumor metastasis, and they expected UCHL1 is as prognostic marker and treatment target for breast and lung cancers [20].

From the enormous single-cell RNA-sequencing data, the researcher must determine to manage and understand the functional meaning of the cell population. In particular, understanding how the gene is related to overall survival of EA patients in the clinical site is useful. Hence, we used the “cBioPortal For CANCER GENOMICS” website and chose “Uterine Corpus Endometrial Carcinoma (EC) (TCGA, Provisional).” Subsequently, we set “Genomic profiles” as “mRNA Expression,” and chose “mRNA Expression z-Score (microarray),” then input the gene name “UCHL1” (http://www.cbioportal.org/). Overall Survival of Kaplan-Meier (K-M) Estimate showed that high UCHL1 expression in endometrial carcinoma patients significantly decreased survival time (Figure 2F). The median months survival in the UCHL1 high group was 48.75 months. The log-rank p value for K-M analysis for correlation between mRNA expression level and patient survival was 1.965 × 10⁻⁴. The Overall Survival of K-M Estimate was not calculated from the EA but EC dataset, however EA of the endometrium is the most common type of EC [21]. Therefore, the result indicates that higher expression of UCHL1 on the E-side somehow affects EA progression, and it supports our hypothesis that cells with HMP are present on the E-side. Whether we chose other data set “Uterine Corpus Endometrial Carcinoma (TCGA, Nature 2013),” the result of Overall Survival of K-M Estimate was also significant (p = 1.06 × 10⁻³). The median months of disease-free in high UCHL1 patients was 12.94 months, and it was significantly earlier by the Disease/Progression-free Kaplan-Meier Estimate. However, the significant correlation was not observed if we chose “mRNA Expression z-Scores (RNA Seq V2 RSEM), z-score threshold ± 2.0” (p = 0.955). There was other useful database to realize the overall survival of EC patients. We used the “THE HUMAN PROTEIN ATLAS” website and input the gene name “UCHL1,” then set “PATHOLOGY ATLAS” (http://www.proteinatlas.org/). The prognostic summary highlighted that UCHL1 was the candidate as the prognostic marker in EC. The 5-year survival in the UCHL1 high or low group was 66 or 86% respectively, and the p score was 4.1 × 10⁻⁵ from the total of 541 female patients.
Transcriptome Analysis

Hydrolase activity and UCHL1 ligase activity may affect the neurodegeneration [17, 18]. In our EA tissue, the relative intensity of UCHL1 expression was higher on the E-side (Figure 2E). The functional role of UCHL1 in human tumor malignancy is still unresolved, but this gene has been reported to be cancer-related in endometrial cancer patients [19]. Goto et al. demonstrated that activation of UCHL1 via hypoxia inducible factor-1 (HIF-1) is the key regulator for the underlying mechanism of tumor metastasis, and they expected UCHL1 is as a prognostic marker and treatment target for breast and lung cancers [20]. From the enormous single-cell RNA-sequencing data, the researcher must determine to manage and understand the functional meaning of the cell population. In particular, understanding how the gene is related to overall survival of EA patients in the clinical site is useful. Hence, we used the “cBioPortal For CANCER GENOMICS” website and chose “Uterine Corpus Endometrial Carcinoma (EC) (TCGA, Provisional).” Subsequently, we set “Genomic profiles” as “mRNA Expression,” and chose “mRNA Expression z-Score (microarray),” then input the gene name “UCHL1” (http://www.cbioportal.org/). Overall Survival of Kaplan-Meier (K-M) Estimate showed that high UCHL1 expression in endometrial carcinoma patients significantly decreased survival time (Figure 2F). The median months survival in the UCHL1 high group was 48.75 months. The log-rank p value for K-M analysis for correlation between mRNA expression level and patient survival was $1.965 \times 10^{-4}$. The Overall Survival of K-M Estimate was not calculated from the EA but EC dataset, however EA of the endometrium is the most common type of EC [21]. Therefore, the result indicates that higher expression of UCHL1 on the E-side somehow affects EA progression, and it supports our hypothesis that cells with HMP are present on the E-side. Whether we chose other data set “Uterine Corpus Endometrial Carcinoma (TCGA, Nature 2013),” the result of Overall Survival of K-M Estimate was also significant ($p = 1.06 \times 10^{-3}$). The median months of disease-free in high UCHL1 patients was 12.94 months, and it was significantly earlier by the Disease/Progression-free Kaplan-Meier Estimate. However, the significant correlation was not observed if we chose “mRNA Expression z-Scores (RNA Seq V2 RSEM), z-score threshold ± 2.0” ($p = 0.955$). There was other useful database to realize the overall survival of EC patients. We used the “THE HUMAN PROTEIN ATLAS” website and input the gene name “UCHL1,” then set “PATHOLOGY ATLAS” (http://www.proteinatlas.org/). The prognostic summary highlighted that UCHL1 was the candidate as the prognostic marker in EC. The 5-year survival in the UCHL1 high or low group was 66 or 86% respectively, and the $p$ score was $4.1 \times 10^{-5}$ from the total of 541 female patients.

As shown in Figure 3, there was significant differences about UCHL1 expression between each side, and immunostaining of UCHL1 showed a similar staining pattern.
pattern in macrophages as well as cancer cells. Also, UCHL1 staining in E-side was relatively higher. These data indicate that microenvironment of tumor tissue might affect the gene properties of immune cells, and gene expression pattern might resemble closely to cancer cells.

The high expression of CXCL3 was not related to the prognosis of EA ($p = 0.987$) by the CANCER GENOMICS, but CXCL3 high group significantly improved 5-year overall survival by the PROTEIN ATLAS. At this moment, there was only one patient whose expression of CXCL8 was high in the CANCER GENOMICS, but 5-year survival of CXCL8 high or low group in EC patients was 79 or 70% respectively, and there was no significant difference. In contrast, higher expression of NFKBIA significantly decreased survival time by “mRNA Expression z-Sore (microarray)” ($p = 0.0246$), but not “mRNA Expression z-Sore (RNA Seq V2 RSEM)” by the CANCER GENOMICS. In contrast, the 5-year overall survival in the PROTEIN ATLAS was not significant. These results indicate that the researcher must use some database to understand how the target genes are related to the prognosis of cancers. Further studies for other HMP-related genes are ongoing in our laboratory.

5. Conclusion

Single-cell sequencing is believed to be a powerful tool to answer unknown biological questions, and researchers may have many expectations to find new insights of their hypotheses. Indeed, bulk-based RNA sequencing is averaged across a cell population, but the method to obtain total RNA is relatively simple and easy. Most importantly, we can detect gene expression profiling of the whole tissue. Of course, as mentioned above, the existence of minor cell populations, such as cancer stem-like cells, may not be detected in bulk-based RNA sequencing data. If the researchers knew the biomarkers of targeted cells in small population, the gene can be detected from the bulk-based RNA sequencing data. But it is unknown which cell expresses and how many cells have the targeted gene because of the averaged data by bulk-based RNA sequencing. Thus, it is better to ponder over which method is aimed at the biological question before choosing more difficult and expensive single-cell RNA sequencing.

Current protocols of dispersing single cells in each tissue are not optimized worldwide; therefore, some cells or cell populations may disappear in the course of isolating single cells from tumor tissue. One of the most important procedures for single-cell RNA sequencing is isolation of single cells from tumor tissues. Mechanical and/or enzymatic cell distributed processes followed by fluorescence-activated cell sorting (FACS), magnetic-activated cell sorting (MACS), or density-gradient method are the current standard [22], but the softness or hardness of tissues differs depending on the tumor. Inappropriate single-cell isolation methods are biased; therefore, more detailed studies are needed to optimize isolation of single cells for each tissue.

Nonetheless, single-cell sequencing is a great tool for detecting heterogeneous subpopulations, cell-to-cell communication, and spatial interactions. Moreover, the many gene expression changes by carcinostatic agents can be monitored. To analyze extensively heterogeneous clinical samples, highly sensitive, low cost, quick, and simple technologies to capture mRNA from a single cell are required. Our newly developed single-cell transcriptome analysis, Nx1-seq, can be a useful tool to understand tumor microenvironments with high sensitivity and low cost. This new approach is a simple method, and it can be used to analyze several hundreds to tens of thousands of cells without specialized equipment. Further, it is easy to
change the size of the microwell for larger or smaller cells. Furthermore, microwells equipped with bar-code beads in the Nunc™ Lab-Tek™ Chamber slide system can be stored for several months before use. Nx1-seq is a powerful approach for characterizing cellular diversity under physiological and pathological conditions. The combined analysis of t-SNE by Seurat and detailed gene profiling can discover new tumor biomarkers or new target genes for regression of tumor tissues. We continue to develop better Nx1-seq devices to satisfy requests from researchers. It is about continued learning on a daily basis.

Acknowledgements

The authors would like to thank Y. Kamide, S. Kurokawa, and H. Sekine for technical assistance with the experiments. We also thank A. Tagata for assistance of accounting work. We greatly appreciate Dr. Y. Takamura for supplying PDMS membrane. Prof. T. Torigoe and Dr. Y. Hirohashi gave us the data of UCHL1 immunohistochemistry and constructive comments, thank you. This work is supported by the Japan Agency for Medical Research and Development (AMED). We would like to thank Editage (http://www.editage.jp) for English language editing.

Conflict of interest

The authors have no conflicts of interest directly relevant to the content of this article.

Author details

Sadahiro Iwabuchi* and Shinichi Hashimoto
Department of Integrative Medicine for Longevity, Graduate School of Medical Sciences, Kanazawa University, Kanazawa, Ishikawa, Japan

*Address all correspondence to: s_iwabuchi@staff.kanazawa-u.ac.jp

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
References


[8] Yuan J, Sims PA. An automated microwell platform for large-scale single cell RNA-seq. Scientific Reports. 2016;6:33883-33892. DOI: 10.1038/srep33883


cancer stem cell plasticity overcomes chemoresistance in triple negative breast cancer. Nature Communications. 2018;9:2897-2904. DOI: 10.1038/s41467-018-05220-6


Section 3

Reference Transcriptomes
Chapter 3

Transcriptome Atlas by Long-Read RNA Sequencing: Contribution to a Reference Transcriptome

Dong Jin Lee and Chang Pyo Hong

Abstract

The recent emergence of long-read transcriptome sequencing has helped improve the overall accuracy of gene prediction compared with that by short-read RNA-Seq. In addition, the technology can offer a more comprehensive view of functional genomics in uncharacterized species with an efficient full-length unigene build and high-precision gene annotation, thus being efficient in developing transcriptome data resources from useful genetic pools. Hence, I will review the applications of long-read RNA isoform sequencing, including the relative merits of the technology, the improvement of the accuracy in gene prediction and gene annotation, and the full-length unigene builds in a new genome; the limitations of the technology will be also discussed. The review will be valuable in collecting data resources for functional genomic studies.

Keywords: functional genomics, gene prediction, long-read RNA sequencing, transcriptome

1. Introduction

Transcriptomics is the study of transcript catalogs in a cell, tissue, or organism for a given developmental stage or physiological condition [1]. The transcriptome indicates the complete set of transcripts that consists of protein-coding messenger RNA (mRNA) and non-coding RNA (ncRNA), including ribosomal RNA (rRNA), transfer RNA (tRNA), and other ncRNAs [2, 3]. In contrast with the relatively stable genome, various factors such as developmental stage, physiological condition, and external environment influence the changes in the transcriptome. The goals of transcriptomics include the annotation of the transcriptome, and the determination of the functional structure of each gene in the genome and the changes in the expression levels of each gene among different transcriptome samples [1, 4, 5].

Transcriptome analysis depends heavily on the availability of high-throughput tools on account of the complexity of the transcriptome. Thus, RNA sequencing (RNA-Seq) has become an important tool for biological studies. RNA-Seq can quantify gene expression spatially and temporally. Although RNA-Seq has enabled the generation of massive amounts of sequence data due to their high-throughput characteristic, their application of short reads makes them poorly suited for genome and transcriptome assembly, and isoform detection. Single-molecule real-time (SMRT) sequencing, a new method to generate long-read sequences developed by...
Chapter 3

Transcriptome Atlas by Long-Read RNA Sequencing: Contribution to a Reference Transcriptome

Dong Jin Lee and Chang Pyo Hong

Abstract

The recent emergence of long-read transcriptome sequencing has helped improve the overall accuracy of gene prediction compared with that by short-read RNA-Seq. In addition, the technology can offer a more comprehensive view of functional genomics in uncharacterized species with an efficient full-length unigene build and high-precision gene annotation, thus being efficient in developing transcriptome data resources from useful genetic pools. Hence, I will review the applications of long-read RNA isoform sequencing, including the relative merits of the technology, the improvement of the accuracy in gene prediction and gene annotation, and the full-length unigene builds in a new genome; the limitations of the technology will be also discussed. The review will be valuable in collecting data resources for functional genomic studies.

Keywords: functional genomics, gene prediction, long-read RNA sequencing, transcriptome

1. Introduction

Transcriptomics is the study of transcript catalogs in a cell, tissue, or organism for a given developmental stage or physiological condition [1]. The transcriptome indicates the complete set of transcripts that consists of protein-coding messenger RNA (mRNA) and non-coding RNA (ncRNA), including ribosomal RNA (rRNA), transfer RNA (tRNA), and other ncRNAs [2, 3]. In contrast with the relatively stable genome, various factors such as developmental stage, physiological condition, and external environment influence the changes in the transcriptome. The goals of transcriptomics include the annotation of the transcriptome, and the determination of the functional structure of each gene in the genome and the changes in the expression levels of each gene among different transcriptome samples [1, 4, 5].

Transcriptome analysis depends heavily on the availability of high-throughput tools on account of the complexity of the transcriptome. Thus, RNA sequencing (RNA-Seq) has become an important tool for biological studies. RNA-Seq can quantify gene expression spatially and temporally. Although RNA-Seq has enabled the generation of massive amounts of sequence data due to their high-throughput characteristic, their application of short reads makes them poorly suited for genome and transcriptome assembly, and isoform detection. Single-molecule real-time (SMRT) sequencing, a new method to generate long-read sequences developed by
PacBio platform, provides an alternative approach to overcome these limitations in sequence length and accelerate improving our understanding of the complexity of the transcripts [6].

In general, the read length of Illumina HiSeq platform is about 100–150 bp, which is relatively short compared to that of PacBio platform (around 10 kb). However, Illumina HiSeq platform has the advantage of generating more accurate reads and high-throughput data. On the other hand, even though its accuracy is lower than that of Illumina HiSeq platform, single-molecule real-time (SMRT) sequencing of PacBio platform, a new method of sequence analysis, was developed and applied to elucidate the genomic structures of difficult to sequence organisms [7] because of its long-reads, which results in the improvement of assembly, gene prediction, and annotation. Using this technique, sequences are analyzed from a single strand of DNA without genomic amplification [9]. PCR-free long-read sequencing enables to help to carry out large complex whole-genomes (i.e., hexaploid wheat and maize).

PacBio sequencing captures sequences during the replication process of the target DNA in real-time. The template, also called a SMRTbell, contains a target double-stranded DNA (dsDNA) ligated with hairpin adaptors at both ends, resulting in a closed and single-stranded circular DNA [8]. When the SMRTbell is loaded into a chip called a SMRT cell, diffusion of the SMRTbell into a sequencing unit called a zero-mode wave guide (ZMW) is carried out [10]. In each ZMW, a single polymerase immobilized at the bottom can bind to adaptors of the SMRTbell [11]. Each of the four nucleotides is fluorescent-labeled. As a nucleotide associates with the template in the active site of the polymerase, a light pulse is produced for base detection. A single polymerase read can be generated up to 40 kb, depending on the library size and sequencing time. The closed-circle form of the SMRTbell can make the reaction repeat until the reaction is terminated after the replication of one strand of the target dsDNA or double-stranded complementary DNA by the polymerase. However, the mean length of full transcripts is 1–3 kb in most plant and animal genomes (e.g., 1.6 kb in Arabidopsis [12], 1.8 kb in rice [13], 2.3 kb in human [14], and 1.2 kb in mouse [15]); thus, the same transcript can be covered multiple times by the long polymerase read. In this scenario, a few reads (called subreads) can be generated from the polymerase read by trimming adaptor sequences. The consensus sequence of multiple subreads in a single ZMW generates a read of insert (ROI) or a circular consensus sequence (CCS) read with higher accuracy. Hence, a protocol of isoform sequencing (Iso-Seq) for long-read transcriptome sequencing that includes library construction, size selection, sequencing, and data processing was developed by PacBio. Iso-Seq allows the direct sequencing of transcripts up to 10 kb, which is particularly useful for the genomes of uncharacterized species.

However, even though PacBio sequencing has an advantage in terms of read length over next-generation sequencing, the throughput of PacBio sequencing is relatively low. A single SMRT cell contains 150,000 ZMWs, each of which can produce one polymerase read with a mean length of 10 kb. Typically, only 35,000–70,000 reads of the 150,000 ZMW wells on a SMRT cell can be produced successfully because of the failure of anchoring a polymerase and loading more than one DNA molecule in a ZMW. Consequently, the typical throughput of the PacBio RS II system is around 0.5–1 Gb per SMRT cell [16]. Recently, PacBio developed another system called Sequel that produces over seven times the reads, with 1,000,000 ZMWs, and yields around 3.5–7 Gb per SMRT cell [17]. Sequel is appropriate for projects such as de novo genome assembly and isoform sequencing of transcriptomes. Another notable problem of PacBio sequencing is the relatively high error rate (around 11–15%) of polymerase reads [18]. Many hybrid sequencing approaches have been attempted to develop a method that has the accuracy of short reads but with the length of PacBio reads [19].
Long-read transcriptome sequencing generates longer and improved transcripts with a high level of assembly completeness and gene annotation. Moreover, it prevents obtaining artifacts such as chimeras, structural errors, incomplete assembly, and base errors [20].

Here, we review the sample preparation, library construction, analytical pipelines, and the result of isoform sequencing (Iso-Seq), as a long-read transcriptome sequencing, in gene prediction and annotation. Furthermore, we will also discuss the relative merits and the limitations of the Iso-Seq technology.

