SESSION 4. HUMAN DISEASE

When DNA sequences are toxic

Inherited vs non-inherited genetic diseases

Inherited disease – The Majority are Mendelian diseases (+ Germline or *De novo* mutation)

Non-inherited disease – Somatic mutations

Genetic vs Epigenetic changes

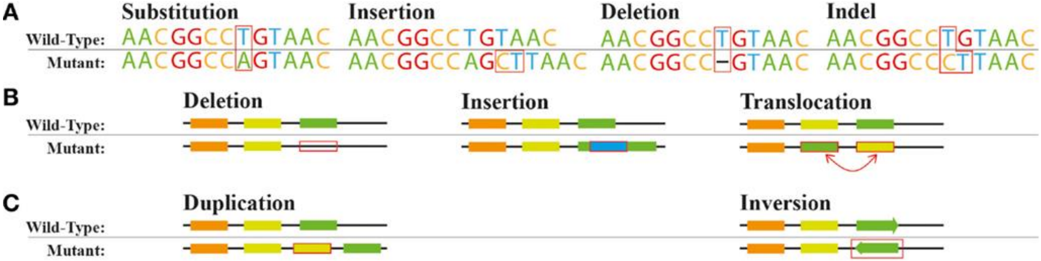
- Genetic change: with DNA change
- Epigenetic change: without DNA change

DNA changes often give rise to diseases.

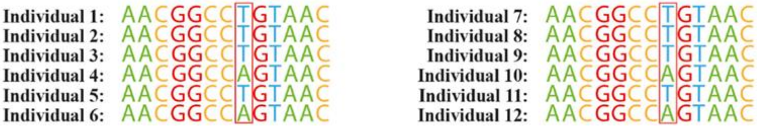
Epigenetic changes could cause diseases

DNA alterations – local and largescale changes

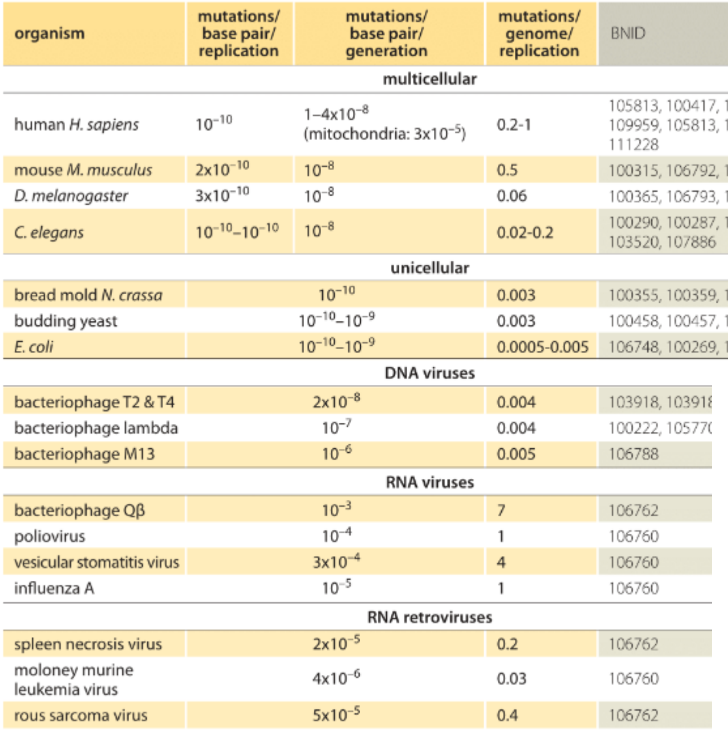
- Local changes
  - Single nucleotide variations (SNVs)
  - Insertion + Deletion (indels)
- Largescale rearrangement
  - Inter/intra chromosomal translocation
  - Inversion
  - Large insertion/deletion
  - Duplication



SNP vs mutation

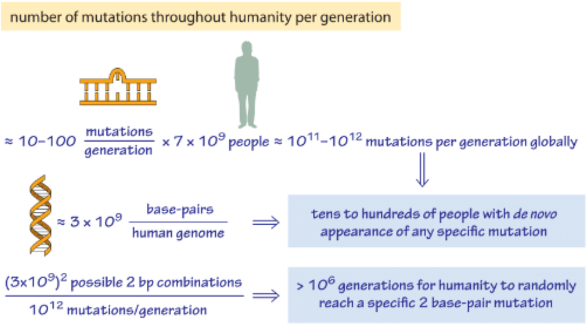

Individual 1: AACGGCCTGTAAC
Individual 2: AACGGCCTGTAAC
Individual 3: AACGGCCTGTAAC
Individual 4: AACGGCCAGTAAC
Individual 5: AACGGCCTGTAAC
Individual 6: AACGGCCAGTAAC
Individual 7: AACGGCCTGTAAC
Individual 8: AACGGCCTGTAAC
Individual 9: AACGGCCTGTAAC
Individual 10: AACGGCCAGTAAC
Individual 11: AACGGCCTGTAAC
Individual 12: AACGGCCAGTAAC

