Title

• Introduction (½ page)

The focus of this research is on the gene DCAF17 which encodes the same named protein DCAF17. This protein has full name DDB1 and CUL4 associated factor 17 and is also known as C2orf37. The reason this gene is selected is the fact that, if some certain mutation happens in this gene, a rare regressive autosomal disease, Woodhouse–Sakati syndrome, occurs in patients. Woodhouse–Sakati syndrome is an endocrine system related disease that is first diagnosed in 1983. The effects of this disease on patients include Hypogonadism, diabetes mellitus, alopecia, mental retardation and electrocardiographic abnormalities. To this date, it has been reported in Eastern Europe, Middle East and India.

In this research, DCAF17 gene's genomic structure was investigated. By looking at similar genes closely related to humans, important regions where a possible mutation would be pathologic were tried to be estimated. The result then compared to an online database's patient genomes.

• Results (2 pages)

Firstly, an amino acid sample of DCAF17 was obtained from NCBI. Then, this similar amino acid sequences were found using blastp on ncbi's site. 794 number of samples were found, while the aim was finding 1000 sequences. From ncbi's site, some additional information was gathered, such as the fact that DCAF17 have only one protein domain, and for humans it is ranged from 31st to 505th amino acid, while having 520 amino acids. The existence of only one protein domain were further tested by using hmmer and cdvist. By using the obtained amino acid sequences, using MUSCLE on MEGA application, a multiple sequence assignment was created. After observing some big gaps, a brief investigation was done on the assignment, from which some sequences were found out to be responsible to a big number of gaps (the threshold was established such that if only one or two sequences were present in one column, they would be responsible for this gap, and if a sequence were responsible for more than 40 such gaps, it would be taken out from the data). The main reason

for doing this was decreasing workload for computer as much as possible. After doing this, the same phenomena were observed, so the steps are repeated. After the second iteration only one sequence above the threshold was discovered, and because its affect would not be much, it was left. By doing these, 9 sequences were left out and the length of the alignment went down from 1673 to 1400.

After creating the multiple sequence assignment, a phylogenic tree was constructed. Maximum likelihood method was used while creating the tree. The clade where homo sapiens were in was selected. The sequences in this clade were taken separately for future analysis.

For further information on this gene, NCBI site's orf finder was used on this gene's reference nucleotide sequence. The reason for that is to be able to learn what a nucleotide mutation would change amino acid sequence. Information about whether they are on positive or negative strand, starting position, ending position were noted to use in future analysis.

In order to gain data on humans, ClinVar database of NCBI was used and a total of 155 gene entries where a mutation happened on DCAF17 gene were collected. From their HGVS code, the location of their mutation was determined. These locations were compared with the ORFs depending on the disease's severity. For genes that are marked with pathogenic and disease's symptoms, 8 out of 8 had their mutation in ORFs. For genes that are marked with only disease's symptoms, (regardless of being pathogenic) this ratio was 125 out of 139. And finally, for all genes 133 out of 152 had a mutation in ORFs.

• Figures (min 4) with legends

• Discussion (½ page)

• Materials and Methods (½-1 page)

- References