CKDGen Consortium: Round 4 Analysis Plan
July of 2016

The purpose of these analyses is to perform trans-ethnic GWAS meta-analyses of densely imputed genotype data to uncover novel loci associated with kidney function related traits and kidney disease.

If the collection of data in your study does not allow for carrying out one or more analyses as outlined in this document, please contact us before proceeding.

Outline of the analysis plan:

Contacts

For any issue related to the CKDGen Consortium and analyses, please contact *Anna Köttgen* (anna.koettgen@uniklinik-freiburg.de, akottge1@jhu.edu) and/or *Cristian Pattaro* (cristian.pattaro@eurac.edu). For questions on the following issues, please include the following person(s):

- Prospective phenotypes: *Carsten Böger* (Carsten.Boeger@ukr.de);
- Phenotype creation software: *Matthias Wuttke* (matthias.wuttke@uniklinik-freiburg.de) and *Mathias Gorski* (Mathias.Gorski@klinik.uni-regensburg.de)
- GWAS and analytical issues: *Alexander Teumer* (ateumer@uni-greifswald.de)
- EPACTS analysis pipeline: *Christian Fuchsberger* (christian.fuchsberger@eurac.edu)

Please upload all result files **by September 30th, 2016**.

1. Genotypes and imputation

We are requesting data using one of the following haplotype reference panels for imputation: Haplotype Reference Consortium (HRC) version 1.1 (preferred for studies of European ancestry individuals) or 1000G phase 3 v5 ALL (if HRC not available or for studies of non-European ancestry). If neither of these panels is available, imputation using the 1000G phase 1 v3 ALL haplotype reference panel or later is acceptable. Imputation should be carried out excluding monomorphic sites and singletons, and including chromosomes 1-22 and X. Data are requested on the forward strand and using NCBI b37 (hg19) coordinates.

If your study has such data already available, please move on to the next section.

If you need to impute your data, you can find details related to data quality control, lift over, strand alignment, helpful links to imputation programs, reference panels and imputation settings in the **Appendix**.

2. Phenotype generation

Along with this analysis plan, we are **distributing** a **script** that will **generate the phenotypes** to be used as the outcome for GWAS (download link in section 2.2). The following steps need to be carried out to generate all GWAS-ready phenotypes:

- **Section 2.1**: **create** an **input file** <input_filename>.txt that contains all necessary variables from your study – **Table 1** lists these variables and their required names.
- **Section 2.2**: **download** the phenotype generation **script and edit** the **parameter file**.
- **Section 2.3**: **run** the **phenotype generation script** to obtain analysis-ready phenotypes.

2.1 Set up the file <filename>.txt that is the input for the phenotype generation script

- For all subjects with imputed genotypes, generate or obtain all variables as described in the "definition" column of **Table 1**. We realize that you may not have all of the biomarkers: provide those that you have.
- Do not transform phenotypes; the provided script will do so automatically.
- **Name the variables exactly as in the "variable name" column of Table 1.**
- Use **tab** as the column **separator** in the **input file.**
- Replace <filename>.txt with a name for your input file
- If you have a phenotype measured at more than one study visit, use the one with the **largest sample size**.

- Code missing values as NA, unless indicated differently in **Table 1** (see column "Definition").
- The script will also run if columns for non-available phenotypes are missing.
- **Age**: age should always be **in years**; for any age variable, **do not round** but use greater precision if available.
- **Prospective studies with more than 2 time-points**: please consult with us to define a baseline and a follow-up point for creating the longitudinal traits.

Table 1: Description, names and definition of required variables for the input file.

