Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function


∗To whom correspondence should be addressed at: Division of Preventive Medicine, Brigham and Women’s Hospital, 900 Commonwealth Ave East, Boston, MA 02215, USA. Tel: +1 6172780821; Email: dchasman@rics.bwh.harvard.edu (Daniel I. Chasman); Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA. Tel: +1 7346156833; Email: cfuchsb@umich.edu (Christian Fuchsberger); Universitätsklinikum Freiburg, Medizin IV, Nephrologie, Berliner Allee 29, 79110 Freiburg im Breisgau, Germany. Tel: +49 76127078050; Email: anna.koettgen@uniklinik-freiburg.de (Anna Köttgen)

†Writing group.

‡These authors jointly contributed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
Paolo Gasparini47,84, Mario Pirastu48, J. Wouter Jukema91,92,93, Nicole M. Probst-Hensch50,51, Florian Kronenberg52, Daniela Toniolo53, Vilmundur Gudnason21,22, Alan R. Shuldiner23,94, Josef Coresh11,95, Reinhold Schmidt96, Luigi Ferrucci26, David S. Siscovick27, Cornelia M. van Duijn18, Ingrid B. Borecki12, Sharon L.R. Kardia97, Yongmei Liu32, Gary C. Curhan2,98, Igor Rudan64, Ulf Gyllensten36, James F. Wilson64, Andre Franke68, Peter P. Pramstaller4, Rainer Rettig99, Inga Prokopenko75,76, Jacqueline Witteman18, Caroline Hayward67, Paul M Ridker1,2, Afshin Parsa24, Murielle Bochud100, Iris M. Heid7,13, W.H. Linda Kao11,95,§, Caroline S. Fox2,15,101,§ and Anna Köttgen14,11,§

1Division of Preventive Medicine, Brigham and Women’s Hospital, 900 Commonwealth Avenue, Boston, MA 02215, USA, 2Harvard Medical School, Boston, MA 02115, USA, 3Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA, 4Center for Biomedicine, European Academy of Bozen/Bolzano, Affiliated Institute of the University of Lübeck, Germany, Viale Druso, 1-39100 Bolzano, Italy, 5Interfaculty Institute for Genetics and Functional Genomics, University of Greifswald, Friedrich-Ludwig-Jahn-Str. 15a, 17487 Greifswald, Germany, 6Department of Internal Medicine II and 7Department of Epidemiology and Preventive Medicine, University Hospital Regensburg, Franz-Josef-Strauss-Allee 11, 93042 Regensburg, Germany 8Institute of Anatomy and Cell Biology, University of Greifswald, Friedrich-Loeffler-Str. 23c, 17487 Greifswald, Germany, 9Department of Neurology, Boston University School of Medicine, 72 East Concord St. B603, Boston, MA 02118, USA, 10Department of Biostatistics, Boston University School of Public Health, 715 Albany St., Boston, MA 02118, USA, 11Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, USA, 12Division of Statistical Genomics, Washington University School of Medicine, 4444 Forest Park Blvd. Box 8506, St Louis, MO 63108, USA, 13Institute of Medical Informatics, Biometry and Epidemiology, and Institute of Epidemiology, Ludwig-Maximilians-Universität, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, 14Renal Division, Freiburg University Clinic, Berliner Allee 29, 79110 Freiburg, Germany, 15Center for Population Studies, NHLBI Framingham Heart Study, 73 Mt Wayte Ave Suite #2, Framingham, MA 01702, USA, 16Division of Nephrology, Brigham and Women’s Hospital, 75 Francis St., Boston, MA 02215, USA, 17Section of Preventive Medicine and Epidemiology, Department of Medicine, Boston University School of Medicine, 761 Harrison Ave, Boston, MA 02118, USA, 18Department of Epidemiology, Erasmus Medical Center, Dr. Molewaterplein 50, PO Box 2040, 3000CA Rotterdam, The Netherlands, 19Centre for Medical Systems Biology, Leiden University, Leiden, The Netherlands, 20Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Ave, CT3, Boston, MA 02118, USA, 21Icelandic Heart Association, Research Institute, Holtasmani 1, 201 Kopavogur, Iceland, 22University of Iceland, Sæmundargötu 2, 101 Reykjavik, Iceland, 23Department of Medicine and 24Division of Nephrology, University of Maryland Medical School, Baltimore, MD, USA 25Departments of Epidemiology and Biostatistics, and Forensic Molecular Biology, University of Rotterdam, Dr Molewaterplein 50-603015, Rotterdam, GE, The Netherlands, 26Clinical Research Branch, National Institute of Aging, 3001 S Hanover Street, Baltimore, MD 21225, USA, 27University of Washington, 1730 Minor Ave. Campus Box 358080, Seattle, WA 98101-1448, USA, 28Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA and 29Departments of Genetics and Developmental Biology, Stanford University, Stanford, CA 94305, USA 30Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA and 31Department of Internal Medicine, Division of Nephrology and Hypertension, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA 32Department of Epidemiology and Prevention, Public Health Sciences, 33Department of Internal Medicine and 34Department of Biostatistical Sciences, Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA 35Department of Nutrition, Harvard School of Public Health, 665 Huntington Avenue, Building 2, Boston, MA 02115, USA, 36Genetics and Pathology, Rudbeck Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden, 37IFB Adiposity Diseases and Department of Medicine, University of Leipzig, Liebigstr. 18, 04103 Leipzig, Germany, 38INSERM UM744, Institut Pasteur, Lille, France, 39Centre for Information-based Medicine and School of Medicine and Public Health, University of Newcastle, Hunter Medical Research Institute, Newcastle, NSW, Australia, 40Institute of Genetics and Biophysics ‘Adriano-Buzzati Traverso’, CNR, Via P. Castellino 111, 80131 Napoli, Italy, 41Department

