Using Population Genetics to Interrogate the Monogenic Nephrotic Syndrome Diagnosis in a Case Cohort

Matthew G. Sampson,* Christopher E. Gillies,* Catherine C. Robertson,* Brendan Crawford,* Virginia Vega-Warner,* Edgar A. Otto,* Matthias Kretzler,†‡ and Hyun Min Kang§

Departments of *Pediatrics and Communicable Diseases, and †Internal Medicine, Division of Nephrology and ‡Department of Computational Medicine and Bioinformatics, University of Michigan School of Medicine, Ann Arbor, Michigan; and §Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan

ABSTRACT

To maximize clinical benefits of genetic screening of patients with nephrotic syndrome (NS) to diagnose monogenic causes, reliably distinguishing NS-causing variants from the background of rare, noncausal variants prevalent in all genomes is vital. To determine the prevalence of monogenic NS in a North American case cohort while accounting for background prevalence of genetic variation, we sequenced 21 implicated monogenic NS genes in 312 participants from the Nephrotic Syndrome Study Network and 61 putative controls from the 1000 Genomes Project (1000G). These analyses were extended to available sequence data from approximately 2500 subjects from the 1000G. A typical pathogenicity filter identified causal variants for NS in 4.2% of patients and 5.8% of subjects from the 1000G. We devised a more stringent pathogenicity filtering strategy, reducing background prevalence of causal variants to 1.5%. When applying this stringent filter to patients, prevalence of monogenic NS was 2.9%; of these patients, 67% were pediatric, and 44% had FSGS on biopsy. The rate of complete remission did not associate with monogenic classification. Thus, we identified factors contributing to inaccurate monogenic classification of NS and developed a more accurate variant filtering strategy. The prevalence and clinical correlates of monogenic NS in this sporadically affected cohort differ substantially from those reported for patients referred for genetic analysis. Particularly in unselected, population–based cases, considering putative causal variants in known NS genes from a probabilistic rather than a deterministic perspective may be more precise. We also introduce GeneVetter, a web tool for monogenic assessment of rare disease.

More than 25 genes have been discovered in which fully penetrant mutations cause monogenic forms of nephrotic syndrome (NS).1 Published reports suggest that monogenic causes are responsible for a substantial portion of NS, and these patients are resistant to immunosuppressive treatment.2–6 These prevalence estimates and clinical correlates can influence decisions to screen patients for monogenic NS as well as the subsequent management. Given this and the increasing availability of sequencing technology, there is growing interest in expanding the use of genetic screening to diagnose monogenic NS.

However, concurrently, results of population–based genetic studies outside of NS are challenging beliefs about the accuracy of pathogenicity prediction of rare variants found in known disease genes.

Received May 8, 2015. Accepted September 6, 2015.
Published online ahead of print. Publication date available at www.jasn.org.
Correspondence: Dr. Matthew G. Sampson, Division of Pediatric Nephrology, University of Michigan School of Medicine, Ann Arbor, MI 48109. Email: mgsamps@med.umich.edu
Copyright © 2016 by the American Society of Nephrology.
as well as their clinical consequences. Exome–wide sequencing studies in thousands of population subjects have identified a substantial prevalence of predicted harmful, protein–altering genetic variants in each human genome. Sequencing studies of genes implicated in monogenic disease in the population and/or healthy controls have shown that the prevalence of putatively causal mutations is orders of magnitude greater than the reported prevalence of the disease.

Thus, three critical concerns for the broad clinical implementation of genetic sequencing in NS are (1) distinguishing bystander or incompletely penetrant variants from causal mutations, (2) setting appropriate expectations for the chance that a patient will be diagnosed with a monogenic form of NS, and (3) determining the clinical consequences for those harboring a causal variant in a known NS gene.

To address these concerns, 21 genes previously identified as harboring causal mutations for steroid resistant nephrotic syndrome (SRNS) (monogenic NS genes) were sequenced in putative controls from the 1000 Genomes Project (1000G) and 312 North American patients with NS recruited from the prospective, longitudinal Nephrotic Syndrome Study Network (NEPTUNE). The prevalence of predicted causal NS mutations in the population was analyzed under a number of filtering parameters. These results were used to devise a more accurate filtering strategy that reduced false attribution of pathogenicity. This filter was subsequently applied to patients to determine the prevalence of monogenic NS in the NEPTUNE cohort and its clinical correlates.

### RESULTS

In total, 312 participants from NEPTUNE (Supplemental Table 1, Table 1) and 61 ancestry-matched subjects from the 1000G underwent targeted sequencing of 21 genes implicated in SRNS (Table 2). A default filtering pipeline was created, similar to previously reported variant–filtering pipelines, to classify variants found in these genes as causal (Table 3). Using the “default filter”, 13 of 312 patients with NS (4.2%) were classified with monogenic NS; 75% (9 of 13) classified as such had putative mutations in NS genes with autosomal dominant transmission. Analysis of the same genes under the same filter in 61 ancestry–matched, sequenced subjects from the 1000G did not identify any putative mutations.

Given the rarity of these types of variants, analyzing >61 subjects would be necessary to accurately determine their population prevalence. Thus, the same analysis was then extended to the approximately 2500 subjects from Phase 3 of the 1000G, with a 5.8% prevalence of subjects with putatively causal variants in monogenic NS genes; 70% of these subjects had putative causal variants in NS genes with dominant transmission. Hereafter, the prevalence of variants implicated as causal in subjects from the Phase 3 1000G is referred to as the background prevalence for a given filtering pipeline.

A 5.8% background prevalence is >550 times higher than the population prevalence of NS (<0.01%), even without considering only monogenic forms of this condition. Assuming that subjects in the 1000G are not enriched for monogenic NS, this implies that a substantial fraction of the population is carrying rare variants in implicated NS genes that are classified as mutations by existing filtering strategies. A proportion of this 5.8% background prevalence is likely caused by false attribution of pathogenicity to rare but harmless variants. It is also likely that there is a greater amount of incomplete penetrance for the NS phenotype than is currently appreciated.

