# Final Report

## Undergraduate Research

2014 Fall Semester

Name: Yuh Jen, Hwong

Student ID: R04945024

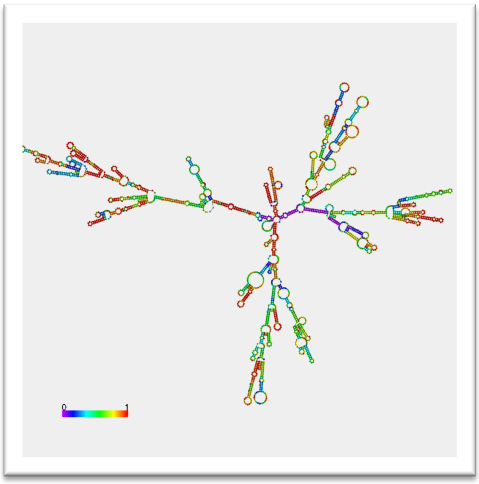
Advisor: Eric Y., Chuang

# Introduction

## *Motivation*

In the very beginning, my partner and I were introduced to the paper, “Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer” in which we acquired knowledge on lncRNA. In the paper, experiment was conducted on four types of cancer in search of their relationship with lncRNA. After reading the paper, we were curious if the same procedure could be applied to other diseases. Therefore, with suggestion and help from our instructors, we set off to run the experiment to our disease of interest individually. My target of interest is Stroke.

## *lncRNA*



Long non-coding RNAs, commonly abbreviated as lncRNA, is a new category of RNA that are non-protein coding transcripts longer than 200 nucleotides. Since lncRNAs do not encode proteins, their functions are closely related to their transcript abundance. A handful of lncRNAs have already been functionally characterized. However, little is known about functions of most lncRNAs in normal physiology or disease. Despite limited knowledge on these transcripts, growing evidence suggest that lncRNAs may also serve as cancer diagnostic or prognostic biomarkers.

## *Disease – Stroke*

Stroke is a loss of brain function due to a disturbance in the blood supply to the brain. It is a medical emergency that can cause permanent neurological damage or death. Overall, two thirds of strokes occurred in those over 65 years old. It is the second most frequent cause of death worldwide in 2011. The causes of stroke may vary. Usually it is classified into two major categories: ischemic and hemorrhagic.

Ischemic stroke is caused by interruption of the blood supply where arteries that connect to the brain becomes blocked or narrowed. These blockages are often caused by blood clots that form in the brain-connected arteries. Another possibility is that blood clots from further away may be swept through the bloodstream and into narrower arteries within the brain. Fatty deposits within the arteries, also called plaque, can form blood clots.

Hemorrhagic stroke is rupture of a blood vessel or an abnormal vascular structure. Arteries in the brain will either leak blood or burst open. The hemorrhaged blood creates pressure for the brain cells and it damages them. Blood vessels then burst or spill blood in the middle of the brain or near the surface of the brain, where blood flows into the space between the brain and the skull. There are several causes for ruptures including hypertension, trauma, blood-thinning medications and aneurysms.

### *Reason of choice*

I chose Stroke because I've heard of this disease since I was small. Many friends of my father and uncles who work as general manager or chairman of a company have suffered from Stroke and I was so worried that dad might one day get one as well. Mother would tell me that this is due to unhealthy diet and high pressure so we should behave well when dad is back from work. Secondly, the quantities of sample on GEO are of a large enough (92) so the outcomes should be meaningful. Also, the samples are beautifully categorized into control and patient sets, so it would be possible for me to figure out which lncRNAs are upregulated or down-regulated in this disease.

