



Multivariate canonical correlation analysis identifies additional genetic variants for chronic kidney disease



Amy J. Osborne¹ ✉, Agnieszka Bierzynska², Elizabeth Colby², Uwe Andag³, Philip A. Kalra⁴, Olivier Radresa³, Philipp Skroblin³, Maarten W. Taal⁵, Gavin I. Welsh², Moin A. Saleem² & Colin Campbell¹ ✉

Chronic kidney diseases (CKD) have genetic associations with kidney function. Univariate genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) associated with estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN), two complementary kidney function markers. However, it is unknown whether additional SNPs for kidney function can be identified by multivariate statistical analysis. To address this, we applied canonical correlation analysis (CCA), a multivariate method, to two individual-level CKD genotype datasets, and *metaCCA* to two published GWAS summary statistics datasets. We identified SNPs previously associated with kidney function by published univariate GWASs with high replication rates, validating the *metaCCA* method. We then extended discovery and identified previously unreported lead SNPs for both kidney function markers, jointly. These showed expression quantitative trait loci (eQTL) colocalisation with genes having significant differential expression between CKD and healthy individuals. Several of these identified lead missense SNPs were predicted to have a functional impact, including in *SLC14A2*. We also identified previously unreported lead SNPs that showed significant correlation with both kidney function markers, jointly, in the European ancestry CKDGen, National Unified Renal Translational Research Enterprise (NURTURE)-CKD and Salford Kidney Study (SKS) datasets. Of these, rs3094060 colocalised with *FLOT1* gene expression and was significantly more common in CKD cases in both NURTURE-CKD and SKS, than in the general population. Overall, by using multivariate analysis by CCA, we identified additional SNPs and genes for both kidney function and CKD, that can be prioritised for further CKD analyses.

Chronic kidney disease (CKD), a major public health burden, affects over 697 million people and causes over one million deaths per year¹. CKD etiology is complex; its occurrence is related to either a primary renal disorder or a complication of a multisystem disorder or comorbidity (secondary CKD)^{2,3}. Estimated glomerular filtration rate (eGFR), used to assess CKD stage, and blood urea nitrogen (BUN or serum urea), are complementary kidney function markers. eGFR is estimated from serum creatinine⁴. BUN measures the nitrogen component of serum urea, the primary metabolite derived from dietary protein and tissue protein turnover^{4,5}.

Common genetic variants are thought to contribute to CKD risk via complex genetic architecture⁶. Genome-wide association studies (GWAS) have identified several common SNPs and loci associated with CKD or kidney function^{7–19}. In May 2019, a trans-ancestry GWAS meta-analysis of 765,348 CKDGen participants, and replication in 280,722 Million Veteran Program participants, reported 264 eGFR-associated loci (256 for European ancestry), of which 147 (134 for European ancestry) were prioritized as likely relevant for kidney function by additional independent association with BUN¹⁷. In August 2019, a transethnic GWAS for eGFR in 280,722

¹Intelligent Systems Laboratory, University of Bristol, Bristol BS8 1TW, UK. ²Bristol Renal, University of Bristol and Bristol Royal Hospital for Children, Bristol BS1 3NY, UK. ³Department of Metabolic and Renal Diseases, Evotec International GmbH, Marie-Curie-Strasse 7, 37079 Göttingen, Germany. ⁴Department of Renal Medicine, Salford Royal Hospital, Northern Care Alliance NHS Foundation Trust, Stott Lane, Salford M6 8HD, UK. ⁵Centre for Kidney Research and Innovation, University of Nottingham, Derby, UK. ✉e-mail: amy.osborne@bristol.ac.uk; c.campbell@bristol.ac.uk

Table 1 | Datasets analysed in this study

Dataset	Data type	Reference	Number of		
			Participants	Chronic kidney disease cases	Single nucleotide polymorphisms
NURTuRE-CKD, NURTuRE-controls	Individual-level genotype	NURTuRE Biobank	2519	2500 (99%)	6,419,966
Salford Kidney Study, NURTuRE-controls	Individual-level genotype	Ali et al. ⁵⁸	1919	1900 (99%)	6,290,407
CKDGen, European ancestry	Genome-wide association study summary statistics	Wuttke et al. ¹⁷	567,460 (eGFR and BUN)	41,395 (7%)	8,346,783
BioBank Japan	Genome-wide association study summary statistics	Kanai et al. ¹⁶	143,658 (eGFR); 139,818 (BUN)	8586 (5%)	5,961,600

Million Veteran Program participants, followed by replication in 765,289 CKDGen participants, confirmed 54 loci and identified 82 previously unreported variants¹⁸. In 2021, by integrating CKDGen and UK BioBank data (predominantly European ancestry), Stanzick et al identified 424 loci associated with eGFR, of which 348 were classified as likely relevant for kidney function based on additional independent association with either BUN or eGFRcys¹⁹. Several loci associated with each of eGFR and BUN were identified by the BioBank Japan GWAS (162,255 Japanese individuals)¹⁶. BUN is affected by other aspects of renal disease, rather than simply filtration rate, but one of its main advantages here is that it is complementary to eGFR²⁰. Therefore it has value in validation of SNPs found to be associated with eGFR. A 2021 study integrating GWAS summary statistics with expression quantitative trait loci (eQTL) data, which links SNPs with gene expression in specific cell types, identified over 182 likely causal kidney function genes²¹.

In a typical univariate GWAS, millions of associations between individual genetic variants and a phenotype of interest are tested using regression models²². Variants that show a statistically significant association with the trait of interest are typically clustered (due to linkage disequilibrium (LD)) in sets of correlated variants²². Canonical correlation analysis (CCA) can simultaneously test for multivariate-based correlation (or co-variance) between multiple SNPs and multiple phenotypic variables²³. CCA was originally introduced in 1926 to find combinations maximally correlated with each other using linear combinations of variables derived from two sets of data objects²⁴. CCA for genotype-phenotype analysis was proposed in 2009²⁵ and subsequently extended for testing the association of multiple SNPs with phenotypes in unrelated individuals²⁶. CCA is symmetric in that the two datasets have equivalent status, whereas multiple multivariate regression, the most similar statistical method to CCA, is asymmetric in that it tests whether each of the responses can be explained by linear combinations of the explanatory variables. CCA therefore allows the identification of multiple SNPs (gene-gene interactions) and pleiotropic mechanisms thought to be the product of complex genetic diseases²⁷. We previously applied CCA to a cardiovascular disease genotype dataset and confirmed already established findings with increased power (*P*-value), and found novel pleiotropic genotype-phenotype associations²³. *MetaCCA*, developed for identifying multivariate relationships from univariate GWAS summary statistics, has shown agreement with CCA results and identified shared SNPs among type 2 diabetes, obesity and coronary artery disease and stroke risk factors, and between CKD and heart disease^{28–31}. However, multivariate methods such as CCA and *metaCCA* have not previously been applied to multiple CKD genomic datasets to look for additional SNPs associated with two kidney function markers, jointly.

Here, to identify additional SNPs associated with two kidney function markers by multivariate methods, we applied CCA and *metaCCA* to two types of CKD genotype-phenotype dataset. We applied *metaCCA* to three publicly available GWAS summary statistics datasets with minor subsets of CKD cases (the CKDGen study (European ancestry) and BioBank Japan)^{16,17,28}. We applied CCA to two individual-level SNP genotype datasets of mostly CKD patients (NURTuRE-CKD and the Salford Kidney Study).

Results

Single nucleotide polymorphisms identified by *metaCCA*

For each of the CKDGen (European ancestry; 567,460 participants) and BioBank Japan (143,658 participants) GWAS summary statistics datasets (for eGFR and BUN), totals of 8,346,783 and 5,837,593 SNPs were analyzed using univariate-SNP *metaCCA* (Table 1, Figs. 1 and 2). Of these, totals of 26,562 (0.3%) and 5513 (0.09%) unique SNPs, respectively, showed a significant correlation with both eGFR and BUN, jointly, using *metaCCA* (Bonferroni-corrected *P*-value < 0.05; Fig. 2). These results reflected their sample sizes (described above) which affected the power to detect SNPs (Table 1). Of these, using the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) program, 472 (1.8%) and 208 (3.8%) were lead SNPs (of 514 and 1657 independent SNPs), for totals of 253 and 99 independent genomic loci, respectively (Supplementary Datasets 1 and 2)³². Lead SNPs were defined by FUMA as independent SNPs (using LD estimation) with the lowest *P*-value in the genomic region³². Of these 472 (1.8%) and 208 (3.8%) identified lead SNPs by *metaCCA* in the CKDGen and BioBank Japan datasets, respectively, totals of 157 (33%) and 48 (23%) lead SNPs were (i) not previously reported as statistically significant for eGFR and BUN by the respective published GWASs^{16,17}, and (ii) showed effect sizes in opposite directions (thus compatible for kidney function) (Datasets 1 and 2). These mapped to 117 and 40 independent genomic loci, of which the closest gene annotations for 75 and all 40, had not previously been reported as closest mapped genes for the lead SNPs reported by Wuttke et al.¹⁷ (CKDGen) and Kanai et al.¹⁶ (BioBank Japan), respectively (Datasets 1 and 2)^{16,17}.

Overlap with previously reported lead SNPs by published GWAS

For each of the CKDGen and BioBank Japan datasets, we compared *P*-values of *metaCCA*-identified SNPs with those previously reported by the respective published GWAS by using quantile-quantile plots (Figs. 3 and 4). In the published CKDGen GWAS by Wuttke et al., SNPs were defined as significant for kidney function if they showed an association with both eGFR (*P*-value < 5.0×10^{-8}) and BUN (one-sided *P*-value < 5.0×10^{-2}), and with eGFR and BUN effect sizes in opposite directions (total of 10,437 SNPs)¹⁷. For the published BioBank Japan GWAS by Kanai et al.¹⁶, the eGFR and BUN *P*-value cut-offs were both < 5.0×10^{-8} (total of 1241 SNPs with effect size direction filter)¹⁶. For our *metaCCA* analyses in this paper, we used a standard genome-wide Bonferroni correction (5e–8) to identify significant multivariate *P*-values for both eGFR and BUN (Methods). When compared to the published CKDGen and BioBank Japan univariate GWAS for eGFR and BUN, multivariate eGFR and BUN *metaCCA* identified 9846 (94%), and 1241 (100%) SNPs, respectively, of those previously reported SNPs and identified additional SNPs for kidney function (Figs. 3 and 4). Of the 26,562 (0.3%) CKDGen and 5513 (0.09%) BioBank Japan dataset SNPs that showed a significant correlation with both eGFR and BUN using *metaCCA*, 5840 (22%) and 2471 (45%) SNPs, respectively, had not been previously reported to show a significant association with eGFR and BUN by the published univariate GWASs by Wuttke et al.¹⁷ and Kanai et al.¹⁶ (Figs. 3 and 4)^{16,17}.

