Cell Reports Medicine, Volume *4*

Supplemental information

Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults

Ruizhi Zheng, Yu Xu, Mian Li, Zhengnan Gao, Guixia Wang, Xinguo Hou, Li Chen, Yanan Huo, Guijun Qin, Li Yan, Qin Wan, Tianshu Zeng, Lulu Chen, Lixin Shi, Ruying Hu, Xulei Tang, Qing Su, Xuefeng Yu, Yingfen Qin, Gang Chen, Xuejiang Gu, Feixia Shen, Zuojie Luo, Yuhong Chen, Yinfei Zhang, Chao Liu, Youmin Wang, Shengli Wu, Tao Yang, Qiang Li, Yiming Mu, Jiajun Zhao, Chunyan Hu, Xiaojing Jia, Min Xu, Tiange Wang, Zhiyun Zhao, Shuangyuan Wang, Hong Lin, Guang Ning, Weiqing Wang, Jieli Lu, Yufang Bi, and for the China Cardiometabolic Disease and Cancer Cohort (4C) Study Group

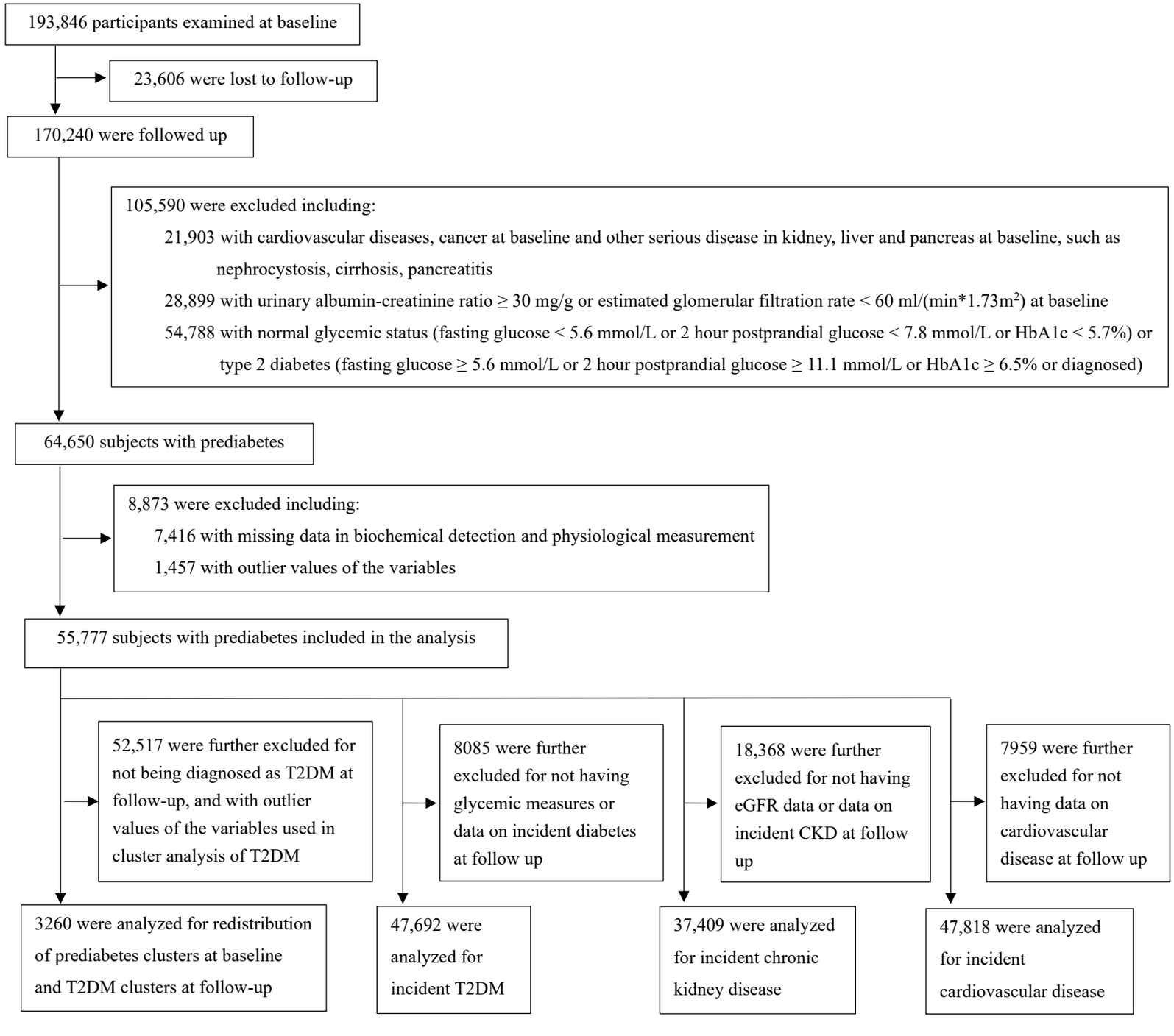
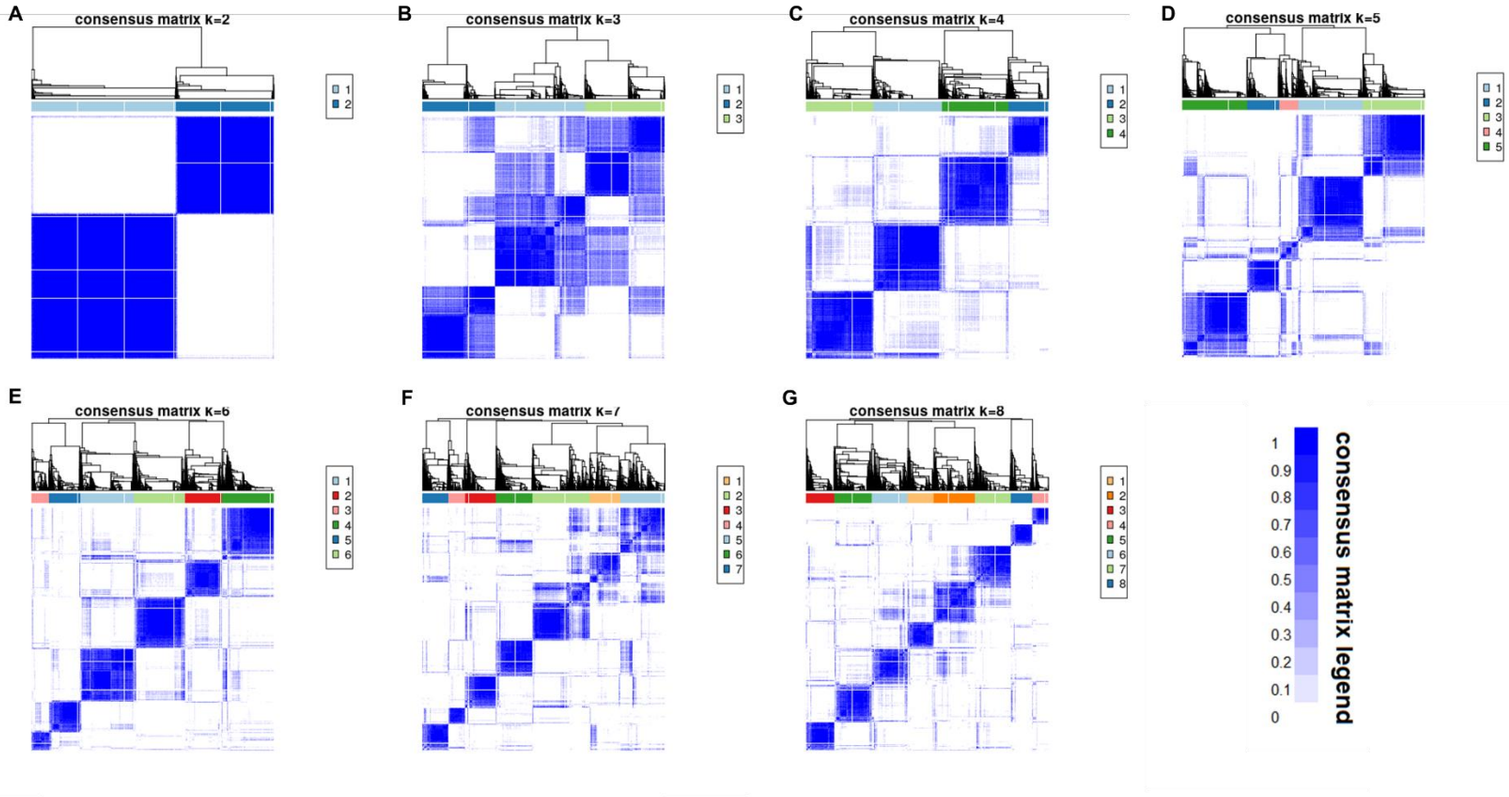
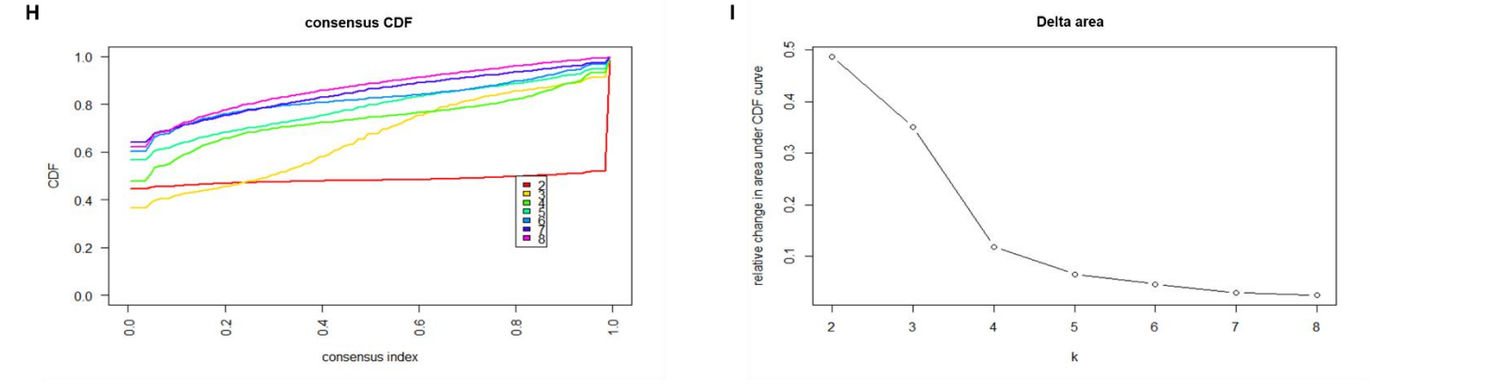


