**Locating SINEs**

**Locating SINEs**

Shira Rotshild

209251438

Shira Rotshild

209251438

Shalhevet Gamliel

316274042

Shalhevet Gamliel

316274042

Ofra Hadad

314733197

Ofra Hadad

314733197

Advisor: Dr. Anat Paskin-Cherniavsky

**1. Abstract**

Since the dawn of history people who have died of diseases have been considered as "dead prematurely", while death caused by aging process is perceived by society as normal, meaning that these people "died in their time". The mystery of aging has always existed and till these days resonates in the minds of many people around the world. In the past, many people have died of infectious diseases that are not related to the aging process and can also attack children and infants. These days science knows very well how to treat such infectious diseases, so life expectancy increased. In general, each person's life can be divided into two main periods: the period of development and growth and the following period in which a process of slow decay, weakening and finally the death. Another type of diseases that usually begins in this second period of life is diseases such as cancer, heart-related diseases, and the circulatory system. The cause of these diseases is much less pronounced and probably related to the aging process, which is characterized, among other things, by cell destruction and weakened body and immune system. Researchers in aging eld want to focus on finding the direct causes of the aging process and fight them. Their purpose is to understand what the causes are and how to neutralize them. In our work we intend to assist bio-informatics researchers in the aging field by developing an effective tool for finding mutations in DNA sequences. Given a collected data that contains genetic sequences of laboratory mice and an algorithm that analyze the data and find irregularities and changes in these sequences, we would like to explore and improve the algorithm. This tool hopefully helps the researchers in trying to understand more clearly the process of the formation of these aging cells in our body, and thus fight the cause of their formation.

**2. Background**

Don't you want to live longer and healthier life? Many actions and measures can help you, such as exercise or calories restriction, to improve your life quality. But that would not make a significant difference. Researchers in the eld of aging believe that in the future we will be able to live significantly longer lives. As opposed to popular belief, they do not take the aging process for granted, but believe that the source of aging lies in an abnormal processes that happen in our bodies, and are probably the cause of most diseases of a later age. Therefore, we should treat the aging process as a disease. Recent research in the field has put the spotlight on several processes in the human body that are suspected of causing aging. There are 7 major processes currently assumed to be the root causes, among them are Mitochondrial DNA Damage, Nuclear DNA Damage, The Failing Innate Immune System, etc. The purpose of the researchers is to examine how these processes are created and what effects they cause in our bodies. It is not entirely clear yet about the activity of these processes and about their relationships with each other, or which of these processes is the first in the chain and influences the rest. These processes all happen at the low-level phase, and therefore they are not involved in more complicated and complex systems of the human body. Therefore, the researchers hope that as soon as they can find out what is the main and primary factor of all, and what is its relation to the other processes, the treatment of those factors will be simpler and hopefully will take less time.

**3. About the research**

one of the processes suspected as causing aging is cumulative damage to cellular DNA. We assist in aging research focused on researching this process (1) . The research is conducted by Dr. Andrei Gudkov of the Roswell Cancer Research Department. Researchers believe that an outbreak of early viruses found dormant in the bodies of mammals is the primary cause of aging. These viruses are expressed as mutations created in the cellular DNA and accumulate as "junk" likely to disrupt the normal functioning of the DNA (2). These mutations are referred to as "retrotransposons" (3) - genetic factors that copy themselves into RNA and then converted back into DNA and inserted into the genome. They have long repetition zones at their edges and so they can be identified. There are various types of retrotransposons, some of which are called SINEs (Short interspersed nuclear elements) (4). These are short sequences of DNA that are not encoded and appear in millions of copies within the DNA. Dr. Gudkov hypothesizes that the appearance of retrotransposons may "poison" the DNA and cause aging cells, which contribute to develop diseases and accelerate the aging process.

**4. The data**

We received a DNA of 2 mice are very close genetically, one young and one old. The DNA is broken into millions of random size pieces. The cutting process is probabilistic, but most of the pieces are between 200 to 300 bases length. Each piece is forming a small part of the original DNA coil, which is made up of 2 strands(5). As part of this breaking process and its randomization, the order of the pieces is not maintained and, in this situation, the complete DNA helix cannot be easily reconstructed. Now for each part, the laboratory staff performs a parallel reading of the two strands. 151 bases are read from each strand. For the first strand, 151 bases will be read from left to right. And for the second strand, 151 bases from right to left. Some overlap will appear between the 2 reads since each piece from the previous step is on average less than 300. At the end of the process, all of the readings from left to right inserted into a file named "R1.fastq" (6) and all readings from right to left inserted into a file named "R2.fastq".

