PLOS Genetics

RESEARCH ARTICLE

New composite phenotypes enhance chronic kidney disease classification and genetic associations

Kim Ngan Tran[1], **Heidi G. Sutherland**[1], **Andrew J. Mallett**[2,3,4], **Lyn R. Griffiths**[1], **Rodney A. Lea**[1]*

1 Centre for Genomics and Personalised Health, Queensland University of Technology, Kelvin Grove, Queensland, Australia, **2** Institute for Molecular Bioscience & Faculty of Medicine, The University of Queensland, St Lucia, Queensland, Australia, **3** Department of Renal Medicine, Townsville University Hospital, Townsville, Queensland, Australia, **4** College of Medicine & Dentistry, James Cook University, Townsville, Queensland, Australia

* rodney.lea@qut.edu.au

OPEN ACCESS

Citation: Tran KN, Sutherland HG, Mallett AJ, Griffiths LR, Lea R (2025) New composite phenotypes enhance chronic kidney disease classification and genetic associations. PLoS Genet 21(5): e1011718. https://doi.org/10.1371/journal.pgen.1011718

Editor: David A. Buchner, Case Western Reserve University, UNITED STATES OF AMERICA

Received: December 2, 2024

Accepted: May 9, 2025

Published: May 23, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1011718

Abstract

Chronic kidney disease (CKD) is a multifactorial condition driven by diverse etiologies that lead to a gradual loss of kidney function. Although genome-wide association studies (GWAS) have identified numerous genetic loci linked to CKD, a large portion of its genetic basis remains unexplained. This knowledge gap may partly arise from the reliance on single biomarkers, such as estimated glomerular filtration rate (eGFR), to assess kidney function. To address this limitation, we developed and applied a novel multi-phenotype approach, combinatorial Principal Component Analysis (cPCA), to better understand the complex genetic architecture of CKD. Using UK Biobank dataset (n = 337,112), we analyzed 21 CKD-related phenotypes, generating over 2 million composite phenotypes (CPs) through cPCA. Nearly 50,000 of these CPs demonstrated significantly higher classification power for clinical CKD compared to individual biomarkers. The top-ranked CP—a combination of albumin, cystatin C, eGFR, gamma-glutamyltransferase, HbA1c, low-density lipoprotein, and microalbuminuria, achieved an AUC of 0.878 (95% CI: 0.873–0.882), significantly outperforming eGFR alone (AUC: 0.830, 95% CI: 0.825–0.835). Genetic association analysis of the ~50,000 high-performing CPs identified all major eGFR-associated loci, except for the *SH2B3* locus rs3184504, a loss-of-function variant, which was uniquely identified in CPs ($p = 3.1_{\times}10^{-56}$) but not in eGFR within the same sample size. In addition, *SH2B3* locus showed strong evidence of colocalization with eGFR, supporting its role in kidney function. These results highlight the power of the multi-phenotype cPCA approach in understanding the genetic basis of CKD, with potential applications to other complex diseases.

Data availability statement: The summary statistics are publicly available in Figshare repository at https://doi.org/10.6084/m9.figshare.26122540.v1. All the scripts used in the study are available in the S2 Text.

Funding: This work was supported by Queensland University of Technology (to KNT) and the Queensland Health (to AJM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Chronic kidney disease (CKD) can result from diverse underlying causes, such as diabetes, high blood pressure, infections, and lifestyle factors. However, most CKD studies rely on single measurements, such as estimated glomerular filtration rate (eGFR), which assesses kidney filtration but may not fully capture the complexity of the disease. Here, we applied a novel approach to explore CKD from a broader perspective. Using a large dataset of over 300,000 individuals, we combined 21 kidney-related health measures into millions of new composite traits, providing a more comprehensive view of kidney function. One of these composite traits resulted from a combination of albumin, cystatin C, eGFR, gamma-glutamyltransferase, HbA1c, low-density lipoprotein, and microalbuminuria, proved to be significantly more effective at identifying CKD than any single measurement. Additionally, we identified key genetic factors associated with CKD, including the *SH2B3* gene. By integrating multiple measurements, our work offers a clearer understanding of the genetic basis of CKD and paves the way for similar approaches to unravel other complex diseases, ultimately aiding in their prevention and treatment.

Introduction

Chronic kidney disease (CKD) is a collective term encompassing a range of heterogeneous diseases characterized by persistent structural or functional kidney abnormalities. CKD is stratified into five stages, culminating in kidney failure, which necessitates consideration of interventions such as kidney transplantation or dialysis. This condition has a high prevalence, affecting approximately 10–15% of the global population, resulting in significant burden on both public health and the economy [1].

Genome-wide association studies (GWAS) investigating CKD have traditionally focused on evaluating kidney function using single biomarkers, such as estimated glomerular filtration rate (eGFR), microalbuminuria, or blood urea nitrogen [2–5]. For example, a robust GWAS analysis of eGFR in a cohort of over 1.2 million individuals identified 634 independent genetic signals, collectively accounting for 9.8% of the eGFR variance [4]. However, a portion of the heritability of CKD remains unexplained. This gap in understanding can be attributed, in part, to the fact that eGFR and other individual biomarkers do not fully capture the underlying causes of CKD nor accurately predict an individual's risk of CKD or progression to kidney failure [6]. For a comprehensive diagnosis and prognosis of systemic CKD, it is recommended to employ a combination of various markers that collectively reflect the diverse alterations occurring over the course of CKD development [7].

Previously, we employed principal component analysis (PCA) on multiple quantitative phenotypes associated with CKD, uncovering a novel susceptibility gene for kidney function that remained undetected in single-phenotype GWASs [8]. In this study, we introduce and implement a new multi-phenotype approach termed combinatorial PCA to further investigate the genetic basis of CKD within the UK Biobank dataset.

Results

In this study, our objective was to identify novel genetic loci associated with CKD through a comprehensive multi-phenotype analysis. We conducted our analyses on the White-British group within the UK Biobank (UKB) prospective cohort study (n = 337,112) to identify composite phenotypes (CPs) with higher performance in CKD classification compared to individual CKD-related biomarkers. Subsequently, we performed GWAS on the identified CPs and replicated the results in the Irish cohort, also within the UKB (n = 11,106).

Best single-markers for CKD classification

Prior to conducting the multi-phenotype analysis, we examined the 21 phenotypes previously linked to CKD (Table 1) in terms of their performance in classifying clinical CKD. This was evaluated by the area under the curve (AUC) of receiver operating characteristic (ROC) curves using the ICD codes for CKD as clinical outcomes. The biomarkers encompassed a range of physiological indicators of CKD risk, including markers of renal function, metabolic parameters, inflammation, lipid profile, and blood pressure. eGFR and CYSC both had the highest discriminatory power among the biomarkers, exhibited by the highest AUCs ranging from 0.825 to 0.842. Although the AUC value of CYSC was slighly higher than that of eGFR, comparing the 2 ROC curves showed no statistically significant difference ($p = 0.256$). Other biomarkers, such as blood urea nitrogen (BUN), uric acid (UA), and glycated hemoglobin (HbA1c), demonstrated moderate discriminatory performance, with AUCs ranging from 0.658 to 0.742. Conversely, other biomarkers such as vitamin D (VITD), calcium (CALC), and diastolic blood pressure (DBP) exhibited low AUC values, ranging from 0.489 to 0.525.

Composite phenotypes perform better than single markers in CKD classification

In this multi-phenotype analysis, we developed and applied a method called combinatorial PCA (cPCA) to identify combinations of biomarkers that outperformed single markers. General steps of the cPCA method are illustrated in Fig 1A. Through cPCA application, a total of 2,097,130 composite phenotypes (CPs) were extracted from all unique combinations of 21 CKD-related biomarkers, and were subsequently evaluated for performance in CKD classification. As a result, we identified 49,734 CPs with significantly better disease classification compared to eGFR as a single biomarker ($p < 2.5e_{\times}10^{-8}$, as adjusted for 2 million independent tests), as assessed using ROC curves and AUC. The top ten CPs with the highest performance are listed in Table 2.

