ember Documentation

Release 0.1.0

Pachter Lab

Sep 18, 2025

CONTENTS:

CHAPTER

ONE

GENERATING ENTROPY METRICS

ember.light_ember.**light_ember**(*h5ad_dir*, *partition_label*, *save_dir*, *sampling=True*, *sample_id_col=None*, *category_col=None*, *condition_col=None*, *num_draws=100*, *save_draws=False*, *seed=42*, *partition_pvals=True*, *block_pvals=False*, *block_label=None*, *n_pval_iterations=1000*, *n_cpus=1*)

Runs the ember entropy metrics and p-value generation workflow on an AnnData object.

This function loads an AnnData *.h5ad* file, optionally performs balanced sampling across replicates, computes entropy metrics for the specified partition, and generates p-values for Psi and Zeta and optionally Psi_block for a block of choice.

Entropy metrics generated:

- Psi : Fraction of infromation explained by partition of choice
- Psi_block : Specificity of infromation to a block
- Zeta : Speicifcty to a partition/ distance of Psi_blocks distribution from uniform

Parameters

- **h5ad_dir** (*str, Required*) – Path to the *.h5ad* file to process. Data should be log1p and depth normalized before running ember. Remove genes with less than 100 reads before running ember.
- **partition_label** (*str, Required*) – Column in *.obs* used to partition cells for entropy calculations (e.g., "celltype", "Genotype", "Age"). Required to run process. If performing calculation on interaction term, first create a column in *.obs* concatnating the two columns of interested with a semicolon (:).
- **save_dir** (*str, Required*) – Path to directory where results will be saved. Required to run process.
- **sampling** (*bool, default=True*) – Whether to perform balanced sampling across replicates before entropy calculation. If True, *sample_id_col*, *category_col*, and *condition_col* must be provided. Sampling should only be False if fast intermediate results are desired or if there are no replicates to sample over. If sampling is set to False but either partition_pvals or block_pvals are set to True then the sampling=False will be overridden as pval generation requires sampling.
- **sample_id_col** (*str, default = None*) – The column in *.obs* with unique identifiers for each sample or replicate (e.g., 'sample_id', 'mouse_id').
- **category_col** (*str, default = None*) – The column in *.obs* defining the primary group to balance across in order to generate a balanced sample of the experiment. (e.g., 'disease_status', 'mouse_strain'). Refer to readme for further explanation on how to select category and condition columns. category_col and condition_col are interchangable. If balanc-

ing across more than 2 variables, generate interaction terms, create a column in *.obs* concatnating the two (or more) columns of interested with a semicolon (:). This way balancing can be done across as many variables as desired.

- **condition_col** (*str, default = None*) – The column in *.obs* containing the conditions to balance within each categoryto generate a balanced sample of the experiment. (e.g., 'sex', 'treatment'). Refer to readme for further explanation on how to select category and condition columns. category_col and condition_col are interchangable. If balancing across more than 2 variables, generate interaction terms, create a column in *.obs* concatnating the two (or more) columns of interested with a semicolon (:). This way balancing can be done across as many variables as desired.
- **num_draws** (*int, default = 100*) – The number of balanced subsets to generate, by default 100.
- **save_draws** (*bool, default=False*) – Whether to save intermediate draws to save_dir.
- **seed** (*int, default = 42*) – The random seed for reproducible draws, by default 42.
- **partition_pvals** (*bool, default=True*) – Whether to compute permutation-based p-values for the *partition_label*. P-values are generated by sampling. If sampling = False and partition_pvals = True, the sampling=False will be overwritten. Calls generate_pavls, which can be called manually after metric generation as well.
- **block_pvals** (*bool, default=False*) – Whether to compute permutation-based p-values for the *block_label*. P-values are generated by sampling. If sampling = False and block_pvals = True, the sampling=False will be overwritten. Calls generate_pavls, which can be called manually after metric generation as well.
- **block_label** (*str, default = None*) – One value in the *.obs* column for partition_label to use for block-based permutation tests. Required if *block_pvals=True*.
- **n_pval_iterations** (*int, default=1000*) – Number of permutations to use for p-value calculation.
- **n_cpus** (*int, default=1*) – Number of CPU cores to use for parallel permutation testing. For this script, performance is I/O-bound and may not improve beyond 4-8 cores.'

