**CARDIoGRAM*Plus*C4D X-chromosome Analysis Plan 2023**

**Multi-ancestry meta-analysis & Information for New Cohorts**

**Note: recent paper summarizing the X-chromosome :** [**https://www.sciencedirect.com/journal/the-american-journal-of-human-genetics/vol/110/issue/6**](https://www.sciencedirect.com/journal/the-american-journal-of-human-genetics/vol/110/issue/6)

**Rationale**

The impressive progress in understanding the genetic architecture of CAD that has been achieved in recent months by collaborative efforts from CARDIoGRAMplusC4D and the Million Veterans Project, enhanced by individual studies as well as international biobanks (e.g. Japan Biobank, UK Biobank), has resulted in unprecedented data availability in the field. Genome-wide summary data (i.e. association statistics) from these efforts is currently generated for all autosomal chromosomes. Combining data from these resources in a transethnic meta-analysis (and related analyses), and expanding upon the contributing cohorts, in advance of public data release is a timely priority that would further enhance our understanding of the transethnic genetic architecture of CAD.

**Primary aims:**

- To undertake the X-chromosome multi-ancestry meta-analysis based on data from the recent CARDIoGRAMplusC4D GWAS (2021) with comparable data from MVP (2021), and including Japan Biobank (2020) and additional cohorts with available genotyping, and comparable European-ancestry only meta-analysis. Ideally, we hope to conduct the X-chromosome analysis for all cohorts that are currently part of the Card+C4D extension.

- To undertake ancestry specific meta-analysis of the X-chromosome based on individual cohort data from the recent (2021) CARDIoGRAMplusC4D GWAS as well as MVP data (2021), Japan Biobank and other biobank data or cohorts in five major ancestral groups currently part of the Card+C4D consortium: Whites, Africans, Hispanics, East Asians, and South Asians.

**Case definition**

**Note: We anticipate using the CAD definition that was used for the autosomal chromosome analysis (filled out the specific definition of CAD that was used, this will help standardize definition of CAD for all cohorts)**

Coronary artery phenotypes (preference is for the following or minor variation to it – please do contact us to discuss any concerns or variations). We wish to avoid overly soft phenotypes as previous work has established the limited value of such cases. Consequently, for de novo studies undertaking new analyses, cases should adhere to the following as closely as possible. However, individual studies should use insights into their data resource to adapt appropriately (and provide details of the phenotype provided).

* Confirmed MI (see below)
* > 50% stenosis in at least one coronary vessel at angiography with validation from hospital records
* Validated history of percutaneous transluminal coronary angioplasty (PTCA) or coronary artery bypass graft surgery (CABG)
* Validated angina, defined as symptoms + confirmation from at least one non-invasive provocation test e.g. scintigraphy or exercise treadmill test (**angina cases based only on self-report angina alone should not be included**).
* Death confirmed to be due to CAD or highly likely to be due to CAD
* Hospitalization or death recorded under ICD10 I21-I25

**Control definition**

- All participants not defined as coronary cases using the most liberal ‘Soft’ definition to be “non-cases” in the analysis.

Note: check if there are cohort that use population-based controls (these cohort often do not check for MI or CAD cases.) (make sure that we don’t have overlapping)

**Association Analysis**

* analysis required from each cohort

Please test additive models using logistic regression, accounting for genotype imputation uncertainty (i.e. SNP probability or dosage). For studies that have already conducted analyses, please discuss the models used (for example, some studies may have already adjusted for age and this is not considered a substantial deviation from the analysis plan). Use study appropriate software to account for (or exclude as appropriate) relatedness.

If undertaking new analyses to contribute then the following approach is advised. If providing previously generated results, please ensure the approach is documented and discuss any major deviations:

**Primary model: the primary regression model should be performed stratified by sex and by ancestry for multi-ancestries cohorts.**

1. Female CAD (female CAD cases vs. female non-CAD controls) = Age (?) + SNP + PCs (at least 10?) + other study specific covariates (leave the age adjustment up to the cohort)
2. Male CAD (male CAD cases vs. male non-CAD controls) = Age (?) + SNP + study specific covariates (e.g. PCs) (leave the age adjustment up to the cohort)

The regression model is simply additive considering activation of the X chromosome as default model where SNPs are coded 0/1 in male and 0/1/2 in female. Dosage will be used to account for genotype imputation uncertainty. The X-chromosome analysis can be performed in PLINK, REGENIE, or SAIGE.

