project 4

(Michael Dymshits)

Identifying regulatory interactions Single cell RNA sequencing data allows the measurement of all genes in many cells at the same time. Genes are often co-regulated, or regulate each other. Correlation, anti-correlation or mutual information between pairs of genes can hint regulatory interactions, and help in reconstructing regulatory networks.

* Background single cell RNA sequencing
* Data – any single cell RNA-sequencing data.
* Filter out genes that are not expressed
* Replace values lower than one by one. log2 transform the data
* Identify pairs of genes that are significantly correlated (alternatively, anti-correlated or have mutual information). Take into account the problem of multiple comparisons
* Test if those gene pairs are also correlated in the other dataset.
* For at least one correlated pair of genes, see if you can find research about interaction or functional relationships between those genes.

Code at: https://github.com/michael135/comp\_bio

# Data sets description.

The data was taken from 2 papers one about melanoma and the second one about leukemia.

Papers details:

Gerber T, Willscher E, Loeffler-Wirth H, Hopp L et al.

Mapping heterogeneity in patient-derived melanoma cultures by single-cell RNA-seq.

Oncotarget 2017

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81383

Li S, Garrett-Bakelman FE, Chung SS, Sanders MA et al.

Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia.

Nat Med 2016 Jul;22(7):792-9.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83533>

First dataset has 307 cells and 19019 genes, the second one has 96 cells and 23737 gens.

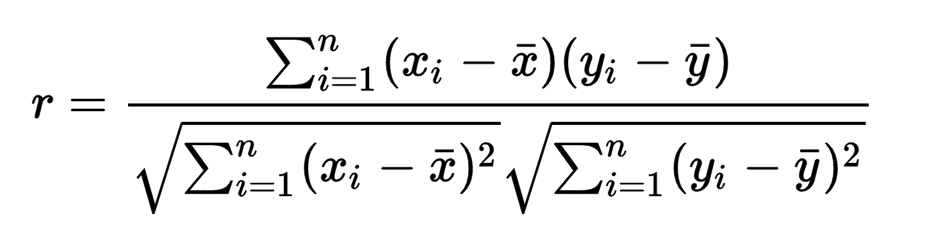
# Preprocessing

1. Normalization – log 2 for each value it it’s higher than 1 and 0 otherwise.
2. We removed the genes which has less than 40 values, which are higher than 1.

After step (2) the amount genes in melanoma dataset reduced to 8785 and leukemia dataset reduced to 894.

# Calculation of Correlation

Correlation coefficient was calculated by equation below, x and y bar a sample means.



using numpy library in python.

We start with a melanoma dataset, and start looking for a gene pairs, which has correlation coefficient higher than 0.85 we found 185 such pairs. Than we searched such pairs in leukemia dataset and printed those pairs in (leukemia) dataset which has correlation coefficient > 0.4. In brackets the correlation value for leukemia.

UQCRHL UQCRH

Corr = 0.83 (0.94)

HLA-DRA CD74

Corr = 0.54 (0.94)

HLA-DRB1 CD74

Corr = 0.54 (0.88)

HLA-DRB1 HLA-DRA

Corr = 0.49 (0.90)

