**Confluence Analyses Plan for Studies Sharing GWAS Summary Statistics**

This document outlines the analysis plan for contributing GWAS summary results to the [Confluence](https://dceg.cancer.gov/research/cancer-types/breast-cancer/confluence-project) breast cancer genome-wide association study (GWAS) by contributing summary GWAS study results to a meta-analysis. Confluence is collaboration among breast cancer consortia to conduct the largest and most diverse breast cancer GWAS to date: it will include over 300,000 participants with breast cancer and 300,000 without, residing across six continents. GWAS summary statistics from the meta-analysis combining all participating consortia and studies will be made available to collaborators upon paper submission and be made available to the broader research community upon publication.

**Study Objectives**

**Primary Goals**

* Conduct a multi-ancestry GWAS to identify genetic susceptibilities for overall breast cancer using the largest datasets available.
* Conduct a multi-ancestry GWAS to identify genetic susceptibilities for estrogen receptor (ER)-specific, and triple-negative (TN) subtypes (optional)
* Utilize the summary statistics for follow-up analyses, including polygenic risk score (PRS) development, heritability estimation, and other downstream analyses.

**Secondary Goals (Optional)**

* Perform ancestry-specific analyses for overall, ER-specific and TN-specific breast cancer risk.

If conducting new analyses is not feasible, existing GWAS summary statistics aligned with these goals are welcome. Ideally, each GWAS (overall, subtype, and/or ancestry-specific) should include at least 500 breast cancer cases and 500 controls. If sample sizes are too small for subtype or ancestry-specific analyses, overall analyses alone are acceptable. We primarily focus on female participants but will also consider male breast cancer GWAS data if available to examine sex differences.

**Analysis Plan**

**1. Overall Breast Cancer**

* Perform case-control GWAS: Cases include all female participants diagnosed with breast cancer (in-situ or invasive) and no history of other invasive cancers. Controls are individuals with no history of invasive cancers.
* Covariates: Include the first 10 PCs for population stratification. (If your existing pipeline uses a different set of covariates for stratification adjustment, that is acceptable.)

**2. Subtype-Specific Analyses (Optional)**

* Compare ER-positive vs. controls, ER-negative vs. controls, and triple-negative (defined as the breast cancer being negative for ER, Progesterone Receptor (PR), and human epidermal growth factor receptor 2 (HER2)) vs. controls.

**3. Ancestry-Specific Analyses (Optional)**

* Conduct separate GWAS by genetically inferred ancestry (methods described in next section.
* Covariates: ancestry-specific PCs 1–10
* Ancestry Definitions
  + Preferred: GrafAnc (see below).
  + Acceptable: Similarity to 1000 Genomes superpopulations (EUR, AFR, AMR, EAS, SAS) or any established in-house definitions

**4. Recommended Software**: Use REGENIE (https://rgcgithub.github.io/regenie/) for binary-trat case-control GWAS. A detailed pipeline regrading implementing REGENIE on breast cancer dataset is provided here: https://github.com/confluence-breast-cancer-consortia/Confluence\_Analysis\_Pipeline/blob/main/README.md. If other tools (e.g., SAIGE, BOLT-LMM, PLINK or any existing GWAS tools) were used previously, results can still be shared. The analysts working with the individual level data shared with the Confluence Project will make their analysis coded available over GitHub. This code will be available to you to help try to harmonize analyses across data resources.

**5. Covariate Adjustment**

* Population Stratification:
  + Multi-ancestry: First 10 PCs from the combined dataset.
  + Ancestry-specific: First 10 PCs within each ancestry group.

**6. Quality Control procedure**

* Variant-level filters
  + Minor allele count (MAC) ≥ 30 (based on combined cases and controls or within each ancestry group for ancestry-specific analyses)
  + Imputation INFO score ≥ 0.2
  + Variant missingness rate ≤ 5%
* Sample-level filters
  + Sample missingness rate ≤ 5%

**Ancestry Determination (Optional)**

Global Ancestry Inference Using GrafAnc: To estimate global ancestry patterns, we propose using the GrafAnc classifier to estimate the global ancestry of each sample. GrafAnc is designed to classify genetic ancestry from human genomic data by calculating genetic distances between individual samples and reference populations. The tool estimates ancestry proportions at both continental and subcontinental levels using a series of genetic distance (GD) scores and other specialized metrics. The full detailed guidance can be found [here](https://github.com/jimmy-penn/grafanc/blob/master/GrafAncDocumentation.md).

Briefly, GrafAnc accepts genotype data in PLINK format (.fam, .bim, .bed) or VCF format (.vcf or .vcf.gz) and works with variants labeled using RS IDs and/or chromosome positions in either GRCh37 or GRCh38. It requires at least 10,000 ancestry SNPs for reliable continental classification and 50,000 SNPs for subcontinental classification.

GrafAnc assigns each sample to a labeled genetic ancestry group, namely, African (AFR), Middle East and North African (MENA), European (EUR), South Asian (SAS), East Asian (EAS), American (AMR), Oceanian (OCN), and Unclassified (UNC).

**5. Deliverables**

1. GWAS Summary Statistics (Prefer TSV/CSV format):
   * Columns: Chromosome, Position, SNP ID (e.g., rsID or another unique variant identifier), Effect Allele, Other Allele, Effect Allele Frequency, Effect Allele Frequency in Controls, Effect Allele Frequency in Cases, Beta (log-odds ratio), Standard Error (SE), p-value, N for cases, N for controls, Imputation Quality (if available), and Hardy-Weinberg Equilibrium (HWE) test results (if available). If you used the REGENIE pipeline as described here: <https://github.com/confluence-breast-cancer-consortia/Confluence_Analysis_Pipeline/blob/main/README.md>, it would be ideal if you could share the direct REGENIE output, as it includes all the required information.
   * For multi-ancestry GWAS summary statistics: Provide ancestry-stratified sample sizes
   * Please specify what genome build is used for the position. The preferred build is HG38.
2. Meta-data: provide information on genotyping platform(s), pre-imputation individual- and variant-level QC exclusions, imputation protocol and post-imputation exclusions, and association analyses (see study description table).

**6. Alternative Options**

**Share Existing GWAS Summary Statistics**

* We are happy to receive any existing breast cancer GWAS files (overall, subtype, ancestry-stratified, etc.).
* Please name the files using the following convention: STUDY\_GROUP\_Type\_out.txt, where STUDY is your study name/acronym and GROUP denotes the study sample (e.g. POOLED for multi-ancestry or EUR, AFR, SA etc. for ancestry-group specific analyses). Type referred to cancer subtypes (e.g. Overall for overall breast cancer, ERpos/ERneg for ER positive and negative breast cancer, TN for triple negative breast cancer)
* Our team can handle additional QC, meta-analysis, and post-processing.

**Minimal Ancestry Definitions**

* If implementing the gnomAD based approach is infeasible, we welcome current operational definitions (e.g., 1000 Genomes based ancestry inference (AFR, AMR, EUR, EAS, SAS) or any in house genetic inferred ancestry approach) and will report these methods accordingly.

**7. Next Steps**

1. Confirm feasibility: Indicate whether new analyses can be conducted or existing results shared before July 15th, 2025
2. The data will be shared through Box or other preferred approaches
3. Authorship: The manuscript plans to use a group authorship. All participating studies will be included as part of the group authorship.