2. Merits of long-read transcriptome sequencing

Long-read transcriptome sequencing such as Iso-Seq generates longer and improved transcripts from a species with a high level of assembly completeness and gene annotation, enabling a comprehensive view of the transcriptome. Conventional methods, such as cDNA cloning and EST sequencing, have limitations with relatively low data coverage. Although deep short-read sequencing (i.e., RNA-Seq) provides good sequencing depth and coverage for genome-wide transcriptome analysis, their short-read length generates assembly incompleteness of transcripts, resulting in high error rate in assembly and unreliable gene annotation. Long-read transcriptome sequencing can also provide experimental verification of predicted gene models in a genome, enable the quality of gene structures predicted and also give the potential to reduce missing gene annotation. For example, missing gene annotation may lead to false interpretation such as gene loss and errors in gene expression profiles that map and quantify RNA-seq reads using predicted gene models. Thus, this technology can be helpful to find full-length (FL) transcripts harboring complete open reading frames (ORFs) and uncover novel splice isoforms as well as novel genes. This can result in the improvement of accuracy of gene prediction with an experimental verification and annotations for aiding in studying gene regulation.

3. Sample preparation and library construction for isoform sequencing

Iso-Seq with the PacBio platform can generate FL cDNA sequences including the 5′ and 3′-UTRs (untranslated regions), as well as the polyA tails of the transcripts. The whole workflow including the experimental protocol and analytical pipelines is illuminated in Figure 1 [10].

3.1 Isolation of total RNA

The samples can be collected from various tissues (i.e., blood, gill, skin, muscle, liver, spleen, intestine, ovary, testis, kidney, heart, and brain of an animal) [21], or from certain developmental stages (developing rabbit at 21, 49, and 84 days of age) [22]. The high quality of RNA with enough purity and integrity is critical to reduce the amplification cycles required in large-scale PCR and improve the sequencing diversity. RNA extraction is usually done through an easy-spin RNA extraction kit, or RNAiso Pure RNA Isolation kit [20–22]. In general, 2–5 μg of total RNA with an RNA integrity number (RIN) greater than 7 is required.

3.2 cDNA synthesis and size partitioning

Isolation of polyA mRNA is required for analyzing the transcripts of protein-coding genes. The Iso-Seq method is flexible and allows different types of RNA
to be sequenced. Alternatively, mRNAs can be selected by polyA enrichment. The first-strand cDNA is amplified with oligo(dT) to enrich RNAs with a polyA tail, including mRNAs and long noncoding RNAs (IncRNAs) for further analysis.

For parallel analysis of RNA samples derived from various tissues, barcode for each sample with unique sequences is alternatively used. For instance, multiplex sequencing was performed to construct a maize transcriptome library from various tissues [23]. However, barcoding samples is not always desired because sequencing efficiency may be reduced by the barcode sequence.

### 3.3 Size partitioning

Size selection for size partitioning, which is the most commonly used method to avoid over-representation of smaller transcripts in sequencing data, allows for more even representation of cDNA of different size ranges, since smaller fragments may load preferentially on the sequencer. Furthermore, the process of second fractionation is recommended to remove any smaller fractions from the first size selection. To enhance PCR amplification, different sizes of the cDNA libraries including <1, 1–2, 2–3, and 3–6 kb are generally constructed to maximally recover transcript diversity and sequence. However, such size selection may bring about missing small size transcripts less than approximately 1 kb. This problem appears to result from technical limitation by size selection in the construction of mRNA sequencing libraries. This can get solved by combinatorial use with short-read RNA-Seq data that are very effective for transcriptome coverage, especially small size of transcripts.
3.4 Library preparation and sequencing

Double-stranded cDNA is not enough for SMRTbell library construction following size selection. PacBio suggests PCR amplification using the KAPA HiFi Enzyme [24] with about 10 cycles. Then, a circularized molecule called a SMRTbell template is transformed from the amplified cDNAs by the SMRTbell Template Prep kit. After the step is completed, the library is ready to be loaded into a SMRT cell and subjected to sequencing on the PacBio platform. There is a compromise between SMRT cell numbers and the sequencing cost. In general, the Iso-Seq protocol recommends 8–50 SMRT cells to retrieve diversity in a tissue.

4. Building full-length transcripts in a genome

Error correction of the raw reads is necessary to improve the assembly quality of the FL transcripts. PacBio provides the Iso-Seq analysis software to perform the procedure by iterative clustering for error correction (ICE) and the Quiver algorithm (https://www.pacb.com/applications/rna-sequencing). Then, various analysis approaches can be applied to overcome the limitation of Iso-Seq, improve assembly quality, and evaluate the quality assessment of the unigenes.

The Iso-Seq raw reads are usually called polymerase reads or continuous long reads (CLRs) and have an average length of 10 kb (Figure 1). Considering the average length of a transcript is 1–2 kb, the same copies of the inserts are contained in a single polymerase that could be split into several subreads by removing the adaptor sequences by PacBio SMRT link analysis [20]. The circular consensus sequences or ROIs are generated from several subreads. The full-length non-chimeric read (FLNC) is defined not only when the polyA tail signal preceding the 30-primer is present, but also when both 50- and 30-cDNA primers are present. To enhance consensus accuracy and remove the redundancy of FLNC without any additional sequence data, ICE and Quiver can be applied [20]. The Iso-Seq classify tool is used for classifying the ROIs into full-length nonchimeric and non-full-length reads by identifying the 50 and 30 adapters used in library preparation. Then, the Iso-Seq cluster tool is used for clustering all the full-length reads, and the consensus sequences produced by the cluster tool are polished using the non-full-length reads through the Quiver algorithm [25]. Additionally, the CD-HIT program [26] is likely to be helpful to cluster the high and low quiver consensus isoforms from ROIs with high sequence identity threshold (i.e. 0.98–0.99) [20, 21].

Iso-Seq reads present a disadvantage with the high frequency of errors of nucleotide indels and mismatches. Thus, the procedure of correcting InDels and mismatches is performed via alignment with reference genomes [27]. To overcome this, a viable alternative approach is to integrate short reads with long reads via hybrid sequencing. For instance, RNA samples prepared from the same samples are sequenced by both PacBio and Illumina HiSeq. The short reads from the Illumina HiSeq are applied to correct the transcript isoforms using LoRDEC tool v0.6 [28]. Then, the corrected isoform sequences are aligned against a reference genome by GMAP aligner [29]. The following analyses are recommended to exclude the sequences with multiple and chimeric alignments. To assess quality of the unigenes, some software such as CEGMA [30] and BUSCO [31] can be applied [20, 21, 32, 33]. The percentages of the transcripts that fully and partially aligned to the conserved proteins are calculated.

FL or longer transcriptome data have been mostly published from large complex or uncharacterized genomes of plant species (Table 1). Although deep short-read transcriptome sequencing (i.e., RNA-Seq) have accumulated over recent year, they are likely to generate low-quality transcripts with a small portion of FL transcripts, prohibiting accurate transcript reconstruction and leading incorrect annotation.
<table>
<thead>
<tr>
<th>Species</th>
<th>No. of transcripts</th>
<th>Mean length (bp)</th>
<th>Identification of novel gene isoforms</th>
<th>Isoform annotation</th>
<th>Alternative splicing events</th>
<th>Gene prediction</th>
<th>Other</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panax ginseng</td>
<td>135,317</td>
<td>3178</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>—</td>
<td>[20]</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>91,881</td>
<td>2388</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>[45]</td>
</tr>
<tr>
<td>Zea mays B73</td>
<td>111,151</td>
<td>3372</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Fusion transcripts</td>
<td>[23]</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>27,860</td>
<td>1042 (full-length ROI)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>[27]</td>
</tr>
<tr>
<td>Trifolium pratense</td>
<td>206,465</td>
<td>2789</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>—</td>
<td>[34]</td>
</tr>
<tr>
<td>Zea mays W64A</td>
<td>166,693</td>
<td>2715</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>Fusion transcripts</td>
<td>[35]</td>
</tr>
<tr>
<td>Allium sativum</td>
<td>36,321</td>
<td>1500</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>—</td>
<td>Association study</td>
<td>[36]</td>
</tr>
<tr>
<td>Populus (P. deltoides × P. euramerica cv. ‘Nanlin895’ )</td>
<td>87,150</td>
<td>2417</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>Fusion transcripts</td>
<td>[37]</td>
</tr>
<tr>
<td>Coffea arabica</td>
<td>95,995</td>
<td>3236</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>—</td>
<td>[38]</td>
</tr>
</tbody>
</table>

Table 1. Transcriptomics studies in plants by isoform sequencing.
vUnlike RNA-Seq data, Iso-Seq data, which are derived from various tissues as many as possible, harbor a large portion of unique FL transcripts. For example, Wang et al. [23] reported that maize yielded 111,151 non-redundant FL transcript isoforms, corresponding to approximately 26,946 genes. In addition, genome coverage of Iso-Seq data is achieved near-saturation. Ultimately, cost-effective long-read transcriptome sequencing can be the gold standard for transcript completeness, characterization of transcriptome, and draft genome annotation. To identify trait-associated transcripts in species for which a reference genome is lacking (i.e., garlic), this approach was used as a reference sequence for scoring the variation in both SNP and expression level in the population [36], reporting the characterization of transcripts (lncRNAs) associated with garlic clove shape traits.

5. Improvement of the efficiency of functional gene prediction and annotation

Completeness of assembled transcripts is closely related to the efficiency of functional gene prediction or annotation, especially in the absence of reference genome information. Because of such advantage, Iso-Seq has been applied in a variety of species [20–22, 32, 33]. In addition, optimized training and prediction settings on the basis of short- and long-read transcriptome data in gene prediction results in increased their sensitivity and precision [39]. In particular, the method is helpful for obtaining comprehensive gene sets for newly sequenced genomes of non-model eukaryotes [39].

To identify the protein coding potential of transcripts, Transdecoder (https://transdecoder.github.io) is generally applied [20, 21, 32, 40]. For example, even though the number of transcripts using Iso-Seq is much smaller than those de nova assembled in previous RNA-seq studies, the transcripts from Iso-Seq show high efficiency in recovering full-length transcripts. ESTScan [41], in addition to Transdecoder, is used to predict coding DNA sequences (CDSs) unless isoforms are annotated in the databases. For example, in the study of Halogeton glomeratus [42], the CDS prediction ratio of transcripts using Iso-Seq (95.09%) is much higher than that of transcripts using Illumina RNA-Seq data (66.86%).

For functional annotation, isoform sequences are used as queries for sequence homology searches in Blast, Blast2GO [43], and InterProScan5 [44] to identify functional annotation terms from the nonredundant protein (NR), non-redundant nucleotide (NT), Gene Ontology (GO), Clusters of Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt, and Interpro databases. For example, when the RNA-Seq data of H. glomeratus were re-annotated with Iso-Seq transcriptome data, the length distribution, functional annotation, and coding sequence quantity of the Iso-Seq transcripts were significantly improved [42]. In particular, with respect to the species distribution of the annotation from the NR database, 98.31% of the annotated isoforms showed the highest similarity to sequences from the three most prevalent species. In addition, Illumina RNA-Seq data were highly mapped to the Iso-Seq transcripts (unigenes). This suggests that long-read, full-length or partial-unigene data with high-quality assemblies are invaluable resources as transcriptomic references in a genome and can be used for comparative analyses in closely related medicinal plants.

6. Conclusion

Transcriptome data generated by Iso-Seq generate longer and improved unigenes with a high level of assembly completeness and gene annotation, enabling a
comprehensive view of the transcriptome. In particular, compared with conventional methods, long-read transcriptome sequencing seems to improve misassembly rate and unreliable gene annotation, thus enabling to elucidate the function of genes associated with traits of interest as well as novel transcripts. A hybrid approach that combines isoform sequencing with full-length transcripts and RNA-Seq capable of fixing sequence error and quantifying gene expression is the optimal solution to study transcriptomes for improving completeness of transcripts, data coverage, and gene annotation.

Acknowledgements

This work was supported by grants from the National Agricultural Genome Center (project No. PJ01349002), Rural Development Administration, Republic of Korea.

Conflict of interest

The author declares no conflict of interest to disclose.

Author details

Dong Jin Lee and Chang Pyo Hong*
Theragen Etex Bio Institute, Suwon, Republic of Korea

*Address all correspondence to: changpyo.hong@theragenetex.com

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
References


[22] Chen SY, Deng F, Jia X, Li C, Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Scientific Reports. 2017;7(1):7648. DOI: 10.1038/s41598-017-08138-z


Section 4

Transcriptome Analysis in Plants
Section 4

Transcriptome Analysis in Plants
Chapter 4

Plant Comparative Transcriptomics Reveals Functional Mechanisms and Gene Regulatory Networks Involved in Anther Development and Male Sterility

Xiangyuan Wan and Ziwen Li

Abstract

Gene transcription and transcriptional regulation are crucial biological processes in all cellular life. Through the next-generation sequencing (NGS) technology, transcriptome data from different tissues and developmental stages can be easily obtained, which provides us a powerful tool to reveal the transcriptional landscape of investigated tissue(s) at special developmental stage(s). Anther development is an important process not only for sexual plant reproduction but also for genic male sterility (GMS) used in agriculture production. Plant comparative transcriptomics has been widely used to uncover molecular mechanism of GMS. Here, we focused on researches of anther developmental process and plant GMS genes by using comparative transcriptomics method. In detail, the contents include the following: (1) we described the commonly used flowchart in comparative transcriptomics; (2) we summarized the comparative strategies used to analyze transcriptome data; (3) we presented a case study on a maize GMS gene, \( \text{ZmMs33} \); (4) we described the methods and results previously reported on gene co-expression and gene regulatory networks; (5) we presented the workflow of a case study on gene regulatory network reconstruction. The further development of comparative transcriptomics will provide us more powerful theoretical and application tools to investigate molecular mechanism underlying anther development and plant male sterility.

Keywords: plant comparative transcriptomics, gene regulatory network, anther development, genic male sterility, molecular mechanism

1. Introduction

Gene transcription is an important biological process by which genetic information stored within DNA molecules is transmitted to RNA molecules according to the "genetic central dogma" in molecular biology [1]. After completion of the human genome project, the researchers began to reveal the transcriptional
Chapter 4

Plant Comparative Transcriptomics Reveals Functional Mechanisms and Gene Regulatory Networks Involved in Anther Development and Male Sterility

Xiangyuan Wan and Ziwen Li

Abstract

Gene transcription and transcriptional regulation are crucial biological processes in all cellular life. Through the next-generation sequencing (NGS) technology, transcriptome data from different tissues and developmental stages can be easily obtained, which provides us a powerful tool to reveal the transcriptional landscape of investigated tissue(s) at special developmental stage(s). Anther development is an important process not only for sexual plant reproduction but also for genic male sterility (GMS) used in agriculture production. Plant comparative transcriptomics has been widely used to uncover molecular mechanism of GMS. Here, we focused on researches of anther developmental process and plant GMS genes by using comparative transcriptomics method. In detail, the contents include the following: (1) we described the commonly used flowchart in comparative transcriptomics; (2) we summarized the comparative strategies used to analyze transcriptome data; (3) we presented a case study on a maize GMS gene, ZmA33; (4) we described the methods and results previously reported on gene co-expression and gene regulatory networks; (5) we presented the workflow of a case study on gene regulatory network reconstruction. The further development of comparative transcriptomics will provide us more powerful theoretical and application tools to investigate molecular mechanism underlying anther development and plant male sterility.

Keywords: plant comparative transcriptomics, gene regulatory network, anther development, genic male sterility, molecular mechanism

1. Introduction

Gene transcription is an important biological process by which genetic information stored within DNA molecules is transmitted to RNA molecules according to the “genetic central dogma” in molecular biology [1]. After completion of the human genome project, the researchers began to reveal the transcriptional
landscape of all genes in a genome to further investigate the functional mechanisms underlying phenotypic variations at a genome-wide transcriptional level. Therefore, biological studies on high-throughput omics data run from the genomic level into the transcriptomic level. Transcriptome data includes biological information of gene transcriptional activities in a certain cell, a tissue, or an individual (a population of cells) and even in a pool of samples under a certain developmental stage, an environmental condition, or an experimental treatment. Compared with other omics data (e.g., data of genome, epigenome, proteome, metabolome, or phenome), the primary characteristic of transcriptome data is that it includes temporal–spatial bioinformation affected by diverse developmental stages, tissue types, and internal/external environment events. Therefore, transcriptome data is more complex than genome data.

Transcriptomic studies usually focus on the transcriptional content and gene regulations in a genome. Gene expression microarray (GEM) is an early developed but still-utilized biotechnology by which the genome-wide transcription information can be obtained for genome-sequenced or transcriptional loci available species. In 1995, Schena et al. monitored expression levels of 48 genes by GEM in *Arabidopsis thaliana* [2], and then GEM was gradually and widely used for the estimation of gene expression levels. Until 2013, the amount of transcripts monitored by one microarray had been reached to more than 285,000 in human transcriptomics studies (the human transcriptome array). GEM is a hybrid-based method, while the sequencing-based method has been developed much faster and became one of the most commonly used biotechnologies in scientific studies and applications related to disease diagnosis [3]. Serial analysis of gene expression (SAGE) proposed by Velculescu et al. [4] and massively parallel signature sequencing (MPSS) reported by Brenner et al. [5] are two earlier developed sequencing-based methods to estimate the transcription information at a genome level. Nowadays, the majority of transcriptome data are generated by the NGS-based RNA sequencing (RNA-seq). RNA-seq technology combining with the following developed comparative transcriptomics analysis flowchart that is mainly based on digital gene expression profile (DGEP) is a commonly used research strategy in biological studies at molecular and genomic levels.

