A clustering approach to improve our understanding of the genetic and phenotypic complexity of chronic kidney disease

A. Eoli*[†1-2], S. Ibing[†1-2], C. Schurmann[1-2,6], G.N. Nadkarni[4-5], H.O. Heyne[1,2,4] & E. Böttinger[1-4]

1. Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany
2. Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York City, NY, USA
3. Department of Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY, USA
4. Windreich Dept. of Artificial Intelligence & Human Health, Icahn School of Medicine at Mount Sinai, New York City, NY, USA
5. The Charles Bronfman Institute of Personalized Medicine, New York City, NY, USA
6. Current address: Bayer AG, Research & Development, Pharmaceuticals, Berlin, Germany

* Corresponding author: andrea.eoli@hpi.de

† Authors contributed equally

Abstract

Chronic kidney disease (CKD) is a complex disorder that causes a gradual loss of kidney function, affecting approximately 9.1% of the world's population. Here, we use a soft-clustering algorithm to deconstruct its genetic heterogeneity. First, we selected 322 CKD-associated independent genetic variants from published genome-wide association studies (GWAS) and added association results for 229 traits from the GWAS catalog. We then applied nonnegative matrix factorization (NMF) to discover overlapping clusters of related traits and variants. We computed cluster-specific polygenic scores and validated each cluster with a phenome-wide association study (PheWAS) on the BioMe biobank (n=31,701). NMF identified nine clusters that reflect different aspects of CKD, with the top-weighted traits signifying areas such as kidney function, type 2 diabetes (T2D), and body weight. For most clusters, the top-weighted traits were confirmed in the PheWAS analysis. Results were found to be more significant in the cross-ancestry analysis, although significant ancestry-specific associations were also identified. While all alleles were associated with a decreased kidney function, associations with CKD-related diseases (e.g., T2D) were found only for a smaller subset of variants and differed across genetic ancestry groups. Our findings leverage genetics to gain insights into the underlying biology of CKD and investigate population-specific associations.

Introduction

Chronic kidney disease (CKD) is a primarily asymptomatic disease characterized by a gradual loss of kidney function over a period extending from several months to years[1]. CKD affects approximately 9.1% of the global population, with a higher prevalence in high-income countries[2]. The leading risk factors for developing CKD are diabetes (40% of cases) and hypertension (29% of cases), followed by heart disease, family history of CKD, and obesity[3]. Other factors, such as exposure to HIV and contaminants, are additionally important in low-income countries[4,5]. The genetic ancestry also plays a crucial role, with increased risk rates of kidney failure in Black/African Americans and Hispanics/Latinos compared to individuals of European ancestry[6]. If left untreated, CKD increases the mortality risk for individuals with cardiovascular disease (CVD) and can result in the complete loss of kidney function[7]. Therefore, early detection is critical for improving quality of life and life expectancy. During the early stages of CKD, cost-effective treatment options are available and can be tailored to the cause of the disease[8].

CKD is defined by a reduced functionality of the kidneys, which limits its filtering capability over a period of at least three months[9]. The main biomarkers for CKD detection include the urinary albumin/creatinine ratio (ACR) and the estimated glomerular filtration rate (eGFR) [10]. While ACR facilitated diagnosing albuminuria – an indicator of kidney damage characterized by an elevated excretion of urinary albumin – the eGFR estimates the filter volume of the glomerulus per unit of time using different biomarkers such as serum creatinine[10]. An abnormal kidney activity is indicated by high ACR values, reduced eGFR, or both.

Over the past few decades, many large-scale genomic studies, such as genome-wide association studies (GWAS), have successfully identified more than 500 independent genetic variants associated with reduced kidney function[11–13]. The association between genetic variants and various phenotypes has been studied, and the results are often shared in publicly available databases, like the GWAS Catalog[14]. The association of one genetic variant with multiple traits can be considered to identify secondary traits associated with a phenotype. This understanding can help elucidate potentially shared disease mechanisms, assuming that genetic variants affecting a shared pathway also have a similar impact on the associated traits.

Soft-clustering methods provide a means to reduce the genetic complexity of a heterogeneous disease while also accounting for shared disease mechanisms. In contrast to hard-clustering approaches like K-means or hierarchical clustering, soft-clustering enables the factorization of high-

dimensional data by identifying overlapping clusters[15]. Non-negative matrix factorization (NMF) is a family of algorithms within multivariate analysis that addresses the dimensionality challenge by extracting meaningful features from a given data set[16,17].

In this study, we aimed to deconstruct the heterogeneity of CKD by identifying its genetic subtypes. First, we collect all published variant-trait associations for variants associated with reduced kidney function and apply soft-clustering using NMF. We used the algorithm's weights to calculate cluster-specific polygenic scores (cPGS) within the BioMe biobank. Finally, we use a phenome-wide association study (cPGS-PheWAS) to validate and interpret the clusters. By deconstructing the complexity of CKD, this methodology contributes new insights into the disease pathways of CDK and enhances our understanding of population-specific differences for CDK.

Results

NMF identified nine clusters of CKD-associated variants

The most frequent CKD-associated secondary traits retrieved from the GWAS Catalog are related to kidney function (e.g., blood urea nitrogen, urea, uric acid, and cystatin C measurements), hemoglobin levels (e.g., hemoglobin measurements, hematocrit, and erythrocyte counts), T2D, body weight (e.g., body height, appendicular lean mass, BMI, BMI-adjusted waist-hip ratio), and pulse pressure (systolic and diastolic blood pressure measurements), among others (see Fig.S1). CKD-associated traits and their associated CKD variants were factorized into nine partly overlapping clusters by conducting NMF. To ensure the results were robust, we repeated the clustering with bNMF and got comparable results (Tab.S1). The top seven traits per cluster are summarised in Fig.1. The 'Reduced lipids' cluster is associated with decreasing blood lipid levels (triglycerides, total cholesterol, use of lipid-lowering medications) and liver enzymes. The top traits of the cluster 'Increased body mass' show a positive association with body weight (appendicular lean mass, body height, and body weight). The clusters 'Increased blood volume' and 'Reduced blood volume' are positively and negatively associated with volemic traits (e.g., mean corpuscular volume and mean corpuscular hemoglobin), respectively. Similarly, clusters 'Increased/Reduced hematocrit' show opposite associations with hemoglobin content (e.g., hematocrit, hemoglobin measurements, red blood cell density, erythrocyte count), and clusters 'Increased/Reduced inflammation' convey opposite associations with markers of inflammation (C-reactive protein) and blood lipids. Lastly, cluster 'Increased urate' is positively associated with kidney function biomarkers like urate, blood/serum urea nitrogen, blood proteins,

and Cystatin C. The complete lists of the top features and variants per cluster, defined as traits and variants in the top decile of the cluster weights of the matrices H and W, are listed in Tab.S2. The matrices H and W are also available as supplementary material. Fig.S2 summarises how the variants are distributed in each cluster, showing their overlaps.