Return type
None

Notes

- Results are saved to *save_dir* as CSV files.
- one csv file with all entropy metrics
- one csv file in a new Psi_block_df folder with psi block values for all blocks in a partition
- Separate file for pvals
- Separate files for each partition
- Alternate file names depending on sampling on or off.

What to expect inside 'entropy_metrics.csv':

- gene_name: All genes in *.var*
- Psi_mean: Psi scores averaged over n draws (between 0 and 1) corresponding to the selected partition for each gene in *.var*.

- Psi_std: Standard deviation of Psi scores across n draws corresponding to the selected partition for each gene in *.var*.
- Psi_valid_counts: Number of valid Psi scores observed across n draws. Only use genes for downstream analysis that have valid counts=num_draws. If valid counts is not close to num_draws, increase threshold for filtering genes with low reads beforehand(recommended <100 reads, increase as needed).
- Zeta_mean: Zeta scores averaged over n draws (between 0 and 1) corresponding to the selected partition for each gene in *.var*.
- Zeta_std: Standard deviation of Zeta scores across n draws corresponding to the selected partition for each gene in *.var*.
- Zeta_valid_counts: Number of valid Psi scores observed across n draws. Only use genes for downstream analysis that have valid counts=num_draws. If valid counts is not close to num_draws, increase threshold for filtering genes with low reads beforehand (recommended <100 reads, increase as needed).

What to expect inside 'Psi_block_df/':

- mean_Psi_block_df.csv : A dataframe of mean Psi_block scores (between 0 and 1) corresponding to the selected partition for each gene in *.var*. Scores are caluclated for all blocks, each column of the dataframe corresponds to one block.
- std_Psi_block_df.csv : A dataframe of standard deviations for Psi_block scores corresponding to the selected partition for each gene in *.var*.Scores are caluclated for all blocks, each column of the dataframe corresponds to one block.

What to expect inside 'pvals_entropy_metrics.csv':

- gene_name: All genes in *.var*
- Psi: Psi scores averaged over n draws (between 0 and 1) generated by light_ember for each gene in *.var*.
- Psi p-value: Permutation based empirical p-values for observed Psi scores for each gene in *.var*.
- Zeta: Zeta scores averaged over n draws (between 0 and 1) generated by light_ember for each gene in *.var*.
- Zeta p-value: Permutation based empirical p-values for observed Zeta scores for each gene in *.var*.
- Psi FDR: Multiple testing corrected q-values for Psi scores.
- Zeta FDR: Multiple testing corrected q-values for Zeta scores.Correction perfromed to include all p-values generated in a single file (Psi and Zeta).

If block_pvals = True and a single block_label is given:

- psi_block: psi_block scores (between 0 and 1) generated by light_ember for each gene in *.var*.
- psi_block p-value: Permutation based empirical p-values for observed psi_block scores for each gene in *.var*.
- psi_block FDR: Multiple testing corrected q-values for psi_block scores. Correction perfromed to include all p-values generated in a single file (Psi, psi_block and Zeta).

CHAPTER

TWO

GENERATING P-VALUES

ember.generate_pvals.**generate_pvals**(*h5ad_dir*, *partition_label*, *entropy_metrics_dir*, *save_dir*, *sample_id_col*, *category_col*, *condition_col*, *block_label=None*, *seed=42*, *n_iterations=1000*, *n_cpus=1*, *Psi_real=None*, *Psi_block_df_real=None*, *Zeta_real=None*)

Calculate empirical p-values for entropy metrics from permutation test results. This function can be called manually or accessed through light_ember with partition_pvals = True or block_pvals = True.

Manual access useful if wanting to generate p-values for multiple blocks and partitions of interest after initial investigation using entropy metrics.

Integrated access with light_ember is easier if investigating only a partition or a block in a partition.

Entropy metrics generated:

- Psi : Fraction of infromation explained by partition of choice
- Psi_block : Specificity of infromation to a block
- Zeta : Speicifcty to a partition/ distance of Psi_blocks distribution from uniform