The discovery of substantial background prevalence under the default filter could have important research and clinical implications in terms of misclassification of pathogenicity and recognition of incomplete penetrance. Thus, with an aim to identify a more stringent filtering strategy that reduced background prevalence without inflating false negative rates, conventional and novel filtering criteria were analyzed against population-level data of known monogenic NS genes.

### The Discordance between Functional Prediction Databases

The use of functional prediction programs in filtering pipelines was examined first. Requiring two of three programs to predict a variant as damaging is currently an accepted approach to pathogenicity filtering of missense variants. The merits of this choice were investigated.

PolyPhen2, SIFT, and MutationTaster pathogenicity predictions of 1220 missense variants found in the SRNS genes

<table>
<thead>
<tr>
<th>Table 2. Genes implicated in SRNS undergoing sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inheritance Mode</strong></td>
</tr>
<tr>
<td>Dominant</td>
</tr>
<tr>
<td>Recessive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Value or Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age, yr</td>
<td>34</td>
</tr>
<tr>
<td>Age range, yr</td>
<td>2–84</td>
</tr>
<tr>
<td>Histologic diagnosis</td>
<td>FSGS (30%), minimal change (22%), membranous (15%), IgA nephropathy (10%), others (23%); (Supplemental Table 1)</td>
</tr>
<tr>
<td>Fraction of pediatric patients</td>
<td>30%</td>
</tr>
<tr>
<td>Continental ancestry</td>
<td>European (48%), African (29%), admixed American (15%), East Asian (9%)</td>
</tr>
</tbody>
</table>
in the subjects in the 1000G were computed. When requiring one, two, or three of these three programs to classify a variant as damaging, 12.6%, 5.8%, or 1.5% background prevalence, respectively, in 1000G was observed (Figure 1). When applying this approach to 312 patients with NS, the prevalence of monogenic diagnosis was 8.7%, 4.2%, or 2.2%, respectively. As expected, the most stringent threshold was most effective in reducing background prevalence but also likely increased false negatives. This was caused by both missing scores for particular variants and discordance between classification methods. Altogether, requiring two of three was found to be a compromise that substantially reduces background prevalence while allowing enough variability in the prediction algorithms to prevent excessive false negative predictions.

**Population Stratification**

Applying an allele frequency (AF) threshold using Exome Variant Server (EVS) data (http://evs.gs.washington.edu/EVS/) is frequently used to exclude noncausal variants. However, because the EVS population is only composed of European Americans and blacks, EVS-based AF thresholds will not robustly remove variants that, although rare in those of African or European ancestries, are common in other ancestries (e.g., East Asian). In addition, some variants predicted to be rare in the 1000G continental populations may, in fact, be common when considering local populations (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). To address this, an AF of <0.5% (dominant gene); <1% (recessive gene) was applied to the background prevalence in the population decreased from 5.8% (n=147) to 4.5% (n=113) and 3.7% (n=95), respectively.

**Incorporating Prior Gene– and Transcript–Specific Knowledge**

Negative selection is the genome-wide process of depletion of variants that are harmful and reduce an organism’s ability to reproduce. The degree of negative selection acting on a gene can differ by transcript. Variants found in transcripts under the most negative selection would be expected to be the most deleterious, whereas healthy individuals may be expected to have variants in less selected transcripts. It was hypothesized that background prevalence could be reduced by considering causal only variants occurring in the transcript of a gene under the highest degree of negative selection. An existing metric was used (TIMS score) to identify the transcript in each gene under the most negative selection.

Among 21 genes, only INF2 and LAMB2 had a transcript under substantially greater negative selection. When considering INF2’s longest transcript (Figure 2) and holding all other filtering criteria constant, the odds of a putative pathogenic variant being present in a patients versus a control are 1.3 (5 of 312 patients; 32 of 2535 subjects in the 1000G). However, when using the most negatively selected transcript of INF2 (ENST00000398337), the odds increased to 24.6 (3 of 312 patients; 1 of 2535 subjects in the 1000G). This is consistent with recent studies reporting INF2 mutations in NS.

**Table 3. Characteristics of the default filter devised to classify variants as causal for NS**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of genes</td>
<td>21</td>
</tr>
<tr>
<td>AF threshold in sequenced samples</td>
<td>&lt;5%</td>
</tr>
<tr>
<td>AF threshold in the EVS</td>
<td>&lt;0.5% (dominant gene); &lt;1% (recessive gene)</td>
</tr>
<tr>
<td>AF threshold in the 1000G</td>
<td>Not used</td>
</tr>
<tr>
<td>Functional prediction threshold</td>
<td>Predicted damaging by two of three of PolyPhen2, SIFT, and MutationTaster (A or D)</td>
</tr>
<tr>
<td>Transcript specificity</td>
<td>Longest transcript</td>
</tr>
<tr>
<td>Conservation score (GERP)</td>
<td>Not used</td>
</tr>
</tbody>
</table>

GERP, genome evolutionary rate profiling.
having monogenic NS can now be sequenced. By using the background prevalence in 21 SRNS genes and making assumptions regarding false positive and false negative rates, simulations were performed to quantify the effect on the accuracy of monogenic classification as a function of 1) the number of genes sequenced or 2) the estimated true prevalence of monogenic NS in the population under study.

Varying Numbers of Genes Analyzed

The default pathogenicity filter was applied to 21 SRNS genes in a theoretical cohort of patients with NS whose predicted prevalence of true monogenic NS was 15% (Figure 3). In this way, 16.5% of participants were classified with monogenic NS. Although 11.7% would have true monogenic NS, 4.9% would be falsely classified as such because of the background prevalence.

The next simulation considered a theoretical scenario in which an additional 21 newly discovered monogenic NS genes were added to a diagnostic panel, with similar background prevalence rates per analysing the initial panel. Now, with this 42-gene panel, 22.9% of subjects were classified with monogenic NS. Although 13.1% would be true positives, 9.8% of subjects would be falsely classified. Thus, because of the rarity of monogenic NS, doubling the number of candidate genes analyzed results in a doubling of false positive classification without significant changes in true positives.

