Kidney Medicine

Type IV Collagen Variants in CKD: Performance of Computational Predictions for Identifying Pathogenic Variants

Cole Shulman, Emerald Liang, Misato Kamura, Khalil Udwan, Tony Yao, Daniel Cattran, Heather Reich, Michelle Hladunewich, York Pei, Judy Savige, Andrew D. Paterson, Mary Ann Suico, Hirofumi Kai, and Moumita Barua

Rationale & Objective: Pathogenic variants in type IV collagen have been reported to account for a significant proportion of chronic kidney disease. Accordingly, genetic testing is increasingly used to diagnose kidney diseases, but testing also may reveal rare missense variants that are of uncertain clinical significance. To aid in interpretation, computational prediction (called in silico) programs may be used to predict whether a variant is clinically important. We evaluate the performance of in silico programs for *COL4A3/A4/A5* variants.

Study Design, Setting, & Participants: Rare missense variants in *COL4A3/A4/A5* were identified in disease cohorts, including a local focal segmental glomerulosclerosis (FSGS) cohort and publicly available disease databases, in which they are categorized as pathogenic or benign based on clinical criteria.

Tests Compared & Outcomes: All rare missense variants identified in the 4 disease cohorts were subjected to in silico predictions using 12 different programs. Comparisons between the predictions were compared with: (1) variant classification (pathogenic or benign) in the cohorts and (2) functional characterization in a randomly selected smaller number (17) of pathogenic or uncertain significance variants obtained from the local FSGS cohort.

Results: In silico predictions correctly classified 75% to 97% of pathogenic and 57% to 100% of benign *COL4A3/A4/A5* variants in public disease databases. The congruency of in silico predictions was similar for variants categorized as pathogenic and benign, with the exception of benign *COL4A5* variants, in which disease effects were overestimated. By contrast, in silico predictions and functional characterization classified all 9 pathogenic *COL4A3/A4/A5* variants correctly that were obtained from a local FSGS cohort. However, these programs also overestimated the effects of genomic variants of uncertain significance when compared with functional characterization. Each of the 12 in silico programs used yielded similar results.

Limitations: Overestimation of in silico program sensitivity given that they may have been used in the categorization of variants labeled as pathogenic in disease repositories.

Conclusions: Our results suggest that in silico predictions are sensitive but not specific to assign *COL4A3/A4/A5* variant pathogenicity, with misclassification of benign variants and variants of uncertain significance. Thus, we do not recommend in silico programs but instead recommend pursuing more objective levels of evidence suggested by medical genetics guidelines.

Complete author and article information provided before references.

Correspondence to M. Barua (moumita.barua@uhn.ca)

Kidney Med. 3(2):257-266. Published online February 10, 2021.

doi: 10.1016/j.xkme.2020.12.007

Chronic kidney disease (CKD) represents a heterogeneous group of disorders that result in irreversible fibrosis over time. Current diagnostic methods often fail to distinguish molecular mechanisms or predict disease course. CKD affects more than 750 million people globally and results in more than 1 million deaths annually. As such, kidney disease is a major health burden with substantive costs.[1,2]

Genomics is emerging as one tool to identify mechanistically relevant CKD subtypes. Using whole-exome sequencing, we have recently reported that pathogenic variants in the *COL4A3/A4/A5* genes are the leading single gene causes (~5%) of focal and segmental glomerulosclerosis (FSGS), a histopathologic entity representing diverse causes.[3] Similarly, pathogenic variants in the *COL4A3/A4/A5* genes have also been reported to account for a significant proportion of CKD.[4] Pathogenic variants in type IV collagen are well known to cause Alport syndrome.[5-10]

The human genome has tremendous sequence variation and the effect of rare nonsynonymous single-nucleotide variants (SNVs) in a disease-associated gene can be unclear. The American College of Medical Genetics (ACMG) has standards based on expert consensus for declaring the pathogenicity of rare variants that are organized into supporting, moderate, strong, and very strong levels of evidence.[11] Some of these criteria include assessment of frequency in population data, type of variant change (eg, null variant), identification of familial cosegregation, presence in clinically ascertained mutation databases, bioinformatics, and functional data.[12] Well-established functional studies that show a deleterious effect are considered strong levels of evidence.[13]

Computational predictions, also known as in silico programs, are one part of clinical variant classification in the diagnostic setting but are considered supportive compared with stronger lines of evidence. These programs

PLAIN-LANGUAGE SUMMARY

Type IV collagen mutations have been reported to account for a significant proportion of chronic kidney disease. As a result, genetic testing is increasingly being used for diagnosis but can uncover DNA changes that are of uncertain clinical significance. To determine whether causative for disease (called pathogenic), DNA changes can be tested with cell and animal models, an approach that is limited by the absence of well-established models for most genes, expense, and time-consuming nature. Alternatively, computational programs can be used to make predictions for pathogenicity. In this report, we begin to define the test characteristics for these computational predictions using bioinformatic and experimental approaches, with results suggesting that programs tend to overestimate the effects of DNA changes.

have been developed to predict the functional effects of rare missense variants. Broadly, the algorithms use different types of variant information, including sequence conservation, protein structure analysis, and meta prediction (using results from multiple programs) for predictions.[14-18]

The predictive performance of in silico programs has been evaluated with computational methods against data sets that contain pathogenic and benign variants obtained from public resources (eg, Universal Protein Resource [Uniprot]), literature, and curated disease databases in which variants in kidney disease genes are not highly represented.[19-26] We evaluate the predictive performance of in silico programs for *COL4A3/A4/A5* missense variants by first comparing with clinically categorized variants deposited in 3 public disease databases and a local FSGS cohort. As a second approach, in silico predictions are compared with functionally characterized missense variants identified in the local FSGS cohort.

METHODS

Whole-Exome Sequencing Analysis

Details on how patients were recruited and exome data analyzed have been previously described.[3] Study participants gave their written informed consent and the study protocol was approved by the Toronto General Hospital's committee on human research (98-UO13). Whole-exome sequencing and data processing were performed by The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada. Exomic capture was achieved with Agilent SureSelect Human All Exon V5. Reads were mapped to the hg19 reference sequence.