Mutation rate during replication



organism	mutations/ base pair/ replication	mutations/ base pair/ generation	mutations/ genome/ replication	BNID
		multicellular		
human *H. sapiens*	10^{-10}	$1–4x10^{-8}$ (mitochondria: $3x10^{-5}$)	0.2-1	105813, 100417, 105095, 108040, 109959, 105813, 110292, 111227, 111228
mouse *M. musculus*	$2x10^{-10}$	10^{-8}	0.5	100315, 106792, 100320
D. melanogaster	$3x10^{-10}$	10^{-8}	0.06	100365, 106793, 100370
C. elegans	$10^{-10}–10^{-10}$	10^{-8}	0.02-0.2	100290, 100287, 109959, 103520, 107886
		unicellular		
bread mold *N. crassa*	10^{-10}		0.003	100355, 100359, 106747
budding yeast	$10^{-10}–10^{-9}$		0.003	100458, 100457, 109959, 110018
E. coli	$10^{-10}–10^{-9}$		0.0005-0.005	106748, 100269, 100263
		DNA viruses		
bacteriophage T2 & T4	$2x10^{-8}$		0.004	103918, 10391
bacteriophage lambda	10^{-7}		0.004	100222, 1057
bacteriophage M13	10^{-6}		0.005	106788
		RNA viruses		
bacteriophage Qβ	10^{-3}		7	106762
poliovirus	10^{-4}		1	106760
vesicular stomatitis virus	$3x10^{-4}$		4	106760
influenza A	10^{-5}		1	106760
		RNA retroviruses		
spleen necrosis virus	$2x10^{-5}$		0.2	106762
moloney murine leukemia virus	$4x10^{-6}$		0.03	106760
rous sarcoma virus	$5x10^{-5}$		0.4	106762

- □ DNA polymerase, its proof-reading, and base-pairing affinity give a 10^{-8} error per bp
- □ DNA repair enzymes fix 99% of the errors → 10^{-10} error per bp



number of mutations throughout humanity per generation

$\approx 10–100 \frac{\text{mutations}}{\text{generation}} \times 7 \times 10^9$ people $\approx 10^{11}–10^{12}$ mutations per generation globally

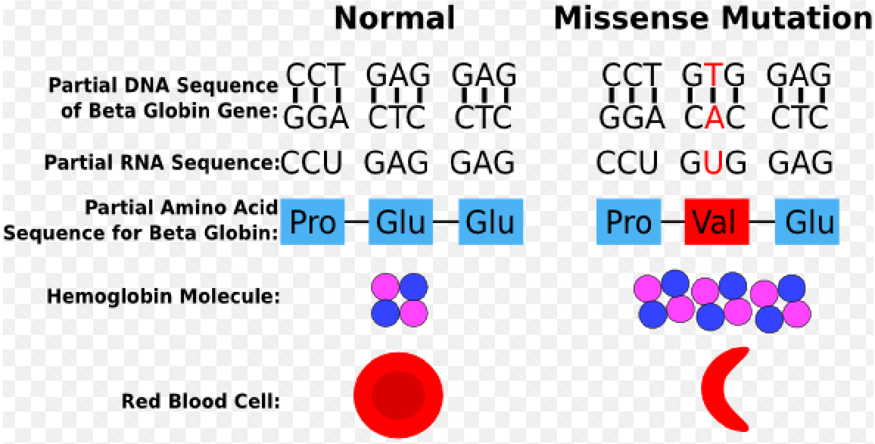
$\approx 3 \times 10^9 \frac{\text{base-pairs}}{\text{human genome}} \Longrightarrow$ tens to hundreds of people with *de novo* appearance of any specific mutation

$\frac{(3x10^9)^2 \text{ possible 2 bp combinations}}{10^{12} \text{ mutations/generation}} \Longrightarrow$ > 10^6 generations for humanity to randomly reach a specific 2 base-pair mutation

A classic example of point mutation is *sickle-cell anemia*

□ Point mutation (GAG:Glu→GUG:Val) affects one of the polypeptide chains of beta-globlin.