Anther is an important organ in sexual plant reproduction. Anther development is a dynamic process from the identity of the stamen to the production of mature pollen grains. During this period, two-thirds of protein-coding genes are transcribed, and more than 6% of them are anther specific (a reanalyzed result based on [6]). Thus, the anther transcriptome is specific and complex compared with transcriptomes of other plant organs. Plant comparative transcriptomics is an effective strategy used to investigate the molecular mechanism underlying anther developmental process. The comparative method based on anther transcriptomes can be performed between different genotypes, different developmental stages, different types of anther cells, and different biotic or abiotic treatments and even between different plant species. Consequently, differentially expressed genes (DEGs) are identified from above comparisons. Based on the comparison results, functionally important coding genes and noncoding transcripts including long noncoding RNAs (lncRNAs), microRNAs (miRNAs), and other small RNAs could be uncovered. However, the goal of plant comparative transcriptomics is not only to identify DEGs but also to reconstruct gene regulatory relationships of the upstream regulators and the downstream regulated targets of the investigated genes. In this review, based on anther transcriptomes, we first summarized the research workflow commonly used in the experimental design and data analyses in plant transcriptomics studies, and then we described several types of comparison strategies in comparative transcriptomics using anther transcriptome data as the analyzed example. In the following
section, we generally discussed gene regulatory and co-expression networks used to investigate the molecular foundation of developing anther in a network-based perspective. Additionally, we described two case studies in our laboratory to explain the detailed analysis processes and applications of comparative transcriptomics in plant GMS gene studies.

2. Comparative analysis using transcriptome data

In comparative transcriptomics, the commonly used pipeline to identify potential functional genes and to reveal the gene functions, as well as to investigate the regulatory relationships between these genes, includes five aspects. They are data preparation, DGEP analysis, DEG analysis, gene set enrichment (GSE) analysis, and gene regulatory network (GRN) analysis, respectively (Figure 1). These five aspects are closely connected in the whole pipeline, and the corresponding analyses mainly depend on data management skills in bioinformatics.

The basic application of comparative transcriptomics is to obtain a transcriptional landscape of the investigated biological sample. It is composed of not only the estimated transcription levels of annotated transcribed loci along the genomes (the known genomic loci with reported or predicted transcription abilities) but also the identification of novel transcribed loci (the stably transcribed loci not annotated or identified in previous studies). More importantly, in current biological studies, transcribed loci identified by researchers include not only the protein-coding genes but also IncRNAs and other noncoding RNAs. Both GEM and RNA-seq technologies can be used to uncover the genome-wide profiles of transcription levels of annotated genes. However, the identification of novel transcribed loci can be only effectively performed by RNA-seq method and the following DGEP analysis. This is one reason why RNA-seq is more commonly used in transcriptomics studies. Moreover, GEM method depends on hybridization probes that are designed based on known whole genome sequence or an appreciable set of sequenced transcripts.
(e.g., expressed sequence tags) of the investigated species, which restrict its application on some species without whole genome information or sequence resource. On the contrary, the sequencing-based method of RNA-seq can be applicable for species without sequenced genomes. This is another reason for the popularity of RNA-seq. In genome available species, RNA-seq data should be firstly mapped to the reference genome (Figure 1).

A gene with its transcription levels significantly different between two groups of samples is defined as a DEG under a certain comparison condition (Figure 1). It is notable that the concept of DEG specially represents the expression changes of protein-coding genes at the earlier stages of expression data analysis. However, along with the rapid development of molecular biology and the deeper understanding on the functional element on the genome, the concept of DEG has been expanded to noncoding transcripts, for example, the differentially expressed (DE) miRNA and the DE lncRNA. Furthermore, if both coding and noncoding transcripts are considered in the comparative analysis of transcriptome data, transcriptional alterations between control and treated samples should be defined as DE transcribed loci or DE loci. Thus, DE loci is a broad concept used to describe transcriptional alterations of genetic element. There are several strategies for comparing transcriptomes from different research subjects to identify DE loci (described in Section 3, “Plant comparative transcriptomics in anther”).

Identified DEG set or DE loci should be appropriately annotated with functional descriptions to determine which biological process or pathway these DEGs are involved in. In comparative transcriptomics, this step is a critical bridge linking transcriptional changes to gene functions and even gene regulation networks. Two commonly utilized methods to annotate DEGs consist of the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway gene enrichment analyses. Both of them belong to GSE analysis (Figure 1). The GO database includes tens of thousands of GO terms, and each GO term contains several genes with the same biological function. Each gene has three functional aspects, including molecular function (the molecular activities of gene products), cellular component (the cellular locations of functional gene products), and biological process (the gene products’ molecular functions with biological process). GSE analysis based on GO database provides some basic functional descriptions for the investigated DEG set. KEGG analysis is a pathway-based enrichment method. The KEGG database has accumulated hundreds of metabolism pathways in plants, animals, and other species. Thus, KEGG analysis can reveal significant pathways the DEGs participated in. GO-based methods can annotate more genes than KEGG-based method, as the GO terms are more flexible and include a larger number of genes. On the other hand, because most metabolic pathways are conserved across species and more significant in biological processes, annotated results obtained from KEGG-based method may be more conserved and stable. In comparative transcriptomics, GO- and KEGG-based analyses are together utilized in gene function studies.

The locations of transcribed loci on the genome, their transcription levels, and the changed expression can be identified through comparative transcriptomics analysis. The detected DEG set represents a functional gene set related to the function of investigated gene, the phenotype variation, the stress resistance ability, or the development process. Furthermore, gene regulation relationships are the underlining molecular mechanism of altered transcriptomes, and novel gene regulatory networks could be uncovered by comparative transcriptomics analysis (Figure 1). Several types of gene regulatory relationships and the reconstructions of gene regulatory networks based on plant comparative transcriptomics are described and discussed in Section 5 (“Gene co-expression and regulatory networks reconstructed by comparative transcriptomics method”).
3. Plant comparative transcriptomics in anther

One of the major subjects of modern molecular biology is to uncover the functions of genes in the genome and reveal the molecular mechanism of phenotypic variation. Gene transcription levels and their changes in different conditions are important information that can reflect the functions and transcriptional regulation relationships of investigated genes. How to estimate the transcription levels of genes and how to obtain the transcriptional landscape of a genome are two major subjects in biological studies on gene expression. DGEP and DEG analyses are powerful tools to solve these questions. In DEG analysis, according to the scientific or application questions, the comparison strategies between investigated biological samples are classified into six types including (1) different genotypes, (2) different developmental stages, (3) different tissues, (4) different cell types, (5) different treatments, and (6) different species (Figure 2). Here, as we mainly focus on comparative transcriptomics analysis on the developmental anther tissues and the interspecies analysis on anther transcriptome data being rare, the third and sixth types will not be discussed.

3.1 Different genotypes

There are two types of genotype-based transcriptome data between wild type (WT) and mutant lines in GMS studies, which are based on whether the causal mutation is known or not (Table 1). For transcriptomes of male sterility (MS) lines with known causal mutations, the comparison of transcriptomes between WT and MS lines will identify many DEGs associated with the function loss or expression change of the investigated mutation locus. If the causal mutation has not been identified from the MS line, comparative transcriptomics analyses will provide the researchers important results related to the unsettled genetic difference, such as how many genes are changed in expression levels in the MS lines and what the functions of these genes are, even though the causal mutation candidates can be inferred from these genes if the researchers have primary mapping results.

3.2 Different developmental stages

The phenotypic differences among tissues and organs (e.g., root, leaf, and flower in plant) due to their differences of transcriptome landscape are well known.

Figure 2. Comparative transcriptomics strategies.
Transcriptome Analysis

<table>
<thead>
<tr>
<th>Plants</th>
<th>MS gene or MS line</th>
<th>Method</th>
<th>Tissue</th>
<th>Data source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>ROXY1 and ROXY2</td>
<td>Microarray</td>
<td>Young inflorescences</td>
<td>SD</td>
<td>[7]</td>
</tr>
<tr>
<td></td>
<td>AMS</td>
<td>Microarray</td>
<td>Anthers</td>
<td>GSE18225</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>AMS</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>AMS</td>
<td>Microarray</td>
<td>Floral buds</td>
<td>SD</td>
<td>[10]</td>
</tr>
<tr>
<td></td>
<td>EMSI</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>EMSI</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[11]</td>
</tr>
<tr>
<td></td>
<td>Ms1</td>
<td>Microarray</td>
<td>Young closed buds</td>
<td>SD</td>
<td>[12]</td>
</tr>
<tr>
<td></td>
<td>Ms1</td>
<td>Microarray</td>
<td>Floral buds</td>
<td>GSE8864</td>
<td>[13]</td>
</tr>
<tr>
<td></td>
<td>ICE1</td>
<td>RNA-seq</td>
<td>Anthers</td>
<td>GSE107260</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>DYT1</td>
<td>Microarray</td>
<td>Anthers</td>
<td>GSE18225</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>CDM1</td>
<td>Microarray</td>
<td>Young floral buds</td>
<td>GSE55799</td>
<td>[15]</td>
</tr>
<tr>
<td></td>
<td>TEK</td>
<td>Microarray</td>
<td>Closed floral buds</td>
<td>GSE56497</td>
<td>[16]</td>
</tr>
<tr>
<td></td>
<td>bHLH010, bHLH08, bHLH091</td>
<td>RNA-seq</td>
<td>Anthers</td>
<td>SRS838170, SRS838173</td>
<td>[17]</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>PTC1</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[18]</td>
</tr>
<tr>
<td></td>
<td>UDT1</td>
<td>Microarray</td>
<td>Anthers</td>
<td>GSE2619</td>
<td>[19]</td>
</tr>
<tr>
<td></td>
<td>OsGAMYB</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[20]</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>Microarray</td>
<td>Spikelets</td>
<td>SD</td>
<td>[21]</td>
</tr>
<tr>
<td></td>
<td>MADS3</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[22]</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>Ms23</td>
<td>RNA-seq</td>
<td>Anthers</td>
<td>GSE90849</td>
<td>[23]</td>
</tr>
<tr>
<td></td>
<td>Ms32</td>
<td>Microarray</td>
<td>Anthers</td>
<td>GSE90968</td>
<td>[23]</td>
</tr>
<tr>
<td></td>
<td>MAC1</td>
<td>Microarray</td>
<td>Anthers</td>
<td>SD</td>
<td>[24]</td>
</tr>
<tr>
<td><em>Triticum aestivum</em></td>
<td>TaMx1</td>
<td>RNA-seq</td>
<td>Anthers</td>
<td>SRP113349</td>
<td>[25]</td>
</tr>
<tr>
<td></td>
<td>TaMx2</td>
<td>RNA-seq</td>
<td>Anthers</td>
<td>SRP092366</td>
<td>[26]</td>
</tr>
<tr>
<td><em>Solanum lycopersicum</em></td>
<td>mi1035</td>
<td>RNA-seq</td>
<td>Floral buds</td>
<td>SD</td>
<td>[27]</td>
</tr>
<tr>
<td></td>
<td>MS line 7B-1</td>
<td>RNA-seq</td>
<td>Anthers</td>
<td>GSE85859</td>
<td>[28]</td>
</tr>
<tr>
<td><em>Brassica napus</em></td>
<td>MS line WSLA</td>
<td>RNA-seq</td>
<td>Young flower buds</td>
<td>SRR2192464, SRR2192489</td>
<td>[29]</td>
</tr>
<tr>
<td></td>
<td>MS line SP25</td>
<td>RNA-seq</td>
<td>Young flower buds</td>
<td>GSE69638</td>
<td>[30]</td>
</tr>
<tr>
<td></td>
<td>MS line TE5A</td>
<td>RNA-seq</td>
<td>Young flower buds</td>
<td>SRP068170</td>
<td>[31]</td>
</tr>
<tr>
<td><em>Citrullus lanatus</em></td>
<td>MS line DAH3615-MS</td>
<td>RNA-seq</td>
<td>Floral buds and flowers</td>
<td>GSE69073</td>
<td>[32]</td>
</tr>
</tbody>
</table>

*SD* indicates the raw data is unavailable, while the up- and downregulated genes are listed in the supplemental data (SD) in references cited.

**Table 1.**
Published studies on anther transcriptome data between WT and MS lines.
Furthermore, it is a developmental process for most types of plant organs from the organ identity (e.g., meristematic cells) to the final mature organ. Thus, how to reveal the dynamic changes of gene transcription levels and how to explain the morphological alterations regulated by gene expression changes are important tasks in plant comparative transcriptomics studies.

Meiosis is an important step in gametophyte generation process and sexual plant reproduction. Morphologic changes during cell meiosis process have been well described by cellular level investigations, while the molecular level alterations and their corresponding gene regulatory networks are not well understood. Plant transcriptomes are a powerful dataset to estimate the gene expression changes and infer the regulatory roles of key genes. Based on GEM technology, Ma et al. investigated maize anther transcriptomes during seven developmental stages and found that transcriptomes during meiosis stages exhibited the lowest complexity [33]. Hollender et al. surveyed the gene transcription profiles of anther of woodland strawberry (Fragaria vesca) from developmental stages 7–12 and identified numerous F-Box genes induced in transcription levels at meiosis stage [34]. Besides, tapetum is the inner cell layer of anther with important functions in anther development and gametocyte maturation. The generation, development, and degradation of tapetum are fine regulated during the anther development, while the regulatory framework and the details are far from complete. Yue et al. identified 243 DEG and 108 stage-specific genes during four anther developmental stages in Hamelia patens [35]. Chen et al. investigated the expression of genes involving in tapetum development of male floral bud during eight developmental stages in Populus tomentosa [36]. Thus, anther transcriptome data during different developmental stages provide valuable data sources for anther development studies. By the combination of comparative transcriptomics and bioinformatics analyses, more key functional genes and the underlying regulatory mechanisms for anther development will be further revealed.

3.3 Different types of anther cells

The cytological structure of anther consists of four cell layers, including the epidermis, endothecium, middle layer, and tapetum, and the archesporial cells are directly surrounded by the tapetum. Thus, the transcriptome data of a whole anther tissue is a mixed gene expression data from diverse cell types with different functions in the anther development process. It is necessary to obtain transcriptional dynamics from different cell layers separately to investigate anther development and the underlying molecular mechanism at a cell type-specific level. Several studies have identified cell layer-specifically expressed genes (e.g., tapetum cells or microgametes). Ma et al. identified 104 MS-related and non-pollen expressed genes most specifically expressed in tapetum by comparative transcriptomics analysis on four diverse MS lines in Brassica oleracea [37]. The other way to obtain cell layer-specific transcriptome in anther is firstly separating the investigated cell layer by laser capture microdissection (LCM) technology and then performing RNA-seq or GEM experiment on the separated samples. This strategy has been successfully used in rice, maize, and woodland strawberry to identify the tapetum- or microgamete-specifically expressed genes and their expression dynamics [34, 38, 39]. A recent published research has investigated maize male meiosis using single-cell RNA sequencing (scRNA-seq) technology on pre-meiotic and meiotic cells from maize anthers, which greatly promoted studies on plant anther scRNA-seq [40]. The comparative studies on transcriptomic dynamics between different types of cells facilitate the deeper understanding of functions of specific cell layers on anther development.
3.4 Different treatments

At the reproductive stage, plant is more sensitive to external environment conditions. The abiotic stresses, such as high temperature, drought, and cold and freezing stresses, will critically affect the developmental process of anther and pollen in flowering plants. Though there have been numerous studies on stress resistance and response in plant, the regulatory pathways of stress response and their cross talk at molecular level should be further investigated for anther development. Additionally, more effective stress-resistant genes should be identified for the purpose of crop improvement. Plant comparative transcriptomics between normal and stress-treated plants provide a wide insight into the stress response mechanisms of plant during sexual reproductive stage. Zhang et al. investigated the genome-wide transcriptional changes of rice panicle under heat treatment (40°C) and found thousands of DEGs participating in transcriptional regulation, transport, cellular homeostasis, and stress response [41]. Studies on photosensitive or thermosensitive GMS lines can also reveal a lot of genes responding to environmental changes.

4. A case study: revealing the molecular functions of a MS gene, ZmMs33, by comparative transcriptomics

The discoveries of genes that play key roles in the development of maize anther provide important genetic resources for the utilization of heterosis in maize. Analysis of functional mechanism of GMS genes can effectively promote researches on anther development biology and deepen our understanding of molecular mechanism controlling sexual plant reproduction [42]. There are several published case studies containing comparative transcriptomics analysis on maize GMS genes in our laboratory, including ZmMs7 [43], ZmMs20 [44], ZmMs30 [45], and ZmMs33 [46, 47]. We used comparative transcriptomics analysis based on developmental anthers of ZmMs33 wild type and ms33–6038 mutant to analyze the transcription changes corresponding to male sterility phenotype and to further investigate the underlying molecular mechanisms of GMS regulated by ZmMs33 gene.

This ms33–6038 mutant is complete male sterility and displays small and pale-yellow anthers (Figure 3A). Transmission electron microscope (TEM) observation and dynamic scanning electron microscopy (SEM) analysis were performed to analyze the phenotypic alteration of anther wall layers, microspores, Ubisch bodies, and exine between wild type and ms33–6038 mutant during anther developmental stages (Figure 3A–C). Maize Zm00001d007714 was identified as ZmMs33 via a map-based cloning approach (Figure 3D). ZmMs33 encodes an esterase that belongs to gene family of glycerol-3-phosphate acyltransferase (GPAT) in maize. To further confirm gene function of Zm00001d007714, a CRISPR/Cas9 system was used to generate targeted knockout lines. Three types of T0-generation maize plants homozygous for null alleles of Zm00001d007714 were observed to be complete male sterility (Figure 3E), suggesting that function loss of Zm00001d007714 is the causal mutation for male sterile phenotype of the ms33 mutant.