§These authors jointly oversaw the work.
of Medical Genetics, University of Lausanne, Rue du Bugnon 27, CH, 1005 Lausanne, Switzerland. 42Swiss Institute of Bioinformatics, Lausanne, Switzerland, 43Department of Clinical Chemistry, Centre for Laboratory Medicine Tampere Finn, University of Tampere and Tampere University Hospital, Medi 2, 3th, floor, PO Box 2000, 33521 Tampere, Finland, 44Institute of Molecular and Cell Biology, Estonian Biocenter, University of Tartu, Riia 23, Tartu, Estonia. 45Estonian Genome Center, University of Tartu (EGCUT), Tiigi 61b, Tartu, Estonia, 46Clinical Research Centre, University of Dundee-Wellcome Trust Centre for Molecular Medicine, Level 7, Ninewells Hospital, Dundee DD1 9SY, Scotland, 47Institute for Maternal and Child Health IRCCS ‘Burlo Garofolo’, Trieste, Italy, 48Department of Population Genetics, CNR-Traversa La Crucca, 3 07040 Reg. Baldinca Li Punti, Sassari, Italy, 49Department of Cardiology, Leiden University Medical Center, Postzone C2-R, PO Box 9600, 2300 RC Leiden, The Netherlands, 50Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, PO Box, Basel 4002, Switzerland, 51University of Basel, Basel, Switzerland, 52Division of Genetic Epidemiology, Innsbruck Medical University, Schoepfstraße 41, 6020 Innsbruck, Austria, 53Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Via Olgettina 58, 20132 Milano, Italy, 54Coronary ARtery Disease Genome-wide Replication And Meta-Analysis (CARDIoGRAM) Consortium, 55International Consortium for Blood Pressure Genomewide Association Studies, 56Candidate Gene Association Resource (CARe) Consortium, 57Welcome Trust Case Control Consortium 2, 58Laboratory of Epidemiology, Demography, and Biometry, NIA-Gateway Building, 3C309, 7201 Wisconsin Ave., Bethesda, MD 20892-9205, USA, 59Human Genetics Center Houston, University of Texas Health Science Center, TX-1200 Herman Pressler Drive, Houston, TX 77030, USA, 60Institute of Molecular Biology and Biochemistry and Department of Neurology, Medical University Graz, Harrachgasse 21 8010, Graz, Austria, 61Division of Nephrology/Tufts Evidence Practice Center, Tufts University School of Medicine-Tufts Medical Center, Boston, MA 02111, USA, 62Abteilung Innere II, Universitätshäklinikum Ulm, Albert-Einstein-Allee 23, Ulm 89081, Germany, 63Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology II, Neuherberg, Germany, 64Center for Population Health Sciences, University of Edinburgh Medical School, UK-Teviot Place, Edinburgh EH8 9AG, Scotland, 65Andrija Stampar School of Public Health, Medical School, University of Zagreb, Zagreb, Croatia, 66Croatian Centre for Global Health, Faculty of Medicine, University of Split, Soltanska 2, Split, Croatia, 67MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine—Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland, 68Institute of Clinical Molecular Biology, Christian-Albrechts University, Schittenhelmstr.12, Kiel 24105, Germany, 69Section for Epidemiology, Institute for Experimental Medicine, Christian-Albrechts-University Kiel and Popgen Biobank, University Hospital Schleswig-Holstein, Niemannsweg 11, Kiel 24105, Germany, 70Clinic for Prosthodontic Dentistry, Gerostomatology and Material Science, University of Greifswald, Rotgerberstr 8, Greifswald 17475, Germany, 71Institute of Pharmacology, University of Greifswald, Friedrich-Loeffler-Str., 23d, Greifswald 17487, Germany, 72Institute of Clinical Chemistry and Laboratory Medicine, Ernst-Moritz Arndt-University Greifswald, Ferdinand-Sauerbruch-Str., Greifswald 17475, Germany, 73Clinic for Internal Medicine A, University of Greifswald, Friedrich-Loeffler-Str. 23a, Greifswald 17475, Germany, 74Institute for Community Medicine, University of Greifswald, Walther-Rathenau-Str. 48, Greifswald 17487, Germany, 75Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK, 76Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, 77Department of Internal Medicine, Erasmus Medical Center, PO Box 1738, 3000 DR Rotterdam, The Netherlands, 78INSERM U897, Université Victor Ségalen Bordeaux 2, ISPED case 11, 146 Rue Léo Saignant, F-33076 Bordeaux, France, 79Centre for Vision Research, University of Sydney, Westmead Millennium Institute, Westmead Hospital, Sydney NSW 2145, Australia, 80Centre for Eye Research Australia (CERA), University of Melbourne, 32 Gisborne Street, East Melbourne, Victoria 3002, Australia, 81INSERM UMRS 1018-CESP Team 10, Univ Paris Sud, Villejuif, France, 82Department of Clinical Physiology, University of Tampere and Tampere University Hospital, 33521-PO Box 2000, Tampere 33521, Finland, 83Department of Clinical Physiology, Research Centre of Applied and Preventive Cardiovascular Medicine, Turku University Hospital, University of Turku, PO Box 52, Turku 20521, Finland, 84University of Trieste, Trieste, Italy, 85Department of Medicine, University Medical Centre Mannheim, Theodor Kutzer Ufer 1-3, Mannheim 68167, Germany, 86Robertson Centre for Biostatistics, University of Glasgow, Glasgow, Scotland, 87Department of Pharmacology and Therapeutics, University College Cork, Cork, Ireland, 88First Department of Internal Medicine, Paracelsus Medical University, Müllner Hauptstraße 48, Salzburg 5020, Austria, 89Department of Internal Medicine, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Rue du Bugnon 46, Lausanne 1011, CH,
In conducting genome-wide association studies (GWAS), analytical approaches leveraging biological information may further understanding of the pathophysiology of clinical traits. To discover novel associations with estimated glomerular filtration rate (eGFR), a measure of kidney function, we developed a strategy for integrating prior biological knowledge into the existing GWAS data for eGFR from the CKDGen Consortium. Our strategy focuses on single nucleotide polymorphism (SNPs) in genes that are connected, in a biological sense, to genes at previously validated eGFR associations. It then requires association thresholds consistent with multiple testing, and finally evaluates novel candidates by independent replication. Among the samples of European ancestry, we identified a genome-wide significant SNP in FBXL20 ($P = 5.6 \times 10^{-8}$) in meta-analysis of all available data, and additional SNPs at the INHBC, LRP2, PLEKHA1, SLC3A2 and SLC7A6 genes meeting multiple-testing corrected significance for replication and overall $P$-values of $4.5 \times 10^{-4}$–$2.2 \times 10^{-7}$. Neither the novel PLEKHA1 nor FBXL20 associations, both further supported by association with eGFR among African Americans and with transcript abundance, would have been implicated by eGFR candidate gene approaches. LRP2, encoding the megalin receptor, was identified through connection with the previously known eGFR gene DAB2 and extends understanding of the megalin system in kidney function. These findings highlight integration of existing genome-wide association data with independent biological knowledge to uncover novel candidate eGFR associations, including candidates lacking known connections to kidney-specific pathways. The strategy may also be applicable to other clinical phenotypes, although more testing will be needed to assess its potential for discovery in general.

INTRODUCTION

Chronic kidney disease (CKD) is a major public health burden, affecting up to 13% of the US population and comparable proportions of other populations worldwide (1,2). The clinical ramifications of CKD extend beyond the kidney to comorbidities including hypertension and cardiovascular disease. CKD is defined via the estimated glomerular filtration rate (eGFR) (3), which is heritable (4,5). The findings from genome-wide association studies (GWAS) of CKD and eGFR have identified multiple loci in biological pathways related to nephrogenesis, podocyte function, angiogenesis, solute transport and metabolic functions of the kidney (6,7). However, the validated associations in the largest GWAS explain only about 1.4% of the variation in eGFR, suggesting that the published data for CKD and eGFR still include a large number of true genetic associations awaiting identification and validation (6).

