



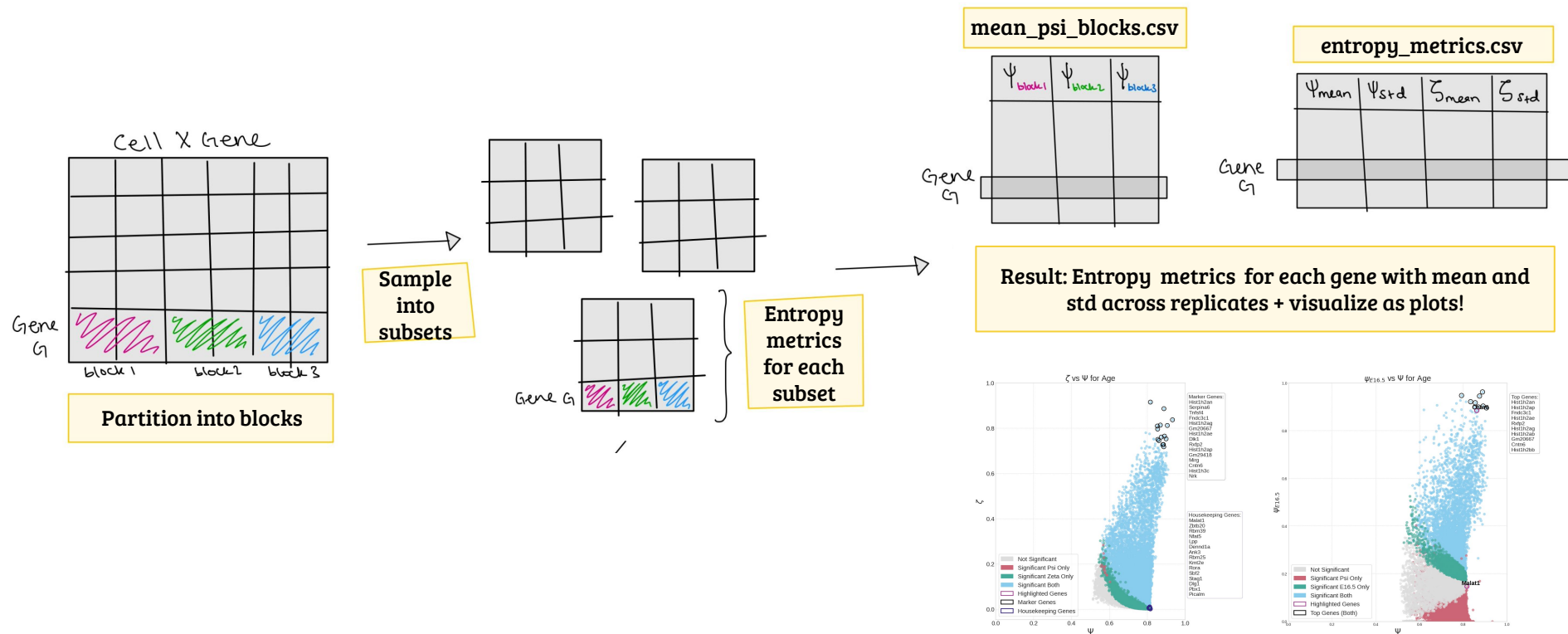
# ember entropy metrics for biological exploration

---

**Nikhila P. Swarna**  
**September 18 2025**

**Pachter Lab**  
**Division of Biology and Bioengineering**  
**California Institute of Technology**

# ember entropy metrics for biological exploration



# ember entropy metrics for biological exploration

**light\_ember** - one stop shop for generating entropy metrics and p-values

## INPUT

- **h5ad\_dir** (path to adata)
- **partition\_label** (col in adata.obs)
- **save\_dir** (path to save results)
- **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 (for parallel processing of sampling)

→ **light\_ember** →

## OUTPUT

- csv file in **save\_dir** with all entropy metrics
- csv file in **Psi\_block\_df** folder with psi block
  - Separate file for pvals
  - Separate files for each partition
  - Alternate file names depending on sampling on or off.