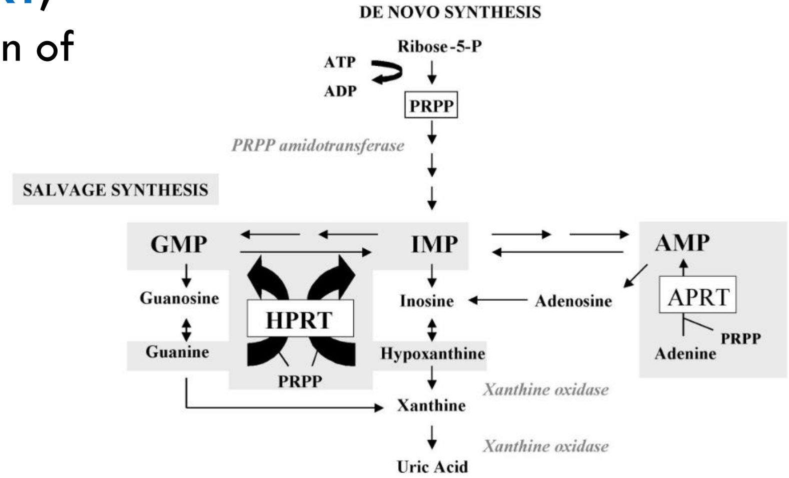
□ 1% of the population in Western Africa have the sickle-cell gene in hetero-type that protects against malaria.



	Normal	Missense Mutation
Partial DNA Sequence of Beta Globin Gene:	CCT GAG GAG GGA CTC CTC	CCT GTG GAG GGA CAC CTC
Partial RNA Sequence:	CCU GAG GAG	CCU GUG GAG
Partial Amino Acid Sequence for Beta Globin:	Pro—Glu—Glu	Pro—Val—Glu
Hemoglobin Molecule:		
Red Blood Cell:		

Rare disease: Lesch- Nyhan syndrome (LNS)

- A disease with a neurological symptoms
- Resulted by mutations in the gene encoding **HPRT**, leading to accumulation of Uric acid in body.

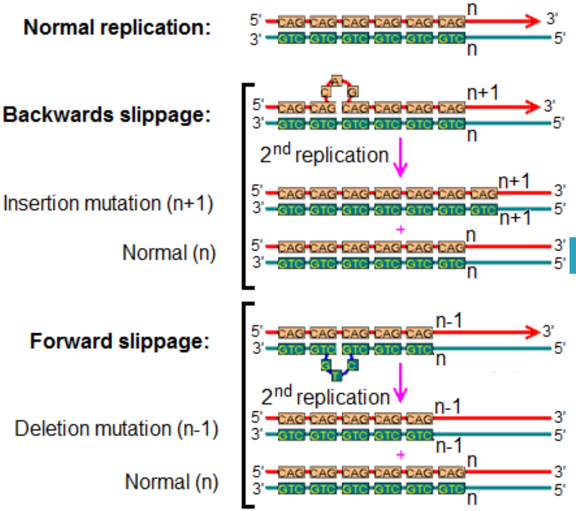


https://youtu.be/DajyHjo1-Ns

Huntington's disease

- The disease was first discovered in 1841 by Charles Oscar Water
- The responsible gene, huntingtin, was first reported in1993
- Unfortunately, we still don't know the function of the protein in cell.
- The protein includes a region of repeated glutamine (polyQ region)
- Normally, less than 26 Qs but in rare cases, greater than 36, giving rise to Hungtington's disease.
- The polyQ stimulates the protein aggregation.
- Caused by expansion of short repeats (CAG)
- Replication slippage may cause the expansion of the repeats.

