Genes Caught In Flagranti: Integrating Renal Transcriptional Profiles With Genotypes and Phenotypes

Yuanfang Guan, PhD,[*,†,‡] Sebastian Martini, MD,[†,§] and Laura H. Mariani, MD, MS[†]

Summary: In the past decade, population genetics has gained tremendous success in identifying genetic variations that are statistically relevant to renal diseases and kidney function. However, it is challenging to interpret the functional relevance of the genetic variations found by population genetics studies. In this review, we discuss studies that integrate multiple levels of data, especially transcriptome profiles and phenotype data, to assign functional roles of genetic variations involved in kidney function. Furthermore, we introduce state-of-the-art machine learning algorithms, Bayesian networks, support vector machines, and Gaussian process regression, which have been applied successfully to integrating genetic, regulatory, and clinical information to predict clinical outcomes. These methods are likely to be deployed successfully in the nephrology field in the near future.

Semin Nephrol 35:237-244

Keywords: Gene regulation, SNP, clinical outcomes, predictions

In the past decade, population genetics has gained tremendous success in identifying single-nucleotide polymorphisms (SNPs) that are correlated to the clinical outcomes of renal diseases. For example, Kopp et al[1,2] showed that SNPs in a locus on chromosome 22 are correlated to the susceptibility of African Americans to focal segmental glomerulosclerosis. Later, it was identified that apolipoprotein L1 (APOL1) in this region is associated with focal segmental glomerulosclerosis in African Americans.3 In addition, a significant number of studies have been published to identify SNPs related to diabetic nephropathy. For example, Pezzolesi et al4 identified a total of 13 SNPs that are associated with type 1 diabetes–related diabetic nephropathy with a P value less than 1 × 10-5. The strongest associated gene, FERM domain containing 3 (FRMD3), was found to be expressed in human kidney.[4]

Although genome-wide association studies have been used widely to understand the genetic basis of complex diseases, the follow-up functional studies of the relevant genes are not standardized. Typically, genome-wide association studies conclude by presenting a list of SNPs and their associated genes, leaving the functional analysis for future work. It has become clear that functional characterization of SNPs is fundamental for interpreting the genetic mechanism of diseases. A particular challenge in this regard is SNPs situated in non–protein coding regions of the genome that may impact regulatory function in a manner that is evident in only a certain functional context. One such context may be a biological signaling cascade or pathway determined by genes whose transcription is synchronized by common regulatory elements within their promoters.

Previous studies have used elegant methods such as luciferase reporter gene assays and electrophoretic mobility shift assays to identify those alleles that alter the promoter activity of cis-genes. Identifying promoter activity–modifying alleles is usually the first step toward the identification of the underlying mechanisms that can be followed by bioinformatics analyses that allow for the identification of potential transcription factors that may be affected by a particular SNP. Several bioinformatics tools, such as transcription factor binding site (TFBS) _SEARCH[5] and MATCH[6] (developed by the TRANSFAC team) can be used to scan the promoter region for potential binding sites, and then the SNP location can be correlated to the TFBS. Super-gel shift assays then can be used to verify these interactions if antibodies specific for that particular transcription factor are available. Additional studies could use immunoprecipitation plus massively parallel DNA sequencing (ChIP-Seq) to test whether these transcription factors indeed are involved in the formation of transcriptional complexes at a certain SNP site.

Complexes of several transcription factors often work in concert, in so-called *promoter modules*, linked to regulatory patterns or pathways involved in developmental, physiological, and pathophysiological responses. Their binding results in an activation or inhibition of target gene expression. These functions often are executed via differentially regulated gene products. These

[*]*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.*
[†]*Department of Internal Medicine, University of Michigan, Ann Arbor, MI.*
[‡]*Department of Computer Science and Engineering, University of Michigan, Ann Arbor, MI.*
[§]*Nephrologisches Zentrum, Medizinische Klinik und Poliklinik IV, Klinikum der Universität München, Ludwig-Maximilians-University Munich, Munich, Germany.*

Financial disclosure and conflict of interest statements: none.
Address reprint requests to Yuanfang Guan, PhD, 100 Washtenaw Ave., 2017 Palmer Commons Building, Ann Arbor, MI 48109-1600. E-mail: gyuanfan@umich.edu

0270-9295/ - see front matter

http://dx.doi.org/10.1016/j.semnephrol.2015.04.003

gene products are regulated at the level of transcription initiation by transcription factors that physically and functionally interact with each other and with regulatory sequences within the DNA.