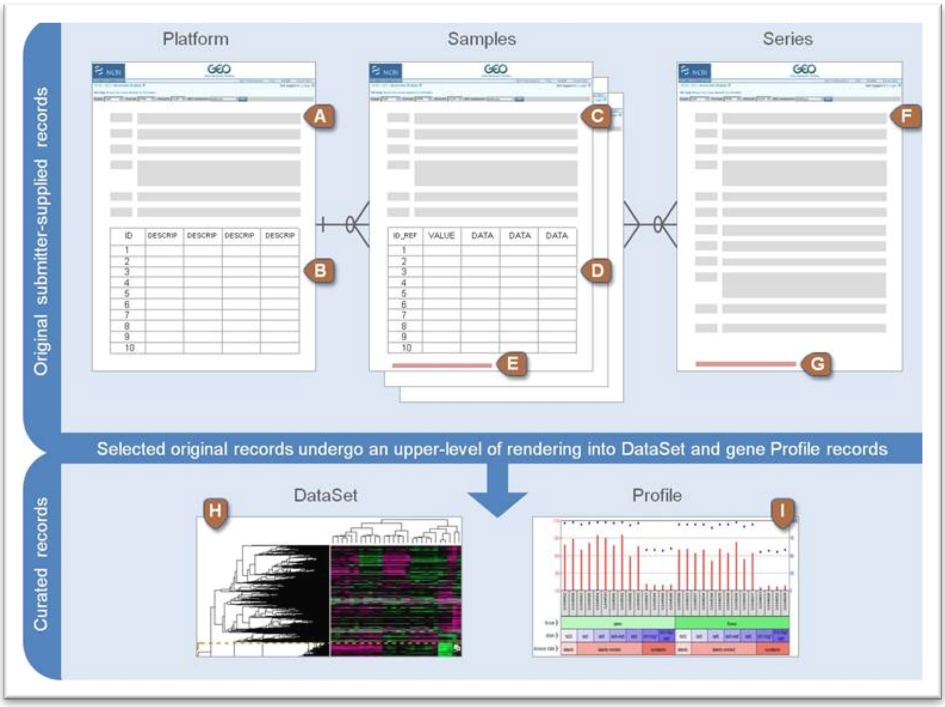
# Method

### *Background description of data used in this experiment*

The dataset I used was Series GSE58294, Gene Expression Following Cardioembolic Stroke. This dataset consists of 92 samples which were done on Platform GPL570, where 23 are controls and 69 being patients. Among the 69 patients, there were each 23 samples for records taken after 3 hours, 5 hours and 24 hours. The Series Matrix File of array-based data was downloaded from GEO NCBI.

## *Why choose to acquire data from GEO?*

Gene Expression Ombibus (GEO) is a global public functional genomics data repository. It archives microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. Therefore its data is not only of variety, but also complete and reliable. Furthermore, it provides user-friendly mechanisms that allow users to query, locate, review and download studies and gene expression profiles of interest. In addition, everything can be freely obtained. Thus, with zero budget, I managed to obtain a set of informative data on certain disease on specific platform efficiently.



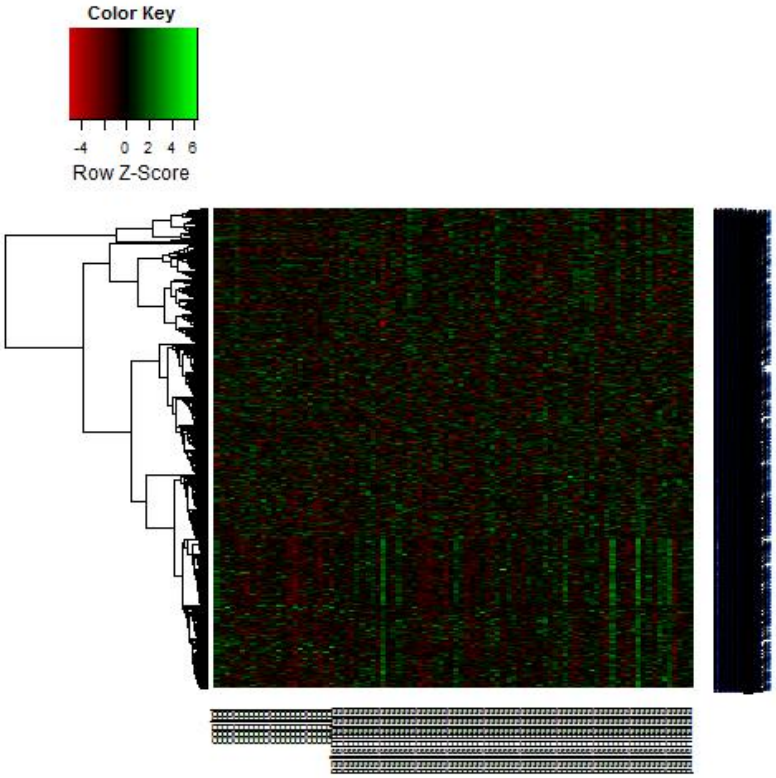

How GEO records are organized

## *Why is GPL570 the chosen platform?*

Before researching on Stroke, I ran through the procedure stated in the paper with the corresponding data on Lung SCC, to make sure that each step is understood correctly and that it would lead to the same result as shown in the paper. The platform on which it was ran on was GPL570. Throughout the procedure, one of the steps was to filter out the probes that consisit of lncRNA Ensembls. Since I picked out the probes that are valid from platform GPL570 before, selecting a dataset done on the same platform would save a step.

