Techniques and Approaches to Genetic Analyses in Nephrological Disorders.

Laurel K. Willig
Children's Mercy Hospital

Follow this and additional works at: https://scholarlyexchange.childrensmercy.org/papers

Part of the Congenital, Hereditary, and Neonatal Diseases and Abnormalities Commons, Genetics and Genomics Commons, Nephrology Commons, and the Pediatrics Commons

Recommended Citation

This Article is brought to you for free and open access by SHARE @ Children's Mercy. It has been accepted for inclusion in Manuscripts, Articles, Book Chapters and Other Papers by an authorized administrator of SHARE @ Children's Mercy. For more information, please contact library@cmh.edu.
Techniques and Approaches to Genetic Analyses in Nephrological Disorders

Laurel K. Willig

1 Center for Pediatric Genomic Medicine, Kansas City, Missouri, United States
2 Division of Nephrology, Children’s Mercy Hospital, Kansas City, Missouri, United States
3 Department of Pediatrics, University of Missouri-Kansas City, Kansas City, Missouri, United States


Abstract

Inherited renal disease is a leading cause of morbidity and mortality in pediatric nephrology. High throughput advancements in genomics have led to greater understanding of the biologic underpinnings of these diseases. However, the underlying genetic changes explain only part of the molecular biology that contributes to disease manifestation and progression. Other omics technologies will provide a more complete picture of these cellular processes. This review discusses these omics technologies in the context of pediatric renal disease.

Introduction

In 1985, researchers identified the first renal disease gene, polycystic kidney disease 1 (PKD1), responsible for autosomal dominant polycystic kidney disease (ADPKD), but renal gene discovery progressed slowly prior to sequencing of the human genome and development of high throughput next generation sequencing (NGS).1,2 Currently, there are more than 1,200 genes associated with renal disease and development, largely owing to advancements in genomic technology.3 This collection of genes includes more than 160 genes associated with rare inherited kidney diseases that contribute greatly to morbidity and mortality in pediatric renal disease.1,4 Investigators and clinicians are using this technology to diagnose Mendelian diseases such as Alport syndrome, search for novel pathogenic genes in diseases such as focal segmental glomerulosclerosis (FSGS), and identify risk genes in hypertension.5–8 These genetic discoveries allow for further elucidation of the mechanistic underpinnings of inherited kidney disease.

However, genomics and NGS technology address only one aspect of the molecular biology. Other high throughput omics technologies attempt to address the missing pieces and complete the biological story. The high throughput methods of epigenomics and transcriptomics utilize basic NGS principals to describe regulation of genomic information, while proteomics and metabolomics use mass spectrometry (MS) to explore the downstream effects. This knowledge promises to improve prognostication and will translate into more effective treatments for children afflicted with these diseases.1,9 However, the genetic complexity of many of these diseases coupled with immature high throughput research methodology may lead to improper application and inaccurate conclusions.10

This review discusses the available omics techniques, highlighting weaknesses and illustrating their use in pediatric nephrology research. It will not review microarray technology but will instead focus on NGS and MS high throughput methods. In addition, it will briefly discuss systems biology for high throughput data analysis and some of the model

Keywords
► next generation sequencing
► nephrology
► proteomics
► genomics
► metabolomics
► epigenomics
► transcriptomics

Received
January 10, 2015
Accepted after revision
February 20, 2015
Published online
August 13, 2015

Copyright © 2016 by Georg Thieme Verlag KG, Stuttgart · New York
systems used to test these integrative approaches in pediatric nephrology.

**General Study Design Considerations for Omics Data**

**Biostatistical Considerations**

Each omics field examines a specific aspect of a molecular pathway and as a result, multiple different techniques exist to study each element (Table 1). These techniques require unique biostatistical methods and software to analyze the data. A detailed review of these methods and software is beyond the scope of this article, but this information may be found elsewhere. However, researchers should consider some basic principles when using omics data. All experiments should apply appropriate power analyses prior to implementation. Power calculations must consider the genomic or biologic space interrogated to determine adjustments in the false discovery rates or p values due to multiple testing. This adjustment relies on understanding the interdependency of the different tests which is often difficult to estimate. Thus, researchers must decide how conservative to be in their adjustment of type I error rates. In genomic studies, researchers must consider the allele frequency, and the genetic and phenotypic studies, researchers must consider the allele frequency, and the genetic and phenotypic interdependency of the different tests which is often difficult to estimate. Thus, researchers must decide how conservative to be in their adjustment of type I error rates. In genomic studies, researchers must consider the allele frequency, and the genetic and phenotypic heterogeneity of the trait of interest when considering power estimates. For omics studies of quantitative traits, power calculations should include consideration of the biological and measurement variability. Finally, the testing software selected must accommodate the massive amount and format of data generated from the experiments and be specific for the type of omics question being answered.

In addition to these common considerations, there are specific considerations unique to each high throughput method. Researchers must consider genomic unit for comparison as this determines the number of tests performed and is important for the choice of downstream analysis. In genomic association studies of rare variants for complex traits, rare variants are often collapsed into units based on gene parameters, linkage disequilibrium, or simply by proximity using sliding windows of a particular length. Unlike genomic studies with low intra-sample variability due to homogeneity of cells and the relatively static nature of DNA, the other omics fields have much more intra-sample and inter-sample variability. Researchers must consider a statistical test’s ability to capture and describe this variability when selecting their statistical approach. Many of these studies are measuring quantitative traits. The addition of more samples will help improve some of the technical biases of the study and decrease variability for the measured trait. The addition of sample replicates will help control for within-sample biases.

