R programming project

Gene clustering based on their expression profiles

Gaëlle Lelandais

gaelle.lelandais@universite-paris-saclay.fr

The file « Mito_Genes.txt »

Mito_Genes.txt

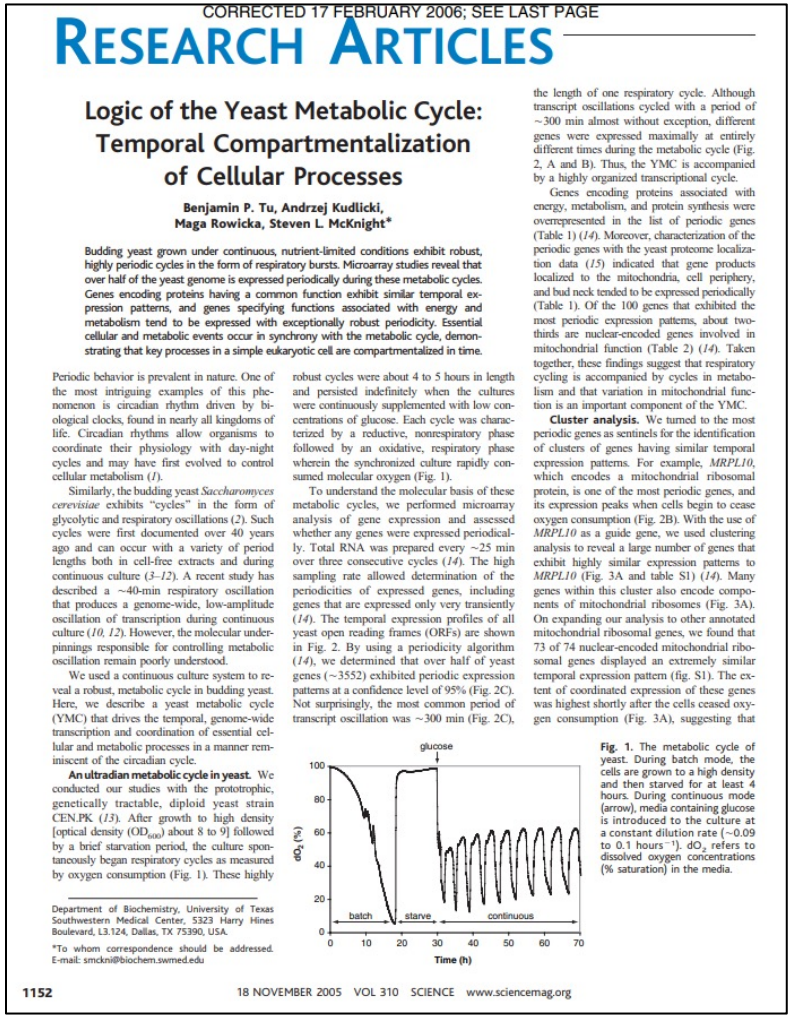
	GENE_ID	GSM77298_T1	GSM77299_T2	GSM77300_T3	GSM77301_T4	GSM77302_T5	GSM77303_T6	GSM77304_T7	GSM77305_T8	GSM77
2	YGR234W|YHB1	33.102	16.733	14.173	13.126	11.343	12.259	13.148	16.114	31.631	33.629	38.644	26.5
3	YEL039C|CYC7	1.23	2.423	1.286	1.631	1.077	1.885	1.866	1.155	0.356	0.934	0.437	0.57
4	YGL187C|COX4	29.48	26.334	22.992	20.055	22.066	18.979	19.251	17.684	8.393	23.196	25.08	22.58
5	YGL191W|COX13	23.448	21.726	22.306	17.664	19.42	14.673	15.564	13.21	8.592	17.213	21.731	17.05
6	YHR051W|COX6	29.857	26.911	26.908	20.949	18.877	17.242	17.073	16.289	11.238	26.492	33.611	24.17
7	YLR038C|COX12	40.713	48.95	51.821	46.777	40.616	31.031	36.191	30.668	24.632	38.945	43.975	39.17
8	YLR395C|COX8	32.766	35.924	43.495	38.335	29.136	27.276	27.436	25.353	22.413	31.165	36.876	33.14
9	YMR256C|COX7	22.186	32.672	43.094	35.976	29.4	25.574	21.982	22.437	22.546	23.559	32.234	32.03
10	YNL052W|COX5A	33.116	22.792	25.244	21.791	20.048	16.707	19.713	19.473	14.695	26.22	27.15	30.61
11	YKR016W|AIM28	5.295	2.892	2.637	2.11	1.604	1.244	1.53	1.573	1.598	2.981	2.92	4.813
12	YNR020C|ATP23	3.107	1.29	0.773	0.73	0.627	0.65	0.788	0.859	0.909	3.184	3.404	3.547
13	YLR201C|COQ9	5.837	7.005	8.955	6.286	5.82	3.836	5.458	4.07	2.23	3.165	5.455	4.051
14	YOR222W|ODC2	2.949	1.375	2.262	2.253	2.692	3.186	3.194	3.8	7.516	4.423	3.829	3.675
15	YGR231C|PHB2	13.843	13.092	7.948	8.232	8.535	8.473	8.412	7.981	3.653	13.012	14.927	10.05
16	YGL219C|MDM34	3.236	3.929	3.499	2.642	2.545	2.602	2.528	1.934	1.424	2.74	3.103	2.498
17	YGR132C|PHB1	8.647	8.854	6.591	5.346	4.214	4.737	4.974	4.276	1.586	6.244	7.402	5.961
18	YJL066C|MPM1	10.074	15.177	8.112	8.271	7.274	5.62	6.58	5.367	0.995	5.848	9.271	5.539
19	YBR003W|COQ1	4.555	2.164	1.661	1.445	1.074	1.057	1.111	1.393	0.437	2.798	3.501	2.651
20	YNR041C|COQ2	2.995	2.834	2.635	1.798	0.84	1.142	1.174	1.385	1.702	1.867	1.397	2.228
21	YPL109C|NA	4.386	6.238	7.104	6.627	5.71	5.81	5.489	3.905	0.625	2.311	2.842	2.686
22	YGL119W|ABC1	0.926	0.649	0.413	0.485	0.571	0.55	0.545	0.697	0.556	0.833	1.121	0.966
23	YGR255C|COQ6	4.404	3.328	3.533	3.228	2.99	2.838	2.762	2.296	0.708	2.702	3.326	2.166
24	YLR056W|ERG3	10.98	3.135	4.723	5.367	7.29	8.694	13.834	17.927	16.873	19.849	17.174	15.40
25	YOL096C|COQ3	3.329	3.726	2.975	2.508	1.796	2.228	2.379	2.329	0.43	2.336	2.65	2.12
26	YDR204W|COQ4	3.974	7.12	4.956	5.768	6.764	5.777	6.125	4.541	0.858	2.619	1.452	1.453

