NEXT-GENERATION SEQUENCING AND SEQUENCE DATA ANALYSIS

Authored By

Kuo Ping Chiu
Associate Research Fellow/Associate Professor
Genomics Research Center, Academia Sinica
National Taiwan University
National Central University
Taiwan

including, without limitation, special, incidental and/or consequential damages and/or damages for lost data and/or profits arising out of (whether directly or indirectly) the use or inability to use the Work. The entire liability of Bentham Science Publishers shall be limited to the amount actually paid by you for the Work.

General:

1. Any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims) will be governed by and construed in accordance with the laws of the U.A.E. as applied in the Emirate of Dubai. Each party agrees that the courts of the Emirate of Dubai shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims).
2. Your rights under this License Agreement will automatically terminate without notice and without the need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.
3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Ltd.
Executive Suite Y - 2
PO Box 7917, Saif Zone
Sharjah, U.A.E.
Email: subscriptions@benthamscience.org

CONTENTS

ABOUT THE AUTHOR

Kuo Ping Chiu is currently an Associate Research Fellow at Academia Sinica with joint appointments with National Taiwan University (NTU) and National Central University (NCU). He received his PhD in Microbiology from UC Davis in 1991 and completed his postdoc at Harvard Medical School on Neurosciences during 1993 - 1996. Kuo Ping's research career is closely associated with biotechnology. His PhD research focused on intracellular amplification of mouse mammary tumor proviral DNA using *in situ* PCR to identify the infected cells, while his postdoctoral training was related to multiple colorimetric labeling of acetylcholine receptor subunit transcripts to study their coordinated expression pattern. These trainings have critical influence on his academic research career and industrial experience. His industrial experience started with a job at Bio-Rad Laboratories where he developed protocols and kits for antimicrobial susceptibility testing using flow cytometry (5/1996-4/1998). Later he switched from wetlab to Bioinformatics and worked for Genome Institute of Singapore on developing Paired-End diTag technology and methods for sequence data analysis (8/2002 - 8/2008). He moved back to Taiwan in 2008 working for Academia Sinica on developing DNA sequencing-related biotechnologies and studying gene expression and regulation in normal and cancer cells. He also teaches sequencing technologies, sequence data analysis and pathway analysis in a number of national universities including NTU, NCU, and National Yang-Ming University (NYMU). He is currently holding three US patents related to paired-end ditag technologies.

FOREWORD

I have known Kuo Ping for almost thirty years. He is very fond of science, especially in developing or employing novel technologies for biological investigations. He and I often share ideas and have worked on common scientific interests, including the project to study the early anti-cancer activity of the medicinal fungus *Antrodia cinnamomea* using Next-Generation Sequencing (NGS) to sequence and compare the transcriptomes of cancer cells treated with or without the extract of *A. cinnamomea*. This work led to the discovery that this medicinal fungus was able to globally collapse the miRNA system.

It is his never-resting mind that has provided him a motivation to start writing this book some years ago, while he was still working for Genome Institute of Singapore when NGS was starting to take shape. During that time, there was nowhere to find a book to introduce advanced sequencing technologies and the potential applications to either students or laymen. This book evolves from his previous work on paired-end ditag technology and has a strong association with NGS which can be considered as the most fascinating technology of the 21st century. I am glad to see him accomplish this hard task. This book seems to be a good textbook for graduate students and a handbook for researchers who are interested in sequencing and biotechnologies.

It is worth mentioning that Kuo Ping is a scientist with a strong passion for outdoor activities. We did some interesting, and sometimes funny, things together. Back in California, we spent some weekends fishing at Pacifica Pier for perch and crabs, and then barbecued and talked over beer in my backyard. Here in Taiwan, we often tour around his "secret gardens" where he grew trees and raised domestic fish in a pond. The most interesting outdoor experience was a trip to find fossil stones in a wild countryside, where we found forgotten creeks running through the hillside. We tied a very heavy fossil stone with rope and carried it with a bamboo stick across water streams, making loud laughter in the wild. I actually believe that his passion for outdoor activities has contributed to the completion of this useful book.

Miao-Lin (Merlin) Hu, Ph.D.
Honorary Distinguished Professor,
National Chung Hsing University
Department of Food Science and Biotechnology
250 Kuo Kuang Road, Taichung 402
Taiwan

PREFACE

Many biological fields, including genetics, immunology, genomics, and epigenomics, can be readily integrated with bioinformatics and sequencing. In fact, sequencing is becoming an indispensable tool for biological and medical investigations, especially for the study of cancer, genetics, epigenetics, immunology, and developmental biology. Foreseeing the potential applications of sequencing technologies, we have initiated a sequence data analysis course at the Institute of Zoology, National Taiwan University to introduce sequencing technologies, genomics, epigenomics, bioinformatics, and biotechnologies to graduate and undergraduate students. At the same time, I started organizing the teaching materials to compose this book. Most of the materials were derived from my research at Academia Sinica and my teaching at National Taiwan University (NTU), National Central University (NCU), and National Yang Ming University (NYMU). Additionally, there are some from my previous work in the United States and Singapore.

Besides college students, this book is also written for people who already have some biological background, or with an interest in knowing more about DNA sequencing and its applications. With this effort, we hope to help develop sequence-associated knowledge among people from all walks of life, including researchers, professionals, and amateurs working in biology-related or -unrelated fields, with or without an ambition to apply sequencing technologies to biological or medical investigations. Extracting biological meaning from large quantities of sequence data is an art. We intend to help readers appreciate the various types of sequencing technologies, and learn how to use sequencing technologies to unravel the mystery of the biological system.

Sequence data analysis is not only an art, but also a tool to help researchers develop a constructive philosophy. When a diploma is awarded to a candidate, it may be treated by the recipient as a degree which can then help the graduate to find a better job, or achieve a higher social status. Some recipients may take it more seriously and use it for further, and possibly more advanced, studies. Similarly, sequence data analysis can be treated as either a course to be accomplished in class, or, alternatively, a process which will be a guide to the further understanding of the life of nanoscopic molecules within a cell. We gain many more experiences through the years of working in a lab, and, with those proficiencies, we build our philosophy. Sequencing and sequence data analysis can help us make sense of the activities which take place on a daily basis in the molecular world of our bodies, even though these activities are impalpable to the naked eye. Being able to “see” things that cannot “normally” be seen can help us build a philosophy, in which we travel between the macroscopic and molecular worlds.

A sincere attitude towards sequence data analysis is essential. Since omics sequence datasets are normally at large volumes, these data cannot be handled by traditional means, but by

computer programs. Frequently we find bugs in computer programs which may produce erroneous or misleading results. A bug is a bug. No matter whether it is big or small, it has to be removed, so as to obtain the real molecular status of interest. Moreover, the author would like to emphasize the importance of practical exercise. Students are strongly encouraged to personally construct sequencing libraries, run sequencing, and analyze sequence libraries or related libraries whenever possible.

Although this book carries a mission, readers can treat it either as a novel, or a science fiction story. The development of sequencing technologies per se is science fiction, isn't it? With limitations in man power and time, the contents of this edition may not be able to completely satisfy serious readers. Your opinions and suggestions are warmly welcomed!

Kuo Ping Chiu, Ph.D.
Genomics Research Center, Academia Sinica
National Taiwan University
National Central University
Taiwan

ACKNOWLEDGEMENTS

I would like to thank Grace Chu-Fang Lo, former Dean of the College of Life Science (CLS) of NTU (National Taiwan University), and Director Pan from the Institute Zoology of NTU for their support on running a course on sequence data analysis at NTU. Although most people have realized the importance of building high throughput sequencing facilities in Taiwan, financial supports remain very limited. Professor Lo was among the few scientists who can fully appreciate the effort to build a NGS sequencing facility at NTU campus. I also like to express my special thanks to Hong Sain Ooi and Prof. Xiaodong Zhao, my longtime friends since we worked together back in Singapore, for their support on RNA-Seq analysis and friendship. Academia Sinica (AS) president CH Wong, AS Vice President CJ Chen, AS Genomics Research Center (GRC) Vice Director Alice LT Yu and Prof. ML Hu from National Chung Hsing University also provided great help on my career at Academia Sinica. Prof. H.T. Yu from the Institute of Zoology of NTU, CLS secretary Ho, my friend Amy Y.M. Chou from biotech industry, Dean Pei from National Pingtung University of Science and Technology and many of my colleagues at AS and NTU have also been very kind and helpful. Thank all of you!

CONFLICT OF INTEREST

The author confirms that this ebook contents have no conflict of interest.

Part 1

Background Introduction

CHAPTER 1

The Ultimate Frontier of Knowledge: the Mysterious Genomes and the Gene Expression and Regulation in the Upstream of Biological Information Flow

Abstract: DNA sequencing is just a tool that we can employ to study biological phenomena. It can be useful for us to review some biological background before we discuss how to use the tool. It will help us to understand what subjects in the biological field DNA sequencing can be applied to and how to apply sequencing technologies in the study. Before we discuss the applications of DNA sequencing, Let's review some molecular biology, starting from the structure and organization of information molecules at the molecular level.

Keywords: Alternative splicing, Central dogma, Chromatin, DNA packaging, Epigenetic modifications, Euchromatin, Gene, Genome, Heterochromatin, MicroRNA, Nucleosome, Transcription factors.

Definition of Terminologies

Genome

A genome is a complete set of genetic material in a cell. Prokaryotic cells do not have nuclei or mitochondria, while eukaryotic cells have both. The genome of an E. coli cell is a circular chromosome, while a eukaryotic cell normally comprises of a set of linear nuclear chromosomes together with a circular mitochondrial chromosome (chrM).

Chromatin

A chromatin is an interphase chromosome, comprising DNA and proteins bound to the DNA. Sometimes chromatin and chromosome are interchangeably used in the book.

Central Dogma

The classic view of gene expression from DNA to RNA, and then to protein.

Gene

In a broad sense, a gene is a genomic region capable of being transcribed into mRNA (protein-coding messenger RNA) or non-coding RNA. In other words, a gene is a piece of molecular information and a transcription unit, prepared and stored in the genome for

Kuo Ping Chiu

being transcribed into mRNA, which has a protein-coding capability, or a non-coding RNA, which potentially has regulatory or structural function. According to gene nomenclature, gene names should be in lower case and italicized, but sometimes the first letter can be capitalized. On the other hand, protein names should have the first letter or the whole name capitalized and should not be italicized.

Transcript
The product of transcription.

Non-Coding RNA
Synonymous to non-protein-coding RNA, a non-coding RNA (ncRNA) is a RNA transcript which is not destined for translation. In contrast, a protein-coding RNA is a message molecule capable of being translated into a protein. ncRNAs include rRNA, tRNA, miRNA (microRNA), siRNA, snRNA, piRNA, etc.

Locus (plural: loci)
Simply means a genetic location.

Homolog
A gene (or protein) homolog is a nucleotide (or amino acid) sequence related to another nucleotide (or amino acid) sequence by at least partial sequence homology.

Ortholog
An ortholog is a common ancestor-derived transcription unit found in multiple species. Normally they have the same, or similar, functions.

Paralog
Gene paralogs are normally generated by gene duplication. Due to the redundancy in structure and function, gene paralogs may diversify through evolution.

Gene Isoforms
Gene isoforms are transcription units encoding proteins with similar functions and thus can be assigned to the same position of a pathway.

Scales Used in the Molecular World
Angstrom (10^{-10} meter, about the diameter of a hydrogen atom), nano-meter (10^{-9} m, or ~10 H atoms), micro-meter (10^{-6} m), *etc*.

INTRODUCTION

Transcription, or gene expression, represents the forefront of biological information flow. Information flow at this stage is carried out predominantly by chromatin-associated activities inside the nucleus. The mRNA molecules are subsequently translocated from the nucleus to the cytoplasm and get translated by ribosomes into proteins. Gene

expression in the upstream is regulated in a hierarchical manner through the interaction between various biological molecules, especially RNAs and proteins. Conceivably, information molecules, such as DNA and RNA, are appropriate for DNA sequencing when studying gene expression and regulation.

MICROSCOPIC STRUCTURE AND ORGANIZATION OF INFORMATION MOLECULES

Chromatins are packed into distinguishable domains with variable degrees of condensation. Each of these domains, either heterochromatins or euchromatins, may occupy a large portion of the chromatin. Moreover, some domains may switch from heterochromatin to euchromatin, or *vice versa* , and are thus defined as facultative heterochromatin. Heterochromatin domains are condensed and more likely to be located underneath the nuclear envelope (Fig. **1**), although some are dispersed across the nuclear matrix. On the other hand, euchromatin domains are loosely packed and are most likely to be found in the interior of the nucleus or near the nuclear pores.

Genes within the heterochromatin domains are transcriptionally inactive, while those in the euchromatin domains are accessible for regulatory proteins and transcription machinery. It is conceivable that genes encompassed in the heterochromatin domains are correlated with tissue type as determined during development and differentiation. Moreover, both the content of heterochromatin-encompassed and the euchromatin-encompassed genes can be dynamically influenced by intracellular and extracellular signals. Induced pluripotent stem cells (iPS) are a good example.

When compared to heterochromatin, euchromatin is gene-rich, more open to transcription machinery, while heterochromatin is condensed chromatin only visible under a microscope. Heterochromatin represents an inactive state of interphase chromatin and exists in every centromere and the inactivated X-chromosome. Some euchromatic regions become heterochromatic in later life (implicated in development and aging). Heterchromatin contains very few genes (most in the facultative heterochromatins) and favors HP1 binding. Heterochromatins can be either facultative (reversible, *e.g.* X inactivation) or constitutive (fixed and irreversible). Inactivation of one of the X chromosome is mediated by epigenetic modification. The inactivated X chromosome varies from one cell to another. Thus, a female “tissue” is a mosaic construct containing a mixture of gene products from both X chromosomes. Activated genes move from the peripheral nuclear region to the interior region, or from the facultative heterochromatin region to the euchromatin region and require chromatin remodelers, all containing multiple subunits, for activation (Narlikar *et al*. 2002).

CHAPTER 2

Evolution of DNA Sequencing Technologies

Abstract: Current nucleotide sequencing focuses on DNA sequencing. As you will see later in this chapter, direct RNA sequencing was de-selected by evolution. Instead, RNA molecules are converted to cDNA and subject to DNA sequencing. Moreover, DNA sequencing can be conducted by a number of sequencing technologies, each of which uses a company- or inventer-defined procedure and sequencing mechanism. By nature, both living organisms and non-living objects are constantly challenged by evolution. DNA sequencing technologies are of no exception. For living organisms, phenotypic variations resulted from genetic alterations are constantly tested by the surrounding environment, which allows the fittest to propagate more efficiently than the others. Similarly, using DNA sequencing technologies that we will discuss later in this chapter as an example, each technology has its pros and cons against one another. Also, their advantages and disadvantages co-evolve with, and depend on their environmental backgrounds. Here, we review the evolution of DNA sequencing technologies to appreciate the evolutionary process eventually leading to the development of Next-Generation Sequencing technologies.

