Revolution of nephrology research by deep sequencing: ChIP-seq and RNA-seq

Imari Mimura[1,2], Yasuharu Kanki[2], Tatsuhiko Kodama[2] and Masaomi Nangaku[1]

[1]*Division of Nephrology and Endocrinology, The University of Tokyo, Tokyo, Japan and* [2]*Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan*

The recent and rapid advent of next-generation sequencing (NGS) has made this technology broadly available not only to researchers in various molecular and cellular biology fields but also to those in kidney disease. In this paper, we describe the usage of ChIP-seq (chromatin immunoprecipitation with sequencing) and RNA-seq for sample preparation and interpretation of raw data in the investigation of biological phenomenon in renal diseases. ChIP-seq identifies genome-wide transcriptional DNA-binding sites as well as histone modifications, which are known to regulate gene expression, in the intragenic as well as in the intergenic regions. With regard to RNA-seq, this process analyzes not only the expression level of mRNA but also splicing variants, non-coding RNA, and microRNA on a genome-wide scale. The combination of ChIP-seq and RNA-seq allows the clarification of novel transcriptional mechanisms, which have important roles in various kinds of diseases, including chronic kidney disease. The rapid development of these techniques requires an update on the latest information and methods of NGS. In this review, we highlight the merits and characteristics of ChIP-seq and RNA-seq and discuss the use of the genome-wide analysis in kidney disease.

Kidney International advance online publication, 28 August 2013; doi:10.1038/ki.2013.321

KEYWORDS: ChIP-seq; epigenetics; genome-wide analysis; RNA-seq

Correspondence: *Masaomi Nangaku, Division of Nephrology and Endocrinology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. E-mail: mnangaku-tky@umin.ac.jp*

Received 17 August 2012; revised 21 May 2013; accepted 13 June 2013

Next-generation sequencing (NGS) has recently become a familiar research tool. NGS not only enables the detection of novel genes and single-nucleotide polymorphisms (SNPs), which have important roles in human disease, but also provides a variety of information about what kinds of functions are associated with genes in these conditions. Genome-wide analysis, including systems biology as well as DNA methylation and microRNA, is well established and has been reviewed in detail elsewhere.[1,2] In the present review, we introduce studies on nephrology and other fields. The mechanisms of acute kidney injury and chronic renal fibrosis are the main topics in nephrology. However, NGS has not been fully used to analyze these mechanisms, because it is still a new technique. NGS provides us a great amount of new information and insights in nephrology. We also discuss how to put the massive amounts of data into practical use in nephrology research using the following methods for genome-wide analysis, particularly ChIP-seq (chromatin immunoprecipitation with sequencing) and RNA-seq.

CONTENTS

Applications of NGS

The technology of sequencing has markedly evolved since Sanger developed the first technical methods for DNA sequencing in 1975. The ever-increasing demands of researchers for more DNA sequencing at lower cost have led to the development of NGS. NGS makes use of the technology that reads short fragments of genomic sequences and then aligns them to known genome sequences, and manufacturers provide machines that can analyze higher resolution, fewer artifacts, greater coverage, and a larger dynamic range than microarrays.[3] These sequencers differ in terms of their operational principle, output file format, and amount of raw data generated, and researchers therefore need to decide which is most appropriate for their particular interests. The most striking characteristic of GS FLX+ by Roche, for example, is that it reads fragments of 700–1000 bp, which are longer compared with those read by other sequencers. This sequencer is thus very useful for finding sequences *de novo* by comparison with a known reference sequence (Refseq), which is an integrated annotated set of sequences that include genomic DNA, transcripts, and proteins.

Products by Illumina (San Diego, CA) are developed on the basis of the principle of sequencing by synthesis using fluorescent probes. A characteristic of GAIIx (Genome Analyzer IIx) and HiSeq by Illumina is the striking amount of raw data that are produced for one sequence (576 Gb for HiSeq2000). They are better suited to the mapping of already known sequences (resequencing), small RNA-seq, and ChIP-seq. They are also suited for the identification of DNA methylation status, as well as SNPs and mutations. Further, these other machines are able to analyze multiple conditional experimental designs at a time and are therefore suitable for time course–dependent experimental samples. SOLiD by Applied Biosystems (Carlsbad, CA) reads DNA fragments amplified on magnetic beads, which are fixed on the slides. It can simultaneously perform not only ChIP-seq, RNA-seq, and MeDIP-seq (methylated DNA immunoprecipitation sequence) but also small RNA-seq, SNPs, and CNV (copy number variant) analysis for multiple lanes. The benefits of PacBio RS include its speed of analysis (about 30 min), and the total sequencing read length of PacBio is about 50 Mbase. Chin *et al.*[4] used this DNA sequencing to determine the genome sequences of two clinical *Vibrio cholera* isolates from the current outbreak in Haiti, indicating its suitability for *de novo* sequencing of microbes.

