Skin Genetics Consortium
Analysis Plan: Cross-disease GWAS
Version 5

Please read this document fully and then if you have any questions or need assistance with your pipeline, you can contact us at info@skingeneticsconsortium.org

Aims

Cross-ancestry and population specific summary statistics per disease per cohort

- In each biobank, perform separate case-control genome-wide association studies (GWAS) for a wide range of common skin diseases and skin disease features
- GWAS should include participants with diverse ancestries
- The GWAS summary statistics that you provide will be used to perform central disease/trait specific meta-analyses and to perform comparative analyses to identify shared and distinct susceptibility loci

PHASE 0

Phenotype list and definitions

The skin diseases and features that will be analysed are listed in the phenotype mapping document, along with the relevant ICD-10, ICD-9, READ and SNOMED code(s). If your cohort uses alternative diagnostic codes from other coding schemes or there are other possibilities to capture additional cases, please let the central team know.

Cases should comprise all participants with at least one relevant coding, in any of the coding systems. If you have missing or limited data for any/all coding systems, please still extract cases based on the data you do have access to.

For ICD-10 and ICD-9, codes will either be labelled as "EXPAND" or "EXACT" in column C. For those with "EXPAND", cases should be classified as an individual with the listed code OR any of its subcodes i.e. for A60, extract A60, A60.0, A60.1 and A60.9. This can be done in R using the command "%like%" from the package "data.table". For example: df[df$codes %like% "A60",]

In contrast where codes have "EXACT" in column C, only individuals with the exact code given, should be considered as cases.

ICD-O codes are used within cancer registries to classify tumours by site, histology and behaviour. The site is defined using ICD10/9 codes, which are listed as Code1 in our phenotype mapping document (all EXPAND, see above). The ICD-O codes (histology and behaviour) are listed in the Code2 column. When using ICD-O, cases should be defined as participants that have entries in the cancer registry that match both Code1 (site) **and** Code2 (histology/behaviour). Note - in some cohorts ICD-O codes may be stored in two separate fields (Histology & Behaviour) separated at the "/" – so please either join the fields with a "/", or ensure you are matching using both fields.

For all other coding systems, please only extract exact matches to the code listed.

Please calculate case numbers for your cohort and share with the central team. We appreciate that each cohort may not have all codes due to differences between coding systems. Therefore, if possible, we would also like to see a breakdown of case numbers by each code used to define phenotypes in your cohort. This would be in an additional column with a semi-colon-separated list with each code and number of cases it identifies. For example:

L40:15;L400:150;L401:13;L402:16;L403:18;L404:80;L407:3;L408:12;L409:90

An example R script is attached which calculates case numbers and the breakdown of codes captured- please note that it would need to be repeated for males and females separately. It also assumes ICD codes are provided like they are in UK Biobank, who have applied some data-cleaning before release (including the removal of dots [.] in the codes), so the code may need editing based on cohort-specific requirements.

Diverse populations

We are aiming to generate cross-ancestry and population specific summary statistics per disease per cohort

As a starting point we would like to evaluate which of the major ancestry population groups each cohort could run separate analysis for. These ancestry population groups (as defined by gnomAD) include: African/African American (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), Middle Eastern (MID), South Asian (SAS). If you think different stratifications should be used for your cohort, please contact us.

Please provide PCA plots for your cohort based on a set of LD-independent common variants and the number of individuals that you can assign to each of these groups (to be included with Phase 1 uploads).

Requests:

To get an idea of which phenotypes will be feasible to GWAS, we ask that you return to us the following information:

- □ Table(s) containing the number of cases and controls for each phenotype and the (optional) breakdown of code numbers for each phenotype (stratified by sex- one file for males, one file for females). If you have no cases for a listed ICD-10 code, please still include this in the table so that we have this information. The table should be saved as a tab-delimited text file (Columns: SGC phenotype label, number of cases, number of controls, code numbers breakdown [optional]). Please ensure to use the exact SGC phenotype label provided in our list to ensure that we can aggregate results correctly.

- □ For every pair of phenotypes, a table showing the number of case individuals with both phenotype codes (stratified by sex, one file for males, one file for females). Diagnoses can be at any point in time, not necessarily at the same time. For example, how many individuals have both codes L40 and L20 in their linked health data. Provide this data for all pairs of phenotypes, to give a co-occurrence table. Please provide this table as a tab-delimited text file.

