Untapped Potential of Big Data in Precision Medicine

Sweta Yadav

Harrisburg University of Science and Technology

Abstract

Precision medicine provides tailored medical treatment using individual's genetic information and the manual process of genome annotation takes lot of time, money and effort. The study analyzes the relationship between time taken for the process of genome annotation by using data analytics methods/tools and time taken for genome annotation by conventional manual process of annotation. It also focuses on the accuracy of results using various analytical methods and in turn the analysis of the effort and labor required for the process using artificial intelligence tools and machine learning algorithms.

Keywords: precision medicine, big data, genome annotation, machine learning

Untapped Potential of Big Data in Precision Medicine

Precision medicine, also known as personalized medicine is a medical approach where patients are prescribed medicines and provided treatment that is most suitable for them depending on their genetic, lifestyle and environmental conditions (Mirnezami, 2012). This approach uses an individual's genomic information to provide targeted treatment, customized to the individual (Kingsmore, 2015). Practitioners of precision medicine use genomic sequencing tools to compute a patient's complete genome to detect the specific genetic alterations that led to the rise and propagation of disease (Sugeir, 2018).

Genomic annotation is a process which can be defined as obtaining meaningful and useful biological information by the interpretation of raw sequence data. Conventionally, this work was completely done manually as human genome annotators used to review the evidence for each gene so as to make decisions on their intron-exon structures. Inspite of the fact, that the results obtained from this kind of manual work produces annotation of very high-quality annotation, there is a need to explore and develop methods for automated genome annotation which does not require so much of effort like the conventional labor-intensive tools and time-consuming methods and eventually would reduce the cost input (Mungall, 2002).

Big Data is the massive amount of structured and unstructured data that are increasingly created by high-performance applications used by various domains like molecular biology, genetics, biochemistry, astronomy, physics, business etc. Different techniques are utilized to analyze big data like statistics, machine learning, data mining, signal processing, visualization techniques etc (Rodriguez-Mazahua, 2016). A machine learning algorithm can be explained as a computational method which is dependent upon statistics, implemented in software and is able to find hidden non-obvious patterns in a dataset, and in addition is able to create consistent and

authentic statistical predictions about similar new data (Chicco, 2017). In an article by Kevin Yip, it is explained that the ability of machine learning to automatically detect patterns in data is essentially important when there is limited or inaccurate expert knowledge and also when the amount of data available is too huge to be handled by the manual process, or when exceptions are encountered for normal cases (Yip, 2013). This is clearly the case for genomics annotation as the size of the data is huge and predictions are required to be made by automatically detecting patterns in data.

The aim of this study is to evaluate the relation between time taken for the process of genome annotation by using data analytics methods/tools and time taken for genome annotation by conventional manual process of annotation. Based on established studies where use of big data analytical methods in various healthcare areas like EHRs, treatment time to patients, real time monitoring, etc has significantly impacted the time taken due to fast data processing, it has been hypothesized that machine learning or data analytics tools will perform genome annotation, required for precision medicine at a faster speed than conventional manual methods. Also, since manual process of genome annotation incorporated human errors, it has also been hypothesized that use of machine learning and data analytics methods for genome annotation is directly proportional to highly accurate results. Furthermore, it has been hypothesized that the effort and labor required for the automated process of genome annotation will be significantly reduced as compared to the manual process of genome annotation.

Literature Review

Precision medicine can be described as novel practice entitled by molecular diagnostics that is extremely different from the traditional approach where all the patients with the similar

condition or disease were treated with the same drug and same dosage (Leff, 2015). In reality, this practice of individual medicine is actually not new and was used in ancient Greece, 2500 years ago by Hippocrates, widely known as “Father of Western Medicine”. In an article by Sykiotis, it has been mentioned that Hippocrates believed in the individuality of disease and the need to give different drugs to different patients. He studied a different aspect of person's constitution, physique, age and also the time of the year to decide how to prescribe the medicine (Sykiotis, 2005). Now it is known that the differences among people are due to variations in their genetic constitution. Also, it is well known that all the patients do not respond in the same way to the same drug. Adverse drug reaction (ADR) is the fourth major cause of death in the United States and it has been reported that “2.74 million adverse drug reactions and 128,000 deaths are the results of prescription drugs. Almost two adverse drug reactions cost $136 billion in a year which is more than the complete cost for diabetes and cardiovascular care and also lead to one in five losses of life or injury to hospitalized patients in a year” (Chandra, 2017).

Precision medicine is based on collecting large amounts of data and this data is produced from high-throughput screening bio-technologies after studying various aspects of complex biomedical systems. Infact, it combines various research areas, the collectin of meaningful data is still a big obstruction because of huge sizes, heterogeneity and different time and space scales of the data. Therefore, the future of precision medicine still bears a big challenge on how to translate large-scale data into new knowledge to be applied separately to each one of us to improve our health (Malod-Dognin, 2018).

Big Data can be explained as data sets which are very large in volume or complexity that standard data processing methods are not sufficient enough to handle them (Baro, 2015). The most well- known definition of Big Data is the 5Vs, which are Volume, Variety, Velocity,

Verification/Veracity, and Value (Giri, 2014). This definition of Big Data can be exposed to technological advancements in the future. Big Data analytics covers collection, manipulation, and analyses of enormous, multiple data sets that consists of a range of data types comprising genomic data and EHRs to show hidden patterns, confusing correlations, and various other intuitions on a Big Data infrastructure. Because of its usefulness and impact, big data analytics is greatly used in various research fields (Chute, 2013).

The electronic patient health record (EHR) is one of the sources of big data comprising information related to medical conditions, treatments, socio-demographics, and genetics, however, the human ability is limited to handle this data without potential decision support. In the field of healthcare, the aim is to generate a dynamic learning infrastructure with real-time knowledge production and to create a system which is preventative, preventive and predictive. As a step to attain this aim, the clinicians need help from computer models to arrange the data, identify patterns, derive results, and set limits for actions. Big data analytics has already made remarkable improvement for new knowledge generation, streamlined public health surveillance and advance clinical care. For example, the EHR has been effectively used for post-marketing surveillance of drugs or medicine and to make advancement in pharmacovigilance (Daniel, 2017). The growing use of EHR systems throughout the world makes it acceptable to capture enormous amounts of clinical data. The following step would be to convert this captured generous size healthcare data into understandable information or knowledge. For this EHR data, natural language processing and data mining are primary components of analytics (Ross, 2014).

To share EHRs among different healthcare providers, various factors are to be considered as First, functional interoperability, which means that data for example medical records can be exchanged without any limitations or problems from one EHR system to another EHR system.

Second, structural interoperability, which allows the data to be exchanged among all systems. Third, semantic interoperability, which means data can be exchanged across multiple systems and this exchanged data can be easily used for different analysis and Fourth is interpretation, which permits clinicians to correctly decipher the health records for example symptoms which contain the same meaning. At present, various global EHR providers, like MEDITECH, Epic, Allscripts, and Cerner have collaborated to achieve an interoperability initiative. All these listed vendors have approved to conquer the drawbacks of the medical record, positive patient engagement, interchangeability, information sharing. This pact is a crucial step towards creating an integrated healthcare system (Lin, 2014).

Disease diagnosis and identification of the illness is in the frontline of big data and machine learning research. One example is the prediction of urinary tract infections (UTI) in the emergency department, which otherwise has reported high diagnostic error rates. In an experiment, six machine learning algorithms were developed for UTI prediction and parameters considered were patient's demographic information, vitals, medication, laboratory results, past medical history and other factors. Prediction results were compared between models and documentation of UTI diagnosis and antibiotic administration. The best performing algorithm accurately diagnosed positive urine culture results and surpassed previously developed models (Taylor, 2018).

Medical imaging data is a kind of big data in the field of medical research. Imaging genomics is a fast- growing branch originated from recent progress in multimodal imaging data and high-throughput omics data. The striking complexity of these datasets poses vital computational challenges. Robert studied methods to subdue the challenges related to integrating, transcriptomic, genomic and neuro- proteomic data (Robert, 2014). To encourage data mining and meta-analyses, there has been arising interest in the integration of neuroimaging data into

frameworks. A central interest in cognitive neuroscience has been making a continuous effort in the past century to understand the human brain. Lately, to understand the complex structure of cerebral cortex, a team of neuroscience researchers used machine learning methods to map the human brain (Glasser, 2016).

Human brain mapping is another remarkable step towards the field of precision medicine. For instance, an essential improvement to develop operational management and therapeutics of neurological and psychiatric disease enables researchers and scientists in collecting and exploring data from almost 100 billion neurons in the brain at a relatively greater dimension and at much fast speed. Because the human brain is responsible to control a number of chronological and spatial scales, this data can be utilized to better understand the working of the brain by gathering all required and related information. Hence, development of high-performance computing tools derived from Big Data framework is essential and crucial to neuroscience for progress in the field of healthcare (Dinoy, 2016).