Figure S1. Participants selection flow diagram, related to Table 1


A
consensus matrix k=2
B
consensus matrix k=3
C
consensus matrix k=4
D
consensus matrix k=5
E
consensus matrix k=6
F
consensus matrix k=7
G
consensus matrix k=8
consensus matrix legend


H
consensus CDF
CDF
consensus index
I
Delta area
relative change in area under CDF curve
k

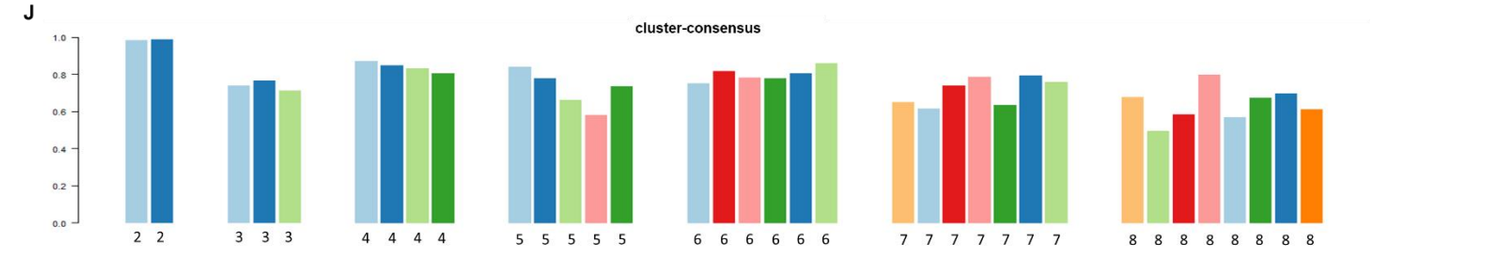

J
cluster-consensus

Figure S2 A-J. Consensus matrix heatmaps using diabetes related factors, consensus cumulative distribution function and cluster consensus score to determine at what number of clusters, related to STAR Methods

(A) *K*=2; (B) *K*=3; (C) *K*=4; (D) *K*=5; (E) *K*=6; (F) *K*=7; (G) *K*=8; (H) The lines by colors indicating the cumulative distribution functions (CDF) of the consensus matrix for each number of clusters; (J) The mean consensus score for different numbers of clusters (*K* ranges from 2 to 8). For *K* = 6, the mean consensus score was 0.75 for cluster 1, 0.82 for cluster 2, 0.78 for cluster 3, 0.78 for cluster 4, 0.80 for cluster 5, and 0.86 for cluster 6.

