Investigating cancer resistance in a Glioblastoma cell line with gene expression data

Hicham Janati

Hicham JANATI

ENSAE ParisTech - 2nd year
Application Internship - 2015 / 2016

Internship report

On

Investigating cancer resistance in a Glioblastoma cell line with gene expression data

INRIA Paris

Laboratoire Jacques-Louis Lions

Supervised by:

Marie Doumic & Jean Clairambault

15 / 06 - 15 / 09

Contents

Abstract

Glioblastoma multiforme is the most lethal brain tumor as 3 to 4 per 100 000 people develop the disease every year. Only less than 9.8% survive five years after diagnosis (as opposed to 89% in breast cancer) [1][2]. Surgery – if feasible – combined with chemotherapy and radiotherapy is the most common treatment extending median survival expectancy from 3 months (without treatment) to about 12-15 months [3]. Drug resistance and heterogeneity in cancer cell populations are believed to explain a large part of our failure in fighting brain tumors [2, 4]. Here, to highlight interesting genes, gene expression profiling is performed on a culture of human cancerous cell-line before and after the chemotherapeutic agent TMZ (Temozolomide) is introduced. Different mathematical procedures are used: count data models (Deseq2) and gene correlation networks (WGCNA). We also developed a software application for analyzing further Single-Cell Real-Time PCR data that will exhibit expression profiles of 96 interesting genes pointed out by the very RNA-seq results. After a simplified biological introduction, this paper explains how data were analyzed using statistical procedures.

1 Introduction and biological background

1.1 Genomic information

What are we? Over a trillion of cells fulfilling some function in order to keep us alive. But do cells *know* they are alive? Maybe, maybe not. But they certainly know what they are supposed to do next. The presence of biological information within cells was not confirmed until Watson and Crick *cracked the code* of the complex DNA molecule. Four molecules called nucleotides are the basic ingredients DNA is built of. Nucleotides are the same across all living creatures, their order however is not. At a slightly bigger scale, DNA can be seen as a large set of sequences called *genes*. Each gene (a specific order of nucleotides) codes for a certain list of proteins, controlling all biological processes such as proliferation (cell division), angiogenesis (extension of blood vessels) or apoptosis (programmed cell death). Unfortunately, this mechanism is not flawless. Gene mutations may occur[1] and therefore alter the instructions contained in genomic information. Mutations are conserved through cell division, which makes them combinable during a lifetime. If proliferation, DNA repair and apoptosis are not well regulated (due to mutations), cells might be at higher risk of becoming cancerous: proliferating rapidly and refusing to die.

Thanks to the tremendous advance in gene expression profiling techniques in the last two decades, we are now able to *see* what *are cells doing* using *Next Generation Sequencing* (NGS). Comparing (using statistical tests) sick and normal tissues in terms of expressed genes for example – called *differential gene expression analysis* – allows to spot differentially expressed genes to be used as biomarkers for the studied disease. Cancer is no exception. Gene expression analysis not only improved diagnosis quality but also offered a better understanding of tumors.

1.2 Glioblastoma

MGMT

Glioblastoma is the most aggressive type of gliomas (tumors developing in glial cells). Glial cells (*glia* from Greek: *glue*) surround neurones to protect and support them. Temozolomide (TMZ) is the most common chemotherapeutic agent used to fight glioblastoma. TMZ damages DNA during cell division and thus triggers the death of

[1] Mutations happen because of environmental factors, mistakes during DNA replication and perhaps other causes. Not all mutations are harmful. Some are responsible for different physical traits between individuals.

cancerous cells. But DNA damage does not always go unnoticed: in about 2/3 of the cases (patients), the action of TMZ is cancelled if the gene MGMT is expressed [5]. But DNA repair provided by MGMT is only one example[2] of drug resistance in chemotherapy. The study we performed on RNA-seq count data aims to highlight other genes related to drug resistance by multiple testing. However, drug resistance mechanisms can be very complex involving many genes interactions [2]. Weighted Gene Co-expression Networks Analysis (WGCNA) complements the former study by clustering genes in representative modules and spotting hidden co-expression patterns.

Heterogeneity

Several studies showed the presence of distinguished subpopulations within cancer cells [4]. Such heterogeneity in tumors is considered a form of resistance since drugs must be adapted to each type of cells. One of the observed clones regroups cancerous cells showing stem characteristics that can acquire drug resistance and repopulate the sample. Stem cells are cells that can differentiate into specialized cells (think for example of embryonic cells). Heterogeneity has often been seen as a form of Darwinian selection [4, 6].

1.3 RNA sequencing

When a cell needs the information contained in a certain gene, it makes a copy of the needed part from the DNA molecule called RNA. Hence, quantity of a specific RNA molecule shows how much a gene is expressed in the sample. In brief, RNA sequencing is basically a fragmentation of all RNA matter found in the cell population[3]. Obtained fragments are then mapped to the genome. Data element (i, j) of gene i in sample j is the number of *reads* (fragments) found in sample j mapped to gene i. Data are analyzed by performing pairwise comparisons of samples for each gene (tens of thousands in general).

1.4 Single Cell Real-Time PCR

Heterogeneity cannot be observed and studied in sample data given that counts are performed on a population of cells. Recent technological advance made this possible by isolating cells in wells. Gene expression profiling is then performed in each cell. Real-Time PCR techniques rely on DNA amplification by PCR *(Polymerase chain reaction)*. The idea is that the faster the chemical reaction, the more abundant the DNA in the isolated cell. Given that a comparison threshold is taken in the first exponential phase of the reaction, the data element (i, j) referring to gene i in cell j is the number of cycles (time) required to reach the defined threshold. Data are therefore given in logarithmic scale (log2 precisely)[7]. The main downside of this technique is the limited number of genes that can be analyzed at once. Here, a Fluidigm Biomark HD 96 × 96 platform was used (96 cells x 96 genes)[4].

Internship context

All data were provided by a CRCNA[a] team led by François Vallette. Several meetings and video calls were held in Nantes in order to address the biological background and interpretation of the obtained results. The Internship took place at LJLL, Paris[b] and was supervised by mathematicians Marie Doumic and Jean Clairambault[c]. At first, the idea behind the project

[2] And the most important one: median survival increases by 7 months in patients with silenced MGMT [5].

[3] Fragmentation is needed because some genes are too long (too many nucleotides) to be identified as a whole. Here, the length of fragments is around 90 nucleotides.

[4] This is not the state-of-the-art science: recently, Single Cell gene expression profiling has been extended to whole genome sequencing: Single Cell RNA-seq [8].

Internship context

was to investigate heterogeneity of Glioblastoma cells using genes expression data with respect to the drug resistance model Chisholm et al. presented in [4]. However, provided single cell data did not contain interesting discriminatory genes. Highlighting such genes became a priority that was addressed using whole genome RNA-seq data. Single Cell experiments involving these genes will be held by Vallette's team later this year. This paper has a twofold purpose:

- highlight interesting genes or groups of genes in RNA-seq data from which a few ones will be studied in single cell experiments later this year.
- develop an application (first *tried* on available single cell data) for further use in future single cell data to investigate tumor heterogeneity.

In *Materials and methods*, we establish the theoretical basement of the statistical tools used in analyzing RNA-seq data and developing the Single Cell application. Readers who are familiar with the software used in biostatistics can move directly to the next section where we first discuss the obtained RNA-seq results before testing our program on the available single cell data.

2 Materials and methods

2.1 RNA-seq: Deseq2 model

Table 1: *First four rows of raw RNA-seq count data.*

	Day 0, no TMZ[5]				TMZ, day 4			TMZ, day 9				TMZ, day 12				TMZ, day 16			
Gene | Sample	d0_1	d0_2	d0_3	d0_4	d4_1	d4_2	d4_3	d9_1	d9_2	d9_3	d9_4	d12_1	d12_2	d12_3	d12_4	d16_1	d16_2	d16_3	d16_4
TSPAN6	602	655	564	296	953	748	1010	576	722	1148	936	794	625	1086	1132	717	345	815	423
TNMD	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
TPM1	1061	989	906	624	1681	2215	1747	1279	1370	2134	1656	1200	783	1739	1906	1326	1098	1399	1198
SCYL3	115	199	175	125	291	342	297	172	180	296	212	172	135	209	268	65	282	207	192

Data description

As mentioned in section 1.3, RNA-seq data are a numerical matrix of non-negative integers representing the number of reads found in a sample mapped to a specific gene. A third and crucial dimension is the number of replicates per biological condition: in general, comparisons are made for each gene between samples of different biological conditions (before drug, after drug). For any statistical inference to be made, data must be generated at least twice [9]. Here, four replicates are analyzed per condition (except for day 4 where one replicate has been removed after quality control which will be discussed further). Table 1 shows the upper part of the raw count data. Shape of the full obtained matrix is (57 905 × 19). dx_y means: *sample of biological condition* day x, *repli-*

[a] Centre de Recherche en Cancérologie Nantes - Angers

[b] Laboratoire Jacques-Louis Lions

[c] Both researchers in the INRIA team MAMBA (Modelling and Analysis for Medical and Biological Applications).

cate number y. For further use, let $\rho(j)$ represent the biological condition of sample j (for instance, $\rho(1) = \rho(2) = \rho(3) = \rho(4) = "d0"; \rho(5) = \rho(6) = \rho(7) = "d4"$).

Models

Consider one sample j of total RNA fragments N_j. The counts distribution obtained from random sampling would be multinomial (the number of genes being finite). Hence, for each gene i, read counts follow a binomial distribution. Now, as mapped fractions to each gene tend to be very small compared to N_j (Here $\mathbb{E}(N_j) \simeq 2$ million), binomial distribution can be approximated by a Poisson distribution (Supplementary material A.1). Poisson distribution has therefore been used for statistical tests between conditions.

The purpose of modeling is to test for differential gene expression (DGE). For each gene, we would like to test whether the difference between gene expression levels in two different conditions is statistically significant. Poisson distribution has been used in many studies to achieve this. However, since its variance is equal to its mean, recent papers showed over-dispersion in real data: a variance higher than the mean [10, 11] (which is also the case in our data, Figure 8). This *extra* variance is believed to be the result of biological and sequencing processes [10, 11, 12].

To account for it, one could use a bayesian method by allowing the Poisson mean to be a random variable and model read counts by marginal distributions [9, 13]. This is basically the intuition that led, with a few more assumptions, to the over-dispersion models: Gamma, Negative binomial and lognormal. One important property of such biological studies is the low number of replicates due to cost and time: 2-3 replicates per condition are very common yet reasonable [11]. The true distribution of individual gene counts remained unclear until M. Gierliński et al. [12] performed a high-replicate data to be confronted with statistical models showing that negative binomial model is the most consistent.

Another element needs to be discussed before introducing the model: the library size N_j. Let Y_{ij} be the number of reads in sample j that were mapped to gene i. For a gene to be differentially expressed, **fractions** of reads over library sizes are compared instead of fragment **counts** to account for different sequencing depth across samples. However, even though ratios $\frac{Y_{ij}}{N_j}$ seem to be a good normalization scheme, genes with high number of reads and high differential expression tend to introduce a bias in library sizes and shade less differentially expressed genes. In any event, the *observed* ratios of counts $\mathbb{E}(Y_{ij})$ are assumed to be proportional to some normalizing size factors s_j which estimation will be presented later on.

Furthermore, due to the low number of replicates, a function linking mean and variance is needed to estimate both parameters. Using such an assumption, *EdgeR* has only one parameter to estimate [13]. *Deseq2* generalizes this relationship to account for more variability and presents an estimation scheme with a better fit.

Deseq2 model

Negative binomial:

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma^2_{ij}) \tag{1}$$

Assumptions on mean and variance:

$$\mu_{ij} = q_{i\rho(j)} s_j \tag{2}$$

$$\sigma^2_{ij} = \mu_{ij} + \alpha_{i\rho(j)} \mu^2_{ij} \tag{3}$$

Where α_i models within-group variability of gene i.

Variance decomposition can be seen as a sum of a sampling term (Poisson) and an *extra* variability term modeling biological variance.

Testing for differential expression of gene i between two conditions ρ and ρ' is basically testing the null $H_0 : q_{i\rho} = q_{i\rho'}$ against the alternative $H_1 : q_{i\rho} \neq q_{i\rho'}$. But first the normalization size factors s_j must be estimated. As it was mentioned earlier, N_j is not a good normalization factor because of the strong influence of highly and differentially expressed genes. Instead, *Deseq2* takes the median of ratios of observed counts (Supplementary section A.2) [10]. The matrix of general term $q_{i\rho(j)}$ is called normalized count matrix.

GLM fit and empirical Bayes shrinkage

A log-linear GLM (Generalized linear model) is used to analyse the experimental design. In this study, the resulting equation is:

$$\log_2\left(q_{ij}\right) = \sum_r x_{jr}\beta_{ir} \tag{4}$$

where the matrix elements x_{jr} indicate which condition the sample j is taken of and the GLM coefficients β_{ir} are used to compute LFCs (log fold change) estimates i.e log ratios of q_{ik}. Testing for differential expression is then exactly the same as testing the hypothesis of null LFC estimates.

Using the model assumptions, the GLM model and the size factors estimators of s_j, we can carry out estimates of the unknown parameters (the mean ratios (LFC) and the dispersion α_i) using Maximum Likelihood Estimates (MLE), but two difficulties lie ahead:

- Number of samples: estimating dispersion parameters α_i with such a few replicates per condition leads to a high level of noise, hence inaccurate DGE tests.
- Heteroscedasticity: genes with low counts tend to have strong variance of LFC estimates (the lower their level, the higher their sensitivity to differential change) which brings up the weak genes to be the most differentially expressed (false positives).

To overcome these issues, *Deseq2* uses two separate techniques that both rely on a Bayesian method in order to shrink estimates to a more *reasonable* trend. Shrinkage is more observed on genes with low counts or high dispersion. Follows in figure 1 a scheme we elaborated here to give a general idea of the concept.

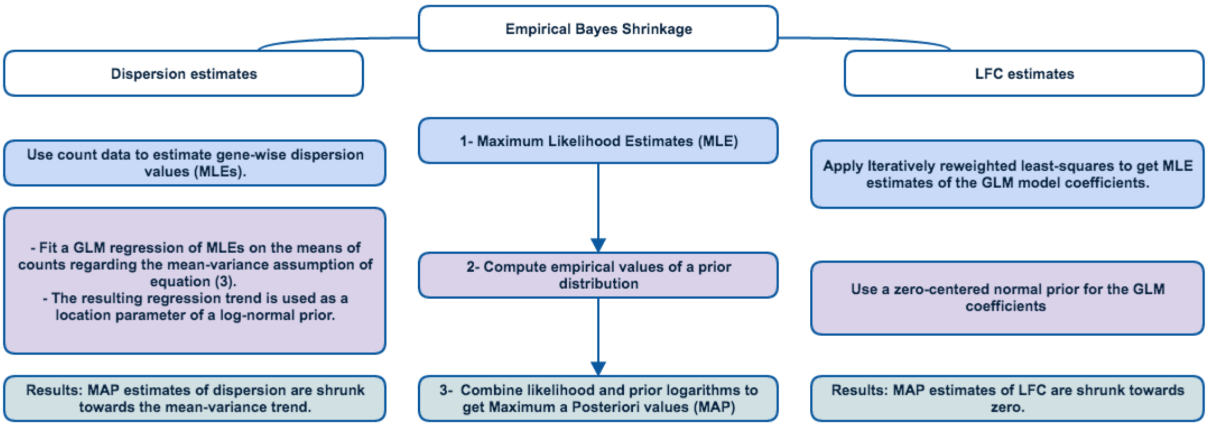

Empirical Bayes Shrinkage
Dispersion estimates
LFC estimates
Use count data to estimate gene-wise dispersion values (MLEs).
1- Maximum Likelihood Estimates (MLE)
Apply Iteratively reweighted least-squares to get MLE estimates of the GLM model coefficients.
- Fit a GLM regression of MLEs on the means of counts regarding the mean-variance assumption of equation (3).
- The resulting regression trend is used as a location parameter of a log-normal prior.
2- Compute empirical values of a prior distribution
Use a zero-centered normal prior for the GLM coefficients
Results: MAP estimates of dispersion are shrunk towards the mean-variance trend.
3- Combine likelihood and prior logarithms to get Maximum a Posteriori values (MAP)
Results: MAP estimates of LFC are shrunk towards zero.

Figure 1: *Bayesian shrinkage procedure cartoon. Estimating dispersion and log fold change (LFC) is done following the same steps that are described in the middle column. First, likelihood estimate are computed. Then a gaussian prior distribution is carried out. Finally using the first two steps, posteriori estimates are calculated.*

The procedure is actually more subtle than that, readers seeking theoretic details should refer to section *Materials and Methods* of [11]. A concrete example in our data is given in Figure 9 that shows the difference between shrunken and unshrunken LFC estimates.

Multiple testing

Wald test

To test for DGE between *days*, a Wald test is used. Recalling that LFCs are log-ratios of normalized counts, testing for equal means is equivalent to testing LFCs' nullity. LFCs can be written as linear combinations of GLM coefficients estimates $\hat{\beta}_{ir}^{\mathrm{MAP}}$: GLM coefficients β_{ir} represent the $\log_2$-expression level gene i in sample r.

Formally, to perform a pairwise comparison between sample a and b regarding expression level of gene i we must test the null $H_0 : q_{ia} = q_{ib}$. Now since genes with no reads are filtered out of the data, q_{ij} are non-zero:

$$\begin{aligned} q_{ia} = q_{ib} &\Leftrightarrow \frac{q_{ib}}{q_{ib}} = 1 \\ &\Leftrightarrow \log_2\Big(\frac{q_{ib}}{q_{ia}}\Big) = 0 \\ &\Leftrightarrow \underset{a \to b}{\mathrm{LFC}} = 0 \\ &\Leftrightarrow \beta_{ib} - \beta_{ia} = 0 \\ &\Leftrightarrow C'_{a,b}\beta_{i.} = 0 \end{aligned}$$

Where $C_{a,b}$ is a vector containing 1 in b's position, -1 in a's position and zero elsewhere.

In general, the standard error of any vector C (called *contrast vector*) is:

$$\mathrm{SE}(C'\beta_{i.}) = \sqrt{C'Cov(\beta_{i.})C} \tag{5}$$

Using (5), the Wald statistic for the null $H_0 : C'_{a,b}B_{i.} = 0$ against $H_1 : C'_{a,b}B_{i.} \neq 0$ is:

$$\frac{C'_{a,b}\beta_{i.}}{\sqrt{C'Cov(\beta_{i.})C}} \sim \mathcal{N}(0,1) \tag{6}$$

False discovery rate correction

Testing multiple hypotheses at once generates a certain number of false rejections that increases with the number of performed tests. Several methods have been introduced to correct the multiple testing problem but limiting FDR (False discovery rate) has proven to be the most suited to biological studies [14].

Let's reproduce the count table of Benjamini and Hochberg (1995). A total number of m null hypothesis are tested of which m_0 are true. The left axis reveals the true nature of the test. The top axis represents the decision made by the statistician. U and S are the numbers of *good calls*. T is the number of false negatives and V is the number of false positives. Capital letters are meant to distinguish variables from constants.

	Not rejected	Rejected	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	$m1$
Total	$m - R$	R	m

Table 2: *Counts of m multiple tests results of which m_0 are true. R is the only observable variable.*

One of the possible corrections is limiting by some threshold α the probability of having one or more false positive. In statistics litterature, it is known as the family-wise error rate (FWER) and formally, using the table's notations:

$$\mathrm{FWER} = \Pr\big(V \geq 1\big) \tag{7}$$

FDR however is defined by the average ratio of false positives:

$$\mathrm{FDR} = \mathbb{E}\bigg(\frac{V}{R}\bigg) \tag{8}$$

FDR has proven to be less conservative than FWER[15] and more appealing to biostatisticians: controlling the proportion of false discoveries is more interesting than the probability of carrying out one or more false positives [14].

Given a testing threshold α, making sure than FDR $\leq \alpha$ is done by using one out of the many algorithmic procedures.The procedure introduced by Benjamini and Hochberg (1995) called *BH procedure* [15] is as follows:

Let $\left(P_i\right)_{1 \leq i \leq m}$ be the corresponding p-values series of the null hypotheses $\left(H_i\right)_{1 \leq i \leq m}$.

1. Order the p-values: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$
2. Find k, the largest i verifying: $P_{(i)} \leq \frac{i}{m}\alpha$
3. Reject $H_{(i)}$ for all $i = 1, 2 \ldots k$

The Benjamini and Hochberg (1995) theorem states that such a procedure ensures that $\mathbb{E}\left(\frac{V}{R}\right) \leq \alpha$ [15].

In practice however, Yekutieli & Benjamini (1999) [16] showed that FDR can be controlled by adjusting the obtained p-values as to account for the multiple testing problem. The adjusted p-value of test i is defined by:

$$p_i^{\text{adj}} = \inf\{\alpha : H_i \text{ is rejected at FDR} = \alpha\} \qquad (9)$$

When computing Deseq2 Wald tests, both *original* p-values and adjusted p-values are returned but only adjusted (FDR) are kept for analysis.

DBSCAN

Why ?

Wald tests highlight (many) differentially expressed genes. DBSCAN clusters these genes by their kinetics to investigate each type of gene evolutions.

Density-based spatial clustering of applications with noise is a clustering algorithm created by [17] that works as follows:

Given a space of elements D, a distance threshold ϵ and a minimal number of points m, DBSCAN pops from point to point regrouping all neighbors laying within ϵ into one cluster if their number is greater or equal than m. If one point has enough neighbors, a cluster is formed. The algorithm goes through all the points from neighbor to neighbor expanding the cluster.

Denoting the neighborhood set of a point i by $N(i)$, we say that a separate point j is reachable from i if a *path* : $k_1, k_2 ... k_n$ between the two exist. Formally:

$$(\exists n \geq 2)(\exists k_1, k_2 ... k_n \in D)$$

$$\begin{cases} (\forall i \in [|1, n-1|]) & \text{Card}(N(k_i) \geq m) \\ (\forall i \in [|1, n-1|]) & k_{i+1} \in N(k_i) \\ j = k_n \text{ and } i = k_1 \end{cases}$$

A data element i is therefore considered part of a cluster C if and only if: i is reachable from an element of C. Reciprocally, noisy elements are those that cannot be reached from any element of D.

DBSCAN is deterministic, it does not require the specification of a number of cluster and it can retrieve uneven sized clusters. Some genes have high mean and variance with no apparent biological reason and can be treated as noisy. For these reasons, DBSCAN is very appropriate for RNA-seq data.

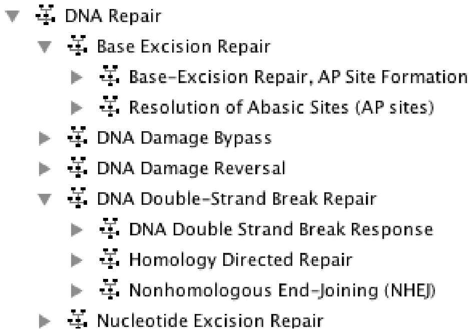
The main downside of the algorithm however is the choice of ϵ, m and the metric distance to use in defining neighborhoods. We compared several distances listed out by [18] and settled for the euclidean distance as it produced the most coherent clustering. The choice of epsilon was based on a heuristic method proposed by the authors of the algorithm. The value of m was decided by a biological argument: gene modules should contain at least 30 genes to take into consideration their interactions as a network. [19].

2.2 Enrichment analysis

Why ?

Get biological insight out of long lists of genes: enrichment analysis is a test for over-representation of categories of biological processes.

In common genes expression studies, long lists of differentially expressed genes are highlighted. The statistical measures used for that purpose can also serve as a tool to rank them by importance (Fold change for e.g). Nonetheless, because of the high number of genes, interpreting the results can be very disconcerting, specially in getting biological insights. The incredible amount of data generated in the past two decades led researchers to develop Gene Set Enrichment Analysis that compares a set of genes grouped by a statistical measure to an *a priori* list of genes known to be involved in the same biological pathway[6]. Pathways can be displayed in a tree of nodes.



- ▼ DNA Repair
  - ▼ Base Excision Repair
    - ▶ Base-Excision Repair, AP Site Formation
    - ▶ Resolution of Abasic Sites (AP sites)
  - ▶ DNA Damage Bypass
  - ▶ DNA Damage Reversal
  - ▼ DNA Double-Strand Break Repair
    - ▶ DNA Double Strand Break Response
    - ▶ Homology Directed Repair
    - ▶ Nonhomologous End-Joining (NHEJ)
  - ▶ Nucleotide Excision Repair

Figure 2: *Example of some Reactome pathway nodes displayed in Cytoscape software.*

Consider a reference list of genes of size N. Suppose we are interested in the pathway node *Regulation of apoptosis* containing K reference genes and that in our set of genes of size n, k genes are involved in *Regulation of apoptosis*. We would like to test if the studied pathway is over-represented in our data.

The most straight forward and most common way to do that is to compute a p-value based on a hypergeometric distribution which models the event of k successes in n draws from a population of size N containing K successes [20]. The p-value of the over-representation test is computed as the probability of drawing k or more successes:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{i}} \qquad (10)$$

The test is applied for each referenced pathway, followed by a multiple testing correction by FDR controlling.

Many bioinformatics projects offer pathway reference libraries, in this study *Reactome* software was used for its simplicity and availability as an R package (ReactomePA [21], DOSE [22]).

2.3 RNA-seq: Weighted Genes Co-expression Network Analysis

Why ?

Model genes co-expression as an interacting network and cluster them in correlated dense modules.

Summary

Testing for differential expression using Deseq2 assumes the independence of genes and treats them separately which is not biologically true: genes are involved in a complex mechanism of regulation and co-expression. They can be seen as a dense modules of a giant network. WGCNA (Weighted Genes Co-expression Network Analysis) [23] de-

[6] Pathway: a series of actions among molecules in a cell that leads to a certain product or a change in the cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move (*National Human Genome Research Institute*)

veloped a network model to avoid the multiple testing problem and to analyze gene groups in modules.

For a weighted network, the $(n \times n)$ symmetric adjacency matrix verifies:

$$\begin{cases} s_{ij} \in [|0,1|] \\ s_{ij} = s_{ji} \\ s_{ii} = 1 \end{cases}$$

One possible adjacency measure is the signed correlation: $s_{ij} = \frac{1+corr(gene_i, gene_j)}{2}$[7]. Relying on this measure, clustering would be based on proximity of pairs of genes instead of dense network interactions. To address that, WGCNA uses the *topological overlap measure* which replaces the pairwise elements s_{ij} by:

$$\text{TOM}_{ij} = \frac{\sum_{k \neq i,j} s_{ik} s_{kj} + aij}{\min(\sum_{k \neq i} s_{ik}, \sum_{k \neq j} s_{kj}) + 1 - s_{ij}} \qquad (11)$$

Two genes *i,j* would be TOM-similar if they interact strongly with the same neighbor genes.

