Online Supplementary Material

Associations between dietary patterns and gene expression pattern in peripheral blood mononuclear cells: a cross-sectional study

Jacob J. Christensen, PhD[1,2], Stine M. Ulven, PhD[2], Magne Thoresen, PhD[3], Kenneth Westerman, PhD[4,5], Kirsten B. Holven, PhD[1,2], and Lene F. Andersen, PhD[2]

[1]Norwegian National Advisory Unit on Familial Hypercholesterolemia, Department of Endocrinology, Morbid Obesity and Preventive Medicine, Oslo University Hospital, Forskningsveien 2B, 0373 Oslo, Norway

[2]Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway

[3]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway

[4]Clinical and Translation Epidemiology Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

[5]Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA

Correspondence: Jacob J. Christensen, PhD, j.j.christensen@medisin.uio.no

Extended Supplementary Methods

In this section, we provide an extension of selected sections of the Methods.

Dietary assessment

We used a food-frequency questionnaire (FFQ) to assess habitual food intake from the preceding year (1). Daily intake of energy, nutrients and foods were computed using the food database AE-07 and KBS software system (version 4.9 2008; KBS), developed at the Department of Nutrition, University of Oslo, Norway. The food database AE-07 is based on the 2006 edition of the Norwegian food composition table supplemented with data from other tables and calculated recipes.

PMBC isolation, RNA extraction, and microarray analysis

We collected PBMCs and extracted RNA according to standardized protocols, as previously described (2). Briefly, we isolated PBMCs from whole blood after an overnight fasting (12 hours) using BD Vacutainer Cell Preparation Tubes (Becton Dickinson, New Jersey, US) according to the manufacturer's instructions. At the last step of the PMBC pipeline, we stored the PBMC pellets at −80 degrees Celsius until further RNA isolation. Next, we extracted total RNA using RNeasy mini kit (Qiagen, Hilden, Germany), treated it with DNase I (Qiagen) according to the manufacturer's instructions, and stored at −80 degrees Celsius. The quantity and quality of RNA were assessed using an ND-1000 Spectrophotometer (Saveen Werner, Carlson Circle, Florida, US) and an Agilent Bioanalyser (Agilent Technology, California, US), respectively. All samples had sufficiently high RIN values (> 8).

The microarray gene expression analyses followed standard Illumina protocol (Illumina Inc., CA, USA). In brief, cRNA was prepared with Ambion's Illumina® TotalPrep RNA Amplification Kit (Thermo Fisher Scientific, MA, USA), using 300 ng total RNA as input material. For each sample, the biotin-labelled cRNA concentrations were measured (NanoDrop, Thermo Fisher Scientific, MA, USA) and 750 ng hybridized overnight, to HumanHT-12 Expression BeadChips (Illumina Inc., CA, USA). After washing, the BeadChips were scanned with the Illumina HiScann instrument (Illumina Inc., CA, USA), according to the manufacturer's instructions. Illumina Genome Studio was used to transform bead level data to probe level intensity values, which were extracted using Illumina's BeadStudio software (gene expression module v3.0.19.0) (Illumina Inc., CA, USA) for bioinformatics analysis.

Following this step, all gene expression data pre-processing was performed in R version 3.6.0 (3). We used the `lumi` package (Bioconductor) to read, normalize and transform the data. There were no outliers based on principal component analysis (PCA), clustering analysis or by examining boxplots and density plots. We used the log2-quantile technique to normalize the samples, and we removed genes that were sufficiently expressed in fewer than 25 samples. Finally, we adjusted for batch effects using the `sva` package (Bioconductor). In total, we were left with 13967 mRNA transcripts for data analysis.

Gene expression clusters

We used WGCNA to identify highly correlated ("co-expressed") clusters of genes (4). The `WGCNA` package (CRAN, Bioconductor) provides a well-established and popular framework to perform the WGCNA analysis (5). The details of the implementation can be found in (5). First, we

determined the "soft thresholding power β" using the `WGCNA::pickSoftThreshold` function. This function creates a co-expression matrix and raises this to the power β to get the adjacency matrix. Balancing the approximate scale-free network properties and network connectivity, we chose β = 3 for both genders. Next, we used the high-level `WGCNA::blockwiseModules` function to create the gene expression clusters in blocks of 5000 mRNA using unsigned networks. Any genes that affiliated with a cluster with fewer than 20 members were assigned to the so-called *grey* cluster. Each cluster was then summarized using the first principal component (the "cluster eigengene"), and genes with low cluster membership were reassigned to another cluster. Finally, by default, cluster eigengenes that strongly correlated (r > 0.85) were merged to avoid redundancy.

To examine stability and validity of the resulting gene expression clusters between genders, we calculated module preservation statistics (6). To do this, we used the `WGCNA::modulePreservation` function; we used women's cluster affiliations as *reference*, and men's as *test*, and extracted the *median rank preservation* and *median rank quality*, as well as the corresponding Z scores. In addition, we extracted the actual cross-tabulation between women's and men's clusters, and the associated P values.

We performed gene ontology (GO) analyses to describe relevant WGCNA gene expression clusters biologically. The GO Consortium provides a comprehensive, computational model of biological systems, and is among the largest resources of gene-specific information (7,8). We used the `biomaRt::useMart(host = "http://jan2019.archive.ensembl.org", dataset = "hsapiens_gene_ensembl")` function to set up a connection to Ensembl, and then the

`biomarRt::getBM` to retrieve various gene annotation, including chromosome, start and end, strand, and GO identifier. We then created a background annotation object for our specific gene set (p = 13967 genes), and used this to compile *topGOdata* objects using the `topGO` package. We did this for all three GO classes: biological process (BP), cellular compartment (CC) and molecular function (MF). Finally, we ran enrichment tests on the *topGOdata* objects and compiled the results into data tables, using the high-level `topGO::runTest(algorithm = "classic", statistic = "fisher")` and `topGO::genTable` functions.

We performed protein-protein interaction (PPI) network analyses using The Protein Interaction Network Analysis (PINA) 2.0 database to link statistical findings with existing biological knowledge. We downloaded manually curated protein–protein interaction data from PINA (http://omics.bjcancer.org/pina/), and created networks based on input of a smaller set of driver genes defined in upstream analyses. Finally, to rank the importance of the proteins, we calculated and applied the betweenness centrality measure of nodes in the resulting networks, using the `tidygraph::centrality_betweenness` function.

Linear models

We report the *total effect* in the manuscript. In sensitivity analyses, we also used the minimal sufficient adjustment set for estimating the *direct effect* of dietary pattern on gene expression. This was considered to be: adiposity (total fat mass, measured by bioelectrical impedance analysis), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), homeostatic model assessment for insulin resistance (HOMA-IR), systolic blood pressure, use of anti-hypertensive medication (two levels: no, yes), kidney function (creatinine), liver function

(alanine aminotransferase, ALAT), familiy history of myocardial infarction (two levels: no, yes), age, and smoking (two levels: never/stopped, seldom/some/regularly). Results from the two types of models were similar (data not shown).

Miscellaneous

All plotting was performed using `tidyverse` tools, in particular `tidyr`, `dplyr` and `ggplot2`, and by use of the following key functions: correlations were calculated using `stats::cor(use = "pairwise.complete.obs", method = "spearman");` distance was calculated using `stats::dist (method = "euclidean");` clustering was performed using `stats::hclust (method = "complete");` dendrograms were transformed to networks using `tidygraph::as_tbl_graph` and the `ggraph` package.

Supplementary References

1. Carlsen MH, Lillegaard IT, Karlsen A, Blomhoff R, Drevon CA, Andersen LF. Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records. Nutrition Journal. 2010;9:37.

2. Ulven SM, Christensen JJ, Nygård O, Svardal A, Leder L, Ottestad I, Lysne V, Laupsa-Borge J, Ueland PM, Midttun Ø, et al. Using metabolic profiling and gene expression analyses to explore molecular effects of replacing saturated fat with polyunsaturated fat-a randomized controlled dietary intervention study. The American Journal of Clinical Nutrition. 2019;109:1239–50.

3. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

4. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005;4:Article17.

5. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC bioinformatics. 2008;9:559.

6. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, et al. Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. PLOS Genetics. 2006;2:e130.

7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene Ontology: Tool for the unification of biology. Nature genetics. 2000;25:25–9.

8. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research. 2019;47:D330–8.

Supplementary Tables

Due to document size constraints, all Supplementary Tables can be found in a separate Excel file.

Supplementary Table 1. *Population characteristics: food intake*.

Supplementary Table 2. *Population characteristics: energy percent*.

Supplementary Table 3. *Population characteristics: nutrient intake*.

Supplementary Table 4. *Loading scores for DP input variables*.

Supplementary Table 5. *Cluster size and variance explained*.

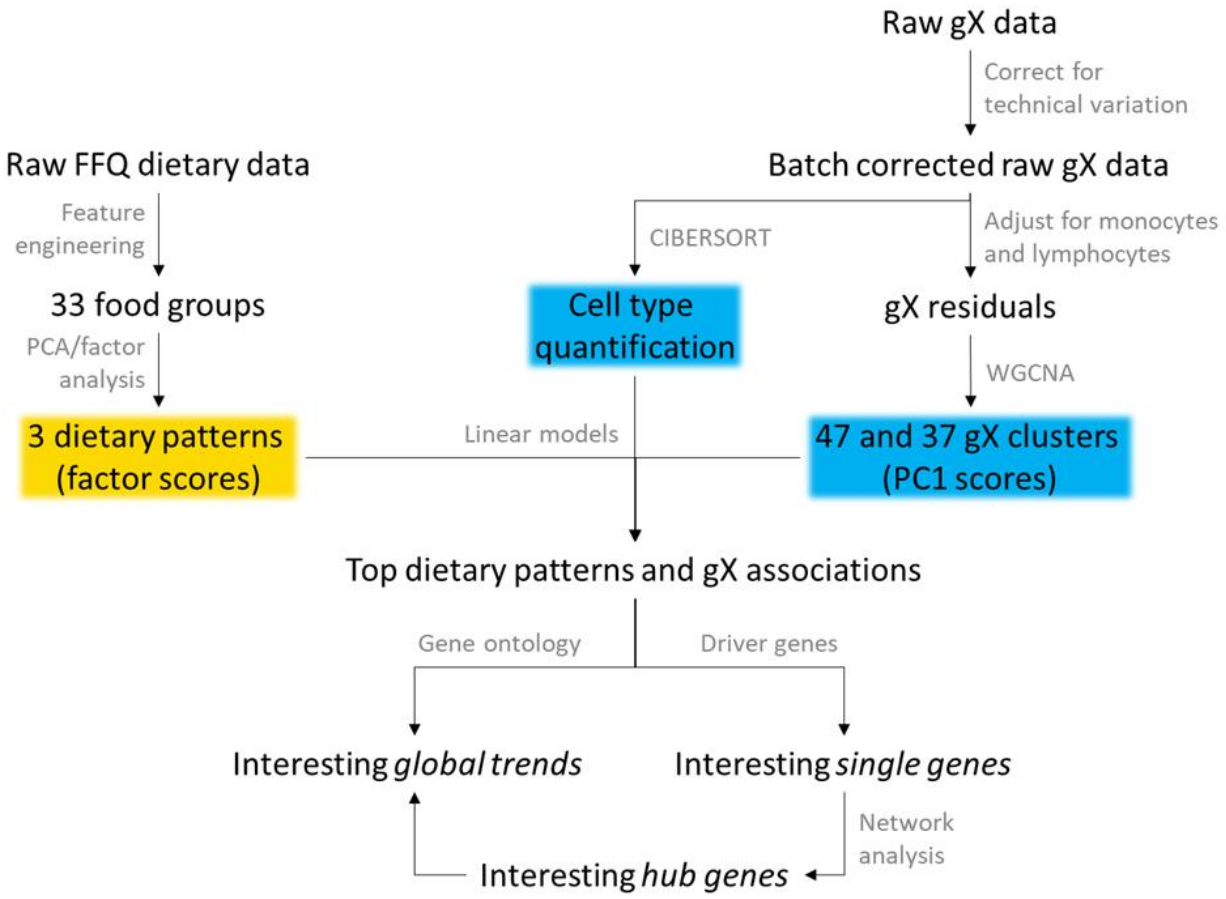
Supplementary Table 6. *Gene ontology terms for for key associations between DPs and WGCNA gene expression clusters*.

Supplementary Table 7. *Driver genes: genes that show strong association with both a DP and a gene expression cluster*.

Supplementary Table 8. *Hub proteins: proteins that physically interact and mediate signals in protein-protein interaction networks*.

Supplementary Figures

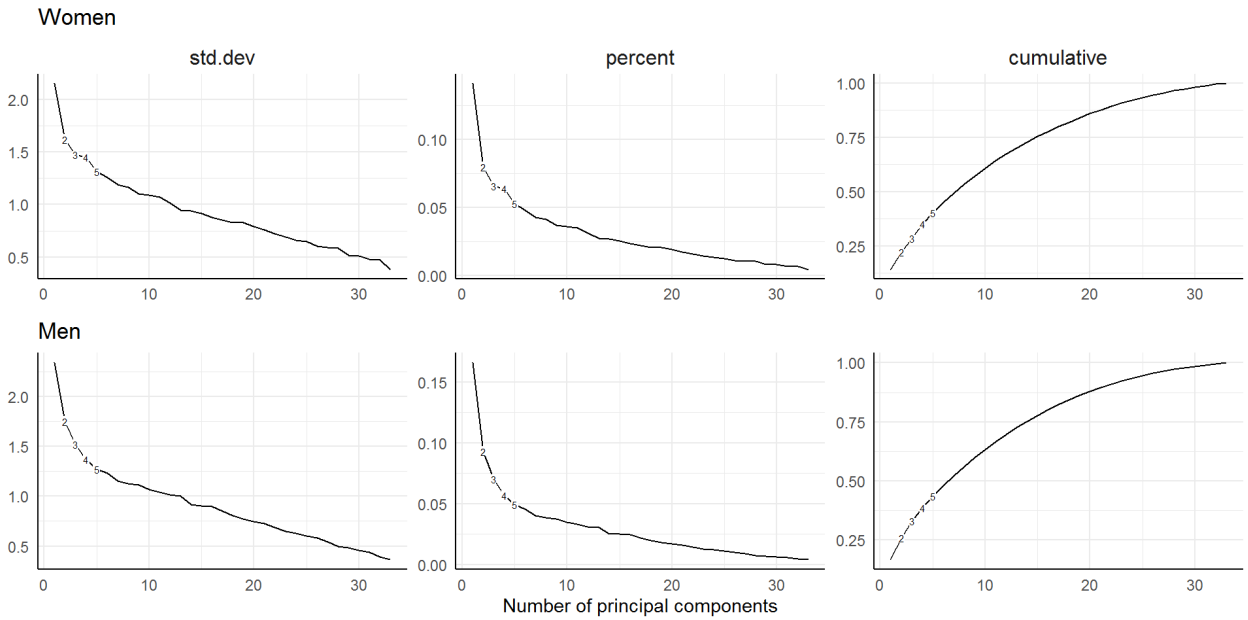
Supplementary Figure 1



Supplementary Figure 1. *Analysis pipeline*. The analysis pipeline consisted of two arms that converged in the center. The first arm (left-hand side of figure) involved feature engineering and dimension reduction analyses for the dietary data, particularly the creation of three dietary patterns. The second arm concerned work related to the gene expression data, and both the creation of 47 and 37 gene expression clusters for women and men, respectively, and an *in silico* flow cytometry cell type quantification. We used linear models and pre-specified DAGs