Big data has been used to reduce waste and inefficiency in the following three healthcare areas: First is Research & development, in which predictive modeling has been used to decrease attrition and generate a leaner, faster and more targeted R & D pipeline for drugs and devices, algorithms and statistical tools have been used to enhance clinical trial design and patient recruitment to better understand and relate treatments to individual patients, as a result it reduces trial failures and accelerates new treatments into the market, and for analysis of patient records and clinical trials to detect follow-on indications and find out adverse effects even before the product is launched in the market. Second is Clinical operations in which comparative effectiveness research is used to understand more clinically relevant and cost-effective ways to diagnose and treat patients. Third is Public health in which disease patterns are analyzed and disease outbreaks

are tracked and information is transmitted to improve public health surveillance and accelerate the patient response, rapid advancement of more precisely targeted vaccines, like selecting the annual influenza strains and converting large amounts of data into actionable information which can be used to understand needs, provide services, and predict and prevent crises, specifically for the welfare of populations (Janet, 2017).

Another important application of big data is in the prediction of epidemic outbreak around the world. Machine learning and artificial intelligence technologies are being used for monitoring and prediction of various epidemic outbreaks by using the data which is collected from satellites, real-time updates on social media and different other sources. In an article, prediction model has been used to predict malaria outbreak in Maharashtra region in India using support vector machines and artificial neural networks by considering parameters like temperature, average monthly rainfall, humidity, total number of positive cases and other data points. The model can predict the outbreak 15- 20 days in advance and there is scope to scale up the model at country level. Early prediction helps the policy makers and health providers to control the mortality and reduce risk of transmission of disease (Sharma, 2015).

Next-generation sequencing (NGS) technologies, like whole-exome sequencing (WES), whole-genome sequencing (WGS) and targeted sequencing are increasingly being more used in medical practice and biomedical study to determine drug and disease-related genetic variants to develop precision medicine. Precision medicine enables clinicians and scientists to determine more accurately as which preventive and therapeutic procedure to a particular disease or condition can work efficiently in different groups of patients depending upon their lifestyle, environmental factors, and genetic make-up (Vassy, 2015). Precision medicine has the enormous capability in the treatment of many diseases (Gray, 2017). It is targeted, individualized care that is modified to suit

each patient based on his or her typical genetic structure and past medical history. In contrast to traditional medicine, where one-size-fits-all, practitioners of precision medicine use genomic sequencing tools to compute a patient's complete genome to detect the specific genetic alterations that led to the rise and propagation of disease (Sugeir, 2018).

The process of elucidating raw sequence data into useful biological information is known as annotation. Genomes are described by annotations and transform raw genomic sequences into biological information by combining computational analyses, adding biological data and biological expertise. Generally, studies carried out at small scale for isolated genes in a laboratory by individual researcher require integration of experimental and computational methods which allow the description of features in a detailed manner. The information which is obtained is narrow but deep. On the other hand, now the best results obtained from the annotation of large eukaryotic genomes give a complete and overall view and summary of the whole genome, but they are shallow and does not describe individual genes completely. The information obtained is broad but superficial. Currently, the annotation of large-scale sequences is not upto standards and is compromised but, preferably, the goal is that the description of the genome should have both breadth and depth (Lewis, 2000)**.** Gene annotation refers to the process of assigning functions to different parts of a newly sequenced genome. Nowadays as advanced and continuously automated technologies are quickly increasing the total number of sequenced genomes, the process of annotation of genes is not even close to advancement or is not even automated and thus careful analysis is required by trained human annotators. Important decisions are made during the process of annotation regarding the specific DNA sequences present in a genome in which DNA sequence contains the biological information. This process includes identification of features in the DNA sequences and most importantly determining as which of these features in the sequences are genes.

Human annotators integrate these generally opposing predictions with sequence alignment and data about gene expression like RNA sequencing data to create a gene model that is best supported (Saville, 2012).

The final goal of all the annotation attempts is to procure a synthesis of alignment-based proof with gene predictions to secure a complete set of gene annotations. Traditionally, this work was completely done manually as human genome annotators used to review the evidence for each gene so as to make decisions on their intron-exon structures. However, the result of all this manual work generates high-quality annotation, it is such a labor-intensive and time taking work that, due to budgetary reasons, small size genome projects are increasingly being forced to depend upon automated annotations (Mungall, 2002). Genome annotation has now advanced in identifying just the protein-coding genes to an even greater emphasis on the annotation of ncRNA genes, pseudogenes, regulatory regions and transposons. In fact, the quality control and management of annotations are also on the edge of competing `for the betterment and almost perfect. As far as, the annotation tools and sequencing technologies are developing, there is always a need for periodic updates to every genome's annotations. Hence, the new developments and advancements in the genome annotation projects need to be updated continuously. Just like parenthood, the responsibilities towards annotations never end with the birth. If annotations are done incorrectly and are incomplete, they poison each and every experiment uses them for any further work. In today’s genomics driven world, providing accurate and up-to-date annotations is simply a must (Yandell, 2012)**.**

Annotation of the human genome and other species should be accurate to support current drug discovery efforts. To validate possible drug targets confronting genomic sequence demands accurate annotation in the first chance to make this procedure valuable and beneficial.

Traditionally, accurate genome annotation has been accomplished through manual curation, by using the experience and knowledge of expert individuals who know the process to annotate sequence by hand. While manual curation can achieve the high level of accuracy, it cannot stride with the throughput of the multi-species sequence which is now very much in demand. Therefore, there is a huge requirement by bioinformatics solutions to develop automatic annotation techniques which can support and supplement the manual curation process of annotation (Rust, 2002).

The branch of machine learning is involved with the development and implementation of computer algorithms which progress and update with experience. The methods of machine learning have been used to a wide range of areas within genetics and genomics. The field of machine learning, in fact, is most useful for the understanding of large genomic data sets and has extensively been used for annotation of different type of genomic sequence elements. For instance, to 'learn' how to recognize the locations of transcription start sites in a genome sequence, machine learning methods can be used. Similarly, algorithms can be trained to detect positioned nucleosomes, splice sites, enhancers, and promoters. (Maxwell, 2015).

Since all the traditional and conventional methods of genome annotation are mainly manual procedures though they produce highly accurate results, there is a great need to develop or explore already existing automatic tools for the process of gene annotation to handle such huge genomic datasets which would significantly impact the field of medicine and healthcare. Machine Learning is a field which is advancing very fast and is based on learning with experience and has already been successfully applied to a variety of areas including healthcare. Hence, machine learning holds huge potential to overcome the limitations of handling large genomic datasets with billions of data points and produce genome annotation which is the most critical step at a much faster speed, with

great accuracy which could also save lot of time, money and effort invested in the conventional manual methods of genome annotation.

The independent variables are the techniques of genome annotation i.e. machine learning algorithms and conventional methods; and dependent variable is the time taken for the process of genome annotation. The hypothesis is that machine learning algorithm for genome annotation in human beings will take less time than conventional manual methods. Also, use of machine learning and data analytics methods for genome annotation in human beings is directly proportional to highly accurate results.

Methods

Participants

The analysis will be done on the datasets containing information about genetic annotation for precision medicine which contains genetic mutation information from different individuals of both sex as male and female, any age and any race or ethnicity. The individual information is represented with the specific ID, used to combine the datasets.

Procedure and Data

Data selected to be used in the study was obtained from kaggle website. A brief description of the dataset is: It is categorized into two different files, training and test. One (training/test_variants) provides the information about the genetic mutations, whereas the other (training/test_text) provides the clinical evidence (text) that our human experts used to classify the genetic mutations. The training data has 3321 records and the test data has 986 records. Both are linked via the ID field as explained below:

1. training_variants – this is a comma separated file consisting of the description of the genetic mutations that is used for training. Fields in the dataset are:
    - ID - the id of the row used to link the mutation to the clinical evidence.
    - Gene - the gene where this genetic mutation is located
    - Variation - the aminoacid change for these mutations.
    - Class - (1-9) the class this genetic mutation has been classified on.
2. training_text – this is a double pipe (||) delimited file which contains the clinical evidence (text) used to classify genetic mutations. Fields in this dataset are:
    - ID - the id of the row used to link the clinical evidence to the genetic mutation.
    - Text - the clinical evidence used to classify the genetic mutation.
3. test_variants – this is a comma separated file containing the description of the genetic mutations used for training. Fields in this dataset are:
    - ID - the id of the row used to link the mutation to the clinical evidence.
    - Gene - the gene where this genetic mutation is located.
    - Variation - the amino-acid change for these mutations.
4. test_text – this is a double pipe (||) delimited file which contains the clinical evidence (text) used to classify genetic mutations. Fields are:
    - ID - the id of the row used to link the clinical evidence to the genetic mutation.
    - Text - the clinical evidence used to classify the genetic mutation.