Keywords: Next-generation sequencing, NGS, Sanger sequencing, Single-molecule sequencing.

Definition of Terminologies

Sequencing

Sequencing, in biological terms, refers to the usage of methodologies and/or instruments to determine the order of building blocks in a macromolecule. DNA sequencing determines the order of deoxyribonucleic acid bases (A, T, G, and C) in a DNA molecule, RNA sequencing determines the order of ribonucleic acid bases (A, U, G, and C) in a RNA molecule, and protein sequencing determines the order of amino acids in a protein molecule. The authors will focus on DNA sequencing using the Next-Generation Sequencing (NGS) technologies.

Sanger Sequencing

In short, Sanger sequencing can be defined as the method that uses di-deoxynucleotides (ddNTPs) in partial termination reactions. This approach is readily distinguishable from the NGS sequencing methods which use engineered substrates instead of ddNTPs. Sanger sequencing can be further categorized into manual Sanger sequencing (using radioactive labeling) and automated Sanger sequencing (using four-colored fluorescent labeling in

Kuo Ping Chiu

conjunction with computerized signal-capturing and processing system).

INTRODUCTION

As one may be aware, the application of sequencing technologies is enormous. Genome sequencing has unraveled the genetic sequences of hundreds of prokaryotic and eukaryotic organisms, and many more are on the way. Genome assemblies are used as references for further biological and medical investigations. So far, a number of prokaryotic and eukaryotic genomes have been sequenced. With the momentum provided by NGS technologies, the number of sequenced genomes is increasing dramatically. Sequencing of specific molecular species produced along the gene expression and regulation cascade play an important role in unraveling the entities of the molecular species and the study of vertical and horizontal interactions between molecules. Undoubtedly, these efforts will strengthen our understanding of certain key subjects in the "biological field", including variation in immune repertoire, genetic diversity, developmental process, and diseases.

CHARACTERISTICS OF DNA SEQUENCING

There are evidently clear characteristics for DNA sequencing. These include, but not limited to, 1) completeness, 2) high resolution, and run in quantum jump fashion. By running DNA sequencing base-by-base, we can completely read through a genome. Moreover, DNA sequencing is able to provide a resolution at, and beyond, the single nucleotide level. For example, DNA sequencing allows access to the locations of SNVs across the whole genome. Furthermore, modifications on nucleotide (*e.g.*, DNA methylation) or amino acids (*e.g.*, methylation, acetylation, phosphorylation, *etc*) can be analyzed. In contrast to hybridization, DNA sequencers perform sequencing in quantum jump fashion.

ADVANCES IN SEQUENCING TECHNOLOGIES

In fact, there was RNA sequencing before the commencement of DNA sequencing. RNA sequencing adopted the following procedure: 1) Label RNA sample (*e.g.,* bacterial phage) with radioactivity, (*e.g.,* P^{32}). 2) Treat sample with chemicals and ribonucleases to hydrolyze RNA at specific residues. 3) Run the sample with 2D gel/membrane. 4) Use the amino acid sequence of the corresponding protein as a reference to help understand the RNA sequence. Overall, the procedure is tedious, making it not a competitive methodology (Metzker, 2010).

DNA sequencing prevailed over RNA and protein sequencing due to, at least, the following reasons:

1. All amino acids (about 20 or so) are encoded only by 4 deoxynucleotides through 3-

base coding. This 4-to-20 encoding mechanism makes it easy to convert a nucleotide/deoxynucleotide sequence to its corresponding amino acid sequence, but not the other way around. As such, deoxynucleotides/nucleotide sequencing makes more sense.

2. Technically, DNA sequencing is much simpler than RNA or protein sequencing. DNA polymerase was the only enzyme required for DNA sequencing and, compared to RNA and protein sequencing, preparation of sequencing reagents and setup of reaction conditions are much easier for DNA sequencing.
3. DNA is much more stable than RNA, and as such, is much easier than dealing with RNA.
4. RNA can be easily converted to cDNA (complementary DNA).
5. In general, DNA molecules, either double-stranded or single-stranded, are structurally simpler than single-stranded RNA molecules, which are likely to loop back and form complex structures. Notice that base sequencing requires the target molecules (either in DNA or RNA form) to be single-stranded; double-stranded structures are expected to interfere the sequencing process.
6. Polymerase chain reaction (PCR), which amplifies DNA molecules, dramatically facilitates DNA sequencing.

MANUAL SANGER SEQUENCING

In nature, DNA replication is conducted by DNA polymerase which uses deoxyribonucleotides triphosphates (dNTP, namely dATP, dTTP, dGTP, and dCTP) as building blocks for the synthesis of the complementary strand. This process requires the replication origins (ori) to be opened by helicase and the synthesis be primed (*i.e.* started from specific sites) by "primers". The synthesized strand can elongate only in 5' to 3' direction.

Following the publication of DNA double helix structure in 1953 by Francis Crick and James Watson, Frederick Sanger (Fig. **1**) invented an enzymatic method for DNA sequencing in 1975 - 1977 (Sanger and Coulson, 1975; Sanger *et al*., 1977). He took advantage of this natural system and invented a method for DNA sequencing which helped him to win a Nobel Prize. During the same period of time Maxam and Gilbert invented a chemical method for DNA sequencing (Maxam and Gilbert, 1977), which was de-selected by competition because of its technical complexity.

For the purpose of DNA sequencing, Sanger made a number of modifications. The procedure can be summarized as the following steps:

1. A single species of DNA sample was prepared in single-stranded form, which would be used as templates for the synthesis of the second DNA strand.
2. A sequence-specific oligonucleotide (oligo) was prepared to serve as the 'sequencing primer'. This primer flanks the upstream (5') of the DNA region to be sequenced. Such

CHAPTER 3

Mechanisms of Next-Generation Sequencing (NGS)

Abstract: DNA sequencing consists of a number of methodologies, each adopts a unique process of sequencing mechanisms. During 1970s, Sanger sequencing survived the competition against other approaches and dominated DNA sequencing for a number of decades. As stimulated by urgent demand of high throughput sequencing approaches by the Human Genome Project and various genome projects that followed, Next-Generation Sequencing evolved to replace Sanger sequencing as the main sequencing approach. NGS consists of three major sequencing platforms (*i.e.*, 454/Roche, Solexa/Illumina and SOLiD/Life Technologies) and each has its own sequencing mechanism. These mechanisms have experienced severe competition, leading to the election of Illumina system by the sequencing market as the main stream sequencing platform. Before we can fully appreciate the reasons leading to the success of the Illumina system, here we analyze and discuss the sequencing mechanisms adopted by these NGS platforms.

Keywords: 454, Bridge amplification, Emulsion PCR, *in situ* PCR, Illumina sequencing, Next-generation sequencing, NGS, Solexa, SOLiD, Sequencing-by-synthesis, Sequencing-by-ligation.

Definition of Terminologies

Target (DNA)

DNA molecules to be sequenced

Sequencing adaptors

Short DNA fragments ligated to the ends of the target DNA. It normally contains both PCR primer-binding sites for PCR amplification and sequencing primer-binding sites for sequencing initiation.

Sequencing library

A library that has been made ready for sequencing. At this point, a pair of sequencing adaptors should have been ligated to the ends of the target molecules.

Sequencing run

A complete sequencing of a (sequencing) library.

Sequencing cycle (or chemistry cycle)

A full process required to complete the incorporation of a nucleotide (for sequencing-by-

Kuo Ping Chiu

synthesis sequencers) or an oligo (for sequencing-by-ligation sequencers) into the elongating strand. (A sequencing run comprises a number of sequencing cycles.)

Massively parallel sequencing
Initially used by 454, to mean a massive number of sequencing reactions simultaneously taking place in massively amplified templates in a synchronized fashion.

INTRODUCTION

Changes Made From Automated Sanger Sequencing to Next-Gen Sequencing

1. No more colony picking
2. Clonal *in situ* PCR amplification of templates becomes essential for all NGS sequencing library preparations. Current NGS sequencing technologies can only sequence DNA templates that have been clonally amplified to thousands of copies in each flow cell channel (Illumina sequencers) or millions of copies on beads (454 and SOLiD machines) by *in situ* PCR, so that signals produced from the massive parallel sequencing reactions in each monoclonal template population can reach beyond the detectable level. Please note that, DNA templates need to be reset to a single-stranded state at two stages: 1) before clonal amplification of templates with *in situ* PCR, and 2) before sequencing.
3. ddNTPs are replaced by ddNTP analogs.
4. Fluorescent signals are still used in Solexa and SOLiD systems, but the 454 system uses light emission as its signal.
5. No electrophoresis is needed. Instead, signals are caught by camera directly on spots.
6. No improvement in sequencing accuracy and the sequence length gets even shorter!
7. However, yield reaches the giga-base level and sequencing cost is dramatically reduced.

In situ PCR for Clonal Amplification of Templates

Strategies of *in situ* PCR for clonal amplification of templates vary among sequencers but can be categorized into two types: 1) solid phase PCR amplification by Solexa system, and 2) emulsion PCR (emPCR or ePCR) by 454 and SOLiD systems (see following sections).

NGS Sequencing Mechanisms

NGS sequencers adopt two types of sequencing mechanisms: by synthesis - using DNA polymerase or by ligation - using DNA ligase (Metzker, 2010). Each category can be further divided into subtypes based on the chemistry undertaken. Since the pros and cons of a NGS sequencer are mainly determined by its sequencing mechanism (Table **1**), it is essential for us to discuss the sequencing mechanisms adopted by each NGS sequencer.

Comparison between different NGS systems can be found in a number of informtive review articles although only a few are listed in the references (Mardis, 2008; Metzker, 2010).

Table 1. Two types of NGS sequencing machines/mechanisms.

Sequencer	454 series	GA/Solexa series	SOLiD series
Inventor/manufacturer	454 Life Sciences -> Roche	Solexa -> Illumina	Applied Biosystems -> Life Technologies
First launch	2005	2006	2007
History	The 454 sequencer was invented by Jonathan Rothberg, the founder of 454 Life Sciences. Dr. Rothberg is the pioneer of NGS machines. Roche acquired 454 Life Sciences in 2007.	Solexa , a spin-off company from Cambridge Univ., launched its first NGS sequencer, Genome Analyzer, in 2006. Solexa was purchased by Illumina in 2007.	SOLiD (supported oligo ligation detection) sequencers are manufactured by Applied Biosystems Inc. (AB or ABI), a longtime leader in DNA sequencing.
Sequencing mechanism	Sequencing by synthesis	Sequencing by synthesis	Sequencing by ligation
Direction of elongation	5' -> 3'	5' -> 3'	3' -> 5'
Direction of reading	5' -> 3'	5' -> 3'	5' -> 3' (on the template strand)
Substrates/ building blocks	dNTPs	dNTP analogs (fluorescently labeled and 3' blocked)	oligos
Delivery of substrates/ building blocks	One by one	All together	All together
Signal	Light (Photon)	Fluorescence	Fluorescence
Imaging	Concurrent w/ the release of signal	Colors taken right after the nt is incorporated	colors taken right after the nt is incorporated
Terminate signal after each cycle?	No	Yes	Yes
One base per cycle?	No	Yes	Yes, if based on sequencing run. No, if based on reaction cycle.
Polymer problem	Yes	No	No
Can read through palindrome?	Yes	Yes	No (major drawback for SOLiD systems!)

I. Sequencing-By-Synthesis

This approach is continuous from the Sanger sequencing method and is adopted by 454

CHAPTER 4

Genome Assembly, the Genomic Era and the Rise of the Omics Era

Abstract: Genome assembly, or genome sequencing, refers to the process or the end product of sequencing genomic fragments of an organism, followed by piecing together, or 'assembling', in scientific terms, the genomic fragments in sequential order to reveal the original genome sequence. To make a genome assembly usable as a reference, annotation (*i.e.*, assigning locations for genes along the chromosome) is also required. The end product of genome sequencing is a complete set of nucleotide sequence(s), of a genome in linear or circular, of DNA or RNA form, depending on the organismic species. It represents the complete genetic makeup determining all molecular potential, entities and activities of that organism. Similar to road maps used for guiding traffic and for helping people to find a person living at a specific address, genome assemblies act as genomic maps (references) to guide us to find genes (eq. persons), regulatory elements, or mutations in specific locations in the genome. Genome assembly aims to generate a genomic map for future studies of that organism and other related organisms.

Keywords: *De novo* genome assembly, Genome assembly, Genome sequencing, Human Genome Project, Omics era, Resequencing.

Definition of terminologies

de novo genome sequencing/assembly

'de novo' is a Latin expression meaning 'from the beginning' or 'anew'. 'de novo genome assembly', or 'de novo genome sequencing', refers to the sequencing followed by the assembly of a genome which has never been done previously. As one can expect, de novo sequencing is much more difficult than resequencing because the latter already has a copy for comparison.

Resequencing

Genome resequencing refers to the sequencing of a genome which has been previously sequenced. A genome is re-sequenced for various reasons, including for improving sequence reliability, or for identifying individual variations, etc.

INTRODUCTION

All genomes evolve from evolution. Each genome represents a track of evolution. With a

Kuo Ping Chiu

complete set of genetic instruction each genome directs the molecular activities associated with the making, the maintenance and the death of an organism, as well as its interaction with other organisms in the environment. A combination of all genomes of all organisms available on Earth represents the outcome of the previous evolutionary process and the genetic blueprint which will guide evolution to shape any future biological field.

PART I. REVIEW OF THE STRATEGIES EMPLOYED FOR GENOME ASSEMBLY

Genome assembly involves technological, molecular and computational strategies to figure out the complete sequence of a genome. Before the automation of Sanger sequencing (see Chapter 1), small DNA sequences were assembled without serious computational algorithms. During that stage, the process relied more on molecular cloning strategies to facilitate the sequencing process. Later, with the advent of sequencing automation (automated Sanger sequencing and next-generation sequencing), sophisticated computational algorithms were developed to deal with large volumes of all kinds of sequence data. A co-evolution between sequencing technologies and computer technologies (hardware and software) thus became a unique feature of the Genomic Era.

A few milestones marked the progress of the Genomic Era. The first genome sequenced was phiX174, a single-stranded circular DNA bacterial phage of 5386 bp in size, by Fred Sanger and colleagues (Sanger 1977, Nature 265:687). In the process, the phage genome was physically mapped by restriction mapping and sequenced by 'plus (viral strand) and minus (complementary strand) method' primed with restriction fragments. Later, a few phages (bacterial viruses) and viruses with genome sizes ranging between a few hundred and a few thousand base pairs were subsequently sequenced and assembled using Sanger sequencing in conjunction with restriction mapping or cosmid cloning.