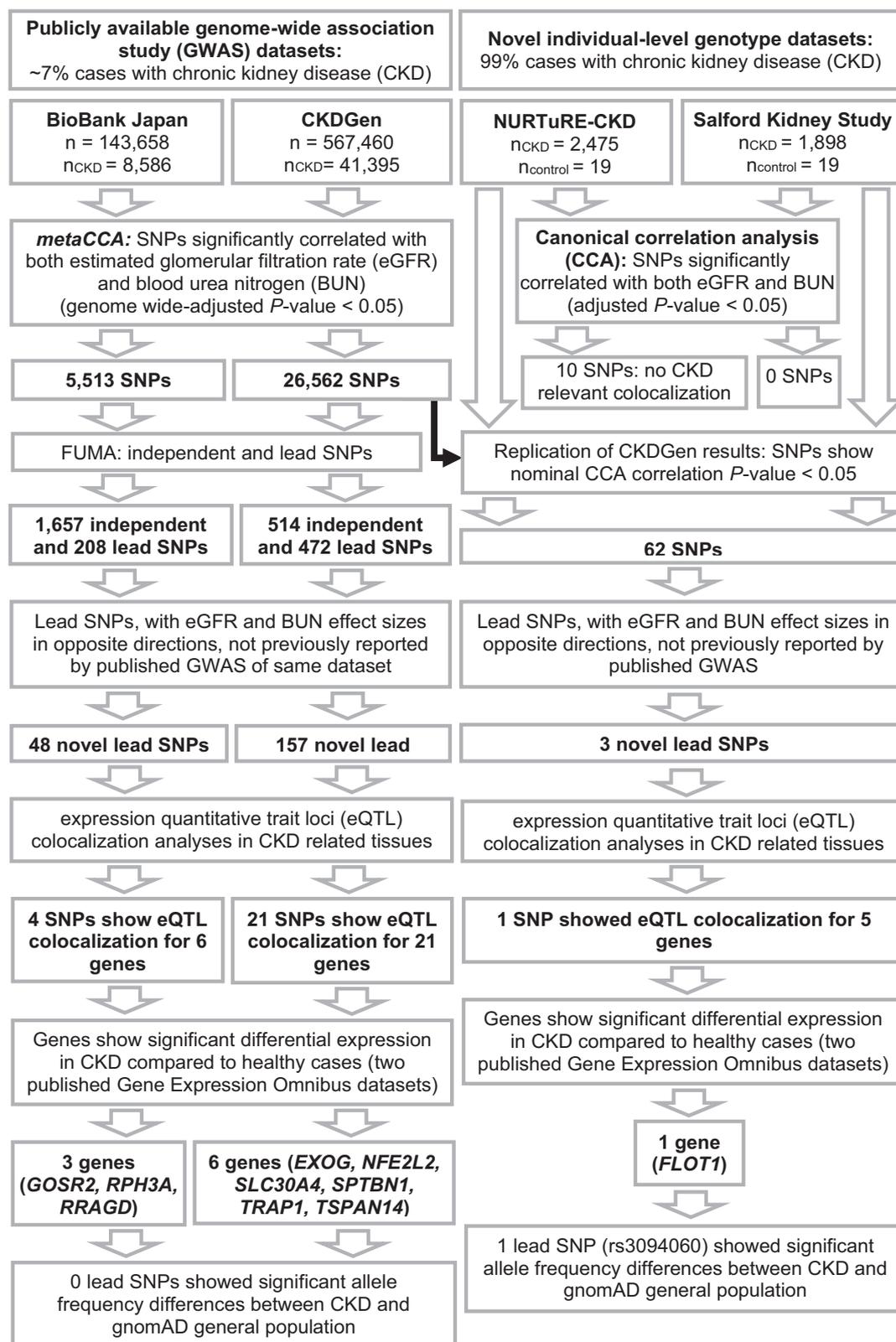


Fig. 1 | Workflow using metaCCA and canonical correlation analysis. Workflow diagram of the analysis method and results obtained for the four datasets.

Previously reported lead SNPs that were described as likely relevant for kidney function were listed in Wuttke et al.¹⁷ for the CKDGen GWAS¹⁷ and Kanai et al.¹⁶ for the BioBank Japan GWAS¹⁶. Of the 122 previously reported lead SNPs in the CKDGen (European ancestry) dataset by Wuttke et al., we found that 113 SNPs (93%) showed a

significant correlation with both eGFR and BUN by metaCCA, of which 78 SNPs (64%) were defined as lead SNPs using FUMA (Supplementary Dataset 3)¹⁷. The nine (7%) SNPs missed by metaCCA appeared to be due to our use of a Bonferroni correction for the joint eGFR and BUN analyses which was more stringent compared to the

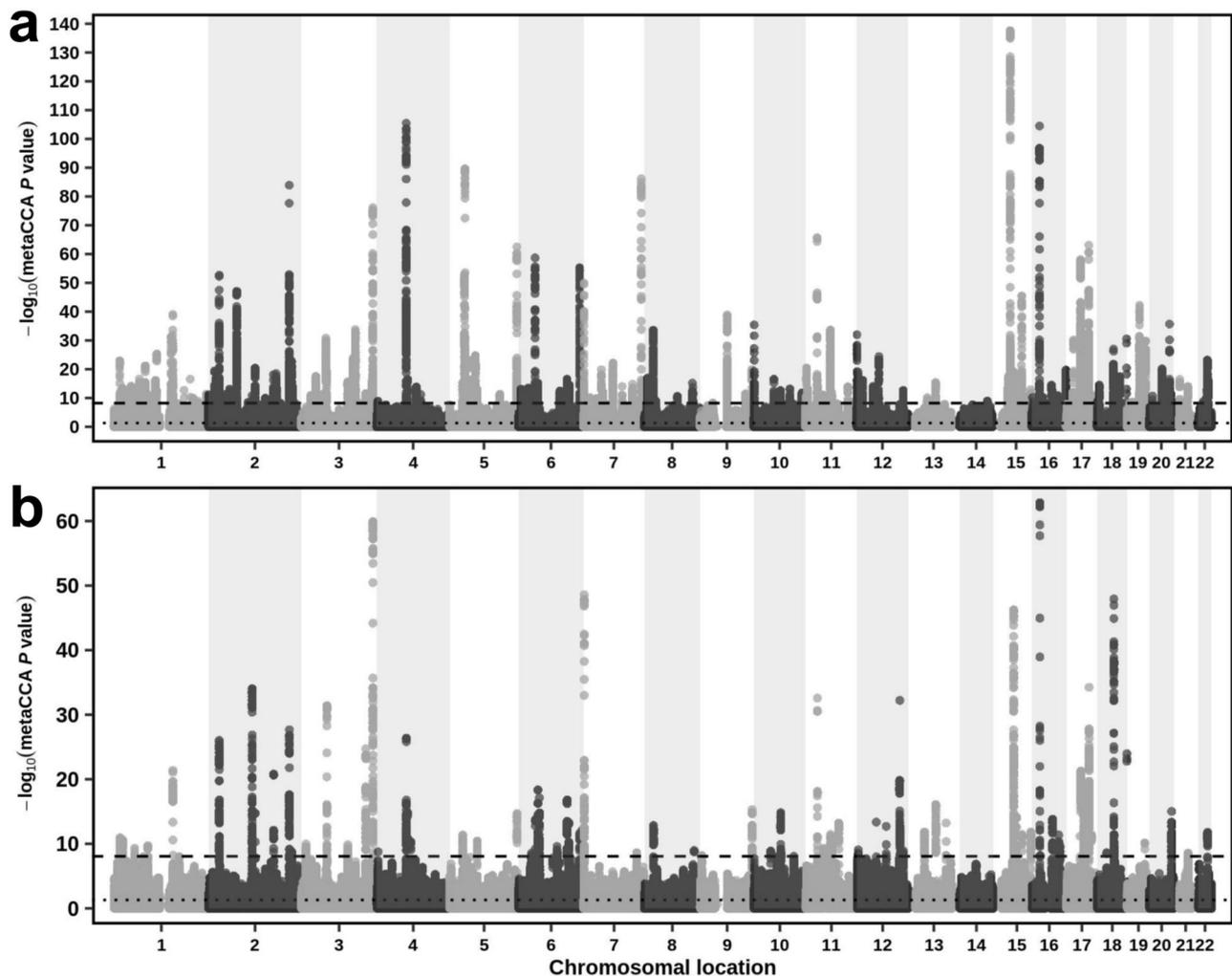


Fig. 2 | Manhattan plots showing *metaCCA* P-value results for single nucleotide polymorphisms in genome-wide association study datasets. Results for univariate single nucleotide polymorphism (SNP) *metaCCA* applied to (a) 8,346,783 CKDGen and (b) 5,837,593 BioBank Japan SNPs, to identify correlations with the two phenotypic variables estimated glomerular filtration rate (eGFR) and blood urea nitrate

(BUN) considered jointly. Each point represents one SNP with the chromosomal number and co-ordinates on the x-axis and corresponding $-\log_{10}$ *metaCCA* P-value on the y-axis. The dashed line denotes the Bonferroni-corrected P-value cut-off of 0.05.

GWAS P-value cut-off of 0.05 (one-sided) used for the univariate BUN analyses by Wuttke et al.¹⁷ (Supplementary Dataset 3). We found an overlap of 8/8 (100%) *metaCCA*-identified SNPs with previously reported lead SNPs by the published BioBank Japan GWAS (Supplementary Dataset 4)¹⁶.

Overlap of additional *metaCCA*-identified SNPs between CKDGen and BioBank Japan

Of the previously unreported 5840 CKDGen and 2471 BioBank Japan SNPs for kidney function identified here by *metaCCA*, 4855 and 2091 SNPs, respectively, were available in both datasets, and of these, there was an overlap of 394 (8% and 19%, respectively) SNPs (Fig. 5A, Supplementary Dataset 5). This overlap of 394 SNPs was significant compared to that expected for the number of SNPs analysed using the hypergeometric distribution (P -value < 0.05 ; Table 2). Of these 394 SNPs, using FUMA, 13 (3%) were defined as lead SNPs for 13 independent genomic loci (Supplementary Dataset 5). The canonical correlation coefficients (r values) for these SNPs were relatively small, as expected, since multiple small effect common genetic variants in aggregate are thought to be required to show a large enough effect on kidney function and/or CKD risk (Supplementary Fig. 1)⁶.