Varying Underlying Prevalence of True Monogenic NS

Diagnostic genetic screening of subjects with a higher pretest probability of monogenic NS improves the accuracy of pathogenic classification. Screening the SRNS gene set in a population with a 2% estimated true prevalence of monogenic NS was simulated. In this scenario, the relative risk of putative mutations being in a patient versus a subject in the 1000G was 1.25 (Figure 4). However, if the prevalence of true NS in the cohort under study was increased to 20%, the relative risk was increased to 3.5.

Systematic Assessment of Parameters to Create a Stringent Pipeline

Ultimately, insights derived from these analyses led to the creation of a stringent filtering strategy (Table 4). Using this "stringent filter" reduced the background prevalence in the 1000G to 1.5%, a 74% decrease compared
Applying the stringent filter devised to classify variants as causal for NS.27–45 This filter correctly identified the causal mutation in these three families, while rejecting common polymorphisms found in each gene. Of 70 established mutations from familial NS, the filter classified 90% as causal as well (63 of 70). Importantly, two of seven of the putative mutations that the filter did not classify as causal, on manual review, were convincingly unlikely to represent causal mutations. Thus, the stringent filter had 93% sensitivity overall when using these mutations as a gold standard. Stratified by inheritance, the filter was 96% and 83% sensitive for published mutations in recessive and dominant SRNS genes, respectively.

**Putative Monogenic NS in Patients under the Stringent Filter**

Applying the stringent filter, 9 of 312 (2.9%) patients with NS in NEPTUNE were classified as monogenic (Table 5). Putative mutations were present in 7 of 21 of the genes sequenced. Seven participants had mutations in dominantly inherited genes. The remaining two were homozygous for NPHS1 and NPHS2 variants; 4 of 10 distinct mutations had been previously reported in the Human Gene Mutation Database (HGMD).

Sixty-seven percent (six of nine) of patients who were monogenic had pediatric onset; 22% (two of nine) had a likely family history of NS. Four participants had a histologic diagnosis of FSGS, and three had minimal change disease. The remaining two in the other glomerulopathy group were diagnosed with IgM nephropathy and immune complex GN. Additional characteristics of participants with putative monogenic versus nonmonogenic NS are presented in Supplemental Table 2. Prevalence estimates of monogenic NS in NEPTUNE stratified by clinical criteria of interest are presented in Table 6.

On the basis of proteinuria measurements obtained longitudinally, 56% of patients with a monogenic diagnosis (five of nine) attained complete remission (CR). This included four of six pediatric-onset patients and one of three adult-onset patients. Additional details in Table 5 show that remission was achieved with a variety of therapeutic strategies, including steroid monotherapy and renin-angiotensin-aldosterone system blockade monotherapy as well as regimens of multiple immunosuppressants. In survival analyses, the chance of achieving CR in those classified with monogenic NS was not significantly different than in those without NS in both unadjusted (hazard ratio, 1.4; $P=0.44$) (Figure 5) and adjusted (hazard ratio, 1.5; $P=0.35$) models.

**Genome-Wide Prevalence of Potential False Monogenic Diagnosis**

The effect of background prevalence on potential false diagnosis extends beyond known NS genes. In subjects of the 1000G, the background prevalence of putative pathogenic mutations was calculated genome-wide using the default pipeline and under both dominant and recessive models. A subset analysis of a gene set known to cause Mendelian diseases under dominant inheritance (hOMIM dominant) was also performed.

Under the dominant model, 83.1% of all genes and 87.7% of hOMIM-dominant genes had putative pathogenic variants in >0.1% of the population (Figure 6). This persisted even after normalizing all genes to a typical gene length. By contrast, only 7.3% of genes under the recessive model had >0.1% of

---

**Table 4. Characteristics of the stringent filter devised to classify variants as causal for NS**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of genes</td>
<td>21</td>
</tr>
<tr>
<td>AF threshold in sequenced</td>
<td>&lt;5%</td>
</tr>
<tr>
<td>samples</td>
<td></td>
</tr>
<tr>
<td>AF threshold in the EVS</td>
<td>&lt;0.1% (dominant gene); &lt;0.1% (recessive gene)</td>
</tr>
<tr>
<td>Functional prediction</td>
<td>&lt;1% in any of 26 populations</td>
</tr>
<tr>
<td>threshold</td>
<td>Predicted damaging by two of three of PolyPhen2, SIFT, and MutationTaster</td>
</tr>
<tr>
<td>Transcript specificity</td>
<td>Most negatively selected</td>
</tr>
<tr>
<td>Conservation score (GERP)</td>
<td>&gt;4</td>
</tr>
</tbody>
</table>

GERP, genome evolutionary rate profiling.
subjects predicted to be monogenic (91% fewer than the dominant model). Thus, especially under a dominant mode of transmission, there is substantial exome-wide background prevalence of predicted harmful, protein-altering variants. Perhaps unexpectedly, although these genes are implicated in Mendelian disease, they do not have fewer pathogenic variants compared with all genes.

**DISCUSSION**

The growing efforts to perform genetic screening in patients with NS are motivated by reports of the clinical significance of deriving a diagnosis of monogenic NS.16 Many of the studies that have shaped the current understanding of monogenic NS were performed in a limited number of genes in patients who, in various ways, were selected for a higher prevalence of monogenic NS. This includes those with steroid resistance, those with congenital or infantile onset, those with familial disease, and those from regions in which there was an elevated degree of consanguinity.48,49 Now, in this genomic era, increasingly diverse populations of affected patients (both demographically and clinically) are undergoing genetic screening of an expanded number of previously implicated NS genes. Given the potential influence on clinical decision making of making this diagnosis, there was motivation to discover whether previous findings about monogenic NS were generalizable in a clinically familiar scenario, namely sporadically affected North American patients with family data that are absent.