**Utilizing Preexisting Datasets**

Many omics studies use previously generated datasets for additional affected individuals, as validation cohorts, or as reference datasets due to the significant costs of generating data and the rarity of some conditions. There are many publicly accessible databases/repositories and ongoing initiatives to establish more (Table 2). However, this pooling of data from multiple sites introduces the possibility of population stratification and systematic differences in high throughput methods and quality control.

Phenotypic data plays an important role in identifying the underlying cause and the affected molecular pathways that cause inherited renal disease. Often phenotypic data stored in databases and repositories are sparse or poorly categorized and thus may confound study results. In attempts to improve omics data sharing and transparency, experts are establishing guidelines for both study publication and

---

**Table 1** Overview of omics data and experimental approaches

<table>
<thead>
<tr>
<th>Technology</th>
<th>Type of data</th>
<th>Common approaches</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td>Gene number, structure, function, and organization</td>
<td>Targeted gene panels, WES, WGS</td>
<td></td>
</tr>
<tr>
<td>Epigenomics</td>
<td>Regulatory elements: transcription factor binding sites, methylation patterns, nucleosome location, chromatin accessibility, histone modification</td>
<td>ChIP-seq, Methyl-seq, MethylC-seq, MeDip-seq, MBD-seq, ChiA-PET, ATAC-seq, FAIRE-seq, DNase-seq, Mase-seq</td>
<td>61,66</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>All RNA species including regulatory RNA, gene expression levels, posttranscriptional modifications</td>
<td>RNA-seq, NET-seq, CAPTURE-seq CLIP-seq, ICLIP, and PAR-CLIP</td>
<td>74,100</td>
</tr>
<tr>
<td>Proteomics</td>
<td>All proteins, protein–protein interaction</td>
<td>LCMS</td>
<td>101</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>Peptides, amino acids, nucleic acids, and organic acids</td>
<td>LCMS</td>
<td>91</td>
</tr>
</tbody>
</table>

Abbreviations: ATAC-seq, assay for transposable-accessible chromatin using sequencing; CAPTURE-seq, RNA capture sequencing; ChiA-PET, chromatin interaction analysis by paired-end tag sequencing; ChIP-seq, chromatin immunoprecipitation sequencing; CLIP-seq, cross-linking immunoprecipitation sequencing; DNase-seq, DNase I sequencing; FAIRE-seq, formaldehyde-assisted isolation of regulatory elements; ICLIP, individual-nucleotide resolution cross-linking immunoprecipitation; LCMS, liquid chromatography-mass spectrometry; MBD-seq, methylated DNA binding domain sequencing; MeDip-seq, methylated DNA immunoprecipitation sequencing; Mase-seq, micrococcal nuclease sequencing; NET-seq, native elongating transcript sequencing; PAR-CLIP, photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; WES, whole exome sequencing; WGS, whole genome sequencing.
database deposition that will hopefully help clarify these issues. Without this clarification, systematic confounding of data introduced by sequencing platforms and bioinformatics pipelines will lead to misinterpretation of results.

### Basics of Next Generation Sequencing

NGS also known as massively parallel sequencing revolutionized the field of genomics. These methods create the foundation for a wide variety of experimental approaches in genomics, epigenomics, and transcriptomics. The basic approach to NGS starts with DNA fragmentation and platform-specific adapter ligation to create a DNA library, which can be amplified later using polymerases if desired. The sequencing reactions then take place in situ on a solid surface containing complimentary sequence to the adapters. This in situ sequencing generates a massive number of sequencing reads simultaneously (i.e., in parallel), vastly improving the efficiency compared with Sanger sequencing.

NGS platforms perform sequencing and nucleotide detection in one fluid process; nucleotides are added, a detection step is performed, a wash step occurs, and the process repeats itself. These improvements decrease the per base sequencing costs and time compared with Sanger sequencing, allowing for larger adoption of NGS in research.

### Sequencing Platforms

All of the NGS high throughput methods require choice of a sequencing platform. Most platforms have built-in flexibility to perform sequencing for genomic, epigenomic, and transcriptomic experiments. While all NGS technologies have relatively high per base sequencing error rates and rely on increased coverage to overcome these errors, recent comparisons of commonly used platforms demonstrate variable performance. Common problems with this technology include short read lengths causing issues with alignment in repetitive regions of the genome, poor coverage in GC rich areas, and amplification bias for those technologies that use targeted enrichment. Combining data from multiple platforms surmounts some of the weaknesses inherent to different sequencing methods.

Modifications to the basic NGS workflow that are employed to study regulatory elements and RNA lead to additional biases largely in the generation of DNA fragments for sequencing and library preparation. For instance, fragmentation of chromatin for CHIP-seq occurs in a biased fashion with euchromatin being sheared more readily than heterochromatin. In RNA sequencing, commonly occurring biases include low complexity of transcripts, uneven transcript coverage, and antisense artifacts. There are methods that may be employed during library preparation and enrichment to help overcome some of the weaknesses.

