# Epigenome-wide association study of kidney function-related traits and serum urate: Analysis plan in individual cohorts

## Contacts

Please email questions to lead contacts:

For project 1 (kidney function): Alexander Teumer (ateumer@uni-greifswald.de) and Pascal Schlosser (schlosser@imbi.uni-freiburg.de)

For project 2 (urate): Adrienne Tin (atin1@jhu.edu) and Anna Köttgen (anna.koettgen@uniklinik-freiburg.de)

## Background

Chronic kidney disease (CKD) can progress to end-stage renal disease and is a major contributor to cardiovascular morbidity and mortality. The biological mechanisms leading to CKD and its progression are incompletely understood. CKD can be defined as the sustained presence of reduced glomerular filtration rate, estimated from serum creatinine concentrations (eGFR) and/or increased concentrations of albumin in the urine (quantified as urinary albumin to creatinine ratio, UACR).

Heritability estimates of eGFR and UACR as well as familial aggregation studies of CKD support a substantial heritable component,of which only a small part is due to classical monogenic diseases. Rather, CKD susceptibility is influenced by DNA sequence variants in many genes, environmental factors, and their interactions. Genome-wide association studies have successfully identified common variants at two genetic loci for UACR and >60 genetic loci that associate with eGFR and CKD. Despite implicating novel pathways in CKD, the lead DNA sequence variants in the associated loci together explained <5% of the eGFR and UACR variance. Integration of chromatin annotation maps with results of the largest GWAS meta-analysis of eGFR to date supports the importance of altered transcriptional regulation as a mechanism contributing to CKD. Thus, the investigation of epigenetic changes that may influence transcription and associate with eGFR and CKD is of particular interest.

Hyperuricemia is the most important risk factor for gout, which is the most common form of inflammatory arthritis. DNA methylation levels in immune cells may be involved in the pathogenesis of gout. Identifying associations between serum urate levels and DNA methylation levels in whole blood or in specific immune cell types may provide insight for future studies on the relation between hyperuricemia and DNA methylation in immune cells. In addition, DNA methylation may integrate adverse environmental influences that modulate the susceptibility to these common complex diseases.

## Aims

We are interested to detect the association of DNA methylation (DNAm) levels in whole blood or specific immune cell types with CKD-related traits and serum urate levels.

This project has two sub-projects.

**Project 1: CKD/eGFR and UACR/microalbuminuria (MA).** We aim to:

* 1. Detect associations between DNAm levels and the different kidney function measures
  2. Detect whether identified differences translate to kidney tissue

**Project 2: Urate**. We aim to:

1. Detect associations between DNAm levels and serum urate

## General guidelines for DNA methylation data

DNA methylation data pre-processing will take place independently in each participating cohort and should contain the following components:

1. Identifying samples with technical issues (e.g. bisulfate conversion, hybridization, staining)
2. Excluding samples with low call rate
3. Excluding samples with low detection p-value
4. Background correction of intensity values
5. Normalization of DNA methylation levels
6. Sex-Check

For exclusions, please use study-specific cutoffs determined in your study. If there are no such cutoffs, please contact us for recommendations. Please prepare a brief methods write-up and fill out the table on DNAm data processing methods (in 2\_QC\_def in the Study\_Summary excel file) to be uploaded with the analysis results.

We will use methylation beta values as the outcome for EWAS.

## Phenotype (i.e. independent variables of interest)

To facilitate cross-referencing to DNA sequence variation association results from GWAS, phenotypes will be generated using a distributed script that is similar to the script used in the CKDGen Consortium Round 4. Pediatric cohorts: please contact us for a separate pediatric analysis plan. The phenotype preparation script can be found at <https://github.com/genepi-freiburg/ckdgen-pheno/tree/ckdgen-ewas-pheno>

Detailed instructions on how to run the script can be found at

<https://github.com/genepi-freiburg/ckdgen-pheno/blob/ckdgen-ewas-pheno/ckdgen-pheno-prep/README>

The script automatically generates the following variables that represent the independent variables of interest (column names):

**eGFR, ml/min/1.73m2 (egfr\_ckdepi\_creat):** calculated from the CKD-EPI equation

**CKD (ckd)**: defined as eGFR <60 mL/min/1.73m2

**ln(UACR) (ln\_uacr):** log transformed urinary albumin to creatinine ratio, mg/g

**MA (microalbuminuria):** defined as 1 if UACR >30 mg/g, 0 if <10 mg/g

**Serum urate (uric\_acid\_serum):** in mg/dL. To convert umol/L to mg/dL, please divide by 59.48. Your mean serum urate should be approximately 5-6.5 mg/dl. Please do not impute missing values of urate.

**Please send the phenotype summary.pdf** that is generated by the script to us for review before starting to run EWAS. We will briefly review the phenotypes to make sure they are ok. If they are okay, please include these phenotypes into the regression models.

