Biology & Physiology:
Exome Data Analysis and Primer Design

Kim Eileen Roggenbuck

October 23, 2020

In the following I will indicate the variants left after each step with a number enclosed in parentheses at the end of each explanation. The filtering process is divided into 2 phases. The first one being the criteria concerning the variant and the second one those affecting the gene.

Firstly, the suitable frequency for the disease (dbSNP database) is filtered out. By excluding the more common variants first, the number of variants is decreased substantially, leaving only those that match the fitting frequency (variants left by other criteria would have to be discarded anyway). Since 1:40.000 people have Leigh Syndrome, filtering out all variants with a frequency higher than 1% (= 0.01) is appropriate (2817). Next, I added one additional step: knowing the disease is autosomal recessive we can conclude that all variants located on an X or Y chromosome can be deleted from the dataset as well (2755). In the third filtering step the data is filtered depending the location of the variant. The variant of interest is affecting the built of a protein, meaning it affects the splice site. This means that variants are kept if they are exonic, exonic- splice and intronic- splice. If one of these locations is included in a variant, it is always kept (1494). The fourth important criterion is that the protein created from the gene is affected by the given variant, making it nonsynonymous. However, there are multiple reasons that can explain why in many cases the column showing the effect on the protein is empty. One reason is that in the case of a deletion or insertion the protein is not changed but exchanged for another protein. This leaves the column empty even though the protein is majorly affected. This is the reason why the criterion was stated in the code as either the column not being empty, or in case of an insertion/deletion the given column of the data frame is disregarded (862).

At this point all criteria regarding the variant are accounted for, leaving three more involving the gene. Leigh syndrome is an autosomal recessive disease, thus only affecting the phenotype if at least two alleles are mutated. This can mean the gene is homozygous (both mutated the same way) of heterozygous (both mutated, but in different ways), therefore variants with heterozygosity, but only one mutated allele, are disregarded in this step (172). Lastly, the protein function this variant impairs must match the description of the phenotype. In Leigh Syndrome the mitochondria is affected, thus only variants on genes with a function involving the mitochondria are kept, leaving SACS, SARM1, SLC25A46, SYNE2 and THEM4. Three of those seem more likely to be responsible for Leigh Syndrome, when looking at their function: SARM1 is connected to inhibited mitochondrial respiration. THEM4 encodes a protein inhibiting phosphorylation and SLC25A46 mutations result in optic atrophy, a condition that many patients with Leigh Syndrome develop as well. Since optic atrophy is one of the many symptoms of Leigh Syndrome, it is chosen as the most likely gene. The variant is in position 880 (AG), so around this position primers are found. The settings were not changed, except the product size range, which is set to 450-550, as suggested in the tutorial. The sequence that is used is the interval (300-1300). Primer3Plus generated the following left primer: gtccagggagtcacacttgg and right primer: atgtgaaggcggtgcaaaac. The left primer is then checked in BLAST and it is found that all results with a perfect coverage (of 100%) correspond to the gene picked:

$$SLC25A46.$$

This means the chosen primers are perfectly usable for this case.