Moreover, to enhance strong correlations at the expense of weak ones, a power transformation is applied to correlation elements s_{ij}, replaced by $a_{ij} = s_{ij}^{\beta}, \beta > 1$.

Modules of genes are constructed by applying hierarchical clustering based on the *TOM* adjacency matrix of equation 11.

Choosing β

The growth of biological systems networks is thought to be characterized by a power-law distribution [24]. Such networks are called *scale-free*: the frequency distribution of the number of node having k connections follows a power law:

$$P(k) \sim k^{-\lambda}.$$

Measuring the coefficient of determination R^2 of the regression model: $\log(P(k))$ versus $\log(k)$ is an essential quality control method of the scale-free topology property.

The power parameter β has a mixed effect on the network quality:

- Low values of β keep the adjacency elements a_{ij} equally distributed on $[|0,1|]$ (hence high connectivity or information) which leads to a poor scale-free fitting index R^2.
- High values of β shrink the adjacency elements a_{ij} to zeros and ones (loss of connectivity) boosting the scale-free topology index R^2.

where the connectivity of a network is defined as the total mean of the final adjacency matrix (TOM):

$$C(A) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} TOM_{ij}$$

A trade-off between scale-free and connectivity must be made to choose β properly. Based on biological results quality, [24, 23] recommend to consider values of β where $R^2 \geq 0.8$.

Eigengene networks

Once modules are obtained by hierarchical clustering, a fictive gene representative is computed for each module by taking the PCA first principal component and calling it *the module Eigengene.* Correlations of eigengenes are plotted in a signed heatmap to analyze co-regulated modules.

Eigengenes cannot be analyzed any further since they *are not real.* To get insight on interesting genes, correlation of genes and the eigegene of their module can be a membership quality measure.

2.4 Single Cell RT-PCR data

This section describes single cell data and presents all the statistical methods that were put in use

[7] Signed in contrast with the unsigned correlation $|corr(gene_i, gene_j)$ that does not distinguish sign of correlation.

in the software application. As normalization and data filtering are performed by the team of Nantes, they will not be discussed here. Interested readers can however find information on the matter in Supplementary section A.3.

Data description

In the performed single cell PCR experiments, a set of 96 genes are analyzed in each of the 96 isolated cells. The experiment is destructive, which means that the comparison of different biological conditions is not assessed on the same cells but relies on the probabilistic distribution of genes expression. For each biological condition, a separate matrix A is generated. Data element A_{ij} is proportional to the required time for the amplification of the quantity of gene j found in the cell i to reach a significance threshold. In other words, the higher A_{ij}, the lower the expression of gene j in cell i.

Because of a small quantity of DNA, some reactions are so slow that the measure of expression is not reliable. A maximum number of cycles C_{max} is used. If the amplification in reaction (i, j) does not reach the significance threshold within C_{max} amplification cycles, the gene j is declared unexpressed in cell i and A_{ij} is assigned the value C_{max} (here $C_{max} = 30$). This decision in not arbitrary and is backed by the work of [25] where single cell and 100 cells measurements were confronted and showed strong concordance.

Assigning $A_{ij} = C_{max}$ to an unexpressed gene j in well i is mandatory when generating the data. However, it is very confusing in analysis because according to our convention, it means that no quantity of gene j was found in the well but mathematically it means that the reaction took an amount of time equal to 30 cycles ! To avoid this ambiguity, we set these values to $+\infty$.

Finally, in order to make data more intuitive, an inverse is applied before data analysis:

$$A_{ij} = C_{max} - A_{ij}$$

which leads to the following:

- $A_{ij} > -\infty$: A_{ij} is proportional to the expression of gene j in well i.
- $A_{ij} = -\infty$: gene j is not expressed in cell i.

Since the amplification cycles are base-2 exponential with respect to the quantity of DNA Q_{ij} [25, 7], we can retrieve an approximation of the number of reads Q_{ij} :

$$Q_{ij} = 2^{Aij}.$$

The final revised categories are:

- if $Q_{ij} > 0$ (or $A_{ij} > -\infty$): A_{ij} is proportional to the expression of gene j in well i.
- if $Q_{ij} = 0$ (or $A_{ij} = -\infty$): gene j is not expressed in cell i.

Bernoulli-Lognormal Model

Let Q_{ij} be the expression level (number of reads) of gene j in cell i and $B_{ij} = \mathbb{1}(Q_{ij} > 0)$. The Bernoulli-Lognormal (BL) model assumptions are:

$$(Q_{ij}|B_{ij} = 1) \sim \log\mathcal{N}(\mu_j, \sigma_j^2) \quad (12)$$

$$(Q_{ij}|B_{ij} = 0) \sim \delta_0 \quad (13)$$

$$B_{ij} \sim \mathcal{B}(\pi_j) \quad (14)$$

where (μ_j, σ_j^2) are the lognormal parameters and δ_0 is the Dirac delta function. $\mathcal{B}(\pi_j)$ denotes a Bernoulli distribution with π_j being the frequency of expression of gene j across all cells.

Bernoulli-lognormal model

Previous work [7, 25, 26] showed that gene expression data Q_{ij} (reads) follow a lognormal distribution for each gene across cells. However, because of the C_{max} bounding and the inverse operation, genes distributions are zero inflated log-gaussians. Our provided single data are consistent with these observations. QQ plots of Supplementary Figures (35 - 39) motivate the adoption of the BL (Bernoulli-lognormal) model by [25]. Figure 3 shows four gene distributions of A_{ij} from our data. Recall that if $(Q_{ij}|B_{ij} = 1) \sim \log\mathcal{N}(\mu_j, \sigma_j^2)$ then $(A_{ij}|B_{ij} = 1) \sim \mathcal{N}(\mu_j, \sigma_j^2)$

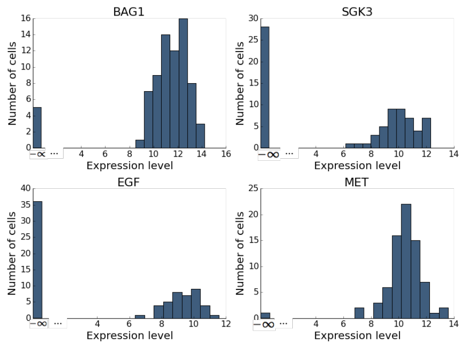


Figure 3: *Expression histograms of four genes at day 0. On the bottom axis, A_i values are used. To visualize both gaussian and bernoulli aspects of the distribution, the count value of non-expressed genes ($A_{ij} = -\infty$) are also plotted on the left side.*

Likelihood-ratio test

Why ?

The frequency and mean of expression are the most biologically significant parameters that determine how a gene is expressed across cells. The likelihood ratio test of the BL model combines both parameters (we assume that the variance is the same across genes).
The distribution of a gene j depends on the triplet (π_j, μ_j, σ_j).

Imagine we would like to highlight the genes with distributions that shifted from a biological condition (0) to another (1). Omitting the gene index j, we could test the null hypothesis:

$$H_0 : (\pi_0, \mu_0) = (\pi_1, \mu_1)$$

against the alternative

$$H_1 : (\pi_0, \mu_0) \neq (\pi_1, \mu_1)$$

This can be carried out using an LRT (Likelihood-ratio test) which is defined as:

$$\Lambda(C) = \frac{\sup\limits_{\theta_0,\theta_1 \in \Theta_0} \mathcal{L}(\theta^0, \theta^1 | Q^0, Q^1)}{\sup\limits_{\theta_0,\theta_1 \in \Theta} \mathcal{L}(\theta^0, \theta^1 | Q^0, Q^1)} \qquad (15)$$

Where $\Theta = \{(\pi_0, \mu_0, \sigma^2, \pi_1, \mu_1, \sigma^2)\}$, $\Theta_0 = \{(\pi_0, \mu_0, \sigma^2, \pi_0, \mu_0, \sigma^2)\}$, $\theta^k = (\pi_k, \mu_k, \sigma^2)$ and Q^k is a gene array under the condition $k \in \{0, 1\}$.

Let I_k be the number of cells in condition k and n_k the number of cells expressing the studied gene. The latter will be denoted by S_k: $S_k = \{i \in [|1, I_k|], {q^k}_i > 0\}$.

f_k being the lognormal density with parameters (μ_k, σ^2), the likelihood taken on one biological condition k is given by (we elaborated a proof in Supplementary information A.4):

$$\begin{aligned} \mathcal{L}(\theta^k | Q^k) &= \prod_{i \in I_k \setminus S_k} (1 - \pi_k) \prod_{i \in S_k} \pi_k f_k({q^k}_i) \\ &= {\pi_k}^{n_k} (1 - \pi_k)^{I_k - n_k} \prod_{i \in S_k} f_k({q^k}_i) \end{aligned}$$

And now, extending the sample Q to both conditions:

$$\mathcal{L}(\theta^0, \theta^1 | Q^0, Q^1) = \prod_{k \in \{0,1\}} {\pi_k}^{n_k} (1 - \pi_k)^{I_k - n_k} \prod_{i \in S_k} f_k({q^k}_i)$$

As the products of the formula above are independent, the maximization problem can be divided into two problems: one regarding $\{\pi_0, \pi_1\}$ and a second on $\{\mu_0, \mu_1, \sigma^2\}$. The (simple) resolution of both problems is given in Supplementary Information A.5.

Finally, Wilks' theorem [27] ensures that under H_0:

$$-2\log\Lambda(C) \underset{I_k\to\infty}{\sim} \chi^2_2 \tag{16}$$

where the degree of freedom of the χ^2 is given by: $\dim(\Theta) - \dim(\Theta_0) = 5 - 3 = 2$

However, even if I_k can be large, some care is needed when comparing the test statistic to the χ^2 distribution as under a condition k, the gaussian side of the distribution Q^k will be assessed on $\pi_k I_k$ observations. We simulated here a set of 1000 genes according to the BL model and under the null H_0 for a large spectrum of π values. Goodness-of-fit tests are performed to compare the test statistic with a χ^2_2 distribution. The results of the simulation (Supplementary Figure B.7) suggest that the LRT test should be carried out for genes with $\pi > 0.12$.

Goodness of fit: 2-samples Kolmogorov-Smirnov test

Why ?

- Test for the asymptotic property (Chi-square) of the LRT tests above.
- Implement it in our software application (Ladybird) in case the Bernoulli-lognormal model is not suited for the data of the user.

The CDF (cumulative distribution function) of a variable X will be denoted by F_X.

Suppose we would like to test for the independency of two continuous variables X and Y.

Let n_x and n_y be the number of observations of X and Y respectively.

Now define the test statistic:

$$D_{n_x,n_y} = \sup_z |\hat{F_X}(z) - \hat{F_Y}(z)|$$

where $\hat{F_X}$ denotes the empirical distribution function of X.

Under the null hypothesis $H_0 : F_X = F_Y$:

$$\lim_{n\to\infty} P\Big(\sqrt{\frac{n_X n_Y}{n_X + n_Y}} D_{n_X,n_Y} \leq t\Big) = H(t)$$

Where the limiting distribution H is given by:

$$H(t) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2}$$

H has been tabulated and the test can easily be computed [27].

Density estimation

Why ?

For a better visualization (and precision): kernel density estimation (KDE) offers a better comparison of distributions across multiple biological conditions than histograms when the underlying distribution is known. Here, the expressed part of single cell data follows a gaussian distribution which is recognizable in KDE plots.

Let $(X_1, \ldots, X_n)$ independent identically distributed (i.i.d) random variables that have a density function with respect the the Lebesgue measure on $\mathbb{R}$.

Let K be an even non-negative function that integers to 1 (kernel) and a parameter h (bandwidth).

The kernel density estimator is defined as:

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)$$

The choice of h is critical as it increases with bias and decreases with variance [28].

Here, a Gaussian kernel is used to approximate the density function. In this case, the optimal bandwidth h^* is given by [29] :

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}}$$

where $\hat{\sigma}$ is the estimated standard-deviation of the data.

Rank magnitude

> **Why ?**
>
> Rank magnitude is a distance that combines both correlations (Pearson coefficient) and ranks. It has proven to be better than Euclidean and correlation measures when dealing with single cell data. [18]

[18] studied the performance of several distances used in clustering analysis in single cell gene expression data. Rank magnitude came out as the most adapted distance given the nature of the data.

Let $x, y \in \mathbb{R}^n$. Let $\hat{x}$ denote the sorted array x in increasing order of values.

$$\text{Define: } \begin{cases} \text{rank}^{\text{min}} = \sum_{i=1}^{n} \hat{y}_i(n - i + 1) \\ \text{rank}^{\text{max}} = \sum_{i=1}^{n} i\hat{y}_i \end{cases}$$

When first introduced by [30], the rank magnitude was an asymmetric coefficient defined as:

$$r\hat{m}(x,y) = \frac{2\sum_{=1}^{n}\left(Rank(x_i)y_i\right) - \text{rank}^{\text{min}} - \text{rank}^{\text{max}}}{\text{rank}^{\text{max}} - \text{rank}^{\text{min}}} \tag{17}$$

The symmetric version of 17 can obtained by taking the mean as in:

$$\text{RM}(x,y) = \frac{r\hat{m}(x,y) + r\hat{m}(y,x)}{2} \tag{18}$$

This measure correlates sequences with ranks and values which can be intuitive in biological data as gene expression values are always compared to some reference (for instance: differential expression tests and significance thresholds in single cell experiments).

2.5 PyQt application for single cell data: Ladybird

Summary

Based on provided single cell data, we developed here a PyQt[8] program (called *Ladybird*) for data analysis that will be used in the Nantes team's labs in upcoming projects.

The application is meant to be suited to any single cell data, hence the need for various parametric and non-parametric procedures that were discussed above. Data visualization techniques include PCA and HCA (Hierarchical clustering analysis) plotted on the side of heatmaps. Heatmaps can be drawn from gene expression data or cells dissimilarity matrix. Some care is needed in choosing the clustering metric as the results can be dramatically different. The paper by [18] provided interesting insight as it studied the performance of several distances and clustering methods on gene expression data. The *Rank magnitude* distance (not implemented in standard Python libraries) proved to be the best among other distances. It combines both the ranks and the values of arrays when computing similarities. We decided to include: Rank magnitude, Euclidien distance and Pearson correlation.

It also possible to plot histograms and compare frequency, mean and variance of expression between groups. Finally, any procedure that generates a list of genes can be saved and used in filtering the data to focus on a specific set of genes.

Technical aspects

Figure 4 shows the main window of Ladybird. The main window includes the menubar, the toolbar and the MDI Area (multiple document interface). The MDI Area holds all Ladybird windows within the main window and is used as *parent* object for

[8] Python library (adapted from C++) specialized in creating professional graphical user interfaces (GUI)

all opened projects.

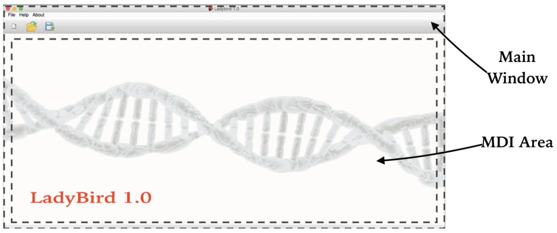
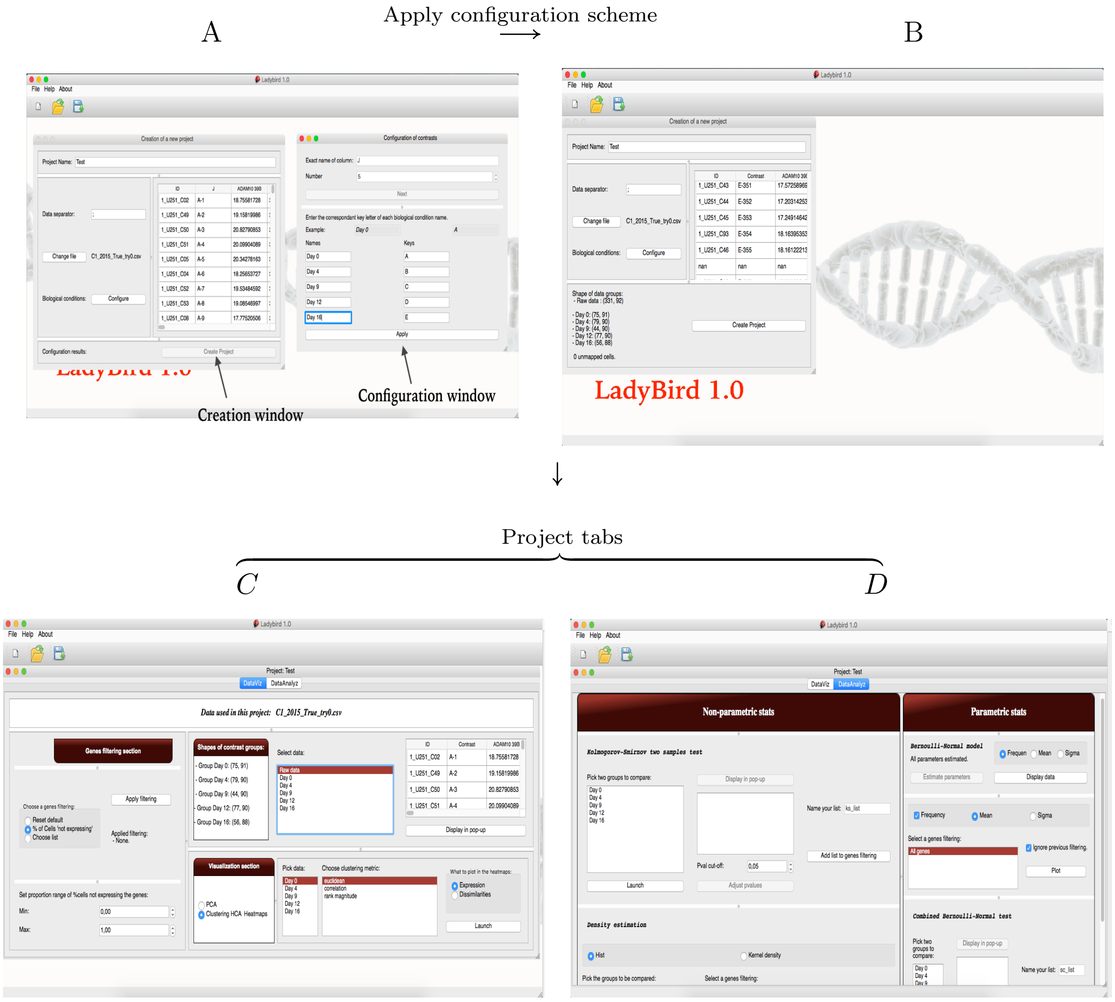


Figure 4: *Screenshot of Ladybird's interface at launch.*

The creation of new project requires configuration steps where biological conditions must be entered. This step of the program is critical as it must deal with empty columns, wrong numerical formats and parsing errors. Figure 5 shows the process of creating and configuring a project. A contrast column is needed to distinguish between cells of different groups. In the given example of the figure, the column *Day* is the contrast column with keys (A-E) denoting respectively days 0, 4, 9, 12 and 16. Every request is tested, error and warning messages pop up if the number of cells is too low for any inference to be made. A short summary of every group is provided before confirming the creation of the project.



Project Name: Test

Data separator: ;

C1_2015_True_try0.csv

Biological conditions:

Configuration results:

ID	J	ADAM10 39B
1_U251_C02	A-1	18.75581728
1_U251_C49	A-2	19.15819986
1_U251_C50	A-3	20.82790853
1_U251_C51	A-4	20.09904089
1_U251_C05	A-5	20.34278163
1_U251_C04	A-6	18.25653727
1_U251_C52	A-7	19.53484592
1_U251_C53	A-8	19.08546997
1_U251_C08	A-9	17.77520506

Exact name of column: J

Number: 5

Enter the correspondant key letter of each biological condition name.

Example: *Day 0* *A*

Names	Keys
Day 0	A
Day 4	B
Day 9	C
Day 12	D
Day 16	E

ID	Contrast	ADAM10 39B
1_U251_C43	E-351	17.57258969
1_U251_C44	E-352	17.20314253
1_U251_C45	E-353	17.24914642
1_U251_C93	E-354	18.16395353
1_U251_C46	E-355	18.16122213
nan	nan	nan

Shape of data groups:
- Raw data : (331, 92)

- Day 0: (75, 91)
- Day 4: (79, 90)
- Day 9: (44, 90)
- Day 12: (77, 90)
- Day 16: (56, 88)

0 unmapped cells.

Data used in this project: C1_2015_True_try0.csv

Choose a genes filtering:
○ Reset default
● % of Cells 'not expressing'
○ Choose list

Applied filtering:
- None.

Set proportion range of %cells not expressing the genes:

Min: 0,00

Max: 1,00

- Group Day 0: (75, 91)
- Group Day 4: (79, 90)
- Group Day 9: (44, 90)
- Group Day 12: (77, 90)
- Group Day 16: (56, 88)

Select data: Raw data, Day 0, Day 4, Day 9, Day 12, Day 16

ID	Contrast	ADAM10 39B
1_U251_C02	A-1	18.75581728
1_U251_C49	A-2	19.15819986
1_U251_C50	A-3	20.82790853
1_U251_C51	A-4	20.09904089

○ PCA
● Clustering HCA Heatmaps

Pick data: Day 0, Day 4, Day 9, Day 12, Day 16

Choose clustering metric: euclidean, correlation, rank magnitude

What to plot in the heatmaps:
● Expression
○ Dissimilarities

Kolmogorov-Smirnov two samples test

Pick two groups to compare: Day 0, Day 4, Day 9, Day 12, Day 16

Pval cut-off: 0,05

Name your list: ks_list

Density estimation

● Hist ○ Kernel density

Pick the groups to be compared:

Select a genes filtering:

Bernoulli-Normal model

All parameters estimated.

● Frequen ○ Mean ○ Sigma

☑ Frequency ● Mean ○ Sigma

Select a genes filtering: All genes

☑ Ignore previous filtering.

Combined Bernoulli-Normal test

Pick two groups to compare: Day 0, Day 4, Day 9

Name your list: sc_list

Figure 5: *From creation to data analysis.* ***A & B:*** *Creation and contrast configuration windows. Once the project is created, the project window opens with two tabs:* ***C:*** *Data filtering and visualization procedures ;* ***D:*** *Data analysis including parametric and non-parametric procedures.*

Once a project is created, the data analysis window shows up with two tabs:

DataViz: Data visualization and filtering window. It includes filtering with expression frequency or specific genes lists, PCA and HCA (clustering) plots.

DataAnalyz: Statistical procedures. Non-parametric: two samples Kolmogorov-smirnov test, histograms, KDE (kernel density estimation). Parametric: Bernoulli-lognormal estimates and LRT test.

Saving

To avoid repetitive tasks such as creation, configuration, and more importantly the loss of interesting findings when closed, Ladybird makes saving previous work possible through:

- Project object: when created, a project object stores configuration settings, genes filtering and new sets of genes generated during data analysis. The project object can be saved locally as **.ldb* file and opened later without going through steps A and B of figure 5.

- PDF printers: each plotting procedure offers the possibility of storing the figures output as a pdf file locally (or .png images).

- Procedures returning data columns can also be saved as .csv files.

3 Results

Reader companion

The explained methods in Materials and methods can be confusing in the sense that many tests and procedures are involved in the analysis of RNA-seq and Single Cell data (which are totally independent). To assist the reader throughout this section (and also for the sake of clarity), we present here an outline explaining the purpose of each paragraph and the steps involved in the analysis.

- RNA-seq data - Deseq2 Model:

  Purpose: *Highlight interesting genes affected by the treatment*

  1. Estimate **dispersion** and **log fold change (LFC)** between days according to the **Negative-Binomial** model. Use them to compute **(multitple) Wald tests** to highlight genes affected by the treatment.
  2. A high number of genes (7031 here) are returned by the test as significant (at 1%).
  3. Cluster this list of genes based on their kinetics using **DBSCAN** algorithm.
  4. To interpret the role of each cluster, **enrichment analysis** (hypergeometric test) is performed on all clusters separately to label them by the most significant and redundant pathway category.

- RNA-seq data - WGCNA - Model:

  Purpose: *Study co-expression of genes by analyzing their correlations and using over-representation tests to label each cluster of co-expressed genes.*

Reader companion

1. Construct Topological overlap dissimilarity matrix which, in brief, considers genes close if they are correlated with many common genes.
2. Cluster the genes based on this measure (Hierarchical clustering analysis) and form modules (by cutting the clustering dendrogram)
3. Perform **enrichment analysis** to discover the biological processes that distinguish clusters one from another other.

- Single-cell RT-PCR data:

  Purpose: *Highlight resistant subclones of cancerous populations. Here, we set another objective: develop a software application that includes visualization techniques that are intuitive for biologists: kernel density estimation, PCA, heatmaps and evolution curves. In the case where data follows the* ***Bernoulli-lognormal model:***

  1. Estimate the triplet (π, μ, σ^2) of each gene according to the **Bernoulli-Lognormal model.**
  2. Compute the **LRT** (likelihood ratio test) of frequency and mean (π, μ) between conditions to highlight genes that are affected by the treatment.
  3. Use these genes to cluster cells.