Variant Calling From FSGS Whole-Exome Sequencing Data

Variants were identified using GATK (version 4.0.5.1).[27] Gene-based annotation features of ANNOVAR were applied (access date, April 16, 2018).[28] The frequency of variants was determined using Genome Aggregation Database (gnomAD; version 2.1.1; access date, March 18, 2019).[26,29-31] Variants in *COL4A3/A4/A5* were categorized as rare if having a minor allele frequency ≤ 0.005 in the ethnically matched population within gnomAD. This cutoff was selected in consideration of the low prevalence of FSGS, estimated at 7 per million for the general population, 20 per million for Africans, and 5 per million for Europeans.[32,33] It was also selected in consideration of inheritance patterns: *COL4A3/A4/A5* is associated with autosomal recessive, dominant, or X-linked recessive disease. Rare missense variants in *COL4A3/A4/A5* were designated as pathogenic if reported in other cases of kidney disease after searching the literature and disease databases ClinVar, ARUP, and LOVD.[3,34]

In Silico Predictions Programs

Rare *COL4A3/A4/A5* missense variants from our FSGS whole-exome sequencing data and disease databases ClinVar, ARUP, and LOVD (accessed October 22, 2019, September 13, 2019, and August 28, 2019, respectively) were identified. All rare SNVs reported in these sources have already been categorized. We in turn submitted the missense variants to 12 in silico programs for predictions (Table S1).[15,24,35-39,40-45,46] A variant was categorized as pathogenic if the majority, selected as 10 or more of 12 programs, categorized the variant as pathogenic using the program's recommended scoring cutoffs.

COL4A Split Luciferase Assay

From our FSGS cohort with whole-exome sequencing data, 9 pathogenic variants and 8 variants of uncertain significance in *COL4A3* and *COL4A5* were randomly selected. We defined pathogenic variants as rare (minor allele frequency < 0.005) and reported in other cases with kidney disease, whereas variants of uncertain significance were defined as any other rare missense variant. To assess heterotrimer formation ability of these missense variants, we used the split complementation Nano-luciferase assay system that we have previously developed.[47] Tagged plasmid constructs of *COL4A4*-FLAG, wild-type or mutant *COL4A3*-SmBiT, and *COL4A5*-LgBiT were generated as described previously.[47] Corresponding SmBiT and LgBiT tags were attached at either the N-terminal or C-terminal of *COL4A3* and *COL4A5*. After mutagenesis, sequences were verified. The *COL4A3/A4/A5* tagged constructs were subsequently cotransfected into human embryonic kidney 293 (HEK293T) cells. Twenty-four hours posttransfection, cells were replated in LumiNunc 96-well white plates (Thermo Fisher Scientific) and cultured in phenol red-free Dulbecco's Modified Eagle Medium (DMEM) containing 10%

fetal bovine serum, 100 U of penicillin and streptomycin, 2 mmol/L of glutamine, and 200 μmol/L of L-ascorbic acid 2-phosphate trisodium salt. After 24 hours, cells (intracellular heterotrimer) and media (secreted heterotrimer) were assayed using Nano-Glo Live Cell Assay reagent and GloMax Navigator system (Promega).

RESULTS

Computational Validation

We identified 70 SNVs in *COL4A3* (29), *COL4A4* (26), and *COL4A5* (15) across 186 adults with FSGS with whole-exome sequencing. Of these, 31 were rare (minor allele frequency < 0.005), of which 30 were missense (14 in *COL4A3*, 10 in *COL4A4*, and 6 in *COL4A5*) and 1 was a stop-gain in *COL4A4*. Characteristics of the sequenced cohort have been published previously.[3]

In parallel, 2,803 nonsynonymous *COL4A3/A4/A5* variants were identified in 125,748 unscreened participants with whole-exome data in gnomAD, a public database of genomic variation. Of these, 2,307 were rare and 2,279 were missense variants.

Rare missense variants in our local FSGS cohort, gnomAD, and Alport databases were each interrogated using 12 in silico programs for predictions (Table S1; Item S1). In the FSGS cohort, for rare missense variants in *COL4A3*, *COL4A4*, and *COL4A5*, 43% (6/14), 40% (4/10), and 33% (2/6) were predicted to be deleterious by at least 10 of 12 programs, respectively (Fig 1; Item S1).

By comparison, for rare missense variants in *COL4A3*, *COL4A4*, and *COL4A5* identified in gnomAD, 35% (301/851), 32% (306/949), and 41% (197/483) were predicted to be deleterious by at least 10 of 12 programs, respectively (Fig 1). gnomAD is a database in which some rare diseases would be even less represented than population estimates given that severe pediatric cases are not included.[48] However, the lack of clinical data to correlate rare variants in gnomAD controls does not enable us to draw conclusions as to the accuracy of these predictions.

We also accessed disease databases in which *COL4A3*, *COL4A4*, and *COL4A5* variants would be deposited, which included ARUP, ClinVar, and LOVD (accessed October 22, 2019, September 13, 2019, and August 28, 2019, respectively; Fig 2). ARUP documented 346 SNVs in *COL4A5*. Three hundred twenty-seven were categorized as pathogenic, with 97% (317/327) concordance to in silico predictions. ARUP does not document SNVs in *COL4A3* or *COL4A4*. In ClinVar, 120 *COL4A3* SNVs were documented. Sixteen were categorized as pathogenic and 75% were assigned (12/16) correctly by in silico programs. For *COL4A4*, 55 SNVs were reported. Nine were classified as pathogenic, with 100% assigned correctly by in silico programs. For *COL4A5*, there were 367 SNVs. Two hundred eighty-seven were categorized as pathogenic, with 90% (258/287) concordance to in silico predictions. In LOVD, 412 *COL4A3* SNVs were catalogued. Of these, 34 were pathogenic and 82% (28/34) were predicted accurately. For *COL4A4*, there were 306 SNVs, of which 49 were classified as pathogenic, with 86% (42/49) concordance to in silico predictions. For *COL4A5*, there were 987 SNVs. Six hundred ninety-nine were classified as pathogenic, with 94% (650/699) concordant in silico predictions. In silico program sensitivity could be overestimated given that they may have been used in the categorization of variants labeled as pathogenic in these disease databases.