Measures

To complete the analysis machine learning algorithm were developed to automatically classify genetic mutations based on clinical evidence (text) across different classes. The models

were trained by using the training sets and then the test dataset was used to make predictions for the classification of genes in different class. After that the accuracy percentage was calculated to check the performance of the developed model. The machine learning models used were as follows:

1. Xgboost (e**X**treme **G**radient **Boost**ing) Model
2. Naive Bayes Model
3. Support Vector Machine

Analysis

The main objective of the analysis is to develop the model that can predict the class of each genetic mutation using the clinical evidence text data. A genetic mutation can be classified into one of the nine different classes.

The first step is to install and load the required libraries required for data analysis. Then the data exploration and cleaning is done to check for any null values, to check the number of observations and variables like the length difference in gene and variation variable is not low which implies that it might not affect the model learning as number of observations, nrow() for training and test is 3321 and 986 which is not that long. Few findings after checking the data were:

- The variation in the type of gene has frequency less than 2 and more than half have frequency less than 5.
- The difference in frequency is because of the difference in number of rows. As it is already checked before from the comparison of number of rows that the difference is small, it is anticipated that the intersection of test and training variation is small and gene interaction is comparatively large. Also, in case of length comparison the unique sum of training and

test data is almost same to the nrow sum, which indicates that the learning might not be significant.

- However, it is anticipated that the gene variables might be significant for learning as considerable interaction is observed in the first 10 gene observations and there is not much difference in the configuration of actual factor set. In this case, the nrow sum of training and test data is different from unique sum. Hence, the variable, variation doesn't seem relevant for learning and ID has unique variation.

Data Cleaning and Data Visualization

Data cleaning and visualization was done in the following steps:

Categorical variables. - frequency of two variables, Gene and Variation were visualized.

Frequency Distribution of Gene and Variation Variable. - As can be seen in Figure 1, the distribution of gene variable compared to variation was not bad and we can use it further analysis. In case of variation variable, most of the classes have frequency atleast once or twice except some classes. In case of class label, the frequency of the classes 3, 8 and 9 have low frequency, so these can be analyzed in more detail while inspecting word and bigram whereas the frequency distribution for other classes seems fair and not thickened to one particular class.

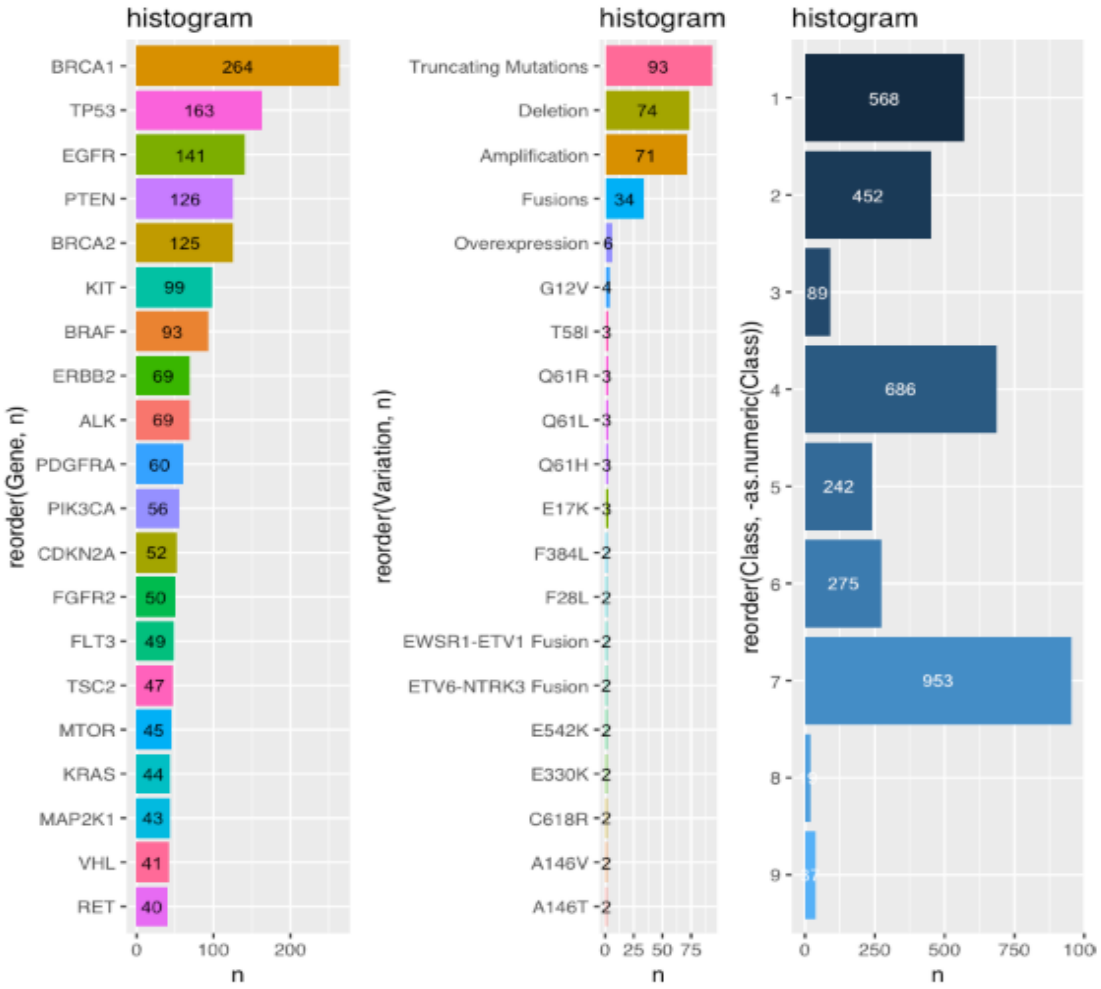

Figure 1. Histogram for frequency distribution of gene, variation and class variables

Frequency Distribution of Gene and Variation Variable by Class. - The classes 3, 5 and 9 were discarded as no row is there and top 10 classes for variation variable were included. As can be seen in Figure 2, there is difference in gene distribution and class distribution when we compare it to Figure 2, there seems lack of observations as in class 1 has 568 rows of training data whereas total sum of top 10 variation and gene is less than 300 each which implies that if top 10 rows doesn't contain gene or variation of some rows, then for prediction of other 300 rows, these variables might not be much useful.