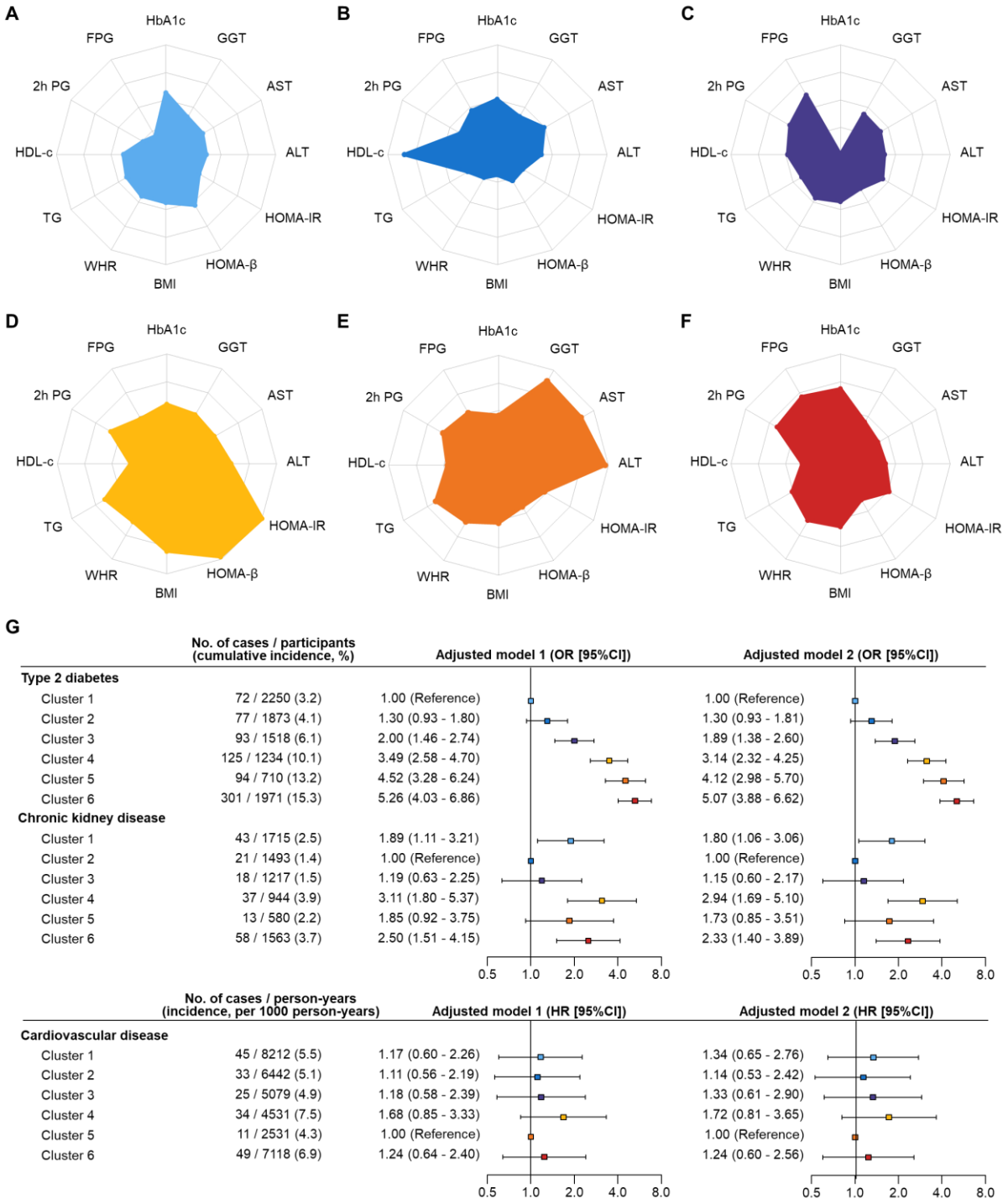


Figure S3 A-G. Characteristics and disease risks of the random sample of participants in diabetes at baseline used to perform consensus clustering algorithm by clusters (n = 11 155), related to STAR Methods

(A) cluster 1; (B) cluster 2; (C) cluster 3; (D) cluster 4; (E) cluster 5; (F) cluster 6; (G) Comparison of incident type 2 diabetes, chronic kidney diseases, and cardiovascular diseases between clusters.