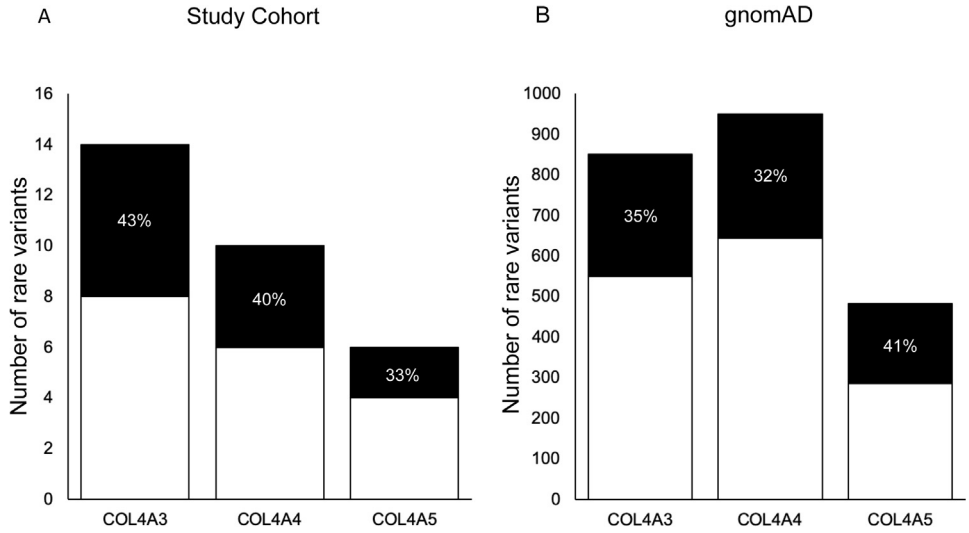

Figure 1. Number of rare missense variants predicted to be pathogenic in: (A) the focal segmental glomerulosclerosis (FSGS) study cohort and (B) Genome Aggregation Database (gnomAD). For rare missense *COL4A3*, *COL4A4*, and *COL4A5* variants in our FSGS cohort, 43% (6/14), 40% (4/10), and 33% (2/6) were predicted to be deleterious by at least 10 of 12 programs, respectively. For rare missense *COL4A3*, *COL4A4*, and *COL4A5* variants identified in gnomAD, 35% (301/851), 32% (306/949), and 41% (197/483) were predicted to be deleterious by at least 10 of 12 programs, respectively.

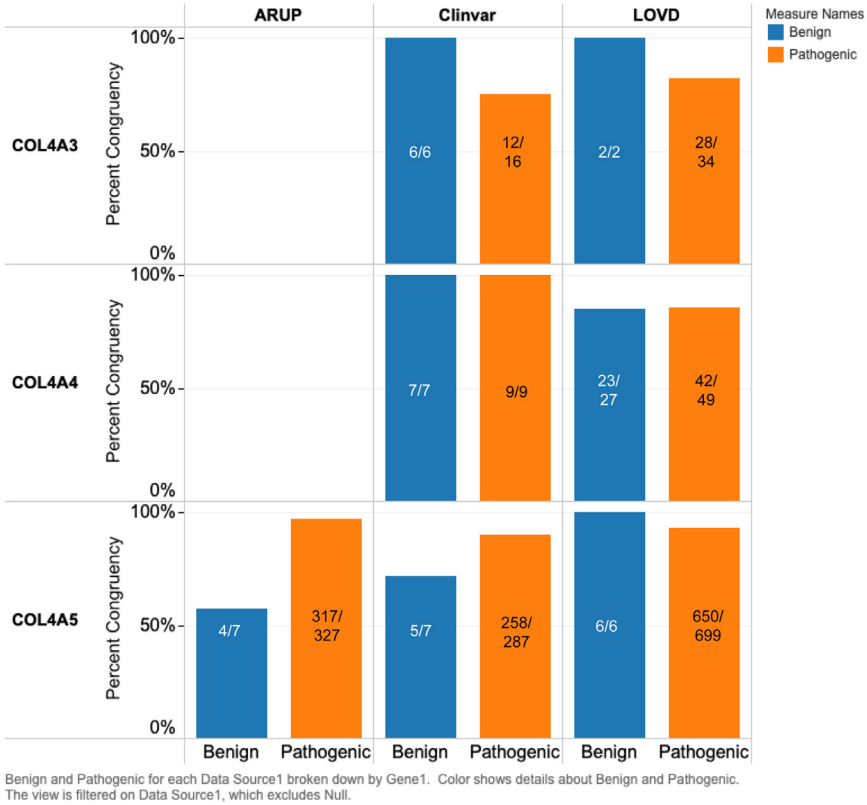

Figure 2. Comparison of *COL4A3*, *COL4A4*, and *COL4A5* in silico predictions with disease database categorization. For ARUP *COL4A5* pathogenic variants, there was 97% (317/327) concordance with in silico predictions. For ClinVar *COL4A3/A4/A5* pathogenic variants, there was 75% (12/16), 100% (9/9), and 89% (258/287) concordance with in silico predictions, respectively. For LOVD *COL4A3/A4/A5* pathogenic variants, there was 82% (28/34), 86% (42/49), and 94% (650/699) concordance. Congruency of in silico predictions was similar for variants categorized as benign, with the exception of *COL4A5* variants documented in ARUP and ClinVar, in which the effects were overestimated by in silico programs, though there were fewer variants to interrogate. In ARUP, 57% (4/7) of *COL4A5* variants were classified correctly by in silico predictions. In ClinVar, 100% (6/6), 100% (9/9), and 71% (5/7) of *COL4A3/A4/A5* variants, respectively, were correctly assigned. Finally, for LOVD, 100% (2/2), 85% (23/37), and 100% (6/6) of *COL4A3/A4/A5* variants were correctly classified.

The congruency of in silico predictions was similar for variants categorized as benign, with the exception of *COL4A5* variants documented in ARUP and ClinVar, in which the effects were overestimated by in silico programs, though there were fewer variants to interrogate (Fig 2). In ARUP, 7 *COL4A5* variants were classified as benign, 57% (4/7) of which were assigned as such by in silico predictions. In ClinVar, 6 *COL4A3*, 7 *COL4A4*, and 7 *COL4A5* variants were classified as benign, with 100%, 100%, and 71% (5/7) concordance with predictions, respectively. In LOVD, 2 *COL4A3* and 6 *COL4A5* variants were classified as benign, with 100% concordance for both. For *COL4A4*, there were 27 benign variants, with 85% (23/27) concordance with predictions.

