How to Cite:
Raipure, S., Awachat, S., Amudhavel, J., & Kalambe, K. (2022). Genome sequencing using machine learning with a special focus on tuberculosis. *International Journal of Health Sciences*, *6*(S2), 9603–9614. https://doi.org/10.53730/ijhs.v6nS2.7512

Genome sequencing using machine learning with a special focus on tuberculosis

Shwetal Raipure
Department of CSE, RCOEM, Nagpur
Corresponding author email: raipuress@rknec.edu

Snehal Awachat
Department of CSE, RCOEM, Nagpur
Email: awachatsr@rknec.edu

Dr. J. Amudhavel
Department of CSE, VIT Bhopal
Email: amudhavel.j@vitbhopal.ac.in

Kavita Kalambe
Department of CSE,RCOEM, Nagpur
Email: kalambeka@rknec.edu

***Abstract*---**Machine learning is becoming increasingly prevalent. However, in the discipline of Bioinformatics and Computational Biology, it is not a popular use case. Machine learning techniques are used in only a few technologies. The majority of the tools are built using deterministic techniques and algorithms. Deoxyribonucleic acid (DNA) is a biological macromolecule composed up of deoxyribonucleic acid. Its main function is to store data. Due to breakthroughs in sequencing technology, DNA sequence data is presently rising at an exponential pace, ushering the study of DNA sequences into the big data age. Machine learning is also a powerful tool for massive processing it learns on its own from large volumes of data. We've talked about machine learning techniques and how they can be used to improve genome sequencing accuracy. In our review we have also discussed about genome sequence for Mycobacterium Tuberculosis. Tuberculosis is because of the bacteria, Tuberculosis caused by Mycobacterium tuberculosis. TB is considered one of the leading the reasons for dying all over the world. MDR-TB is a form of germs that cause tuberculosis that is not susceptible to anti-TB medications such as isoniazid (INH) and rifampin (RMP). To predict MDR TB, many machine learning algorithms have been widely used. In this article, we look at various Machine Learning Methodologies for predicting MDR-TB. Various techniques for estimating features, as well as the

execution of various Models of machine learning have all been investigated. In addition, in recent decades, the use of unique machine learning system models for separating the functional flaws of MDR-TB has been discussed.

***Keywords*---**mayobacterium tuberculosis, genome sequencing, machine learning.

Introduction

Celera Genomics and the International Human Genome Sequencing Consortium have released a new paper the first draughts of human genetic material in 2001, and genomics is a science that deals with the study of DNA was forever changed[1]. While the euchromatic fraction of the genome was effectively covered by these draughts and subsequent revisions, heterochromatin, as well as a slew of other factors complicated areas was left incomplete or incorrect. Whole genome sequencing (WGS), referred to as full Genome sequencing, full genome sequencing, or whole genome sequencing are all terms used to describe the process of sequencing a person's genome, is the method assessing the completeness, all of an organism's DNA sequence, or almost all of it, may be read at the same time [2]. This comprises chromosomal DNA sequencing, mitochondrial DNA sequencing, and chloroplast DNA sequencing in plants. Deoxyribonucleic acid (DNA) is a biological macromolecule composed up of deoxyribonucleic acid (DNA). Its main function is to store data. Due to breakthroughs in sequencing technology, DNA sequence data is presently rising at an exponential pace, ushering the study of DNA sequences into the big data age. In recent years, the Yang and his team have a thorough examination of development and prospective difficulties in the realm of data mining DNA sequences by analyzing their related biological application background and importance [3]. Pattern recognition capabilities of artificial neural network-based learning systems are widely known, and their deep architectures—known as deep learning (DL)—have recently been effectively employed to address many complicated pattern recognition tasks. A meta-analysis was undertaken, and the obtained resources were rigorously reviewed by Mahmud and colleagues, to see how DL—especially its diverse architectures—has made a contribution and has been a part used in the mining of biological data relevant to those three kinds of people [4]. Researchers and medical specialists may benefit from Sequence analysis of the human genome and modern artificial intelligence tools to better comprehend COVID-19 genetic variations or SARS-CoV-2 genetic variants. COVID-19 the sequencing of a person's genome is analyzed critical for understanding the virus's source, behaviour, as well as structure, and it may be of assistance. The technology aids in the extraction of critical information derived from the sequencing of the genomes of various viruses. Ahmed and his colleagues do extraction of fundamental data for comparative data analysis from COVID-19 genome sequences and additional genome sequences, such as nucleotide tri-nucleotide composition and frequency compositions, amino acid counts, genome alignment, and DNA similarity information. With a 97 percent accuracy rate, percent for COVID-19, 96 percent for SARS, and 95 percent for Sequences of the MERS and Ebola genomes, the algorithm produces good classification results[5].