Subsequently, RNA-seq was performed using anther tissues during developmental stages 5–9 to obtain a comprehensive transcriptional profile of WT and ms33–6038. Three biological samples were collected at each developmental stage for sequencing. After data preparation and transcription level estimation, we compared similarities of transcriptional profiles of protein-coding genes by principal component analysis (PCA) (Figure 3F) and found good repeatability among three biological repeats.
3.4 Different treatments
At the reproductive stage, the plant is more sensitive to external environmental conditions. Abiotic stresses, such as high temperature, drought, and cold and freezing stresses, will critically affect the developmental process of anther and pollen in flowering plants. Though there have been numerous studies on stress resistance and response in plants, the regulatory pathways of stress response and their cross talk at the molecular level should be further investigated for anther development. Additionally, more effective stress-resistant genes should be identified for the purpose of crop improvement. Plant comparative transcriptomics between normal and stress-treated plants provide a wide insight into the stress response mechanisms of plants during sexual reproductive stages. Zhang et al. investigated the genome-wide transcriptional changes of rice panicles under heat treatment (40°C) and found thousands of DEGs participating in transcriptional regulation, transport, cellular homeostasis, and stress response [41]. Studies on photosensitive or thermosensitive GMS lines can also reveal a lot of genes responding to environmental changes.

4. A case study: revealing the molecular functions of a MS gene, ZmMs33
The discoveries of genes that play key roles in the development of maize anther provide important genetic resources for the utilization of heterosis in maize. Analysis of functional mechanisms of GMS genes can effectively promote researches on anther development biology and deepen our understanding of molecular mechanisms controlling sexual plant reproduction [42]. There are several published case studies containing comparative transcriptomics analysis on maize GMS genes, including ZmMs7 [43], ZmMs20 [44], ZmMs30 [45], and ZmMs33 [46, 47]. We used comparative transcriptomics analysis based on developmental anthers of ZmMs33 wild type and ms33–6038 mutant to analyze the transcriptional changes corresponding to male sterility phenotype and to further investigate the underlying molecular mechanisms of GMS regulated by ZmMs33 gene.

This ms33–6038 mutant is complete male sterility and displays small and pale-yellow anthers (Figure 3A). Transmission electron microscope (TEM) observation and dynamic scanning electron microscopy (SEM) analysis were performed to analyze the phenotypic alteration of anther wall layers, microspores, Ubisch bodies, and exine between wild type and ms33–6038 mutant during anther developmental stages (Figure 3A–C).

Maize Zm00001d007714 was identified as ZmMs33 via a map-based cloning approach (Figure 3D). ZmMs33 encodes an esterase that belongs to the gene family of glycerol-3-phosphate acyltransferase (GPAT) in maize. To further confirm gene function of Zm00001d007714, a CRISPR/Cas9 system was used to generate targeted knockout lines. Three types of T0-generation maize plants homozygous for null alleles of Zm00001d007714 were observed to be complete male sterility (Figure 3E), suggesting that function loss of Zm00001d007714 is the causal mutation for male sterile phenotype of the ms33 mutant.

Subsequently, RNA-seq was performed using anther tissues during developmental stages 5–9 to obtain a comprehensive transcriptional profile of WT and ms33-6038. Three biological samples were collected at each developmental stage for sequencing. After data preparation and transcription level estimation, we compared similarities of transcriptional profiles of protein-coding genes by principal component analysis (PCA) (Figure 3F) and found good repeatability among three biological repeats.

Finally, we identified DEGs between WT and mutant and between adjacent developmental stages, separately. We found that the amount of DEGs between WT and mutant at stages 5–7 was significantly smaller in magnitude than that at
stages 8a–9 (Figure 3G), indicating that ms33 mutant transcriptomes are significantly divergent from WT starting from stage 8a. The transcriptome landscapes of WT were similar to those of ms33 mutant at stages 5–7. Besides, DEG amounts were various between adjacent developmental stages. It is worth noting that the DEG amount between WT and mutant exceeded that between adjacent stages from stage 8a–9. This result implied that the transcriptomes were significantly changed at the later three stages. Therefore, we compared the transcriptomes between genotypes at the former three and the later three stages, separately. In contrast to a limited number of DEGs (only two genes) shared by the former three stages, there were thousands of shared DEGs at the later three stages. GSE analysis based on KEGG database suggested that the upregulated gene set was firstly enriched in the function of biosynthesis of secondary metabolites, while the downregulated genes were significantly related to the photosynthesis process. This pathway enrichment analysis partly represents the alterations in metabolisms and physiological activities closely associated with the transcriptional changes caused by function defect of ms33.

5. Gene co-expression and regulatory networks reconstructed by comparative transcriptomics method

Though DEGs are mainly identified by pairwise comparisons between transcriptomes of tissues, stages, or treatment conditions and can reflect most of the transcriptional changes between two sets of samples, these transcriptional alterations are not sufficient to explain the detailed molecular mechanism underlying tissue-specific development processes and stress–resistant pathways. Moreover, the molecular functions of genes act under GRNs. All the biological processes of growth, development, stress response, and reproduction are regulated by GRNs. The prediction of gene regulatory relationships and the reconstruction of the GRNs by using the transcriptome data are also the major aims in transcriptomics studies, except for the DGEP and DEG analyses.

5.1 Gene co-expression analysis

Function-related genes tend to co-express in a cell, either to form a complex or to involve in the same biological pathway. Thus, the similar pattern of gene expressions can be used as an indicator to predict gene functions. Gene co-expression (GCE) analysis is a powerful tool to discover important functional genes in biological processes including anther development. A relatively early study identified two functional GMS genes, POLYKETIDE SYNTHASE A (PKSA) and PKSB, through detecting co-expressed genes with ACOS5, a GMS gene belonging to fatty acyl-CoA synthetase gene family, based on microarray data in A. thaliana [48]. Similarly, ABORTED MICROSPORES (AMS) gene was reported participating in the pollen wall formation in rice by the analyses of 98 co-expressed genes with AMS in flower development [49]. GCE analysis can be also used to investigate the biological functions and the regulatory targets of a gene. This genome-wide analysis on GCE networks has been performed based on microarray data from A. thaliana anther tissues, and 254 complete GCE groups containing 10,513 anther-transcribed genes were revealed [50]. Another microarray-based GCE network was reconstructed in A. thaliana anther by using 10,797 genes expressed in anther/flora, and transcriptional landscape of GMS mutant was included in the stable examination of this newly constructed network [51]. In rice, microarrays from WT
anther tissue across stages 2–14 and nine GMS lines were integrated to reconstruct a big GCE network containing more than 9000 genes and 0.4 million pairs of co-expression relationships [52].

RNA-seq data-based GCE network analysis was performed in anther when high-throughput sequencing technology was developed. In woodland strawberry, stages 1–12 floral samples dissected by LCM or hand, including stages 6–12 anther tissues, were sequenced by RNA-seq. Gene co-expression network analysis was used to reconstruct GCE networks in strawberry’s flower development, and 23 modules were discovered from the GCE networks including 4584 pollen-specific genes [34]. These genome-wide GCE networks are useful for characterization of genes associated with anther development and floral reproduction.

5.2 TF-encoding gene regulatory network

Genes with their products forming one protein complex, genes encoding transcription factor (TF) and TF target genes, and genes functioning in the same metabolic pathway or stress-resistant process often tend to be co-expressed in a cell. Therefore, the expression-associated genes in GCE network may be not directly functionally linked. A more accurate and robust gene regulatory network is needed for both the biological function and network researches at molecular and genome levels. One way to improve the gene regulatory network is to introduce gene regulatory types into the network. Several TF gene regulatory networks (TF-GRN), also called as transcriptional regulatory network (TRN), were reconstructed based on expression patterns of TF-encoding genes and TF target genes from transcriptome data. One TF-GRN comprised 19 TFs and their 101 target genes involving in A. thaliana pollen development [53]. Another GRN of early anther development was constructed by interactively analyzing transcriptome data from three GMS lines of TF-encoding gene knockout mutants [9]. In the maize genome, there are 2298 TF-encoding genes identified which belonged to 56 diverse families [54]. A total of 3078 TF-encoding genes belonging to 59 families are predicted in silico analysis in rice genome [55]. These TF databases, combining with increased amount of transcriptome data from mutants of TF-encoding genes and other omics data (e.g., Chip-seq, DAP-seq), provide abundant data for the reconstruction of TF-GRN with increased credibility, applicability, and completeness.

5.3 miRNA target gene regulatory network

Both transcriptional and posttranscriptional regulations are crucial in controlling the normal development and stress-resistant process in cellular life. The miRNA-mediated regulation model on target genes is a well-studied posttranscriptional gene regulation pathway that plays important roles in floral identification and the following development of flower organs [56–58] as well as male fertility [59, 60]. Beyond numerous case studies on functional miRNAs in anther development and GMS genes [61–64], the expression profile of miRNAs and the regulatory networks were investigated to elevate our understanding on the transcriptional regulatory mechanism of miRNAs. GRNs between miRNA and their target genes have been constructed via flower/anther transcriptomics in the model plant species, A. thaliana, and some other plants [65–68]. Furthermore, comparative transcriptomics analysis on small miRNAs has been commonly used as a research method to reveal the transcriptional alterations between fertility and sterility lines in economic and food plant species, such as maize [45], tomato [69], cotton [70, 71], wheat [72, 73], pine [74], lycium [75], watermelon [32], and Brassica campestris [76].
5.4 ceRNA-miRNA regulatory network

It is well known that miRNAs are crucial regulators on gene expressions that control key biological functions including anther development, since miRNA was firstly found in nematodes in 1993 [77]. It is noteworthy that a novel type of gene regulatory model, the competing endogenous RNA (ceRNA) hypothesis, was recently proposed [78]. According to the ceRNA hypothesis, some endogenous transcripts have abilities to adsorb miRNA molecules; subsequently, the expression levels of miRNA target genes can be derepressed [78, 79]. A typical ceRNA in plant, a long noncoding RNA, IPSI, was found in *A. thaliana*. It could completely sponge miRNA *ath-miR399* and indirectly increase the transcription levels of an important gene involved in phosphate homeostasis [80]. The following studies revealed that transcripts of protein-coding genes, pseudogene, transposable elements, simple sequence repeat, and circular RNAs have molecular functions as ceRNAs [79, 81, 82], indicating that the ceRNA-miRNA relationship is an essential gene regulatory mechanism in the growth and development of plants and animals. Consequently, it is necessary to introduce ceRNA regulators into GRN construction. Here, we present our recent study on reconstructing ceRNA regulatory network mainly based on RNA-seq and small RNA-seq transcriptomes from developmental maize anther.

6. A case study: reconstructing ceRNA-miRNA target gene regulatory networks using transcriptome data of maize anther

Here we summarized the research progress of one recently completed research related to the ceRNA-mediated GRN in our laboratory. Generally speaking, this is the first study introducing ceRNA regulation into miRNA target gene regulatory pathway for deeply dissecting the mechanism of anther development and sexual plant reproduction at a network level. This provides a fresh example for GRN research by plant comparative transcriptomics and has dual significance in both theoretical and practical senses. It may also provide new thoughts and strategies for further transcriptome-based GRN studies.

It is well known that gene expressions are controlled by the GRN in cellular life. Newly found regulatory patterns (e.g., miRNA pathway and epigenetic modification) have enhanced our understanding on the GRN. Recently, “ceRNA hypothesis” was proposed as a novel type of gene regulatory relationship and was found to participate in different development and stress response processes of organisms by a number of case studies. However, the network level study on ceRNA regulatory functions is still rare, which limited our deep understanding on the GRN. In addition, studies on the GRN of sexual plant reproduction and male sterility are crucial for both fundamental biological significance and applications in plant hybrid breeding and seed production. We investigated ceRNA-miRNA target gene regulatory network in maize anther developmental process by plant comparative transcriptomics method. Six steps were performed from raw sequencing data preparation to the finally constructed GRN (**Figure 4**). **Firstly**, we performed RNA- and small RNA-seq using anther tissues at three developmental stages from two maize lines to obtain a relative broad transcriptional landscape in anther development and transcribed loci that are stably expressed in maize species. **Secondly**, we identified stably transcribed loci based on the maize reference genome and estimated their transcription levels. In this step, we only used shared transcription loci identified from RNA-seq data between two maize lines (**Figure 4A**). Notably, these transcribed loci were divided into five groups such as protein-coding genes, IncRNAs, transposable elements, and unassigned loci. **Thirdly**, we identified known miRNAs
Our recent study on reconstructing ceRNA regulatory network mainly based on is necessary to introduce ceRNA regulators into GRN construction. Here, we present mechanism in the growth and development of plants and animals. Consequently, it [82], indicating that the ceRNA-miRNA relationship is an essential gene regulatory sequence repeat, and circular RNAs have molecular functions as ceRNAs [79, 81, transcription of protein-coding genes, pseudogene, transposable elements, simple gene involved in phosphate homeostasis [80]. The following studies revealed that ath-miR399 and indirectly increase the transcription levels of an important miRNA, was found in a long noncoding RNA, IPSI A. thaliana. It could completely sponge levels of miRNA target genes can be derepressed [78, 79]. A typical ceRNA in plant, transcripts have abilities to adsorb miRNA molecules; subsequently, the expression recently proposed [78]. According to the ceRNA hypothesis, some endogenous regulatory model, the competing endogenous RNA (ceRNA) hypothesis, was firstly found in nematodes in 1993 [77]. It is noteworthy that a novel type of gene control key biological functions including anther development, since miRNA was further transcriptome-based GRN studies. In this step, we only used shared transcription loci identified stably transcribed loci based on the maize reference genome and estimated their separation to the finally constructed GRN (Figure 4). The analysis workflow of small RNA-seq data was performed from raw sequencing data prep- and predicted potential novel miRNAs that may be involved in maize anther development. Sequenced small RNA data were obtained from the same samples that were used in RNA-seq. A matched dataset (e.g., matched RNA and small RNA sequenced dataset here) is important in experimental design and more powerful to reveal the investigated biological questions. Though the analysis workflow of small RNA-seq data is similar to that of RNA-seq data in general (Figure 1), there are some differences between them. In our analysis, we reanalyzed two sets of published small RNA data to compare with their results from our own sequenced data for credible known and potential novel miRNAs involved in maize anther development [23, 83] (Figure 4B). This is an important check method to confirm the stability of research
results and conclusions. Fourthly, we predicted ceRNA-miRNA interaction pairs and miRNA target gene regulatory pairs by computational approach (Figure 4C). Bioinformatics analysis in this step is mainly based on genome sequence but not the transcriptomes. Fifthly, we reconstructed ceRNA-miRNA target gene regulatory networks by predicted interaction pairs and transcription correlation patterns from transcriptomics data (Figure 4D). It is well known that miRNAs could repress the transcription levels of their target genes. Additionally, ceRNA was demonstrated to negatively regulate the transcription levels of matched miRNAs. The negatively associated gene pairs in transcription levels may be more credible in mutual interactions. By integrating ceRNA-miRNA and miRNA target gene interactions, we reconstructed ceRNA-miRNA target gene regulatory networks in maize anther. Finally, we generally investigated the functional significance of genes in the regulatory network by GO enrichment analysis. In these networks, we found a number of well-studied genes and miRNA target gene pairs involved in maize anther development and male sterility, suggesting that the ceRNA-miRNA target gene regulatory networks contribute to anther development in maize. Besides, GO analysis of target genes in the network revealed that they are functionally enriched in flower development process (Figure 4E) [84].

7. Conclusions

Here, we summarized major points in comparative transcriptomics analysis from the commonly utilized workflow to the closely related research cases and from the single gene-based function analysis to GRN-based gene function investigation. In GMS gene studies, the research experiments using comparative transcriptomics method to investigate key functional genes and the genome-wide GRNs in developmental anther will facilitate our systematical understanding on the biological processes and molecular regulatory networks for anther development and sexual plant reproduction. More importantly, case studies illustrated here have a general meaning on technologies and methodologies for functional researches of other biological pathways and processes. With the fast advancement of sequencing technology, plant comparative transcriptomics has achieved considerable development. However, our understanding on the transcriptional dynamics and gene regulatory relationships of biological processes are far from being completed. Consequently, more efforts are needed for the further improvement of comparative transcriptomics in plant biological studies.

Acknowledgements

The research in our lab was supported by the National Key Research and Development Program of China (2018YFD0100806, 2017YFD0102001, 2017YFD0101201), the National Transgenic Major Program of China (2018ZX0800922B, 2018ZX0801006B), the National Natural Science Foundation of China (31,771,875, 31,871,702), the Fundamental Research Funds for the Central Universities of China (06500060), and the “Ten Thousand Plan” of National High-level Talents Special Support Plan (For Xiangyuan Wan).

Conflict of interest

The authors declare that they have no conflict of interest.
Transcriptome Analysis

results and conclusions.

Fourthly, we predicted ceRNA-miRNA interaction pairs and miRNA target gene regulatory pairs by computational approach (Figure 4C). Bioinformatics analysis in this step is mainly based on genome sequence but not the transcriptomes.

Fifthly, we reconstructed ceRNA-miRNA target gene regulatory networks by predicted interaction pairs and transcription correlation patterns from transcriptomics data (Figure 4D). It is well known that miRNAs could repress the transcription levels of their target genes. Additionally, ceRNA was demonstrated to negatively regulate the transcription levels of matched miRNAs. The negatively associated gene pairs in transcription levels may be more credible in mutual interactions. By integrating ceRNA-miRNA and miRNA target gene interactions, we reconstructed ceRNA-miRNA target gene regulatory networks in maize anther.

Finally, we generally investigated the functional significance of genes in the regulatory network by GO enrichment analysis. In these networks, we found a number of well-studied genes and miRNA target gene pairs involved in maize anther development and male sterility, suggesting that the ceRNA-miRNA target gene regulatory networks contribute to anther development in maize. Besides, GO analysis of target genes in the network revealed that they are functionally enriched in flower development process (Figure 4E) [84].

7. Conclusions

Here, we summarized major points in comparative transcriptomics analysis from the commonly utilized workflow to the closely related research cases and from the single gene-based function analysis to GRN-based gene function investigation. In GMS gene studies, the research experiments using comparative transcriptomics method to investigate key functional genes and the genome-wide GRNs in developmental anther will facilitate our systematical understanding on the biological processes and molecular regulatory networks for anther development and sexual plant reproduction. More importantly, case studies illustrated here have a general meaning on technologies and methodologies for functional researches of other biological pathways and processes. With the fast advancement of sequencing technology, plant comparative transcriptomics has achieved considerable development. However, our understanding on the transcriptional dynamics and gene regulatory relationships of biological processes are far from being completed. Consequently, more efforts are needed for the further improvement of comparative transcriptomics in plant biological studies.