Additional single nucleotide polymorphisms identified by *metaCCA* and CCA

Next, we investigated whether any of the CKDGen *metaCCA*-identified SNPs that showed eGFR and BUN effect sizes in opposite directions (14,045 of 26,562 SNPs) showed replicated kidney function associations in each of the NURTuRE-CKD ($n = 2494$ including 19 healthy participants) and SKS ($n = 1917$ including a different set of 19 healthy participants) individual-level SNP genotype datasets by using CCA (Fig. 1). Of these 14,045 CKDGen *metaCCA*-identified SNPs, 12,711 SNPs were available for analysis in both the NURTuRE-CKD and SKS datasets (Table 2). Of these 12,711 SNPs, 62 (0.5%) SNPs showed nominally significant CCA correlation with both eGFR and BUN in both the NURTuRE-CKD and SKS datasets (P -value < 0.05 ; Table 2, Fig. 5B). This overlap of 62 *metaCCA* and CCA-identified SNPs between the CKDGen (12,711 SNPs), NURTuRE-CKD (272,655) and SKS (268,630) datasets was statistically significant compared to that expected by chance for the number of SNPs analysed by using the hypergeometric distribution (P -value < 0.05 ; Table 2). Of these 62 SNPs (of which rs1398018 and rs9992101 were non-imputed), FUMA analyses showed six lead SNPs which were all imputed SNPs (Supplementary Dataset 6). Of these, three were previously unreported SNP associations for kidney

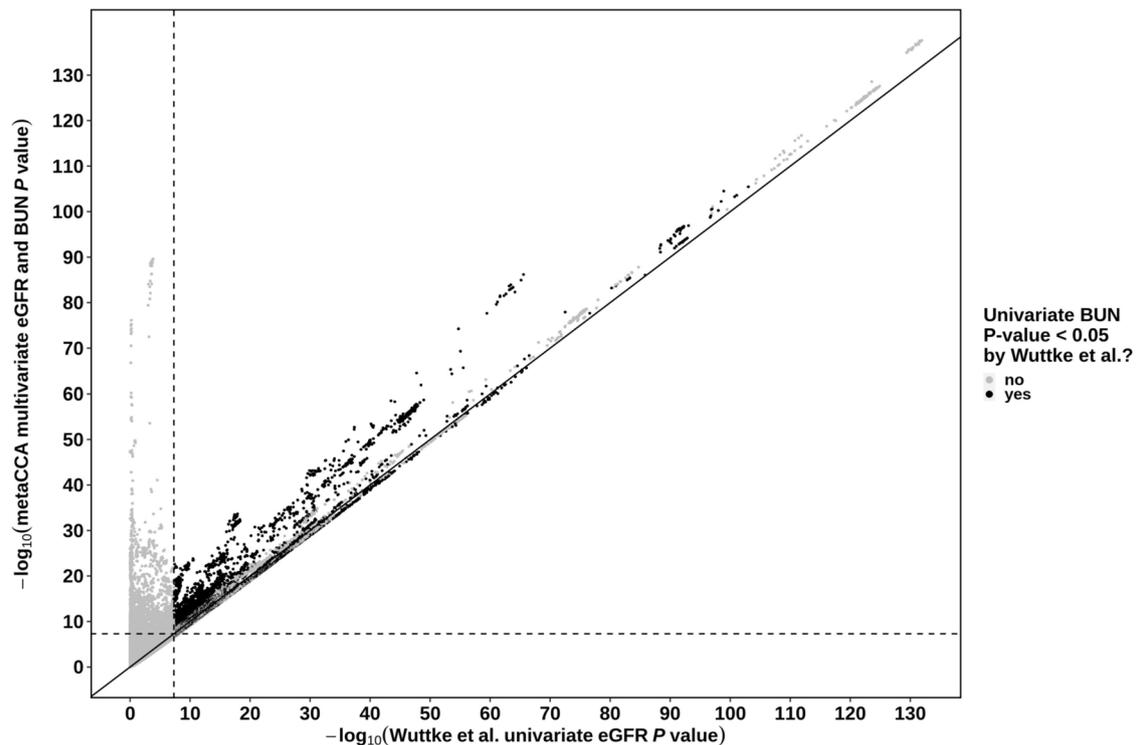


Fig. 3 | Comparison of CKDGen multivariate *metaCCA* *P* values with those previously reported by published univariate analysis. Shown are 8,346,783 CKDGen single-nucleotide polymorphisms (SNPs) we analysed using *metaCCA*, which were previously reported to show a significant association with (i) both estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN) (black

points), or (ii) only eGFR and not BUN (grey points), in the published CKDGen univariate genome-wide association study by Wuttke et al.¹⁷. The horizontal and vertical dashed lines show the genome-wide statistical significance cut-off equivalent to 0.05 for the *metaCCA* multivariate eGFR and BUN test, and the univariate eGFR test by Wuttke et al.¹⁷, respectively.

function based on a lack of any published associations for kidney function in the GWAS Catalog or by Wuttke et al.¹⁷ (Supplementary Dataset 6).

Colocalisation of *metaCCA*, CCA and eQTL signals

For the four sets of previously unreported lead SNPs identified for kidney function by *metaCCA* and CCA (in CKDGen, in BioBank Japan, in both CKDGen and BioBank Japan, and in CKDGen, NURTuRE-CKD and SKS), to determine whether the genomic signals were colocalised with the primary eQTL signals for closest mapped genes, we assessed colocalisation using the Bayesian test “coloc” (Methods)^{33,34}. Colocalisation was defined as a high posterior probability that a single shared variant is responsible for both signals with posterior probability of colocalisation (PP4) ≥ 0.8 ^{33,34}. For eQTL data, we analysed all available datasets in the European Bioinformatics Institute (EBI) eQTL Catalogue and two NephQTL datasets for glomerular and tubular cells^{35,36}.

For the 157 previously unreported lead SNPs we identified using *metaCCA* in the CKDGen dataset, a total of 21 (13%) SNPs showed colocalisation for 21 genes in CKD relevant tissues and cell-types (kidney cortex, glomerular or tubular cells, immune cells, liver or blood) and other tissues (Fig. 6, Supplementary Dataset 7). Of these 21 genes, only *CDK12*, *LINC00243* and *SLC7A9* showed colocalisation in only CKD-related tissues or cell-types (Fig. 6).

For the 28 previously unreported lead SNPs we identified using *metaCCA* in the BioBank Japan dataset, a total of four (14%) SNPs showed colocalisation for six genes in CKD relevant tissues or cell-types (Fig. 6, Supplementary Dataset 8).

Of the 62 SNPs we identified using both *metaCCA* in the CKDGen dataset and CCA in the NURTuRE-CKD and SKS datasets, one (33%) of the three previously unreported lead SNPs showed colocalisation signals for five genes in CKD relevant tissues and cell-types (Fig. 7, Supplementary Dataset

9). Of these five genes, only two (*IER3* and *RPL23AP1*) showed colocalisation in only CKD-related tissues or cell-types (Fig. 7).

For the 13 previously unreported lead (of 394) SNPs we identified using *metaCCA* in both the CKDGen and BioBank Japan datasets, two (15%) SNPs showed colocalisation signals for three genes in CKD relevant tissues and cell-types (Supplementary Dataset 10). Of these, three genes (*C1orf213*, *TCEA3*, *THBS3*) showed colocalisation in only CKD-related tissues or cell-types (Fig. 7).

Overall, for the *metaCCA* or CCA-identified previously unreported lead SNPs, the fraction of these SNPs that showed colocalisation with their closest mapped genes (loci), thus pointing to a shared underlying SNP associated with both kidney function and gene expression, ranged from 13–33%, as described above. This was greater than the fraction of 16 out of 228 (7%) replicated loci that showed colocalisation reported by Wuttke et al.¹⁷ for the CKDGen dataset¹⁷.

Differential gene expression from published chronic kidney disease datasets

Using the Gene Expression Omnibus to R (GEO2R) web-application, CKD case-control differential gene expression profiles were analyzed for two available published datasets (“expression data from uremic patients and 20 healthy controls”, and “development of gene expression profiles in human chronic kidney disease”)^{37–39}. The 21, six, five and three genes that showed colocalisation with *metaCCA*-identified previously unreported lead SNPs in the CKDGen, the BioBank Japan, the CKDGen, NURTuRE-CKD and SKS, and both the CKDGen and BioBank Japan datasets, respectively, were analysed for differential gene expression (Methods). Totals of six (*EXOG*, *NFE2L2*, *SLC30A4*, *SPTBN1*, *TRAP1* and *TSPAN14*), three (*GOSR2*, *RPH3A*, *RRAGD*), one (*FLOT1*), and one (*THBS3*) genes, respectively, showed significant \log_2 (fold change) either equal to or greater than 1, or equal to or less than -1 , between CKD and

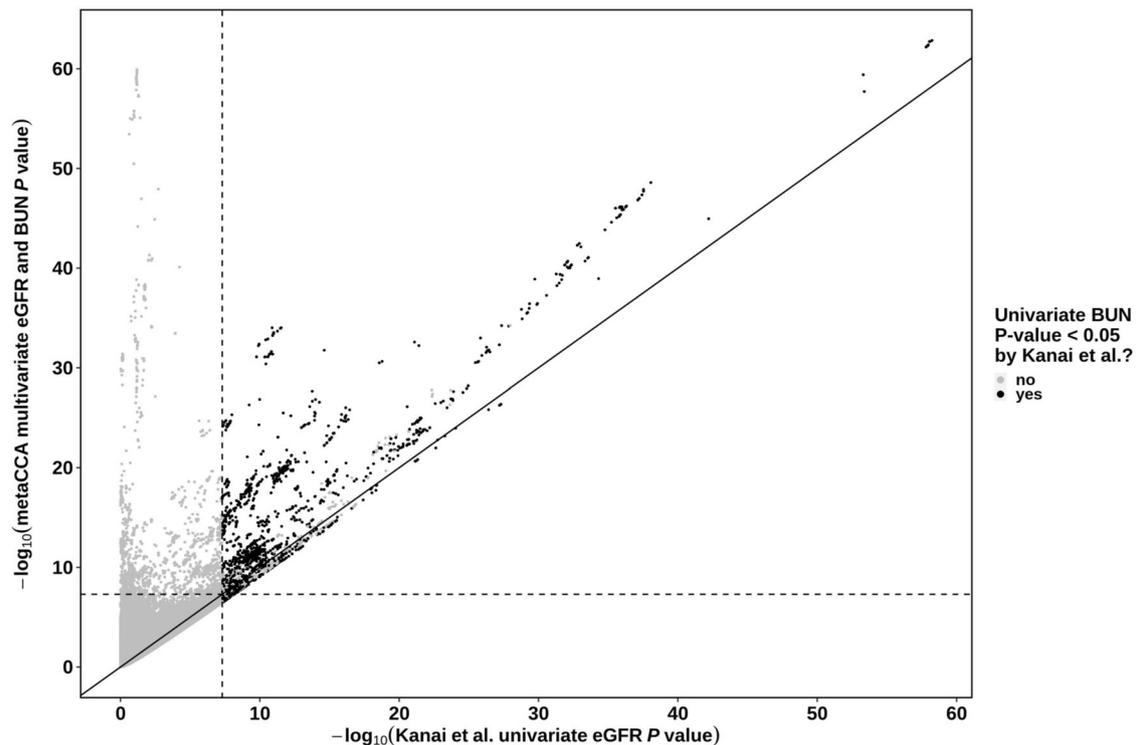


Fig. 4 | Comparison of BioBank Japan multivariate *metaCCA* *P* values with those previously reported by published univariate analysis. Shown are 5,837,593 BioBank Japan single-nucleotide polymorphisms (SNPs) we analysed using *metaCCA*, which were reported to show a significant association with (i) both estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN) (black points), or (ii)

only eGFR and not BUN (grey points), in the published Biobank Japan univariate genome-wide association study by Kanai et al.¹⁶. The horizontal and vertical dashed lines show the genome-wide statistical significance cut-off equivalent to 0.05 for the *metaCCA* multivariate eGFR and BUN test, and the univariate eGFR test by Kanai et al.¹⁶, respectively.