A report suggests that one in silico classifier called M-CAP outperforms popular scores such as SIFT, PolyPhen-2, and CADD in its ability to separate pathogenic from benign variants.[45] As a result, analysis of variants from disease databases was performed using M-CAP only. None of the benign variants in the disease databases were correctly classified, either as a result of incorrect categorization as pathogenic or by not generating an output (Fig S1). The accuracy for classification of pathogenic variants was much better, ranging from 89% to 96%. Additionally, a receiver operating curve for each of the 12 in silico programs was generated using the disease database type IV collagen variants and their in silico scores (Fig S2). When each curve is examined, we find that the score cutoff that maximizes the true-positive rate while minimizing the false-positive rate does not coincide with the in silico programs' recommendations. For instance, we find that the cutoff for SIFT should be approximately less than 0.004, whereas the recommended cutoff is less than 0.05 (Fig S2). As a result, variants with scores between 0.004 and 0.05 are being predicted as pathogenic, leading to false positives.

Congruency in classification between in silico programs was also explored (Fig 3). Most programs had similar prediction scores when comparing with each other except for FATHMM and M-CAP.

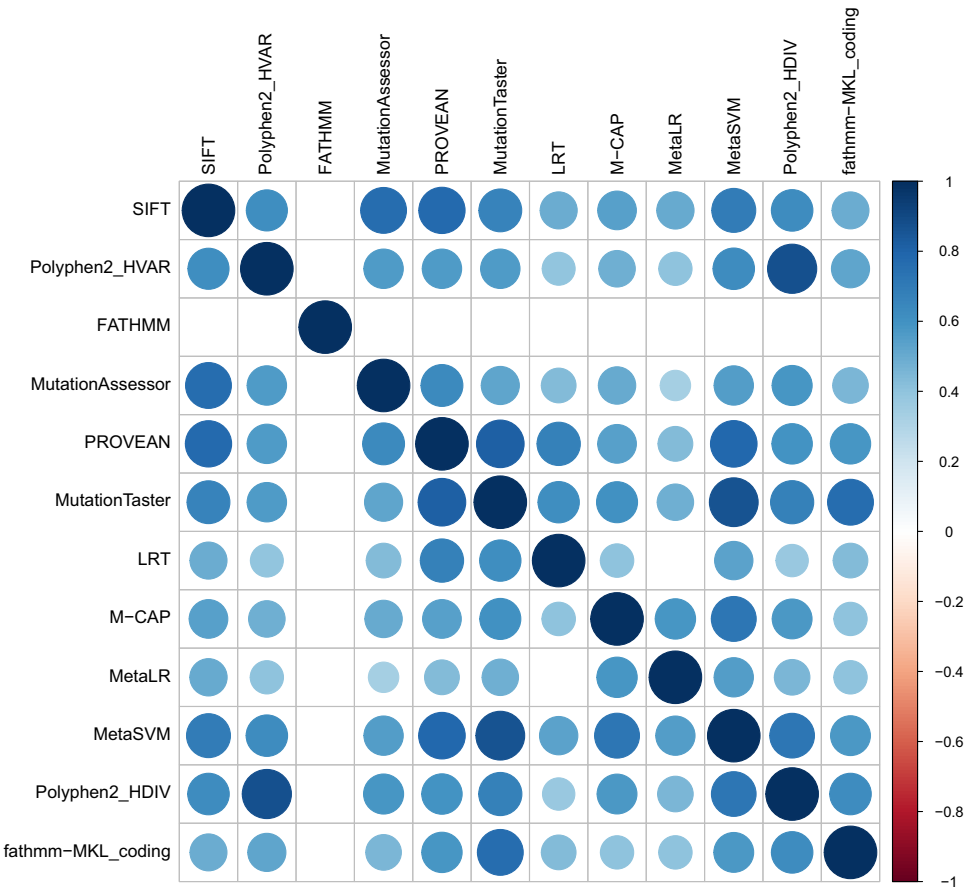

Figure 3. Spearman correlation coefficient heatmap comparing results of various prediction models. Most programs had similar prediction scores when comparing with each other except for FATHMM and M-CAP. Dark blue signifies a strong direct correlation while dark red signifies a strong indirect correlation. Squares that are lighter in color signify a weak correlation between the results of the 2 prediction models. Figure created using the corrplot package available in Rstudio.

Functional Validation

We evaluated 9 pathogenic missense variants in *COL4A3* and the X-linked *COL4A5* identified in the local FSGS cohort. For *COL4A3* and *COL4A4*, the mode of inheritance has traditionally been reported as recessive, but next-generation sequencing studies have reported about 20% to 30% of patients with dominant disease.[49-52] Of 3 pathogenic heterozygous missense variants in *COL4A3* (ie, rare variant reported in other individuals with kidney disease), all were predicted to be deleterious by at least 10 of 12 in silico programs (Table 1; Item S2). Of 6 pathogenic variants in *COL4A5*, all were predicted to be deleterious by at least 10 of 12 in silico programs (Table 1). Under normal conditions, *COL4A3*, *COL4A4*, and *COL4A5* each encodes a protein that heterotrimerizes and is secreted into the glomerular basement membrane. To determine the secretory behavior of the *COL4A3* and *COL4A5* mutants, we used an assay system that quantified the intracellular trimerization and trimer secretion of *COL4A3/4/5*.[47] Using this split luciferase complementation assay, all *COL4A3* and *COL4A5* pathogenic variants were found to have secretory defect with the N-terminal tagged versions of *COL4A3* and *COL4A5* (Fig 4A and C). The pathogenic variants could form trimers intracellularly but could not be efficiently secreted. By contrast, this was not always observed for C-terminal tagged versions of *COL4A3* and *COL4A5* (Fig 4B and D). We speculate that this could be due to heterotrimer formation being initiated at the noncollagenous (NC1) domain of the C-terminal region of collagen and terminates at the N-terminal region. The fusion of the monomers initially at the C-terminal region brings the reporter tags closer together to produce luminesce regardless of whether the trimer is completely formed.