**5. Previous works**

At first researches have worked with RNA because searching in the DNA has great time complexities. The RNA search has several disadvantages-it difficult to catch when the SINE active in the RNA, SINE that in the RNA not necessarily will insert into the DNA,etc. Therefore, it is more accurate to perform the search on the DNA, that why they develop this algorithm which shortens the complexity of time in DNA search.

We continue a project that implement the algorithm (7). The search process begins with looking for all the SINEs appearances in the DNA - with 22% error (considering possible changes such as insertion, deletions, and exchanges when searching-we call it edit distance). The second stage decides which of the SINEs is new by comparing barcodes - 36 length prefix from each SINE. They decided the best edit distance is 3, and that why they split the barcode to 4 windows. To compare all the barcodes, we use a data structure of buckets, where the key to each bucket is one of the possible combinations of the DNA letters( A,T,C,G,N) size 9 - the max number of buckets is 5^9 ( the structure is building while running so it's actually smaller). Now we run over all the barcodes, and for each barcode we take all the sub-strings size 9 (28 windows) and put the barcode in all the buckets that their key match to one of the windows. The next stage, we take all the barcodes of the potential new SINEs, and each one we split to 4 (window size 9) and compare to the barcodes at the match buckets. If the barcode appears more than one time, the search stops and the barcode classified as inherited, else its classified as new SINE. When they ran the code on the data (mouses DNA) the results were inconsistent and didn’t match the researcher believe and the RNA conclusions.

**6. Project Goal**

The general goal is to help the researchers to verify their hypothesis by developing a computational tool that can find those SINEs sequences in DNA and classify them as those which created in postnatal DNA, and those which have been inherited. With this tool, the researches will be able to check whether there is a correlation between the number of instances of SINEs, and the age of the mouse. We hope that this tool will help the researches validate Dr. Gudkov's hypothesis - that the DNA sequence of an adult mouse contains more SINEs segments than that of a young mouse, which indicates that the process of retrotransposons is directly related to the aging process. In our project we focused on exploring the reason to the previous project results, were the assumptions about the data incorrect? Or maybe the code is wrong? Or is the hypothesis incorrect?!

**7 Research Phases**

**7.1 Obtaining the distribution of the data.**

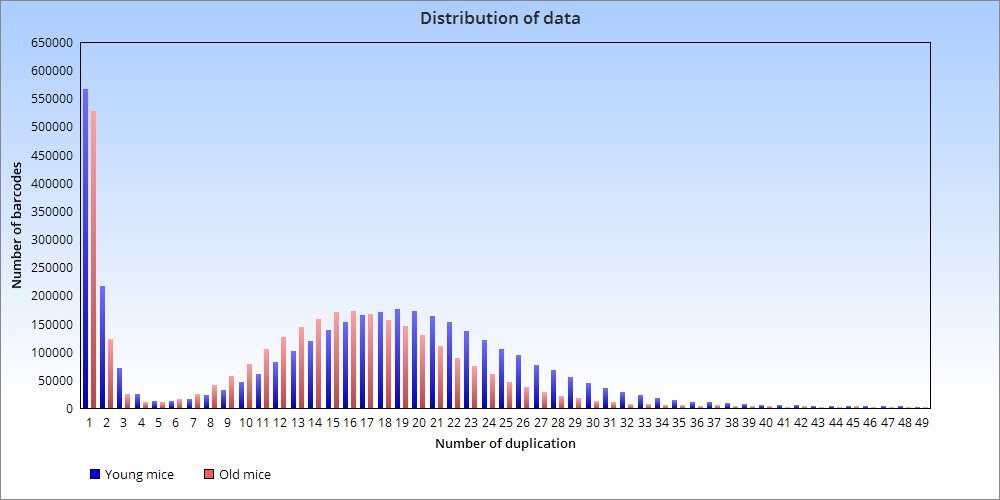
First, we wanted to understand the distribution of the data.

One of the assumptions the researchers gave is that the data has about cover number 30 )each sequence appears an average of 30 times)

To check the assumption, we added to the code a counter, for each barcode, that check how many times the barcode appears in the data.