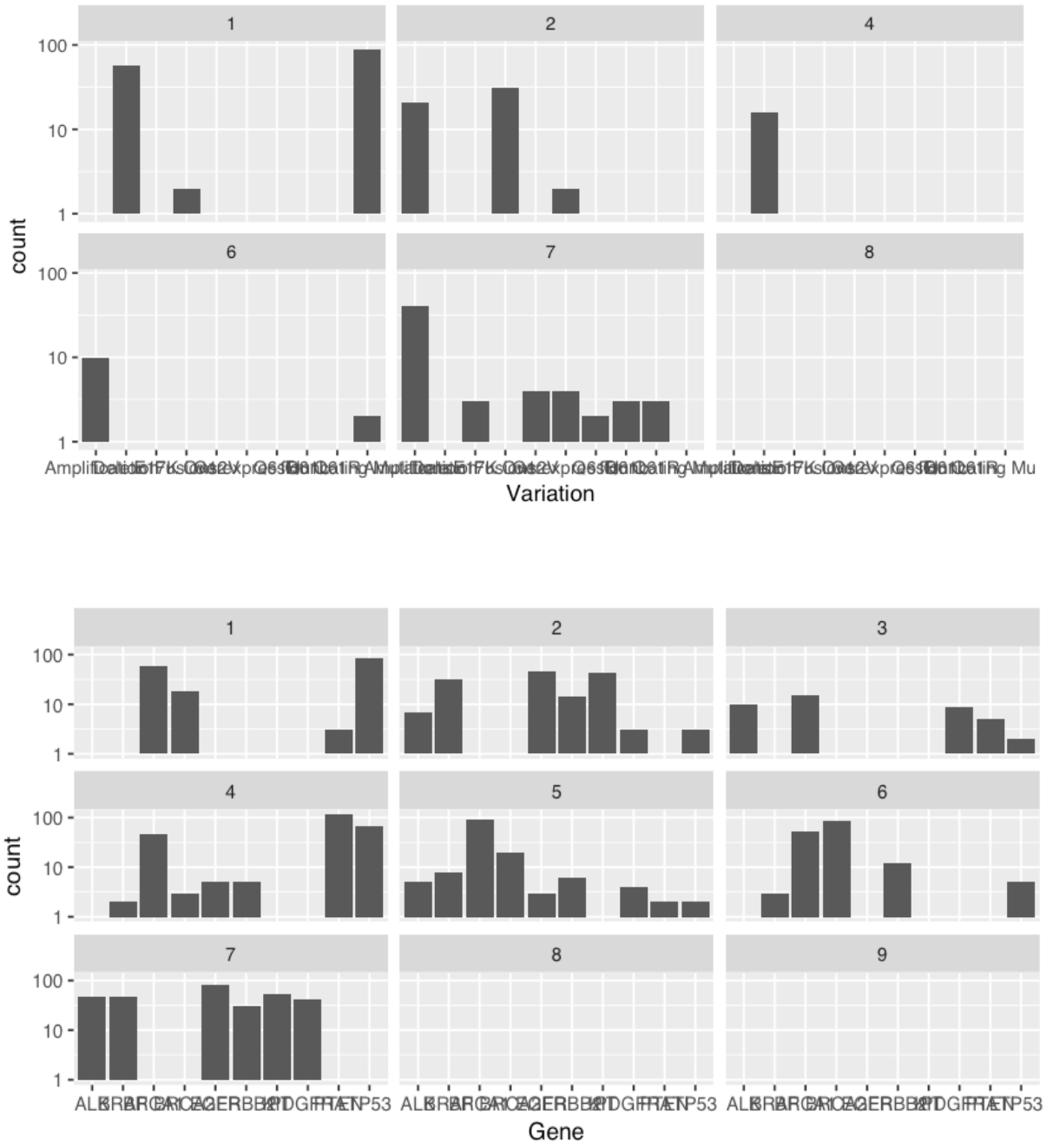

Figure 2. Variation and gene variable frequency distribution by class. After sorting in descending order n=10 was used rather than using top 10 rows as they all have the similar frequency.

Comparing test and training data for frequency distribution of gene and variation variable. - As can be seen from Figure 3, the training and test data doesn't show same

patterns for frequency distribution which implies that gene and variation variable might be very informative for prediction model for the test data.

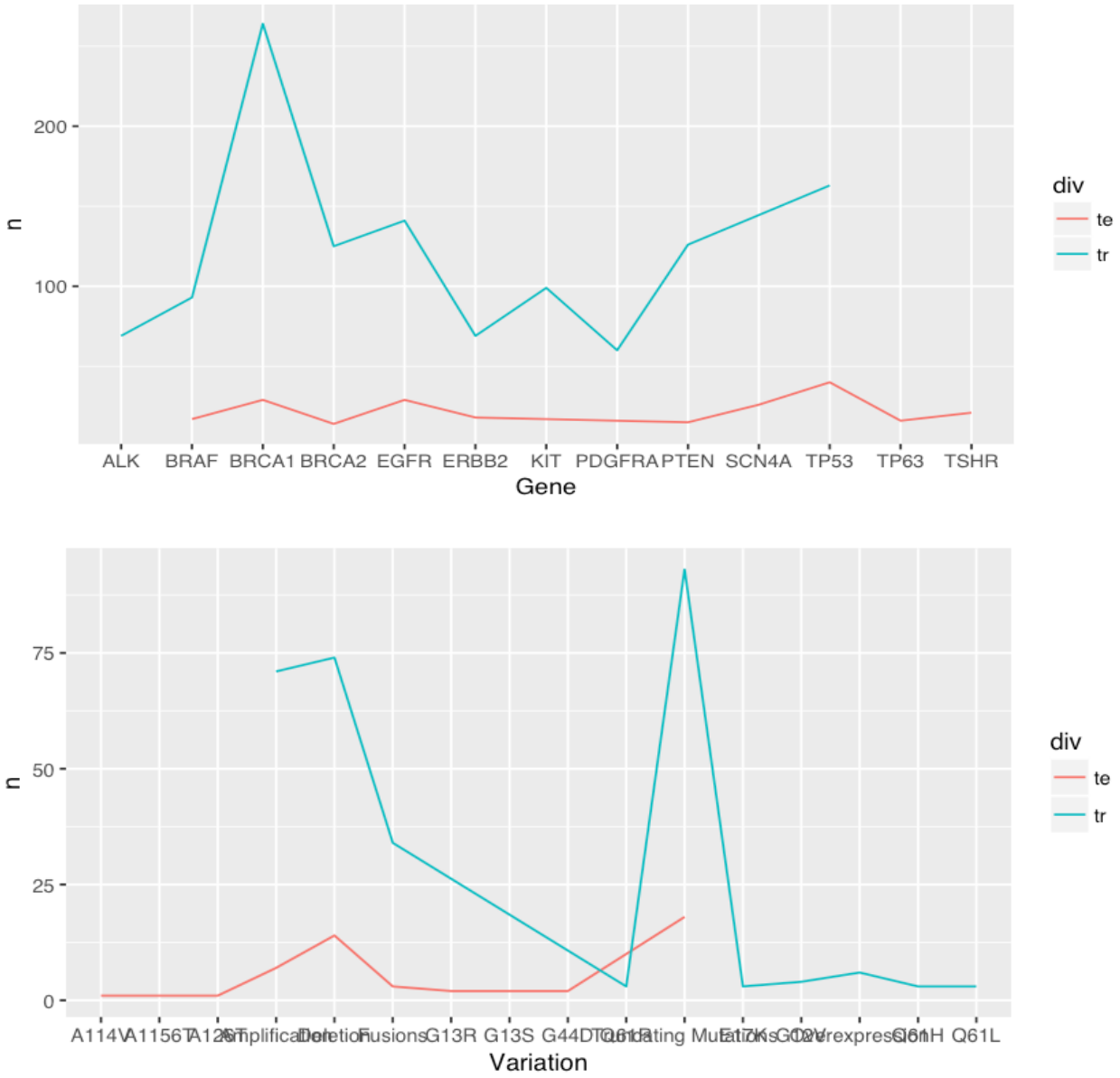

Figure 3. Comparing test and training data for frequency distribution of gene and variation variable. After sorting in descending order n=10 was used rather than using top 10 rows as they all have the similar frequency.

Clinical text data for non-semantics (sentences, words and word or text length) – the attributes which do not have any semantic information were searched within the text data and removed by analyzing the patterns of length of text, total number of sentences and words.

Removing Null values. - it was found that in training data, there were 5 rows which doesn't have any semantics text and none was found in test data.

Length of the text by class. - As can be seen from Figure 4, the classes 1,4 and 7 have more amount of information as compared to other classes and classes 3,8 and 9 have the least information. It was also observed that classes 3,8 and 9 also have less frequency. Hence, length of text might be important variable for further classification.

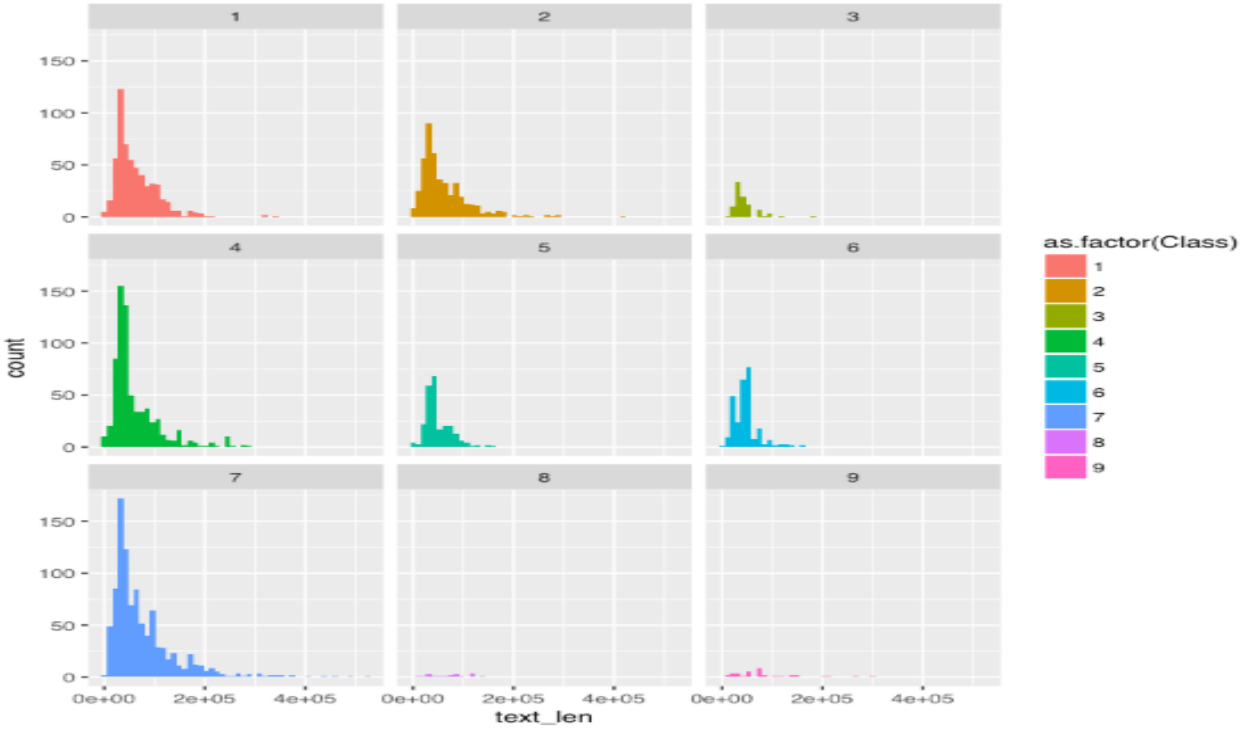

Figure 4. Length of text distribution across different class.

