Providing An Optimized Model to Detect Driver Genes From Heterogeneous Cancer Samples, Using Restriction in Subspace Learning

Ali Reza EBADI
Islamic Azad University Sanandaj Branch
Ali Soleimani (a.soleimani.uni@iaumalard.ac.ir)
Islamic Azad University Malard Branch https://orcid.org/0000-0001-8878-4046
ABDULBAGHI GHADERZADEH3
Islamic Azad University Sanandaj Branch

Research

Keywords: Subspace learning, driver genes, heterogeneous cancer, Non-negative matrix factorization

Posted Date: November 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-112114/v1

Providing an Optimized Model to Detect Driver Genes from Heterogeneous Cancer Samples, Using Restriction in Subspace Learning

ALI REZA EBADI1, *ALI SOLEIMANI2, ABDULBAGHI GHADERZADEH3

1Department of Electronic and Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
E-mail: ebadi.phdstudent@iausdj.ac.ir,
**2Department of Computer Engineering, College of Technical and Engineering, Malard Branch, Islamic Azad University Tehran, Iran*
**E-mail:* a.soleimani.uni@iaumalard.ac.ir
3Department of Electronic and Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
E-mail: b.ghaderzadeh@iausdj.ac.ir

Abstract

Extracting the drivers from genes with mutation, and segregation of driver and passenger genes are known as the most controversial issues in cancer studies. According to the heterogeneity of cancer, it is not possible to identify indicators under a group of associated drivers, in order to identify a group of patients with diseases related to these subgroups. Therefore, the precise identification of the related driver genes using artificial intelligence techniques is still considered as a challenge for researchers. In this research, a new method has been developed using the subspace learning method, unsupervised learning, and with more constraints. Accordingly, it has been attempted to extract the driver genes with more precision and accurate results. The obtained results show that the proposed method is more efficient for more genes compared to other methods. The p-value of the proposed method was 9.21e-7 better than previous methods for 200 genes. The results show that the p. value of the proposed method is about 2.7 times less than the driver sub method, indicating that the proposed method can identify driver genes in cancerous tumors with greater accuracy and reliability.

Keyword: Subspace learning, driver genes, heterogeneous cancer, Non-negative matrix factorization

1. Introduction

Cancer is one of the deadliest diseases, and according to the estimation of the American Cancer Association in 2019, about 1762450 people has cancer worldwide, of them about 606, 880, individuals have died. Cancer is the second leading cause of death among all diseases [1]. One of the reasons for the abnormal tumor growth, is the rate of DNA mutation in the driver genes, which consequently causes mess in the function of the cancerous cell of a tumor. Due to this reason, having integrated information on this field helps establishing cancer detection and treatment strategies [2]. The large genome changes is one of the causes of cancer, using the second-generation technology of DNA sequencing and analysis, which would significantly contribute to the biological understanding of diagnosis and treatment of cancer. This insight helps us examining each type of change in the somatic genome, and also facilitates the detection of mutant genes in cancer samples [3]. In this study, although we have been able to identify all mutant genes in the tumor, many of these mutant genes have no effect on the tumor development, which are known as passenger genes. Accurate and direct identification of whether this passenger gene has an impact on the development of the tumor or not, still remains a challenge. So, one of the major works in the field of cancer research is identifying the passenger gene from the driver's gene in cases with cancer [4, 5]. One common way to deduce driver genes, is a hypothesis that "the driver's mutated genes are primarily among the large groups of sample mutated tumor genes". Therefore, based on this

