**International Endometriosis Genomics Consortium (IEGC-ESP)**

**Study Level Analysis Plan: Imputation & GWAS analyses of individual datasets**

**Analysis Plan Authors** Nilufer Rahmioglu, Kathryn Terry, Reedik Magi, Mette Nyegaard, Andrew Morris, Krina Zondervan, Stacey Missmer

*Defining Endometriosis Physiologic Sub-Phenotypes and Subsequent Cancer and Comorbidities Risk Through Discovery of Novel Genetic Variant – ESP (*USAMRAA W81XWH-20-PRMRP-IIRA)

Principal Investigator: Stacey Missmer, Sub-site PIs: Kathryn Terry, Krina Zondervan

**Rationale**

Endometriosis has an estimated heritability of ~50% 1,2, with ~26% attributable to common genetic variants3. Our most recent collaborative GWAS meta-analysis of endometriosis identified 49 distinct association signals with a sample size of 60,674 cases and 701,926 controls4. These analyses showed that the larger effect sizes observed for stage III/IV disease is driven by ovarian endometriosis highlighting potential genetic sub-types of endometriosis. While we also tested the association of these 49 lead SNPs with pain and surgical features, no genome-wide level association analyses were conducted apart from stage-based sub-groups and infertility. Moreover, our genetic correlation analysis illustrated the significant genetic correlation between overall endometriosis and chronic pain and inflammatory conditions. Each of these explorations relied on the 49 SNPs associated with endometriosis overall. Analyses attempting to discover loci associated with subphenotype-defined endometriosis cases that may not be associated with endometriosis overall were not explored.

New endometriosis GWAS datasets have become available since, bringing the total of endometriosis cases to ~100K with doubling of the original numbers of case with subphenotypic details available - facilitating further discovery. The aim of this current round of meta-analysis is to conduct de-novo phenotype-stratified GWAS: (1) dissecting ovarian, deep and superficial disease, (2) investigating cases with different pelvic pain symptomatology, and (3) exploring adenomyosis (endometriosis of the uterus) associated signals and their sharing with endometriosis and/or its sub-types. Furthermore, building upon the observational associative and genetic evidence illustrating comorbidity between endometriosis and other female reproductive, cardio-metabolic, inflammatory, autoimmune, chronic pain conditions and cancers4-11, we will further investigate the genetic correlation and shared genetic basis between these conditions and endometriosis and its sub-types. For identified genetic variants we will conduct functional annotation and pathway enrichment analyses to understand the underlying potential mechanisms for endometriosis and its subtypes. Depending on findings we anticipate that these discoveries will result in one, possibly two, comprehensive manuscripts.

1. Treloar SA, O'Connor DT, O'Connor VM, Martin NG. Genetic influences of endometriosis in an Australian twin sample. Fertil.Steril.

1999;71(4):701-10.

2. Saha R, Pettersson HJ, Svedberg P, Olovsson M, Bergqvist A, Marions L, et al. Heritability of endometriosis. Fertil Steril 2015;104(4):947-52.

3. Lee SH, Harold D, Nyholt DR, Goddard ME, Zondervan KT, Williams J, et al. Estimation and partitioning of polygenic variation captured by

common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet 2013;22(4):832-41.

4. Rahmioglu N, Mortlock S, Ghiasi M, Moller PL, Stefansdottir L, Galarneau G, et al. The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. Nat Genet 2023;55(3):423-436.

5. Lu Y, Cuellar-Partida G, Painter JN, Nyholt DR, Morris AP, Fasching PA, et al. Shared genetics underlying epidemiological association

between endometriosis and ovarian cancer. *Hum Mol Genet* 2015;24(20):5955-64.

6. Rahmioglu N, Macgregor S, Drong AW, Hedman AK, Harris HR, Randall JC, et al. Genome-wide enrichment analysis between endometriosis and obesity-related traits reveals novel susceptibility loci. *Hum Mol Genet* 2015;24(4):1185-99.

7. Kvaskoff et al. Endometriosis and cancer. A systematic review and meta-analysis. Hum Reprod Update 2021

8. Farland et al. Laparoscopically oncfimred endometriosis and risk of incident stroke: A prospective cohort study. Stroke 2022

9. Mu et al. Association between endometriosis and hypercholesterolemia or hypertension. Hyptertension 2017

10. Shafrir et al. Co-occurrence of immune-mediated conditions and endometriosis among adolescents and adult women. Am J Reprod Immuol 2021

11. Kvaskoff et al., Endometriosis: a high-risk population for major chronic diseases? Hum Reprod Update 2015.

**Time-line**

It is assumed in this round that each study will have imputed genotypes to a recent reference panel appropriate to their population (HRC for European ancestry; 1000G v3 for non-European ancestry). Each study will test for association with endometriosis (Minimum N>=300 endometriosis cases) and its subtypes (Minimum N>=100 cases with surgical subtypes, rASRM stage and infertility; Minimum N>=50 cases with pain subtypes / adenomyosis) where available using the analysis plan attached. The summary statistics from the GWAS will be uploaded to the dropbox link provided to each group.