3.1 RNA-seq

Descriptive analytics

Table 3: *First four rows of cleaned and normalized RNA-seq count data.*

	Day 0, no TMZ[9]				TMZ, day 4			TMZ, day 9				TMZ, day 12				TMZ, day 16			
Gene | Sample	d0_1	d0_2	d0_3	d0_4	d4_1	d4_2	d4_3	d9_1	d9_2	d9_3	d9_4	d12_1	d12_2	d12_3	d12_4	d16_1	d16_2	d16_3	d16_4
TSPAN6	762.3	644.6	625.1	419.1	850.2	616.6	881.4	591.2	696.7	788.4	710.1	752.0	889.3	970.2	811.8	921.3	377.3	734.1	502.2
TPM1	1343.4	973.3	1004.2	883.5	1499.7	1825.9	1524.6	1312.8	1322.0	1465.6	1256.3	1136.6	1114.1	1553.6	1366.8	1703.8	1200.9	1260.1	1422.2
SCYL3	145.6	195.8	194.0	177.0	259.6	281.9	259.2	176.5	173.7	203.3	160.8	162.9	192.1	186.7	192.2	83.5	308.4	186.4	227.9
C1orf112	320.3	256.9	237.2	167.1	721.7	563.0	733.9	517.3	568.4	686.8	634.2	540.8	495.1	495.8	461.8	460.0	423.3	471.1	441.6

RNA-sequencing was performed on five different time points, four times each. *day 0* is the control condition (no Temozolomide). The remaining conditions are 4, 9, 12 and 16 days after the drug was introduced in the cell-line culture. Table 3 shows the normalized version of the data from where genes with extreme low counts (0 or 1) are filtered out. We find a matrix with 21 325 genes. Samples filtering excluded the fourth replicate of the second biological group (4 days after TMZ) most likely because of technical and/or sample-intrinsic problems; the fourth replicate showed:

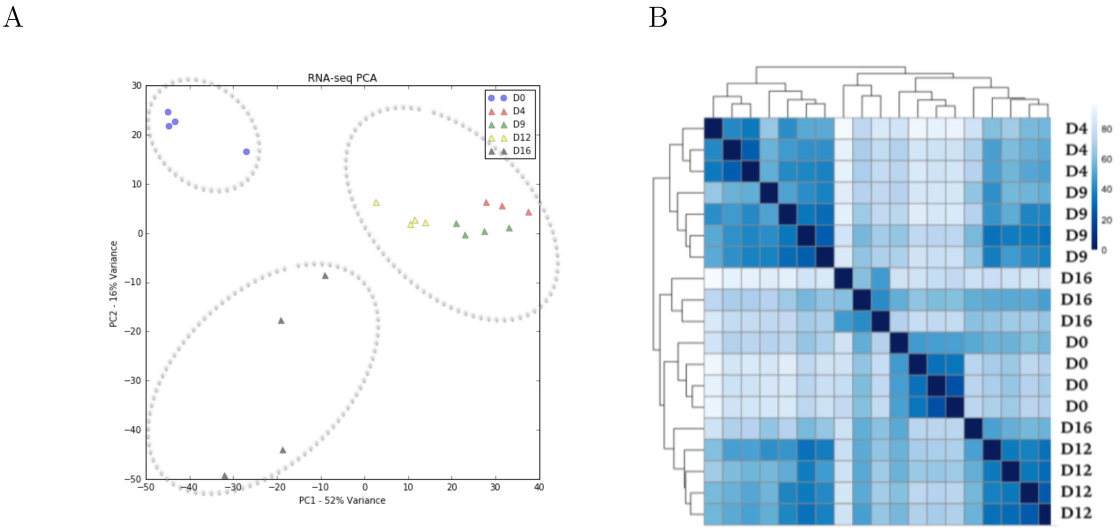


Figure 6: ***A.*** *PCA across samples.* ***B:*** *Hierarchical clustering performed on samples and plotted on top of a heatmap of euclidian distances.*

- abnormal read counts and GC content[10] differences compared to the other three replicates of the same condition.
- inferior sequencing quality than the remaining samples.

After vertical and horizontal filtering, dimension reduction procedures should make samples clusterized together. The variance of the log-count data being higher with lower counts would introduce a bias in determining PCA components. To address this, Deseq2 paper [11] introduced the *regularized log* transformation that removes the experiment-wide trend of variance over mean. The transformation is only used for visualization tools and is presented in [11].

Figure 6 shows PCA and HCA (hierarchical clustering analysis) of the normalized and cleaned data. The PCA plot explaining 68% of the variance reveals three separate *stages* going first from Day0 to (Day4, Day9, Day12) and then to Day16.

The four corners of the similarity heatmap confirms the proximity between samples of conditions 4, 9 and 12. Day 16 replicates are not as close as the other conditions which can be also seen in Figure A as they are more dispersed than the others.

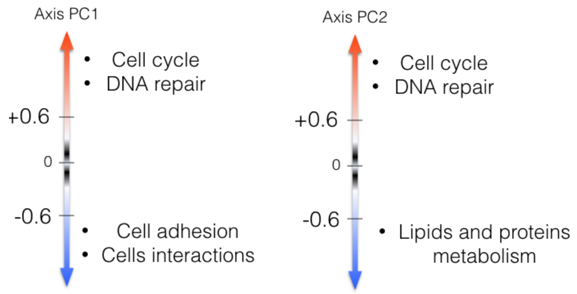


Figure 7: *PCA Axes interpretation in terms of correlated pathways. Groups of genes highly correlated are tested using Reactome pathways database.*

The PCA axes show interesting insight regard-

[10] DNA is composed of four types of molecules called nucleotides: Guanine(G), Cytosine(C), Adenine(A) and Thymine(T). To form the double-helix, G bonds with C and A bonds with T. GC content is the ratio $\frac{G+C}{A+T+G+C}$ which can bias gene expression profiling [31]

ing the group of genes they are strongly correlated with. The correlation of each gene with both axis is computed. Correlations higher than 0.6 (in module) are analyzed using over-representation tests of pathways. Figure 7 shows the categories of pathways positively and negatively correlated with the PCA axes.

Combining the PCA plot and Figure 7, we see that throughout the full expriment, the coordinates on the second axis decrease, which means that genes linked with DNA repair and proliferation are enabled meanwhile protein metabolism related genes are not. Axis one however goes up from stage 1 to stage 2 but then decreases to stage 3. The increase could be the result of the degradation of the extra-cellular matrix (hence, cells loss of adhesion and interactions) which was indeed observed during the experiments. The decrease is however not very clear and will be discussed after further analysis.

Supplementary Figure B.1 shows all pathways categories and their associated p-values.

Model estimates

Let Y_{ij} be the number of reads found in sample j and mapped to gene i assumed NB (Negative binomial, Deseq2 model) distributed with mean μij and variance σ^2_{ij}. The mean-variance trend assumption takes into account technical and biological variance:

$$\sigma^2_{ij} = \mu_{ij} + s_{ij}\mu^2_{ij}$$

where s_{ij} are size factors absorbing the differences between samples (about size factors: Supplementary Information A.2).

Figure 8 shows the presence of over-dispersion. The means of estimated gene-wise variance (red) follow the trend assumption (blue). The NB (negative binomial) model is suited for our data (*Materials and methods*, section 2.1). The fitted model here is a GLM[11]. First, dispersion estimation is carried out by shrinking MLE[12] estimates towards to mean-variance trend assumption. Dispersion estimates that are found far above (or below) the trend are considered not following the prior distribution and are not shrunk. For the sake of brevity, the dispersion plot is presented as Supplementary Figure B.3.

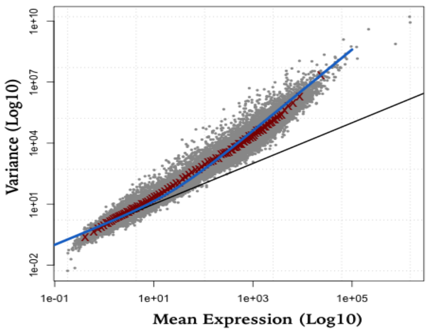


Figure 8: *EdgeR mean-variance plot. Gray: raw variance estimates of all genes in the filtered dataset on each condition. Red: Mean of the raw variance estimates per gene. Blue: Regression curve obtained when regressing variance estimates on mean expression estimates.* ***Black****: "Variance = Mean" curve (Poisson).*

Testing differential expression between two days A and B of gene i is performed by testing the null hypothesis of the correspondant $\underset{a\to b}{LFC}$ estimate (log fold change). LFCs estimation is carried out using a GLM model and a Bayesian shrinkage (section 2.1). The idea behind the shrinkage procedure is to avoid genes with low counts to have infinite or very high fold change ratios. Figure 9 compares shrunken and unshrunken LFC estimates of Day 4 over Day 0 by plotting LFCs against mean counts. Bayesian shrinkage improved considerably the test results as the genes with less than 10 reads no longer showed significant change. The remaining plots are provided in Supplementary figure B.2.

[11] Generalize linear model

[12] Maximum likelihood estimates

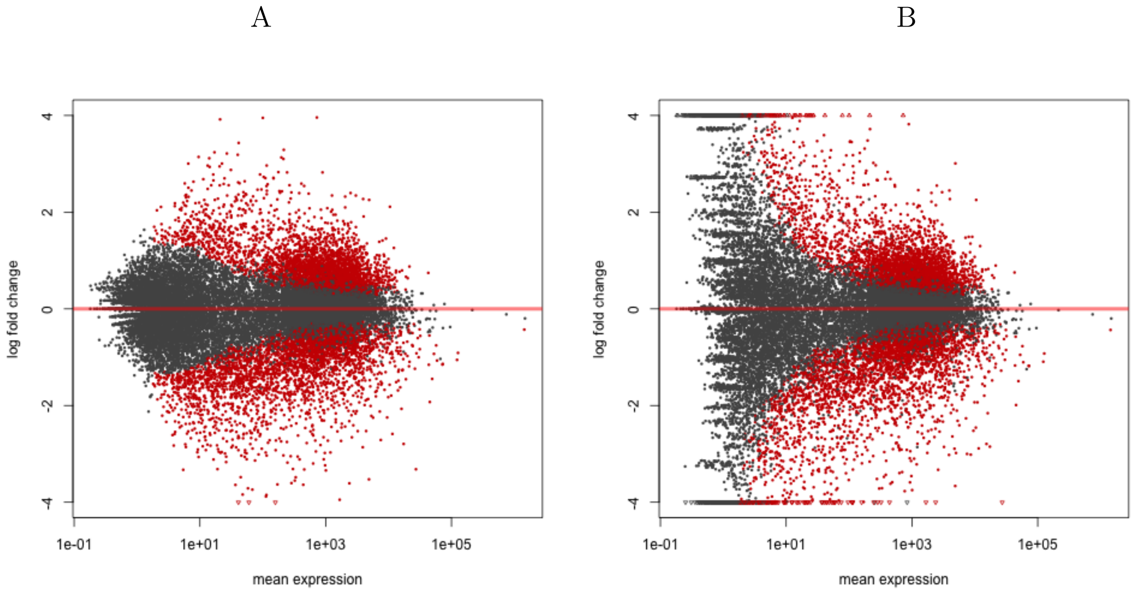


Figure 9: *Plots representing LFCs against mean counts; Day 4 over Day 0. Each dot is a gene. Red dots are differentially expressed at 0.01 significance (Wald test).* ***A:*** *Shrunken LFCs.* ***B:*** *Un-shrunken LFCs. The bayesian shrinkage reduces dramatically the number of false positives. Genes with low counts have no longer high LFCs.*

Wald tests

First, successive pair-wise comparisons are performed using Deseq, Deseq2 and edgeR. Deseq2 is the upgraded version of Deseq introducing GLM models and Bayesian shrinkage, it is expected to be less stringent on tests decisions. EdgeR however uses nearly the same model as Deseq2 with a few differences on how normalization is performed [10, 11, 13]. These similarities are shown in Figure 10.

EdgeR and Deseq2 do not only agree on the number of DE (differentially expressed) genes but also on the set of differentially expressed genes: a mean of 94% overlapping genes is found when comparing the four couples of sets. The first and the last transitions (0 to 4, 12 to 16) seem to be the most important as they witnessed the most significant change in gene expression. Which was also observed in data visualization plots.

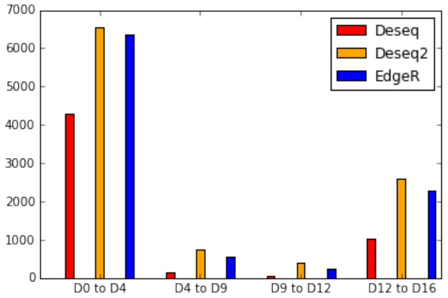


Figure 10: *Number of differentially expressed (DE) genes at 0.01 significance using Wald tests to compare the four pairs of successive days across packages.*

In what follows, only Deseq2 tests will be analyzed.

Figure 11 generalizes Wald tests to all possible pairs of conditions showing how can the period 4-12 be considered as a transition state between *Day 0* and *Day 16* which witness the most significant change in gene expression.

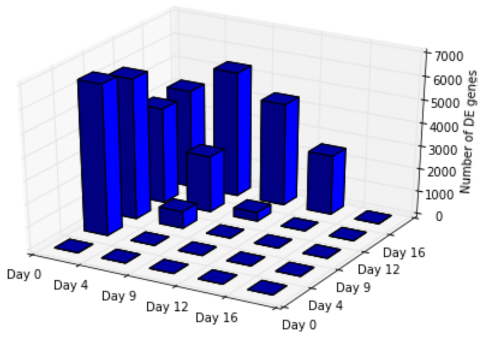


Figure 11: *Number of differentially expressed genes at 0.01 significance using Wald tests to compare all possible pairwise combinations in a 3D truncated symetrical matrix*

Biological insight

Pathways: time-course

Since tests yielded thousands of differentially expressed genes, interpreting the results by looking up each gene could be daunting or even misleading since genes are known to interact with each other. Reactome[13] pathways over-representation tests are performed, first on the three main evolutions and then on separate clusters of genes based on their kinetics.

Figure 12 shows how significantly are proliferation (Cell cycle, Mitosis, S phase, G1/S Transition, DNA replication) and DNA repair (Activation of ATR) involved in the first and last pairs of days. The statistical proximity between samples Day 4, 9 and 12 observed in the PCA plot and tests histograms is now sustained by biological evidence: during the phase Day 4 - Day 12, only three pathways are significant at 10% and all of them are related to metabolism of lipids (Supplementary figure B.4).

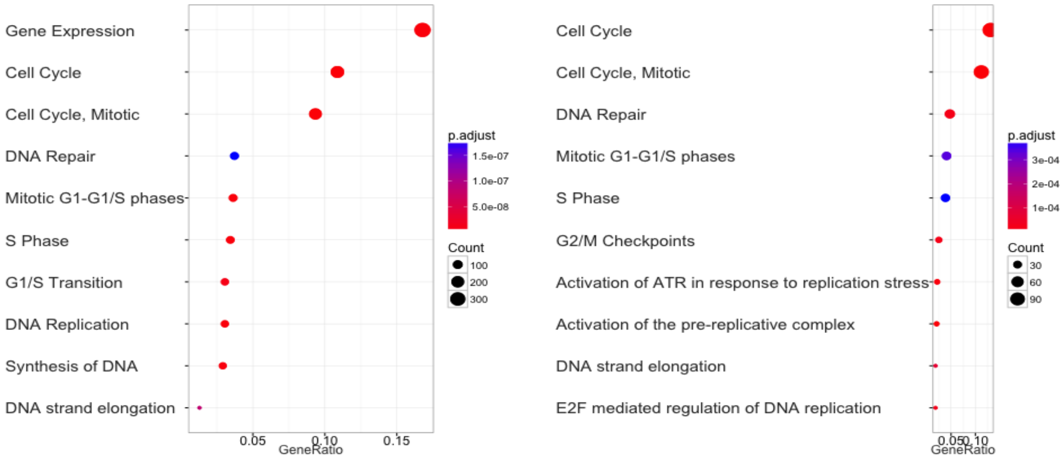


Figure 12: *Over representation tests; first 15 categories of pathways highlighted in three main evolutions.* ***Left:*** *Day 0 to Day 4.* ***Right:*** *Day 12 to Day 16.* ***p.adjust:*** *FDR adjusted p-values.* ***Gene Ratio:*** *proportion of mapped genes to the each pathway.* ***Count:*** *number of genes. Tests comparing Day 4 to Day 12 are not shown here as they yielded poor results: too vague categories related to metabolism of lipids. They can however be consulted in Supplementary figure B.4.*

Pathways: genes kinetics

Genes having at least one significant change are the subject of the following analysis. In order to take into account slow evolutions, signif-

[13] Reactome: an online database of reactions, pathways and biological processes.

icant and non-successive changes are used to fill in the blanks. For example, if a gene's level "statistically remains constant" between 0 and 4 and 4 and 9, the log fold change between 0 and 9 is studied. The remaining non significant log fold changes are replaced by zero. The cumulative sum of log fold changes per gene is computed to display the genes' significant evolution. Based on this dataset, a clustering is performed to spot groups of genes that behaved similarly throughout the experiment. DBSCAN[14] algorithm is used for this purpose for its ability to detect uneven-sized clusters of genes based on their proximity (Euclidean distance).

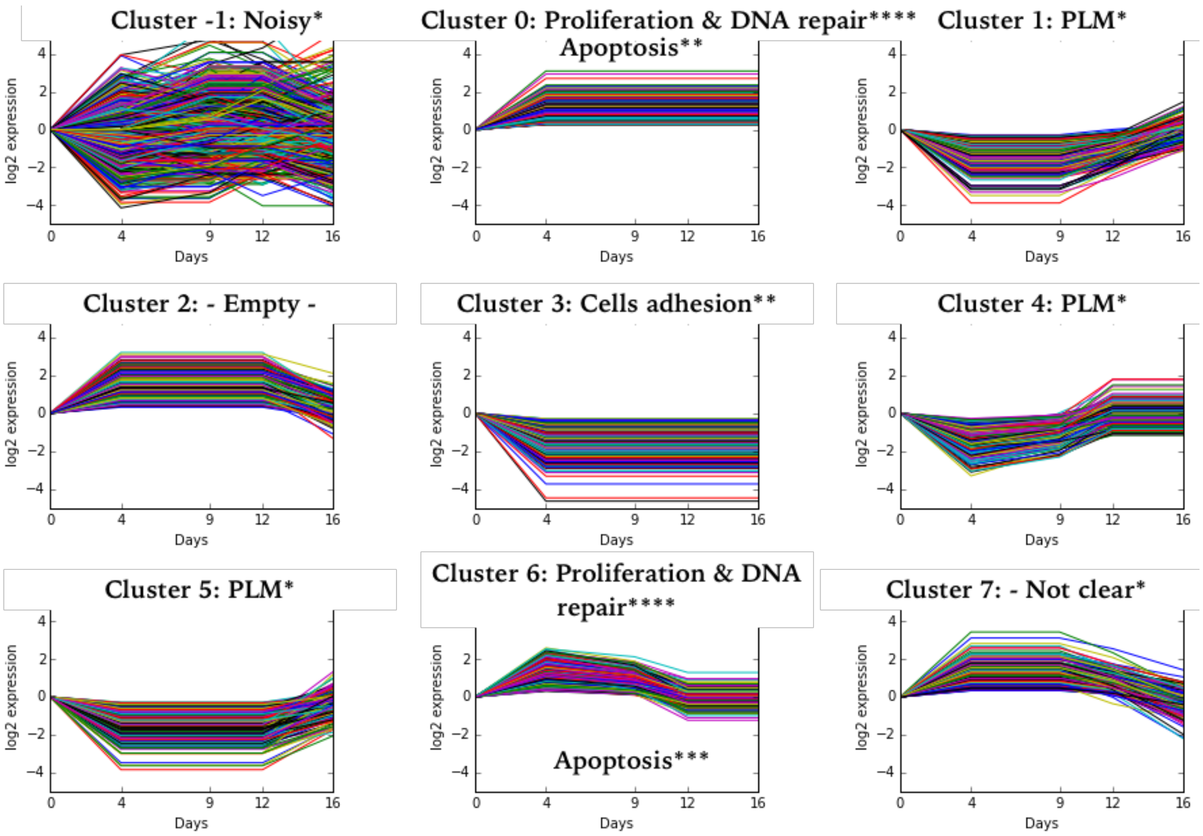


Figure 13: *Clusters of genes returned by DBSCAN algorithm. Clustering is applied on LFC estimates. For the comparison to be "fair", the evolution previous to the experiment is considered null. Cluster -1 regroups the genes that didn't have enough neighbors to form a cluster. Their evolution is too specific to represent a general behavior. For each cluster, pathways enrichment test is computed. The major pathway category is displayed on top of each figure. The number of stars denotes the level of significance: p-value* $= 10^{-\#*}$ *where* $\#*$ *is the number of stars.* ***PLM:*** *protein and lipid metabolism.*

The main categories of pathways displayed in Figure 13 reveal an interesting concordance of gene behavior and gene function. All of the three clusters that are related to protein and lipid metabolism show a significant decrease followed by a significant increase. This can be seen as a response to stress induced by the chemotherapeutic drug. Clusters 0, 3 and 6 confirm the interpretation of the PCA axis 1 (Figures 6, 7) as the coordinates of the axis follow the same pattern of

[14] Density-based spatial clustering of applications with noise, see *Materials and methods*

figure: increase from stage one[15] to stage two[16] and then decrease to stage three[17]. The increase is the combined effect of:

- Genes related to DNA repair and cell cycle go up and they are positively correlated with the first axis.
- Genes related to Cells adhesion and communication go down and they are negatively correlated with the first axis

The following decrease of the first axis coordinates seems to be the result of the decrease observed in cluster 6.

Cluster 2 didn't show any specific pathway category: there was no match between the cluster of genes and the pathways database.

Cluster 7 lists some pathway categories without any obvious general category.

It is important to note that even if cluster -1 is considered *noisy*, it contains the key gene in Glioblastoma MGMT. Which is not surprising since MGMT is one of the very few genes (EMB, TSPAN8, MMP7, CAPN6) that increase significantly in the last three evolutions.

Details of these tests can be consulted in Supplementary Figures B.5.

Table 5 suggests a list of genes as potentially interesting in further research. Genes of the three clusters -1, 0, 3 and 6 are ranked by their variability (sum of their absolute LFCs).

	GENE	INTEL
Cluster -1	MMP3	Breakdown of ECM[18] - Tumor initiation [32]
	MGP	Migration - ECM breakdown [33]
	CD22	Regulation of immune response [32] - Effect on GBM survival [34]
	CASS4	Local adhesion integrity, and cell spreading [32]
	SLAMF7	Both innate and adaptive immune response [32]
	HAS3	Abnormal biological processes such as transformation and metastasis. [35]
	TMEM140	Pprognosis of glioma by promoting cell viability and invasion [36]
	MGMT	- cf *Introduction*
	DNER	Notch2 activation [32] - regulates glioblastoma-derived neurosphere cell differentiation and tumor propagation [37]
	AP3B2	neuron-specific functions [32]
	RGS7	-
	CCDC64	Regulator of neurite outgrowth [32]
	LTA	Member of the tumor necrosis factor family - Plays a role in apoptosis [32]
	DBMT1	Interaction of tumor cells and the immune system [32] - Tumor suppressor gene [38]

[15] Day 0
[16] Days 4, 9 and 12
[17] Day 16
[18] Extra-cellular matrix

Cluster 0	INPP4B	Promotes melanoma cell proliferation independently of Akt through activation of PI3K/SGK3 signallin [39]
	EFEMP1	Suppresses malignant glioma growth and exerts its action within the tumor extracellular compartment [40]
	C1QTNF9-AS1	non-coding RNA gene [32]
	KCNK2	Two-pore-domain background potassium channel protein family [19] [32]
	TNFSF13	Tumor necrosis factor [32]
Cluster 3	CCR7	Migration of memory T cells to inflamed tissues - Stimulates dendritic cell maturation[32]
	LEMD1	-
	PTP4A3	Target for inhibition of cell proliferatin, migration and invasion [42]
	FER1L4	Suppresses cancer cell growth [43]
	JPH3	Intracellular ion channels [32]
Cluster 6	DNAH12	-
	CPNE4	Membrane trafficking, mitogenesis and development [32]
	CDC6	Cell cycle, initiation of DNA replication [32]
	BTLA	modulation of T cell responses [32]
	SFN	Potent apoptotic effects and invasion inhibition effects [44]

Table 5: *Strongest genes of DBSCAN clusters, ranked by sum of absolute LFCs.*

3.2 RNA-seq: WGCNA

WGCNA[20] is a model of gene co-expression networks. In brief, it clusters in modules genes with very similar connections. The similarity of the connections is soft: it takes into consideration both which set of genes is involved and the strength of their correlation. WGCNA was applied on normalized RNA-seq data (Deseq2 normalization scheme).

The power parameter β was assessed by the scale-free topology criterion. Figure 14 shows that the data hardly form a scale-free network: the fitting index R^2 reaches 0.8 when the mean connectivity of the network shrinks down to almost zero.

So as not to lose all the network's information, we settled for $R^2 = 0.7$ i.e $\beta = 6$. One possible explanation of this failure is the high variability of genes between biological conditions: only a few gene hubs share very strong correlations, the number of nodes with k connections $P(k)$ does not follow an exponential distribution.

The hierarchical clustering cut threshold is set at a height that ensures module sizes greater than

[19] Gliomas display enhanced glycolysis and heightened acidification of the tissue interstitium. Two-pore potassium channels (K2Ps) is one of the brain tumors pH-sensitive ion channels [41].

[20] Weighted genes co-expression network analysis, see *Material and methods*

30 genes. Figure 15 shows the hierarchical tree and a eigengene correlation heatmap.

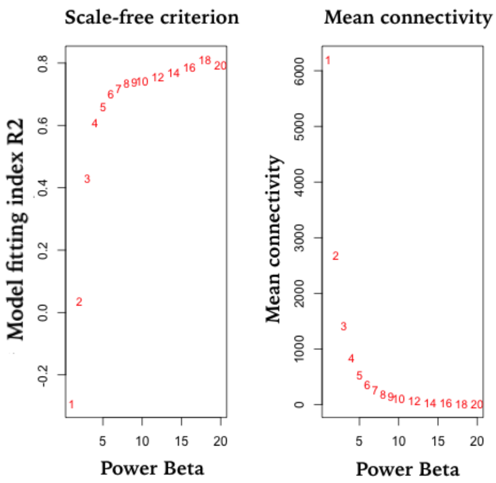


Figure 14: ***Left:*** *R^2 of the regression model: $\log(P(k))$ on $\log(k)$ which is supposed to be close to 1 for a reasonable value of β. By reasonable, we mean that β should not be too large so that the mean connectivity still ensures a certain level of network information.* ***Right:*** *Mean connectivity of the network, computed as the mean of the topological overlap matrix.*

Unlike differential expression clusters obtained using the Deseq2 model, over-representation tests applied on each eigenmodule yielded outrageous pathways categories. For instance, the two *a priori* interesting modules in the middle and the right showing strong positive correlations showed a mixture of pathways: proliferation, DNA repair, protein metabolism, apoptosis regulation, extracellular matrix organization with the same order of significance.

In such cases where data are heterogenous, WGCNA authors [19] recommend to build independent networks on each group and study *consensus* modules. This approach is similar to that of differential expression insofar as it compares data separately pairwise. In WGCNA, the comparisons are performed on module correlations to answer the question: to what extend co-regulated genes conserve their co-expression and how does it evolve ?

Our data do not qualify for this study since correlations with less than 15 samples would be too noisy [19].

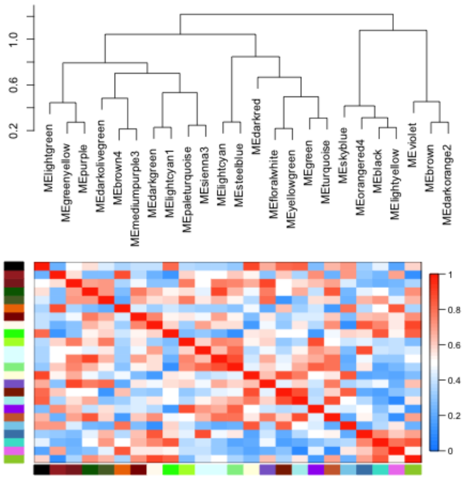


Figure 15: ***Top:*** *Hierarchical tree of the eigenmodules, colors are arbitrary.* ***Bottom:*** *Eigengenes correlation heatmap. The modules do not show any obvious proximity even if they are close in the clustering tree.*

3.3 Single Cell application

Previous single cell experiments (by the CRCNA team of Nantes[21] involved 96 genes that are directly or indirectly linked with cancer in general. We present in this section how our PyQt application *LadyBird* was used to analyze the provided data.

