To quantify and compare the variants in epilepsy-related genes between two different human populations.

1. Background:

Epilepsy is one of the most common and severe neurological genetic disorders worldwide (Forsgren et al., 2005, Wang et al., 2017, Thijs et al., 2019). Over 70 million people have been affected by this complex disease, and infants and elders are the most vulnerable to the disorder. When epilepsy is not treated correctly, it could even cause fatality. Since the treatment is not cost-effective, the fatality rate becomes comparatively higher in low-income and middle-income countries (Thijs et al., 2019). Also, Several studies reported that it is primarily found in the European population because of the acquired mutations (Forsgren et al., 2005). Since it is a genetic disorder, it quickly runs in the family and affects the lineage. Over 30 autosomal rare dominant mutations have been found within families.

Several studies reported the specific genes associated with the disease Epilepsy in humans (Pal et al., 2010). Over 84 genes are responsible for this disorder (Wang et al., 2017). However, the aim of this study is to compare the epilepsy mutations from at least 20 genes present in two different populations and quantify their results. In order to determine the fate of the rare and common alleles present in two different populations, we have formulated the following objectives.

Objective1: To quantify and compare the rare alleles present in the European population and compare that locus with the common alleles in the *#newlysequencedpopulation*.

Objective 2: To examine whether the allele frequency in the *#newlysequencedpopulation* is two times higher than that of the European population.

Objective 3: To perform the fisher test to determine the statistical significance of those variants

2. Methods:

I have chosen the complete genomes of 91 European populations (GBR – Great Britain) from the 1000 genome dataset and 108 newly *#newlysequencedpopulation* genomes from one of the tiny *#islands*.

2.1.Workflow:

To process the whole genome data and extract the genes which are potentially deleterious enough to cause epilepsy, I have created a series of computer programs. The programs were written in Perl Scripting language and Unix AWK commands. The whole genome data of *#Islanders* were given in the binary file format of the VCF file. This whole-genome dataset was initially mapped to the reference genome assembly of hg18, whereas the 1000 Genome data I had was mapped to the hg19 reference genome assembly. However, to compare two different datasets, both have to be mapped to the same reference genome to differentiate their coordinates and to determine the common and rare alleles. Hence, I wrote a set of computer programs which are described below.

2.2.To cross-map the Islanders coordinates from hg18 to hg19 assembly:

First, I chose and created a text file that contained the names of around 20 genes that were potentially reported to be the cause of the Epilepsy disorder. Then, I acquired the complete gene list annotations, including each gene's start and end position in a human genome, from the *GENCODE* portal, which was exclusively for the hg18 assembly. Hence, I gave both the said files as input into *Perl script 1,* to obtain the start and end positions of the epilepsy-related gene (Output 1) (Step 1 in Figure 1). Once I acquired the coordinates for epilepsy-related genes, I passed the same file (Output1) to *Perl script 2* along with the VCF file containing all variants of *#Islanders* genomes which was mapped against hg18. Then, by using *Perl Script 3*, I converted the file type from VCF to GTF by primarily keeping the Epilepsy gene coordinates. Once this was done, I gave this GTF file as input to the assembly converter tool *Liftover.* Using all of the above, I acquired a list of coordinates and variants linked to the Epilepsy genes used in relation to hg19.