Similarly, 8 variants of uncertain significance in *COL4A3* and *COL4A5* identified in our FSGS cohort were selected, comparing in silico predictions with functional characterization. Of 4 variants of uncertain significance in *COL4A3*, 3 were predicted to be deleterious by at least 10 of 12 in silico programs (Table 1). Of 4 variants of uncertain significance in *COL4A5*, 2 were predicted to be deleterious by at least 10 of 12 in silico programs (Table 1). Using the split luciferase complementation assay, only 1 variant of uncertain significance in *COL4A5*

Table 1. Comparison of Functional Annotation With In Silico Predictions for Pathogenic *COL4A3* and *COL4A5* Variants Identified in the FSGS Cohort

Gene	No. of Pathogenic Variants	No. Predicted Deleterious by 10/12 Programs	No. of Variants With Secretory Defect	Congruence
COL4A3	3	3	3	100%
COL4A5	6	6	6	100%

Note: All pathogenic *COL4A3* and *COL4A5* variants were categorized as such as a result of being identified in other kidney disease cases. All pathogenic variants demonstrated a secretory defect with functional characterization and were correctly assigned by in silico predictions.
Abbreviation: FSGS, focal segmental glomerulosclerosis.

was found to have a secretory defect using N-terminal tagged proteins, though not to the degree observed for the definitely pathogenic variants (Fig 4C). Any data point under the −50 line was considered as a significant secretory defect. Thus, there was poor concordance between in silico predictions and functional characterization, with the former potentially overestimating the functional characteristics of missense variants (Table 2).

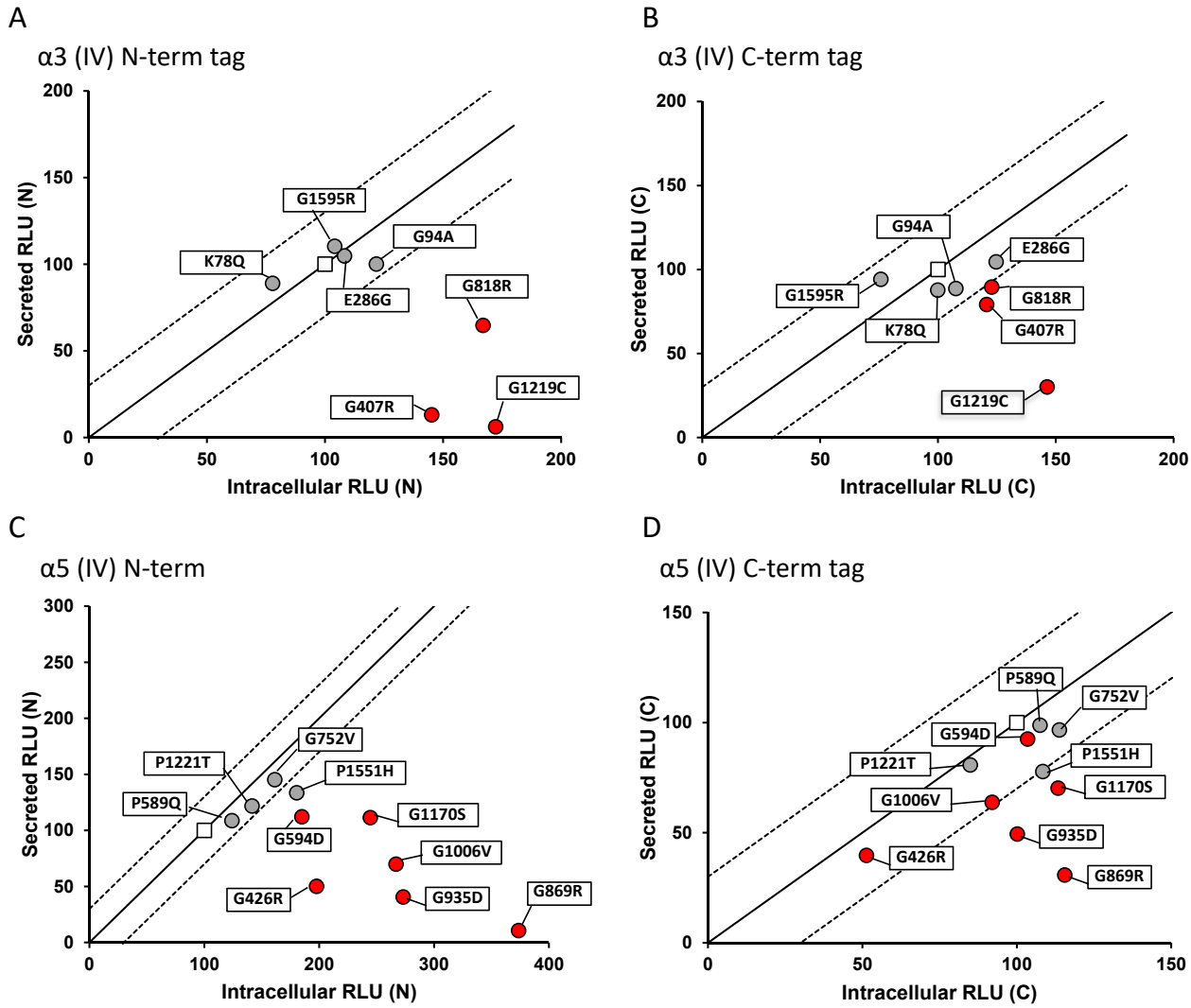

Figure 4. Functional characterization of *COL4A3* and *COL4A5* using the split-luciferase assay. Scatterplots of the intracellular/secreted relative light unit (RLU) ratio from human embryonic kidney 293 (HEK293T) cells expressing (A, B) mutant *α3* chain or (C, D) mutant *α5* chain compared to wild type (WT) using N-terminal and C-terminal split-luciferase tagged constructs. Pathogenic *α3* and *α5* chain mutants showed clearer secretory defect with N-terminal tagged constructs. Solid line: Y = X, dotted line: Y = X + 50, Y = X − 50. Square, WT; red circles are pathogenic variants and grey circles are variants of uncertain significance. Any data point under the −50 line was considered as a significant secretory defect. Experiments were performed in triplicate. Data presented are representative of 2 independent experiments.