Framework of length of text, number of sentences and number of words by class and correlation between all three of these. - Boxplot and correlation matrix was made to visualize the framework of the three variables and to study the correlation between them. As can be seen from Figure 5 and Figure 6, there is very high correlation between these variables and the distribution for length of the text seems distinct and this dissimilarity in the length of text is because of the different words utilized, however, we it can't be interpreted that short words used have less amount of knowledge and meaning than the long words. Hence, to develop the model, only number of words variable was included.

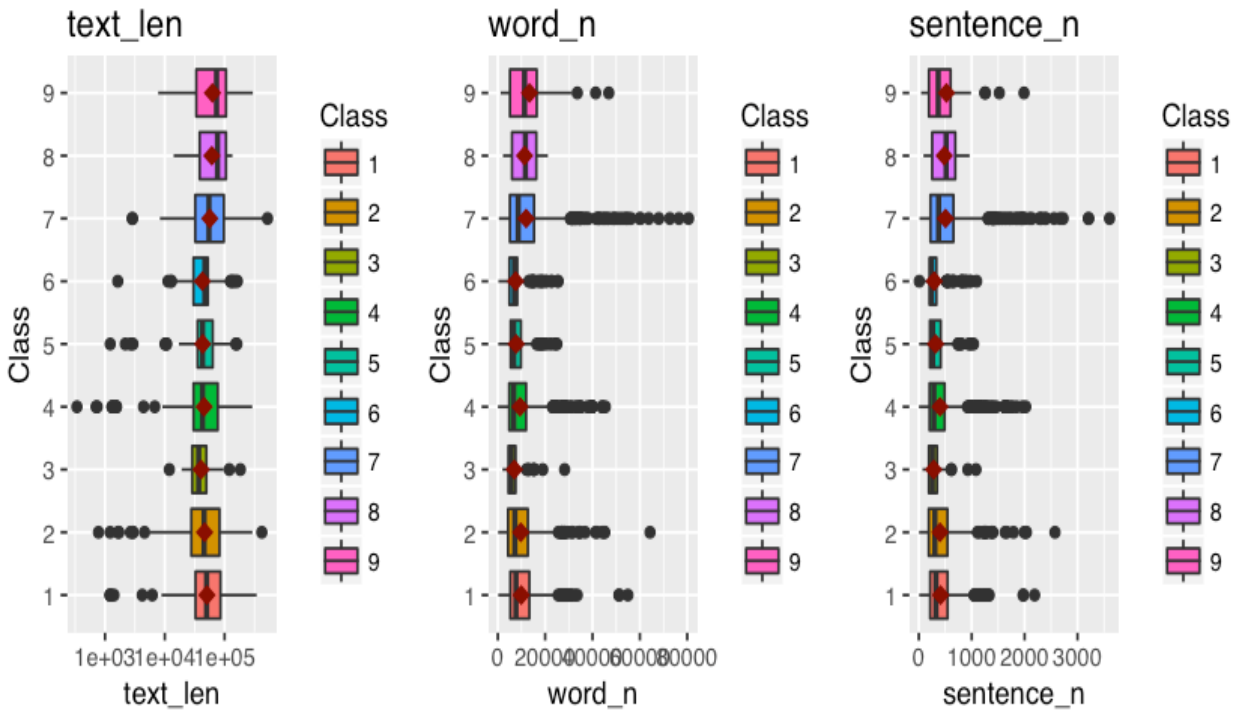

Figure 5. Boxplot for distribution of length of text, number of word and number of sentences.

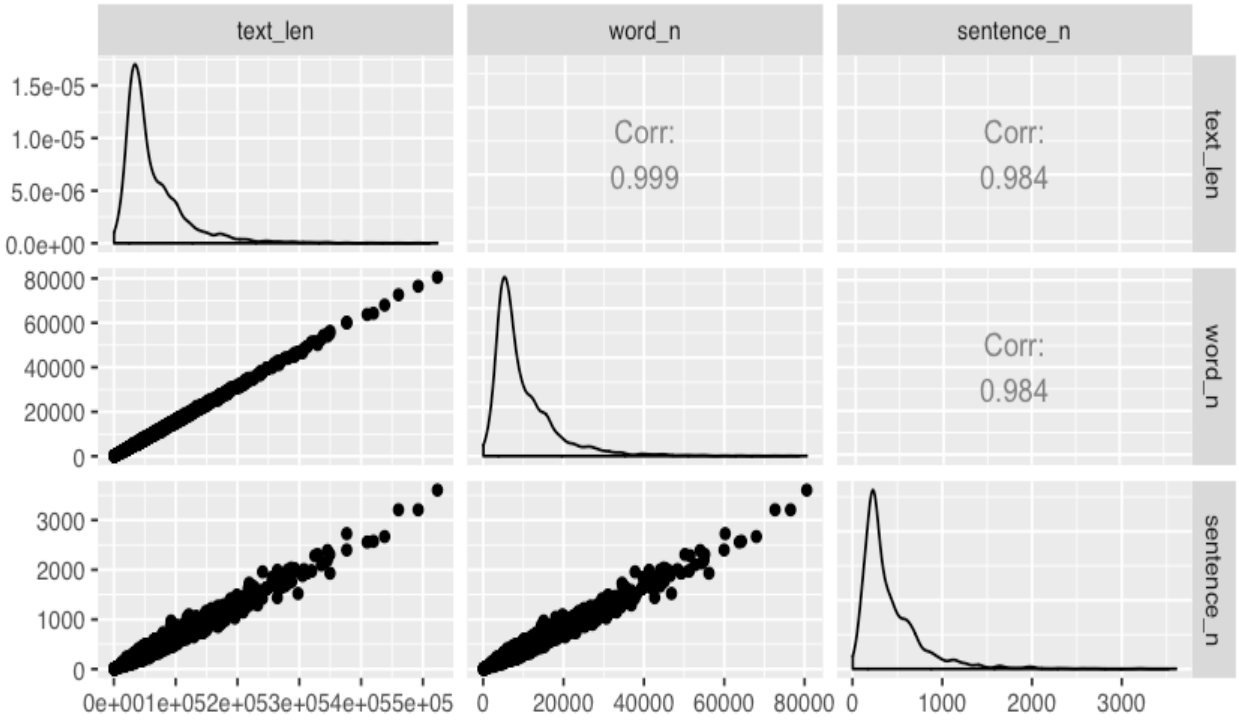

Figure 6. Correlation matrix for length of text, number of words and number of sentences

Clinical text for semantics (words) - Word is considered meaningful if it has least semantic unit i.e in English language, lemmatized or stemmed words are regarded to be least meaningful. The analysis for the word in text data was carried out in following 2 steps: First step was the pre-processing to remove the stop words and stemming and then transforming the word into numeric type. The complete text was separated into single words, called as stemming and pull out the most commonly used words from the complete text and these extracted 20 words are called as stop words. For the second step, words extracted in the first step are utilized and 2 sparse matrices consisting of rows as IDs and columns as stemmed words are created. One matrix has the information about each word and the second matrix has the tf-idf value for each word.

Extracting the stop words. - From the training and test datasets, the top 20 words were extracted by checking with the list from package tidytextR and the stop words were discarded. It was observed that in the test and training datasets, top 20 words were similar and resembled the stop words like 'for', 'by', 'the', 'a' and so on. Also, as can be seen from Figure 7, top 20 words distribution across different classes seems almost alike, hence we can discard these top 20 words and also stop words are removed from tidytextR package.