Given the modest effect sizes of the common genetic variants identified by GWAS and the stringent statistical significance levels in GWAS ($P < 5 \times 10^{-8}$) (8), increasing sample size has been the conventional approach for providing statistical power to detect these additional loci. However, the sample size for GWAS may be limited by the availability of studies with the phenotype of interest. Thus, identification and validation of true but modest associations may require novel methods, including prioritizing single nucleotide polymorphism (SNP) associations through independent biological knowledge. One such approach may be to restrict genome-wide analysis to a more targeted search, focused on genes that are connected, in a biological sense, to genes at previously validated GWAS loci, here for eGFR association.
The objective of this study was, therefore, to identify novel CKD and eGFR-associated candidate loci through systematic application of prior biological knowledge within an existing eGFR GWAS, followed by external replication for eGFR association in separate studies, and association testing with other related phenotypes. To do so, our goal was to develop an algorithmic strategy focused on candidate genes that were connected to previously validated eGFR loci through pairwise gene similarities derived from natural language processing of PubMed abstracts in Gene Relationships Across Implicated Loci (GRAIL) (9,10) as well as a pairwise metric of biological relatedness derived from the gene ontology (GO) gene set classifications (11).

RESULTS

Information about cohorts and baseline characteristics of all study populations of European ancestry totaling 130 600 individuals are presented in the Supplementary Material, Tables S1, S3 and S4 (Women’s Genome Health Study, WGHS, and first meta-analysis) and Supplementary Material, Tables S2, S5 and S6 (external replication cohorts = second meta-analysis).

Figure 1 depicts a multistep strategy for identifying novel candidate associations with eGFR on the basis of inferred biological connections to 24 seed genes implicated by previous GWAS as, in most cases, the nearest gene to a replicated eGFR SNP (see Methods); biological connections were inferred from similarities in published scientific abstracts (GRAIL) and GO hierarchies. The 24 seed genes used as the input to the strategy are presented in Supplementary Material, Table S7. As described in the Methods section, the number of top-ranked candidate genes considered for each seed gene was optimal for \( N = 25 \), i.e. the number of candidates yielding a minimal overall false discovery rate (FDR) in the first replication sample (Supplementary Material, Tables S8–S11). In addition, restricting candidate genes to On-line Mendelian Inheritance in Man (OMIM) rather than RefSeq yielded lower overall FDR estimates in the first replication sample (see Materials and Methods). Applying this optimized method to each of the 24 seed genes resulted in 33 candidate SNP associations that were located in one of 416 unique candidate genes and associated with eGFR at \( P < 0.05 \) after gene-wide correction for the number of SNPs (see Methods). These 33 SNPs were tested in the first replication meta-analysis and it was estimated that 0.07, i.e. 7\%, represented null hypotheses, implying that an estimated 31 \( [33 \times (1 - 0.07)] \) of the associations were true (Supplementary Material, Table S8). In contrast, when 24 randomly chosen seeds were used instead of the 24 validated GFR seeds, the estimated fraction of hypotheses representing the null in the first replication meta-analysis was a median 0.58 (inter-quartile range 0.44–0.78), i.e. 58\%. Furthermore, this estimate did not vary according to the number of candidate genes selected for each seed gene (Supplementary Material, Table S8), emphasizing the relative estimated enrichment of true associations when the algorithm was applied to the authentic seed genes. Examining more than 25 candidates per seed gene led to an estimated smaller fraction of true associations but a greater total number overall (Supplementary Material, Table S8). Not restricting the candidate genes to OMIM resulted in lower estimates for the fraction of true associations (compare Supplementary Material, Tables S8–S11). Very similar results were observed when considering SNPs within 10 000 bp of each candidate gene instead of 1000 bp (see Methods, compare Supplementary Material, Tables S8–S11).

With the optimized algorithm, namely investigating SNPs within 1000 bp of each of the 25 most related OMIM genes per seed gene, 10 candidate SNPs were selected for replication in a second meta-analysis (Step 4, Fig. 1) consisting of 18 independent cohorts totaling 56,246 samples. Five of the ten SNPs were selected because of an association with eGFR at \( P < 0.05 \) in the first replication meta-analysis and with the same direction of effect as in the WGHS discovery sample derived from applying the optimized algorithm with the GRAIL metric (Supplementary Material, Table S8, bold). The remaining five SNPs had the smallest \( P \)-values similarly from applying the algorithm using the GO pairwise similarity metric as opposed to the GRAIL metric.

In the independent, second replication, six of the ten SNPs met the required FDR standard of \( q \)-value <0.05 (Table 1). They had one-sided \( P \)-values ranging from \( 5.6 \times 10^{-2} \) to \( 6.9 \times 10^{-3} \) and combined \( P \)-values (discovery and all replication) from \( 4.5 \times 10^{-4} \) to \( 5.6 \times 10^{-9} \). These SNPs mapped in or near to LR2P on chromosome 2, PLEKHA1 on chromosome 10, SLC3A2 on chromosome 11, INHBC on chromosome 12, SLC7A6 on chromosome 16 and FBXL20 on chromosome 17. This last SNP, rs7208487, reached conventional genome-wide significance in the combined analysis (\( p_{\text{comb}} = 5.6 \times 10^{-9} \)). The six SNPs, as well as three of the remaining four
considered for independent replication that had FDR > 0.05 in the second independent replication, had direction of effect concordant with the WGHS discovery sample \( P \)-value for directional consistency (binomial distribution) = 0.01, Table 1 and Supplementary Material, Table S12). The GRAIL words and GO terms defining the connections between the seed and candidate genes are shown in Table 2, and detailed information about each of the six replicated genes is provided in Box 1.

Regional association plots using imputed SNP data from the combined WGHS and first CKDGen replication meta-analysis confirmed that each replicating SNP was among the SNPs with the lowest \( P \)-values within each candidate gene, as expected by design (Supplementary Material, Fig. S1A–F). At the LRP2 locus, the replicated GFR-associated SNP rs10490130 is one of two highly significant SNPs in the region; the second SNP (rs6433115), which has a smaller \( P \)-value in the first meta-analysis than rs10490130, is located in a separate block of linkage disequilibrium (LD) and was not selected for replication because it was not genotyped in the WGHS (Supplementary Material, Fig. S1A).

To further support and characterize the six novel replicating SNPs, we assessed the association with eGFR in an existing genome-wide association study among African-Americans from the CARe Consortium \( (N > 7300, \text{Table } 3) \) (12). In this independent sample, the SNPs at PLEKHA1 and FBXL20 showed nominally significant association with eGFR, and the minor allele at all six SNPs had the same direction of effect on eGFR compared with the discovery and replication cohorts of European ancestry (directional consistency test, \( P = 0.031 \)). Moreover, when looking at the entire gene region rather than at the index SNP only, the most significantly associated SNP at three loci (PLEKHA1, FBXL20, SLC7A6) met nominal significance after correction for the number of independent SNPs in each region evaluated, while the most significantly associated SNP at LRP2 narrowly missed the threshold for locus-wide significance but showed a larger effect than the LRP2 index SNP among European ancestry individuals (Table 3).

We also evaluated whether the six index SNPs were associated in cis with the expression of a nearby transcript in several tissues (eQTL, see Methods, Table 4). Of the six SNPs, intronic rs4751890 in the PLEKHA1 gene was the only eQTL SNP for its assigned candidate gene, displaying a significant association with PLEKHA1 transcript levels in blood \( (P = 8 \times 10^{-4}) \) and lymphocytes \( (P = 2 \times 10^{-3}) \). In blood (13), it was the strongest expression-related SNP for this transcript among all SNPs. Increasing copies of the minor allele of rs4751890 were associated with decreased levels of PLEKHA1 expression (13) and with increased eGFR (Table 1). Among the other SNPs, rs6499166 in SLC7A6, rs7208487 at FBXL20 and rs489381 at SLC3A2 were associated with gene expression in cis but none was an eQTL for one of the candidate genes (Supplementary Material, Table S13).