Acknowledgements

The research in our lab was supported by the National Key Research and Development Program of China (2018YFD0100806, 2017YFD0102001, 2017YFD0101201), the National Transgenic Major Program of China (2018ZX0800922B, 2018ZX0801006B), the National Natural Science Foundation of China (31,771,875, 31,871,702), the Fundamental Research Funds for the Central Universities of China (06500060), and the "Ten Thousand Plan" of National High-level Talents Special Support Plan (For Xiangyuan Wan).

Conflict of interest

The authors declare that they have no conflict of interest.

Abbreviations

- ceRNA: competing endogenous RNA
- DE: differentially expressed
- DGEP: digital gene expression profile
- GCE: gene co-expression
- GEG: differentially expressed gene
- GEM: gene expression microarray
- GMS: genic male sterility
- GO: gene ontology
- GRN: gene regulatory network
- GSE: gene set enrichment
- KEGG: Kyoto encyclopedia of genes and genomes
- LCM: laser capture microdissection
- LncRNA: long noncoding RNA
- miRNA: microRNA
- MPSS: massively parallel signature sequencing
- MS: male sterility
- NGS: the next-generation sequencing
- RNA-seq: RNA sequencing
- SAGE: serial analysis of gene expression
- scRNA-seq: single-cell RNA sequencing
- SEM: scanning electron microscopy
- TEM: transmission electron microscope
- TF: transcription factor
- TF-GRN: TF gene regulatory network
- TRN: transcriptional regulatory network
- WT: wild type

Author details

Xiangyuan Wan¹,²* and Ziwen Li¹,²*

1 Biology and Agriculture Research Center, University of Science and Technology Beijing, Beijing, China

2 Beijing Engineering Laboratory of Main Crop Bio-Tech Breeding, Beijing International Science and Technology Cooperation Base of Bio-Tech Breeding, Beijing Solidwill Sci-Tech Co. Ltd., Beijing, China

*Address all correspondence to: wanxiangyuan@ustb.edu.cn and liziwen@ustb.edu.cn

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
References


buds and flowers of male sterile and fertile lines in watermelon. BMC Genomics. 2015;16:914


[40] Nelms B, Walbot V. Defining the developmental program leading to meiosis in maize. Science. 2019;364:52-56


glycerol-3-phosphate acyltransferase. Theoretical and Applied Genetics. 2018;131:1363-1378


[57] Li X. Next-generation sequencing sheds new light on small RNAs in plant reproductive development. Current Issues in Molecular Biology. 2018;27:143-170

[58] Li ZF, Zhang YC, Chen YQ. miRNAs and IncRNAs in reproductive development. Plant Science. 2015;238:46-52


Chapter 5

Transcriptome Analysis for Abiotic Stresses in Rice (Oryza sativa L.)

Ashutosh Kumar and Prasanta K. Dash

Abstract

Rice, a model monocot system, belongs to the family Poaceae and genus *Oryza*. Rice is the second largest produced cereal and staple food crop fulfilling the demand of half the world's population. Though rice demand is growing exponentially, its production is severely affected by variable environmental changes. The various abiotic factors drastically reduce the rice plant growth and yield by affecting its different growth stages. To fulfill the growing demand of rice, it is imperative to understand its molecular responses during stresses and to develop new varieties to overcome the stresses. Earlier, the microarray experiments have been used for the identification of coexpressive gene networks during various conditions in crop plants. Though the microarray experiments provided very useful information, the unviability of genome-wide information did not provide complete information about the regulatory gene networks involved in the stress response. The advancement of molecular techniques provided breakthrough to understanding the complex regulatory gene networks and their signaling pathways during stresses. The high-throughput RNA sequencing data have opened the floodgate of transcriptome data in rice. Here we have summarized some of the transcriptome data for abiotic molecular responses in rice, which further help to understand their complex regulatory mechanism.

Keywords: abiotic stresses, cold stress, drought, micronutrients, rice, RNA-Seq, salt stress, submergence, trace element stress, transcriptome

1. Introduction

Rice is the most important staple food crop across the globe and is a model monocot system [1]. It is the second largest produced cereal fulfilling the demand of half of the world's population. Rice belongs to family Poaceae and genus *Oryza*. Two species *Oryza sativa* (Asian rice) and *Oryza glaberrima* (African rice) out of 23 species have been cultivated worldwide [2]. The *O. sativa* is native to tropical and subtropical southern and southeastern Asia, while *O. glaberrima* is grown only in South Africa. A third species, *O. rufipogon*, has also been grown in South Asian, Chinese, New Guinean, Australian, and American farms. In Asia, *O. sativa* is separated into three subspecies according to its geographical environment: indica, japonica, and javanica. The variety indica refers to the tropical and subtropical varieties grown throughout South and Southeast Asia and Southern China. The variety japonica is grown in temperate areas of Japan, China, and Korea, while javanica varieties are...
Chapter 5

Transcriptome Analysis for Abiotic Stresses in Rice
(Oryza sativa L.)

Ashutosh Kumar and Prasanta K. Dash

Abstract

Rice, a model monocot system, belongs to the family Poaceae and genus Oryza. Rice is the second largest produced cereal and staple food crop fulfilling the demand of half the world’s population. Though rice demand is growing exponentially, its production is severely affected by variable environmental changes. The various abiotic factors drastically reduce the rice plant growth and yield by affecting its different growth stages. To fulfill the growing demand of rice, it is imperative to understand its molecular responses during stresses and to develop new varieties to overcome the stresses. Earlier, the microarray experiments have been used for the identification of coexpressive gene networks during various conditions in crop plants. Though the microarray experiments provided very useful information, the unviability of genome-wide information did not provide complete information about the regulatory gene networks involved in the stress response. The advancement of molecular techniques provided breakthrough to understanding the complex regulatory gene networks and their signaling pathways during stresses. The high-throughput RNA sequencing data have opened the floodgate of transcriptome data in rice. Here we have summarized some of the transcriptome data for abiotic molecular responses in rice, which further help to understand their complex regulatory mechanism.

Keywords: abiotic stresses, cold stress, drought, micronutrients, rice, RNA-Seq, salt stress, submergence, trace element stress, transcriptome

1. Introduction

Rice is the most important staple food crop across the globe and is a model monocot system [1]. It is the second largest produced cereal fulfilling the demand of half world’s population. Rice belongs to family Poaceae and genus Oryza. Two species Oryza sativa (Asian rice) and Oryza glaberrima (African rice) out of 23 species have been cultivated worldwide [2]. The O. sativa is native to tropical and subtropical southern and southeastern Asia, while O. glaberrima is grown only in South Africa. A third species, O. rufipogon, has also been grown in South Asian, Chinese, New Guinean, Australian, and American farms. In Asia, O. sativa is separated into three subspecies according to its geographical environment: indica, japonica, and javanica. The variety indica refers to the tropical and subtropical varieties grown throughout South and Southeast Asia and Southern China. The variety japonica is grown in temperate areas of Japan, China, and Korea, while javanica varieties are
grown alongside of indica in Indonesia (http://agropedia.iitk.ac.in/?q=content/ botanical-classification-rice).

Rice is an annual plant, even though in tropical areas, it is cultivated perennially. It is self-pollinated (wind pollination) tropical C3 grass that evolved in a semi-aquatic, low-radiation habitat having arenchymatic tissues [3]. Rice is cultivated in more than 100 countries, with a total harvested area till 2017 is of approximately 165 million hectares, and produced ~700 million tons (503.9 million tons of milled rice) (http://www.fao.org/3/i92433EN/i9243en.pdf). About 91% of the rice in the world is grown in Asia (nearly 640 million tons) where 60% of the world’s population lives. Rice is also cultivated in Sub-Saharan Africa and Latin Americas, and evenly poised in the Eastern and Western Asia. China and India, which account for more than one-third of global population, supply over half of the world’s rice. The China produces ~30% of total world rice production followed by India (21%), Indonesia (9%), and Bangladesh (6%). On the other hand, rest of Asia, Americas, and Africa produce 37, 5, and 3%, respectively, of the total world rice production [4]. However, demand of the rice is still growing day by day, as the world population is mounting exponentially.

To fulfill the demand of growing population, yield needs to be increased by the application of agricultural as well as biotechnological approaches.

Rice production is severely affected by changing environment including extreme variability in temperature and rainfall pattern along with other factors [5]. The abiotic stresses including drought, high salinity, high or low temperatures, flooding, high light, ozone, low nutrient availability, mineral deficiency, heavy metals, pollutants, wind and mechanical injury, drastically reduce the rice plant growth and yield by affecting it during different growth stages [6]. However, rice has very antagonistic character about tolerances and susceptibilities to abiotic stresses, as compared to other crops. It is very well known that rice paddy grows in standing water containing soil and can tolerate submergence at levels that would kill other crops. However, it is moderately tolerant to salinity and soil acidity but highly susceptible to drought and cold. Drought influences all physiological processes involved in plant growth and development [5]. Drought at vegetative stage can moderately reduce yield, but entire yield is lost if it occurs during pollen meiosis or fertilization [7]. The high salt concentration disrupts the ability of roots for efficient water uptake, leading to perturbation of crucial metabolic reactions inside the cell restricting plant growth and yield potential [8]. Low temperature reduces germination, causes poor establishment, delays phenological development, and increases spikelet sterility [9], and other physiological and metabolite changes causing low yield [10]. Furthermore, rice can tolerate partial submergence as paddy rice or deepwater rice because it is very well adapted to waterlogged conditions as it has well-developed aerenchyma that facilitates oxygen diffusion and prevents anoxia in roots [11–13]. However, it was damaged when submerged partially or completely for a relatively longer period [14] due to the shortage of oxygen during submergence. The response of plants to low oxygen stress comprises complex biochemical and genetic programs that include the differential expressions of a large number of genes. Importantly, abiotic stress conditions not only harm the crop but also influence the manifestation and extent the pathogen infection, attack of insects, and growth of weeds [6]. Though rice has superior response to abiotic stresses, development of their improved tolerant germplasm is indispensable [11]. Besides abiotic stress, the deficiency of micronutrients also affects the crop production.

The crop plants are very sensitive and respond to environmental stimuli through signal perception. The plant responds accordingly for a specific environmental stimulus instigating specific physiochemical changes. These physiochemical changes or adaptions are administered by complex molecular regulatory mechanism of involving various sensors regulated by transcriptional factors/regulators. Various studies have been carried out for understanding the regulatory mechanism of plants during stress.
conditions. Earlier, CIPK genes (OsCIPK01–OsCIPK30) in the rice genome were studied for their transcriptional responses to various abiotic stresses [15]. The results showed that 20 OsCIPK genes were differentially induced by at least one of the stresses, including drought, salinity, cold, polyethylene glycol, and abscisic acid treatment. Most of the genes induced by drought or salt stress were also induced by abscisic acid treatment but not by cold. A few CIPK genes containing none of the reported stress-responsive cis-elements in their promoter regions were also induced by multiple stresses [15]. The proteins possessing A20/AN1 zinc-finger, named SAP gene family in rice and Arabidopsis, were inducible by one or the other abiotic stresses indicating that the OsSAP gene family is an important component of stress response in rice [16]. In addition, the role of SAP gene family in abiotic stress conditions was established by expression profiling under abiotic stress conditions. Seven Expansin A (ExpA) mRNAs were accumulated in leaves of deepwater rice, and their abundance was upregulated by submergence [17]. Similarly, the drought response in rice incites a signaling cascade through osmolyte synthesis that involves perception and translation of drought signal [18, 19].

Earlier, microarray experiments have been used for expression analysis of multiple genes during various conditions in different tissues for crop plants. The microarray experiments helped to identify the coexpressive genes during a stress condition [20–23]. Though the microarray experiments provided very useful information, the unviability of genome-wide information about the transcripts did not provide the complete information about the regulatory gene networks involved in the stress response. Nowadays, the availability of high-throughput techniques, achieved through advancement of molecular techniques, provided breakthrough in the understanding of complex regulatory gene networks and their signaling pathways involved in stress responses [24]. The techniques are comprised of whole genome transcriptome analyses, small RNA sequencing analysis (RNA-Seq), proteomic analyses, epigenetic sequencing analysis, and metabolomic analyses [25]. These high-throughput techniques use sequence-based approaches instead of hybridization-based approaches (like microarray), which require known genomic sequences, rather able to determine the transcript sequences directly from new genomes, able to map and quantify them [26, 27]. The RNA-Seq has superiority among these techniques due to its in-depth coverage of genome, global expression of transcripts, and also providing detailed information about alternative splicing and allele-specific expressions [27]. The inception of RNA-Seq technique has reformed the perception of complex and dynamic nature of the genomes, further helps to comprehensively elucidate the complex regulatory gene networks pertaining to different physiological and developmental stages of plants [28]. Currently, the various transcriptome analyses of rice genome, accomplished through RNA-Seq, during various abiotic stresses have generated enormous data. Further, these data have been able to decipher the complex regulatory gene networks in rice during various abiotic stresses which helped to understand the adaptive physiological measures taken by rice at cellular level and ascertain the development of tolerant rice varieties. Here, we are describing some of the different transcriptome studies carried out to understand the molecular responses in rice genome during various abiotic stresses.

2. Transcriptome data for submergence/flooding

Flooding is considered as a major threat to the rice crops, as irregular flash floods are very common in the Southeast Asia (major rice producing region), severely affecting the rice productivity [29]. Rice produces high yields, when it is grown in water-logged rice paddies. It can tolerate partial submergence as paddy rice or deepwater rice. However, it is damaged when submerged for a relatively longer
period [14] due to the slow diffusion of oxygen in water fails to match the demands of respiration [30] resulting an anaerobic metabolism and energy crisis [12]. Also, in deepwater rice, energy generation through fermentative metabolism, aerenchyma development in parenchymal tissues that improves access to O2, activation of ethylene promoted gibberellic acid (GA)-mediated internode elongation cause foliage to shoot up above the water surface for gas exchange and restricting growth and conserving available energy until floodwater recedes [12, 13]. Similarly, flood-tolerant rice varieties have developed the capacity to generate ATP without the presence of oxygen and/or to develop specific morphologies that improve the entrance of oxygen [31]. Moreover, the phytohormonal regulation revealed that gibberellin (GA) has negative effects on submergence tolerance, whereas paclobutrazol (PB), chemical inhibitor of GA, acted contrary to GA [32]. The transcriptome analysis between GA- and PB-treated samples and control identified 3936 differentially expressed genes largely associated with the stress response, phytohormone biosynthesis and signaling, photosynthesis, and nutrient metabolism. It was observed that the PB improved the rice survival during submergence through sustaining the photosynthesis capacity and by dropping nutrient metabolism [32].

Despite knowledge of adaptive mechanisms and regulation at the gene and protein level, our understanding of the mechanisms behind plant responses to submergence is still limited. Even in flood-intolerant species, such as Arabidopsis thaliana, many genes are triggered in response to flooding stress [33, 34]. The response of plants to low oxygen stress comprises complex biochemical and genetic programs that include the differential expressions of a large number of genes (Table 1). Gene expression is altered under low oxygen stress, and the existence of anaerobic response elements (AREs) along with their binding factors has already been reported [35]. Eventually, a SUB1 locus and three ethylene response factors (ERFs) were identified within the locus in tolerant rice varieties (e.g., FR13A), whereas SUB1 is a major determinant of tolerance [36]. Introduction of the SUB1A gene into submergence-intolerant rice variety significantly increased its flooding tolerance, thus demonstrating the importance of the SUB1 locus for flooding tolerance [36]. Two different types of molecular mechanisms are adapted by rice ecotypes to survive under stress, SUB1A-mediated “quiescence strategy” [37, 38] and “escape strategy” induced by SNORKEL1/2 [13]. The submergence response in rice consists of the differential expression of genes related to gibberellin biosynthesis, trehalose biosynthesis, anaerobic fermentation, cell wall modification, and transcription factors that include ethylene-responsive factor genes [39]. Though the regulatory mechanism in rice during submergence response has been comprehensively studied, the genome-wide gene expression as well as allelic variation among the cultivars for specific quantitative traits remained elusive. One of the studies was conducted in six rice genotypes to estimate the coleoptile elongation rates during submergence [39]. The result postulated that the coleoptile elongation was augmented by transcriptional regulation. Further, the reason for the variation in anaerobic germination was due to the allelic variation caused by the small-to-large deletions in the coding region of susceptible varieties [39].