[21] Centre de Recherche en Cancérologie Nantes - Angers

Example of analysis

After filtering out genes and wells where technical problems occurred, the following data shapes are obtained:

Day 0: 75 cells × 89 genes

Day 4: 79 cells × 89 genes

Day 9: 44 cells × 89 genes

Day 12: 77 cells × 89 genes

Day 16: 56 cells × 87 genes

PCA on single cell data (Figure 16) shows a similar proximity between Day 0 and Day 16 to that of RNA-seq data. PCA within groups however did not reveal any particular clones of cells as we were hoping to see. Clustering (HCA) using the recommended [18] distance (*rank magnitude*, *Materials and methods*) trees and dissimilarity heatmaps were also in favor of a dynamic homogenous population changing throughout the experiment as an entire group.

Perhaps most genes do not exhibit strong discriminatory expression patterns but what if some of them do and their influence is shadowed by the *global* expression behavior?

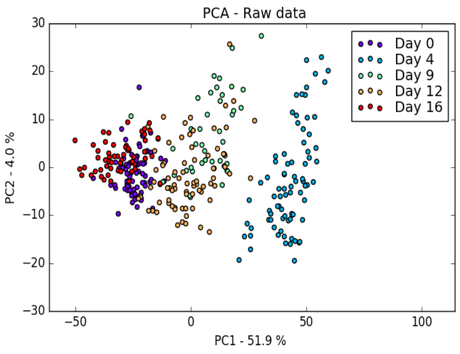


Figure 16: *Ladybird: PCA on raw single cell data.*

We argue here that genes responsible for drug tolerance are effected by the treatment through time [45]. To select these genes, we use the Bernoulli-lognormal model described in *Materials and methods*. The QQ[22] plots in Supplementary Figures B.6 suggest that the said model is suited to our data.

After estimating model parameters (π, μ, σ) of each gene at each time, we perform the following pairwise LRT (likelihood ratio test) tests of the Bernoulli lognormal model, as defined in *Materials and methods* and eliminate genes with a p-value greater than 0.001:

$$\begin{cases} H_0 : (\pi_0, \mu_0) = (\pi_4, \mu_4) \\ H_0 : (\pi_4, \mu_4) = (\pi_9, \mu_9) \\ H_0 : (\pi_9, \mu_9) = (\pi_{12}, \mu_{12}) \\ H_0 : (\pi_{12}, \mu_{12}) = (\pi_{16}, \mu_{16}) \end{cases}$$

The four tests combined filtered out 60 genes or so per group. Applying PCA to the filtered set of genes led to the emergence of two separate subpopulations of cells in Day 4 (Figure 17 - A). To investigate what genes are responsible for that division, we plotted an expression heatmap (Figure 17 - B). EGF and ADAMTS are some of the few genes showing very different expression patterns across cells. Indeed, when computing gene correlations with the PCA components, we find that the strongest correlations (in module) are -0.98, -0.39 and 0.3 that respectively correspond to EGF, EMP1 and ADAMTS.

Let's have a look at the distribution of theses genes throughout the experiment. From Figure 18, we notice that:

- The distribution shift seems to be the same across genes, even if their functions are different: EGF and EMP1 are believed to promote and regulate cell growth; ADAMTS codes *ECM protease* (enzyme that breaks protein chains).

- The discriminatory genes of the heatmap (ADAMTS and EGF) are not expressed in a subpopulation.

[22] Quantile-quantile plots

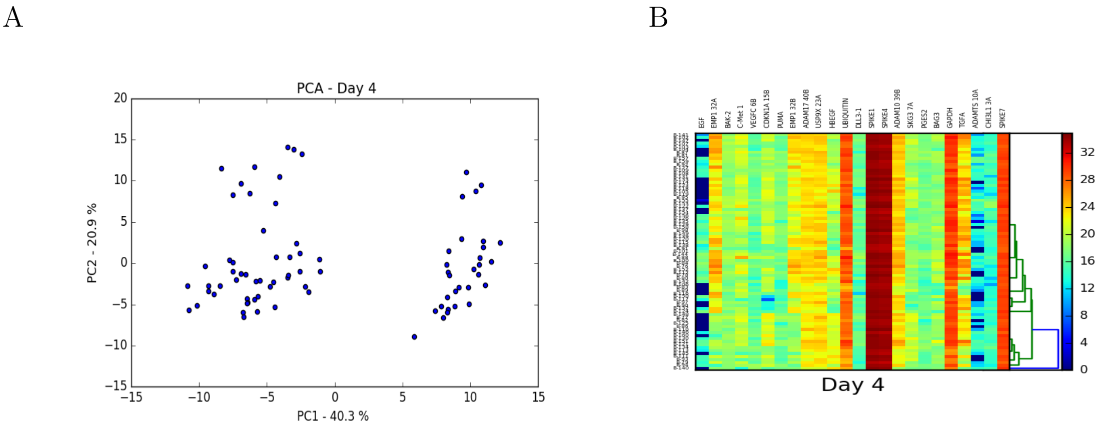


Figure 17: *Ladybird: After LRT genes filtering, on group of day 4:* ***A.*** *PCA showing two clones of cells. The value on the axes labels is the part of total variance explained by the correspondant axis.****B:*** *Hierarchical clustering using Rank Magnitude distance and expression heatmap. Rows represent IDs of wells, they are not relevant to the study. Columns are genes after filtering. The color bar values correspond to centered expression values.*

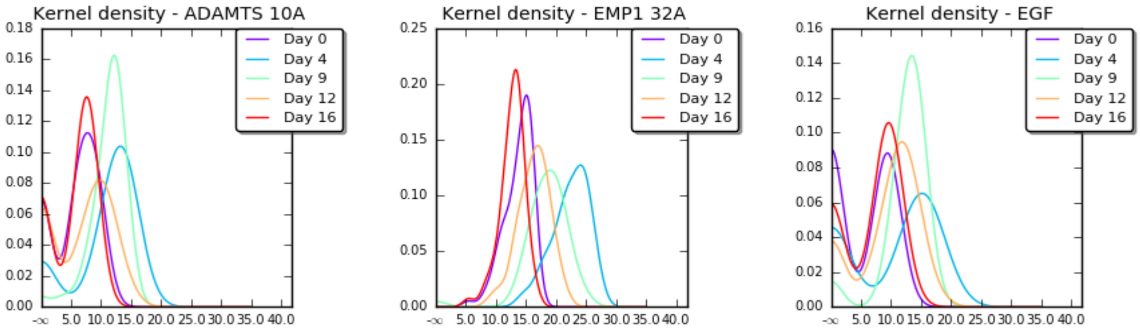


Figure 18: *Ladybird: Kernel density estimation of the genes EGF, EMP1, ADAMTS*

Normalization review

The first remark leads to the question: what if the observed shift is not biologically interesting but is rather due to a technical artifact ?

Supplementary Figures B.8 show that the same distribution shift in the majority of the remaining genes. The variation (if due to technical artifacts) *should* have been eliminated by the normalization scheme. Distributions suggest that it may not be the case.

Normalization across groups was performed by an IGBMC team using Spikes (Supplementary information A.3) and is not carried out by LadyBird.

Spikes are known DNA quantities added in each well. Their variation through time is due to experimental bias and is not relevant to the study. Figure 19 shows the distribution of spikes before and after normalization. In Figure 19 B, we can see that distribution of Spike1 did not change throughout the experiment, it should have not been part of the normalization scheme. The observed bias is even higher after normalization in all spikes.

Data visualization without modeling

Whether the observed distributions shift is due to relevant biological processes or normalization bias is a tough question to answer since we do not have any experimental measures or observations to confirm or deny the genes expression evolution.

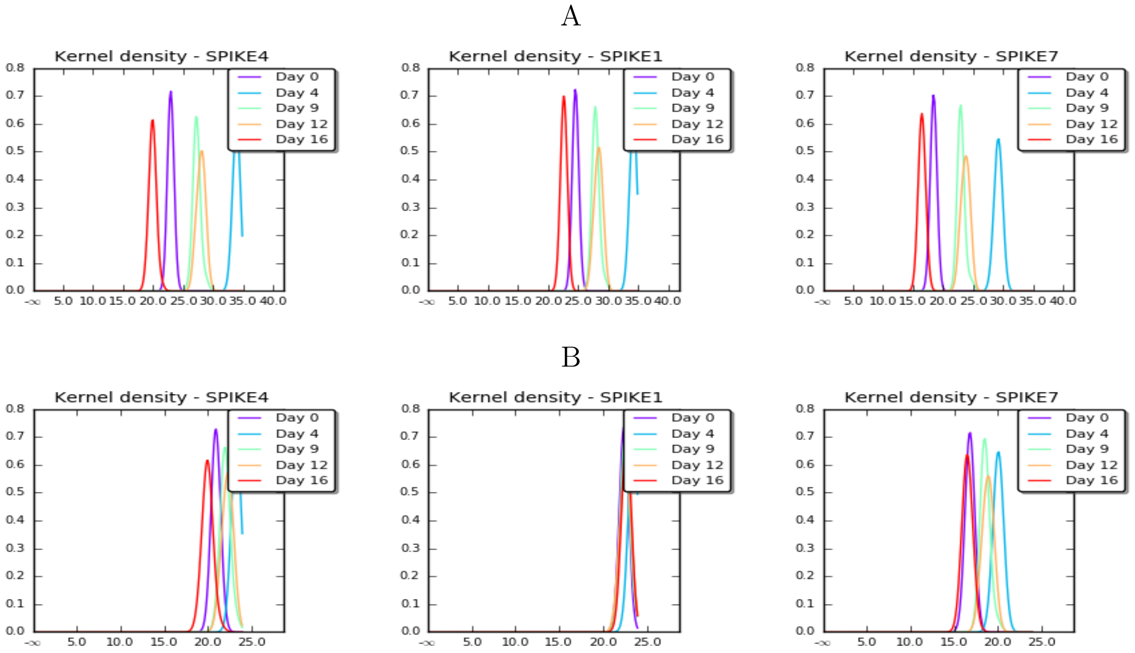


Figure 19: *Ladybird: Kernel density estimation of Spikes.* ***A:*** *Normalized data* ***B*** *Raw data. The distributions shift observed in B is an artifact. Spikes vary more in normalized data than in raw data !*

Yet normalization does not interfere within groups. We can still focus on the first observation above and keep genes with a frequency of expression not too low, nor too high. It is important to note that this filtering does not rely on the bernoulli-lognormal model.

Filtering $0.1 \leq \pi \leq 0.9$ kept around 25 genes per group. PCA did not show any particular subpopulations of cells. In Heatmaps and HCA of Figure 20 we can see that the discrete part of the genes distributions led to an overfitting: cells have such different expression patterns that they formed too many clusters with very few members.

The results of these different approaches suggest that provided data cannot be analyzed by comparing groups: normalization needs to be improved. Clustering cells and performing data visualization techniques without labels is very ambitious: measures and preliminary observations are necessary in single cell studies as they can reveal interesting behaviors (or at least eliminate some scenarios).

Other Ladybird functionalities

The previous section showed how LadyBird can be used in practice. But what if the data violate the bernoulli-lognormal assumption ? One could still perform Kolmogorov-Smirnov 2-samples test to compare biological conditions. It is also possible to plot the evolution of mean, frequency and variance of expression through time. Histograms and Kernel density estimation could bring some insight on the genes distribution evolution. Supplementary Figures B.9 show screenshots of Ladybird main windows.

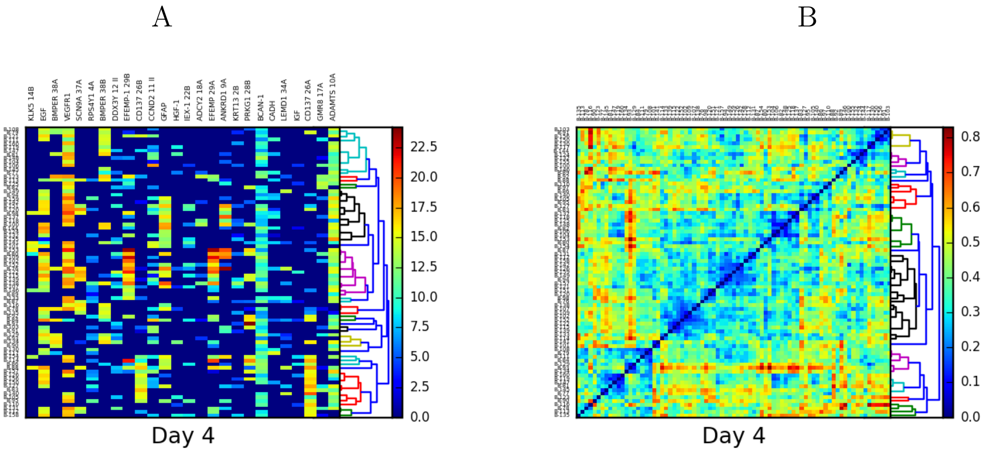


Figure 20: *Ladybird: HCA clustering using rank magnitude distance on Day 4 data* ***A:*** *Expression heatmap.* ***B:*** *Dissimilarities heatmap*

4 Discussion

RNA-seq

Investigating drug resistance in Glioblastoma remains a difficult task because of the multiple and various mechanisms it operates through which are only partially understood. Here, DBSCAN clustering and Reactome tests suggest combined scenarios where after inducing TMZ (Temozolomide), GBM (Glioblastoma) cells undergo reversible and irreversible transformations:

reversible :

- Protein and lipid metabolism: GBM cells abandon protein and lipid metabolism temporarily due to stress. Concerned clusters contain some reference housekeeping genes (CHMP, PGK1, VPS29) [46].
- Proliferation, DNA repair, apoptosis: a fraction of genes linked with these processes increase and decrease significantly during the experiment.

irreversible :

- Cells adhesion: GBM cells go through breakdown of extracellular matrix and do not retrieve their adhesion which was observed in the microscope weeks after inducing TMZ.
- Proliferation, DNA repair, apoptosis: another fraction of these genes change significantly only in the beginning and remain at a high level of expression.

Yet we cannot state that the observed signs of evolution correspond to real genes expression kinetics: genes can participate in both positive and negative regulation of biological processes, their expression level is not *a priori* a perfect predictor of what happens within cells.

Another downside of our approach is the low mapping rate (10%-15%) of genes with Reactome pathways database: 7031 genes are declared differentially expressed in our study and only 6750 genes have been annotated in the reactome database. For example, all suggested target genes in Table 5 (except DNER, DBMT1, CDC6) were not found in Reactome, suggesting that many other hidden genes could be potential key genes in fighting GBM resistance. Such experiments must be performed on different cell cultures in order to narrow down possibilities and eliminate outliers.

We do not recommend the use of WGCNA on few samples with such strong biological differences between groups. Ideally, when differential expression is studied, one should build a specific network for each group (with at least 20 samples [23]) and compare *consensus modules* that are shared between groups. The study would then be about the evolution of correlations.

Single-cell

The main purpose of the single cell experiments was to detect (and explain) the heterogeneity of tumor cells: detect a drug-resistant population and investigate the biological processes responsible for the emergence of such cells. But it is very unlikely to be achieved without a prior knowledge on the selected genes in the experiment. Here, genes were chosen based on insight that does not involve resistance in particular.

[2] is a valuable review of all published literature about drug resistance in GBM. Mechanisms enlisted are: hypoxia, drug efflux, DNA repair, miRNAs and stem-like cells. Each set of particular genes interactions promote a certain mechanism of resistance. Unfortunately, provided single cell data contains only one gene per mechanism or none at all. Given that subpopulation clones in GBM are not precisely characterized, we suggest more convenient experimental techniques such as single cell RNA-seq where the whole genome is sequenced or a high-complexity barcoding. Barcoding technique allows to track cancer subclones at a resolution of 1 in a million without prior knowledge of the underlying biological mechanisms [47]. Otherwise, with RT-PCR technique the 96 genes should be picked with the intention of investigating a particular drug resistance mechanism instead of *discovering resistance mechanisms.*

The evolution of the genes mentioned in the GBM resistance review [2] can be found in Supplementary Figure B.10.

To compare genes distribution between conditions, normalization needs to be readdressed. The use of median in the normalization formula is not justified: Spikes display a very narrow spectrum of values and no outliers can be found in Spikes measurements. PCR related literature states that there is no need in normalizing the data as its inherently normalized by isolating cells. It is important to note that normalization must be studied with care as it can skew results.

About Ladybird: any software application for data analysis must provide a vast spectrum of possibilities as it is practically impossible to anticipate every purpose of every project. The Bernoulli-Lognormal model is very suited to our provided data but what if it is not the case with future studies ? The non-parametric procedures are to the rescue to test and study the evolution of the distributions from a condition to another. At the time of writing, Ladybird is convenient for investigating RT-PCR single cell data which does not rely on the analysis of external information other than gene expression measurements.

Tutorials of Ladybird can be found at *this link* [48]

Experimental protocol

The high number of genes highlighted by RNA-seq can be diminished by improving the experimental protocol. We make two suggestions.

- Perform the same measures (on the five days) on another culture of cells in which no TMZ was induced. These samples would provide a better control sample for each day, eliminating the variability of genes that is not due to TMZ. We could then apply the model *DyNB* [49] that analyzes the data as a time series (Dynamic negative binomial).
- Induce TMZ a second time. Acquired drug resistance is a tumor that initially responded to a treatment but is no longer sensitive to the drug [2]. Repeating the experi-

ment after Day 16 would lead to more accurate results by eliminating irrelevant genes.

Conclusion

Our work provides a general view and example of use of the most common tools in analyzing gene expression data: Deseq2 and WGCNA in RNA-seq data. Single cell RT-PCR data however lack a general consensus modeling in the literature. We developed for that purpose an open-source framework to visualize and perform statistical procedures on single cell data called *LadyBird*.

Our analysis of RNA-seq data showed interesting results. We suggest a list of potential target genes linked with resistance in GBM, some of which have not been linked with glioblastoma in previous studies. These findings concern the U251-MG cell line and need to be confirmed by future work.

Abbreviations

TMZ: Temozolomide **WGCNA:** Weighted genes correlation network analysis **RT-PCR:** Real-time polymerase chain reaction **NGS:** Next Generation Sequencing **DGE:** Differentia gene expression **GLM:** Generalize linear model **LFC:** Log fold change **MLE:** Maximum likelihood estimates **FDR:** False discovery rate **DBSCAN:** Density-based spatial clustering of applications with noise **TOM:** Topological overlap measure **CDF:** Cumulative distribution function **HCA:** Hierarchical clustering analysis **PCA:** Principal component analysis **NB:** Negative binomial **BL:** Bernoulli lognormal **QQ:** Quantile-quantile **LRT:** Likelihood ratio test

Acknowledgements

I would like to thank my internship supervisors Marie Doumic and Jean Clairambault for allowing me to be part of their research team and for their valuable directions. I am also grateful for the team of François Vallette who welcomed me in their labs in Nantes and helped me understand the required biological knowledge in this project.

References

[1] Jackson D. Hamilton et al. Glioblastoma multiforme metastasis outside the CNS: Three case reports and possible mechanisms of escape. *American Society of Clinical Oncology*, 2014.

[2] Catherine P. Haar et al. Drug resistance in glioblastoma: A mini review. *HHS Public Access*, 2012.

[3] Anthony H. and V. Schapira. *Neurology and Clinical Neuroscience*. Mosby Elsevier, 2007.

[4] Chisholm et al. Emergence of drug tolerance in cancer cell populations: An evolutionary outcome of selection, nongenetic instability, and stress-induced adaptation. *Cancer Research*, 2015.

[5] Monika E. Hegi et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, 2005.

[6] Felix Schmidt and Thomas Efferth. Tumor heterogeneity, single-cell sequencing, and drug resistance. *Pharmaceuticals*, 2015.

[7] Anders Ståhlberg et al. RT-qPCR work-flow for single-cell data analysis. *Elsevier*, 2012.

[8] Matthias Farlik et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Elsevier*, 2015.

[9] Jeffrey Martin Zhide Fang and Zhong Wang. Statistical methods for identifying differentially expressed genes in RNA-seq experiments. *Cell and Bioscience*, 2012.

[10] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 2010.

[11] Wolfgang Huber Michael Love and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 2014.

[12] Marek Gierlińsk et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, 2015.

[13] Aaron T. L. Lun Yunshun Chen and Gordon K. Smyth. Differential expression analysis of complex RNA-seq experiments using edger. *Statistical Analysis of Next Generation Sequence Data, Somnath Datta and Daniel S Nettleton (eds), Springer, New York, pages 51–74. Statistical Analysis of Next Generation Sequence Data, Somnath Datta and Daniel S Nettleton (eds), Springer, New York, pages 51–74.*, 2014.

[14] S. Dudoit Y. Ge and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *The Walter and Eliza Hall Institute of Medical Research, Australia*, 2003.

[15] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[16] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, pages 171–196., 1999.

[17] Jiirg Sander Martin Ester, Hans-Peter Kriegel and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Institute for Computer Science, University of Munich*, 1996.

[18] Ricardo JGB Campello Pablo A Jaskowiak and Ivan G Costa. On the selection of appropriate distances for gene expression data clustering. *Bioinformatics*, 2014.

[19] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *Bioinformatics*, 2008.

[20] M Achermann and K Strimmer. A general modular framework for gene set enrichment analysis. *Bioinformatics*, 2009.

[21] G. Yu and He Q. ReactomePA: an r bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12:477–479., 2016.

[22] Yan G Yu G, Wang L and He Q. DOSE: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 2015.

[23] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC Bioinformatics*, 2007.

[24] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 2005.

[25] Andrew McDavid et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, 2012.

[26] Martin Bengtsson et al. Gene expression profiling in single cells from the pancreatic islets of langerhans reveals lognormal distribution of mRNA levels. *Genome Research*, 2005.

[27] Larry A. Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer, 2004.

[28] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2006.

[29] B.W. Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, 1986.

[30] Campello R.J.G.B. and Hruschka E.R. On comparing two sequences of numbers and its applications to clustering analysis. *Elsevier*, 2009.

[31] Davide Risso et al. GC-Content normalization for RNA-seq data. *BMC Bioinformatics*, 2011.

[32] Genecards: The human gene database, 1997.

[33] Sonja Mertsch et al. Matrix gla protein (MGP): an overexpressed and migration-promoting mesenchymal component in glioblastoma. *BMC Cancer*, 2009.

[34] Elodie Vauléon et al. Immune genes are associated with human glioblastoma pathology and patient survival. *BMC Medical Genomics*, 2012.

[35] Hee-Jin Kwak Jong Bae Park and Seung-Hoon. Role of hyaluronan in glioma invasion. *Cell adhesion and migration*, 2008.

[36] Bin Li et al. TMEM140 is associated with the prognosis of glioma by promoting cell viability and invasion. *Journal of hematology and oncology*, 2015.

[37] Peng Sun et al. DNER, an epigenetically modulated gene, regulates glioblastoma-derived neurosphere cell differentiation and tumor propagation. *Stem cells*, 2009.

[38] Pang JC et al. Mutation analysis of DMBT1 in glioblastoma, medulloblastoma and oligodendroglial tumors. *International Journal of cancer*, 2003.

[39] Meng Na Chi et al. INPP4B is upregulated and functions as an oncogenic driver through SGK3 in a subset of melanomas. *Oncotarget*, 2015.

[40] Hu Y et al. EFEMP1 suppresses malignant glioma growth and exerts its action within the tumor extracellular compartment. *Mol cancer*, 2015.

[41] Avinash Honasoge and Harald Sontheimer. Involvement of tumor acidification in brain cancer pathophysiology. *Frontiers in Physiology*, 2013.

[42] Wang L et al. PTP4A3 is a target for inhibition of cell proliferatin, migration and invasion through akt/mTOR signaling pathway in glioblastoma under the regulation of mir-137. *Brain Research*, 2016.

[43] Tian Xia et al. Long noncoding RNA FER1L4 suppresses cancer cell growth by acting as a competing endogenous RNA and regulating PTEN expression. *Scientific Reports*, 2015.

[44] Zhen Zhang. Sulforaphane induces apoptosis and inhibits invasion in U251MG glioblastoma cells. *Springer*, 2016.

[45] Jason P. Glotzbach et al. An information theoretic, microfluidic-based single cell analysis permits identification of subpopulations among putatively homogeneous stem cells. *Plos One*, 2011.

[46] Eisenberg E and Levanon EY. Human housekeeping genes, revisited. *Trends Genetics*, 2014.

[47] Hyo eun C Bhang et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature America*, 2015.

[48] Github page www.janatih.github.io, October 2016.

[49] Tarmo Äijö et al. Methods for time series analysis of RNA-seq data with application to human th17 cell differentiation. *Bioinformatics*, 2014.

[50] Papoulis. *Probability Random Variables and. Stochastic Processes,.* McGraw-Hill Higher Education, 2002.

Supplementary material

A Information

A.1 Binomial approximation

Theorem: Poisson limit theorem

Let $X \sim \mathcal{B}(n,p)$.
If $n \to \infty$ and $p \to 0$ such that $np \to \lambda, \lambda \in \mathbb{R}_+^*$.
Then X can be approximated by a random variable $Y \sim \mathcal{P}(\lambda)$.

Proof [50]:

Supposing the statement's assumptions, let's show that X and Y have asymptotic equivalent mass probability functions.

Using Sterling's factorial approximation $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ and asymptotic analysis properties (since the denominator is nonzero for all $n \geq k$):

$$\Pr(X = k) = \frac{n!}{(n-k)!k!}p^k\left(1-p\right)^{n-k} \underset{n\to\infty}{\sim} \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}k!}p^k\left(1-p\right)^{n-k}$$

$$\underset{n\to\infty}{\sim} \frac{n^n p^k\left(1-p\right)^{n-k}}{\left(n-k\right)^{n-k}e^k k!}$$

And using $np \to \lambda$:
$$\frac{n^n p^k\left(1-p\right)^{n-k}}{\left(n-k\right)^{n-k}e^k k!} \underset{n\to\infty}{\sim} \frac{\lambda^k\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{k}{n}\right)^n e^k k!}$$

Finally, since $\left(\forall x \in R\right) \quad \left(1+\dfrac{x}{n}\right)^n \underset{n\to\infty}{\sim} e^x$:

$$\frac{\lambda^k\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{k}{n}\right)^n e^k k!} \underset{n\to\infty}{\sim} \frac{\lambda^k}{k!}e^{-\lambda} = \Pr(Y = k)$$

□

A.2 Deseq2 normalization

Deseq2 estimates size factors by comparing samples in order to take into consideration highly expressed and differentiated genes in the normalization scheme. To do so, let m be the number of samples in the full data. Define a theoretic reference sample gene expression level by taking the geometric mean across samples:

$$G_{mean}(Y_{i.}) = \left(\prod_{k=1}^{k=m} y_{ik}\right)^{\frac{1}{m}}$$

For each sample j, define the *weight* $\hat{s_j}$ by the median (across genes) of the ratios of observed counts to the theoretic sample expression level:

$$\hat{s_j} = \underset{i}{\text{median}}\left(\frac{Y_{ij}}{G_{mean}(Y_{i.})}\right)$$

A.3 Single cell data normalization and quality control

Due to technical issues (cells not well isolated; remaining cell fragments in wells ..) some reactions are filtered out of the data.