We analyzed the phenotypic components of the 49,734 CPs that exhibited statistically significantly better performance in CKD classification compared to eGFR (Fig 1B, 1C and 1D). The top ranked CP was represented by albumin (ALB), CYSC, eGFR, gamma glutamyl-transferase (GGT), HbA1c, low density lipoprotein (LDL), and microalbuminuria (MA) (AUC = 0.878, 95%CI = 0.873-0.882). Among the other combinations, CYSC and eGFR were consistently present, with HbA1c appearing in nearly all instances. Other notable phenotypes included MA, ALB, and LDL, with appearances ranging from 75% to 55% across the combinations. Regarding pairs or triples of phenotypes, as expected, the most frequent combinations included CYSC, eGFR, and HbA1c: CYSC-eGFR pairs were present in all combinations, while CYSC-HbA1c, eGFR-HbA1c, and CYSC-eGFR-HbA1c were found in 99% of combinations.

We analyzed the relationship between the CP extracted from {eGFR, CYSC, ALB, HbA1c, GGT, LDL, and MA} and CKD using a logistic regression model: CKD status ~ CP + age + sex. In this model, the odds ratio (OR) for CP was 0.419 (95% CI: 0.413 – 0.425), indicating that each one-unit increase in CP corresponds to a 58.1% decrease in CKD odds. Fig 2 presents the PCA biplot, while the loadings of each biomarker contributing to the CP are detailed in S1 Table.

Genetic associations of the identified CPs

The cPCA analysis identified a total of 49,734 CPs with significantly higher AUCs than that of eGFR. Although each of the CPs might have distinct underlying genetic bases due to their constituent biomarkers, they all shared a common

Table 1. Twenty one kidney function related phenotypes selected from the UK Biobank dataset.

No.	Phenotypes	Abbr.	Relation to kidney function	AUC	95% CI
1	Cystatin C	CYSC	CYSC levels associated with kidney function [9].	0.837	0.832-0.842
2	Creatinine-based eGFR (CKD-EPI Creatinine Equation - 2021)	eGFR	Marker of kidney function [10].	0.830	0.825-0.835
3	Urea	UREA	Higher UREA levels associated with adverse renal outcomes [11].	0.735	0.729-0.742
4	Urate	UA	Higher UA associated with new and progressive CKD [12].	0.681	0.675-0.687
5	Glycated haemoglobin	HbA1c	Higher HbA1c associated with increased risk of CKD or CVD [13].	0.664	0.658-0.67
6	C-reactive protein	CRP	Higher CRP associated with CKD incidence [14].	0.634	0.628-0.64
7	Body mass index	BMI	Higher BMI associated with increased risk of CKD [15–17].	0.631	0.625-0.637
8	LDL direct	LDL	Higher LDL associated with increased risk of CVD in non-dialysis CKD patients [18].	0.628	0.621-0.635
9	HDL cholesterol	HDL	Both low and high HDL associated with adverse outcomes in patients with CKD [19].	0.628	0.621-0.634
10	Apolipoprotein B	APOB	Higher APOB associated with lower eGFR, increased ESRD risk [20–22].	0.602	0.595-0.609
11	Apolipoprotein A	APOA1	Higher APOA1 associated with lower CKD prevalence [23].	0.600	0.593-0.606
12	Albumin	ALB	Associated with reduced kidney functions in HIV-infected individuals and elders [24,25].	0.590	0.584-0.597
13	Triglycerides	TRIG	Associated with CKD stages [26].	0.588	0.582-0.594
14	Gamma glutamyltransferase	GGT	Higher GGT associated with increased risk of ESRD [27].	0.582	0.576-0.589
15	Systolic blood pressure	SBP	Lower SBP associated with ESRD and increased mortality in CKD patients [28].	0.572	0.566-0.579
16	Microalbuminuria	MA	Biomarker for kidney injury [29].	0.564	0.56-0.568
17	Haematocrit percentage	HCT	Lower HCT associated with declined kidney function and increased risk of ESRD [30,31].	0.549	0.542-0.556
18	Phosphate	PHOS	High PHOS associated with increased CVD risk and mortality in patients with or without CKD [32].	0.521	0.515-0.528
19	Vitamin D	VITD	Lower VITD associated with adverse outcomes and mortality in CKD patients [33].	0.518	0.512-0.525
20	Calcium	CALC	Lower CALC associated with rapid CKD progression [34].	0.505	0.498-0.512
21	Diastolic blood pressure	DBP	Lower DBP associated with increased mortality in CKD patients [28,35].	0.496	0.489-0.503

ESRD: end-stage renal disease.

https://doi.org/10.1371/journal.pgen.1011718.t001

kidney-function-related genetic component, owing to their superior discriminatory power in CKD classification compared to eGFR. To identify the shared genetic loci associated with kidney function across these CPs, we looked for loci that consistently reached genome-wide significance ($p < 5_{\times}10^{-8}$) across all 49,734 CPs.

This analysis yielded 80 loci consistently identified in all the identified CPs ($p = 5_{\times}10^{-8}$) and as shown in Fig 3. most of these loci were also observed in our eGFR GWAS. However, 5 loci – *CST3, SH2B3*, *FTO*, *SEMA3F-AS1*, and *AC128707.1* - were not identified in the eGFR GWAS and were instead found in GWASs of other individual phenotypes. *SH2B3* was found in 12 out of the 21 individual-phenotype GWASs, *FTO* was found in 7 and *SEMA3F-AS*1 in 5, and *AC128707.1* was found to be genome-wide significant in 2 GWASs, and *CST3* was only found to be genome-wide significant in the GWAS of CYSC.

Among these loci, 75 were found to overlap with those identified in the GWAS of eGFR, while 5 loci were not. Instead, these 5 loci were discovered in GWASs of other individual phenotypes, each represented by a distinct color.