We ran the code on all the barcodes, unlike the previous project, that used only the potential new barcodes-barcodes without duplication in the DNA with distance 0.

Here are the results for the liver tissue withSINE B1**.**



In the histogram we see that most of the barcodes cover number is less than 30 and bigger in the young mouse.

In addition, we see from the results, that like the previous project result, there are more new SINEs in the young mouse than in the older mouse- the opposite of the research expectations.

We noticed that the number of SINEs in general, is larger in the young mouse than in the older mouse- therefore, it can be concluded that more information was lost in the older mouse than in the younger mouse.

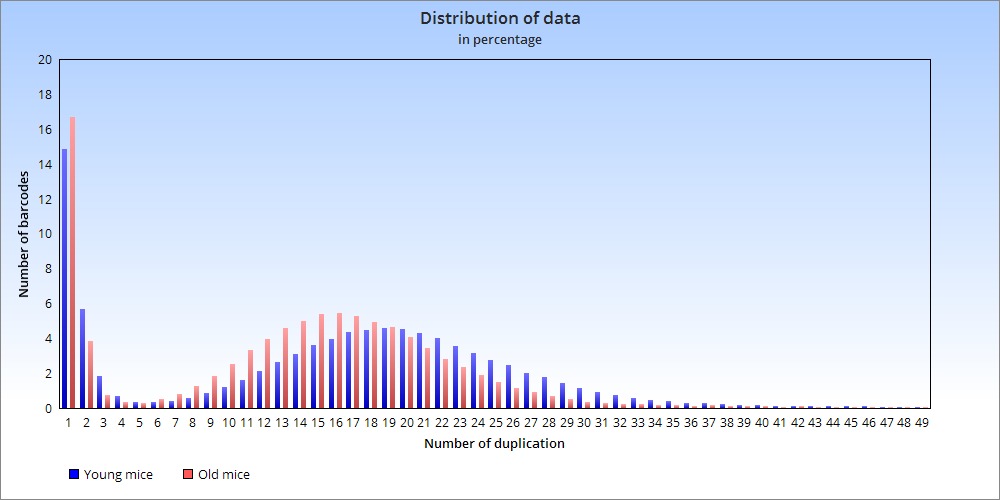
This is perhaps one of the reasons for the unexpected and inconsistent results obtained in the previous project.

**7.2 Obtaining distribution in percentage**

Due to the results we got, we understand that we need to look differently on the results.

Therefore, we added in the code a calculation of the distribution in percentage.

Here are the results of distribution in percentage for the liver tissue with SINE B1.

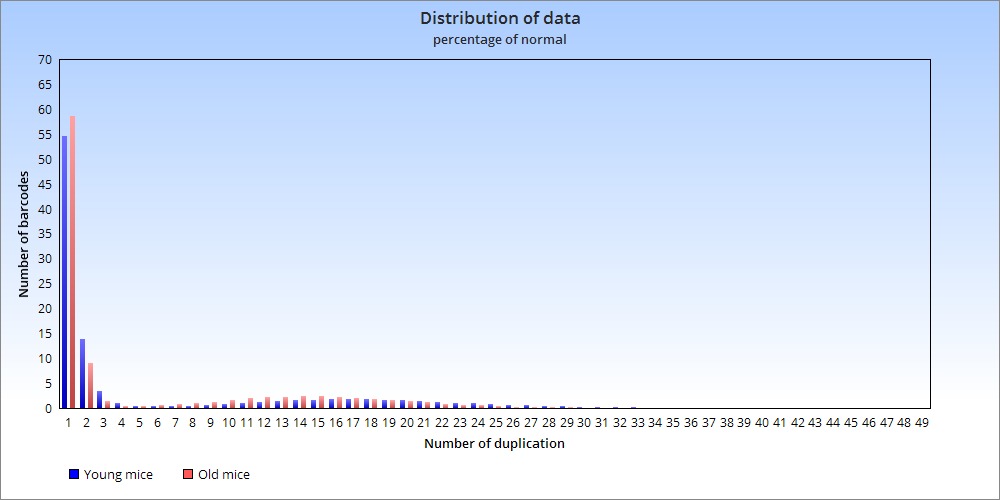


We can see now that there are more new SINEs at the older mouse than the younger one- in line with the research expectations

In our perception every SINE duplication is the same as the original SINE, that why we decided to normalize the histogram.

