ReadFR - SNP Quality Control Program

User Manual and Technical Guide

Program Version: 1.0 - GenomeQC SNP Quality Control Pipeline
Release Date: February 13, 2026
Author: Dr. DEUKMIN LEE
Organization: Hankyong National University
Department: Department of Animal Science
Email: dhlee@hknu.ac.kr

Table of Contents

---

1. Overview

What is ReadFR?

ReadFR is a high-performance Fortran program designed for comprehensive quality control (QC) of SNP genotype data from Illumina FinalReport files. It processes large-scale genomic datasets, applies multiple QC criteria, and generates standardized output formats compatible with GBLUP and other genomic analysis tools.

Key Capabilities

- **Scale**: Processes 10,000+ SNPs and 1,000+ animals simultaneously
- **Speed**: O(1) hash table-based animal lookup
- **Flexibility**: Configurable QC thresholds via parameter file
- **Output**: BLUPF90-compatible GENO format
- **Reliability**: Comprehensive error checking and reporting

Main Features

Feature	Description
Hash Table Technology	O(1) animal lookup time
Memory Efficiency	Dynamic memory allocation
QC Filters	GC Score, R-Intensity, GT Score, Cluster Separation, Call Rate
Multi-parameter Support	Multiple SNP chips and versions
Flexible Configuration	Parameter file-based settings
Detailed Reporting	QC statistics and filtering results

---

2. System Requirements

Hardware Requirements

Component	Minimum	Recommended
CPU	2 cores	4+ cores
RAM	4 GB	8 GB+
Storage	100 MB	500 MB+
Network	Optional	For data transfer

Software Requirements

Software	Minimum Version	Recommended
OS	Linux (CentOS 7+)	Ubuntu 20.04 LTS+
Compiler	gfortran 4.8	gfortran 9.0+
Tools	make, ar	make, ar, gdb

Data Requirements

- **FinalReport File**: Illumina SNP genotyping output
- **Pedigree File**: Animal information and relationships
- **MAP File**: SNP physical positions and marker information
- **DATA File** (optional): Additional animal information

---

3. Installation

Quick Installation (System-wide)

```
cd DKBLUPF90
sudo ./install.sh
```

User Directory Installation (No root required)

```
cd DKBLUPF90
PREFIX=$HOME/.local ./install.sh
```

```
export PATH="$HOME/.local/bin:$PATH"
export LD_LIBRARY_PATH="$HOME/.local/lib:$LD_LIBRARY_PATH"
```

Verification

```
ReadFR --help
which ReadFR
```

See **INSTALL.md** for detailed installation instructions.

---

4. Quick Start Guide

Step 1: Prepare Parameter File

Create a parameter file (named `parameter`):

```
# SNP File information
SNPFILE: PorcineSNP60_SJ_51th_595sp_DY_FinalReport.txt
HEADER: 10
DELIM: TAB
NO_VARIABLES: 11
2 ANIMAL_ARN
5 SNP_ID
9 CHR
10 POS
11 ALLELE1_AB
12 ALLELE2_AB
25 R_INTENSITY 0.4 2.0
27 GC_SCORE 0.65
30 GT_SCORE 0.50
31 CLUSTER_SEP 0.30
99 CALL_RATE 0.70

# MAP File
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
PEDFILE: PED_Total.txt
HEADER: 0
DELIM: TAB
```

```
NO_VARIABLES: 7
1 BREED
2 ID
3 ARN
4 SIRE
5 DAM
6 SEX
7 BDATE

# Output prefix
OUTPUTPREFIX: GENO_QC
```

Step 2: Prepare Input Files

1. **FinalReport File**: Export from GenomeStudio
2. **MAP File**: SNP information
3. **PED File**: Animal pedigree information

Step 3: Run Program

```
ReadFR parameter
```

Step 4: Check Output

```
ls -lh GENO_QC_*.geno
head -2 GENO_QC_YYYYMMDD_00.geno
```

---

5. Parameter File Configuration

Basic Structure

```
# Comments start with #
KEYWORD: value1 value2 ...
```

File Keywords

SNPFILE Illumina FinalReport file path

```
SNPFILE: path/to/FinalReport.txt
```

PEDFILE Pedigree file path

```
PEDFILE: path/to/pedigree.txt
```

MAPFILE SNP map information file path

```
MAPFILE: path/to/snp_map.txt
```

DATAFILE (Optional) Additional animal information file

```
DATAFILE: path/to/additional_data.txt
```

OUTPUTPREFIX Prefix for output files

```
OUTPUTPREFIX: GENO_QC
# Output: GENO_QC_YYYYMMDD_00.geno, GENO_QC_YYYYMMDD_01.geno, ...
```

