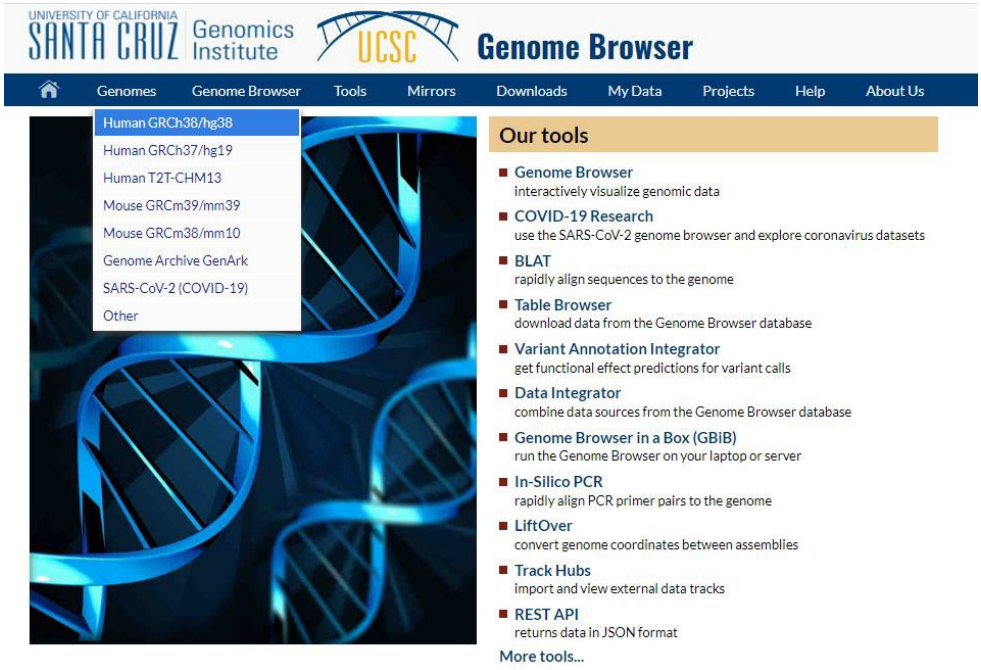
第 1 次作业报告

Exploring the Human Genome

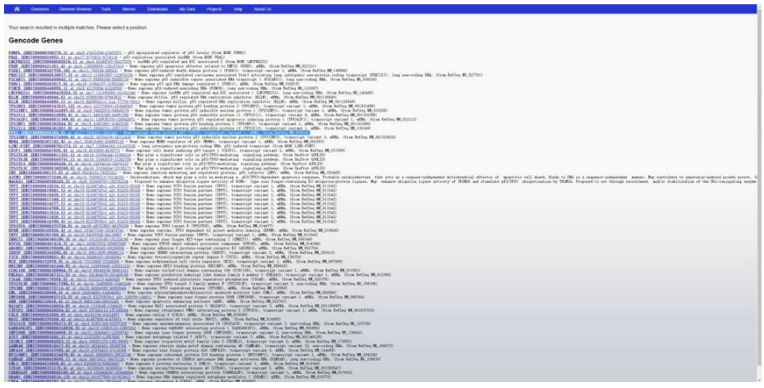
崔晏菲 自硕 21 2021210976

一、UCSC Genome Browser 网站的使用 [1]

输入网站 https://genome.ucsc.edu/，进入 Genome Browser 的主界面。在左上角 Genomes 菜单栏选择希望查看的基因组，如下图所示，以 Human GRCh38/hg38 为例。



点击后，进入 Human GRCh38/hg38 的查看界面。在搜索框中输入与肺癌密切相关的基因 p53，搜索结果如下图所示，选择 TP53，进入该基因的查看界面。



[1] 这一部分的图例均来自于网站 Genome Browser 的截图

TP53 的查看结果如下图所示。对于图中每一条蓝色的线，较粗的蓝色部分代表外显子，箭头部分的线代表内含子，同时箭头的方向代表转录的方向，将鼠标停留在线上，可以查看相应外显子和内含子的名称。不同的线代表不同的转录方式。