**Deadline of summary statistics upload:** **July 29th, 2024**

In the meantime, any analysis queries should be directed via e-mail to the IEGC Central Analysis Team at the following address: nilufer.rahmioglu@wrh.ox.ac.uk

**Analysis Plan for individual datasets**

This document includes the following analysis protocols to be conducted for each dataset:

Phenotype definitions................................................................................................................................................2

Imputation protocol...................................................................................................................................................3

Uploading imputation QC data..................................................................................................................................8

Association analyses..................................................................................................................................................9

Uploading association results and descriptive QC data...........................................................................................11

**IEGC Consortium**

**Phenotype definitions**

***Case/control definitions***

We rely on the expertise of each group to define case and control groups most appropriate for their datasets and population sampling schemes. We aim to be inclusive here, with definitions spanning from self-reported endometriosis to endometriosis extracted from medical records to claims / national ICD code documentation to radiologic imaging captured cases to surgically confirmed cases (and mixtures). However, it is important that inclusion/exclusion criteria are **well described** **for each dataset**. Each group is asked to complete the **Descriptive/QC spreadsheet sent along with this analysis plan** with the case and controls definitions.

**(1) Overall endometriosis (N=1 GWAS)**

**Cases**

Endometriosis cases can be defined as 1) surgically or imaging confirmed based on medical records; 2) self-reported surgically or imaging confirmed; 3) self-reported (no known surgical or imaging validation); 4) ICD codes (ICD10: N801-N809 or ICD9: 617.1-617.9) in medical records. Describe in the Descriptives/QC spreadsheet the sampling frame from which cases were sampled (e.g. population-based study/biobank; gynaecology clinic; infertility clinic).

**Controls**

Exclude any controls with reported history of endometriosis (if known). Please provide the control definition including the sampling frame (e.g. population-based unscreened or only imaging screened controls; women without endometriosis-surgically confirmed) in the Descriptives/QC spreadsheet. Where a clinical control group is used, provide details on its specific definition (e.g. infertility presenting patients). Also note the sex breakdown if you include males as part of a population-based control group (NB for X-chromosome analyses, male controls will need to be excluded).

**(2) Endometriosis Sub-phenotypes (N=11 GWAS)**

**Cases**

We have defined 5 macro surgical sub-types of endometriosis that will be defined by surgical- and/or ICD-based coding of endometriosis (Table 1): (1) Any case with endometrioma, (2) Any case with deep lesions, (3) Only superficial peritoneal lesions, (4) rASRM Stage I/II disease, (5) rASRM Stage III/IV disease. Please see Table 1 for the case definitions.

We have defined 5 pain-specific sub-types of endometriosis that will be based upon EPHect-based questionnaire data or other defined self-reported questionnaire data and/or ICD based coding of these pain types (Table 1): (1) Any endo case with dyspareunia, (2) Any endo case with acyclical, i.e. non-menstrual pelvic pain, (3) Any endo case with severe dysmenorrhea, (4) Any endo case with gastro-intestinal (GI) pain/IBS symptoms, (5) Any endo case with gynaecological pelvic pain (Dyspareunia, acyclical pain and/or severe dysmenorrhea). Please see Table 1 for the case definitions.

We also have defined an infertile endometriosis sub-type that will be based upon EPHect-based questionnaire data or other defined self-reported questionnaire data and/or ICD based coding female-specific infertility.

**Controls**

Controls exclude anyone with believed history with endometriosis - whether self-reported, working clinical diagnosis, or imaging or surgical confirmation (if known). Keep the same control set utilised for overall endometriosis GWAS for sub-phenotype analyses.

**(3) Adenomyosis (N=1 GWAS)**

**Cases**

Adenomyosis cases will be defined as 1) with hysterectomy and should have no history of endometriosis; 2) ICD codes for adenomyosis with hysterectomy codes but no endometriosis history (ICD10: N80.0 and those with hysterectomy history but no N80.1-9 or ICD9: 617.0 and those with hysterectomy history but no 617.1-9).