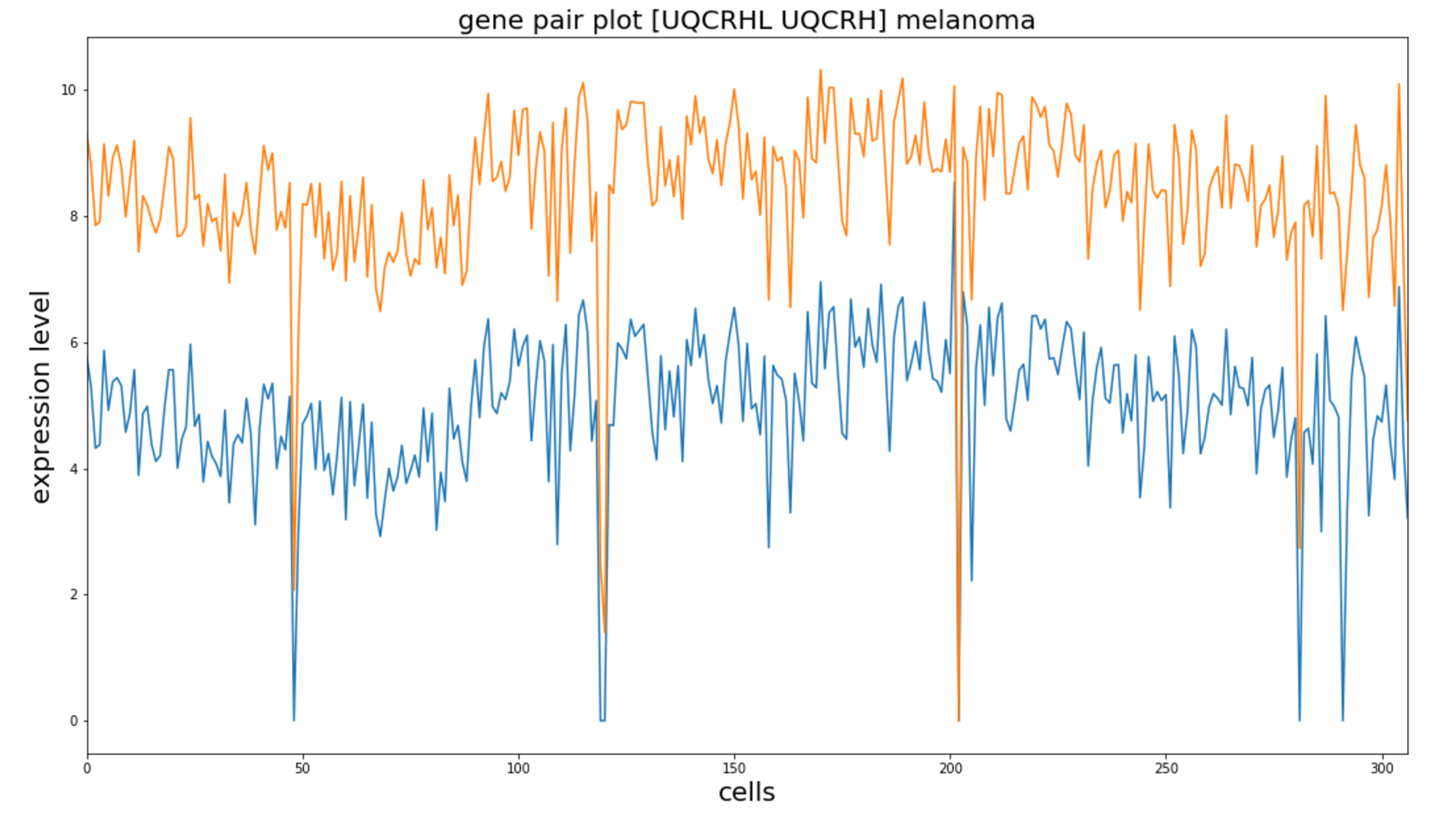
HLA-DRB1 CD74

Corr = 0.53 (.88)