For the duplication 2 we count every SINE twice (although it points to one place in the DNA) therefor we will divide by 2, and so on.

Here are the results of distribution in percentage of normal for the liver tissue with SINE B1.



we can see in the histogram that more than 50% of the barcodes are new in both mice.

The results are not so reasonable, so it is not certain that it is correct to normalize the histogram. We have not reached a final conclusion about normalization, it remains an open question

**7.3 crossing the data**

Because we saw a lot of information lost, we wanted to get more accurate results, and check if the SINEs we found as new are new, or they are inherited that their duplications got lost. We added to the code a test that compares the barcodes of the young and the old mouse. We took the barcodes that we found as new in the young mouse, and check if they appear in the old mouse data structure (And vice versa). If so, the SINE is probably inherited, therefore we will not continue to consider it as new.

The 567,222 new SINEs of the young crossing the old data

|  |  |
| --- | --- |
| **Number of duplications** | **Number of barcodes** |
| **0** | **96.13414148252359** |
| **1** | **3.096318548998452** |
| **2** | **0.5320668098204935** |
| **3** | **0.11300690029653292** |
| **4** | **0.034554372009548286** |
| **5** | **0.015690505657396926** |
| **6** | **0.009696379900638551** |
| **7** | **0.00634672138950887** |
| **8** | **0.004760041042131652** |
| **9** | **0.005641530124007884** |

The 526,829 new SINEs of the old crossing the young data

|  |  |
| --- | --- |
| **Number of duplication** | **Number of barcodes** |
| **0** | **95.27417814888702** |
| **1** | **3.4001545093379444** |
| **2** | **0.8530282121902932** |
| **3** | **0.24163438231380582** |
| **4** | **0.08143059702484108** |
| **5** | **0.03340742442044762** |
| **6** | **0.018981491147981602** |
| **7** | **0.014805563095425649** |
| **8** | **0.01006019030843025** |
| **9** | **0.009300930662510986** |

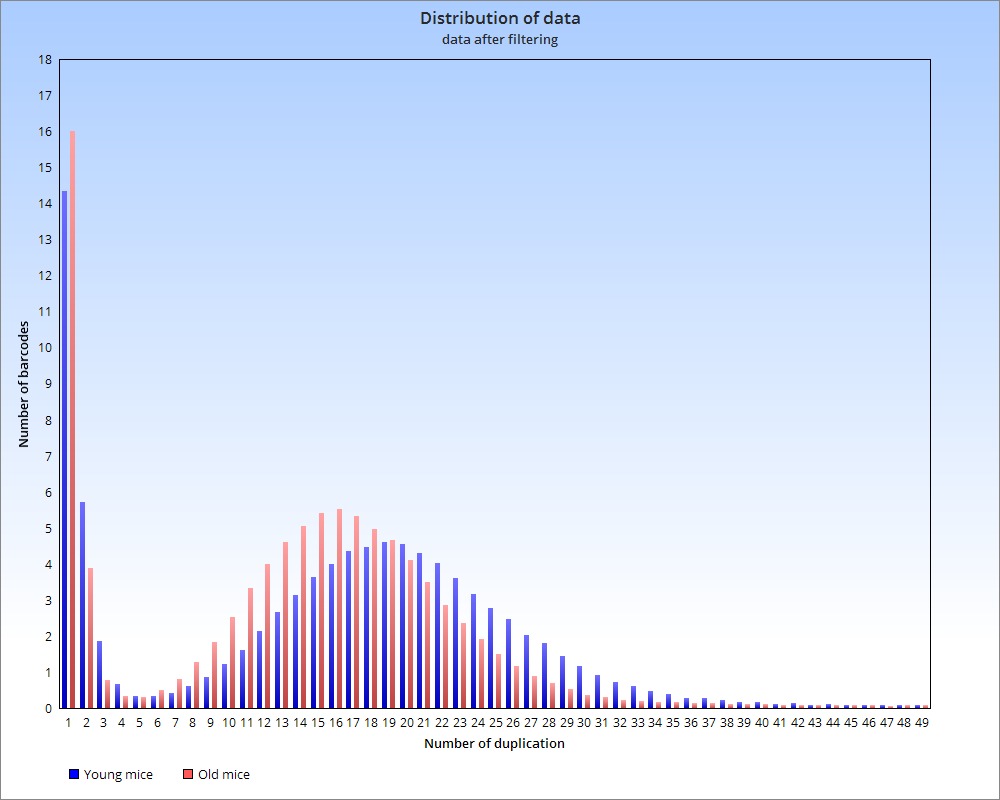
the 526,829 new SINEs of the old crossing the young data