Parameters

- **h5ad_dir** (*str, Required*) – Path to the *.h5ad* file to process. Data should be log1p and depth normalized before running ember. Remove genes with less than 100 reads before running ember.
- **partition_label** (*str, Required*) – Column in *.obs* used to partition cells for entropy calculations (e.g., "celltype", "Genotype", "Age"). Required to run process. If performing calculation on interaction term, first create a column in *.obs* concatnating the two columns of interested with a semicolon (:).
- **entropy_metrics_dir** (*str, Required*) – Path to csv with entropy metrics to use for generating pvals.
- **save_dir** (*str, Required*) – Path to directory where results will be saved.
- **sample_id_col** (*str, Required*) – The column in *.obs* with unique identifiers for each sample or replicate (e.g., 'sample_id', 'mouse_id').
- **category_col** (*str, Required*) – The column in *.obs* defining the primary group to balance across in order to generate a balanced sample of the experiment. (e.g., 'disease_status', 'mouse_strain'). Refer to readme for further explanation on how to select category and condition columns. category_col and condition_col are interchangable. If balancing across more than 2 variables, generate interaction terms, create a column in *.obs* concatnating the two (or more) columns of interested with a semicolon (:). This way balancing can be done across as many variables as desired.

- **condition_col** (*str, Required*) – The column in *.obs* containing the conditions to balance within each category to generate a balanced sample of the experiment. (e.g., 'sex', 'treatment'). Refer to readme for further explanation on how to select category and condition columns. category_col and condition_col are interchangable. If balancing across more than 2 variables, generate interaction terms, create a column in *.obs* concatnating the two (or more) columns of interested with a semicolon (:). This way balancing can be done across as many variables as desired.
- **block_label** (*str, default=None*) – Block in partition to calucate p-values for. Default set to None, program will continue generating p-values for only Psi and Zeta.
- **seed** (*int, default=42*) – The random seed for reproducible draws, by default 42.
- **n_iterations** (*int, default = 1000*) – Number of iterations to calulate p-vales. Default set to 1000. Note that doing fewer than 1000 iterations is a good choice to get first pass p-values but for reliable p-values 1000 iterations is recommended. Larger than 1000 will give you more relibale p-values but will increase runtime significantly.
- **n_cpus** (*int, default=1*) – Number of cpus to use to perfrom p-value calculation. Default set to 1 assuming no parallel compute power on local machine. User can input -1 to use all available cpus but one.
- **Psi_real** (*pd.Series, default=None*) – Observed Psi values for each gene. Used by light_ember, not necessary for user use.
- **Psi_block_df_real** (*pd.Dataframe, default = None*) – Observed Psi_block values for all blocks in chosen partition. Used by light_ember, not necessary for user use.
- **Zeta_real** (*pd.Series, default=None*) – Observed Zeta values for each gene. Used by light_ember, not necessary for user use.

Return type
None

Notes

What to expect inside 'pvals_entropy_metrics.csv':

- gene_name: All genes in *.var*
- Psi: Psi scores averaged over n draws (between 0 and 1) generated by light_ember for each gene in *.var*.
- Psi p-value: Permutation based empirical p-values for observed Psi scores for each gene in *.var*.
- Zeta: Zeta scores averaged over n draws (between 0 and 1) generated by light_ember for each gene in *.var*.
- Zeta p-value: Permutation based empirical p-values for observed Zeta scores for each gene in *.var*.
- Psi FDR: Multiple testing corrected q-values for Psi scores.
- Zeta FDR: Multiple testing corrected q-values for Zeta scores.Correction perfromed to include all p-values generated in a single file (Psi and Zeta).

if block_pvals = True and a single block_label is given:

- psi_block: psi_block scores (between 0 and 1) generated by light_ember for each gene in *.var*.
- psi_block p-value: Permutation based empirical p-values for observed psi_block scores for each gene in *.var*.
- psi_block FDR: Multiple testing corrected q-values for psi_block scores. Correction perfromed to include all p-values generated in a single file (Psi, psi_block and Zeta).

CHAPTER

THREE

PLOTTING FUNCTIONS

This section details the various plotting functions available in the *ember.plots* module.

3.1 Psi vs. Zeta scatter plots

ember.plots.**plot_partition_specificity**(*partition_label*, *pvals_dir*, *save_dir*, *highlight_genes=None*, *fontsize=18*, *custom_palette=None*)

Generate a Zeta vs. Psi scatter plot to visualize partition-specific genes.

This function reads p-value data, colors genes based on their statistical significance for Psi and Zeta scores, and highlights top "marker" and "housekeeping" genes. Allows for custom highlighting of a user-provided gene list. Fontsize and color pallette can be customized.