A separate eQTL analysis on kidney biopsies from 81 individuals in two patient cohorts did not show significant

---

### Table 1. Genome-wide and suggestive SNPs identified using GRAIL and GO metrics of biological relatedness to seed genes

<table>
<thead>
<tr>
<th>SNP information</th>
<th>GRAIL</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>rs10490130</td>
<td>rs4751890</td>
</tr>
<tr>
<td>Position (chromosome:basepair)</td>
<td>chr2:169807356</td>
<td>chr10:124151780</td>
</tr>
<tr>
<td>Minor allele frequency (WGHS)</td>
<td>C/A</td>
<td>C/T</td>
</tr>
<tr>
<td>Candidate gene</td>
<td>LRP2</td>
<td>PLEKHA1</td>
</tr>
<tr>
<td>Seed gene</td>
<td>DAB2</td>
<td>PIP5K1B</td>
</tr>
<tr>
<td>Candidate gene rank in GRAIL</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Gene size (bp)</td>
<td>235504</td>
<td>57647</td>
</tr>
<tr>
<td>Minor allele frequency (WGHS)</td>
<td>0.08</td>
<td>0.40</td>
</tr>
<tr>
<td>Number of SNPs in gene in WGHS</td>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>Best gene-wide corrected P-value</td>
<td>3.0E-02</td>
<td>4.9E-02</td>
</tr>
<tr>
<td>Combined WGHS and first replication meta-analysis (( N = 74 ) 354)</td>
<td>4.4E-02</td>
<td>7.0E-03</td>
</tr>
<tr>
<td>Combined second independent replication meta-analysis (( N = 56 ) 246)</td>
<td>6.3E-04</td>
<td>4.1E-04</td>
</tr>
<tr>
<td>All samples combined (( N = 130 ) 600)</td>
<td>-0.008</td>
<td>0.003</td>
</tr>
</tbody>
</table>

\( ^a \)Minor allele (A1) and major allele (A2).  
\( ^b \)Effect and standard error for the effect of each additional copy of the minor allele.  
\( ^c \)Rank of candidate gene in similarity to the seed gene by GRAIL metric among all non-seed genes in OMIM.
Table 2. Gene function for genome-wide significant and suggestive associations

<table>
<thead>
<tr>
<th>Candidate</th>
<th>seed</th>
<th>N GRAIL words</th>
<th>N (%) words for 50% of GRAIL score</th>
<th>GRAIL words/GO terms connecting seed and candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRP2</td>
<td>DAB2</td>
<td>1940</td>
<td>10 (0.52%)</td>
<td>Megalin, dab2, disabled, endocytic, receptor, epithelial, 2, endocytosis, binding, lipoprotein</td>
</tr>
<tr>
<td>SLC7A6</td>
<td>SLC7A9</td>
<td>996</td>
<td>9 (0.90%)</td>
<td>4f2hc, transport, amino, transporter, dibasic, acid, barcelona, kidney, acids</td>
</tr>
<tr>
<td>SLC3A2</td>
<td>SLC7A9</td>
<td>1124</td>
<td>7 (0.62%)</td>
<td>Cystinuria, 4f2hc, cystine, 4f2, transport, amino, transporter</td>
</tr>
<tr>
<td>PLEKHA1</td>
<td>PIK3K1B</td>
<td>819</td>
<td>6 (0.73%)</td>
<td>Ptdins, phosphatidylinositol, 4, bisphosphate, phosphate, actin</td>
</tr>
<tr>
<td>FBXL20 (from GO)</td>
<td>UBE2Q2</td>
<td>NA</td>
<td>NA</td>
<td>Modification-dependent protein catabolic process, post-translational protein modification, regulation of protein metabolic process</td>
</tr>
<tr>
<td>INHBC (from GO)</td>
<td>VEGFA</td>
<td>NA</td>
<td>NA</td>
<td>Kidney-specific (1): kidney development, not kidney-specific (43), e.g. mRNA stabilization, primitive erythrocyte differentiation, cell maturation, cell migration</td>
</tr>
</tbody>
</table>

association between the index SNPs or highly correlated SNPs and gene expression in cis (see Methods, Supplementary Material, Tables S13 and S14).

To evaluate whether the associations with eGFR reflected effects on kidney function more broadly, we investigated association of the six replicating SNPs with other renal phenotypes in recently published genome-wide scans (14). The Supplementary Material, Table S15, shows that none of the SNPs showed significant associations with either microalbuminuria (MA) or the urinary albumin-to-creatinine ratio (both N = 63 153) after correction for multiple testing. Variants in moderate LD (r² > 0.6) with intronic rs7208487 in FBXL20 are significantly associated with the concentrations of urinary histidine and tyrosine among a selection of urinary metabolites examined (15).

As effects on kidney function may also have ramifications for blood pressure phenotypes or incident cardiovascular disease, associations for the six SNPs were examined in GWAS data from the International Consortium for Blood Pressure (ICBP) (systolic and diastolic blood pressure) and the CARDIoGRAM (myocardial infarction) Consortia. No associations were observed for any of the SNPs (Supplementary Material, Table S15).

Finally, we interrogated the NHGRI GWAS catalog for associations of variants at the loci identified here with additional phenotypes (Supplementary Material, Table S16). The region containing INHBC on chromosome 12 was identified in association with serum urate concentrations (16) and rs3741414 in the 3′-UTR of INHBC identified in our study is in high LD with the top urate-associated SNP (r² = 0.84). The allele associated with higher urate levels is associated with lower eGFR, in agreement with the known relationship between serum urate levels and kidney function.

**DISCUSSION**

We identify six novel associations for eGFRcrea, one genome-wide significant and the remaining five highly suggestive, through a systematic strategy that uses data from existing meta-analyses of GWAS to select and replicate new SNP associations not identified in an initial genome-wide scan. This is accomplished through the selection of SNPs in genes that are connected to genes containing previously confirmed SNP associations on the basis of existing biological knowledge. Our approach may represent a way to identify novel associated genomic regions for other complex phenotypes without the need for further increase in the sample size. Although more testing will be needed to assess its utility in general, the approach can be applied whenever there is access to genome-wide SNP data in a single study and results from a GWAS meta-analysis for primary replication. All of the GFR associations at the six novel genomic regions satisfy conventional statistical thresholds for overcoming multiple hypothesis testing, and rs7208487 in FBXL20 meets the standard genome-wide significance threshold (P < 5 × 10⁻⁸) in the combined discovery and replication meta-analyses totaling 130 600 samples. However, the power for detecting the other associations even in the combined sample at genome-wide significance ranged only from 0.01 to 0.21. It may have been further diminished by demographic differences between the discovery cohort (WGHS) and the replication meta-analyses, thus emphasizing the challenges for establishing genome-wide significance for these SNPs and other weak but possibly true replicating associations with GFR even with a large sample.