PheWAS replicated biological pathways of most clusters across ancestries

Each cluster was examined by conducting a cPGS-PheWAS with 988 quantitative traits and 832 binary traits on the four Bio*Me* cohorts (ALL, AFR, AMR, EUR). Except for the 'Increased body mass' and the 'Reduced blood volume' clusters, we replicated at least 3 of the clusters' top 5 traits (Fig.2). The level of significance was reached more frequently across ancestries (ALL) than when replicating on the individual ones (AMR, AFR, EUR) (Tab.S3). In addition to the replicated traits, significant associations with decreasing eGFR were seen in clusters 'Increased urate' (β=-0.04 [-0.06 - -0.03], p-value=6.7e-09) and 'Reduced hematocrit' (β=-0.05 [-0.07 - -0.04], p-value=4.0e-12) (Tab.S3). 'Reduced hematocrit' was also nominally associated with an increased risk for chronic renal failure (OR=1.11 [1.05-1.16], p-value=1.2e-04) and with the curated phenotype 'diabetic and hypertensive CKD' (OR=1.27 [1.11-1.46], p-value=6.6e-04). Besides showing negative associations with disorders of lipoid metabolism, cluster 'Increased inflammation' shows strong negative associations with Alzheimer's disease (OR=0.60 [0.52-0.7], p-value=1.5e-11) and dementias (OR=0.77 [0.71-0.84], p-value=1.0e-09) (Fig.S3). Regarding the individual ancestries, EUR showed the strongest associations when replicating on binary traits, with an increased risk for "visual disturbances" (OR=1.51 [1.27-1.79], p-value=2.1e-06) in the cluster 'Reduced inflammation,' while AFR showed the strongest associations when replicating on quantitative traits, with the strongest association being for the LDL-HDL ratio (β=-0.14 [-0.17 - -0.11], p-value=9.3e-21) in the 'Increased inflammation' cluster. Fig.2 summarizes which of the top traits of each cluster have been replicated, while the complete list of cPGS-PheWAS results by ancestry is stored in Tab.S3.

Increased body mass

Effect	Trait	Activity
↑	Appendicular lean mass	14.36
↑	Body height	13.09
↑	Body weight	7.27
↑	Lymphocyte count	3.64
↓	Heel bone mineral density	3.49
↑	Erythrocyte count	2.93
↑	Educational attainment	2.84

Reduced lipids

Effect	Trait	Activity
↓	Triglyceride meas.	10.00
↓	Total cholesterol meas.	7.08
↓	HMG CoA reductase inhibitor use meas.	5.42
↓	Serum alanine aminotransferase meas.	5.16
↓	Serum gamma-glutamyl transferase meas.	4.96
↑	Red blood cell distribution width	4.69
↓	Serum albumin meas.	3.85

Increased blood volume

Effect	Trait	Activity
↑	Mean corpuscular volume	17.95
↑	Mean corpuscular hemoglobin	16.19
↑	Mean corpuscular hemoglobin conc.	12.93
↓	Erythrocyte count	7.09
↓	Reticulocyte meas.	6.06
↑	Mean reticulocyte volume	5.87
↓	Red blood cell density meas.	4.72

Reduced blood volume

Effect	Trait	Activity
↓	Mean corpuscular volume	14.70
↓	Mean corpuscular hemoglobin	10.28
↓	Mean corpuscular hemoglobin conc.	6.48
↑	Erythrocyte count	5.28
↑	Red blood cell density meas.	4.60
↓	HDL cholesterol meas.	3.38
↓	Systolic blood pressure	3.29

Increased hematocrit

Effect	Trait	Activity
↑	Hematocrit	16.36
↑	Hemoglobin meas.	14.99
↑	Red blood cell density meas.	12.99
↑	Erythrocyte count	9.92
↓	Total cholesterol meas.	3.82
↑	Aspartate aminotransferase meas.	3.77
↑	Neutrophil count	3.64

Reduced hematocrit

Effect	Trait	Activity
↓	Hematocrit	16.79
↓	Hemoglobin meas.	15.51
↑	Blood urea nitrogen meas.	14.69
↓	Red blood cell density meas.	13.26
↓	Erythrocyte count	7.35
↑	Systolic blood pressure	3.83
↑	Uric acid meas.	3.00

Increased inflammation

Effect	Trait	Activity
↑	C-reactive protein meas.	20.74
↑	Apolipoprotein A 1 meas.	7.91
↓	Familial hyperlipidemia	6.81
↑	VLDL cholesterol meas.	5.95
↓	Phospholipids:total lipids ratio	4.12
↑	Red blood cell distribution width	3.55
↑	Cholesteryl ester and HDL meas.	3.41

Reduced inflammation

Effect	Trait	Activity
↓	Triglyceride meas.	7.60
↓	C-reactive protein meas.	6.77
↑	Sex hormone-binding globulin meas.	6.32
↓	Serum gamma-glutamyl transferase meas.	4.97
↓	Serum albumin meas.	4.78
↓	Lipid meas., HDL cholesterol meas.	4.48
↓	Uric acid meas.	4.41

Increased urate

Effect	Trait	Activity
↑	Urate meas.	21.53
↑	Uric acid meas.	14.08
↑	Blood urea nitrogen meas.	8.95
↑	Blood protein meas.	5.79
↑	Cystatin C meas.	3.86
↑	Serum urea meas.	2.39
↓	HDL cholesterol meas.	1.67