- □ A text file containing a brief description of study ascertainment and the source of data used to establish skin disease and trait phenotypes (we will need more extensive methodological details later, in Phase 1). We also request that you inform us if you have the required data to run age of onset GWAS for our phenotypes.

PHASE 1

Genetic data generation and QC

GWAS analyses should be performed using genome-wide imputed data based on a suitable large reference panel (or using whole genome sequence data). QC checks and filtering appropriate to your dataset should be performed to ensure high quality data. Our minimum recommended QC steps if you have non-imputed data are:

- □ Filter for variants with call rates ≥ 98%
- □ Filter for samples with call-rates ≥ 95%
- □ Check concordance of MAF (per ancestry) with reference population MAF (exclude markers with high discordance).

Upon upload you will be prompted to complete a form with details of your data generation, imputation and QC steps. Please flag to us via email if your pre-imputation QC steps differ significantly from our recommended specifications above. We also request that, if possible, variants are positioned using hg38 coordinates.

We recommend that multiple individuals with genetic relationships are all retained in GWAS analyses, provided that an appropriate association testing method is employed to account for this.

For each GWAS analysis please include at least all variants with a minor allele frequency of more than 0.5%. Rarer variants can also be included if available. For imputed datasets, please include variants with an imputation quality score (INFO/R^2) down to 0.3.

Association testing should be performed for variants on chromosomes 1-22 and X.

Where possible, it would also be valuable to perform association testing on imputed classical HLA alleles. Please contact us if you have processes in place for association testing with two- and four-digit resolution or would be able to work with us to do this.

Association testing

Binary (case-control) association tests should be performed separately for each disease/trait. We strongly recommend using a software that implements a mixed model (or similar) which will account for imbalanced data, population stratification and relatedness, such as REGENIE, fastGWA-GLMM or SAIGE. We can provide support for this.

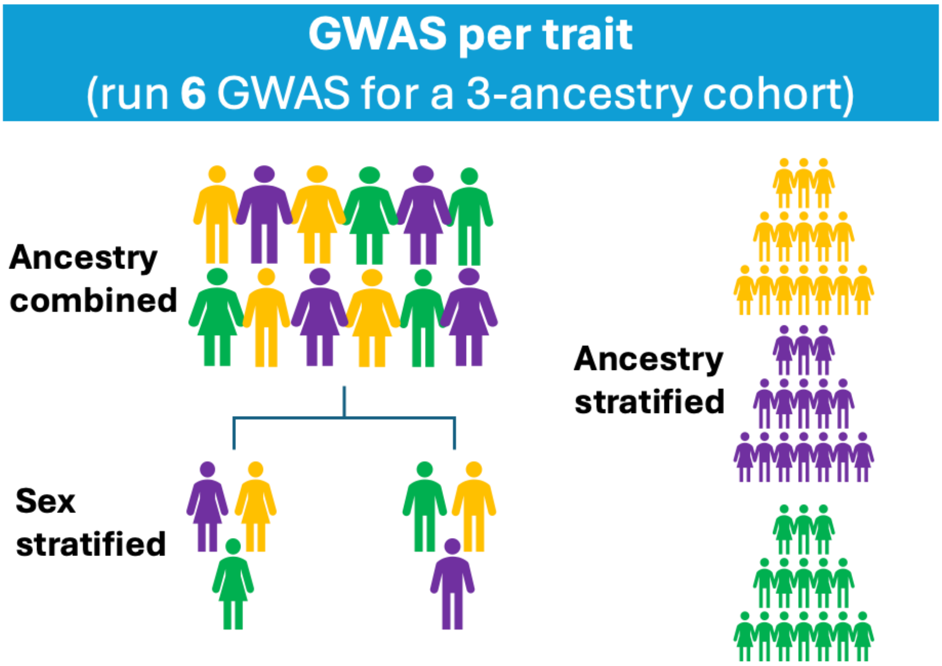
As described above (in the phase 0 section), cases should include any participant with at least one occurrence of a relevant diagnostic code. There should be no exclusion criteria applied to controls, meaning that for each GWAS the control group should include all participants with no occurrences of the relevant codes.