Motivation

Whole exome sequencing (WES) allows researchers to examine all of the human genome's protein coding sequences. This method allows researchers to look at cancer-related genetic mutations that are mostly found in exotic areas. WES provides high-throughput outcomes at a low cost. Here, Bartha and his colleagues looked at analysis methods that would allow WES data to be used in clinical and research settings[6]. WES can detect SNVs (single nucleotide variants) and copy number variations (CNVs) are two types of genetic variations (CNVs) in the beginning, and the information gathered from both approaches can be merged and used further.

Tuberous sclerosis (TSC) is a pediatric disease with an autosomal dominant inheritance pattern. The outcomes of genetic tests on a TSC in a newborn are discussed. TSC was discovered echocardiogram and magnetic resonance imaging in a newborn who had no clinical indications of the condition. The diagnosis was confirmed using next-generation sequencing (NGS) and multiple ligation-dependent probe amplification (MLPA) of the TSC1 and TSC2 gene exons. The MLPA despite the fact that the findings were negative, NGS revealed a heterozygous T residue insertion c.2165 (exon 17) to c.2166 (exon 18) (exon 17) in the TSC1 gene, a mutation that causes a loss of function. These findings were validated by means of Sanger sequencing. In the infant, this genetic alteration was evident, while the genotypes of the parents were wild-type, suggesting a mutation had occurred spontaneously [7]. TSC is inherited as an autosomal dominant trait disease caused by latent alterations in the TSC1 or TSC2 genes. TSC is characterized by neurocutaneous syndrome and benign hamartoma development.

Related Work

There are a few tens of gene mutations with extra concrete treatments that should be investigated in adulthood, including those with restrictive/substitutive medicines, therapies that change signaling pathways, and channelopathies. Phenotyping adult disorders, early diagnosis, and the creation of focused therapeutics all need further investigation. [1].

There were 18 patients with tuberous sclerosis complex (mean [SD] age: 48 years [9.54]; 13 women [72 percent]), 16 patients with frontotemporal dementia (60 [6.93] years; 7 women [44 percent]), and 18 healthy controls (63 [3.85] years; 9 women [50 percent]). When the tuberous sclerosis–associated neuropsychiatric disorders checklist and cognitive test results from the tuberous sclerosis complex and frontotemporal dementia populations were examined, there were no significant differences [2].It is necessary to develop an integrated diagnosis that includes both histology and molecular findings based on genetic and genomic markers.[3].Integrative clinical sequencing programs that are used at the point of care have the potential to enhance cancer patient treatment [4].TSC is characterized by hamartomas in many organ systems, including the brain, heart, skin, lungs, and kidneys, which may lead to learning impairments, epilepsy, behavioral issues, and renal failure. Rhabdomyoma has a highly significant etiological diagnosis [5].In TSC1, heterozygous nonsense mutations were

discovered, as well as a heterozygous intronic variation in TSC2. Two heterozygous missense mutations in polycystic kidney and hepatic disease 1 (PKHD1) were discovered in one patient, confirming polycystic kidney disease type 4. One patient had a heterozygous missense mutation in solute carrier family 12 member 5 (SLC12A5), which has been associated to idiopathic generalized epilepsy type 14. One patient had a heterozygous nonsense variation of ring finger protein 213 (RNF213), which is linked to Moyamoya disease type susceptibility [6].