Variable description	Variable name that must be used in input file	Definition
Participant identifier	iid	Use your **unique** participant identifier
Age at time of serum creatinine measurement	age_screa	Age **in years** at the visit at which serum creatinine was measured. For studies with prospective data, this refers to baseline visit.
Age at time of urine biomarker measurements	age_urine	Age in **years** at the visit at which urinary biomarkers were measured. If the time point it the same as for serum creatinine, leave empty or copy/paste the same values.
Age at time of BUN or urea measurement	age_bun_urea	Age **in years** at the visit at which serum urea/BUN (whichever is available) was measured. If the time point is the same as for serum creatinine, leave empty or copy/paste the same values.
Age at time of serum urate measurement	age_uric_acid	Age **in years** at the visit at which serum uric acid (a.k.a. urate) was measured. If the time point it the same as for serum creatinine, leave empty or copy/paste the same values.
Male sex	male	**Code**: 1 = male, 0 = female
African ancestry	black	**Code**: 1 = African or African American ancestry, 0 = any other ancestry
Blood creatinine	screa	As measured, possible units: mg/dl or µmol/l (units will be specified in the parameter file, see section 2.2)
Urinary creatinine	ucrea	As measured, possible units: mg/dl or µmol/l (units will be specified in the parameter file, see section 2.2)
Urinary albumin	ualb	use/**convert to unit mg/l** **Handling of values below** the **limit of detection (LOD)** of the albumin assay**:** • **Do not set** them **to missing**. • If numerical values below LOD are available, leave them as they are, but specify the assay LOD in the script parameter file as described in the next section. • If they are reported as "<LOD" (e.g.: "<3"), drop the "<" operator, leave numerical value of the LOD (e.g.: 3), and specify the assay LOD in the script parameter file. • Ensure that the variable has only numeric values.
Blood urea	urea	use/**convert to unit mmol/l**. Your study will likely only have either blood urea OR blood urea nitrogen (BUN). Please use whichever is available.
Blood urea nitrogen (BUN)	bun	use/**convert to unit mg/dl**. Your study will likely only have either blood urea OR BUN. Please use whichever is available.
Blood urate (i.e. uric acid)	uric_acid	As measured, possible units: mg/dl or µmol/l (units will be specified in the parameter file, see section 2.2)

Diabetes at the time of serum creatinine measurement	diabetes_screa	**Code**: 1 = diabetes; 0 = no diabetes; NA = missing information Preferred definition: fasting plasma glucose ≥126 mg/dl (7.0 mmol/L) OR treatment for diabetes. If fasting glucose is not available: non-fasting glucose ≥200 mg/dl (11.0 mmol/L) OR treatment for diabetes If glucose is not available: self-reported diabetes status
Diabetes at the time of urinary biomarker measurement	diabetes_urine	Defined as above, but use time point of urinary biomarkers. If the time point it the same as for serum creatinine, leave empty or copy/paste the same values.
Hypertension at the time of serum creatinine measurement	htn	**Code**: 1 = hypertension; 0 = no hypertension; NA = missing information Preferred definition: systolic BP ≥ 140 mm Hg OR diastolic BP ≥ 90 mm Hg OR treatment for hypertension If measured BP not available: self-reported hypertension
Gout	gout	**Code**: 1 = gout; 0 = no gout or missing information; **note**: set missing to 0! Preferred definition: self-reported gout If self-report not available: gout defined based on ICD-coding (ICD-9 code 274.0, 274.1, 274.8, or 274.9; ICD-10 M10.0, M.10.3, M.10.4, M10.9) from hospital discharge records and/or death certificates. If ICD codes not available: intake of gout specific medication within the last month: allopurinol, febuxostat, probenecid, benzbromarone, or colchicine
Blood creatinine at follow up	screa_fu	**Only for prospective studies:** as measured at the follow-up visit for serum creatinine, must be in the same unit as blood creatinine at baseline
Age at time of follow-up creatinine measurement	age_fu	**Only for prospective studies:** age in years at the follow-up visit for serum creatinine

2.2 Download the phenotype generation script and edit the parameter file

Please start by **downloading the phenotype generation script** from
https://github.com/genepi-freiburg/ckdgen-pheno/

On the page, click on the green “clone or download” button, then select “download ZIP”.

To successfully run the phenotype generation script, the following additional information is needed:

- **Assay used for measurement of blood creatinine** (Jaffe vs. enzymatic). For studies with prospective data: you need this information for baseline and follow-up.
- **Year of blood creatinine measurement**. For studies with prospective data: you need this information for baseline and follow-up.
- The lower **limit of detection** (LOD) for the assay used to measure urinary albumin (see **Table 1** for how to code values < LOD in the urinary albumin field)

With this information, please **edit** the parameter file **“params-template”**:

- Rename the parameters file to include your study name, e.g. "cp params-template params-gckd.txt"

- Edit the parameter file by replacing the example values with your study's information (detailed instructions are found in the params-template file).

