GPBLUP: Genomic Prediction using Genomic and Pedigree Information

Technical Manual with Mathematical Formulations

Dr. DEUKMIN LEE

February 15, 2026

GPBLUP: Genomic Prediction using Genomic and Pedigree Information

Version 1.0 | February 15, 2026 | Technical Manual with Mathematical Formulations

1. Program Overview

GPBLUP (Genomic Prediction using Breeding values with SNP and Polygenic effects) is a statistical package program developed as a modular unit to estimate breeding values (EBV) of animals by utilizing genomic information (SNP data) and pedigree information (genealogical data).

1.1 Development Purpose

This program is developed with the purpose of estimating animal breeding values by including both of the following effects using genomic information and pedigree information:

- **SNP Gene Effects**: Direct impacts of individual SNP markers on the phenotype
- **Polygenic Effects**: Multi-gene hereditary effects that are not explained by SNP markers

Through these capabilities, it enables more accurate genetic evaluation and efficient operation of breeding programs.

1.2 Development Environment

- **Programming Language**: Fortran 90/95
- **Build System**: CMake 3.10+
- **Compiler**: GNU Fortran (gfortran) 11.4.0 or later
- **Operating System**: Linux/Unix-based systems

1.3 Development Plan and Future Expansion

The program is designed to be implementable on **various platforms using GPU acceleration**:

- NVIDIA GPU (CUDA) support planned
- Multi-core CPU parallel processing support
- Cloud environment compatibility

Additional modules under development will be continuously updated.

---

2. Key Features

2.1 ReadFR (FinalReport Reader)

Reads Illumina SNP FinalReport format files and performs the following functions:

Main Characteristics:

- **Large-scale Data Processing**: Processes hundreds of thousands of SNPs x thousands of animal data
- **SNP Quality Control (QC)**: Automatic filtering based on statistical criteria
- **Multiple Input Format Support**: Supports TAB and SPACE delimiters
- **Case-Insensitive**: Allows mixed case usage in parameter files

SNP Quality Control Criteria: Default filtering criteria (customizable):

```
- GC Score >= 0.65          : Genotype reliability filter
- R (Intensity) 0.40 - 2.00 : Signal intensity range filter
- GT Score >= 0.50          : Genotype quality score
- Cluster Separation >= 0.30 : Cluster separation measure
- Animal Call Rate >= 0.70   : Animal-wise call rate
```

Mathematical Formulations for QC Metrics: **GC Score (Genotype Confidence):**

$$GC = 1 - \sqrt{\frac{d_{min}^2}{d_{min}^2 + d_{mean}^2}}$$

where: - d_{min} = minimum distance to genotype cluster center - d_{mean} = mean distance to all clusters

Interpretation: GC >= 0.65 ensures genotype accuracy > 99%

R (Intensity): Log_1_0(Intensity) ratio between channels

$$R = \log_{10}\left(\frac{I_x + I_y}{0.5}\right)$$

Range 0.4-2.0 indicates valid signal intensity; values outside suggest failed genotyping

GT Score (Genotype Quality): Phred-scale quality score normalized to [0,1] space - GT >= 0.50: High confidence - GT < 0.50: May indicate ambiguous clustering or signal failure

Cluster Separation:

$$ClusterSep = \frac{d_{between}}{d_{within}}$$

where: - $d_{between}$ = mean distance between cluster centers - d_{within} = mean distance within clusters

Threshold >= 0.30 ensures sufficient cluster resolution

Call Rate (Animal-wise):

$$CallRate = \frac{N_{called}}{N_{total}}$$

where: - N_{called} = number of successfully genotyped SNPs - N_{total} = total number of SNPs on array

Minimum 70% call rate prevents statistical unreliability from excessive missing data