**Biological and methodological context of findings**

While the biological functions of some of the novel candidate genes are more certain than others, the connections with the seed genes emphasize roles in solute transport, endocytosis, posttranscriptional modification and development (Table 2). For some of the genes, a role in the kidney is known and dysfunctions of the encoded proteins could plausibly be linked to reduced eGFR. This is the case for the genes encoding subunits of an amino acid transport complex expressed in kidney, SCL7A6 and SLC3A2, and for LRP2, encoding megalin, which has a known function in the reabsorption of albumin and other low molecular weight proteins from the urine. Rare mutations in LRP2 have been described as a cause of Donnai–Barrow and facio-oculo-acoustico-renal syndromes (17). Conversely, PLEKHA1 and FBXL20 were connected to their seed genes by terms that are not specific to the kidney (Table 2). The lack of a prior connection to kidney function for these last two candidates, PLEKHA1 and FBXL20, underscores the potential for identifying novel eGFR-associated genes and thus advancing biological understanding by the strategy we describe. These genes would not have been identified at genome significance in an initial meta-analysis nor
Box 1. Known functions of genome-wide significant and suggestive genes for eGFR.

**FBXL20**: little published evidence links F-box and leucine-rich repeat protein 20 to renal function, except an expression pattern that includes kidney. The gene product has been reported to have a role in regulatory pathways involving ubiquitination (57), which have also been suggested for its seed gene, **UBE2Q2**.

**INHBC**: identified through its seed gene **VEGFA**, which may be related to renal function through developmental processes. The **INHBC** gene encodes the beta C chain of inhibin (58), which belongs to the TGF-beta superfamily and together with other subunits forms activin complexes. Activins have a role in the regulation of hormone secretion as well as cell differentiation and growth including branching of the kidney, but the specific role of the subunit encoded by **INHBC** is not well studied.

**LRP2**: encodes megalin, which as part of a complex with cublin and amnionless has a known function in the reabsorption of albumin and other low molecular weight proteins from the urine (59). **LRP2** is connected to its seed gene, **DAB2**, through protein–protein interaction (60). Megalin localizes to the proximal renal tubule; the mechanism by which it is connected to eGFR is presently unclear but could include protein internalization and subsequent renal damage (59). Rare mutations in **LRP2** have been identified in patients with Donnai–Barrow and facio-oculo-acoustico-renal syndromes (17). We previously identified variants in **CUBN**, encoding for cublin, as associated with UACR (14), but the GFR-associated SNP in **LRP2** identified here did not show association with UACR or MB. This implies that variation in different components of the megalin complex associates with different measures and mechanisms of kidney disease.

**PLEKHA1**: a potential link to renal physiology is its expression in renal tubule cells. It is reported to localize to the plasma membrane and binds to phosphatidylinositol-3, 4-bisphosphate (61), which suggests a role in signaling consistent with the phosphatidylinositol-4-phosphate-5-kinase activity encoded by its seed gene, **PIP5K1B**. Recently, a role in the regulation of insulin sensitivity has been reported for the **PLEKHA1**-encoded protein (62). Polymorphisms in the **PLEKHA1** gene are associated with age-related macular degeneration (63–65).

**SLC7A6** and **SLC3A2**: the gene products of both **SLC7A6** and **SLC3A2** interact, and their co-expression facilitates the uptake or exchange of amino acids such as arginine, leucine and glutamine for example in the kidney (66). The protein encoded by **SLC3A2** is expressed at high levels in both freshly isolated and cultured podocytes, kidney-specific cells (67). Both genes were identified based on the seed genes **SLC7A9**.

would they have been investigated in a classical candidate gene approach that focused on biological pathways specific to the kidney. It is worth noting, however, that single SNPs rather than genes were replicated, and follow-up experimental evidence is needed to confirm that the genes that suggested the associations are indeed the causal genes in the region.

In conventional genome-wide genetic analysis, where the evidence for association is solely statistical, SNPs selected for replication typically rank among the top few hundred according to P-value. This corresponds approximately to the top <0.008% in a study based on HapMap (r2<2) imputed genotypes at ~2.6 million SNPs. In contrast, the SNPs selected for replication by our prioritization approach had clearly lower ranks, and none of the 10 loci advanced for a second replication here was among the top 10 with associations just above genome-wide significance thresholds (p>5×10−8). Among 146215 genotyped WGHS SNPs mapping within 1000 bp of Refseq genes, the ranks of the six replicated SNPs ranged from the top 0.10 to 2.33% (1148–3409th, Table 5), and 0.23–6.23% (all RefSeq) or 0.24–6.31% (OMIM only) for the gene-wise corrected P-values used in the SNP selection algorithm. Thus, these SNPs would have been unlikely to be selected for replication on the basis of P-value alone.

Comparison with other approaches

The strategy used here can be compared and contrasted with previously reported approaches for integrating prior biological information into genome-wide association analysis. First, the pairwise gene similarity matrix used in the development of the strategy and for SNP selection is derived from GRAIL directly (10). GRAIL computes P-values for the scores of binary connections between genes tagged by independently selected candidate SNPs, including sub-genome-wide associations where it has been proven effective in identifying novel replicating SNPs (9). This application of GRAIL may be useful in resolving the ambiguity in assignment of the gene underlying an observed genome-wide association. Our strategy differs in prioritizing genes on the basis of gene-wide corrected rather than uncorrected association P-values. We also relied on a discovery and first replication approach guided by the FDR. This two-stage approach may have diminished overall power for individual associations, but also provided empirical support for proceeding to the second replication step. There was little benefit of the two-stage approach in identifying potentially heterogeneous associations that would otherwise have been missed, since heterogeneity was modest among the six novel associations in the meta-analysis combining all samples (I² range 0–37%). A final difference between our strategy and typical GRAIL applications was the selection of candidates on the basis of the rank of the gene–gene score rather than a fixed threshold of the quantitative gene score. By design, functional inferences in GRAIL may be most accurate with a quantitative rather than ordinal interpretation of the gene–gene scores, especially for the top connections. However, we note that the gene–gene scores between the top 25th ranked candidates and each of the seed genes were comparable having a mean (SD) of 0.24 (0.12) and none was >0.65 (within a maximal range of 0–1).