# Results

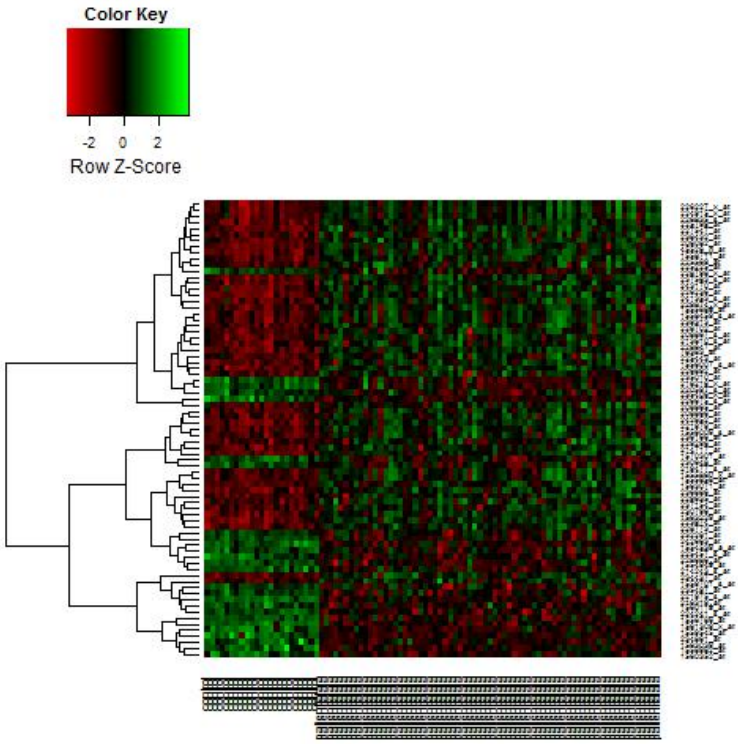
*HeatMaps (x-axis – samples, y-axis – probes) and Tables*




Heatmap 1 – Control vs. Patients after 3hr, 5hrs, 24hrs
Before t-test, all probes included
Number of probes – 3052

*Top 6 probes with smallest p-value and its respective lncRNA*

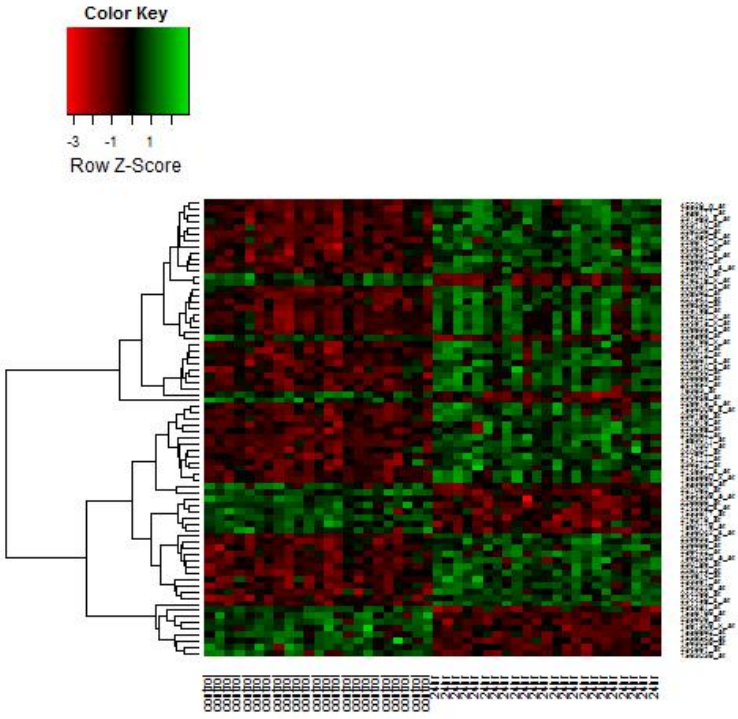
| **Probe** | **lncRNA Ensembl** | **p-value** |
|---|---|---|
| 216319_at | ENSG00000236673 | 7.309974*10^(-23) |
| 45526_g_at | ENSG00000262621 | 2.189319*10^(-21) |
| 1558786_at | ENSG00000260787 | 1.649436*10^(-18) |
| 241353_s_at | ENSG00000232533 | 1.938365*10^(-18) |
| 227547_at | ENSG00000260257 | 5.597988*10^(-17) |
| 1561309_x_at | ENSG00000230010 | 6.353367*10^(-17) |




