Hands-On Exercises #1

1. **Use head and tail to extract lines 2-4 from hgmd_brca1.txt**

```
head –4 hgmd_brca1.txt | tail –3
```

```
CM014520 ATG–AGG  Met1Arg  M1R     2T>G     Ovarian cancer    Sekine (2001) Clin Cancer Res 7:3144      PubMed 11595708
CM960163 ATGg–ATT Met1Ile  M1I     3G>T     Breast cancer    Couch (1996) Hum Mutat 8:8       PubMed 8807330
CM041678 ATG–ACG  Met1Thr  M1T     2T>C     Breast and/or ovarian cancer ?    Abkevich (2004) J Med Genet 41:492       PubMed 15235020
```

2. **How many BRCA1 variants were published in 2003? How many of these involve both ovarian and breast cancer?**

   40 variants were published in 2003.
   8 variants involve both ovarian and breast cancer.

   ```
   grep "(2003)" hgmd_brca1.txt | wc –l

   grep "(2003)" hgmd_brca1.txt | grep –i ovar
     | grep –i breast | wc –l
   ```

   If we leave off parentheses, we pick up an extraneous match. With the parentheses, we need quotes, since parentheses have a special meaning to the Unix shell.

3. **Use man to learn what the -w option does in the grep command. Try it.**

   -w forces the pattern to match an entire word.

   ```
   grep    Ovar     hgmd_brca1.txt | wc –l
   grep –w Ovar     hgmd_brca1.txt | wc –l
   grep –w Ovarian hgmd_brca1.txt | wc –l
   ```

Hands-On Exercises #2

1. **Are all TF binding sites on the "+" (forward) strand?**

   Yes, they are all on the _+ strand.
   Here's one way to confirm (there are several others):
   ```
   cut –f6 tfbs.txt | uniq
   ```

2. **What are the highest scoring binding sites for c-Myc? the lowest scoring?**

   The highest scoring binding site is 886; the lowest scoring site is 103.
   ```
   grep c-Myc tfbs.txt | sort –k5,5nr | head
   grep c-Myc tfbs.txt | sort –k5,5n  | head
   ```

3. **How many binding sites for NFKB have the highest score? the second highest?**

   6 binding sites have the high score of 1000. 1 binding site has the second-highest score of 792.
   ```
   cut –f4,5 tfbs.txt | grep NFKB  | sort –k2,2nr
     | uniq –c | head –2
   ```

4. **Going back to hgmd_brca1.txt, how many variants are there for each type of cancer?**

   ```
   cut –f6 hgmd_brca1.txt | sort | uniq –c

     2 Breast and colorectal cancer
   117 Breast and/or ovarian cancer
    58 Breast and/or ovarian cancer ?
   190 Breast cancer
    42 Breast cancer ?
     1 Breast cancer, early-onset
     1 Cervical cancer, decreased risk, association with
     1 Endometrial cancer
     1 Endometriosis, association with ?
    22 Ovarian cancer
     6 Ovarian cancer ?
     1 Phenotype
     5 Prostate cancer
   ```