



Oct 29, 2021

The Simulated Random Assignment of Missense Mutations Throughout a Gene of Interest Can Determine Whether Missense Mutations Found in That Gene in a Population of Tumor Genomes Are Non-Randomly Distributed

Richard L Cullum¹, David J Riese II¹¹Auburn University

1

dx.doi.org/10.17504/protocols.io.bwtwpepe David Riese II
Auburn University

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Human malignancies result from the accumulation of genetic and epigenetic changes to normal cells. In many malignancies, gain-of-function mutations in oncogenes and loss-of-function mutations in tumor suppressor genes drive tumorigenesis and tumor progression. The identification of tumor driver mutations and the genes that host such mutations is critical for the molecular staging and targeted therapy of malignancies.

Since tumor driver mutations cause tumorigenesis or tumor progression, the proliferation of tumor cells selects for these mutations. Thus, in a gene that hosts tumor driver mutations, there will be a non-random distribution of mutations across the gene, as mutations that provide a selective advantage for the tumor cells will predominate over mutations that do not provide a selective advantage for the tumor cells. Consider a particular gene in a population of tumor genomes; the total number of coincident missense mutations in that gene, defined here as two or more missense mutations that affect a particular codon, will be greater than the total number of coincident missense mutations that arise through random assignment of missense mutations across the gene.

Consequently, here we use the R Statistical Computing environment to simulate the random assignment of missense mutations across a user-specified gene. The number of randomly assigned missense mutations is defined by the user and should be equal to the total number of missense mutations observed in the desired gene in the collection of tumor genomes of interest. Based on the simulated random assignment of missense mutations, the R code then determines the total number of simulated coincident and non-coincident mutations. This simulation is repeated a user-defined number of times, and the average number of simulated coincident and non-coincident mutations is calculated from the set of simulations.

The R code then uses a Chi-square test to determine whether the observed number of coincident mutations (in the gene of interest in a collection of tumor genomes) significantly exceeds the average number of simulated coincident mutations. A positive result indicates that the gene hosts a non-random distribution of missense mutations and suggests that the gene hosts tumor driver mutations.

We have used this R code to analyze mutations in the *ERBB4* receptor tyrosine kinase gene that are found in The Cancer Genome Atlas (TCGA) dataset. Our analysis indicates that the number of coincident mutations observed in *ERBB4* in the TCGA dataset is statistically greater than the number of coincident mutations that arise from the simulated random assignment of missense mutations across the *ERBB4* gene. This finding indicates that the distribution of missense mutations in *ERBB4* in the TCGA dataset is non-random.

DOI

dx.doi.org/10.17504/protocols.io.bwtwpepe

Richard L Cullum, David J Riese II 2021. The Simulated Random Assignment of Missense Mutations Throughout a Gene of Interest Can Determine Whether Missense Mutations Found in That Gene in a Population of Tumor Genomes Are Non-Randomly Distributed. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.bwtwpepe>



oncogenes, tumor suppressor genes, identification, tumor drivers, mutations, simulation

document ,

Jul 22, 2021

Oct 29, 2021

51798

DISCLAIMER:

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to [protocols.io](https://dx.doi.org/10.17504/protocols.io.bwtwpepe) is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with [protocols.io](https://dx.doi.org/10.17504/protocols.io.bwtwpepe), can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Human malignancies result from the accumulation of genetic and epigenetic changes to normal cells. In many malignancies, gain-of-function mutations in oncogenes and loss-of-function mutations in tumor suppressor genes drive tumorigenesis and tumor progression. The identification of tumor driver mutations and the genes that host such mutations is critical for the molecular staging and targeted therapy of malignancies.

Since tumor driver mutations cause tumorigenesis or tumor progression, the proliferation of tumor cells selects for these mutations. Thus, in a gene that hosts tumor driver mutations, there will be a non-random distribution of mutations across the gene, as mutations that provide a selective advantage for the tumor cells will predominate over mutations that do not provide a selective advantage for the tumor cells. Consider a particular gene in a population of tumor genomes; the total number of coincident missense

mutations in that gene, defined here as two or more missense mutations that affect a particular codon, will be greater than the total number of coincident missense mutations that arise through random assignment of missense mutations across the gene.

