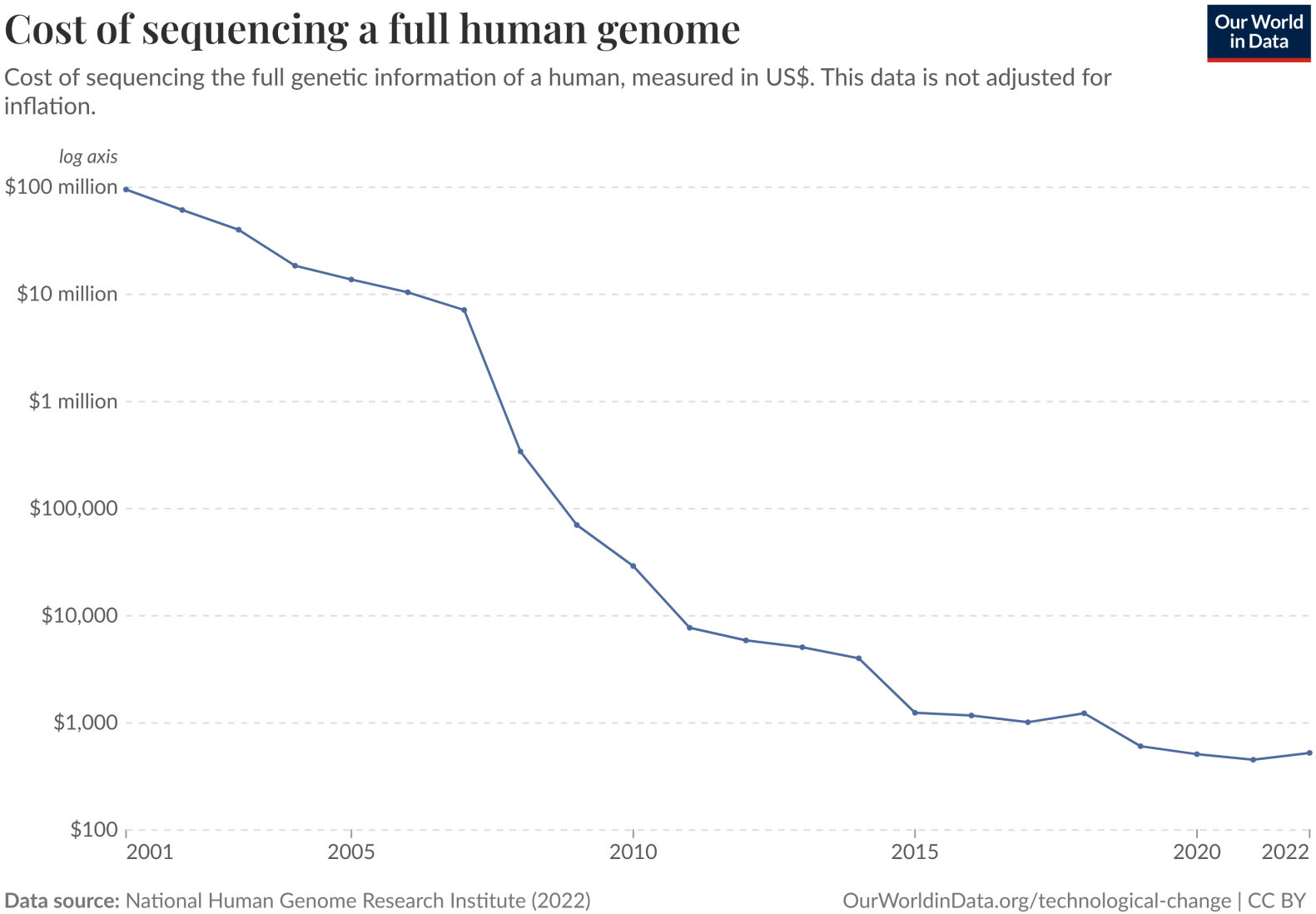
Temporal Structure in the Yeast Transcriptome

Period Discovery, Gene Clustering, and Semantic Annotation

Jackson Pond, Isaac Smith, Henry Tullis, Wyatt Wimmer

April 18, 2026

Motivation


Cost of sequencing a full human genome
Our World in Data
Cost of sequencing the full genetic information of a human, measured in US$. This data is not adjusted for inflation.
log axis
$100 million
$10 million
$1 million
$100,000
$10,000
$1,000
$100
2001
2005
2010
2015
2020
2022
Data source: National Human Genome Research Institute (2022)
OurWorldinData.org/technological-change | CC BY