## Covariates

The following variables will be used as covariates (additional independent variables) in at least one of the models. Please note that the phenotype generation script only generates the main predictor variables, and that the covariates below need to be generated by yourself and included in addition.

|  |  |  |
| --- | --- | --- |
| **Covariates** | **Kidney function models** | **Serum urate models** |
| **Male:** 0=female, 1=male | **X** | **X** |
| **Age:** continuous | **X** | **X** |
| **Prevalent diabetes (DM)**   * Preferred definition: fasting plasma glucose ≥ 126 mg/dl (7.0 mmol/L) OR treatment for diabetes * If fasting glucose is not available: casual glucose (non-fasting) ≥ 200 mg/dl (11.0 mmol/L) OR treatment for diabetes * If glucose is not available: self-reported diabetes status * If an alternative definition was used: please specify this in your methods write-up | **X** |  |
| **Prevalent hypertension (HTN)**   * Preferred definition: systolic blood pressure ≥ 140 mm Hg OR diastolic blood pressure ≥ 90 mm Hg OR treatment for hypertension * If measured blood pressure is not available: self-reported hypertension * If an alternative definition was used: please specify in your methods write-up | **X** |  |
| **BMI:** kg/m2, continuous | **X** | **X** |
| **Current smoking:** 0=no, 1=yes | **X** | **X** |
| **Population structure**   * If applicable, use the genetic PCs as a fixed effect as you have been doing in previous GWAS | **X** | **X** |
| **Estimated or measured white blood cell type proportions (wbcprop):**   * Use measured cell type composition or estimated proportions (best performing measure in your study, if not evaluated, use Houseman; example code can be found below) | **X** | **X** |
| **Technical covariates:**   * according to the cohorts QC and preprocessing pipeline, e.g:   + Control probe PCs (for studies using CPACOR)   + Study center   + Processing batch   + Array (Sentrix\_ID)   + Position on array (Sentrix\_Position) | **X** | **X** |
| **eGFR, ml/min/1.73m2** |  | **X** |
| **HDL-cholesterol** |  | **X** |
| **Systolic blood pressure (SBP)** |  | **X** |
| **C-reactive protein (CRP) levels (log transformed)** |  | **X** |
| **Triglycerides (TG) levels (log transformed)** |  | **X** |

Serum urate levels are highly associated with lipid levels. Preliminary analysis in African Americans in the ARIC study have identified urate-associated DNA methylation sites that are also associated with triglycerides (TG) and C-reactive protein (CRP) levels. Controlling for additional covariates for the analysis of serum urate may help to isolate DNA methylation sites that could be more specifically linked to serum urate levels

## Definition of other variables in population characteristics

Medications used to treat gout:

* If your study uses Anatomical Therapeutic Chemical (ATC) codes to record medication use, please include all medications in the Antigout preparations (M04A) category.
* Otherwise, please include the following medications: allopurinol, tisopurine, febuxostat, probenecid, sulfinpyrazone, benzbromarone, isobromindione, lesinurad, colchicine, cinchophen, urate oxidase, pegloticase The ATC code of these medications are included in the Gout\_med\_reference sheet in the Study Summary excel file for your reference.
* If other medications are classified as urate lowering medications in your study, please specify them in the methods write-up.

## Sample exclusion

### Sample exclusion for the analysis of serum urate

If information on serum urate lowering/gout medication is available (see above), please exclude participants taking such medications.

## Statistical Analyses

1. Please use the same sample size for models 1 and 2 for each trait, e.g. Model\_CKD\_1 and Model\_CKD\_2 should have the same sample size, and Model\_UACR\_1 and Model\_UACR\_2 should have the same sample size. **Samples in model 1 thus need complete information for all covariates in model 2.**
2. Please analyse participants from different ethnic groups separately.
3. Please analyse DNA methylation data from different DNA methylation arrays separately.

Models for kidney traits

**Model\_CKD\_1:**

methylation(beta values, [0,1]) ~ as.factor(CKD) + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates

**Model\_CKD\_2:**

methylation(beta values, [0,1]) ~ as.factor(CKD) + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates + as.factor(DM) + as.factor(HTN) + BMI + as.factor(current smoking)

**Model\_eGFR\_1:**

methylation(beta values, [0,1])~ eGFR + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates

**Model\_eGFR\_2:**

methylation(beta values, [0,1])~ eGFR + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates + as.factor(DM) + as.factor(HTN) + BMI + as.factor(current smoking)

**Model\_UACR\_1:**

methylation(beta values, [0,1])~ log(UACR) + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates

**Model\_UACR \_2:**

methylation(beta values, [0,1]) ~ log(UACR) + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates + as.factor(DM) + as.factor(HTN) + BMI + as.factor(current smoking)

**Model\_MA\_1:**

methylation(beta values, [0,1])~ as.factor(MA) + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates