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

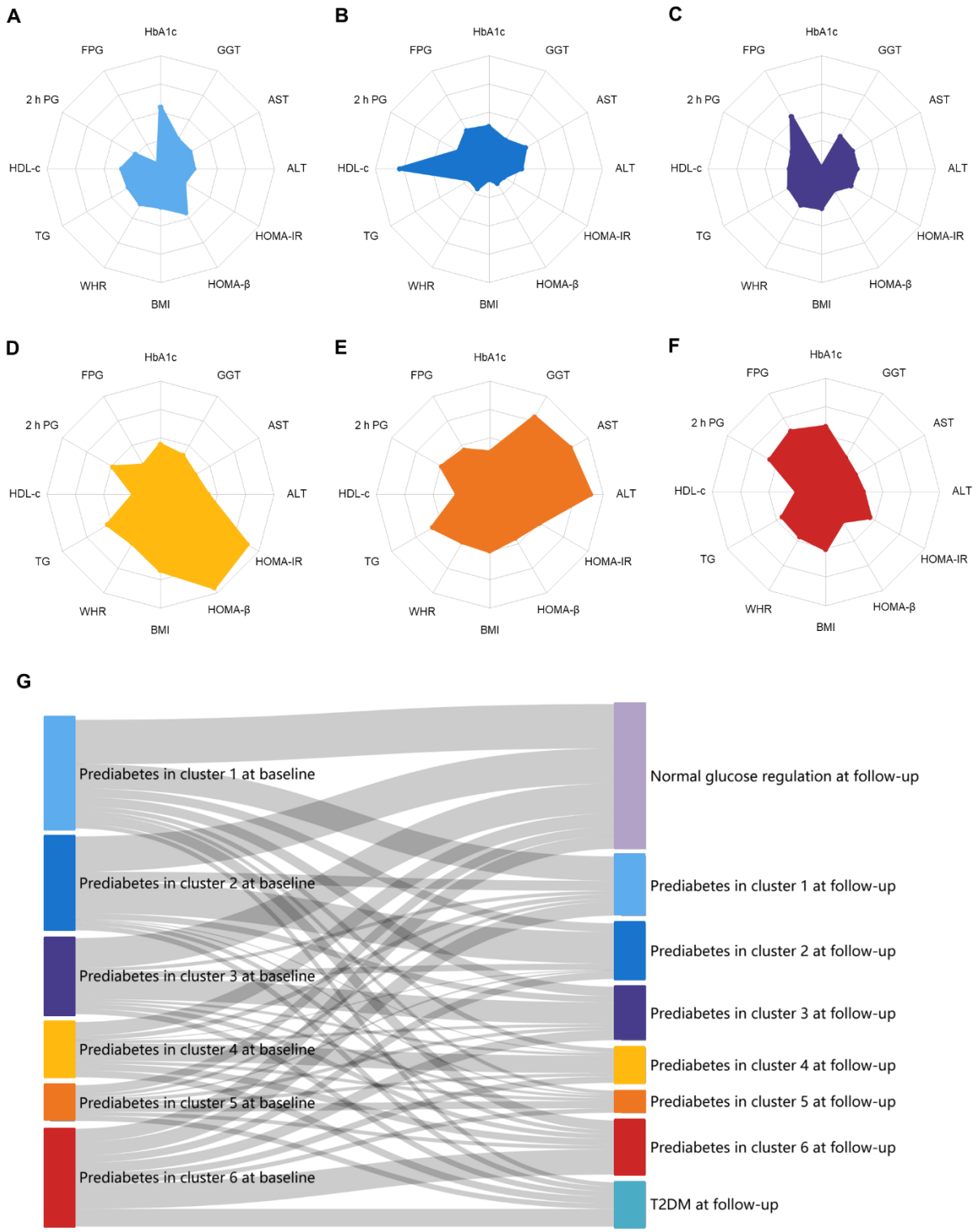


Figure S4 A-G. Distribution of the cluster feature variables among the individuals with prediabetes at the follow-up (A-F) and cluster migration pattern from prediabetes clusters at baseline to normal glucose regulation, prediabetes clusters and type 2 diabetes clusters at follow-up (G), related to Figure 5

(A) cluster 1; (B) cluster 2; (C) cluster 3; (D) cluster 4; (E) cluster 5; (F) cluster 6

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose;

HOMA-β, homoeostasis model assessment of β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

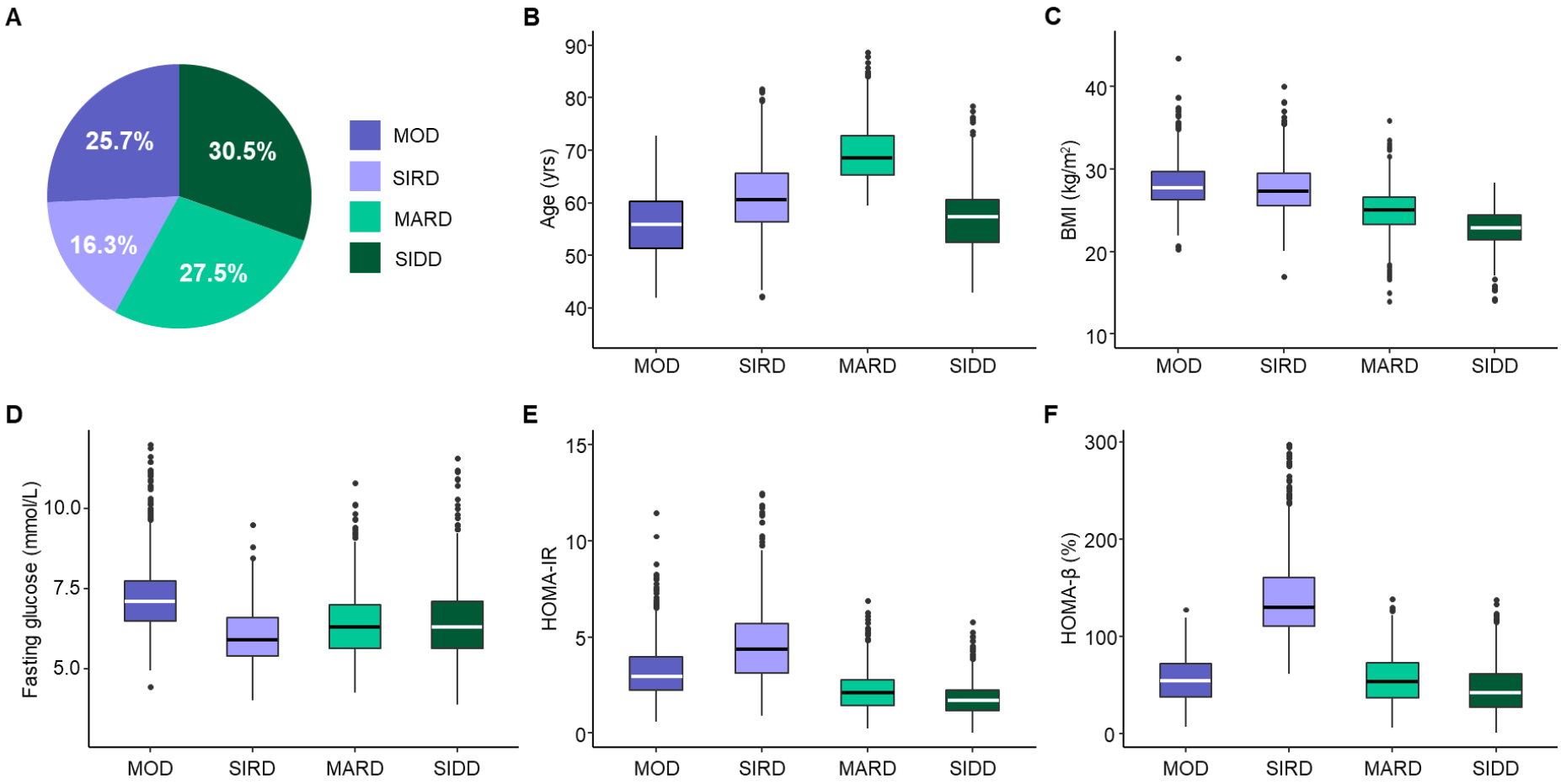


Figure S5 A-F. Characteristics of the subjects developed type 2 diabetes at follow-up clustered by using Ahlqvist-diabetes-classes, related to Figure 5