### Bioinformatics Platforms

In addition to the variety of sequencing platforms, multiple bioinformatics pipelines exist to filter the raw data and create the data files used for analysis. In the most simplistic terms, NGS bioinformatics pipelines consist of an aligner and a variant caller. During alignment, each read is aligned to the best-fit reference sequence using a statistical algorithm and is assigned a mapping quality score. These alignments are then fed to the variant caller for assignment of variant presence and zygosity. In-depth comparisons have revealed significant differences in the overall output as well as the

---

**Table 2** Common and nephrologic databases with available omics datasets

<table>
<thead>
<tr>
<th>Database</th>
<th>Data</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes</td>
<td>Low coverage human genome datasets</td>
<td>102</td>
</tr>
<tr>
<td>MGI Mouse Genome</td>
<td>Integrated genetic, genomic, and biologic information from laboratory mice</td>
<td>103</td>
</tr>
<tr>
<td>Human Epigenome Project</td>
<td>DNA methylation in human tissues</td>
<td>104</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Functional genomic elements in multiple human tissues</td>
<td>105</td>
</tr>
<tr>
<td>Mouse Reference Transcriptome Project</td>
<td>Gene expression data on multiple mouse tissues</td>
<td>106</td>
</tr>
<tr>
<td>PRIDE</td>
<td>Proteomic datasets from large number of organisms</td>
<td>107</td>
</tr>
<tr>
<td>Human Metabolome Database</td>
<td>Detailed metabolites in the human body</td>
<td>108</td>
</tr>
</tbody>
</table>

**Nephrology**

<table>
<thead>
<tr>
<th>Database</th>
<th>Data</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Renal cDNA Bank</td>
<td>Gene expression data from renal biopsies</td>
<td>109</td>
</tr>
<tr>
<td>EURenOmics</td>
<td>Biorepository and phenotypic data for rare kidney diseases</td>
<td>110</td>
</tr>
<tr>
<td>GUDMAP</td>
<td>Gene expression datasets for genitourinary tract of mice</td>
<td>111,112</td>
</tr>
<tr>
<td>NEPTUNE</td>
<td>Biorepository and phenotypic data on patients with proteinuria</td>
<td>113</td>
</tr>
</tbody>
</table>

Abbreviations: ENCODE, Encyclopedia of DNA Elements; GUDMAP, Genitourinary Development Molecular Anatomy Project; NEPTUNE, Nephrotic Syndrome Study Network; PRIDE, PRoteomics IDEntiﬁcations database.
computational complexity and cost for different bioinformatics pipelines (Table 3). Epigenomic and transcriptomic analyses require additional data processing steps introducing additional levels of complexity to the bioinformatics pipeline.

### Genomics

#### Approach
The researcher generally selects one of three approaches for sequencing, namely, targeted panels, whole-exome

### Table 3 Overview of next generation sequencing techniques

<table>
<thead>
<tr>
<th>Platform</th>
<th>Description</th>
<th>Strengths</th>
<th>Weaknesses</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina HiSeq/Miseq</td>
<td>Reversible terminator sequencing</td>
<td>• Less GC biases in GC poor areas&lt;br&gt;• More uniform coverage&lt;br&gt;• Better coverage in simple repeat areas&lt;br&gt;• High sensitivity&lt;br&gt;• Low sample-to-sample variation</td>
<td>• GC biases in GC rich areas&lt;br&gt;• Short read length</td>
<td>35,36,114</td>
</tr>
<tr>
<td>Complete genomics</td>
<td>DNA nanoball sequencing</td>
<td>• Less GC bias in GC rich areas&lt;br&gt;• Highest total number of bases covered&lt;br&gt;• High sensitivity</td>
<td>• GC biases in GC poor areas&lt;br&gt;• Less uniform genome coverage&lt;br&gt;• Large amount of DNA required</td>
<td>36,114</td>
</tr>
<tr>
<td>ABI/SOLiD 4</td>
<td>Ligation</td>
<td>• Most specific in SNP calling</td>
<td>• Significant GC bias&lt;br&gt;• Less uniform coverage and smallest percentage of genome coverage</td>
<td>36,114</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>Proton detection</td>
<td>• Fast&lt;br&gt;• Longer read length than many technologies</td>
<td>• Higher error rates&lt;br&gt;• GC biases</td>
<td>35,36</td>
</tr>
<tr>
<td>Alignment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burrows Wheeler Aligner</td>
<td>Burrows–Wheeler algorithm based</td>
<td>• Fast&lt;br&gt;• Memory efficient&lt;br&gt;• Quality score to discard poorly aligned reads</td>
<td>• Sensitivity highly dependent on quality of sequence</td>
<td>40,115</td>
</tr>
<tr>
<td>Novoalign</td>
<td>Hash table-based approach</td>
<td>• Good for long indels if mapping quality is good&lt;br&gt;• Performs well with variety of variant callers</td>
<td>• Sensitivity highly dependent on quality of sequence</td>
<td>115</td>
</tr>
<tr>
<td>GSNAP</td>
<td>Hash table-based approach—genome</td>
<td>• Genome indexing tool&lt;br&gt;• Detects complex variants and splicing in individual reads</td>
<td>• Higher error percentage on mapped reads requiring filtering</td>
<td>115</td>
</tr>
<tr>
<td>Bowtie</td>
<td>Burrows–Wheeler algorithm based</td>
<td>• Highest quality reads&lt;br&gt;• Fast</td>
<td>• Low SNP detection rate</td>
<td>40,115</td>
</tr>
<tr>
<td>Variant Caller</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATK</td>
<td>Gold standard</td>
<td>• Gold standard for SNP calling&lt;br&gt;• Most accurate for variants&lt;br&gt;• High sensitivity and positive predictive value</td>
<td>• Time intensive</td>
<td>43,44</td>
</tr>
<tr>
<td>SAM tools</td>
<td>Manipulates alignments in SAM format</td>
<td>• Accurate at lower read depths</td>
<td>• Bias toward transition events&lt;br&gt;• Performs poorly using multiple sample algorithm</td>
<td>43,44</td>
</tr>
</tbody>
</table>

