DIAGRAM v.4 1000 Genomes (March 2012 release) Imputation and Meta-Analysis Data Sharing Document and planning for GOT2D imputation

Version 31-aug-2012. Inga Prokopenko, Clement Ma, Christian Fuchsberger, Laura Scott, Mark McCarthy

Following our initial 1000 Genomes reference panel imputation within DIAGRAM, we are planning another round of imputation using the 1000G reference panel and subsequently the GOT2D reference panel to identify novel variant associated with T2D. Our goal is to perform imputation across the genome (all chromosomes) for the 1000 Genomes reference panel and to prepare for the GOT2D based imputation.

Main features (more details below)

1. 1000G Imputation reference haplotype panel. For the new round of 1000G imputation we will use ALL haplotypes (e.g. from all ethnicities), excluding monomorphic and singleton sites, all chromosomes (1-22, X), from the phase one integrated variant release v3, March 2012. This reference panel was agreed upon after extensive testing and discussion within the GIANT consortium (including many DIAGRAM members) to enable the best overall imputation quality within and across multiple ethnic groups.

2. Imputation using genome chunks. It is much faster to impute chunks of the genome rather than whole chromosomes.

3. SNP-T2D analysis method or software for common and for less frequent variants. For common variants, the most often used tests (Wald and Likelihood ratio) have good type 1 error rate for meta-analysis. Thus, for the initial analysis please use your existing logistic regression software on all SNPs. In contrast, for less frequent variants, the score test and Firth logistic regression tests show (often substantially) better control of type 1 error rate than Wald or Likelihood ratio tests. We will distribute analysis software that can run the score and Firth logistic regression tests in MID September.

TIMELINES (imputation from large-scale panels generated from re-sequencing and analysis)

PHASE I (1000 Genomes, March 2012 reference panel):
14th of September - genotype data is pre-phased
28th of September – cohorts are imputed to March 2012 1000Genomes
13th of October – association results are submitted

PHASE II (GoT2D reference panel):
15th of October – GoT2D haplotypes are released
MID Nov or~ 4 weeks after GoT2D haplotypes release – imputation in cohorts is done
Start Dec or ~5-6 weeks after GoT2D haplotypes release – association analysis results are submitted

Sample and genotype exclusion criteria and quality control for pre-phasing of GWAS data

1. Use your existing exclusion criteria for sample call rate, gender checks and sample heterogeneity.
2. Use your existing exclusion cut-offs for the SNPs call rate, HWE.
3. You can use your existing minor allele frequency cut-offs for SNP inclusion. The choice of MAF cut-point should reflect the accuracy of the genotyping or lower frequency SNPs.

4. **Lift-over your genotype data to NCBI genome build 37 (hg19)** to match current releases of 1000 Genome Project data.
5. Make sure the SNP alleles are aligned to "+" strand of the reference genome.

Imputation strategy

We ask you to follow agreements about imputation strategy, decided in collaboration with GIANT consortium.

1. Follow the Minimac and IMPUTE protocols for imputation

IMPUTE2: use standard settings with pre-calculated Build 37 recombination maps
http://genome.sph.umich.edu/wiki/Impute2:_GIANT_1000_Genomes_Imputation_Cookbook
Minimac: use standard settings which calculate recombination maps on-the-fly
http://genome.sph.umich.edu/wiki/Minimac:_GIANT_1000_Genomes_Imputation_Cookbook

2. Use the pre-made 1000G reference haplotype panels for Minimac and IMPUTE.

IMPUTE2:
https://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute_macGT1.tgz
readme file:
https://mathgen.stats.ox.ac.uk/impute/README_1000G_phase1integrated_v3_macGT1.txt
More details about this panel are given on the following page:
https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html

Minimac:
http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-03-14.html

3. Imputation by genome chunks is standard in IMPUTE (see the software website for more detail). 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 web protocol.

4. Provide log files from chromosome 20 after running: Please provide command line output for chromosome 20 to facilitate the identification of imputation quality issues caused by, for example, SNP identifier mapping problems between target and reference panel, lift over and parameter setting problems.

For Minimac: by saving and uploading the screen output during imputation for all chunks on chromosome 20.
Minimac [parameters] | the studyname.chr20.log

For IMPUTE2: by uploading all log files for chromosome 20 imputation step.