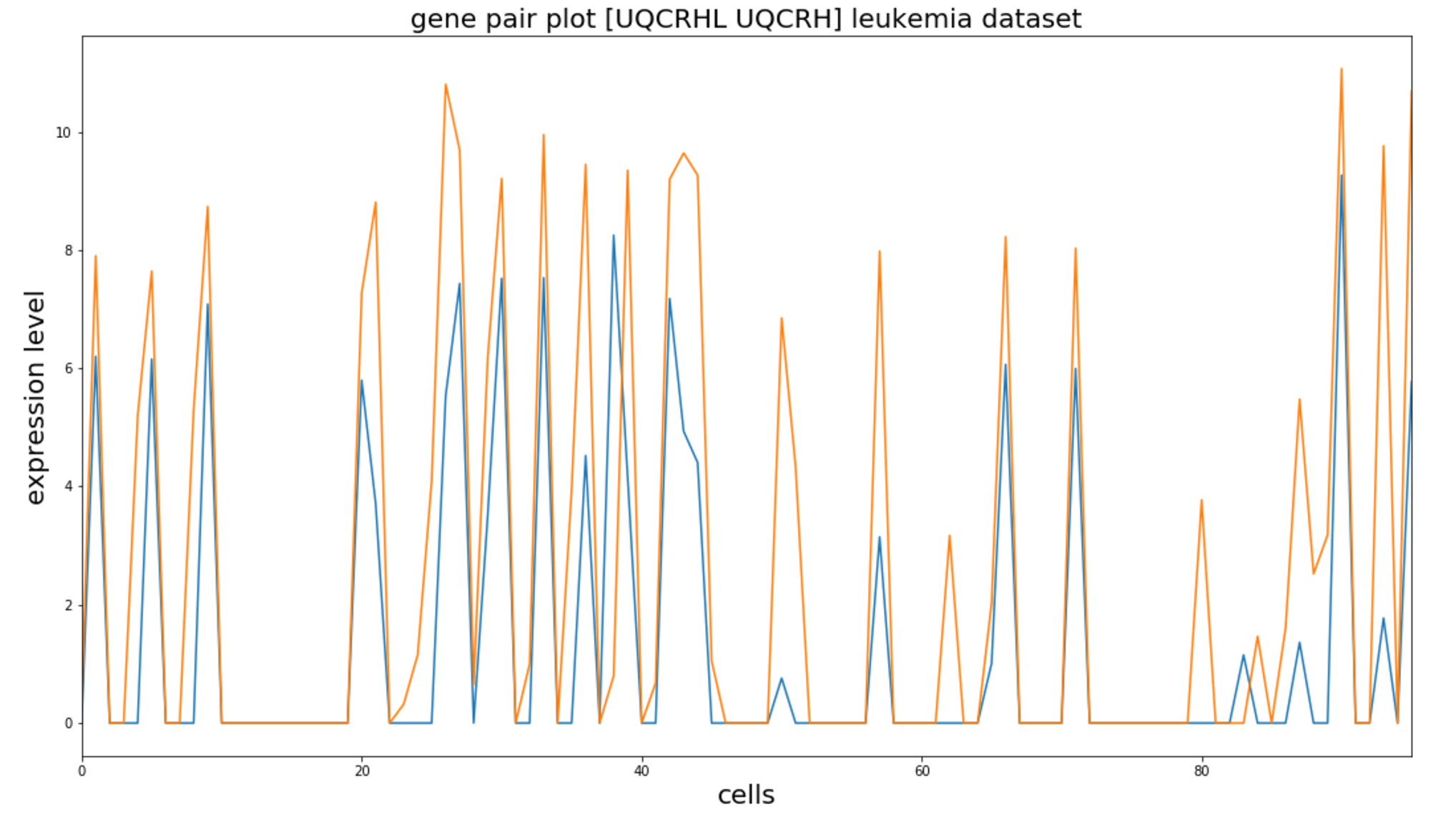
In our opinion the first pair of genes has the most meaningful correlation in both datasets. In next plots we will presents the cells value per each gene for both data sets.

## First pair of genes UQCRHL, UQCRH

Melanoma dataset (R=0.94)