Recently, a study on SUB1A-1 genotypes is carried to understand the molecular mechanism pertaining to the physiological function upon desubmergence through transcriptomic analysis [29]. The results enumerated around 1400 genes that were differentially expressed to recover from the stress to preserve the plastid integrity, and the genes regulating the cell division, chromatin structure, and signaling associated with starch catabolism [29]. They also found that the rice plants recover shoot transcriptome significantly to the control state and return to homeostasis during the 24-h recovery period. It also regulated the GA-responsive starch metabolism
<table>
<thead>
<tr>
<th>Abiotic stress condition</th>
<th>Gene/s responsible for tolerance</th>
<th>Downstream key gene/s</th>
<th>Physiological functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submergence</td>
<td>SUB1A, SNORKEL1/2</td>
<td>ERs regulating genes of GA-responsive starch metabolism, anaerobic fermentation, cell wall modification, JA-mediated internode elongation, and biotic responsive</td>
<td>Quiescence strategy to stop all physiological functions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Escape strategy to supersede water level</td>
</tr>
<tr>
<td>Drought</td>
<td>DREBs (DREB1A-D/ CBF1-4 and DREB2)</td>
<td>ABA-responsive genes, LEA, NAC, DBF, α-linolenic acid metabolic pathway genes, osmolyte biosynthesis genes, phospholipid metabolism genes; water channel protein, sugar and proline transporters, and detoxification enzyme-encoding genes; and signaling molecule-encoding genes</td>
<td>Stomatal closure, repression of cell growth, photosynthesis and activation of respiration and production of phytohormone ABA</td>
</tr>
<tr>
<td>Salt</td>
<td>SOS1, NHX, HKT2, CAX1, AKT1, KCO1, TPC1, CLC1, NRT1, CDPK7, MAPK5, CaMBP, GST, LEA, V-ATPase, OSAPI, and HBP1B</td>
<td>Genes related to antioxidants, transcription factors, signaling, ion and metabolic homeostasis and transporters</td>
<td>Imbalance in ion homeostasis of cells at plasma membrane and sequestration of vacuolar ion, and stomatal closure which causes higher leaf temperature and reserve shoot elongation</td>
</tr>
<tr>
<td>Cold</td>
<td>CBF1, DREB1A, and DREB1B</td>
<td>ABA-responsive genes, ABF, NAC, NACRS containing genes, ERF922, WRKY25, and WRKY74, gene related to signal transduction, phytohormones, antioxidant system and biotic stress</td>
<td>Altered chlorophyll content and fluorescence causing reduction in photosynthesis, increases content of ROS and malondialdehyde causing oxidative damage to cells</td>
</tr>
<tr>
<td>Cadmium (Cd)</td>
<td></td>
<td>Cd-responsive transporters, ROS-scavenging enzymes, chelators, and metal transporter-encoding genes and many drought stress-related genes</td>
<td>Fatal damage to rice seedlings during their development</td>
</tr>
<tr>
<td>Phosphorus (P)</td>
<td>RNA transport and mRNA monitoring path genes</td>
<td>Important for energy transfer, signal transduction, photosynthesis, and respiration</td>
<td></td>
</tr>
<tr>
<td>Manganese (Mn)</td>
<td>TFs, transporters, transferase protein genes, catalytic protein encoding genes, WRKY, and potassium transporter-related genes, Aux/IAA family, and sodium transporter-related genes</td>
<td>Important for catalyzing the water-splitting reaction of oxygen-evolving complex in photosystem II (PSII), acts as cofactor that activates different enzymes, such as Mn-superoxide dismutase and others, to protect against oxidative stresses</td>
<td></td>
</tr>
<tr>
<td>Alkaline stress</td>
<td>Alkali-responsive genes</td>
<td>Alkaline resistant genes, TFs related to hormone signal transduction and secondary metabolite biosynthesis pathways</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Regulatory role of different abiotic stress-responsive genes based on RNA-Seq analysis.
indirectly through SUB1A and downstream regulatory network to resume the photosynthesis [29]. Similar studies have also been carried between two contrasting deepwater growth rice cultivars [40]. The RNA-Seq analysis was conducted from different tissues, shoot base region, including basal nodes, internodes, and shoot apices of seedlings at two developmental stages. The study elucidated the possible role of jasmonic acid-mediated internode elongation and expression of biotic stress-related genes during submergence response [40].

3. Transcriptome data for drought stress

One of the major abiotic stresses that severely affect the rice production is drought stress. Drought stress causes a series of physiological and biochemical changes which included stomatal closure, repression of cell growth, photosynthesis, and activation of respiration along with production of the phytohormone abscisic acid (ABA) [41]. In response to the drought stress, ABA triggers stomatal closure and induces expression of stress-related genes (Table 1) [41]. However, some of drought-related genes were not expressed by the external ABA treatment. Therefore, the drought response is either of ABA-independent or of ABA-dependent or both inducible gene regulatory system networks [42]. These regulatory networks are the amalgamation of interaction between transcription factors and their respective promoter cis-elements. It was observed that the promoters of ABA-dependent genes have ABA-responsive element (ABRE) and, dehydration- and cold-responsive element (C-repeat/DRE) [42]. The transcription factors, which specifically bind to ABRE are known as DREBs, trigger the expression of ABA-responsive genes [43], which further encode AP2 domain-containing transcription factors regulating the stress-related genes in an ABA-independent manner [44]. The DREB gene family has two groups DREB1/CFB and DREB2, whereas DREB1/CFB consists of DREB1A (CBF3), DREB1B (CBF1), DREB1C (CBF2), and DREB1D (CBF4). However, five DREB homologs were identified in rice, OsDREB1A, OsDREB1B, OsDREB1C, OsDREB1D, and OsDREB2A [45, 46]. These gene-encoded proteins are classified into two: the first group belongs to the functional proteins included chaperones, late embryogenesis abundant (LEA) proteins, osmotin, antifreeze proteins, mRNA-binding proteins, enzymes for osmolyte biosynthesis, water channel proteins, sugar and proline transporters, and detoxification enzymes; the second group is of regulatory proteins (signal transduction and stress-responsive) including various transcription factors, protein kinases, protein phosphatases, enzymes involved in phospholipid metabolism, and other signaling molecules such as calmodulin-binding protein [22, 41]. Interestingly, it was found that many of these proteins, especially DREBs, are also involved in transcriptional regulation of stress-response mechanism during cold and salt stresses [46, 47].

The rice is the only crop which is grown in the waterlogged fields and it has very low water-use efficiency [48]. Therefore, it is imperative to decipher the molecular regulatory mechanism to increase the water usage efficiency of rice or the drought tolerance. Nowadays, the drought stress is continuously affecting the rice productivity due to the harsh environmental condition. The transcriptome studies proved to be the boom for researchers due to its global genomes depth and all at once allele mining among different rice genotypes. Earlier, a transcriptome analysis between drought-tolerant and drought-sensitive cultivars was carried out for the identification of novel genetic regulatory mechanisms [48]. This study suggested that the upregulation of genes related to carbon fixation, glycolysis gluconeogenesis, and flavonoid biosynthesis, whereas the downregulation of genes associated with starch and sucrose metabolism during drought. Further, they also found the upregulation...
of genes associated with α-linolenic acid metabolic pathway in tolerant genotype during the stress which supported the previous findings. Consecutively, the analysis of consensus cis-motif among the coexpressed drought-induced genes led to the identification of novel cis-motifs [48]. Similar comparative studies have been carried out between tolerant and susceptible rice cultivars and in other crops to understand the regulatory mechanisms during drought [49–51]. Their result suggested that 801 transcripts differentially expressed in tolerant cultivar including the TFs NAC and DBP, and thioredoxin involved in phenylpropanoid metabolism [49].

To sustain the drought condition, the roots have a very important role. To understand the molecular regulation in rice seedling roots (4-weeks old) during drought condition, comparative RNA-Seq analysis has been carried out between wet and dry soil conditions [52]. This analysis suggested that 68% of identified genes were novel, and also found that the one of the enzymes RING box E3 ligases from ubiquitin-proteasome pathway was induced by drought. Interestingly, it was found that the OsPhyB represses the activity of ascorbate peroxidase and catalase-mediating reactive oxygen species (ROS) processing machinery required for drought tolerance of roots in soil condition, contrary to the previous results [52].

4. Transcriptome data for salt stress

Some of the abiotic stresses are complementary to each other such as the drought and salt, drought and cold stresses, etc., affecting the rice productivity. It is evident that excessive loss of water from the soil evaporation due to drought causes salt accumulation in soil. The salinity is defined as deposition of sodium chloride from natural accumulation or irrigation in soil. It causes imbalance in ion homeostasis of cells regulated by ion influx and efflux at the plasma membrane and sequestration of vacuolar ion [8]. The salt stress affects stomatal closure causing increased leaf temperature and reserved shoot elongation [53]. Studies on the salinity tolerant in rice have shown the regulation of genes related to antioxidants, transcription factors, signaling, ion and metabolic homeostasis, and transporters (Table 1) [54]. The identified important class of genes regulated during a salt stress in rice are OsSOS1, OsNHX1 (Na+/H+ antiporters), OsHKT2;1 (Na+/K+ symporter), OsCAX1 (H+/Ca2+ antiporter), OsAKT1 (K+ inward-rectifying channel), OsKCO1 (K+ outward-rectifying channel), OsTPC1 (Ca2+ permeable channel), OsCLC1 (Cl− channel), OsNRT1;2 (nitrate transporter), OsCDPK7, OsMAPK5, CaMBP (calmodulin motif binding protein), GST (glutathione-S-transferase II), LEA (late embryogenesis abundant protein), V-ATPase (vacuolar ATP synthase 16KD proteolipid subunit), OSAP1 (zinc finger protein), and HBP1B (histone binding protein, TF) [55–63]. The salt stress response mechanism is moreover of complex physiological process pertaining to metabolic and morphological changes, which is comprehensively studied, but in rice, the molecular regulatory mechanism to salt tolerance is elusive [64]. Some of the transcriptome analyses have been completed in conjugation with the drought stress to understand the salt tolerance in rice [46, 49, 59]. Earlier, a comparative study has been carried out between salt tolerant and susceptible rice cultivars to understand the regulatory mechanisms [49]. The result suggested higher expression of bHLH and C2H2 TF family members, which might be regulating the genes associated with wax and terpenoid metabolism pathways [49]. Similarly, to understand the salinity stress, a comparative leaf transcriptome analysis at three time points on rice seedlings has been completed [65]. They identified 1375 novel genes, whereas 286 differentially expressed genes exclusively found in tolerant cultivar. They validated two genes: disease resistance response protein 206 and TIFY10A to understand the molecular response to salinity stress [65].
5. Transcriptome data for cold stress

The cold stress is defined according to the temperature affecting the plant growth and development which ranges 0–15°C (chilling stress) and <0°C (freezing stress) [66]. The tropical origin of rice makes it more susceptible to cold, critically affecting reproductive stages and grain quality leading to yield reductions [67]. The cold stress affects chlorophyll content and fluorescence causing reduction in photosynthesis, increases content of reactive oxygen species (ROS) and malondialdehyde (MDA) causing oxidative damage to cells in rice [68]. The molecular regulation of cold stress is identified in conjugation of drought stress (Table 1) [45]. Many stress-inducible genes are regulated via ABA-independent pathway, characteristically having a cis element responsible for dehydration (DRE) as well as low-temperature-induced expression. The low-temperature-inducible genes possess C-repeat (CRT) and low-temperature-responsive element (LTRE). The DRE-binding proteins encoding genes CBF1, DREB1A, and DREB1B were induced by cold stress [46]. During cold stress, ABA also accumulates and initiates the ABA signaling cascade, which regulates the ABA-responsive genes through ABRE and the ABRE-binding bZIP transcription factor ABF [69]. The OsNAC gene transduces the ABA signal through an ABRE in its promoter and regulates the expression of NACRS-containing genes to control cold tolerance in rice [67]. Further, to understand comprehensively the regulation of genes during cold stress, a transcriptome study is carried out between weedy and cultivated rice [70]. The analysis suggested that some typical cold stress-related genes were of basic helix-loop-helix (bHLH) gene and leucine-rich repeat (LRR) domain genes, and several genes associated with phytohormones like abscisic acid (ABA), gibberellic acid (GA), auxin, and ethylene [70]. Similarly, the wild rice, O. longistaminata, tolerates nonfreezing cold temperatures, is used for the identification of molecular mechanisms in response to low temperature in its shoots and rhizomes at seedling and reproductive stages using transcriptome analysis [71]. They found photosynthesis-pathway-related genes were prevalent in shoots, whereas metabolic pathways and the programmed cell death process-related genes were expressed only in rhizomes. Further, they found that the TFs CBF/DREB1, AP2/EREBPs, MYBs, and WRKYs were synergistically expressed in shoots, whereas OsERF922, OsNAC9, OsWRKY25, OsWRKY74, and eight antioxidant enzymes encoding genes were expressed in rhizomes during cold stress. The cis-regulatory element analysis suggested the enrichment of ICE1-binding site, GATA element, and W-box in both tissues. And the highly expressed genes in shoots were associated with photosynthesis, whereas signal transduction-related genes were highly expressed in rhizomes [71].

Furthermore, a transcriptome analysis is performed in germination phase for contrasting cultivars of rice in cold stress [72], suggesting the higher expression of gene related to signal transduction, phytohormones, antioxidant system, and biotic stress during germination in cold stress [72].

6. Transcriptome data for trace element stress

The rice is the staple food fulfilling the dietary needs of a large population around the world. Besides dietary energy and proteins, it also contains trace elements (Li, B, Al, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Sr, Mo, Cd, Ba, Pb, and Bi) in low amounts [73]. Some of these trace elements Se, Mo, Cr, Mn, Fe, Co, Cu, Zn are micronutrients that help in proper functioning of human biological systems, while nonessential heavy elements such as Pb, As, Cd, Hg are referred as toxins for consumption [73, 74]. However, the trace elements in rice are invariably increasing
either due to the use of agrochemicals or irrigation with contaminated water. The deficiency or accumulation of these trace elements in soil hampers plant growth and development. On the other hand, their biofortification helps to add nutrition supplement. Henceforth, the detailed study about the effects of these trace elements on the rice is indispensable. There are many reports about trace element stresses on rice achieved through transcriptome studies (Table 1).

The higher concentration of heavy metal cadmium (Cd) severely hampers the rice growth. Therefore, to understand the molecular mechanism during Cd stress, transcriptome analysis has been completed by exposing rice to higher concentrations of Cd [75]. They found constitutively expressed genes were less affected by low Cd concentrations, whereas high Cd concentration causes fatal damage to rice seedlings during their development. They also found some novel Cd-responsive transporters encoding genes [75]. Previously, they found the upregulation of many genes related to ROS-scavenging enzymes, chelators, and metal transporters during Cd exposure along with upregulation of many drought stress-related genes [76].

Phosphorus (P) is an essential trace element required for proper plant growth and development where it plays an important role in energy transfer, signal transduction, photosynthesis, and respiration [77]. A comparative transcriptome study has been carried out in leaf and root tissues during phosphorus stress to elucidate their molecular mechanisms [78]. The transcriptome analysis suggested that many differentially expressed TFs and functional genes were uniquely involved in multiple regulatory pathways (including RNA transport and mRNA monitoring path) during phosphorus deficiency tolerance [78].

Manganese (Mn) is an essential trace element which plays an important role in catalyzing the water-splitting reaction of oxygen-evolving complex in photosystem II (PSII). It also acts as a cofactor that activates different enzymes, such as Mn-superoxide dismutase and others, to protect against oxidative stresses in plants [79]. However, higher Mn affects the physiological and biochemical pathways associated with plant growth and development. Therefore, to decipher the molecular mechanisms in leaves of Mn-sensitive rice exposed to high Mn stress, transcriptome analysis has been done [79]. The analysis suggested that a large number of TFs, transporters, transferase proteins, catalytic proteins encoding genes were differentially expressed having a major role in primary and secondary metabolisms. Further, it was found that the WRKY family and potassium transporter-related genes were significantly upregulated, whereas Aux/IAA family and sodium transporter-related genes were strongly downregulated [79].

7. Transcriptome data for other stresses

Besides common abiotic stresses, some other stresses are also studied with the help of transcriptome analysis. A transcriptome study has been carried out for alkaline stress caused by alkaline NaHCO₃ and Na₂CO₃ [80]. The study reported the identification of 926 differentially expressed important alkali-responsive genes including 28 alkaline-resistant genes and 74 transcription factor genes. These genes were related to hormone signal transduction and secondary metabolite biosynthesis pathways [80].

The RNA-Seq or transcriptome analysis has tremendous potential to divulge the complex molecular machinery of plant regulatory response during stress conditions. However, this large number of transcriptome data of abiotic stresses in rice has contributed significantly to rice researchers. It helped to understand complete molecular mechanism pertaining to their physiological and biochemical changes. Such data mining could be a high impact methodical source for identification of candidate gene through integration of functional genomics approach. This will also
help to establish the hierarchical relationships between specific signaling components and downstream effector genes to cope up the stress conditions.

**Acknowledgements**

PKD acknowledges ICAR-NASF and ICAR-NPTC for funding and support of research work at NRC on Plant Biotechnology.

**Author details**

Ashutosh Kumar* and Prasanta K. Dash  
ICAR-National Institute for Plant Biotechnology, PUSA, New Delhi, India

*Address all correspondence to: kr.ashutosh@yahoo.com
References


[16] Vij S, Tyagi AK. Genome-wide analysis of the stress associated protein (SAP) gene family containing A20/AN1 zinc-finger(s) in rice and their phylogenetic relationship with


[50] Gupta P, Dash PK. Molecular details of secretory phospholipase A2 from flax (Linum usitatissimum L.) provide insight into its structure and function. Scientific Reports. 2017;7(1):11080


[61] Xiong L, Yang Y. Disease resistance and abiotic stress tolerance in rice are inversely modulated by an abscisic acid-inducible mitogen-activated protein kinase. The Plant Cell. 2003;15(3):745-759


[63] Kumari S, Sabharwal VP, Kushwaha HR, Sopory SK, Singla-Pareek SL, Pareek A. Transcriptome map for seedling stage specific salinity stress
response indicates a specific set of genes as candidate for saline tolerance in *Oryza sativa* L. Functional & Integrative Genomics. 2009;9(1):109-123


[74] Sebastian A, Prasad MNV. Trace element management in rice. Agronomy. 2015;5:30


Abstract

The novel powerful technique is used for a study of combinatorial and statistical properties of transcriptome sequences. The main approach stands on the study of distribution of nucleotide triplet frequency dictionaries obtained from the conversion of transcriptome sequences. The distribution is revealed through PCA presentation and elastic map technique. The transcriptomic data of Siberian larch (Larix sibirica Ledeb.) and Siberian pine (Pinus sibirica Du Tour) were studied. The transcriptomes exhibit unusual symmetries. The octahedral structure exhibiting rotational symmetry in transcriptome contig distribution was found for L. sibirica, while mirror symmetry was found for P. sibirica. The octahedron structure seems to be universal for plants.

Keywords: Chargaff’ s parity, order, structuredness, mirror symmetry, rotational symmetry

1. Introduction

A discovery of an order and new structures in genetic entities is an up-to-date scientific problem. Indeed, the amount of primary genomic data shows the daily growth for billions of megabases. The symbol sequences from four-letter alphabet ℵ = A, C, G, T (with few variations in some nucleotide sequences; say, U substitutes T in RNAs).