Giving the complexity of the X-chromosome structure, additional analysis will be required for extensive quality controls.

Assuming inactivation of the X-chromosome (male code as 0/2)

1. Male CAD (male CAD cases vs. male non-CAD controls) = Age (?) + SNP + study specific covariates (e.g. PCs) (leave the age adjustment up to the cohort)

**Analysis required for X-chromosome QC**

* + - 1. Rate of heterozygosity in male controls only
      2. Rate of heterozygosity in females controls only
         1. Flag SNPs with high rate of heterozygosity and investigate whether these SNPs are located in the PAR chromosomal region (PAR1, PAR2, PAR3)
      3. Provide information on how they did QC on the individuals levels
      4. Check for information about phenotypic and genotypic se

How imputation was done and QC of imputation Minor allele frequencies in controls separately in males and females

Test for HWE in female control only

1. Association testing should be done assuming activation and inactivation of the X-chromosome. This mean 2 coding :
   1. Model 1: SNP coded as 0,1,2 (assume activation of the x-chr), analysis stratify by sex
      1. Analysis in female as 0,1,2
      2. Analysis in male as 0,1
   2. Model 2 SNP code as 0,2 (assume inactivation of the X chr)
      1. Analysis in female as 0,1,2 (model above)
      2. Analysis in male as 0,2 (standard X-chr analysis model)
2. Variable to request for each model (total of 3 models should be run; 2 model for male and 1 model for female). 3 separate summary statistic should be provided. 2 for male and 1 for female. List of variables to request from the summary statistic for each cohort.

|  |  |  |  |
| --- | --- | --- | --- |
| Column header | Description | Format | Examples |
| SNP | SNP label for the variant | Illumina identifier from the annotation file | rs693 chr2:7819 chr:pos:A1:A2 |
| CHR | Chromosome on which SNP resides | Numeric for chromosomes 1-22; [current upload of autosomes only] | 1 |
| POS | Position of SNP on chromosome | Basepairs on human genome build used | 34000345 |
| EFFECT\_ALLELE | Allele at this site to which the effect has been estimated  (please check software documentation before labelling a1,a2 as effect allele) | Capital letter (A,C,G,T) | A |
| NON\_EFFECT\_ALLELE | Other allele at this site (please check software documentation before labelling a1,a2 as non-effect allele) | Capital letter (A,C,G,T) | G |
| N\_TOTAL | Total number of cases and controls analyzed | Numeric, integer | 1243 |
| N\_CASES | Total number of cases analyzed | Numeric, integer | 1243 |
| N\_CONTROLS | Total number of controls analyzed | Numeric, integer | 1243 |
| N0\_CASES | Number of homozygous cases with zero copies of the EFFECT\_ALLELE | Numeric, integer or float with 3 digits to the right of the decimal (imputed) | 623  745.234 |
| N1\_CASES (the number of heterozygote for males should be 0 or close to 0) | Number of heterozygous cases with one copy of the EFFECT\_ALLELE | Numeric, integer or float with 3 digits to the right of the decimal (imputed) | 623 for female and 0 or very low for male |
| N2\_CASES | Number of homozygous cases with two copies of the EFFECT\_ALLELE | Numeric, integer or float with 3 digits to the right of the decimal (imputed) | 623  745.234 |
| N0\_CONTROLS | Number of homozygous controls with zero copies of the EFFECT\_ALLELE | Numeric, integer or float with 3 digits to the right of the decimal (imputed) | 623  745.234 |
| N1\_CONTROLS (the number of heterozygote for males should be 0 or close to 0) | Number of heterozygous controls with one copy of the EFFECT\_ALLELE | Numeric, integer or float with 3 digits to the right of the decimal (imputed) | 623 for female and 0 very low for male |
| N2\_CONTROLS | Number of homozygous controls with two copies of the EFFECT\_ALLELE | Numeric, integer or float with 3 digits to the right of the decimal (imputed) | 623  745.234 |
| N\_missingness cases | Number of missingness in cases | Numeric, integer or float with 1 or more digits to the right of the decimal (imputed) | 22 |
| N\_missingness Controls | Number of missingness in controls | Numeric, integer or float with 1 or more digits to the right of the decimal (imputed) | 22 |
| EAF\_ALL | Allele frequency of the EFFECT\_ALLELE in all | Frequency with 3 digits to the right of the decimal | 0.354 |
| EAF\_CASES | Allele frequency of the EFFECT\_ALLELE in cases analyzed | Frequency with 3 digits to the right of the decimal | 0.354 |
| EAF\_CONTROLS | Allele frequency of the EFFECT\_ALLELE in controls analyzed | Frequency with 3 digits to the right of the decimal | 0.354 |
| BETA | Estimate of the allelic effect, defined as the natural logarithm of the odds ratio, ln(OR) | Numeric float with 3 digits to the right of the decimal | 0.203 |
| SE | Estimated standard error on the estimate of the allelic effect, uncorrected for genomic control | Numeric float with 4 digits to the right of the decimal | 0.5611 |
| PVAL | Significance of the variant association, uncorrected for genomic control | Scientific E notation with 3 digits to the right of the decimal | 3.244E-10 |
| INFO\_TYPE | Type of information provided in the INFO column | 0 = SNP is genotyped  1 = “r2\_Hat” from MACH2DAT  2 = “proper\_info” from SNPTEST  3 = “INFO” from PLINK | 1 |
| INFO | Measure of information content for the imputed SNP result (range 0-1) (**autosomal only**) | Numeric float with 3 digits to the right of the decimal (set to missing if genotyped) | 0.483  . |
| HWE | P-value of the HWE test (female controls only) | numerical | 3E-06 |