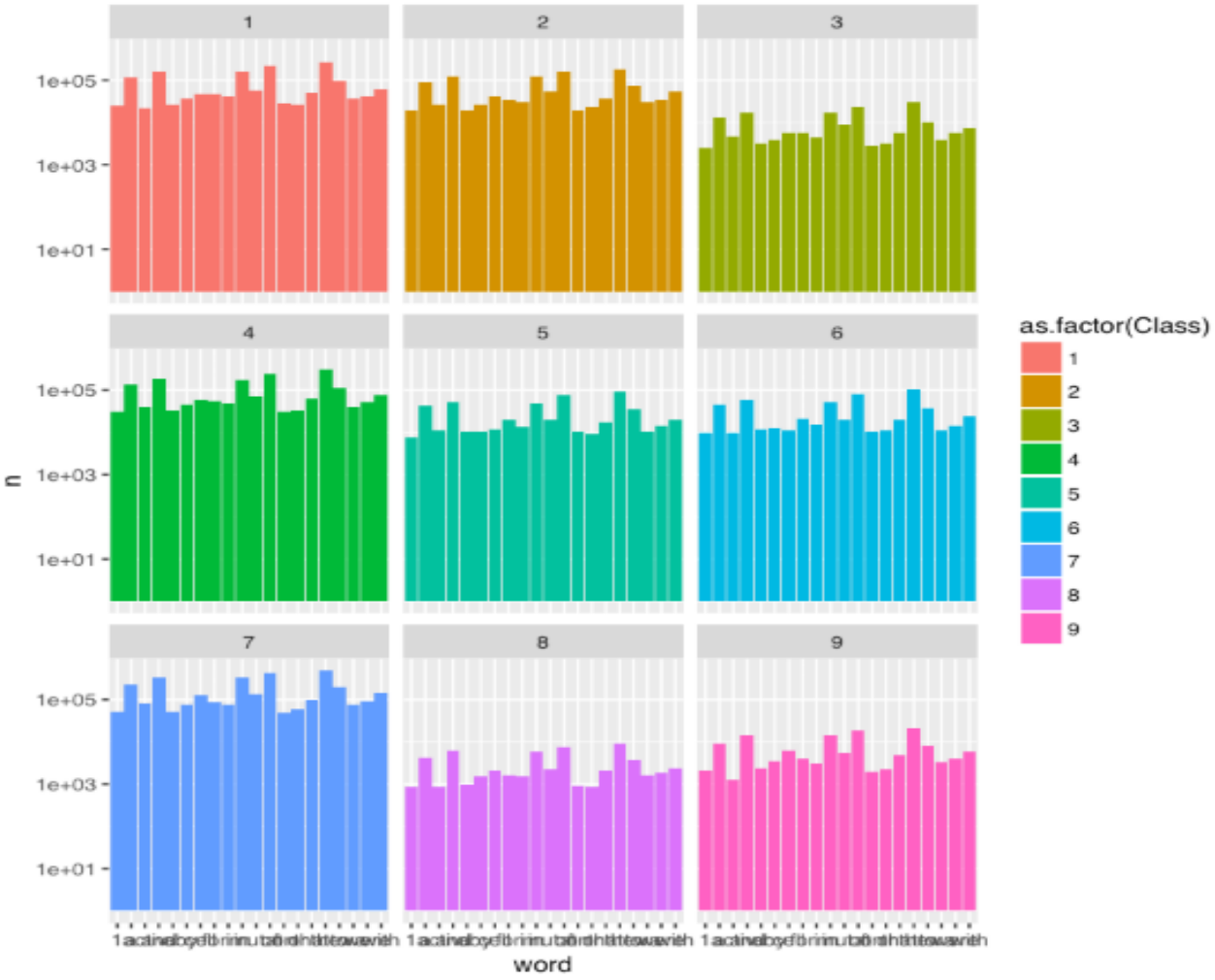

Figure 7. Top 20 words distribution across different class

Discarding non-useful words with tf-idf value. - A word that is classified in a class and contains less information would generally have low value for tfidf. When total number of words are more than it gets more complex to carry out the analysis using different computing tools, hence, it is important to discard the non-useful words which have very low tfidf values in class distribution. As can be seen from Figure 8, the distribution of sorted word with low tf-idf value is almost similar across all the classes. Again, class 3, 8 and 9 show low frequency due to the low length of text.

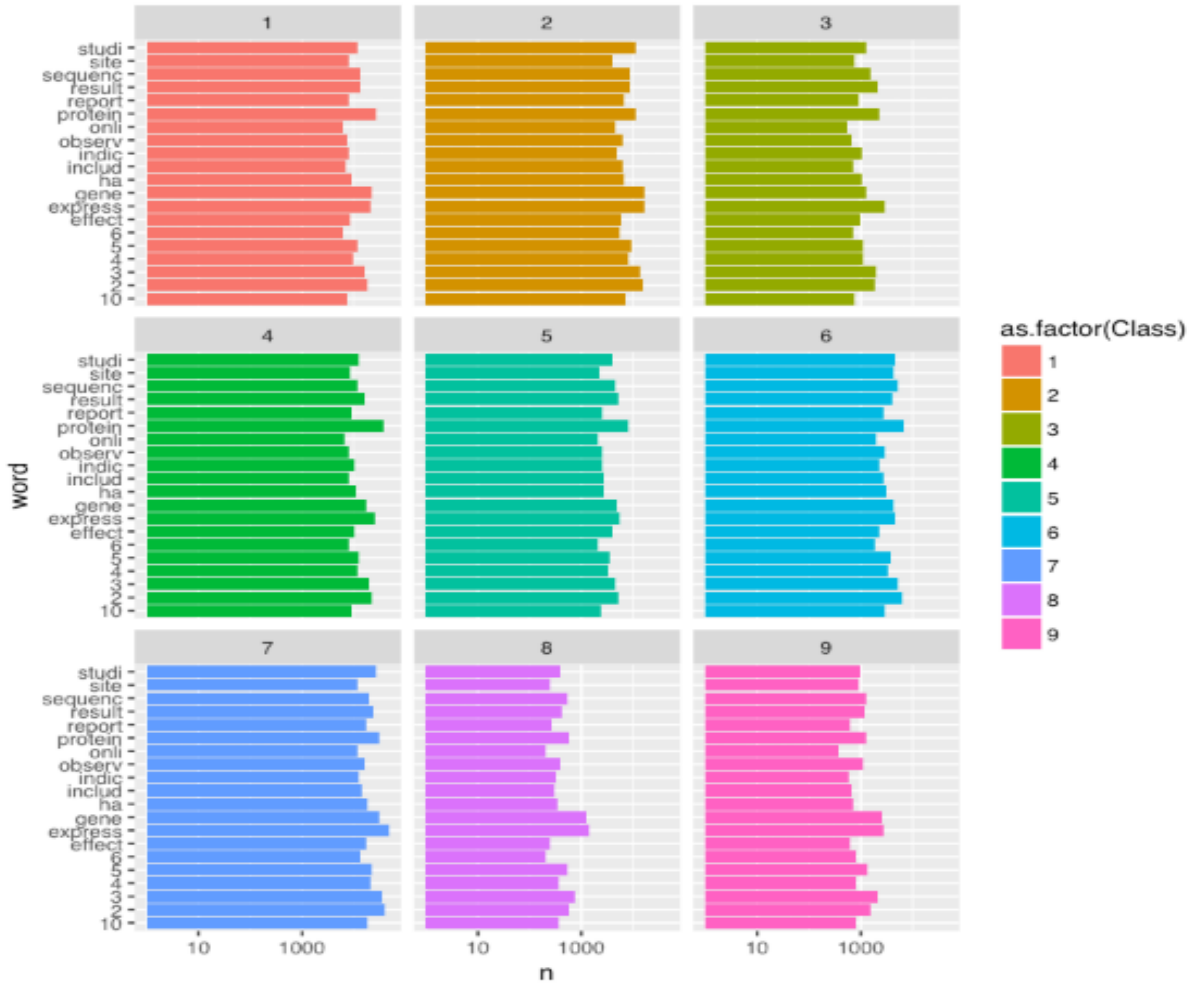

Figure 8. Words with low tf-idf value distribution across different class

Determining useful words with word frequency across different class. - The distribution of top 20 most common words across different class was analyzed in rest of the words after the previous sorting. As can be seen from Figure 9, the word distribution looks unusual except few words and have distinct frequencies across different class which implies that while creating model these might be differential variable.