In 1995, the *H. influenzae* Rd genome of 1,830,137 bp in size was published by Fleischmann and colleagues. It was achieved by whole-genome shotgun sequencing followed by assembly with TIGR assembler implemented with advanced computational methods and algorithms. Besides the shotgun sequences, paired reads were employed to help define the order of contigs and the sizes of gaps within scaffolds (Fleischmann *et al*., 1995). This work was in fact a pilot project to test the hypothesis that an entire genome of several Mb in size can be sequenced by whole-genome shotgun sequencing and assembled by an integrated assembler. In practice, computational methods were developed to create contigs assemblies from 300 – 500 bp cDNA shotgun sequences and read pairs (eq. paired-ends) were used to create scaffolds, within which the number and the sizes of gaps could be estimated. This project represented a milestone for whole-genome shotgun sequencing.

In 2000, the whole-genome shotgun assembly of the *Drosophila melanogaster genome* of

~120 Mb was published by Myers *et al*. It was achieved by shotgun sequencing of bacterial artificial chromosomes (BACs) and assembling by Celera Assembler (Myers *et al*., 2000).

Technological impacts on genome assembly

A number of factors made critical impacts on the evolution of genome assembly. Here the author would like to emphasize the impacts made by 1) NGS and by 2) oligo-mediated technologies. As you might have already sensed, NGS also heavily relies on oligos (*e.g.*, for *in situ* PCR amplification of templates and for priming sequencing reactions). However, it would make it easier for our discussion if we temporarily treat it as a stand-alone technology founded on its non-oligo-based attributes.

Next-Generation Sequencing (NGS)

In 2005, Jonathan Rothberg invented the 454 sequencer, the first NGS machine, in Connecticut, USA. Simultaneously, he created the terminology "Next-Generation Sequencing", or NGS for short, to distinguish his approach from Sanger sequencing (manual and then automated) used by ABI sequencers manufactured in the San Francisco Bay Area. The subsequent marketing of Solexa and SOLiD sequencers helped to shape up the era of NGS. Notice that NGS employs *in situ* PCR, instead of cosmid or BAC cloning, for clonal amplification of templates and that NGS sequencers abandoned di-deoxynucleotide-based Sanger sequencing method, and claimed a number of advantages including low cost, high speed, high throughput, and high yield. However, short read length is a common problem across all NGS platforms and demands another wave of innovation in assembly algorithms (Miller *et al*., 2010) .

When it comes to genome assembly, a number of factors need to be taken into account. These include, but not limited to, genome size, repetitive sequences, assembler, whether there is/are relative genome(s) already available. The most common problems are the short read length and the repetitive sequences interspersed across the genome and those located in the telomere and centromere regions. Long repetitive sequences are more likely to be present in telomere and centromere regions, while short repetitive sequences are more likely to be present in microsatellites

The procedure of the sequencing library construction for a NGS-based genome assembly turned out to be much simpler than that of using Sanger sequencing. Basically, it contains the following steps: 1) collect genomic DNA (gDNA), 2) sonicate gDNA randomly with a sonicator or non-randomly with restriction enzymes (REs). The former is preferred because no discrimination is involved. 3) select DNA fragments within a specific range by running agarose gel electrophoresis, followed by gel excision with a sharp razor blade and DNA extraction from agarose gel, 4) end-repair, 5) construct sequencing library: ligation to sequencing adaptor either by blunt-end ligation or sticky-end ligation

CHAPTER 5

Laboratory Setup and Fundamental Works

Abstract: This chapter aims to share some previous experiences in laboratory setup and bioinformatics exercises with readers and hopefully, by using this chapter as a mediator, to reduce problems which may be encountered by some readers, especially those who haven't had a chance to personally use sequencers for their studies, and those who have just begun to acquire a taste of sequencing and/or genomics studies. There are many big laboratories and sequencing centers which are able to give you some thoughts and useful opinions. Please consult these resources if possible.

Keywords: Drylab, Sequencing Libraries, Wetlab.

Definition of Terminologies

Sequencing Library

A "sequencing library" is defined as the library actually being sequenced by a sequencer. Since sequencing of an unknown target of DNA molecules has to start from a known and well-defined region, ligation of target DNA molecules to a pair of sequencer-dependent sequencing adaptors is essential for making a sequencing library – also see Chapter 6 for the definition of 'sequencing libraries' and classifications.

Pre-Library

A "pre-library" refers to any type of library built before the sequencing library. It is so defined just to distinguish all the other libraries from the sequencing library. Thus, a pre-library can be the total RNA library, the mRNA library, or the cDNA library used to construct the sequencing library.

Wetlab and Drylab

Here, "wetlab" refers to the lab division involved in the preparation of materials, pre-libraries, sequencing libraries as well as all sorts of bench works in the lab. In contrary to a wetlab, a drylab refers to the lab division involved in data analysis.

INTRODUCTION

Sequencing has myriad applications in many fields, either directly or indirectly related to biology, and its impact on our lives is expected to be enormous and profound. Thanks to those scientists, working either at industry or academia, who have become personally engaged in the progress from manual Sanger sequencing to automated Sanger sequencing, and from automated Sanger sequencing to next-generation sequencing and

Kuo Ping Chiu

single-molecule sequencing. Without their efforts, sequencing technologies wouldn't have been able to move forward so expeditiously and sequencing wouldn't be as efficient as we have observed today. However, as shown in the past few decades of human history, sequencing has never been a stand-alone technology and there are many sequencing technologies/platforms and technologies involved, especially computer sciences and biotechnologies.

We have seen a coevolution between sequencing technologies and computer sciences over the years, and this trend is expected to continue for many more decades to come. Without computerization, sequencing automation wouldn't be possible, and without the improvement in speed and capacity of computers and computer-associated devices, the handling of sequence data would have been severely hampered. Furthermore, the development of computer software has added another driving force. Therefore, the role of computer software in genomics investigations is expected to become much more important during the Post Genomic Era.

We have also witnessed a coevolution between sequencing technologies and biotechnologies including those directly or indirectly related to sequencing. In a broad sense, biotechnologies refer to lab methods or methods associated with laboratory equipment. Some lab procedures, such as protein, RNA, or DNA preparations, although also with clear procedures and purposes, are not always recognized as biotechnologies. Some technologies, such as PCR, X-ray crystallography, and sequencing technologies, as each have a well-defined procedure and purpose, making them stand out to be recognized as technologies or biotechnologies, when used for biological investigations. Methods like DNA or RNA isolation may only be considered as laboratory procedures, instead of biotechnologies. However, the fact is, no matter if it is a procedure for DNA or RNA isolation, or a method for molecular cloning for sequencing library construction, it may exert some kind of impact on sequencing efficacy and should not be ignored. For instance, an improved RNA isolation procedure may significantly enhance the quality of a transcriptome library and sequence data. Thus, understanding the rationale and detail of, at least, some key sequencing-related biotechnologies will help you identify the key points which may go wrong, so you can prevent these potential problems from happening by double checking these steps.

Genome sequencing during the genomic era has generated a tremendous amount of sequence data and the data-producing speed is still increasing at an exponential rate during the post genomic era. Undoubtedly, high throughput DNA sequencing will result in a convergence of biology-related fields, asking every biological phenomena and diseases to be traced back to a single nucleotide level, which is then able to provide the highest resolution and has a direct link to genetic mutations and epigenetic modifications.

To cope with the revolutionary situation, a different concept for a laboratory setup and

management is desired. First of all, besides a wetlab setup to handle cell culture and experiments (and sequencing in some cases), it is strongly recommended to recruit at least one or more bioinformaticians to work in the lab to handle sequence data, setup and maintain server or computer software and help sequence data analysis.

Here, I list a few things which may be of interest. However, if you are not interested, or are already an experienced lab leader, please skip this chapter.

I. WETLAB

Setting up a Wetlab

The setup of a genomics laboratory is similar to that of a regular research lab, except some may be luckier than others to have sequencers and reagents for their sequencing library construction and sequencing. Some years ago, sequencer manufacturers started to make medium-sized sequencers for medium-sized laboratories, so that we have seen many more labs equipped with these types of sequencers, although not too technical, but workable for certain purposes such as for clinical samples or diagnostics. Very likely this type of sequencer will be as popular as the PCR machine, when everyone in the lab will be able to prepare sequencing libraries and run sequencing independently. For the time being, it is strongly recommended to assign a specialist to take care of the sequencer(s) and sequencing.

Protocols to Use

For a number of reasons, it seems inadequate to describe protocols used for a sequencing library construction in detail, unless under certain special circumstances. First of all, there are a number of protocols currently available for NGS sequencers (Meyer and Kircher, 2010), and each has pros and cons of its own (Head *et al*., 2014; van Dijk *et al*., 2014), making it sometimes difficult to choose one from so many. Secondly, protocols used for making a sequencing library are not only library type-dependent, but also sequencer-dependent. As such, before we ask which protocol to use to make a sequencing library, we have to ask which sequencer will be used to sequence the library. Moreover, not only does every sequencer manufacturer tend to make their own protocols for their own sequencers, some reagent providers and sequencer users also join together, making sequencing protocols extremely diversified. We expect to experience a selection process to take place to de-select some less competitive ones. At the same time, competition between sequencers will also cause some protocols to disappear with their attached machines. For all these reasons, I won't try to sell any NGS protocols unless it's for special reasons.

However, it is important to understand the basic structure of some protocols, especially those used for sequencing library construction. Here, I would like to outline and discuss

CHAPTER 6

Sequencing Libraries and Basic Procedure for Sequencing Library Construction

Abstract: A sequencing library is the library prepared to be put on the sequencer for sequencing. As such, sequencing library construction is essential for sequencing. Since sequencing of an unknown target DNA molecules has to start from a known and well-defined region, and every NGS sequencer manufacturer uses unique sequencing primers for its own machines, ligation of target DNA molecules to a pair of adaptors is not only essential but also sequencer-dependent, for the making of a sequencing library. This chapter carries a mission to clarify various types of sequencing libraries and the general procedure for their constructions.

Keywords: Emulsion PCR, ePCR, *in situ* PCR, Mate pair, MP, Paired-end , Paired-end ditag, PE, PED, Sequencing libraries.

INTRODUCTION

There is a saying "garbage in, garbage out". The quality of a sequencing library directly affects the quality of the sequences output from the sequencer, and it is thus required to maintain sequencing libraries at a high quality (Head *et al*., 2014). Most sequencing labs or sequencing service providers construct sequencing libraries (Meyer and Kircher, 2010), but some don't. To ensure sequence quality, it's strongly recommended for you to construct your own libraries, unless you are not yet familiar with the procedure, or you have a reliable partner to do the job for you.

CLASSIFICATION OF SEQUENCING LIBRARIES

To facilitate communication, it's helpful to classify sequencing libraries into distinguishable categories. Here are three sets of criteria, or bases, summarized from sequencer manufacturers' conventions together with my personal experiences, which I think should be suitable to serve our purpose: A) wetlab technologies and sequencing design, B) cellular origin of the target molecules, and C) gene expression and regulation.

A. Based on wetlab technologies and sequencing design, all non-miRNA molecules of various cellular origins can be constructed into three types of sequencing libraries: 1) *(shotgun) fragment (SF)library*, 2) *Paired-End (PE) library*, and 3) *Paired-End Ditag, or Mate-Paired (PED/ MP), library*. SF sequencing library is the most straightforward sequencing approach which requires only a single sequencing primer.

Kuo Ping Chiu

Because the target DNA molecules can be ligated to sequencing adaptors in either direction, both strands of the target molecule can be read by the sequencing primer. The chromosomal origin and strand can only be identified by mapping (alignment) of the sequence read against the reference genome assembly. PE sequencing library is defined by "forward (F)" and "reverse (R)" sequencing primers used in the sequencer. PE sequencing strategy is one way to increase sequence length if F read and R read have an overlap. The overlap sequence allows us to merge them into a single read. For this application, the inserted (target DNA) length has to be less than the total length (sum) of F and R reads. PED sequencing library has the 5' tag and 3' tag of the same target molecule, and are linked directly through the wetlab procedure to form a ditag prior to the sequencing library construction. Generally speaking, the span, or distance, variation between paired reads of a PED library is much larger than that of a PE library. This is especially true for mRNA-derived libraries. Since miRNAs are small molecules, fragment sequencing is thus the most suitable choice for miRNA sequencing. Molecules over a hundred bp in size, no matter that it is genomic, transcriptomic, ChIP-EM, ChIP-TFBS, or immuno-library, can be sequenced by either fragment, PE, or PED approach.

B. Based on the cellular origin of the target molecules, sequencing libraries can be categorized into 1) *genomic library*, 2) *transcriptome library*, 3) *miRNA library*, 4) *ChIP-mediated epigenetic modification (ChIP-EM) library*, 5) *ChIP-mediated Transcription factor binding site (ChIP-TFBS) library*, 6) *immuno-sequencing library*, and others. Genomic libraries are constructed from genomic DNA fragments (or genomic cDNA fragments if the target organism has a RNA genome). Transcriptome libraries are built from cDNA libraries derived from mRNA transcripts. Similarly, miRNA libraries are constructed from the complementary DNA molecules of miRNAs. Both ChIP-EM and ChIP-TFBS libraries use antibodies (Abs) each of which recognizes a specific protein to enrich chromatin fragments directly or indirectly bound by the Ab-recognized protein. After the removal of proteins, DNA portions can be constructed into a sequencing library and the genomic locations associated with the Ab-recognized protein can be sorted out. Immuno-sequencing libraries can use DNA (or cDNA) from many different sources, depending on research interest and experimental design.

C. Based on gene expression and regulation, sequencing libraries can be split into three major groups: 1) *genomic library* (*e.g.* genomic libraries for genome assembly or the study of recombination or variation of immune genes); 2) *expression library* (*e.g.* mRNA transcriptome and miRNA transcriptome libraries); and 3) *regulation library* (*e.g.* ChIP-EM (epigenetic modification), ChIP-TFBS (transcription factor binding site), and miRNA library). A non-coding miRNA library can be considered as either a transcription library or a regulation library, because miRNAs are transcribed in a way similar to mRNAs and play a role to (negatively) regulate mRNAs.

GENERAL PROCEDURE FOR SEQUENCING LIBRARY CONSTRUCTION

Step I. Libraries Constructed Prior to Sequencing Libraries

Here, upstream wetlab procedure is defined as the bench work conducted prior to sequencing library construction. For example, before a transcriptome sequencing library can be constructed, normally a mRNA and/or cDNA library needs to be constructed; similarly, before a ChIP-EM sequencing library can be constructed, an antibody-enriched ChIP library has to be constructed beforehand, and so on.

DNA prepared during the pre-library stage is fragmented into smaller fragments to fulfill the requirement of NGS sequencers. DNA fragments within a desired size range are selected (normally by gel excision or beads) to make a sequencing library. Such size selection is essential. DNA molecules exceeding the desired length will be likely to tangle with other templates (during cluster generation), causing the sequencer to be unable to distinguish clusters from each other (during sequencing). This problem will result in difficulties in base-calling for tangled clusters, and the so-generated ambiguous reads will be automatically discarded by the sequencer before output. Overall, the yield will be compromised.