**Controls**

Controls with hysterectomy history and no history of endometriosis and/or adenomyosis diagnosis

***Covariates***

Other than ancestry-informative covariates (principal components representing population substructure), in our relatively young datasets it is unlikely we need to adjust for additional covariates as they would not act as confounders (i.e. factors associated and preceding both case/control status AND genotype status in a population). If you intend to include additional covariates to corrections for ancestry in your association analyses, please upload analyses with and without this correction, and define your variables in the Descriptives/QC spreadsheet.

**Table 1.** Case and control definitions for each GWAS.

|  |  |  |
| --- | --- | --- |
| **GWAS** | **Definition of Cases** | **Definition of Controls** |
| Overall endometriosis vs. controls | 1) **Surgically confirmed** endometriosis; 2) **Medical records** - ICD codes (ICD10: **N80.1-N80.9** or ICD9: **617.1-617.9**); 3) **Self-reported** endometriosis diagnosis. | Controls exclude anyone with history of endometriosis (if known) |
| Any case with endometrioma vs. controls | 1) **Surgically confirmed** endometrioma; 2) **Medical records** - ICD codes (ICD10: **N80.1** or ICD9: **617.1**) | Controls exclude anyone with history of endometriosis (if known) |
| Any case with deep lesions vs. controls | 1) **Surgically confirmed** deep lesions; 2) **Medical records** - ICD codes (ICD10: **N80.4, N80.5** or ICD9: **617.4, 617.5**) | Controls exclude anyone with history of endometriosis (if known) |
| Only superficial lesions vs. controls | 1) **Surgically confirmed** superficial lesions but no deep lesion or endometrioma; 2) **Medical records** - ICD codes (ICD10: **N80.3 but no N80.1, N80.2, N80.4, N80.5, N80.6, N80.7, N80.8, N80.9** or ICD9: **617.3 but no 617.1, 617.2, 617.4, 617.5, 617.6, 617.7, 617.8,617.9**) | Controls exclude anyone with history of endometriosis (if known) |
| rASRM stage I/II vs. controls | 1) **Surgically confirmed** stage I/II disease; 2) **Medical records** - ICD codes (ICD10: **N80.3 but no N80.1, N80.2, N80.4, N80.5, N80.6, N80.7, N80.8, N80.9** or ICD9: **617.3 but no 617.1, 617.2, 617.4, 617.5, 617.6, 617.7, 617.8,617.9)** | Controls exclude anyone with history of endometriosis (if known) |
| rASRM stage III/IV vs. controls | 1) **Surgically confirmed** stage III/IV disease; 2) **Medical records** - ICD codes (ICD10: **N80.1, N80.4, N80.5**) or ICD9: **617.1, 617.4, 617.5)** | Controls exclude anyone with history of endometriosis (if known) |
| Any endo case with dyspareunia vs. controls | 1) **EPHect questionnaire or other self-reported questionnaire**: Pelvic pain during or within 24 hours after sexual intercourse: categorized as ever having experienced dyspareunia and also reported worst ever dyspareunia-specific pain severity  2) **Medical records** - ICD codes (ICD10: **N94.1** or ICD9: **625.0**) | Controls exclude anyone with history of endometriosis (if known) |
| Any endo case with acyclical pelvic pain vs. controls | 1) **EPHect questionnaire or other self-reported questionnaire**: Chronic recurrent pelvic pain experienced at any time throughout the menstrual cycle): categorized as ever having experienced acyclic pelvic pain and also reported current acyclic pelvic pain severity; 2) **Medical records** - ICD codes (ICD10: **R10** or ICD9: **789.0**). | Controls exclude anyone with history of endometriosis (if known) |
| Any endo case with severe dysmenorrhea vs. controls | 1) **EPHect questionnaire or other self-reported questionnaire**: Severe dysmenorrhea rating pain 7-10 on the numerical rating scale; 2) **Medical records** - ICD codes (ICD10: **N94.4, N94.5, N94.6** or ICD9: **625.3**) | Controls exclude anyone with history of endometriosis (if known) |
| Any endo case with GI pain/IBS symptoms vs. controls | **1) EPHect questionnaire or other self-reported questionnaire:**Self-reported clinician diagnosed IBS**or**any pain with bowel movements in the past 12 months**or**answered ≥1 day a week during the past ≥3 months acyclical pelvic pain or discomfort associated with 2 or more of the following criteria: (i) related to defecation, (ii) associated with a change in stool frequency, and/or (iii) associated with a change in stool characteristics;**2) Medical records**- ICD codes (ICD10:**K58.0, K58.1, K58.2, K58.8, K58.9 or ICD9: 564.1**). | Controls exclude anyone with history of endometriosis (if known) |
| Any endo case with any pelvic pain vs. controls | 1) **EPHect questionnaire or other self-reported questionnaire**: pelvic pain during or within 24 hours after sexual intercourse (Dyspareunia), and/or, chronic recurrent pelvic pain experience at any time throughout the menstrual cycle (Acyclical pain) and/or, severe dysmenorrhea in last 3 months determined using the numerical rating scale (0-10) for pain severity 7-10; 2) **Medical records** - ICD codes (ICD10: **N94.1, N94.4, N94.5, N94.6, R10** or ICD9: **625.0, 625.3, 789.0**) | Controls exclude anyone with history of endometriosis (if known) |
| Infertile endometriosis cases vs. controls | 1) **Self-reported**. Definition: non-conception after 12 menstrual cycles with regular unprotected intercourse; 2) **Medical records** - ICD codes (ICD10: **N97.0, N97.1, N97.2, N97.8, N97.9** or ICD9: **628.0-628.9**) | Controls exclude anyone with history of endometriosis (if known) |
| Adenomyosis cases with hysterectomy history and no endometriosis diagnosis vs. controls (with hysterectomy and no adenomyosis and/or endometriosis history) | 1) **Hysterectomy confirmed** adenomyosis but no history of endometriosis; 2) **Medical records** - ICD codes (ICD10: **N80.0 and those with hysterectomy history but no N80.1-9** or ICD9: **617.0 and those with hysterectomy history but no 617.1-9**). | Controls with hysterectomy history and no history of endometriosis and/or adenomyosis |