下图展示了 TP53 基因相关的统计图表。第一行的柱状图代表 TP53 基因在不同组织中表达量，点开后有更清晰的箱线图。第二行代表了 ENCODE 记录的顺式调控元件候选者，点开后可以查看 ENCODE 的具体方法。第三行代表组蛋白标记的富集情况，点开后会显示组蛋白标记的详细解释。第四行的峰值图展现了 100 个脊椎动物用 PhyloP 方法分析得到的保守区域。下边的几行黑色图表分别展示了恒河猴、小鼠、狗、大象、鸡、爪蟾、斑马鱼对应片段的保守区域。最后展现了单核苷酸多态性的情况。

之后的表格有各种各样的菜单栏和选项框，就不再赘述。

二、人类基因数据分析 [2]

1. 数据来源

本次作业的数据来源为网站 https://www.gencodegenes.org/human/中的 gencode.v41.annotation.gtf.gz。

下载链接为：https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_41/gencode.v41.annotation.gtf.gz。

所用编程语言为 Python。

[2] 本部分代码见 homework_code.ipynb 文件

2. 实验原理与步骤

(a) 实验数据读取

本次实验使用 python 中 gtfparse 包进行数据的读取。读取后，以第一行为例展示该数据，如下图所示。数据共有 26 列，包含其类型、所在染色体、起止位置等信息。

```
print(data.iloc[0])
```

0.4s

```
seqname                                      chr1
source                                     HAVANA
feature                                      gene
start                                       11869
end                                         14409
score                                         NaN
strand                                          +
frame                                           0
gene_id                          ENSG00000223972.5
gene_type       transcribed_unprocessed_pseudogene
gene_name                                  DDX11L1
level                                           2
hgnc_id                                 HGNC:37102
havana_gene                OTTHUMG00000000961.2
transcript_id
transcript_type
transcript_name
transcript_support_level
tag
havana_transcript
exon_number
exon_id
ont
protein_id
ccdsid
artif_dupl
Name: 0, dtype: object
```

(b) 计算蛋白编码基因的数量

通过代码逐层筛选出编码蛋白的基因。第一步，根据”seqname”的值选择相应染色体上的数据。第二步，根据”feature”的值选出基因数据。第三步，根据”gene_type” 的值选择出蛋白编码基因。相关代码下图所示。

```
# 计算每个染色体上的蛋白质编码基因
for chr in chr_num:
    chr_data = data[data["seqname"] == chr]
    chr_gene = chr_data[chr_data["feature"] == "gene"]
    chr_protein_gene = chr_gene[chr_gene["gene_type"] == "protein_coding"]
    protein_coding_genes[chr] = len(chr_protein_gene)
```

(c) 计算外显子、内含子、基因的长度分布及相关的统计量

基因长度计算：end-start+1 即为每个基因的长度。

外显子长度计算：在原始数据中，该基因对应位置下的 feature 为 exon 的数据，计算其长度。

内含子长度计算：基因长度减去外显子长度即为内含子长度。

相关的统计量使用 numpy 中的 median()、mean()等函数计算即可。直方图使用 matplotlab.pylot 中的 hist()等函数计算即可。本部分代码如下：

```
def cal_gene_len(gene, ind):
    len = gene["end"][ind] - gene["start"][ind] + 1
    return len

def cal_dis(gene_data, chr_data):
    genes_len_list, exons_len_list, introns_len_list = [], [], []
    for i in gene_data.index:
        gene_len = cal_gene_len(chr_data, i)
        exon_len = 0
        if(chr_data["feature"][i + 1] == "transcript"):
            exon_end = chr_data["start"][i + 1]
            trains_len = cal_gene_len(chr_data, i + 1)
            i += 2
            while(chr_data["feature"][i] != "gene" and chr_data["feature"][i] != "transcript"):
                if(chr_data["feature"][i] == "exon"):
                    if(chr_data["start"][i] >= exon_end):
                        exon_len += cal_gene_len(chr_data, i)
                        exon_end = chr_data["end"][i]
                    else:
                        break
                i += 1
            genes_len_list.append(gene_len)
            exons_len_list.append(exon_len)
            introns_len_list.append(gene_len - exon_len)
    return genes_len_list, exons_len_list, introns_len_list
```