Per disease/trait, GWAS should be performed for:

- All participants (ancestry-combined, sex-combined)
- Male participants (ancestry-combined)
- Female participants (ancestry-combined)
- Ancestry stratified (if applicable, sex-combined).

We recommend only running GWAS where there is a minimum count of 30 cases.

The figure below shows the number of GWAS we expect to be run per disease/trait, in a cohort with three ancestries.


GWAS per trait
(run **6** GWAS for a 3-ancestry cohort)

Principal component analysis should be performed and an appropriate number of PCs (number determined by each cohort), included as covariates alongside sex. Where a cohort plans to retain related participants, PCs should be calculated on a maximal unrelated subset of participants and projected onto the whole cohort.

For cohorts with multiple ancestries, PCs should be calculated based on the whole cohort for the ancestry-combined analyses. For the ancestry stratified analyses, PCs should then be calculated within ancestry groups.

GWAS data output

For each GWAS, please return genome-wide summary statistics including at least the following columns:

Columns	Required	Format
Chromosome	Yes	1-24
Position	Yes	Integer, hg38 coordinates if possible
Variant identifier	No	rsid number
Effect allele	Yes	ACTG or -
Non-effect allele	Yes	ACTG or -
Beta (log odds ratio)	Yes	Numeric
Standard error	Yes	Numeric
P-value	Yes	Scientific notation to 1-4 decimal places.
Effect allele frequency	Yes	Numeric
Imputation quality score	Yes	Numeric

Please format your GWAS outputs to be compressed tab-delimited files and name them programmatically to aid QC and the detection of errors. For example, “UKBiobank_PSOR_EUR_male.txt.gz” to represent a GWAS of psoriasis in European males from UK Biobank cohort. The upload process will require that the correct phenotype labels (as provided with the full phenotype list) are used in the meta data (below), so we recommend using these labels within filenames to reduce the possibility of data being mismatched.

In addition to the GWAS results files, the upload process will require consistently formatted metadata files to be uploaded (one per GWAS summary stats file). The table below shows the format, and an example .json file will be shared. This file should be named to correspond with the summary statistics file, e.g. “UKBiobank_PSOR_EUR_male.json”

Field : Example	**Format**	**Info**
"cohort": *UK_Biobank*	Text string	Needs to match cohort name on data portal
"dataSetName": *UKBiobank_PSOR_EUR_male*	Text string	GWAS file name: no hard formatting requirements but please name sensibly to aid QC checks, given the large number of GWAS being uploaded
"phenotype": *PSOR*	Text string from our list	Must match our provided list of phenotype names
"codesUsed": *L40, L40.0, L40.1, L40.2, L40.3, L40.4, L40.7, L40.8, L40.9, 696, 696.1, 696.8*	Series of codes (text strings)	List the diagnosis codes used to extract cases from linked health data. Include only the codes for which at least one case was identified.
"sex": *Male*	3 options	All, Male, or Female,
"ancestry": *EUR*	Text string	'Combined' or specified ancestral group (e.g. gnomAD groupings: AFR, AMR, EAS, EUR, MID, SAS, or suitable name for alternative ancestral group).
"cases": *5000*	Integer	
"controls": *100000*	Integer	
"maleProportionCases": *0.40*	Numeric	Should be 1 for male only and 0 for female only
"maleProportionControls": *0.47*	Numeric	Should be 1 for male only and 0 for female only
"assocTestSoftwareAndVersion": *regenie v4.1*	Text string	Software and version used for association testing

How to upload data

The Skin Knowledge Portal team have developed a private data repository to which you can upload all your files. Information on how to access this will be communicated in due course.

Summary checklist of items to be returned

Phase 0

- □ Case and control numbers for each phenotype
- □ Co-occurrence matrix
- □ A brief description of study ascertainment and the source of data used to establish phenotypes.
- □ Total participant numbers for each ancestry group
- □ A description of your established cohort pipeline for performing multi-ancestry and admixed analyses

Phase 1

- □ Completed cohort information form on website
- □ GWAS summary statistics for each trait/ancestry/sex combination
- □ Correctly formatted metadata .json file for each GWAS run
- □ PCA plots by ancestry group