Abbreviations: GATK, genome analysis toolkit; GSNAP, Genomic Short-read Nucleotide Alignment Program; SAM, sequence alignment/map; SNP, single nucleotide polymorphism; SOLiD, sequencing by oligonucleotide ligation and detection.
sequencing (WES), or whole-genome sequencing (WGS) (Table 4). Targeted gene panels sequence a select number of genes and are often customized by the investigator to include candidate genes. These genes may have been previously implicated in disease, part of the same biologic pathways as disease genes, or essential to the function of the tissue or cell of interest.47 WES is technically a targeted panel, but, due to its inclusion of all protein coding regions of the genome, it allows for a more all-inclusive approach than a targeted panel. It can examine known genes of interest as well as novel candidates. WGS identifies variants at almost all sites in the genome. It is the most inclusive testing for the identification of disease-causing variants, as it includes non-coding sequence important to gene regulation.48 Often family data including linkage analysis or prior genome-wide association studies results are utilized in data analysis to improve results in genomic testing.48,49

In analyzing data from such experiments, researchers often consider previous evidence for genetic pathogenicity. Notably, gene pathogenicity and variant pathogenicity are two different things and must be considered independently by investigators. Recent studies have shown that the literature is rife with false-positive genetic associations and many poorly validated pathologic variants are propagated in the literature.50,51 These misclassifications have led to recent attempts to establish more rigorous guidelines for variant classification including study design improvements, increased data sharing, and more comprehensive databases.

Applications

In nephrology, researchers have applied targeted NGS panels and WES to study both Mendelian disease and recent more complex diseases.1 For example, FSGS, a leading cause of steroid-resistant nephrotic syndrome in children, is a genetically complex disease. Brown et al utilized linkage analysis in a large pedigree of individuals affected by familial FSGS which identified 17 areas on multiple chromosomes that had a logarithm of odds score of >1. They then used WES in five affected individuals and one unaffected individual to identify a mutation in LIM homeobox transcription factor 1 β (LMX1B). They discovered another family with a mutation in LMX1B in their validation cohort of 91 patients with familial FSGS.52 Other studies used target gene panels to identify pathogenic or likely pathogenic mutations in patients with steroid-resistant nephrotic syndrome.7 Other diseases undergoing extensive study with NGS include congenital anomalies of the kidney and urinary tract (CAKUT), nephronophthisis and other polycystic kidney disease, IgA nephropathy, and atypical hemolytic uremic syndrome.5,53–56

Epigenomics

Epigenomics is the study of the regulatory elements that control gene expression using high throughput technology. These elements are heritable and also reversible at the cellular level. Regulatory elements control many developmental processes and modification of these elements cause disease.57,58 As these elements have diverse physical and chemical properties, many different methods have been developed to optimize the epigenome mapping resolution and overcome weaknesses inherent to the individual methods (Table 4).59–61 Commonly studied elements include histone modifications and DNA methylation. Histone modifications are important in the regulation of chromatin architecture and transcription. More than 100 different types of histone modification exist, including acetylation, methylation, and phosphorylation, all of which have been associated with transcription regulation.62–65 Chromatin immunoprecipitation sequencing (CHIP-seq) is the most common technique used to study histone modifications and common issues with this technique include antibody affinity, GC biases, and sample bias due to biased fragmentation.61,63,66 DNA methylation refers to the methylation of cytosine especially in CpG islands. DNA methylation plays a role in tumorigenesis and aging. More importantly, it is instrumental in mammalian cell differentiation and development.63,67 Researchers use multiple different methods to describe the DNA methylome.

Approach

Unlike genomics which studies DNA specifically, epigenomic techniques must describe the architecture of chromatin, the molecules that alter that architecture and how these traits differ by cell type and change with environmental exposures and/or time. Thus, investigators have developed multiple different approaches to help address these questions. These techniques suffer from the same sequencing biases as those of genomic studies plus additional ones associated with the prescreening procedures (Table 4). Researchers may utilize these methods with either observational or experimental designs. Traditional epidemiology study designs need modifications but may be used with epigenomic methods. Some specific observational study designs include birth cohorts, twin studies, and prenatal cohorts.68,69 Experimental approaches often examine changes in regulatory elements before and after toxin or drug exposure in functional assays.