For clinical researchers, we consider that the GAIIx or HiSeq by Illumina or SOLiD by Applied Biosystems is the most suitable of these sequencers, on the basis of their ability to read massive amounts of data under several experimental conditions at a time and at a relatively lower cost. They can also analyze small RNA fragments by making cDNA library with reverse transcription. These attributes will strongly facilitate the understanding of not only genetics but also epigenetics of human diseases.[5]

ChIP-seq

Methods for cell fixation. Paraformaldehyde is used as a standard method of fixing DNA and DNA-binding proteins, such as transcriptional factors for immunoprecipitation. There are two general methodologies for ChIP experiments: chromatin fixation with paraformaldehyde (XChIP) and native chromatin prepared by nuclease digestion of cell nuclei (Native ChIP (NChIP)).[6] Selection for a particular experiment should be dependent on the advantages and disadvantages of each. In NChIP, DNA is cut into fragments not with a sonicator but with an endonuclease, such as micrococcal nuclease. The advantages of NChIP are as follows: (1) immunoprecipitation is highly efficient and (2) the specificity of antibody binding to unfixed chromatin is more predictable. The disadvantages of NChIP are as follows: (1) general inapplicability to non-histone proteins, because only proteins that bind tightly to DNA can be studied in NChIP and (2) the risk of protein rearrangement during chromatin preparation and precipitation.[6]

As an example of its use, we performed ChIP-seq by NChIP using di-methyl H3K9 (H3K9me2) antibody after culturing human umbilical vein endothelial cells under normoxia and 1% hypoxic conditions for 24 h (Figure 1a). As a result, we succeeded in detecting the enrichment of H3K9me2 using NChIP but not with XChIP. MageA2 is a positive control gene, which is enriched in H3K9m2 (Figure 1b).[7] We assume that the repressive mark H3K9me2, which is known to characterize the heterochromatin region, is more difficult to fractionate than the open chromatin, which is modified with active marks such as H3K4me1 and H3K4me3.

Analysis of raw data. Methods for analyzing the raw data of ChIP-seq using GAIIx by Illumina are shown in Figure 2. The DNA fragments between 200 and 600 bp are immunoprecipitated by specific antibody (Figure 2, Step 1) and sequenced by NGS (Figure 2, Step 2). NGS provides raw data as sequence reads (Figure 2, Step 3). The total number of sequence reads is between 10,000,000 and 100,000,000. Information of genomic position (Refseq) is added to the raw data files, a step called alignment (Figure 2, Step 4).

Next, the files with information on genome position are changed into bed files using analysis tools, a step called mapping (Figure 2, Step 5).[8,9] The bed file data are formatted as enrichment peaks on Refseq and are viewable in browsers, as shown in Figure 1.

Compared with ChIP-chip analysis, ChIP-seq has a wide variety of merits. One of the greatest improvements of ChIP-seq over ChIP-chip is genome-wide analysis of DNA fragments collected by ChIP. ChIP-chip has limitations of regions for mapping of DNA fragments, because probes cannot cover all the genome. In contrast, the genome coverage of ChIP-seq is not limited by the ranges of probe sequences fixed on the array. Another improvement is the base-pair resolution. Although ChIP-chip requires a large number of probes, arrays also have fundamental limitations due to the uncertainties in the hybridization process, while the read of GC-rich regions of ChIP-seq is also difficult to read. In addition, ChIP-seq does not suffer from hybridization step.

In contrast, one of the disadvantages of ChIP-seq is PCR bias, because it needs PCR amplification to detect visible signals. Finally, ChIP-seq requires only a small amount of DNA, which is necessary for analysis, while ChIP-chip requires a few micrograms of DNA for hybridization.

Another disadvantage of ChIP-seq is high cost. The overall cost of ChIP-seq includes machine-maintenance fee and reagent cost. In addition, we need to perform biological replicate experiments of both ChIP-seq and ChIP-chip in order to ensure reliability of the data.[10] These replicate experiments need further high costs.

Using ChIP-seq data in nephrology. ChIP-seq has been widely used in fields related to nephrology, with a number of outstanding outcomes. Sun *et al.*[11] performed ChIP-seq experiments using RNA polymerase II antibody in five adult mouse tissues, including the kidney. They clarified that 6384 promoters are tissue specific among 12,270 novel promoters. ChIP-seq analysis of RNA polymerase II made it possible to identify the novel tissue-specific promoters, including the kidney.

Susztak *et al.* performed ChIP-seq of the active histone modification H3K4me3 using human biopsy samples of