2.3 Running the phenotype generation script

Please read the **README** file distributed with the script. With the files generated as outlined in sections 2.1 and 2.2, run the phenotype generation script on the command line, passing the parameters file.
Example:

./ckdgen-pheno-prep.sh params-gckd.txt

Windows or Mac versus Unix environment: to avoid problems with line breaks when you run the script in a Unix environment, use the following commands to convert the files prior to running the script:

- "dos2unix <input_filename.txt>" for input files generated under Windows
- "mac2unix <input_filename.txt>" for input files generated using MAC

The script will calculate analysis-ready phenotypes for use in the GWAS. Output is written both to the screen and to a log file. Please **examine** the **output carefully** (i.e. the .log and the .errors.csv files). If there are problems, try to adjust the parameters in the parameters file or the data in your input file according to the error messages.
Note that if your study only has one stratum of variables for which stratified phenotypes are generated, this stratum will be named "*_overall" by the script (e.g., if all participants are male, you will only have "uric_acid_overall" in the output.

If you need assistance, do not hesitate to contact us.

Upload both **summary output files** (**both .summary.pdf and .summary.txt,** e.g. ckdgen-pheno-SHIP-1-201606210921.summary.pdf and ckdgen-pheno-SHIP-1-201606210921.summary.txt) generated by the script along with the GWAS results data as detailed in **section 5** (do not upload the .phenotype.txt file).

3. Running of GWAS

3.1 General instructions

- Please contribute whichever traits listed in **Table 2** are available from your study.
- Minimum sample size for binary phenotypes: for a given stratum, **at least 100 cases and 100 controls** are required. Do not run GWAS if your sample size is smaller.
- GWAS should be run assuming an **additive genetic model**.
- Run GWAS by chromosome and upload **one file per chromosome**.
- **No GC correction**: do not apply genomic control correction to GWAS results.
- **Multi-ethnic studies**: perform ancestry-specific analyses, **do not combine ethnicities**. In case of difficult dissection of the ethnic groups, please get in touch with us.
- **X chromosome**: perform **sex-stratified GWAS for the X** chromosome. Male genotypes should be coded 0/2 in the non-pseudoautosomal region of the X (see **Appendix** for details on the genotype imputation of this region).

3.2 GWAS

All GWAS will be based on the two following models:

Continuous phenotypes:	analysis-ready phenotype* ~ SNP + study-specific covariates** + PCs#
Binary phenotypes:	analysis-ready phenotype* ~ SNP + age and/or sex## + study-specific covariates** + PCs#

where:

*__Analysis-ready phenotype:__ Use the output variables generated by the script (**section 2.3**)

**__Study-specific covariates__ reflecting characteristics of the study design, e.g., different recruitment centers.

#__PCs__: genetic principal components; each study should account for population stratification or family/pedigree substructure using the most appropriate method such as PC adjustment or linear mixed models based on kinship coefficients, respectively.

##__Age and/or sex__: as specified for each trait in **Table 2**.

Table 2: Overview of all requested GWAS (not all phenotypes may be available in your study)

Outcome: analysis-ready phenotype (output from script)	Description of outcome	Regression model	Covariates to be included in the GWAS
eGFR_overall	Age- and sex-adjusted residuals of ln(eGFR)	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFR_nonDM	Age- and sex-adjusted residuals of ln(eGFR) among those without diabetes	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFR_DM	Age- and sex-adjusted residuals of ln(eGFR) among those with diabetes	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
creatinine_overall	Age- and sex-adjusted residuals of ln(crea)	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)