Figure 1 - Top seven CKD-associated secondary traits per cluster

Cluster	CKD-associated trait	Dir	Activity	OR	Coeff	95% CI	p-value	Signif
IBM	Body height	↑	13.09		0.04	[0.03 - 0.05]	2.6E-17	✓
	Body weight	↑	7.27		0.02	[0 - 0.03]	7.6E-03	
Reduced lipids	Total cholesterol meas.	↓	7.08		-0.03	[-0.04 - -0.02]	3.8E-06	✓
	Serum alanine aminotransferase meas.	↓	5.16		-0.02	[-0.03 - -0.01]	4.8E-03	
	Serum gamma-glutamyl transferase meas.	↓	4.96		-0.03	[-0.04 - -0.01]	1.7E-03	
	Serum albumin meas.	↓	3.85		-0.02	[-0.03 - -0.01]	3.0E-03	
	Total blood protein meas.	↓	2.97		-0.03	[-0.04 - -0.02]	6.7E-06	✓
	LDL cholesterol meas. (phospholipids:total lipids)	↑	2.29		-0.04	[-0.05 - -0.02]	4.7E-08	✓
Incr. BV	Mean corpuscular volume	↑	17.95		0.04	[0.03 - 0.05]	2.8E-10	✓
	Mean corpuscular hemoglobin	↑	16.19		0.04	[0.03 - 0.05]	4.2E-11	✓
	Mean corpuscular hemoglobin conc.	↑	12.93		0.02	[0.01 - 0.03]	9.0E-05	
	Erythrocyte count	↓	7.09		-0.03	[-0.04 - -0.02]	4.8E-06	✓
Reduced BV	Mean corpuscular volume	↓	14.70		-0.03	[-0.04 - -0.02]	1.2E-07	✓
	Mean corpuscular hemoglobin	↓	10.28		-0.03	[-0.04 - -0.02]	2.8E-06	✓
	Mean corpuscular hemoglobin conc.	↓	6.48		0.05	[0.01 - 0.1]	2.5E-02	
	Erythrocyte count	↑	5.28		0.02	[0.01 - 0.04]	9.9E-05	
	Red blood cell density meas.	↑	4.60		0.02	[0.01 - 0.03]	8.2E-04	
Incr. hematocrit	Hematocrit	↑	16.36		0.03	[0.02 - 0.04]	5.7E-07	✓
	Hemoglobin meas.	↑	14.99		0.03	[0.02 - 0.04]	2.1E-06	✓
	Red blood cell density meas.	↑	12.99		0.03	[0.02 - 0.05]	1.4E-08	✓
	Erythrocyte count	↑	9.92		0.03	[0.02 - 0.04]	2.6E-08	✓
	Aspartate aminotransferase meas.	↑	3.77		0.08	[0.03 - 0.12]	1.2E-03	
	Leukocyte count	↑	3.03		0.03	[0.01 - 0.04]	2.8E-05	
	Thyroid stimulating hormone meas.	↑	1.46		-0.04	[-0.05 - -0.02]	1.1E-07	✓
Reduced hematocrit	Hematocrit	↓	16.79		-0.03	[-0.05 - -0.02]	6.5E-08	✓
	Hemoglobin meas.	↓	15.51		-0.03	[-0.04 - -0.02]	7.0E-08	✓
	Blood urea nitrogen meas.	↑	14.69		0.06	[0.04 - 0.07]	1.8E-21	✓
	Red blood cell density meas.	↓	13.26		-0.03	[-0.04 - -0.02]	6.9E-07	✓
	Erythrocyte count	↓	7.35		-0.03	[-0.04 - -0.02]	2.1E-06	✓
	Systolic blood pressure	↑	3.83	1.27		[1.11 - 1.46]	6.6E-04	
	Uric acid meas.	↑	3.00		0.05	[0.03 - 0.07]	8.9E-06	✓
	RAS-acting agents use meas.	↑	2.09	1.05		[1.01 - 1.08]	5.2E-03	
Incr. inflammation	C-reactive protein meas.	↑	20.74		0.08	[0.06 - 0.1]	4.7E-12	✓
	Familial hyperlipidemia	↓	6.81	0.89		[0.87 - 0.92]	1.0E-15	✓
	HDL cholesterol meas.	↑	5.95		-0.10	[-0.11 - -0.08]	2.6E-50	✓
	VLDL cholesterol meas. (phospholipids:total lipids)	↓	4.12		-0.11	[-0.12 - -0.09]	6.9E-57	✓
	LDL cholesterol meas., phospholipids:total lipids ratio	↑	3.17		-0.09	[-0.11 - -0.06]	9.7E-13	✓
	Type 2 diabetes mellitus	↑	2.41	1.11		[1.02 - 1.2]	1.8E-02	
	Free cholesterol meas., VLDL cholesterol meas.	↓	1.77		-0.10	[-0.12 - -0.09]	7.0E-55	✓
Reduced inflammation	C-reactive protein meas.	↓	6.77		-0.03	[-0.06 - -0.01]	3.2E-03	
	Serum gamma-glutamyl transferase meas.	↓	4.97		-0.05	[-0.06 - -0.03]	1.5E-07	✓
	Platelet count	↓	4.12		-0.06	[-0.1 - -0.02]	1.9E-03	
	Fatty acid meas.	↓	3.69	0.95		[0.92 - 0.98]	6.0E-04	
	Calcium meas.	↓	3.59		-0.02	[-0.03 - -0.01]	6.4E-04	
	Alkaline phosphatase meas.	↓	3.49		-0.03	[-0.04 - -0.01]	2.2E-05	✓
	Total blood protein meas.	↓	3.34		-0.02	[-0.04 - -0.01]	2.2E-04	
	Glucose meas.	↑	3.14		0.03	[0.01 - 0.04]	3.8E-03	
	Platelet crit	↓	3.22	-0.02		[-0.03 - -0.01]	7.7E-04	
	Blood protein meas.	↓	2.66		-0.02	[-0.04 - -0.01]	2.2E-04	✓
Incr. urate	Urate meas.	↑	21.53		0.06	[0.03 - 0.08]	1.0E-06	✓
	Uric acid meas.	↑	14.08		0.06	[0.03 - 0.08]	1.0E-06	✓
	Blood urea nitrogen meas.	↑	8.95		0.03	[0.02 - 0.04]	2.4E-07	✓
	Blood protein meas.	↑	5.79		0.04	[0.02 - 0.05]	1.7E-09	✓