Second, there is overlap with gene set methods that test for an over-abundance of significant associations mapping to pre-specified collections of genes with related function termed “gene sets” (18–20). These gene sets may be inferred from the literature and often include the GO gene collections,
Table 3. Interrogation of genome-wide and suggestive loci in CARe African–American sample (12)

<table>
<thead>
<tr>
<th>Index SNP ID</th>
<th>Closest gene</th>
<th>Locus best SNP</th>
<th>chrom position (build 36)</th>
<th>A1/A2</th>
<th>A1 beta(SE), P-value</th>
<th>N. indep. SNPs</th>
<th>N. P-threshold</th>
<th>Significant (gradient Y)</th>
<th>LD-CELsquare (squared r)</th>
<th>A1/A2</th>
<th>A1 beta(SE), P-value</th>
<th>N. indep. SNPs</th>
<th>N. P-threshold</th>
<th>Significant (gradient Y)</th>
<th>LD-CELsquare (squared r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10490130</td>
<td>LRP2</td>
<td>rs4668121</td>
<td>2:169684351</td>
<td>T/G</td>
<td>0.45</td>
<td>44</td>
<td>1.0E-03</td>
<td>Yes</td>
<td>0.156</td>
<td>0.002</td>
<td>7.38</td>
<td>0.09</td>
<td>0.015</td>
<td>No</td>
<td>0.156</td>
</tr>
<tr>
<td>rs4751890</td>
<td>PLEKHA1</td>
<td>rs7919241</td>
<td>10:124152898</td>
<td>A/C</td>
<td>0.22</td>
<td>17</td>
<td>0.05</td>
<td>No</td>
<td>0.119</td>
<td>0.014</td>
<td>3.8E-03</td>
<td>0.0094</td>
<td>0.013</td>
<td>Yes</td>
<td>0.119</td>
</tr>
<tr>
<td>rs489381</td>
<td>SLC3A2</td>
<td>rs2282538</td>
<td>11:62356308</td>
<td>T/C</td>
<td>0.11</td>
<td>8</td>
<td>0.02</td>
<td>Yes</td>
<td>0.0142</td>
<td>0.007</td>
<td>1.3E-02</td>
<td>0.055</td>
<td>0.011</td>
<td>No</td>
<td>0.0142</td>
</tr>
<tr>
<td>rs6499166</td>
<td>SLC7A6</td>
<td>rs9935022</td>
<td>16:66836780</td>
<td>G/A</td>
<td>0.11</td>
<td>14</td>
<td>0.05</td>
<td>No</td>
<td>0.097</td>
<td>0.002</td>
<td>7.1E-03</td>
<td>0.0114</td>
<td>0.017</td>
<td>No</td>
<td>0.097</td>
</tr>
<tr>
<td>rs7208487</td>
<td>FBXL20</td>
<td>rs4795358</td>
<td>17:34826591</td>
<td>G/T</td>
<td>0.33</td>
<td>3</td>
<td>0.02</td>
<td>Yes</td>
<td>0.007</td>
<td>0.001</td>
<td>6.3E-03</td>
<td>0.0103</td>
<td>0.015</td>
<td>No</td>
<td>0.007</td>
</tr>
<tr>
<td>rs3741414</td>
<td>INHBC</td>
<td>rs2242578</td>
<td>12:56139420</td>
<td>T/C</td>
<td>0.10</td>
<td>2</td>
<td>0.005 (0.0004), 0.0019 (0.0003)</td>
<td>No</td>
<td>0.0207</td>
<td>0.014</td>
<td>1.0E-03</td>
<td>0.0077</td>
<td>0.016</td>
<td>No</td>
<td>0.0207</td>
</tr>
</tbody>
</table>

aA1 = minor allele, A2 = major allele.
bN = 738 for all association tests. cMost significant SNP within 25 kb of the transcribed region of the candidate gene. dFrom LD pruning in the CARe sample using the PLINK algorithm with window of 50, window shift of 5 SNPs, and VIF = 2. eLD between index SNP and locus best SNP. 

which were transformed here for a gene–gene similarity metric alternative to the GRAIL metric. Some gene set methods further adopt the strategy of establishing the significance for each gene through gene-wide multiple hypothesis correction as was done in our strategy (19).

Finally, recent applications have begun incorporating protein–protein interaction data to find connections among candidate genes, allowing not only binary but also multivalent connections (21). Although protein–protein interaction data may be considered more reliable than text-based inference as in GRAIL, the total number of genes in the human genome explored for protein–protein interaction is currently smaller than the number explored through the literature. In preliminary analyses, a multivalent strategy based on GRAIL similarity scores was not encouraging, in part because the GRAIL connections among our GFR seed genes were weak (data not shown).

Strengths and limitations

Our findings, which are supported by independent replication, highlight both novel candidate associations for eGFR and the biological context through which they may act. The latter aspect of the approach is typically not conferred by analysis based solely on statistical considerations. One concern with this approach may be a potential bias introduced by the reliance on prior biological knowledge as represented in the scientific literature. While some of our results may be influenced by the fact that some genes are better studied than others, it is interesting to observe that biological information directly related to eGFR does not account for all of the identified connections. For instance, the connections of the seed genes to PLEKHA1 and FBXL20 were more closely related to their inferred molecular properties than to kidney function (Table 2). As genome-wide analyses of gene expression continue to supply public databases with new, tissue-specific and unbiased snapshots of the functional state, integration of prior biological information into genome-wide association analysis can be expected to become less biased and increasingly revealing.

Searching for new candidate genes through inferred biological connections was more efficient using the previously validated eGFR-associated seeds compared with genes chosen at random, emphasizing the gains in specificity provided by the method. However, we note that the enrichment of true associations as estimated by the QVALUE algorithm applied to results from the first replication may have an upward bias. In spite of an estimated alternative hypothesis rate of 0.93 (1–0.07), only 6 of 33 candidates in the first replication had P < 0.05 in SNP selection with the GRAIL metric, and the lower estimated rate of the null hypotheses compared with estimates using randomly chosen seed genes appears to derive from overall excess of smaller rather than significant P-values (e.g. of 33 SNPs, 9 had P < 0.1, 15 had P < 0.2, median P = 0.34; see Supplementary Material, Table S8 for estimated rate and inter-quartile range for random seed genes). With this caveat in mind, it is noteworthy that the estimated yield of true associations when substituting randomly selected genes for the previously validated seed genes was still appreciable (Supplementary Material,
Lymph (lymphoid cells) and blood (peripheral blood mononuclear cells).

Table 4. eQTL* analysis for rs4751890 at PLEKHA1

<table>
<thead>
<tr>
<th>Index SNP</th>
<th>Closest gene</th>
<th>Tissue</th>
<th>eSNP</th>
<th>Best eSNP in that tissue</th>
<th>P-value of the best eSNP for the given transcript</th>
<th>LD r2 index SNP and best eSNP</th>
<th>Array ID</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4751890</td>
<td>PLEKHA1</td>
<td>Blood</td>
<td>7.95E-05</td>
<td>Same SNP</td>
<td>2.6E-04</td>
<td>n/a</td>
<td>HSG00267573</td>
<td>(13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lymph</td>
<td>1.90E-03</td>
<td>rs6585827</td>
<td>1.5E-11</td>
<td>0.605</td>
<td>GI_31377719-S</td>
<td>(35)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blood</td>
<td>6.20E-08, 2.30E-06</td>
<td>rs6585827</td>
<td>0.605</td>
<td>3930541, 6650324</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lymph (lymphoid cells) and blood (peripheral blood mononuclear cells).

Table S8), possibly suggesting abundant sub-genome-wide SNP associations with eGFR, as has been observed also in genome-wide studies of other clinical phenotypes (e.g. height (22,23)). Some of these additional candidate associations with eGFR may be validated by replication through the statistical power of larger replication samples.

Finally, we cannot exclude the possibility that—despite prior evidence connecting the seed genes to genes identified here—one or more of the causal variants presumed to underlie the observed associations influence the function not of one of the six novel genes but instead neighboring genes not explicitly implicated by our strategy.