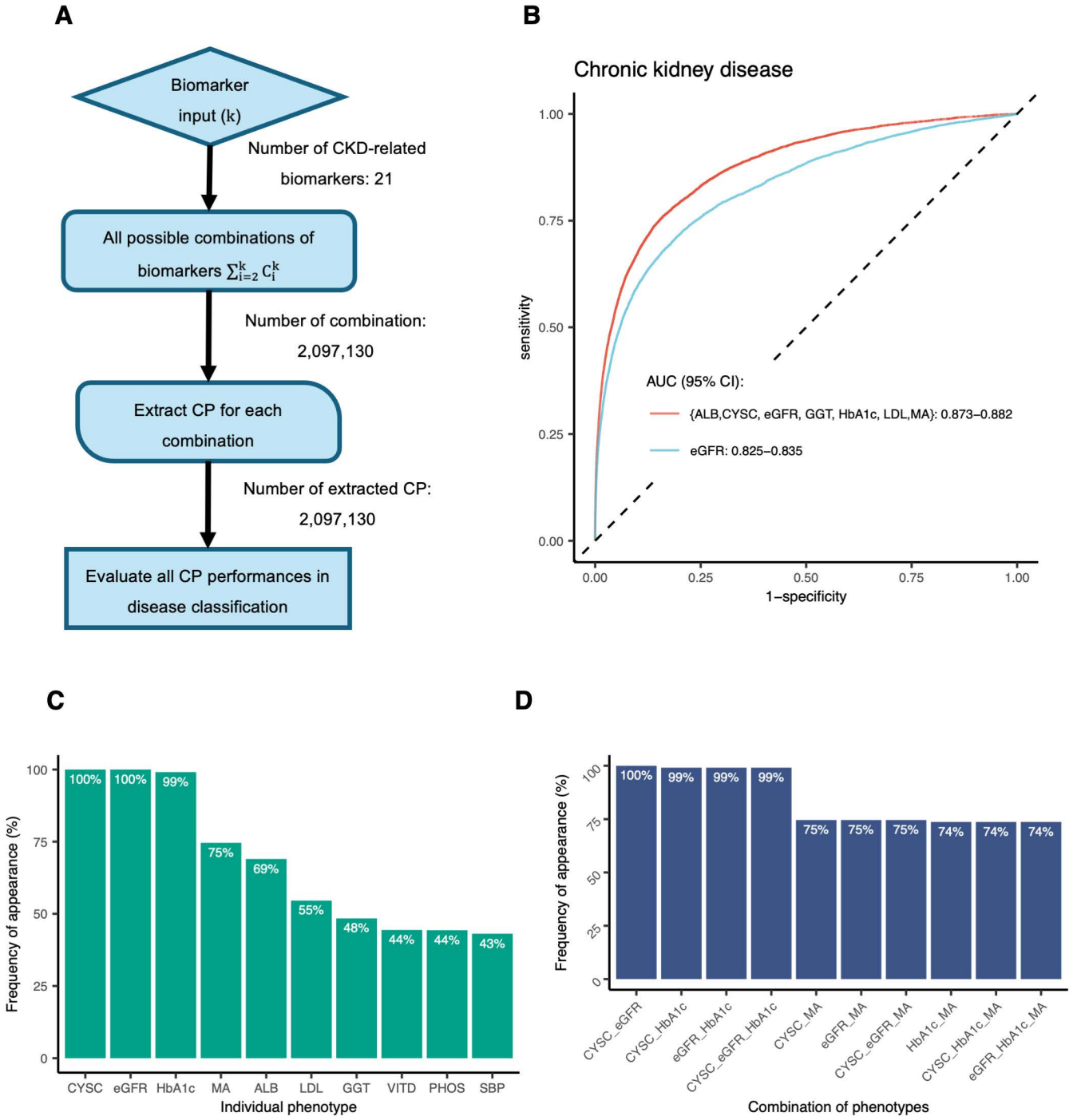

Fig 1. Combinatorial Principal Component Analysis (cPCA). (A) Flowchart of the cPCA method. **(B)** The ROC curve of the top CP extracted from eGFR, CYSC, MA, HbA1c, LDL, ALB, and GGT in comparison to the ROC curve of eGFR in terms of CKD classification. **(C)** Top 10 of the phenotypes that appeared most frequently in the ~50,000 CPs. **(D)** Top 10 of the phenotypes pairs or triples that appeared most frequently in the ~50,000 CPs.

https://doi.org/10.1371/journal.pgen.1011718.g001

Table 2. Top 10 CPs with the highest AUCs for CKD classification.

No.	Combination which CP extracted from	AUC	95% CI	P-values
1	{ALB, CYSC, eGFR, GGT, HbA1c, LDL, MA}	0.878	0.873-0.882	$3.7_{\times}10^{-149}$
2	{ALB, CYSC, eGFR, GGT, HbA1c, LDL, MA, PHOS}	0.878	0.873-0.882	$3.4_{\times}10^{-146}$
3	{ALB, CYSC, eGFR, GGT, HbA1c, LDL, MA, PHOS, VITD}	0.877	0.873-0.882	$9.1_{\times}10^{-141}$
4	{ALB, CYSC, eGFR, GGT, HbA1c, LDL, MA, VITD}	0.877	0.873-0.882	$2.6_{\times}10^{-140}$
5	{ALB, CALC, CYSC, DBP, eGFR, GGT, HbA1c, LDL, MA}	0.876	0.872-0.881	$1.2_{\times}10^{-134}$
6	{ALB, CYSC, eGFR, HbA1c, LDL, MA, PHOS, SBP}	0.876	0.872-0.881	$8.4_{\times}10^{-142}$
7	{ALB, CALC, CYSC, eGFR, HbA1c, LDL, MA, PHOS, SBP}	0.876	0.872-0.881	$5.5_{\times}10^{-137}$
8	{ALB, CALC, CYSC, eGFR, HbA1c, LDL, MA, SBP}	0.876	0.872-0.88	$2.9_{\times}10^{-138}$
9	{ALB, APOB, CYSC, eGFR, GGT, HbA1c, MA}	0.876	0.872-0.88	$2.6_{\times}10^{-145}$
10	{ALB, CALC, CYSC, DBP, eGFR, GGT, HbA1c, LDL, MA, VITD}	0.876	0.872-0.88	$1.3_{\times}10^{-135}$

Each CP's ROC curve was compared with the eGFR's ROC curve (AUC = 0.830, 95% CI: 0.825-0.835) and P-value was derived based on the boostrap method with 2000 resamples.

https://doi.org/10.1371/journal.pgen.1011718.t002

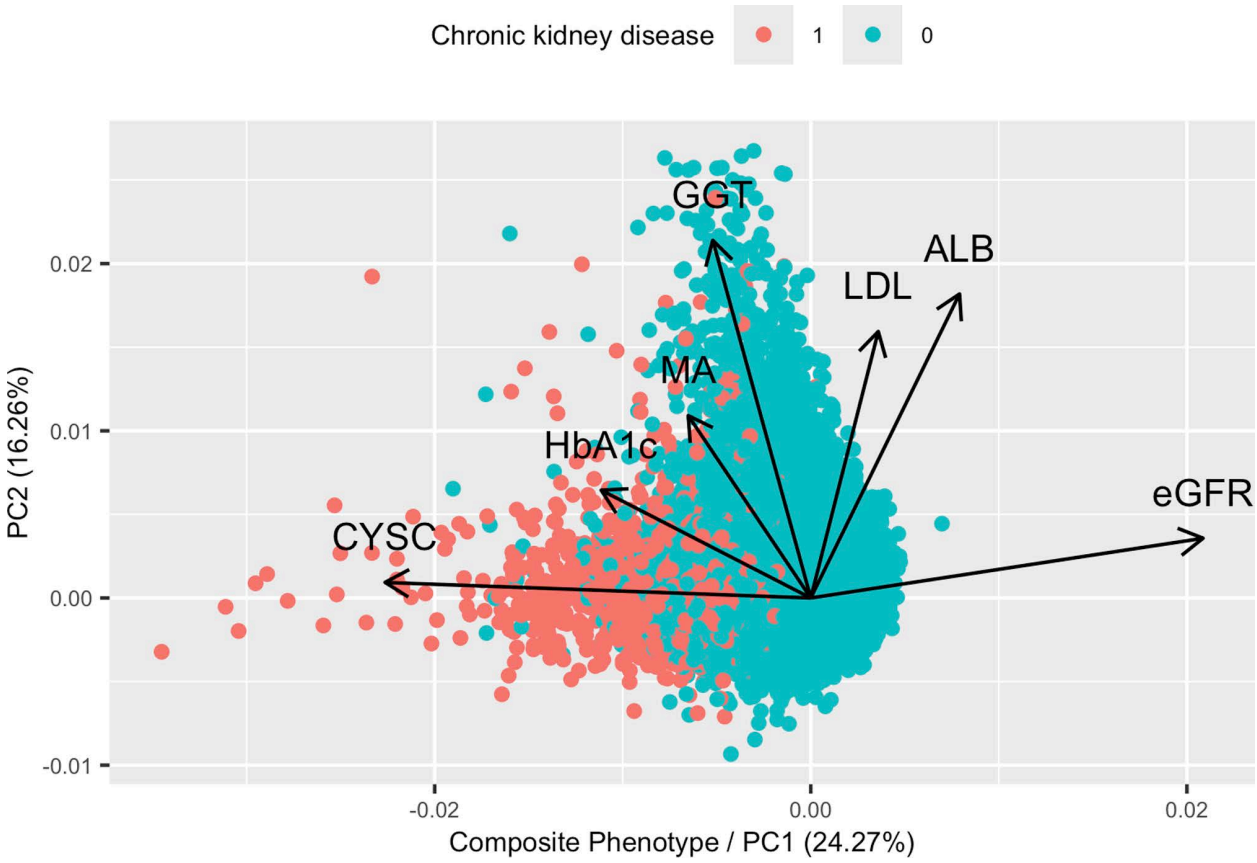

Fig 2. PCA illustrating the PCA scores and the trait loadings on the CP extracted from eGFR, CYSC, MA, HbA1c, LDL, ALB, and GGT. Samples were labeled with chronic kidney disease status (1:case; 0:non-case).