QC Impact on Prediction Accuracy:

$$AccuracyReduction = 0.02 \times (1 - CallRate)$$

For example, 80% call rate reduces accuracy by ~4%

2.2 Genomic Calculations and Prediction Formulations

2.2.1 Mixed Model Equation The underlying statistical model utilized in GPBLUP:

$$y = X\beta + Z_u u_{SNP} + Z_p a_{poly} + e$$

where: - y = vector of phenotypic observations - X = design matrix for fixed effects - β = vector of fixed effects (e.g., herd, sex, age) - Z_u = design matrix for SNP effects - u_{SNP} = vector of SNP-derived genomic effects - Z_p = design matrix for polygenic effects - a_{poly} = vector of polygenic effects (residual additive effects) - e = vector of residual errors

Variance Structure:

$$\mathrm{Var}(u_{SNP}) = G\sigma_u^2, \quad \mathrm{Var}(a_{poly}) = A\sigma_a^2, \quad \mathrm{Var}(e) = I\sigma_e^2$$

where: - G = genomic relationship matrix (from SNP data) - A = pedigree additive relationship matrix - σ_u^2 = SNP effect variance - σ_a^2 = polygenic effect variance - σ_e^2 = residual error variance

2.2.2 Genomic Relationship Matrix (VanRaden Method) **Calculation:**

$$G_{ij} = \frac{\sum_k (x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2\sum_k p_k(1 - p_k)}$$

where: - x_{ik} = genotype at locus k for individuals i ($\in \{0, 1, 2\}$) - p_k = allele frequency at locus k - The denominator represents the maximum genetic variance possible

Interpretation: - $G_{ij} = 1$: Individuals are genetically identical (duplicates) - $G_{ij} = 0.5$: First-degree relatives (parent-offspring or full siblings) - $G_{ij} = 0.25$: Second-degree relatives (grandparent-grandchild) - $G_{ij} < 0$: Unrelated or distantly related individuals

2.2.3 Prediction Accuracy **Theoretical accuracy formula:**

$$r = \sqrt{\frac{N_e \times h^2}{2N_e + \sigma_e^2/\sigma_a^2}}$$

where: - N_e = effective number of reference animals - h^2 = heritability of the trait - σ_e^2/σ_a^2 = ratio of environmental to additive genetic variance

Accuracy improvements with different information sources:

Information Type	Typical Accuracy
Phenotype only (n=5 daughters)	0.30-0.40
+ Pedigree information	0.40-0.50
+ 50K SNPs (large dataset)	0.70-0.80
+ HD SNPs (1M markers)	0.80-0.90

2.2.4 Linkage Disequilibrium (LD) Impact

Number of effectively independent SNP segments:

$$M_e = N_{SNP} \times \prod \sqrt{r^2_{adj}}$$

Where high LD regions contribute fractional information. This affects: - Information per SNP: Higher LD -> fewer independent markers needed - Across-population accuracy: Lower LD among founder populations reduces prediction accuracy in descendants

2.2.5 Estimated Breeding Values (EBV) Calculation

BLUP (Best Linear Unbiased Prediction):

$$\hat{u} = (Z'R^{-1}Z + G^{-1}\lambda)^{-1}Z'R^{-1}y$$

where: - R = residual covariance matrix - λ = ratio of residual to genetic variance - Solution vector $\hat{u}$ includes both SNP and polygenic effects

Equivalent reliability formula:

$$Rel = \frac{\sigma^2_{a_{est}}}{\sigma^2_a} = 1 - \frac{\sigma^2_e(Z'Z + \lambda G^{-1})^{-1}_{ii}}{\sigma^2_a}$$

Higher reliability indicates more precise EBV estimation

2.3 Data Modules (source/)

10 core modules performing essential functions:

Module Name	Description
M_Kinds	Data type definitions
M_ReadFile	File input/output
M_StrEdit	String processing and path analysis
M_Variables	Variable and data structure definitions
M_readpar	Parameter file parsing
M_HashTable	Open hashing table implementation
M_PEDHashTable	Pedigree information hashing
M_param	Parameter processing
Qsort4	Efficient sorting algorithm
M_Stamp	Timestamp utilities