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabete

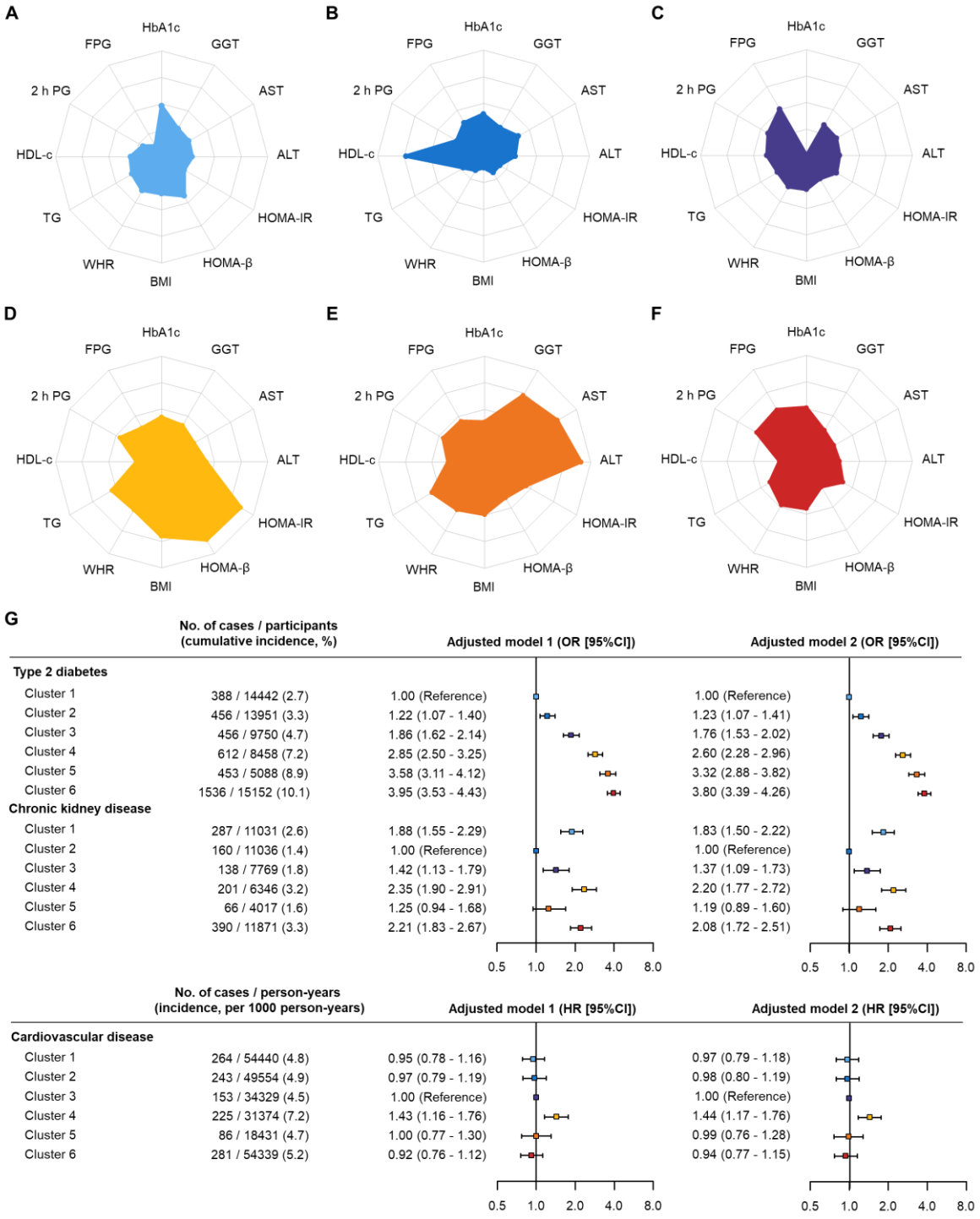


Figure S6 A-G. Characteristics and disease risks of the participants without diabetes defined as fasting plasma glucose < 7.0 mmol/L, 2-hour post-load plasma glucose < 11.1 mmol/L, HbA1c < 6.5% and having not been diagnosed with diabetes at baseline by clusters, related to Figure 2 and 4

(A) cluster 1; (B) cluster 2; (C) cluster 3; (D) cluster 4; (E) cluster 5; (F) cluster 6; (G) Comparison of incident type 2 diabetes, chronic kidney diseases and cardiovascular diseases between clusters.

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

Table S1. Cluster centers using K-means clustering by analysis datasets, related to Figure 2

	WHR	BMI	FPG	2 h PG	HbA1c	HOMA-IR	HOMA-β	ALT	AST	GGT	TG	HDL-c
Clustering analysis in participants with prediabetes at baseline												
Cluster 1	-0.033	-0.103	-0.855	-0.618	0.301	-0.437	0.085	-0.307	-0.273	-0.238	-0.186	-0.316
Cluster 2	-0.695	-0.849	-0.116	-0.307	0.057	-0.664	-0.612	-0.252	0.009	-0.217	-0.504	1.138
Cluster 3	-0.115	-0.175	0.412	0.091	-1.396	-0.172	-0.397	-0.211	-0.168	-0.157	-0.206	-0.027
Cluster 4	0.396	0.961	-0.071	0.232	0.139	1.453	1.449	0.109	-0.08	0.03	0.451	-0.411
Cluster 5	0.396	0.378	0.197	0.235	-0.06	0.263	0.104	1.659	1.217	1.045	0.622	-0.13
Cluster 6	0.289	0.214	0.747	0.675	0.627	0.108	-0.331	-0.197	-0.233	-0.122	0.08	-0.359
Clustering analysis in participants without diabetes at baseline												
Cluster 1	-0.047	-0.098	-0.788	-0.533	0.299	-0.402	0.137	-0.274	-0.252	-0.223	-0.181	-0.260
Cluster 2	-0.682	-0.815	-0.055	-0.271	0.043	-0.620	-0.604	-0.247	-0.011	-0.217	-0.484	1.054
Cluster 3	-0.119	-0.185	0.374	0.110	-1.143	-0.163	-0.386	-0.208	-0.169	-0.155	-0.200	-0.022
Cluster 4	0.419	0.965	-0.038	0.219	0.121	1.441	1.409	0.120	-0.082	0.044	0.479	-0.436
Cluster 5	0.415	0.346	0.188	0.205	-0.009	0.212	0.030	1.583	1.230	1.028	0.579	-0.115
Cluster 6	0.282	0.202	0.536	0.487	0.371	0.045	-0.304	-0.212	-0.248	-0.137	0.046	-0.385
Clustering analysis in participants with prediabetes at the follow-up												
Cluster 1	-0.081	-0.145	-1.009	-0.385	0.481	-0.433	0.194	-0.270	-0.246	-0.244	-0.174	-0.120
Cluster 2	-0.545	-0.870	0.018	-0.221	-0.035	-0.700	-0.711	-0.281	-0.033	-0.278	-0.574	1.203
Cluster 3	-0.050	-0.103	0.449	-0.256	-1.149	-0.260	-0.497	-0.215	-0.211	-0.166	-0.158	-0.297
Cluster 4	0.335	0.860	-0.263	0.295	0.169	1.502	1.693	0.105	-0.099	0.034	0.453	-0.447
Cluster 5	0.305	0.351	0.204	0.317	-0.038	0.371	0.186	1.516	1.306	1.212	0.598	-0.293
Cluster 6	0.221	0.373	0.698	0.552	0.573	0.199	-0.252	-0.159	-0.244	-0.111	0.177	-0.384

