



R programming project

Gene clustering based on their expression profiles

Gaëlle Lelandais

gaelle.lelandais@universite-paris-saclay.fr

The file « Mito_Genes.txt »

	Gene_ID	GSM77298_T1	GSM77299_T2	GSM77300_T3	GSM77301_T4	GSM77302_T5	GSM77303_T6	GSM77304_T7	GSM77305_T8	GSM77306_T9	GSM77307_T10	GSM77308_T11	GSM77309_T12	GSM77310_T13	GSM77311_T14	GSM77312_T15	GSM77313_T16	GSM77314_T17	GSM77315_T18	GSM77316_T19	GSM77317_T20	GSM77318_T21	GSM77319_T22	GSM77320_T23	GSM77321_T24	GSM77322_T25	GSM77323_T26	GSM77324_T27	GSM77325_T28	GSM77326_T29	GSM77327_T30	GSM77328_T31	GSM77329_T32	GSM77330_T33	GSM77331_T34	GSM77332_T35	GSM77333_T36	GSM77334_T37	GSM77335_T38	GSM77336_T39	GSM77337_T40	GSM77338_T41	GSM77339_T42	GSM77340_T43	GSM77341_T44	GSM77342_T45	GSM77343_T46	GSM77344_T47	GSM77345_T48	GSM77346_T49	GSM77347_T50	GSM77348_T51	GSM77349_T52	GSM77350_T53	GSM77351_T54	GSM77352_T55	GSM77353_T56	GSM77354_T57	GSM77355_T58	GSM77356_T59	GSM77357_T60	GSM77358_T61	GSM77359_T62	GSM77360_T63	GSM77361_T64	GSM77362_T65	GSM77363_T66	GSM77364_T67	GSM77365_T68	GSM77366_T69	GSM77367_T70	GSM77368_T71	GSM77369_T72	GSM77370_T73	GSM77371_T74	GSM77372_T75	GSM77373_T76	GSM77374_T77	GSM77375_T78	GSM77376_T79	GSM77377_T80	GSM77378_T81	GSM77379_T82	GSM77380_T83	GSM77381_T84	GSM77382_T85	GSM77383_T86	GSM77384_T87	GSM77385_T88	GSM77386_T89	GSM77387_T90	GSM77388_T91	GSM77389_T92	GSM77390_T93	GSM77391_T94	GSM77392_T95	GSM77393_T96	GSM77394_T97	GSM77395_T98	GSM77396_T99	GSM77397_T100	GSM77398_T101	GSM77399_T102	GSM77400_T103	GSM77401_T104	GSM77402_T105	GSM77403_T106	GSM77404_T107	GSM77405_T108	GSM77406_T109	GSM77407_T110	GSM77408_T111	GSM77409_T112	GSM77410_T113	GSM77411_T114	GSM77412_T115	GSM77413_T116	GSM77414_T117	GSM77415_T118	GSM77416_T119	GSM77417_T120	GSM77418_T121	GSM77419_T122	GSM77420_T123	GSM77421_T124	GSM77422_T125	GSM77423_T126	GSM77424_T127	GSM77425_T128	GSM77426_T129	GSM77427_T130	GSM77428_T131	GSM77429_T132	GSM77430_T133	GSM77431_T134	GSM77432_T135	GSM77433_T136	GSM77434_T137	GSM77435_T138	GSM77436_T139	GSM77437_T140	GSM77438_T141	GSM77439_T142	GSM77440_T143	GSM77441_T144	GSM77442_T145	GSM77443_T146	GSM77444_T147	GSM77445_T148	GSM77446_T149	GSM77447_T150	GSM77448_T151	GSM77449_T152	GSM77450_T153	GSM77451_T154	GSM77452_T155	GSM77453_T156	GSM77454_T157	GSM77455_T158	GSM77456_T159	GSM77457_T160	GSM77458_T161	GSM77459_T162	GSM77460_T163	GSM77461_T164	GSM77462_T165	GSM77463_T166	GSM77464_T167	GSM77465_T168	GSM77466_T169	GSM77467_T170	GSM77468_T171	GSM77469_T172	GSM77470_T173	GSM77471_T174	GSM77472_T175	GSM77473_T176	GSM77474_T177	GSM77475_T178	GSM77476_T179	GSM77477_T180	GSM77478_T181	GSM77479_T182	GSM77480_T183	GSM77481_T184	GSM77482_T185	GSM77483_T186	GSM77484_T187	GSM77485_T188	GSM77486_T189	GSM77487_T190	GSM77488_T191	GSM77489_T192	GSM77490_T193	GSM77491_T194	GSM77492_T195	GSM77493_T196	GSM77494_T197	GSM77495_T198	GSM77496_T199	GSM77497_T200	GSM77498_T201	GSM77499_T202	GSM77500_T203	GSM77501_T204	GSM77502_T205	GSM77503_T206	GSM77504_T207	GSM77505_T208	GSM77506_T209	GSM77507_T210	GSM77508_T211	GSM77509_T212	GSM77510_T213	GSM77511_T214	GSM77512_T215	GSM77513_T216	GSM77514_T217	GSM77515_T218	GSM77516_T219	GSM77517_T220	GSM77518_T221	GSM77519_T222	GSM77520_T223	GSM77521_T224	GSM77522_T225	GSM77523_T226	GSM77524_T227	GSM77525_T228	GSM77526_T229	GSM77527_T230	GSM77528_T231	GSM77529_T232	GSM77530_T233	GSM77531_T234	GSM77532_T235	GSM77533_T236	GSM77534_T237	GSM77535_T238	GSM77536_T239	GSM77537_T240	GSM77538_T241	GSM77539_T242	GSM77540_T243	GSM77541_T244	GSM77542_T245	GSM77543_T246	GSM77544_T247	GSM77545_T248	GSM77546_T249	GSM77547_T250	GSM77548_T251	GSM77549_T252	GSM77550_T253	GSM77551_T254	GSM77552_T255	GSM77553_T256	GSM77554_T257	GSM77555_T258	GSM77556_T259	GSM77557_T260	GSM77558_T261	GSM77559_T262	GSM77560_T263	GSM77561_T264	GSM77562_T265	GSM77563_T266	GSM77564_T267	GSM77565_T268	GSM77566_T269	GSM77567_T270	GSM77568_T271	GSM77569_T272	GSM77570_T273	GSM77571_T274	GSM77572_T275	GSM77573_T276	GSM77574_T277	GSM77575_T278	GSM77576_T27
--	---------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	--------------