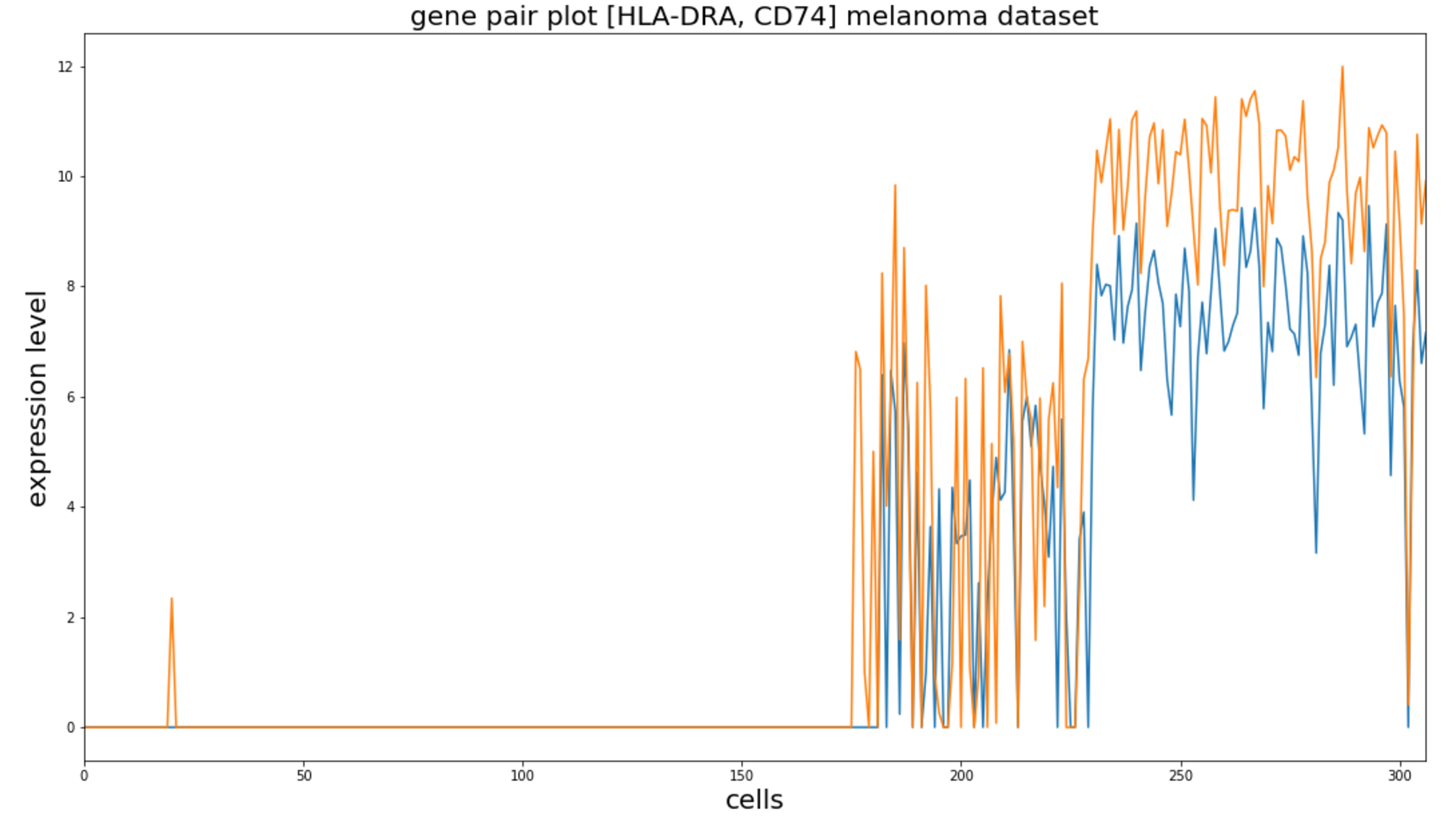


Leukemia dataset (R=0.83)

