**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.  Putative Genes  Retroposons/  Pseudogenes  Pseudogenes  Ortholog with  Dog  Pseudogenes  Ortholog with  Mouse  Pseudogenes  Sequence Unavailable  (non-coding genes)  Pseudogenes  Candidate Orphan Genes  Ensembl BioMart  Paralog with Human  Pseudogenes  Pfam Domains  Pseudogenes  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 were six 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 paralogous genes within Human 5. Removal of known pfam genes 6. Removal of genes that absent the protein sequence (i.e. non-coding genes) |

|  |
| --- |
| Results |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Chr** | **# Genes** | **Retrotranspon** | **Mouse**  **Ortholog** | **Dog**  **Ortholog** | **Human**  **Paralog** | **Pfam** | **All ORFan** | **Protein Coding ORFan** |
| 1 | 5194 | 897 | 1858 | 2070 | 2020 | 1886 | 1731 | 44 |
| 2 | 3971 | 751 | 1205 | 1375 | 1351 | 1207 | 1503 | 33 |
| 3 | 3010 | 616 | 1029 | 1140 | 1114 | 995 | 977 | 13 |
| 4 | 2505 | 561 | 706 | 789 | 797 | 701 | 935 | 11 |
| 5 | 2868 | 625 | 801 | 936 | 916 | 805 | 1112 | 27 |
| 6 | 2863 | 639 | 930 | 1032 | 1024 | 967 | 932 | 29 |
| 7 | 2867 | 577 | 865 | 982 | 989 | 893 | 1062 | 43 |
| 8 | 2353 | 468 | 621 | 711 | 709 | 629 | 995 | 11 |
| 9 | 2242 | 461 | 719 | 787 | 742 | 719 | 815 | 25 |
| 10 | 2204 | 422 | 682 | 773 | 764 | 679 | 835 | 14 |
| 11 | 3235 | 473 | 1182 | 1279 | 1167 | 1201 | 1247 | 26 |
| 12 | 2940 | 490 | 969 | 1061 | 1031 | 972 | 1170 | 26 |
| 13 | 1304 | 295 | 309 | 350 | 372 | 300 | 547 | 10 |
| 14 | 2224 | 324 | 637 | 733 | 751 | 683 | 959 | 109 |
| 15 | 2152 | 287 | 596 | 627 | 727 | 549 | 997 | 27 |
| 16 | 2511 | 252 | 789 | 856 | 761 | 793 | 1237 | 18 |
| 17 | 2995 | 310 | 1067 | 1178 | 1078 | 1100 | 1280 | 37 |
| 18 | 1170 | 207 | 255 | 288 | 317 | 248 | 580 | 6 |
| 19 | 2926 | 263 | 1179 | 1230 | 1245 | 1361 | 1095 | 25 |
| 20 | 1386 | 198 | 509 | 567 | 539 | 499 | 499 | 8 |
| 21 | 835 | 143 | 194 | 237 | 223 | 206 | 405 | 14 |
| 22 | 1318 | 156 | 426 | 438 | 436 | 455 | 589 | 18 |
| X | 2359 | 723 | 736 | 776 | 954 | 773 | 553 | 10 |
|  | 57432 | 10138 | 18264 | 20215 | 20027 | 18621 | 22055 | **584** |

584 orphan genes were filtered out by the bioinformatics pipeline. Yin and Fischer (2008) mentioned that ORFan genes are short sequences. Many candidate ORFan genes found are short and contain repetitive sequences(duplications) as well.

Reference:

Yin, Y. and Fischer, D. (2008). Identification and investigation of ORFans in the viral world. *BMC Genomics*, 9(1), p.24.