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 prototypic, 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

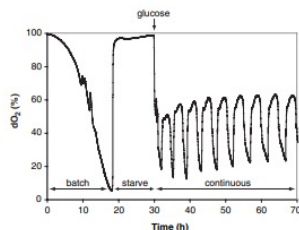


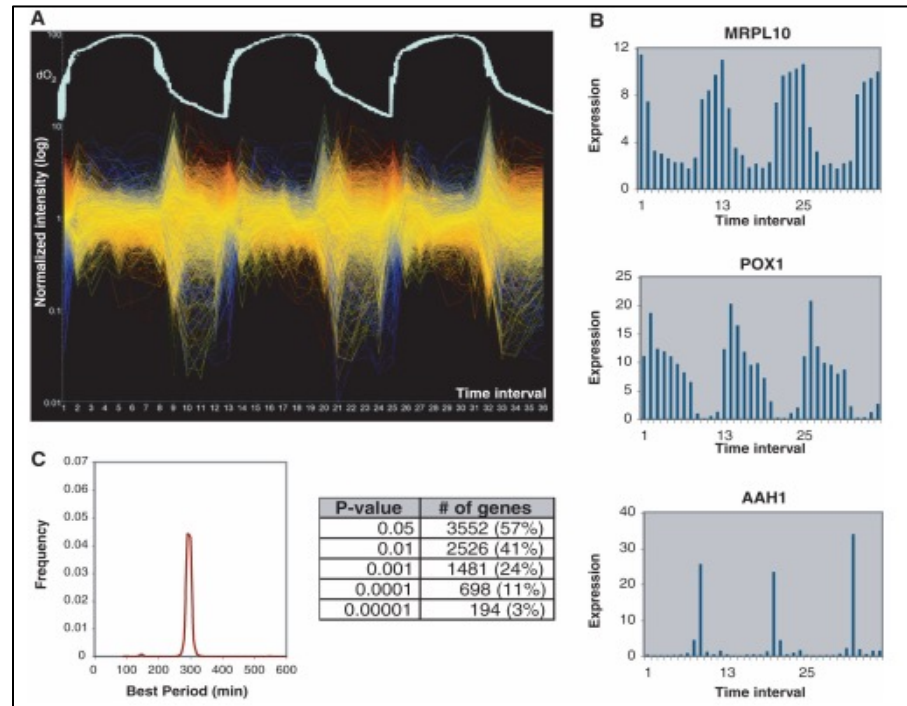
Fig. 1. The metabolic cycle of yeast. During batch mode, the cells are grown to a high density and then starved for at least 4 hours. During continuous mode (arrow), media containing glucose is introduced to the culture at a constant dilution rate (~0.09 to 0.1 h⁻¹). dO_2 refers to dissolved oxygen concentrations (% 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: smcknight@biochem.utswmed.edu