hypothesis, many scientific studies have been driven using computational methods of identifying driver genes among the mutated gene groups in this field, [6]. In OncodriveCLUST, specific genes that tend to cluster mutations throughout the protein sequence, were identified, which indicated that these genes have a particular bias toward their dependent gene sequences. Moreover, based on this hypothesis, in this method, a number of genes that had high mutation frequencies were the driving candidate genes, which were later found to have no significant effect on tumor growth [7]. The MutSigCV method solves one of the challenges in identifying driver genes. Previous methods have identified a list of driver genes, but because of mutation heterogeneity, some of them have not been identified properly. Therefore, by the use of this method, this problem has been solved [8]. Because cancer is a heterogeneous disease, there are many different subtypes for one type of cancer, and the driver genes of each subgroup may be different from the other genes. If a mutated gene acts as a driver gene for several specimens in a subgroup, it can be identified as the driver gene of the subgroup and also can be used as a criterion for separating subgroups [9]. Considering the genomic diversity and heterogeneity of subgroups of specific genes in a group, which their driver is small part of samples, so they are rarely changing among all the samples [10]. Other methods are also used to identify a rare mutation except for mutation frequency, such as modifying the amino acid of the flanking sequence [11]. Another method based on optimizing SpeMDP and the maximum matrix weight, is used to identify the driver genes. In this method, the genome data of twelve different types of heterogeneous cancer are used to form a common biological path, and finally the genes in this common path, are used as candidate genes [12]. All the results of the previous methods are encountered the problem that methods are suitable for the idea mode. It is appropriate when all subgroup information are available, which are not mostly available in many cases. The accurate extraction of the driver genes, without providing the subgroup information to find the exact treatment of cancer and personal medicine, still remains as a challenge [10]. To solve the problem of inaccessibility of information, the margin writing of the subgroups, as Driversub method was proposed. Correspondingly, in this method, an unsupervised learning method was used, which needs no information about subgroup [13]. One of the challenges in analyzing the results of this method is that the available data has noise and there is still discarded data, which consequently affects the accuracy of the results. Therefore, in this study, we have tried to overcome this problem by developing this method. In this article, we achieved better accurate for this method via developing the driversub method, and by applying more restrictions on the data. To achieve this goal, robust adaptive graph regularized non-negative matrix factorization method has been used, and by applying less weight to noisy as well as the discarded data, and giving more weight to clean data, we have tried to improve the accuracy of the results. This article used the Cancer Genome Atlas (TCGA) program and Cancer Gene Gensus (CGC).

2- Methods

2-1 Subspace learning

Due to the lack of subgroup information, we have used an unsupervised learning method. To do this, a subspace learning framework has been used [14]. Afterward, we displayed the marginal writing information of a gene, as a vector, so that the mutation data of the gene with high dimensions was converted into a small subspace with smaller dimensions. Gene mutation input data was converted to a binary matrix. The mutation vector of each gene is

$$X = [x_1, x_2, \ldots x_i, \ldots x_p],$$

where p is equal to the total number of genes. The input matrix contains p-genes and n-samples, and each entry of this matrix indicates whether the i^{th} gene has been mutated in the j sample or not [15]. The output matrix was $Z=[\, z_1 , z_2 , z_3 , ...z_i \, ...z_p \,]$, which was compressed space with less dimensions, so that k << n, where k is the dimension of the output matrix Z [16]. Low-dimensional output matrices in vector space can be better suited for the computational analysis. Although the output matrix can well represent the mutation index of the input matrix, the main challenge is that there is no indication to show that the investigated gene is from which one of the subgroups. In fact, there is no general criterion for matching a gene with a subgroup. Due to the fact that the sub-space dimensions can determine the hidden features related to each gene, and the sub-space output dimensions are almost able to determine the indicators related to each subgroup. However, there is no guarantee that the dimensions of the subspecies matrix represent those indicators related to that subgroup, so it can be used to determine whether the special checked genes is relevant to that subgroup [17]. Based on the two hypotheses proposed in the driversub method, the values of the output vector indices can be used as criteria for evaluating the driver's genes. Also, in the second hypothesis, the values of the output vectors can be used indicators for determining whether a gene belongs to a specific subgroup.

However, to increase the guarantee of the first hypothesis in the driversub method, the regularization of L1 norm was used to ensure that the output vectors are sparse [18]. Because the values of the espresso of output vector index are large, the output vectors will more tend to be inclined to match the coordinates of the dimensions subspace. The axes of dimensions of subspace can be used as an indicator to recognize if a gene belongs to a particular subgroups.