Genetic Modelling and Analysis

FOUR separate analyses should be performed for each dataset: your traditional analysis using logistic regression **without/with adjustment for BMI**, and an analysis for variants with low minor allele counts **without/with adjustment for BMI**:

1. For all variants perform an analysis on a log-additive scale, i.e. using logistic regression. Use the allele count for genotype data and the allele dosages or genotype probabilities for imputed data. This analysis can use the Wald, likelihood ratio or score test to evaluate the significance.

A. Do not include adjustment for BMI.

B. Include adjustment for BMI.

2. For **variants with low minor allele counts** Wald and likelihood-ratio based tests have deflated and inflated type 1 errors, respectively. We are working on an analysis package that will run the score and Firth bias corrected logistic regression. We will distribute the package shortly.

A. Do not include adjustment for BMI.

B. Include adjustment for BMI.

ANALYSIS options and notes:

- **Include covariates** of AGE, SEX, and population genetic structure or geographic ascertainment (if applicable).
- For the **X chromosome**, ensure that males are coded as homozygous for the allele they carry, and adjust for sex in addition to any other cohort-specific covariates.
- **Do not adjust summary statistics** (p-value, OR, standard error) for genomic control.
- Please submit the **results for all imputed SNPs** (regardless of imputation quality, and for all genotyped SNPs).
- Before submitting results of an analysis you may want to use the **SNP name checking tool** to verify the SNP names are valid. You can access the tool in the "GWAS result QC and SNP harmonization" section of http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook.

Data File General content Requirements:

- Data should be reported on **NCBI BUILD 37**, and in all cases the build information should be stated in the file. This is extremely important with multiple releases of the 1000Genomes data that might differ significantly by many parameters.

- All numeric data can be specified in either scientific or decimal notation and should be specified to 4 or 6 decimal places, as indicated below. P-values should be specified to 4 **significant** digits.

- Remove monomorphic SNPs and SNPs for which no association data is available.

Overview of Checklist of items to include for submission

1. Final check that the provided files are properly formatted
2. Check that all items in the provided "Checklist" have been appropriately reviewed
3. Data File can contain both, imputed and genotyped SNPs in one file. Alternatively 2 files can submitted, if you directly genotyped and imputed SNPs separately) uploaded to the FTP website. In either case only one copy of the SNP association results should be submitted for each SNP.
4. An email sent to Inga Prokopenko (inga@well.ox.ac.uk) indicating the upload has been sent

Data File Requirements

- White-space (not tab) delimited text file, one row per genotyped or imputed SNP, the first row with a header with the labels given below, with the requested information in the following columns (see below).

- The format for the file name should be:

DIAGRAMv4_QQQ_XXX_1000G _KKK_TTT_YYY_ZZZ.txt
OR
DIAGRAMv4_QQQ_XXX_adjBMI_1000G _KKK_TTT_YYY_ZZZ.txt

where:

QQQ: either "SNPs", "gSNPs" or "iSNPs". Use "SNPs" for the file, which contains both imputed and genotyped SNPs, "gSNPs" for the file which contains only genotype data, "iSNPs" for the file which contains imputed data.

XXX indicates a uniquely identifiable STUDY NAME: (e.g. WTCCC, DGI, DGDG, FUSION, ERGO, DUNDEE, NHS, FHS, TYROL, EUROSPAN etc.)

"_adjBMI" – should be present for the model with BMI adjustment

KKK indicates date of the 1000Genomes map, Month and year (defined as MMMYY) of the 1000Genomes map that was used e.g. MAR12

TTT indicates the test used to evaluate significance
SCR = Score
WLD = Wald
LHR = Likelihood ratio
FBC = Firth Bias Corrected Logistic Regression

YYY indicates the DATE of file generation (MMDDYY format, e.g. 021710 – apologies in advance to our European colleagues)

ZZZ indicates the name + other initials of the uploader (e.g., BFV, LJS, ABC, etc.)

- Please use the period character "." to denote missing data or data not available.
- For calculating allele frequencies of imputed SNPs, please provide mean of the DOSAGES/2 rather than calculating frequencies based on genotypes called at pre-specified threshold.

