**Comparative Genomic Studies of ORFan genes in Mammalian Genomes**

|  |
| --- |
| Purpose |
| Investigate the function/s, if any, of the ORFan gene sequences identified in previous studies from the literature, with a special focus of Clamp *at el* study. However, the initial goal is to find out all the orphan genes of the human genome. |

|  |
| --- |
| Method |
| Human Genome GRCh38.p7 was used for the analysis. Two approaches were followed to find orphan genes: pre-filter(blue) and without filtering(green).  Putative Genes  Retroposons/  Pseudogenes  Pseudogenes  Ortholog with  Dog  Pseudogenes  Ortholog with  Mouse  Pseudogenes  Sequence Unavailable  (non-coding genes)  Pseudogenes  Candidate Orphan Genes  ORFan Finder tool  Ensembl BioMart  Orphan Genes   |  | | --- | | NOTE: This analysis was completely conducted only for chromosome 1 of the human genome. Once this approach is approved, it can be repeated for the other chromosomes. |   Ensembl BioMart[[1]](#footnote-1) online tool was used to retrieve all the genes along their protein sequences. Following parameters were used in the BioMart search:  *Dataset*: Homo sapiens genes (GRCh38.p7)  *Filters*: Chromosome: 1  *Attributes*: Peptide, Ensembl Gene ID, Ensembl Transcript ID  Approach 1 – Direct Search (shows in green colour)  Human chromosome ‘1’ consists of 5,194 genes with a total of 17,296 exons. Out of that, 9,439 transcripts lack of protein sequences; therefore, they were removed from the dataset and rest of the protein sequences were input to the *ORFanFinder*. No filtering was applied.  Approach 2 – Pre-filtering (shows in blue colour)  A bioinformatics pipeline was developed using R language with *BiomaRt* package to filter genes. *BiomaRt* is the official filtering tool to query ensemble databases. There are four main steps in the filtering process:   1. Removal of Retrotransposons/Pseudogenes 2. Removal of orthologous genes with Dog 3. Removal of orthologous genes with Mouse 4. Removal of genes that absent the protein sequence (i.e. non-coding proteins)   As the final step, the candidate genes were submitted separately to *ORFanFinder* tool[[2]](#footnote-2) to look for orphan genes. Following parameters were used for the analysis:  *NCBI Taxonomy ID*: 9606  *Program*: blastp  *Use Premade Database*: nr  *E-value*: 1e-3 |

|  |
| --- |
| Results |
| While running the pipeline in the second approach, it was found that 897 genes in the human chromosome 1 was annotated its gene type as “processed pseudogenes” (retrotransposons), therefore, these genes were excluded from the candidate gene list. Further, 1,874 genes that are orthologous with Dog and 2,047 genes that are orthologous with Mouse. Following Venn diagram shows the distribution of orthologues genes among Human, Dog and Mouse.    Human genes that are orthologous to Dog and Mouse were excluded from the candidate list which could narrow down to 2,271 candidate genes. Further, 2,053 genes were had no protein sequences. Those genes were randomly selected and manually searched in the *ensembl* database. All of them were non-protein coding genes (RNA forming), therefore, they were excluded from the candidate list, narrow downed to 218 genes.    Two candidate gene list from two approaches were separately queried against the *ORFanFinder* tool and their individual results are illustrated below:   |  |  | | --- | --- | | Direct Search Approach | Pre-filtered Approach | | ../../../../Desktop/Screen%20Shot%202016-11-29%20at%2010.42. | ../../../../Desktop/Screen%20Shot%202016-11-29%20at%2010.42. |   Direct Search approach shows 199 strict orphan genes (genome-specific genes) while pre-filtered approach (more accurate approach) shows only 10 strict orphan genes. However, surprisingly, *ORFanFinder* did not show any species-specific orphan genes in their results.  Same pipeline was applied to chr 1 – 22 and chr X of Human Genome and the statistics are illustrated in the table below:   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Chr | Human  Genes | Retrotranspon/ Psuedogenes | Mouse  Orthologs | Dog  Orthologs | Filtered candidate | Candidates with protein sequence | | 1 | 5194 | 897 | 1874 | 2047 | 2271 | 218 | | 2 | 3971 | 751 | 1194 | 1384 | 1858 | 104 | | 3 | 3010 | 616 | 1030 | 1139 | 1271 | 77 | | 4 | 2505 | 561 | 709 | 804 | 1159 | 55 | | 5 | 2868 | 625 | 805 | 917 | 1364 | 99 | | 6 | 2863 | 639 | 941 | 1012 | 1232 | 136 | | 7 | 2867 | 577 | 872 | 979 | 1303 | 109 | | 8 | 2353 | 468 | 625 | 698 | 1166 | 37 | | 9 | 2242 | 461 | 725 | 796 | 987 | 66 | | 10 | 2204 | 422 | 691 | 773 | 1026 | 54 | | 11 | 3235 | 473 | 1180 | 1271 | 1465 | 101 | | 12 | 2940 | 490 | 967 | 1060 | 1441 | 113 | | 13 | 1304 | 295 | 307 | 345 | 679 | 34 | | 14 | 2224 | 324 | 655 | 708 | 1155 | 180 | | 15 | 2152 | 287 | 584 | 616 | 1257 | 70 | | 16 | 2511 | 252 | 789 | 817 | 1566 | 220 | | 17 | 2995 | 310 | 1050 | 1185 | 1558 | 151 | | 18 | 1170 | 207 | 259 | 282 | 674 | 18 | | 19 | 2926 | 263 | 1228 | 1232 | 1698 | 484 | | 20 | 1386 | 198 | 514 | 564 | 626 | 39 | | 21 | 835 | 143 | 198 | 225 | 471 | 42 | | 22 | 1318 | 156 | 431 | 442 | 695 | 59 | | X | 2359 | 723 | 741 | 780 | 857 | 147 | |

|  |
| --- |
| Future work |
| Filtered candidate genes from the two approaches can be input to *ORFanFinder* to investigate orphan genes in other chromosomes. Since that process is very time consuming, it is better to confirm this approach is valid in order to continue search for other chromosomes. |

1. http://asia.ensembl.org/biomart/martview [↑](#footnote-ref-1)
2. http://cys.bios.niu.edu/orfanfinder/index.php [↑](#footnote-ref-2)