Normalization across conditions is performed using added Spikes. Spikes are known quantities of DNA introduced in the samples and analyzed. The IGBMC team provided the following normalization scheme:

Let A^q be the dataset under condition q and the columns of spikes $(S_1^q, \dots, S_p^q)$ where p is the number of spikes (here $p = 3$). For each condition q:

$$A^q = \alpha \frac{A^q}{median(S_1^q, \dots, S_p^q)}$$

where α is an arbitrary constant used to disperse the data for the sake of convenience.

A.4 Likelihood details

Likelihood of the Bernoulli-Lognormal model

Under the assumptions:

$$(Q_{ij}|B_{ij} = 1) \sim \log\mathcal{N}(\mu_j, \sigma_j^2) \tag{19}$$

$$(Q_{ij}|B_{ij} = 0) \sim \delta_0 \tag{20}$$

$$B_{ij} \sim \mathcal{B}(\pi_j) \tag{21}$$

$$\mathcal{L}(\theta^k|Q^k) = \prod_{i \in I_k \setminus S_k} (1 - \pi_k) \prod_{i \in S_k} \pi_k f_k({q^k}_i)$$

Where I_k is the number of cells under condition k, $S_k = \{i, {Q^k}_i > 0\}$ and $n_k = \text{Card}(S_k)$.

proof: The CDF (cumulative distribution function) of δ_0 is simply $\mathbb{1}_{\mathbb{R}^+}$. The CDF of a lognormal distribution will be denoted by F, and its density function by $f.\mathbb{1}_{{\mathbb{R}^*}_+}$. λ and δ_0 are respectively Lebesgue's and Dirac's measures.

Let's compute the density function of Q_i as the Radon-Nikodym derivative.

Let $t \in \mathbb{R}$. Omitting the condition k and the gene j:

$$\begin{aligned}
P(Q_i \leq t) &= P(Q_i \leq t|B_i = 1)P(B_i = 1) + P(Q_i \leq t|B_i = 0)P(B_i = 0) \\
&= \pi F(t) + (1-\pi)\mathbb{1}(t \geq 0) \\
&= \int_{]-\infty,t]} \pi f.1_{\mathbb{R}^*} d\lambda + \int_{]-\infty,t]} (1-\pi)\delta_0 d\delta_0 \\
&\overset{*}{=} \int_{]-\infty,t]} \big(\pi f.1_{\mathbb{R}^*} + (1-\pi)\delta_0\big) d(\lambda + \delta_0) \\
&= \int_{]-\infty,t]} h \, d(\lambda + \delta_0)
\end{aligned}$$

The passage (*) is justified by:

1.

$$\begin{aligned}
0 \leq \int_{]-\infty,t]} \pi f.1_{\mathbb{R}^*} d\delta_0 &= \int_{]0,t]} \pi f d\delta_0 \\
&\leq \max(f)\delta_0(]0,t[) \\
&= 0
\end{aligned}$$

2.

$$\begin{aligned}
\int_{]-\infty,t]} (1-\pi)\delta_0 d\lambda &= \int_{\{0\}} (1-\pi) d\lambda \\
&= 0
\end{aligned}$$

One could easily verify that $h \geq 0$, and with respect to the measure $\lambda + \delta_0 : \int h = 1$.

h is therefore the density function of Q_j relative to the reference measure $\lambda + \delta_0$ which can be written as, for any variable Q_i^k (f_k density function of a $\log \mathcal{N}(\mu_k, \sigma^2)$) :

$$h_i^k(x) = \begin{cases} \pi_k f_k(x)\mathbb{1}(x > 0) \text{ if } i \in S_k. \\ 1 - \pi_k \text{ else.} \end{cases}$$

And finally, using the definition of the likelihood function and the categorical form of h above :

$$\begin{aligned}
\mathcal{L}(\theta^k|Q^k) &= \prod_{i \in I_k} {h_i}^k(q_i^k) \\
&= \prod_{i \in I_k \setminus S_k} (1-\pi_k) \prod_{i \in S_k} \pi_k f_k({q^k}_i) \\
&= {\pi_k}^{n_k}(1-\pi_k)^{I_k - n_k} \prod_{i \in S_k} f_k({q^k}_i)
\end{aligned}$$

□

A.5 LRT details

As defined in Section 2.1:

$$\Lambda(C) = \frac{\underset{\theta\in\Theta_0}{\sup}\mathcal{L}(\theta^0,\theta^1|Q^0,Q^1)}{\underset{\theta\in\Theta}{\sup}\mathcal{L}(\theta^0,\theta^1|Q^0,Q^1)} \tag{22}$$

Using the Likelihood formula (Q^0 and Q^1 are supposed independent):

$$\mathcal{L}(\theta^0,\theta^1|Q^0,Q^1) = \prod_{k\in\{0,1\}} {\pi_k}^{n_k}(1-\pi_k)^{I_k-n_k} \prod_{i\in S_k} f_k({Q^k}_i)$$

we get by separating the independent ratios:

$$\begin{aligned}\Lambda(C) &= \frac{\underset{\theta\in\Theta_0}{\sup}\mathcal{L}(\theta^0,\theta^1|Q^0,Q^1)}{\underset{\theta\in\Theta}{\sup}\mathcal{L}(\theta^0,\theta^1|Q^0,Q^1)}\\ &= \frac{\underset{\pi_0}{\sup}\quad \pi_0^{n_0+n_1}(1-\pi_0)^{I_0+I_1-n_0-n_1}}{\underset{\pi_0,\pi_1}{\sup}\quad \pi_0^{n_0}(1-\pi_0)^{I_0-n_0}\pi_1^{n_1}(1-\pi_1)^{I_1-n_1}} \frac{\underset{\mu_0,\sigma^2}{\sup}\prod_{i\in S_0 U S_1} f_0(q_i)}{\underset{\mu_0,\mu_1,\sigma^2}{\sup}\prod_{i\in S_0} f_0({q^0}_i)\prod_{i\in S_1} f_1({q^1}_i)}\end{aligned}$$

Where Q is the concatenated array (Q^0,Q^1).

Taking out the log of q to get normal distributions with the change of variables: $a_i = log(q_i)$ (over S_k, $q_i > 0$):

$$\Lambda(C) = \frac{\underset{\pi_0}{\sup}\quad \pi_0^{n_0+n_1}(1-\pi_0)^{I_0+I_1-n_0-n_1}}{\underset{\pi_0,\pi_1}{\sup}\quad \pi_0^{n_0}(1-\pi_0)^{I_0-n_0}\pi_1^{n_1}(1-\pi_1)^{I_1-n_1}} \frac{\underset{\mu_0,\sigma^2}{\sup}\prod_{i\in S_0 U S_1} g_0(a_i)}{\underset{\mu_0,\mu_1,\sigma^2}{\sup}\prod_{i\in S_0} g_0({a^0}_i)\prod_{i\in S_1} g_1({a^1}_i)}$$

where g_k is the normal density function with mean and variance (μ_k,σ^2).

Resulting in four maximization problems carried out using 1st and 2nd derivatives of log-likelihood functions:

1. $\underset{\pi_0}{\sup}\quad \pi_0^{n_0+n_1}(1-\pi_0)^{I_0+I_1-n_0-n_1}$

$$\pi_{max} = \frac{n_0+n_1}{I_0+I_1}$$

2. $\underset{\pi_0,\pi_1}{\sup}\quad \pi_0^{n_0}(1-\pi_0)^{I_0-n_0}\pi_1^{n_1}(1-\pi_1)^{I_1-n_1}$

$$\begin{cases}\pi_{0max} = \frac{n_0}{I_0}\\ \pi_{1max} = \frac{n_1}{I_1}\end{cases}$$

3. $\sup\limits_{\mu_0,\sigma^2} \prod_{i \in S_0 U S_1} g_0(a_i)$

$$\begin{cases} \mu_{max} = \frac{1}{n_0+n_1} \sum_{i \in S_0 U S_1} a_i \\ \sigma^2_{max} = \frac{1}{n_1+n_0} \sum_{i \in S_0 U S_1} (a_i - \mu_{max})^2 \end{cases}$$

4. $\sup\limits_{\mu_0,\mu_1,\sigma^2} \prod_{i \in S_0} g_0({a^0}_i) \prod_{i \in S_1} g_1({a^1}_i)$

$$\begin{cases} \mu_{0max} = \frac{1}{n_0} \sum_{i \in S_0} {a_i}^0 \\ \mu_{1max} = \frac{1}{n_1} \sum_{i \in S_1} {a_i}^1 \\ \sigma^2_{max} = \frac{1}{n_0} \sum_{i \in S_0} ({a_i}^0 - \mu_{0max})^2 + \frac{1}{n_1} \sum_{i \in S_1} ({a_i}^1 - \mu_{1max})^2 \end{cases}$$

B Tables and figures

B.1 PCA axes correlations

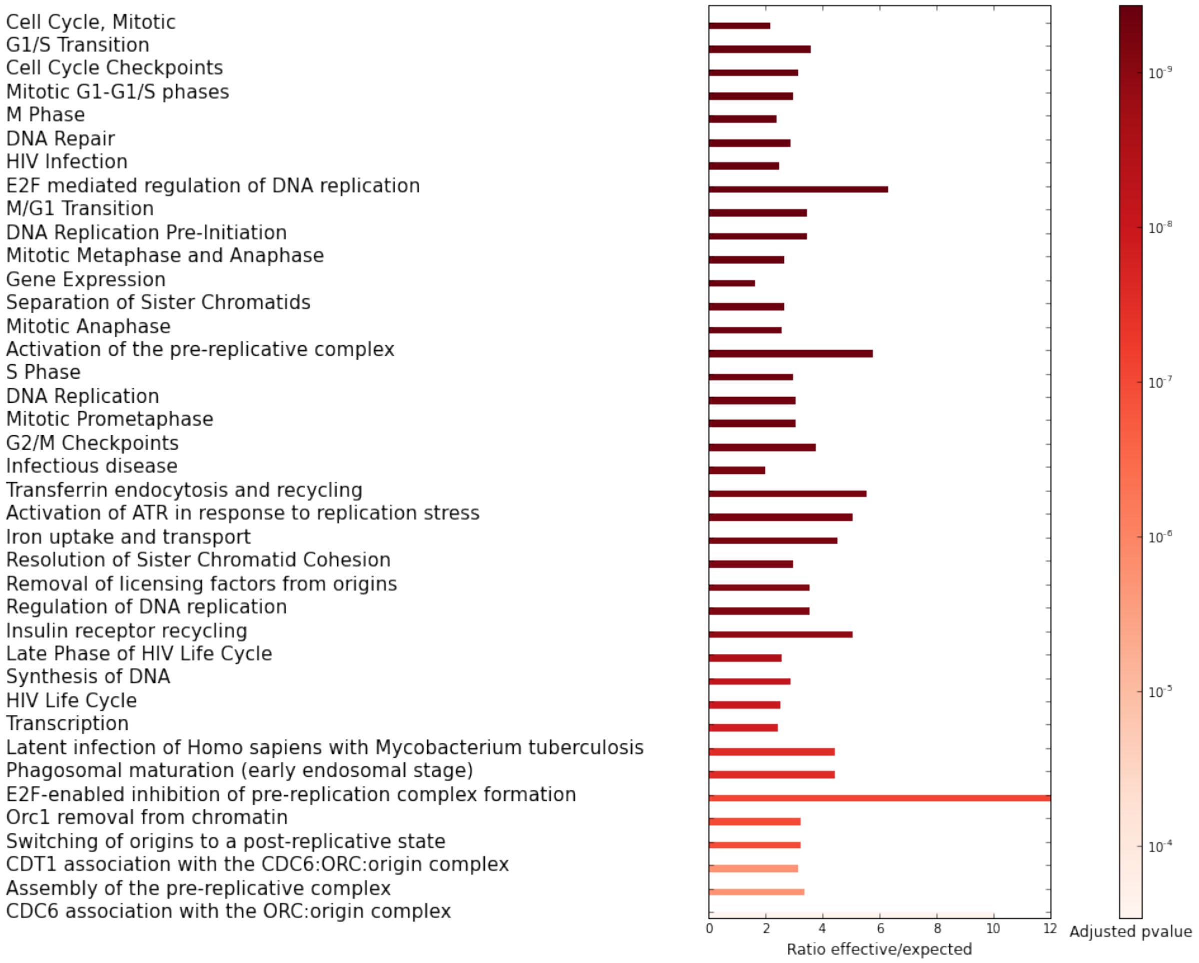


Figure 21: *Significant pathway categories (at p-value = 0.001) positively correlated (+0.6) with PCA first axis PC1.*

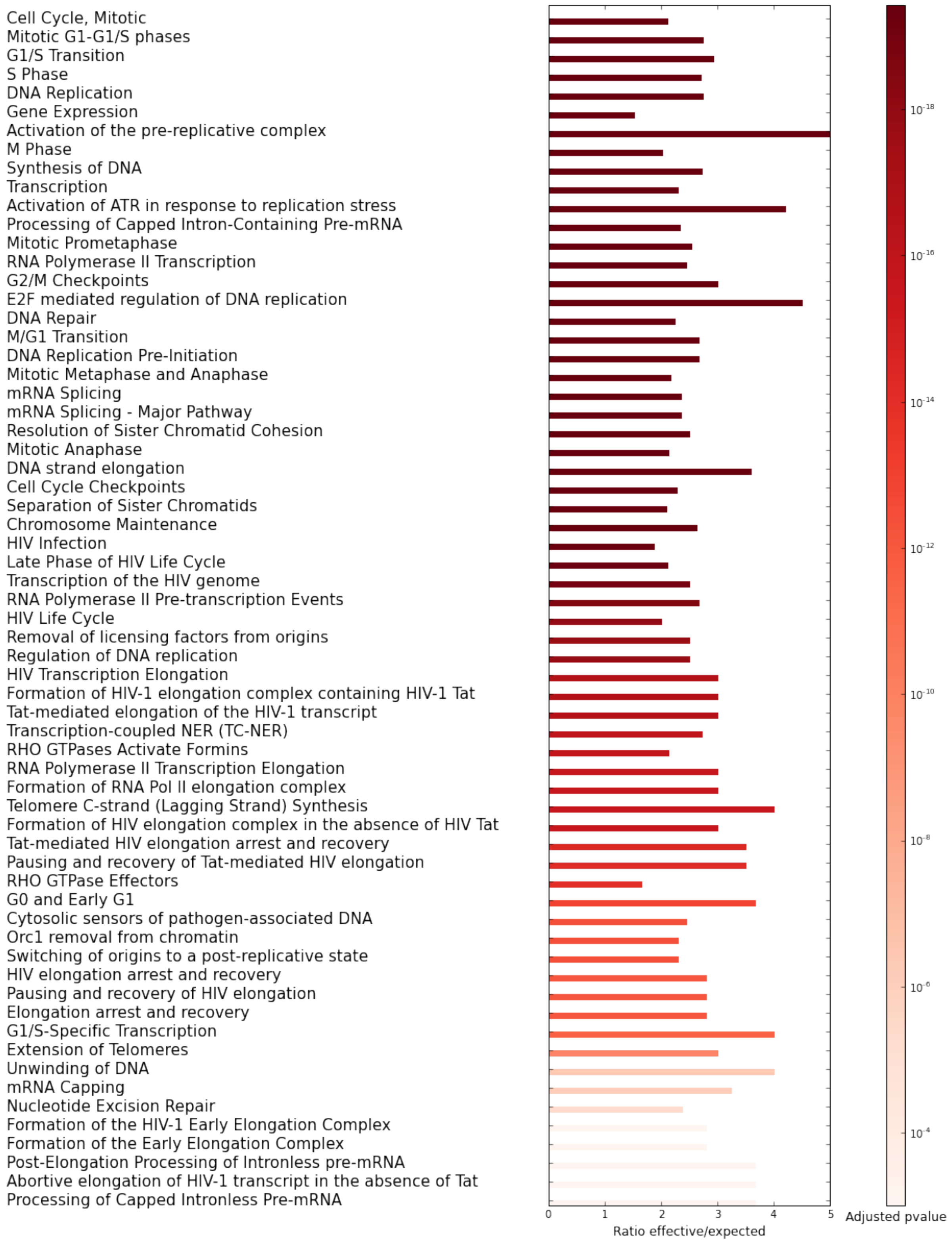


Figure 22: *Significant pathway categories (at p-value = 0.001) positively correlated (+0.6) with PCA second axis PC2.*

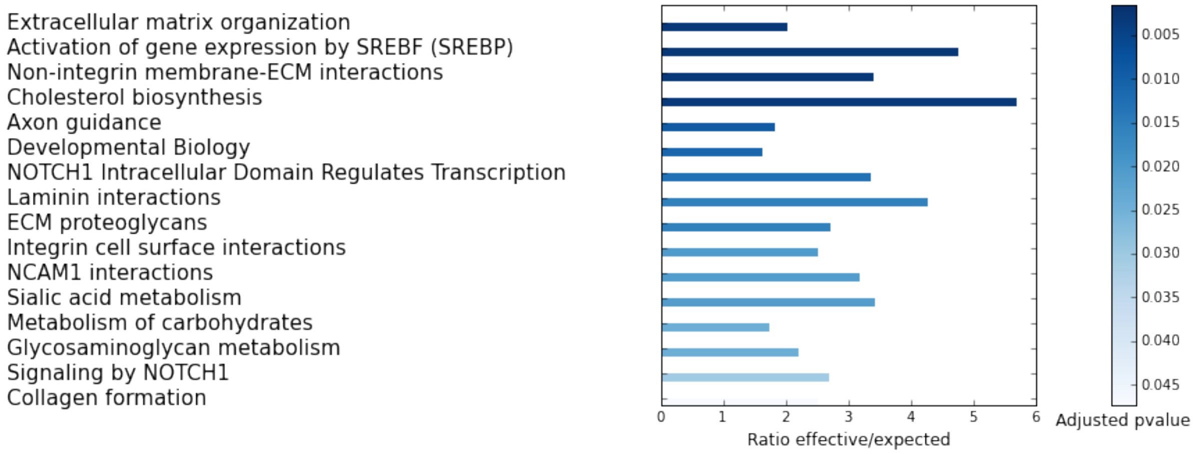


Figure 23: *Significant pathway categories (at p-value = 0.05) negatively correlated (-0.6) with PCA first axis PC1.*

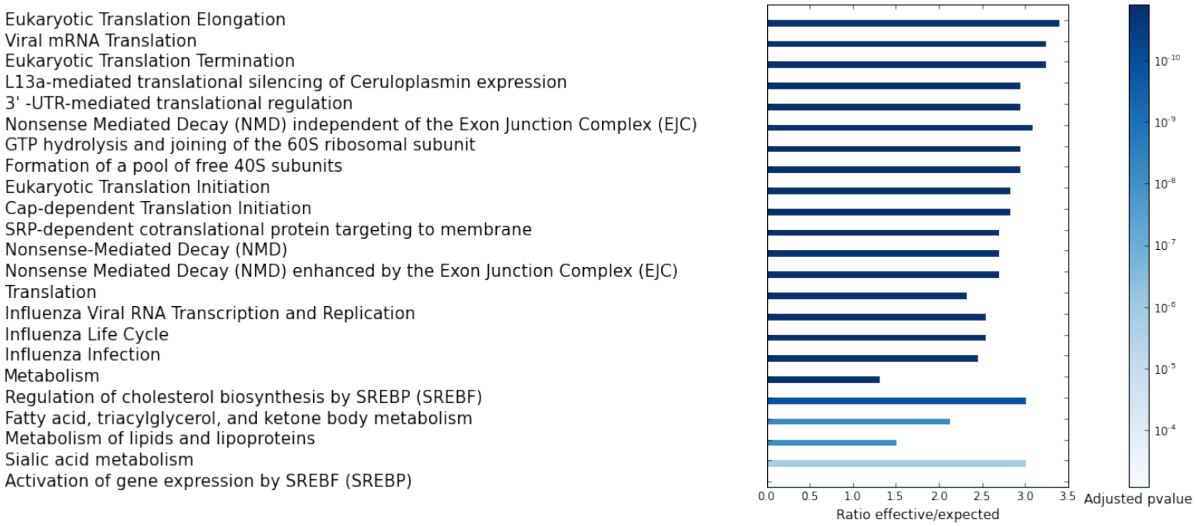
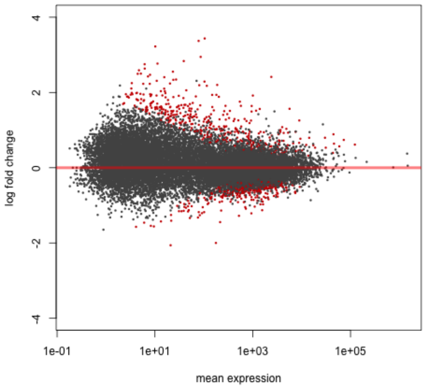


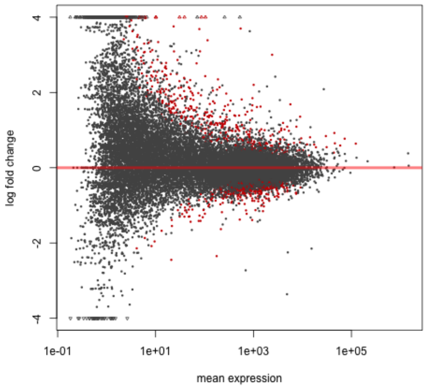
Figure 24: *Significant pathway categories (at p-value = 0.001) negatively correlated (-0.6) with PCA second axis PC2.*

B.2 Log fold change Bayesian shrinkage

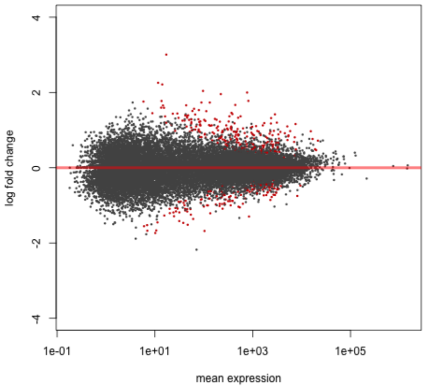
Shrunken LFCs — Un-shrunken LFCs (MLE estimates)

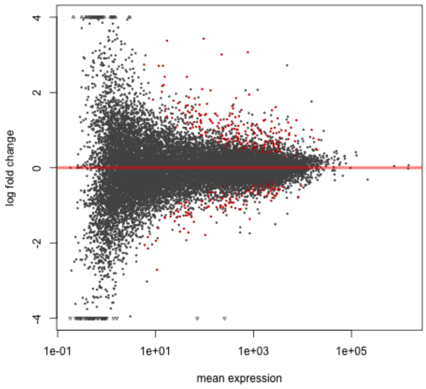
Day 9 / Day 4



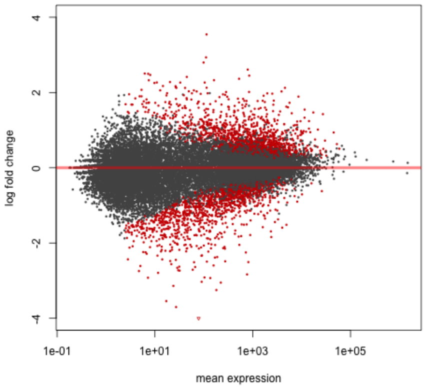


Day 12 / Day 9





Day 16 / Day 12



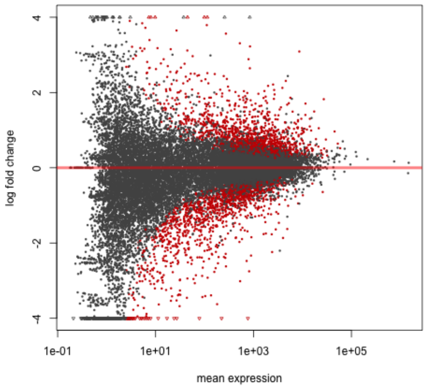


Table 6: *MA plots representing LFCs against mean counts. Each dot is a gene. Red dots are differentially expressed at 0.01 significance (Wald test).*

B.3 Dispersion Bayesian shrinkage

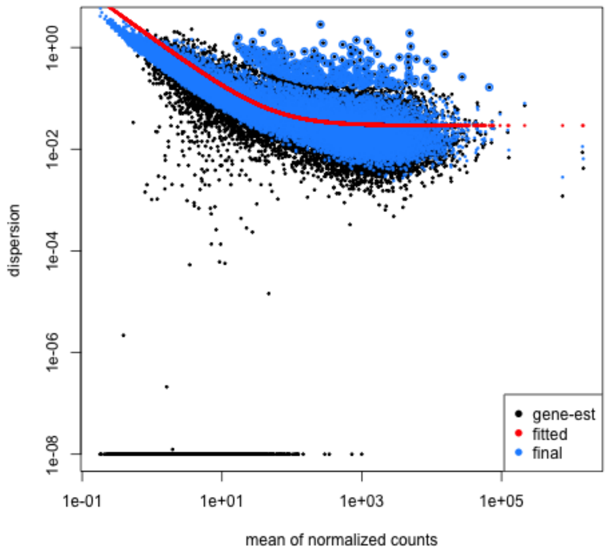


Figure 25: *Plot of dispersion estimates over mean of normalized read counts.* ***Black:*** *First MLE dispersion estimates.* ***Red:*** *Fitted model-assumption trend curve regressing dispersion over mean counts.* ***Blue:*** *Final estimates after shrinkage; except outliers that do not seem to follow the trend assumption (black-blue dots).*

B.4 Enrichment test between Day 4 and day 12

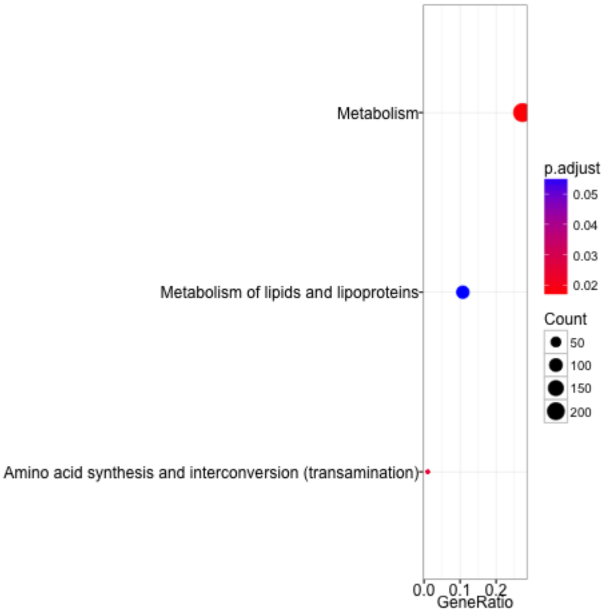


Figure 26: *Over representation test highlighting significant pathways between day 4 and day 12. Compared to the other evolutions, the phase 4-12 can be considered stable as the categories are general and the p-values are not as significant.*

B.5 DBSCAN clusters enrichment

Enrichment results of each cluster obtained in DBSCAN clustering of genes.