Motivation

STANFORD BIO-X

WELCOME TO BIO-X

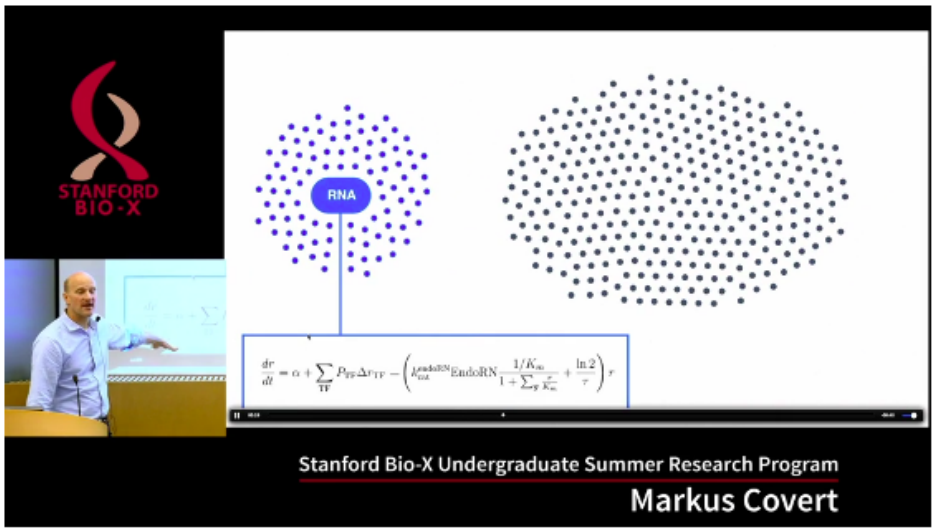
Sup

Search this

ABOUT | PEOPLE | RESEARCH | HIGHLIGHTS | VIDEOS | EVENTS | GE

Events

USRP 2023 - Saving Lives, Saving the Planet - Via Simulation and Computational Modeling



Saving Lives, Saving the Planet - Via Simulation and Computational Modeling

2023 STANFORD BIO-X USRP TALKS - WEDNESDAY, AUGUST 23, 2023

Events

- Upcoming Events
- Archive
- USRP 2023 - Savi Saving the Planet Simulation and C Modeling
- Talk Videos

COURSES AND WOR

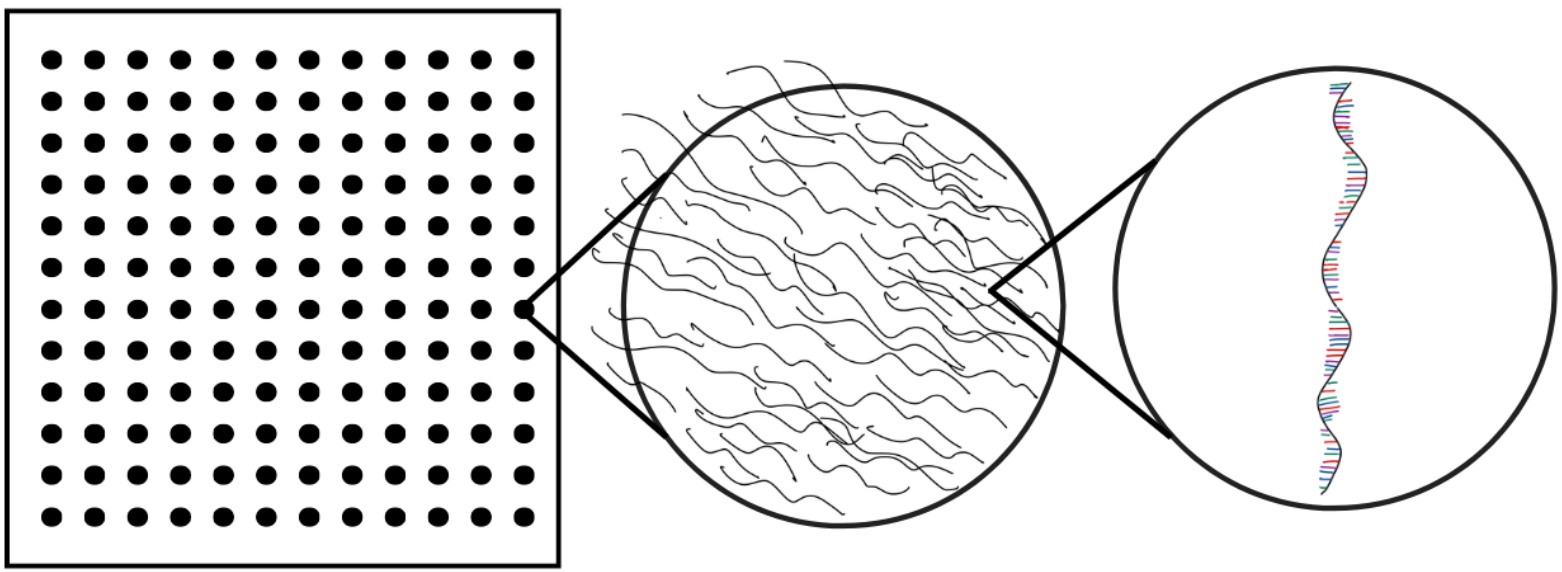
The Yeast Metabolic Cycle

- *Saccharomyces cerevisiae* (budding yeast) exhibits a spontaneous, synchronized **metabolic oscillation** when grown in continuous culture.
- Thousands of genes are expressed periodically in different phases
- Many genes are co-expressed in groups we call **transcriptional modules**

Two Central Questions

1. What is the period of the metabolic cycle, and what are its modules/phases?
2. Which genes drive each module/phase, and what biological processes do they represent?