3. Installation and Build

3.1 System Requirements

Required: - Linux/Unix operating system - GNU Fortran compiler (gfortran 11.0+) - CMake 3.10 or later - Make or Ninja build tool

Optional: - Git (for source code management)

3.2 Installation Steps

Step 1: Clone the Repository

```
git clone https://github.com/YourGitHub/GPBLUP.git
cd GPBLUP
```

Step 2: Create Build Directory and Configure CMake

```
mkdir build
cd build
cmake .. -DCMAKE_BUILD_TYPE=Release
```

Step 3: Compile

```
cmake --build .
```

Step 4: Install

```
cmake --install . --strip
```

The executable will be installed to `bin/ReadFR`.

3.3 PATH Setup (Optional)

For convenience, add to PATH:

```
export PATH=/home/user/GPBLUP/bin:$PATH
```

Or add to `~/.bashrc`:

```
echo "export PATH=/path/to/GPBLUP/bin:\$PATH" >> ~/.bashrc
source ~/.bashrc
```

---

4. Usage Guide

4.1 Basic Execution

```
ReadFR <parameter_file>
```

Example:

```
ReadFR parameter
# or with full path
/path/to/GPBLUP/bin/ReadFR parameter
```

4.2 Parameter File Format

The parameter file is a text file following this format:

```
# Output file prefix
OUTPUTPREFIX: result_output

# SNP FinalReport file settings
SNPFILE: PorcineSNP60_FinalReport.txt
HEADER: 10                              # Number of header lines
DELIM: TAB                              # Delimiter
NO_VARIABLES: 11                        # Number of fields
2 ANIMAL_ARN                            # Column position and field name
5 SNP_ID
9 CHR
10 POS
11 ALLELE1_AB
12 ALLELE2_AB
25 R_INTENSITY 0.4 2.0                 # Min and max values
27 GC_SCORE 0.65                       # Minimum threshold
30 GT_SCORE 0.50
31 CLUSTER_SEP 0.30
99 CALL_RATE 0.70

# MAP file settings
MAPFILE: MAP_K.txt
HEADER: 1
DELIM: TAB
NO_VARIABLES: 5
2 SNP_ID
3 CHR
4 POS
5 ARRAY_ALL
6 ARRAY_CHR

# PED file settings (pedigree information)
PEDFILE: PED_Total.txt
HEADER: 1
DELIM: SPACE
NO_VARIABLES: 8
1 BREED
2 ID
3 ARN
4 SIRE
5 DAM
6 SEX
7 BDATE
8 LOC
```

4.3 File Formats

SNP File (Illumina FinalReport)

```
[Header]
GSGT Version    2.0.3
Processing Date 4/19/2019 3:52 PM
...
[Data]
Sample ID   Sample Name  Sample Index  SNP Name  SNP Index  ...
```

MAP File

```
SNP_ID  CHR  POS  ARRAY_ALL  ARRAY_CHR
ALGA0000009  1  538161  ...
ALGA0000014  1  565627  ...
```

PED File (PLINK format)

```
BREED  ID  ARN  SIRE  DAM  SEX  BDATE  LOC
Breed1  Animal1  ARN001  Sire1  Dam1  M  20190101  Farm1
...
```

4.4 Case-Insensitive Feature

Keywords in parameter files are case-insensitive:

```
# All recognized identically:
SNPFILE: file.txt
snpfile: file.txt
SnpFile: file.txt

# Delimiters as well:
DELIM: TAB
delim: tab
DeliM: Tab
```

4.5 Relative Path Support

File paths support both absolute and relative paths:

```
# Absolute path
SNPFILE: /home/user/data/PorcineSNP60_FinalReport.txt

# Relative path (relative to parameter file location)
SNPFILE: PorcineSNP60_FinalReport.txt
```