**IEGC Consortium: Imputation protocol**

This document describes the protocol for imputation of GWAS studies. The protocol will be implemented in four stages:

1. **Initial sample and variant quality control.** For imputation, it is crucial that only high quality common or low-frequency variants in the scaffold are included.

2. **Lift-over of GWAS scaffold to NCBI build 37 and matching to reference panel.** Scaffold annotation must be updated to NCBI build 37 prior to analysis and all variants must be aligned to the forward DNA strand. Variants that do not match with the reference panel in terms of position, alleles, or frequency must be excluded.

3. **Pre-phasing of the GWAS scaffold.** There are considerable computational advantages to prephasing the GWAS scaffold, prior to imputation.

4. **Imputation into the pre-phased GWAS scaffold.** To avoid mismatching of variant IDs and allelic coding it is crucial to harmonise imputation strategies.

The stringent quality control of the GWAS scaffold in steps 1 and 2 are strongly recommended. However, studies that, having already pre-phased their scaffold in previous imputation efforts, do not wish to repeat this process, this is okay. Please describe in the Descriptives/QC spreadsheet. Prephasing and imputation (steps 3 and 4) can be performed locally or via the “imputation server” hosted at the University of Michigan. Please be reminded that the use of the imputation server requires adequate permissions to upload individual-level genotype data to an external server.

**1. Initial sample and variant quality control**

Sample and variant quality control are of the responsibility of individual studies, with their own filters defined and applied as appropriate. Standard quality control protocols should remove samples with low call rate (e.g. < 95%), extreme heterozygosity (e.g. outside mean±(3xSD)), a mismatch of gender with X chromosome variants, duplicates, first or second degree relatives (unless explicitly undertaking a family-based association study or explicitly accounting for relatedness in downstream association analyses), or individuals of outlying ancestry. For

imputation only high quality common or low-frequency variants in the scaffold should be included. After sample quality control, exclude all variants with call rate <95% in cases or controls (unless more stringent filters have already been employed), with Hardy-Weinberg equilibrium exact p<10-6 in cases or controls (for autosomes only, unless more stringent filters have already been employed, founders only in family-based association studies), or MAF<1% (estimated across cases and controls combined, founders only in family-based association

studies).

**2. Liftover of GWAS scaffold to NCBI build 37 and matching to the reference panel**

Before pre-phasing, the GWAS scaffold must be mapped to NCBI build 37 of the human genome. If the scaffold is not mapped to NCBI build 37, we would recommend the use of strand files and scripts developed by the University of Oxford, which can be downloaded from: <https://www.chg.ox.ac.uk/~wrayner/strand/> .Any queries regarding this process can be addressed to the IEGC Central Analysis Team.