Microarrays



Dataset

- **Source:** Tu et al. (2005), *Science* — publicly available via GEO (GSM77298).
- **Assay:** Yeast whole-transcriptome microarray.
- **Features:** 9,725 RNA transcripts.
- **Time points:** 36 samples at ≈25-minute intervals (≈15 hours total).
- Each time point is an independent microarray chip measuring relative transcript abundance across the *same* oscillating population.
- Author's analysis: preprocessing, maximum likelihood, sentinel clustering, $k = 3$ means

Preprocessing Pipeline

We applied three normalization steps before analysis.

1. **Per-chip median normalization.** Each chip is divided by its median intensity, removing global staining and loading differences between arrays.
   - Assumption: true median expression is constant across time.
   - Result: *relative* (not absolute) expression.
2. **arcsinh transform** Compresses the dynamic range (similar to $\log$, but handles zeros).

$$\tilde{x} = \operatorname{arcsinh}(x) = \ln\left(x + \sqrt{x^2 + 1}\right)$$

3. **Per-gene z-score** (for clustering only). Each gene's time series is standardized to zero mean and unit variance so genes with very different baseline expressions are comparable.

Approach: Clustering Time Points

Key Insight

If the metabolic cycle has period T, time points separated by T should look alike — they should fall in the *same cluster*.

Strategy: Cluster the 36 *time points* (not genes) using a Gaussian Mixture Model (GMM), then read off the period from the repeating cluster sequence.

Dimensionality reduction (necessary — $n = 36 \ll d = 9725$):

- Approach 1: Keep top-n highest-variance genes, sweep $n \in [200, 1100]$.
- Approach 2: Project all genes to 5 PCA components; select k by held-out log-likelihood (first 24 train, last 12 test).
- Both approaches recover $k = 5$ clusters.

Result: Period = 12 Time Steps, 300 seconds

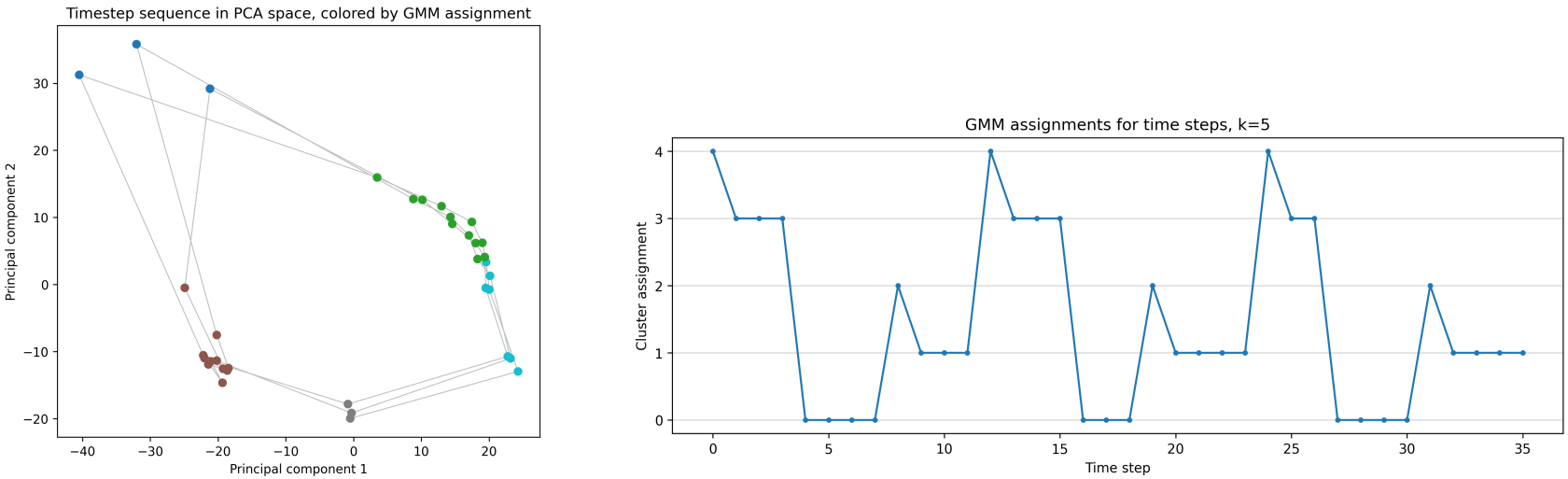


Figure: Left: 36 time points in PCA space, colored by GMM cluster. Right: Cluster assignment vs. time step, showing three complete repetitions of the 4-phase cycle.