Table 2. Comparison of Functional Annotation With In Silico Predictions for Variants of Uncertain Significance in *COL4A3* and *COL4A5* Identified in the FSGS Cohort

Gene	No. of Variants of Uncertain Significance	No. Predicted Deleterious by 10/12 Programs	No. of Variants With Secretory Defect	Congruence
COL4A3	4	3	0	0%
COL4A5	4	2	1	50%

Note: Only 1 variant of uncertain significance in *COL4A5* was found to have a secretory defect, which was accurately predicted by in silico predictions. However, 1 variant of uncertain significance without evidence of a secretory defect was also predicted to be deleterious.
Abbreviation: FSGS, focal segmental glomerulosclerosis.

To further determine the functional nature of 8 variants of uncertain significance in *COL4A3* and *COL4A5*, clinical characteristics for these patients were obtained (Tables S2 and S3). Many individuals with Alport syndrome have hematuria and basement membrane abnormalities. In our cohort, microscopic hematuria data were reported for 9 patients with pathogenic variants and 8 patients with a variant of uncertain significance in *COL4A3* and *COL4A5*. Microscopic hematuria was observed in 4 of 9 patients with pathogenic variants and 2 of 8 patients with a variant of uncertain significance (Tables S2 and S3). For the 2 patients with variants of uncertain significance (*COL4A3* p.G1595R and *COL4A5* p.P589Q) and hematuria, neither variant was characterized as functionally deleterious (Fig 4).

DISCUSSION

Our results demonstrate that in silico predictions correctly classified most pathogenic *COL4A3/A4/A5* variants catalogued in ClinVar, ARUP, and LOVD. In silico predictions performed similarly for benign variants with the exception of *COL4A5* (concordance in ARUP and ClinVar with predictions and classification was 57% [4/7] and 71% [5/7], respectively) but there were also far fewer benign variants to interrogate in these disease databases. Our second approach of correlating in silico predictions with functional testing showed that both accurately classified all pathogenic *COL4A3/A4/A5* missense variants in the FSGS cohort. These variants were labeled as pathogenic because they are rare and already reported as disease-causing in other individuals with kidney disease, which are considered strong lines of evidence (ACMG criteria PS1 and PS4; Item S2).[11] By contrast, in silico predictions overestimated the effects of *COL4A3/A5* variants of uncertain significance when compared with functional characterization. A variant of uncertain significance was defined as a rare variant that did not satisfy ACMG criteria for definite pathogenicity. Interestingly, interrogation of *COL4A3/A4/A5* variants found in gnomAD predicted a high percentage to be deleterious, but the lack of clinical data for correlation prevents us from making any conclusion with these data.

Genomics facilitates clinically meaningful classification of CKD but sequencing can reveal rare SNVs for which the relationship to disease is unclear. The ACMG has standards based on expert consensus for declaring pathogenicity wherein in silico predictions are considered only supporting compared with higher levels of evidence that are deemed moderate, strong, or very strong.[11] Well-established functional studies that show deleterious effect is an example of one criterion considered strong level of evidence. Against this background, we provide an assessment of in silico programs using both computational and experimental approaches.

Using the Nano-luciferase complementation system, we have recently quantified trimerization of 9 typical glycine substitutions in *COL4A5* that differ in disease progression, finding a correlation between in vitro results and phenotype.[53] In the data presented here, we observe that the pattern of heterotrimer formation and secretion for pathogenic mutants differed slightly between N-terminal and C-terminal tagged constructs. The N-terminal tagged pathogenic mutants showed clearer secretory defect. We postulate that this could be as a result of heterotrimer formation initiating at the NC1 domain at the C-terminal region of collagen and terminates at the N-terminal region. The fusion of the monomers initially at the C-terminal region could bring the reporter tags closer together to produce luminesce regardless of whether the heterotrimer is completely formed. Therefore, the luciferase reporter attached at the N-terminal region, that is, the N-tagged constructs, may better reflect the state of trimer folding.

The 12 prediction models used in this study can be categorized as solely conservation based: (SIFT Polyphen2-HVAR, Polyphen2-HDIV, MutationAssessor, PROVEAN, and LRT) and multifeatured algorithms (FATHMM, M-CAP, MetaLR, MetaSVM, FATHMM-MKL, MutationTaster; Table S1). Conservation-based models select homologous sequences to create multiple sequence alignments across species (MSA) and use the sequence and predicted structure-based features of the MSA to predict pathogenicity with variants in more conserved areas predicted to be deleterious. The multifeatured algorithms integrate other information, such as epigenomic signals (FATHMM-MKL and MutationTaster), allele frequencies (FATHMM, MetaLR, and MetaSVM), or the results of other prediction algorithms (M-CAP, MetaLR, and MetaSVM). Eleven of the 12 prediction models are trained using databases including UniProt[54] (PolyPhen2-HDIV, PolyPhen2-HVAR, FATHMM, PROVEAN, MetaLR, and MetaSVM), Human Gene Mutation Database[55] (FATHMM, FATHMM-MKL, MutationTaster, and M-CAP), ExAC[56] (M-CAP), Ensembl[57] (LRT), 1000 Genomes Project[58] (MutationTaster and FATHMM-MKL), and COSMIC[59] (MutationAssessor). SIFT

was trained using known variants of the *E coli* LacI gene that have been individually mutated and functionally tested.[60,61]

M-CAP has been previously reported to outperform popular pathogenicity classifiers but our results demonstrate that it unreliably categorizes the small number of benign type IV collagen variants in disease databases by incorrectly assigning pathogenicity or not generating an output. M-CAP already uses 9 established pathogenicity likelihood scores included in our scoring system: SIFT13, PolyPhen-2, CADD15, MutationTaster20, MutationAssessor21, FATHMM22, LRT23, MetaLR16, and MetaSVM16.[45] It incorporates 7 established measures of base pair, amino acid, genomic region, and gene conservation: RVIS24, PhyloP25, PhastCons26, PAM250, BLOSUM62, SIPHY28, and GERP29. Additionally, M-CAP introduces 298 new features derived from multiple-sequence alignment of 99 primate, mammalian, and vertebrate genomes to the human genome30. However, previous reports seeking to demonstrate superiority of one classifier over others are all limited by the veracity of variant assignment in test databases and in which kidney gene variants contribute a small proportion.