Figure 2 - Replication of cluster traits

cPGSs suggest clear differences between genetic ancestries

We extracted the cluster weights of the W matrix and used them to calculate cluster-specific polygenic scores (cPGS) for participants of the Bio*Me* cohort. Fig.4 shows the standardized polygenic score distributions for all NMF clusters across the Bio*Me* cohort (ALL) and the individual continental populations EUR (n=7,447), AMR (n=5,336), and AFR (n=5,660). A normal distribution was observed for the cluster 'Increased urate' (EUR, AMR, and AFR; Anderson-Darling test, all p-values available in

Tab.S5). Although polygenic scores are expected to have a normal distribution[18], the other eight clusters present either a skewed tail (e.g., 'Increased hematocrit') or several peaks in their cPGS distributions (e.g., 'Reduced inflammation'). As illustrated in Fig.3, the peaks are caused by a few variants with relatively high cluster weights (the complete list of cluster weights for the top variants of each cluster is available in Tab.S2). For example, the top variant in cluster 'Increased inflammation' (rs429358, mapped gene: APOE) weighs 4.6, while the second one (rs17050272) weighs 0.2. In Fig.4, we can also observe how this variant is more frequent in participants of inferred EUR ancestry. Similarly, the top variant of 'Reduced inflammation' (rs1260326, mapped gene: GCKR) weighs 5.7 and seems to be more frequent in the AFR population, while the second one (rs4418728) weighs 0.9. This unbalance in weight creates the three peaks of the distributions: the lower peak includes the scores of individuals without the top variant (0 copies), the middle one the heterozygous (1 copy), and the higher peak includes scores of participants with two copies of the top variant. Other ancestry-specific differences are visible in the distributions of four clusters and are significant when testing with the Mann-Whitney test (all p-values available in Tab.S4). This suggests that some variants appear with different frequencies in people that do not share similar ancestry: 'Increased inflammation' (all combinations), 'Reduced inflammation' (all combinations), 'Reduced lipids' (EUR vs. AFR), and 'Increased body mass' (EUR vs AFR and AFR vs AMR).

Cluster	SNPs	Median	Mean	Q_{90} weight	Max
Increased body mass	154	0	0.07	0.24	2.41
Reduced lipids	84	0.04	0.14	0.24	3.52
Increased blood volume	114	0.01	0.08	0.13	1.7
Reduced blood volume	205	0.01	0.07	0.09	2.29
Increased hematocrit	144	0.01	0.09	0.34	2.44
Reduced hematocrit	195	0.06	0.11	0.19	2.05
Increased inflammation	172	0.01	0.05	0.04	4.6
Reduced inflammation	170	0.02	0.09	0.12	5.73
Increased urate	214	0.04	0.11	0.31	1.47

Figure 3 - Summary statistics of the weights of each cluster

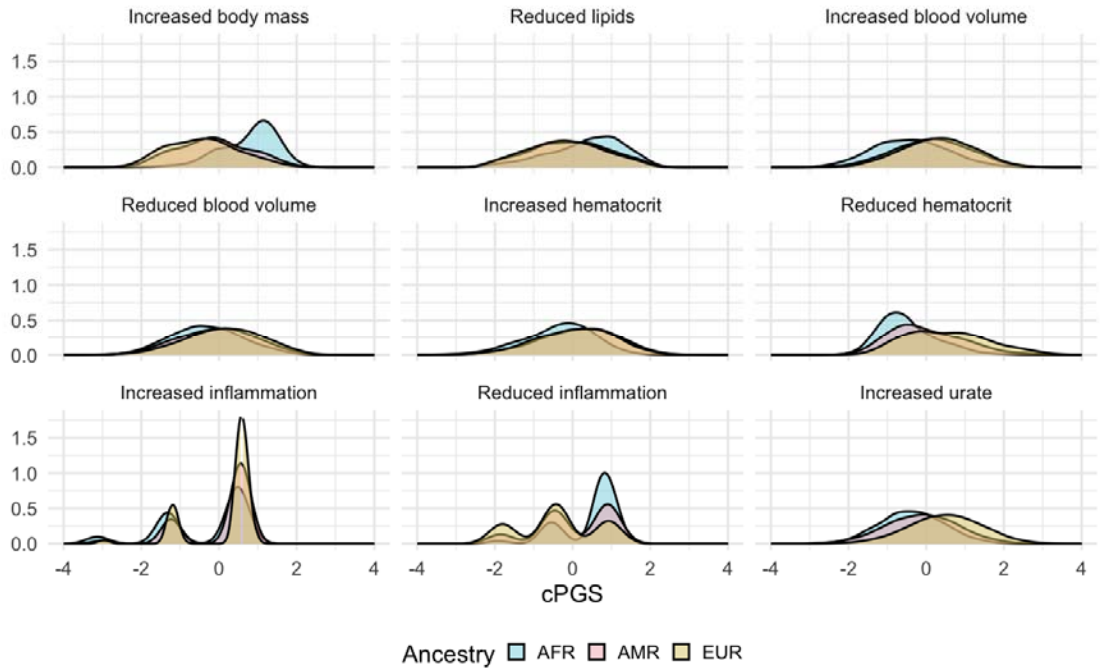

Figure 4 - Standardized cluster-specific polygenic scores (cPGS) per genetic population

Distribution of participants across clusters

As the cPGS are calculated separately per cluster, each Bio*Me* participant might have high polygenic scores in multiple clusters. Therefore, to understand the cluster overlap in terms of relative risk, we checked how many individuals belonged to the top decile of 1 or more clusters. 58% (18,431/31,701) of the whole Bio*Me* cohort (ALL) were at high risk in at least 1 cluster. Of these, 60.2% were in the top decile for only 1 cluster, while 37% were at risk for 2-3 clusters (Fig.S4).