Once the scaffold has been lifted over to build 37, you should check if the SNP IDs, alleles, and frequencies match those with the reference panel. Allele frequencies for all SNPs in the scaffold can be obtained using the command:

plink --bfile fname --freq --out fname.freq

where fname is the root of the filename for the GWAS scaffold in binary PLINK format (BED/BIM/FAM). Please note that for family-based association studies, allele frequencies should be calculated for founders only, which is the default for PLINK, provided that familial relationships are defined in the fam file.

To reduce the errors due to strand misalignment, duplicates, or genotyping issues with the scaffold, a stringent quality control should be applied. To match the scaffold against the HRC or 1000G reference panel, removing or updating SNPs that do not agree in terms of position, alleles, or frequency (in the relevant ethnic group), we would recommend the use of the script HRC-1000G-check-bim.pl developed by the University of Oxford, which are freely available for download from: <https://www.chg.ox.ac.uk/~wrayner/tools/> .

Dependent on the reference panel chosen, the script requires the 1000G legend file which can be downloaded from the same website (1000GP\_Phase3\_combined.legend); or the HRC reference (unzipped tab limited; currently v1.1 HRC.r1- 1.GRCh37.wgs.mac5.sites.tab) which can be downloaded from the Haplotype Reference Consortium Website (http://www.haplotype-reference-consortium.org/site).

The script can be run using a command of the form:

For the 1000G:

perl HRC-1000G-check-bim.pl –b fname.bim -f fname.freq –r 1000GP\_Phase3\_combined.legend -g –p ethnicity

where ethnicity by: AFR for African (American); AMR for Hispanic; EAS for East Asian; EUR for European; or SAS for South Asian.

For the HRC:

perl HRC-1000G-check-bim.pl –b fname.bim -f fname.freq –r HRC.r1-1.GRCh37.wgs.mac5.sites.tab –h

The script will create a series of command files to do the following: (i) exclude SNPs that do not match on chromosome, position, and alleles (including multi-allelic variants); (ii) exclude SNPs not present in the reference panel (including PAR X chromosome regions); (iii) update chromosome and position to match the reference panel; (iv) align the scaffold to the forward strand; (v) fix REF alleles to the reference panel; (vi) remove A/T or G/C SNPs with MAF >40% in the reference panel (for the relevant ethnic group); (vii) remove all SNPs with an allele frequency difference >20% between the scaffold and reference panel (for the relevant ethnic group); and (viii) remove duplicates that may be introduced with the chromosome and position update applied in (iii).

To run these command, use the following commands:

chmod 770 Run-plink.sh

./Run-plink.sh

This command will clean the scaffold (i.e. remove problematic variants and correct errors) and generate separate binary PLINK format files for each chromosome: fname-updated-chr1, fname-updated-chr2, etc. Any queries regarding this process can be addressed to the IEGC Central Analysis Team.

**For non-European cohorts, please follow the 1000G imputation protocol provided below in points 3 and 4 if pre-phasing and imputation will be performed locally, or in point 5 if the Michigan “imputation server” will be used instead. For the European populations, please use the HRC reference imputation detailed in point 5.**

**3. Local pre-phasing GWAS scaffold with SHAPEIT**

We recommend the use of SHAPEIT for pre-phasing the GWAS scaffold prior to imputation. SHAPEIT is computationally efficient, and its output can be used with a variety of downstream imputation tools. Full details of the SHAPEIT software can be found at:

https://mathgen.stats.ox.ac.uk/genetics\_software/shapeit/shapeit.html

Please ensure that you use the most up to date version of the software: SHAPEIT v2 (r837). The genetic map files required for pre-phasing can be downloaded from the IMPUTE2 website at:

https://mathgen.stats.ox.ac.uk/impute/1000GP\_Phase3.html

To pre-phase chromosome 1, for example, a command line of the following form can be used:

shapeit --input-bed \

fname-updated-chr1.bed \

fname-updated-chr1.bim \

fname-updated-chr1.fam \

--input-map genetic\_map\_chr1\_combined\_b37.txt \

--output-max \

fname-updated-chr1.haps.gz fname-updated-chr1.sample \

--output-log shapeit.chr1.log

A similar command line is used for all autosomal chromosomes. For the X chromosome, the binary PLINK FAM file must contain the sex of each individual in the fifth column. The basic autosomal command line above should then be modified with the additional option --chrX. Any queries regarding this process can be addressed to the EC Central Analysis Team.