Consequently, here we use the R Statistical Computing environment to simulate the random assignment of missense mutations across a user-specified gene. The number of randomly assigned missense mutations is defined by the user and should be equal to the total number of missense mutations observed in the desired gene in the collection of tumor genomes of interest. Based on the simulated random assignment of missense mutations, the R code then determines the total number of simulated coincident and non-coincident mutations. This simulation is repeated a user-defined number of times, and the average number of simulated coincident and non-coincident mutations is calculated from the set of simulations.

The R code then uses a Chi-square test to determine whether the observed number of coincident mutations (in the gene of interest in a collection of tumor genomes) significantly exceeds the average number of simulated coincident mutations. A positive result indicates that the gene hosts a non-random distribution of missense mutations and suggests that the gene hosts tumor driver mutations.

We have used this R code to analyze mutations in the *ERBB4* receptor tyrosine kinase gene that are found in The Cancer Genome Atlas (TCGA) dataset. Our analysis indicates that the number of coincident mutations observed in *ERBB4* in the TCGA dataset is statistically greater than the number of coincident mutations that arise from the simulated random assignment of missense mutations across the *ERBB4* gene. This finding indicates that the distribution of missense mutations in *ERBB4* in the TCGA dataset is non-random.

Dataset Retrieval

Our approach uses mutational data from a set of human tumor genomes. As an example we have used human tumor genome data from the National Cancer Institute Genomic Data Commons ([NCI-GDC](#)). Specifically, we have used *ERBB4* mutation data from the entire The Cancer Genome Atlas (TCGA) project (12,922 cases). This dataset can be found [here](#) and contains 720 different *ERBB4* mutations. The dataset was downloaded in July, 2021, and was incorporated into a single sheet of the attached Microsoft Excel workbook [TCGA ERBB4 All Mutations.V1.xlsx](#).

Dataset Transformations

This sheet was edited to remove all mutations except for the 414 unique *ERBB4* missense mutations (total of 480 *ERBB4* mutations). Moreover, all columns were removed except for the column that describes each mutation (the "Consequences" column) and the column that describes the incidence of each mutant allele (the "# Affected Cases Across the GDC"). For example, the edited spreadsheet indicates that the *ERBB4* missense mutation A4E occurs once across the 12,922 cases of the TCGA dataset ("GDC"). The resulting Microsoft Excel workbook is attached

[TCGA ERBB4 Missense.V1.xlsx](#).

These data were transformed as follows, resulting in the attached Microsoft Excel workbook

[TCGA ERBB4 Missense Transformed.V1.xlsx](#). The affected codon was extracted from the

"Consequences" column (Column A) in a three-step process. (1) Three columns were inserted to the

right of the "Consequences" column. Then the first 16 characters stored in each cell of the "Consequences" column were removed using the Excel command `+RIGHT(A2,LEN(A2)-16)` and the result was stored in the "Affected Codon 1" column (Column B). (2) Next, the last character stored in each cell of the "Affected Codon 1" column was removed using the Excel command `+LEFT(B2,LEN(B2)-1)` and the result was stored in the "Affected Codon 2" column (Column C). (3) Because the values in the cells of the "Affected Codon 2" column are stored in text format, these values were copied into the "Affected Codon 3" column (Column D) using the EDIT/PASTE SPECIAL/VALUES function. The entire "Affected Codon 3" column was selected and the values in these cells were transformed into numerical format using the TEXT TO COLUMNS function as described [here](#).