Our study highlights several limitations and opportunities for future investigation. Estimating in silico program accuracy using disease databases relies on robust categorization and underscores a need for consistency in variant annotation. The sensitivity of in silico programs could be overestimated given that they may have been used in the categorization of variants labeled as pathogenic. In disease databases, there were far fewer variants classified as benign compared with pathogenic. However, to address these limitations, we have pursued more laborious functional characterization on randomly selected type IV collagen variants from the FSGS cohort as an additional line of evidence.

With respect to functional characterization, we include data to support our conclusions, but only a small number of missense variants were characterized. We use the arbitrary cutoff of ±50 from wild-type data, but characterizing more pathogenic and benign variants would better define a threshold. As per standard convention throughout the literature, we characterize the effects of single variants on the reference haplotype, but there are several common haplotypes documented in the 1000 Genomes Project (Table S4). Studying the effects of single variants on different haplotype backgrounds could provide important information regarding interaction effect between haplotype and mutation. Second, our assay will identify mutations that are associated with secretory defects, but this is a simplification of disease pathogenesis that does not account for the complexities involving extracellular type IV collagen network formation. For instance, a previous report suggests that ~20% of *COL4A5* mutations have detectable heterotrimers in the glomerular basement membrane, suggesting alternate disease mechanisms.[62]

Recent reports demonstrate that pathogenic variants in *COL4A3/A4/A5* account for a significant and unappreciated proportion of patients with Alport syndrome in CKD.[3,5-10,63] Sequencing is increasingly being used to obtain mechanistically relevant diagnoses but often generates rare missense variants that remain of uncertain clinical significance. In silico predictions have been developed to aid in categorizing variants. We show here that computational approaches including M-CAP, which was reported to outperform other classifiers, are sensitive but not sufficiently specific to confidently assign *COL4A3/A4/A5* variant pathogenicity. Thus, we do not recommend any in silico program in the consideration of type IV collagen variant categorization, but instead pursuing more objective levels of evidence suggested by medical genetic guidelines.

SUPPLEMENTARY MATERIAL

Supplementary File 1 (PDF)

Figure S1: Comparison of COL4A3, COL4A4, and COL4A5 in silico predictions by M-CAP with disease database categorization.

Figure S2: Receiver operator curves for the 12 in silico programs using scores generated from type IV collagen variants obtained from disease databases.

Table S1: Twelve In Silico Predictions Programs Used.

Table S2: Clinical Characteristics of FSGS Patients With COL4A3 Variants.

Table S3: Clinical Characteristics of FSGS Patients With COL4A5 Variants.

Table S4: Common COL4A3 Haplotypes Identified in Europeans in the 1000 Genomes Project.

Supplementary File 2 (XLS)

Item S1: COL4A3/A4/A5 Variants Identified in FSGS Cohort and Predictions by In Silico Programs.

Supplementary File 3 (XLS)

Item S2: Variants Selected for Functional Testing.

ARTICLE INFORMATION

Authors' Full Names and Academic Degrees: Cole Shulman, Emerald Liang, BSc, Misato Kamura, BSc, Khalil Udwan, MD, PhD, Tony Yao, BSc, Daniel Cattran, MD, Heather Reich, MD, PhD, Michelle Hladunewich, MD, MSc, York Pei, MD, MSc, Judy Savige, MB, PhD, Andrew D. Paterson, MD, PhD, Mary Ann Suico, PhD, Hirofumi Kai, PhD, and Moumita Barua, MD.

Authors' Affiliations: Division of Nephrology, University Health Network (CS, EL, KU, TY, DC, HR, MH, YP, MB); Toronto General Hospital Research Institute, Toronto General Hospital, Toronto, Canada (CS, EL, KU, TY, DC, HR, MH, YP, MB); Department of Molecular Medicine, Graduate School of Pharmaceutical Science, Kumamoto University, Kumamoto, Japan (MK, MAS, HK); Institute of Medical Sciences, Toronto, Canada (DC, HR, MH, YP, MB), Department of Medicine, Toronto, Canada (DC, HR, MH, YP, MB); University of Melbourne, Melbourne, Australia (JS); Division of Epidemiology and Biostatistics, Dalla Lana School of Public Health (ADP); and Genetics and Genome Biology, Research Institute at Hospital for Sick Children, Toronto, Canada (ADP).

Address for Correspondence: Moumita Barua, MD, 200 Elizabeth Street, 8NU-855, Toronto General Hospital, Toronto, ON, Canada M5G 2C4. E-mail: moumita.barua@uhn.ca

Authors' Contributions: Research idea and study design: CS, EL MK, ADP, HK, MB; patient ascertainment and DNA archiving: KU,

DC, HR, MH, YP; acquired LOVD variant data: JS; computational analysis: CS, EL MB, TY; prepared figures: CS, EL, MB; collection and analysis of functional data: MK, MS; supervision/mentorship: HK, MB (equal contribution). CS, EL, and MK contributed equally to this work. Each author contributed important intellectual content during manuscript drafting or revision and accepts accountability for the overall work by ensuring that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved.

Support: Dr Barua received a NephCure Kidney International-Neptune Ancillary Studies Grant in 2016, Health Research Grant (14-04) from Physician's Services Inc in 2015, a McLaughlin Accelerator Award in 2019, and support from the Can-SOLVE CKD Network (https://www.cansolveckd.ca/) and the Toronto General Hospital Foundation.

Financial Disclosure: The authors declare that they have no relevant financial interests.

Acknowledgements: We thank the study participants.

Peer Review: Received September 24, 2020. Evaluated by 2 external peer reviewers, with direct editorial input by the Editor-in-Chief. Accepted in revised form December 13, 2020.