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

Table S2. Cohen's d estimates of cluster comparisons of variables in prediabetes at baseline, related to Figure 3

Comparison groups	WHR	BMI	HOMA-IR	HOMA-β	ALT	AST	GGT	HbA1c	FPG	2 h PG	HDL-c	TG
Cluster 1 vs Cluster 2	0.841	1.020	0.466	1.100	-0.114	-0.550	-0.051	0.364	-0.948	-0.371	-1.790	0.618
Cluster 1 vs Cluster 3	0.103	0.094	-0.486	0.730	-0.200	-0.215	-0.195	2.460	-1.740	-0.842	-0.375	0.032
Cluster 1 vs Cluster 4	-0.535	-1.290	-2.710	-1.680	-0.763	-0.384	-0.623	0.228	-0.990	-1.010	0.125	-0.831
Cluster 1 vs Cluster 5	-0.541	-0.609	-1.180	-0.026	-2.980	-2.370	-1.880	0.518	-1.380	-1.020	-0.244	-1.030
Cluster 1 vs Cluster 6	-0.402	-0.409	-0.966	0.671	-0.233	-0.085	-0.283	-0.518	-2.150	-1.580	0.056	-0.383
Cluster 2 vs Cluster 3	-0.734	-0.938	-0.884	-0.423	-0.084	0.332	-0.129	1.860	-0.644	-0.434	1.350	-0.609
Cluster 2 vs Cluster 4	-1.370	-2.310	-2.930	-2.920	-0.637	0.161	-0.511	-0.102	-0.050	-0.585	1.810	-1.400
Cluster 2 vs Cluster 5	-1.390	-1.670	-1.520	-1.220	-2.740	-1.750	-1.670	0.144	-0.356	-0.580	1.440	-1.620
Cluster 2 vs Cluster 6	-1.240	-1.440	-1.340	-0.596	-0.115	0.475	-0.210	-0.800	-1.040	-1.110	1.780	-0.957
Cluster 3 vs Cluster 4	-0.633	-1.360	-2.020	-2.470	-0.551	-0.168	-0.382	-1.810	0.572	-0.151	0.476	-0.814
Cluster 3 vs Cluster 5	-0.641	-0.699	-0.615	-0.804	-2.560	-2.010	-1.510	-1.550	0.263	-0.151	0.125	-0.981
Cluster 3 vs Cluster 6	-0.502	-0.503	-0.438	-0.136	-0.030	0.135	-0.078	-2.740	-0.428	-0.658	0.415	-0.401
Cluster 4 vs Cluster 5	<-0.001	0.652	1.290	1.520	-1.860	-1.780	-1.200	0.219	-0.289	-0.003	-0.347	-0.162
Cluster 4 vs Cluster 6	0.132	0.889	1.710	2.590	0.545	0.307	0.320	-0.642	-0.962	-0.498	-0.066	0.428
Cluster 5 vs Cluster 6	0.133	0.204	0.221	0.775	2.700	2.270	1.580	-0.909	-0.663	-0.491	0.286	0.600

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

Table S3. Comparisons of incident type 2 diabetes, chronic kidney diseases, and cardiovascular disease between clusters by using different clusters as reference groups, related to Figure 4