1. Fragmentation of DNA, or RNA, by enzymatic reaction or sonication (preferred)
2. End-repair target DNA
3. Size-select target DNA using magnetic beads (Step II and III are interchangeable.)
4. dA-tailing of the size-selected target DNA

Step II. Ligation of Sequencing Adaptors to Target DNA Molecules

Ligation of target DNA to adaptors is essential for all NGS sequencers. Subsequently, the adaptor-ligated target molecules are denatured and subjected to ePCR (for 454 and SOLiD machines) or cluster amplification (for Solexa machines). Notice that, it is important to quantify the amount of DNA molecules whenever possible, and DNA quantification immediately before ligation to sequencing adaptors is critically important, because overcrowded templates will compromise resolution and thus reduce the yield.

1. Ligate (sequencing) adaptors to target DNA molecules using kits recommended by your sequencer manufacturer
2. Purify the ligated DNA
3. Quality evaluation of the ligated constructs using a Bioanalyzer (Fig. **1**).

Note: This step is able to detect the amount of "free" P1 hP2 adaptors in the preparation, but cannot tell whether these adaptors are incorporated into the construct.

CHAPTER 7

Paired-End (PE), Mate Pair (MP) and Paired-End Ditag (PED) Sequencing

Abstract: Information regarding the distance between paired reads enhances the accuracy of genome assembly and sequence-to-genome mapping, making paired-end indispensable strategies for DNA sequencing. The most commonly used paired-end sequencing strategies are Paired-End (PE) sequencing and Paired-End Ditag (PED) sequencing. Similarity in terminologies frequently causes confusion. This chapter is set out to clarify these terminologies and then, using PED as an example, to illustrate how a biotechnology can be sequentially developed.

Keywords: bPED, ChIP-EM, ChIP-HM, ChIP-TFBS, *in situ* PCR, mbPED, PE sequencing, PED sequencing.

Definition of Terminologies

PE *sequencing: Paired-End sequencing*
PED *sequencing: Paired-End Ditag sequencing*
PET *sequencing: Paired-End diTag sequencing*

INTRODUCTION

Paired-End diTag (PET) directly links the 5' terminal tags (~18-20 bp each) of genomic DNA fragments or cDNA molecules to their corresponding 3' terminal tags for high throughput sequencing (HTP), has led to a number of important discoveries (Birney *et al*., 2007; Carninci *et al*., 2005; Ng *et al*., 2005; Zhao *et al*., 2007), including fusion gene identification. To move one-step further, we recently invented a robust method which adopts barcoded adaptors to generate barcoded Paired-End Ditag (bPED) libraries from genomic and transcriptomic libraries. Various bPED libraries, each labeled with a unique internal barcode, can be combined to form a multiplex barcoded Paired-End Ditag (mbPED) library for ultra high-throughput (UHTP) sequencing. These paired-end ditag cloning strategies produce ditag libraries at the lab bench and the ditag libraries can be sequenced as fragment libraries by a single sequencing primer. On the other hand, Paired-End (PE) sequencing is conducted on sequencer using two sequencing primers.

PART I. PE SEQUENCING

Paired-End sequencing uses a forward sequencing primer to sequence the initial set of

Kuo Ping Chiu

template clusters, convert the initial clusters *in situ* into a complementary set of template clusters, and then uses a reverse sequencing primer to sequence the complementary template clusters. Two sets of reads are then paired based on X- and Y-coordinates. Illumina's PE sequencing is a popular sequencing approach for current NGS sequencing.

Illumina PE sequencing includes four steps: ***I)*** *library preparation,* ***II)*** *cluster generation,* ***III)*** *sequencing, and* ***IV)*** *data analysis*. Library preparation involves almost all wetlab work, cluster generation and sequencing can now be performed automatically on MiSeq sequencer, and, similar to all other sequencers, data analysis mainly focuses on initial data processing. These steps will be discussed in more detail in the following sections.

I. Library Preparation

The procedure of sequencing library preparation has been a common practice for sequencing labs and sequencing providers. Besides the common procedures for library construction, Illumina also intensify R&D and collaborative effort in developing new protocols to improve accuracy. For related information, please refer to the Sequencing-by-synthesis section of Chapter 2 and Illumina's protocols.

II. Cluster Generation by *in situ* PCR

The amount of DNA required for constructing a sequencing library depends on the type of library to be prepared. For making a transcriptome library, MiSeq requires a total of 600 micro-liter of 12 pM (pico-molar) single-stranded (ss) target DNA. Such diluted concentration was optimized to ensure the separation of templates. About one third is loaded into the chamber of sequencing slide for *in situ* amplification to generate a set of cluster library from which a total of around 15 million raw reads is produced. Thus, the input molecule number to the output cluster ratio is calculated to be about 100 fold excess. Notice that, only the DNA molecules with both ends ligated to sequencing adaptors can be amplified to generate clusters, and the successful rate is another issue.

Prior to cluster amplification, the desired size range has to be determined to prevent cross-over between clusters, which would otherwise impair the specificity of signal. Ambiguous cluster signals are expected to have low quality values and will be discarded eventually. Also, prior to sequencing, the kit of sequencing reagents needs to be determined, so that the sequence length from each primer can be pre-determined. It is better to calculate the overlapped length beforehand.

1. (Two types of oligos, each containing a specific PCR primer, are covalently pre-attached on the surface of a flowcell.) Single-stranded templates, which were already ligated with sequencing adaptors and then denatured, are randomly distributed across the "lawn" of anchored oligos. Optimized concentration allows each template to be well-separated from one another.

2. 1st annealing: Similar to regular PCR reactions, templates are annealed to their complementary template strand
3. 1st elongation: DNA polymerase extends the complementary strand (in "outbound" 5'-to-3' direction) from the PCR primer-defined location until the 5' end of the template
4. 1st denaturation separates the double-stranded structure leaving the oligo-primed strand covalently attached to the surface of the flowcell, while the original templates are washed away from the chamber.
5. 2nd annealing: "Bridge" amplification starts from the second round of PCR because now both PCR primers are all covalently attached on the surface of the flowcell, DNA molecules form "bridges" across the lawn when their free (3') ends anneal to the accessible complementary PCR primers covalently attached to the proximal region of the lawn.
6. 2nd elongation: DNA polymerase makes the complementary strands.
7. 2nd denaturation: denaturation separates the dsDNA molecules and form two ssDNA strands, both of which are covalently attached to the lawn surface at their 5' ends.
8. Continuous PCR cycles eventually generate tens of million copies of clusters on the surface of the flowcell.
9. To achieve a clear sequencing signal, only one strand per cluster is kept for sequencing. This is accomplished by enzymatic digestion of the dsDNA molecules with DNA glycosylaselyase Endonuclease VIII which specifically cleavesU site using USER (Uracil-Specific Excision Reagent; Uracil DNA glycosylase (UDG) + DNA glycosylaselyase Endonuclease VIII). The cleaved strands are washed off the chamber and only the "forward templates" are leave on the lawn for sequencing.
10. Notice that, to prevent unwanted extension from the 3' ends of the sequencing templates. The 3' ends are blocked.

III. Sequencing

Here, an outline is introduced. For detailed procedure please consult Illumina protocol.

Forward Sequencing

1. Before sequencing, clusters are captured tile-by-tile, or panel-by-panel, with camera and their locations "mapped" against X- and Y-coordinates for future referencing to identify the origins of color signals produced during sequencing.
2. Sequencing progresses through "inbound" DNA synthesis primed by sequencing primers bound near the 3' ends of the templates.
   Simplification: This inbound sequencing orientation is universally true for sequencing-by-synthesis sequencers, no matter it is conducted inside the flowcell (as adopted by Illumina) or on a magnetic beads (as adopted by Roche 454). This is an intrinsic character defined by the nature of *in situ* PCR amplification for clonal expansion of the

CHAPTER 8

Genome Sequencing and Assembly

Abstract: In chapter 4, we reviewed the background and history of genome sequencing and assembly. In this chapter, we will focus more on the technical and experimental issues. In science, the term "whole genome sequencing and assembly" is often used interchangeably with "genome sequencing and assembly", "genome sequencing" and "genome assembly", because genome assembly normally refers to the assembly of an entire genome and genome sequencing is normally followed by assembling sequence reads to produce a complete set of chromosomal sequences for the genome of interest. For convenience, genome assembly and genome sequencing are preferred and will be used more often than the others throughout our discussion. Since whole genome sequencing and assembly is a complicate process, involving multiple alternatives and methodologies, it is not possible to cover every detail. We will go through some concepts and NGS-associated procedures so that readers can get some idea of how the genome assembly is achieved. Serious readers are recommended to consult previous reports published by sequencing laboratories.

Keywords: Contig, *De novo* genome assembly, Genome assembly, Genome sequencing, Scaffold.

INTRODUCTION

The beginning of the 21st century was marked by numerous genome releases. In 1992, Craig Venter founded the Institute for Genomic Research (TIGR) to sequence microbial genomes using whole-genome shotgun sequencing strategy. In 1998, he founded another organization, Celera Genomics, to sequence the human genome using techniques his team developed. Since then, a number of genomes have been sequenced and assembled. These assembled genomes are available in many famous genome centers across the world.

Among the most commonly used reference genomes are the human genome and the mouse genome, both of which were published during the first few years of the 21st century. Key publications include 1) Initial sequencing and analysis of the human genome (Lander *et al*., 2001); 2) The sequence of the human genome (Venter *et al*., 2001); 3) Initial sequencing and comparative analysis of the mouse genome (Mouse Genome Sequencing *et al*., 2002); 4) Finishing the euchromatic sequence of the human genome (International Human Genome Sequencing, 2004); and 5) Quality assessment of the human genome sequence (Schmutz *et al*., 2004).

Besides, both *Drosophila melanogaster* and *Arabidopsis thaliana* genomes were

published in 2000 (Arabidopsis Genome, 2000; Myers *et al*., 2000), while the *C. elegans* genome was first released in 1998 (as the first animal genome release) (Scienc 282:2012-2018) and re-sequenced in 2008 (Hillier *et al*., Nature Methods 5:183).

The haploid human genome consists of an approximately 3 Gb sequence distributed in 23 nuclear chromosomes and a mitochondrial chromosome. About 20,000 - 25,000 protein-coding genes and at least hundreds of noncoding genes involved in post-transcriptional regulation have been found in the human genome. This kind of sequence information can be best acquired by whole genome sequencing coupled with whole transcriptome sequencing. Genome assembly aims to build a genetic map to be used as a reference genome for future biological studies. Moreover, reference genome sequences can be used to facilitate the assembly of related genomes.

CLASSIFICATION OF GENOME ASSEMBLY

In general, genome assembly can be categorized into *de novo* genome assembly (or *de novo* genome sequencing) and re-sequencing. ***De novo genome assembly*** refers to the sequencing of the DNA or RNA genome of an organism which has never been sequenced before, followed by using assembler software or by implementing a processing pipeline to piece together the sequence reads, so to generate a "genomic map" for that target organism. Then, sequence-to-gene annotation is conducted to associate chromosomal locations with genes and other genomic entities such as enhancers, promoters, DNA motifs, *etc*. Such assembled genome can then be used in the future for the studies of that organism as well as its related species. On the other hand, **re-sequencing** refers to the sequencing of a genome of an individual, of which a reference genome has previously been built. With the increasing number of genomes having been assembled, more and more "relative genomes" are becoming available, and these relative genomes can be used as references to facilitate the assembly process of a similar genome. This type of genome assembly can be categorized as **"semi-*de novo* genome assembly"**.

WHAT SEQUENCING STRATEGIES TO TAKE?

Among the NGS machines, a combination of 454 (*e.g.* Flex) with Illumina's sequencers (*e.g.* HiSeq series) is frequently used for genome assembly. This strategy takes advantage of 454's long reads and HiSeq's high quality and high yield. Usually, paired-end ditag, paired-end, and, sometimes, fragment libraries are combined to achieve the accuracy of the assembly process. (SOLiD machines, which use sequencing-by-ligation mechanism, are not suitable for genome assembly, simply because these machines are incapable of reading through palindromic structures.)

GENERAL PROCEDURE FOR *DE NOVO* GENOME SEQUENCING

Strategies for genome assembly have experienced a dramatic change since the launch of

Human Genome Project in 1990. The major technical impact resulted from the introduction of NGS technologies starting from 2005. The revolutionary NGS technologies were later accompanied by a wave of software evolution attempting to deal with the ever-increasing huge amounts of short read sequence data produced by various NGS machines. To make a long story short, here I would like to use the shotgun sequencing strategy, which is commonly taken by scientists working on genome assembly, as an example to illustrate how a genome can be assembled (Fig. **1**).

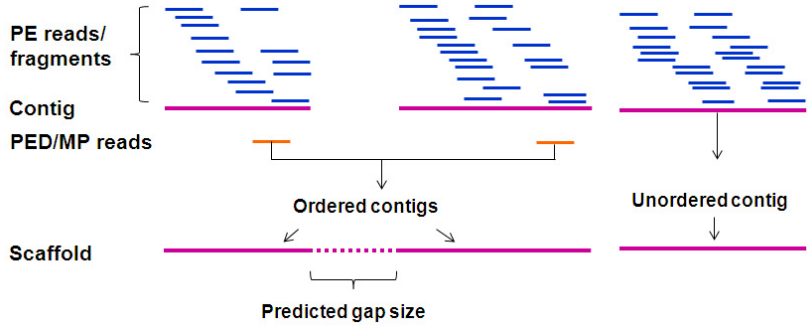

Fig. (1). The first part of *de novo* genome assembly. Sequence reads, either paired-end or fragment reads, are assembled into contigs, which are then ordered and assembled into scaffolds. Based on additional information, gaps between ordered contigs can be predicted.

The general procedure for *de novo* genome assembly is outlined below:

A. ***Preparation of sequencing libraries***

1. Isolate genomic DNA (gDNA) from an organism. White blood cells in the blood is frequently used for gDNA isolation. There are a number of kits/methods that can be used to isolate genomic DNA from prokaryotic or eukaryotic genome. To obtain these commercial products or protocols, please consult experienced personnel, Maniatis Manual, or use Google search. Once you have genomic DNA ready, you can continue to make a few types of sequencing libraries for genome assembly. The common ones include fragment library for shotgun short-read sequencing and Paired-End library for ditag sequencing.

   DNA quantification

   Keep in mind to trace the quantity of your sample frequently. Do quantify your DNA before and after sonication, and before *in situ* PCR amplification. Initial DNA quantification, which does not have to be very accurate, can use NanoDrop or bioanalyzer.Normally, there is no difficulty to obtain a sufficient amount of genomic DNA, because a quantity at only a micro-gram scale is required for each library. More accurate quantification of DNA can be done by bioanalyzer or fragment analyzer which display DNA molecules based on size. The most accurate quantification can be done by qPCR, if necessary.
2. Sonicate gDNA to make gDNA fragments.