**4. Local imputation up to 1000G reference panel**

We recommend the use of IMPUTE2 or minimac3 for imputation after pre-phasing of the GWAS scaffold. These tools have similar performance in terms of imputation quality and speed and can both can be used with data formats generated by SHAPEIT. Imputation should be performed for all samples simultaneously (i.e. do not divide by endometriosis case-control status).

*(a) Imputation with IMPUTE2*

Full details of the IMPUTE2 software can be found at:

https://mathgen.stats.ox.ac.uk/impute/impute\_v2.html#home

Please ensure that you use the most up to date version of the software: IMPUTEv2.3.2. The required 1000 Genomes Phase 3 reference panel in the appropriate format can be downloaded from the IMPUTE2 website at:

https://mathgen.stats.ox.ac.uk/impute/1000GP\_Phase3.html

We recommend imputing small chunks of each chromosome at a time to minimise computational burden and memory requirements. Experience suggests that chunks of 1Mb-5Mb are reasonable, depending on the number of individuals in your sample. For example, to impute the first Mb of chromosome 1 (i.e. from 1 to 1,000,000 base pairs), a command line of the following form can be used:

impute2 -use\_prephased\_g \

-m genetic\_map\_chr1\_combined\_b37.txt \

-h 1000GP\_Phase3\_chr1.hap.gz \

-l 1000GP\_Phase3\_chr1.legend \

-phase -known\_haps\_g fname-updated-chr1.haps.gz \

-o\_gz -outdp 5 -Ne 20000 -int 1 1000000 -buffer 500 \

-o fname-updated-chr1.001.impute

A similar command line is used for the remaining chunks of chromosome 1, and for all other autosomal chromosomes. For the X chromosome, (1) the options -chrX and -sample\_g must be used to provide a sample file with sex information (generated by SHAPEIT), (2) a file with ids of all male controls must be provided with –exclude\_samples option to exclude all the males from the control data for X chromosome analysis. For example, to impute the first Mb of chromosome X, a command line of the following form should be used:

impute2 -use\_prephased\_g \

-m genetic\_map\_chrX\_nonPAR\_combined\_b37.txt \

-h 1000GP\_Phase3\_chrX\_NONPAR.hap.gz \

-l 1000GP\_Phase3\_chrX\_NONPAR.legend \

-phase -known\_haps\_g fname-updated-chr23.haps.gz \

-exclude\_samples\_g males.txt \

-chrX –sample\_g fname-updated-chr23.sample \

-o\_gz -outdp 5 -Ne 20000 -int 1 1000000 -buffer 500 \

-o fname-updated-chr23.001.impute

On completion of the imputation for all chunks on the same chromosome, files should be concatenated and converted into vcf format using PLINKv1.9.

plink --oxford fname.impute fname.sample --recode vcf --out fname

*(b) Imputation with minimac*

We recommend the use of minimac4 for 1000G imputation because it is more computationally efficient than previous versions and requires less memory. Full details of the minimac4 software can be found at:

<https://github.com/statgen/Minimac4>

Please ensure that you use the most up to date version of the software: minimac4 v4.1.6. The 1000 Genomes Phase 3 reference panel in an appropriate format is required. We recommend the use of the VCF formatted reference files (for autosomes and chromosome X), which can be downloaded from the minimac3 website at:

http://genome.sph.umich.edu/wiki/Minimac3#Reference\_Panels\_for\_Download

Then you can convert this to minimac4 acceptable format:

minimac4 --update-m3vcf reference.m3vcf.gz > reference.msav

The minimac4 software requires pre-phased genotype files in VCF format, which can be generated with SHAPEIT using commands of the form:

shapeit -convert \

--input-haps fname-updated-chr1 \

--output-vcf fname-updated-chr1.vcf

This command should be repeated for each chromosome. The resulting files should then by

bgzipped and tabixed, for example:

bgzip fname-updated-chr1.vcf

tabix fname-updated-chr1.vcf.gz

The individual chromosome files can then be concatenated into a single VCF and tabixed with

the commands:

bcftools concat -Oz \

fname-updated-chr1.vcf.gz \

fname-updated-chr2.vcf.gz \

… \

fname-updated-chr22.vcf.gz > fname-updated.vcf.gz

tabix fname-updated.vcf.gz

We recommend imputing small chunks of each chromosome at a time to minimise computational burden and memory requirements. Experience suggests that chunks of 1Mb-5Mb are reasonable, depending on the number of individuals in your sample. **For the X chromosome, please ensure that male controls are excluded prior to analysis.**

**Note: If any centre prefers to run EAGLE for pre-phasing and BEAGLE for imputation that is also possible.**

**5. Pre-phasing and imputation with the University of Michigan imputation server**

The imputation server is available at:

https://imputationserver.sph.umich.edu/start.html

A user account with the server must be created, with detailed instructions provided on the server webpage upon log-in at the help page. Brief instructions can also be found bellow.