Normal text file | length : 147 182 lines : 628 | Ln : 2 Col : 1 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS

EN - Software demo.R

```
5  # Source of data:
6  # The dataset are presented in Tu et al. (2005)
7
8  # Data reading
9  expData = read.table("Mito_Genes.txt", header = T, row.names = 1)
10
11 # Get number of genes and number of experiments
12 nrow(expData)
13 ncol(expData)
```

Original publication



CORRECTED 17 FEBRUARY 2006; SEE LAST PAGE

RESEARCH ARTICLES

Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes

Benjamin P. Tu, Andrzej Kudlicki, Maga Rowicka, Steven L. McKnight*

Budding yeast grown under continuous, nutrient-limited conditions exhibit robust, highly periodic cycles in the form of respiratory bursts. Microarray studies reveal that over half of the yeast genome is expressed periodically during these metabolic cycles. Genes encoding proteins having a common function exhibit similar temporal expression patterns, and genes specifying functions associated with energy and metabolism tend to be expressed with exceptionally robust periodicity. Essential cellular and metabolic events occur in synchrony with the metabolic cycle, demonstrating that key processes in a simple eukaryotic cell are compartmentalized in time.

Periodic behavior is prevalent in nature. One of the most intriguing examples of this phenomenon is circadian rhythm driven by biological clocks, found in nearly all kingdoms of life. Circadian rhythms allow organisms to coordinate their physiology with day-night cycles and may have first evolved to control cellular metabolism (*1*).