1. Overall cohort information: this file should contain a description of the cohort including some information that are specific to the X-chr

|  |  |
| --- | --- |
| **Header** | **Description** |
| Cohort | Name of the cohort |
| Genotype | Genotype platform (ex: ILLUMINA, Affymetrix) |
| Year of genotyping | Year of genotyping of the cohort if known |
| DISEASE/TRAIT | How the case were ascertained (case/controls definition) |
| SAMPLE SIZE | Sample size and ancestry description for stage 1 of GWAS (summing across multiple Stage 1 populations, if applicable) |
| Ancestry specific sample size | Sample size for each ancestry if available |
| SNP\_X | Number of SNPs on X chromosome on the genotype array |
| QC | QC platform (or pipeline used for genotyped QC) |
| GENOTYPE | Genotype platform (ex: ILLUMINA, Affymetrix) |
| M\_SAMPLE\_SIZE | Number of male samples (NA: not mentioned) |
| Ancestry specific M\_sample | Number of male samples (NA: not mentioned) for each ancestry |
| F\_SAMPLE\_SIZE | Number of female samples (NA: not mentioned) |
| Ancestry specific F\_sample | Number of female samples (NA: not mentioned) for each ancestry |
| Imputation | Algorithm and parameter used for imputation of the x-chr |

**Meta-analysis**

Once all the data are QCed and we have clean summary statistics from each cohort, we will proceed for meta-analysis in multiple steps:

* 1. Ancestry specific meta-analysis
     1. Sex-specific meta-analysis: Meta-analysis will be first conducted within each sex separately using GWAMA
     2. Sex combined meta-analysis. Here we will perform a meta-analysis of both sexes while testing for:
        1. difference in effect by sex (proxy for activation or inactivation of the X-chr)
        2. Sex interaction (this test interaction between sex and SNPs effect in a meta-analysis)
  2. Multi-ancestry meta-analysis; Here the analysis will consist of using multi-ancestry approaches for meta-analysis of all ancestry using both GWAMA and MR-Mega

QC after meta analysis

QC to be done once we have all the data

* + - 1. Difference in minor allelic frequencies between male and females: SNPs with significant difference in MAF between sex will be excluded from the meta-analysis
      2. Test for HWE in female control only (exclude SNPs with p(HWE) <1e-06
      3. Flag SNPs that are known to be problematic and compare their allelic frequencies across cohorts of the same ancestry to make sure they are consistent (we have a list of SNPS that are known to be problematic)
      4. Create a reference sample set of the X chromosome SNPs for each ancestry group based on whole genome sequencing (Here we can use Gnomad, TopMed, or UKBB WGS) and use this as baseline for QC of the sex chromosome.
      5. Rate of heterozygotes in male controls
  1. If there is a high rate of heterozygosity in Par1 region for some cohort, we have two options:
     1. Ask for QC of genotyped data and re-imputation before conducting the analysis
     2. Remove all the SNPs with high rate of heterozygosity from the analysis for the cohort with the issues