Huntington’s disease



Normal replication:

Backwards slippage:

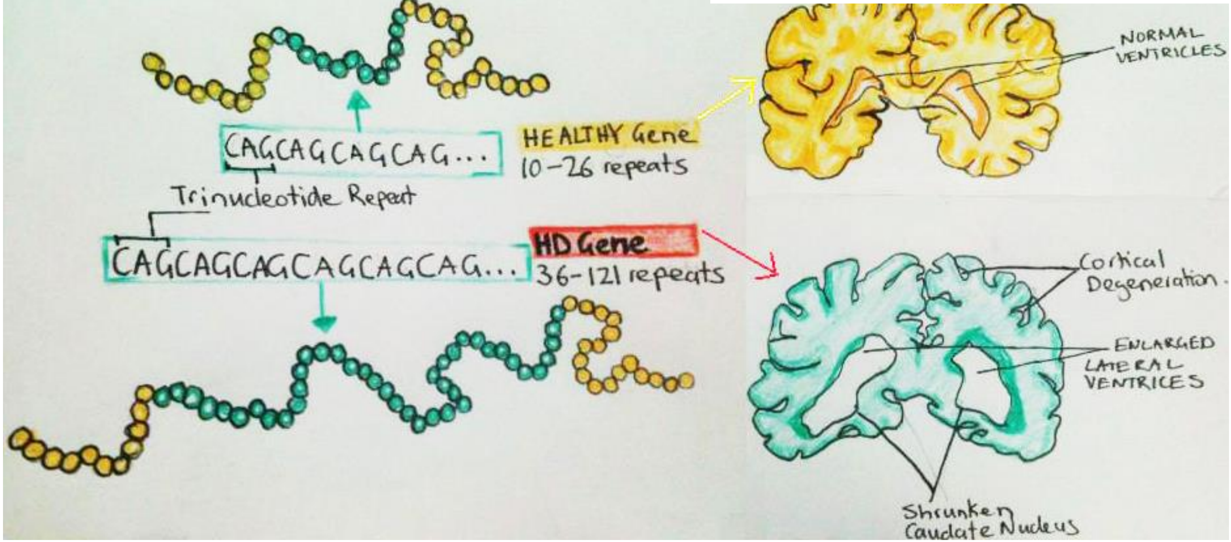
Insertion mutation (n+1)

Normal (n)

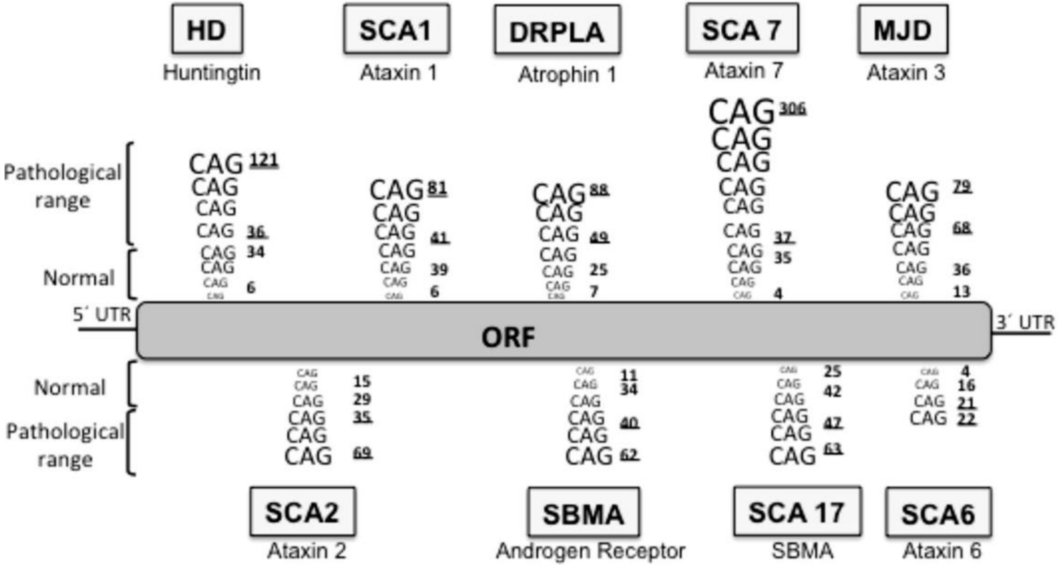
Forward slippage:

Deletion mutation (n-1)