Similarly, the budding yeast *Saccharomyces cerevisiae* exhibits "cycles" in the form of glycolytic and respiratory oscillations (*2*). Such cycles were first documented over 40 years ago and can occur with a variety of period lengths both in cell-free extracts and during continuous culture (*3–12*). A recent study has described a ~40-min respiratory oscillation that produces a genome-wide, low-amplitude oscillation of transcription during continuous culture (*10, 12*). However, the molecular underpinnings responsible for controlling metabolic oscillation remain poorly understood.

We used a continuous culture system to reveal a robust, metabolic cycle in budding yeast. Here, we describe a yeast metabolic cycle (YMC) that drives the temporal, genome-wide transcription and coordination of essential cellular and metabolic processes in a manner reminiscent of the circadian cycle.

An ultradian metabolic cycle in yeast. We conducted our studies with the prototrophic, genetically tractable, diploid yeast strain CEN.PK (*13*). After growth to high density [optical density (OD_{600}) about 8 to 9] followed by a brief starvation period, the culture spontaneously began respiratory cycles as measured by oxygen consumption (Fig. 1). These highly robust cycles were about 4 to 5 hours in length and persisted indefinitely when the cultures were continuously supplemented with low concentrations of glucose. Each cycle was characterized by a reductive, nonrespiratory phase followed by an oxidative, respiratory phase wherein the synchronized culture rapidly consumed molecular oxygen (Fig. 1).

To understand the molecular basis of these metabolic cycles, we performed microarray analysis of gene expression and assessed whether any genes were expressed periodically. Total RNA was prepared every ~25 min over three consecutive cycles (*14*). The high sampling rate allowed determination of the periodicities of expressed genes, including genes that are expressed only very transiently (*14*). The temporal expression profiles of all yeast open reading frames (ORFs) are shown in Fig. 2. By using a periodicity algorithm (*14*), we determined that over half of yeast genes (~3552) exhibited periodic expression patterns at a confidence level of 95% (Fig. 2C). Not surprisingly, the most common period of transcript oscillation was ~300 min (Fig. 2C), the length of one respiratory cycle. Although transcript oscillations cycled with a period of ~300 min almost without exception, different genes were expressed maximally at entirely different times during the metabolic cycle (Fig. 2, A and B). Thus, the YMC is accompanied by a highly organized transcriptional cycle.

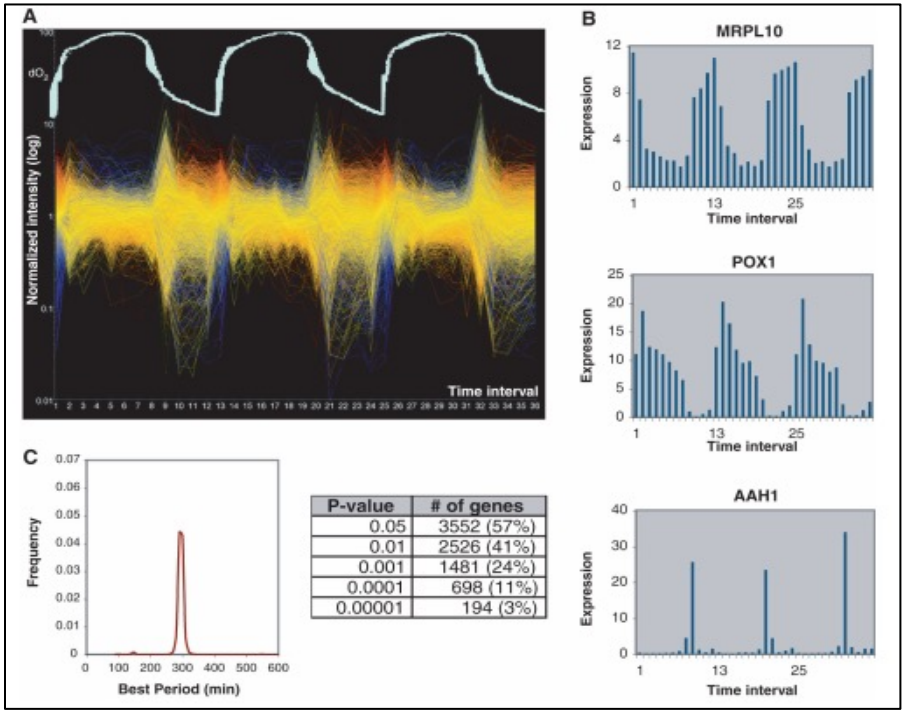
Genes encoding proteins associated with energy, metabolism, and protein synthesis were overrepresented in the list of periodic genes (Table 1) (*14*). Moreover, characterization of the periodic genes with the yeast proteome localization data (*15*) indicated that gene products localized to the mitochondria, cell periphery, and bud neck tended to be expressed periodically (Table 1). Of the 100 genes that exhibited the most periodic expression patterns, about two-thirds are nuclear-encoded genes involved in mitochondrial function (Table 2) (*14*). Taken together, these findings suggest that respiratory cycling is accompanied by cycles in metabolism and that variation in mitochondrial function is an important component of the YMC.