Discussion

CKD is typically defined as a progressive loss of kidney function over time. Although numerous genetic variants have been identified as associated with CKD, their relationship to disease pathways remains largely unclear. The work described here is the most comprehensive assessment of how variants associated with CKD can be grouped according to different CKD-related factors. Specifically, we included variant-trait associations of 322 CKD SNPs and 229 related metabolic traits from publicly available GWAS datasets. By analyzing these associations with NMF, a factorization approach that allows for minimal overlap between groups, we identified 9 clusters of CKD variants and associated traits.

CKD is commonly recognized as a heterogeneous condition with various underlying causes and risk factors, which are unlikely to represent a single disease process. This complexity is also reflected by the associated traits retrieved from published GWAS, which are related to kidney function, hemoglobin levels, T2D, body weight, and pulse pressure, among others. Attempting to deconvolute CKD's genetic heterogeneity and differentially grouping these traits, the nine clusters we identified represented different aspects of CKD. For example, the 'Increased urate' cluster, whose clustering weights represent abnormal levels of urinary metabolites like urate, blood/serum urea nitrogen, blood proteins, and Cystatin C, is related to decreasing kidney function. In normal conditions, such blood metabolites are excreted by the kidneys, but in CKD they accumulate and exert a detrimental biological activity[19,20]. A second cluster, which we summarised as 'Increased inflammation,' was strongly clustered around rising serum C-reactive protein (CRP) concentrations. CRP is a common inflammatory biomarker in chronic diseases like CKD, diabetes, and cardiovascular diseases[21–23]. In line with that, patients with CKD commonly experience chronic inflammatory states[24]. These states tend to worsen as the disease progresses toward end-stage renal disease and are reflected, or even modulated[25], by increasing CRP levels[26–28].

We then studied the genotype-phenotype correlation to demonstrate the utility of the clusters. We could replicate most of the top-weighted features on quantitative traits (i.e., biomarkers), while the validation on binary traits (i.e., diagnoses) was less robust and required additional clinical interpretation. For the clusters of 'Increased urate' and 'Increased inflammation,' the top traits were confirmed by the PheWAS. CKD is also associated with dyslipidemia comprising high levels of triglycerides and LDL-cholesterol, and low levels of HDL-cholesterol and apolipoprotein A1[29]. We could observe similar associations in clusters 'Increased inflammation,' 'Reduced lipids,' and 'Reduced inflammation.' Notably, we found multiple significant associations for cluster 'Increased inflammation' with reduced risk of dementias. Glycerophospholipids play an essential role in neural membranes[30,31], and their levels are directly correlated with serum triglycerides and inversely correlated with total cholesterol and eGFR[32].

A limitation of this study is the need for more genetic diversity in the GWAS Catalog, which mainly consists of studies performed on the European population. This European bias is well described in the literature and has important implications for disease risk prediction across global populations[33]. Despite this lack of genetic diversity, we could still validate our results in Bio*Me*, a biobank enriched for populations with non-European ancestries. We were most powered when jointly analyzing across ancestries (ALL), while signals replicated in different ancestral groups with some group-specific

differences. This result suggests that, although most CKD risk factors converge across ancestral groups, ancestry-specific studies are essential. Another two limitations are the filtering rules used to select traits and variants for the algorithm's input matrix and the possible existence of non-additive interactions between risk factors that we did not consider in this study. Lastly, one of the input CKD studies, the PAGE study[34], was also conducted using Bio*Me* data. However, this should not impact the results since we are not looking at CKD case/control scenarios but at CKD subtypes.

Understanding the biological pathways that lead to CKD is essential to improve clinical management. For example, some clusters group similar traits but with opposite effect directions (e.g., 'Increased hematocrit' and 'Reduced hematocrit'), while others suggest potentially protective effects (e.g., against dyslipidemia in cluster 'Increased inflammation'). This behavior might indicate that CKD can affect the same metabolic pathways differently, confirming the genetic complexity of the disease. Additionally, the clusters have a limited degree of overlap and, as each represents a specific set of variants, participants might be high risk (i.e., in the top decile of the polygenic score) for more than one cluster. This additive disease model, similar to the mutational signatures in cancer, suggests a possible interplay of genetic susceptibility to multiple disease-causing mechanisms[35].

In summary, by clustering genetic variants associated with CKD, we identified clusters with distinct trait associations, likely representing mechanistic pathways involved in CKD. We confirmed the validity of these clusters phenotypically. Further clinical investigations could explore whether individuals with a common disrupted pathway also share similar complications, a comparable rate of disease progression, or a different treatment response. In the future, classifying patients with CKD using their genotype may improve care by offering a more personalized and genetically informed clinical plan.

Methods

Trait-variants selection

We identified and aligned the alleles of 508 independent genetic variants associated either with decreased kidney function (defined as low eGFR levels for at least three months) or with CKD (using ICD-9/10 codes) from the most recent GWAS and GWAS meta-analyses[11,13,34,36,37] (Fig.5a). We then used the R package *LDlinkR* (R version 4.2.1) to retrieve all proxy SNPs in linkage disequilibrium (r^2 >= 0.6) with the lead variants, across all available 1000G human populations[38,39] and used the GWAS Catalog database to link the proxy SNPs to 805 associated traits (as of July 30th, 2022) [14]. We excluded gender-specific GWAS and GWAS performed on less than 100 individuals. Additionally, as

we are interested in secondary features associated with CKD, we excluded GWAS of traits directly related to eGFR or CKD (e.g., "Mild to moderate chronic kidney disease," "Estimated Glomerular Filtration Rate"). We kept trait-variant associations with a significance threshold of less than 1×10^{-6} using a Bonferroni correction for all 2,401 associations in our data set. To reduce sparsity in the data, we excluded traits associated with less than five variants; this threshold was empirically defined by comparing the clustering results of traits associated with up to 15 CKD variants. We standardized effect sizes across all GWAS by dividing the regression coefficient beta (B) by the standard error, using the GWAS summary statistic results. Traits and variants were then arranged as a matrix with the standardized effect sizes (β) as values. Tab.S5 contains, for each input CKD variant, the list of CKD-associated secondary traits extracted from the GWAS Catalog and the corresponding exclusion criteria for those excluded during the filtering steps.