# ember entropy metrics for biological exploration

## generate\_pvals

- manual access to generate pvals after initial investigation using **light\_ember**

### INPUT

- **h5ad\_dir** (path to adata)
- **partition\_label** (col in adata.obs)
- **entropy\_metrics\_dir** (path to light\_ember output files)
- **save\_dir** (path to save results)
- **sample\_id\_col**
- **category\_col**
- **condition\_col**
- **block\_label**=None
- **seed** = 42
- **n\_iterations**=1000
- **n\_cpus**=1



## generate\_pvals



### OUTPUT

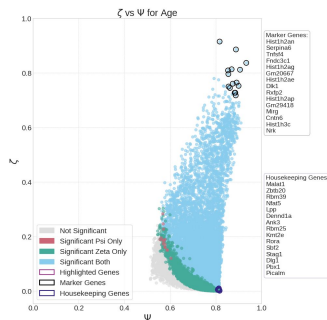
- csv file in save\_dir with entropy metrics and corresponding p-values and FDR q-values
  - Separate files for each partition

# ember entropy metrics for biological exploration

## plot\_partition\_specificity

### INPUT

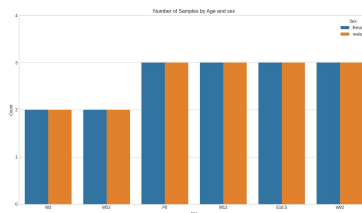
- partition\_label
- pvals\_dir
- save\_dir
- highlight\_genes
- fontsize
- color\_palette



## plot\_sample\_counts

### INPUT

- h5ad\_dir
- save\_dir
- sample\_id\_col
- category\_col
- condition\_col
- fontsize

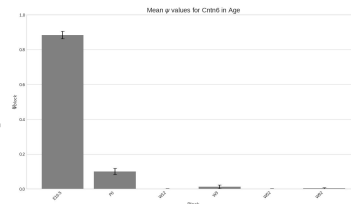


## ember.plots

## plot\_psi\_blocks

### INPUT

- gene\_name
- partition\_label
- psi\_block\_df\_dir
- save\_dir
- fontsize



## plot\_block\_specificity

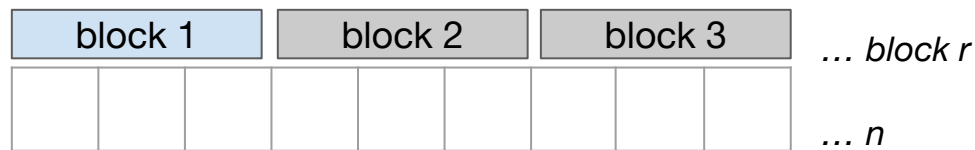
### INPUT

- partition\_label
- block\_label
- pvals\_dir
- save\_dir
- highlight\_genes
- fontsize
- color\_palette



# Defining entropy metrics for biological exploration

For a given gene in a count matrix that can be partitioned into  $r$  blocks (based on sex, strain, cell type, tissue, etc), we introduce **3 measures of specificity**:



- Psi ( $\Psi$ )
- Psi<sub>block</sub> ( $\Psi_{block}$ )
- Zeta ( $\zeta$ )

$\Psi$	$\Psi_{block}$	$\zeta$
Fraction of information explained by partitioning	Specificity to a block	Specificity to a partition

# Ψ- Information Fraction by Partition

The fraction of information explained by using a particular partition on gene  $g$ 's counts is given by:

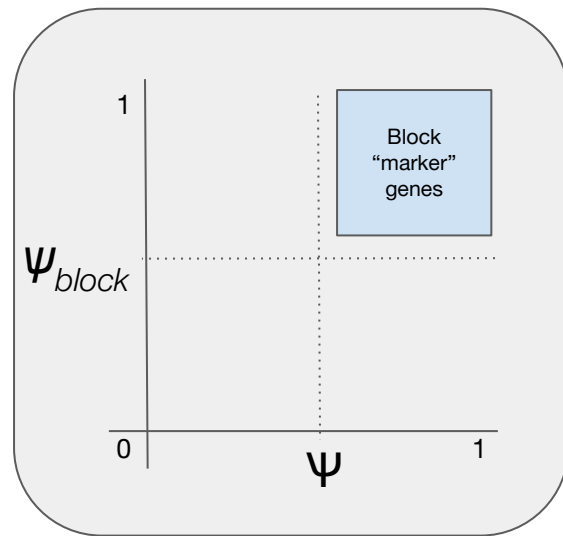
$$\Psi = \frac{E_W}{E_T} = 1 - \frac{E_B}{E_T}$$