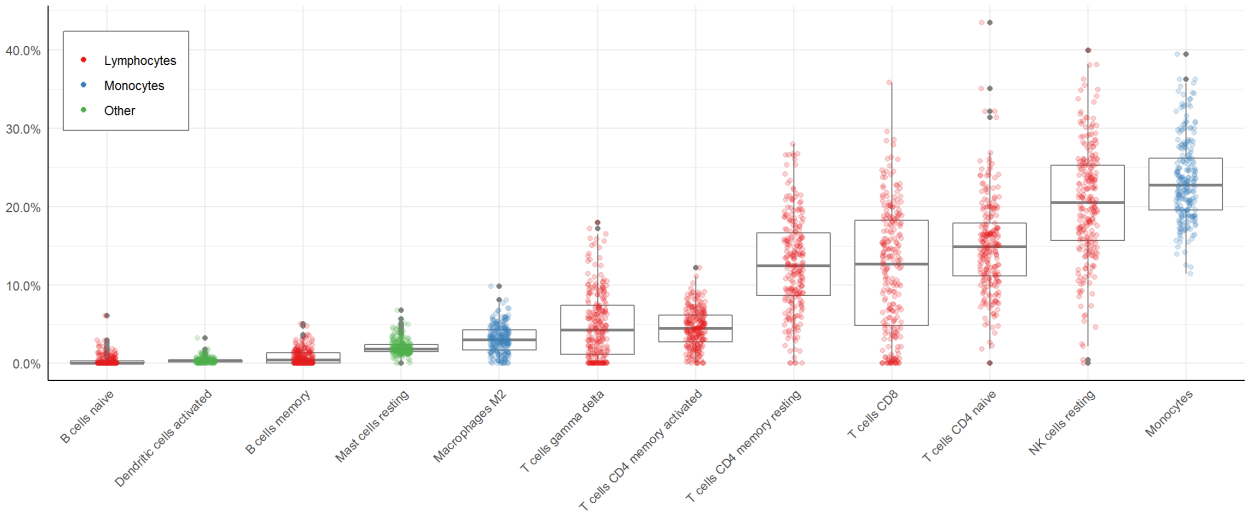
(Supplementary Figure 4) to evaluate the associations between the dietary and gene expression sides. Abbreviations: FFQ, food frequency questionnaire; gX, gene expression; PC1, principal component 1; WGCNA, weighted gene correlation network analysis.

Supplementary Figure 2



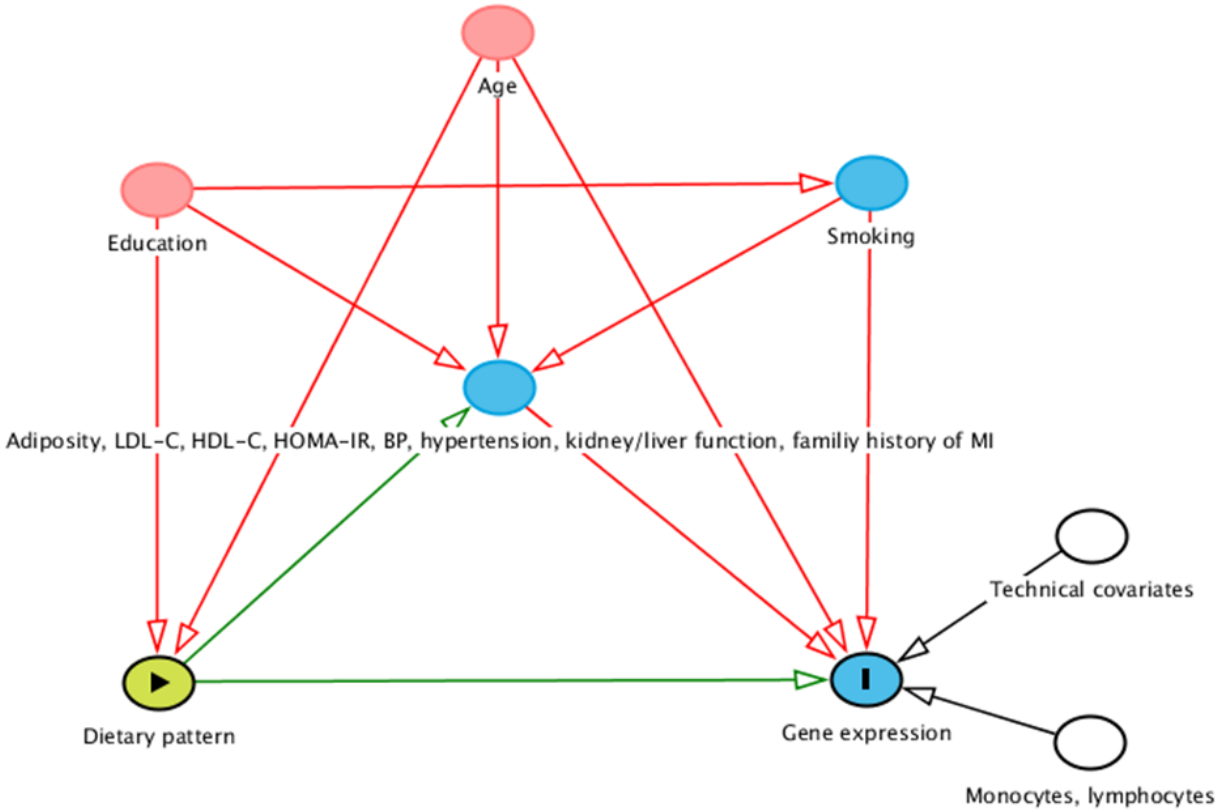
Supplementary Figure 2. *Evaluation of PCA analysis of dietary data*. The figure shows the standard deviation ('std.dev'), percent explained variance ('percent') and cumulative explained variance ('cumulative') for all principal components, derived from a PCA analysis for women (upper row) and men (lower row), respectively. The number of components needed to explain a cumulative total variance at 20, 25, 30, 35, and 40 % are highlighted by integers along the lines.

Supplementary Figure 3



Supplementary Figure 3. *CIBERSORT-precited leukocyte subset distribution*. Abbreviations: NK, natural killer; CD, cluster of differentiation.