Application

Researchers have identified a role for both histone modification and DNA methylation in many renal diseases.70 For instance, genome-wide studies of methylation in subjects with ADPKD reveal hypermethylation of PKD1 and other genes involved in ion transport and cell adhesion. This hypermethylation led to decreased expression of these genes and suggests a role for epigenetics in cystogenesis that may explain the patchy nature of cyst formation.71 Histone modification may also play a role in cyst formation. Knocking out histone deacetylase 5 in pkd2 −/− mice led to a decrease in cyst formation, suggesting a role for histone deacetylases in cystogenesis.72 These examples show that even in diseases with well-defined genetic pathogenicity, epigenetic mechanisms may improve understanding of disease progression and suggest avenues for new therapeutics. Other active areas of epigenetic study in nephrology include renal development, chronic kidney disease progression, and autoimmunity.26,70,73
Table 4 Next generation sequencing–based approaches and considerations for implementation

<table>
<thead>
<tr>
<th>Genomic</th>
<th>Description</th>
<th>Genomic target</th>
<th>Considerations</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted gene panels</td>
<td>DNA library preparation, followed by targeted enrichment with probes designed against only genes of interest, MPS</td>
<td>DNA</td>
<td>• Less potential for novel gene discovery &lt;br&gt;• Missing coverage in genes of interest due to gaps in probe design or biases during enrichment</td>
<td>116</td>
</tr>
<tr>
<td>WES</td>
<td>DNA library preparation, followed by targeted enrichment with probes designed for entire exome (many commercially available exome kits), MPS</td>
<td>DNA</td>
<td>• Missing coverage in genes of interest due to gaps in probe design or biases during enrichment &lt;br&gt;• Potential for incidental findings</td>
<td>39,116</td>
</tr>
<tr>
<td>WGS</td>
<td>DNA library preparation and MPS</td>
<td>DNA</td>
<td>• Less coverage than whole exome or targeted panels &lt;br&gt;• Potential for incidental findings</td>
<td>39,116</td>
</tr>
</tbody>
</table>

Epigenomic

| ChIP-seq | Cross-link proteins to DNA, followed by fragmentation, immunoprecipitation, amplification of DNA, and MPS | Protein–DNA interactions, Histone modification | • Chromatin configuration introduces sample-specific biases <br>• Amplification biases | 61,66     |
| MNase-seq | Use of micrococcal nuclease to digest DNA until it hits an obstruction like nucleosome, then size selection of fragments the size of nucleosomes, amplification of DNA and MPS | Nucleosome identification and quantification | • Cleavage biases for AT-rich sequences <br>• Size selection biases <br>• Variable TF binding characteristics affects enzymatic cleavage | 66        |
| DNase-seq | DNase I digests unprotected DNA, followed by size selection, amplification, and MPS | Maps open chromatin | • Efficiency for TF binding sites varies with fragment size <br>• Cleaves DNA in a sequence-specific manner <br>• Variable TF binding characteristics affects enzymatic cleavage | 61,66     |
| ATAC-seq | Transposase cuts DNA while simultaneously incorporating adapters into areas of open chromatin followed by PCR and MPS | Factor occupation, nucleosome position, and chromatin accessibility | • Cleavage biases based on sequence <br>• Mitochondrial DNA contamination <br>• Need for deep sequencing | 66,117    |
| FAIRE-seq | Cross-linking of protein and DNA with formaldehyde in vivo followed by sonication, phenol-chloroform extraction and MPS | Chromatin accessibility | • Low signal-to-noise ratio making computation difficult <br>• Need for very deep sequencing <br>• Fixation efficiency affects results | 61,66,118 |
| MethylCseq | DNA fragmentation followed by ligation of methylated adapters, size selection using | DNA methylation | • Bias toward GC rich areas of genome but not necessarily CpG islands | 119       |

(Continued)
Table 4 (Continued)

<table>
<thead>
<tr>
<th>Description</th>
<th>Genomic target</th>
<th>Considerations</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>electrophoresis, bisulfate conversion, amplification and MPS</td>
<td>Long-range chromatin interactions</td>
<td>• Mappability of short reads</td>
<td></td>
</tr>
</tbody>
</table>
| ChIA-PET | Cross-linking protein and DNA, followed by sonication, immunoprecipitation, ligation of linkers with MME1 endonuclease, digestion with MME1, pull down of bio- 
tinylated linkers with streptavidin beads, amplification of DNA tags, and mapping to reference genome | Long-range chromatin interactions | • Limited by available antibodies |
| | | | • Inapplicable to repetitive regions of genome |
| | | | • Identifies transcription factor binding sites but cannot identify the specific proteins in the complex |
| | | | • Multiple different analysis programs providing different results |

Transcriptomics

Transcriptomics encompasses all testing that identifies and quantifies RNA species. While microarray has traditionally been used for gene expression, RNA-seq’s use has increased dramatically over the past few years due to its low background noise, ability to identify novel transcripts, a large dynamic range, and accuracy. The development of new high throughput methods has also enabled researchers to examine RNA: protein interactions providing further information about RNA’s role in gene expression.

Approach

Common experimental designs that utilize NGS technology include RNA-seq and cross-linking immunoprecipitation sequencing (CLIP-seq) (Table 4). In RNA-seq, the sequencing process is similar to genomic sequencing but with some significant distinctions. RNA-seq’s input is total RNA. Typically, during library preparation, RNA is separated by type using oligonucleotides attached to magnetic beads. Reverse transcription converts RNA species into cDNA which is then sequenced using NGS technology. RNA-seq utilizes NGS-based techniques to identify the spectrum of RNA species, identify allelic imbalance, catalog alternative splicing patterns and posttranslational modifications, as well as quantify RNA. Study designs may be similar to genomic studies when identification of variants and transcripts are of interest. RNA-seq is also commonly implemented for gene expression studies allowing for similar epidemiological approaches and experimental designs to that of other regulatory elements.