healthy controls in these two published gene expression datasets (adjusted *P*-value < 0.05; Supplementary Datasets 11 and 12). This fold change meant that the expression of the gene was increased or decreased in CKD cases relative to healthy cases by a multiplicative factor of at least 2. The *P*-values were adjusted using the default Benjamini and Hochberg false discovery rate method because it provided a good balance between discovery of statistically significant genes and limitation of false positives³⁹. In summary, we identified 11 previously unreported lead SNPs for multivariate kidney function that showed eQTL colocalisation with 11 genes which also showed differential expression between CKD and healthy individuals (Table 3). These 11 genes included *TRAP1* which is on the Genomics England congenital anomalies of the kidney and urinary tract gene panel (<https://panelapp.genomicsengland.co.uk/panels/234/>). These 11 genes showed significant functional enrichment for several processes related to oxidative-stress induced apoptotic signaling pathway and toxin catabolic processes (Benjamini-Hochberg adjusted *P*-value < 0.05, Supplementary Dataset 13).

Chronic kidney disease allele frequency analyses

For the 11 previously unreported *metaCCA* and/or CCA identified lead SNPs that showed colocalisation with 11 differentially expressed genes in CKD, SNP allele frequencies in CKD cases in each of the NURTuRE-CKD and SKS datasets were computed and compared to gnomAD (general population) allele frequencies (Supplementary Dataset 14)⁴⁰. Of these 11 SNPs, only one SNP (rs3094060), which colocalised with *FLOT1* gene expression, showed a significant difference greater than 15% in allele frequency (AF) between the gnomAD general population (non-Finnish European), and the same type of CKD cases in both the NURTuRE-CKD and SKS datasets (Chi-square test with Yates' correction *P*-value < 0.05; Supplementary Dataset 14)⁴⁰. This was observed for the membranous nephropathy (MN) CKD cases only (Supplementary Dataset 14).

Previously unreported lead missense single nucleotide polymorphisms identified using *metaCCA*

Of the 157 and 28 previously unreported lead SNPs identified in each of CKDGen and BioBank Japan, six and five, respectively, encoded missense variants, of which missense variants in *SLC14A2* were identified in both datasets (Supplementary Dataset 15). Of these two missense SNPs, p.Arg896His identified in CKDGen was predicted to be “pathogenic”, “deleterious” and “probably damaging” by several in silico variant prediction tools (also with a high CADD score) which suggested a negative effect on protein structure and function (Supplementary Dataset 15). However, neither of these *SLC14A2* missense variants showed a difference greater than 15% in allele frequency (AF) between the gnomAD general population (non-Finnish European) and the same CKD cases in both the NURTuRE-CKD and SKS datasets (Supplementary Dataset 14). *SLC14A2* was not previously reported for both eGFR and BUN by the published GWASs^{16,17}.

Discussion

GWASs have identified several common SNPs and loci associated with CKD or kidney function biomarkers^{7–19}. GWAS methods typically measure the association between each SNP and one phenotype using regression methods²². Multiple small effect common genetic variants may contribute to kidney function and/or CKD risk, in a polygenic manner⁶. CCA can identify joint correlations between multiple SNPs and multiple phenotypic variables simultaneously²³. By this, CCA allows the identification of multiple SNPs (gene-gene interactions) and pleiotropic mechanisms, which are thought to be the product of complex genetic diseases²⁷. Earlier studies suggest greater power can be achieved by using CCA or *metaCCA*, leading to novel findings^{23,28}. Multivariate methods such as CCA and *metaCCA* have not previously been applied to genomic datasets with either major and minor CKD subsets to look for additional SNPs associated with two kidney function markers, jointly.

In this study, to identify SNPs associated with multivariate kidney function (both eGFR and BUN), we applied CCA to two individual-level CKD genotype datasets (NURTURE-CKD and SKS) and *metaCCA* to two GWAS summary statistics datasets, CKDGen and BioBank Japan. These CKDGen and BioBank Japan datasets each had a minor subset of CKD cases. For all four datasets, baseline eGFR and BUN measurements or GWAS summary statistics were available. We identified several previously unreported replicated SNPs that showed significant correlation with both eGFR and BUN using CCA and *metaCCA*. We compared our findings with published univariate-GWAS SNPs, assessed replication across datasets, eQTL colocalisation, differential gene expression between CKD cases and healthy individuals, allele frequency analyses, and missense variant effect predictions.

For the two GWAS summary statistics datasets, using univariate-SNP *metaCCA*, we identified many SNPs that showed a significant correlation with both eGFR and BUN jointly. Of the 122 previously reported lead SNPs for both eGFR and BUN in the published CKDGen GWAS (European ancestry)¹⁷, our *metaCCA* results showed a high replication rate of 93%. For

the 7% (nine) missed SNPs, this was likely due to the stricter *P*-value threshold we used for *metaCCA*. For the BioBank Japan dataset, we found 100% overlap between our *metaCCA*-identified SNPs with the eight previously reported lead SNPs by the published BioBank Japan univariate kidney function GWAS¹⁶. Overall, this showed that *metaCCA* had successfully identified previously reported SNPs associated with kidney function in the same datasets.

In the eQTL colocalisation analyses, in addition to kidney cell-types, the immune system was also seen as a CKD relevant cell type to include because some subsets of lymphocytes produce cytokines that can induce or reduce renal inflammation⁴¹. Furthermore, several human leukocyte antigens encoded at the major histocompatibility complex are associated with increased or decreased risk of renal failure⁴², and low levels of some T-cell subsets in peripheral blood were associated with renal outcome in CKD⁴³. We also included the liver as CKD-relevant because some genetic diseases are associated with both kidney and liver disease, e.g. hepatorenal fibrocystic disease⁴⁴.

Across all four *metaCCA*/CCA analyses (the CKDGen, the BioBank Japan, both CKDGen and BioBank Japan, and the CKDGen, NURTURE-CKD and SKS analyses), we identified a total of 11 genes (*EXOGL*, *FLOT1*, *GOSR2*, *NFE2L2*, *RPH3A*, *RRAGD*, *SLC30A4*, *SPTBN1*, *THBS3*, *TRAP1*, *TSPAN14*) that colocalised with the identified 11 previously unreported lead SNPs for multivariate kidney function, in CKD-relevant tissues, and also showed differential expression between CKD and healthy individuals in two published GEO datasets. These 11 genes showed significant functional enrichment for the negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway. Inflammatory cytokines associated with oxidative stress promote the damage of renal tissues by inducing apoptosis, necrosis, and fibrosis and may play an important role in the pathogenesis and progression of CKD^{45,46}. These 11 genes included *TRAP1* (tumor necrosis factor receptor-associated protein 1) which is on the Genomics England congenital anomalies of the kidney and urinary tract gene panel (<https://panelapp.genomicsengland.co.uk/panels/234/>).

Replicated in both the CKDGen and BioBank Japan datasets, we identified a total of 394 previously unreported SNPs for both eGFR and BUN (multivariate kidney function) by using *metaCCA*. Using the hypergeometric statistical test, this SNP overlap between the European and Japanese population datasets was significantly greater than that expected by chance for the numbers of SNPs analysed. Of these 394 SNPs, 13 were lead SNPs for 13 independent genomic loci, of which two SNPs showed colocalisation signals for three genes, in CKD relevant tissues and cell-types only. Of these three genes, the *THBS3* gene also showed significant differential gene expression between CKD and healthy controls in published gene expression datasets. This suggested that SNPs affecting *THBS3* expression were associated with both kidney function and CKD. However, we did not find any significant association of the identified *THBS3* lead SNP (rs2974937) with CKD in our two European ancestry CKD cohorts using AF analyses, compared to the gnomAD non-Finnish European ancestry general population. In the GWAS Catalog, *THBS3* had previously been associated with univariate kidney function by other GWAS by other SNPs, but not both eGFR and BUN in the same study⁴⁷. This could be due to the SNP only affecting a subset of CKD patients, for example tubular

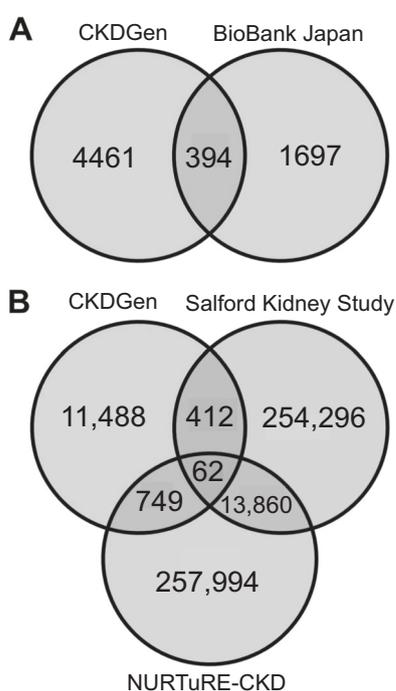


Fig. 5 | Venn diagrams showing *metaCCA* and canonical correlation analysis-identified single nucleotide polymorphism dataset overlaps. The overlaps in single nucleotide polymorphisms (SNPs) identified for both kidney function variables by (A) *metaCCA* for the CKDGen and BioBank Japan datasets, for SNPs not previously reported by the published respective dataset genome-wide association studies (Wuttke et al.¹⁷ for CKDGen and Kanai et al.¹⁶ for BioBank Japan), and (B) *metaCCA* for the CKDGen dataset and canonical correlation analysis for the NURTURE-chronic kidney disease and Salford Kidney Study datasets.