1152

18 NOVEMBER 2005 VOL 310 SCIENCE www.sciencemag.org



Mito_Genes.txt - Bloc-notes

FichierEditionFormatAffichageAide

3.98	4.363							
YNL284C	MRPL10	11.474	7.477	3.293	3.052	2.657	2.317	2.289
1.778	2.705	7.674	8.428	9.739	11.01	6.926	3.536	2.93
1.868	2.2	1.837	2.321	7.398	9.679	10.001	10.264	10.673
5.315	3.231	2.095	2.22	1.794	2.239	2.436	8.099	9.148
9.463	10.032							
YNL005C	MRP7	9.791	5.706	3.36	3.436	2.677	3.022	2.793
1.717	4.311	8.278	9.312	4.853	9.125	4.943	3.357	3.829

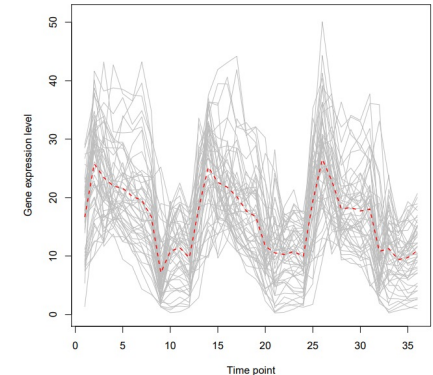
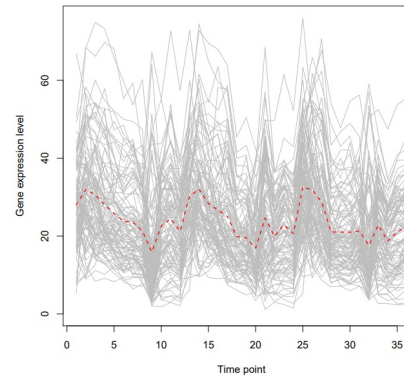
Aim of the project