https://doi.org/10.1371/journal.pgen.1011718.g002

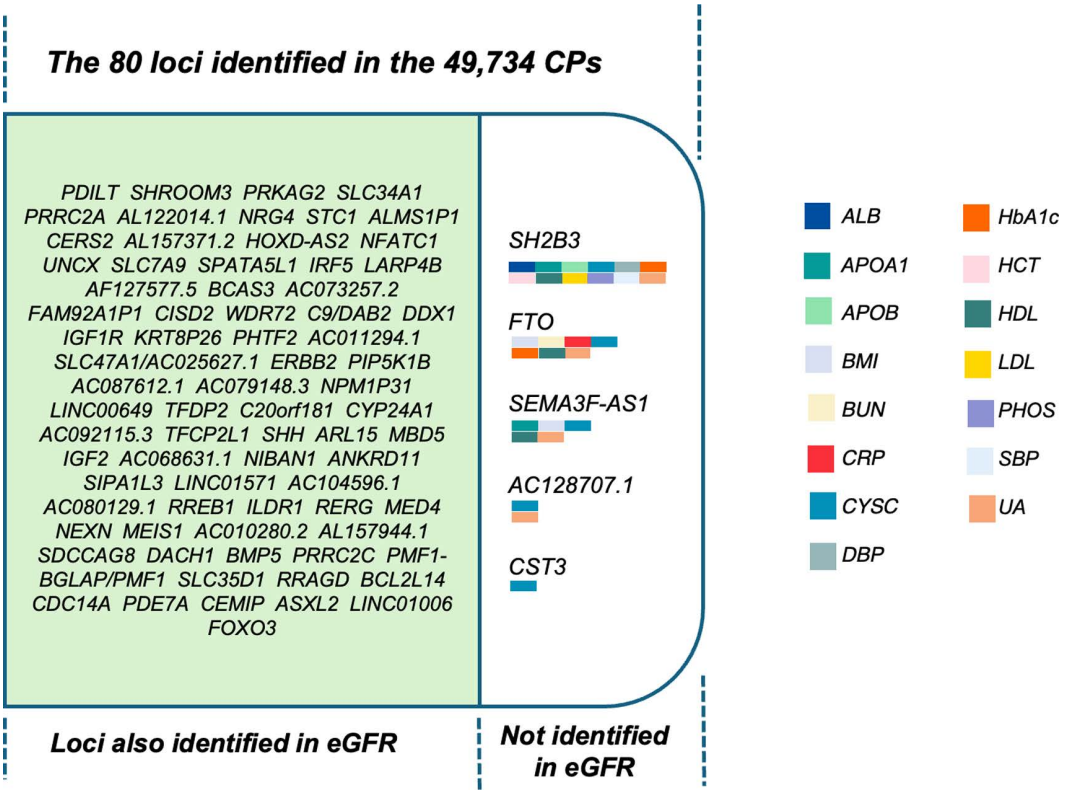

Fig 3. The 80 genetic loci identified in all the ~50,000 CPs.

https://doi.org/10.1371/journal.pgen.1011718.g003

Replication in Irish cohort and validation against large-scale eGFR GWAS

We utilised the independent cohort of Irish ethnicity in the UKB dataset for our replication analysis. In the discovery group, i.e., the British cohort, the CP extracted from the combination of {eGFR, CYSC, ALB, HbA1c, GGT, LDL, and MA} was among those that had the highest AUCs for CKD classification and at the same time had the least number of phenotypes (Table 2). Therefore, we selected this combination of phenotypes to generate a new CP for the replication cohort. As a result, the new CP in the replication cohort also had significantly better performance in CKD classification compared to those of individual phenotypes (Fig 4). Out of the 5 loci identified through the multi-phenotype approach, *CST3* and *SH2B3* were replicated in the Irish cohort as outlined in Table 3.

The CP was extracted from the combinations of phenotypes {eGFR, CYSC, ALB, HbA1c, GGT, LDL, and MA}.

In addition to the replication in the independent Irish cohort, we also looked at the genetic associations with eGFR in larger GWAS to examine whether these identified loci could be identified. We compared them against the large-scale GWAS of creatinine-based eGFR conducted in 1.2 million individuals [4]. Overall, the effect directions were consistent across the five loci (Table 4). However, only the *SH2B3* locus reached genome-wide significance ($p \leq 5 \times 10^{-8}$). Although not genome-wide significant, the *SEMA3F-AS1* and *AC128707.1* loci reached suggestive significance thresholds ($p \leq 1 \times 10^{-5}$), while the *CST3* and *FTO* loci showed no evidence of association ($p \geq 0.05$).

Colocalization analysis

To further investigate the genetic architecture underlying the composite phenotypes, we conducted a colocalization analysis to determine whether the loci identified through the multi-phenotype cPCA approach share causal variants with known

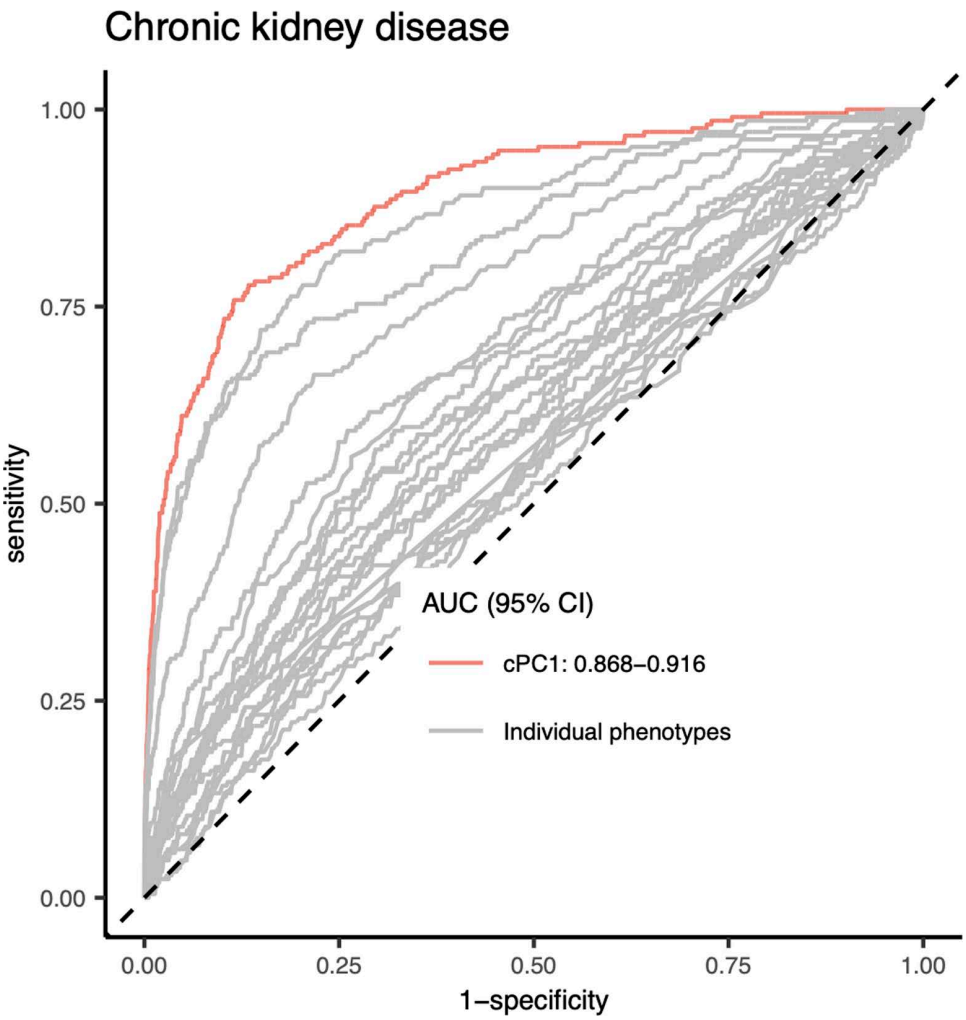

Fig 4. ROC curves for CKD classification of the CP and the 21 CKD-related phenotypes in the replication Irish cohort.

https://doi.org/10.1371/journal.pgen.1011718.g004

Table 3. Association results of the identified kidney-function loci in the discovery British group and the replication Irish group.