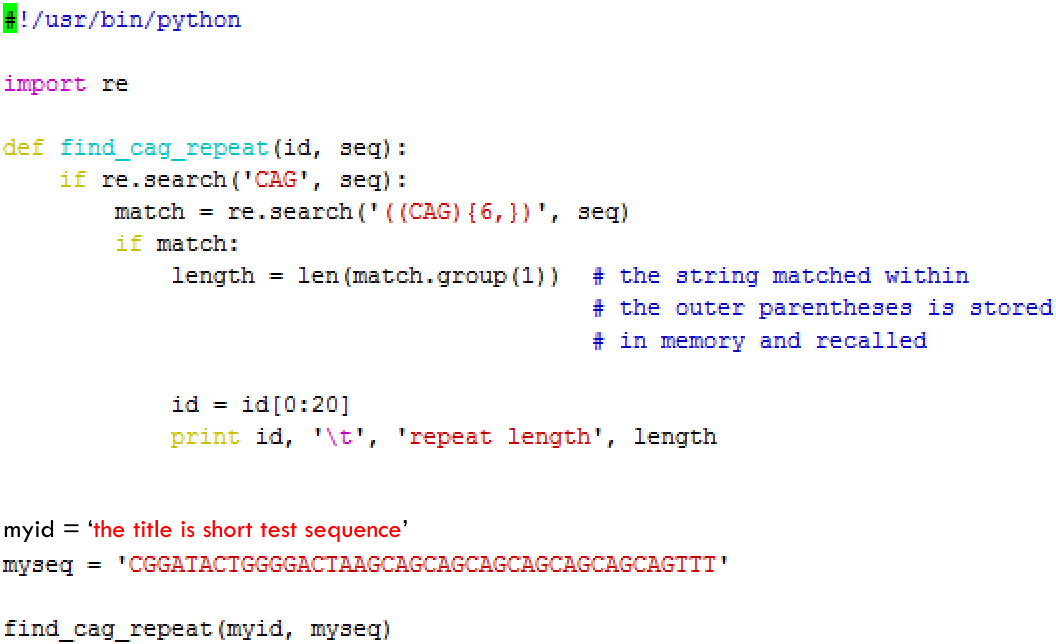
Normal (n)



Other CAG repeat-related genes



find_cag_short.py

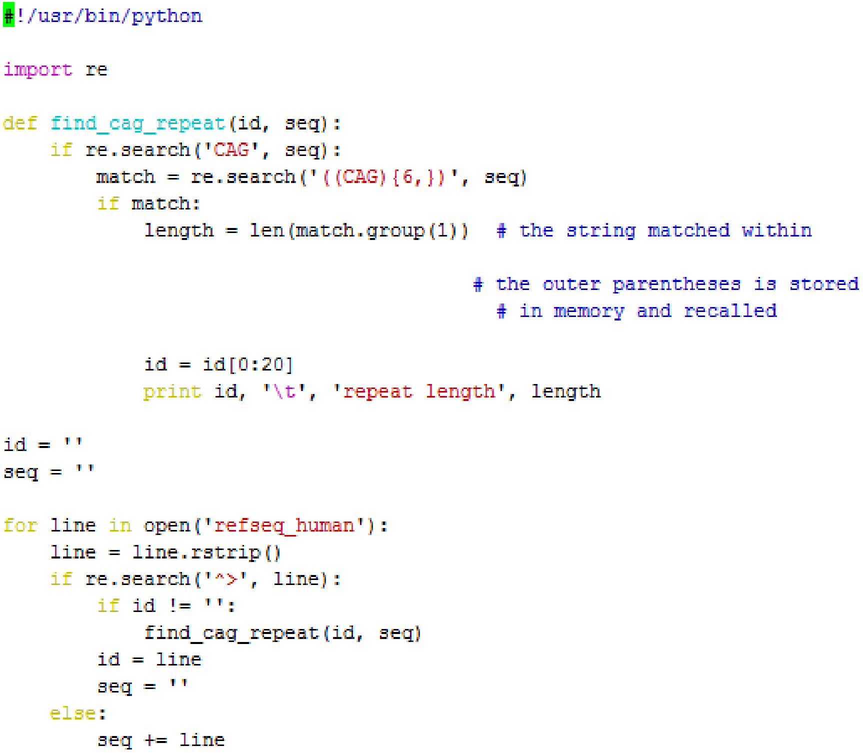


```
#!/usr/bin/python

import re

def find_cag_repeat(id, seq):
    if re.search('CAG', seq):
        match = re.search('((CAG){6,})', seq)
        if match:
            length = len(match.group(1))  # the string matched within
                                          # the outer parentheses is stored
                                          # in memory and recalled

            id = id[0:20]
            print id, '\t', 'repeat length', length


myid = ‘the title is short test sequence’
myseq = 'CGGATACTGGGGACTAAGCAGCAGCAGCAGCAGCAGCAGTTT'

find_cag_repeat(myid, myseq)
```

find_cag.py

FASTA file

```
>geneID1
atatacacacatacccacagcagcag
>geneID2
acagagacagcagcagcagcagacag
```



```
#!/usr/bin/python

import re

def find_cag_repeat(id, seq):
    if re.search('CAG', seq):
        match = re.search('((CAG){6,})', seq)
        if match:
            length = len(match.group(1))  # the string matched within

                                        # the outer parentheses is stored
                                          # in memory and recalled

            id = id[0:20]
            print id, '\t', 'repeat length', length

id = ''
seq = ''

for line in open('refseq_human'):
    line = line.rstrip()
    if re.search('^>', line):
        if id != '':
            find_cag_repeat(id, seq)
        id = line
        seq = ''
    else:
        seq += line
```

