GlobalHumanTree.DescriptiveStat.R

popadin

2019-11-07

```
################################
##################
################################

rm(list=ls(all=TRUE))

############ Syn mut
Mut = read.table("../../Body/3Results/fulltreeCodons.csv", header = TRUE, sep = ';'); # "../../Body/3Re
table(Mut$note)
```

```
##
##       gaps non-coding     normal
##        787     261486     315003
```

```
# gaps non-coding     normal
# 787     261486     315003
# Why gaps? they should be in coding or non-coding? strange names

Pc= Mut[Mut$note == 'normal',] # filter out everything except protein-coding mutations:
AncAa = data.frame(table(Pc$ancestral_aa)) #
AncAa = AncAa[order(-AncAa$Freq),]
AncAa
```

```
##    Var1  Freq
## 11  Leu 33959
## 16  Ser 30725
## 18  Thr 24688
## 15  Pro 22731
## 1   Ala 20650
## 20  Tyr 16987
## 8   Gly 16294
## 17 Stop 16120
## 9   His 16095
## 21  Val 15195
## 10  Ile 13964
## 3   Asn 12502
## 12  Lys 10685
## 13  Met 10543
## 14  Phe  9725
## 6   Gln  8739
## 4   Asp  7838
## 2   Arg  7823
## 19  Trp  7397
## 7   Glu  7101
## 5   Cys  5242
```

```
VectorOfAa=AncAa$Var1; VectorOfAa
```

```
##  [1] Leu  Ser  Thr  Pro  Ala  Tyr  Gly  Stop His  Val  Ile  Asn  Lys  Met
```

```
## [15] Phe  Gln  Asp  Arg  Trp  Glu  Cys
## 21 Levels: Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe ... Val
```

```
table(Pc$derived_aa) # why  Ambiguous, Asn/Asp, Gln/Glu, Leu/Ile? => because of ambiguous derived AA. W
```

```
##
##       Ala Ambiguous       Arg       Asn   Asn/Asp       Asp       Cys
##     20209      5343      7651     12507       104      7857      5146
##       Gln   Gln/Glu       Glu       Gly       His       Ile       Leu
##      8944         3      7088     15794     16514     14064     33631
##   Leu/Ile       Lys       Met       Phe       Pro       Ser      Stop
##        66      7374     10468      9098     23518     30264     15805
##       Thr       Trp       Tyr       Val
##     25240      7197     16654     14464
```

```
PcQuestion= Pc[Pc$derived_aa == 'Ambiguous',]
head(PcQuestion,10)
```

```
##     first      second position ref_pos ancestor descendant nuc_ref_in_ali
## 4   28409  AM_PR_0048     7971    7774    ccGtc      ccRtc              G
## 6   28409  AM_PR_0048    10537   10305    taAcc      taMcc              A
## 16  17837  EU_IT_0648     3647    3487    ccTaa      ccWaa              T
## 396   208  AF_ET_0044     8223    8026    aaGcc      aaRcc              G
## 666 23041  EU_ES_0678    11453   11219    ccTag      ccYag              T
## 748  1664 XX_XX_10750     9407    9179    gtAag      gtRag              A
## 931 10076  EU_ES_0683     7599    7406    ccTac      ccYac              T
## 932 10076  EU_ES_0683     7959    7762    agGaa      agRaa              G
## 934 10076  EU_ES_0683    10537   10305    taAcc      taMcc              A
## 935 10076  EU_ES_0683    13711   13469    ctAgc      ctRgc              A
##     gene_info gene_start pos_in_codon     synonymous ancestral_aa
## 4   mRNA_COX2       7585            2 non-synonymous          Arg
## 6    mRNA_ND3      10058            2 non-synonymous          Asn
## 16   mRNA_ND1       3306            2 non-synonymous          Leu
## 396 mRNA_COX2       7585            2 non-synonymous          Ser
## 666  mRNA_ND4      10759            1 non-synonymous         Stop
## 748 mRNA_ATP6       8526            2 non-synonymous         Stop
## 931 mRNA_COX1       5903            1 non-synonymous          Tyr
## 932 mRNA_COX2       7585            2 non-synonymous          Gly
## 934  mRNA_ND3      10058            2 non-synonymous          Asn
## 935  mRNA_ND5      12336            1 non-synonymous          Ser
##     derived_aa   note
## 4    Ambiguous normal
## 6    Ambiguous normal
## 16   Ambiguous normal
## 396  Ambiguous normal
## 666  Ambiguous normal
## 748  Ambiguous normal
## 931  Ambiguous normal
## 932  Ambiguous normal
## 934  Ambiguous normal
## 935  Ambiguous normal
```

```
PcQuestion= Pc[Pc$derived_aa == 'Asn/Asp' | Pc$derived_aa == 'Gln/Glu' | Pc$derived_aa == 'Leu/Ile',]
head(PcQuestion,10)
```

```
##       first     second position ref_pos ancestor descendant nuc_ref_in_ali
```

```