This study was designed to determine the spectrum of genetic variation independently in patients and controls under the same filtering strategies rather than solely screening control databases to detect the presence of putative mutations found in patients. In doing so, it was appreciated that, under a typical pathogenicity filter, there was a substantial background prevalence of protein-altering variants in the population (e.g., 5.8% under the default filter). The majority of the background prevalence in the NS gene set resulted from putative pathogenic variants in genes with a dominant inheritance model. The lower background contributed by recessive genes under stringent filtering conditions suggests that finding two predicted harmful variants in the same recessive NS gene in a patient is less likely to be caused by chance. Thus, without accounting for the background prevalence and seeking filtering strategies to minimize it, sequencing studies for rare diseases are at risk for inaccurately classifying variants as causal mutations.

**Table 5. Characteristics of patients with NS classified with monogenic NS**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Ancestry</th>
<th>Histology</th>
<th>Gene</th>
<th>Genotype</th>
<th>Amino Acid Change</th>
<th>HGMD</th>
<th>Pediatric</th>
<th>Remission</th>
<th>Immunosuppressive Agent</th>
<th>RAAS Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AFR</td>
<td>MCD</td>
<td>NPH51</td>
<td>HOM_ALT</td>
<td>p.A765V</td>
<td>No</td>
<td>Yes</td>
<td>Complete</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>AMR</td>
<td>FSGS</td>
<td>TRPC6</td>
<td>HET</td>
<td>p.N765S</td>
<td>No</td>
<td>Yes</td>
<td>Complete</td>
<td>Steroid</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>AFR</td>
<td>MCD</td>
<td>CD2AP</td>
<td>HET</td>
<td>p.M496I</td>
<td>Yes</td>
<td>Yes</td>
<td>Complete</td>
<td>Steroid</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>EUR</td>
<td>IgM nephropathy</td>
<td>WT1</td>
<td>HET</td>
<td>p.C350R</td>
<td>No</td>
<td>Yes</td>
<td>Complete</td>
<td>Steroid, tacrolimus, MMF</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>EUR</td>
<td>ICGN</td>
<td>INF2</td>
<td>HET</td>
<td>p.E220K</td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>EUR</td>
<td>FSGS</td>
<td>NPH52</td>
<td>HOM_ALT</td>
<td>p.A284V</td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>EUR</td>
<td>MCD</td>
<td>INF2</td>
<td>HET</td>
<td>p.A13T</td>
<td>Yes</td>
<td>No</td>
<td>Complete</td>
<td>Steroid</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>EUR</td>
<td>FSGS</td>
<td>TRPC6</td>
<td>HET</td>
<td>p.D798H</td>
<td>No</td>
<td>No</td>
<td>Partial</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>AMR</td>
<td>FSGS</td>
<td>INF2</td>
<td>HET</td>
<td>p.D374N</td>
<td>No</td>
<td>No</td>
<td>Partial</td>
<td>Steroid, CNI</td>
<td>Yes</td>
</tr>
</tbody>
</table>

AFR, African 1000G continental ancestry; AMR, American 1000G continental ancestry; EUR, European continental ancestry; MCD, minimal change disease; ICGN, immune complex GN; HOM_AL, homozygous for the alternative allele; HET, heterozygous variant; MMF, mycophenolate mofetil; CNI, calcineurin inhibitor; RAAS, renin-angiotensin-aldosterone system.

**Table 6. Prevalence of monogenic NS by selected subsets**

<table>
<thead>
<tr>
<th>Description of Subsets</th>
<th>Classified Monogenic (%; Fraction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>2.9 (9/312)</td>
</tr>
<tr>
<td>Pediatric</td>
<td>6.3 (6/95)</td>
</tr>
<tr>
<td>Pediatric + FSGS</td>
<td>6.7 (2/30)</td>
</tr>
<tr>
<td>Pediatric + FSGS + no CR</td>
<td>6.3 (1/16)</td>
</tr>
<tr>
<td>Pediatric + FSGS + CR</td>
<td>7.1 (1/14)</td>
</tr>
<tr>
<td>Pediatric + FSGS or MCD</td>
<td>5.5 (4/73)</td>
</tr>
<tr>
<td>Pediatric + no CR</td>
<td>5.6 (2/36)</td>
</tr>
<tr>
<td>Pediatric + CR</td>
<td>6.8 (4/59)</td>
</tr>
<tr>
<td>Adult</td>
<td>1.4 (3/217)</td>
</tr>
<tr>
<td>Adult + FSGS</td>
<td>2.7 (2/75)</td>
</tr>
<tr>
<td>Adult + FSGS + no CR</td>
<td>3.8 (2/53)</td>
</tr>
<tr>
<td>Adult + FSGS + CR</td>
<td>0.0 (0/22)</td>
</tr>
<tr>
<td>Adult + FSGS or MCD</td>
<td>3.2 (3/95)</td>
</tr>
<tr>
<td>Adult + no CR</td>
<td>1.5 (2/133)</td>
</tr>
<tr>
<td>Adult + CR</td>
<td>1.2 (1/84)</td>
</tr>
<tr>
<td>All + FSGS</td>
<td>3.8 (4/105)</td>
</tr>
<tr>
<td>All + FSGS + no CR</td>
<td>4.3 (3/69)</td>
</tr>
<tr>
<td>All + FSGS + CR</td>
<td>2.8 (1/36)</td>
</tr>
<tr>
<td>All + FSGS or MCD</td>
<td>4.2 (7/168)</td>
</tr>
<tr>
<td>All + no CR</td>
<td>2.4 (5/207)</td>
</tr>
<tr>
<td>All + CR</td>
<td>3.5 (5/143)</td>
</tr>
</tbody>
</table>

MCD, minimal change disease.
This background prevalence was useful to aid in simulations, showing for one, that increasing the pretest probability (i.e., the estimated prevalence) of true monogenic NS in the population under study substantially improves the accuracy of classification. This supports the classic approach of incorporating clinical, laboratory, and histologic data to enrich the true prevalence of monogenic disease. Also, increasing the number of NS genes screened in patients (particularly those with dominant transmission) also increases the false discovery rate. Recognizing that sequencing increased numbers of genes in patients with a low pretest probability of monogenic NS increases inaccurate monogenic diagnosis seems particularly relevant at this time, because expanded genetic screening for NS in the clinical and research arenas may enrich for both of these factors.