Continuous-Density HMM Prediction

- A **Gaussian mixture HMM (CDHMM)** treats the hidden metabolic phase as a latent Markov chain and models observations with a GMM emission distribution.
- Hidden states correspond to the 4 metabolic phases recovered earlier.
- Trained on the first 24 time steps; evaluated on the final 12.
- Prediction: given observed gene expression, infer the current phase and project forward.

CDHMM plot

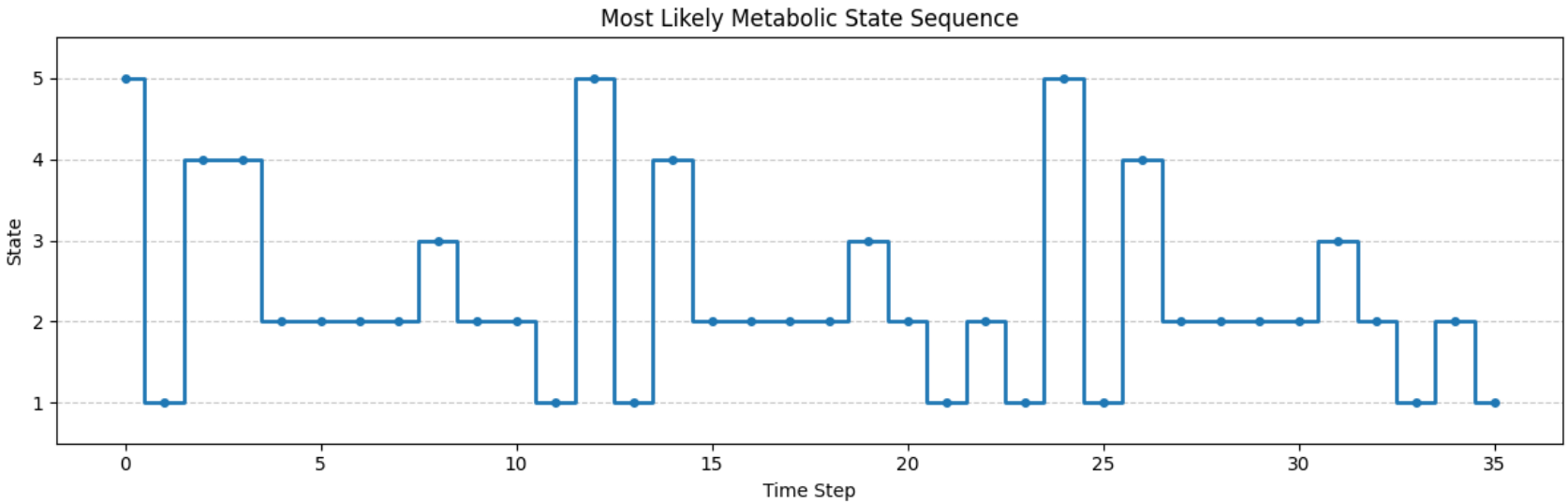


Figure: Most likely metabolic phase vs. time

Period Discovery — Summary

- Every dimensionality-reduction and GMM/CDHMM variant found a **12-timestep period**.

$$T = 12 \times 25\,\text{min} = 300\,\text{min}$$

- Cluster assignments are near-deterministic (high posterior probabilities), indicating well-separated phases.
- The 36-point experiment spans exactly **3 complete cycles**.
- Independent confirmation of the period reported by Tu et al.

Identifying Periodic Genes

With period $T = 12$ established, which genes actually oscillate?

Permutation test on lag-12 autocorrelation:

1. Compute ACF$(x_g, \tau = 12)$ for each gene g.
2. Shuffle gene g's time series 100 times; compute ACF for each shuffle.
3. Accept gene as periodic if its true ACF is at or above the **95th percentile** of the shuffled distribution.

Result: 5,243 of 9,725 genes ($\approx$54%) are identified as periodic at lag 12.