Table 2 | Dataset overlap statistics for single nucleotide polymorphisms identified for multivariate kidney function

Intersections between <i>metaCCA</i> or canonical correlation analysis identified single nucleotide polymorphisms	Total number of single nucleotide polymorphisms	Degree	Observed overlap	Expected overlap	Fold enrichment	<i>P</i> -value
CKDGen (4855) & BioBank Japan (2,091)	5,053,995	2	394	2.897	136.01	2.23E–308
CKDGen (12,718) & NURTURE-CKD (275,787)	5,410,817	2	812	647.917	1.26	9.54E–11
CKDGen (12,955) & Salford Kidney Study (294,406)	5,845,881	2	487	652.430	0.75	1
CKDGen (12,711) & NURTURE-CKD (272,655) & Salford Kidney Study (268,630)	5,355,804	3	62	32.457	1.91	2.50E–6

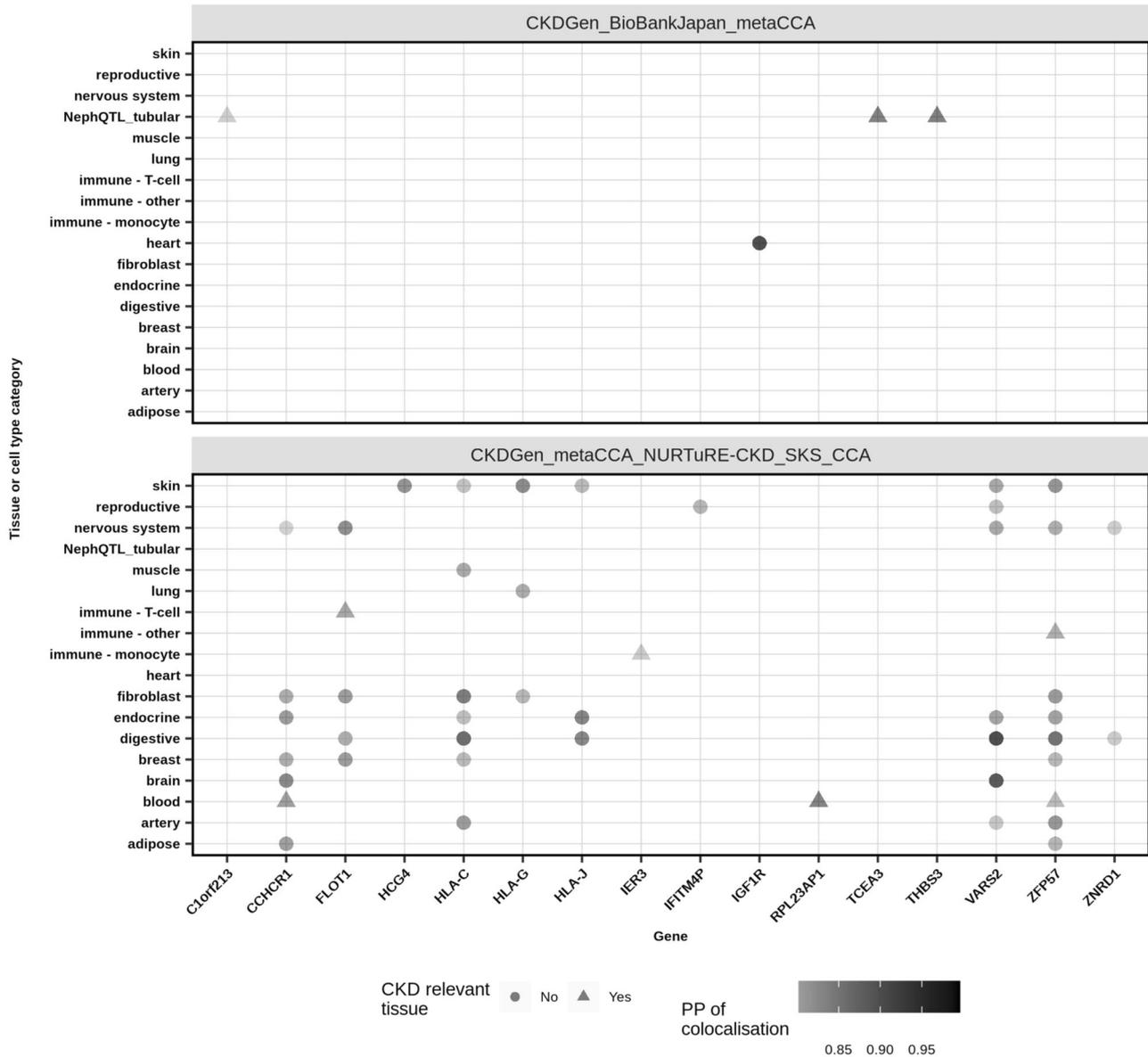


Fig. 7 | Single nucleotide polymorphism expression quantitative trait loci colocalisation analyses for multiple dataset intersections. For the previously unreported lead single nucleotide polymorphisms (SNPs) identified using *metaCCA* in both the CKDGen and BioBank Japan datasets (13 SNPs), and in the CKDGen,

NURTuRE-CKD and SKS datasets (three SNPs), significant colocalisation results using published expression quantitative trait loci datasets, from the European Bioinformatics Institute and NephQTL, are shown by tissue type and gene.

hemodialysis treatment in end stage renal disease patients⁴⁹. These reported potential immune system involvements of *FLOT1* would align with clinical features of membranous nephropathy which is an auto-immune type of kidney disease. In summary, using CCA and *metaCCA*, with eQTL colocalisation and differential gene analyses, we identified a previously unreported kidney function SNP which colocalised with *FLOT1* expression and was significantly more common in MN CKD patients compared to the general population. Overall, this suggested that this SNP may contribute to MN manifestation by affecting *FLOT1* gene expression, possibly by immune system perturbation. *FLOT1* has not previously been associated with kidney function from GWAS in the GWAS Catalog database thus was considered a previously unreported gene association for kidney function, along with the colocalised rs3094060 lead SNP.

For each of the CKDGen and BioBank Japan datasets, although not replicated SNPs, with the addition of orthogonal datasets including eQTL (in CKD-relevant tissues) and CKD gene expression datasets (from the

GEO), we also identified nine previously unreported lead SNPs that showed colocalisation with nine genes. These nine genes (*EXOG*, *NFE2L2*, *SLC30A4*, *SPTBN1*, *TRAP1* and *TSPAN14* in CKDGen, and *GOSR2*, *RPH3A*, and *RRAGD* in BioBank Japan) showed significant differential expression in CKD compared to healthy controls. *RPH3A* lacked any kidney function associations in the GWAS Catalog, thus appeared to be a previously unreported SNP finding for kidney function. *RPH3A* expression was enhanced in several human proteinuric diseases, and was also altered in mouse and human proteinuric disease⁵⁰. Furthermore, combined genomic and metabolomic analyses have previously associated increased urinary albumin excretion in the general population with *RPH3A* polymorphism^{51,52}. However, none of the lead SNPs identified for these nine genes showed AF differences greater than 15% between CKD cases (in each of the NURTuRE-CKD and SKS) and the general population gnomAD dataset. This was despite showing eQTL colocalisation with genes in CKD-relevant tissues which also showed significant differential gene expression between CKD and healthy controls in published gene expression datasets.

Table 3 | Previously unreported lead single nucleotide polymorphisms identified for multivariate kidney function with gene associations

Dataset analysis	Gene symbol	Single nucleotide polymorphism	Gene association method	Published single nucleotide polymorphism univariate GWAS ^a associations with kidney function (PubMed identifier)
CKDGen	<i>EXO2</i>	rs9838792	eQTL ^d colocalisation	eGFR ^e (31152163, 33462484)
CKDGen, NURTURE-CKD ^b and SKS ^c	<i>FLOT1</i>	rs3094060	eQTL colocalisation	none
BioBank Japan	<i>GOSR2</i>	rs3851786	eQTL colocalisation	eGFR (35710981), BUN ^f (31152163, 29403010, 34594039, 36329257)
CKDGen	<i>NFE2L2</i>	rs34468415	eQTL colocalisation	eGFR (31152163, 30604766, 31451708, 33462484, 31015462, 35710981, 34272381, 31451708)
BioBank Japan	<i>RPH3A</i>	rs11614295	eQTL colocalisation	none
BioBank Japan	<i>RRAGD</i>	rs6907843	eQTL colocalisation	eGFR (31152163, 35710981, 34272381), GFR ^g (29403010)
BioBank Japan	<i>SLC14A2</i>	rs1484873, p.Ile132Val	missense variant prediction	eGFR (35710981, 34272381), BUN (31152163, 29403010, 34594039, 36329257, 34272381)
CKDGen	<i>SLC14A2</i>	rs41301139, p.Arg896His	missense variant prediction	eGFR (35710981, 34272381), BUN (31152163, 29403010, 34594039, 36329257, 34272381)
CKDGen	<i>SLC30A4</i>	rs2453531	eQTL colocalisation	eGFR (33462484)
CKDGen	<i>SPTBN1</i>	rs168505	eQTL colocalisation	eGFR (31152163, 30604766, 31152163, 35710981)
CKDGen and Bio-Bank Japan	<i>THBS3</i>	rs2974937	eQTL colocalisation	eGFR (33462484, 31451708), BUN (34594039, 31152163)
CKDGen	<i>TRAP1</i>	rs1635404	eQTL colocalisation	eGFR (31152163, 31451708, 35710981, 34272381)
CKDGen	<i>TSPAN14</i>	rs7087356	eQTL colocalisation	eGFR (31152163, 35710981, 34272381), BUN (34272381)

^aGenome-wide association study.

^bChronic kidney disease.

^cSalford Kidney Study.

^dExpression quantitative trait loci.

^eEstimated glomerular filtration rate.

^fBlood urea nitrogen.

^gGlomerular filtration rate.

Possible reasons for these observations are the same as those described above for the *THBS3* lead SNP.

In addition, we identified a total of 11 previously unreported lead missense SNPs, of which missense variants in *SLC14A2* were seen in both the CKDGen and BioBank Japan dataset *metaCCA* analyses. Of these two *SLC14A2* variants, one (p.Arg896His) was predicted as “likely pathogenic”, “deleterious” and “probably damaging” by several *in silico* variant prediction tools (with a high CADD score). This suggested p.Arg896His may affect kidney function in both European populations by altering *SLC14A2* protein structure and/or function. *SLC14A2* was not previously reported for both eGFR and BUN by the published univariate kidney function GWASs, but has been associated with eGFR and BUN in other studies in the GWAS Catalog^{16,17}. *SLC14A2* is a urea transporter which plays an important role in urine concentration; in knockout mice lacking urea transporters, urine volumes were increased^{53,54}. Volume overload is a risk factor for mortality in CKD and end stage renal disease patients^{55,56}. However, neither of these two identified *SLC14A2* missense SNPs showed AF differences greater than 15% between CKD cases (in each of the NURTURE-CKD and SKS) and the general population gnomAD dataset. Again, possible reasons for this are described above for the *THBS3* lead SNP.