File Format Keywords

HEADER Number of header lines to skip

```
HEADER: 10    # Skip first 10 lines
```

DELIM Delimiter type: TAB, SPACE, or comma

```
DELIM: TAB      # Tab-delimited
DELIM: SPACE    # Space-delimited
```

NO_VARIABLES Number of variables/columns to use

```
NO_VARIABLES: 11    # Use first 11 columns
```

Field Mapping

Each field is defined by: `column_number FIELD_NAME [threshold1 threshold2]`

QC Thresholds Some fields accept additional threshold values:

```
25 R_INTENSITY 0.4 2.0        # min=0.4, max=2.0
27 GC_SCORE 0.65              # min=0.65
30 GT_SCORE 0.50              # min=0.50
31 CLUSTER_SEP 0.30           # min=0.30
99 CALL_RATE 0.70             # min=0.70
```

Complete Example

```
# ==========================================
# SNP File Configuration
# ==========================================
SNPFILE: GenomeStudio_Export.txt
HEADER: 10
DELIM: TAB
NO_VARIABLES: 11
2 ANIMAL_ARN
5 SNP_ID
9 CHR
10 POS
11 ALLELE1_AB
12 ALLELE2_AB
```

```
25 R_INTENSITY 0.4 2.0
27 GC_SCORE 0.65
30 GT_SCORE 0.50
31 CLUSTER_SEP 0.30
99 CALL_RATE 0.70

# ============================================
# MAP File Configuration
# ============================================
MAPFILE: SNP_positions.txt
HEADER: 1
DELIM: TAB
NO_VARIABLES: 5
2 SNP_ID
3 CHR
4 POS
5 ARRAY_ALL
6 ARRAY_CHR

# ============================================
# Pedigree File Configuration
# ============================================
PEDFILE: Animal_pedigree.txt
HEADER: 0
DELIM: TAB
NO_VARIABLES: 7
1 BREED
2 ID
3 ARN
4 SIRE
5 DAM
6 SEX
7 BDATE

# ============================================
# Output Configuration
# ============================================
OUTPUTPREFIX: Analysis_Result
```

---

6. Input Data Formats

FinalReport File Format

Illumina GenomeStudio export (tab-delimited):

```
Header Line 1
Header Line 2
```

```
...
Header Line 10 (usually contains column names)
Animal_ARN  X-Coordinate  Y-Coordinate  ...  SNP_ID  ...  R-Intensity  GC_Score  GT_Score  Clus
ARN001      12345         67890         ...  snp_1   ...  1.85         0.78      0.95      0.45
ARN002      11234         68901         ...  snp_1   ...  1.92         0.81      0.92      0.48
```

MAP File Format

SNP position information (tab-delimited):

```
SNP_ID      CHR    POS        ARRAY_ALL   ARRAY_CHR
snp_1       1      12345678   1           1
snp_2       1      23456789   2           2
snp_3       2      34567890   3           1
```

PED File Format

Animal pedigree information (tab or space-delimited):

```
BREED   ID       ARN      SIRE   DAM    SEX  BDATE
Duroc   PK001    ARN001   0      0      2    20200101
Duroc   PK002    ARN002   ARN001 0      1    20210315
```

---

7. Quality Control Criteria

GC_SCORE Filter

Purpose: Illumina genotyping quality metric
Valid Range: 0.0 - 1.0
Recommended Threshold: 0.65

R_INTENSITY Filter

Purpose: Overall signal intensity
Valid Range: 0.0 - 3.0
Recommended Range: 0.4 - 2.0
Note: Requires TWO values: min and max

GT_SCORE Filter

Purpose: Genotype clustering quality
Valid Range: 0.0 - 1.0
Recommended Threshold: 0.50

CLUSTER_SEP Filter

Purpose: Cluster separation quality
Valid Range: 0.0 - 1.0
Recommended Threshold: 0.30

CALL_RATE Filter

Purpose: Animal-level call rate
Valid Range: 0.0 - 1.0
Recommended Threshold: 0.70

Default Thresholds

Criterion	Default Value	Recommendation
GC_SCORE	0.65	0.65-0.75
R_INTENSITY min	0.4	0.3-0.5
R_INTENSITY max	2.0	1.8-2.2
GT_SCORE	0.50	0.50-0.60
CLUSTER_SEP	0.30	0.25-0.35
CALL_RATE	0.70	0.90-0.95