Periodic vs. Non-Periodic Genes in PCA

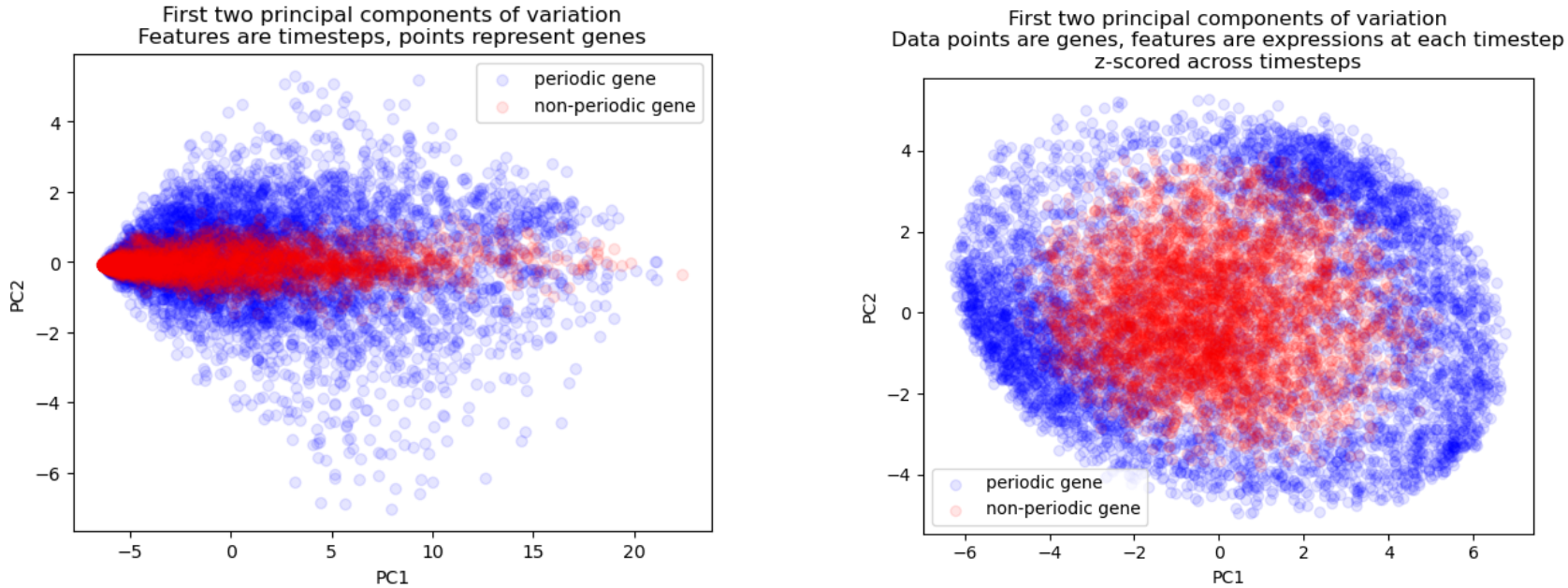


Figure: PCA of all 9,725 genes across 36 timesteps. Left: arcsinh expression (PC1 captures mean expression level). Right: After per-gene z-score normalization, periodic and non-periodic genes separate clearly in both PCs.

Classical Time Series Decomposition

For each periodic gene, decompose its standardized time series:

$$x_g(t) = \underbrace{m_g(t)}_{\text{trend}} + \underbrace{s_g(t)}_{\text{seasonal}} + \underbrace{\varepsilon_g(t)}_{\text{residual}}$$

Trend estimated via centered moving average (24 interior time points).

Clustering the seasonal components:

- Fit GMMs over $k = 2, \ldots, 12$; select k by AIC/BIC.
- $k = 8$ **minimizes BIC** and gives a local AIC minimum.
- 7 clusters show distinct temporal expression profiles; 1 cluster is a catch-all for noisy genes.