In the NURTURE-CKD dataset, we also identified a further 10 SNPs that showed genome-wide significant CCA correlation with both baseline eGFR and BUN, jointly. However none of these SNPs showed colocalisation with genes in CKD-relevant tissues thus were not investigated further.

In summary, by applying a multivariate statistical approach, CCA, to four independent CKD datasets, we identified both previously reported and unreported SNPs associated with kidney function, beyond those reported by published univariate-trait GWAS methods. We identified a total of 11 previously unreported lead SNPs that showed eQTL colocalisation with 11 genes in CKD relevant tissues, that also showed significant differential expression between CKD and healthy individuals. Of these, two genes (*FLOT1* and *RPH3A*) were previously unreported SNP kidney function

gene associations. Furthermore, the SNP colocalised with *FLOT1* gene expression (rs3094060) showed significant association with MN CKD cases in both NURTURE-CKD and SKS using AF analyses. Overall, by using multivariate analysis by CCA, we identified several previously unreported SNPs and genes for both kidney function and CKD, that can be prioritized for further CKD analyses.

There were some limitations to our study. Firstly, it was possible that identified SNPs were merely correlated with the risk-modifying variants, as with other genome-wide tag-SNP array datasets. However, by adding orthogonal information on the SNPs such as published eQTL SNP data, as well as the FUMA SNP independence tool, this provided additional supportive evidence of likely functional SNPs for further investigation. Secondly, since serum urea is affected by dilution/concentration of plasma (volume status), diet on a day-to-day basis and treatment, it is not as reliable an estimator of kidney function as eGFR. However, since it is complementary to eGFR²⁰, it has value in validation of kidney function SNPs found to be associated with eGFR. Thirdly, the NURTURE-CKD and SKS datasets were considerably smaller than the CKDGen and BioBank Japan datasets, thus affecting power to detect, however power analyses showed it was theoretically possible to detect small effect SNPs using CCA. It was not possible to amalgamate the NURTURE-CKD and SKS datasets with the GWAS summary statistics datasets as they were of different data-types (individual level genotype data and GWAS summary statistics data, respectively). Finally, all three GWAS summary statistics datasets may have also contained SNPs with some missing genotype data beyond the GWAS quality control checks reported in the published studies. These factors may have lead to effect size variability between the datasets.

Methods

NURTURE-CKD and Salford Kidney Study participants

The NURTURE-CKD cohort has linked genotype-phenotype data, and all participants provided written informed consent⁵⁷. The study was approved

by the South Central—Berkshire Research Ethics Committee, abides by the principles of the Declaration of Helsinki and is registered at ClinicalTrials.gov (NCT04084145)⁵⁷. The Salford Kidney Study (SKS) dataset, which has been described previously, received ethical approval from the North West Greater Manchester South Research Ethics Committee (REC15/NW/0818) and written informed consent was obtained from all patients⁵⁸. For the genome-wide genotyping, blood samples were collected, stored as whole blood or centrifuged to separate plasma or serum, aliquoted and frozen at -80°C . Deoxyribonucleic acid was extracted from the frozen whole blood. Of the NURTURE cohort, 2903 CKD and 99 control participants were genotyped⁵⁷. Of the Salford Kidney Study (SKS) cohort, 2409 participants were genotyped⁵⁸. NURTURE-CKD and SKS cohort participants had non-dialysis dependent CKD since they were recruited if their eGFR was below $60\text{ ml/min}/1.73\text{ m}^2$ (all SKS patients) or if urine albumin to creatinine ratio was $>30\text{ mg/mmol}$ in those with eGFR above $60\text{ ml/min}/1.73\text{ m}^2$. End-stage kidney disease and renal replacement therapy were exclusion criteria.

Single nucleotide polymorphism genotype data processing

For each of the NURTURE (including 99 controls) and SKS datasets, SNPs were genotyped using the Illumina Global Screening array v2.0 with additional multi-disease content and 2k custom sequences. After excluding duplicated probes using BCFTools version 1.9 (using htlib 1.9), totals of 671,485 and 672,412 variants remained, respectively⁵⁹. The following steps were carried out using Plink version 2⁶⁰. Variants were excluded if the minor allele frequency (MAF) < 0.01 (162,483 and 170,554 variants excluded, respectively), missing SNP genotype call rate $\geq 1.5\%$ (19,149 and 23,889 SNPs excluded, respectively), Hardy-Weinberg assumptions violated with P -value $< 0.001\%$ (116 and 71 SNPs excluded, respectively) and if they were located on mitochondrial or sex chromosomes (21,133 and 19,426 SNPs excluded, respectively). From the SKS dataset only, a total of 135 overlapping samples with NURTURE-CKD were excluded. Further samples were excluded if they were known first or second-degree relatives of another participant (13 and 0 excluded), showed gender mismatches (11 and 33 excluded), showed $>10\%$ low call rate SNPs (12 and 32 samples excluded, respectively), or showed any cryptic relations using KING cut-off of 0.177 (four and 51 samples excluded, respectively). To avoid potential confounding of results due to different genetic ancestries in the dataset, and to match with the genetic ancestries of the CKDGen dataset, non-European ancestry samples were excluded by using principal component analysis (PCA). This was computed using the The 1000 Genomes Project (1000GP), human genome build 38 (hg38) reference dataset and the plinkQC package in R, by adapting the published R script called “Processing 1000 Genomes reference data for ancestry estimation” on the plinkQC website (<https://meyer-lab-cshl.github.io/plinkQC/articles/AncestryCheck.html>) for hg38 use^{61,62}. After combining the sample and reference datasets and running PCA, the plinkQC “evaluate_check_ancestry” function was used to select individuals of European descent⁶¹. This function uses principal components 1 and 2 to find the center of the known European ancestry reference samples. It then labels study samples as non-European if their Euclidean distance from the center falls outside the radius specified by the maximum Euclidean distance of the reference samples multiplied by a scaling factor. A scaling factor of 2 was chosen as it was the minimum value that produced a radius that included all the 1000GP European ancestry reference samples (Supplementary Figure 2). We identified a small number of 10 NURTURE-CKD samples that self-declared as White British but were outside of the European ancestry reference radius using PCA (Supplementary Fig. 2). To avoid any potential confounding, these 10 were excluded as non-European based on the reference population. Overall, using this strategy, 350 NURTURE-CKD, 19 NURTURE-controls and 108 SKS non-European ancestry samples were excluded (Supplementary Fig. 3). Our adapted script is available here:

https://github.com/AmyJaneOsborne/CCA_scripts/PCA_ancestry_hg38_genotype_datasets.sh⁶¹. No batch effects were seen using PCA. Since CCA cannot handle any missing data, SNPs with any missing data were excluded. Remaining for analysis were 2505 of 2903 NURTURE-CKD, 80 of 99 NURTURE-controls and 2078 of 2409 SKS samples. For the NURTURE-CKD (plus 80 NURTURE-controls) dataset, 468,604 variants remained before SNP imputation. For the SKS dataset plus NURTURE-controls dataset, 458,472 variants remained before SNP imputation. All calculations were performed in a 64-bit Linux conda environment.

Kidney function data

For each participant, the first eGFR and serum urea measurements taken on the same date, either on exactly or on the closest after the cohort recruitment date were used. Of the 2505 European ancestry NURTURE-CKD, 2078 SKS and 80 NURTURE-control samples, totals of 2475, 1898, and 38, respectively, had available eGFR and serum urea data on the same date. eGFR was calculated using the CKD-Epidemiology Collaboration (CKD-EPI) equation (without ethnicity adjustment). For NURTURE-CKD, the ranges of the eGFR and serum urea values were $3.3\text{--}139\text{ ml/min}/1.73\text{ m}^2$ and $1.7\text{--}63.7\text{ mmol/L}$, respectively. For NURTURE-controls, the ranges of the eGFR and urea values were $53\text{--}95\text{ ml/min}/1.73\text{ m}^2$ and $3.5\text{--}8\text{ mmol/L}$, respectively. For SKS, the ranges of the eGFR and urea values were $6\text{--}88\text{ ml/min}/1.73\text{ m}^2$ and $3\text{--}47.7\text{ mmol/L}$, respectively. Urea was converted to BUN by dividing by a factor⁶³. BUN measurements were standardized using the common log transformation and Z-score, and eGFR measurements were standardized using the rank-based inverse normal transformation, as described previously. In the CCA input data matrix, the CKD cases ($n = 2475$ and $n = 1898$, respectively) and controls (a different set of $n = 19$ for each dataset) were each encoded as binary indicator variables, 1 and 0, respectively. Kidney function variables were not adjusted for CKD status since they were correlated with CKD status and controlling for CKD status may have introduced collider bias.

Single nucleotide polymorphism imputation

For the NURTURE and SKS datasets, ungenotyped SNPs were imputed using Beagle version 5.4 with The 1000 Genomes Project hg38 genotype dataset as the reference dataset^{64,65}. The 1000 Genomes Project hg38 Variant Call Format files were downloaded from the 1000 Genomes Project website (<https://www.internationalgenome.org/data-portal/data-collection/grch38>)⁶⁴. Imputation is based on the fact that physically close markers are likely inherited together in a cluster and therefore result in the non-random association of alleles (linkage disequilibrium). Imputation in Beagle is performed by using a Hidden Markov Model to identify the most likely path through the haplotype cluster based on the non-missing genotypes present⁶⁶. The Beagle R^2 accuracy score approximates the squared correlation between the best estimate genotype (i.e. the allele dosage with the highest posterior probability) and the true genotype⁶⁶. This is estimated from the posterior genotype probabilities when the true genotypes are not observed⁶⁶. Unix scripting was used to run Beagle and Plink commands. After imputation, there were 6,419,966 NURTURE-CKD and 6,290,407 SKS variants for analysis (including the NURTURE-controls in each dataset).

Genome-wide association study summary statistics on kidney function

European ancestry CKDGen eGFR and BUN GWAS summary statistics were downloaded from the CKDGen Consortium website on 28/05/2020¹⁷. For eGFR and BUN, there were 567,460 samples including 41,395 CKD cases (7%), as described on their website (<https://ckdgen.imbi.uni-freiburg.de/#Wuttke2019data>)¹⁷.

Published BioBank Japan GWAS summary statistics (6,108,953 SNPs) for eGFR (bbj-a-60) and BUN (bbj-a-11) for 143,658 and 139,818 individuals, respectively, were downloaded from the Medical Research Council Integrative Epidemiology Unit OpenGWAS project website (<https://gwas.mrcieu.ac.uk/datasets/bbj-a-60/>) on 20/04/2021^{16,67}. The East Asian

ancestry dataset contained approximately 8586 CKD cases (5%). The eGFR range was 17.2–132.9 ml/min/1.73 m²^{16,68}.

