Package: GenomeScaleEmbeddings (via r-universe)

June 4, 2026

Title Exploring Genome Scale Embeddings Parquet Files

Version 0.0.0.9000

Description Some explorations of the genomics embedding from the paper ``Incorporating LLM Embeddings for Variation Across the Human Genome" <https://arxiv.org/html/2509.20702v1>

Depends R (>= 4.4.0)

Imports duckdb, duckplyr, ggplot2, knitr, jsonlite, houba, bigPCAcpp, httr2

License LGPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.3

Remotes HervePerdry/houba, fbertran/bigPCAcpp

Config/pak/sysreqs libssl-dev xz-utils

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

Date/Publication 2025-11-30 12:20:16 UTC

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

RemoteRef HEAD

RemoteSha 6a7f6dd578e5828dc781dc86c300d31b851c2472

Contents

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

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` *Copy remote parquet files into a local DuckDB database file using explicit URLs*

Description

Copy remote parquet files into a local DuckDB database file using explicit URLs

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

---

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 *Get PCA scores from houbaPCA result*

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

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 *Quick summary for houba info mmatrix*

Description

Quick summary for houba info mmatrix

Usage

```
infoSummary(infoMat)
```

Arguments

infoMat	houba mmatrix

Value

List with dim, unique chroms, and example rsids

IterateEmbeddingsMatrixBatches *Iterate over embeddings as matrix batches from a local DuckDB file*

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

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 *Plot PCA dimensions using ggplot2, colored by annotation*

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)

---

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)

---

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

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

Value

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

Index