KBeagle

KBeagle is an R package (version ≥ 3.6.3) that extends Beagle (version 5.4) with K-Means clustering and multi-threading capabilities to improve genotype imputation accuracy and computational efficiency.

1.System Requirements:

R: 3.6.3 or higher

TASSEL: 5.0 (included in package)

2.Installation

Clone the repository:

git clone https://github.com/99Xingyu-Guo/KBeagle.git

3. Required Files

All these files must be in the same directory as your data:

tasseladmin-tassel-5 (TASSEL 5.0 package)

Impute.subgroup.r (core imputation script)

KBeagle.R (main package script)

beagle.22Jul22.46e.jar (Beagle v5.4 executable)

gapit_functions.txt (GAPIT v3.5 functions)

vcf2hpm2_wy_v1.pl (format conversion script)

data_NA.txt (example test file)

4. Usage

Run KBeagle in R:

```
source("KBeagle.R")
data_NA <- read.table("data_NA.txt")
KBeagle(data_NA)
```

5. File Descriptions

(1) Core Scripts

Impute.subgroup.r: Handles numerical → HapMap → VCF → numerical conversion pipeline

KBeagle.R: Main interface for end-to-end imputation

beagle.22Jul22.46e.jar: Beagle v5.4 Java executable

vcf2hpm2_wy_v1.pl: Perl script for VCF → HapMap conversion

(2) Dependencies

gapit_functions.txt: Original, unmodified GAPIT v3.5 functions

tasseladmin-tassel-5: Complete TASSEL 5.0 package for format conversion

(3) Example Dataset

data_NA.txt provides a small test file to validate pipeline execution.

6. Version Information

Component	Version
R	≥3.6.3
Beagle	5.4
TASSEL	5.0
GAPIT	3.5

7. Troubleshooting

Memory issues: Adjust -Xmx parameter in Java calls

File not found: Ensure all scripts and data are in same directory

Permission denied: Make Perl scripts executable (chmod +x *.pl)

8. Explanation

During the operation of KBeagle, some intermediate files will be generated. These intermediate files will be deleted upon successful completion of the run. Before imputation, the missing values "NA" in the original file will be converted to the placeholder "1" to create the clustering file. The clustering file is only used for clustering. Later, the imputation is performed after dividing the original file based on individual names.