**Heatmap 2** – Control vs. Patients 3hr, 5hrs, 24hrs

Probes with p-value < 1*10^(-18)

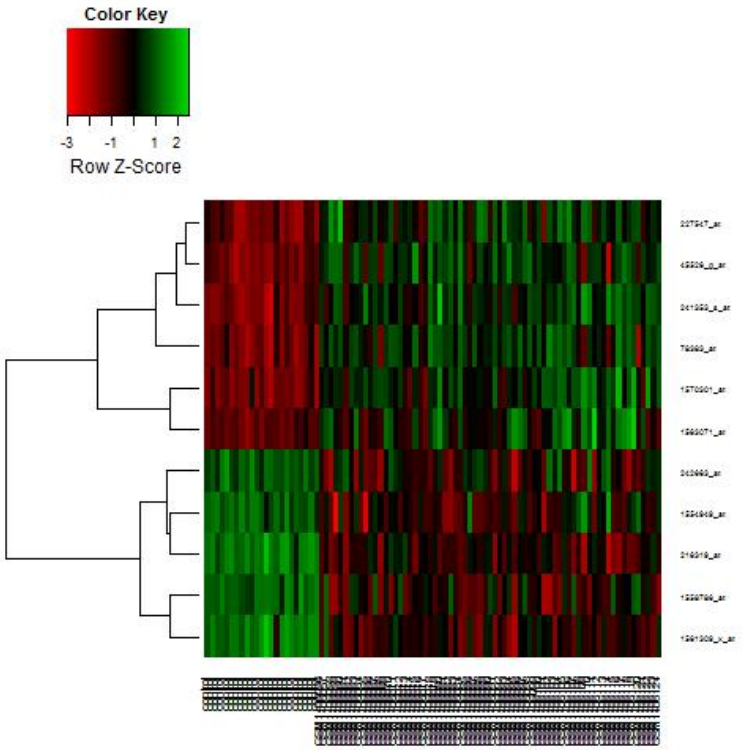
Number of probes – 75




**Heatmap 3** – Control vs. Patients 24hrs

Probes with p-value < 1*10^(-6)

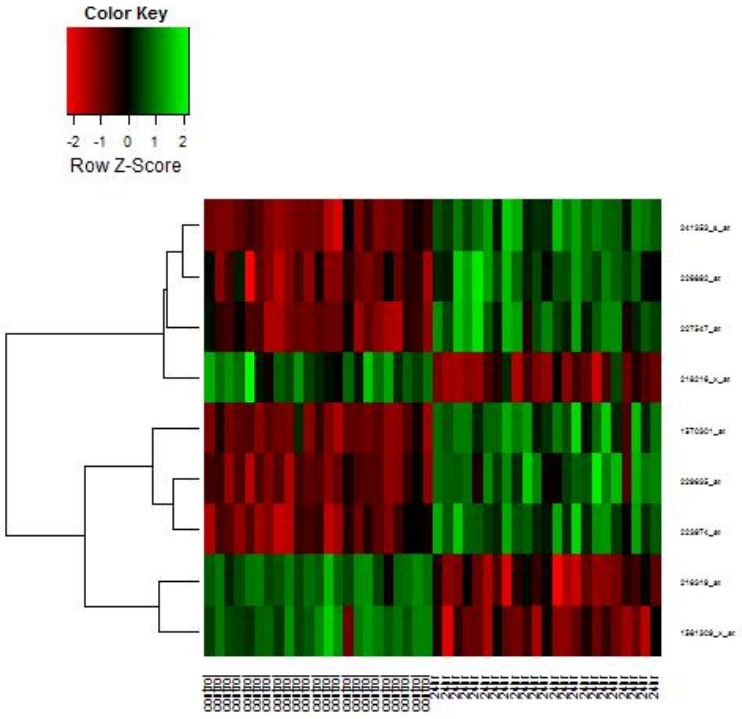
Number of probes – 74




**Heatmap 4** – Control vs. Patients 3hr, 5hrs, 24hrs

Probes with p-value < 1*10^(-15)