Defining the consequences of regulatory variants on gene expression in complex diseases is still in its infancy also because individual TFBS often are not sufficient for regulatory functions. Their contributions to transcriptional regulation can be assessed only in the appropriate regulatory context (ie, the regulatory relationships may change across different tissues and diseases). Bioinformatics tools and techniques involving disease-relevant pathways,[7,8] transcriptional covariance, protein–protein networks,[9–12] and phylogenetic conservation[13] have helped to select genes belonging to a certain functional context. With genetic mapping of expression quantitative trait loci studies becoming available for complex renal diseases, these expression quantitative trait loci will be linked directly to the physical location of transcripts differentially expressed in kidney diseases and support promoter modeling approaches as described in the following example.

BIOINFORMATICS TOOLS HELP UNCOVER THE FUNCTIONAL CONTEXT OF A DIABETIC NEPHROPATHY–ASSOCIATED SNP LOCATED IN THE PROMOTER REGION OF THE GENE FRMD3

Previous genome-wide association studies (GWAS) have reported that rs1888747 was associated significantly with diabetic nephropathy.[4] This SNP is located at the noncoding region near FRMD3 and was found to be bound by transcription factors.[14] By using tubulointerstitial gene expression data from kidney biopsies also from patients with diabetic nephropathy, 581 messenger RNAs that co-express with FRMD3 were identified (Fig. 1).[14] These genes are strongly enriched in the bone morphogenetic protein signaling pathway. In parallel, in silico comparison of sequence variants with and without the risk allele identified a potential homeodomain factor TFBS covering the SNP position. As confirmed by electrophoretic mobility shift assays, this homeodomain factor binding site was absent in the presence of the non–risk allele in the FRMD3 promoter. A set of 4 transcription factors including the homeodomain factor defined by a certain order and distance of each TF, the promoter module, was identified using in silico bioinformatics tools. A genome-wide search then showed that the promoter framework was enriched among bone morphogenetic protein genes as well as the FRMD3 promoter sequence with the risk allele. This led to the hypothesis that the DN risk allele rs1888747 brings FRMD3 under the control of a proposed transcriptional regulatory module and inhibits renal expression of FRMD3. These findings not only detect a transcriptional regulatory pattern affected by the candidate SNP but also connect known DN-associated pathways to the GWAS-derived candidate gene, providing further insight into the pathophysiology of DN that ultimately could lead to individual risk assessments and selection of targeted therapies.

THE CHRONIC KIDNEY DISEASE PATHWAY NETWORK: CROSS-TALK AMONG MULTIPLE MOLECULAR MECHANISMS

Defining the pathophysiology of chronic kidney disease, which affects more than 20 million individuals in the United States, is critical to identifying predictors of the disease course and potential therapeutic targets. Although several mechanisms have been connected with the development of chronic kidney disease (CKD), the CKDGen and CHARGE consortia were able to use GWAS to identify genetic risk factors for renal function decline. Expanding on the concept outlined previously of a single molecular pathway affected by a SNP driving the disease process, the following example provide insight into one possible systems genetics perspective on CKD, in which multiple data types are integrated to identify a hierarchy relating candidate genes to co-expressed transcripts as functional relationships. This concept is in line with comprehensive studies in model organisms that show genes and pathways in dense interrelationships, with multiple genes mapping onto multiple pathways that likely contribute or affect CKD. Genotypic, transcriptomic, and clinical data are linked by performing a pathway–cross-talk analysis of the gene sets linked to glomerular filtration rate (GFR).

The 40 candidate genes identified by the meta-analysis by the CKDGen and CHARGE consortia were located in proximity (±60 kb) of 16 SNPs strongly associated with renal function decline. The majority of these transcripts (29 in total) were found to be expressed in renal gene expression profiles of 157 subjects with one of the nine different chronic renal diseases (focal and segmental glomerulosclerosis, membranous glomerulonephritis, minimal change disease, diabetic nephropathy, hypertensive nephropathy, IgA nephritis, lupus nephritis, thin-membrane disease, or were from histologically unaffected parts of tumor nephrectomies).

These overlapping genes from the genetic (GWAS) study and the transcriptomic study described earlier were examined further for co-expression patterns[15] (Fig. 2). Thereby, the 18 genes were used as seeds to retrieve additional genes correlated with them, resulting in co-expressed gene sets. The biological signaling cascades or pathways enriched among the co-expressed gene sets were identified, and thus linking each gene