The findings show that phase analysis based on targeted double-stranded cDNA sequencing is helpful for discovering compound heterozygous mutations and provides information on allelic expression [7].The approach involved first developing a list of genome regions known to be similar to other regions and then using them to teach a machine learning algorithm how to recognize them. Researchers then used the algorithm to spot mutations in different tissues—2,658 samples from the Pan-Cancer Whole Genome Analysis dataset [8].A genome sequencing is performed 320 Chinese children with epilepsy were studied, and single-nucleotide and copy number variations were evaluated in all samples. The probands' entire lineage and clinical data were established and followed up on [9].

Only two modifications were made to the clinical diagnostic criteria published in 2013: the wording "multiple cortical tubers and/or radial migration lines" was changed with "multiple cortical tubers and/or radial migration lines," and sclerotic bone lesions were restored as a minor criterion. The genetic diagnostic criteria were confirmed, with current studies underlining the fact that some people with TSC are genetically mosaic for TSC1 or TSC2 mutations. Increased focus on early screening for electroencephalographic abnormalities, greater monitoring and treatment of TSC-related neuropsychiatric problems, and new pharmaceutical approvals all contributed to changes in surveillance and management criteria [10].We provide a clinical example with TSC with a request for genetic counseling on reproductive concerns. Initial examination of peripheral blood revealed no mutations; however, mosaicism for a potentially pathogenic frameshift variation in TSC2 was observed at a rate of 15% in renal angiomyolipoma tissue [11].

ML algorithms for Genome sequencing

Ensemble

Ensemble techniques are used to train many machines learning models to address the same issue. Unlike ensemble, a single classifier approaches attempt to create a group mix a variety of models. Ensemble learning is also known as committee learning or multiple systems for learning classifiers [19]. Traditionally, Models of learning may be combined, there are three methods to do this: by average, vote, or learning model.

Evaluation Metrics

To compare the models in this study, seven measures are used: accuracy, precision, sensitivity, specificity, F1-score, AUC ROC, and F1-macro. To

comprehend these measures, A confusion matrix's composition must be defined:, false positive (FP), false negative (FN), true positive (TP) and true negative (TN)

Accuracy to be considered a is a performance metric that indicates how many samples were correctly classified in relation to the whole, that is, the ratio between the sum of FP and FN and the sum of all samples Equation (1).

$$a = \frac{FP+FN}{FP+TP+FN+TN} \quad (1)$$

Precision to be considered p indicates the correct classifications among all classified as positive by the model, that is, the ratio between FP and the sum of FP and TP Equation (2).

$$p = \frac{FP}{FP+TP} \quad (2)$$

Sensitivity to be considered s indicates the correct classifications among all expected cases as correct, that is, the ratio between TP and the sum of FP and TN Equation (3).

$$s = \frac{FP}{FP+TN} \quad (3)$$

Specificity as considered as S indicates how well the classifier can identify correctly the negative cases, that is the ratio between FN and the sum of FN and TP Equation (4).

$$S = \frac{FN}{FN+TP} \quad (4)$$

The P1-score metric, used in the feature selection step, is defined as the harmonic mean between precision and sensitivity, as presented in Equation (5). Note that, if TP = 0, all positive samples are misclassified, and if FP = FN = 0, there is a perfect classification.

$$P1 - score = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \quad (5)$$

The P1-macro average (P1-macro) is a variant of the P1-score, composed of the average of the P1-score of the positive class and the P1-score of the negative class Equation (6). The more the model hits the prediction in both classes (positive and negative), the P1-macro tends to indicate, in general, a degree of a model correctness without bias by balanced or imbalanced the data set.