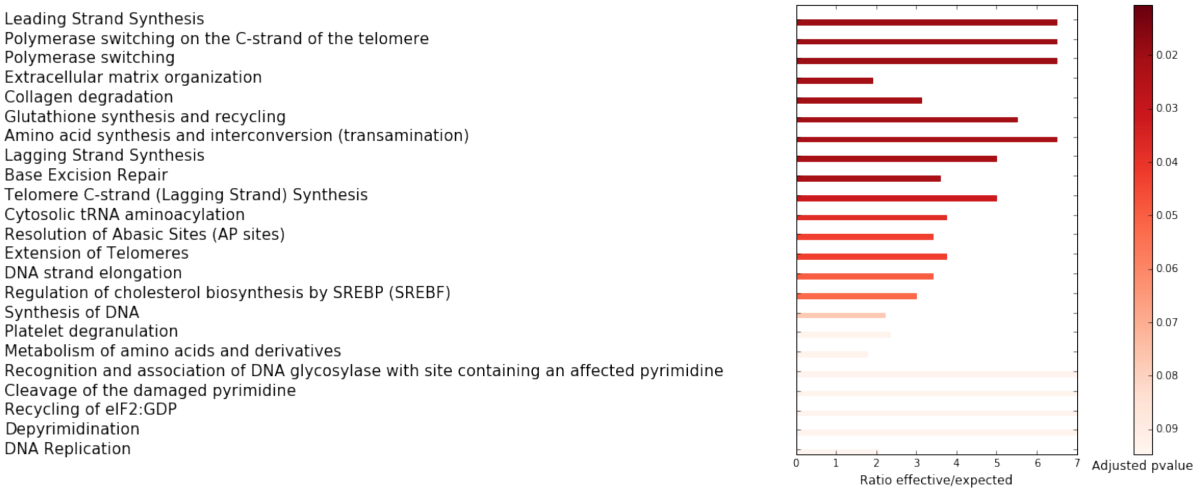


Figure 27: *Significant pathway categories (at p-value = 0.01) of cluster -1.*

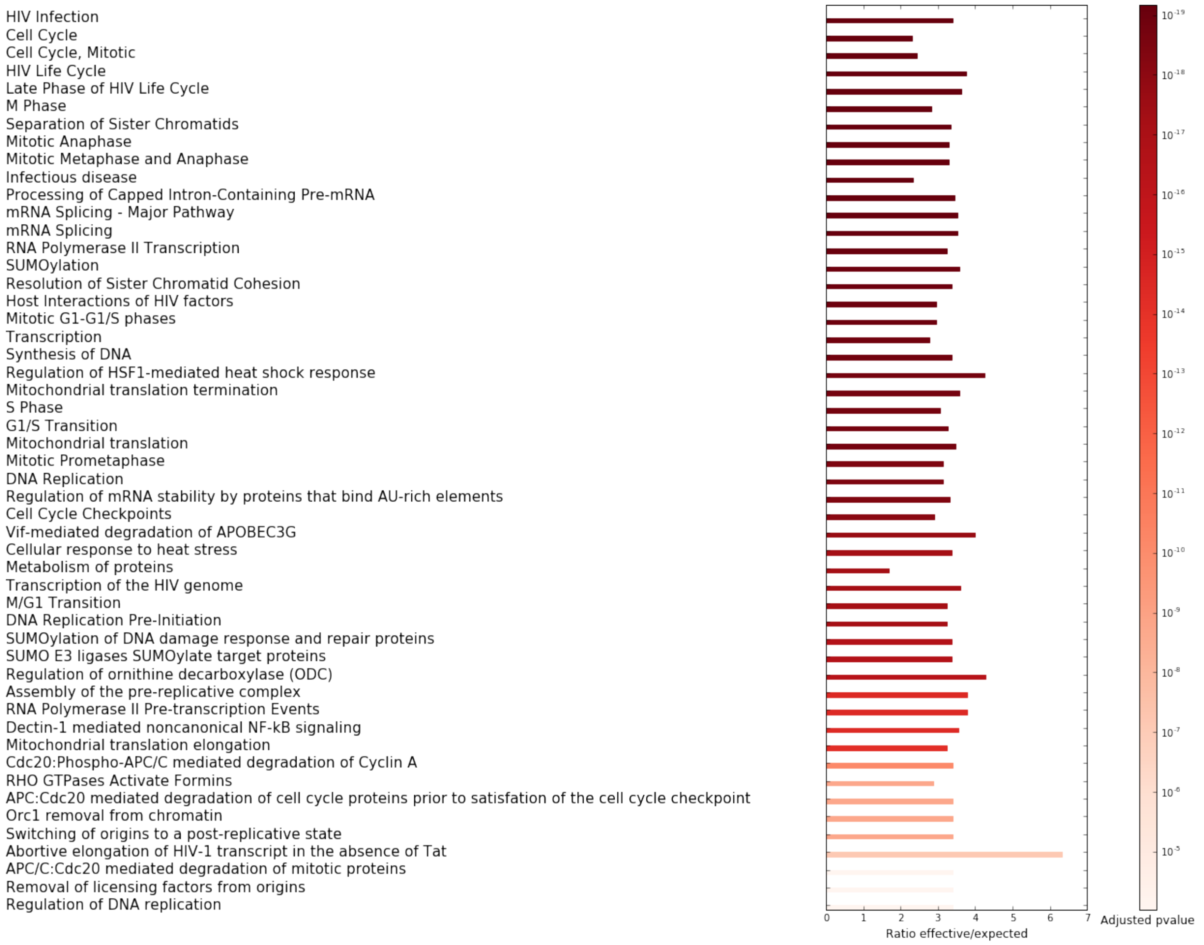


Figure 28: *Significant pathway categories (at p-value = 0.001) of cluster 0.*

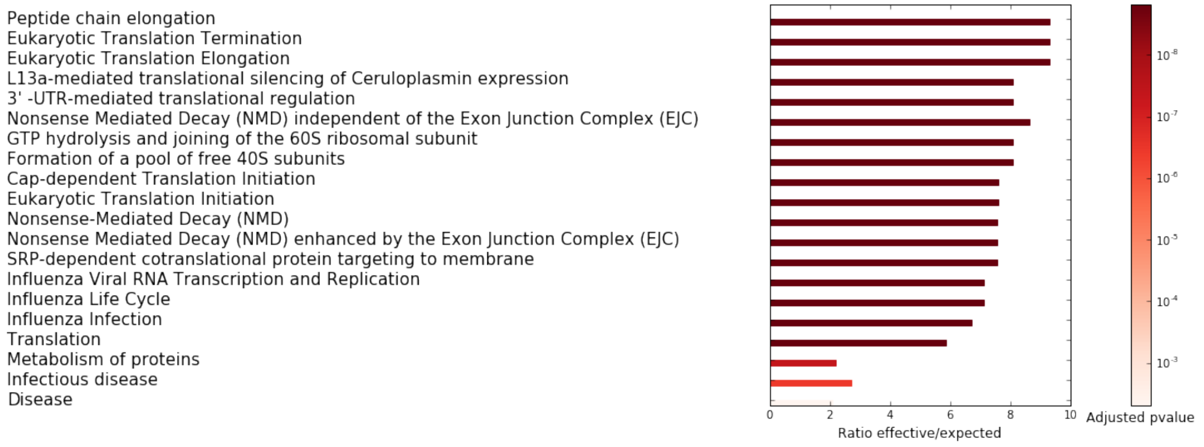


Figure 29: *Significant pathway categories (at p-value = 0.01) of cluster 1.*

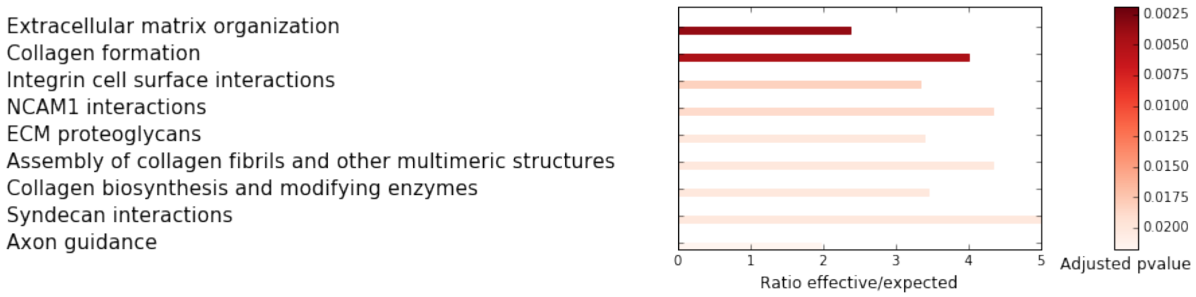


Figure 30: *Significant pathway categories (at p-value = 0.05) of cluster 3.*

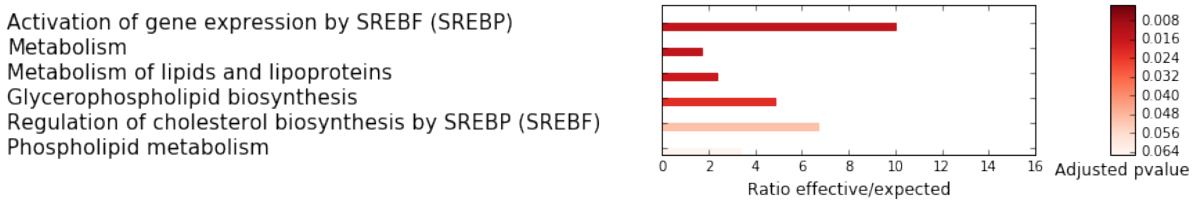


Figure 31: *Significant pathway categories (at p-value = 0.1) of cluster 4.*

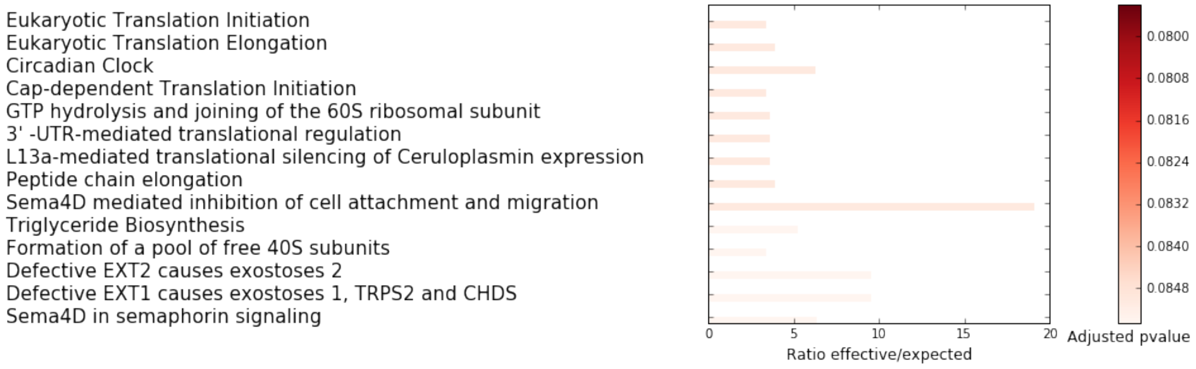


Figure 32: *Significant pathway categories (at p-value = 0.1) of cluster 5.*

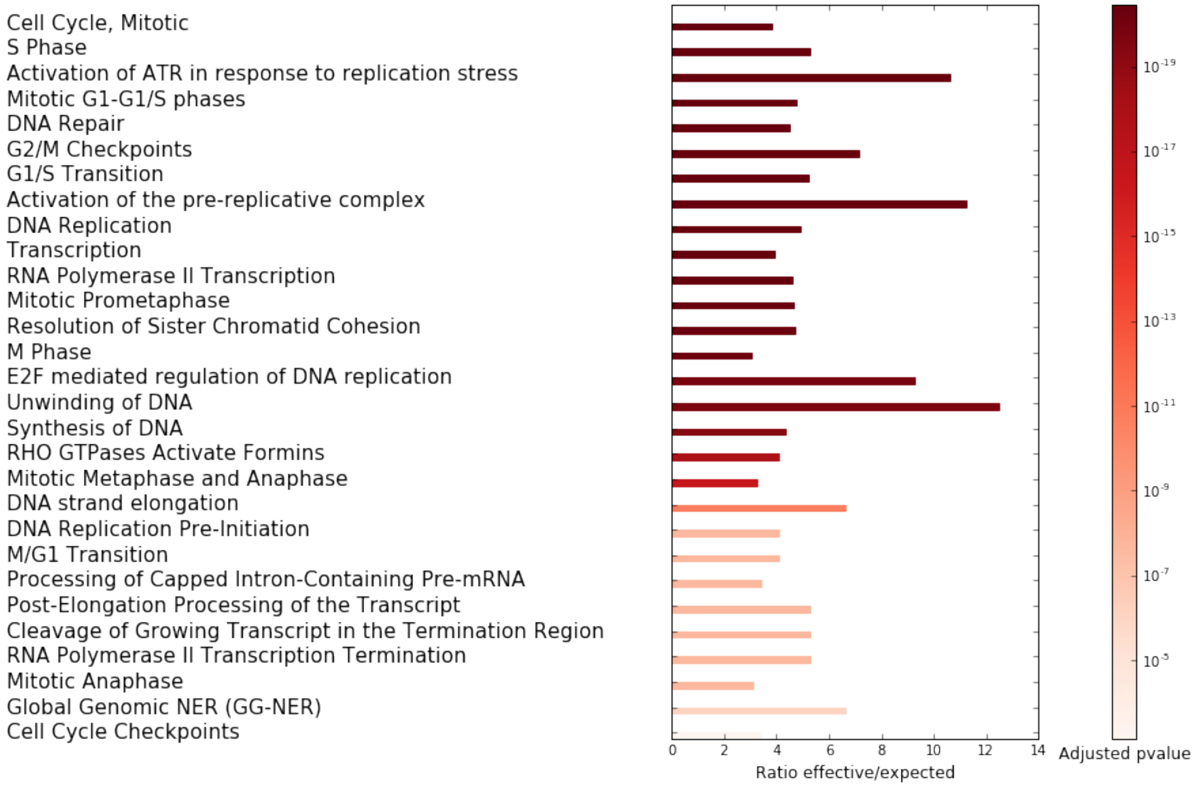


Figure 33: *Significant pathway categories (at p-value = 0.001) of cluster 6.*

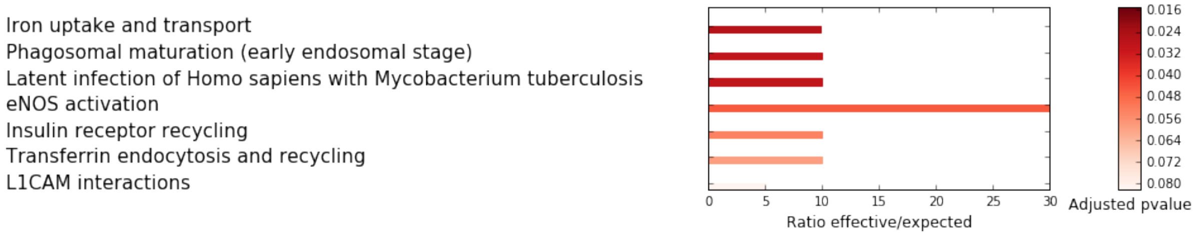


Figure 34: *Significant pathway categories (at p-value = 0.01) of cluster 7.*

B.6 QQ plots of Single cell data

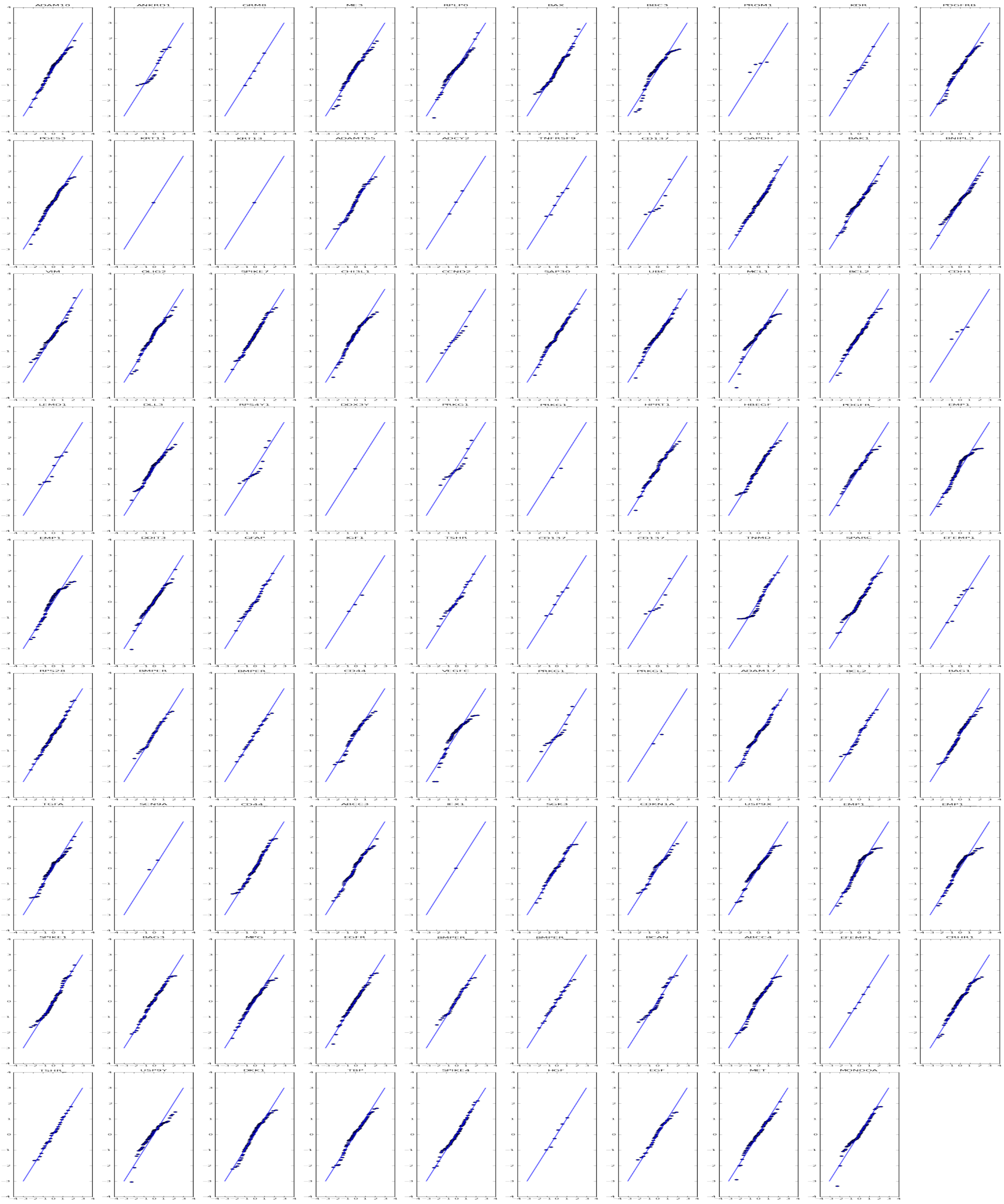


Figure 35: *Quantile-Quantile plots of single cell data. Quantiles of genes with non-zero frequency of expression* $(A_{ij}|B_{ij} = 1)$ *are plotted against quantiles of gaussian distributions. Biological condition = day 0.*

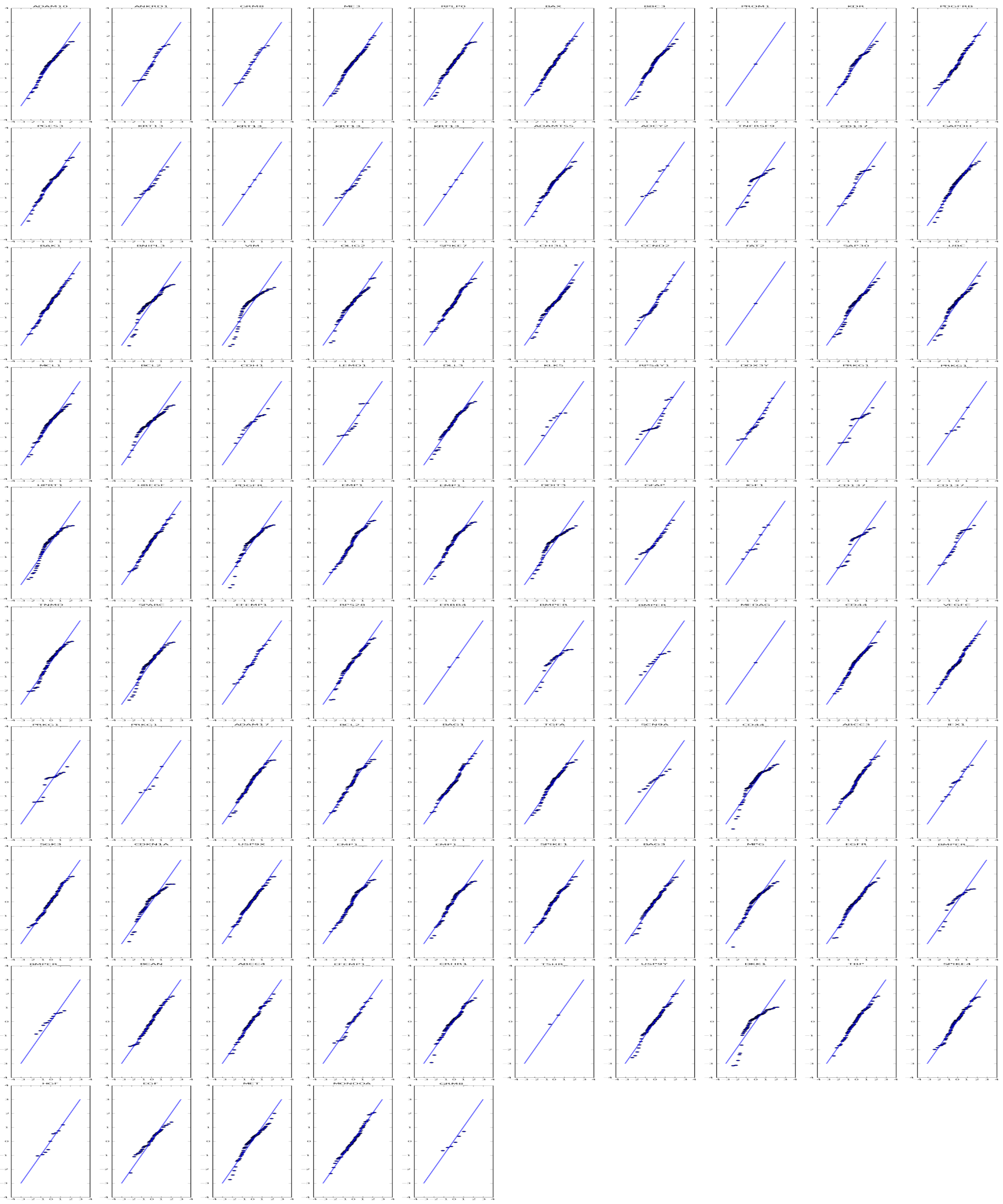


Figure 36: *Quantile-Quantile plots of single cell data. Quantiles of genes with non-zero frequency of expression* $(A_{ij}|B_{ij} = 1)$ *are plotted against quantiles of gaussian distributions. Biological condition = day 4.*

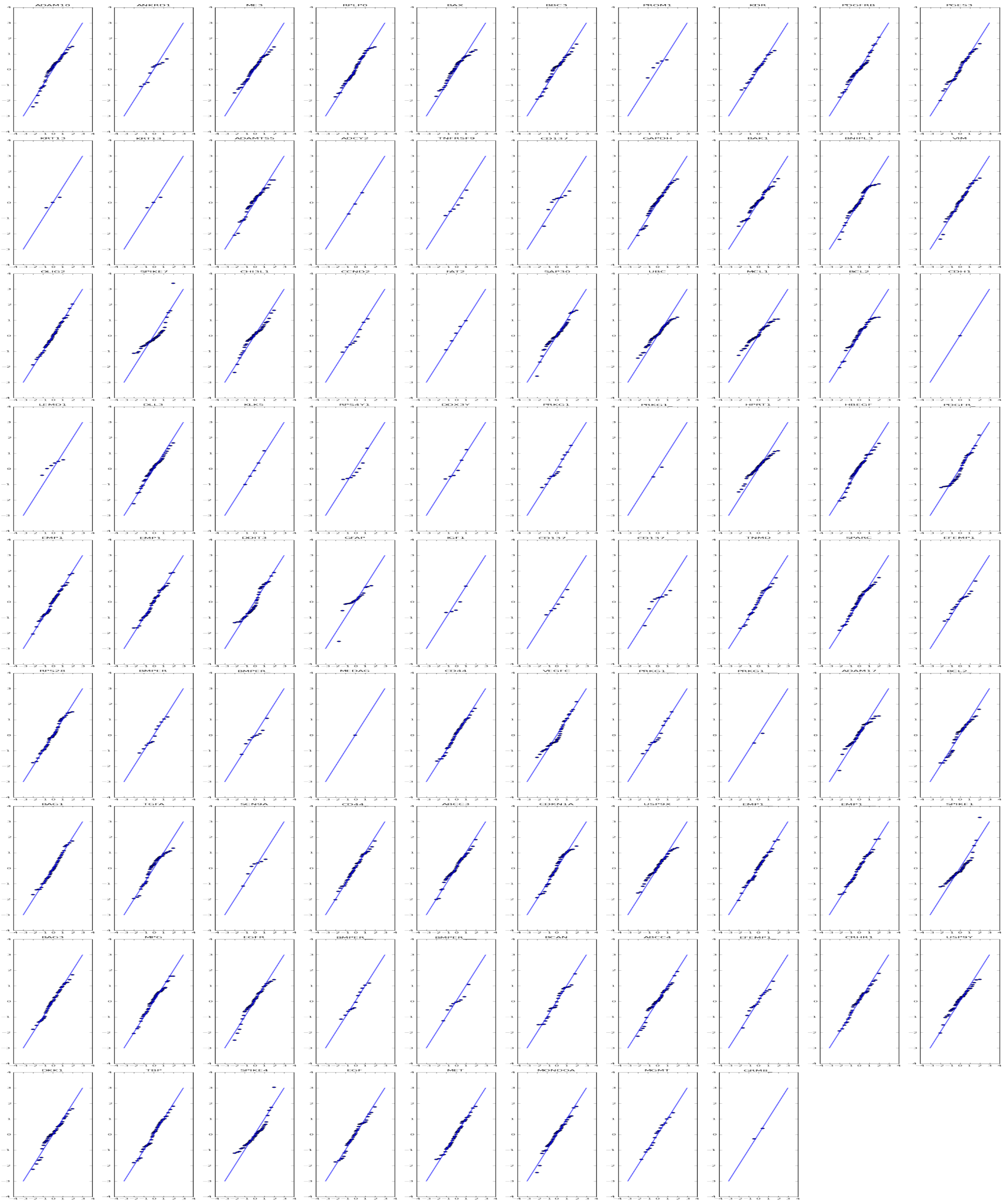


Figure 37: *Quantile-Quantile plots of single cell data. Quantiles of genes with non-zero frequency of expression* $(A_{ij}|B_{ij} = 1)$ *are plotted against quantiles of gaussian distributions. Biological condition = day 9.*