REFERENCES

1. Klarenbach SW, Tonelli M, Chui B, Manns BJ. Economic evaluation of dialysis therapies. *Nat Rev Nephrol*. 2014;10(11):644-652.
2. Canadian Institute for Health Information. *National Health Expenditure Trends, 1975 to 2015*. Ottawa, ON: Canadian Institute for Health Information; 2015.
3. Yao T, Udwan K, John R, et al. Integration of genetic testing and pathology for the diagnosis of adults with FSGS. *Clin J Am Soc Nephrol*. 2019;14:213-223.
4. Groopman EE, Marasa M, Cameron-Christie S, et al. Diagnostic utility of exome sequencing for kidney disease. *N Engl J Med*. 2019;380(2):142-151.
5. Kashtan CE. Alport syndrome. An inherited disorder of renal, ocular, and cochlear basement membranes. *Medicine (Baltimore)*. 1999;78(5):338-360.
6. Voskarides K, Pierides A, Deltas C. COL4A3/COL4A4 mutations link familial hematuria and focal segmental glomerulosclerosis. Glomerular epithelium destruction via basement membrane thinning? *Connect Tissue Res*. 2008;49(3):283-288.
7. Papazachariou L, Papagregoriou G, Hadjipanagi D, et al. Frequent COL4 mutations in familial microhematuria accompanied by later-onset Alport nephropathy due to focal segmental glomerulosclerosis. *Clin Genet*. 2017;92(5):517-527.
8. Xie J, Wu X, Ren H, et al. COL4A3 mutations cause focal segmental glomerulosclerosis. *J Mol Cell Biol*. 2014;6(6):498-505.
9. Malone AF, Phelan PJ, Hall G, et al. Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. *Kidney Int*. 2014;86(6):1253-1259.
10. Pierides A, Voskarides K, Athanasiou Y, et al. Clinico-pathological correlations in 127 patients in 11 large pedigrees, segregating one of three heterozygous mutations in the COL4A3/ COL4A4 genes associated with familial haematuria and significant late progression to proteinuria and chronic kidney disease from focal segmental glomerulosclerosis. *Nephrol Dial Transplant*. 2009;24(9):2721-2729.
11. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.
12. 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RN, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
13. Starita LM, Ahituv N, Dunham MJ, et al. Variant interpretation: functional assays to the rescue. *Am J Hum Genet*. 2017;101(3):315-325.
14. Christenhusz GM, Devriendt K, Dierickx K. Disclosing incidental findings in genetics contexts: a review of the empirical ethical research. *Eur J Med Genet*. 2013;56:529-540.
15. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. 2011;39(17):e118.
16. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*. 2006;7:61-80.
17. Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum Mutat*. 2009;30(5):703-714.
18. Cooper GM, Goode DL, Ng SB, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods*. 2010;7(4):250-251.
19. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*. 2011;32(4):358-368.
20. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*. 2011;39(database issue):D214-D219.
21. Wu CH, Apweiler R, Bairoch A, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*. 2006;34(database issue):D187-D191.
22. Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. *Hum Mutat*. 2013;34(1):42-49.
23. Morrison AC, Voorman A, Johnson AD, et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet*. 2013;45(8):899-901.
24. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24(8):2125-2137.
25. Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M. Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity. *Genomics*. 2019;111(4):869-882.
26. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat*. 2016;37(3):235-241.
27. Van der Auwera GA, Carneiro MO, Hartl C, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protocols Bioinforms*. 2013;43. 11.10.1-11.10-33.
28. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
29. Karczewski KJ, Francioli LC, Tiao G. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *Nature*. 2019. https://doi.org/10.1038/s41586-020-2308-7.

30. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46:D1062-D1067.
31. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat*. 2011;32:894-899.
32. Kitiyakara C, Kopp JB, Eggers P. Trends in the epidemiology of focal segmental glomerulosclerosis. *Semin Nephrol*. 2003;23(2):172-182.
33. Kitiyakara C, Eggers P, Kopp JB. Twenty-one-year trend in ESRD due to focal segmental glomerulosclerosis in the United States. *Am J Kidney Dis*. 2004;44(5):815-825.
34. Crockett DK, Pont-Kingdon G, Gedge F, Sumner K, Seamons R, Lyon E. The Alport syndrome COL4A5 variant database. *Hum Mutat*. 2010;31(8):E1652-E1657.
35. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31:3812-3814.
36. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248-249.
37. Ramensky V. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 2002;30:3894-3900.
38. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*. 2013;34:57-65.
39. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015;31:1536-1543.
40. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol*. 2007;8:R232.
41. Schwarz JM, Cooper DN, Schuelke M, Seelow D. Mutationtaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014;11:361-362.
42. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31:2745-2747.
43. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. 2012;7:e46688.
44. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res*. 2009;19:1553-1561.
45. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. 2016;48:1581-1586.
46. Kim S, Jhong JH, Lee J, Koo JY. Meta-analytic support vector machine for integrating multiple omics data. *BioData Mining*. 2017;10:2.
47. Omachi K, Kamura M, Teramoto K, et al. A split-luciferase-based trimer formation assay as a high-throughput screening platform for therapeutics in Alport syndrome. *Cell Chem Biol*. 2018;25(5):634-643 e634.
48. Song W, Gardner SA, Hovhannisyan H, et al. Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genet Med*. 2016;18(8):850-854.
49. Fallerini C, Dosa L, Tita R, et al. Unbiased next generation sequencing analysis confirms the existence of autosomal dominant Alport syndrome in a relevant fraction of cases. *Clin Genet*. 2014;86(3):252-257.
50. Moriniere V, Dahan K, Hilbert P, et al. Improving mutation screening in familial hematuric nephropathies through next generation sequencing. *J Am Soc Nephrol*. 2014;25(12):2740-2751.
51. van der Loop FT, Heidet L, Timmer ED, et al. Autosomal dominant Alport syndrome caused by a COL4A3 splice site mutation. *Kidney Int*. 2000;58(5):1870-1875.
52. Pescucci C, Mari F, Longo I, et al. Autosomal-dominant Alport syndrome: natural history of a disease due to COL4A3 or COL4A4 gene. *Kidney Int*. 2004;65(5):1598-1603.
53. Kamura M, Yamamura T, Omachiet K, et al. Trimerization and genotype–phenotype correlation of COL4A5 Mutants in Alport SYNDROME. *Kidney Int Rep*. 2020;5(5):718-726.
54. Bateman A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47:D506-D515.
55. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD®): 2003 update. *Hum Mutat*. 2003;21:577-581.
56. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285-291.
57. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46:D754-D761.
58. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
59. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;47:D941-D947.
60. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*. 2001;11(5):863-874.
61. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. *J Mol Biol*. 1994;240(5):421-433.
62. Kashtan CE, Kleppel MM, Gubler MC. Immunohistologic findings in Alport syndrome. *Contrib Nephrol*. 1996;117:142-153.
63. Groopman E, Goldstein D, Gharavi A. Diagnostic utility of exome sequencing for kidney disease. Reply. *N Engl J Med*. 2019;380(21):2080-2081.