UACR_overall	Inverse normal transformed age- and sex-adjusted residuals of ln(UACR)	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
UACR_DM	Inverse normal transformed age- and sex-adjusted residuals of ln(UACR) among those with diabetes	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
UACR_nonDM	Inverse normal transformed age- and sex-adjusted residuals of ln(UACR) among those without diabetes	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
bun_overall	Age- and sex-adjusted residuals of ln(bun) [calculated from urea]	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
uric_acid_overall	Age- and sex-adjusted residuals of uric acid	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
uric_acid_men	Age-adjusted residuals of uric acid among men	linear	If needed: study-specific covariates (e.g., study site,PCs, etc.)
uric_acid_women	Age-adjusted residuals of uric acid among women	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
CKD_overall	CKD as generated from script	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
CKD_DM	CKD as generated from script among those with diabetes	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
CKD_nonDM	CKD as generated from script among those without diabetes	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
MA_overall	MA as generated from script	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
MA_DM	MA as generated from script among those with diabetes	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
MA_nonDM	MA as generated from script among those without diabetes	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
Gout_overall	Gout as generated from script	logistic	**age, sex**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
Gout_men	Gout as generated from script among men	logistic	**age**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
Gout_women	Gout as generated from script among women	logistic	**age**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFRdecline	**Only for studies with prospective data** Age-, sex- and baseline eGFR-adjusted residuals of eGFRdecline	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFRdecline_DM	**Only for studies with prospective data** Age-, sex- and baseline eGFR-adjusted residuals of eGFRdecline in diabetes	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFRdecline_nonDM	**Only for studies with prospective data** Age-, sex- and baseline eGFR-adjusted residuals of eGFRdecline in non-diabetes	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFRdecline_CKD	**Only for studies with prospective data** Age-, sex- and baseline eGFR-adjusted residuals of eGFRdecline in CKD	linear	If needed: study-specific covariates (e.g., study site, PCs, etc.)
Rapid3	**Only for studies with prospective data** Rapid3 as generated from script	logistic	**age, sex, baseline eGFR**; if needed: study-specific covariates (e.g., study site, PCs, etc.)

Rapid3_DM	**Only for studies with prospective data** Rapid3 as generated from script among those with diabetes	logistic	**age, sex, baseline eGFR**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
Rapid3_ nonDM	**Only for studies with prospective data** Rapid3 as generated from script among those without diabetes	logistic	**age, sex, baseline eGFR**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
iCKD25	**Only for studies with prospective data** Incident CKD as generated from script	logistic	**age, sex, baseline eGFR**; if needed: study-specific covariates (e.g., study site, PCs, etc.)

GWAS software options:

A) [**recommended**] **EPACTS** pipeline

We recommend that all studies use the EPACTS pipeline. The full pipeline, adapted to the CKDGen needs and including specifications of the options, is reported here:

https://ckdgen.eurac.edu/mediawiki/index.php/CKDGen_Round_4_EPACTS_analysis_plan

By following this pipeline, results will be automatically ready for upload, with no additional formatting required. You can skip to section 5 for result upload.

B) **SNPtest** v2 or later

If you use SNPtest v2 or later, specify the "-frequentist 1", "-method expected", "-call_thresh 0.0001" and "–use_raw_phenotypes" options for analyses of the autosomes and the X chromosome. **Results must be formatted as described in Section 4**.

C) **Other** pipelines

If you prefer using your own pipeline, this is fine. In this case, **results must be formatted as described in Section 4**.

3.3 X chromosome analyses

Only for the phenotypes listed in **Table 3**, run the analyses on chromosome X **for males and females separately**, using one of the analysis pipelines listed in section 3.2. We recommend using the same software for autosome and X chromosome analyses. For imputation of X chromosome genotypes, see **Appendix**.

Table 3: Overview of requested GWAS from the X chromosome

Outcome: analysis-ready phenotype (output from script)	**Covariates to be included in the GWAS**
eGFR_overall	If needed: study-specific covariates (e.g., study site, PCs, etc.)
creatinine_overall	If needed: study-specific covariates (e.g., study site, PCs, etc.)
UACR_overall	If needed: study-specific covariates (e.g., study site, PCs, etc.)
bun_overall	If needed: study-specific covariates (e.g., study site, PCs, etc.)
uric_acid_overall	If needed: study-specific covariates (e.g., study site, PCs, etc.)
CKD_overall	**age**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
MA_overall	**age**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
Gout_overall	**age**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
eGFRdecline	**Only for studies with prospective data** If needed: study-specific covariates (e.g., study site, PCs, etc.)
Rapid3	**Only for studies with prospective data** **age, baseline eGFR**; if needed: study-specific covariates (e.g., study site, PCs, etc.)
iCKD25	**Only for studies with prospective data** **age, baseline eGFR**; if needed: study-specific covariates (e.g., study site, PCs, etc.)