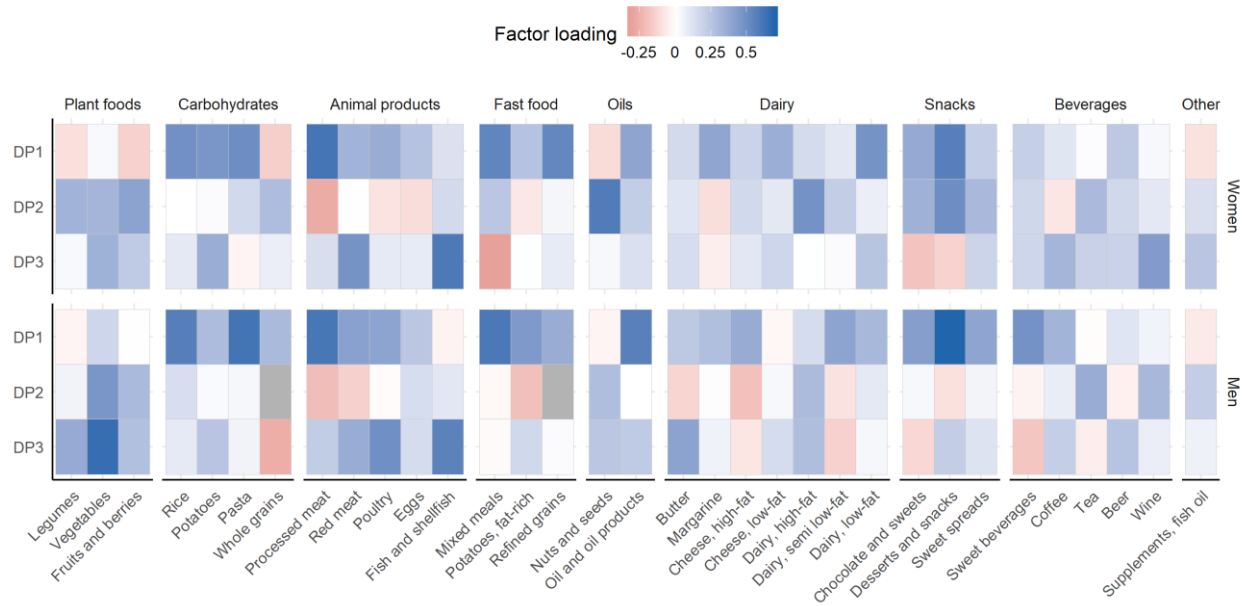
Supplementary Figure 4



Supplementary Figure 4. *Directed acyclic graph (DAG) used in model development*. Relationship between variables evaluated in model development. Dietary pattern is the exposure; gene expression is the outcome; blue circles are ancestors of the outcome; red-pink circles are ancestors of both the exposure and the outcome; green arrows indicate causal paths; and red-pink arrows indicate biasing paths. Minimal sufficient adjustment sets for estimating the *total effect* of dietary pattern on gene expression were age and education (three levels: lower, middle, higher). For comparison, the minimal sufficient adjustment sets for estimating the *direct effect* of dietary pattern on gene expression were considered to be: adiposity (total fat mass,

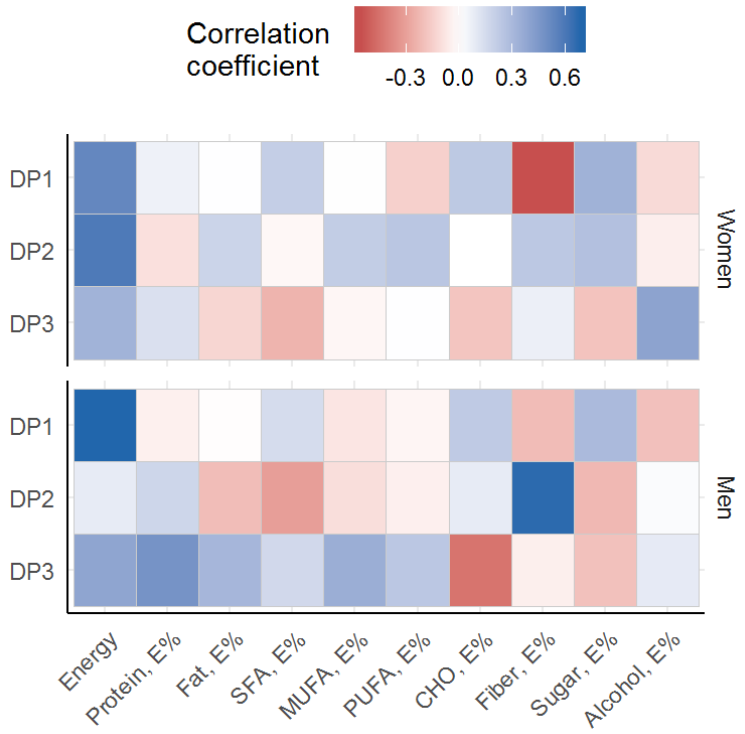
measured by bioelectrical impedance analysis), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), homeostatic model assessment for insulin resistance (HOMA-IR), systolic blood pressure, use of anti-hypertensive medication (two levels: no, yes), kidney function (creatinine), liver function (alanine aminotransferase, ALAT), familiy history of myocardial infarction (two levels: no, yes), age, and smoking (two levels: never/stopped, seldom/some/regularly). Because the two types of models were similar, we herein report results from the total effect model for both types of outcomes (data not shown). Note that for all models, technical covariates were considered in upstream batch correction. Percentage of total leukocyte count of monocytes and lymphocytes (which make up the pool of PBMC subsets) were adjusted for in the gene expression pre-processing pipeline (as shown in Supplementary Figure 1), prior to WGCNA only. We used the open-access dagitty.net/dags web-resource to evaluate these relationships. Abbreviations: BP, blood pressure; HDL-C, high-density lipoprotein cholesterol; HOMA-IR, homeostatic model assessment for insulin resistance; LDL-C, low-density lipoprotein cholesterol; MI, myocardial infarction.

Supplementary Figure 5



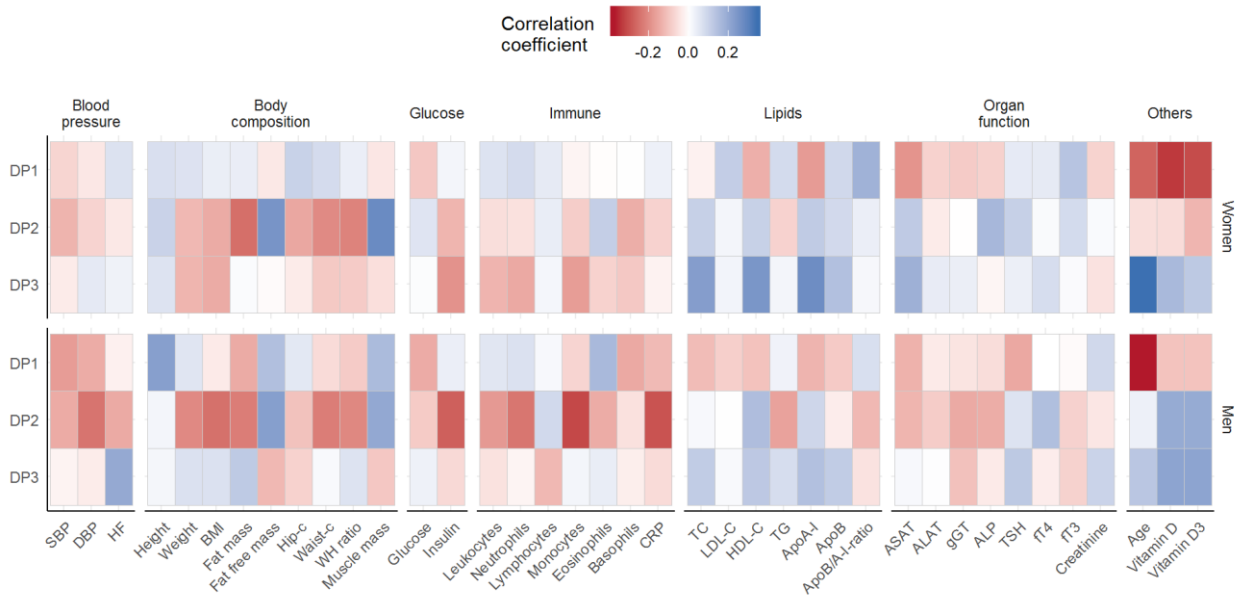
Supplementary Figure 5. *DP loading for different foods for women and men*. The figures display all DP loadings for different foods for women and men, corresponding to Figure 1 and Supplementary Table 4. Food groups *whole grains* and *refined grains* displayed high absolute values for DP2 for men: 0.86 and -0.72, respectively. We therefore removed these to highlight the fine details in the other DP-food pairs. Abbreviations: DP, dietary pattern.

Supplementary Figure 6



Supplementary Figure 6. *Correlation between DP scores and macronutrient intake for women and men*. The figures display the Spearman's rho (ρ) inter-correlation between DP scores (Figure 1) and macronutrient intake for women and men, which are displayed on the x and y axes, respectively. Abbreviations: DP, dietary pattern; E%, energy percent.