- For genotyped data, please provide all data that passed quality control thresholds.
- Please, provide either one (all SNPs) or two files (one containing analysis of imputed data, and a second for analysis of genotype data).
- For imputed data, please provide imputed results for all non-genotyped SNPs regardless of imputation accuracy.
- If the estimated SE on the effect size (in of the allelic odds ratio) is greater than 10, we recommend setting the effect size statistics (BETA, SE) to the missing data value (the period character, ".").
- Generally, we acknowledge that some details of the recommended formats may not be available or may not be readily sharable (for example, BAYES_FACTOR, or N0/N1/N2 exact genotype counts). In those cases, *we recommend including* **all columns** as a general practice, but replacing the information with the missing data character (".").
- We strongly plead for/will only accept files with no deviations from the file naming convention, order of the columns, or information provided in the columns (no extra information, please). Please check the following items before you submit to the FTP repository. To help ensure data integrity and the sanity of the person responsible for initially processing the data, data that appears to be mal-formed will be returned for reformatting.

Data Exchange Repository and Login Information

Information on upload to Broad FTP site (to upload data):

1. Connect via ftp to ftp.broadinstitute.org
2. Username is "anonymous"; password is your email address
3. Upload to the /incoming directory
4. Once complete, send an email to Inga Prokopenko (inga@well.ox.ac.uk) and indicate:
    - Your name
    - The name of the file you uploaded
    - Your study affiliation (e.g. DGI, WTCCC).
5. Once the email is received, we will endeavour to post them on a password protected website.

--> Please note that even though many users can see the names of files you have uploaded, these files cannot be copied or viewed by other users who log in anonymously.
--> Files uploaded here will be purged in 1-2 month intervals, or once files from a large number of groups have been received.

Information on how to access the DIAGRAM website:

1. Connect to www.broadinstitute.org/~bvoight/DIAGRAM
2. Username: DIAGRAM_analyst
Password: 4GzPM1uf

--> If you share login information with other people, please, let Ben Voight (bvoight@upenn.edu) and Inga Prokopenko (inga@well.ox.ac.uk) know as he is keeping an informal tally on which he provides the login information for.
--> Please note that the login information is subject to periodic change.

Tom Cruise-Blue-Jean-Awesome Funtime Checklist (complete before submission):

__ Have you used new SNPID naming system?
__ Have you removed monomorphic SNPs and SNPs for which no association data is available?
__ Have you separated columns by white-space (not tab delimited), and not by comma or another non-white-space delimiter?
__ Have you excluded quotation marks in file (i.e. entries are not encapsulated by quotation marks – Excel does this if you ask it to)
__ Have you included a header in the data file?
__ Are your entries in the same order as requested below?
__ Are missing entries filled with the pre-approved missing data character (i.e. the period character ".")?
__ Have all control character sequences been removed from the file (e.g., ^M)?
__ Have blank lines been removed from the file?
__ Have extra blank columns been removed from the file?
__ Have you checked that all rows have the expected number of columns?
__ Does data provided for all entries in each column match what is expected in the given column (e.g. ALLELES should only have ACGT; p-value columns will be numeric from [0-1] with "." acceptable, etc.)
__ Have you checked that all SNPs are provided and that the parsed data set contains the expected number of SNPs that you calculated statistics on?
__ Have you checked that imputed SNPs are denoted as "1", and genotyped SNPs denoted as "0" for the IMP? column?
__ Have you checked that the number of genotyped and imputed SNPs matches what you analyzed?
__ Have all duplicated entries been removed and only unique entries exist for every SNP id in the file?
__ Have you checked that (a) an appropriate number of significant figures are printed, (b) are in scientific notation (e.g. 1.2 x 10^{-10}), and (c) are not truncated?
__ Is the standard error always non-negative?
__ Have you had another person verify that the data file you have produced conforms to the rules?