RNA-seq has several weaknesses including transcript bias introduced by sample preparation (3’ bias from poly A tail selection), the necessity of deep sequencing to capture lower expressed genes, and possibility of base errors introduced during reverse transcription.

RNA binding proteins (RBP) regulate posttranscriptional expression. There are now several different high throughput methods for studying RBPs including UV individual-nucleotide resolution cross-linking immunoprecipitation (iCLIP), RNA co-immunoprecipitation, and photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; TF, transcription factor; WES, whole exome sequencing; WGS, whole genome sequencing.

Abbreviations: ATAC-seq, assay for transposable-accessible chromatin using sequencing; CAPTURE-seq, RNA capture sequencing; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; ChIP-seq, chromatin immunoprecipitation sequencing; CLIP-seq, cross-linking immunoprecipitation sequencing; DNase-seq, DNase I sequencing; FAIRE-seq, formaldehyde-assisted isolation of regulatory elements; ICLIP, individual-nucleotide resolution cross-linking immunoprecipitation; LC-MS, liquid chromatography-mass spectrometry; MDB-seq, methylated DNA binding domain sequencing; MeDIP-seq, methylated DNA immunoprecipitation sequencing; MNase-seq, micrococcal nuclease sequencing; MPS, massive parallel sequencing; NET-seq, native elongating transcript sequencing; PAR-CLIP, photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; TF, transcription factor; WES, whole exome sequencing; WGS, whole genome sequencing.
proteins and RNA are digested leaving only a peptide bound to the RNA. Then reverse transcription is performed and when the RNA transcriptase encounters the peptide, a mutation occurs or transcription prematurely stops. Bioinformatics pipelines then determine the sites of these mutations and where the truncated RNA aligns to identify the location of RBPs.78

**Application**

Both RNA-seq and CLIP-seq are relatively new technologies and remain underutilized in pediatric nephrology. Recently, investigators employed RNA-seq to describe the monocytome transcriptome in systemic lupus erythematosus (SLE) with associated glomerulonephritis. The monocytome was chosen for this study because of its important role in the pathogenicity of SLE. In this case–control design, investigators discovered that several developmental genes were underexpressed in SLE, suggesting that monocytome in SLE are more differentiated compared with controls. Total transcription of protein coding regions and antisense RNA were decreased in SLE subjects. While most of the noncoding RNAs were decreased in SLE, pri-miRNAs which prevent transcription were upregulated corresponding to decreased expression observed in protein coding genes.79 Other transcriptomic studies have revealed a potential role for miRNA in lupus and have even begun to examine the urine transcriptome in hopes of identifying potential biomarkers for disease.80 Preliminary studies of RNA-seq in ADPKD, acute kidney injury, and CAKUT have demonstrated promise for elucidation of disease pathogenicity.81–83 One can easily imagine multiple uses for these methods in the study of inherited renal disease.

**Mass Spectrometry–Based Approaches**

For the high throughput study of proteomics and metabolomics, MS has become the leading experimental method. It has relatively high sensitivity and specificity. MS uses mass to charge ratio (m/z) to identify substances. Typically a compound is dissolved in a solvent and ionized leading to its fracture into charged fragments. These fragments are then separated using a combination of electric and magnetic fields and then detected as a spectrum based on their m/z. The spectra are compared with other spectra in databases to identify the compounds. Additionally, MS provides information about the relative abundance of the compounds. Different ionization methods and analyzers exist, each with their own strengths and weaknesses (Table 5).59,84 Notably, analyzers are often used in tandem to allow improved separation and identification of material.84

**Proteomics**

Proteomics relies heavily on high throughput methods such as MS to catalog all proteins in a cell or biological system. Proteomic experiments are difficult due to the issues mapping spectra and peptides to unique proteins, wide dynamic range in abundance of proteins within the cell, and posttranslational modifications.85 It is estimated that only about two-thirds of the proteins present in humans have been detected using MS.59

**Approach**

The most common MS method utilized in proteomics is a bottom-up approach using trypsin digestion and liquid chromatography (LC-MS/MS).84 Studies may aim to simply identify proteins in a sample or more commonly investigators are searching for quantitative differences between samples. In quantitative experiments, one must consider both the biological variation and technical variation given MS run variability. Two main approaches exist: one aims to characterize the relative or absolute concentration of proteins and the other aims to compare the changes in average abundance between conditions. Downstream analysis of this data including class discovery and class prediction often occurs. Class discovery attempts to identify functional protein classes or biological samples with similar protein profiles. Often systems biology approaches are applied to accomplish this. Class prediction is the use of protein abundance profile to predict disease status and is the first step in establishing biomarkers.85

**Application**

To date, most studies of proteomics in pediatric nephrology have focused on biomarker discovery. For instance, patients with posterior urethral valves (PUV) are now often diagnosed prenatally, but clinicians struggle with the extent of kidney damage that the neonate will manifest. Klein et al86 examined the power of urinary proteins to predict renal disease progression in PUV, first taking an untargeted approach to identify urine proteins that were differentially expressed in patients with severe kidney disease/end-stage renal disease and those without end-stage renal disease at 2 years. Then they validated 26 peptides in a separate cohort, limited these tests further, and tested this algorithm in an independent cohort.86 Follow-up studies have utilized this early work to start predicting outcomes for fetuses with PUV with the ultimate goal of using this information to drive fetal intervention.87 These studies demonstrate some of the guiding principles put forward in clinical proteomic applications.27 Other nephrologic areas of interest for proteomics include acute kidney injury, CAKUT, allograft rejection, and renal Fanconi syndrome.88