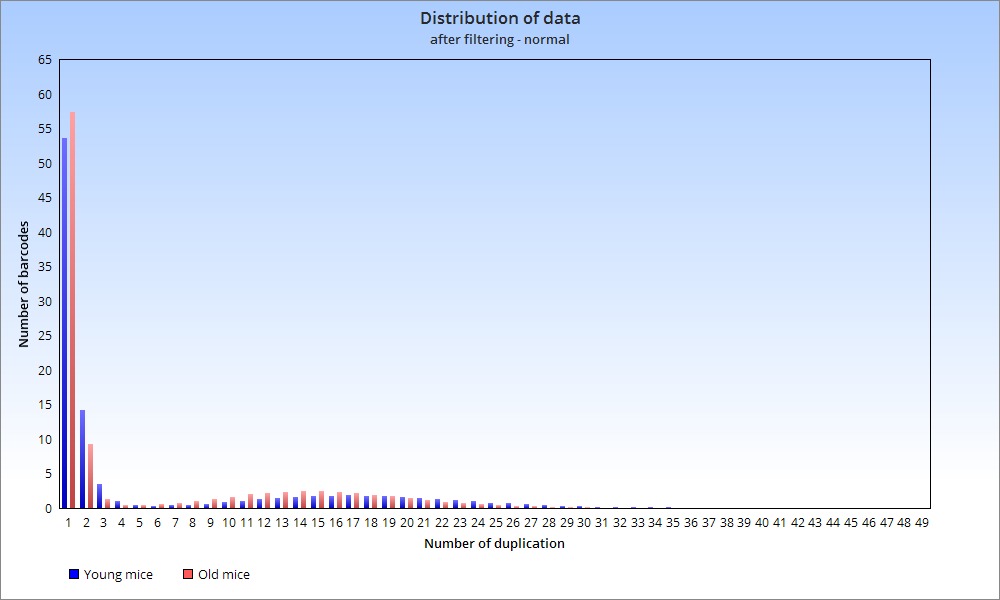
|  |  |
| --- | --- |
| **Number of duplications** | **Number of barcodes** |
| **0** | **95.27417814888702** |
| **1** | **3.4001545093379444** |
| **2** | **0.8530282121902932** |
| **3** | **0.24163438231380582** |
| **4** | **0.08143059702484108** |
| **5** | **0.03340742442044762** |
| **6** | **0.018981491147981602** |
| **7** | **0.014805563095425649** |
| **8** | **0.01006019030843025** |
| **9** | **0.009300930662510986** |

We can see that most of the SINE we considered as new still new.

As we assumed before, in the older mouse we lose more information, and we can see here, that more new SINEs discover as inherited in the old one.

And here are the results after updating the data





We can see from the histogram that in the adult (in the tissue of the liver) there are more new SINEs than in the young.

It therefore appears that we are on the right path to prove the researchers' hypothesis.

Of course, to strengthen the claim, we need to run on more tissues and more SINEs.

**8. Future work**

**\*** Run the algorithm on more tissue from other organs

\* Run on different SINEs (B2, B4)

\* Do crossing between organs

**\*** DNA has two sides, we ran the algorithm on one side, we would like to test on the other side as well.

**Bibliography**

1. https://docs.google.com/document/d/1fN6mBN-lIVcPIny8SHd8hyrukiLgo4sJZqc2Dl7Ua8/edit

2. https://www.youtube.com/watch?v=L3irS84U8rs

3. https://en.wikipedia.org/wiki/Retrotransposon

4. https://www.illumina.com/documents/products/illumina\_sequencing\_introduction.pdf

5. https://www.illumina.com/documents/products/illumina\_sequencing\_introduction.pdf

6.https://en.wikipedia.org/wiki/FASTQ\_format

7.Binyamin Yitzhak and Ayelet Yerushalmi.

An algorithmic framework for testing the contribution of retro-transposon activity to degenerative agin