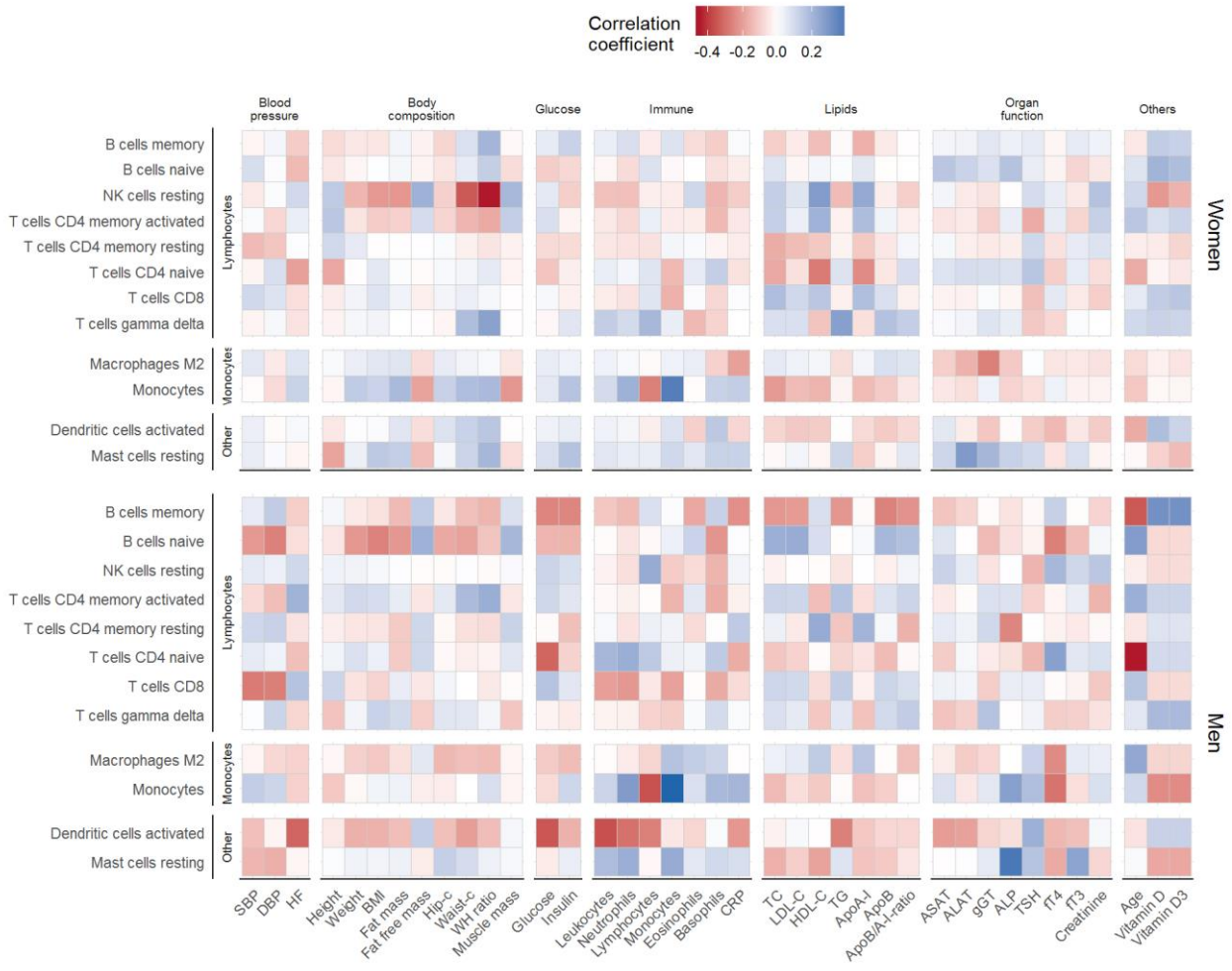
Supplementary Figure 7



Supplementary Figure 7. *Associations between DP scores and clinical variables for women and men*. The figures display the Spearman's rho (ρ) inter-correlation between DP scores (Figure 1) and clinical variables for women and men, which are displayed on the x and y axes, respectively. Abbreviations: Age, Age; ALAT, Alanine aminotransferase; ALP, Alkaline phosphatase; ApoA-I, Apolipoprotein A-I; ApoB, Apolipoprotein B; ApoB/A-I-ratio, Apolipoprotein B-apolipoprotein A-I ratio; ASAT, Aspartate aminotransferase; Basophils, Basophils; BMI, Body mass index; Creatinine, Creatinine; CRP, C-reactive protein; DBP, Diastolic blood pressure; DP, dietary patterns; Eosinophils, Eosinophils; Fat free mass, Fat free mass; Fat mass, Fat mass; fT3, Free triiodothyronine; fT4, Free thyroxine; gGT, gamma-Glutamyltransferase; Glucose, Glucose; HDL-C, HDL cholesterol; Height, Height; HF, Heart frequency; Hip-c, Hip circumference; Insulin, Insulin; LDL-C, LDL cholesterol; Leukocytes, Leukocytes; Lymphocytes, Lymphocytes; Monocytes, Monocytes; Muscle mass, Muscle mass; Neutrophils, Neutrophils; SBP, Systolic blood pressure;

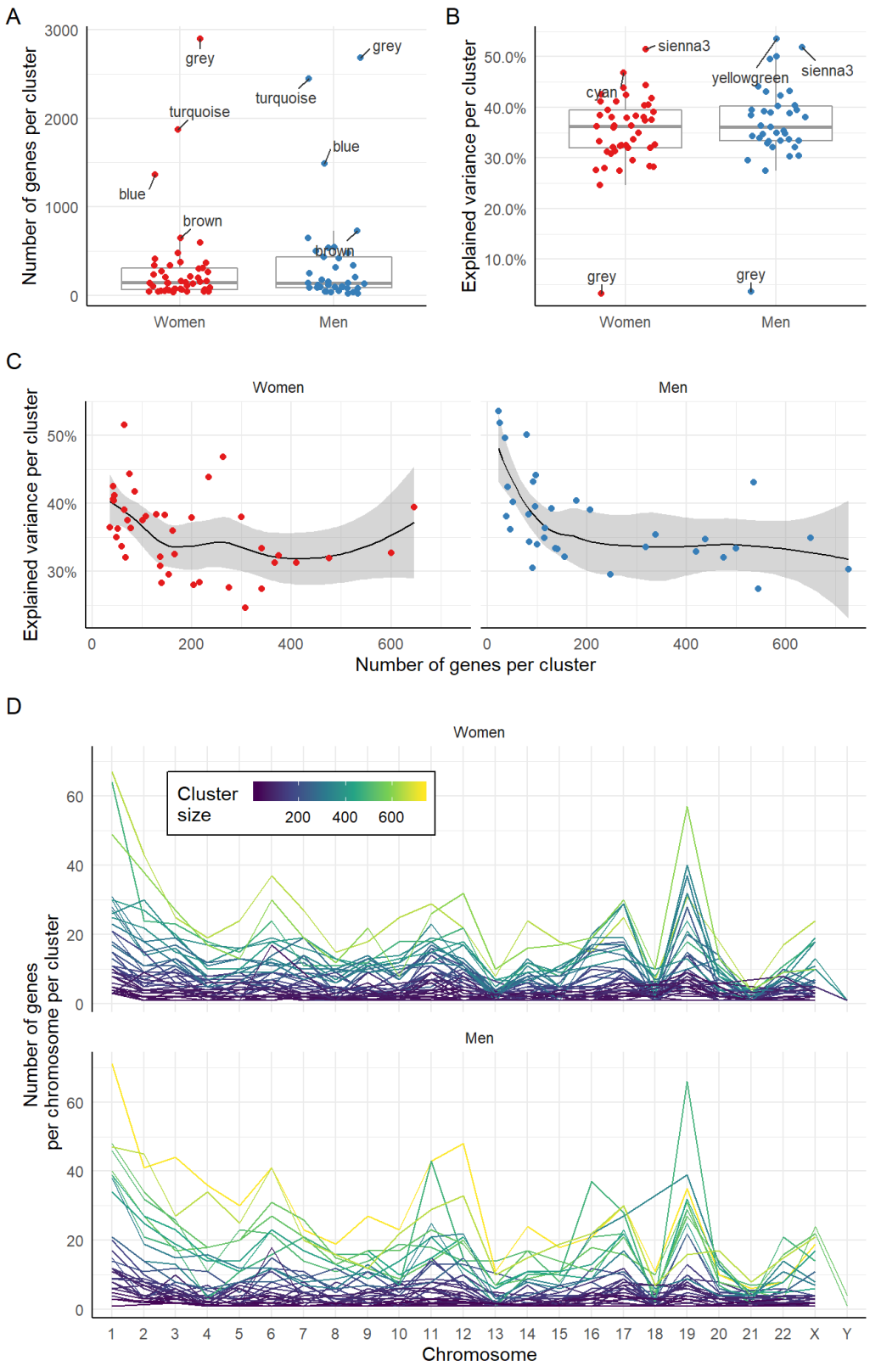
TC, Total cholesterol; TG, Triglycerides; TSH, Thyroidea stimulating hormone; Vitamin D, Vitamin D; Vitamin D3, Vitamin D3; Waist-c, Waist circumference; Weight, Weight; WH ratio, Waist-hip ratio.