**Metabolomics**

Metabolomics is the high throughput study of small molecules such as peptides, amino, nucleic, and organic acids in a biological system. MS and nuclear magnetic resonance (NMR)-based techniques provide high throughput, sensitive, and specific information about the metabolome. These experiments often examine potential biomarkers using blood, urine, and serum. This review focuses on MS-based metabolomics, but thorough reviews of NMR may be found elsewhere.89 The metabolome is a heterogeneous mixture of molecules with different physical and chemical properties preventing any one MS analytic method from becoming the gold standard. MS-based metabolomic experiments have many similar limitations and biases as are found in proteomics. These include
appropriate sample preparation, high analytic variability, and destruction of material during experimentation.\textsuperscript{90}

**Approach**

Most studies attempt to identify metabolites and quantify their abundance. Often researchers use this information to detect difference between two conditional states using classic epidemiological approaches and may also employ experimental techniques especially in the area of pharmacology. Targeted approaches are commonly implemented in the study of drugs and other clinical biomarkers. This type of approach may utilize knowledge of metabolic pathways to study changes in biological networks. Targeted approaches allow for increased sensitivity and specificity of metabolite identification and improved quantification of products. Untargeted approaches are usually hypothesis generating and require further studies for validation.\textsuperscript{91} Multiple in vitro study approaches attempt to identify protein–metabolite interactions.\textsuperscript{59}

**Application**

Metabolomics especially on urine samples promises to provide biomarkers and metabolic profiles that help identify pathologic processes and potentially aid in prognosis in a noninvasive manner. For example, researchers recently examined the urine metabolome of pediatric renal transplant patients who underwent renal biopsy either as part of their surveillance or due to increased creatinine. The results revealed that a specific metabolic profile could be used to separate patients into three categories: acute T-cell rejection, borderline rejection, and no rejection. Baseline kidney function, inflammation, and pyuria did not affect the reliability of this metabolic profile.\textsuperscript{92} This study demonstrates the promise of metabolic profiling in diagnosing renal disease states in a noninvasive manner. Other recent studies have examined the metabolome as a possible source of biomarkers in acute kidney injury and IgA nephropathy.\textsuperscript{93}

---

**Table 5** Overview of mass spectrometry methods

<table>
<thead>
<tr>
<th>Ionization methods</th>
<th>Description</th>
<th>Strengths</th>
<th>Weaknesses</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESI</td>
<td>High voltage applied to liquid, creating aerosol</td>
<td>• Leaves the molecule intact</td>
<td>• Prone to ion suppression</td>
<td>84,121</td>
</tr>
<tr>
<td>MALDI</td>
<td>MALDI matrix and sample co-crystallize and then matrix absorbs energy and sample desorbs into gas</td>
<td>• Rapid sample prep • More tolerant of salts and detergents</td>
<td>• Sensitivity and peak resolution suffer from sub-optimal crystal formation • Not used for absolute quantification</td>
<td>84</td>
</tr>
</tbody>
</table>

| Analyzers | | | | |
| Triple quadrupole | Tandem quadrupole MS using time varied electric current using DC and RF voltage | • Good reproducibility • Low cost | • Limited resolution | 84 |
| Orbitrap | Uses electrostatic force balanced by centrifugal force to trap ions | • Low cost • High space charge capacity | • Slow scan speed • Less mass resolution than FT-IRC | 122 |
| Time of flight | M/Z determined by TOF through vacuum tube | • Detects largest molecules • Fast • Compatible with MS imaging | • -Less mass resolution than others | 84 |
| FT-IRC | Homogeneous application of magnetic field inducing cyclotron motion determined by m/z | • Highest mass resolving power • Highest mass accuracy | • Slow • Not used in quantification | 84 |

Abbreviations: ATAC-seq, assay for transposase-accessible chromatin using sequencing; CAPTURE-seq, RNA capture sequencing; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; ChIP-seq, chromatin immunoprecipitation sequencing; CLIP-seq, cross-linking immunoprecipitation sequencing; DNase-seq, DNase I sequencing; FAIRE-seq, formaldehyde-assisted isolation of regulatory elements; iCLIP, individual-nucleotide resolution cross-linking immunoprecipitation; LC-MS, liquid chromatography-mass spectrometry; MeDIP-seq, methylated DNA binding domain sequencing; Methyl-seq, methylated DNA immunoprecipitation sequencing; MNase-seq, micrococcal nuclease sequencing; MPS, massive parallel sequencing; NET-seq, native elongating transcript sequencing; PAR-CLIP, photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; TF, transcription factor; WES, whole exome sequencing; WGS, whole genome sequencing.
Integration of Data

The development of high throughput omics methods has shifted the burden of genetic disease from data generation to data analysis. Each field has made important discoveries in isolation, but the richness of these data lies in its integration to more fully depict biological systems. While NGS technologies allow exploration of the genomic footprint, regulation, and transcription, the translation of this information into cellular function in health and disease requires exploration of proteins and metabolites. This incorporation of multiple different high throughput methods to describe biological systems currently underpins many principles of systems biology and has stimulated bioinformatics tools and statistical strategies to connect data from these large datasets.94