Cluster analysis. We turned to the most periodic genes as sentinels for the identification of clusters of genes having similar temporal expression patterns. For example, *MRPL10*, which encodes a mitochondrial ribosomal protein, is one of the most periodic genes, and its expression peaks when cells begin to cease oxygen consumption (Fig. 2B). With the use of *MRPL10* as a guide gene, we used clustering analysis to reveal a large number of genes that exhibit highly similar expression patterns to *MRPL10* (Fig. 3A and table S1) (*14*). Many genes within this cluster also encode components of mitochondrial ribosomes (Fig. 3A). On expanding our analysis to other annotated mitochondrial ribosomal genes, we found that 73 of 74 nuclear-encoded mitochondrial ribosomal genes displayed an extremely similar temporal expression pattern (fig. S1). The extent of coordinated expression of these genes was highest shortly after the cells ceased oxygen consumption (Fig. 3A), suggesting that

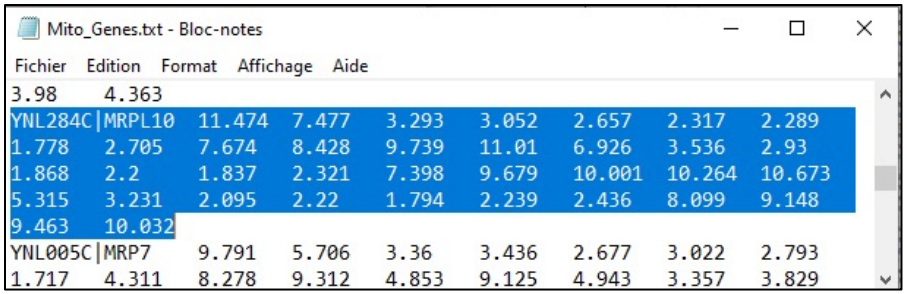
Fig. 1. The metabolic cycle of yeast. During batch mode, the cells are grown to a high density and then starved for at least 4 hours. During continuous mode (arrow), media containing glucose is introduced to the culture at a constant dilution rate (~0.09 to 0.1 $hours^{-1}$). dO_2 refers to dissolved oxygen concentrations (% saturation) in the media.

Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, L3.124, Dallas, TX 75390, USA.

*To whom correspondence should be addressed. E-mail: smckni@biochem.swmed.edu



P-value	# of genes
0.05	3552 (57%)
0.01	2526 (41%)
0.001	1481 (24%)
0.0001	698 (11%)
0.00001	194 (3%)



Aim of the project

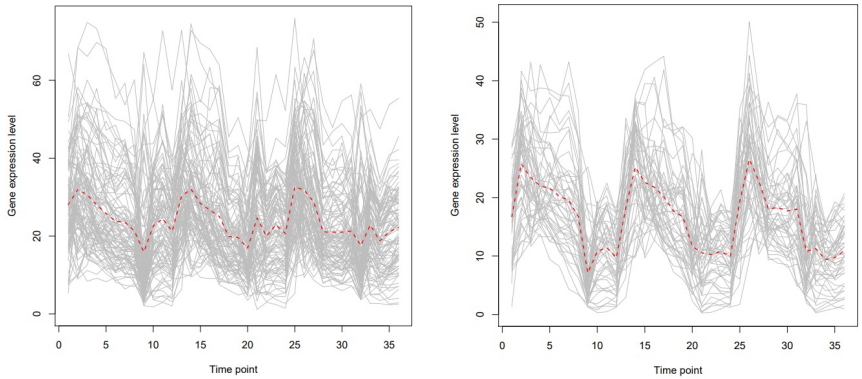
Write a script that groups the genes according to their expression profiles

