Genomic Computing Evaluation

Assignment 1: The Genome Browser

Fabrizio Frasca

April 21, 2017

1 UCSC Genome Browser

(a)

Go to the genomic region chr6:45,296,054-45,518,819 of the Human assembly GRCh37/h19.

Which genes do you see in this region? On which strand are they?

This region of human DNA contains part of the **SUPT3H** gene and part of the **RUNX2** gene. The former is present in 4 variants and the latter is present in 3 variants in the UCSC database.

SUPT3H is on the **reverse strand** (arrows pointing towards left), whereas the RUNX2 is on the **forward strand** (arrows pointing towards right).

(b)

Enable GC Percent [dense] from Mapping and Sequencing tracks and CpG Islands [show] from Regulation tracks.
Now zoom into the region chr6:45,330,000-45,400,000.

Does the GC composition reach a peak in correspondence of some important regulatory element?

Yes, it does. In particular, as shown in figure 1, two significant peaks are observed: one corresponds to the promotorial region of the SUPT3H gene and the other one corresponds to the promotorial region of one of the variants of RUNX2 gene.

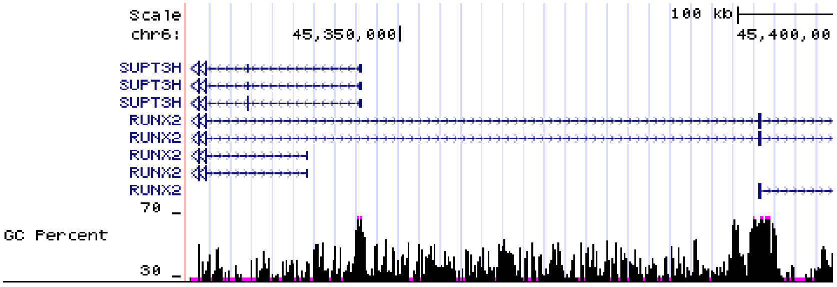


Figure 1: Peaks for the GC signal

In which region do you notice the most regulatory activity? Does it involve a CpG island? (If yes, report Coordinates, Chromosome band, Genomic size of the first (along the genome) of them)

The regions interested by regulatory activities the most are the ones detected above: **chr6:45,343,511-45,348,384** ca. for the promotorial region of SUPT3H and **chr6:45,387,920-45,392,851** ca., corresponding to the promoter of one of the variants for the RUNX2 gene.

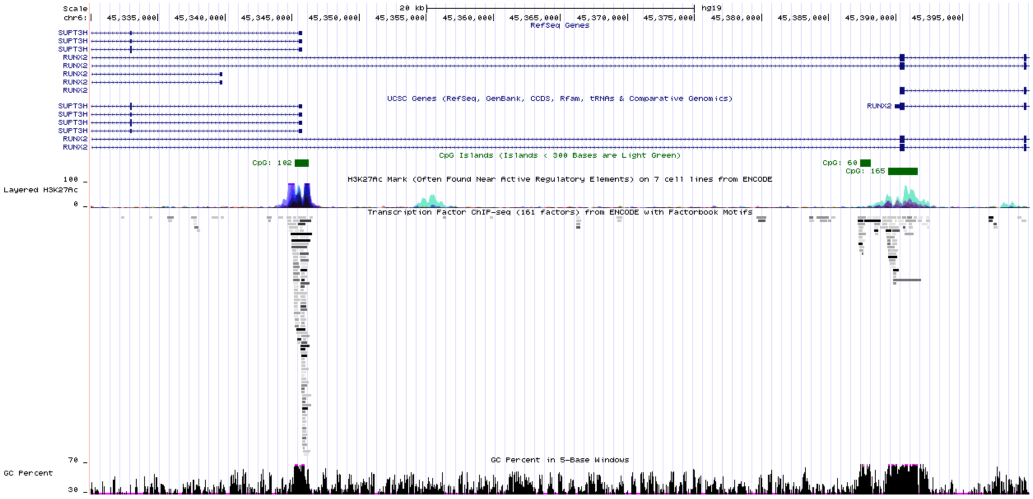

Scale
chr6:
20 kb
hg19
45,335,000
45,340,000
45,345,000
45,350,000
45,355,000
45,360,000
45,365,000
45,370,000
45,375,000
45,380,000
45,385,000
45,390,000
45,395,000
RefSeq Genes
SUPT3H
RUNX2
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 102
CpG: 60
CpG: 165
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
70
GC Percent
30
GC Percent in 5-Base Windows

Figure 2: Regions involved in regulatory activities

These regions are promisingly characterized by regulatory activities because of the conjugate presence of CpG islands, strong evidence of H3K27 acetylation and the binding of several transcription factors, as shown by the relative tracks in figure 2.