The used attribute function in this method is as follows:

$$\min_{w,z_i} \sum_{i=1}^{p} \mathrm{P}x_i - wz_i \mathrm{P}_2^2 + \lambda_z \sum_{i=1}^{p} \mathrm{P}z_i \mathrm{P}_1$$

s.t W≥0 and $Z_i \geq 0$, $\forall_i = 1, ...p$

Here, λ_i is the coefficient of regulator of sparse value. Also, one of the problems of the space Learning method is an overflowing problem [19]. To overcome this problem, Frobenius norm regularization has been used in the driversub method, which changed the attribute function as follows.

$$\min_{W,z_i} \sum_{i=1}^{p} \mathrm{P}x_i - Wz_i \mathrm{P}_2^2 + \lambda_z \sum_{i=1}^{p} \mathrm{P}z_i \mathrm{P}_1 + \lambda_w \mathrm{P}W \mathrm{P}_F^2$$

s.t $W \geq 0$ and $z_i \geq 0$ $\forall_i = 1, ..., p$

What has been forgotten is that in calculating the similarity between the Samples, the Gaussian kernel function can also be used, which is as follows:

$$s_{i,j} = e^{\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$$

$s_{i,j}$ is the similarity between i and j samples . While the Euclidean distance is used to compute the difference between two different samples, real trait space of sample including noise and large amounts of unrelated features, which play no role in similarity, but they can be used for similarity. To enhance the accuracy and precision, a number of irrelevant and disconnected attributes should be eliminated. It was shown that the attributes that have an impact, will have more weight in the distance calculation. Therefore, in order to achieve this goal, it is necessary to learn an M matrix to obtain the exact distance, so we have used M matrix in this article. Herein, we get the distance as follows:

$$\|x_i - x_j\|_M^2 = (x_i - x_j)^T M (x_i - x_j)$$

In this article, we have attempted to reduce noise by combining driversub. The methods as well as applying more restrictions on the obtained samples. Robust adaptive graph regularized NMF (RAGNMF) was also used, which is as follow:

$$\min_{w,z,W,M} Tr[M(x - wz_i)W(x - wz_i)^T] + \lambda Tr(z^T l_s z) + \alpha \|W\|_F^2 + \beta \|M\|_F^2$$

s.t $w \geq 0, z \geq 0, W \geq 0, M \geq 0$

2-2 Optimization

To solve the desired method, a duplicate updating method was used, which is as follows
By keeping W, M constant, the values of w and z were calculated as follows:

$$w_{ir} = w_{ir} \frac{(MxWz)_{ir}}{(MwzWz^t)_{ir}}$$

$$z_{jr} = z_{jr} \frac{(Wx^t Mw)_{jr}}{(Wz^t w^t Mw + \lambda l_s z^t)_{jr}}$$

In order to update the W value, by keeping values of M, w, and z constant, the following relationships were obtained.

$$\min_W = Tr[E^M W] + \alpha \|W\|_F^2$$

s.t $W_i \geq 0$, $\sum_{i=1}^{n} W_i = C_i$

Where E^M is as follows:

$$E^M = (x - wz)\, M\, (x - wz)^T$$

Function (10) can be converted to the following equation:

$$\min \sum_{\substack{i=1 \\ M}}^{n} (W_i + \frac{E_i^M}{2\sigma})^2$$

s.t $W_i \geq 0$, $\sum_{i=1}^{n} W_i = C_i$

Also, by keeping the values of W, z, and w constant, the value of M was calculated as follows:

$$\min_M Tr[E^W M] + \beta \, \mathrm{P} M \, \mathrm{P}_F^2$$

s.t $M_i \geq 0$, $\sum_{i=1}^{m} M_i = C_i$

In this study, to solve this equation, we used the Accelerated Gradient Method, which was earlier used in [21]. The steps of performing this work are shown in the algorithm1.1:

Algorithm 1.1

Start

Input

$X = [x_1, x_2, ...x_i, ...x_p]$ # Mutation vector of each gene

W, M initialize is Identity matrix # is similarity between i,j genes and i,j sample in ordering

Output

$Z = [z_1, z_2, z_3, ...z_i.....z_p]$, W

Begin

1. The calculated Z according to the formula (6)
2. The calculated W according to the formula (5)
3. The Updated W and M according to the formula (7),(10)
4. The Updated Attribute function according to the formula (4)

End

The results indicate that the used method in this article with a high ability to predict and deduce driver genes was shown to be better than previous methods.