**Model\_MA \_2:**

methylation(beta values, [0,1])~ as.factor(MA) + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates + as.factor(DM) + as.factor(HTN) + BMI + as.factor(current smoking)

## Models for serum urate analysis

**Model\_urate\_1:**

methylation(beta values, [0,1])~ serum urate + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates

**Model\_urate\_2:**

methylation(beta values, [0,1])~ serum urate + as.factor(male) + age + genetic PCs + wbcprop(%, measured or estimated) + technical covariates + BMI + as.factor(current smoking) + HDL-C + SBP + log(CRP) + log(TG) + eGFR

## Data reporting

1. Regression result files

Whitespace or tab delimited files; one file per trait and model; one row per probe. First row is a header as defined in the table below. Please use exactly the header names shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Column Header | probeID | BETA | SE | P\_VAL | N\_samp | mean | SD |
| Description | Probe name | Effect size | Standard error of BETA | P value for the probe | Number of samples | Mean of DNA methylation beta values | SD of DNA methylation beta values |
| Format | string | numeric | numeric | numeric | numeric | numeric | numeric |
| Example 1 | cg2356825 | 2.6897 | 0.8456 | 1.9971e-5 | 517 | 0.4568 | 0.1865 |
| Example 2 | ch.10.1345297R | 1.3443 | 0.3421 | 1.231e-9 | 1232 | 0.8913 | 0.3214 |

* File naming scheme:
* TRAIT\_MODEL\_ETHNICITY\_STUDY\_ARRAY\_DATE.txt.gz
* TRAIT: CKD, eGFR, UACR, MA, urate
* MODEL: Model\_CKD\_1, Model\_CKD\_2, Model\_eGFR\_1, Model\_eGFR\_2, Model\_UACR\_1, Model\_UACR\_2, Model\_MA\_1, Model\_MA\_2, Model\_urate\_1 or Model\_urate\_2
* ETHNICITY: EA: European ancestry, AA: African Americans, SA: South Asian, EAS: East Asian
* STUDY is a short identifier of your study name
* ARRAY: EPIC, 450k or 27k
* DATE is the date on which the file was prepared. Please use the following format “DDMMYYYY”.
* Data file contents requirements:
* If you have missing data in your results file, please do not leave any cells blank. Missing data should be denoted by NA.
* Please provide all numeric data with at least 4 decimal places. P-values should be specified to at least 4 relevant digits.
* Please do not adjust your P-values for multiple testing.
* Example code for exporting from R can be found below.

1. Summary table

Please fill in the attached Study Summary excel file (3\_results) of summary statistics of the regression results.

File naming scheme: summary\_ETHNICITY\_STUDY\_ARRAY\_DATE.xlsx

1. QC and pre-processing

Please fill in the attached Study Summary excel file (2\_QC\_def) of QC and pre-processing steps taken.

1. Summary statistics of variables

Please fill in the attached Study Summary excel file (1\_population) of outcome variables and covariates for each model (as datasets might vary).

1. Methods write-up read me file

Please upload a methods write-up file with the following information:.

* A brief write-up of the methods in DNA methylation data pre-processing
* Alternative definition of prevalent diabetes if applicable
* Alternative definition of prevalent hypertension if applicable
* Alternative definition of gout medication if application

## Data exchange

Data can be uploaded to Freiburg University servers. Please email schlosser@imbi.uni-freiburg.de when you are ready to upload your data and we will provide you with instructions for uploading data.

## Example code

### Preprocessing with CPACOR

Details of the CPACOR pipeline are provided in:

Lehne B, Drong AW, Loh M et al. *A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biology 2015; 16: 37. Available from:* [*http://genomebiology.com/2015/16/1/37*](http://genomebiology.com/2015/16/1/37)

Scripts and documentation are provided in the Supplement:

<https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0600-x/MediaObjects/13059_2015_600_MOESM2_ESM.zip>

### Houseman estimation of WBCs (both 450k and EPIC arrays):

# install packages (to do only once)

source("http://bioconductor.org/biocLite.R")

biocLite("minfi")

biocLite("IlluminaHumanMethylationEPICmanifest") # for EPIC arrays only

biocLite("IlluminaHumanMethylation450kmanifest")

biocLite("FlowSorted.Blood.450k")

biocLite("IlluminaHumanMethylation450kanno.ilmn12.hg19")

# load package

require(minfi)

# read arrays

RGset <- read.metharray(sampleIDs, verbose=TRUE)

# Calculating Houseman Cell Type Counts and save them

counts <- estimateCellCounts(RGset)

write.csv(counts,file="celltypeCounts\_default.csv"),row.names=T,quote=F)

### Export command within R for regression results file:

# character vectors describing the current analysis:

# trait, model, ethnicity, study, array

out\_file = paste0(trait, model, ethnicity, study, array, format(Sys.time(),format="%d%m%Y"), ".txt")

write.table(results\_dataframe, out\_file, quote=F, row.names=F,sep="\t")