Supplementary Figure 8



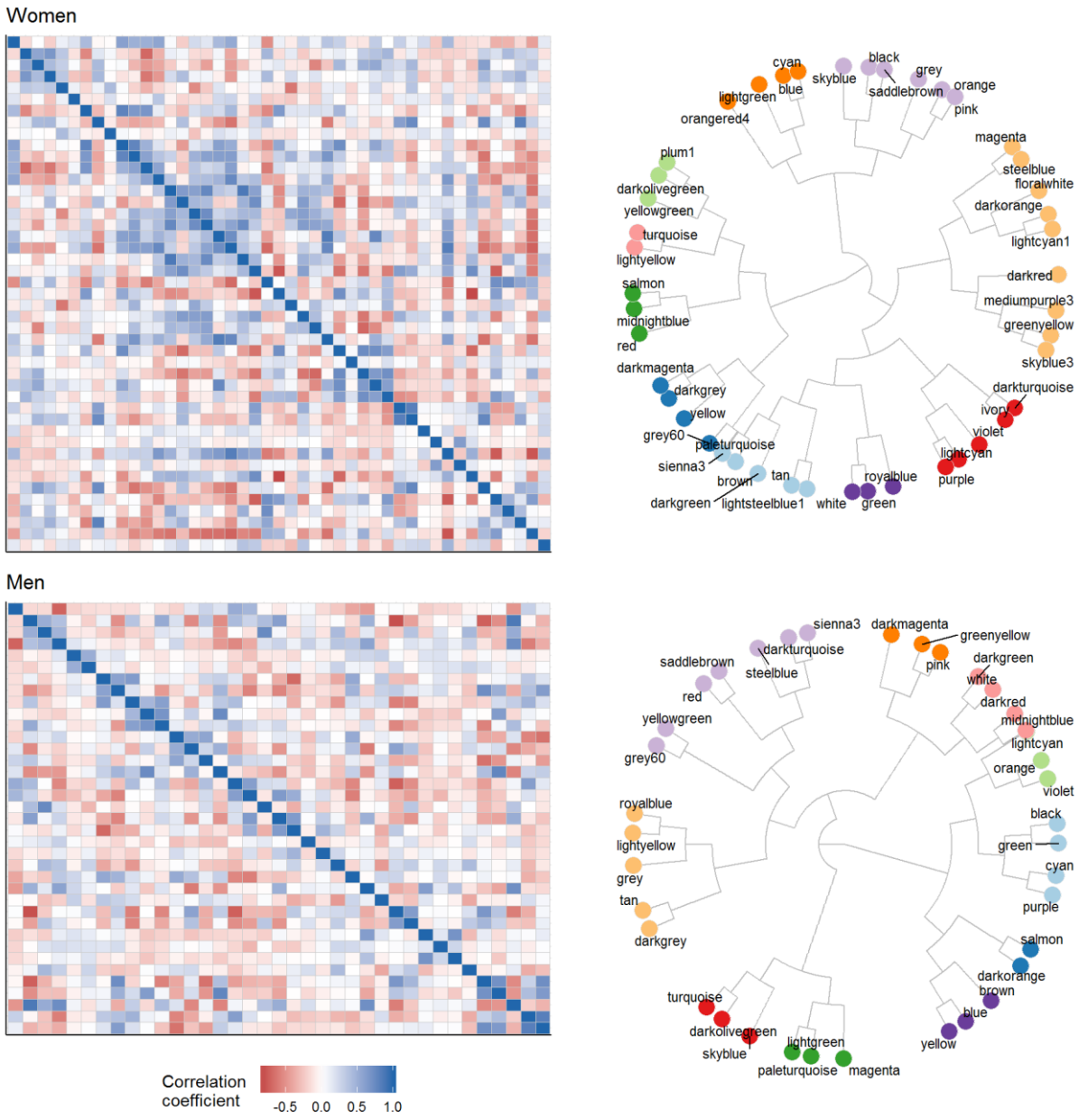
Supplementary Figure 8. *Associations between CIBERSORT-predicted leukocyte subset distribution and clinical variables*. The figures display the Spearman's rho (ρ) inter-correlation between relative level of leukocyte cell subsets for women and men, which are displayed on the x and y axes.

Supplementary Figure 9



Supplementary Figure 9. *Cluster descriptives*. The figure displays A) the number of genes per cluster, B) explained variance per cluster, C) the explained variance as a function of cluster size, and D) the number of genes per chromosome per cluster. C) and D) display clusters less than 1000 genes only, to emphasize details. Supplementary Table 5 lists exact size and variance explained for each gender and cluster.