2-3 Analysis of results

In this study, we used breast cancer data (Cancer Genome Atlas Network and others, 2012), which included somatic mutations of 507 samples and 12233 genes that can be downloaded from the cBioPortal database [22]. By default, we considered the dimensions of k subspace as 4. In the present study, we have used Python 3.7 to implement this method. Moreover, we used Gsea Msigdb web-based software to analyze the results. Firstly, we calculated the mutation score of each gene from the output vector obtained from the learning subspace, and then arranged it in descending order. Thereafter, we separated the top 500 genes with the highest mutation scores, and then selected them as the candidate for driver genes. Finally, we compared the results with the Benchmarks on Msigdb. The results show that the candidate genes derived from the method in this study, overlap well with the previous investigated gene sets in the curated gene sets, so that, for 200 candidate driver genes, it has less p-value than the previous method.

Fig1.1 the Msigdb database showed that the top 200 driver candidate genes overlap well for q-values (<0.05) with the curated gene sets.

The proposed method in this study has less p-value than the previous methods for the most 200 driver sample genes.

Fig1.2 comparing p-values of different methods showed that the proposed method can predict the genes of the driver candidate more realistically due to the limitations used in the method.

method	p-value
MutSigCV	8.35e-02

OncodriveCLUST	1.23e-02
DriverSub	1.46e-06
proposed method	9.21e-07

Table 1.1 Comparison of p-values between the previous and the proposed methods

In this method, BRCA1, BRCA2, ERBB2, PIK3CA, TP53, and KDM6A genes were introduced as driver candidate genes, which were also common in previous methods. The genes introduced by the proposed method, had a good overlap. In addition, the genes MYO10, ISTN1, EPHA4, SLIT2, WRN, DOP1B PLXNA2, and TCHH were introduced using the proposed method. Due to the elimination of overflow and suspension in the proposed method, the predicted genes were significantly different from the previous methods. Figure 1.3 shows the heat map diagram of seven genes with the highest score subspace (z) with k = 4 in the proposed method, which showed the heterogeneity of the mutation of specific genes in each one of the subgroups.

Fig1.3 Overlap of the number of suggested genes with Curated gene sets (Misgdb)

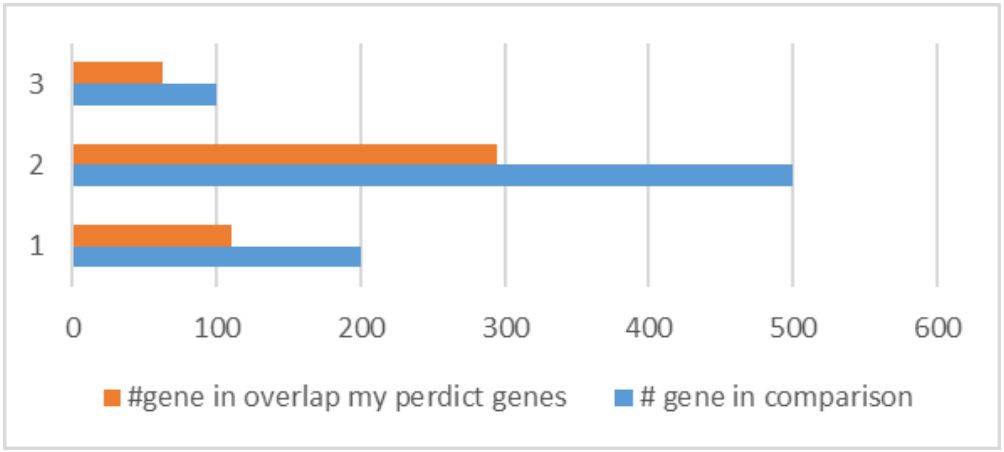
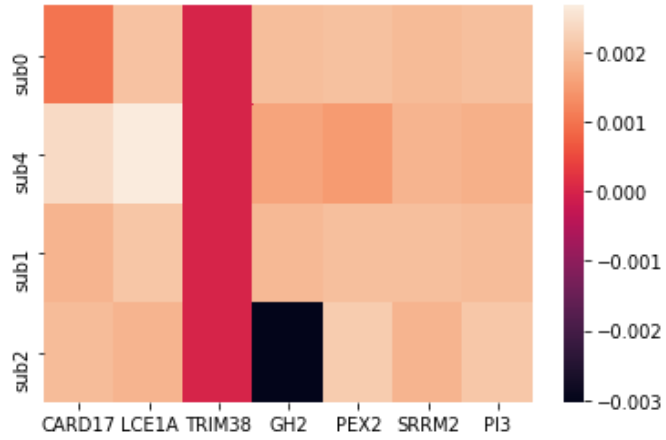


Fig1.4 The rate of mutation of the first seven genes on breast cancer data with the highest scores in each subgroup



In Fig1.4, it is shown that gene mutations in each subgroup were heterogeneous.