The first CpG island encountered by proceeding in the direction of the forward strand is the one corresponding to the promotorial region of the SUPT3H region. In particular, it has:

coordinates:	chr6:45345186-45346261
chromosome band:	6p21.1
genomic size:	1076

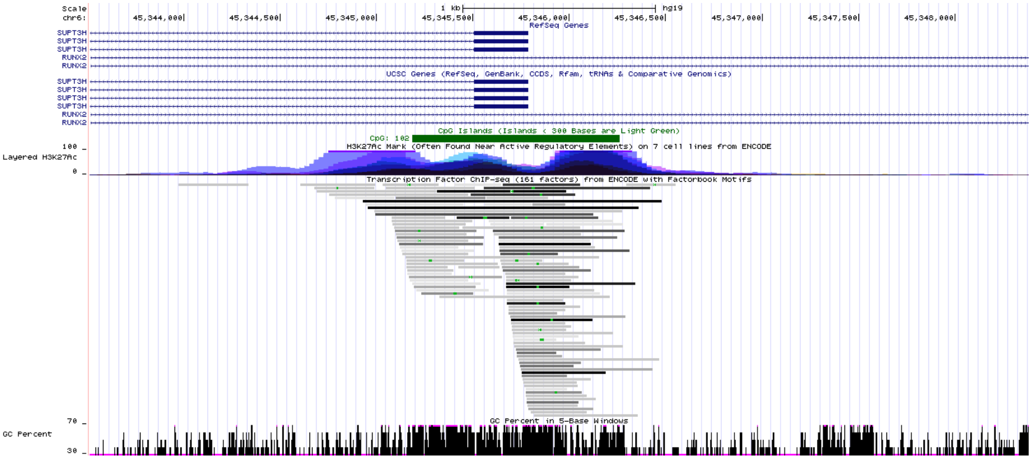

Scale
1 kb
hg19
chr6:
45,344,000
45,344,500
45,345,000
45,345,500
45,346,000
45,346,500
45,347,000
45,347,500
45,348,000
RefSeq Genes
SUPT3H
SUPT3H
SUPT3H
RUNX2
RUNX2
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SUPT3H
SUPT3H
SUPT3H
SUPT3H
RUNX2
RUNX2
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 102
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
70
GC Percent
30
GC Percent in 5-Base Windows

Figure 3: The first CpG island, along with other 'regulatory tracks' in the same region

2 BLAT

Consider the following FASTA file:

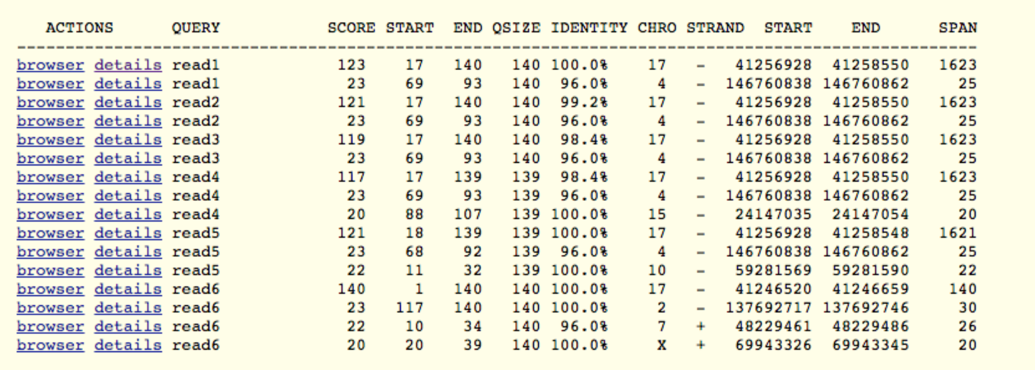
```
> read1
ACCACATATTTTGCAAATTTTGCATGCTGAAACTTCTCAACCAGAAGAAAGGGCCTTCACAG
TGTCCTTTATGTAAGAATGATATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAA
CTTGTTGAAGAGCTA
> read2
ACCACATATTTTGCAAATTTTGCATGCTGATACTTCTCAACCAGAAGAAAGGGCCTTCACAGT
GTCCTTTATGTAAGAATGATATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAAC
TTGTTGAAGAGCTA
> read3
ACCACATATTTTGCAAATTTTGCATGCTGATACTACTCAACCAGAAGAAAGGGCCTTCACAGT
GTCCTTTATGTAAGAATGATATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAAC
TTGTTGAAGAGCTA
> read4
ACCACATATTTTGCAAATTTTGCATGCTGATACTACTCAACCAGAAGAAAGGGCCTTCACAGT
GTCCTTTATGTAAGAATGATATAACCAAAAGGAGCCTAAAGAAAGTACGAGATTTAGTCAACT
TGTTGAAGAGCTA
> read5
ACCACATATTTTGCAAATTTGCATGCTGAAACTTCTCAACCAGAAGAAAGGGCCTTCACAGTG
TCCTTTATGTAAGAATGATATAACCAAAAGGAGCCTACAAGAAAGTACGAGATTTAGTCAACT
TGTTGAAGAGCTA
> read6
```

```
ATGAATGTAGAAAAGGCTGAATTCTGTAATAAAAGCAAACAGCCTGGCTTAGCAAGGAGCCAA
CATAACAGATGGGCTGGAAGTAAGGAAACATGTAATGATAGGCGGACTCCCAGCACAGAAAAA
AAGGTAGATCTGAA
```

Use BLAT to map these sequences onto Human assembly GRCh37/h19.

Does each read map in a single region?

No, it does not. For instance, read 1 maps on chr17:41256928-41258550 and chr4:146760838-146760862 regions at the same time, see figure 4.



```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START     END       SPAN
---------------------------------------------------------------------------------------------------
browser details read1             123    17  140   140 100.0%     17   -    41256928  41258550  1623
browser details read1              23    69   93   140  96.0%      4   -   146760838 146760862    25
browser details read2             121    17  140   140  99.2%     17   -    41256928  41258550  1623
browser details read2              23    69   93   140  96.0%      4   -   146760838 146760862    25
browser details read3             119    17  140   140  98.4%     17   -    41256928  41258550  1623
browser details read3              23    69   93   140  96.0%      4   -   146760838 146760862    25
browser details read4             117    17  139   139  98.4%     17   -    41256928  41258550  1623
browser details read4              23    69   93   139  96.0%      4   -   146760838 146760862    25
browser details read4              20    88  107   139 100.0%     15   -    24147035  24147054    20
browser details read5             121    18  139   139 100.0%     17   -    41256928  41258548  1621
browser details read5              23    68   92   139  96.0%      4   -   146760838 146760862    25
browser details read5              22    11   32   139 100.0%     10   -    59281569  59281590    22
browser details read6             140     1  140   140 100.0%     17   -    41246520  41246659   140
browser details read6              23   117  140   140 100.0%      2   -   137692717 137692746    30
browser details read6              22    10   34   140  96.0%      7   +    48229461  48229486    26
browser details read6              20    20   39   140 100.0%      X   +    69943326  69943345    20
```

Figure 4: Results for the alignment from BLAT

On the base of the alignment score and sequence identity, which genome region do the reads belong?

It is fairly evident that all the alignments with high score map to the same genome region and they are characterized by very high identity as well. The region is **chr17:41256928-41258550** for all the reads 1-5 and **chr17:41246520-41246659** for read 6, which is definitely similar. Other alignments have pretty good identity, but they must be discarded since their score is poor.

Which gene this reads come from?

By clicking on the “browser” link for one of the mapping with highest score, we can realize the interested gene is **BRCA1**.