No.	Gene	Chr.	Position	rsID	A1	A2	British (n = 296,372)			Irish (n = 11,206)		
							Beta	SE	P	Beta	SE	P
1	*CST3*	20	23,569,186	rs2405392	T	C	0.190	0.003	0	0.223	0.018	$1.90_{\times}10^{-35}$*
2	*SH2B3*	12	111,884,608	rs3184504	T	C	-0.045	0.003	$7.15_{\times}10^{-53}$	-0.061	0.015	$5.74_{\times}10^{-5}$*
3	*FTO*	16	53,818,834	rs56313538	G	A	-0.026	0.003	$2.30_{\times}10^{-18}$	-0.037	0.016	0.017
4	*SEMA3F-AS1*	3	50,174,197	rs2624847	G	T	-0.024	0.003	$6.76_{\times}10^{-13}$	0.011	0.017	0.505
5	*AC128707.1*	12	78,807,411	rs7311712	T	C	0.019	0.003	$1.22_{\times}10^{-10}$	-0.004	0.015	0.800

The phenotypic outcomes were CP extracted from {eGFR, CYSC, ALB, HbA1c, GGT, LDL, and MA} for both the British group and the Irish group, respectively.

* p-values < 0.01 as accounted for multiple corrections.

https://doi.org/10.1371/journal.pgen.1011718.t003

Table 4. Validation against the large GWAS of creatinine-based eGFR (n = 1,201,909).

No.	Gene	Chr.	Position	rsID	A1	A2	GWAS meta-analysis for eGFR		
							Beta	SE	P
1	*CST3*	20	23,569,186	rs2405392	T	C	6.00×10^{-4}	3.00×10^{-4}	0.04875
2	*SH2B3*	12	111,884,608	rs3184504	T	C	-0.0016	3.00×10^{-4}	6.72×10^{-10}
3	*FTO*	16	53,818,834	rs56313538	G	A	-2.00×10^{-4}	3.00×10^{-4}	0.534
4	*SEMA3F-AS1*	3	50,174,197	rs2624847	G	T	-0.0014	3.00×10^{-4}	4.86×10^{-6}
5	*AC128707.1*	12	78,807,411	rs7311712	T	C	0.0013	3.00×10^{-4}	5.70×10^{-7}

https://doi.org/10.1371/journal.pgen.1011718.t004

eGFR-associated loci. This analysis aimed to validate the relevance of the identified loci to kidney function and assess their potential pleiotropic effects.

Among the two loci uniquely identified using cPCA and successfully replicated—*CST3* and *SH2B3*—only *SH2B3* showed strong evidence of colocalization with eGFR GWAS signals (Table 5). In contrast, *CST3* was associated with composite phenotypes but did not share a causal variant with eGFR, suggesting its role as a biomarker rather than a direct causal gene for CKD.

Discussion

CKD is a common term to describe a range of diseases characterized by impaired kidney structure, or reduced kidney function over time. Because there is an incomplete understanding of the genetics for different CKD subtypes, the identification of effective drug targets has been hindered. Research has tended to focus on eGFR or other single CKD-related biomarkers, yet this approach could be inadequate for capturing the underlying CKD etiology or pathophysiology. Since CKD is associated with many individual phenotypes, we reasoned that a multi-phenotype analytical approach may identify novel genetic loci relevant to CKD. Specifically, we designed a combinatorial PCA algorithm (cPCA), the aim of which was to extract relevant composite phenotypes for accurate CKD classification. This involved iteratively exploring all possible combinations of the 21 input phenotypes to identify composite phenotypes that outperformed individual biomarkers in CKD classification. Over 2 million phenotypic combinations were analyzed, resulting in the identification of nearly 50,000 composite phenotypes with significantly higher AUCs than eGFR or any individual phenotype.

The primary objectives of cPCA are to identify optimal combinations of biomarkers that collectively enhance disease classification, outperforming individual biomarkers alone, and to identify genetic loci associated with the target disease by leveraging multi-trait GWAS approaches. Regarding the second objective, cPCA shares similarities with multivariate GWAS and other joint analyses of multiple traits, including PCA-based methods (e.g., combined-PC [36] or adaptive principle component test [37]). However, unlike traditional multivariate methods that require a predefined set of traits, cPCA systematically explores and selects optimal biomarker combinations, making it a distinct and more flexible approach.

CYSC, eGFR, HbA1c, MA, ALB, LDL, and GGT were the most frequently observed phenotypes, appearing in 75% to 48% of those combinations. The frequent presence of HbA1c, ALB, LDL, and GGT alongside well-established CKD

Table 5. Colocalization probabilities for each locus are presented across five categories.

Loci	No causal variant	Causal variant for the CP only	Causal variant for eGFR only	Two distinct causal variants	One common causal variant
SH2B3	1.39×10^{-39}	0.1745	7.12×10^{-41}	0.008	**0.817**
CST3	3.41×10^{-291}	**0.970**	3.66×10^{-293}	0.010	0.020

Probabilities range from 0 to 1, with higher values indicating stronger support for a given colocalization scenario.

https://doi.org/10.1371/journal.pgen.1011718.t005

phenotypes such as CYSC, eGFR, and MA highlighted the overlap between kidney function and other aspects of human health, including blood glucose levels, cardiovascular health, and liver function [38–40]. Furthermore, we observed that although BUN and UA, which are highly correlated with eGFR, exhibited higher performance in CKD classification than HbA1c, MA, and others, they appeared less frequently in the ~ 50,000 combinations (BUN: 22.4% and UA: 0%). This suggests that cPCA could mitigate multicollinearity by ensuring the inclusion of independent phenotypes that are not highly correlated.

Consequently, we analysed the genetic associations and identified 80 loci that were consistently reached genome-wide significant in all the ~ 50,000 CPs. There were 5 loci that were not identified in eGFR but consistently identified in all the ~ 50,000 CPs. They were *CST3, SH2B3*, *FTO*, *SEMA3F-AS1*, and *AC128707.1*. *CST3* and *SH2B3* were successfully replicated in an independent cohort. *CST3*, encoding Cystatin-C (CYSC), a key marker of kidney function, appears to have been driven primarily by the presence of CYSC in all the combinations (Fig 1C). As *CST3* was not found in any other individual-trait GWAS except for CYSC's, it is more likely to be a biomarker gene than a causal gene influencing CKD pathology.

By contrast, *SH2B3*, encoding a cytokine-signaling regulator, was found to be genome-wide significant in a total of 12 single-trait GWASs including those for ALB, APOA1, APOB, CYSC, DBP, HbA1c, HCT, HDL, LDL PHOS, SBP, and UA. This was consistent with the fact that the index SNP rs3184504, which is also a missense SNP of *SH2B3*, has been found to be associated with multiple phenotypes and diseases relating to blood pressure, blood cells, cholesterol levels, as well as cardiovascular diseases and type-1 diabetes. The effect allele T had an effect size of -0.045 (SE = 0.003) in the GWAS of CP extracted from {eGFR, CYSC, ALB, HbA1c, GGT, LDL, and MA} (Tables 3 and 4), indicating that each additional copy of the T allele is associated with a decrease in CP. Since higher CP values correspond to lower CKD odds, this suggests that the *SH2B3* variant increases CKD risk by reducing CP. This finding is consistent with previous studies showing that animal models homozygous for the T allele, generated using CRISPR-Cas9, exhibited higher blood pressure and exacerbated kidney dysfunction compared to control mice [41].