Computing background prevalence also provided an outcome metric to calibrate a number of different filtering strategies on prevalence of monogenic diagnosis in subjects in the 1000G and patients. For instance, filters that achieved a background prevalence matching prevalence estimates of NS in the population resulted in an absence of patients with NS being classified with monogenic NS. However, only requiring one method to predict a variant as pathogenic or using AF filters of 1% substantially boosted background prevalence, including for recessive genes. The stringent filter devised reduced background prevalence but also, had face validity. Although the 1.5% background prevalence for NS under the stringent filter is far greater than would be predicted under a Mendelian paradigm, it is consistent with reports of incomplete penetrance for other conditions with monogenic causes.\(^{50,51}\)

Recently, the Exome Aggregation Consortium (ExAC) released the ExAC Browser (http://exac.broadinstitute.org). ExAC provides summary variant calls of whole-exome sequencing data from 60,076 unrelated individuals worldwide. ExAC provides an increasingly accurate estimate of allele frequencies worldwide. In this study, our inferences of putative causal variants were comparable when using ExAC-based AF thresholds versus our existing strategy of combining an EVS AF of \(<0.1\%\) with the 1000G local population AF threshold of \(<1\%\). Moving forward, ExAC should be a powerful resource for the NS community to identify optimal filtering strategies for determining causality of variants for NS. In particular, its increased power to detect rare variants may contribute to our understanding of the prevalence of putative disease-causing mutations in the general population.

Applying the stringent filter to patients in NEPTUNE resulted in a number of notable observations; 2.9% (9 of 312) of patients in NEPTUNE were classified with monogenic NS. However, in NEPTUNE, as in many clinical care settings, parental and sibling sequence data were unavailable to perform segregation analysis, raising the potential that some of these putative causal variants are present in healthy family members. There was a question of whether the NEPTUNE cohort, reflecting a patient population of all comers de novo diagnosed in North American academic medical centers, fully accounts for some of the differences in prevalence estimates compared with previous studies.\(^{2,4,25,52,53}\) Thus, subset analyses were also performed, stratifying patients by more traditional classifications, such as pediatric onset, FSGS histology, or SRNS. Still, the group that one may predict to have the highest prevalence of monogenic diagnosis, children with FSGS who have not achieved CR, had a monogenic prevalence of only 6.3%.

Of equal interest to the prevalence of monogenic NS in the cohort was the observation that 56% (five of nine) of participants classified with monogenic NS achieved CR. This includes one participant with FSGS and two participants whose
causal variants were previously reported in the HGMD. In regard to specific therapies used to achieve CR, it was notable that there was no specific strategy used. CR was achieved with monotherapy with ACE inhibitors or steroids, ACE inhibitors and steroids, and in one participant, steroids, calcineurin inhibitors, and mycophenolate. Furthermore, there was no significant association between monogenic status and failure to achieve CR.

A potential reason for this observation is that the harmful effects of these rare variants are potentiated or protected by environmental or genetic backgrounds that differ in this population versus those previously reported. Another possibility is that NEPTUNE’s prospective, longitudinal design may have captured CR events that are not ascertained or not appreciated in a cross-sectional or retrospective design. Finally, it is possible that the filtering parameters were still not stringent enough and that functional studies and family segregation analysis would have eliminated all participants who achieved CR. For a number of reasons, including having a child with a homozygous variant in *NPHS1* (p.A765V) who achieved CR and a participant with an *INF2* variant previously reported in a family study with SRNS,\(^{36}\) this explanation seems less likely.

Although variable expressivity of the SRNS phenotype in patients with monogenic NS has certainly been reported,\(^{4,29,54–57}\) the immunosuppressive-resistant phenotype in monogenic NS has been of greater focus.\(^{2,5,21,22,25,58–60}\) Arguably, the strongest motivation for clinicians and families to perform genetic screening is the belief that this diagnosis would allow avoidance of immunosuppressant agents that would only harm with no chance of benefit. Altogether, this study’s findings suggest a previously unrecognized degree of variable expressivity of putatively fully penetrant variants in monogenic NS genes.

Thus, at least in this population of sporadically affected North American patients, making a monogenic diagnosis of NS is challenged by background prevalence, particularly in dominant genes, and incomplete penetrance of the SRNS phenotype. Like any screening test, variant prediction filters have imperfect sensitivity and specificity. Under the stringent filter, there was 1.5% background prevalence in the 1000G, suggesting that some variants, even if found in patients, may be false positives. In regard to false negatives, when using 92 published mutations as a gold standard, the sensitivity was not 100%. This was more of a factor in genes with a dominant mode of transmission. Thus, care must be taken to consider the output of this filter in the context of other complementary information about both the variants and affected patients/research participants under consideration. Future sequencing studies of hundreds to thousands of longitudinally phenotyped patients with sporadic and familial NS and controls with proteinuria and renal functional data should increase our accuracy of these prediction filters.

Monogenic prediction may also be challenged by the existence of more complex modes of inheritance in NS, such as alleles that modify the effect’s causal variants, triallelic or bigenic heterozygosity,\(^{61,62}\) or risk alleles that exist independently of rare coding variants in known monogenic NS genes.\(^{63–66}\) These complex forms of genetic architecture should continue to be the focus of study as well. Altogether, these factors seemingly challenge the ability to make treatment decisions solely on the basis of a subject’s genetic sequence. Thus, it may be more accurate to reconsider predicted monogenic mutations as probabilistic risk factors.

There are a number of advantages of a probabilistic model. First, quantifying the pathogenicity of variants discovered along a continuous spectrum may more accurately reflect their effect on disease. Second, consideration of factors additional to variant prediction could allow more informed adjustment of the probability of a patient having monogenic disease. Although more complex than a deterministic, dichotomous diagnosis, quantifying relative risk of a variant being pathogenic for NS may improve the precision of clinical care.
For rare diseases in general, whether for family counseling, clinical management, or study design, there is strong motivation to perform genetic screening and optimize the accuracy and interpretation of a monogenic diagnosis. This study showed that systematically devising a variant filtering strategy calibrated to genetic variation in the population can improve the accuracy and potential clinical use of genetic screening for monogenic forms of NS in a group increasingly undergoing diagnostic sequencing. This study also showed that the background prevalence of rare, protein–altering variants exists genome wide, implying that benefits of accounting for its effect could extend beyond NS.