Identifying mRNAs with CAG repeats

- Regular expression

  (CAG){6}

  (CAG){6,9}

  (CAG){6,}

  (CAG)*

  (CAG)+

- Match group

‘((CAG){6,})’

The match within the outer parenthesis is stored.

→ match.group(1)

‘((CAG){6,})(.*)(.)$’

→ match.group(1), match.group(2), match.group(3)

Function : y=f(x)

```
def function_name(arg1, arg2):
        ans=0
        for i in xrange(int(arg1),int(arg2)): ans+=1
        return ans

print function_name(10, 20)
```

```
[jwnam@biglab-master Session4]$ python function_test.py
10
```

Write standard output and pipe

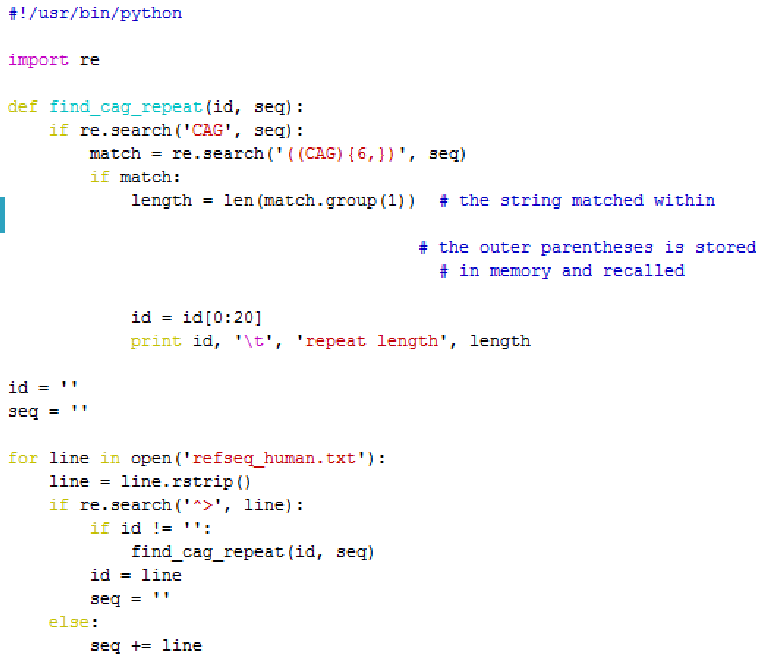
```
python function_test.py >test.out
```

→ Write a standard output of the python script into a file ‘test.out’

```
python function_test.py |wc
```

```
[jwnam@biglab-master Session4]$ python function_test.py |wc
      1       1       3
```

find_cag.py



```python
#!/usr/bin/python

import re

def find_cag_repeat(id, seq):
    if re.search('CAG', seq):
        match = re.search('((CAG){6,})', seq)
        if match:
            length = len(match.group(1))  # the string matched within

                                           # the outer parentheses is stored
                                            # in memory and recalled


            id = id[0:20]
            print id, '\t', 'repeat length', length

id = ''
seq = ''

for line in open('refseq_human.txt'):
    line = line.rstrip()
    if re.search('^>', line):
        if id != '':
            find_cag_repeat(id, seq)
        id = line
        seq = ''
    else:
        seq += line
```

```
[jwnam@biglab-master Session4]$ python find_cag.py
>gi|157151758|ref|NM      repeat length 36
>gi|116875847|ref|NM      repeat length 18
>gi|223646108|ref|NM      repeat length 39
>gi|114431247|ref|NM      repeat length 33
>gi|125346191|ref|NM      repeat length 27
>gi|154350223|ref|NM      repeat length 21
>gi|157168352|ref|NM      repeat length 18
>gi|154350245|ref|NM      repeat length 21
>gi|209862781|ref|NM      repeat length 27
```