We studied an order and structuredness over a set of sequences representing the transcriptome of Siberian larch (Larix sibirica Ledeb.) and Siberian pine (Pinus sibirica Du Tour), also known as Siberian cedar. Transcriptome represents sequences of expressed genes and corresponds to the mRNA molecule isolated from biological cells or tissues. Obviously, whether a transcriptome exhibits structuredness or not heavily depends on the concept of a structuredness to be revealed and analyzed. One may face a huge number of patterns claimed to be structural units; a number of papers report on newly discovered structures in genomes [1].

There are two approaches to discuss structuredness in a set of symbol sequences (transcriptome nucleotide sequences, in our case). The first implies that one seeks for inhomogeneities in the mutual distribution of the sequences form the ensemble under consideration. Of course, to do it, one must introduce a metrics to measure...
Chapter 6

Revealing the Symmetry of Conifer Transcriptomes through Triplet Statistics

Sadovsky Michael, Putintseva Yulia, Biryukov Vladislav and Senashova Maria

Abstract

The novel powerful technique is used for a study of combinatorial and statistical properties of transcriptome sequences. The main approach stands on the study of distribution of nucleotide triplet frequency dictionaries obtained from the conversion of transcriptome sequences. The distribution is revealed through PCA presentation and elastic map technique. The transcriptomic data of Siberian larch (Larix sibirica Ledeb.) and Siberian pine (Pinus sibirica Du Tour) were studied. The transcriptomes exhibit unusual symmetries. The octahedral structure exhibiting rotational symmetry in transcriptome contig distribution was found for L. sibirica, while mirror symmetry was found for P. sibirica. The octahedron structure seems to be universal for plants.

Keywords: Chargaff’s parity, order, structuredness, mirror symmetry, rotational symmetry

1. Introduction

A discovery of an order and new structures in genetic entities is an up-to-date scientific problem. Indeed, the amount of primary genomic data shows the daily growth for billions of megabases. The symbol sequences from four-letter alphabet $\mathbb{A} = \{A, C, G, T\}$ (with few variations in some nucleotide sequences; say, U substitutes T in RNAs).

We studied an order and structuredness over a set of sequences representing the transcriptome of Siberian larch (Larix sibirica Ledeb.) and Siberian pine (Pinus sibirica Du Tour), also known as Siberian cedar. Transcriptome represents sequences of expressed genes and corresponds to the mRNA molecule isolated from biological cells or tissues. Obviously, whether a transcriptome exhibits structuredness or not heavily depends on the concept of a structuredness to be revealed and analyzed. One may face a huge number of patterns claimed to be structural units; a number of papers report on newly discovered structures in genomes [1].

There are two approaches to discuss structuredness in a set of symbol sequences (transcriptome nucleotide sequences, in our case). The first implies that one seeks for inhomogeneities in the mutual distribution of the sequences form the ensemble under consideration. Of course, to do it, one must introduce a metrics to measure
the difference between any two sequences; there are various ways to do it [2–4]. An alignment might be such a measure [5, 6] (see also much more prominent approach presented in [7, 8]). Alternatively, the second approach implies the search for inhomogeneities within a sequence, e.g., through the comparison of the formally identified fragments of a sequence.

Regardless the specific approach to seek for structuredness, one must introduce a way to measure the difference between the objects to be analyzed. Alignment [9–11] is the most widespread approach here. An alternative idea to search a structure and order in symbol sequences is to transform them into frequency dictionary [12–15]. A frequency dictionary could be defined in various ways, but basically it is a list of all the strings of a given length accompanied with a frequency of each string (a detailed description is given below). A transformation of a symbol sequence into a frequency dictionary provides a mapping of a set of sequences into a metric space. Hence, one may apply all the tools for analysis.

As soon, as a structure in ensemble of sequences, or over a sequence is defined, the question arises toward the properties of those structures. Probably, symmetry of such structures is the most fundamental and basic one. Again, there could be various notions of the symmetry. The first concept of the symmetry aims to figure out structures that seem to remain similar, when some simple transformations in a proper space are provided. First of all, a rotational symmetry of a cluster structure [3, 4] or mirror symmetry [16, 17] must be mentioned here.

Few words should be said toward the symmetry. Here we shall consider two notions of that issue. The first is a well-known rotational, mirror, or similar symmetry observed in the distribution of the contigs converted into triplet frequency dictionary as they are distributed in the relevant Euclidean space (where the triplets are the coordinates). The second issue is measured through the proximity (or deviation) to Chargaff’s parity rules, to be observed for various entities, both natural (these are contigs) and artificial (kernels or arithmetic means of the frequency of identical triplets counted over an ensemble of contigs).

2. Material and methods

2.1 Transcriptome nucleotide sequence data

The transcriptomes of Siberian larch and Siberian pine were originally sequenced under the project on the whole genome sequencing of Siberian larch [18, 19]. The sequence data of *L. sibirica* and *P. sibirica* were obtained using Illumina MiSeq sequencer at the Laboratory of Forest Genomics of the Siberian Federal University. The RNA was isolated from buds [19].

2.1.1 *L. sibirica* bud transcriptome

For the purposes of our study, we have selected the bud transcriptome of *L. sibirica*; we have taken into consideration the transcripts longer than 600 bp. The longest one in the transcriptome is as long as 10,795 bp, with average length \( \langle L \rangle = 1243.4 \) bp and standard deviation \( \sigma(L) = 717.9 \) bp.

The total number of sequences in the transcriptome is 12,353 transcripts. The histograms of the distribution of the transcriptome sequence entries over their length are presented in Figure 1. Evidently, the distribution resembles Poisson distribution quite strongly. There are 7573 transcripts in the transcriptome bearing a single CDS (maybe in various directions). Four thousand thirty-eight transcripts
Figure 1.
Distribution of L. sibirica contigs over the length (left) and P. sibirica (right).

<table>
<thead>
<tr>
<th>#</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. sibirica</td>
<td>3049</td>
<td>738</td>
<td>175</td>
<td>61</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P. sibirica</td>
<td>962</td>
<td>226</td>
<td>41</td>
<td>14</td>
<td>3</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

N—number of CDS in a transcript.

Table 1.
Distribution of number of CDS per transcript.

have two or more CDS in them; the distribution of number of CDS in transcripts is shown in Table 1. Finally, in 742 transcripts no CDS have been found.

2.1.2 P. sibirica bud transcriptome

We used bud transcriptome from Pinus sibirica obtained from witch’s broom (i.e., morphologically different part of a tree). It might be considered as a disease. Again, we have selected the transcripts longer that 600 bp that yields 4675 entries in the transcriptome, 3003 among them have a single CDS.

There are as many as 426 transcripts with no CDS detected in them. Surprisingly, there are no transcripts in the transcriptome with CDS belonging to both strands, simultaneously. The distribution of number of CDS found in a transcript is shown in Table 1. On the contrary to L. sibirica transcriptome, P. sibirica transcriptome contains no transcript without CDS.

2.2 Triplet frequency dictionary

Triplet frequency dictionary \(W_{(3,t)}\) is the list of all 64 triplets found within a sequence under consideration, where each entry (triplet) \(\omega\) is assigned with the frequency \(f_\omega\) of the triplet \(\omega\). The reading frame move \(t\) could be chosen arbitrary and depends on the specific problem to be solved. Everywhere further we use \(t = 1\) or \(t = 3\); for \(t = 1\) we use the notation of \(W_3\), unless it makes a confusion.

A frequency dictionary \(W_{(3,t)}\) unambiguously maps a sequence into a point in 64-dimensional metric space. Strongly speaking, \(W_{(3,t)}\) with \(t > 1\) maps a subsequence into the point of the metric space, not the sequence entirely; further we shall discuss this point in more detail. Next, the dimension of the space is 63, not 64; this fact follows from the linear constraint:

\[
\sum_{\omega=AAA} f_\omega = 1. \tag{1}
\]
This constraint allows to exclude any triplet from the analysis, thus changing 64-dimensional space for 63-dimensional, where all variables are linearly independent [20].

Formally speaking, any triplet could be excluded. Practically, one must eliminate the triplet with the least standard deviation figure determined over the set of frequencies under consideration. Indeed, suppose a triplet \( \omega^* \) yields the standard deviation equal to zero, as determined over a set of dictionaries, it means, all dictionaries in the set have the same frequency, for this triplet \( f_{\omega^*}^j = \text{const}, \forall j \) (here \( j \) enlists the dictionaries in the set). Such invariance makes the dictionaries (and the sequences standing behind) indistinguishable, from the point of view of the triplet. The choice of a triplet with minimal standard deviation for the exclusion provides the elimination of the variable contributing least of all in distinguishability of the entities.

### 2.2.1 Metric choice

The list of triplets accompanied with the frequency of each entry makes frequency dictionary \( W_{(3,1)} \); let \( t = 1 \), at the moment. Hence, a dictionary is a point in metric space; obviously, one may define metrics in a number of ways, in such space. For the purposes of further analysis, we use the Euclidean metrics:

\[
\rho(W_i^j, W_j^k) = \sqrt{\sum_{\omega=\text{AAA}}^{\text{TTT}} (f_i^{j\omega} - f_j^{k\omega})^2}.
\] (2)

Some other metrics might be used, as well. Here \( i \) and \( j \) index two different dictionaries (sequences, respectively).

### 2.3 Chargaff’s imparity index

To begin with, we bring to mind the well-known complementarity pattern established by E. Chargaff in 1952 [21, 22]; it consists in a strong equality of A’s and T’s numbers (C’s and G’s numbers, respectively) counted over DNA molecule. Of course, some minor violations may take place due to mutations; meanwhile the accuracy of this equality is very high. This fact is also known as the first Chargaff’s parity rule.

The second Chargaff’s parity rule stipulates that

\[
n_A \approx n_T \quad \text{and} \quad n_C \approx n_G,
\] (3)

if counted within a single strand. The accuracy of (3) is rather high but varies for different taxa.

Surprisingly, similar to (3) relations are observed for oligonucleotides counted over a single strand. Let us now introduce some rigorous definitions and notions.

Definition 1. Consider a string \( \omega = \nu_1\nu_2...\nu_q \) be an oligonucleotide of the length \( q \), where \( \nu_j \) is nucleotide occupying the \( j \)-th position. Palindrome is the word \( \omega^* = \nu_1^*\nu_2^*...\nu_{q-1}^*\nu_q^* \) read equally in the opposite direction: \( \nu_j = \nu_{q-j}^* \).

Definition 2. Two strings \( \omega \) and \( \overline{\omega} \) make the complementary palindrome, if they are read equally in the opposite directions, with respect to Chargaff’s complementarity rule:

\[
A \leftrightarrow T \quad C \leftrightarrow G.
\]
Hence, \( \forall j, 1 \leq j \leq q \text{ } \nu_j \mapsto \nu_{q-j+1}^* \). Here are some examples of complementary palindromes:

\[
\text{ACT} \Leftrightarrow \text{AGT}, \quad \text{ACTGG} \Leftrightarrow \text{CCAGT}, \quad \text{ACGT} \Leftrightarrow \text{ACGT}.
\]

So, the generalized second Chargaff’s rule stipulates equality (or proximity, to be exact) of frequencies of two strings comprising complementary palindrome [23–33]. Surely, one hardly could expect to get the absolute equality of the frequencies of any two strings comprising complementary palindrome. There is a number of reasons standing behind the violation of such absolute equality; they range from purely combinatorial [25–27, 34] and/or finite sampling effect to biological peculiarities [24, 28, 30, 33].

To reveal the difference between genetic entities or biological objects, one must introduce a measure of the violation of the generalized second Chargaff’s rule; one may do it in various ways; we use the discrepancy index:

\[
\mu(W^q_{\bar{\omega}}, W^q_{\omega}) = 4^{-q} \cdot \sqrt{\sum_{\omega \in \Omega} (f_\omega - f_{\bar{\omega}})^2}.
\]

Here \( \Omega \) is the set of strings of the length \( q \) observed in two sequences (\( i \) and \( j \), respectively), \( \omega \) enlists all the strings, and \( \overline{\omega} \) is the string complementary palindromic to \( \omega \). Normalization factor \( 4^{-q} \) is introduced to equalize the figures (4) observed for various \( q \).

The index (4) measures the discrepancy between two dictionaries \( W^i_q \) and \( W^j_q \). Meanwhile, this index could be applied for a single frequency dictionary \( W_q \):

\[
\mu^*(W_q) = 2 \cdot 4^{-q} \cdot \sqrt{\sum_{\omega \in \Omega^*} (f_\omega - f_{\omega^*})^2}.
\]

Here the complementary palindromic couples are combined from the strings belonging to the same frequency dictionary \( W_q \).

The discrepancy measure (4) looks like Euclidean distance, while it is not. More exactly, it could be considered as a metrics in Euclidean space. To do it, one must reconsider a point in a couple, changing it for the dual one that is a complementary palindrome.

The inner discrepancy measure (5) definitely is not a distance, since it characterizes a single object, not a couple.

### 2.4 \( W_{(3,3)} \) and \( W_3 \) dictionaries

This is a very common fact that a genome comprises coding and noncoding regions. Basically, they differ in the statistical properties manifested in triplet frequency dictionaries. One might detect some minor difference in \( W_3 \) composition developed for coding vs. noncoding regions. Significantly greater difference between these two types of genome parts is observed for \( W_{(3,3)} \) dictionaries [2–4].

Dictionary \( W_3 \) is uniformly defined, for any sequence. The situation differs for \( W_{(3,3)} \) dictionaries. Consider a sequence \( \xi \) of the length \( N \). Starting to cover the sequence with the frames of the length 3 moving along the sequence with the step 3, one may get three different dictionaries, in dependence to the location of the start point. The starts may be located at the first nucleotide of a sequence, at the second nucleotide, and at the third nucleotide; thus, three different triplet frequency dictionaries \( W_{(3,3)} \) could be obtained.
The key difference between coding and noncoding regions consists in the deviations between these three dictionaries. In other words, let the sequence \( \mathcal{L} \) falls entirely into a noncoding region of a genome. One may develop three triplet frequency dictionaries \( W_{3}^{j}, 0 \leq j \leq 2 \) corresponding to three positions of the reading frame shift (these are 0, 1, and 2). The key issue is that these three dictionaries:

1. Differ significantly if developed for coding and noncoding regions.
2. Differ each other, if developed for a coding region.
3. Differ between them negligibly, if developed for a noncoding region.

In other words, consider a set \( W_{3}^{j} \), \( 0 \leq j \leq 2 \) developed over a noncoding region and a set \( \tilde{W}_{3}^{j} \), \( 0 \leq j \leq 2 \) developed over a coding region. Then, \( \forall j \) the difference between \( W_{3}^{j} \) is rather small, when expressed in any way (as Euclidean distance, entropy, mutual entropy, etc.; see also [7, 8]), but the difference between \( \tilde{W}_{3}^{j} \) is significantly greater. Besides, \( \forall i, j \) the difference between \( W_{3}^{i} \) and \( \tilde{W}_{3}^{j} \) manifests apparently. These deviations in statistical properties of such triplet frequency stand behind the Hidden Markov Model methodology [35, 36].

We shall explore structuredness in transcriptomes through the analysis of those triplet dictionaries developed over the individual transcripts.

### 2.5 Relative phase

To reveal the inner structuredness of a (bacterial) genome, Gorban and coauthors have introduced special construction that might be called tiling [2–4]. The idea was to cover a genome (considered as a symbol sequence from ) with a set of overlapping and ordered windows called tiles. All tiles are of the same length \( L \) (\( L = 603 \) in [2–4, 16, 17]); the tiles are located along a sequence with the permanent step \( P \). In the papers mentioned above, \( P = 11 \), and the choice of the specific figures of \( L \) and \( P \) is determined by the specific task of a research.

A subsequence identified by a specific tile is then converted into frequency dictionary \( W_{3} \), and the inner structuredness of a genome is represented through the distribution of the points corresponding to tiles, in 64-dimensional (or 63-dimensional) metric space.

This structuredness is basically determined by the so-called relative phase of a tile. It may:

1. Fall completely into a coding region.
2. Fall completely outside a coding region.
3. Contain a border between coding and noncoding regions.

In any chance, the relative phase indicates whether the start of a tile coincides with a start of a coding region or not. There are following combinations determining the relative phase index:

1. Start of a coding region coincides to the start of a tile. In this case relative phase \( \delta = 0 \).
2. Start of a coding region does not coincide to the start of a tile, and the reminder of the division of the distance (expressed in number of nucleotides) from the start of the tile, and the start of coding region is 1. Then \( \delta = 1 \) in this case.

3. Finally, the start of a coding region falling inside the tile does not coincide to the start of a tile, and the remainder is 2. Then \( \delta = 2 \) in this case.

For any tile covering a noncoding region, \( \delta = 4 \), by definition.

It should be stressed that genes (or coding regions) may take place in opposite strands; in such capacity, the relative phase index must be defined for leading strand and lagging one, separately, where the reminder of the division must be determined for the difference between the last symbol of a tile and the last nucleotide of a gene annotated in a sequence as located in the lagging strand. Thus, seven figures of the relative phase index \( \delta \) are possible: \( F_0, F_1, \) and \( F_2 \) for the tiles containing coding regions from the leading strand; \( B_0, B_1, \) and \( B_2 \) for the tiles containing coding regions from the lagging strand; and, finally, \( J \) labeling the tiles covering noncoding regions, only.

For genome tiling (see [2–4, 16, 17]), the labeling of tiles with the relative phase index is based on genome annotation.

### 2.5.1 Transcriptome relative phase

The situation is slightly different for transcriptome (and the transcriptomes of \( L. \) sibirica Ledeb. and \( P. \) sibirica Du Tour, specifically). First of all, we did not develop any tiling, for transcripts; reciprocally, the transcripts themselves have been considered as tiles. It means that each transcript was converted into \( W_{(3, 3)} \) frequency dictionary as a whole, with no dissection into tiles.

Each frequency dictionary corresponding to a specific transcript was labeled with relative phase index; the labeling procedure was pretty close to that one described above, with few exceptions. We used TransDecoder™ software to find the start of a coding region within a transcript, as well as the strand location of CDS.