Conclusions

Our strategy’s potential yield of true associations as judged by the FDR highlights the vast repository of modest but true associations remaining to be discovered in existing genome-wide scans of eGFR. The six novel loci, including one genome-wide significant association in FBXL20, are assigned to at least five distinct biological processes. They extend the role of genome-wide analysis in understanding the genetic underpinnings influencing kidney function. Thus, our findings provide a starting point for functional studies, which may lead to additional insights into kidney disease pathophysiology and ultimately contribute to the potential for attenuating the burden of kidney disease.

MATERIALS AND METHODS

We developed a multistep strategy to identify novel eGFR-associated variants in genes that have a biological connection to one or more of 24 genes within previously identified and replicated genome-wide significant eGFR-associated loci (Fig. 1, Step 0). Within each of these loci, the gene closest to the SNP with the lowest P-value in the region was chosen, except for the NAT8/LALMS1 and ANXA9/LASS2 regions, where the second closest gene was chosen on the basis of pre-existing biological evidence: rare mutations in ALMS1 lead to a Mendelian disorder featuring kidney disease (24), and LASS2 (alias CERS2) −/− mice show renal abnormalities (25). These 24 genes, termed ‘seed’ genes, form the starting point of the strategy and are listed in Supplementary Material, Table S7. In outline, the strategy consisted of specifying an algorithm for identifying SNP associations (Fig. 1, Steps 1 and 2), exploring parameters of this algorithm to optimize the estimated yield of true associations with eGFR, selecting candidate SNPs for replication using the optimized parameters and finally replicating the candidate SNPs in two different datasets including an entirely independent replication meta-analysis of 56 246 individuals (Steps 3 and 4). The study populations, genotyping and statistical analyses are described in detail below.

Study populations

(1) The source dataset, from which the 24 seed genes were derived, consisted of 20 study populations that formed the CKDGen Consortium at the time of their publication (6).

(2) The primary sample for discovery of eGFR-associated SNPs in genes implicated by the seed genes was the WGHS, a large population-based cohort of female health care professionals aged ≥45 at baseline (26). As described elsewhere, 21 940 WGHS participants of verified, self-reported European ancestry had information on whole genome genotype data and baseline serum creatinine measures for estimating eGFR available as summarized in the supplement (Supplementary Material, Table S1 summarizes the WGHS cohort, Supplementary Material, Table S3 shows WGHS baseline characteristics).

(3) Since the publication describing the 24 seed genes (6), the CKDGen Consortium had expanded from the 20 original cohorts to include 26 cohorts with information on genome-wide genotypes and renal phenotypes. Information about baseline characteristics of the 25 cohorts, excluding the WGHS, of European ancestry totaling 52 414 individuals that were used as the first replication sample to verify SNPs associated in the WGHS are listed in the supplement (Supplementary Material, Table S1 summarizes the cohorts, Supplementary Material, Table S3 shows baseline characteristics).

(4) The 18 cohorts of European ancestry totaling 56 246 individuals that comprised the second, independent replication meta-analysis are described in the supplement (Supplementary Material, Table S2 summarizes the cohorts, Supplementary Material, Table S5 shows baseline characteristics).

Phenotype definition

In each cohort, serum creatinine was measured as described in the supplement (Supplementary Material, Tables S1 and S2)
and in previous publications (6). To minimize inter-laboratory variation in serum creatinine measurements among the study cohorts, serum creatinine was calibrated to the National Health and Nutrition Examination Study (NHANES) standards in all discovery and replication studies as described previously (27–29). The GFR was then estimated from the calibrated serum creatinine using the four-variable MDRD study equation (30).

Genotype information

Information on genotyping, imputation and statistical analysis methods for all participating cohorts is presented in Supplementary Material, Table S4 (WGHS and first replication sample) and Supplementary Material, Table S6 (second independent replication sample).

Metrics of biological relatedness between pairs of genes

GRAIL infers relatedness between genes on the basis of shared informative words extracted from PubMed abstracts (10). The GRAIL relatedness information is summarized in a sparse matrix, with genes in rows and informative words in columns, so that each gene is represented as a vector of the weights of the informative words. Each of the gene vectors was first normalized to have a dot product with itself equal to 1. The pairwise similarity score between two genes was calculated as the dot product of their normalized word vectors, i.e. a number between 0 and 1. The analysis presented here used the GRAIL relatedness matrix constructed from PubMed abstracts published through 2006, which largely predates information from GWAS to avoid influencing the results by previously published information from GWAS.

The GO (http://www.geneontology.org) is a well-established resource for the annotation of gene products across species. The functional relatedness between pairs of genes was derived from the semantic similarity between GO terms associated with each gene as described in (11). GO terms from all three ontologies (biological process, cellular component or molecular function) were considered in the calculation of the similarity value. Since gene ontologies are directed acyclic graphs (DAG), the relationship between the two GO terms reflects the relative locations of these terms in the DAG graphs as well as their semantic relations (class-subclass relation or partial ownership relation) with their ancestor terms.

Gene selection algorithm

Preliminaries

All analyses used gene assignments from Refseq downloaded in March 2010 from the UCSC genome browser (http://genome.ucsc.edu). Gene names were mapped to Entrez gene IDs for use with GRAIL using information downloaded from NCBI. The OMIM database was downloaded in April 2010 from NCBI. Mapping of SNPs to genic regions used the March 2006 build 36 reference sequence for the human genome (hg18).
Algorithm
The algorithm used to identify eGFR-associated SNPs consisted of three steps. First, for each seed gene, all genes in OMIM or RefSeq (excluding the seed genes) were ranked ordered according to the GRAIL or GO metrics of relatedness to the seed gene. A fixed number \( N \) of the most related genes was designated as a set of candidate genes for further consideration (Fig. 1, Step 1). Second, genotyped SNPs from the WGHS mapping within a fixed number of base pairs (bp) of the transcribed region of each candidate gene were examined for association with eGFR in the WGHS genome-wide scan. Only SNPs genotyped on the Illumina HumanHap300 chip were used in order to keep the amount of LD between the SNPs small. The SNP with the lowest \( P \)-value in each candidate gene was retained if its association with eGFR in the WGHS had \( P < 0.05 \) after a stringent Bonferroni correction for the number of genotyped SNPs mapping to that same gene (Fig. 1, step 2). Third, SNPs meeting gene-wide significance in the WGHS were then examined for association with eGFR in the first replication meta-analysis, which did not include the WGHS (Fig. 1, step 3). The decision to identify associations using a discovery (WGHS) and replication (meta-analysis of 25 CKDGen cohorts) strategy rather than a combined larger discovery meta-analysis (26 cohorts, including WGHS) in order to select SNPs for follow-up in external replication samples was based on several considerations, including the large size of the WGHS, the relatively low LD among its genotyped HumanHap300 chip and its epidemiologic homogeneity, the latter potentially allowing the detection of otherwise heterogeneous associations. Moreover, splitting the available samples between discovery and first replication addressed the need for an intermediate replication sample to optimize the algorithm using the FDR (next section). Statistical significance thresholds and cutoffs for advancement of SNPs to the next stage are provided in the section ‘Association analysis and Statistical Significance for Discovery and Replication’.