$$P1 - macrao = \frac{1}{m}\sum_{i=1}^{m} P1 - score_i \quad (6)$$

Method

Principal Component Analysis

Principal component analysis (PCA) is a popular method for analysing data. It is intended that fewer variables can be utilised to understand the majority of

variables in the original data, and that the data's key feature components can be recovered.Suppose the sample set Y includes m samples, and each sample is k-dimensional vector. At the same time, the sum of these l samples is 0 as shown in Equations (7) and (8).

$$Y_{k\times l} = (y_1, y_2, \dots, y_k) \quad (7)$$

$$\sum_{i=0}^{l} y_i = 0 \quad (8)$$

Suppose the new coordinate system is $V_{n\times n} = (\omega_1, \omega_2, \dots, \omega_n)$ after the transformation of projection, where ω_i is an orthonormal basis. The original data sample is projected to a new coordinate system. The projection rule is shown in Equation (9).

$$Z_{n\times m} = W_{n\times n}^{T} \times X_{n\times m} \quad (9)$$

The variance of these samples after projection should be increased in order to separate all samples as far as feasible after projection. As a result, the improved objective function appears in Equations (10), where I is the unit vector.

$$\max_{W} tr\,(W^T X X^T W) s.t. W^T W = I \quad (10)$$

The problem is solved using the Lagrange multiplier technique, and the goal function is as follows.

$$J(W) = tr\big(W^T X X^T W + \lambda(W^T W - I)\big) \quad (11)$$

Equation shows how to get the derivative of the given equation. (12)

$$XX^T W = \lambda \mathrm{W} \quad (12)$$

The preceding equation shows that in order to determine the eigenspace, $W_{n\times n}$, the covariance matrix's related eigenvalues and eigenvectors should be determined.

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq t_1 \quad (13)$$

Then, the eigenspace $W_{n\times k} = (\omega_1, \omega_2, \cdots, \omega_n)(k < n)$ It is possible to determine a set of k eigenvectors. Noise is often associated with the information contained in the deleted segment. As a result, removing this piece of data may enhance the experimental effect to some amount.

Kernel Principal Component Analysis

Kernel principal component analysis (KPCA) can mine the nonlinear information in the data set better than PCA. In KPCA, a kernel function is introduced and utilized to compute the input data's kernel matrix K. The kernel function is chosen to be a Gaussian kernel, hence the kernel matrix K is described as

$$K_{i\times j} = e^{\|X_i - X_j\|^2/2\sigma^2} \quad (14)$$

Then, eigenvalues and eigenvectors of the kernel matrix K are calculated. After arranging the eigenvalues from the largest to the smallest, the reconstruction threshold t1 should be set to determine the eigenspace W. In our experiment, the reconstruction threshold t1 is set 95% and 99% for both PCA and KPCA.

Multiset Canonical Correlation Analysis

Because the PCA and KPCA algorithms can only analyze a certain kind of sample set. A multiset canonical correlation analysis (MCCA) approach is utilized to get lipids that better discriminate three samples. MCCA is a method for analyzing the connection between two or more sets of data [20]. The primary notion behind MCCA is that when the correlation coefficient is high, the correlation coefficient is low β between several sample sets is maximum, the typical variable w_i the sample set that corresponds to each sample set is located. The objective function is defined as where the number of sample sets is u and each sample set has N samples

$$\arg max\ \beta = \sum_{\substack{k,l=1 \\ k\neq l}}^{u} W_k^T \sum_{ij} w_l (k \neq l) s.t. \sum_{k=1}^{u} w_l = 1 \quad (15)$$

Where$\sum_{ij} = x_k^T \cdot x_l$.