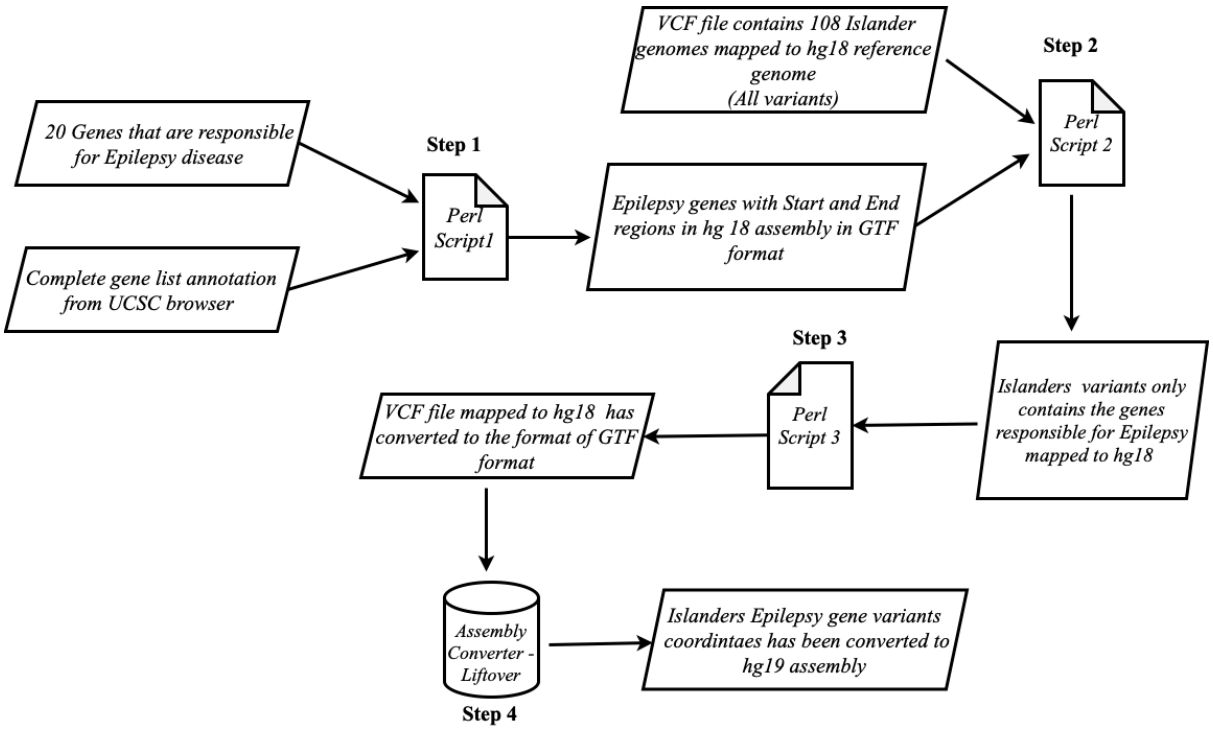

Figure 1: The complete genomic assembly conversion process from hg18 to hg19 in Islander's population

2.3. To obtain the allele frequencies of rare and common alleles:

Pipeline: I developed a pipeline to accomplish four primary tasks: counting alleles, calculating allele frequency, determining rare alleles and determining major alleles (Figure 2). The counting allele's function was used to count all the alleles in each coordinate. For example, it will sum up all the homozygous dominant (AA), homozygous recessive(aa), and heterozygous dominant alleles (Aa). Then, it will calculate the allele frequency of both P (A) and Q (a)

P+Q=1. If the allele frequency of reference/alternate in each position is more than 5%, it will be counted as a major allele. If the allele frequency of reference/alternate in each position is less than 1%, the program considers it as a rare allele.

The pipeline mentioned above was almost similar to both datasets, with minor script changes. Hence, the pipeline will be executed twice with two input files – One for Islanders and the other for Europeans. In the Islanders dataset, the Major and rare allele frequency sites of all the variants related to Epilepsy genes will be obtained as an output from the pipeline. In contrast, the second output file will contain the variants with a deleterious score of more than 20. To determine the deleterious score, the deleterious quantifying method CADD has been used in comparison with the coordinates.