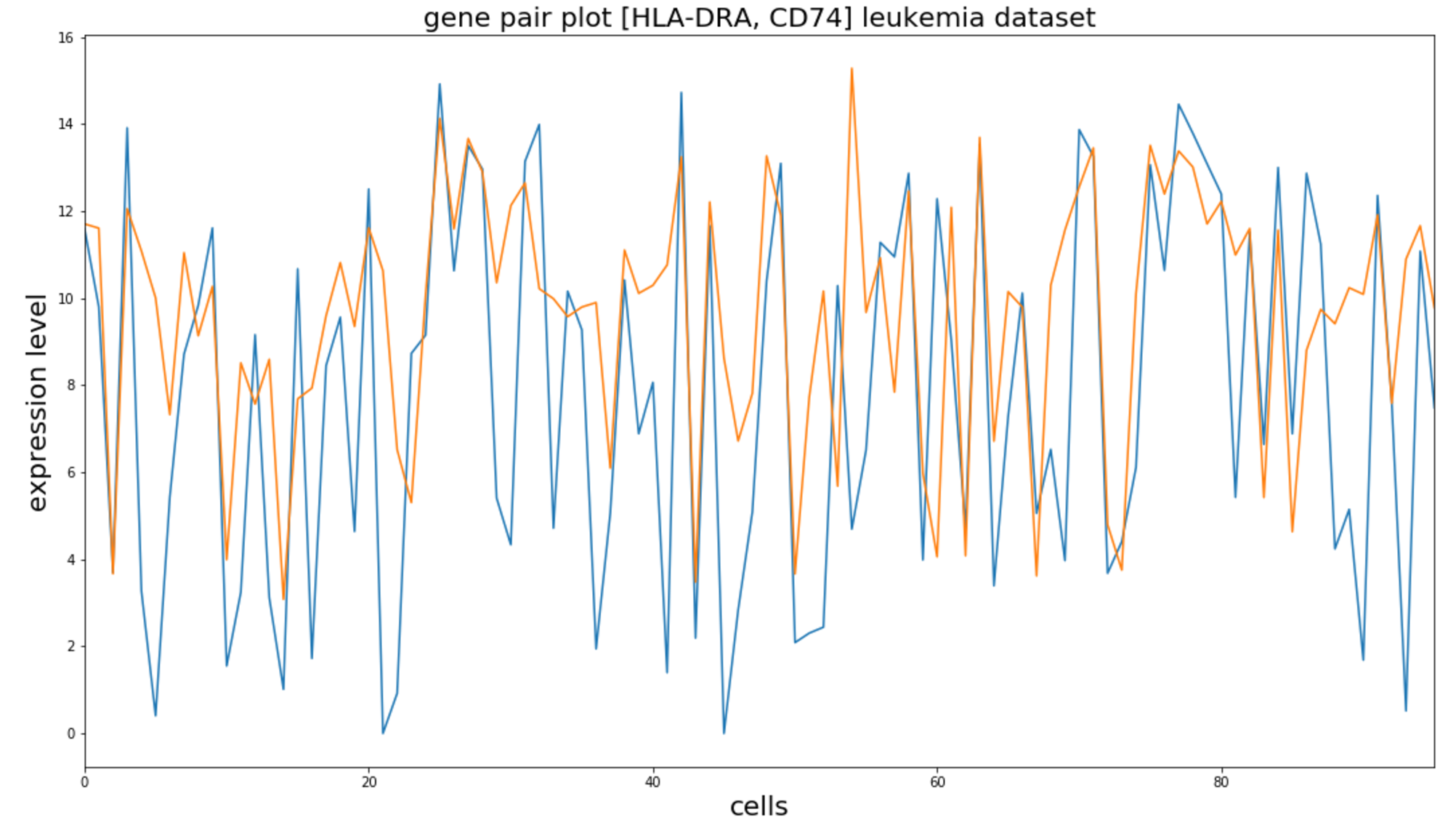


## Second pair of genes HLA-DRA, CD74

Melanoma data set (R=0.94):

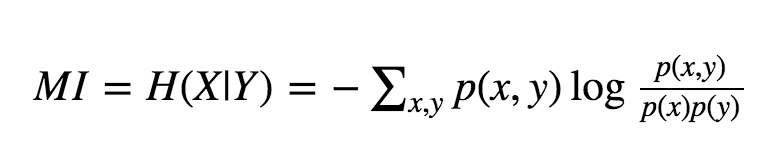


Leukemia data set (R=0.54):



# Mutual Information

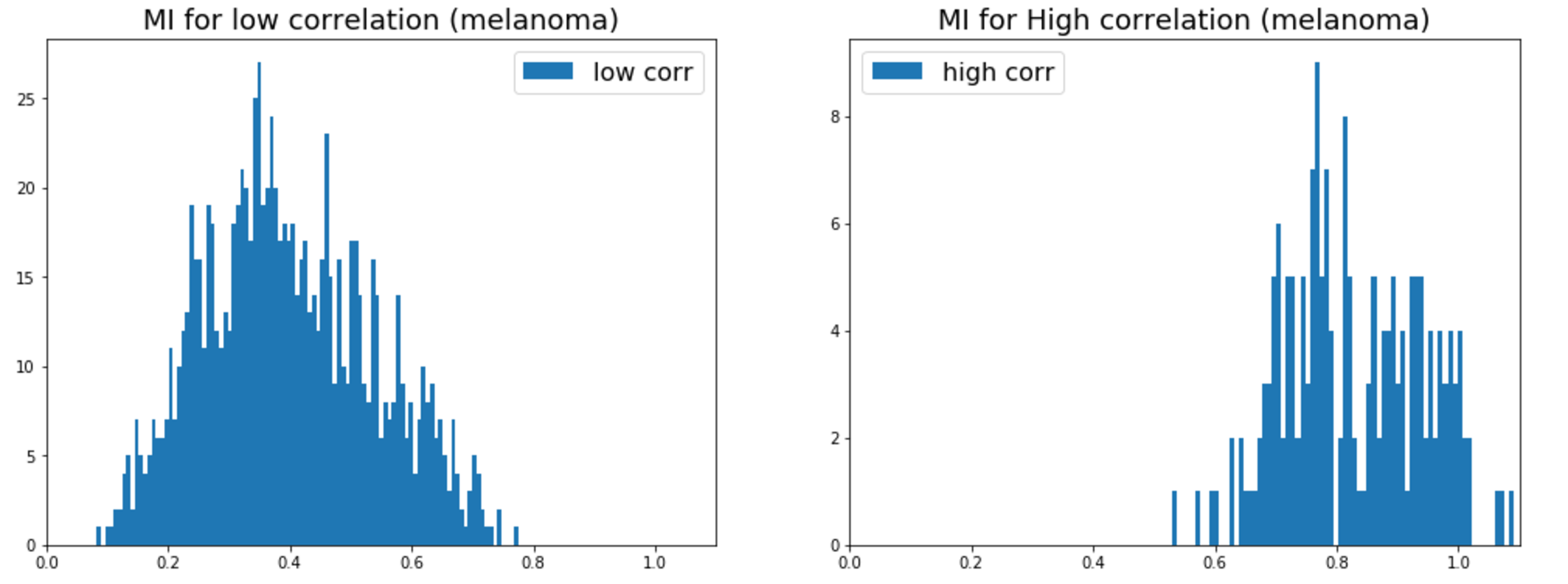
Mutual Informoation (MI) is presented in the below equation