The frequency of each mutation in the TCGA dataset was extracted from the "# Affected Cases Across the GDC" column (Column E) in a two-step process. (1) The last 8 characters stored in each cell of the "# Affected Cases Across the GDC" column were removed using the Excel command `+LEFT(E2,LEN(E2)-8)` and the result was stored in the "Allele Frequency 1" column (Column F). (3) Because the values in the cells of the "Allele Frequency 1" column are stored in text format, these values were copied into the "Allele Frequency 2" column (Column G) using the EDIT/PASTE SPECIAL/VALUES function. The entire "Allele Frequency 2" column was selected and the values in these cells were transformed into numerical format using the TEXT TO COLUMNS function as described [here](#). The total number of *ERBB4* missense mutations was determined by summing the values found in the "Allele Frequency 2" column using the Excel command `+SUM(G2:G415)`. A total of 480 *ERBB4* missense mutations are found in the TCGA dataset.

The mutations found in the spreadsheet were manually assigned to the coincident and non-coincident mutations categories and were assigned to the "Coincident Mutations" column (Column H) and the "Non-Coincident Mutations" column (Column J) as appropriate. One way that a particular mutation can be assigned to the "Coincident Mutations" category is if it occurs more than once in the 12,922 cases of the TCGA. For example, the A16V mutation occurs twice in the TCGA dataset, so the spreadsheet indicates that two coincident mutations reside at codon 16 of *ERBB4*. The other way that a particular mutation can be assigned to the "Coincident Mutations" category is if it occurs at a codon that is also the site of a different mutation. For example, the A4E and A4T mutations occur once each in the TCGA dataset. Thus, the spreadsheet indicates that two coincident mutations reside at codon 16 of *ERBB4*, one for the A4E mutation and one for the A4T mutation.

The total number of coincident mutations was calculated by summing the values in the "Coincident Mutations" column using the Excel command `+SUM(H2:H415)`. Parenthetically, two coincident mutations occur at *ERBB4* codons 4, 16, 32, 72, 78, 114, 130, 190, 201, 211, 217, 225, 280, 286, 293, 401, 422, 488, 516, 521, 522, 537, 586, 682, 686, 697, 731, 735, 741, 774, 785, 800, 813, 822, 836, 838, 840, 868, 910, 947, 975, 991, 1003, 1020, 1023, 1053, 1063, 1080, 1090, 1100, 1166, 1180, 1242, and 1303 (54 codons accounting for a total of 108 coincident mutations); three coincident mutations occur at *ERBB4* codons 47, 103, 196, 303, 452, 544, 572, 573, 662, 671, 713, 751, 759, 906, 1002, 1043, 1102, 1187, 1223, and 1304 (20 codons accounting for a total of 60 coincident mutations); four coincident mutations occur at *ERBB4* codons 50, 798, and 922, (3 codons accounting for a total of 12 coincident mutations); seven coincident mutations occur at *ERBB4* codon 106 (1 codons accounting for a total of 7 coincident mutations); eight coincident mutations occur at *ERBB4* codon 711 (1 codons accounting for a total of 8 coincident mutations). Thus, the TCGA dataset of 12,922 cases contains a total of 195 coincident *ERBB4* mutations distributed across 79 different *ERBB4* codons.


The total number of non-coincident mutations (285 non-coincident mutations) was calculated by summing the values in the "Non-Coincident Mutations" column using the Excel command +SUM(J2:J415). Codons that are affected by coincident mutations were manually recorded in the "Codons With Coincident Mutations" column (Column I). The total number of codons (79 codons) that are affected by coincident mutations was calculated by summing the values in the "Codons With Coincident Mutations" column using the Excel command +SUM(I2:I415). The manual assignment of mutations to the "Coincident Mutations" and "Non-Coincident Mutations" columns was checked by adding the value of each cell in the "Coincident Mutations" column to the value of the adjacent cell in the "Non-Coincident Mutations" column. This sum was recorded in the "Allele Frequency Audit 1" column. This value was subtracted from the corresponding value in the "Allele Frequency 2" column, with the difference (which should be zero) recorded in the "Allele Frequency Audit 2" column.

The total number of missense mutations (480 - from the "Allele Frequency 2" column) and the total number of coincident mutations (285 - from the "Coincident Mutations" column) are utilized in the simulated assignment of random missense mutations in *ERBB4*.