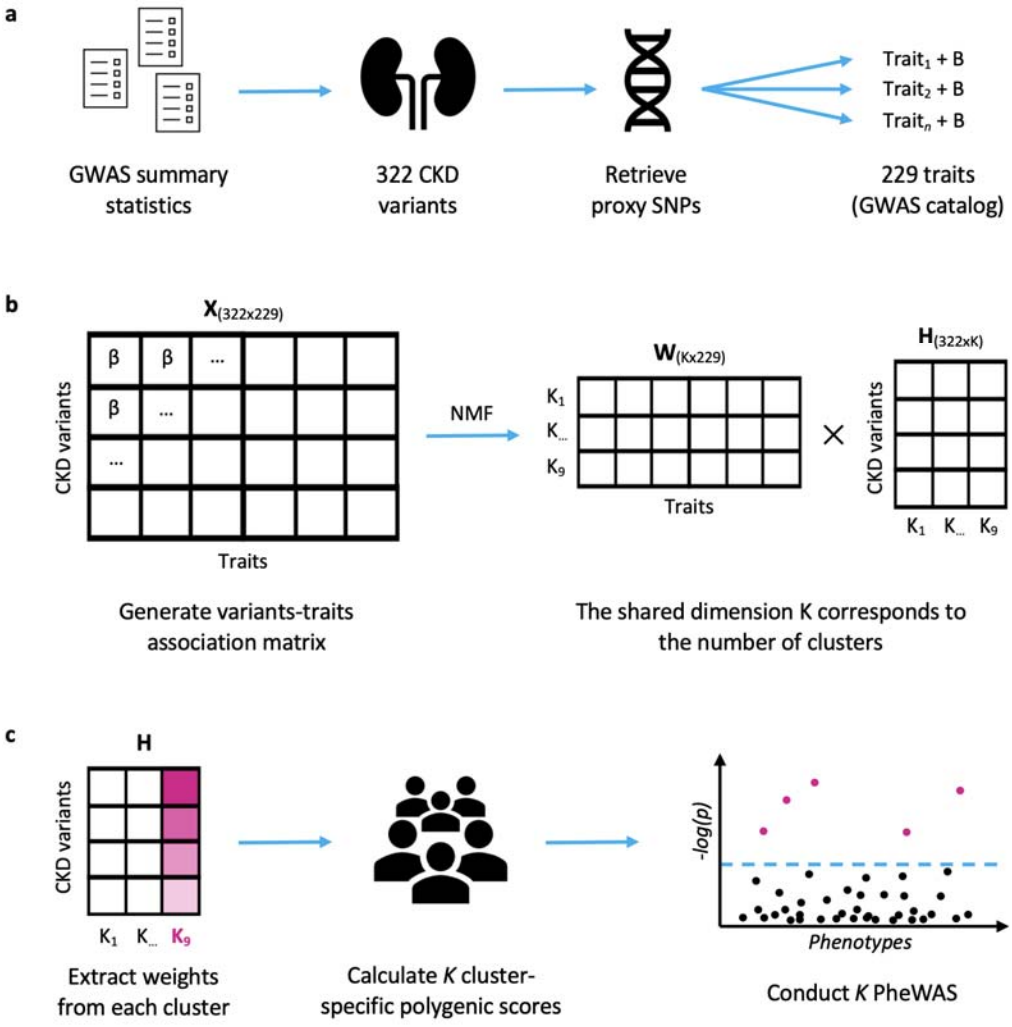

Figure 5 - Methods overview

NMF

NMF factorizes the input matrix of trait-variant associations (X, of dimensions 229x322) into a matrix of traits (H, 229xK) and one of variants (W, Kx322), so that HxW ≈ X[16] (Fig.5b). The factorization rank K corresponds to the number of clusters. We implemented NMF using the R package *ButchR* with 10,000 iterations, 30 random initiations, and the convolution threshold set to 80[40]. The number of expected clusters was set between 2 and 20. ButchR suggests the optimal K based on six cluster evaluation metrics, like the mean silhouette width and the Frobenius error. If two or more K were presented, we considered results with the highest mean silhouette width and the lowest Frobenius error, as suggested by Alexandrov et al[35]. As additional validation, we also performed a Bayesian version of NMF[17], using the code provided by Udler et al[41]. bNMF was run 1,000 times with up to 200,000 iterations in each run.

Cluster-specific polygenic scores

The results of clustering provide cluster-specific weights for each variant and trait. We used *PLINK* and the variant cluster weights to calculate cluster-specific polygenic scores (cPGS) of the Bio*Me* biobank participants[42]. cPGS were standardized within each cluster. The normality of each cPGS distribution was tested with the Anderson-Darling method. Differences between ancestry-specific distributions were tested with the Mann-Whitney test.

Validation cohort (Bio*Me*)

We validated our results using the genetic and linked electronic health records (EHR) data of 31,701 Bio*Me* biobank participants[43] (Fig.5c). As a fine-scale population structure can improve the risk prediction of complex diseases within genetic groups[44], we inferred the genetic ancestry of the Bio*Me* participants. We then performed a Principal Component Analysis (PCA) using *PLINK*, excluding relatives above 2nd-degree (kinship method, estimated using *KING*[45]) and variants with minor allele frequency below 0.05[42,46]. We trained a random forest classifier to infer the genetic ancestry of Bio*Me* participants using the 1000 Genomes labels as reference[47]. The labeled ancestries are Admixed American (AMR, n=5,336), African (AFR, n=5,660), European (EUR, n=7,447), South Asian (SAS, n=613), and East Asian (EAS, n=728). For sub-population-specific analyses, we removed participants with mixed ancestry (defined as having a random forest probability ≤ 0.5) and outliers by only including the quantiles 0.25-0.90[48] (n=11,404).