4. Content and formatting of GWAS output for studies not using EPACTS

Studies using the **EPACTS** association **pipeline** can **skip to section 5**. All other studies:

a. General instructions for the GWAS summary statistics files for submission

- **Exclude** variants with invalid associations (**missing beta or missing SE**).
- **Do not pre-filter** on **allele frequency or imputation quality**.
- If your study automatically pre-filters on these, please let us know.

b. Formatting of the GWAS summary statistics files

- Submitted summary files should be **tab-delimited**.
- **Missing** information should be coded "**NA**".
- Include one row per variant (SNPs or indels).
- Include columns for **chr and pos (b37)**. Because of different imputation reference panels across cohorts, this **is crucial information** for variant harmonization.
- Include **all columns** shown in **Table 4 and use** the name exactly as in the "**column name**" column.

Table 4: Column headers for GWAS summary files for studies not using the EPACTS pipeline

Column name	Description
RSID	Variant identifier. Use the variant identifier exactly as it is represented in your results. We will convert these to a common ID during data cleaning and meta-analysis.
chr	**Mandatory.** Chromosome number. Use "X" for chromosome X.
position	**Mandatory.** Physical position for the reference sequence (only build 37/hg19)
coded_all	**Mandatory.** Coded allele, also called modeled or effect allele (in example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G). Use A/C/G/T or the applicable indel allele coding as present in your results data (do not recode alleles to R/D/I).
noncoded_all	**Mandatory.** The alternate allele. Use A/C/G/T or the applicable indel allele coding as present in your results data (do not recode alleles to R/D/I).
beta	Beta estimate from genotype-phenotype association, at least 5 decimal places, variants with missing beta estimates should be excluded.
SE	Standard error of beta estimate, to at least 5 decimal places, variants with missing SE should be excluded.
Pvalue	p-value of test statistic, "NA" if not available
AF_coded_all	Allele frequency of the coded allele, "NA" if not available
n_total	Total sample with phenotype and genotype for variant
used_for_imp	1/0 coding; 1=used for imputation, 0=not used for imputation
IQ	Imputation quality: take r2 values if minimac was used for imputation, info values if IMPUTE2 was used

5. Upload instructions and timeline

a. What to upload

i. Studies using EPACTS

- **All *epacts.gz files** (GWAS results for each chromosome); use naming convention indicated in **Table 5**.
- The **two .summary.pdf and .summary.txt files** from the phenotype generation script; *e.g.: ckdgen-pheno-SHIP-0-201606210921.summary.txt, ckdgen-pheno-SHIP-0-201606210921.summary.pdf*; do not upload individual level data files.
- The **.info file** generated during imputation. Please combine information for the different chunks of all chromosomes into one file.

ii. Studies not using EPACTS

- Formatted **GWAS results files**; use naming convention indicated in **Table 5**.
- The **two .summary.pdf and .summary.txt files** from the phenotype generation script; *e.g.: ckdgen-pheno-SHIP-0-201606210921.summary.txt, ckdgen-pheno-SHIP-0-201606210921.summary.pdf*; do not upload individual level data files.

b. Naming convention of GWAS and other files

Please name all files to be uploaded (GWAS results/GWAS summary statistics and imputation quality files) as follows:

A_B_C_D_E_F.<original_file_extension>.

where

Table 5: Instructions for file naming convention.

A	Your study's name
B	Ethnicity; use EA for European ancestry, AA for African American, AFR for African, EA for East Asian, SA for South Asian, HIS for Hispanic, IA for Indian ancestry or as applicable
C	The analyzed study trait, e.g. "eGFR_overall"
D	Imputation reference panel, use "1KGPph1v3", "1KGPph3v5" or "HRC", as applicable
E	Chromosome, use "chrXX" for autosomes (e.g.: "chr03") and "chrX_F" and "chrX_M" for X chromosome analyses on females and males, respectively.
F	Date, use YYYYMMDD

Ex: ARIC_AA_eGFRoverall_1KGPph3v5_chr20_20160530.txt, ARIC_AA_1KGPph3v5_20160530.info

Output files from the phenotype generation script does not need to be renamed.

c. Where to upload

Upload all output to:

https://ckdgen.eurac.edu/upload/

User name: ckdgenR4

Password: ExcitingScience!

Notice: file size limit is 4GB.

d. Timeline

Please upload all files by September 30th, 2016.