3. Discussion and Conclusion

Extraction of subgroups of driver genes is one of the most important cases in personal medicine and heterogeneity in cancer. One of the problems in this regard is the lack of subspace margin information. Due to this reason, in this method, we have used the subspace learning method and the unsupervised learning method. Due to the used method in this paper, more restrictions were applied on the distance between the input vector (X) and the output vector (z) in the subgroups, which was done by applying more weight to the samples that were more effective, and giving less weight to those that had no effect, and then applying it to the Euclidean distance between the two input and output vectors' subspace. Herein, we attempted to extract the subgroups of the driver's genes more accurately. The results show that the proposed method can extract the driver genes more accurately and realistically compared to the previous methods. The point is that, in future work, better results can be obtained for more accurate extraction of the driver genes by combining somatic mutation with transcriptome and epigenome, as well as copy number alternations.

Conflict of Interest

* Disclosure of potential conflicts of interest

All Authors declares that they have no conflict of interest.

* Research involving human participants and/or animals

This article does not contain any studies with human participants or animals performed by any of the authors.

* Informed consent

There was no human involved.

REFERENCES

1. Siegel, R. L., et al. (2019). "Cancer statistics, 2019." CA: a cancer journal for clinicians 69(1): 7-

2.Bailey, M. H., et al. (2018). "Comprehensive characterization of cancer driver genes and mutations." Cell 173(2): 371-385. e318.

3.Meyerson, M., et al. (2010). "Advances in understanding cancer genomes through second-generation sequencing." Nature Reviews Genetics 11(10): 685-696.

4.De, S. and S. Ganesan (2017). "Looking beyond drivers and passengers in cancer genome sequencing data." Annals of Oncology 28(5): 938-945.

5.Vogelstein, B., et al. (2013). "Cancer genome landscapes." science 339(6127): 1546-1558.

6.Tokheim, C. J., et al. (2016). "Evaluating the evaluation of cancer driver genes." Proceedings of the National Academy of Sciences 113(50): 14330-14335.

7.Tamborero, D., et al. (2013). "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes." Bioinformatics 29(18): 2238-2244.

8.Lawrence, M. S., et al. (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes." Nature 499(7457): 214-218.

9.Alizadeh, A. A., et al. (2015). "Toward understanding and exploiting tumor heterogeneity." Nature medicine 21(8): 846.

10.Cyll, K., et al. (2017). "Tumour heterogeneity poses a significant challenge to cancer biomarker research." British journal of cancer 117(3): 367-375.

11.Carter, H., et al. (2009). "Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations." Cancer research 69(16): 6660-6667.

12.Zhang, J. and S. Zhang (2017). "Discovery of cancer common and specific driver gene sets." Nucleic acids research 45(10): e86-e86.

13.Xi, J., et al. (2018). "Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network." BMC bioinformatics 19(1): 214.

14.Zheng, R., et al. (2019). "SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation." Bioinformatics 35(19): 3642-3650.

15.Hofree, M., et al. (2013). "Network-based stratification of tumor mutations." Nature methods 10(11): 1108-1115

16.Wang, K., et al. (2015). "Joint feature selection and subspace learning for cross-modal retrieval." IEEE transactions on pattern analysis and machine intelligence 38(10): 2010-2023.

17.Jolliffe, I. T. and J. Cadima (2016). "Principal component analysis: a review and recent developments." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065): 20150202.

18.Ramirez, C., et al. (2013). "Why ℓ1 is a good approximation to ℓ0: A geometric explanation." Journal of Uncertain Systems,7(3), 203–207.