Algorithm optimization
Three aspects of the algorithm as implemented with the GRAIL metric were optimized using the estimated overall FDR, designated as \( \pi_0 \), applied to the \( P \)-values of SNP associations in the first replication meta-analysis using the QVALUE software with the default settings (Fig. 1, step 3). As a first aspect, the fixed number \( N \) of most related genes considered for each of the seed genes was varied from 10 to 200. As a second aspect, restricting the candidate genes to the subset of genes in GRAIL with OMIM annotation \((n = 10\,978)\) was compared with the use of the complete set of genes in GRAIL and also in RefSeq genes \((n = 19\,698)\). As a third aspect, the mapping of SNPs to the candidate genes was varied, considering SNPs within either 1000 or 10,000 bp of the transcribed region of each candidate gene. As presented in Supplementary Material, Tables S8–S11, the performance of the algorithm applied to the 24 seed genes was compared across all combinations of the three aspects by estimating the overall fraction of true associations in the first replication meta-analysis as \( 1 - \pi_0 \). Further, these estimated fractions of true associations were compared with the distribution of the estimated fractions of true associations derived by applying the algorithm to 24 genes chosen at random instead of the seed genes over 50 iterations. A substantial effect on enrichment for the estimated overall fraction of true associations was observed when the number \( N \) of most related candidate genes was varied: the algorithm was optimized as judged by a minimum of \( \pi_0 \) for the \( P \)-values in the first replication meta-analysis when the top 25 genes were examined (Supplementary Material, Tables S8–S11). Restricting candidate genes in GRAIL to OMIM also substantially minimized \( \pi_0 \), possibly because the literature derived GRAIL metric may be more informative for genes in OMIM rather than RefSeq genes in general. There was little difference when SNPs within either 1000 bp or 10,000 bp of each candidate gene were examined. When investigating the 25 most seed-gene-related OMIM genes, examining SNPs within 1000 bp yielded a slightly larger estimated number of true associations compared with SNPs within 10,000 bp (Supplementary Material, Tables S8 and S10). For the main part of the paper, the results are therefore presented based on the algorithm configured to investigate SNPs within 1000 bp of the 25 candidate genes from OMIM most related to the seed genes. The same configuration was used in applying the algorithm with the pairwise similarity metric derived from GO instead of GRAIL.

Association analysis and statistical significance for discovery and replication
In the WGHS, SNP associations with eGFR were performed by linear regression assuming an additive relationship between the number of inherited minor alleles of each SNP and the natural logarithm of eGFR, adjusted for age. In the first replication analysis of CKDGen cohorts excluding the WGHS, cohort level association was performed within each study for each SNP with the additive assumption applied to either experimental genotype information or imputed genotype information using dosage data, adjusting for age and sex. Meta-analysis was performed by inverse variance weighing assuming fixed effects, applying genomic control both at the study level and overall to offset potential inflation of the test statistic due to incidental confounding (6). The value of the final genomic control parameter \( \lambda \) in the first meta-analysis was 1.09.

In the second replication analysis in the 18 external cohorts only, significance of candidate associations was judged with one-sided \( P \)-values based on the direction of association in the discovery step. The FDR was applied to these \( P \)-values.

The thresholds for statistical significance at each step of the procedure were as follows: Candidate SNPs were identified in the WGHS (step 1) on the basis of gene-wide Bonferroni-corrected \( P \)-values \(< 0.05 \). Candidate SNPs were further advanced in the first replication step (overall second step) with \( P < 0.05 \) and effect in same direction as WGHS. SNPs were considered nominally replicated on the basis of an FDR \( q \)-value \(< 0.05 \) (QVALUE software, (31)) as applied to one-sided \( P \)-values from the second independent replication analysis. Genome-wide significance was specified as \( P < 5 \times 10^{-8} \) in analysis combining all samples.

Power
Power for SNP associations was estimated assuming an additive relationship between the number of inherited alleles and
mean log eGFR, using effect sizes estimated from the first replication meta-analysis to diminish the effect of the winner’s curse.

**Association of replicated SNPs with additional phenotypes**

Associations with coronary artery disease were examined in a dataset derived from meta-analyses of SNP associations with the urinary albumin-to-creatinine ratio (UACR) and MA among 63,153 European ancestry participants from the CDKGen Consortium (14).

Associations with coronary artery disease were examined in a dataset derived from meta-analyses of SNP associations in a study of >22,000 cases and >64,000 controls of European ancestry in the CARDIOGRAM Consortium (32). Associations with systolic and diastolic blood pressure were examined in a dataset derived from meta-analyses of SNP associations in a study of >69,000 individuals of European ancestry in the ICBP Consortium (28). Associations with concentrations of metabolites in urine were examined using the publicly available SNP associations published recently (15). LD for proxy SNPs was estimated using SNAP (33) as applied to genotypes in the HapMap (CEU population, r22) (34).

**Expression SNP analysis**

Published associations between gene transcript levels and the genotype of nearby SNPs in cis for a wide spectrum of tissue/cell types were interrogated to assess the potential of the replicating candidate SNPs for eGFR to influence gene expression. The tissues and cell types with available data were: fresh lymphocytes (35), fresh leukocytes (36), leukocyte samples in individuals with celiac disease (37), lymphoblastoid cell lines (LCL) derived from asthmatic children (38), HapMap LCL from three populations (39), a separate study on HapMap CEU LCL (40), peripheral blood monocytes (41,42), adipose (13,43) and blood samples (13), 2 studies including prefrontal cortex, visual cortex and cerebellum (Emilsson, personal communication), liver (43,45), osteoblasts (46), skin (47) and additional fibroblast, T cell and LCL samples (48). For each tissue or cell type, the citation describes the study-specific statistical criterion for establishing significant SNP associations. Finally, we queried a database of normal kidney cortex tissue samples from two patient cohorts (49,50). The Affymetrix 6.0 genome-wide chip was used for genotyping, and genotype calling was performed using Affymetrix’s GTC Software. To evaluate an association with the expression of transcripts in cis, SNPs with call rates >90% were related to the expression probes measured on the Affymetrix U133 set, using RefSeq annotation (Affymetrix build a30). The P-values were computed using linear multivariable regression in each cohort and then combined using Fisher’s combined probability test (49). Pairwise linkage disequilibrium was obtained from SNAP (33) with the CEU HapMap release 22 as the reference.

**Programming**

All programming was performed in R (51), including the FDR estimation, which used the default settings from the R-package QVALUE (31). Association testing in the WGHS was performed with PLINK (52). Association testing in the replication cohorts from both stages was performed as described in Supplementary Material, Tables S4 and S6. Quality control of GWAS result files was performed with GWAtoolbox (53). Meta-analysis was performed with METAL (Release February 2010 (54)). Plots of locus-wide association were prepared with LocusZoom (55).

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG Online.

**ACKNOWLEDGEMENTS**

Acknowledgements and funding information are listed in the Supplementary Material, Tables S1 and S2.

**Conflict of Interest statement**

None declared.

**REFERENCES**