Temporal Expression Modules

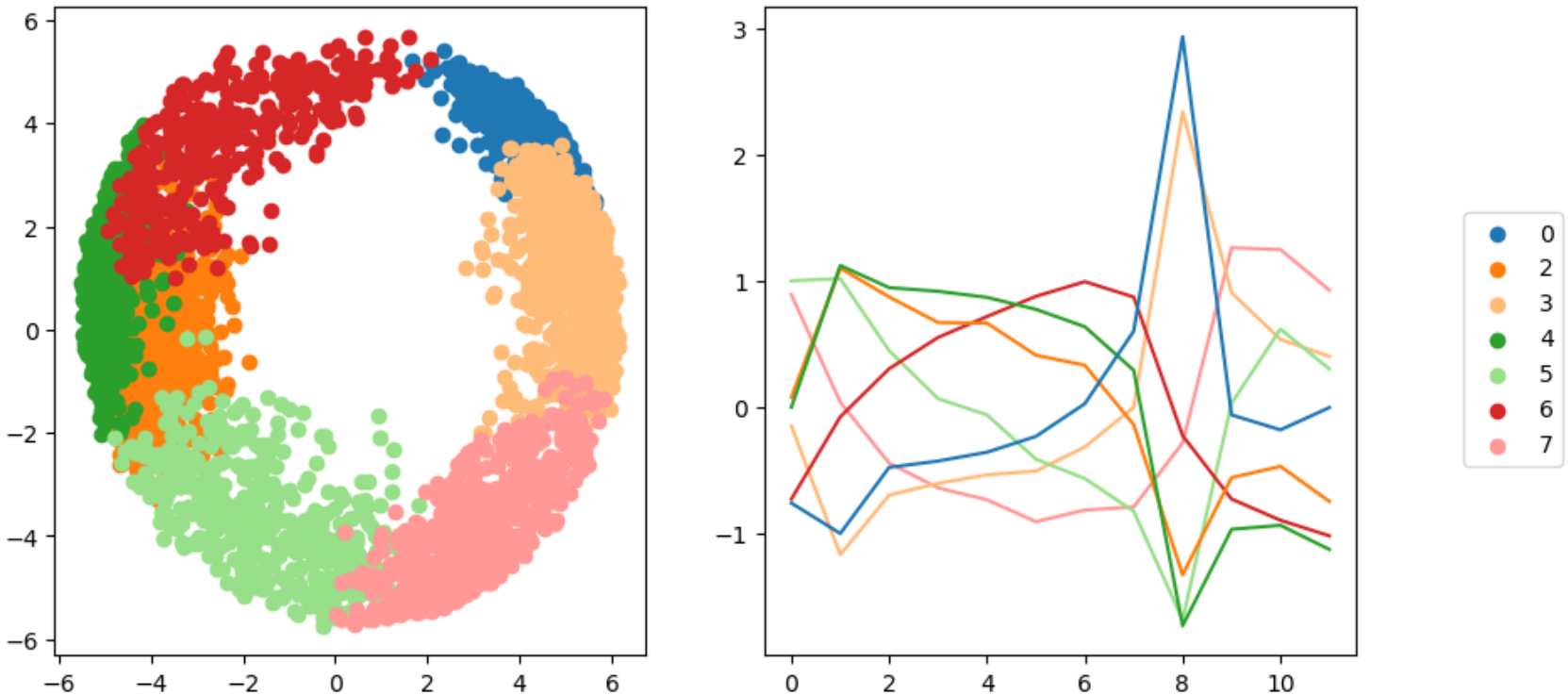


Figure: Left: UMAP of genes by seasonal component, colored by GMM cluster. Right: Mean seasonal profile for each of the 7 coherent clusters, showing peaks and troughs at distinct phases of the 12-step cycle.

ARMA Forecasting on Individual Genes

What we attempted:

- ARMA(p, q) and ARIMA(p, d, q) models fit to multivariate or univariate gene's time series.
- **Train:** first 24 time steps (2 full cycles). **Test:** last 12 time steps (1 held-out cycle).
- Order (p, q) selected per gene by AIC.

We attribute breakdown to periodicity - violates covariance stationary assumption

What Are the Gene Clusters *Doing*?

Clustering recovers *when* groups of genes are expressed. To understand *why*, we need biological annotation.

Approach: Latent Dirichlet Allocation (LDA) on gene-level metadata (GO terms, functional annotations).

- Each gene is a "document"; its annotations are "words".
- LDA discovers $K = 10$ latent **functional topics**.
- Topic weights per gene are then correlated with the time-series clusters to ask: *which biological processes dominate each temporal module?*

LDA Topics (labeled by Claude)

Topic	Short Name
1	Cytoplasmic Translation & Ribosome Structure
2	Cell Cycle & DNA Damage Checkpoint Kinases
3	Proteasome & ATP-Dependent Protein Degradation
4	RNA Pol II Transcription & Chromatin Remodeling
5	ER Membrane & Transmembrane Proteins
6	Mitotic/Meiotic Chromosome Segregation
7	Mitochondrial Metabolism & Oxidative Phosphorylation
8	Vesicular Transport & Endomembrane System
9	Retrotransposon Biology (Ty Elements)
10	Ribosome Biogenesis & rRNA Processing