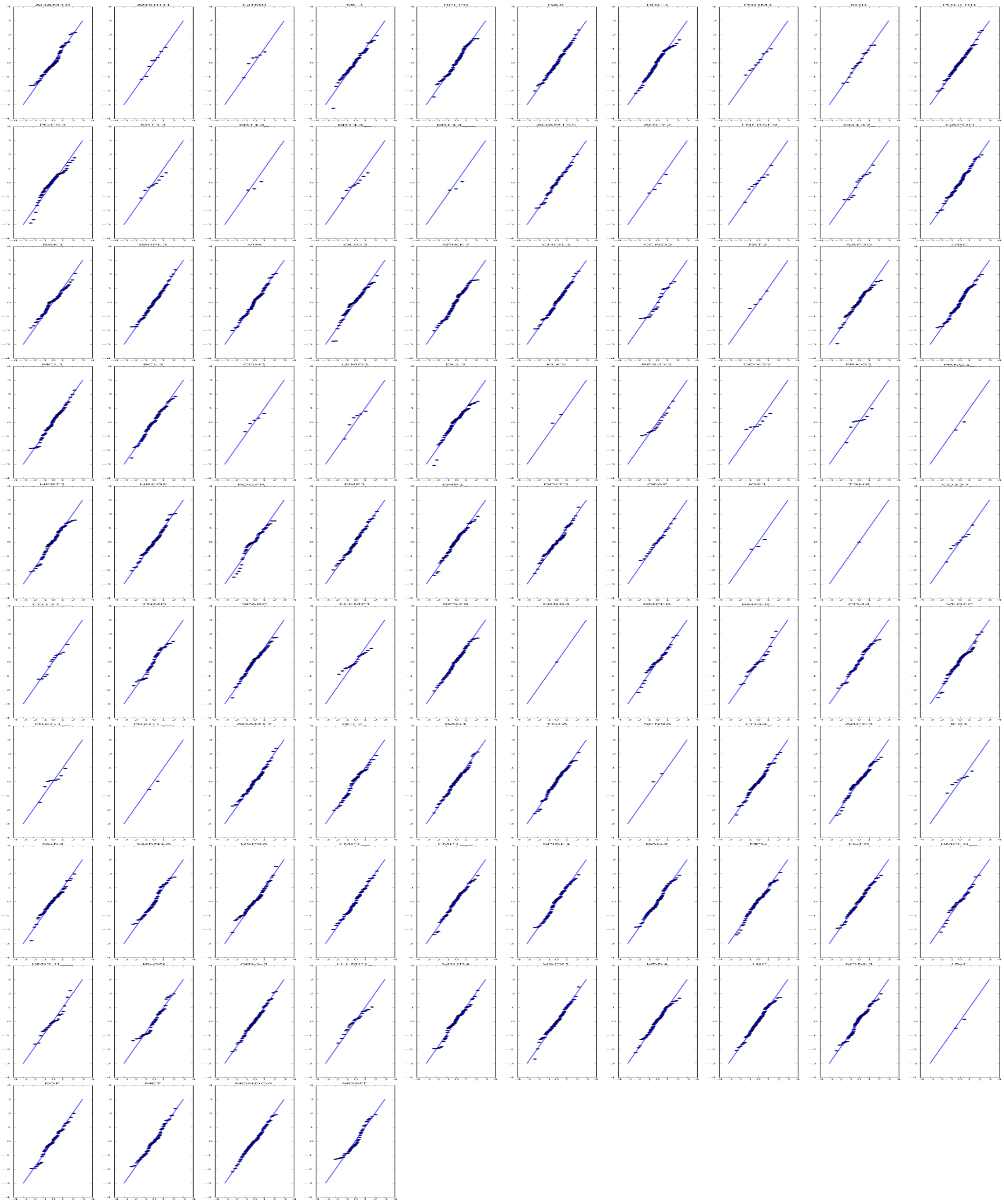


Figure 38: *Quantile-Quantile plots of single cell data. Quantiles of genes with non-zero frequency of expression* $(A_{ij}|B_{ij} = 1)$ *are plotted against quantiles of gaussian distributions. Biological condition = day 12.*

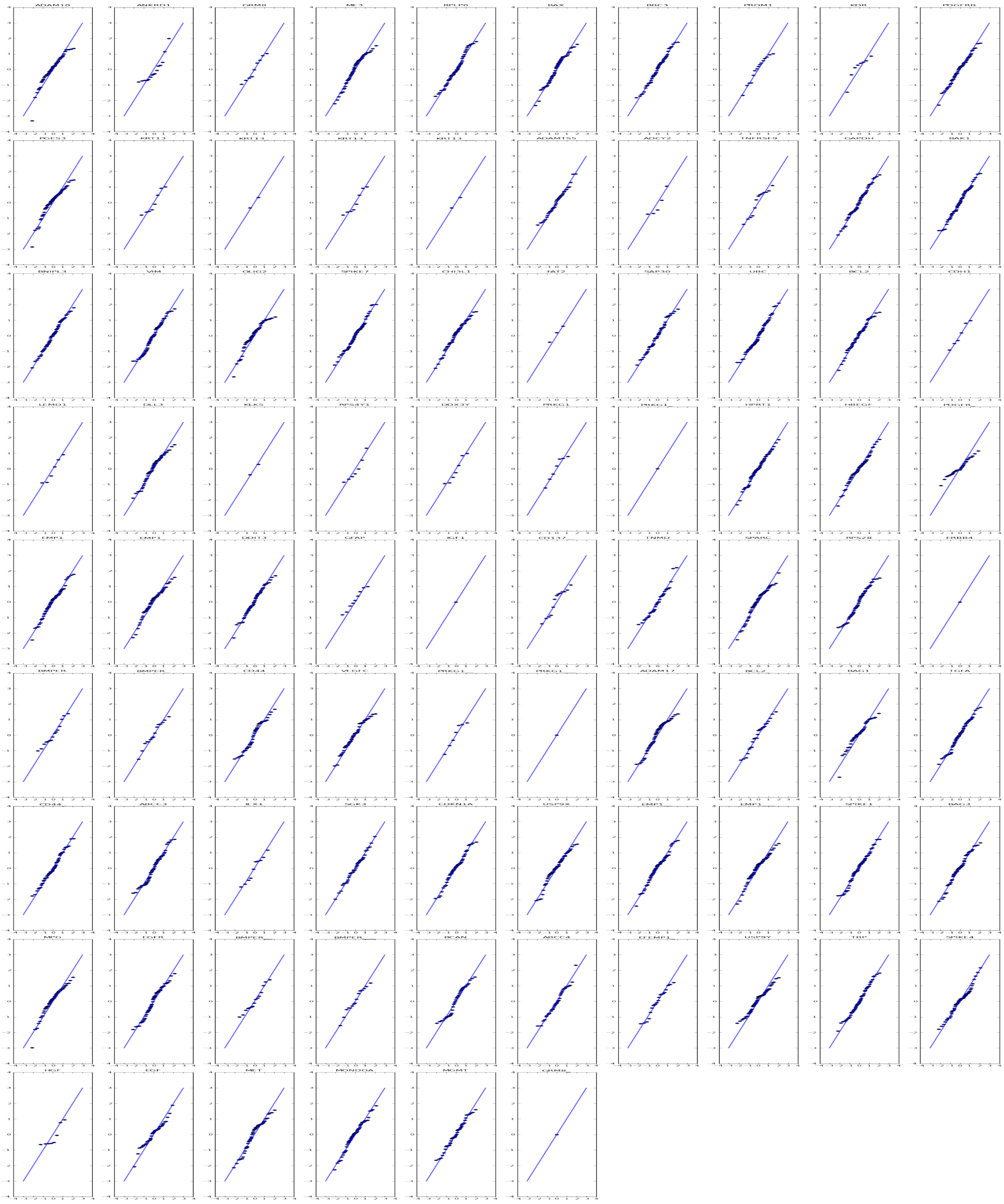


Figure 39: *Quantile-Quantile plots of single cell data. Quantiles of genes with non-zero frequency of expression* $(A_{ij}|B_{ij} = 1)$ *are plotted against quantiles of gaussian distributions. Biological condition = day 16.*

B.7 LRT simulation

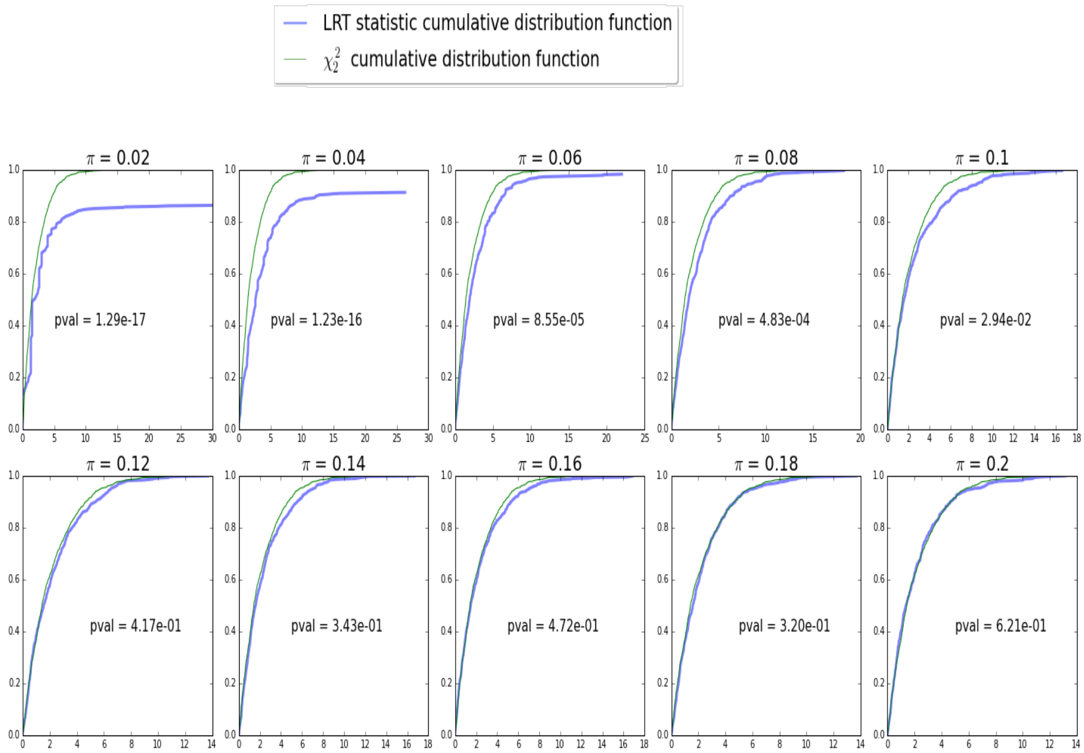
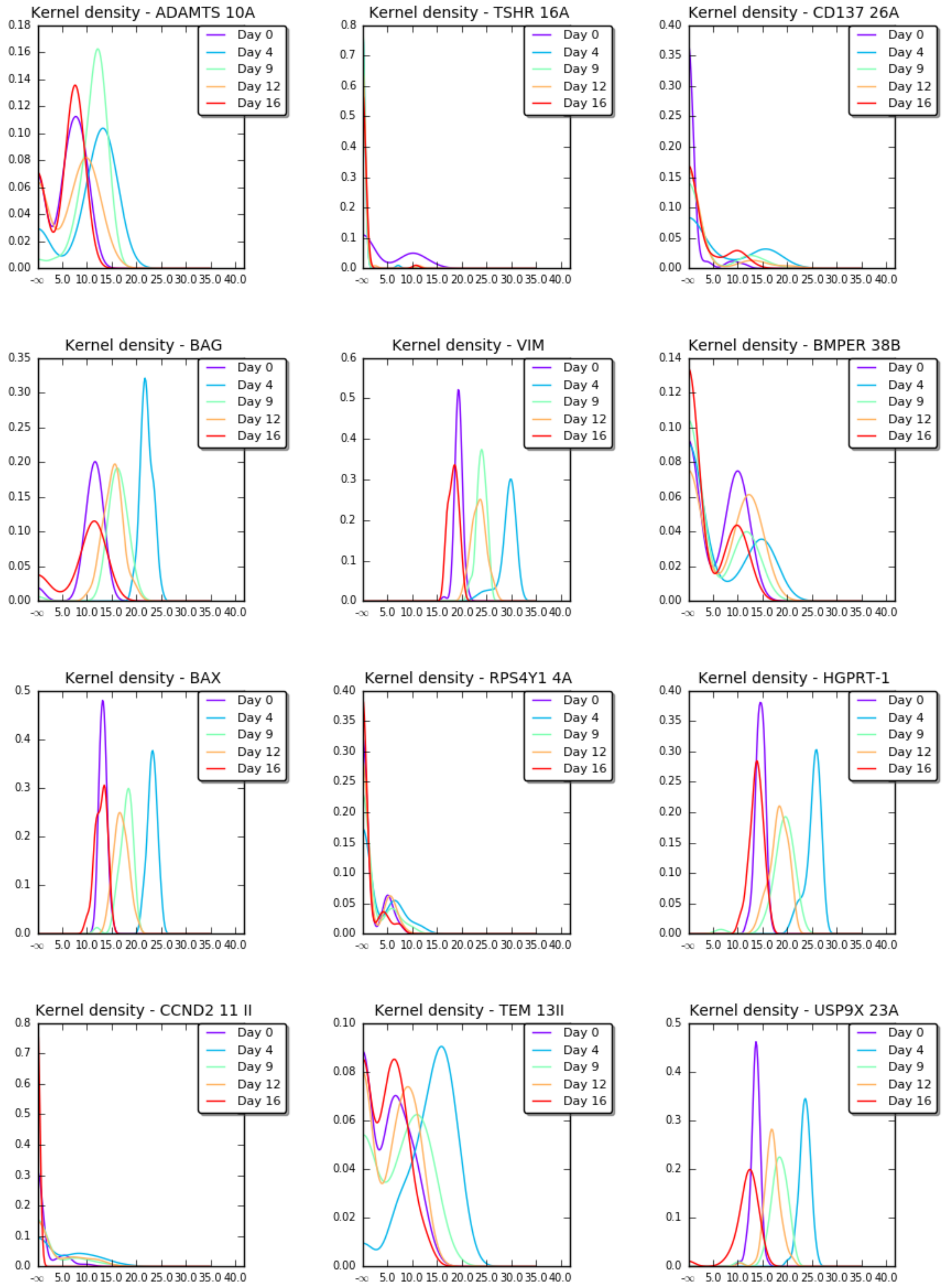
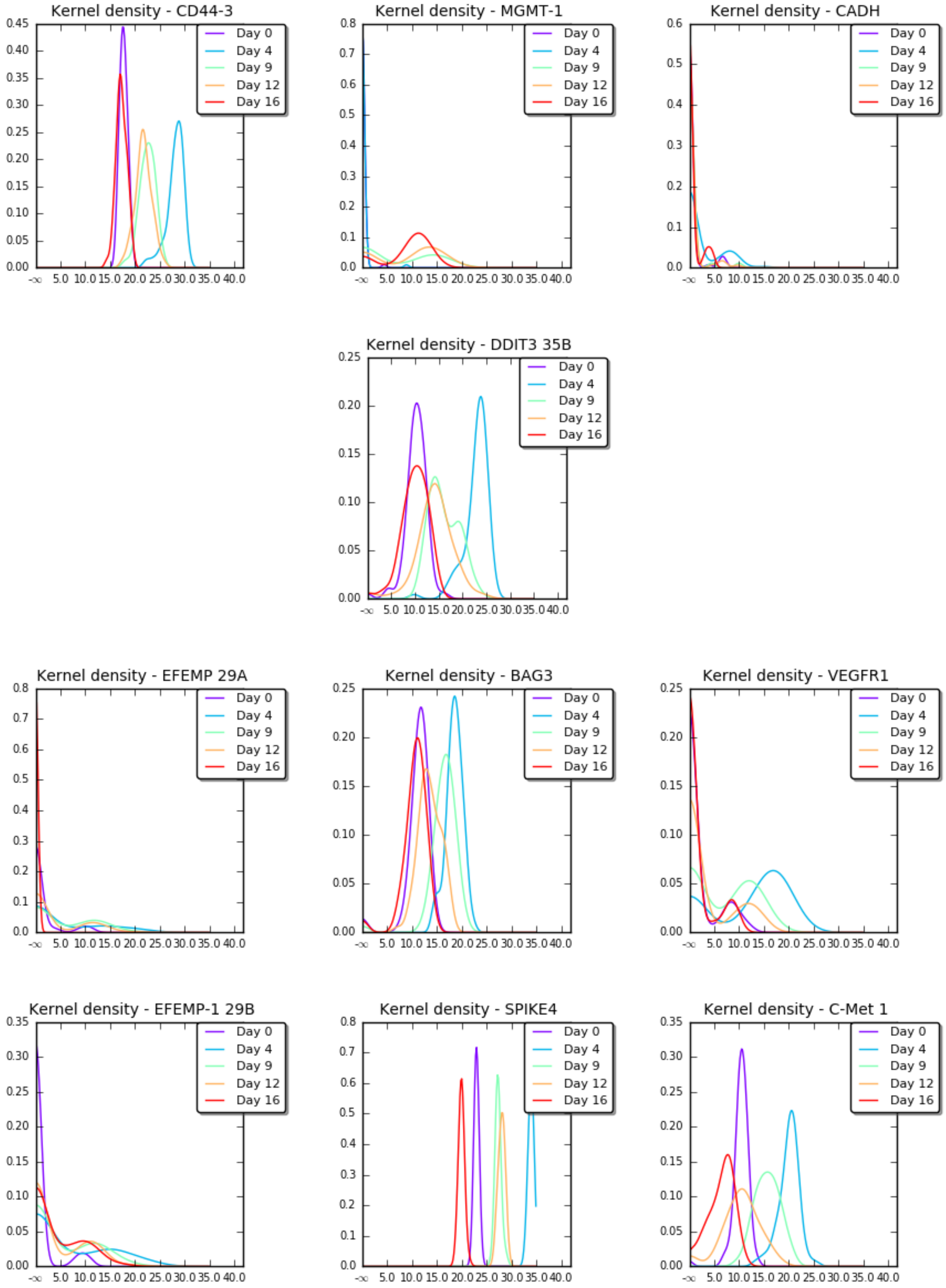


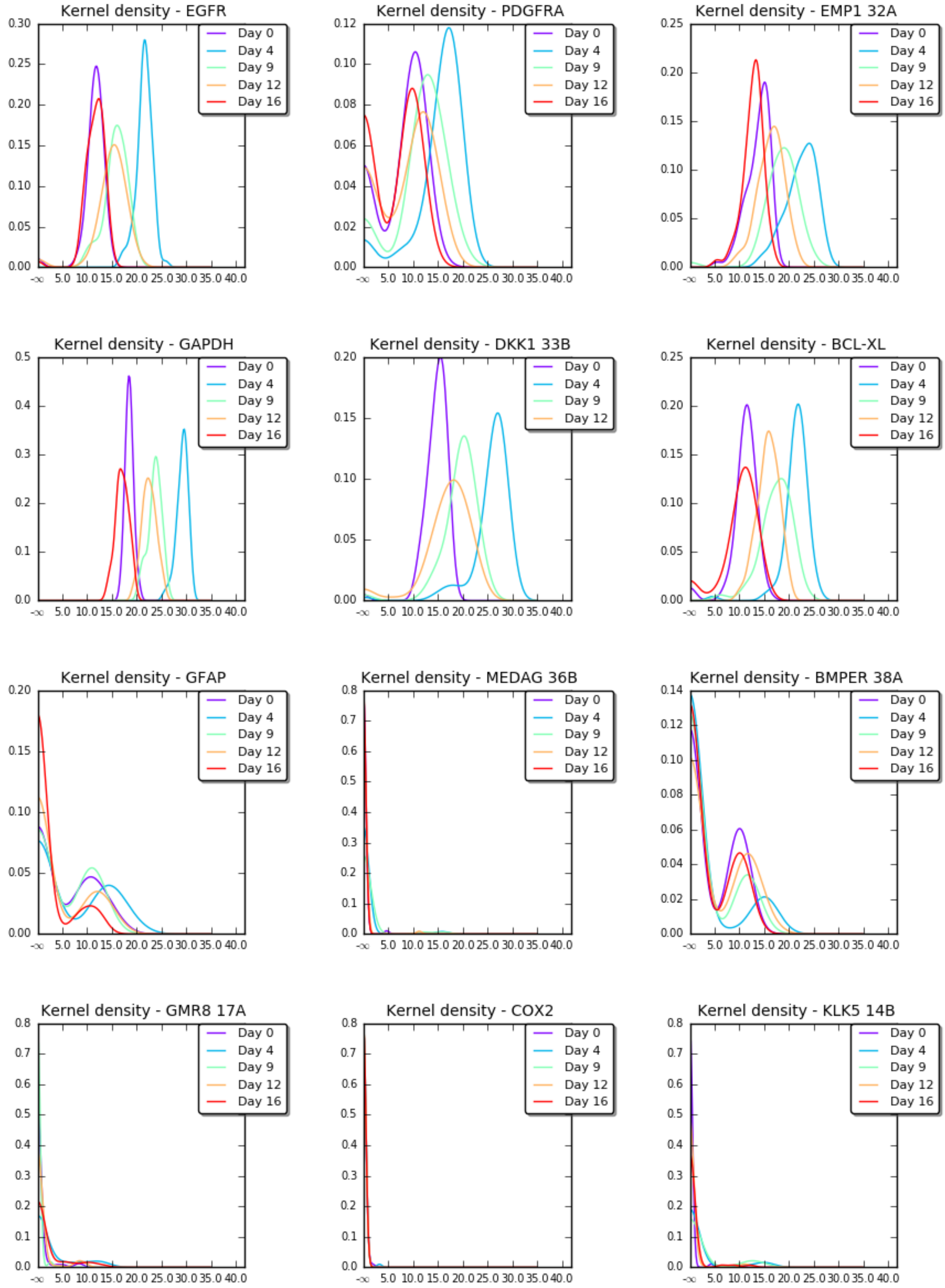
Figure 40: *For different values of π, 500 genes are simulated (bernoulli-lognormal) under two conditions with the same triplet (π, μ, σ^2). The resulting statistic distribution is plotted against the empirical distribution of a χ^2_2. A goodness-of-fit test (Kolmogorov-Smirnov) is used to assess for the significance of the similarity between CDFs (Empirical to be precise). the obtained p-value is written within each plot. These results suggest that for $\pi < 0.12$ the distribution of the test statistics cannot be considered generated from a χ^2_2.*