CHAPTER 9

Exome Sequencing: Genome Sequencing Focusing on Exonic Regions

Abstract: While whole genome sequencing (WGS) remains costly and requires intensive labor and elaborate analytical tools for assembly, whole exome sequencing (WES) is relatively cheaper and easier. Compared to WGS, WES can be considered as an efficient approach when the protein-coding regions are the only concern, because this type of sequencing focuses on the exon regions and its desired sequencing depth can be easily reached. WES is frequently confused with transcriptome analysis because both types of libraries contain solely the exonal sequences. However, the former is generated from genomic DNA fragments, while the latter from expressed mRNA molecules. Readers are asked to distinguish the differences between these two libraries beforehand.

Keywords: Whole exome sequencing, WES, Whole genome sequencing, WGS.

Definition of Terminologies

Exome

Exome is a scientific term representing the collection of exonal sequences.Thus, exome is part of the genome, and exomics is part of the genomics.

INTRODUCTION

There are ~180,000 protein-coding *exons* in the human genome. These exons occupy approximately 1% of the human genome and yet harbor about 85% disease-causing "genetic" mutations, which can be directly detected by next-gen sequencing (Gilissen *et al*., 2012). Whole exome sequencing (WES) application was first reported in 2009 (Choi *et al*., 2009; Ng *et al*., 2009). By combining exome capture, or (exome) target enrichment, approaches with NGS technologies, WES has become a robust approach for the identification sequence variations responsible for common and rare Mendelian diseases. WES enhances the study of SNPs (single nucleotide polymorphisms, or single nucleotide variations (SNVs)) and indels (insertions and deletions) between diseased and normal tissues and leads to the identification of disease genes.

Whole exome sequencing heavily relies on specific selection of exonic DNA from fragmented genomic DNA preparations (of an individual) (see figure shown below). There are a number of exome capture approaches: multiplex PCR which uses multiple

Kuo Ping Chiu

pairs of PCR primers in a single polymerase chain reaction to amplify multiple exons from a genomic DNA preparation, molecular inversion probe (MIP), microarray hybridization capture on glass slides as pioneered by Roche NimbleGen (Sequence Capture Human Exome 2.1M Arrays), and in-solution capture on beads adopted by Illumina (TruSeq Kit), and also by Roche NimbleGen (SeqCap EZ Exome Library Kit) (Mamanova *et al*., 2010). These exome capture methods bypass the intronic regions, and frequently the 5' and 3' untranslated regions (UTRs) as well, to selectively enrich the exonic DNA for next-generation sequencing (Fig. **1**).

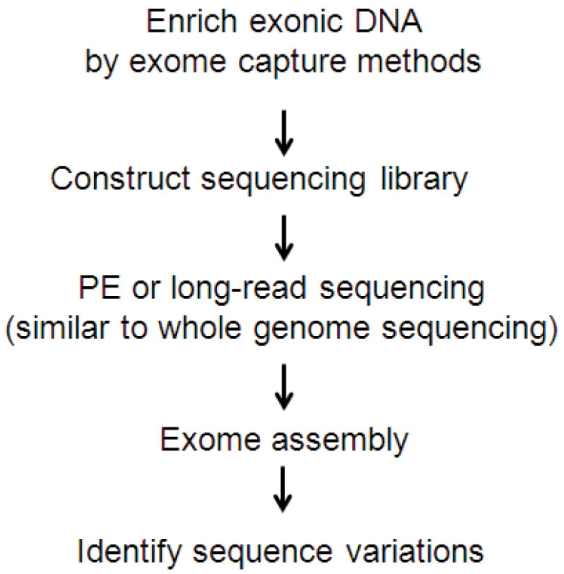

Fig. (1). Experimental procedure of exome sequencing.

SEQUENCING LIBRARY CONSTRUCTION

Procedure for preparing an exome sequencing library remains very similar to that for other types of sequencing, except that an exome capture method has to be decided for initial exonic DNA preparation.

As mentioned above, there are a number of target enrichment kits made commercially available by Roche NimbleGen (using microarray for hybridization capture and in-solution sequence capture), Illumina, and Agilent, *etc*. These kits made the sequencing library construction for exome sequencing simpler and straightforward.

Multiplex PCR is not feasible for whole exome sequencing. It is only suitable for the amplification of limited number of exons, because bias increases along with the number of exons under amplification due to potential imbalanced amplification efficiency between exons. PCR bias can result from the difference in length, primer specificity, enzymatic activity, the amount of dNTPs provided, *etc*. However, multiplex PCR method

can still be considered when limited number of exons are to be examined.

Exome target enrichment strategies for NGS was described by Mamanova *et al*., (Mamanova *et al*., 2010).

SEQUENCE DATA ANALYSIS

The first step in sequence data processing is to remove (discard) unwanted reads from raw data. The unwanted reads include low-quality reads, contaminated sequences such as sequences partially or fully occupied by vector or adaptor sequences, repetitive sequences (*e.g.* $polyA_9/T_9/G_9/C_9$ and sequences with recognizable repetitive patterns which may cause these sequences to map to numerous genomic locations), sequences from rRNAs or tRNAs when these species are not under investigation, and questionable reads (*e.g.* intronic sequences when exome sequencing is the research focus).

Since the sequenced DNA fragments are supposed to be exon-originated while certain level of uncertainty remains, we need to calculate the percentage of reads mapped to the exonic regions over the total reads. Considering the possibility of non-specific capture, 100% is not possible, but normally over 80% is expected. Sequences mapped to non-exonic regions are normally discarded from sequence pool. Sequences mapped to exonic regions are aligned with software to identify the genetic mutations. For current status of NGS capability, at least 20-fold coverage is regularly desired.

Processing and analysis of WES data for the identification of novel gene mutations associated with rare Mendelian diseases was well illustrated in the review article by Gilissen and colleagues (Gilissen *et al*., 2012). As described by the authors, the number of variant calls can vary significantly depending on the capture method and the sequencing platform been used. Roughly the initial number of variant calls can reach tens of thousands per case. Through the routine procedure of sequence quality control to remove the low quality sequences followed by filtering out the ones located in the non-coding regions, most false positive calls can be removed, causing the number drop down to about a few thousands. This number can be further reduced by removing know variants, making the number of variant calls drops for another 10-folds. Now, working on a few hundred genes seems to be much easier comparing to the initial number, but the reduction itself requires fine-tuning using certain case-dependent strategies, including linkage strategy, homozygosity strategy, double-hit strategy, overlap strategy, *de novo* strategy, and candidate strategy, *etc*. (Gilissen *et al*., 2012).

SINGLE NUCLEOTIDE POLYMORPHISM

Single nucleotide polymorphism (SNP, pronounced *snip*) is a sequence variation which may occur naturally between individuals of the same species. SNPs are found not only in the intergenic and intronic regions, but also in the exonic regions. These genetic

CHAPTER 10

Transcriptome Analysis

Abstract: Transcriptome analysis, or transcriptome sequencing, concerns the transcript sequences transcribed from the genome of a specific cell type at specific time and growth conditions. Previous studies have clearly demonstrated that, besides messenger RNA (mRNA), the transcribed RNA sequences also contain large amounts of ribosomal RNA (rRNA), transfer RNA (tRNA) and small-sized non-coding RNA (ncRNA). Transcriptome analysis focuses mainly on mRNA, and sometimes, certain types of ncRNA species which may be co-isolated with mRNA when gene expression and regulation are of the major concern. From transcriptome sequencing, a number of biological information can be retrieved. These include gene expression level, transcriptome landscape across the entire genome, Gene Ontology, pathway, *etc*. Notice that transcriptome analysis normally refers to the whole transcriiptome analysis of a cell population. The result is in fact a combination of millions of potentially diversified single-cell transcriptomes.

Keywords: Gene Ontology, GO, KEGG, Pathway, RPKM, Transcriptome.

Definition of Terminologies

Ttranscriptome

The RNA products of gene expression from a cell or a cell population at a particular status and time point constitute a "transcriptome". In a broad sense, a transcriptome should include mRNAs to be translated into a "proteome", rRNAs to be used for making ribosomes for protein synthesis, tRNAs to be used in carrying amino acids for protein synthesis, and many kinds of small-sized non-coding RNAs, including so-called small RNAs (sRNAs, ~50-250 nucleotides), microRNAs (miRNAs, ~ 17-25 nucleotides), and *Piwi-interacting RNA (*piRNA, typically 24-32 nucleotides*). Since the roles of rRNAs and tRNAs have been well-studied and defined, most current transcriptomic investigations focus on mRNA transcriptomes and non-coding small-sized RNA molecules, especially sRNAs, miRNAs, and piRNAs, and the like.*

INTRODUCTION

There are about 24,000 protein-coding genes in the human genome. In a differentiated cell population, only about half of these human genes are expressed, or transcribed, into mRNAs, while single cells express even fewer genes. The expression of protein-coding genes is accompanied by the expression of a few thousand non-coding genes which

Kuo Ping Chiu

selectively degrade or inhibit the translational efficiency of their target mRNAs in response to the environmental cues. The expression of protein-coding and non-coding genes are fine-tuned in a coordinated fashion to balance physiological conditions in a homeostatic manner.

Gene expression, or transcription, involves multiple steps of hierarchical regulation. During cell differentiation, genes, and frequently, their corresponding regulatory elements, are packaged in various manners based on their prospects of future usage. Those packaged in heterochromatin domains are likely to become unusable; those maintained in the facultative heterochromatin domains may still be likely to be used; and those in the euchromatin domains will be translated. Notice that the arrangement of heterochromatin, euchromatin, and facultative euchromatin is tissue type-dependent because different tissues use different sets of genes. Compact packaging in the heterochromatin domains makes the encompassed genes and their corresponding regulatory elements inaccessible by transcription factors (TFs), cofactors, or transcriptional machinery. Epigenetic modifications further separate the euchromatin domains, and thus their encompassed genes, into various states of readiness for transcription. Extracellular molecular signals (*e.g.*, hormones from the endocrine system, cytokines from the immune system, and extracellular ligands) regulate intracellular states and activate corresponding TFs and their cofactors. Activated TFs recruit their cofactors, including histone modifiers such as HDAC and methyltransferase, and bind to their target motifs in the euchromatin. These events result in limited alteration in epigenetic modifications, leading to a transcriptional activation and/or repression of certain sets of responsive genes.Taken together all these molecular events taking place in the microscopic world, one can understand that transcriptional activation and repression are highly regulated.

IMPACT OF GENOME ASSEMBLY ON TRANSCRIPTOME ANALYSIS

Before genome assemblies were made available during the turn of the twenty-first century, transcriptome analysis was heavily relying on microarray, SAGE (serial analysis of gene expression), and EST (expressed sequence tag), especially microarray. Once being a well commercialized and well-established technology, microarray extracts gene-specific sequences from mRNA to make thousands or tens of thousands of probes. With the assistance of instruments, some of which are semi-automated, probes are hybridized to the transcripts present in transcriptome libraries. After a few washes to remove nonspecific hybridization, signal intensity of a hybridized spot is subtracted by the signal instensity of the nearest control (background). By so doing, the level of expressed transcript is quantified as the hybridization intensity, most likely to be a non-integer value. On the other hand, both SAGE and EST rely on sequencing to profile transcriptomes and virtual databases built from known gene-specific sequences to correlate the expressed transcripts with their corresponding genes. The expression levels

can be presented as integers.

It should be noted that originally these technologies could only detect the expression of known genes while novel genes were ignored. This was because the design of microarray probes relied on known genetic sequences and the setup of the SAGE virtual database also relied on known sequences. Genome assemblies now provide complete sets of genomic sequence information, including known genes and previously unknown (novel) sequences for us to compare with our local transcriptome data. Accordingly, experimental and analytical strategies have to be changed to integrate genome assembly information in order to enhance transcriptome analysis.

This chapter focuses on mRNA transcriptome of protein-coding genes. The study of gene expression of a cell population, tissue, organ, or an organism relies on whole transcriptome sequencing. Whole transcriptome sequencing allows us to discover transcriptional (gene expression) fluctuations such as the upregulated and downregulated genes, grouping of the fluctuated genes by Gene Ontology, and biological pathway analysis (Carninci *et al*., 2005; Chiu *et al*., 2007; Consortium *et al*., 2007; Ng ., 2005).

CONSTRUCTION OF TRANSCRIPTOMIC SEQUENCING LIBRARIES

For the construction of a whole transcriptome library, it is essential to isolate total RNA from cells, followed by isolation of mRNA from the total RNA and reverse transcription using oligo-dT to generate a cDNA library from mRNAs. To achieve optimal sequence specificity and sequencing efficiency, fragmentation of target molecules is essential and can be done on mRNA or cDNA.

We strongly recommend mRNA fragmentation for a number of reasons, in particular, the following. First, the efficiency of reverse transcription of mRNA molecules into cDNA molecules is size-dependent; long mRNA molecules require more time to complete the process. Thus, this process favors short mRNAs and is likely to introduce bias. Second, if there is degradation in the isolated mRNA, only the regions containing poly-As can be converted into cDNA by reverse transcription. This potential problem will create another layer of bias leading to a sequence-to-genome mapping mistakenly concentrated on the last few exons. To minimize these potential sources of bias, mRNA fragmentation is advised.

PAIRED-END DITAG SEQUENCING *VS.* SHOTGUN FRAGMENT SEQUENCING OF TRANSCRIPTOME LIBRARIES

As mentioned previously, a transcriptome library can be built into either a shotgun fragment (SF) sequencing library or a paired-end ditag (PED) sequencing library (Fig. **1**). The former will be subjected to either fragment sequencing (using a single (set) of sequencing primer(s)) or paired-end (PE) sequencing (using forward and reverse

CHAPTER 11

Single Cell Sequencing (SCS) and Single Cell Transcriptome (SCT) Sequencing

Abstract: Sometimes genomic and transcriptomic information of single cells, instead of those produced from cell populations, are desired. Obtaining such information relies on single cell sequencing (SCS) and single cell transcriptome (SCT) sequencing.

Although SCT sequencing is in fact part of SCS, they are readily distinguishable not only in research objective, but also in experimental procedure and bioinformatic approach. We will first review the history and achievements that have been made in these fields, and then discuss an experimental procedure of SCT sequencing to gain more insight into the subject.

Keywords: SCS, Single cell Sequencing, Single cell transcriptome sequencing, SCT, Transcriptome.

INTRODUCTION

Single cell sequencing (SCS) refers to the sequencing and analysis of the genomic or transcriptomic sequences of single prokaryotic or eukaryotic cells. It possesses unprecedented potential in resolving genetic substructures and the variations in genomic and transcriptomic profiles at both single cell and molecular levels. During the past few years, studies on SCS have shown promising results, especially in cancer research and transcriptome analysis. Foreseeing its great potential applications in biological studies, SCS was chosen as Method of the Year in 2013 by *Nature Methods*.

