

Package: GenomeScaleEmbeddings (via r-universe)

October 19, 2026

Title Exploring Genome-Scale Embeddings in 'Parquet' Files

Version 0.0.0.9000

Description Provides tools to explore genome-scale variant embeddings stored in 'Parquet' files, including data access, matrix conversion, principal component analysis, summaries, and visualization. The datasets accompany Niu et al. (2025), ``Incorporating Large Language Model (LLM) Embeddings for Variation Across the Human Genome" <<https://arxiv.org/abs/2509.20702>>.

Depends R (>= 4.4.0)

Imports bigmemory, bigPCAcpp, DBI, dplyr, duckdb, ggplot2, houba, jsonlite

License LGPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

Remotes HervePerdry/houba, fbertran/bigPCAcpp

Config/roxygen2/version 8.0.0

Config/pak/sysreqs xz-utils

Repository <https://sounkou-bioinfo.r-universe.dev>

Date/Publication 2026-07-21 20:11:18 UTC

RemoteUrl <https://github.com/sounkou-bioinfo/GenomeScaleEmbeddings>

RemoteRef HEAD

RemoteSha c5b01269e2ad03da66d450bda8e3aa077e1912fa

Contents

attachHoubaBigMatrix	2
CopyParquetToDuckDB	2
correlatePCWithPosition	3

DatasetParquetUrlList	4
embeddingSummary	4
getPcaScores	5
houbaPCA	5
infoSummary	6
IterateEmbeddingsMatrixBatches	6
OpenRemoteParquetView	7
plotPcaDims	7
plotPCSpatialCorrelation	8
print.HoubaEmbeddings	8
writeEmbeddingsHoubaFromDuckDB	9
Index	10

attachHoubaBigMatrix	<i>Attach a houba file and return the bigmemory::big.matrix</i>
----------------------	---

Description

Attach a houba file and return the bigmemory::big.matrix

Usage

attachHoubaBigMatrix(houba_file)

Arguments

houba_file Path to houba file (without .desc extension)

Value

bigmemory::big.matrix object

CopyParquetToDuckDB	<i>Copy remote parquet files into a local DuckDB database file</i>
---------------------	--

Description

Copy remote parquet files into a local DuckDB database file

Usage

```
CopyParquetToDuckDB(  
  db_path = "local_embeddings.duckdb",  
  urlList = DatasetParquetUrlList(),  
  table_name = "embeddings",  
  overwrite = FALSE  
)
```

Arguments

db_path	Path to the local DuckDB database file
urlList	Character vector of parquet URLs
table_name	Name of the table to create in the DuckDB database
overwrite	Whether to overwrite the existing database

Value

Invisibly, the path to the DuckDB database file.

`correlatePCWithPosition`

Compute correlation between PC scores and genomic position, per chromosome

Description

Compute correlation between PC scores and genomic position, per chromosome

Usage

```
correlatePCWithPosition(pc_scores, info_df, pc = 1, method = "spearman")
```

Arguments

pc_scores	Matrix of principal component scores (variants x PCs)
info_df	Data frame with variant info (must contain 'chrom' and 'pos')
pc	Which principal component to correlate (default: 1)
method	Correlation method (default: 'spearman')

Value

Named vector of correlation values per chromosome

`DatasetParquetUrlList` *List of the huggingface datasets for the paper "Incorporating LLM Embeddings for Variation Across the Human Genome"*

Description

List of the huggingface datasets for the paper "Incorporating LLM Embeddings for Variation Across the Human Genome"

Usage

```
DatasetParquetUrlList()
```

Value

A character vector of dataset names

`embeddingSummary` *Quick summary for houba mmatrix*

Description

Quick summary for houba mmatrix

Usage

```
embeddingSummary(embMat)
```

Arguments

`embMat` `houba mmatrix`

Value

List with `dim`, `colMeans`, `rowMeans`

getPcaScores	<i>Get PCA scores from houbaPCA result</i>
--------------	--

Description

Get PCA scores from houbaPCA result

Usage

```
getPcaScores(houbaPCA_res)
```

Arguments

houbaPCA_res List returned by houbaPCA (with 'pca' and 'houbaM')

Value

Matrix of principal component scores (variants x PCs)

houbaPCA	<i>PCA using bigPCAcpp on houba mmatrix or bigmemory::big.matrix</i>
----------	--

Description

PCA using bigPCAcpp on houba mmatrix or bigmemory::big.matrix

Usage

```
houbaPCA(  
  embMat = "local_embeddings.houba",  
  center = TRUE,  
  scale = TRUE,  
  ncomp = 15  
)
```

Arguments

embMat	houba mmatrix path or bigmemory::big.matrix
center	logical, whether to center columns
scale	logical, whether to scale columns
ncomp	number of principal components to compute

Value

List with PCA result object and houbaM big.matrix

infoSummary	<i>Quick summary for houba info mmatrix</i>
-------------	---

Description

Quick summary for houba info mmatrix

Usage

infoSummary(infoMat)