Input data file → Output data files

Mito_Genes.txt

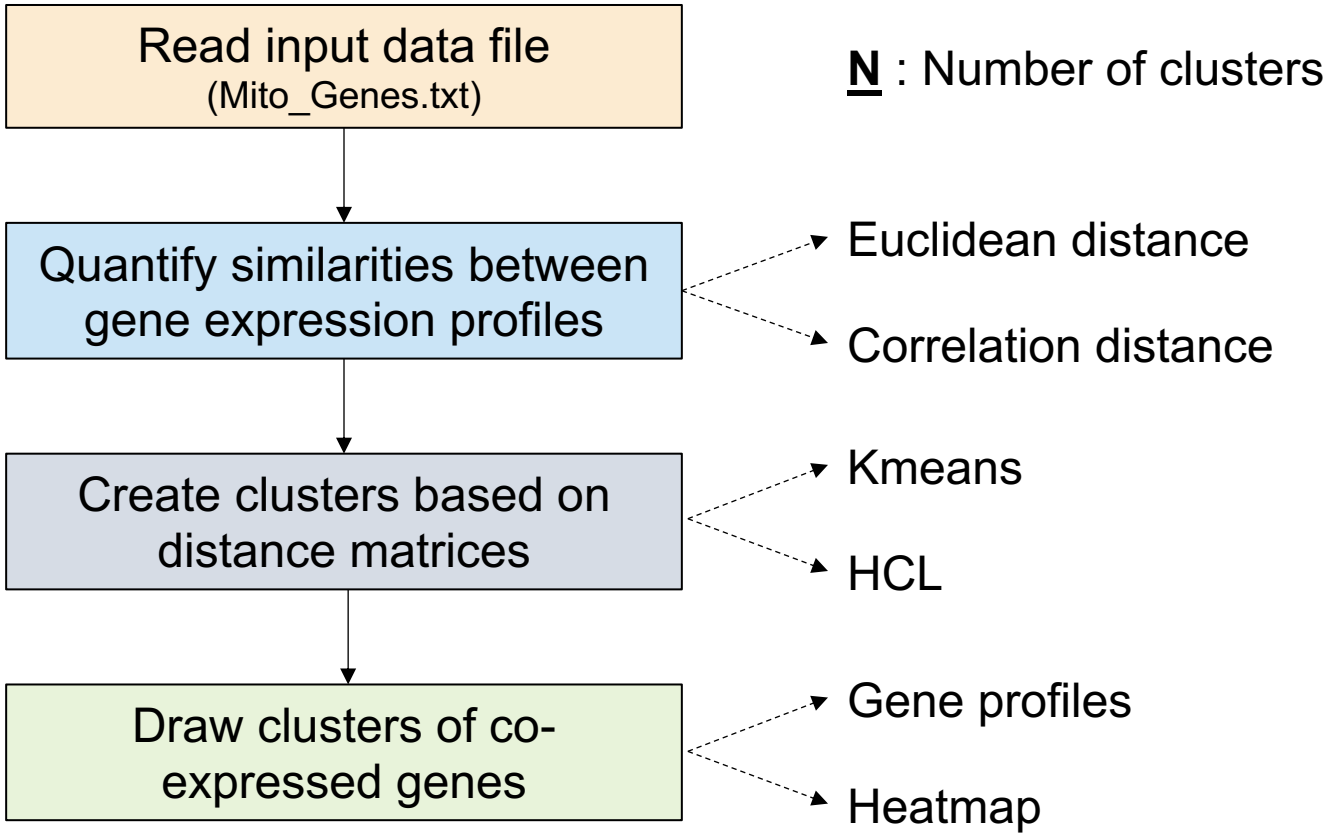
	GENE_ID	GSM77298_T1	GSM77299_T2	GSM77300_T3	GSM77301_T4	GSM77302_T5	GSM77303_T6	GSM77304_T7	GSM77305_T8	GSM7
2	YGR234W|YHB1	33.102	16.733	14.173	13.126	11.343	12.259	13.148	16.114	31.631	33.629	38.644	26.5
3	YEL039C|CYC7	1.23	2.423	1.286	1.631	1.077	1.885	1.866	1.155	0.356	0.934	0.437	0.57
4	YGL187C|COX4	29.48	26.334	22.992	20.055	22.066	18.979	19.251	17.684	8.393	23.196	25.08	22.5
5	YGL191W|COX13	23.448	21.726	22.306	17.664	19.42	14.673	15.564	13.21	8.592	17.213	21.731	17.0
6	YHR051W|COX6	29.857	26.911	26.908	20.949	18.877	17.242	17.073	16.289	11.238	26.492	33.611	24.1
7	YLR038C|COX12	40.713	48.95	51.821	46.777	40.616	31.031	36.191	30.668	24.632	38.945	43.975	39.1
8	YLR395C|COX8	32.766	35.924	43.495	38.335	29.136	27.276	27.436	25.353	22.413	31.165	36.876	33.1
9	YMR256C|COX7	22.186	32.672	43.094	35.976	29.4	25.574	21.982	22.437	22.546	23.559	32.234	32.0
10	YNL052W|COX5A	33.116	22.792	25.244	21.791	20.048	16.707	19.713	19.473	14.695	26.22	27.15	30.6
11	YKR016W|AIM28	5.295	2.892	2.637	2.11	1.604	1.244	1.53	1.573	1.598	2.981	2.92	4.81
12	YNR020C|ATP23	3.107	1.29	0.773	0.73	0.627	0.65	0.788	0.859	0.909	3.184	3.404	3.54