Genotype files must be provided in VCF format for each chromosome. Some available tools to create these files include PLINK v1.9, VCFtools or vcfCooker (see “Help” option on the server webpage for further details). Example commands for PLINK v1.9 (https://www.cog-genomics.org/plink/) are as follows:

plink --bfile fname-updated-chr1 \

--recode vcf --out fname-updated-chr1

The VCF files need then to be sort and compress:

vcf-sort fname-updated-chr1.vcf | \

bgzip -c > fname-updated-chr1.vcf.gz

Validity of the VCF files can be checked using the python script checkVCF which compares the reference alleles in the VCF file with the reference genome. The script can be downloaded from https://github.com/zhanxw/checkVCF and run with a command of the form:

checkVCF.py –r human\_g1k\_v37.fasta \

–o fname-updated-chr1-check fname-updated-chr1.vcf.gz

Problems in the alleles will be reported in the log files. An empty fname-updated-chr1-check.ref file indicates that the reference alleles in the VCF file are in accordance with the reference genome and the VCF file can therefore be uploaded to the server for imputation. If fnameupdated-chr1-check.ref is not empty, additional steps are required to force/flip the reference alleles.

Example of .ref file:

MismatchRefBase 1:752566:G-A/G

MismatchRefBase 1:1018704:A-G/A

This indicates that, for the first variant, the reference allele is “G” but in the VCF file the reference allele is “A” and the alternate allele is “G”. A file force-ref.txt with the variants for which the references alleles need to be flipped should be created and then used in PLINK to force the reference allele. For example, given that the rsid of 1:752566 is rs3094315, and 1:1018704 is rs9442372, the force-ref.txt for the above example should be:

rs3094315 G

rs9442372 A

These alleles can then be flipped and a new VCF file generated by using:

plink --bfile fname-updated-chr1 \

--recode vcf –-a2-allele force-ref.txt \

--out fname-updated-chr1

The VCF files needs then to be sorted and compressed as before:

vcf-sort fname-updated-chr1.vcf | \

bgzip -c > fname-updated-chr1.vcf.gz

The script checkVCF.py should then be run again on the new VCFs, checking that fname-updated-

chr1-check.ref is empty. These steps need to be repeated for each of the other chromosomes. The resultant VCF files can then be uploaded to the server for imputation.

On the imputation server “Run” tab, choose the following options:

- Reference panel: 1000G Phase 3 v5 or HRC r1.1 2016

- Phasing: SHAPEIT

- Population: EUR

Once the imputation is finished, new \*.vcf formatted files, containing both genotyped and imputed variants, can be downloaded from the server.

**Uploading imputation QC data**

Once the imputation analyses are finished, please upload the following imputation quality check statistics, tables and plots to your respective dropbox folder:

**- From HRC imputation servers:** The quality controls statistics, quality control/imputation reports and reference allele frequency plot between uploaded samples vs. reference panel.

**- From in-house imputation from IMPUTE2:** \*.info files per chromosome that include the alleles for each variant, allele frequency and imputation quality.

**- From in-house imputation from minimac4:** The quality controls statistics, quality control/imputation reports and reference allele frequency plot between uploaded samples vs. reference panel

Files should be uploaded into the dropbox folder provided for each group. Please see the email sent to you for the link to access the dropbox folder. Any questions please contact us at nilufer.rahmioglu@wrh.ox.ac.uk

**IEGC Consortium**

**Association analysis**

We recommend the use of SNPTEST v2 or EPACTS or similar (Please specify in the Descriptives/QC spreadsheet) for association analysis.

**(a) Association analysis with SNPTEST:**

Association analyses can be performed on imputed \*.gen formatted files, from IMPUTE2 local imputation or \*.vcf formatted files from the Michigan imputation servers or minimac3 local imputation. However, since \*.vcf format is more complex the SNPTEST command line should include -genotype\_field option to specify, which field SNPTEST should read the genotype data. Moreover, within each \*.vcf file, format definition must be given for all fields, e.g. ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

Outprefix for the output files should be created as:

Endo\_cohort\_initials\_date

where:

Cohort: cohort acronym

Initials: initials of the analyst name

Date: analysis date in format DDMMYYYY

If the data is split by chromosome and/or into smaller chunks, then also please add chromosome and chunk numbers into the file name (Endo\_cohort\_chr\_chunk\_initials\_date), otherwise each new run will overwrite the previous filenames.