Although it was not genome-wide significant in our eGFR GWAS (beta = -0.00856, p = $8.47_{\times}$10-5, n = 337,112), rs3184504-T has been shown to reach significance in larger GWAS cohorts (beta = -0.0016, p = $5.3_{\times}$10-8, n = 1,031,620) [4]. In our study this variant was robustly identified using a comparatively smaller cohort (~300,000 subjects) and successfully replicated in an even smaller sample size (n = 11,206). This finding emphasizes the increased power and effectiveness of the method (cPCA) in identifying genetic associations for complex phenotypes like CKD.

However, the cPCA approach also has its limitations. While it could identify genes with strong discriminatory power, it struggled to differentiate between biomarkers and causal genes. *CST3* is likely a biomarker reflecting kidney function rather than a gene driving CKD pathology, whereas *SH2B3* appears to play a more direct functional role. This limitation stemmed from the inclusion of biomarkers that reflect disease status but may not be directly involved in the underlying disease mechanisms (e.g., Cystatin-C).

Another limitation of this study is the reliance on ICD codes for identifying CKD outcomes. ICD codes often exhibit low sensitivity in detecting CKD, leading to potential underreporting of cases [42]. For instance, a study found that ICD-10 codes had a sensitivity ranging from 25% to 51% for detecting CKD stages G3-5, depending on the specific codes used [43]. This underreporting may introduce selection bias toward individuals with more severe disease, thereby limiting the generalizability of our findings. Nevertheless, the use of ICD-based CKD diagnoses remains common in large-scale biobank studies, particularly when serum creatinine-based eGFR values are unavailable for long-term CKD classification. While this approach likely enriches the cohort for more advanced CKD cases, it maintains high specificity—reported between 82% and 99%—ensuring that cases included in our analysis truly represent CKD rather than transient reductions in kidney function [43,44].

In conclusion, our multi-phenotype approach highlighted the increased power of cPCA in identifying CKD-related loci. Future studies should incorporate functional validation and explore a broader range of phenotypes to better understand the genetic architecture of CKD.

Methods

Ethics statement

Ethical approval for the UKB study was obtained from the North West Multi-Centre Research Ethics Committee, and all participants provided written informed consent. This research has been conducted utilizing the UK Biobank Resource under Application Number 86460.

Research cohort

The UK Biobank (UKB) [45] is a longitudinal cohort study designed to investigate the interplay between genes, the environment, and health. The study includes over 500,000 participants aged 40–69 years, recruited between 2006 and 2010 from 22 assessment centers across England, Scotland, and Wales. Participants provided detailed information about their health and lifestyle, donated samples of blood, urine, and saliva for long-term storage and analysis, and underwent various physical measurements, including height, weight, spirometry, blood pressure, and heel bone density.

We selected White-British samples, constituting the largest ethnic group within the UKB dataset, for the discovery cohort based on both the 'Ethnic background' and 'Genetic ethnic grouping' data. This approach allowed us to accurately identify individuals who self-identified as 'White British' and exhibited very similar genetic ancestry profiles, as determined by a principal components analysis of their genotypic data. Additionally, we excluded individuals whose genetic sex differed from their self-identified sex, those with sex chromosome aneuploidy, or those who were not included in the genetic principal components analysis conducted by the UKB research team. The final sample size was 337,112. Finally, we included individuals from the Irish ethnicity within the UKB dataset for the replication analyses (n = 11,106). The data processing steps were performed similarly to those used for the discovery cohort.

Phenotype data

In total, 21 biomarkers relevant to chronic kidney disease (CKD) were included in this study (Table 1). These phenotypes were assessed based on the correlations of the measurements with prevalence of CKD, CKD stages, kidney function, and an increased risk of adverse outcomes in individuals with CKD. All measurements were collected at baseline for all participants. Details of the assay manufacturers, analytical platforms, and analysis methodologies can be found at https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/biomarker-data. Quantitative measures outside their respective analytical ranges were treated as missing data. For urine albumin, which is essential for calculating the urine albumin-to-creatinine ratio (UACR), we implemented a multi-step procedure to address cases where urine albumin levels fell below the detection threshold of 6.7 mg/L for a significant number of participants (see S1 Text). Estimated GFR (eGFR) was calculated using the creatinine-based CKD-EPI-2021 equation without race coefficient. [10] Samples with more than 30% missing data points were excluded. Remaining missing phenotypic values were imputed to obtain a complete dataset using the R package missMDA v1.11 [46], ensuring that the imputed values had no effect on the principal component analysis (PCA) results. PCA was performed using the R package FactoMineR v1.34 [47].

Genotype data

Genome-wide genotyping was conducted on all UKB participants using the UK Biobank Axiom Array. Approximately 850,000 variants were directly measured, while over 90 million variants were imputed using the Haplotype Reference Consortium and UK10K + 1000 Genomes reference panels. Imputation data were stored in the compressed and indexed BGENv1.2 format. We converted the data from BGEN format into binary PGEN files and performed quality control procedures within PLINK2.0 [48]. The criteria for selecting variants were: (1) autosomal variants; (2) missing rate of less than 5%; (3) not significantly deviated from Hardy-Weinberg equilibrium (p-value = $10_{\times}$10-10); (4) minor allele frequency (MAF) of at least 0.01; and (5) imputation score of more than 0.8. After quality control, we retained 12.7 million SNPs for subsequent analysis.

CKD clinical outcome data

Health-related outcome data are available in death, hospital, and primary care records. Using the ICD-10 and ICD-9 codes (International Classification of Diseases, tenth and ninth editions), we categorized individuals diagnosed with chronic kidney disease, renal failure, renal sclerosis, chronic glomerulonephritis, nephritis, nephropathy, hypertensive chronic kidney disease, hypertensive heart and kidney disease, diabetes with renal complications, kidney replaced by transplant, disorders resulting from impaired renal function, or unspecified disorders of the kidney and ureter as CKD cases.

Combinatorial principal component analysis (cPCA)

Principles: We developed an multi-phenotype approach called combinatorial PCA (cPCA) to identify combinations of biomarkers that collectively offer improved discriminatory power in disease classification compared to individual biomarkers alone. In cPCA, various combinations with varying numbers of biomarkers are generated from a fixed set of input biomarkers. The number of possible combinations generated can be calculated as $\sum_{i=2}^{k} C_i^k$.The first principal component, denoted as CP, is then extracted to represent each combination. CP serves as a comprehensive biomarker signature, representing the maximum variance direction within the biomarker combination. Finally, the performance of each CP in disease classification is evaluated and compared to that of single biomarkers.

Implementation Details: To systematically explore and identify potential superior components for CKD classification beyond conventional biomarkers, we applied cPCA to a set of 21 CKD-related phenotypes. Initially, we generated 2,097,130 unique combinations out of the 21 phenotypes. These combinations encompassed all possible subsets of the 21 phenotypes with varying numbers, ranging from 2 to the complete set of 21. For each combination, we extracted the first principal component as CP, resulting in 2 million CPs.

Here, we only select the first principal component (PC1) because of several reasons. First, PC1 represents the direction of maximum variance within the biomarker combination, capturing the most substantial and dominant patterns of variability in the data. In disease classification, capturing this primary variance is crucial, as it likely reflects the strongest underlying biological signals linked to disease risk. Higher-order principal components (e.g., PC2, PC3) often capture less significant variance, which may be more influenced by noise, measurement error, or irrelevant biological variation. By focusing on PC1, we emphasize the most reliable signal, which could improve the robustness of the disease classification model. Additionally, by reducing the dimensionality to PC1, the method simplifies the analysis, making it computationally more efficient, especially when the number of generated combinations is large.