For each of the three GWAS summary statistics datasets, only SNPs in chromosomes 1–22 were selected since these were the only ones available in the CKDGen dataset. SNPs with both eGFR and BUN statistics available were selected based on matching chromosome, position, reference and effect alleles. For the CKDGen and BioBank Japan datasets, any SNPs with AF < 0.01 were excluded from the *metaCCA* analyses. For each dataset, 8,346,783 and 5,837,593 SNPs remained for analysis, respectively.

Statistical analysis using canonical correlation analysis

CCA was used to compute canonical correlations between each SNP with both eGFR and BUN by using the “cancel” function in R v3.6.0, based on CCA scripts published by Seoane et al, Ferreira et al, and Tang et al.^{23,25,26}. To analyze the NURTURE and SKS datasets, we used our scripts available from Github (https://github.com/AmyJaneOsborne/CCA_scripts). In the CCA matrix, the same 26 controls were used for each of the NURTURE and SKS dataset analyses. The controls were encoded as ‘0’ and the cases as ‘1’, as an additional binary indicator variable.

In CCA, linear combinations of two sets of variables, or views of the same object, X and Y, with the highest correlations are found. This corresponds to finding vectors $\mathbf{a} \in \mathbb{R}^G$ and $\mathbf{b} \in \mathbb{R}^P$ that maximize:

$$r_1 = \frac{(X\mathbf{a}_1)^T(Y\mathbf{b}_1)}{\|X\mathbf{a}_1\| \|Y\mathbf{b}_1\|} = \frac{\mathbf{a}_1^T \sum_{XY} \mathbf{b}_1}{\sqrt{\mathbf{a}_1^T \sum_{XX} \mathbf{a}_1} \sqrt{\mathbf{b}_1^T \sum_{YY} \mathbf{b}_1}} \quad (1)$$

This value r_1 is called (the first) *canonical correlation* between X and Y, and the corresponding vectors \mathbf{a}_1 and \mathbf{b}_1 are called (the first) canonical weights. To identify the variables that provide a large contribution to the observed canonical correlation, the magnitudes of the canonical weights are used. These are found by firstly computing the matrix *K* from the covariance matrices:

$$K = \sum_{XX}^{-1/2} \sum_{XY} \sum_{YY}^{-1/2} \quad (2)$$

Then, a Singular Value Decomposition (SVD) is applied, to decompose the matrix into eigenvectors and eigenvalues. The canonical correlation values can then be computed directly from the matrix *K*, and the canonical weights are determined based on the eigenvectors. To test the significance of each canonical correlation r , which equals the maximum correlation between the variant and the phenotypes, the test statistic was computed based on *Wilks’ Lambda*, as described previously²³. For both CCA and *metaCCA*, for each test, the *N* parameter was set to the total number of samples analysed. Results were visualized using Manhattan plots. For *univariate*-SNP analyses, to account for multiple testing, the *P*-values were adjusted for multiple comparisons using a Bonferroni correction based on the number of SNPs analyzed. A cut-off of adjusted *P*-value < 0.05 was used to determine significance. For NURTURE-CKD and SKS, our approximate power analysis based on univariate analysis suggested it was possible to identify a significant SNP with an effect size (canonical correlation r) of 0.12 or 0.14, respectively, with 80% power (Table 4). The sample size required to identify CCA correlation r of 0.1 with 90% power with two variables has been reported as approximately 1000 samples in Helmer et al.⁶⁹

Statistical analysis using metaCCA

For each of the two GWAS summary statistics datasets, the *metaCcaGp* function provided by the *metaCCA* package in R v3.6.0 was used to compute the canonical correlation between each SNP and both eGFR and BUN^{28,70}. Our scripts for analyzing the two GWAS summary statistics datasets using *metaCCA* are available from Github (https://github.com/AmyJaneOsborne/CCA_scripts)²⁸. Before running *metaCCA*, SNPs were removed if they had a standard error of 0 for the eGFR or BUN beta coefficients, had reference and alternative alleles not containing ‘A’, ‘C’, ‘G’, or ‘T’, or if they were SNPs that were duplicates. A mathematical explanation of *metaCCA* is provided by Cichonska et al.²⁸ After running *metaCCA*, any SNPs that showed a significant *metaCCA* *P*-value were assessed for kidney function relevance based on the published eGFR and BUN GWAS summary statistics. SNPs were excluded if the eGFR and BUN effect sizes (with respect to the same effect allele) were in the same direction, because they were unlikely to be relevant for kidney function as previously described¹⁷.

Single nucleotide polymorphism annotation

SNPs were annotated and prioritized by using the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) program³². FUMA “SNP2GENE” was used to find lead (most significant, independent) SNPs for each locus genomic region, using default r^2 thresholds of 0.6 to define independent significant SNPs and 0.1 to define lead SNPs³². The HUGO Gene Nomenclature Committee (HGNC) online multi-symbol checker was used to verify gene symbols⁷¹. g:Profiler g:GOST and stringDB were used for functional enrichment analyses^{72,73}. The effects of any missense SNPs were predicted by using four in silico variant prediction tools including FATHMM-XF, Combined Annotation Dependent Depletion (CADD), Sorting Intolerant From Tolerant (SIFT) and PolyPhen-2 (run by using the ensembl Variant Effect Predictor online tool)^{74–77}. To test for significant overlap between dataset results, the hypergeometric test provided by the package “hyper” (v1.5.4) via Bioconductor in R (v4.0.2) was used⁷⁸.

Colocalisation of CCA and eQTL signals

We applied Bayesian colocalisation analyses by using the R package ‘coloc’ (cran.r-project.org/web/packages/coloc)^{33,34}. We applied the COLOC function, which uses Approximate Bayes Factor computations, to lead SNPs identified in CKDGen and BioBank Japan by *metaCCA*, and in NURTURE-CKD by CCA, and to overlapping lead SNPs identified across 2 or 3 datasets. We used default priors that a random variant in the region is associated with either kidney function *metaCCA*/CCA or eQTL individually (prior probabilities = 1×10^{-4}), and set the prior probability that the random variant is causal to both kidney function *metaCCA*/CCA and eQTL (prior probability = 1×10^{-6}). As recommended by the authors of the method, we defined the variants as colocalised when the posterior probability of a colocalised signal (PP4) was >0.8. We used all eQTL data available in the EBI eQTL Catalogue database which included Gtex v8 (https://www.ebi.ac.uk/eql/Data_access/). These were downloaded for analysis by adapting an R script available from the EBI public eQTL Catalogue resources (https://github.com/kauralasoo/eQTL-Catalogue-resources/blob/master/tutorials/tabix_use_case.html)³⁵. We followed the coloc example in this script thus included the lead SNP plus surrounding SNPs within 200 kB as input. The single-SNP matrixEQTL results for NephQTL glomerular and tubular cells were downloaded from the NephQTL2 browser (https://hugeampkpn.org/research.html?pageid=nephqtl2_about_118)⁷⁹.

Table 4 | Power analysis for two individual level genotype datasets

Dataset analysis	Participants (n)	Power	Significance level (Bonferroni-corrected)	r
NURTURE-chronic kidney disease (CKD) and NURTURE-controls, <i>univariate</i> -single nucleotide polymorphism	2494 including 19 controls	80%	5E–8	0.12
Salford Kidney Study (SKS) and NURTURE-controls, <i>univariate</i> -single nucleotide polymorphism	1917 including 19 controls	80%	5E–8	0.14

Allele frequency analyses

For any shortlisted SNPs that showed a difference of at least 10% in allele frequency between the gnomAD general population and each of the same CKD groups in NURTURE-CKD and SKS, a Chi-square test with Yates' correction was used to test for significant association of the SNP with CKD cases⁴⁰. This was analyzed using the online GraphPad QuickCalcs tool (<https://www.graphpad.com/quickcalcs/contingency1/>).

Differential gene expression analyses

Using the Gene Expression Omnibus GEO2R application, the differential gene expression between CKD cases and healthy controls was computed for the "GSE66494: development of gene expression profiles in human chronic kidney disease" and "GSE37171: expression data from uremic patients and 20 healthy controls (normals)" datasets^{37,38}. Uremia is the build-up of toxins in blood which occurs when the kidneys stop working and is a sign of CKD. The R scripts generated to reproduce the results are available from Github: https://github.com/AmyJaneOsborne/CCA_scripts. Genes were considered to show significant differential expression between the CKD cases and controls where the $\log_2(\text{fold change}) \geq 1$ or $\log_2(\text{fold change}) \leq -1$ and adjusted *P*-value based on default Benjamini & Hochberg false discovery rate method < 0.05 ⁵⁹.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets analysed during the current study are available in the CKDGen Consortium repository (<http://ckdgen.imbi.uni-freiburg.de/>)⁷, The 1000 Genomes Project Phase 1 genotype repository (<http://www.cog-genomics.org/plink/1.9/resources#1kg>)⁶⁰ and the EBI GWAS Catalog (<https://www.ebi.ac.uk/gwas/studies/>). Genotype-phenotype data access for NURTURE-CKD is available by application to the Kidney Research UK NURTURE Biobank resource (<https://nurturebiobank.org/information-for-researchers/>). Genotype-phenotype data access for the SKS is available by application to the Salford Kidney Study (<https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/salford-kidney-study/>).

Code availability

To analyze the NURTURE-CKD dataset we used our scripts available from Github: https://github.com/AmyJaneOsborne/CCA_scripts, based on scripts used for our previous study²³ using Canonical Correlation Analysis and also downloadable from Github: <https://github.com/jseoane/gaCCA>. Our scripts for analysing the three GWAS summary statistics datasets (see Methods) are available from Github: https://github.com/AmyJaneOsborne/CCA_scripts, and were based on *metaCCA*²⁸.