Topic–Cluster Associations

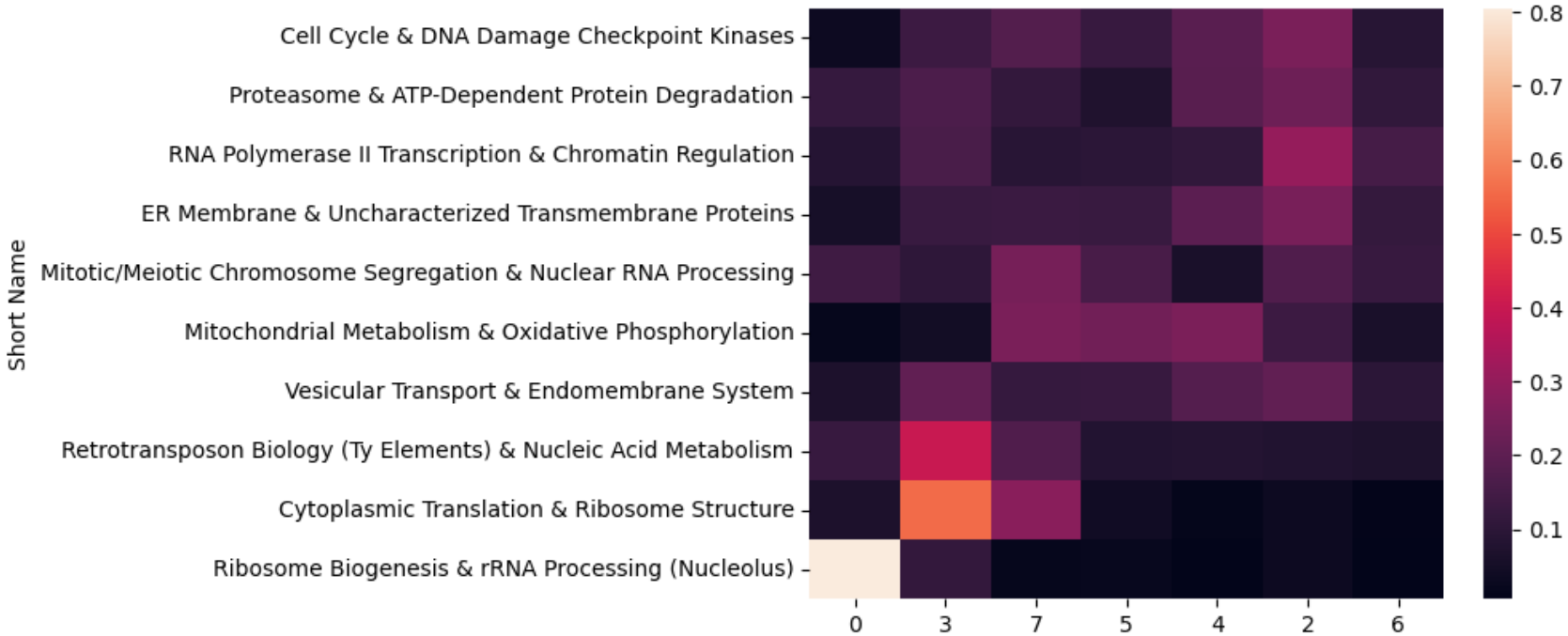
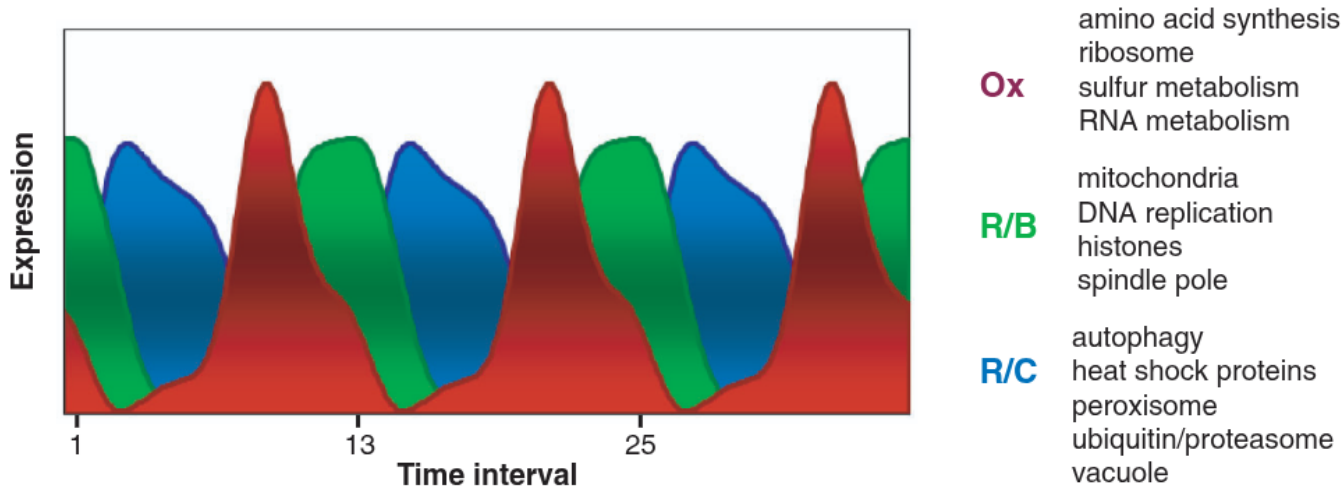
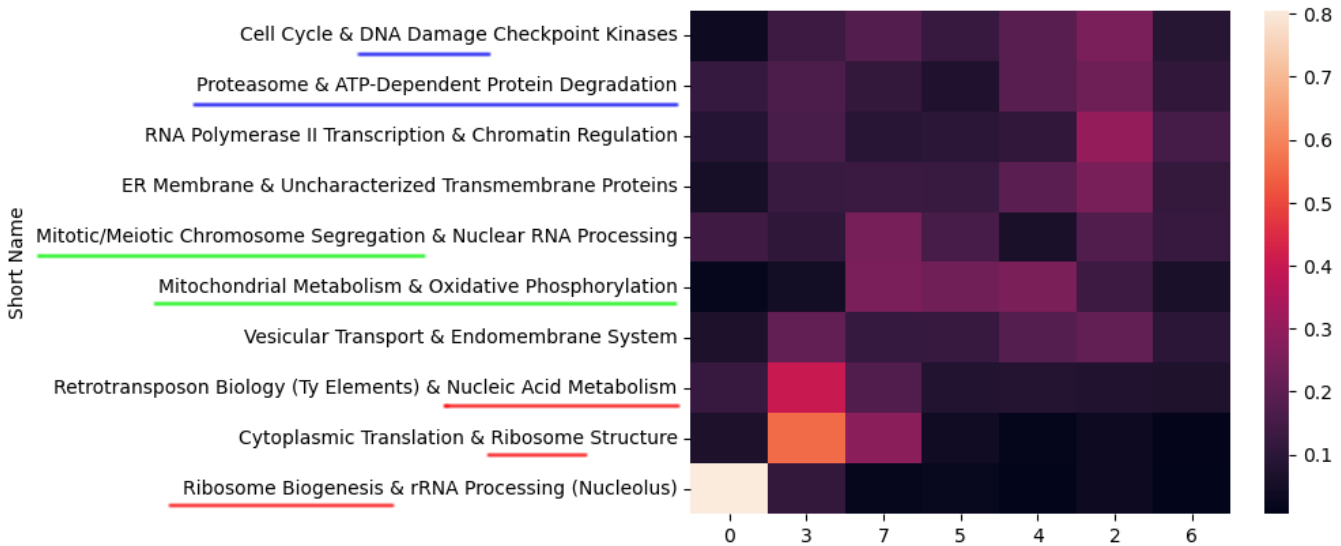