13	YLR201C|COQ9	5.837	7.005	8.955	6.286	5.82	3.836	5.458	4.07	2.23	3.165	5.455	4.05
14	YOR222W|ODC2	2.949	1.375	2.262	2.253	2.692	3.186	3.194	3.8	7.516	4.423	3.829	3.675
15	YGR231C|PHB2	13.843	13.092	7.948	8.232	8.535	8.473	8.412	7.981	3.653	13.012	14.927	10.0
16	YGL219C|MDM34	3.236	3.929	3.499	2.642	2.545	2.602	2.528	1.934	1.424	2.74	3.103	2.49
17	YGR132C|PHB1	8.647	8.854	6.591	5.346	4.214	4.737	4.974	4.276	1.586	6.244	7.402	5.96
18	YJL066C|MPM1	10.074	15.177	8.112	8.271	7.274	5.62	6.58	5.367	0.995	5.848	9.271	5.53
19	YBR003W|COQ1	4.555	2.164	1.661	1.445	1.074	1.057	1.111	1.393	0.437	2.798	3.501	2.65
20	YNR041C|COQ2	2.995	2.834	2.635	1.798	0.84	1.142	1.174	1.385	1.702	1.867	1.397	2.22
21	YPL109C|NA	4.386	6.238	7.104	6.627	5.71	5.81	5.489	3.905	0.625	2.311	2.842	2.686	
22	YGL119W|ABC1	0.926	0.649	0.413	0.485	0.571	0.55	0.545	0.697	0.556	0.833	1.121	0.96
23	YGR255C|COQ6	4.404	3.328	3.533	3.228	2.99	2.838	2.762	2.296	0.708	2.702	3.326	2.16
24	YLR056W|ERG3	10.98	3.135	4.723	5.367	7.29	8.694	13.834	17.927	16.873	19.849	17.174	15.4
25	YOL096C|COQ3	3.329	3.726	2.975	2.508	1.796	2.228	2.379	2.329	0.43	2.336	2.65	2.12
26	YDR204W|COQ4	3.974	7.12	4.956	5.768	6.764	5.777	6.125	4.541	0.858	2.619	1.452	1.45