The Specificity of Information to Block, denoted by  $\psi_{block}$ , is the contribution of each block to  $\Psi$ :

$$1 = \boxed{\psi_1} + \boxed{\psi_2} + \dots + \boxed{\psi_r}$$

$$\frac{p_1 E_1}{E_W} + \frac{p_2 E_2}{E_W} + \dots + \frac{p_r E_r}{E_W}$$

block 1			block 2			block 3			... block $r$
									... $n$

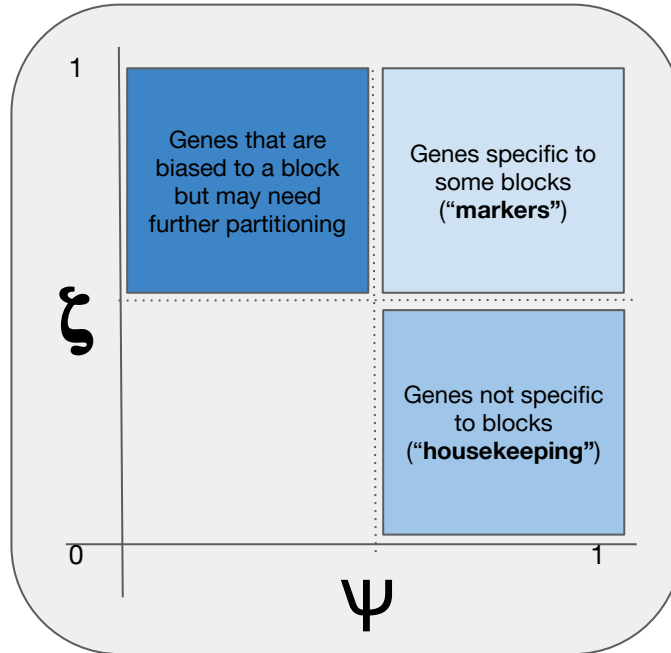


# $\zeta$ - Specificity of Information to Partition

The specificity of information to a partition ( $\zeta$ ) is given by:

$$\zeta = 1 - \frac{H(\psi_{\text{blocks}})}{\log r}$$

Comparison of the SIB distribution  
to the uniform distribution





# How to select category and condition

---

Category - Mouse strain

Condition - Sex

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mouse strain	Strain A				Strain B				Strain C				Strain D			
Sex	Male		Female		Male		Female		Male		Female		Male		Female	

**Number of unique draws**

$$\begin{aligned} &= (\text{Number of replicates per category-condition group})^{(\text{Number of category-condition groups})} \\ &= 2^8 = \mathbf{256} \end{aligned}$$

One example draw:

Sample	1	3	5	7	9	11	13	15
Mouse strain	Strain A		Strain B		Strain C		Strain D	
Sex	Male	Female	Male	Female	Male	Female	Male	Female

# How to select category and condition

---

Category - Cell line  
Condition - Gene perturbation

Sample	1	2	3	4	5	6	7	8	9	10	11	12
Cell line	Cell line A						Cell line B					
Gene perturbation	Wildtype		Overexpression		Knockout		Wildtype		Overexpression		Knockout	

**Number of unique draws**

$$\begin{aligned} &= (\text{Number of replicates per category-condition group})^{(\text{Number of category-condition groups})} \\ &= 2^6 = \mathbf{64} \end{aligned}$$

One example draw:

Sample	1	3	5	7	9	11
Cell line	Cell line A			Cell line B		
Gene perturbation	WT	OE	KO	WT	OE	KO

# How to select category and condition

Category - Mouse strain  
Condition - Sex

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mouse strain	Strain A				Strain B				Strain C				Strain D			
Sex	Male	Female			Male		Female		Male			Female		Male		Female

$$\begin{aligned} &\text{Number of unique draws} \\ &= \prod (\text{Number of replicates per category-condition group}) \\ &= 1 \cdot 3 \cdot 2 \cdot 2 \cdot 3 \cdot 2 \cdot 2 \cdot 1 = \mathbf{144} \end{aligned}$$

One example draw:

Sample	1	2	5	7	9	12	14	16
Mouse strain	Strain A		Strain B		Strain C		Strain D	
Sex	Male	Female	Male	Female	Male	Female	Male	Female

Note: #1 and #16 will appear in every draw since there are no replicates for these groups