To this end, the web application GeneVetter (http://genevetter.org) was created to allow investigators of any phenotype to quantify the background prevalence of predicted casual variants in the 1000G within their genes of interest under default or custom filtering strategies. In addition, GeneVetter also allows users to analyze sequence data from their own subjects without transmitting genotype information off their servers. Overall, GeneVetter should aid users in their own sequencing studies by improving accuracy in implicating variants as causal as well as interpreting the results published from other reports.

CONCISE METHODS

Study Participants
Participants were enrolled in NEPTUNE at their first clinically indicated biopsy for suspicion of having minimal change disease, FSGS, or membranous nephropathy. From each enrolled participant, blood, renal biopsy tissue, and urine along with highly detailed demographic and clinical data were collected. Patients are followed four times in the first year and biannually thereafter. Unlike other studies, sporadic patients without ascertainment by family history or age of onset were collected; 325 participants with available DNA for sequencing were included in this study. Of the participants, 13 were removed, because they failed quality control for a total of 312 participants. Per the NEPTUNE protocol, CR was defined as a reduction of the urine protein-to-creatinine ratio to $<0.3$. Partial remission was defined as a reduction of the urine protein-to-creatinine ratio to between 0.3 and 3.5 and a 50% reduction compared with measurement at the eligibility visit.

Matching 1000G Controls
The first 256 subjects with NS were also genotyped using Illumina Exome Chip. Genotypes were called by GenomeStudio software with default cluster. After excluding single-nucleotide polymorphisms with $<95\%$ call rates, all samples had $>98\%$ call rates. These genotypes were combined with the Exome Chip genotypes of 2089 samples from the 1000G across all overlapping markers and produced principal components using the EPACTS software package (http://genome.sph.umich.edu/wiki/EPACTS). The five closest remaining participants with NS and their computed centroid were then iteratively identified. Next, the closest subjects in the 1000G to this centroid were identified by Euclidian distance in the space of the first four principal components, where each principal component was weighted by its eigenvalue. When the number of remaining subjects was $<32$, centroids were computed using two samples instead of five, and the final three participants with NS were matched directly to the closest remaining subjects in the 1000G. This was repeated until 61 controls were selected. The number of participants with NS per centroid was reduced as the number of remaining subjects became small, because the remaining participants were relatively far from each other. One control was removed after sequencing because of poor sequencing quality.

Targeted Sequencing of 21 Genes
Microfluidic PCR paired with next generation sequencing of all exons of 21 monogenic NS genes was performed using methods previously described.

Sequence Alignment and Variant Calling
ATOMIC Targeting (GATK) best practices (https://www.broadinstitute.org/gatk/guide/best-practices; accessed 7/13/2015) were followed for preprocessing, sequence alignment, and variant calling. First adapter sequences were trimmed using cutadapt 1.8.1 BWA mem$^{72}$ (0.7.12) software was used to align the raw sequence data. Duplicated reads were not removed or marked, because it is impossible to distinguish PCR duplicates from independent segments with the microfluidic PCR data. GATK (3.4) was applied to recalibrate base quality scores, perform indel realignment, and detect single-nucleotide polymorphisms and short indels. GATK’s Unified Genotyper was used for variant calling instead of Haplotype Caller, because some known variants were missed with Haplotype Caller. The default settings for the Unified Genotyper were used, except that the down sampling was turned off. GATK best practices recommend using hard variant filters for low-coverage data. However, greater sensitivity and specificity in terms of Sanger–sequenced confirmed variants was achieved using a Support Vector Machine (SVM)$^{73}$ with a radial basis function kernel. The SVM was trained using the R package e1071 (http://cran.r-project.org/web/packages/e1071; R interface to libsvm)$^{74}$ with the following features: read position rank sum, call rate, quality by depth, mean allele balance for heterozygous subjects, strand odd ratio, and inbreeding coefficient. The impute R package (http://www.bioconductor.org/packages/release/bioc/html/impute.html), which uses the k nearest neighbor algorithm, was used to impute missing values. A site was labeled a negative training example if it failed at least two criteria of the default GATK best practices hard filter or was in the 1000G Phase 3 or ExAC v0.3 and not labeled PASS. A site was labeled a positive training example if it was labeled PASS in the 1000G or ExAC. All remaining sites were labeled as positive or negative training examples on the basis of their presence or absence in dbSNP b138. Sites passing the SVM filter were further refined at the genotype level. For heterozygous genotypes, at least five alternative alleles, an allele balance of $\pm 10\%$, and a genotype quality of 40 were required. For nonreference homozygous genotypes, the same parameters were required, with the exception of allele balance.
Functional Annotation of Variants
The functional consequences of exonic variants were annotated using the SnpEff, version 3.5 software package with the GenCODE v9 database.46 Each missense and nonsense variant was annotated using 19 functional scores annotated in the dbNSFP 2.5.76 Variants were annotated as being in the HGMD77 if they matched the chromosome, position, and amino acid change for some HGMD entry.

Variant Filtering Parameters Considered
The following criteria per each variant or individual were considered when devising variant filtering criteria: (1) maximum AF in the EVS across European and black samples; (2) functional prediction results from SIFT, PolyPhen2, and MutationTaster as annotated in dbNSFP; (3) maximum continental AF and population AF calculated from the 1000G after leaving one sample out, which was calculated by [allele count −1]/[total allele count −2] within each continental or population group; (4) minimum genomic evolutionary rate profiling + score78 threshold; and (5) the longest versus the most negatively selected transcript. Variants with AF ≥5% in the sequenced pool of samples were excluded in all of the analyses. The individual thresholds are as stated in each section or are not used if not mentioned.