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

Input data file

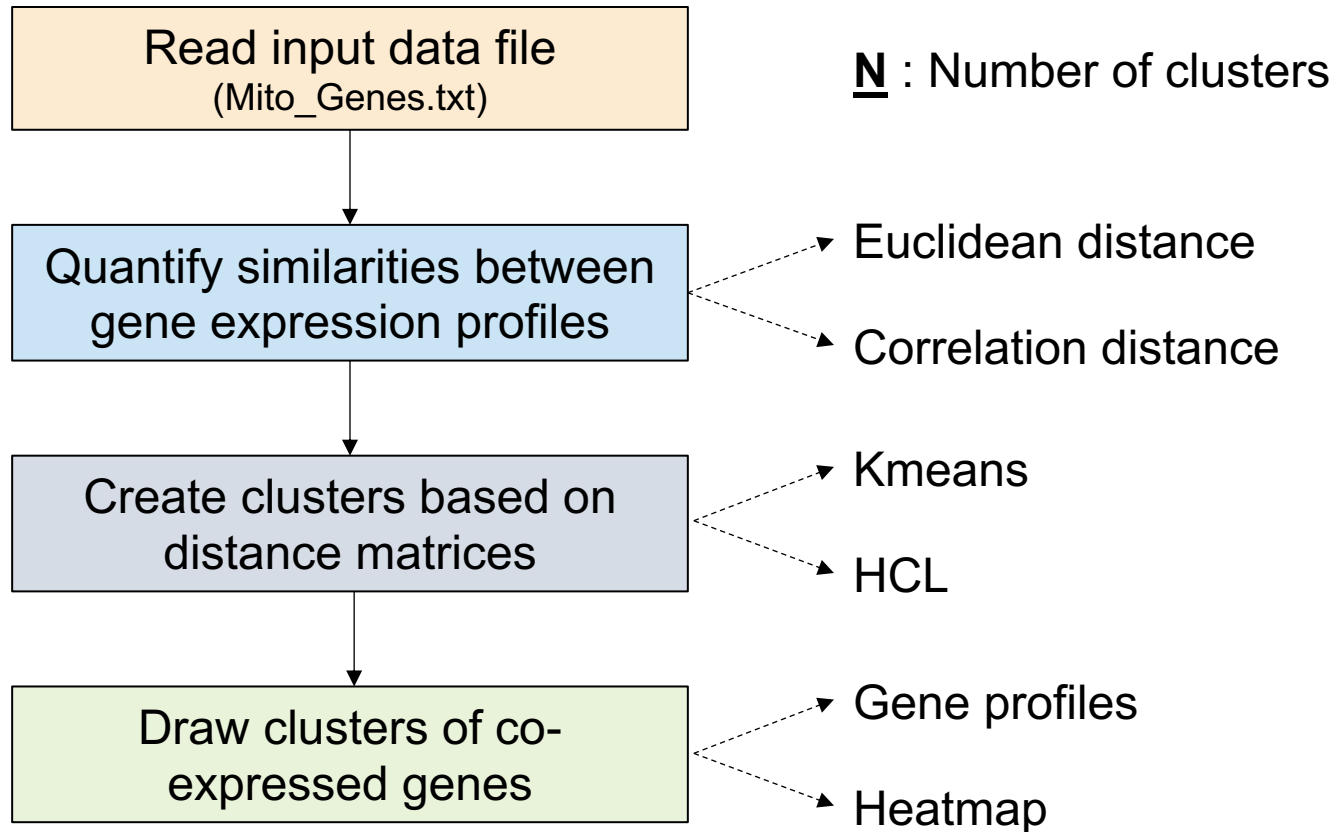
→ Output data files

1	GENE_ID	GSM77298_T1	GSM77299_T2	GSM77300_T3	GSM77301_T4	GSM77302_T5	GSM77303_T6	GSM77304_T7	GSM77305_T8	GSM77306_T9
2	YGR234W YHB1	33.102	16.733	14.173	13.126	11.343	12.259	13.148	16.114	31.631
3	YEL039C CYC7	1.23	2.423	1.286	1.631	1.077	1.885	1.866	1.155	0.934
4	YGL187C COX4	29.48	26.334	22.992	20.055	22.066	18.979	19.251	17.684	8.393
5	YGL191W COX13	23.448	21.726	22.306	17.664	19.42	14.673	15.564	13.21	8.592
6	YHR051W COX6	29.857	26.911	26.908	20.949	18.877	17.242	17.073	16.289	11.238
7	YLR038C COX12	40.713	48.95	51.821	46.777	40.616	31.031	36.191	30.668	24.632
8	YLR395C COX8	32.766	35.924	43.495	38.335	29.136	27.276	27.436	25.353	22.413
9	YMR256C COX7	22.186	32.672	43.094	35.976	29.4	25.574	21.982	22.437	22.546
10	YNL052W COX5A	33.116	22.792	25.244	21.791	20.048	16.707	19.713	19.473	14.695
11	YKR016W AIN28	5.295	2.892	2.637	2.11	1.604	1.244	1.53	1.573	1.598
12	YNR020C ATP23	3.107	1.29	0.773	0.73	0.627	0.65	0.788	0.859	0.909
13	YLR201C COQ9	5.837	7.005	8.955	6.286	5.82	3.836	5.458	4.07	2.23
14	YOR222W ODC2	2.949	1.375	2.262	2.253	2.692	3.186	3.194	3.8	7.516
15	YGR231C PHB2	13.843	13.092	7.948	8.232	8.535	8.473	8.412	7.981	3.653
16	YGL219C MDM34	3.236	3.929	3.499	2.642	2.545	2.602	2.528	1.934	1.424
17	YGR132C PHB1	8.647	8.854	6.591	5.346	4.214	4.737	4.974	4.276	1.586
18	YJL066C HFM1	10.074	15.177	8.112	8.271	7.274	5.62	6.58	5.367	0.995
19	YBR003W COQ1	4.555	2.164	1.661	1.445	1.074	1.057	1.111	1.393	0.437
20	YNR041C COQ2	2.995	2.834	2.635	1.798	0.84	1.142	1.174	1.385	1.702
21	YPL109C IRA	4.386	6.238	7.104	6.627	5.71	5.81	5.489	3.905	0.625
22	YGL116W ABC1	0.926	0.649	0.413	0.485	0.571	0.55	0.545	0.697	0.556
23	YGR255C COQ6	4.404	3.328	3.533	3.228	2.59	2.838	2.762	2.296	0.708
24	YLR056W ERG3	10.90	3.135	4.723	5.367	7.29	8.694	13.834	17.927	16.873
25	YOL096C COQ3	3.329	3.726	2.975	2.508	1.796	2.228	2.379	2.329	0.43
26	YDR204W COO4	3.974	7.12	4.956	5.768	6.764	5.777	6.125	4.541	0.858