When you finish uploading, please inform us with an email to *ckdgenconsortium@gmail.com*, indicating your study and your name.

e. Cohort-specific information: funding, acknowledgements

Complete cohort-specific information sheet (https://docs.google.com/spreadsheets/d/11pGt-_LvGVT6OLShsbtcET8gRBRtgbnIUbK-XEBbqMw/edit?usp=sharing), including

i. authors and affiliations
ii. acknowledgements
iii. study information
iv. genotyping information
v. author contributions
vi. conflict of interest

Thank you very much for your participation in the CKDGen Consortium analyses!

Appendix

Additional information regarding genotyping and imputation

1.1 Imputation Reference Haplotype Panels

The preferred haplotype reference panel for imputation is the Haplotype Reference Consortium (HRC) panel for European ancestry studies, and the 1000G phase 3 v5 ALL panel (excluding monomorphic sites and singletons, including chromosomes 1-22, and X) for studies of non-European ancestry, or if HRC is not available. We will also accept other densely imputed data, when the reference haplotype panel was a 1000G phase 1 v3 ALL panel or later.

1.2 Sample and Variant Quality Control

Each study is responsible for their own QC using appropriate filters. Standard procedures include removing samples of low genotyping call rate, mismatch between genotypic and phenotypic sex, excess heterozygosity, first-degree relatives for non-family based studies and outlying genetic ancestry. Prior to imputation, please ensure usage of high-quality variants by filtering the SNPs on your existing exclusion criteria for call rate, minor allele frequency and HWE p-value.

1.3 Lift over of genotype data to NCBI b37 (hg19)

To match current releases of the reference haplotype panels, please lift over your genotype data so that they have b37 coordinates before starting imputation. You can find information on how to perform the liftover here http://genome.sph.umich.edu/wiki/LiftOver.

1.4 Align all SNPs to the positive (+) strand

To match the imputation reference panels, all SNPs need to be expressed relative to the + strand of the human reference genome sequence before imputation.
Useful resources:

- Wrayner files for strand flipping: http://www.well.ox.ac.uk/~wrayner/strand/index.html
- Genotype Harmonizer: https://github.com/molgenis/systemsgenetics/wiki/Genotype-Harmonizer
- checkVCF: https://github.com/zhanxw/checkVCF

1.5 Resources for imputation

Please follow one of these protocols for two-step imputation and use standard settings:

1000 Genomes imputation phase 3 v5 ALL, recommended for studies of non-European ancestry or if HRC is not available:

- IMPUTE2: http://genome.sph.umich.edu/wiki/Impute2:_GIANT_1000_Genomes_Imputation_Cookbook
- Minimac: http://genome.sph.umich.edu/wiki/Minimac:_GIANT_1000_Genomes_Imputation_Cookbook
- Minimac3: http://genome.sph.umich.edu/wiki/Minimac3_Imputation_Cookbook

HRC (Haplotype Reference Consortium) version 1.1 (please, contact us if you are using previous versions), recommended for European-ancestry studies, using the available imputation servers:

- Michigan Imputation Server: https://imputationserver.sph.umich.edu/ (Eagle/minimac3)
- Sanger imputation service: https://imputation.sanger.ac.uk/ (Eagle/PBWT)

Recommended options (default): Reference Panel: HRC r1.1 2016; Phasing: Eagle (phased output); Population (for the allele frequency check): EUR; Mode: Quality Control & Imputation

1.6 Use pre-generated reference haplotype panels

The following 1000G Phase 3 version 5 pre-formatted reference haplotype panels are recommended for genotype imputation:

- **Impute2**: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
- **Minimac3**: ftp://share.sph.umich.edu/minimac3/G1K_P3_M3VCF_FILES_WITH_ESTIMATES.tar.gz

1.7 X chromosome imputation

See instructions in the Impute2, minimac and HRC cookbooks, section 1.5 above. Males should be coded 0/2 on the non-pseudoautosomal region of the X. The non-pseudoautosomal region spans between 2,699,521 and 154,931,043 (build: hg19) base-pairs (http://genome.sph.umich.edu/wiki/Minimac#X_Chromosome_Imputation).

1.8 Genotype imputation by chunks

Imputation by genome chunks is standard in IMPUTE (see the software website for more details). As there are considerable time savings, imputation by genome chunks should also be used for Minimac imputation (2500 marker chunks, with 500 marker overhang on each side of the chunk). See section on "Further Time Savings" in Minimac cookbook.