Modeling disease outcomes as a function of cluster-specific polygenic scores

For each cluster, the cPGS were associated with the phenotypes available in the Bio*Me* data set by performing a phenome-wide association study (cPGS-PheWAS). We fitted linear regression models to analyze 988 quantitative traits (e.g., laboratory results) and logistic regression models for 832 binary traits with cPGS as independent variables, adjusting for sex, age, and the first ten genetic principal components (*stats* R package[49]). Binary traits included Phecodes mapped to ICD-9 and ICD-10 codes (a Phecode is considered if at least two relevant diagnostic codes were present in a patient's EHR) [50] and curated phenotypes[51]. Controls were identified as the reference category. Traits were only considered if present or measured in at least 100 biobank participants. The model parameters were standardized using the *effectsize* R package (refit method) [52]. Standardized coefficient estimates (linear regression) and odd ratios (logistic regression) were reported with the corresponding 95% confidence intervals. The Bonferroni method was used to adjust for multiple testing, and the alpha threshold was defined as 2.7e-05 (0.05/(988+832)). We then compared the PheWAS results with the traits in the top decile of NMF's trait weights.

References

1. Luyckx, V. A., Tonelli, M. & Stanifer, J. W. The global burden of kidney disease and the sustainable development goals. *Bull. World Health Organ.* **96**, 414-422D (2018).
2. Bikbov, B. *et al.* Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **395**, 709–733 (2020).
3. Couser, W. G., Remuzzi, G., Mendis, S. & Tonelli, M. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney Int.* **80**, 1258–1270 (2011).
4. Ekrikpo, U. E. *et al.* Chronic kidney disease in the global adult HIV-infected population: A systematic review and meta-analysis. *PLOS ONE* **13**, e0195443 (2018).
5. Jha, V. *et al.* Chronic kidney disease: global dimension and perspectives. *The Lancet* **382**, 260–272 (2013).
6. Saran, R. *et al.* US Renal Data System 2015 Annual Data Report: Epidemiology of Kidney

Disease in the United States. *Am. J. Kidney Dis.* **67**, A7–A8 (2016).

7. Levey, A. S. & Coresh, J. Chronic kidney disease. *The Lancet* **379**, 165–180 (2012).
8. Ruggenenti, P., Cravedi, P. & Remuzzi, G. Mechanisms and Treatment of CKD. *J. Am. Soc. Nephrol.* **23**, 1917–1928 (2012).
9. Levey, A. S. *et al.* Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int.* **67**, 2089–2100 (2005).
10. Cockwell, P. & Fisher, L.-A. The global burden of chronic kidney disease. *The Lancet* **395**, 662–664 (2020).
11. Gorski, M. *et al.* Meta-analysis uncovers genome-wide significant variants for rapid kidney function decline. *Kidney Int.* **99**, 926–939 (2021).
12. Teumer, A. *et al.* Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. *Nat. Commun.* **10**, 4130 (2019).
13. Morris, A. P. *et al.* Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nat. Commun.* **10**, 29 (2019).
14. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
15. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
16. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126 (1994).
17. Févotte, C. & Idier, J. Algorithms for Nonnegative Matrix Factorization with the β-Divergence. *Neural Comput.* **23**, 2421–2456 (2011).
18. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic

risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

19. Gagnebin, Y. *et al.* Exploring blood alterations in chronic kidney disease and haemodialysis using metabolomics. *Sci. Rep.* **10**, 19502 (2020).

20. Vanholder, R. *et al.* Review on uremic toxins: Classification, concentration, and interindividual variability. *Kidney Int.* **63**, 1934–1943 (2003).

21. Stenvinkel, P. & Lindholm, B. C-Reactive Protein in End-Stage Renal Disease: Are There Reasons to Measure It? *Blood Purif.* **23**, 72–78 (2005).

22. Kalantar-Zadeh, K., Stenvinkel, P., Pillon, L. & Kopple, J. D. Inflammation and nutrition in renal insufficiency. *Adv. Ren. Replace. Ther.* **10**, 155–169 (2003).

23. Hansson, G. K. Inflammation, Atherosclerosis, and Coronary Artery Disease. *N. Engl. J. Med.* **352**, 1685–1695 (2005).

24. Stenvinkel, P. *et al.* Strong association between malnutrition, inflammation, and atherosclerosis in chronic renal failure. *Kidney Int.* **55**, 1899–1911 (1999).

25. Li, Z. *et al.* C-reactive protein promotes acute renal inflammation and fibrosis in unilateral ureteral obstructive nephropathy in mice. *Lab. Invest.* **91**, 837–851 (2011).

26. Stuveling, E. M. *et al.* C-reactive protein is associated with renal function abnormalities in a non-diabetic population. *Kidney Int.* **63**, 654–661 (2003).

27. Menon, V. *et al.* C-reactive protein and albumin as predictors of all-cause and cardiovascular mortality in chronic kidney disease. *Kidney Int.* **68**, 766–772 (2005).

28. Muntner, P. *et al.* The Prevalence of Nontraditional Risk Factors for Coronary Heart Disease in Patients with Chronic Kidney Disease. *Ann. Intern. Med.* **140**, 9 (2004).

29. Theofilis, P., Vordoni, A., Koukoulaki, M., Vlachopanos, G. & Kalaitzidis, R. G. Dyslipidemia in Chronic Kidney Disease: Contemporary Concepts and Future Therapeutic Perspectives. *Am. J. Nephrol.* **52**, 693–701 (2021).

30. Farooqui, A. A., Horrocks, L. A. & Farooqui, T. Glycerophospholipids in brain: their metabolism, incorporation into membranes, functions, and involvement in neurological

disorders. *Chem. Phys. Lipids* **106**, 1–29 (2000).

31. Frisardi, V., Panza, F., Seripa, D., Farooqui, T. & Farooqui, A. A. Glycerophospholipids and glycerophospholipid-derived lipid mediators: A complex meshwork in Alzheimer's disease pathology. *Prog. Lipid Res.* **50**, 313–330 (2011).

32. Chen, H. *et al.* Combined Clinical Phenotype and Lipidomic Analysis Reveals the Impact of Chronic Kidney Disease on Lipid Metabolism. *J. Proteome Res.* **16**, 1566–1578 (2017).

33. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

34. Lin, B. M. *et al.* Genetics of Chronic Kidney Disease Stages Across Ancestries: The PAGE Study. *Front. Genet.* **10**, 494 (2019).

35. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).

36. 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).

37. Lin, B. M. *et al.* Whole genome sequence analyses of eGFR in 23,732 people representing multiple ancestries in the NHLBI trans-omics for precision medicine (TOPMed) consortium. *EBioMedicine* **63**, 103157 (2021).

38. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants: Fig. 1. *Bioinformatics* **31**, 3555–3557 (2015).

39. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front. Genet.* **11**, 157 (2020).

40. Quintero, A. *et al.* ShinyButchR: Interactive NMF-based decomposition workflow of

genome-scale datasets. *Biol. Methods Protoc.* **5**, bpaa022 (2020).

41. Udler, M. S. *et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLOS Med.* **15**, e1002654 (2018).

42. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

43. BioMe BioBank Program | Icahn School of Medicine. *Icahn School of Medicine at Mount Sinai* https://icahn.mssm.edu/research/ipm/programs/biome-biobank.

44. Belbin, G. M. *et al.* Toward a fine-scale population health monitoring system. *Cell* **184**, 2068-2083.e11 (2021).

45. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

46. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

47. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

48. Quality Control (QC) | Pan UKBB. https://pan-dev.ukbb.broadinstitute.org/docs/qc.

49. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2013).

50. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111 (2013).

51. Nadkarni, G. N. *et al.* Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu. Symp. Proc. AMIA Symp.* **2014**, 907–916 (2014).

52. Ben-Shachar, M., Lüdecke, D. & Makowski, D. effectsize: Estimation of Effect Size

Indices and Standardized Parameters. *J. Open Source Softw*. **5**, 2815 (2020).

Acknowledgment

We would like to thank Miriam Udler, who kindly provided us with the script of the bNMF algorithm, which was used in this study to confirm the clusters identified by NMF. This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Additionally, this work was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880, which allowed us to use Mount Sinai Data Warehouse (MSDW) data. Regarding HPI.MS resources, funding was provided by the Hasso Plattner Foundation (HPF). Additionally, the research leading to these results has received funding from the Horizon 2020 Programme of the European Commission under Grant Agreement No. 826117 (Smart4Health). The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai BioMe Biobank. We also thank all of our recruiters who have assisted in data collection and management, and we are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Contributions

Conceptualization, C.S., E.B.; Methodology, C.S., S.I.; Data analysis, A.E., S.I.; Writing of the original draft, A.E.; Tables and Figures, A.E.; Supervision, S.I., H.O.H., E.B.; Project administration, E.B.; Funding acquisition, E.B., H.O.H. All authors reviewed and edited the manuscript. All authors have read, discussed, and approved the manuscript, its analyses, and interpretations.

Data availability statement

All publicly available data (input variants, trait-variant associations) used to support the findings of this study are included in this published article (and its Supplementary Information files) and are also available from the cited publications and GWAS Catalog. Additional data generated for the analysis steps, including source code and intermediate results, are available from the corresponding author upon reasonable request. The data used to validate the findings of this study are available from BioMe biobank (https://icahn.mssm.edu/research/ipm/programs/biome-biobank), but restrictions apply to their availability. To access the data, please reach out to biomebiobank@mssm.edu.

Additional Information

Competing Interests Statement

Claudia Schurmann is a paid employee of Bayer AG, Pharmaceuticals. All other authors do not have any competing interest.

Figure Legends

Figure 1 - Top seven CKD-associated secondary traits per cluster (also available as LaTeX code)

The top seven secondary traits per cluster are shown with their effect direction (Effect columns) and respective cluster weights (Activity columns). 'HDL' is high-density lipoprotein, 'VLDL' is very low-density lipoprotein, 'meas.' is measurement, and 'conc.' is concentration.

Figure 2 - Replication of cluster traits (also available as LaTeX code)

The table lists traits replicated with the PheWAS on the ALL cohort for each cluster. 'Dir' is the trait effect direction, 'activity' is the trait cluster weight, 'OR' is the standardized odds ratio (binary traits cPGS-PheWAS), 'Coeff' is the standardized coefficient estimate (quantitative traits cPGS-PheWAS), '95% CI' are the 95% confidence intervals. The last column specifies whether the p-value reaches the Bonferroni significance level. 'HDL' is high-density lipoprotein, 'VLDL' is very low-density lipoprotein, 'RAS' is the renin-angiotensin system, 'meas.' is measurement, and 'conc.' is concentration. Regarding the cluster names, IBM is 'increased body mass,' and BV is the short version for 'blood volume.'

Figure 3 - Summary statistics of the weights of each cluster (also available as LaTeX code)

'SNPs' indicates the number of CKD variants with a weight > 0. The minimum weight in all clusters is 1e-45. 'Q_{90} weight' is the minimum weight of the SNPs in the cluster's top decile.

Figure 4 - Standardized cluster-specific polygenic scores (cPGS) per genetic population

The figure compares the standardized cPGS distributions between inferred ancestries of the Bio*Me* participants. The x-axis represents the units of standard deviation (or z-scores). AFR, AMR, and EUR refer to the sub-cohorts of individuals with inferred African, Ad Mixed American, and European ancestry, respectively.

Figure 5 - Methods overview

a, We selected 322 independent CKD-associated variants from the summary statistics of published GWAS. For each of them, we retrieved all independent proxy SNPs in linkage disequilibrium (r^2 >=

0.6) and (from the GWAS Catalog) 229 proxy-associated traits with their respective effect size (B). **b**, We standardised the effect sizes across all GWAS (β) and generated an association matrix X of dimensions 229x322. NMF factorizes X into a matrix of traits (W) and one of variants (H), which share a dimension K (i.e., the number of clusters). **c**, We extracted the weights of each cluster from the H matrix and used them to calculate cluster-specific polygenic scores (cPGS) of 31.701 Bio*Me* participants. After standardizing the cPGS, we conducted a cPGS-PheWAS for each cluster to validate their respective top traits, which were extracted from the W matrix.