✓ 0.3s

```
chr_genes_len = []
chr_exons_len = []
chr_introns_len = []
for chr in chr_num:
    chr_data = data[data["seqname"] == chr]
    chr_genes = chr_data[chr_data["feature"] == "gene"]
    chr_protein_genes = chr_genes[chr_genes["gene_type"] == "protein_coding"]
    genes_len_list, exons_len_list, introns_len_list = cal_dis(chr_protein_genes, chr_data)
    chr_genes_len.append(genes_len_list)
    chr_exons_len.append(exons_len_list)
    chr_introns_len.append(introns_len_list)
```

✓ 8.3s

3. 实验结果

(a) 蛋白编码基因的数量

每条染色体的蛋白编码基因数量如下：

'chr1': 2062,
'chr2': 1247,
'chr3': 1077,
'chr4': 754,
'chr5': 882,
'chr6': 1049,
'chr7': 930,
'chr8': 696,
'chr9': 774,
'chr10': 730,
'chr11': 1318,
'chr12': 1037,

'chr13': 322,
'chr14': 615,
'chr15': 600,
'chr16': 856,
'chr17': 1185,
'chr18': 266,
'chr19': 1474,
'chr20': 546,
'chr21': 219,
'chr22': 446,
'chrX': 856,
'chrY': 63,
'chrM': 13

(b) 内含子、外显子、基因长度的统计量

内含子相关的统计量如下表所示：

	median	mean	max	min
chr1	23417.5	56797.21047526673	1551826.0	0.0
chr2	42842.0	86982.32718524459	1899225.0	0.0
chr3	40967.0	97521.24512534818	1737351.0	0.0
chr4	47326.5	99194.09946949602	1500409.0	0.0
chr5	42387.5	88024.61337868481	1552524.0	0.0
chr6	22802.0	70724.44995233556	1987155.0	0.0
chr7	31372.5	86661.94946236559	2304321.0	0.0
chr8	35242.0	89570.19252873563	2059435.0	0.0
chr9	25603.0	68249.49224806202	2298442.0	0.0
chr10	40465.0	88907.21780821918	1851063.0	0.0
chr11	14945.5	51161.89984825493	2172396.0	0.0
chr12	28678.0	66641.34522661523	1258398.0	0.0
chr13	45742.0	104436.4596273292	1472119.0	0.0
chr14	26215.0	67849.78699186991	1692507.0	0.0
chr15	38344.0	75145.49166666667	956969.0	0.0
chr16	19303.5	48635.66355140187	2472710.0	0.0
chr17	16083.0	39081.92405063291	1159990.0	0.0
chr18	53388.5	111655.23308270676	1185526.0	0.0
chr19	14769.0	23813.70895522388	484462.0	0.0
chr20	23325.0	53369.32417582418	2052835.0	0.0
chr21	24111.0	66517.03196347032	1215019.0	0.0
chr22	21675.0	44695.36771300448	760513.0	0.0
chrX	16288.5	63152.73130841122	2241819.0	0.0
chrY	24164.0	55850.88888888889	737289.0	0.0
chrM	0.0	0.0	0.0	0.0

