**User Guide**

**FAETH Score Calculation Pipeline (Lite Version)**

**Overview**

The FAETH Score Calculation Pipeline is based on genome-wide restricted maximum likelihood (GREML) analysis using multiple Genomic Relationship Matrices (GRMs). These matrices are derived from subsets of sequence variants, partitioned using functional and/or evolutionary annotations. For more details, refer to the methodology described by Xiang et al. (2019) ([bioRxiv](https://www.biorxiv.org/content/10.1101/601658v2)).

**Input Requirements**

Ensure all required files are placed in their respective directories. Below are the descriptions of the necessary inputs:

**Phenotype and Animal ID Data**

* **Directory**: 2\_Data/

1. **pheno.txt**  
   A text file containing phenotypes for FAETH score calculation.  
   **Format**:
2. ANIMAL\_A | ANIMAL\_B | PHENOTYPE 1 | PHENOTYPE 2 | ...
   * Columns:
     + ANIMAL\_A and ANIMAL\_B: Duplicate animal IDs (required for MTG2 software compatibility).
     + Phenotype columns: Observations (missing values should be coded as NA).
   * Example file included in the pipeline.
3. **pheno.fam**  
   A PLINK .fam file for identifying animals in the GRM files.  
   **Format**:
4. ANIMAL\_ID | ANIMAL\_ID | 0 | 0 | 0 | -9

**Target SNP GRMs and High-Density GRM**

* **Directory**: 3\_GRM/

1. **Target GRMs (G.grm) and Map Files (genotypes.bim)**
   * GRMs must represent subsets of SNPs classified as "relevant" based on functional annotation maps.
   * **HD GRM**:
     + Derived from high-density SNP arrays or randomly selected SNPs (~600K SNPs as in Xiang et al., 2019).
   * **GRM Files**: Standard binary format from [GCTA software](https://cnsgenomics.com/software/gcta/#Overview). File must be named “G.grm.bin”. (Including the GRM file for HD map)
   * **Map File Format (PLINK .bim file)**:
     + CHR | CHR\_POS | 0 | POS(bp) | ALLELE\_A | ALLELE\_B
     + File must be named “genotypes.bim”
   * File Organization:
     + GRMs stored in separate folders within 3\_GRM/.
     + Folder naming:
       - High-Density GRM folder: HD/
       - Other GRM folders: Custom names.
2. **Whole Genome Sequence Map (WGS.bim)**
   * **Directory**: 2\_Data/
   * Contains SNP positions from the whole-genome sequence dataset.
   * Must match SNP positions in genotype.bim files from the GRM folders.
   * **Format**: Same as genotypes.bim.

**Pipeline Execution**

The pipeline consists of three R scripts, to be executed in the following order:

**1. GNSW\_FAETH\_VCE.R**

* **Objective**: Estimates variance components for all GRMs (target + HD GRM) across traits.
* **Inputs**: Files in 2\_Data/ and 3\_GRM/ directories.
* **Outputs**: Written to the 4\_VCE/ directory, including results for every GRM-trait combination.

**2. GNSW\_FAETH\_Extract\_h2.R**

* **Objective**: Extracts per-SNP heritability from variance component estimation results.
* **Inputs**: Results from GNSW\_FAETH\_VCE.R.
* **Outputs**: Written to the 5\_FAETH/ directory.

**3. GNSW\_FAETH\_Score\_Out.R**

* **Objective**: Aggregates per-SNP heritability data to calculate FAETH scores across the genome.
* **Inputs**: Results from GNSW\_FAETH\_Extract\_h2.R.
* **Outputs**: Final FAETH scores written to the 5\_FAETH/ directory. pFAETH\_Score.txt file contains FAETH scores for all variants present in the whole-genome sequence file provided.

**Software Dependencies**

* **MTG2**: Used for variance component estimation ([MTG2 Documentation](https://sites.google.com/site/honglee0707/mtg2)).
  + Binaries provided in /bin directory
* **GCTA**: For GRM file generation and processing ([GCTA Overview](https://cnsgenomics.com/software/gcta/#Overview)).
  + Not required by the pipeline, but GRM formats must be prepared from GCTA.

**Key Notes**

* Ensure input files are formatted correctly and placed in the appropriate directories before running the pipeline.
* The pipeline may fail if:
  + Chromosome or position information is mismatched between files.
  + Missing or incorrectly formatted input files are provided.