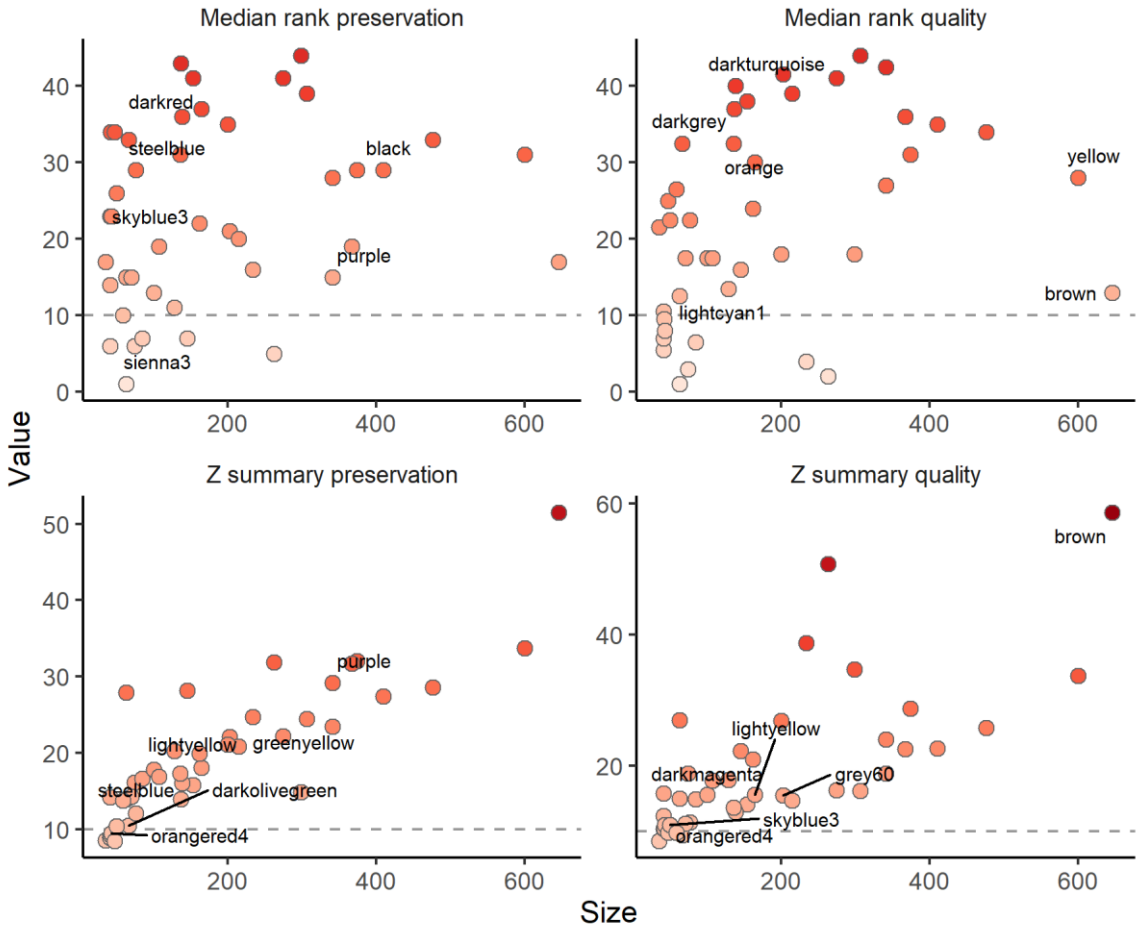
Supplementary Figure 10



Supplementary Figure 10. *Associations between module eigengenes (MEs) for all clusters*. The figures display the Spearman's rho (ρ) inter-correlation between cluster eigengenes (equivalent to principal component 1 for that cluster) for women (upper panels; 40 clusters) and men (lower panels; 27 clusters), which are displayed on the x and y axes. All rows and columns are ordered

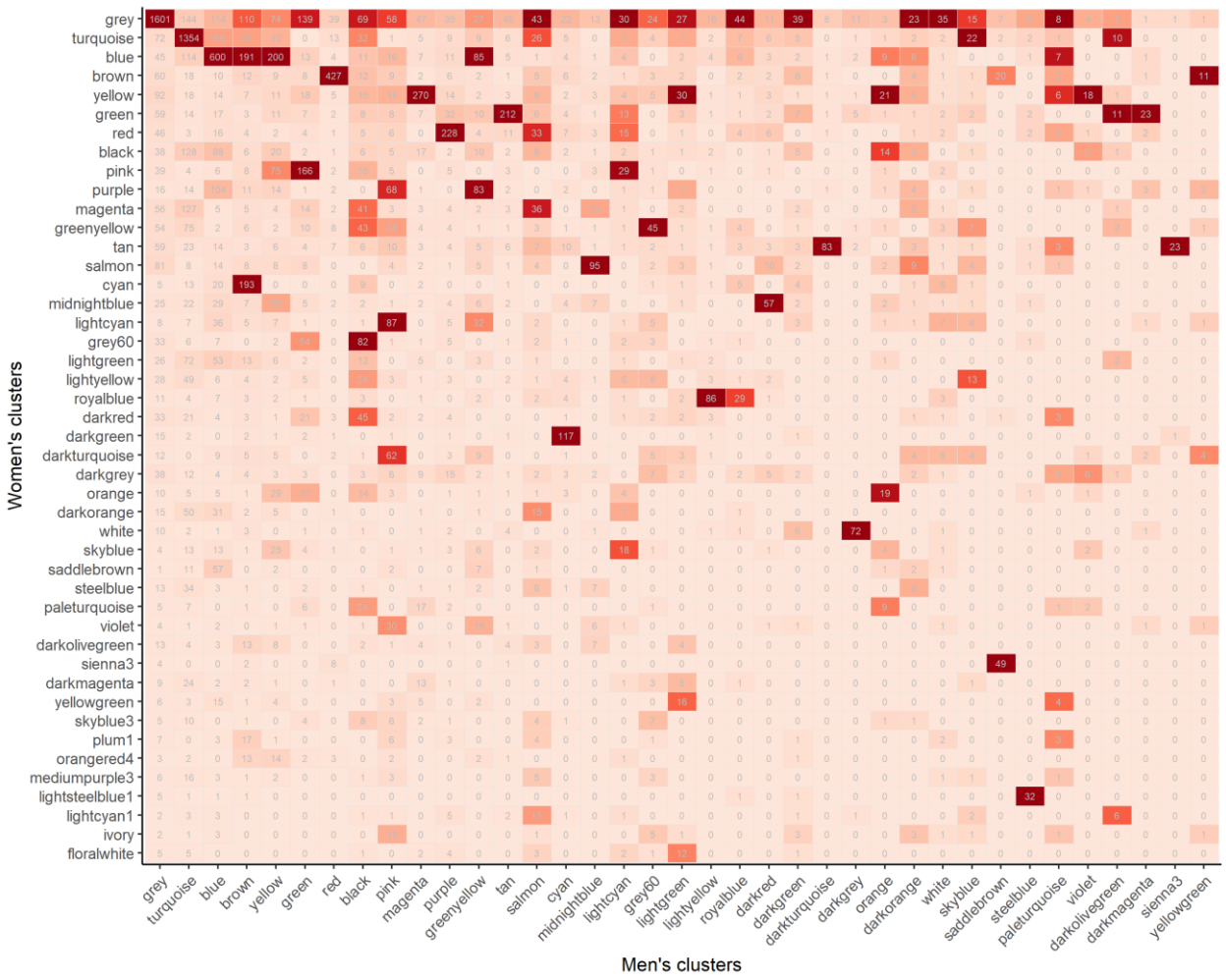
by Euclidean distance and hierarchical clustering (complete linkage). The right-hand side plots are networks representations of the correlation heatmaps (for each gender), which highlights the closeness of the different gene expression clusters.