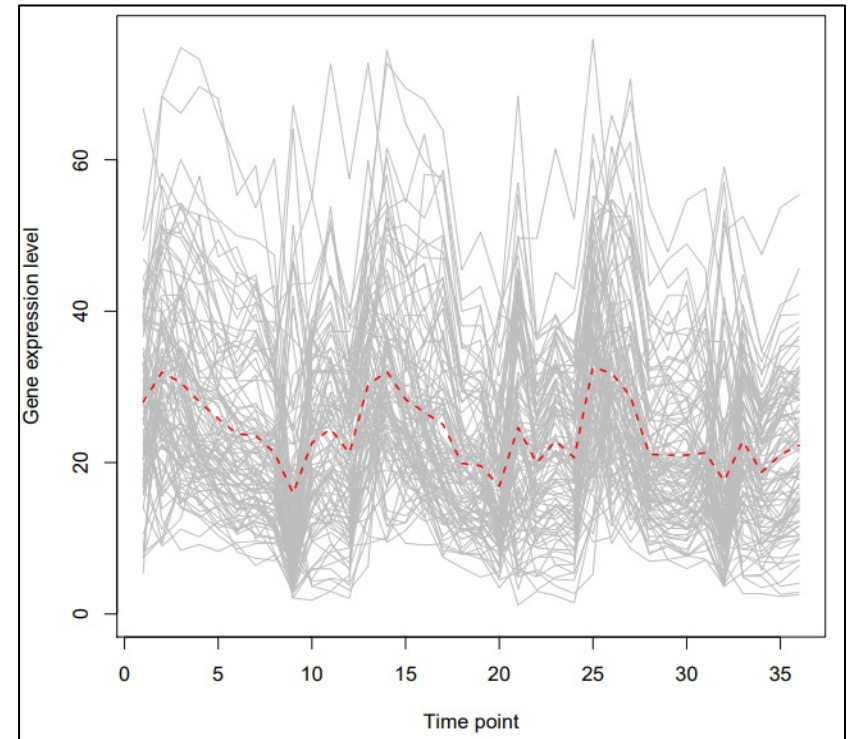
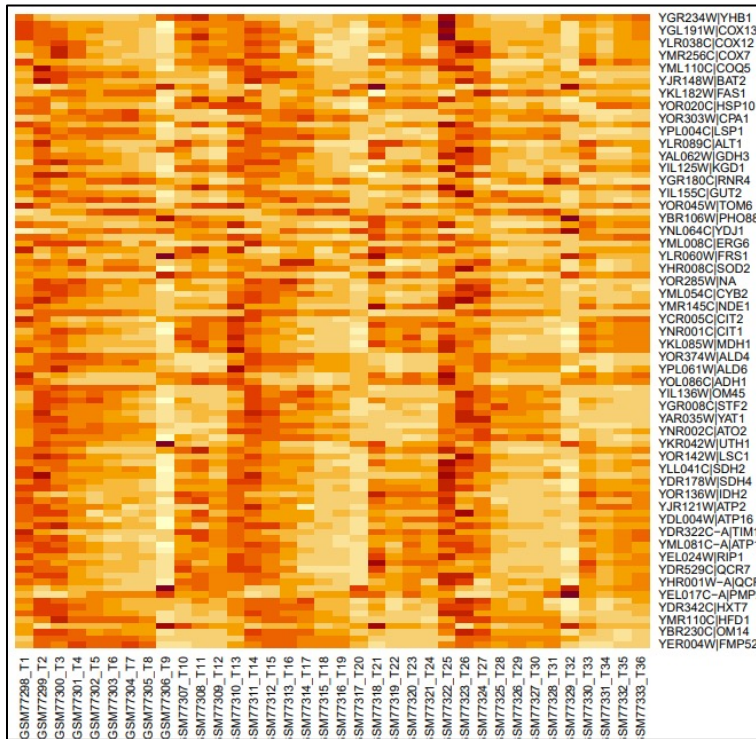


A set of gene clusters, obtained with **different algorithms** (HCL, kmeans, etc.) and several **distances** (Euclidian, correlation, etc.)

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

script_clustering_level0.R

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

script_clustering_level1.R

- **N clusters**
- **Choose distance**
- **Choose algorithm**

script_clustering_level2.R

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

script_clustering_level3.R