---

8. Output Formats

Output File Naming

`[PREFIX]_[YYYYMMDD]_[SequenceNumber].geno`

Example: `GENO_QC_20260213_00.geno`

GENO File Format

Header line:

```
Animal_ID BREED SIRE DAM SEX BDate LOC GENO(1-76756)
```

Data lines:

```
PK001  Duroc  0  0  2  20200101  DY  0 1 1 0 0 1 2 -1 ...
PK002  Duroc  ARN001  0  1  20210315  DY  2 0 0 1 1 0 1 9 ...
```

GENO File Specifications

- **Genotype Coding**: 0 (homozygous 1/1), 1 (heterozygous 1/2), 2 (homozygous 2/2), 9 (missing)
- **Column Order**: Fixed (as shown in header)
- **Format**: Space-delimited
- **Encoding**: Plain text, gzip-compressed available

---

9. Advanced Usage

Case-Insensitive Parameter Handling

The program accepts parameter files with flexible casing:

```
# All accepted formats:
snpfile: data.txt
SNPFILE: data.txt
SnpFile: data.txt
SNPFile: data.txt
```

Field Name Flexibility

Field names support multiple formats:

```
# All equivalent:
25 R_INTENSITY 0.4 2.0
25 R-INTENSITY 0.4 2.0
25 r_intensity 0.4 2.0
25 r-intensity 0.4 2.0
```

Large Dataset Processing

For datasets with >100,000 animals:

```
# Increase stack memory
ulimit -s unlimited

# Run with monitoring
time ReadFR parameter

# Check memory usage
top -p $(pgrep -f "ReadFR parameter")
```

Parallel Batch Processing

Process multiple files sequentially:

```
for file in *.txt; do
    echo "Processing $file..."
    sed "s|SNPFILE:.*|SNPFILE: $file|" parameter.template > parameter_${file}
    ReadFR parameter_${file}
    mkdir -p results_${file}
    mv GENO_QC_*.geno results_${file}/
done
```

---

10. Troubleshooting

Issue: "Input parameter file" error

Cause: No parameter file provided
Solution:

```
ReadFR parameter
# Not: ReadFR
```

Issue: "cannot open file" for input data

Cause: File path error or file not found
Solution:

```
# Check file exists
ls -l /path/to/file

# Use absolute paths in parameter file
# Relative paths are OK from same directory
```

Issue: Segmentation fault during execution

Cause: Memory issue or corrupted data
Solution:

```
# Check available memory
free -h

# Reduce dataset size for testing
# Check file integrity
file yourfile.txt
```

Issue: Unexpected filtering results

Cause: Incorrect threshold values or data format
Solution:

```
# Verify thresholds in parameter file
# Check data value ranges in source files
# Test with default thresholds first
```

---

11. Case Studies

Case Study 1: Large-Scale Commercial Genotyping

Scenario: 1000 pigs, 60K SNPs
Processing Time: ~5 minutes
Output Size: 15 MB

Parameter Configuration:

```
SNPFILE: Commercial_GenomeStudio_Export.txt
SNPFILE: Commercial_Pedigree.txt
MAPFILE: SNP60k_map.txt
OUTPUTPREFIX: Commercial_QC

# Strict QC for downstream analysis
GC_SCORE: 0.75
R_INTENSITY: 0.5 1.9
CALL_RATE: 0.95
```

Case Study 2: Research Project with Multiple Chips

Scenario: Mixed SNP50K and SNP60K
Processing Approach: Separate parameter files per chip

SNP50K_parameter:

```
SNPFILE: Chip_50K_FinalReport.txt
MAPFILE: SNP50k_map.txt
OUTPUTPREFIX: Analysis_50K
```

SNP60K_parameter:

```
SNPFILE: Chip_60K_FinalReport.txt
MAPFILE: SNP60k_map.txt
OUTPUTPREFIX: Analysis_60K
```

---

12. References

Related Publications

- Lee, D. (2025). "Hash table-based genome data processing for large-scale genomic selection." Journal of Genomics.
- GenomeStudio User Guide. Illumina, Inc.
- BLUPF90 Documentation. University of Georgia.

External Resources

- Illumina Genotyping
- BLUPF90 Suite
- Fortran Documentation

Support

For technical support: - Email: dhlee@hknu.ac.kr - Institution: Hankyong National University - Department: Department of Animal Science

---

Appendix: Version History

Version	Date	Changes
1.0	Feb 13, 2026	Initial release

License

Document Information - Title: ReadFR SNP Quality Control Program - User Manual - Version: 1.0 - Date: February 13, 2026 - Author: Dr. DEUKMIN LEE - Organization: Hankyong National University

End of Document