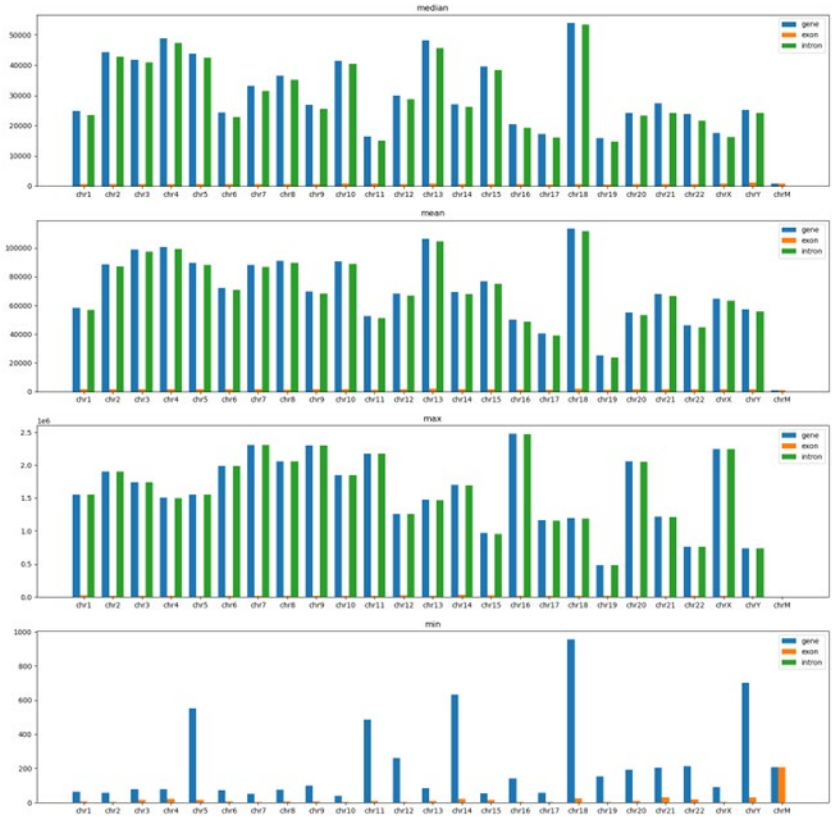
外显子长度的统计数据如下表所示

	median	mean	max	min
chr1	654.0	1557.4752667313287	27923.0	5.0
chr2	613.0	1626.4546912590217	20879.0	3.0
chr3	584.0	1521.4614670380688	17064.0	15.0
chr4	599.0	1555.5358090185675	16162.0	21.0
chr5	607.5	1678.6666666666667	13488.0	15.0
chr6	623.0	1570.5938989513822	18797.0	7.0
chr7	601.0	1532.8333333333333	16366.0	4.0
chr8	560.0	1318.0617816091954	20241.0	7.0
chr9	705.5	1542.6330749354006	15789.0	6.0
chr10	766.0	1788.2027397260274	15857.0	4.0
chr11	728.5	1392.103186646434	19504.0	8.0
chr12	571.0	1486.387656702025	23524.0	4.0
chr13	821.5	1918.5465838509317	16122.0	8.0
chr14	606.0	1442.6878048780488	33681.0	20.0
chr15	572.0	1643.8583333333333	21679.0	15.0
chr16	571.5	1379.5992990654206	14219.0	4.0
chr17	554.0	1290.5130801687765	21079.0	3.0
chr18	597.5	1934.0601503759399	16433.0	25.0
chr19	541.0	1214.4138398914517	16219.0	3.0
chr20	665.5	1575.448717948718	12236.0	9.0
chr21	588.0	1506.2739726027398	13170.0	31.0
chr22	594.0	1512.8744394618834	18445.0	18.0
chrX	786.0	1575.6869158878505	13748.0	3.0
chrY	1161.0	1463.857142857143	7368.0	31.0
chrM	784.0	876.5384615384615	1812.0	207.0