The following equation may be found using the Lagrange multiplier approach for the objective function:

$$(C - D)\, w = \beta D w \quad (16)$$

$$\text{Where} \quad C = \begin{pmatrix} x_1 x_1^T \cdots x_1 x_N^T \\ \vdots \quad \ddots \quad \vdots \\ x_N x_1^T \cdots x_N x_N^T \end{pmatrix} and\ D = \begin{pmatrix} x_1 x_1^T \cdots & 0 \\ \vdots & \ddots \quad \vdots \\ 0 & \cdots x_N x_N^T \end{pmatrix}$$

Then, using the usual variable, the influential lipids may be identified w_i

Feature Selection

PCA and KPCA may be used to recreate data, as seen in Equation (17).

$$Y_{k\times m} = W_{n\times k}^T \times X_{n\times m} \quad (17)$$

After dimension reduction, the resulting data is $Y_{k\times m} = (y_1, y_2, \cdots, y_m)$. The n-dimensional data in the original data X is reduced to k-dimensional data in this manner. The resultant Y is already data in another spatial dimension, not the original data's lipid information.The location information of numerous basis space components, for starters $w_{j\times p}$which have the most impact on the new information Y is enumerated, and the lipids in the original data are matched. Then, the

frequency of each original data in the same eigenvector is calculated, and weights are added according to its eigenvalues. Finally, the frequencies and weights of the original data counted by all eigenvectors are multiplied, and the product is summed if it is the same original data. The goods are sorted in decreasing order, with the most expensive at the top k as a consequence, lipids have a higher influence on acne..

In Equation (17), each element in Y is calculated as shown in Equation (18):

$$\begin{bmatrix} y_{1\times1} & y_{1\times2} & \cdots & y_{1\times m} \\ y_{2\times1} & y_{2\times2} & \cdots & y_{2\times m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ y_{k\times1} & y_{k\times2} & \cdots & y_{k\times m} \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} w_{j\times1}.x_{j\times1}. & w_{j\times1}.x_{j\times2}. & \cdots & w_{j\times1}.x_{j\times m}. \\ w_{j\times2}.x_{j\times1}. & w_{j\times2}.x_{j\times2}. & \cdots & w_{j\times2}.x_{j\times1}. \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ w_{j\times k}.x_{j\times1}. & w_{j\times k}.x_{j\times2}. & \cdots & w_{j\times k}.x_{j\times m}. \end{bmatrix} \quad (18)$$

ML application in MDR-TB

Previously, there were a variety of methods for predicting MDR-TB infection. The majority of methods rely on Deep learning methods and genome sequencing, such as CNN, are also used for the purpose of forecasting. This research attempts to predict MDR-TB from a dataset that includes factors such as age, gender, alcoholism, tobacco use, HIV infection, and if the patient is using first-line medications such as isoniazid, rifampicin, pyrazinamide, and ethambutol are some of the antibiotics used. Other characteristics include whether or if the patient is on second-line drugs, such as Group A, Group B, or Group C. The dataset's result is divided into four categories: Defaulted, died, had treatment completed, and was cured. The class label has passed away indicates whether or not the patient died as a result of MDR-TB. The term "defaulted label" refers to a patient who has ceased taking a prescription recommended by a doctor for a variety of reasons. The term "treatment completed" denotes that the patient has completed his or her treatment finished the prescribed dosage but is still suffering from MDR-TB. Finally, the label Cured indicates that the patient is free of MDR-TB. In this study, various machine learning and ensemble techniques in the database are proposed.

Algorithms for Analysis

Logistic Regression

The process of estimating the parameters of a logistic model is known as logistic regression (or logit regression) (the coefficients in the linear combination). In binary logistic regression, there is a single binary dependent variable, coded by an indicator variable, with two values labeled "0" and "1," and the independent variables may be either binary variables (two classes, coded by an indicator variable) or continuous variables (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.

Binary variables are widely used in statistics to model the probability of a certain class or event taking place, such as the probability of a team winning, of a patient being healthy, etc. (see § Applications), and the logistic model has been the most commonly used model for binary regression since about 1970. Binary variables can be generalized to categorical variables when there are more than two possible values (e.g. whether an image is of a cat, dog, lion, etc.), and the binary logistic regression generalized to multinomial logistic regression.