Current SCS approaches rely on both PCR-based DNA or cDNA amplification and next-generation sequencing. NGS is required because it is the only sequencing approah allowing us to obtain an in-depth coverage. Conventional NGS technologies, however, require large amounts of genetic material (a few nanograms or micrograms per library) as the input for sequencing. Such quantities are many orders of magnitude higher than a single cell can provide, making PCR amplification also an essential element for single cell sequencing.

So far, SCS applications mainly focus on single cell genome sequencing and single cell transcriptome sequencing. The former was highlighted by the pilot studies on cancer evolution by Navin and Hou *et al*. (Hou *et al*., 2012; Navin *et al*., 2011), while the latter by the works on development and cancer transcriptome by Tang and Ramskold *et al*.

Kuo Ping Chiu

(Ramskold *et al*., 2012; Tang *et al*., 2009). In the paper published in 2011, Navin and colleagues reported their work of using single nucleus sequencing (SNS) to study copy number variation in breast cancer (Navin *et al*., 2011). By using single cells from the same cancer origin, they were able to determine the genetic lineages, and thus the evolutionary substructures, in a cancer. Their results suggested that tumors grow by punctuated clonal expansion, instead of gradual tumor progression. The next year, Hou and colleagues published the multiple displacement amplification (MDA) method for single cell genomic DNA amplification and its application in the study of the genes involved in essential thrombocythemia (ET) evolution (Hou *et al*., 2012). Results suggested that ET patient carries a distinct set of mutations and a monoclonal origin of ET cancer cells.

For single cell transcriptome analysis, certain methods to improve single cell cDNA amplification were already reported prior to the advent of NGS technologies. These include the work published by Eberwine *et al*., in 1992 and that published by Kurimoto *et al*., in 2006 (Eberwine *et al*., 1992; Kurimoto *et al*., 2006). The NGS-based single cell transcriptome sequencing was first published in 2009 by Tang *et al*., (Tang *et al*., 2009). In a work published by Ramskold *et al*., in 2012, they reported an elegant method, called Smart-Seq, for reverse transcription and cDNA amplification for SCT analysis (Ramskold ., 2012). The method has been commercialized by Clontech in making "the SMARTer Ultra Low RNA Kit". For technical control, they used defined quantities of total RNA, which could be correlated to defined numbers of cells. In fact, to compensate insufficient qantity of input material for next-generation sequencing, we have been rountinely using defined amounts of total RNA or mRNA for transcriptome analyses since years ago. Our results demonstrated the feasibility of this modification and also indicated that, as expected, mRNA works better than total RNA.

Most people have a concern about the technical variations which may impose bias to SCT analysis. Certainly, many factors may introduce bias into single cell transcriptome analysis. SCT bias may result from variations in personal technical skills, and may confuse the real transcriptional variation among individual cells (Ramskold *et al*., 2012). To minimize the influence of variability in personal technical skill, increasing the number of SCTs and using internal controls such as house keeping genes and previously studied expression patterns of certain genes is recommended.

Experimental procedure for SCT sequencing

To help readers further understand how SCS is conducted, here I would like to present and discuss the experimental procedure that we are using in the lab. Since SCS per se is a broad subject, it would be better off for us to limit the scope by focusing on SCT analysis. For SCS applications in the study of genetic sequence variations, cancer gene identification and cancer evolution, please consult the works published by Navin and

How *et al*., For more useful information about SCT experiments please consult the original procedure published by Ramskold and colleagues (Ramskold *et al*., 2012). We have compared Ramskold's method with others and found that this method is the most reproducible. The experimental procedure outlined below mainly follows this method but with certain modifications.

Basically, the procedure can be divided into two parts: 1) generation and amplification of cDNA and 2) sequencing library construction. We will go through the general ideas. Please consult the original procedure for more detailed information.

Generation and amplification of cDNA

Cells are trypsinized (if needed), washed, and kept in a buffer (*e.g.* PBS) before single cell isolation. Single cell isolation can be done by mouth pipetting or by other methods of micromanipulation. We prefer mouth pipetting because it is very mild and does not cause cell damage. Each single cell is first kept in ≤ 1 uL of 1X PBS and then lysed by adding 4 uL of hypotonic lysis buffer which contains RNase inhibitors together with other ingredients (see the original protocol). The first strand cDNA synthesis is primed by CDS primer (5′–AAGCAGTGGTATCAACGCAGAGTACT(30)VN–3′, where 'V' stands for non-T). The degenerated base (V) is so designed to enhance binding of the oligo to the beginning of the polyA stretch. Without it, the oligo may "slip" within the polyA region. The MMLV reverse transcriptase (RT) possesses a terminal transferase activity and would automatically add a few extra C (polyC tail) to the 3′ end of the first strand cDNA. The reaction solution also contains SMARTer II A oligo (5′-AAGCAGTGGTATCAA-CGCAGAGTACATrGrGrG-3′, where 'r' stands for ribonucleotide base) which is able to anneal to the polyC tail in the first strand cDNA, the SMARTer II A oligo thus creates an extended template, allowing the MMTV RT to continue the first strand cDNA synthesis throughout the second template.Since the sequence in the SMARTer II A oligo is known, PCR primers are designed and used not only to make the second strand of cDNA, but also for the amplification of cDNA.

For "multiple-cell" transcriptome analyses, MCF-7 cDNA was pre-amplified for 12 cycles when using 1 ng of total RNA (~100 MCF-7 cells), or 15 cycles when using 100 pg of total RNA (~10 MCF-7 cells), or 18 cycles when using 10 pg of total RNA (~1 MCF-7 cell). The cDNA samples prepared from MCF-10A or MCF-7 single cells were amplified for 23 cycles and 20 cycles, respectively, because MCF-10A transcribes much less RNA than MCF-7. The PCR-amplified cDNA should appear as a distinct peak of ~500–5,000 bp under Bioanalyzer (Fig. **1**).

CHAPTER 12

ChIP-TFBS Analysis

Abstract: Eukaryotic gene expression is tightly controlled by a cascade of regulatory mechanisms. At the sequence level, gene expression is regulated by *cis*-acting DNA motifs that are able to recruit trans-acting transcription factors (TFs) for positive or negative regulation of local gene expression. The genome-wide mapping of transcription factor binding sites (TFBS) becomes a crucial strategy for the study of gene expression regulation. Here in this chapter we will discuss the preparation of ChIP-TFBS sequencing libraries and the analysis of ChIP-TFBS sequence data.

Keywords: ChIP, ChIP-TFBS, Chromatin immunoprecipitation, Motif, TF, Transcription factor.

Definition of Terminologies

Transcription Factor (TF)

A transcription factor is a protein that binds to specific DNA motifs in the genome and works together with other proteins (including co-factors, helicase and RNA polymerase) to enhance, or block, the transcription (expression) of genes.It is a well-studied mechanism for the regulation of gene expression.

TFBSs*: Transcription factor binding sites*

INTRODUCTION

ChIP-TFBS (Chromatin IP-mediated transcription factor binding site) analysis is a common practice in the study of transcriptional regulation (Park, 2009; Pepke *et al*., 2009). (Lin *et al*., 2007; Loh *et al*., 2006; Zeller *et al*., 2006). Frequently, transcriptional regulation of gene expression is initiated by ligand binding to target receptors on a cell membrane followed by a cascade of molecular signaling to induce transcriptional activation of a group of genes and a simultaneous repression of another group of genes. In other words, ligands, such as hormones or cytokines, bind to their target receptors on a cell membrane to initiate a cascade of signal transduction, leading to the activation of one or more specific TF, which enter(s) the nucleus and bind(s) to its/their accessible motifs in the genome. A TF, by recruiting its co-factors and transcriptional machinery, in turn activates the expression of one group of target genes and simultaneously represses another group of genes. During this process, TF binding plays a key role in determining the specificity for both transcriptional activation and repression.

Kuo Ping Chiu

For TFBS analysis, chromosomes need to be fragmented into smaller pieces. However, prior to sequencing library construction, regional contacts between DNA and proteins need to be maintained. Moreover, to date ways to probe transcription factor binding sites (which are DNA or RNA sequences dispersed across the genome) without using antibody (Ab) to "fish" the DNA-protein complexes of interest have not been developed. To address these various complications, we need to fix cells (either in culture, tissue section, or other kind of cell sample), isolate the chromatin (containing both DNA and associated proteins), fragment the chromatin, enrich the chromatin fragments of interest, then remove proteins and construct a sequencing library on the DNA fragments that are hypothesized to be previously bound by the transcription factor under investigation. This process is shared by ChIP-EM (epigenetic modification) which will be introduced in the nextchapter.

Experimental Procedure

First of all, trypsinized cultured cells or tissue samples must be fixed by a cross-linking fixative (Fig. **1**). Precipitative fixatives are not suitable for this purpose. Normally, ~0.5-2% paraformaldehyde, or formalin, is used. This fixation step is intended to preserve the original state of the chromatin. After fixation, cells can be collected into a test tube for chromatin isolation. Once the chromatin is collected, it is fragmented into smaller pieces, usually a few hundred to a few thousand base pairs in size. For this purpose, sonication, instead of enzymatic digestion, is recommended, as the former generates chromatin fragments by random breakage caused by mechanical force, while the latter produces RE site-defined fragments. To minimize bias and artifacts, chromatin fragments resulting from random breakage are preferred.

After sonication, the desired chromatin fragments are enriched by a specific antibody which recognizes, and binds to, the TF of interest. (Remember, only fixed chromatin fragments can be enriched by an antibody!). ChIP-TFBS analysis relies on antibodies against specific epitopes in the transcription factor. This step enriches TF-bound ChIP fragments. All proteins are then erased (removed) from the enriched DNA fragments, which are subsequently end-repaired, tailed with an 'A', ligated to sequencing adaptors, denatured, annealed to anchored oligos for *in situ* PCR amplification, and then subjected to sequencing. A ChIP-TFBS library can be sequenced as a (shotgun) fragment library, a PE (paired-end) library, or a PED (paired-end ditag) library, depending on the investigator's preference. However, to avoid confusion resulting from low coverage, a PED library is strongly recommended. Sequences produced by the sequencer are expected to harbor the genomic locations bound by the TF under investigation.

The key to the success of ChIP-TFBS analysis relies on using the appropriate fixatives and antibodies at the appropriate concentrations and under the appropriate conditions. The cross-linking fixative preserves DNA/RNA and bound proteins in their original

locations, but still allows the bound proteins to be removed/degraded in a later step. It is not difficult to determine which fixative to use. There are two types of fixatives (either cross-linking fixatives or precipitative fixatives) and the cross-linking fixative paraformaldehyde has long been known to best serve this purpose. On the other hand, it can be difficult to define the optimal fixation conditions because the efficiency of fixation is affected not only by sample type (*e.g.*, tissue culture, tissue section, *etc*.) and material sources (*e.g.*, microorganism, fish, human, *etc*.), but also by the thickness and conditions of preservation (*e.g.*, fresh tissue section, liquid nitrogen-preserved tissue section, paraffin-embedded tissue sections, *etc*.). As such, optimization of fixation conditions may be required for each experiment. Similarly, the antibody chosen plays a key role in determining the specificity of binding. Since the specificity is influenced by a number of factors (including the Ab itself, its concentration, salt concentration, temperature and the stringency of wash conditions, *etc*.), optimization of hybridization and wash conditions are normally required for each antibody.

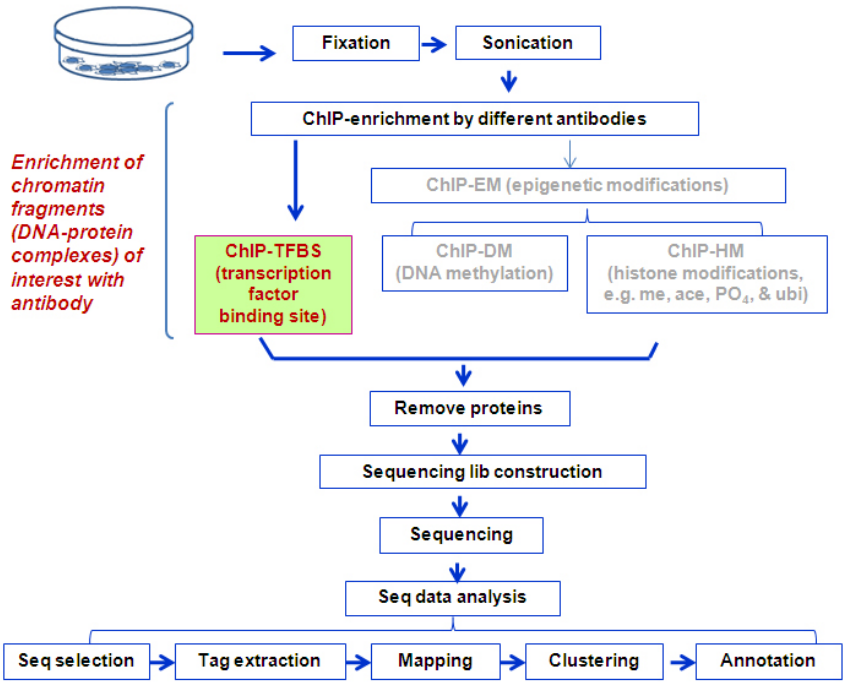

Fig. (1). Experimental procedure for ChIP-TFBS analysis.

Sequence Data Analysis

The initial sequence processing is basically the same for all types of sequence data. That is, the raw sequences need to be cleaned up (including decontamination; trimming of vector sequences, if any; removal of low-quality regions/sequences, *etc*.) and sometimes re-organized. The quality reads are then mapped against the reference genome to identify their genomic origins (chromosomal locations). Aligned reads are grouped into clusters

CHAPTER 13

ChIP-EM Libraries

Abstract: Epigenetic modifications (EMs) refer to the external modifications on DNA that do not alter coding specificity. EMs include DNA methylations (DMs) and histone modifications (HMs). This chapter will focus on HMs. We will discuss how ChIP-EM libraries can be made and what can be expected from the sequence data analysis. There are many laboratories working in this field and many reports have been published. Readers are recommended to consult the previous reports for further understanding of this subject.

Keywords: Chromatin immunoprecipitation, ChIP, EM, Epigenetic modifications.

Definition of Terminologies

Epigenetic Modifications

Epigenetic modifications (EM) refer to the mitotically and/or meiotically heritable, but biochemically reversible, on-chromatin modifications, which confer phenotypic plasticity by coordinating the expression of multiple genes in a 3-dimensionally dispersed but functionally correlated manner without entailing any change in DNA sequence. These modifications include 1) the use of different histone variants, 2) the methylation, acetylation, phosphorylation, ubiquitination, or sumoylation of nucleosomal histone proteins, and 3) DNA methylations. This terminology was first described by Conrad Waddington in his paper entitled "The epigenotype" published in 1942 (Waddington, 2012).