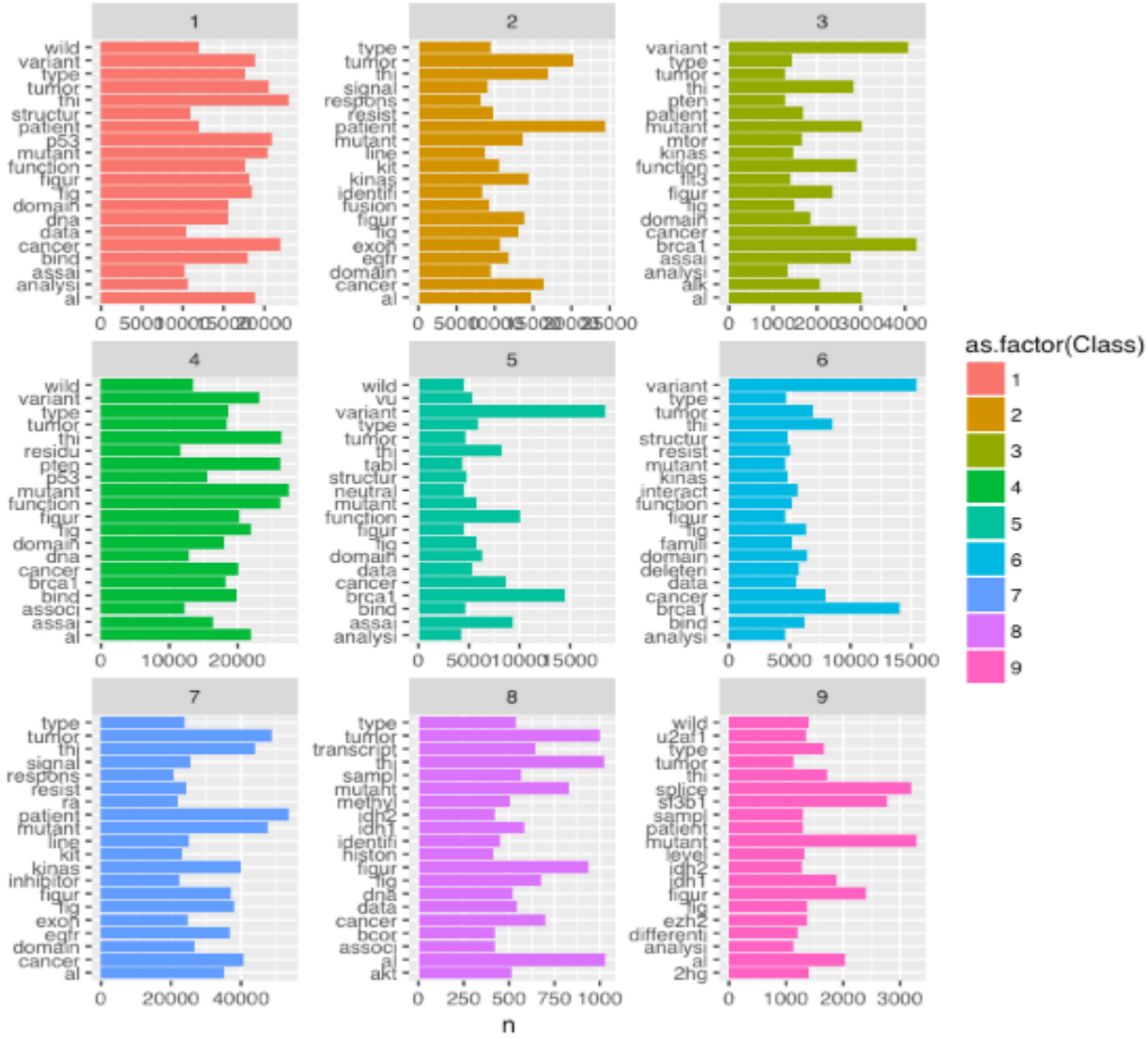

Figure 9. Top 20 words distribution across different class

Determining useful words with word tf-idf across different class. - Before this top 20 words were sorted by frequency across different class and now it is done by tf-idf. As can be observed from Figure 10, distribution of words is different than the previous distribution by frequencies, so for now will save both of these and will evaluate more later.

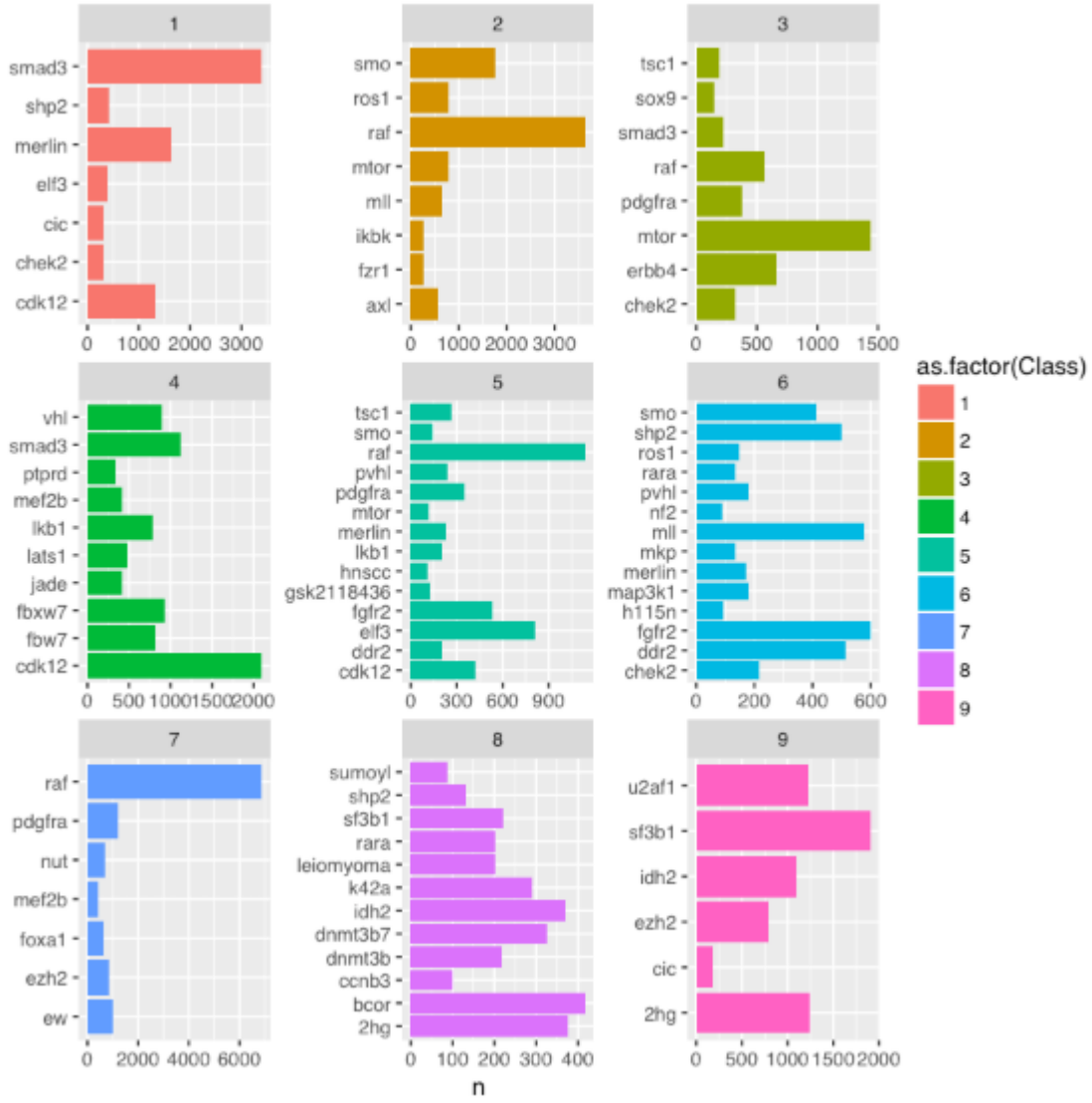

Figure 10. Words with top 20 tf-idf value distribution across different class

Data Analytics Tools/ Models

Following models were created to predict the class for genomic mutation:

1. **Xgboost Model**

   Xgboost is short form of eXtreme Gradient Boosting and is similar to gradient boosting framework but more efficient. It has both linear model solver and tree learning algorithms. This model was created in three steps:

   a) Data cleaning and pre-processing: The data was preprocessed to be used in xgboost using the information acquired from the visualization done in above steps. The data was transformed such that it gets convenient to attain the required results by using functions constructed using the xgboost library.

   b) Defining required functions: The purpose of these functions is to amend the input variables depending upon the frequency/tf-idf and word/bi-gram for models and finally evaluating the results with the help of tables.

   c) Different input variables used as word/bigram and n/tf-idf and interpreting results: After varying input variables, the results were compared to check the accuracy of all the models and it was found that the accuracy values were highest with bigram and n i.e. 62 percent as compared to other models like word and tdf-idf have 59.73 percent, word and n have 61.9 percent and bigram and tf-idf have 55.8 percent. Infact, the final performance of the model was not good enough and not much difference was found inother models developed with word and bigram.

2. **Naive Bayes Algorithm**

   Naive Bayes classifiers are simple probabilistic classifiers that are based on using Bayes theorem with strong independence assumptions between the features. The model was

developed following steps as pre-processing the data and then training the model and applying naïve bayes classifier and finally evaluating the result and creating the confusion matrix. The validation accuracy for the model was 50 percent which indicates that the results are not highly accurate. The confusion matrix of the classifier across all the 9 class is shown in Figure 11.