基因长度的统计数据如下表所示

	median	mean	max	min
chr1	24843.5	58354.68574199806	1551957.0	63.0
chr2	44366.0	88608.7818765036	1899594.0	57.0
chr3	41750.0	99042.70659238625	1743270.0	78.0
chr4	48832.5	100749.63527851459	1506192.0	78.0
chr5	43754.0	89703.28004535148	1553083.0	553.0
chr6	24360.0	72295.04385128694	1987247.0	72.0
chr7	33077.5	88194.78279569892	2304997.0	51.0
chr8	36592.5	90888.25431034483	2059620.0	75.0
chr9	26915.5	69792.12532299741	2298757.0	99.0
chr10	41408.5	90695.4205479452	1851181.0	39.0
chr11	16425.0	52554.00303490137	2172911.0	485.0
chr12	29890.0	68127.73288331727	1258480.0	261.0
chr13	48171.0	106355.00621118012	1475062.0	84.0
chr14	27026.0	69292.47479674796	1697919.0	634.0
chr15	39590.0	76789.35	969299.0	54.0
chr16	20478.5	50015.26285046729	2473539.0	141.0
chr17	17153.0	40372.437130801685	1161878.0	57.0
chr18	53975.0	113589.2932330827	1195707.0	956.0
chr19	15899.0	25028.122795115334	485249.0	153.0
chr20	24231.0	54944.7728937729	2057829.0	192.0
chr21	27326.0	68023.30593607305	1216867.0	204.0
chr22	23823.5	46208.242152466366	760616.0	213.0
chrX	17577.0	64728.418224299065	2241933.0	90.0
chrY	25175.0	57314.74603174603	741999.0	702.0
chrM	784.0	876.5384615384615	1812.0	207.0

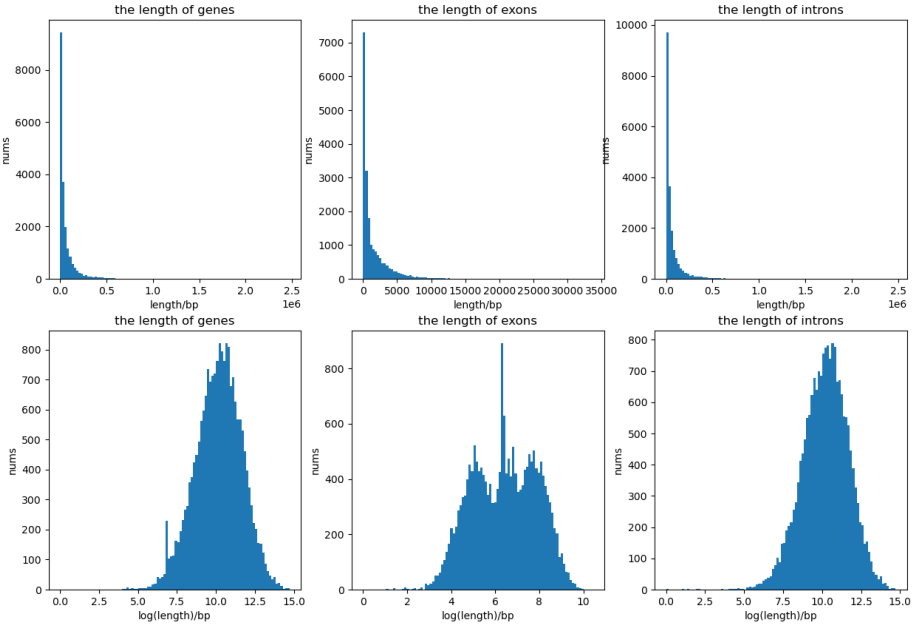
内含子、外显子、基因长度统计量的柱状图如下图所示。



(c) 内含子、外显子、基因长度的分布直方图

以所有染色体的长度分布为例，如下图所示，其余图像见 figures 文件夹。其中上面的图的横坐标为碱基数，下面图的横坐标为碱基数的对数。

从图中可以看出，染色体上大部分基因长度和内含子长度小于$2 \times 10^5 bp$，大部分外显子长度小于$10^4 bp$，外显子的长度显著低于基因的长度和内含子的长度。



4.实验总结

本实验的前提是数据来源准确。

在计算外显子时，默认数据为顺序排列，即一行 feature 为 gene 的数据下，feature 为 exon 的数据均为该基因外显子的数据。若数据出现混乱，则会导致结果不准确。

在计算内含子长度时，直接用基因长度减去外显子长度，但这样的结果不仅包含了内含子的长度，还包括了非翻译区、启动子、终止子的长度，导致内含子结果偏大。

并且，在所有的染色体中，均存在内含子长度最小值为 0 的情况，理论上这是不应该出现的，这样的情况进一步增大了误差。