Decision Tree

Many fields of machine learning, including regression and classification, use Decision Trees. A decision tree is used to visually and plainly illustrate decision making and decisions while undertaking analysis. It makes decisions using a tree model. Decision trees are created using an algorithmic approach that seeks alternative ways to segment a data set based on certain conditions. It's a predictive modelling tool that covers a variety of topics. It employs a supervised learning strategy.

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Decision Trees are a supervised non-parametric learning process that can be used for both regression and classification. The goal of Decision Tree is to construct a prototype that assumes the value of a target by using data attributes to train decision rules. The majority of decision rules are in the form of if-else statements. The rules become more complex as the tree grows deeper, and the model fits better.

Random Forest

Random forests, also known as random choice forests, are an ensemble learning approach for classification, regression, and other tasks that works by building a large number of decision trees during training. Random Forest is a learning algorithm that is supervised. Random Forest is used for classification, regression, and a variety of other tasks by constructing a large number of decision trees, with the final tree being the mode or mean prediction of individual trees.Overfitting of decision trees to the training set is corrected using Random Forest. Random Forest regress or regression tasks can be solved with utilising Random Forest. Instead than looking for the most significant feature, it looks for the best trait among the random subsets. The outcomes would be extremely varied, but in general, they would be a superior model.

Support Vector Machine (SVM)

Support Vector Machine is a traditional machine learning approach that may still assist in the categorization of large amounts of data. It may be especially useful in a large data setting for multidomain applications. Each data item in a Support Vector Machine is plotted as a point in n-dimensional space. Each property indicates a different coordinate's value. The hyper-plane, which separates two classes, is used for classification. The coordinates of each observation are the

support vectors. SVM is a frontier that is used to distinguish between two classes: hyper-plane and line.

K-Nearest Neighbor (KNN)

The KNN technique can be used to solve classification and regression problems. It is incredibly simple and quick to construct and analyse, but due to its greedy character, it occasionally overlooks shorter paths that can be easily recognised by human intuition. The k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning approach invented by Evelyn Fix and Joseph Hodges in 1951 and extended by Thomas Cover. It is used in the categorization and regression of data. If the last few steps of traverse are comparable in length to the first few stages, traversal is fair; if it is bigger, better paths exist. Another way is to use an algorithm to determine whether or not a path is good.

Artificial Neural Networks (ANN)

Artificial neurons are a set of linked units or nodes in an ANN that roughly replicate the neurons in a biological brain. Each link may send a signal to other neurons, much like synapses in a human brain. We employ Artificial Neural Networks to model exceedingly complex non-linear functions (ANN). The ANN approach is a biologically inspired analytical technique. The multi-layer perceptron (MLP) is a typical architecture that uses the back-propagation technique to train. A neural network is a collection of input and output links with a corresponding weight. Adjusting the weights is the most important step in anticipating the proper class label of input through iterative learning. ANN is commonly employed in classification and prediction problems because of its great noise tolerance and ability to classify previously undiscovered patterns.

Extra Trees

Another sort of Bagging model is extra trees. The training sample dataset is used to create random trees. The concept allows for the creation of a large number of tree sets as well as a random state. The sklearn.ensemble package contains the Extratree Classifier module.

AdaBoost

Boosting is a process that involves running a series of machine learning models to try to fix the problems that the preceding model had. The output model makes predictions based on each submodel's accuracy. In a dataset, AdaBoost performs weighted instances

Conclusion

Only a few technologies employ machine learning techniques. The majority of the tools are made with deterministic methods and methodologies. Deoxyribonucleic acid DNA (deoxyribonucleic acid) is a biological macromolecule composed up of deoxyribonucleic acid (DNA). Its main function is to save information. Due to breakthroughs in sequencing technology, Data on DNA sequences is growing at

an exponential pace right now, ushering the study of DNA sequences has moved into the age of big data. The term "machine learning" refers to the process of also a powerful technique for digesting large volumes of data and learning on its own. Machine learning approaches have been discussed, as well as how they may be applied to improve genome sequencing accuracy. In our review, we also talked about the Mycobacterium Tuberculosis genome sequence. Mycobacterium Tuberculosis is the microorganism that causes tuberculosis. MDR-TB has been identified as a severe hazard that is spreading at an alarming rate. To deal with it, many ways are being taken. According to the findings, many technologies such such as genomic sequencing, prediction systems based on machine learning, and many more volunteer organisations such as are all contributing to the elimination of MDR-TB.