Concordance Estimated between Functional Prediction Methods
In total, 1220 variants in 21 SRNS genes identified were identified in the subjects of the 1000G. The pathogenicity score for each of these variants was computed by each of Polyphen2, SIFT, and MutationTaster. The top 25% most damaging from each method were then brought forward for evaluation of concordance of classification. This represented 225 variants per program, which because of overlap, comprised 493 distinct variants.

Population Stratification
For each individual, the number of rare (EVS AF <0.5% or <0.1%) and deleterious (deleterious in two of three functional prediction) variants in 21 SRNS genes is calculated. The average number of rare and deleterious variants per person is then calculated and stratified by continental ancestry. Variants were further filtered by continental AF (AF <1%) or population AF (AF <1%) in the 1000G, leaving one sample out in AF calculations to avoid bias. Average numbers of predicted pathogenic variants per each of these populations are then recalculated.

Transcript Filtering
Analyses focused on the longest transcript per gene, except where specified otherwise. Only transcripts that were labeled protein coding from protein-coding genes for autosomal chromosomes were considered. To ensure fair comparison between transcripts, 1718 (2.3%) of 74,569 transcripts that had entries in dbNSFP were removed.

Simulation Methods
The ratio of patients who are true dominant monogenic dominant to patients who are true dominant recessive was assumed to be three. This ratio is the ratio observed in our cohort. The fraction of subjects who are true monogenic was assumed to be 0.15. Subjects who are true monogenic were generated as \( B(1, 0.116) + B(1, 0.039) \), where \( B(1, p) \) is a binomial distribution with one trial and probability \( p \). Patients who were false monogenic were generated by \( B(1, \text{no. of dominant genes} \times \text{observed background rate per dominant gene}) + B(1, \text{no. of recessive genes} \times \text{observed background rate per recessive gene}) \). The observed background rate for dominant genes was 103/2535/7 = 0.0058. The observed background rate for recessive genes was 45/2535/14 = 0.0013. The subjects predicted to be monogenic had a false negative rate of 0.25; therefore, 0.75 of the subjects who are true monogenic were predicted to be monogenic, and all of the subjects who are false monogenic were predicted to be monogenic. For the calculation where we doubled the genes from 21 to 42, we decreased the false negative rate to 15% to account for subjects who were previously classified as a negative simply because their disease gene was not sequenced. The fractions of true positives, true negatives, false positives, true negatives, and false negatives were calculated using 1 million simulated subjects.

 Genome–Wide Background Prevalence
All coding variants in 19,054 genes observed in the 1000G were classified as pathogenic or benign on the basis of stringent filtering criteria (maximum ancestry-specific EVS minor allele frequency <0.1%; 1000G MAF <5%; loss of function or nonsynonymous and passing the two of three functional filter). For the entire gene set, the background prevalence of predicted pathogenic mutations in the 1000G was calculated under a dominant and recessive model (homozygous or two heterozygous variants). This analysis was repeated in a gene set of 390 genes implicated in dominant Mendelian diseases (hOMIM dominant).47 A background prevalence of 1% for a given gene indicates that 1% of subjects of the 1000G carry a rare and deleterious variant in this gene.

Time to Event Analyses of CR
A Cox proportional hazards model was used to determine an association between time to first CR of proteinuria and the status of a subject being classified as monogenic. One model was estimated without adjustment, and the other was adjusted for cohort, baseline age, and sex.

ACKNOWLEDGMENTS

M.G.S. is a Carl Gottschalk Research Scholar of the American Society of Nephrology and is supported by the National Institute of Diabetes and Digestive and Kidney Diseases Grant 1K08-DK100662-01. M.K. is supported by Nephrotic Syndrome Study Network Consortium (NEPTUNE) Grant U54DK083912 and George M. O’Brien Kidney Research Core Center at the University of Michigan Grant P30-DK081943. The NEPTUNE is a part of the National Center for Advancing Translational Sciences (NCATS) Rare Disease Clinical Research Network (RDCRN), which is supported through a collaboration between the Office of Rare Diseases Research (ORDR), the NCATS, and the National Institute of Diabetes, Digestive, and Kidney Diseases. The RDCRN is an initiative of the ORDR and the NCATS.
Additional funding and/or programmatic support for this project have also been provided by the University of Michigan, NephCure Kidney International, and the Halpin Foundation.

DISCLOSURES
None.