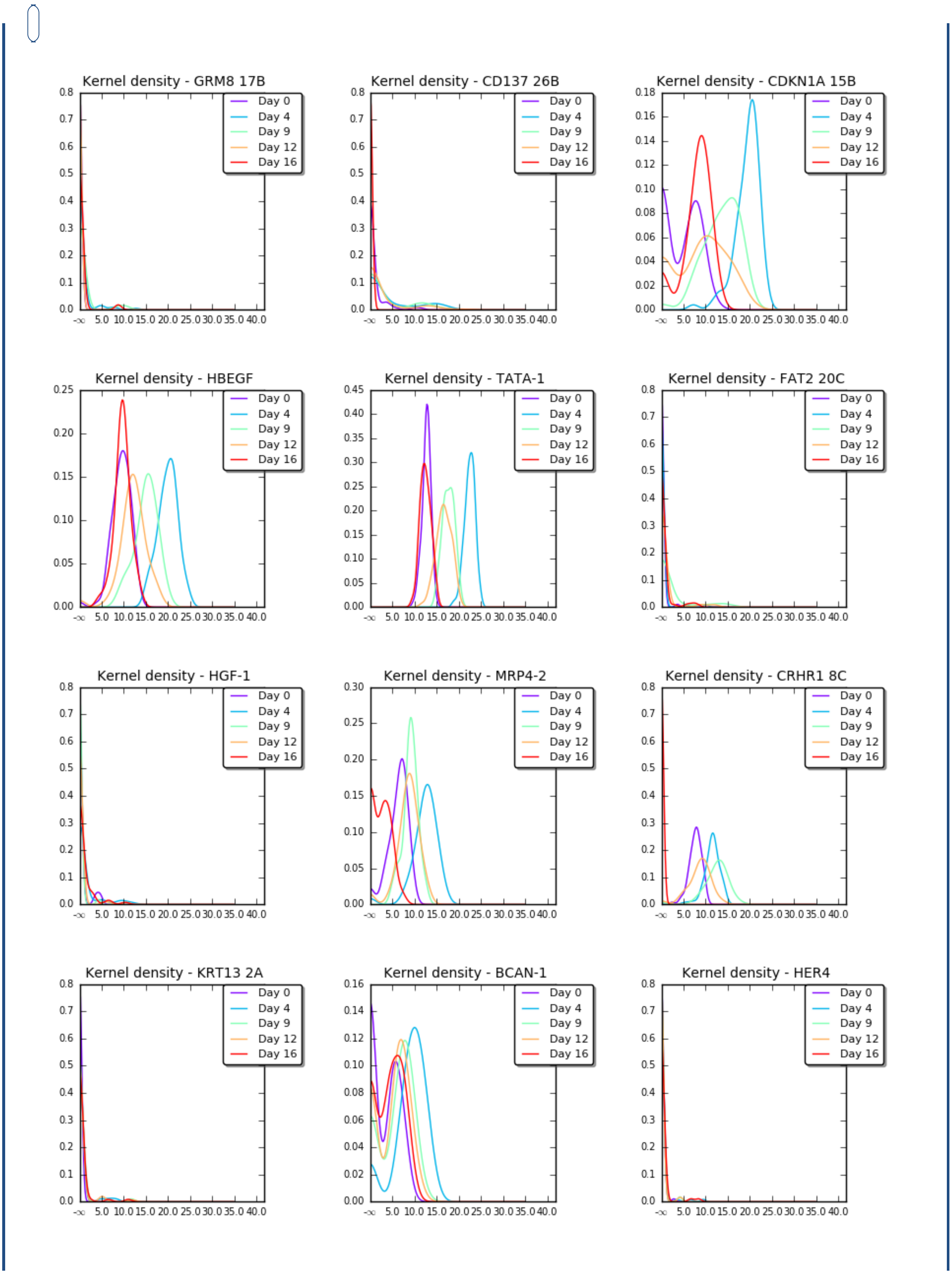
B.8 Single cell: Kernel Density plots

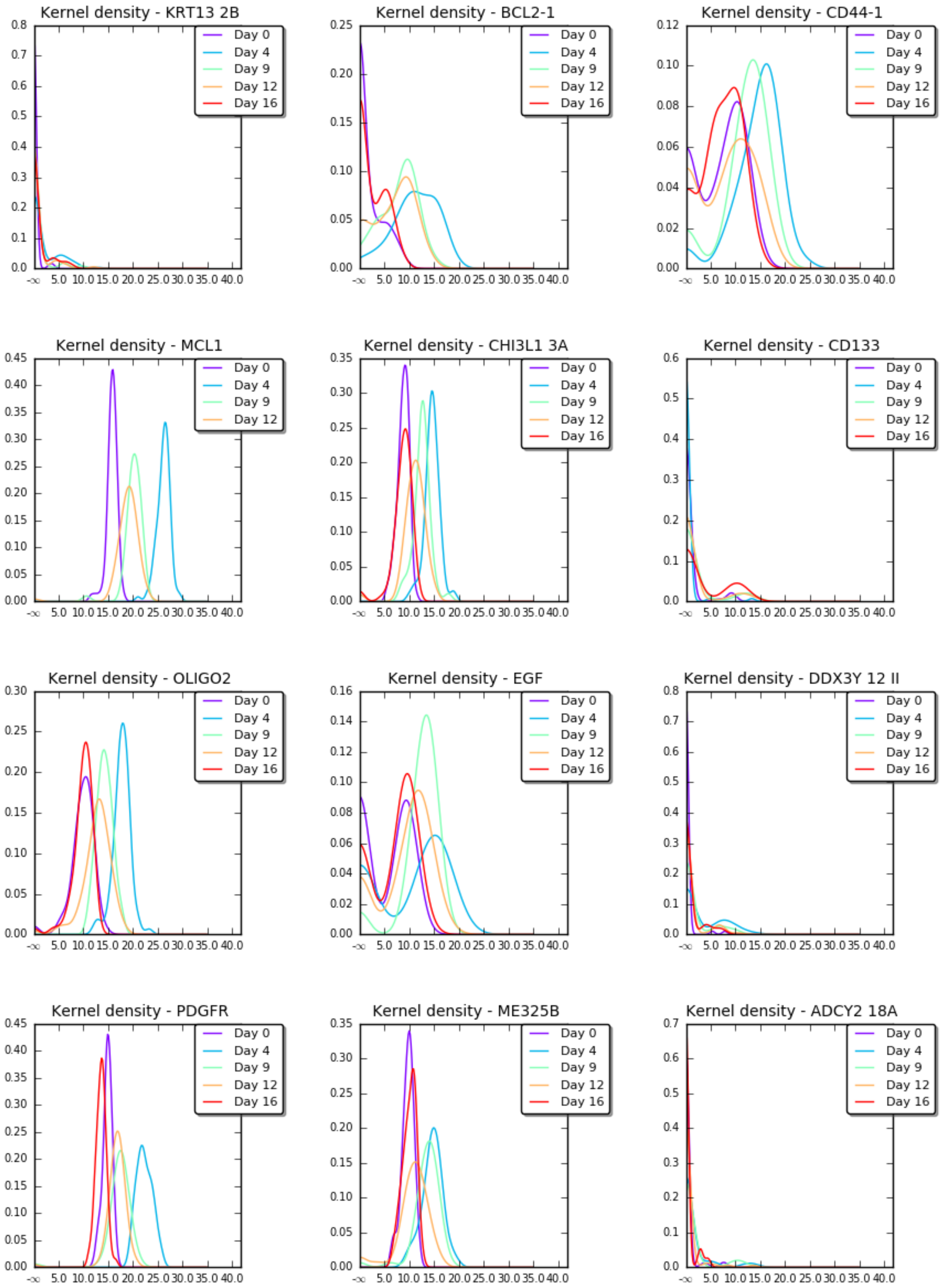
Kernel density estimations of genes after LRT filtering.

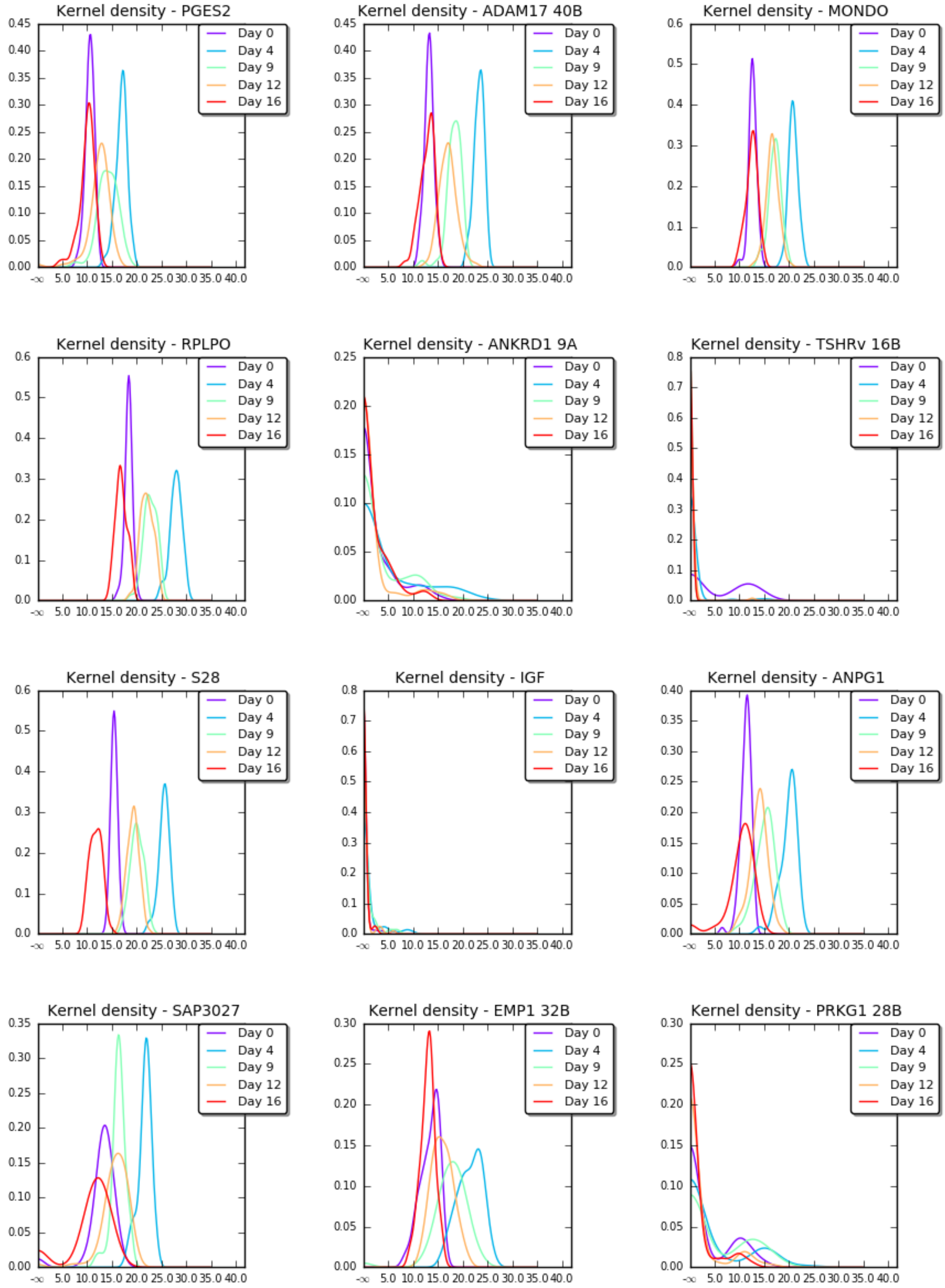


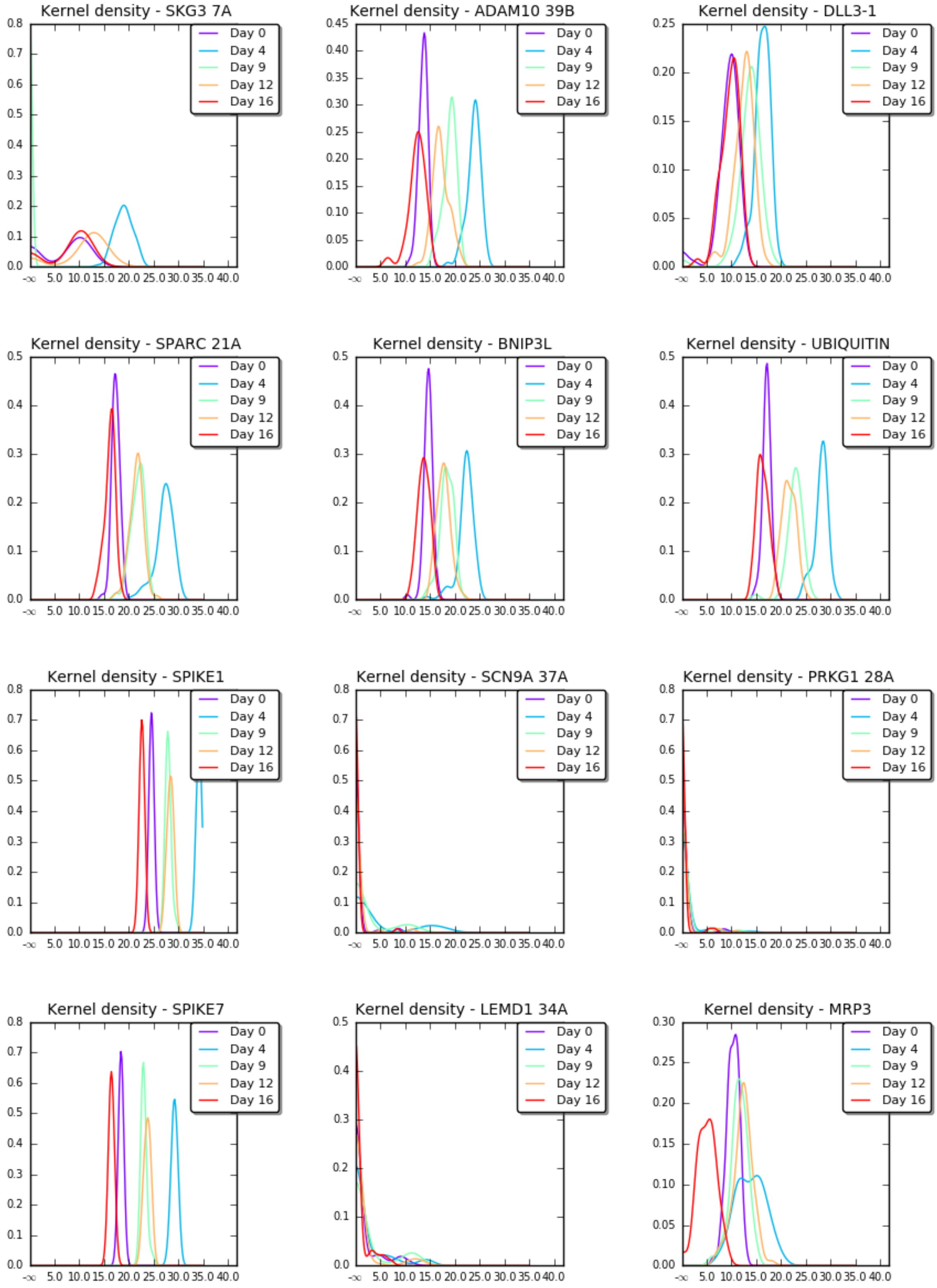

Kernel density - CD44-3
Kernel density - MGMT-1
Kernel density - CADH
Kernel density - DDIT3 35B
Kernel density - EFEMP 29A
Kernel density - BAG3
Kernel density - VEGFR1
Kernel density - EFEMP-1 29B
Kernel density - SPIKE4
Kernel density - C-Met 1
Day 0
Day 4
Day 9
Day 12
Day 16
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
-∞
5.0
10.0
15.0
20.0
25.0
30.0
35.0
40.0

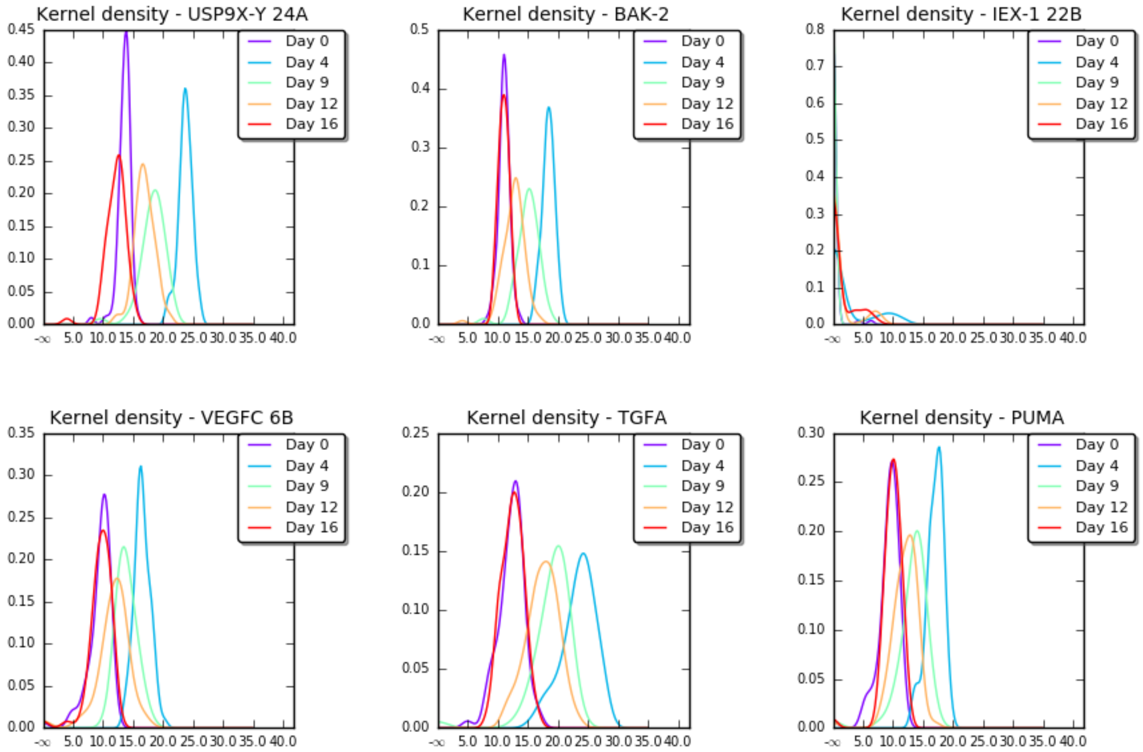

Kernel density - EGFR
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - PDGFRA
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - EMP1 32A
0.25
0.20
0.15
0.10
0.05
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - GAPDH
0.5
0.4
0.3
0.2
0.1
0.0
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - DKK1 33B
0.20
0.15
0.10
0.05
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Kernel density - BCL-XL
0.25
0.20
0.15
0.10
0.05
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - GFAP
0.20
0.15
0.10
0.05
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - MEDAG 36B
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - BMPER 38A
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - GMR8 17A
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - COX2
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16
Kernel density - KLK5 14B
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-∞ 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0
Day 0
Day 4
Day 9
Day 12
Day 16


Kernel density - GRM8 17B
Kernel density - CD137 26B
Kernel density - CDKN1A 15B
Kernel density - HBEGF
Kernel density - TATA-1
Kernel density - FAT2 20C
Kernel density - HGF-1
Kernel density - MRP4-2
Kernel density - CRHR1 8C
Kernel density - KRT13 2A
Kernel density - BCAN-1
Kernel density - HER4
Day 0
Day 4
Day 9
Day 12
Day 16
-∞
5.0
10.0
15.0
20.0
25.0
30.0
35.0
40.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.05
0.15
0.20
0.25
0.30
0.35
0.40
0.45


Kernel density - KRT13 2B
Kernel density - BCL2-1
Kernel density - CD44-1
Kernel density - MCL1
Kernel density - CHI3L1 3A
Kernel density - CD133
Kernel density - OLIGO2
Kernel density - EGF
Kernel density - DDX3Y 12 II
Kernel density - PDGFR
Kernel density - ME325B
Kernel density - ADCY2 18A
Day 0
Day 4
Day 9
Day 12
Day 16
-∞
5.0
10.0
15.0
20.0
25.0
30.0
35.0
40.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.02
0.04
0.06
0.08
0.12
0.14
0.16


Kernel density - PGES2
Kernel density - ADAM17 40B
Kernel density - MONDO
Kernel density - RPLPO
Kernel density - ANKRD1 9A
Kernel density - TSHRv 16B
Kernel density - S28
Kernel density - IGF
Kernel density - ANPG1
Kernel density - SAP3027
Kernel density - EMP1 32B
Kernel density - PRKG1 28B
Day 0
Day 4
Day 9
Day 12
Day 16
-∞
5.0
10.0
15.0
20.0
25.0
30.0
35.0
40.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Kernel density - SKG3 7A
Kernel density - ADAM10 39B
Kernel density - DLL3-1
Kernel density - SPARC 21A
Kernel density - BNIP3L
Kernel density - UBIQUITIN
Kernel density - SPIKE1
Kernel density - SCN9A 37A
Kernel density - PRKG1 28A
Kernel density - SPIKE7
Kernel density - LEMD1 34A
Kernel density - MRP3
Day 0
Day 4
Day 9
Day 12
Day 16
-∞
5.0
10.0
15.0
20.0
25.0
30.0
35.0
40.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45



Kernel density - USP9X-Y 24A

Kernel density - BAK-2

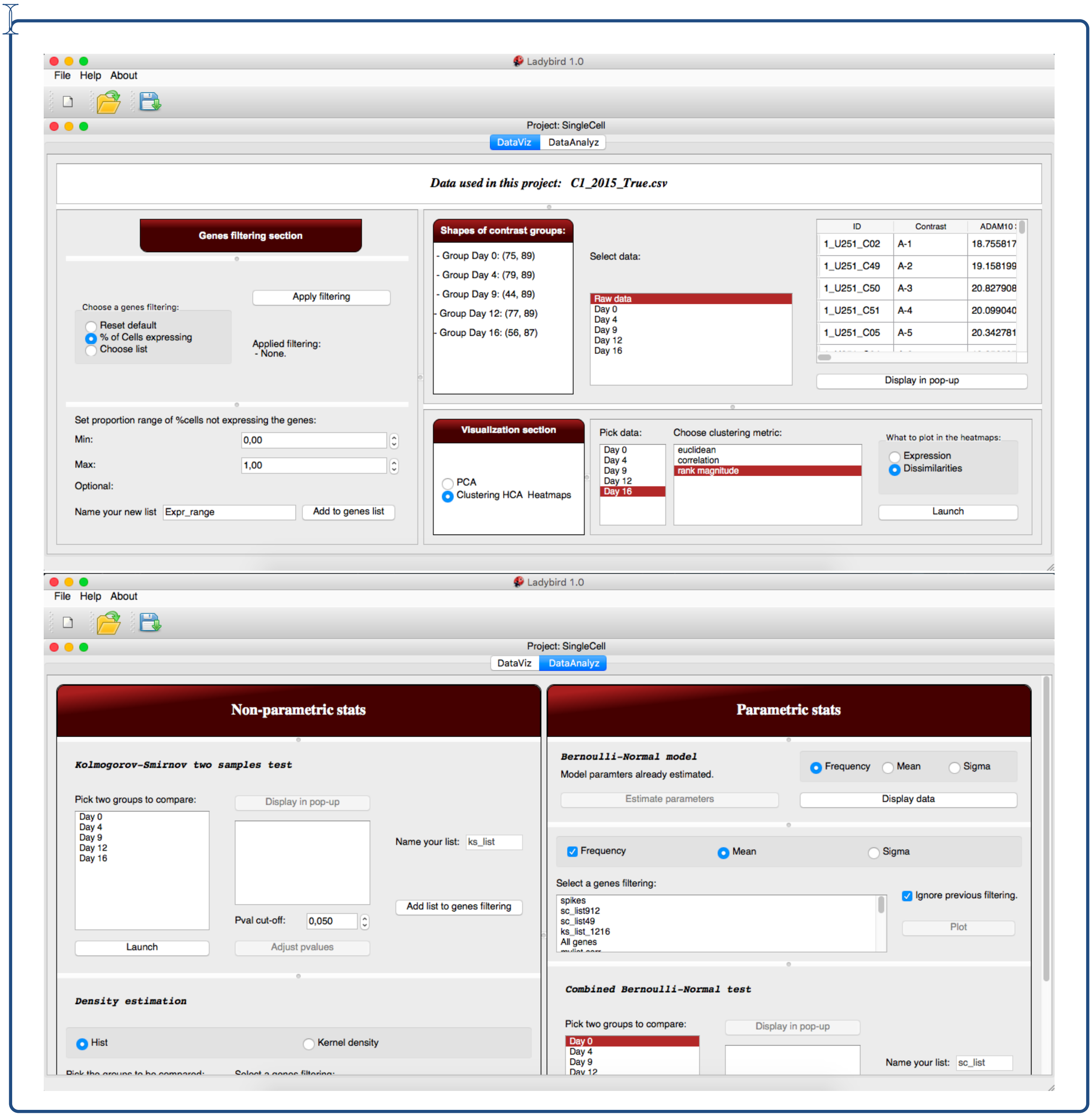
Kernel density - IEX-1 22B

Kernel density - VEGFC 6B

Kernel density - TGFA

Kernel density - PUMA

B.9 Single cell: LadyBird screenshots



Data used in this project: C1_2015_True.csv

Applied filtering:
- None.

Set proportion range of %cells not expressing the genes:

Min: 0,00

Max: 1,00

Optional:

Name your new list: Expr_range

- Group Day 0: (75, 89)
- Group Day 4: (79, 89)
- Group Day 9: (44, 89)
- Group Day 12: (77, 89)
- Group Day 16: (56, 87)

ID	Contrast	ADAM10
1_U251_C02	A-1	18.755817
1_U251_C49	A-2	19.158199
1_U251_C50	A-3	20.827908
1_U251_C51	A-4	20.099040
1_U251_C05	A-5	20.342781

Name your list: ks_list

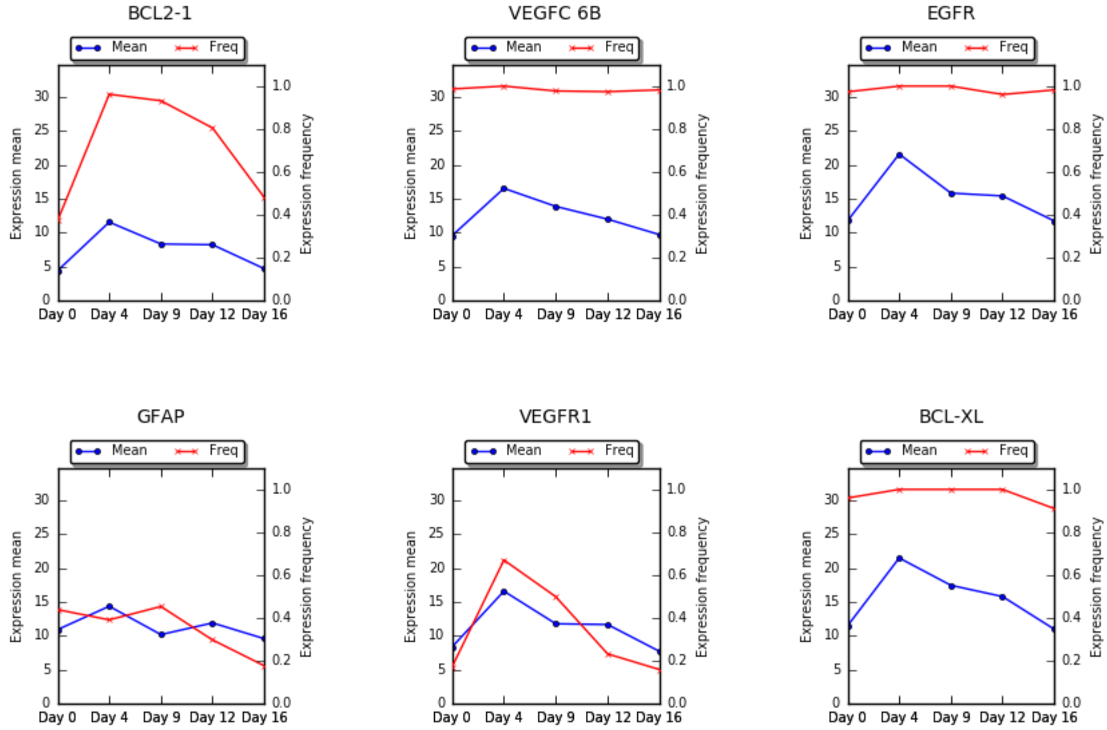
Pval cut-off: 0,050

Model paramters already estimated.

● Frequency ○ Mean ○ Sigma
Estimate parameters
Display data
☑ Frequency ● Mean ○ Sigma
Select a genes filtering:
spikes
sc_list912
sc_list49
ks_list_1216
All genes
☑ Ignore previous filtering.
Plot
Combined Bernoulli-Normal test
Pick two groups to compare:
Day 0
Day 4
Day 9
Day 12
Display in pop-up

Name your list: sc_list

B.10 Single cell: evolution of genes mean and frequency of expression that were highlighted in [2], generated by LadyBird.


BCL2-1
VEGFC 6B
EGFR
GFAP
VEGFR1
BCL-XL
Mean
Freq
Expression mean
Expression frequency
Day 0
Day 4
Day 9
Day 12
Day 16