Parameters

- **partition_label** (*str, Required.*) – The label for the partition being plotted, used in the plot title.
- **pvals_dir** (*str, Required.*) – Path to the input CSV file containing p-values and scores (Psi, Zeta, FDRs). The CSV must have gene names as its index column.
- **save_dir** (*str, Required.*) – Path where the output plot image will be saved.
- **highlight_genes** (*list[str], default=None.*) – A list of gene names to highlight and annotate on the plot, by default None.
- **fontsize** (*int, default=18.*) – The base font size for plot labels and text, by default 18.
- **custom_palette** (*list[str], default=None.*) – A list of 7 hex color codes to customize the plot's color scheme. If None, a default palette is used. Please provide list in this order ['significant by psi', 'significant by zeta', 'highlight genes', 'significant by both', 'cirlce markers', 'circle housekeeping genes', 'significant by neither']

Return type
None

3.2 Psi vs. psi_block scatter plots

ember.plots.**plot_block_specificity**(*partition_label*, *block_label*, *pvals_dir*, *save_dir*, *highlight_genes=None*, *fontsize=18*, *custom_palette=None*)

Generate a psi_block vs. Psi scatter plot to visualize block-specific genes.

This function reads p-value data, colors genes based on their statistical significance for Psi and psi_block scores, and highlights the top genes significant in both metrics. Allows for custom highlighting of a user-provided gene list. Fontsize and color pallette can be customized.

Parameters

- **partition_label** (*str, Required.*) – The label for the partition, used in the plot title.
- **block_label** (*str, Required.*) – The label for the block variable (e.g., a cell type or condition).
- **pvals_dir** (*str, Required.*) – Path to the input CSV file containing p-values and scores. The CSV must have gene names as its index column.
- **save_dir** (*str, Required.*) – Path where the output plot image will be saved.
- **highlight_genes** (*list[str], default=None.*) – A list of gene names to highlight and annotate on the plot, by default None.
- **fontsize** (*int, default = 18.*) – The base font size for plot labels and text, by default 18.
- **custom_palette** (*list[str], default=None.*) – A list of 6 hex color codes to customize the plot's color scheme. If None, a default palette is used. Provide list of colors int his order: ['significant by psi', 'significant by psi_block', 'highlight genes', 'significant by both', 'cirlce markers', 'circle housekeeping genes', 'significant by neither']

Return type
None

3.3 Descriptive bar plot of sample counts

ember.plots.**plot_sample_counts**(*h5ad_dir*, *save_dir*, *sample_id_col*, *category_col*, *condition_col*, *fontsize=18*)

Generate a bar plot showing the number of unique individuals per category and condition.

This function reads an AnnData object from an .h5ad file in backed mode, calculates the number of unique individuals for each combination of a given category and condition, and visualizes these counts as a grouped bar plot. Fontsize can be customized.

Parameters

- **h5ad_dir** (*str, Required*) – Path to the input AnnData (.h5ad) file.
- **save_dir** (*str, Required*) – Path to directory to save the output plot image.
- **sample_id_col** (*str, Required*) – The column name in adata.obs that contains unique sample IDs.
- **category_col** (*str, Required*) – The column name to use for the primary categories on the x-axis.
- **condition_col** (*str, Required*) – The column name to use for grouping the bars (hue).
- **fontsize** (*int, default = 18.*) – The base font size for plot labels and text, by default 18.

Return type
None

3.4 psi_blocks bar plots with error bars

ember.plots.**plot_psi_blocks**(*gene_name*, *partition_label*, *psi_block_df_dir*, *save_dir*, *fontsize=18*)

Generates and saves a bar plot of mean psi block values with error bars.

This function reads two CSV files from a specified directory: one for mean psi block values and one for standard deviations. It plots the mean values for a specific gene as a bar plot with corresponding standard deviation error bars. Fontsize can be customized.

Parameters

- **gene_name** (*str, Required*) – The name of the gene (row) to select and plot from the CSV files.
- **partition_label** (*str, Required*) – The partition label used to find the correct files (e.g., 'Genotype').
- **psi_block_df_dir** (*str, Required*) – Path to the directory containing the mean and std CSV files. Files must be named 'mean_Psi_block_df_{partition_label}.csv' and 'std_Psi_block_df_{partition_label}.csv'.
- **save_dir** (*str, Required*) – Path to directory to save the output plot image.
- **fontsize** (*int, default=18.*) – The base font size for plot labels and text, by default 18.

Return type

None

PYTHON MODULE INDEX

e

INDEX