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Type 2 diabetes						
Reference: cluster 1	1.00 (reference)	1.44 (1.24 - 1.66)	1.80 (1.55 - 2.08)	3.00 (2.61 - 3.45)	3.69 (3.18 - 4.28)	4.93 (4.37 - 5.57)
Reference: cluster 2	0.70 (0.60 - 0.81)	1.00 (reference)	1.25 (1.09 - 1.44)	2.09 (1.83 - 2.39)	2.57 (2.20 - 2.97)	3.44 (3.06 - 3.85)
Reference: cluster 3	0.56 (0.48 - 0.64)	0.80 (0.69 - 0.92)	1.00 (reference)	1.67 (1.46 - 1.90)	2.05 (1.78 - 2.36)	2.74 (2.45 - 3.08)
Reference: cluster 4	0.33 (0.29 - 0.38)	0.48 (0.42 - 0.55)	0.60 (0.53 - 0.68)	1.00 (reference)	1.23 (1.07 - 1.41)	1.64 (1.48 - 1.82)
Reference: cluster 5	0.27 (0.23 - 0.32)	0.39 (0.34 - 0.45)	0.49 (0.42 - 0.56)	0.81 (0.71 - 0.93)	1.00 (reference)	1.34 (1.19 - 1.51)
Reference: cluster 6	0.20 (0.18 - 0.23)	0.29 (0.26 - 0.33)	0.37 (0.33 - 0.41)	0.61 (0.55 - 0.68)	0.75 (0.66 - 0.84)	1.00 (reference)
Chronic kidney disease						
Reference: cluster 1	1.00 (reference)	0.55 (0.44 - 0.69)	0.73 (0.58 - 0.92)	1.26 (1.01 - 1.54)	0.69 (0.51 - 0.93)	1.22 (1.00 - 1.45)
Reference: cluster 2	1.82 (1.50 - 2.27)	1.00 (reference)	1.33 (1.00 - 1.73)	2.29 (1.80 - 2.91)	1.25 (0.91 - 1.74)	2.22 (1.81 - 2.75)
Reference: cluster 3	1.36 (1.10 - 1.71)	0.75 (0.58 - 0.99)	1.00 (reference)	1.72 (1.30 - 2.19)	0.94 (0.68 - 1.30)	1.66 (1.32 - 2.07)
Reference: cluster 4	0.79 (0.65 - 0.98)	0.44 (0.34 - 0.56)	0.58 (0.46 - 0.74)	1.00 (reference)	0.55 (0.40 - 0.75)	0.97 (0.80 - 1.18)
Reference: cluster 5	1.45 (1.08 - 1.96)	0.80 (0.58 - 1.10)	1.06 (0.77 - 1.47)	1.83 (1.34 - 2.49)	1.00 (reference)	1.77 (1.32 - 2.37)
Reference: cluster 6	0.82 (0.69 - 0.97)	0.45 (0.36 - 0.56)	0.60 (0.48 - 0.75)	1.03 (0.85 - 1.25)	0.57 (0.42 - 0.76)	1.00 (reference)
Cardiovascular disease						
Reference: cluster 1	1.00 (reference)	1.00 (0.81 - 1.23)	1.05 (0.84 - 1.31)	1.47 (1.19 - 1.81)	1.02 (0.77 - 1.36)	1.04 (0.86 - 1.26)
Reference: cluster 2	1.00 (0.82 - 1.23)	1.00 (reference)	1.05 (0.83 - 1.33)	1.47 (1.18 - 1.83)	1.03 (0.76 - 1.38)	1.04 (0.85 - 1.28)
Reference: cluster 3	0.95 (0.76 - 1.19)	0.95 (0.75 - 1.20)	1.00 (reference)	1.40 (1.10 - 1.77)	0.97 (0.72 - 1.32)	0.99 (0.79 - 1.24)
Reference: cluster 4	0.68 (0.56 - 0.84)	0.68 (0.55 - 0.85)	0.72 (0.57 - 0.91)	1.00 (reference)	0.70 (0.52 - 0.94)	0.71 (0.58 - 0.87)
Reference: cluster 5	0.98 (0.73 - 1.30)	0.98 (0.73 - 1.31)	1.03 (0.76 - 1.41)	1.43 (1.07 - 1.93)	1.00 (reference)	1.01 (0.76 - 1.35)
Reference: cluster 6	0.96 (0.79 - 1.17)	0.96 (0.78 - 1.18)	1.01 (0.81 - 1.27)	1.41 (1.15 - 1.74)	0.99 (0.74 - 1.31)	1.00 (reference)

The comparisons of incident type 2 diabetes and chronic kidney diseases between clusters used logistic regression and odds ratios were presented. The comparisons of cardiovascular disease between clusters used cox proportional hazards model and hazard ratios were presented.

Table S4. Transitions from prediabetes clusters at baseline to normal glucose regulation, prediabetes clusters and Ahlqvist-diabetes-classes at follow-up, related to Figure 5

Prediabetes at baseline	NGR at follow-up	Prediabetes at follow-up						T2DM at follow-up
		Cluster 1 at follow-up	Cluster 2 at follow-up	Cluster 3 at follow-up	Cluster 4 at follow-up	Cluster 5 at follow-up	Cluster 6 at follow-up	
Cluster 1 at baseline	3263 (40.7%)	1791 (22.4%)	641 (8.0%)	697 (8.7%)	338 (4.2%)	246 (3.1%)	718 (9.0%)	315 (3.9%)
Cluster 2 at baseline	2596 (37.5%)	741 (10.7%)	2324 (33.6%)	451 (6.5%)	54 (0.8%)	106 (1.5%)	288 (4.2%)	361 (5.2%)
Cluster 3 at baseline	2200 (39.3%)	253 (4.5%)	455 (8.1%)	1586 (28.3%)	180 (3.2%)	190 (3.4%)	343 (6.1%)	393 (7.0%)
Cluster 4 at baseline	987 (24.2%)	326 (8.0%)	41 (1.0%)	213 (5.2%)	1279 (31.4%)	229 (5.6%)	495 (12.2%)	501 (12.3%)
Cluster 5 at baseline	643 (24.4%)	220 (8.4%)	108 (4.1%)	242 (9.2%)	235 (8.9%)	528 (20.1%)	274 (10.4%)	381 (14.5%)
Cluster 6 at baseline	962 (13.3%)	1014 (14.1%)	528 (7.3%)	748 (10.4%)	471 (6.5%)	327 (4.5%)	1854 (25.7%)	1309 (18.1%)

Prediabetes at baseline	T2DM at follow-up			
	MOD	SIRD	MARD	SIDD
Cluster 1 at baseline	59 (18.7%)	33 (10.5%)	99 (31.4%)	124 (39.4%)
Cluster 2 at baseline	26 (7.2%)	14 (3.9%)	89 (24.7%)	232 (64.3%)
Cluster 3 at baseline	73 (18.6%)	31 (7.9%)	107 (27.2%)	182 (46.3%)
Cluster 4 at baseline	155 (30.9%)	240 (47.9%)	70 (14.0%)	36 (7.2%)
Cluster 5 at baseline	142 (37.3%)	97 (25.5%)	80 (21.0%)	62 (16.3%)
Cluster 6 at baseline	383 (29.3%)	117 (8.9%)	452 (34.5%)	357 (27.3%)

The number of prediabetes who had taken follow-up examination were 8009, 6921, 5600, 4071, 2631 and 7213 in prediabetes cluster 1 to cluster 6 after excluding those for having missing data in biochemical detection and physiological measurement, and outlier values of the variables.

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabetes, NGR, normal glucose regulation.

