RNAseqParser

December 2, 2019

[4]:
```
import re
```

[5]:
```
string = '\nensGene\tgeneSymb\tESC.RPKM\tMES.RPKM\tCP.RPKM\tCM.
 ↪RPKM\nENSMUSG00000000134\tTcfe3\t14.92599\t6.080252\t7.205497\t5.
 ↪5972915\nENSMUSG00000000708\tKat2b\t9.379815\t0.37079784\t1.1033436\t5.
 ↪6754346\nENSMUSG00000000948\tSnrpn\t40.668293\t14.529371\t13.403415\t23.
 ↪01873\nENSMUSG00000001054\tRmnd5b\t43.369095\t7.0136724\t14.050683\t11.
 ↪829396\nENSMUSG00000001366\tFbxo9\t7.6720843\t6.9369035\t6.499769\t6.
 ↪778531\nENSMUSG00000001482\tDef8\t24.153797\t15.451096\t15.014166\t13.
 ↪819534\nENSMUSG00000001542\tEll2\t8.156232\t3.5004125\t3.5680292\t2.
 ↪2641196\nENSMUSG00000001627\tIfrd1\t28.733929\t16.701181\t15.508437\t12.
 ↪778727\nENSMUSG00000001642\tAkr1b3\t4.319858\t1.9163351\t1.2716209\t0.
 ↪82428175\nENSMUSG00000001687\tUbl3\t28.78591\t9.088697\t9.046656\t20.
 ↪373514\nENSMUSG00000002227\tMov10\t29.740297\t3.2102342\t6.25411\t9.
 ↪091757\nENSMUSG00000002635\tPdcd2l\t30.69546\t18.50777\t15.635618\t15.
 ↪247209\nENSMUSG00000002660\tClpp\t93.85232\t51.403442\t32.20393\t33.
 ↪370808\nENSMUSG00000002767\tMrpl2\t86.59501\t61.894024\t50.002293\t51.
 ↪35253\nENSMUSG00000002963\tPnkp\t8.918158\t5.5222096\t6.193148\t6.
 ↪496989\nENSMUSG00000002983\tRelb\t7.0391517\t1.501116\t1.7450844\t2.
 ↪5017977\nENSMUSG00000003032\tKlf4\t41.70846\t7.747598\t4.1997404\t6.
 ↪5344357\nENSMUSG00000003662\tCiao1\t15.639003\t11.429388\t9.724962\t11.
 ↪069197\nENSMUSG00000003813\tRad23a\t30.253717\t16.276289\t15.284632\t21.
 ↪372665\nENSMUSG00000004285\tAtp6v1f\t30.517672\t23.897362\t24.671564\t25.
 ↪907063\nENSMUSG00000004568\tArhgef18\t13.561201\t6.151879\t5.004999\t6.
 ↪8743706\nENSMUSG00000004667\tPolr2e\t91.243706\t51.02243\t36.53202\t33.37132'
```

[6]:
```
newlist = []
for i in re.finditer("(ENSMUSG\d+\\t\w+\t\d+\.\d+\\t\d+\.\d+)(\t\d+\.\d+)",␣
 ↪string):
    newlist.append(i.group(2))

newlist
```

[6]:
```
['\t7.205497',
 '\t1.1033436',
 '\t13.403415',
 '\t14.050683',
```

```
 '\t6.499769',
 '\t15.014166',
 '\t3.5680292',
 '\t15.508437',
 '\t1.2716209',
 '\t9.046656',
 '\t6.25411',
 '\t15.635618',
 '\t32.20393',
 '\t50.002293',
 '\t6.193148',
 '\t1.7450844',
 '\t4.1997404',
 '\t9.724962',
 '\t15.284632',
 '\t24.671564',
 '\t5.004999',
 '\t36.53202']
```

[7]:
```
newlist2 = []
for j in re.finditer("\t\w{4,}", string):
    newlist2.append(j.group())
newlist2
```

[7]:
```
['\tgeneSymb',
 '\tTcfe3',
 '\tKat2b',
 '\tSnrpn',
 '\tRmnd5b',
 '\tFbxo9',
 '\tDef8',
 '\tEll2',
 '\tIfrd1',
 '\tAkr1b3',
 '\tUbl3',
 '\tMov10',
 '\tPdcd2l',
 '\tClpp',
 '\tMrpl2',
 '\tPnkp',
 '\tRelb',
 '\tKlf4',
 '\tCiao1',
 '\tRad23a',
 '\tAtp6v1f',
 '\tArhgef18',
 '\tPolr2e']
```

[8]:
```
newlist2[0]
```

[8]:
```
'\tgeneSymb'
```

[9]:
```
def RNAseqParser(input):
    genes = []
    exp_data = []
    output = ""
    for i in re.finditer("(\t)(\w{4,})", string):
        genes.append(i.group(2))
    for j in re.finditer("(ENSMUSG\d+\\t\w+\t\d+\.\d+\\t\d+\.\d+)(\t\d+\.\d+)",␣
↪string):
        exp_data.append(j.group(2))
    del genes[0]
    for k in range(len(genes)):
        output += genes[k] + exp_data[k] + "\n"
    print(output)
```

[10]:
```
RNAseqParser(string)
```

```
Tcfe3   7.205497
Kat2b   1.1033436
Snrpn   13.403415
Rmnd5b  14.050683
Fbxo9   6.499769
Def8    15.014166
Ell2    3.5680292
Ifrd1   15.508437
Akr1b3  1.2716209
Ubl3    9.046656
Mov10   6.25411
Pdcd2l  15.635618
Clpp    32.20393
Mrpl2   50.002293
Pnkp    6.193148
Relb    1.7450844
Klf4    4.1997404
Ciao1   9.724962
Rad23a  15.284632
Atp6v1f 24.671564
Arhgef18        5.004999
Polr2e  36.53202
```