RESULTS FORMAT TABLE (white-space text file)
one row per genotyped or imputed SNP, with the following columns:

Column header	Description	Format	Examples
SNPID	SNP ID label	Chromosome:position:type, where type is 'SNP':'CNV':'INDEL' … (use capitals); i.e. CHR:POSTION:SNP (e.g. 1:100982347:SNP). Use '1' … to '22' for the respective autosomal chromosomes Use '23' for the NON-pseudoautosomal regions of the X---chromosome Use '25' for the pseudoautosomal regions of the X---chromosome	11:12222555 :CNV
RSID	dbSNP rsID label (if easily available)	rs number from dbSNP, or set to missing if no dbSNP number exists	rs693
STRAND	Orientation of the site to the human genome strand used	+ (Should be aligned to forward strand)	+
BUILD	Build of the genome on which the SNP is oriented	Numeric	37.1

CHR	Chromosome on which SNP resides	Use '1' ... to '22' for the respective autosomal chromosomes Use '23' for the NON-pseudoautosomal regions of the X---chromosome Use '25' for the pseudoautosomal regions of the X---chromosome	1
POS	Position of SNP on chromosome	Base pairs on human genome build used	34000345
EFFECT_ALLELE	Allele at this site to which the effect has been estimated	Capital letter (A,C,G,T)	A
NON_EFFECT_ALLELE	Other allele at this site.	Capital letter (A,C,G,T)	G
N_CASES	Total number of cases analysed	Numeric, integer	1243
N_CONTROLS	Total number of controls analysed	Numeric, integer	1243
N0_CASES	Number of homozygous cases with zero copies of the EFFECT_ALLELE	Numeric, integer or float with 4 digits to the right of the decimal (imputed)	623 745.2345
N1_CASES	Number of heterozygous cases with one copy of the EFFECT_ALLELE	Numeric, integer or float with 4 digits to the right of the decimal (imputed)	6235 745.2345
N2_CASES	Number of homozygous cases with two copies of the EFFECT_ALLELE	Numeric, integer or float with 4 digits to the right of the decimal (imputed)	623 745.2345
N0_CONTROLS	Number of homozygous controls with zero copies of the EFFECT_ALLELE	Numeric, integer or float with 4 digits to the right of the decimal (imputed)	623 745.2345
N1_CONTROLS	Number of heterozygous controls with one copy of the EFFECT_ALLELE	Numeric, integer or float with 4 digits to the right of the decimal (imputed)	623 745.2345
N2_CONTROLS	Number of homozygous controls with two copies of the EFFECT_ALLELE	Numeric, integer or float with 4 digits to the right of the decimal (imputed)	623 745.2345
EAF_CASES	Allele frequency of the EFFECT_ALLELE in cases analysed	Frequency with 4 digits to the right of the decimal	0.3546
EAF_CONTROLS	Allele frequency of the EFFECT_ALLELE in controls analysed	Frequency with 4 digits to the right of the decimal	0.3546
HWE_P_CASES	Exact HWE p-value for the cases analysed, only if genotyped data is used in analysis	Scientific E notation with 4 digits to the right of the decimal (set to missing if imputed)	1.122E-02 .
HWE_P_CONTROLS	Exact HWE p-value for the controls analysed, only if genotyped data is used in the analysis	Scientific E notation with 4 digits to the right of the decimal (set to missing if imputed)	1.123E-02 .
CALL_RATE	Call rate for this SNP	Frequency 4 digits to the right	0.9936

	across cases and controls, only for genotyped data	of the decimal. Set equal to 1.000 if IMP? = 1.	
BETA	Estimate of the allelic effect, defined as the natural logarithm of the odds ratio, ln(OR)	Numeric float with 6 digits to the right of the decimal (set to missing if poorly estimated)	0.203666 .
SE	Estimated standard error of the ln (OR) estimate of the allelic effect, uncorrected for genomic control	Numeric float with 6 digits to the right of the decimal (set to missing if poorly estimated)	0.561166 .
PVAL	Significance of the variant association, uncorrected for genomic control	Scientific E notation with 4 digits to the right of the decimal (set to missing if BETA and SE are missing)	3.244E-10 .
IMPUTED	Is the SNP imputed?	0 = Genotyped 1 = Imputed	1
INFO_TYPE	Type of information provided in the INFO column	0 = SNP is genotyped 1 = "r2_Hat", e.g. RSQR column in mach2dat output 2 = "proper_info" from SNPTEST 3 = "INFO" from PLINK 4 = Information from other imputation software	1
INFO	Measure of information content for the imputed SNP result (range 0-1)	Numeric float with 4 digits to the right of the decimal (set to missing if imputed)	0.4832 .