Arguments

infoMat houba mmatrix

Value

List with dim, unique chroms, and example rsids

IterateEmbeddingsMatrixBatches	<i>Iterate over embeddings as matrix batches from a local DuckDB file</i>
--------------------------------	---

Description

Iterate over embeddings as matrix batches from a local DuckDB file

Usage

```
IterateEmbeddingsMatrixBatches(  
  chunk_size = 1e+05,  
  db_path = "local_embeddings.duckdb",  
  table_name = "embeddings"  
)
```

Arguments

chunk_size Number of rows per batch
db_path Path to the local DuckDB database file
table_name Name of the table in the DuckDB database

Value

A list of embedding batches

OpenRemoteParquetView	<i>Open remote parquet files as a DuckDB VIEW and return as tibble (minimal, http(s) only)</i>
-----------------------	--

Description

Open remote parquet files as a DuckDB VIEW and return as tibble (minimal, http(s) only)

Usage

```
OpenRemoteParquetView(  
  urlList = DatasetParquetUrlList(),  
  view_name = "embeddings",  
  db_path = tempfile(fileext = ".duckdb"),  
  unify_schemas = FALSE  
)
```

Arguments

urlList	Character vector of parquet URLs
view_name	Name of the DuckDB view to create
db_path	Path to DuckDB database file (default: temporary file, null for in-memory)
unify_schemas	Whether to unify schemas across files

Value

dplyr tibble referencing the DuckDB view

plotPcaDims	<i>Plot PCA dimensions using ggplot2, colored by annotation</i>
-------------	---

Description

Plot PCA dimensions using ggplot2, colored by annotation

Usage

```
plotPcaDims(pc_scores, info_df, annotation_col = "chrom", dim1 = 1, dim2 = 2)
```

Arguments

pc_scores	Matrix of principal component scores (variants x PCs)
info_df	Data frame with variant info (must contain annotation column)
annotation_col	Name of column in info_df to color by (e.g. 'gwas')
dim1	First PC dimension to plot (default: 1)
dim2	Second PC dimension to plot (default: 2)

Value

A ggplot object.

plotPCSpatialCorrelation

Plot spatial correlation between PC scores and genomic position, faceted by chromosome

Description

Plot spatial correlation between PC scores and genomic position, faceted by chromosome

Usage

```
plotPCSpatialCorrelation(pc_scores, info_df, pc = 1)
```

Arguments

pc_scores	Matrix of principal component scores (variants x PCs)
info_df	Data frame with variant info (must contain 'chrom' and 'pos')
pc	Which principal component to plot (default: 1)

Value

Invisibly, NULL. Called for its plotting side effect.

print.HoubaEmbeddings *Print a Houba embeddings object*

Description

Print a Houba embeddings object

Usage

```
## S3 method for class 'HoubaEmbeddings'
print(x, ...)
```

Arguments

x	A HoubaEmbeddings object.
...	Additional arguments (currently ignored).

Value

x, invisibly.

`writeEmbeddingsHoubaFromDuckDB`

*Write embeddings to houba mmatrix and return info as data.frame
from a local DuckDB file*

Description

Write embeddings to houba mmatrix and return info as data.frame from a local DuckDB file

Usage

```
writeEmbeddingsHoubaFromDuckDB(  
  dbPath = "local_embeddings.duckdb",  
  tableName = "embeddings",  
  embeddingCol = "embedding",  
  batchSize = 1e+05,  
  embeddingDim = 3072,  
  embeddingFile = gsub("\\\\.duckdb$", ".houba", dbPath),  
  overwrite = FALSE  
)
```

Arguments

<code>dbPath</code>	Path to the local DuckDB database file
<code>tableName</code>	Name of the table in the DuckDB database
<code>embeddingCol</code>	Name of the embeddings column
<code>batchSize</code>	Number of rows per batch
<code>embeddingDim</code>	Dimension of each embedding vector
<code>embeddingFile</code>	Path for houba mmatrix file
<code>overwrite</code>	Whether to overwrite existing houba file

Value

An object of class 'HoubaEmbeddings' with mmatrix, info data.frame, and houba file path

Index

`attachHoubaBigMatrix`, [2](#)
`CopyParquetToDuckDB`, [2](#)
`correlatePCWithPosition`, [3](#)
`DatasetParquetUrlList`, [4](#)
`embeddingSummary`, [4](#)
`getPcaScores`, [5](#)
`houbaPCA`, [5](#)
`infoSummary`, [6](#)
`IterateEmbeddingsMatrixBatches`, [6](#)
`OpenRemoteParquetView`, [7](#)
`plotPcaDims`, [7](#)
`plotPCSpatialCorrelation`, [8](#)
`print.HoubaEmbeddings`, [8](#)
`writeEmbeddingsHoubaFromDuckDB`, [9](#)