Figure: Heatmap of LDA topic weight vs. time-series cluster. Columns ordered temporally by phase of the cluster cycle's peak. Each cell shows the fraction of genes in a cluster assigned to a given functional topic, revealing which biological processes are activated in each temporal module.

Key Biological Findings

Combining the temporal clusters with LDA topic weights confirms and extends Tu et al.'s interpretation of phases:





Summary

1. **Period:** GMM clustering of time points robustly recovers $T = 12$ steps $= 300$ minutes across all dimensionality-reduction strategies.
2. **Periodic genes:** 5,243/9,725 genes ($\approx$54%) oscillate with this period (permutation test, $\alpha = 0.05$).
3. **Temporal modules:** Seasonal decomposition + GMM identifies 7 coherent clusters of co-expressed genes, each with a distinct phase profile.
4. **Forecasting:** CDHMM models capture the periodic dynamics and predict held-out expression.
5. **Function:** LDA links each temporal module to specific biological processes — oxidative metabolism, translation, replication, degradation — giving a functional portrait of the yeast metabolic cycle.

Thank You

Questions?

Data: Tu et al. (2005) *Science*, GEO accession GSM77298.
Code: `https://github.com/jackdpond/hmm_gene_model`

Backup: GMM Model Selection

- Held-out log-likelihood evaluated over 100 random restarts for $k \in \{2, \ldots, 7\}$.
- $k = 5$ consistently maximized held-out likelihood.
- Tied covariance used (full covariance is ill-conditioned for $n = 24$, $d = 5$).

Backup: Permutation Test Details

For each gene g with time series $\mathbf{x}_g$:

$$r_g = \frac{1}{n-\tau} \sum_{t=1}^{n-\tau} \frac{(x_g(t) - \bar{x}_g)(x_g(t+\tau) - \bar{x}_g)}{s_g^2}, \quad \tau = 12$$

Null distribution: 100 random permutations of $\mathbf{x}_g$.
Reject null if $r_g \geq$ 95th percentile of permuted $r_g^{(b)}$.

Backup: LDA Generative Model

For each gene g:

1. Draw topic proportions $\boldsymbol{\theta}_g \sim \text{Dir}(\alpha)$.
2. For each annotation term w:
   - 2.1 Draw topic $z \sim \text{Categorical}(\boldsymbol{\theta}_g)$.
   - 2.2 Draw term $w \sim \text{Categorical}(\boldsymbol{\phi}_z)$.

Inference via EM; $K = 10$ topics manually selected. Topic names assigned by prompting Claude with top-20 terms per topic.