The relative phase index for transcripts containing a single CDS was determined in completely the same way, as described above. The transcripts bearing no CDS, if any, have been labeled with index \( J \). Finally, the problem arose from the transcripts bearing several CDS: obviously, a relative phase index is defined ambiguously for such transcripts. In such capacity, we labeled the transcripts with multiple CDS with special figure \( M \) of the relative phase index.

Finally, we have calculated the standard deviation for each triplet, over the entire set of transcripts; that is CGT with \( \sigma_{\text{CGT}} = 0.005586 \), so we excluded this triplet from the set of variables to cluster the transcripts. Reciprocally, the triplet with \( \sigma_{\text{TGA}} = 0.014924 \) yields the maximal figure of the standard deviation.

Similar figures determined for \( P. \) sibirica are \( \sigma_{\text{GCG}} = 0.005658 \) and \( \sigma_{\text{TGA}} = 0.014936 \), correspondingly; the former stands for the minimal standard deviation figure, and the latter stands for the maximal one. Hence, in cedar transcriptome, we have excluded GCG triplet. Remarkably, the triplets with the largest standard deviation figures coincide, for these two genetic entities.

### 3. Results

Previously, seven cluster symmetric patterns have been reported [2–4], in bacterial genomes. Later, similar (but not equivalent) structures were found in chloroplast genomes [16, 17]. First of all, the tiles corresponding to specific relative phase
tend to aggregate into clusters apparently seen in the projection into three principal components with the largest eigenvalues. The points corresponding to specific strand (either leading or a lagging one) perform a triangle, in the frequency space; the points corresponding to noncoding regions tend to gather into a ball-like structure located in the central part of the pattern.

The patterns described in [2–4, 16, 17] are provided by the interplay of two triangles and the central ball. The triangles comprise the points corresponding to specific strand. There are two basic symmetries found in these triangles: the former is a shift (rotational) symmetry peculiar for bacterial genomes [2–4], and the latter is mirror symmetry peculiar for chloroplasts [16, 17]. The ball comprise the points corresponding to the tiles with noncoding regions inside (chloroplast genomes have one more cluster called tail; meanwhile, it is not important at the moment).

Whether a pattern would have four or seven clusters depends on GC content of a genome, for bacteria [2–4]. This figure almost completely determines the mutual location of the planes comprising the triangles formed by the clusters belonging to the same strand. There are some exclusions from this rule, for cyanobacteria. Chloroplasts exhibit mirror symmetry in the strand-specific triangles, so they always have a four-beam structure, where the triangles occupy the same plane with obligatory coincidence of F2 and B2 phases [16, 17].

3.1 Phase index coloring agreement

To make the presentation of results clearer, let us fix the color and label mark usage for transcripts to be shown in figures everywhere further. Indeed, we should distinguish eight different phases in the figures: F0, B0, F1, B1, F2, B2, mult, and noCDS.

To do that, we shall use the following labels: all phases of F0 through F2 of transcripts from the leading strand are marked with triangles; all phases of B0 through B2 of transcripts from the lagging strand are marked with diamonds; mult transcripts are marked with teal squares; finally, the transcripts where no CDS have been found are labeled with brown circles.

Besides, the relative phases of single CDS transcripts are colored in the following manner: F0 is purple triangle, F1 is lime triangle, and F2 is yellow triangle; reciprocally, B0 is magenta diamond, B1 is azure diamond, and B2 is sand diamond.

We should say few words concerning the distribution of the transcripts with several CDS detected in them. For both transcriptomes, the distribution of such transcripts in the 63-dimensional space seems to be very homogeneous; in other words, these transcripts do not form any specific cluster, neither they are attracted to any other given one provided by the transcripts with specific (and unambiguous) relative phase index. The same is true for both studied transcriptomes. Later we discuss this point in more detail, while here we fix that the points representing such multi-CDS transcripts are erased from the pictures illustrating the results.

Thus, the clusters formed by transcripts of the same relative phase index are located in two parallel planes (in the space of three principal components with the largest eigenvalues). This observation holds true for L. sibirica transcriptome, while P. sibirica transcriptome exhibits some deviations from this pattern. We should discuss it later in more detail.

3.2 L. sibirica transcriptome octahedron

Unlike the tiles developed for a genome, the transcripts of a transcriptome exhibit an ultimate pattern, that is, octahedron. The rectangular triangles, $\Delta ABC$ and $\Delta A\beta\gamma$, in Figure 2 occupy the position in two orthogonal planes. Note, these
triangles do not comprise the clusters from the same strand; on the contrary, phases over the octahedron are distributed in the manner shown in Figure 2 (right).

Figure 3 shows the distribution of L. sibirica transcripts with relative phase values ranging from $F_0$ to $B_2$; they are colored as described above. This is the distribution in 63-dimensional space (see Section 2.5.1) shown as the projection into two-dimensional plane determined by the first and the second principal components (Figure 3, left) and by the second and the third principal components (Figure 3, right); this right image is rotated for $\pi/4$ angle clockwise.

The transcriptome shown in this figure exhibits clear and unambiguous octahedral pattern in cluster location. It is evident that $F_0$ to $F_2$ phases lay out in a plane and vice versa: the phases from the lagging strand are also laid out in a plane, and these two planes are parallel. It should be stressed that this pattern is observed in the metric space defined by the eigenvectors of the covariation matrix; in other words, the clear and apparent octahedron pattern is observed in affinely transformed space, not in the original one determined by triplet frequency.

Let us now consider the distribution of the points corresponding to noCDS and mult indexes. These two types of sequences differ drastically, in terms of their dispersion over the pattern. The transcripts bearing several CDS (see Table 1) are rather long. The distribution of $W_{(3, 3)}$ of such transcripts is shown in Figure 4; it should be stressed that this is the mutual distribution of all the points, with the

---

**Figure 2.**
Typical distribution of L. sibirica transcripts in 63-dimensional space.

**Figure 3.**
The distribution of L. sibirica transcripts; phases noCDS and mult are erased.
complete set of phase indexes; the only point in this Figure is that the points corresponding to phases \( F_0 \) through \( B_2 \) are erased.

Also, this figure shows the distribution of the transcripts where no CDS have been found (brown circles). The cluster comprising these transcripts is rather remarkable: the transcripts where no CDS have been found behave themselves (in terms of clustering in 63-dimensional triplet frequency space) pretty close to the fragments falling completely into noncoding regions of a genome, when a complete genome is sliced into a set of tiles [2–4, 16, 17]. This observation indirectly (while rather hard) proves the total lack of any CDS in such sequences; otherwise, the corresponding frequency dictionaries never could be gathered in a ball centered at the pattern.

The transcripts with several CDS inside are distributed over the pattern almost homogeneously, including the central spot where the transcripts without CDS are concentrated. Apparently, this fact follows from the multiplicity of CDS in these transcripts: an interplay of different CDS located within a transcript may yield an effective value of its phase index ranging from \( F_0 \) to \( B_2 \), and the impact of those CDS is expected to be rather random.

### 3.3 \( P. \) sibirica transcriptome octahedron

Let us now focus on the peculiarities of the transcriptome of \( P. \) sibirica. First of all, this transcriptome (at least, the part taken into analysis) is less abundant, in comparison to \( L. \) sibirica transcriptome. This fact may impact the pattern of the triplet frequency dictionary distribution, while one may expect the effect to be negligible, since the length distribution of the transcripts of \( P. \) sibirica is similar to that one observed for \( L. \) sibirica (see Figure 1) and the portion of multi-CDS transcripts in these two transcriptomes are quite similar (see Table 1).

To begin with, Figure 5 shows the clustering pattern observed for this transcriptome; the technology of the development of the pattern is absolutely the same, as in Figures 3 and 4. The strongest difference between this transcriptome and the \( L. \) sibirica one consists in the significant deformation of the octahedron observed over \( P. \) sibirica transcriptome; Figure 6 illustrates this point.

At the first glance, the pattern shown in Figure 5 looks like a tetrahedron, while it is not. In proper projection, the pattern looks like a hexagon; adding the subset of multi-CDS transcripts, one gets the same pattern almost homogeneously covered by the point corresponding to the subset.
4. Discussion

The patterns provided by the distribution of considerably short fragments of a genome may tell a lot to a researcher [2–4, 16, 17]. For bacteria, GC content seems to be the key factor determining the details of the pattern [2–4]. That is not so for chloroplasts, mitochondria, and cyanobacteria [16, 17]. The results presented above show that GC content has nothing to do with a pattern observed over a transcriptome. Hence, a question arises toward the key factor determining the specific type of a pattern. Yet, there is no simple and brief answer, while Chargaff’s parity rule discrepancy may be quite informative here.

We have determined Chargaff’s rule discrepancy measure (5) figure $\mu^*$ for all six clusters observed in L. sibirica and P. sibirica transcriptomes; Table 2 shows them. The variation of these figures $\mu^*$ is very smooth, and the clusters are pretty close to each other, in terms of the discrepancy $\mu$ (see Eq. (5)). This fact opposes to similar observations carried out over bacterial, chloroplast, and mitochondrial genomes [16, 17]: these later exhibit significant (more than 10 times) difference in the discrepancy figures calculated for the clusters. It should be said that, unlike transcriptomes, chloroplast genomes exhibit three-beam patterns, where a beam (i.e., a cluster) comprises the fragments from forward and backward strands, simultaneously. There is no such combination, for transcriptomes.

Let us now focus on a few more details on Chargaff’s imparity index, itself. The index value differs for different length $q$ of words. Thus, a question arises toward the reference figures for this index. Suppose, the index is determined over the frequency dictionaries derived from both strands; in such capacity, it must be equal to zero.
**Table 2.**
Discrepancy measure (5) figures $\mu^*$ for two transcriptomes (upper part) and cluster radii, for the same phases (lower part).

Calculating the index (4) over a single strand, one may clearly understand to what extend a strand looks like the opposite one, in terms of the word frequency [23–25]. For random non-correlated sequence with $f_A = f_T$ and $f_C = f_G$ ($\mu_q = 0$). Hence, $V_q, H_q$ figures remain the same, if the discrepancy $\mu_1$ is fixed [23].

Unlike $\mu^*$ figures, the radii of these six clusters exhibit quite diverse behavior. The radius of a cluster is an average distance from the center (that is arithmetic mean) determined over the cluster to each point from the cluster. Lower part of Table 1 shows the radii figures. The radii figures are apparently different, for the transcriptomes under consideration. $F_0$ and $F_1$ phases for *L. sibirica* show extremely high values. These figures may not be explained by the excess of the cluster abundance of *L. sibirica* in comparison to *P. sibirica*. Again, the variation of the radii for *L. sibirica* is evidently greater than for *P. sibirica*, and this fact correlates to the mirror symmetry of *P. sibirica* transcriptome, since it is typical for simpler and less diverse genetic system.

Inter-cluster discrepancy measure $\mu$ is of great interest, for both cases; Table 3 shows these indexes. Careful examination of Table 3 allows to identify three couples of relative phase indexes with distinctively lower figure of (4), namely, the couples:

\[ F_0 \leftrightarrow B_2 \quad F_1 \leftrightarrow B_0 \quad F_2 \leftrightarrow B_1. \] (6)

**Table 3.**
Discrepancy measure (4) figures $\mu$ determined within each of the two transcriptomes.
Evidently, the phases in these couples yield two different types of symmetry: the first one is shift, and the second symmetry is mirror. The situation is opposite for *P. sibirica* transcriptome: the couples with the least Chargaff’s discrepancy measure (4) are the following:

\[
F_0 \Leftrightarrow B_1 \quad F_1 \Leftrightarrow B_0 \quad F_2 \Leftrightarrow B_2. \tag{7}
\]

To make the situation with symmetries clear, we show the clusters over the elastic map shown in the so-called *inner coordinates*; Figure 7 presents the transcriptomes.

Such mirror symmetry has been previously reported for chloroplast genomes [16, 17] (see also [23, 37, 38]); yet, there were no other but the chloroplast genomes exhibiting such mirror symmetry, and *L. sibirica* transcriptome is the next one in this point.

Definitely, the coincidence of these two symmetrical patterns does not mean that *L. sibirica* transcriptome is identical to a chloroplast genome in all other properties. Probably, plants differ from other eukaryotic organisms and bacteria in the symmetry type; currently, no eukaryotic genome is found with mirror symmetry. Shift symmetry observed for *P. sibirica* transcriptome poses a question toward the origin of the symmetry type change: whether it results from some essential biological difference between these two pine species or it is a manifestation of the genomic transformation in witch’s broom cells. To answer the question, more studies are necessary.

The most amazing thing in transcriptome statistical properties is that it yields an octahedral pattern, unlike bacteria, organelle, and other genetic entities (say, yeast genomes). Another point is that the pattern does not depend on the length of transcripts taken into consideration: we have examined separately the subsets of transcripts as long as \(200 \leq N \leq 600\) bp, \(600 \leq N \leq 2500\) bp, and those longer \(3000\) bp. All these subsets yield similar pattern, with very minor variation mainly manifesting in cluster density.

One can easily see two major peculiarities differing a transcriptome from the sets of tiles described above (see [2–4, 16, 17] for details). These are:

- Total absence of the (rather extended) noncoding regions.
- Elimination of introns from the statistical analysis of sequences.
Of course, the first item from this list is quite arguable: a number of transcripts where no CDS has been detected bring a direct and unambiguous disproof of it. Thus, the question arises, whether these transcripts are similar, in some sense, to the fragments of genome comprising purely noncoding regions of the latter.

We have examined the first hypothesis through the simulation of noncoding regions. To do that, we have added a number of $W_{(3,3)}$ frequency dictionaries obtained from the tiles covering the noncoding parts of genomes of several other organisms. All the tiles were as long as 603 bp and contained noncoding regions, exclusively. The number of dictionaries (the points, in other words) varied from one third to one half of the total number of transcripts in the set. By assumption, this addition simulated a genome.

Upon addition, we expected to see a pattern similar to that one observed in bacteria, organelle, or other eukaryotic organisms; the octahedron pattern appeared to be stronger. Figure 8 obviously disproves this hypothesis: it shows the same transcriptome ($L. sibirica$) with eliminated transcripts bearing no CDS, where a set of $W_{(3,3)}$ dictionaries borrowed from three different genomes is added.

![Figure 8](image_url)

*Three noncoding data points added to L. sibirica transcriptome; nothing happened.*
consequently. Obviously, such simulation of a genome does not break down the observed pattern of transcript distribution. Yet, one more option should be examined: what happens if the natural noncoding regions are used to simulate a genome? In other words, the pattern might be sensitive to the noncoding regions from the original genome, only. This point still awaits for examination.

The impact of introns on the alteration of the observed pattern is less evident. Moreover, one faces greater difficulties in revealing it. One might want to compare the distributions developed over \( W_{3,3} \) and \( W_3 \) dictionaries, in this case; yet, this problem needs careful investigation and falls beyond the scope of this paper.

5. Conclusions

Systematic comparison of (rather short) fragments of permanent length formally identified within a genome reveals a symmetry in the distribution of the triplet frequency dictionaries obtained over those fragments; originally this effect has been found on bacterial genomes. Later similar (while rather different in a number of essential details) behavior has been found for chloroplasts and mitochondria genomes. The general pattern of the distribution looks like a superposition of two triangles where the vertices correspond to the fragments of the same relative phase. In simple words, it corresponds to a reading frame shift, in case of a translation-like processing of DNA sequence.

A transcriptome itself might be considered as a set of those fragments, with few exclusions. Firstly, the lengths of transcripts are different and may affect the expected pattern. Secondly, there are no fragments in a transcriptome corresponding to those obtained from noncoding (intergenic) regions of a genome. This fact results in ultimate possible configuration of the clusters corresponding to the transcripts with the same relative phase index, that is, octahedron. All these patterns could be seen in the space of three principal components with the largest eigenvalues. The \( L. sibirica \) transcriptome yields almost perfect octahedral pattern, while the \( P. sibirica \) transcriptome differs rather significantly, with planes comprising the clusters from the same strand to be located almost in parallel. This deformation might result from the biology: we studied the \( P. sibirica \) transcriptome obtained not from a normal tree, but from a witch’s broom bud; the latter is known for extremely deviated morphology that may not avoid serious genetic alteration in its genome.

Acknowledgements

The data used in this study were obtained under the grant 14.Y26.31.0004 from the Russian Government. The authors also thank Serafima Novikova from Siberian Federal University for the helpful discussion.

Conflict of interest

The authors declare no conflict of interest.
Author details

Sadovsky Michael¹,²*, Putintseva Yulia², Biryukov Vladislav²
and Senashova Maria¹

¹ Institute of Computational Modeling SB RAS, Krasnoyarsk, Russia
² Siberian Federal University, Institute of Fundamental Biology and Biotechnology, Krasnoyarsk, Russia

*Address all correspondence to: msad@icm.krasn.ru

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. [CC BY]

Transcriptome Analysis

References


Revealing the Symmetry of Conifer Transcriptomes through Triplet Statistics

DOI: http://dx.doi.org/10.5772/intechopen.88696
References


[33] Nikolaou C, Almirantis Y. Mutually symmetric and complementary triplets:


Transcriptome analysis is the study of the transcriptome, of the complete set of RNA transcripts that are produced under specific circumstances, using high-throughput methods. Transcription profiling, which follows total changes in the behavior of a cell, is used throughout diverse areas of biomedical research, including diagnosis of disease, biomarker discovery, risk assessment of new drugs or environmental chemicals, etc. Transcriptome analysis is most commonly used to compare specific pairs of samples, for example, tumor tissue versus its healthy counterpart.

In this volume, Dr. Pyo Hong discusses the role of long RNA sequences in transcriptome analysis, Dr. Shinichi describes the next-generation single-cell sequencing technology developed by his team, Dr. Prasanta presents transcriptome analysis applied to rice under various environmental factors, Dr. Xiangyuan addresses the reproductive systems of flowering plants and Dr. Sadovsky compares codon usage in conifers.