Normal text file | length : 147 182 | lines : 628 | Ln : 2 Col : 1 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS

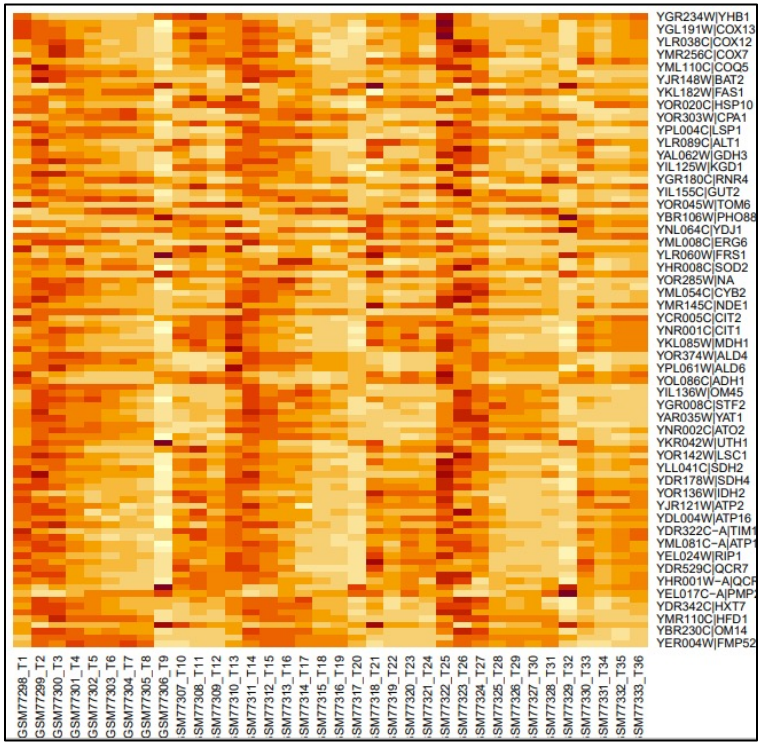


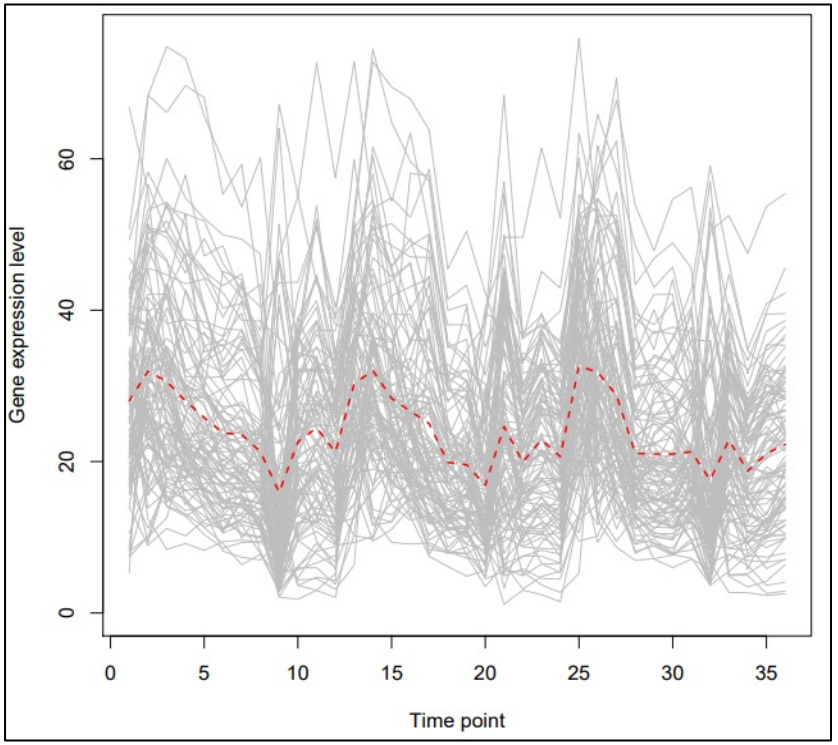
A set of gene clusters, obtained with different algorithms (HCL, kmeans, etc.) and several distances (Euclidian, correlation, etc.)

Functionalities to be coded



Heatmap and/or GeneProfiles





List of R functions to help you

plot( )	heatmap( )	as.matrix( )
read.table( )	paste( )	as.dist( )
lines( )	kmeans( )	hclust( )
print( )	dist( )	Others …
cutree( )	cor( )	

Increasing the complexity, step by step

Level 0 < Level 1 < Level 2 < Level 3

- Four clusters
- Kmeans & Euclidean
- Gene profiles
- HCL & Euclidean
- Gene profiles
- Kmeans & Correlation
- Gene profiles
- HCL & Correlation
- Gene profiles

- N clusters
- Kmeans & Euclidean
- Gene profiles
- HCL & Euclidean
- Gene profiles
- Kmeans & Correlation
- Gene profiles
- HCL & Correlation
- Gene profiles

- N clusters
- Choose distance
- Choose algotithm

- Create functions
- Create main
- Run clustering with different parameters

script_clustering_level0.R | script_clustering_level1.R | script_clustering_level2.R | script_clustering_level3.R