Number of probes – 11




**Heatmap 5** – Control vs. Patients 24hrs

Probes with p-value < 1*10^(-11)

Number of probes – 9

# *Data Analysis*

### *Heatmap 1*

From the figures above, we can see that even with clustering, the first heatmap is very jumbled. It is almost impossible to acquire any information from the picture. This is because we subjectively assume that only a few number of lncRNAs will affect a certain disease. The others will just behave randomly. Therefore, in order to retrieve useful information, we need to cut down the number of probes and leave only those that show significant difference between control and patient group.

### *Heatmap 2*

In Heatmap 2, only probes with p-value < 0.00000001 are selected to draw a heatmap, in which only 75 probes are selected. We can see that an obvious line lays between the margin of control and patient. However, the result isn't grouped beautifully, where usually the colors will be checkered in four quadrants. Possible causes will be addressed in the Discussion section.

### *Heatmap 3*

Later on, I plotted a heat map where only patients after 24 hours are compared to the control group. My assumption was that there will be slight differences in data records taken at different time durations, and that the longer the duration, the more significant the difference will be. However, the resulted map seems similar to Heatmap 2. Although an obvious line separates the control from the patient group, the colors did not distribute nicely into four quadrants either.

### *Heatmap 4 and Heatmap 5*

I moved on to filter the probes down to approximately ten. There are 11 probes for combined patients, and 9 left for patients after 24 hours. We can observe that the colors in Heatmap 5 are obviously grouped more neatly than Heatmap 4, despite the one probe that have the colors inversed.

*The 9 probes in Heatmap 5 in order from top to bottom*

| Probe | lncRNA Ensembl | Stroke patient vs Control |
|---|---|---|
| 1561309_x_at | ENSG00000230010 | Up-regulation |
| 1570301_at | ENSG00000254288 | Up-regulation |
| 216319_at | ENSG00000236673 | Up-regulation |
| 218216_x_at | ENSG00000256028 | Down-regulation |
| 223974_at | ENSG00000262001 | Up-regulation |
| 226892_at | ENSG00000231025 | Up-regulation |
| 227547_at | ENSG00000260257 | Up-regulation |
| 229635_at | ENSG00000251442 | Down-regulation |
| 241353_s_at | ENSG00000232533 | Down-regulation |

# Discussion

1. *How are probes that show useful information picked out?*

To effectively obtain probes that are relevant, t-test is applied to every row of the data. In the end, by setting the margin of p-value, we will be able to cut down the number of probes by only acquiring probes with p-value smaller than the margin. The smaller the margin value is, the fewer number of probes will be pick out and the probes will be the ones that contrast most between the two test groups.

2. *Why isn't Heatmap 2 in a regular pattern?*

One major cause is due to the feature of the disease. Stroke is a generic disease that does not have a specific cause. Factors of stroke include hypertension, trauma, blood-thinning medications, aneurysms… and so on. We are only informed that the patients chosen as sample in the dataset are victims of Cardioembolic Stroke. Since further information is unknown, we may assume that their causes of Stroke vary. Hence, it is predictable that different lncRNAs may have different effects on different patients. Therefore the heatmap did not appear in a regular pattern.

# Reference


- http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58294
- http://compbio.tongji.edu.cn/~duz/lncRNA/lncRNA_data_repository.html
- http://ccb.montana.edu/uploads/researcherfiles/McInnerney/MicroArrayPlatforms.pdf
- http://en.wikipedia.org/wiki/Stroke
- http://en.wikipedia.org/wiki/Long_non-coding_RNA
- http://www.cuilab.cn/lncrnadisease
- http://www.nature.com/nsmb/journal/v20/n7/full/nsmb.2591.html
- Neil Wu and Weiann Wang