Supplementary Figure 11



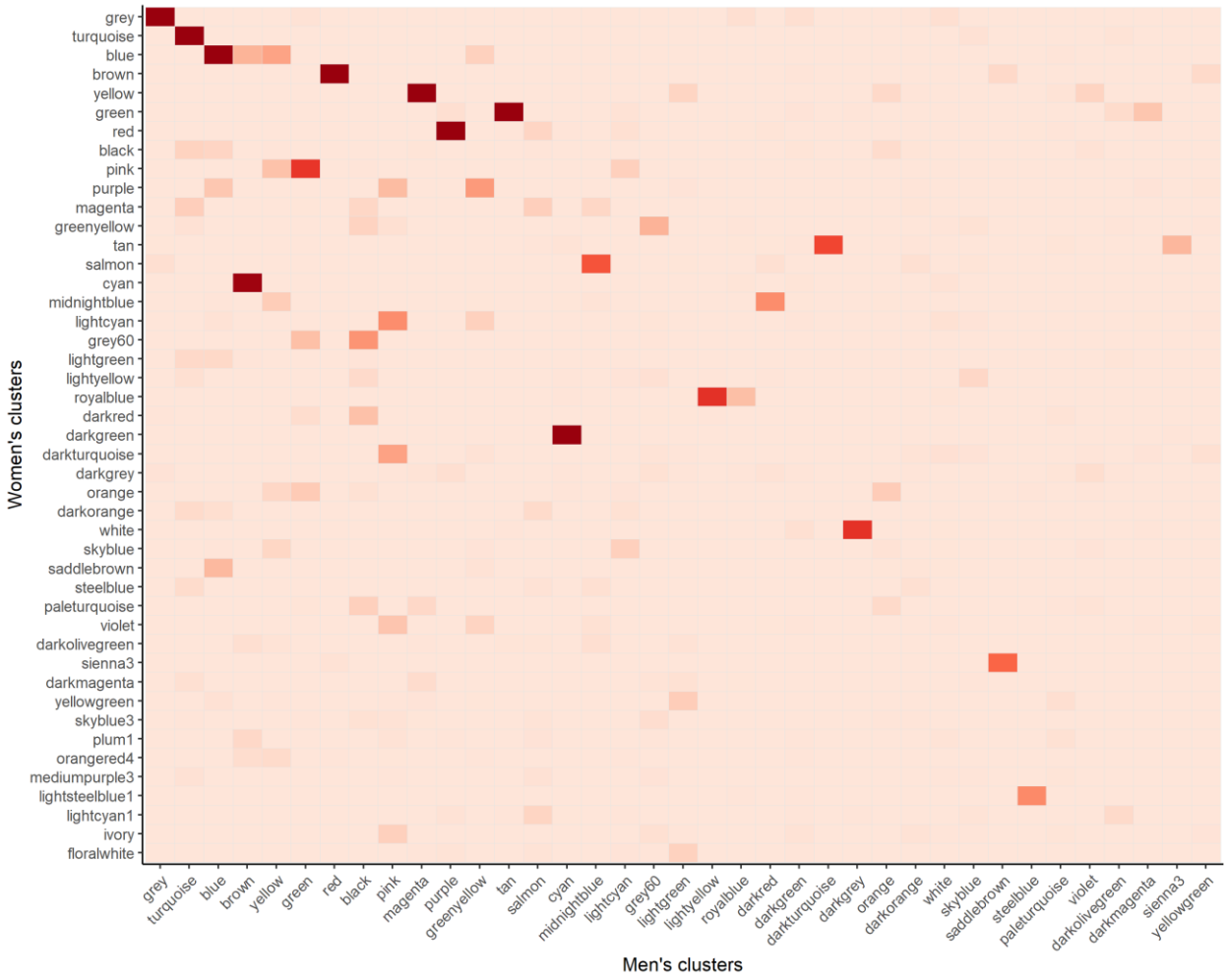
Supplementary Figure 11. *Module preservation and quality across module size*. The figure displays module *median rank preservation* and *quality* in a comparison between women's and men's gene expression clusters, determined by the WGCNA algorithm, as well as the corresponding Z scores. A random selection of clusters are highlighted on the plot. The dashed grey line denotes the threshold for whether the cluster is preserved or not. For all panels: the higher the y value, the more preserved the cluster. Note that many clusters display moderate-to-large module preservation, although not all.

Supplementary Figure 12



Supplementary Figure 12. *Direct comparison of content of women's and men's clusters: counts*. Cross-tabulation of intersected genes for all combinations of women's and men's clusters. Note that for most of the women's clusters, the majority of the genes fall in *one or a few* of the men's clusters, which indicates preservation between genders, in accordance with Supplementary Figure 11. The color is scaled to maximize color intensity for numbers within each column, that is, for the distribution of counts for *men's* clusters.

Supplementary Figure 13



Supplementary Figure 13. *Direct comparison of content of women's and men's clusters: P values*. Cross-tabulation of P values for the intersected genes for all combinations of women's and men's clusters. The figure must be seen in conjunction with Supplementary Figure 12: tiles that light up generally overlap between the genders. The color is not scaled; it is plotted directly as -log10(P value) to highlight the significance level.

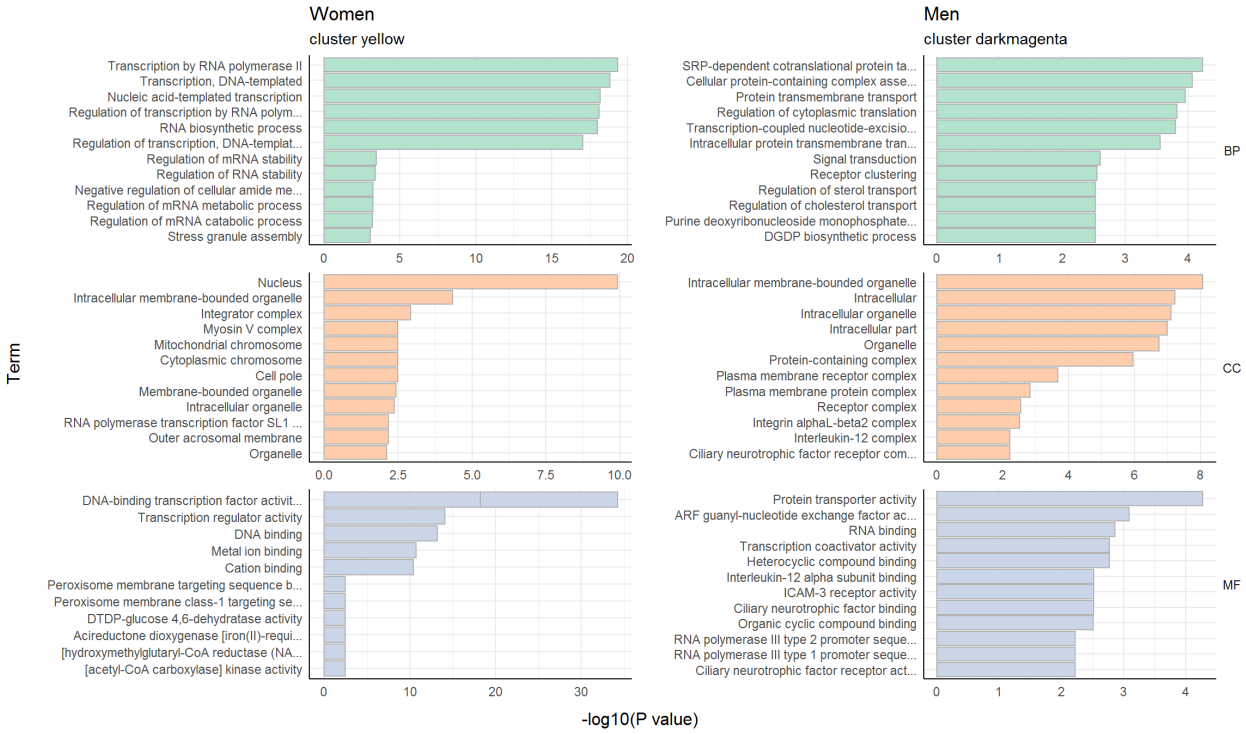
Supplementary Figure 14

Supplementary Figure 14. *Associations between module eigengenes and clinical variables for women and men*. The figures display the Spearman's rho (ρ) inter-correlation between module eigengenes (equivalent to principal component 1 for that gene cluster) and clinical variables for women and men. The rows are ordered by cluster size. Abbreviations: ALAT, Alanine aminotransferase; ALP, Alkaline phosphatase; ApoA-I, Apolipoprotein A-I; ApoB, Apolipoprotein B; ApoB/A-I-ratio, Apolipoprotein B-apolipoprotein A-I ratio; ASAT, Aspartate aminotransferase; BMI, Body mass index; CRP, C-reactive protein; DBP, Diastolic blood pressure; fT3, Free triiodothyronine; fT4, Free thyroxine; gGT, gamma-Glutamyltransferase; HDL-C, HDL cholesterol; HF, Heart frequency; Hip-c, Hip circumference; LDL-C, LDL cholesterol; ME, module eigengene; SBP, Systolic blood pressure; TC, Total cholesterol; TG, Triglycerides; TSH, Thyroidea stimulating hormone; Waist-c, Waist circumference; WH ratio, Waist-hip ratio.

Supplementary Figure 15

Supplementary Figure 15. *Relative abundance of monocytes according to DP2 score*. Scatterplot between DP2 score (Vegetarian diet) and CIBERSORT-predicted monocyte level. Abbreviations: DP, dietary pattern.

Supplementary Figure 16



Supplementary Figure 16. *GO terms for the top DP-associated clusters*. Bar plots of the top 12 most significantly enriched GO terms within biological processes (BP), cellular compartment (CC) and molecular function (MF). The GO terms are derived from analyses of the separate clusters, with the original gene expression dataset as background. For a comprehensive list of detail, see Supplementary Table 6.