exon_counting

August 21, 2022

1 Counting exons in LncRNAs Vs Annnotated protein-coding genes in a GFF

[1]:
```
import pandas as pd
import numpy as np
```

1.0.1 There will be a better way to do this but it works for counting exons in LncRNA gtf

- Repeat command below incrementing exon_number "1" to exon_number "2" etc and record

cat A1163_ML_UX_LncRNA_T_TPM_CUTOFF_6_8.gtf|grep 'exon_number "1" '|wc -l * Then in excel compute how many LncRNAs contain how many exons * Slightly more convoluted is counting how many exons in protein coding genes from a GFF and this is shown below * Then can compare the number of single exon containing transcripts to multi-exon transcripts for LncRNAs versus annotated protein-coding genes using Chisquare test

[2]:
```
#counting exons in protein-coding genes in a GFF

df=pd.read_csv("Aspergillus_fumigatusa1163.ASM15014v1.53.
 ↪gff3",sep=("\t"),comment="#", header=None)

dfmRNA=df.copy()[df.iloc[:,2]=="mRNA"]
protcodtranscripts=dfmRNA.iloc[:,8].str.split("ID=transcript:", expand=True)[1].
 ↪str.split(";", expand=True)[0].tolist()
df["transcript"]=df.iloc[:,8].str.split("Parent=transcript:", expand=True)[1].
 ↪str.split(";", expand=True)[0]
dfprotcod=df.copy()[df.transcript.apply(lambda x:x in protcodtranscripts)]

dfprotcod=dfprotcod[dfprotcod.iloc[:,2]=="exon"]
numberexons=dfprotcod.groupby(by="transcript").agg("count")[8].tolist()
unique_labels, unique_counts = np.unique(numberexons, return_counts=True)
labels_histogram = dict(zip(unique_labels, unique_counts))
```

[3]:
```
labels_histogram
```

[3]:
```
{1: 2043,
 2: 2899,
 3: 2123,
 4: 1339,
 5: 711,
 6: 335,
 7: 210,
 8: 128,
 9: 68,
 10: 32,
 11: 15,
 12: 10,
 13: 7,
 14: 1,
 15: 3,
 18: 1,
 19: 2,
 20: 1,
 26: 1}
```

[4]:
```
len(numberexons)
```

[4]:
```
9929
```

[ ]:
```

```