## 6756   1860 XX_XX_7147     4213    4053    caCtc      caMtc             C
## 9800  34077 XX_XX_7195    15656   15410    ccCtt      ccMtt             C
## 12709  5670 EU_IT_1120    14134   13892    acAtt      acMtt             A
## 12851 15941 AF_GM_0009    14803   14559    ccGac      ccRac             G
## 15011  5876 EU_IT_0710     9444    9216    ccAat      ccRat             A
## 19093 18441 EU_ES_0686     7428    7235    ccGat      ccRat             G
## 21057 26967 AS_JP_0554    12152   11916    ttCtc      ttMtc             C
## 21060 26967 AS_JP_0554    12748   12507    taGac      taRac             G
## 24493 32541 AF_SL_0005     7967    7770    gaAac      gaRac             A
## 31149 12944 AF_SL_0029    14803   14559    ccGac      ccRac             G
##       gene_info gene_start pos_in_codon     synonymous ancestral_aa
## 6756   mRNA_ND1       3306            1 non-synonymous          Leu
## 9800  mRNA_CYTB      14746            1 non-synonymous          Leu
## 12709  mRNA_ND5      12336            1 non-synonymous          Ile
## 12851  mRNA_ND6      14148            1 non-synonymous          Asp
## 15011 mRNA_COX3       9206            1 non-synonymous          Asn
## 19093 mRNA_COX1       5903            1 non-synonymous          Asp
## 21057  mRNA_ND4      10759            1 non-synonymous          Leu
## 21060  mRNA_ND5      12336            1 non-synonymous          Asp
## 24493 mRNA_COX2       7585            1 non-synonymous          Asn
## 31149  mRNA_ND6      14148            1 non-synonymous          Asp
##       derived_aa   note
## 6756     Leu/Ile normal
## 9800     Leu/Ile normal
## 12709    Leu/Ile normal
## 12851    Asn/Asp normal
## 15011    Asn/Asp normal
## 19093    Asn/Asp normal
## 21057    Leu/Ile normal
## 21060    Asn/Asp normal
## 24493    Asn/Asp normal
## 31149    Asn/Asp normal
```

```
## why problematic only in the descendant???? each branch has ancestor and descendant?

## filter out all problems:
nrow(Pc)
```

```
## [1] 315003
```

```
PcGold = Pc[Pc$ancestral_aa %in% VectorOfAa & Pc$derived_aa %in% VectorOfAa,]
nrow(PcGold)
```

```
## [1] 309487
```

```
## syn / nons => too many nons
table(PcGold$synonymous)
```

```
##
## non-synonymous     synonymous
##         200702          108785
```

```
## COX1, ND4, ND4L looks good; - others - not really
ToStop = PcGold[PcGold$ancestral_aa != 'Stop' & PcGold$derived_aa == 'Stop',]; ToStop$synonymous = 'ToS
FromStop = PcGold[PcGold$ancestral_aa == 'Stop' & PcGold$derived_aa != 'Stop',]; FromStop$synonymous =
NoStop = PcGold[PcGold$ancestral_aa != 'Stop' & PcGold$derived_aa != 'Stop',];
PcGold = rbind(NoStop,ToStop,FromStop)
```

```
T1=data.frame(table(PcGold$synonymous, by = PcGold$gene_info))
names(T1)=c('MutType','Gene','Freq')
T1=T1[grep('mRNA',T1$Gene),]; nrow(T1) # 60
```

```
## [1] 60
```

```
## only ND4L and ND3 look good,
Syn = T1[T1$MutType == 'synonymous',]; Syn$Syn = Syn$Freq; Syn = Syn[,grep("Gene|Syn", names(Syn))]
Nons = T1[T1$MutType == 'non-synonymous',]; Nons$Nons = Nons$Freq; Nons = Nons[,grep("Gene|Nons", names
ToStop = T1[T1$MutType == 'ToStop',]; ToStop$ToStop = ToStop$Freq; ToStop = ToStop[,grep("Gene|ToStop",
FromStop = T1[T1$MutType == 'FromStop',]; FromStop$FromStop = FromStop$Freq; FromStop = FromStop[,grep(
MutTypes = merge(Syn,Nons, by = 'Gene', all = TRUE);
MutTypes = merge(MutTypes,ToStop, by = 'Gene', all = TRUE);
MutTypes = merge(MutTypes,FromStop, by = 'Gene', all = TRUE);
MutTypes
```

```
##              Gene   Syn  Nons ToStop FromStop
## 1       mRNA_ATP6  6586 11754   1542     1481
## 2       mRNA_ATP8  1093  3586    158      160
## 3  mRNA_ATP8&ATP6   417  1062    142      155
## 4       mRNA_COX1 26362  7919    152      177
## 5       mRNA_COX2   864 15026    555      543
## 6       mRNA_COX3  3766 13656    797      819
## 7       mRNA_CYTB 12154 21877    749      762
## 8        mRNA_ND1 15847 10343     32       33
## 9        mRNA_ND2  4443 24669    938      891
## 10       mRNA_ND3  6221  3406      0        0
## 11       mRNA_ND4  8765 18874   2211     2268
## 12      mRNA_ND4L  4796   940      0        0
## 13  mRNA_ND4L&ND4     4     0      2        2
## 14       mRNA_ND5  8324 37040   2127     2200
## 15       mRNA_ND6  3212 10706    469      479
```