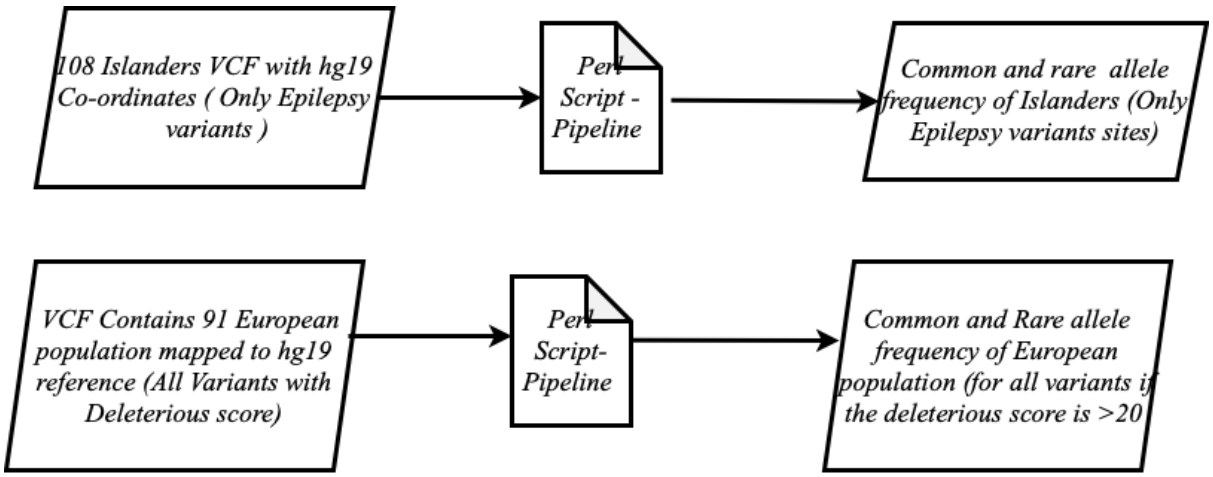

Figure 2: To obtain the allele frequencies of rare and common alleles:

2.4. To compare the allele frequencies of rare and common alleles in both datasets:

In order to obtain the rare variants in the European genome but common in the islander population, a Perl script has been used with input files which I obtained as an output through the pipeline. Also, the same Perl script compares how many variants allele frequencies are two-fold higher in the Islander population compared to the European population.

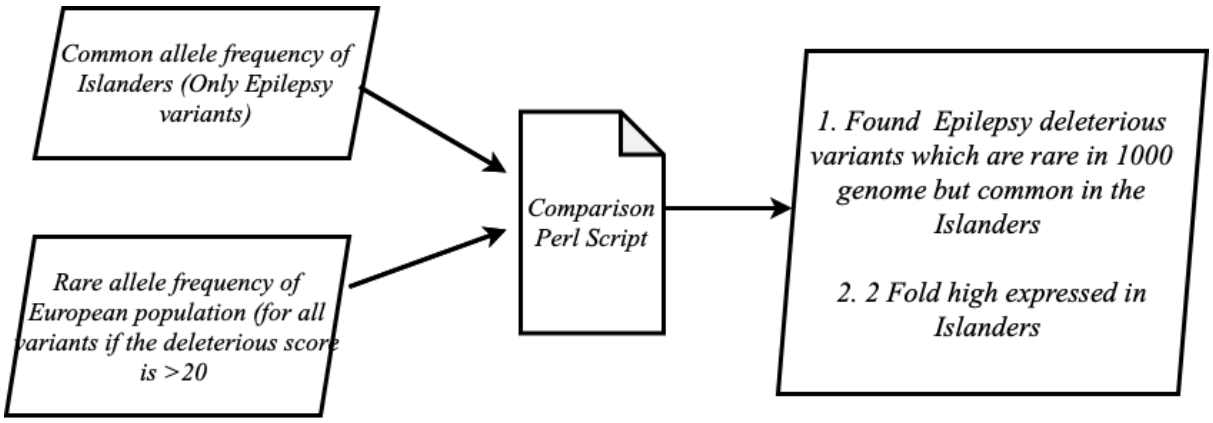

Figure 3: To compare the Allele frequencies from two different dataset

2.5.Statistical test:

The fisher exact test is to observe whether the result we obtained above is statistically significant. In order to attain this, plink will be used.