Software Development and Execution

The software utilized here was developed using the [R Software Environment For Statistical Computing](#). Specifically, the R code was developed and executed using [R for macOS version 3.6.3nn](#) and [R Studio Desktop](#) for macOS version 1.4.1106. The R code was executed using a 2019 16-inch Apple MacBook Pro laptop computer equipped with a single 8-core, 2.3 GHz Intel Core i9 CPU, 16 GB RAM and macOS Big Sur 11.4.

Overview of Software

The R code is attached  [ERBB4-TCGA.Rmd](#). The following is a summary of the salient features of the software.

This program requests that the user specify the length (in amino acids) of the gene of interest. It then asks the user to specify the total number of missense mutations found in the tumor genome dataset of interest. The user also enters the total number of non-coincident missense mutations (codons that are mutated only a single time in the tumor genome dataset). These data are stored as the following variables:

- *NumOfResiduesInGeneOfInterest* (the length of the gene of interest)
- *NumOfMutations* (total number of missense mutations observed in the gene in the tumor genome dataset)
- *ObsFoundOnce* (number of codons that are altered by only a single missense mutations in the tumor genome dataset)
- *TotalObsCoincidence* (total number of missense mutations that affect codons that are altered by two or more missense mutations - this number is calculated by subtracting *ObsFoundOnce* from *NumOfMutations*)

Next, the code uses a random number generator to simulate the random assignment of missense mutations across the gene of interest. The number of randomly assigned missense mutations in each simulation trial is equal to *NumOfMutations*.

The code then scans the randomly assigned missense mutations in order to record the number of non-

coincident mutations (number of codons that are affected by only a single mutation). This value is stored in the variable *Occurs1x*.

The process of repeating the simulated random assignment of missense mutations and recording the number of non-coincident mutations is repeated according to a user-defined number stored in the variable *NumOfSimulationTrials*.

The variable *Occurs1x* is incremented during each simulation trial. Thus, following the completion of the user-defined number of stimulation trials, *Occurs1x* is divided by *NumOfSimulationTrials* to get *Avg1x*, the average number of non-coincident mutations over the set of trials. *Avg1x* is subtracted from *NumOfMutations* to get *CoincidentMutations*, the average number of coincident mutations over the set of trials.

The number of observed non-coincident and coincident missense mutations in the set of tumor genomes is compared to the number of non-coincident and coincident missense mutations averaged from the simulated, randomly generated set of missense mutations. The Pearson's Chi-squared test with Yates continuity correction is used to determine if the number of coincident mutations observed in the set of tumor genomes significantly exceeds the number of coincident mutations averaged from the randomly generated set of missense mutations.

The software outputs the number of observed non-coincident and coincident missense mutations in the set of tumor genomes, the number of non-coincident and coincident missense mutations averaged from the simulated, randomly generated set of missense mutations, and the results of the Pearson's Chi-squared test with Yates continuity correction.

Results of Sample Data Analysis

The results of the sample data analysis is attached [TCGA ERBB4 Results.xlsx](#). The number of non-coincident and coincident missense mutations observed in *ERBB4* in the TCGA dataset is 285 and 195, respectively, yielding a total of 480 missense mutations. The expected number of non-coincident and coincident missense mutations predicted from 10,000 simulation trials is 332.8 and 147.2, respectively. We compared the observed results against the expected results using a 2x2 Pearson's Chi-squared test with Yates continuity correction. This yielded a Chi-squared statistic of 9.951 and a p-value of 0.00161 (df=1). Thus, we reject the null hypothesis that the observed number of coincident mutations in *ERBB4* in the TCGA dataset is equal to the number of coincident mutations expected from the simulated random distribution of missense mutations across *ERBB4*.

Because the number of observed coincident *ERBB4* mutations is significantly greater than the number of predicted coincident *ERBB4* mutations (resulting from the random assignment of *ERBB4* mutations), we conclude that the occurrence of *ERBB4* mutations in the TCGA is non-random. The observed non-random occurrence of *ERBB4* mutations in tumor samples may reflect selection for *ERBB4* mutations. Thus, these data predict that some *ERBB4* mutations may function as tumor drivers. Functional analysis of individual *ERBB4* mutations is warranted to test that prediction.