19.Li, Z., et al. (2015). "Robust structured subspace learning for data representation." IEEE transactions on pattern analysis and machine intelligence 37(10): 2085-2098.

20.He, X., et al. (2019). "Robust adaptive graph regularized non-negative matrix factorization." IEEE Access 7: 83101-83110.

21.Huang, J., et al. (2015). A new simplex sparse learning model to measure data similarity for clustering. Twenty-Fourth International Joint Conference on Artificial Intelligence.

22.Gao, J., et al. (2013). "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal." Science signaling 6(269): pl1-pl1.

Figures

Figure 1

The Msigdb database showed that the top 200 driver candidate genes overlap well for q-values (<0.05) with the curated gene sets.

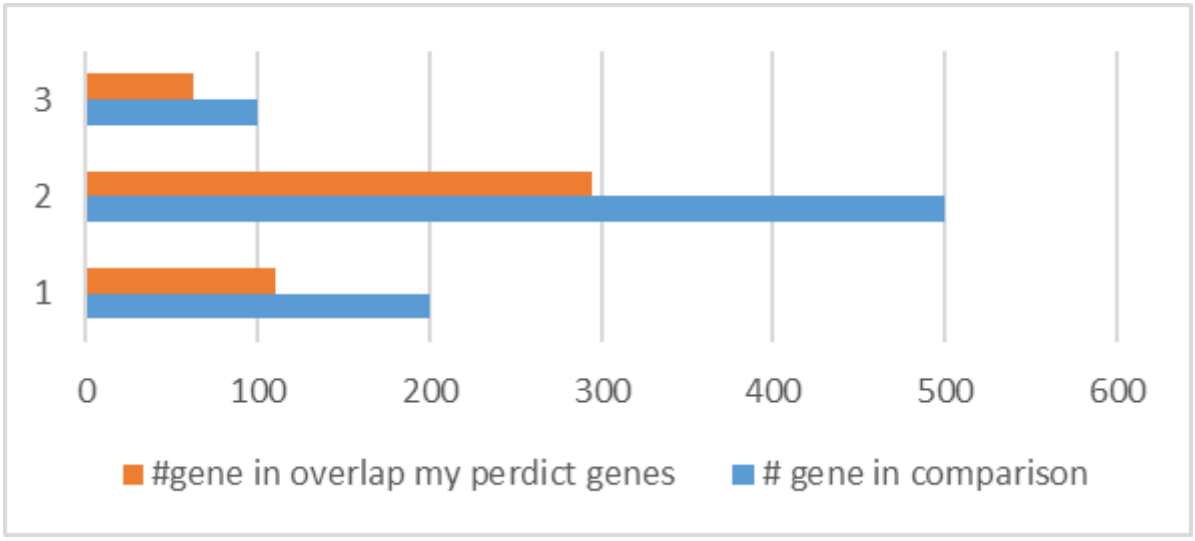


Figure 2

Overlap of the number of suggested genes with Curated gene sets (Misgdb)

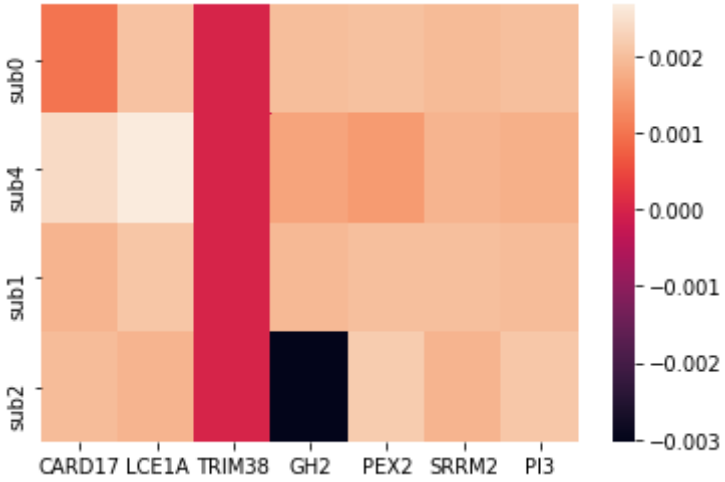


Figure 3

The rate of mutation of the first seven genes on breast cancer data with the highest scores in each subgroup