REFERENCES
21. Kunze P, Henikoff S, Ng PC: Predicting the effects of coding non-
synonymous variants on protein function using the SIFT algorithm. Nat
Protoc 4: 1073–1081, 2009
22. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: MutationTaster
evaluates disease-causing potential of sequence alterations. Nat
Methods 7: 575–576, 2010
23. Sampson MG, Gillies CE, Ju W, Kretzler M, Kang HM: Negative selection of
deleterious variants in genes associated with Mendelian forms of
nephrotic syndrome. Presented at the American Society of Nephrology
Kidney Week, San Diego, CA, 2012
24. Sampson MG, Kretzler M, Shaw AS, Kang HM: Gene-level in-
vestigations with kidney, lung, and skin disease.
–
B, Hildebrandt F, Bruckner-Tuderman L, Laube GF: Integrin a3 muta-
1514, 2012
Salvati L, Hurd TW, Vega-Warner V, Killen PD, Raphael Y, Ashraf S,
Ovunc B, Schoeb DS, McClain HM, Arikir R, Vlangos CN,
Gbadegesin R, Hinkes B, Saisawat P, Trevisson E, Doimo M, Casarin A,
Pertegato V, Giorgi G, Proksch H, Rötg A, Nürnberg G, Becker C,
Wang S, Ozaltin F, Topaloglu R, Bakaloglu A, Bakaloglu SA, Müller D,
Beissert A, Mir S, Berdeli A, Varpizen S, Zenker M, Matejas V, Santos-
S, Mundel P, Reiser J, Nürnberg P, Clarke CF, Wiggins RC, Faul C,
Hildebrandt F: FOCC6 mutations in human patients produce nephrotic
2011
29. Hinkes B, Wiggins RC, Gbadegesin R, Vlangos CN, Seelow D,
Hennies HC, Goyal M, Wahrham BL, Schachter AD, Mudumana S,
Drummond I, Kerjašchki D, Waldherr R, Dietrich A, Ozaltin F,
Bakaloglu A, Cleper R, Basel-Vanagante L, Pohl M, Griesel M, Tsygin
AN, Soylu A, Müller D, Sörlis CS, Bunney TD, Katan M, Liu J, Attanasio M,
O’toole JF, Hasselbacher K, Mucha B, Otto EA, Arikir K, Kispert G,
Kelley GM, Smrcka AV, Gudermann T, Holzman LB, Nürnberg P, Hildebrandt
F: Positional cloning uncovers mutations in PLCE1 responsible for a
nephrotic syndrome variant that may be reversible. Nat Genet 38:
1397–1405, 2006
30. Winn MP, Conlon PJ, Lynn KL, Farrington MK, Creazzo T, Hawkins AF,
Daskalakis N, Kwan SY, Ebersviller S, Burchette JL, Pericak-Vance MA,
Howell DN, Vance JM, Rosenberg PB: A mutation in the TRPC6 cation
channel causes familial focal segmental glomerulosclerosis. Science
308: 1801–1804, 2005
Dahan K, Gubler MC, Naiedt P, Antinac C: NPHS2, encoding the
glomerular protein podocin, is mutated in autosomal recessive steroid-
Johnson RL, Lee B: Mutations in LMX1B cause abnormal skeletal
patterning and renal dysplasia in nail patella syndrome. Nat Genet 19:
47–50, 1998
33. Pelletier J, Bruening W, Kashtan CE, Mauer SM, Manivel JC, Striegel JE,
Houghton DC, Junien C, Habib R, Fouser L, Fine RN, Silverman BL,
Haber DA, Housman D: Germline mutations in the Wilms’ tumor sup-
pressor gene are associated with abnormal urogenital development in
34. Sanna-Cherchi S, Burgess KE, Nees SN, Caridi G, Weng PL, Daganno M,
Bodria M, Carrea A, Allegretta MA, Kim HR, Perry BJ, Giggia M, Clark
LN, Kisselv S, Cusi D, Gesualdo L, Allegri L, Scolari F, D’Agati V,
Shapiro LS, Pecoraro C, Palomero T, Ghiggeri GM, Gharavi AG: Exome
sequencing identified MYO1E and NEFL1 as candidate genes for hu-
mans autosomal recessive steroid-resistant nephrotic syndrome. Kidney
Int 80: 389–396, 2011
35. Mcle I, latropoulos P, Donadelli R, Calabria A, Maranta R, Cassidy P,
Buelli S, Tomasoni S, Piras R, Krendel M, Bettoni S, Morigi M,
Delledonne M, Pecoraro C, Abbate I, Capobianchi MR, Hildebrandt F,
Otto E, Schaefer F, Macciardi F, Ozaltin F, Emre S, Ibsirlofilo T, Benigni
A, Remuzzi G, Noris M; PodoNet Consortium: MYO1E mutations and
36. Brown EJ, Schlöndorff JS, Becker DJ, Tsukaguchi H, Tonna SJ, Uscinski
AL, Higgs HH, Henderson JM, Pollak MR: Mutations in the formation gene
IN2F cause focal segmental glomerulosclerosis. Nat Genet 42: 72–76,
2010
37. Kaplan JM, Kim SH, North KN, Rennke H, Correia LA, Tong HQ, Mathis
BJ, Rodríguez-Pérez JC, Allen PG, Beggs AH, Pollak MR: Mutations in
ACTN4, encoding alpha-actinin-4, cause familial focal segmental glo-
38. Kestilä M, Lenkkeri U, Quinzi CM, Kankk T, Rodenburg RJ, Naini A,
Dimauro S, Hirano M: Leig syndrome with nephropathy and CoQ10
deficiency due to decaprenyl diposphat synthase subunit 2 (PDSL2)
S, Schumacher V, Royer-Pokora B, Wühl E, Cochat P, Bouvier R, Kraus
K, Mark M, Adlon H, Dötsch J, Rascher W, Maruniak-Chudek I, Lennert T,
Neuman LM, Reis A: Human lamin beta2 deficiency causes con-
genital nephrosis with mesangial sclerosis and distinct eye abnormali-
40. Boerkoel CF, Takashima H, John J, Yan J, Stankiewicz P, Rosenberg
Illies F, Joseph M, Kattila I, Lama G, Loirat C, McLeod DR, Milford DV,
Petty EM, Rodrigo F, Saraiva JM, Schmidt B, McLeod DR, Milford DV,
Petty EM, Rodrigo F, Saraiva JM, Schmidt B, Smith GC, Spranger J, Stein A,
Thiele H, Tizard J, Weksberg R, Lupski JR, Stockton DW: Mutant chromatin remodelling protein
SMARCAL1 causes Schimke immuno-osseous dysplasia. Nat Genet 30:
215–220, 2002
Fathy HM, Schoeb D, Chemin G, Lyons RH, Yilmaz E, Hildebrandt F:
Exome sequencing reveals cubilin mutation as a single-gene cause of
42. Ozaltin F, Ibsirlioglu T, Taskiran EZ, Baydar DE, Kaymaz F, Buyukcelik
M, Kicil BD, Balat A, latropoulos P, Asan E, Akarsu NA, Schaefer F,
Yilmaz E, Bakaloglu A, PodoNet Consortium: Disruption of PTPRO
causes childhood-onset nephrotic syndrome. Am J Hum Genet 89:
139–147, 2011
Miranda MC: A nonsense mutation in the LIMP-2 gene associated with
progressive myoclonic epilepsy and nephrotic syndrome. Hum Mol
Genet 17: 2238–2243, 2008
44. Diomedi-Camassei F, Di Giandomenico S, Santorelli FM, Caridi G,
Piemonte F, Montini G, Ghiggeri GM, Murer L, Barisoni L, Pastore A,
71. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMNetjournal 17: 10–12, 2011
77. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133: 1–9, 2014