Subsequently, we evaluated the performance of each CP in CKD classification and compared it to that of CYSC, which served as the best single marker for CKD classification. To validate the efficacy of the identified combinations, we partitioned the dataset into a training set (70%) and a test set (30%). Notably, cPCA was exclusively performed on the training set, encompassing the 2 million combinations. The performance evaluation involved comparing the ROC curves (Receiver Operating Characteristic curves) of each CP against those of individual phenotypes. Confidence intervals and derived p-values for the calculated AUCs (Area Under the Curve) were computed using bootstrap methods with 2000 stratified bootstrap replicates, implemented within the R package pROC [49]. Combinations exhibiting significantly higher AUCs compared to eGFR were further validated using the independent test set. The cPCA script written in R is available in S2 Text.

Genome-wide analyses

Genome-wide association studies (GWAS) were performed by fitting linear models (for quantitative traits) or logistic models (for binary traits) implemented in PLINK2.0 [48]. All the input phenotypes were inverse-normal transformed prior to GWAS. Age, sex, and the first 20 genetic principal components were integrated into the models as covariates. SNP-based

heritability and genetic correlation were estimated based on the GWAS summary statistics using linkage disequilibrium score regression (LDSC) v1.0.1 [50].

Colocalization analysis

To assess whether the loci identified through the multi-phenotype cPCA approach share causal variants with known eGFR-associated loci, we performed colocalization analysis using the coloc 5.2.3 R package [51]. This method estimates the probability that the same causal variant underlies both traits by evaluating their association signals at each locus.

For each locus identified in the cPCA GWAS and successfully replicated (*CST3* and *SH2B3*), we retrieved summary statistics from the eGFR GWAS and performed colocalization analysis across a ± 500 kb window centered on the lead SNP. The coloc Bayesian framework calculates the posterior probabilities for five hypotheses: no causal variant in either trait (PP0), causal variant in the composite phenotype GWAS only (PP1), causal variant in the eGFR GWAS only (PP2), two distinct causal variants (PP3), and one shared causal variant (PP4; evidence of colocalization). A locus was considered to exhibit strong colocalization if PP4 > 0.7, indicating a high probability that the same causal variant influences both traits.

Supporting information

S1 Table. Trait loadings for the CP extracted from {eGFR, CYSC, ALB, HbA1c, GGT, LDL, and MA}.
(DOCX)

S1 Text. Method for handling urine albumin measurements below the detection limit.
(DOCX)

S2 Text. Scripts for conducting cPCA and downstream genomics analyses.
(DOCX)

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 86460.

Author contributions

Conceptualization: Kim Ngan Tran, Lyn R. Griffiths, Rodney Lea.

Data curation: Kim Ngan Tran.

Formal analysis: Kim Ngan Tran.

Funding acquisition: Lyn R. Griffiths.

Investigation: Kim Ngan Tran, Rodney Lea.

Methodology: Kim Ngan Tran, Andrew J. Mallett, Rodney Lea.

Project administration: Lyn R. Griffiths.

Supervision: Lyn R. Griffiths, Rodney Lea.

Validation: Heidi G. Sutherland, Andrew J. Mallett.

Visualization: Kim Ngan Tran.

Writing – original draft: Kim Ngan Tran.

Writing – review & editing: Kim Ngan Tran, Heidi G. Sutherland, Andrew J. Mallett, Lyn R. Griffiths, Rodney Lea.

References

1. Levin A, Tonelli M, Bonventre J, Coresh J, Donner J-A, Fogo AB, et al. Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. Lancet. 2017;390(10105):1888–917. https://doi.org/10.1016/S0140-6736(17)30788-2 PMID: 28434650
2. Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet. 2019;51(6):957–72.
3. Köttgen A, Pattaro C. The CKDGen Consortium: ten years of insights into the genetic basis of kidney function. Kidney Int. 2020;97(2):236–42. https://doi.org/10.1016/j.kint.2019.10.027 PMID: 31980069
4. Stanzick KJ, Li Y, Schlosser P, Gorski M, Wuttke M, Thomas LF, et al. Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. Nat Commun. 2021;12(1):4350. https://doi.org/10.1038/s41467-021-24491-0 PMID: 34272381
5. Teumer A, Li Y, Ghasemi S, Prins BP, Wuttke M, Hermle T, et al. Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. Nat Commun. 2019;10(1):4130. https://doi.org/10.1038/s41467-019-11576-0 PMID: 31511532
6. Cañadas-Garre M, Anderson K, Cappa R, Skelly R, Smyth LJ, McKnight AJ, et al. Genetic Susceptibility to Chronic Kidney Disease - Some More Pieces for the Heritability Puzzle. Front Genet. 2019;10:453. https://doi.org/10.3389/fgene.2019.00453 PMID: 31214239
7. Rysz J, Gluba-Brzózka A, Franczyk B, Jabłonowski Z, Ciałkowska-Rysz A. Novel biomarkers in the diagnosis of chronic kidney disease and the prediction of its outcome. Int J Mol Sci. 18(8).
8. Tran NK, Lea RA, Holland S, Nguyen Q, Raghubar AM, Sutherland HG, et al. Multi-phenotype genome-wide association studies of the Norfolk Island isolate implicate pleiotropic loci involved in chronic kidney disease. Sci Rep. 2021;11(1):19425. https://doi.org/10.1038/s41598-021-98935-4 PMID: 34593906
9. Benoit SW, Ciccia EA, Devarajan P. Cystatin C as a biomarker of chronic kidney disease: latest developments. Expert Rev Mol Diagn. 2020;20(10):1019–26. https://doi.org/10.1080/14737159.2020.1768849 PMID: 32450046
10. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. N Engl J Med. 2021;385(19):1737–49. https://doi.org/10.1056/NEJMoa2102953 PMID: 34554658
11. Seki M, Nakayama M, Sakoh T, Yoshitomi R, Fukui A, Katafuchi E, et al. Blood urea nitrogen is independently associated with renal outcomes in Japanese patients with stage 3-5 chronic kidney disease: a prospective observational study. BMC Nephrol. 2019;20(1):115. https://doi.org/10.1186/s12882-019-1306-1 PMID: 30940101
12. Oluwo O, Scialla JJ. Uric Acid and CKD Progression Matures with Lessons for CKD Risk Factor Discovery. Clin J Am Soc Nephrol. 2021;16(3):476–8. https://doi.org/10.2215/CJN.10650620 PMID: 33055190
13. Hernandez D, Espejo-Gil A, Bernal-Lopez MR, Mancera-Romero J, Baca-Osorio AJ, Tinahones FJ, et al. Association of HbA1c and cardiovascular and renal disease in an adult Mediterranean population. BMC Nephrol. 2013;14:151. https://doi.org/10.1186/1471-2369-14-151 PMID: 23865389
14. Fox ER, Benjamin EJ, Sarpong DF, Nagarajarao H, Taylor JK, Steffes MW, et al. The relation of C--reactive protein to chronic kidney disease in African Americans: the Jackson Heart Study. BMC Nephrol. 2010;11:1. https://doi.org/10.1186/1471-2369-11-1 PMID: 20078870
15. Ejerblad E, Fored CM, Lindblad P, Fryzek J, McLaughlin JK, Nyrén O. Obesity and risk for chronic renal failure. J Am Soc Nephrol. 2006;17(6):1695–702. https://doi.org/10.1681/ASN.2005060638 PMID: 16641153
16. Lu JL, Molnar MZ, Naseer A, Mikkelsen MK, Kalantar-Zadeh K, Kovesdy CP. Association of age and BMI with kidney function and mortality: a cohort study. Lancet Diabetes Endocrinol. 2015;3(9):704–14. https://doi.org/10.1016/S2213-8587(15)00128-X PMID: 26235959
17. Herrington WG, Smith M, Bankhead C, Matsushita K, Stevens S, Holt T, et al. Body-mass index and risk of advanced chronic kidney disease: Prospective analyses from a primary care cohort of 1.4 million adults in England. PLoS One. 2017;12(3):e0173515. https://doi.org/10.1371/journal.pone.0173515 PMID: 28273171
18. De Nicola L, Provenzano M, Chiodini P, D'Arrigo G, Tripepi G, Del Vecchio L, et al. Prognostic role of LDL cholesterol in non-dialysis chronic kidney disease: multicenter prospective study in Italy. Nutr Metab Cardiovasc Dis. 2015;25(8):756–62.
19. Nam KH, Chang TI, Joo YS, Kim J, Lee S, Lee C, et al. Association Between Serum High-Density Lipoprotein Cholesterol Levels and Progression of Chronic Kidney Disease: Results From the KNOW-CKD. J Am Heart Assoc. 2019;8(6):e011162. https://doi.org/10.1161/JAHA.118.011162 PMID: 30859896
20. Zhao W, Li J, Zhang X, Zhou X, Xu J, Liu X, et al. Apolipoprotein B and renal function: across-sectional study from the China health and nutrition survey. Lipids Health Dis. 2020;19(1):110. https://doi.org/10.1186/s12944-020-01241-7 PMID: 32460759
21. Zhao W-B, Zhu L, Rahman T. Increased serum concentration of apolipoprotein B is associated with an increased risk of reaching renal replacement therapy in patients with diabetic kidney disease. Ren Fail. 2020;42(1):323–8. https://doi.org/10.1080/0886022X.2020.1745235 PMID: 32242489
22. Kwon S, Kim DK, Oh K-H, Joo KW, Lim CS, Kim YS, et al. Apolipoprotein B is a risk factor for end-stage renal disease. Clin Kidney J. 2020;14(2):617–23. https://doi.org/10.1093/ckj/sfz186 PMID: 33623687
23. Goek O-N, Köttgen A, Hoogeveen RC, Ballantyne CM, Coresh J, Astor BC. Association of apolipoprotein A1 and B with kidney function and chronic kidney disease in two multiethnic population samples. Nephrol Dial Transplant. 2012;27(7):2839–47. https://doi.org/10.1093/ndt/gfr795 PMID: 22287661
24. Lang J, Katz R, Ix JH, Gutierrez OM, Peralta CA, Parikh CR, et al. Association of serum albumin levels with kidney function decline and incident chronic kidney disease in elders. Nephrol Dial Transplant. 2018;33(6):986–92. https://doi.org/10.1093/ndt/gfx229 PMID: 28992097