INTRODUCTION

Chromatin structures are molecular complexes made primarily of both DNA and proteins. Histones are the most prevalent protein species in chromatin. By forming positively charged histone cores, each of which is made up of 2 copies of (H2AH2BH3H4) and left-handedly wrapped by ~147 bp of negatively charged DNA, histone proteins play a key role in DNA packaging. As noted in Chapter 1, there are three types of chromatin - heterochromatin, euchromatin, and facultative euchromatin - which occupy distinctive segments and form scattered euchromatin (loosely packaged) or heterochromatin (densely packaged) islands in the nuclear genome. Certain amino acid residues in the N-terminal protrusions, or so-called "tails", of histones are posttranslationally modified by a few types of small moieties. These modifications, including acetylation, methylation, phosphorylation, ubiquitination, sumoylation, ADP ribosylation, deamination, and proline

Kuo Ping Chiu

isomerization, affect the degree of DNA packaging and thus influence DNA binding by various proteins (including TFs, cofactors, RNA polymerase, helicase, and topoisomerase), which in turn influences the expression of genes in the modification regions. In fact, functional characterization has also implicated histone modifications in multiple biological processes, including DNA replication, DNA repair, apoptosis, embryogenesis, cell cycle regulation, and embryonic and neuronal development (Arnaudo and Garcia, 2013; Graff *et al*., 2011; Hirabayashi and Gotoh, 2010; Kouzarides, 2007). There are a total of about 60 residues in each histone core which can be modified by at least one moiety per site. This raises the question of how such complex modifications in nucleosomal histones regulate gene expression and participate in so many biological functions.

THE LAW OF UNCERTAINTY AT THE EPIGENETIC MODIFICATION LEVEL

In physics, Heisenberg's uncertainty principle follows the fomula $SD_x \times SD_y \geq h/2$, where SD_x stands for the standard deviation of position, SD_y stands for the standard deviation of momentum, and h stands for Plank's constant. Uncertainty exists in every object under study due to the influence introduced by the interaction between objects. When extrapolated from the quantum level to the molecular level, and further to the cellular level or higher, the law of uncertainty in physics introduces uncertainty to the biological system as all biological phenomena obey chemical and physical laws. Since we cannot precisely define the position of an electron, we cannot precisely define the shape or position of a protein. In the strict sense, we more or less adhere to the concept of probability when we describe a biological information.

In Biology, the Law of Uncertainty is Shown in Genetic Mutations (*e.g.*, SNVs), as Well as at the Level of Epigenetic Modifications

Epigenetic modifications (as well as TFBSs) are very dynamic. Moreover, there are antagonistic and synergistic interactions between epigenetic modifications and, with so many modifiable amino acid residues (~60 per nucleosome) and some residues that may have multiple types of modifications, the histone modifications alone are both spacially and temporally overcrowded, further amplifying the antagonistic and synergistic effects among EMs. Antibodies and mass spectrometry are the most commonly used methods for studying EMs. Like all other methods, both of these methods have their intrinsic limitations. For example, antibodies have their limitations (uncertainties) in specificity and sensitivity, while mass spectrometry has its limitations (uncertainties) in fragmentation and sensitivity. Both limit the accuracy of EM investigations.

DNA methylations occur most frequently in the cytosines of CpG islands, some of which are found in promoter regions of genes and some in the intragenic or intergenic regions.

CpG islands are typically about 300 – 3,000 bp in size. About 40% of mammalian genes have CpG islands in the promoter, some of which are methylated. DNA methylation in CpG islands has frequently been found to result in transcriptional repression, although there are some obvious exceptions. Cross-talk between DNA methylation and histone modification has also been observed (Cedar and Bergman, 2009; Fischle *et al*., 2003).

Epigenetic regulation seems to be a combinatorial effect of multiple types of modifications in DNA and histone proteins. These modifications, which may or may not induce alterations in DNA packaging, are able to synchronize the expression of hundreds or thousands of genes dispersed in large genomic segments and simultaneously modulate the expression of limited numbers of genes in the same or similar regions.

However, order can be created from disorder. Here we will focus on EM study by antibodies, *i.e.*, by the ChIP-EM approach.

EXPERIMENTAL PROCEDURE

Mapping of the genetic modification sites is made possible by using antibodies, each of which reacts to a specific type of modification in a specific location of the histone tails, to enrich the chromatin fragments containing the modification of interest (Zhao *et al*., 2007). Sequencing of the DNA portion of the enriched chromatin fragments followed by mapping sequences against the appropriate reference genome, the genome-wide locations of the modification can be identified. This type of information can be correlated to other information for a comprehensive understanding of the cellular status (Figs. **1** and **2**).

Construction of ChIP-EM Sequencing Libraries

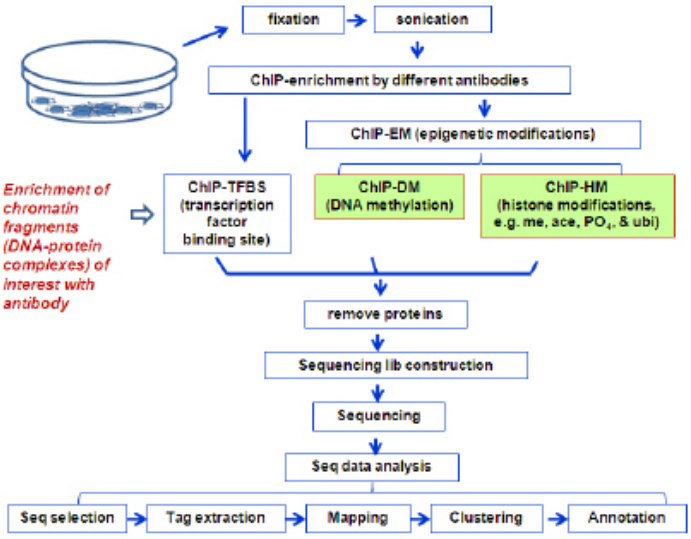

Fig. (1). Procedure for ChIP-EM (epigenetic modification) analysis.

CHAPTER 14

MicroRNA Analysis

Abstract: MicroRNAs (miRNAs) negatively regulate mRNA species by binding to the 3' untranslated region (3' UTR) in mRNA through nucleotide complementarity which allows limited number of nucleotide mismatches to fine tune the target specificity and the degree of represion. Although miRNAs have been intensively studied for decades, most of their targets and functions remain unknown. Furthermore, many of the miRNAs that have been studied are known to target multiple mRNAs. These properties seriously impede the progress of miRNA analysis. Analysis of miRNAs normally relies on commercial kits for miRNA isolation and sequencing library preparation. This chapter will serve as a general introduction of miRNA analysis. Most of the experimental procedure and sequence data analysis discussed in this chapter can also be found in the paper entitled "global assessment of *Antrodia cinnamomea*-induced microRNA alterations in hepatocarcinoma cells" published in 2013.

Keywords: 3' UTR, MicroRNA, MiRNA, Non-coding RNA.

INTRODUCTION

miRNAs are small single-stranded non-coding RNAs of ~18–24 nucleotides that have been postulated to post-transcriptionally regulate up to 50% of genes in both plants and animals (Bushati and Cohen, 2007; Friedman *et al*., 2009; Jovanovic and Hengartner, 2006; Kaikkonen *et al*., 2011). The biogenesis of miRNA follows a distinct pathway. Similar to mRNAs, most miRNA genes are transcribed by RNA polymeras II to generate primary miRNA (pri-miRNA) which also contains a 5'cap and a 3' polyA. A complex of Drosha and DGCR/Pasha cleaves the pri-RNA to produce ~70nt hairpin-shaped precursor miRNA (pre-miRNA), which is subsequently transported by Exporin-5 to the cytoplasm (Bohnsack *et al*., 2004; Du and Zamore, 2005; Yi *et al*., 2003), where the pre-miRNA is cleaved by a complex of Dicer and TRBL/Loquacious, releasing the double stranded ~21nt (miRNA-miRNA* duplex) mature miRNA. In most cases, the miRNA* strand is degraded, whereas the 5' end of miRNA is incorporated into an RNA-induced silencing complex (RISC) with Argonaut proteins to regulate its target mRNA. Through binding to the 3'UTR of its target gene, a miRNA can either degrade the target mRNA or repress its translation (Bushati and Cohen, 2007; Catto *et al*., 2011; Du and Zamore, 2005). We have recently reported the global downregulation of miRNA by *Antrodia cinnamomea* fungus (Chen *et al*., 2013).

Kuo Ping Chiu

EXPERIMENTAL PROCEDURE

Part 1: Construction of Sequencing Libraries

Preparation of the starting material

First prepare a batch of total RNA, which is expected to contain a certain amount of small RNA/miRNA

1. Determine the quantity of total RNA and the quality of small RNA.
   The quantity of total RNAcan be determined by using RNA 6000 Nano Kit in NanoDrop, while the quality of small RNA can be evaluated by, for example, RIN (RNA Integrity Number) as estimated by software. The RIN value reflects the degree of RNA degradation. If RIN > 6, then continue to the next step; otherwise, prepare a fresh sample.
2. Calculate the *percentage* of small RNA in your total RNA sample.
   Quantify the amount of 10-40 miRNA using Small RNA Chip. (% of miRNA= (amount of 10-40 miRNA divided by the amount of total RNA)*100.)
   If miRNA% ≥ 0.5%, then skip this step. That is, you can use the total RNA directly without further isolation. Otherwise, enrich the miRNA (using a commercial kit).
   We strongly recommend to ignore the miRNA% and proceed with miRNA enrichment even if the percentage is higher than the threshold given in A above, because non-miRNA species may complicate downstream reactions.
3. Enrich small RNA
   If miRNA% < 0.1%, flashPAGE Fractionator together with the flashPAGE Cleanup Kit can be used, and the size distribution of the "purified" sample is expected to be within 10-40 nt. If miRNA% ranges between 0.1-0.5%, PureLink miRNA Isolation Kit can be used for enrichment, and the size distribution of the enriched sample is expected to be within 10-200 nt.
4. Evaluate both the quality and quantity of the enriched small RNA sample
   The quality and quantity can be estimated by Agilent 2100 Bioanalyzer or fragment analyzer (AATI). (% of miRNA= (amount of 10-40 miRNA divided by the total amount of enriched RNA sample)*100).
5. Determine the quantity of input

Type of input	**Estimated conc. of miRNA w/ size between 10-40 nt (on Small RNA Chip)**	**RNA input**
Total RNA	5 – 100 ng/mL	≤ 1 mg
Enriched small RNA (size within 10-200 nt	1 – 100 ng/ mL	≤ 1 mg
flashPAGE-enriched small RNA (size within 10-40 nt)	1– 100 ng/ mL	0.1 mg

Part 2: Construction of the Amplified Small RNA Library

The general procedure includes the following steps: 1) hybridization and ligation of small RNA molecules to reverse transcriptase (RT) adaptors; 2) reverse transcription; 3) purification of cDNA; 4) size selection of cDNA (*e.g.* by gel excision); 5) amplification of cDNA by PCR; 6) purification of PCR-amplified DNA; and 7) assessment of the yield and size distribution of the amplified DNA (*e.g.* using Fragment Analyzer™ (AATI)); and 8) sequencing library construction.

Part 3: miRNA and mRNA Data Processing and Analysis

This section aims to give the readers a general idea, and an example, regarding how the miRNA sequence data is processed and analyzed.

For each miRNA library, the qualified sequence reads will be retained for downstream analysis by meeting the criterion of a Phred Quality Score of QV ≥ 20, which is equivalent to 99% accuracy. To process four libraries (2 hr or 4 hr for each, treated or untreated) of sequences with lengths of 35 bp generated from a sequencer, we built an in-house pipeline using shell script combined with custom-made perl script. To avoid having to trim adapters by using duplicated datasets, each identical sequence was clustered and assigned a unique tag with the bp count of the sequence (*e.g.*, tag_100). Cutadapt (Martin, 2011) was used to trim ligated adapters with fewer than 3 mismatches. Then, polyN was cleaned and identical reads shorter than 16 bp and longer than 30 bp were removed to match the range adopted by miRBase using custom made perl script.

To identify known and novel miRNAs, we eliminated tRNA and rRNA sequences from the libraries by mapping sequence reads against Rfam (http://rfam.sanger.ac.uk/) and tRNA (http://lowelab.ucsc.edu/GtRNAdb/) using the Bowtie program (Langmead *et al*., 2009). Sequence reads mapped to those databases without mismatch were removed from the libraries. In addition, repetitive sequences in the libraries that mapped to Repbase (http://www.girinst.org/repbase/) were also removed, after which miRBase (http://www.mirbase.org/) was employed to identify the known miRNAs from the libraries. The remaining sequence reads were mapped to the UCSC hg19 database and only sequence reads successfully mapped to the UCSC hg19 with fewer than 3 mismatches were considered to be novel miRNAs. In our miRNA profile, we only used detected known miRNA with at least 2 reads or more for subsequent analysis.

DISCLOSURE

Part of this chapter has been previously published in PLoS ONE 8(12): e82751. doi:10.1371/journal.pone.0082751. 2013

CHAPTER 15

Application of NGS in the Study of Sequence Diversity in Immune Repertoire

Abstract: During evolution, the immune system evolved as a defense mechanism to protect organisms against pathogens. Since pathogens in the environment are extremely diverse and unpredictable, strategies taken by the immune system have to be highly diversified in order to mount an effective response. At the molecular level, the sequence diversity present in the variable regions of antibody-coding and TCR-coding genomic sequences is eventually reflected in the amino acid sequences of their encoded proteins as seen in the circulation system and on the surface of immune cells. At the cellular level, B cells, T cells, dendritic cells, and many other immune cells have to interact coordinately with one another in order to foster the maturation of an immune response (*e.g.*, affinity maturation) against pathogenic attack. With the advent of NGS technologies, the complexity of the immune system can now be studied in greater detail.

Keywords: Phage display, Single chain variable fragment, ScFv, Single domain antibody, SdAb, TCR, VDJ recombination.

INTRODUCTION

With next-generation sequencing, we are now able to study the immune repertoire (*e.g.*, VDJ recombination events during B cell maturation) by high throughput sequencing of genomic sequences or mRNAs from various stages of B-cell or T-cell development. The amino acid sequences of their corresponding proteins can be deduced from their corresponding nucleotide sequences. Previously most studies focused on understanding immune processes by deducing function from structure (structure -> function). We now can add 'sequence' to the upstream of the process (sequence-> structure-> function). Details are illustrated in the following sections.