3. Result:

The results were observed for the Epilepsy related deleterious alleles in European and Islander populations separately. The European deleterious alleles associated with the Epilepsy gene are around 6040 variants. Also, Islander deleterious alleles associated with the Epilepsy gene are observed to be 517 variants.

Since the variants in association with the Epilepsy gene present in each population are very small, there are no variants found. Also, I performed the test by mapping against the alleles and their positions and created the results in the matrix for the major/rare allele frequency between European and Islander populations in all possible combinations (please refer to table 1). Hence, to report the results for the objective1, the rare alleles which are present in the European population are not commonly found in the *#Islanders* population. Given that there are 0 variants for objective 1, objective 2 – Twofold frequency and Objective 3 - the Fisher association test results are also negative. I believe, they need more phenotypes present to compute the results and perform the Fisher test.

	Common alleles - Norfolk Island (108 genomes)	**Rare alleles -** European genomes from 1000 genome dataset
Common alleles- Norfolk Island (108 genomes)	515 variants	**0 variants found**
Rare alleles Frequency - European genomes from 1000 genome dataset	0	29 variants

Table 1: Allele and its Frequency comparison between two populations

4. Discussion:

Based on the results, it can be said that the alleles which are rare and deleterious in the large population (European) are also not found to be common in the small population (Islander).

5. Conclusion:

To conclude, the locus of the rare deleterious alleles present in the European population is also not commonly found in the *#Islanders* population.

6. Reference:

Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. Lancet. 2019 Feb 16;393(10172):689-701. doi: 10.1016/S0140-6736(18)32596-0. Epub 2019 Jan 24. PMID: 30686584.

Wang J, Lin ZJ, Liu L, Xu HQ, Shi YW, Yi YH, He N, Liao WP. Epilepsy-associated genes. Seizure. 2017 Jan;44:11-20. doi: 10.1016/j.seizure.2016.11.030. Epub 2016 Dec 6. PMID: 28007376.

Forsgren L, Beghi E, Oun A, Sillanpää M. The epidemiology of epilepsy in Europe - a systematic review. Eur J Neurol. 2005 Apr;12(4):245-53. doi: 10.1111/j.1468-1331.2004.00992.x. PMID: 15804240.

Pal DK, Pong AW, Chung WK. Genetic evaluation and counseling for epilepsy. Nat Rev Neurol. 2010 Aug;6(8):445-53. doi: 10.1038/nrneurol.2010.92. Epub 2010 Jul 20. PMID: 20647993.

Tools used in this project:

1. GENECODE file for gene annotation from UCSC browser

2. Perl Scripts

Scripts:

108 Genomes Input Vcf file - Hg38
****

1. Downloaded the comprehensive gene annotation file (GTF/GFF) from

http://ftp.ensembl.org/pub/current_gtf/homo_sapiens/GRch38 and input them in to

Perl PerlScript1.pl

2. Used BCFTools to convert the binary file to vcf for file reading

singularity exec /RDS/Q1233/singularity/bcftools.sif bcftools view NorflokIorfloakIsland_completedata.bcf | bcftools view > islanders.vcf all 22 chromosome in one file

Perl PerlScript2.pl is to extract the Islanders gene from All genetic variants

3. perl PerlScript3.pl is to convert from VCF to GTF format

4. Liftover performed for all the Epilepsy variants from Islanders data (hg18 to hg 19) https://grch37.ensembl.org/Homo_sapiens/Tools/AssemblyConverter?db=core/ -

5. Perl pipeline_for_islanders.pl – The pipeline calculates allele count, allele frequency, rare, and major allele frequency

7. 1000 genome data - Extracted Europeans with Cscore for all 22 chromosome – The pipeline was similar to the above with minor changes.

perl Pipeline_All_Chr_for_Europeans.pl (All chromosome – hg19+extract the highly deleterious sites-15)