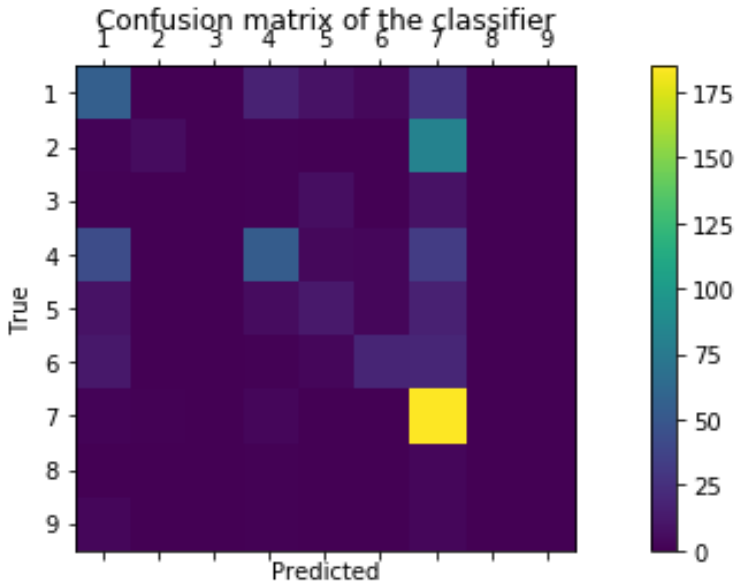

Figure 11. Confusion matrix of the naïve bayes classifier for all classes

3. Support Vector Machines

Support vector machine (SVM) are supervised learning models which work by locating observations in a plane and then differentiating them by creating borders among them. The models are related with learning algorithms which can perform data analysis on classification data. SVM works best for text classification as it can tackle complex and large concepts and text classification contains several features which infers lots of information. When the model was created to classify the gene text across different class, it gave validation accuracy of 60 percent. The cross-validation scores are shown in Figure.

```
Validation Accuracy: 60
[[ 71   0   0  24  10   1   8   0   0]
 [  5  29   0   1   4   0  52   0   0]
 [  2   1   2   2   5   0   6   0   0]
 [ 35   0   1  84   2   3  12   0   0]
 [  9   1   0   6  15   2  15   0   0]
 [  8   1   1   5   2  26  12   0   0]
 [  2  13   0   0   3   0 173   0   0]
 [  0   0   0   1   0   0   2   0   1]
 [  0   1   0   1   0   0   0   0   5]]
```

Figure 11. Cross validation scores for SVM.

Results and Discussion

The present study focused on solving some important key points with the help of machine learning algorithms which could make significant difference in the field of precision medicine. The first hypothesis tested was the accuracy of the gene annotation and classification of genetic mutations in different classes. This was done by developing machine learning algorithms like xgboost model, naive bayes classifier and support vector machines and finally checking the model accuracy to determine the level to which the classification was done correctly. The analysis indicates that with the xgboost model the maximum accuracy was 62%, even if the input variables were changed and tested again there was not much difference in the accuracy percentage. The naïve bayes gave almost 50 percent accuracy and with the support vector machines it was 60 percent. This suggests that the results obtained were not satisfactory as the mutation were not classified accurately. Hence, the first hypothesis doesn't stand its ground for accurate classification of mutations but there is still scope to develop the model and increase the accuracy of the results.

The next hypothesis tested was that about the total time taken and effort for the process of genome annotation by using machine learning algorithm as compared to the conventional manual process. The evaluation of the test data after training the model, whether it was xgboost, bayes

classifier or support vector machines hardly took any time as compared to several years of manual hard work taken by clinical experts and pathologists to study, interpret and classify genetic mutations on the basis of clinical findings based text. This can completely revolutionalise the process of genome annotation and classification of mutation as it would guarantee faster treatment solutions to so many patients and that too highly accurate which could save lot of lives, time, money and effort.

Limitations of this Study

The major challenge faced while conducting the analysis was that the training dataset only had 3322 observations which resulted in poor training of the model as generally the deep learning algorithms works best when the number of observations are large. To overcome this limitation the model training was kept simple and some data augmentation was done to increase the size and then train the models.

Another important challenge was that the type of genes in the test and training samples were very distinct and varied and also limited information was present in the gene or mutation sets, so text data required lot of consideration.

Conclusion and Future Study

This study adds to a growing corpus of research focused on the use of big data medicinal field and especially focusing on precision medicine. The findings suggest that improving the machine learning algorithms like support vector machine, xgboost and naïve bayes classifiers can be used for automatic classification of mutation and gene annotation which can make a huge change by reducing the tedious and time consuming manual process and provide effective tailored treatment to patients depending upon their genetic information. Accepting the fact, that machine learning algorithms and statistical tools are far from exactly reproducing the work of clinical experts and pathologists, future studies could fruitfully explore this issue further by improving the existing models and by developing and testing more machine learning algorithms.

References

Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Research International. 9.

Chandra, R. (2017). The role of pharmacogenomics in precision medicine. Medical Laboratory Observer, 49(9), 8.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. BioData Mining, 10, 35.

Chute, C.G., Ullman-Cullere, M., Wood, G.M., Lin, S.M., He, M., Pathak, J. (2013). Some experiences and opportunities for big data in translational research. Genet. Med. 15, 802–809.

Daniel Richard Leff, & Guang-Zhong Yang. (2015). Big Data for Precision Medicine. Engineering, 1(3), 277-279.

Dinov, I.D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. GigaScience. 5, 12.

Giri, K. & Lone, A. (2014). Big Data -Overview and Challenges. International Journal of Advanced Research in Computer Science and Software Engineering. 4.

Glasser Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M. (2016). A multi-modal parcellation of the human cerebral cortex. Nature. 536, 171–178

Gray, M., Lagerberg, T., & Dombrádi, V. (2017). Equity and Value in ‘Precision Medicine’. The New Bioethics, 23(1), 87-94.

Janet G. B., Debra R., & Ian P. (2017). Big Data in the Era of Health Information Exchanges: Challenges and Opportunities for Public Health. Informatics (Basel), 4(4), 39.

Kingsmore, S., Petrikin, J., Willig, L., & Guest, E. (2015). Emergency medical genomes: A breakthrough application of precision medicine. Genome Medicine, 7(1), 82.

Leff D.R. and Yang G.Z. (2015). Big Data for Precision Medicine. Engineering. 1(3):227 279. doi.org/10.15302/J-ENG-2015075.

Lewis, S., Ashburner, M., & Reese, M. (2000). Annotating eukaryote genomes. Current Opinion in Structural Biology, 10(3), 349-54.

Lin, C.P., Stephens, K.A., Baldwin, L.M., Keppel, G.A., Whitener, R.J., Echo-Hawk, A., Korngiebel, D. (2014). Developing governance for federated community-based EHR data sharing. AMIA Jt. Summits Transl. Sci. Proc. 71–76.

Malod-Dognin, Petschnigg, & Pržulj. (2018). Precision medicine — A promising, yet challenging road lies ahead. Current Opinion in Systems Biology, 7, 1-7.

Maxwell W.L., & William S.N. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-32.

Mirnezami R, Nicholson J, Darzi A. (2012). Preparing for precision medicine. N Engl J Med, 366(6):489–91.10.1056/NEJM. p1114866.

Mungall, C. J. (2002). An integrated computational pipeline and database to support wholegenome sequence annotation. Genome Biol. 3, 081.

Robert R.K., Joel S.R., Mark B.G., & Angus C.N. (2014). Decoding neuroproteomics: Integrating the genome, translatome and functional anatomy. Nature Neuroscience, 17(11), 1491-9.

Rodriguez-Mazahua, L., Rodriguez-Enriquez, C., Sanchez-Cervantes, J., Cervantes, J., GarciaAlcaraz, J., & Alor-Hernandez, G. (2016). A general perspective of Big Data: Applications, tools, challenges and trends. 72(8), 3073.

Ross, M. K., Wei, W., & Ohno-Machado, L. (2014). "Big Data" and the Electronic Health Record. Yearbook of Medical Informatics, 9(1), 97–104.

Rust, Mongin, & Birney. (2002). Genome annotation techniques: New approaches and challenges. Drug Discovery Today, 7(11), S70-S76.

Saville K. and McNeil G. (2012). An introduction to the gene annotation process, from beginning to end, using a simple example from Drosophila erecta. GEP, 1-46.

Sharma, V., Kumar, A., Panat, L., Karajkhede, G., Lele, A. (2015). Malaria outbreak prediction model using machine learning. Int. J. Adv. Res. Comput. Eng. Technol, 4, 4415–4419.

Sugeir, & Naylor. (2018). Critical Care and Personalized or Precision Medicine: Who needs whom? Journal of Critical Care, 43, 401-405.

Sykiotis, GP, Kalliolians, GD, Papavassiliou, AG. (2005). Pharmacogenetic principles in the Hippocratic writings. J Clin Pharmacol. 45(11):1218–1220.

Taylor, R., Moore, C., Cheung, K., & Brandt, C. (2018). Predicting urinary tract infections in the emergency department with machine learning. PLoS ONE, 13(3), E0194085.

Vassy, J.L., Korf, B.R. (2015). Green, R.C. How to know when physicians are ready for genomic medicine. Sci. Transl. Med. 7, 287-219.

Yandell M. & Ence D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329-42.

Yip KY, Cheng C, and Gerstein M. (2013). Machine learning and genome annotation: a match meant to be? Genome Biol, 14(5):205.