Received: 4 September 2023; Accepted: 20 February 2024;

Published online: 09 March 2024

References

- GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **395**, 709–733 (2020).
- Canadas-Garre, M. et al. Genetic Susceptibility to Chronic Kidney Disease - Some More Pieces for the Heritability Puzzle. *Front. Genet.* **10**, 453 (2019).
- Yu, Z. et al. Polygenic Risk Scores for Kidney Function and Their Associations with Circulating Proteome, and Incident Kidney Diseases. *J. Am. Soc. Nephrol.* **32**, 3161–3173 (2021).
- Dossetor, J. B. Creatininemia versus uremia. The relative significance of blood urea nitrogen and serum creatinine concentrations in azotemia. *Ann. Intern. Med.* **65**, 1287–1299, (1966).
- Fox, C. S. et al. Genomewide linkage analysis to serum creatinine, GFR, and creatinine clearance in a community-based population: the Framingham Heart Study. *J. Am. Soc. Nephrol.* **15**, 2457–2461 (2004).
- Tin, A. & Kottgen, A. Genome-Wide Association Studies of CKD and Related Traits. *Clin. J. Am. Soc. Nephrol.* **15**, 1643–1656 (2020).
- Kottgen, A. et al. Multiple loci associated with indices of renal function and chronic kidney disease. *Nat. Genet.* **41**, 712–717 (2009).
- Kottgen, A. et al. New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* **42**, 376–384 (2010).
- Chambers, J. C. et al. Genetic loci influencing kidney function and chronic kidney disease. *Nat. Genet.* **42**, 373–375 (2010).
- Pattaro, C. et al. Genome-wide association and functional follow-up reveals new loci for kidney function. *PLoS Genet.* **8**, e1002584 (2012).
- Okada, Y. et al. Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* **44**, 904–909 (2012).
- Pattaro, C. et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* **7**, 10023 (2016).
- Mahajan, A. et al. Trans-ethnic Fine Mapping Highlights Kidney-Function Genes Linked to Salt Sensitivity. *Am. J. Hum. Genet.* **99**, 636–646 (2016).
- Gorski, M. et al. 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. *Sci. Rep.* **7**, 45040 (2017).
- Li, M. et al. SOS2 and ACP1 Loci Identified through Large-Scale Exome Chip Analysis Regulate Kidney Development and Function. *J. Am. Soc. Nephrol.* **28**, 981–994 (2017).
- Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
- Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
- Hellwege, J. N. et al. Mapping eGFR loci to the renal transcriptome and phenotype in the VA Million Veteran Program. *Nat. Commun.* **10**, 3842 (2019).
- Stanzick, K. J. et al. Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat. Commun.* **12**, 4350 (2021).
- Thio, C. H. L. et al. Genome-Wide Association Scan of Serum Urea in European Populations Identifies Two Novel Loci. *Am. J. Nephrol.* **49**, 193–202 (2019).
- Sheng, X. et al. Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nat. Genet.* **53**, 1322–1333 (2021).
- Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 59 (2021).
- Seoane, J. A., Campbell, C., Day, I. N., Casas, J. P. & Gaunt, T. R. Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS Comput. Biol.* **10**, e1003876 (2014).
- Hotelling, H. Relations between two sets of variates*. *Biometrika* **28**, 321–377 (1936).
- Ferreira, M. A. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25**, 132–133 (2009).
- Tang, C. S. & Ferreira, M. A. A gene-based test of association using canonical correlation analysis. *Bioinformatics* **28**, 845–850 (2012).
- Larson, N. B. et al. Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *Eur. J. Hum. Genet.* **22**, 126–131 (2014).
- Cichonska, A. et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* **32**, 1981–1989 (2016).
- Jia, X. et al. Multivariate analysis of genome-wide data to identify potential pleiotropic genes for type 2 diabetes, obesity and coronary artery disease using MetaCCA. *Int J. Cardiol.* **283**, 144–150 (2019).

30. Wang, Z. et al. Identification of pleiotropic genes between risk factors of stroke by multivariate metaCCA analysis. *Mol. Genet. Genom.* **295**, 1173–1185 (2020).
31. Li, H. et al. Exploring the Pleiotropic Genes and Therapeutic Targets Associated with Heart Failure and Chronic Kidney Disease by Integrating metaCCA and SGLT2 Inhibitors' Target Prediction. *Biomed. Res. Int.* **2021**, 4229194 (2021).
32. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
33. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
34. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in localisation analyses. *PLoS Genet.* **16**, e1008720 (2020).
35. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
36. Gillies, C. E. et al. An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. *Am. J. Hum. Genet.* **103**, 232–244 (2018).
37. Scherer, A. et al. Alteration of human blood cell transcriptome in uremia. *BMC Med. Genom.* **6**, 23 (2013).
38. Nakagawa, S. et al. Molecular Markers of Tubulointerstitial Fibrosis and Tubular Cell Damage in Patients with Chronic Kidney Disease. *PLoS One* **10**, e0136994 (2015).
39. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
40. Chen, S. et al. A genomic mutational constraint map using variation in 76156 human genomes. *Nature.* **625**, 92–100 (2024).
41. Cao, C., Yao, Y. & Zeng, R. Lymphocytes: Versatile Participants in Acute Kidney Injury and Progression to Chronic Kidney Disease. *Front. Physiol.* **12**, 729084 (2021).
42. Robson, K. J., Ooi, J. D., Holdsworth, S. R., Rossjohn, J. & Kitching, A. R. HLA and kidney disease: from associations to mechanisms. *Nat. Rev. Nephrol.* **14**, 636–655 (2018).
43. Xiong, J. et al. T-Lymphocyte Subsets Alteration, Infection and Renal Outcome in Advanced Chronic Kidney Disease. *Front. Med.* **8**, 742419 (2021).
44. Park, E. et al. Hepatorenal fibrocystic diseases in children. *Pediatr. Nephrol.* **31**, 113–119 (2016).
45. Greiber, S., Muller, B., Daemisch, P. & Pavenstadt, H. Reactive oxygen species alter gene expression in podocytes: induction of granulocyte macrophage-colony-stimulating factor. *J. Am. Soc. Nephrol.* **13**, 86–95 (2002).
46. Gyuraszova, M., Gurecka, R., Babickova, J. & Tothova, L. Oxidative Stress in the Pathophysiology of Kidney Disease: Implications for Noninvasive Monitoring and Identification of Biomarkers. *Oxid. Med. Cell Longev.* **2020**, 5478708 (2020).
47. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
48. Zhan, Z., Ye, M. & Jin, X. The roles of FLOT1 in human diseases (Review). *Mol. Med. Rep.* **28**, 212 (2023).
49. Dai, H., Zhou, J. & Zhu, B. Gene co-expression network analysis identifies the hub genes associated with immune functions for nocturnal hemodialysis in patients with end-stage renal disease. *Medicine* **97**, e12018 (2018).
50. Rastaldi, M. P. et al. Glomerular podocytes possess the synaptic vesicle molecule Rab3A and its specific effector rabphilin-3a. *Am. J. Pathol.* **163**, 889–899 (2003).
51. Marrachelli, V. G. et al. Genomic and metabolomic profile associated to microalbuminuria. *PLoS One* **9**, e98227 (2014).
52. Hwang, S. J., Yang, Q., Meigs, J. B., Pearce, E. N. & Fox, C. S. A genome-wide association for kidney function and endocrine-related traits in the NHLBI's Framingham Heart Study. *BMC Med. Genet.* **8**(1), S10 (2007).
53. Fenton, R. A., Chou, C. L., Stewart, G. S., Smith, C. P. & Knepper, M. A. Urinary concentrating defect in mice with selective deletion of phloretin-sensitive urea transporters in the renal collecting duct. *Proc. Natl Acad. Sci. USA* **101**, 7469–7474 (2004).
54. Fenton, R. A. et al. Renal phenotype of UT-A urea transporter knockout mice. *J. Am. Soc. Nephrol.* **16**, 1583–1592 (2005).
55. Zoccali, C. et al. Chronic Fluid Overload and Mortality in ESRD. *J. Am. Soc. Nephrol.* **28**, 2491–2497 (2017).
56. Kuma, A. et al. Inhibition of urea transporter ameliorates uremic cardiomyopathy in chronic kidney disease. *FASEB J.* **34**, 8296–8309 (2020).
57. Taal, M. W. et al. Associations with age and glomerular filtration rate in a referred population with chronic kidney disease: Methods and baseline data from a UK multicentre cohort study (NURTuRE-CKD). *Nephrol. Dial. Transplant.* **38**, 2617–2626 (2023).
58. Ali, I., Donne, R. L. & Kalra, P. A. A validation study of the kidney failure risk equation in advanced chronic kidney disease according to disease aetiology with evaluation of discrimination, calibration and clinical utility. *BMC Nephrol.* **22**, 194 (2021).
59. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience.* **10**, giab008 (2021).
60. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
61. Meyer, H. V. plinkQC: Genotype quality control in genetic association studies. meyer-lab-cshl/plinkQC: plinkQC 0.3.2 (v0.3.2). *Zenodo* (2020) <https://doi.org/10.5281/zenodo.3934294>.
62. Meyer, H. V. et al. Genetic and functional insights into the fractal structure of the heart. *Nature* **584**, 589–594 (2020).
63. Hosten, A. O. In *Clinical Methods: The History, Physical, and Laboratory Examinations* (eds. H. K. Walker, W. D. Hall, & J. W. Hurst) (Butterworths Copyright © 1990, Butterworth Publishers, a division of Reed Publishing, 1990).
64. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
65. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
66. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
67. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human genome. *Elife* **7**, e34408 (2018).
68. Zheng, J. et al. Trans-ethnic Mendelian-randomization study reveals causal relationships between cardiometabolic factors and chronic kidney disease. *Int J. Epidemiol.* **50**, 1995–2010 (2022).
69. Helmer, M. et al. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Commun Biol.* **7**, 217 (2024).
70. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2021) <https://www.R-project.org/>.
71. Braschi, B. et al. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **47**, D786–D792 (2019).
72. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
73. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
74. Rogers, M. F. et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).

75. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
76. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
77. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
78. Federico, A. & Monti, S. hypeR: an R package for geneset enrichment workflows. *Bioinformatics* **36**, 1307–1308 (2020).
79. Han, S. K. et al. Mapping genomic regulation of kidney disease and traits through high-resolution and interpretable eQTLs. *Nat. Commun.* **14**, 2229 (2023).

Acknowledgements

We acknowledge support via the Medical Research Council (MRC) award (MR/R013942/1). A.B. was funded by a Kidney Research UK non-clinical post-doctoral Fellowship.

Author contributions

A.J.O. wrote the programs to prepare and analyze the data, and analyzed the data, prepared the figures and wrote the manuscript, A.J.O. and C.C. designed the analyses and C.C. wrote part of the background and methods sections, M.A.S. wrote parts of the discussion, E.C., G.I.W., M.A.S., M.W.T. and P.K. managed the collection of the NURTuRE and SKS participant clinical and genetic datasets. A.B., G.I.W., M.A.S., M.W.T., C.C. and P.K. oversaw the project. Evotec (U.A., O.R. and P.S.) provided called genotype SNP array data for the NURTuRE-CKD and SKS cohorts. All authors commented on and provided improvements to the paper.

Competing interests

M.W.T. reports consulting fees from Boehringer Ingelheim, honoraria from Bayer and support to attend conferences from Bayer and a leadership role in

the International Society of Nephrology. No competing interests to declare from the other authors.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41540-024-00350-8>.

Correspondence and requests for materials should be addressed to Amy J. Osborne or Colin Campbell.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024