References

1. Álvaro Beltrán-Corbellini 1, Ángel Aledo-Serrano 1*, Rikke S. Møller 2,Eduardo Pérez-Palma 3, Irene García-Morales 1,4, Rafael Toledano 1,5 and Antonio Gil-Nagel, "Landscape for Developmental and Epileptic Encephalopathies", Received: 14 September 2021 Accepted: 27 January 2022 Published: 17 February 2022.
2. Andy J. Liu, MD; Adam M. Staffaroni, PhD; Julio C. Rojas-Martinez, MD, PhD; Nicholas T. Olney, MD; Carolina Alquezar-Burillo, PhD;Peter A. Ljubenkov, MD; Renaud La Joie, PhD; Jamie C. Fong, MS; Joanne Taylor, MS; Anna Karydas, BA; Eliana Marisa Ramos, PhD; Giovanni Coppola, MD; Adam L. Boxer, MD, PhD; Gil D. Rabinovici, MD; Bruce L. Miller, MD; Aimee W. Kao, MD, "Association of Cognitive and Behavioral Features Between Adults With Tuberous Sclerosis and Frontotemporal Dementia".
3. Bongaarts, A. (2021). Molecular features of low-grade developmental brain tumours: Focusing on subependymal giant cell astrocytomas in tuberous sclerosis complex.
4. Kumar-Sinha C, Chinnaiyan AM. Precision oncology in the age of integrative genomics. Nat Biotechnol. 2018;36(1):46-60. doi:10.1038/nbt.4017
5. Chen L, Jiang Y, Wang J. Fetal cardiac rhabdomyoma due to paternal mosaicism of TSC2: A case report. Medicine (Baltimore). 2020;99(35):e21949. doi:10.1097/MD.0000000000021949.
6. Kovesdi, E.; Ripszam, R.;Postyeni, E.; Horvath, E.B.; Kelemen,A.; Fabos, B.; Farkas, V.; Hadzsiev, K.;Sumegi, K.; Magyari, L.; et al. Whole Exome Sequencing in a Series of Patients with a Clinical Diagnosis of Tuberous Sclerosis Not Confirmed by Targeted TSC1/TSC2 Sequencing.Genes 2021, 12, 1401. http://doi.org/,10.3390/genes12091401
7. Ura, H.; Togi, S.; Niida, Y.Targeted ,Double-Stranded cDNA Sequencing-Based Phase Analysis to Identify Compound Heterozygous Mutations and Differential Allelic Expression. Biology 2021, 10, 256. https://doi.org/10.3390/ biology10040256
8. Bob Yirka," Using machine-learning to find mutations in similar genome sequences of cancer samples", 20 July 2021. https://phys.org/news/2021-07-machine-learning-mutations-similar-genome-sequences.html
9. Dongfang Zou,1,† Lin Wang,2,† Jianxiang Liao,1,† Hongdou Xiao,2,† Jing Duan,1 Tongda Zhang,2 Jianbiao Li,2 Zhenzhen Yin,2 Jing Zhou,2 Haisheng Yan,2 Yushan Huang,2 Nianji Zhan,2 Ying Yang,2 Jingyu Ye,2 Fang Chen,2

Shida Zhu,2 Feiqiu Wen3 and Jian Guo," Genome sequencing of 320 Chinese children with epilepsy: a clinical and molecular study", BRAIN 2021: 144; 3623–3634