25. Lang J, Scherzer R, Tien PC, Parikh CR, Anastos K, Estrella MM, et al. Serum albumin and kidney function decline in HIV-infected women. Am J Kidney Dis. 2014;64(4):584–91. https://doi.org/10.1053/j.ajkd.2014.05.015 PMID: 25059222
26. Zubovic SV, Kristic S, Prevljak S, Pasic IS. Chronic Kidney Disease and Lipid Disorders. Med Arch. 2016;70(3):191–2.
27. Lee DY, Han K, Yu JH, Park S, Heo J-I, Seo JA, et al. Gamma-glutamyl transferase variability can predict the development of end-stage of renal disease: a nationwide population-based study. Sci Rep. 2020;10(1):11668. https://doi.org/10.1038/s41598-020-68603-0 PMID: 32669624
28. Agarwal R. Blood pressure components and the risk for end-stage renal disease and death in chronic kidney disease. Clin J Am Soc Nephrol. 2009;4(4):830–7. https://doi.org/10.2215/CJN.06201208 PMID: 19339424
29. Glassock RJ. Is the presence of microalbuminuria a relevant marker of kidney disease?. Curr Hypertens Rep. 2010;12(5):364–8. https://doi.org/10.1007/s11906-010-0133-3 PMID: 20686930
30. Iseki K, Ikemiya Y, Iseki C, Takishita S. Haematocrit and the risk of developing end-stage renal disease. Nephrol Dial Transplant. 2003;18(5):899–905. https://doi.org/10.1093/ndt/gfg021 PMID: 12686662
31. Chen TK, Estrella MM, Astor BC, Greene T, Wang X, Grams ME, et al. Longitudinal changes in hematocrit in hypertensive chronic kidney disease: results from the African-American study of kidney disease and hypertension (AASK). Nephrol Dial Transplant. 2015;30(8):1329–35.
32. Vervloet MG, Sezer S, Massy ZA, Johansson L, Cozzolino M, Fouque D, et al. The role of phosphate in kidney disease. Nat Rev Nephrol. 2017;13(1):27–38. https://doi.org/10.1038/nrneph.2016.164 PMID: 27867189
33. Kim CS, Kim SW. Vitamin D and chronic kidney disease. Korean J Intern Med. 2014;29(4):416–27. https://doi.org/10.3904/kjim.2014.29.4.416 PMID: 25045287
34. Janmaat CJ, van Diepen M, Gasparini A, Evans M, Qureshi AR, Ärnlöv J, et al. Lower serum calcium is independently associated with CKD progression. Sci Rep. 2018;8(1):5148. https://doi.org/10.1038/s41598-018-23500-5 PMID: 29581540
35. Mitka M. Low Diastolic Blood Pressure and Chronic Kidney Disease Are Associated With Increased Mortality. JAMA. 2013;310(12):1215–6.
36. Aschard H, Vilhjálmsson BJ, Greliche N, Morange P-E, Trégouët D-A, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. Am J Hum Genet. 2014;94(5):662–76. https://doi.org/10.1016/j.ajhg.2014.03.016 PMID: 24746957
37. Bu D, Zhang S, Li N. Analyzing Multiple Phenotypes Based on Principal Component Analysis. Acta Math Appl Sin Engl Ser. 2022;38(4):843–60. https://doi.org/10.1007/s10255-022-1019-2
38. Hassanein M, Shafi T. Assessment of glycemia in chronic kidney disease. BMC Med. 2022;20(1):117. https://doi.org/10.1186/s12916-022-02316-1 PMID: 35414081
39. Jankowski J, Floege J, Fliser D, Böhm M, Marx N. Cardiovascular Disease in Chronic Kidney Disease: Pathophysiological Insights and Therapeutic Options. Circulation. 2021;143(11):1157–72. https://doi.org/10.1161/CIRCULATIONAHA.120.050686 PMID: 33720773
40. Fabrizi F, Messa P, Basile C, Martin P. Hepatic disorders in chronic kidney disease. Nat Rev Nephrol. 2010;6(7):395–403. https://doi.org/10.1038/nrneph.2010.37 PMID: 20386560
41. Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, et al. New loci associated with kidney function and chronic kidney disease. Nat Genet. 2010;42(5):376–84. https://doi.org/10.1038/ng.568 PMID: 20383146
42. Geng T, Chen J, Lu Q, Wang P, Xia P, Zhu K, et al. Nuclear magnetic resonance–based metabolomics and risk of ckd. Am J Kidney Dis. 2024;83(1):9–17.
43. Bothe T, Fietz A-K, Schaeffner E, Douros A, Pöhlmann A, Mielke N, et al. Diagnostic Validity of Chronic Kidney Disease in Health Claims Data Over Time: Results from a Cohort of Community-Dwelling Older Adults in Germany. Clin Epidemiol. 2024;16:143–54. https://doi.org/10.2147/CLEP.S438096 PMID: 38410416
44. Paik JM, Patorno E, Zhuo M, Bessette LG, York C, Gautam N, et al. Accuracy of identifying diagnosis of moderate to severe chronic kidney disease in administrative claims data. Pharmacoepidemiol Drug Saf. 2022;31(4):467–75.
45. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z PMID: 30305743
46. Josse J, Husson F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. J Stat Soft. 2016;70(1). https://doi.org/10.18637/jss.v070.i01
47. Lê S, Josse J, Husson F. FactoMineR: AnRPackage for Multivariate Analysis. J Stat Soft. 2008;25(1). https://doi.org/10.18637/jss.v025.i01
48. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852
49. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77 PMID: 21414208
50. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291–5. https://doi.org/10.1038/ng.3211 PMID: 25642630
51. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383. https://doi.org/10.1371/journal.pgen.1004383 PMID: 24830394