Example command line in SNPTEST with .gen files:

-data [input.impute.gz] [input.sam] \

-o [outprefix] \

-frequentist 1 \

-method score \

-pheno Endo \

-cov\_names [COV1], [COV2]

Example command line in SNPTEST with .vcf files:

-data [input.impute.gz] [input.sam] \

-genotype\_field GP \

-o [outprefix] \

-frequentist 1 \

-method score \

-pheno Endo \

-cov\_names [COV1], [COV2]

SNPTEST counts the number of lines in the file before it starts analysis. VCF files can be very large and this can take a long time.  Therefore, you can include in the metadata a line of the form ##number-of-variants=n and SNPTEST will skip the file counting and use the provided number instead.

**(b) Association analysis with EPACTS:**

Association analyses should be performed on imputed \*.vcf formatted files, which we got from the Michigan imputation server or from the local imputation. The files must be gzipped using bgzip tool and then indexed using tabix tool (both programs are part of SAMTOOLS software <http://www.htslib.org>).

bgzip fname.vcf

tabix -p vcf fname.vcf.gz

We test for association of each variant passing initial quality control under an additive model using FIRTH test in EPACTS (http://genome.sph.umich.edu/wiki/EPACTS). Adjustment for study specific covariates (e.g. principal components to reduce the population stratification) can be done.

Endo ~ SNP (+PCs)

Outprefix for the output files should be created as:

Endo\_cohort\_initials\_date

where:

Cohort: cohort acronym

Initials: initials of the analyst name

Date: analysis date in format DDMMYYYY

If the data is split by chromosome and/or into smaller chunks, then also please add chromosome and chunk numbers into the file name (Endo\_cohort\_chr\_chunk\_initials\_date), otherwise each new run will overwrite the previous filenames.

Example command line in EPACTS:

${EPACTS\_DIR}/epacts single \

--vcf [input.vcf.gz] --ped [input.ped] \

--sepchr (if VCF is separated by chromosome) --pheno GDM \

--cov [COV1] --cov [COV2] --test b.firth \

--out [outprefix] --run [# of parallel jobs]

**Sub-phenotype association analysis**

If your dataset includes the sub-phenotypes on the cases, you can utilise the -exclude\_samples option along with a respective .txt file including case IDs you would need to exclude from your overall endometriosis case group.

Outprefix for the output files should be created as:

omaEndo\_cohort\_initials\_date

deepEndo\_cohort\_initials\_date

superfEndo\_cohort\_initials\_date

stI\_IIEndo\_cohort\_initials\_date

stIII\_IVEndo\_cohort\_initials\_date

infertEndo\_cohort\_initials\_date

anypainEndo\_cohort\_initials\_date

dyspaEndo\_cohort\_initials\_date

acyclpEndo\_cohort\_initials\_date

gipainEndo\_cohort\_initials\_date

sevdysmEndo\_cohort\_initials\_date

adeno\_cohort\_initials\_date

Example SNPTEST command with .gen files:

-data [input.impute.gz] [input.sam] \

-o [outprefix] \

-frequentist 1 \

-method score \

-pheno Endo \

-exclude\_samples samples.txt \

-cov\_names [COV1], [COV2]

samples.txt should include only list of individual IDs to be excluded from the analysis with no header.

Any questions please contact us at nilufer.rahmioglu@wrh.ox.ac.uk

**Uploading association results and descriptive QC data**

Once the association analyses are finished, please upload the raw results files **without filtering them for MAF or imputation quality** as listed below (**Dropbox details sent in a separate email**):

(1) All association output files:

For SNPTEST analysis:

Endo\_cohort\_initials\_date.snptest.gz

For EPACTS analysis:

Endo\_cohort\_initials\_date.single.b.firth.epacts.gz

Endo \_cohort\_initials\_date.single.b.firth.epacts.top5000

Endo \_cohort\_initials\_date.b.firth.epacts.qq.pdf

Endo \_cohort\_initials\_date.b.firth.epacts.mh.pdf

(2) The completed Excel spreadsheet (file circulated with this analysis plan) with descriptive and QC information and covariates adjusted for in the association analysis for your cohort

Endo \_cohort\_initials\_date\_InfoTables.xlsx

Files should be uploaded into the dropbox folders provided for each group. Please see the email sent to you for your dropbox link. Any questions please contact us at nilufer.rahmioglu@wrh.ox.ac.uk