Investigators have developed multiple methods in attempts to integrate this omics information. There are two basic approaches to integration: unsupervised and supervised methods. Unsupervised methods ask one basic question. Is there a pattern in this data? It is a hypothesis generating technique. These processes tend to incorporate more of the raw data. These tests do not necessarily contribute knowledge to biological processes but instead provides fodder for a supervised project. Supervised integration starts from a hypothesis and often requires further experimentation. Often supervised approaches limit the dimensionality of the data. There are multiple ways to carryout either a supervised or unsupervised project. These include clustering, feature selection, prediction analysis, text mining, and pathway analysis.59 While these methods attempt to address integration of omics data into a functional biologic picture, more methodological work in both systems biology and omics is required to build sophisticated hierarchal models of genetic disease.94

Functional Assays

To date, many of the systems biology approaches have utilized functional models to obtain the appropriate datasets for integration. Functional models are also used to validate disease hypotheses generated through these techniques. In nephrology, in vitro methods using cell line, ex vivo methods with murine kidney explants, and in vivo methods using model organisms exist for the study of genomic processes (Table 6). Cell lines exist for almost all cells of the glomerulus and renal tubules, and cell-based methods such as cellular reporter assays including three-dimensional cell culture or tubulogenesis assays work well for renal diseases that primarily involve only one cell type (e.g., nephronophthisis or ADPKD).4,95,96 Ex vivo manipulation of murine renal explants or in vivo methods such as conditional gene targeting holds promise for studying more complex genetic diseases such as CANG.4 Animal models commonly used with gene targeting methods include mice and zebrafish.97–99 Recent advances in the use of induced pluripotent stem cells directly from affected patients’ fibroblasts promise to provide a relatively easy genetic model system for studying the effect of specific alleles on renal differentiation and development but have yet to be available for use in inherited renal disease.4 Researchers may also choose to apply omics methods to these model systems, especially if the relevant human tissue is inaccessible or when trying to study only one particular cell type within an organ system. Without more functional studies into genes and pathways identified using omics data, the ability to accurately predict pathogenic genetic mutations and describe the resultant molecular rearrangements identified in omics data will be incomplete.

**Table 6** Functional assays for examining molecular pathways in renal disease

<table>
<thead>
<tr>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-interfering RNA assays</td>
<td>Knockdown of candidate genes in transfected cells using siRNAs that are homologous to the genes of interest</td>
</tr>
<tr>
<td>3D renal spheroid</td>
<td>Knockdown of candidate genes using siRNA in murine collecting duct cell lines looking for change in shape of cilia which should be spherical</td>
</tr>
<tr>
<td>Morpholinos</td>
<td>Antisense oligonucleotides complimentary to mRNA of interest bind mRNA preventing protein synthesis</td>
</tr>
<tr>
<td>Inducible Cre-Lox–based animal models</td>
<td>Site-specific recombinase technology that introduces deletions, insertions, translocations, and inversions into DNA. Targets specific cell types or can be triggered by specific external stimulus</td>
</tr>
<tr>
<td>CRISPR/Cas9 systems</td>
<td>May target and conditionally knockdown any gene in a mouse using guide RNA and Cas9 endonuclease to cut DNA at predetermined site of interest with the introduction of small indels or different base pairs during the repair</td>
</tr>
</tbody>
</table>
methodological development in the area of data processing and statistical analysis will likely improve interpretation of these experimental results. In order for translation of this research into clinical application, more coordinated efforts are needed in the field. Datasets that include not only genetic information but also protein and metabolite information will allow for more integrated data analysis further illuminating pathologic processes. Researchers need to focus on development of high throughput functional models to keep up with the demand that these omics experiments generate. Standardized methods of data sharing and more widely shared datasets will provide greater power to support genetic findings. Finally, greater education and awareness of these methodologies including their limitations among pediatric nephrologists will lead to well-designed research projects in a larger number of pediatric renal disease.

References


Douglas AP, Vance DR, Kenny EM, Morris DW, Maxwell AP, McKnight AJ. Next-generation sequencing of the mitochondrial genome and association with IgA nephropathy in a renal transplant population. Sci Rep 2014;4:7379


84 Lin SY, Hsu WH, Lin CC, Chen CJ. Mass spectrometry-based proteomics in Chest Medicine, Gerontology, and Nephrology: subgroups omics for personalized medicine. Biomedicine (Taipei) 2014;4:25

85 Käll I, Vitke O. Computational mass spectrometry-based proteomics. PLOS Comput Biol 2011;7(12):e1002277


97 Conesa A, Mortazavi A. The common ground of genomics and systems biology. BMC Syst Biol 2014;8(Suppl 2):S1


103 Blake JA, Bult CJ, Epipig JT, Kadin JA, Richardson JE; Mouse Genome Database Group. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res 2014;42(Database issue):D810–D817


115 Hatem A, Bozdağ D, Toland AE. Çatalyurek ÜV. Benchmarking short sequence mapping tools. BMC Bioinformatics 2013; 14:184


121 Annesley TM. Ion suppression in mass spectrometry. Clin Chem 2003;49(7):1041–1044


This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.