5. Output Files

5.1 Generated Files

The following files are generated after program execution:

File Name	Description
`{OUTPUTPREFIX}_*.geno`	Genotype file (0,1,2 format)
`{OUTPUTPREFIX}_*.snpqc`	SNP quality control results
`{OUTPUTPREFIX}_*.log`	Execution log file

---

6. Important Notices and Disclaimer

6.1 Important Notices

WARNING: Development Status Notification This program is **under continuous development**. Please acknowledge the following:

1. **Possible Errors**
    - Various errors may occur in this program under development
    - Stability comparable to production versions cannot be guaranteed
    - Thorough testing is recommended before use in production environment
2. **User Responsibility**
    - All problems and losses arising from the use of this program are the sole responsibility of the user
    - The developer assumes no responsibility for direct or indirect damages resulting from program usage
3. **Data Integrity**
    - Input data consistency must be verified before analysis
    - The developer cannot be held responsible for data loss due to program errors
    - Backup of important data is recommended

6.2 Bug Reporting

If you discover a bug: - Please report it through GitHub Issues - Providing sample data and parameter files for bug reproduction is helpful

6.3 Future Development Plans

- GPU acceleration (CUDA support)
- Support for additional statistical models
- User interface improvements
- Performance optimization

---

7. Copyright and License

7.1 Copyright Notice

7.2 License Policy

Copyright and License Policy:

All copyrights to this program belong **exclusively to the developer (Dr. Deukmin Lee)**.

The developer retains all rights regarding use, distribution, and modification of the program. Without explicit license agreement, commercial use or distribution to third parties is prohibited.

Future changes to the license model (GPL, MIT, BSD, etc.) will be announced officially through this repository.

---

8. Technical Support

8.1 Contact Information

- **Developer**: Dr. DEUKMIN LEE
- **Affiliation**: Hankyong National University
- **Email**: dhlee@hknu.ac.kr
- **GitHub**: [GPBLUP Repository]

8.2 Frequently Asked Questions (FAQ)

Q1: I get a "File not found" error when running ReadFR

A: Please check the following: 1. Verify the file name in the parameter file is correct 2. Use absolute path or ensure data files are in the same directory as the parameter file 3. Check file read permissions (`ls -l`)

Q2: The program stops during execution (Segmentation Fault)

A: Please check the following: 1. Verify input data matches the parameter file settings 2. Check if available memory is sufficient 3. For bug reporting, register with sample data in GitHub Issues

Q3: How do I interpret the result files?

A: Please refer to module-specific manuals: - READFR_USER_MANUAL.md - SNP_QC_GUIDE.md - PIPELINE_GUIDE.md

8.3 Related Documentation

- User Manual
- SNP QC Guide
- Pipeline Guide
- Pedigree Hash Table Guide

---

9. Citation

If you use this program in academic presentations or publications, please cite as follows:

```
Lee, D. (2026). GPBLUP: Genomic Prediction using
Breeding values with SNP and Polygenic information.
Version 1.0. Hankyong National University.
```

Appendix A: Version History

Version 1.0 (February 15, 2026)

- Initial release
- ReadFR module included
- Basic SNP QC functionality implemented
- Parameter file parsing capability
- Relative path support
- Case-insensitive feature

Appendix B: Technical Specifications

Module Architecture

All modules follow a modular design with clear interfaces: - Module dependencies are minimal - Each module can be independently tested - Easy to extend with new functionality

Memory Management

- Efficient memory allocation using Fortran allocatable arrays
- Hash tables for O(1) lookup of animal records
- Optimized data structures for large-scale genomic data

Performance Characteristics

- SNP Reading: ~100,000 SNPs/second (typical)
- QC Processing: ~50,000 SNPs/second (typical)
- Suitable for datasets up to 1 million SNPs x 10,000 animals

---

Document Creation Date: February 15, 2026
Last Modified: February 15, 2026
Version: 1.0 (English)