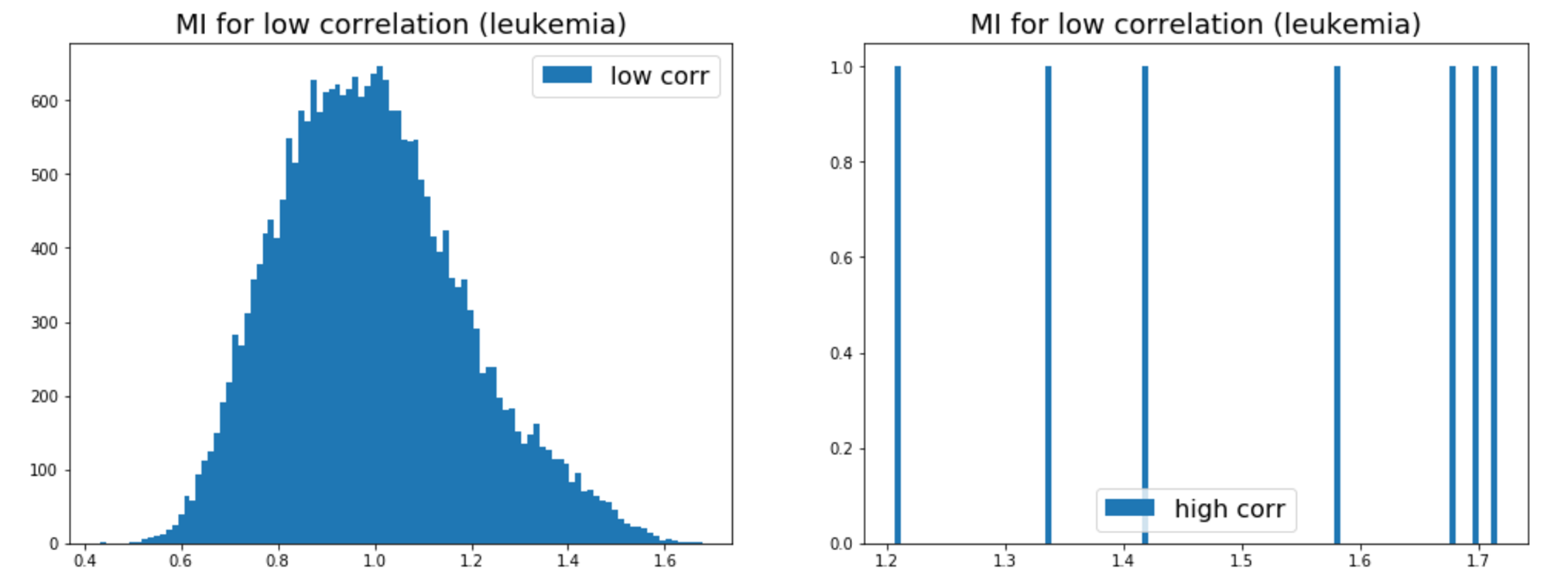
It can be seen as the distance (Kullback Leibler) between joint distribution on product of two marginal distributions.

Since it was a time consuming to calculate the MI for each pair of genes we calculated mutual information just for genes with high correlation and for sample of genes with a low correlation.

Next we will present plot for each data set.



The next plot is for leukemia dataset.



In the melanoma dataset we can see the meaningful difference between gene pairs with a high and low correlation.

From the leukemia we don’t see the meaningful difference between high and low correlation.

Some challenges in MI calculation. In order to calculate p(x) and p(y) we calculate histograms, for calculation of histograms we need to define the number of bins. Different number of bins may produce very different histograms and the histograms define the MI.