# Package: castgen (via r-universe)

May 27, 2026

**Title** Estimate Sample Size for Population Genomic Studies

**Version** 1.0.2

**Description** Estimate sample sizes needed to capture target levels of genetic diversity from a population (multivariate allele frequencies) for applications like germplasm conservation and breeding efforts. Compares bootstrap samples to a full population using linear regression, employing the R-squared value to represent the proportion of diversity captured. Iteratively increases sample size until a user-defined target R-squared is met. Offers a parallelized R implementation of a previously developed 'python' method. All ploidy levels are supported. For more details, see Sandercock et al. (2024) <doi:10.1073/pnas.2403505121>.

**License** Apache License (>= 2)

**URL** `https://github.com/alex-sandercock/castgen`

**BugReports** `https://github.com/alex-sandercock/castgen/issues`

**Imports** doParallel (>= 1.0.17), dplyr (>= 1.1.2), foreach (>= 1.5.2), parallel (>= 4.0.0), Rdpack (>= 0.7), stats, utils, vcfR (>= 1.15.0)

**Suggests** testthat (>= 3.0.0)

**Encoding** UTF-8

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.3.2

**Config/pak/sysreqs** libicu-dev

**Repository** https://alex-sandercock.r-universe.dev

**Date/Publication** 2025-04-02 20:01:46 UTC

**RemoteUrl** https://github.com/alex-sandercock/castgen

**RemoteRef** HEAD

**RemoteSha** 2a218145cea3ebebad9d8a54fdcbb688bf883e46

# Contents

capture_diversity.Gmat . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
capture_diversity.VCF . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**Index** **6**

---

## `capture_diversity.Gmat` *Estimate Minimum Number of Individuals to Sample to Capture Population Genomic Diversity (Genotype Matrix)*

---

### Description

This function can be used to estimate the number of individuals to sample from a population in order to capture a desired percentage of the genomic diversity. It assumes that the samples are the columns, and the genomic markers are in rows. Missing data should be set as NA, which will then be ignored for the calculations. All samples must have the same ploidy. This function was adapted from a previously developed Python method (Sandercock et al., 2023) (https://github.com/alex-sandercock/Capturing_genomic_diversity/)

### Usage

```
capture_diversity.Gmat(
  df,
  ploidy,
  r2_threshold = 0.9,
  iterations = 10,
  sample_list = NULL,
  parallel = FALSE,
  batch = 1,
  save.result = FALSE,
  verbose = TRUE
)
```

### Arguments

| | |
|---|---|
| `df` | Genotype matrix or data.frame with the numeric count of alternate alleles (0=homozygous reference, 1 = heterozygous, 2 = homozygous alternate) |
| `ploidy` | The ploidy of the species being analyzed |
| `r2_threshold` | The ratio of diversity to capture (default = 0.9) |
| `iterations` | The number of iterations to perform to estimate the average result (default = 10) |
| `sample_list` | The list of samples to subset from the dataset (optional) |
| `parallel` | Run the analysis in parallel (True/False) (default = FALSE) |
| `batch` | The number of samples to draw in each bootstrap sample iteration (default = 1) |
| `save.result` | Save the results to a .txt file? (default = FALSE) |
| `verbose` | Print out the results to the console (default = TRUE) |

### Value

A data.frame with minimum number of samples required to match or exceed the input ratio

### References

A.M. Sandercock, J.W. Westbrook, Q. Zhang, & J.A. Holliday, A genome-guided strategy for climate resilience in American chestnut restoration populations, Proc. Natl. Acad. Sci. U.S.A. 121 (30) e2403505121, https://doi.org/10.1073/pnas.2403505121 (2024).

### Examples

```
#Example with a tetraploid population
set.seed(123)
test_gmat <- matrix(sample(0:4, 100, replace = TRUE), nrow = 10)
colnames(test_gmat) <- paste0("Sample", 1:10)
rownames(test_gmat) <- paste0("Marker", 1:10)
test_gmat <- as.data.frame(test_gmat)

#Estimate the number of samples required to capture 90% of the population's genomic diversity
result <- capture_diversity.Gmat(test_gmat,
                                 ploidy = 4,
                                 r2_threshold = 0.90,
                                 iterations = 10,
                                 save.result = FALSE,
                                 parallel=FALSE,
                                 verbose=FALSE)

#View results
print(result)
```

---

## `capture_diversity.VCF` *Estimate Minimum Number of Individuals to Sample to Capture Population Genomic Diversity (VCF)*

### Description

This function can be used to estimate the number of individuals to sample from a population in order to capture a desired percentage of the genomic diversity. VCF files can be either unzipped or gzipped. All samples must have the same ploidy and the VCF must contain GT information. This function was adapted from a previously developed Python method (Sandercock et al., 2024) (https://github.com/alex-sandercock/Capturing_genomic_diversity/)

### Usage

```
capture_diversity.VCF(
  vcf,
  ploidy,
```

```
  r2_threshold = 0.9,
  iterations = 10,
  sample_list = NULL,
  parallel = FALSE,
  batch = 1,
  save.result = TRUE,
  verbose = TRUE
)
```

## Arguments

| | |
|---|---|
| `vcf` | Path to VCF file (.vcf or .vcf.gz) with genotype information |
| `ploidy` | The ploidy of the species being analyzed |
| `r2_threshold` | The ratio of diversity to capture (default = 0.9) |
| `iterations` | The number of iterations to perform to estimate the average result (default = 10) |
| `sample_list` | The list of samples to subset from the dataset (optional) |
| `parallel` | Run the analysis in parallel (True/False) (default = FALSE) |
| `batch` | The number of samples to draw in each bootstrap sample iteration (default = 1) |
| `save.result` | Save the results to a .txt file? (default = TRUE) |
| `verbose` | Print out the results to the console (default = TRUE) |

## Value

A data.frame with minimum number of samples required to match or exceed the input ratio

## References

A.M. Sandercock, J.W. Westbrook, Q. Zhang, & J.A. Holliday, A genome-guided strategy for climate resilience in American chestnut restoration populations, Proc. Natl. Acad. Sci. U.S.A. 121 (30) e2403505121, https://doi.org/10.1073/pnas.2403505121 (2024).

## Examples

```
#Example with a diploid vcf

# Example vcf
vcf_file <- system.file("diploid_example.vcf.gz", package = "castgen")

#Estimate the number of samples required to capture 95% of the population's genomic diversity
result <- capture_diversity.VCF(vcf_file,
                                ploidy = 2,
                                r2_threshold = 0.95,
                                iterations = 10,
                                save.result = FALSE,
                                parallel=FALSE,
                                verbose=FALSE)

#View results
```

```
print(result)
```

# Index

`capture_diversity.Gmat`, 2
`capture_diversity.VCF`, 3