PART I. SEQUENCING THE IMMUNE REPERTOIRE

A. Characterization of a Natural Antibody Repertoire

Antibody repertoires are highly plastic and can be directed to produce antibodies with broad chemical diversity and extremely high selectivity. The work published by Weinstein *et al*., in 2009 represents a great example to demonstrate how an immune repertoire can be analyzed comprehensively and thoroughly by deep sequencing

Kuo Ping Chiu

(Weinstein *et al*., 2009). In the work, the authors studied the variable domain of the antibody heavy chain and analyzed the VDJ usage using mRNA samples isolated from zebrafish. In principle, similar approach can be applied to the study of TCR variable regions. Moreover, by using the genomic DNA, instead of mRNA, it can be applied to study the recombination events inside the nucleus. The experimental procedure adopted by Weinstein *et al*. is summarized below.

Experimental Procedure

First, multiple wild type zebrafish were collected and each fish was homogenized in the presence of TRIzol. Total RNA was then extracted from the fish and mRNA was subsequently isolated from the total RNA. Then, cDNA libraries were synthesized using superScript™ III reverse transcriptase. PCR amplification using 27 forward primers located within the consensus leader sequences for 39 functional V gene segments together with reverse primers located within the first 100bp of the IgM and IgZ constant domain were conducted to capture the entire complementarity-determining region 3 (CDR3), which should contain the vast majority of antibody diversity. PCR amplified fragments were then subjected to next-generation sequencing using 454 FLX.

The sequencer output a total of 640 million bases from 14 zebrafish, equivalent to 28,000-112,000 useful sequence reads per zebrafish. By sequence alignment to a reference genome, sequence reads were mapped to V and J segments in the genome (success rate reached 99.6%; failures were mainly caused by similarity in the V segment). Alignment to the D segment was determined within the VJ region to all reads using a clustering algorithm, The success rate was 69.6%, and many of the unassignable reads had D segments mostly deleted.

Their work resulted in a number of discoveries. For example, they found that 50-86% of all possible VDJ combinations were used and that zebrafish shared a similar frequency distribution of VDJ usage. Moreover, there was a correlation of VDJ patterns between individuals. They also demonstrated an evidence of convergence, as indicated by the fact that different individuals may make the same antibody.

B. Sequencing scFv and sdAb for Therapeutics

Recent immune engineering using synthesized DNA sequences, which may contain degenerate and constant regions, has been able to create chimeric antibodies for therapeutic purposes. This strategy opened up a new dimension for the usage of next-generation sequencing (Chang *et al*., 2014; Hsu *et al*., 2014).

For example, a single chain variable fragment (scFv) is a simplified fusion antibody produced by directly linking the variable regions of a heavy chain and a light chain through a “linker peptide” of ~25 aa in length. The order of V_H and V_L is interchangeable,

making it either N-(V_H)-(linker peptide)-(V_L)-C or N-(V_L)-(linker peptide)-(V_H)-C in structure. Since both the variable regions of the heavy chain and the light chain are present in its structure, an scFv can bind its antigen with high specificity. Moreover, the structure and function of their coding sequence can be easily tested using phage display libraries. As such, this scFv provides a convenient approach for quick Ab production and testing for therapeutics.

Selection of high-affinity variable sequences using phage display screening

Phage display screening is a powerful approach for selecting high-affinity variable sequences from an immune repertoire for sequencing. A general workflow is shown in Fig. (**1**).

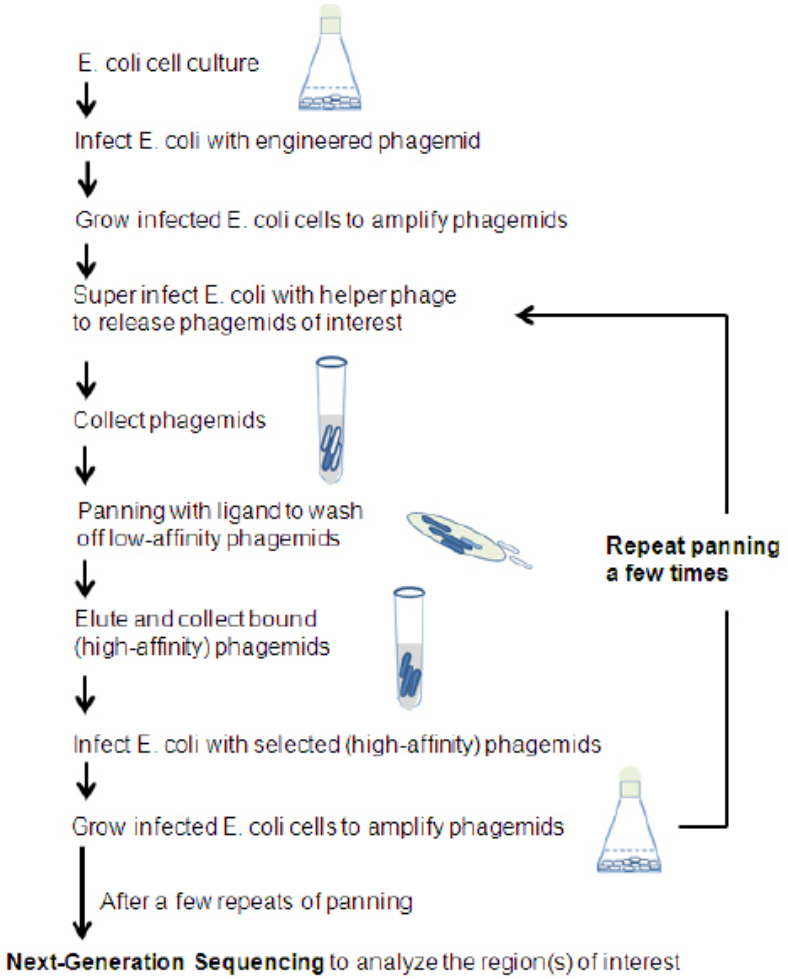

Fig. (1). Phage display screening for high-affinity phagemids. Phage display screening enriches high-affinity (or low-affinity if the counterpart is desired) phagemids. The procedure consists of a few transductions, growth of bacterial cultures and a series of panning used to separate low-affinity phagemids from its high-affinity counterpart.

CHAPTER 16

Galaxy Pipeline for Transcriptome Library Analysis

Abstract: Next Generation Sequencing (NGS) provides researchers with an unprecedented opportunity to produce a large volume of DNA sequences quickly, and is one of the fundamental methods for high-throughput genomic studies. Currently, the most widely-used NGS platforms are Illumina, Roche 454 and SB SOLiD. These platforms differ in the chemistry used in the sequencing process and the length of sequencing read generated. Each platform has its own strengths and weaknesses. In particular, the required length of the sequence read to be generated plays an important role when designing an experiment. For example, a longer read length would be needed in the assembly of a novel genome, while throughput-maximizing PED-based techniques would be better-suited when shorter reads will suffice.

Keywords: ChIP-Seq, Galaxy, RNA-Seq.

INTRODUCTION

Previous chapters have provided readers with the basic concepts behind various NGS platforms and their applications. In this chapter, we present tools that are widely used for data analysis. Over the years, a large number of tools have been created for analyzing NGS data. Many of these tools require creating local databases, and familiarity with the Unix/Linux operating system, which poses a huge challenge to non-computational scientists, is needed. Galaxy was designed to overcome these issues by providing an open, web-based platform that integrates various databases and tools together with a simple interface for performing NGS data analysis.

GALAXY INTERFACE

Galaxy is a popular pipeline for RNA-Seq (Trapnell *et al*., 2012). It is available from http://usegalaxy.org. The interface is simple and intuitive, as shown in Fig. (**1**). This is the Analyze Data window where users can perform data analysis. The interface is organized into four main sections: the top Menu bar, Tools panel (left column), Operation panel (middle column), and History panel (right column). The Tools panel provides the users a list of available algorithms that are useful for processing and analyzing NGS data. The Operation panel is used as the input interface for tools and information display. The History panel keeps track of all operations, parameters, and input and output datasets.

Kuo Ping Chiu

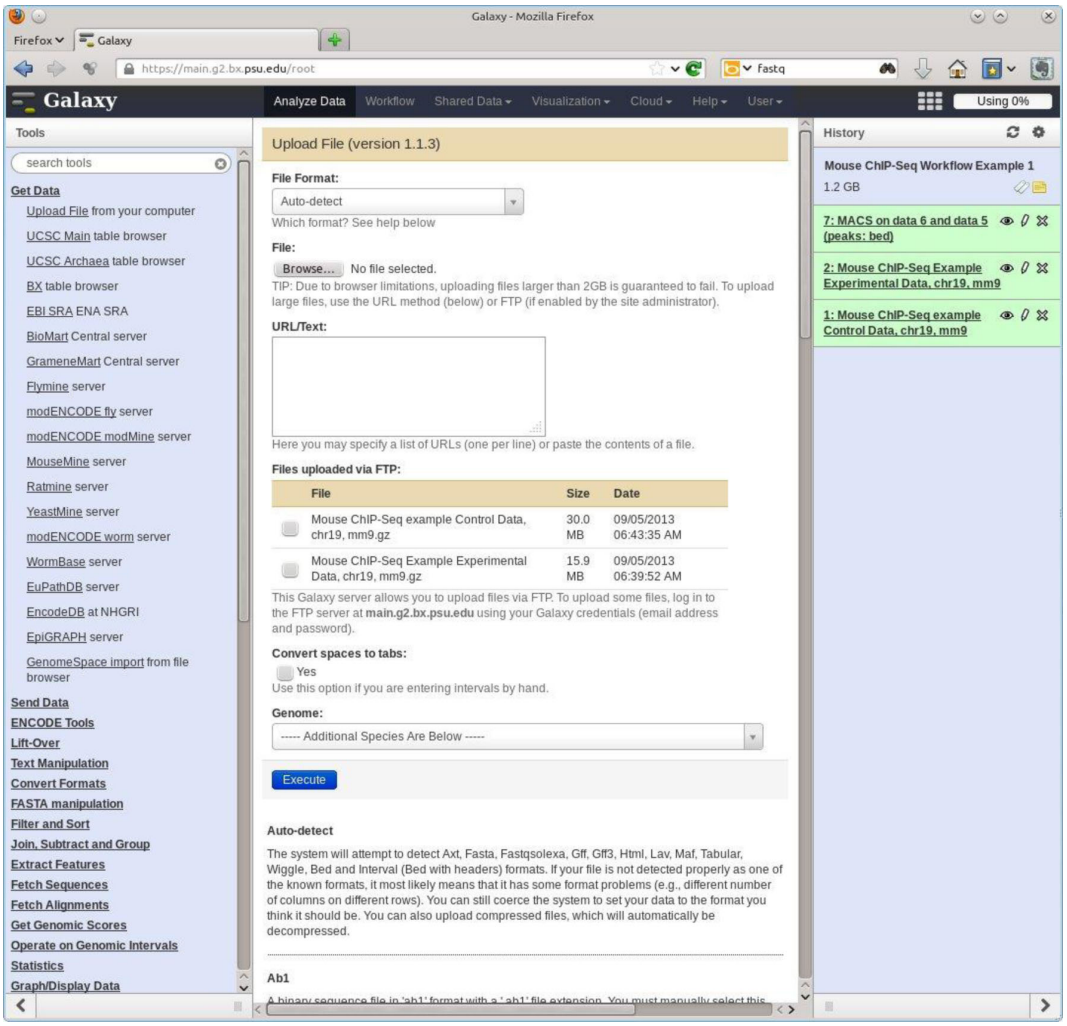

Fig. (1). Galaxy interface is organized into four sections: Menu bar (top), Tools panel (left), Operation panel (middle), and History panel (right). The menu bar allows users to switch between various functional interfaces. The Tools panel contains all the available data analysis integrated into Galaxy. The middle panel is used as the input interface for an algorithm and to display information or results. Pictured here is the Upload File interface. The History panel keeps track of the datasets and operations. There are three datasets in the current history. Datasets 1 and 2 are the input datasets, and dataset 7 is the result of applying a workflow program to the two input datasets.

There are a large number of tools integrated into Galaxy. These tools are grouped into various sections, according to their functions. For example, the Get Data section contains a list of tools for getting data into Galaxy. These tools allow users to upload data from their computer or retrieve information from public resources such as the UCSC Table Browser and the BioMart server.

In the following sections, we describe the functionality of Galaxy for analyzing ChIP-Seq data. ChIP-Seq is a method used to study protein-DNA interactions, which it does by

replacing the microarray ("chip") used in the ChIP-chip technique with sequencing ("Seq"). These methods have been widely used to identify the DNA-binding sites of transcription factors and the locations of histone modification.

FASTQ FORMAT

Many sequencing platforms can be used for sequencing. While they vary in the ways they encode the quality scores, as exemplified below, these sequencing platforms produce data in the standard Fastq format. And, unlike the old Fasta format, the scores of all nucleotides in a read are also included. Each read is represented in four lines (Fig. **2a**): 1) the header line, which begins with an '@' character and is followed by a sequence identifier; 2) the raw sequence letters; 3) the third line begins with a '+' character and is optionally followed by the same sequence identifier; and 4) the fourth line gives the quality scores.

a

```
@HWI-ST_0216:7:1101:1742:2161#CGATGG/1
AGCCGGGACCCCGCCGATCACACCGGAGATGGTCCGTCGCGCGCTCGATGAGGACTGACGCGGGTGGCGCTGCTCGACGTCAACGCATTGGTCGCGCTGG
+HWI-ST_0216:7:1101:1742:2161#CGATGG/1
___cccc^beeZc^fa[_U^Z^eZ`eggUbHHM\\`a_[a^aaWVOY_b]]bbb_X^`GTZaaaTZa[]aaaOSYWLOX]OGT]_aX^ab_]]]O[][aa
@HWI-ST_0216:7:1101:1679:2241#ATGGCA/1
GCACCAACATTGATGAGCTGCTACTGGAACAACCCGGAGGCCACCGCGGAGGCGTTCGCAGGCGGCTGGTTCCATTCTGGGGATCTGGTTCGTATGGACT
+HWI-ST_0216:7:1101:1679:2241#ATGGCA/1
_^_eeeee^cgaegd`ehgg`fhegghbaffddPbec`egddgfefhfgc`^^a^^W^aV_cZacaTZabcbbbc_b`]`b__a^_bb_bba[[b[X_`b
```

b

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.......................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                        |    |       |                              |                     |
33                       59   64      73                            104                   126
 0.....................26...31.......40
                         -5....0........9.............................40
                              0........9.............................40
                                 3.....9.............................40
 0.2...................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Fig. (2). An example of Fastq format **(a)**. Each sequence is represented in four lines, sequence header, sequence, quality header, and quality line. Various schemes can be used to encode quality scores **(b)**. The numeric scores are converted to the corresponding ASCII code.

As noted, each platform uses different schemes to encode the quality scores. The initial scheme was introduced by the Sanger Institute; it encodes a Phred quality score from 0 to 93 using ASCII 33 to 126. The currently available encoding schemes are shown in Fig. (**2b**). Please note that the analysis modules in Galaxy require Sanger quality scores.

SUBJECT INDEX