Table S5. Ahlqvist-diabetes-classes for the subjects with type 2 diabetes at follow up, related to Figure 5

Clusters	Men		Women		Overall	
	N	%	N	%	N	%
MOD	358	29.8	480	23.3	838	25.7
SIRD	200	16.7	332	16.1	532	16.3
MARD	336	28.0	561	27.2	897	27.5
SIDD	306	25.5	687	33.3	993	30.5

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabetes

Table S6. Cluster center using K-means clustering in subjects with type 2 diabetes at follow up, related to Figure 5

	Age	BMI	FPG	HOMA-IR	HOMA-β
Men					
MOD	-0.831	0.428	0.430	0.098	-0.287
SIRD	-0.144	0.718	-0.356	0.936	1.619
MARD	1.032	0.086	-0.178	-0.301	-0.231
SIDD	-0.070	-1.076	-0.327	-0.643	-0.521
Women					
MOD	-0.396	0.935	0.398	0.289	-0.209
SIRD	0.081	0.524	-0.455	0.893	1.731
MARD	1.064	-0.334	-0.164	-0.343	-0.271
SIDD	-0.635	-0.644	-0.061	-0.497	-0.478

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabetes; BMI, body mass index; FPG, fasting plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance

Methods S1. Details of specimen testing and data collection, related to Table 1

Fasting and post-load glucose concentrations were measured at local hospitals using the glucose oxidase or hexokinase method. Triglyceride (TG), total cholesterol (TC), low density lipoprotein-cholesterol (LDL-c), high density lipoprotein-cholesterol (HDL-c), aspartate aminotransferase (AST), alanine transaminase (ALT), and glutamyl transferase (GGT) were tested using an auto-analyser (ARCHITECT ci16200, Abbott Laboratories, Abbott Park, IL) at the central laboratory in the Shanghai Institute of Endocrine and Metabolic Diseases (certified by the National Glycohemoglobin Standardization Program and the College of American Pathologists Laboratory Accreditation Program). Finger capillary whole-blood samples were collected using the Hemoglobin Capillary Collection System (Bio-Rad Laboratories, Hercules, CA, USA) and were shipped and stored at 2°C to 8°C. HbA1c was measured within 4 weeks of blood collection by high-performance liquid chromatography using the VARIANT II Hemoglobin Testing System (BioRad Laboratories) at the central laboratory Serum insulin was measured by an autoanalyser (ARCHITECT ci16200, Abbott Laboratories, Chicago, IL, USA). Urinary albumin concentrations were measured at the central laboratory by immunonephelometry using Siemens BNII nephelometers (Siemens Healthcare Diagnostics, Marburg, Germany). The lower limit of detection is 2.13 mg/L. The intra-assay and inter-assay coefficients of variation for urinary albumin were 2.1% and 2.3%, respectively. Urinary creatinine concentrations were measured at the central laboratory by an enzymatic method (ADVIA Chemistry XPT System; Siemens Healthcare, Erlangen, Germany). The intra-assay and inter-assay coefficients of variation for urinary creatinine were 1.1% and 1.3%, respectively. Albuminuria was assessed using albumin-to-creatinine ratio (ACR) based on morning spot urine. Insulin resistance was estimated by the homoeostasis model assessment of insulin resistance (HOMA-IR) index: fasting insulin (μU/mL) × fasting glucose (mmol/L)/22.5 [1]. β-cell function was estimated by the homoeostasis model assessment β of cell function (HOMA-β) index: (20 × fasting insulin [μU/mL])/(fasting glucose [mmol/L] – 3.5) [1].

Blood pressure was measured by the trained staff. Before blood pressure measurement, participants were advised to avoid alcohol, coffee, tea, smoking, and exercise at least 30 minutes. The appropriate cuff was used depending on the subject's arm circumference. An automated electronic device (OMRON Model HEM-725 FUZZY, Omron Company, Dalian, China) was used to measure blood pressure of seated participants three times consecutively at 1-min intervals after a ≥ 5-min rest. The three readings were averaged for analysis.

Standardized questionnaires were used to collect participants' demographic characteristics, dietary, and lifestyle risk factors. Education level was classified by using 9 years of education as the cutoff. Current smoking

was defined as having smoked at least 1 cigarette per day for the past 6 months. Current drinking was defined as drinking alcohol at least once a week for the past 6 months. The International Physical Activity Questionnaire was used to assess physical activity [2]. Moderate and vigorous physical activity was defined as ≥ 150 min/week of moderate-intensity physical activity, or 75 min/week of vigorous aerobic activity, or an equivalent combination of moderate-intensity and vigorous aerobic activities. A food frequency questionnaire was used to collect habitual dietary intake by asking the consumption frequency and portion size of typical food items during the previous 12 months, and a dietary quality score was categorized as high (≥ 4.5 cups per day) and low (< 4.5 cups per day) based on the intake of fruits and vegetables. Nighttime sleep duration was defined as the time space between bedding and waking up.

Methods S2. Consensus clustering algorithm, related to STAR Methods

Consensus clustering is a method of unsupervised cluster has been widely used for high-dimensional data [3]. The clustering algorithm is to maximize the number of clusters meanwhile maintaining high cluster consensus. We set a prespecified number of clusters K=2, 3, …, 8, for each number of clusters, the consensus clustering algorithm created a random subset that included 80% of the data records without replacement and repeated 100 times. For each random subset, K-means (Euclidean distance based) algorithm was performed and each individual was assigned to one of the clusters. After running 100 times, the frequencies of any pair of two individuals were calculated, which were clustered together under each scenario of K and constructed a $N \times N$ matrix of participants' pairwise consensus value (N is the sample size). The cluster membership was determined by applying a hierarchical clustering algorithm using the consensus matrix as a measure of similarity. In the consensus matrix, consensus values range from 0 (never clustered together) to 1 (always clustered together) were marked by white to bright blue. The consensus matrix is ordered by the consensus clustering which is displayed as a dendrogram atop the heatmap. The cluster memberships are marked by colored rectangles between the dendrogram and heatmap with a legend above the graphic.

The optimum number of clusters was ascertained by reviewing the consensus matrix heatmap, cumulative distribution function (CDF) (range 0 – 1) plot and the within-cluster consensus scores. The CDF plot showed the area under the CDFs for each K, and at what number of clusters, the CDF reached an approximate maximum, thus consensus and cluster confidence was at a maximum at this K. The relative change in area under the CDF curve comparing K and K – 1 also provide the suggestions of the optimum number of clusters. The cluster consensus score, ranged between 0 and 1, was defined as the average consensus value for all pairs of individuals belonging to the same cluster. A value closer to one indicated better cluster stability.

We performed consensus clustering analysis on the random sample containing 20% of the whole participants in prediabetes (n = 11 155). Consensus clustering analysis was done using the ConsensusClusterPlus function (maximum K = 8, replication = 100, proportion of random subset = 0.8, Euclidean distance-based K-means algorithm) in the ‘ConsensusClusterPlus’ package in R version 4.0.3 (http://www.r-project.org).

References

[1] Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985; 28: 412-9.

[2] Craig CL, Marshall AL, Sjostrom M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003; 35: 1381-95.

[3] Stefano MP, Tamayo; Jill, Mesirov; Todd, Golub. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning 2003; 52: 91-118.