Homework 3

David Angeles Albores, bi183

January 25, 2018

Problem 1. How many total bases of sequence are there in the transcriptome? How many transcripts are there? What is the longest transcript in the human genome?

There are 294306140 bases coding for 180869 transcripts, of which the largest is ENST00000589042.5. See attached code.

Problem 2. Find the string S whose Burrows-Wheeler transform is: ACTTCCCGGAAAAA$TTAA

I implemented the inverse Burrows-Wheeler transform in a Jupyter notebook in Python (see code). The string S is GATTACACACAGATTACA$.

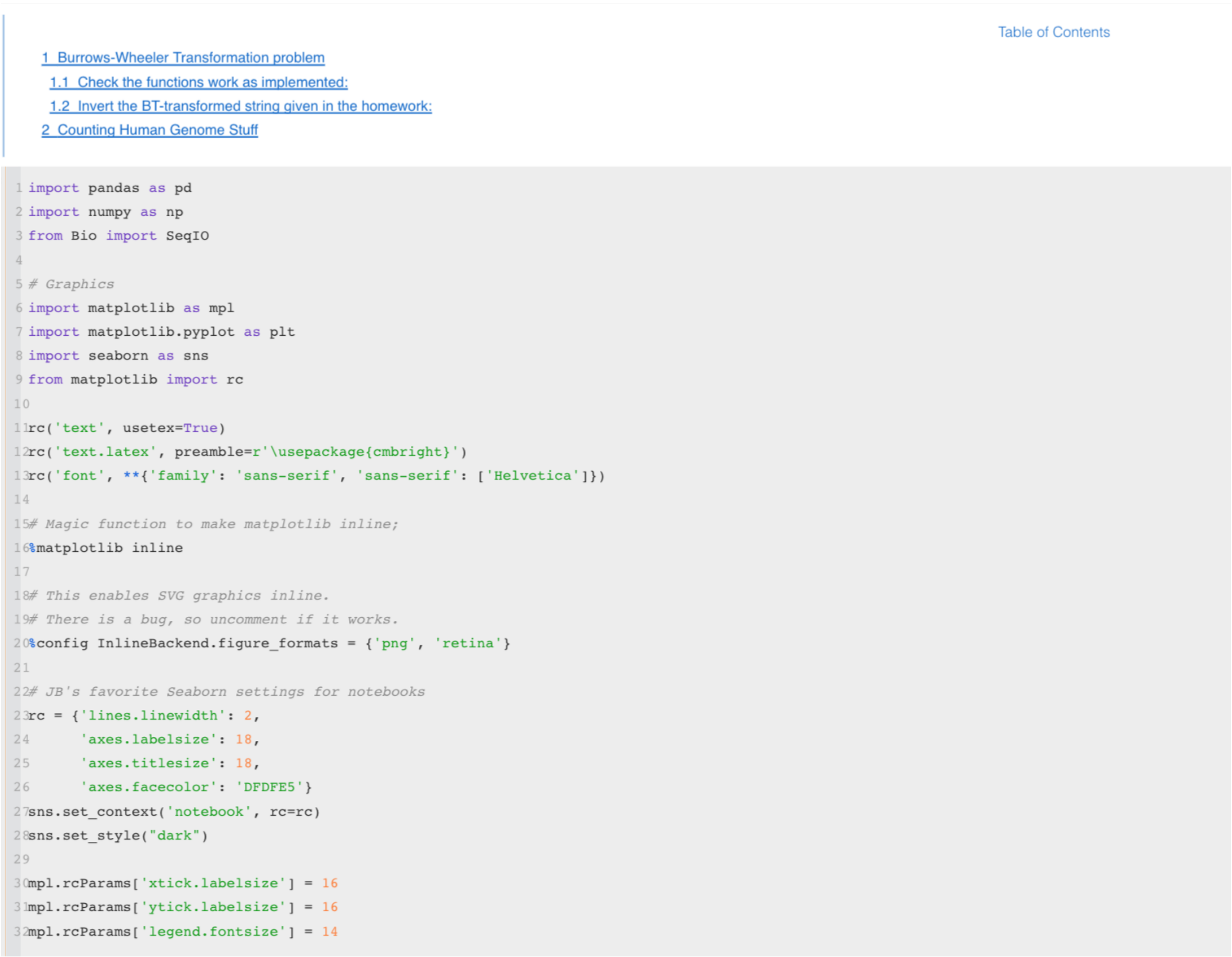


Table of Contents

```python
import pandas as pd
import numpy as np
from Bio import SeqIO

# Graphics
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import rc

rc('text', usetex=True)
rc('text.latex', preamble=r'\usepackage{cmbright}')
rc('font', **{'family': 'sans-serif', 'sans-serif': ['Helvetica']})

# Magic function to make matplotlib inline;
%matplotlib inline

# This enables SVG graphics inline.
# There is a bug, so uncomment if it works.
%config InlineBackend.figure_formats = {'png', 'retina'}

# JB's favorite Seaborn settings for notebooks
rc = {'lines.linewidth': 2,
      'axes.labelsize': 18,
      'axes.titlesize': 18,
      'axes.facecolor': 'DFDFE5'}
sns.set_context('notebook', rc=rc)
sns.set_style("dark")

mpl.rcParams['xtick.labelsize'] = 16
mpl.rcParams['ytick.labelsize'] = 16
mpl.rcParams['legend.fontsize'] = 14
```

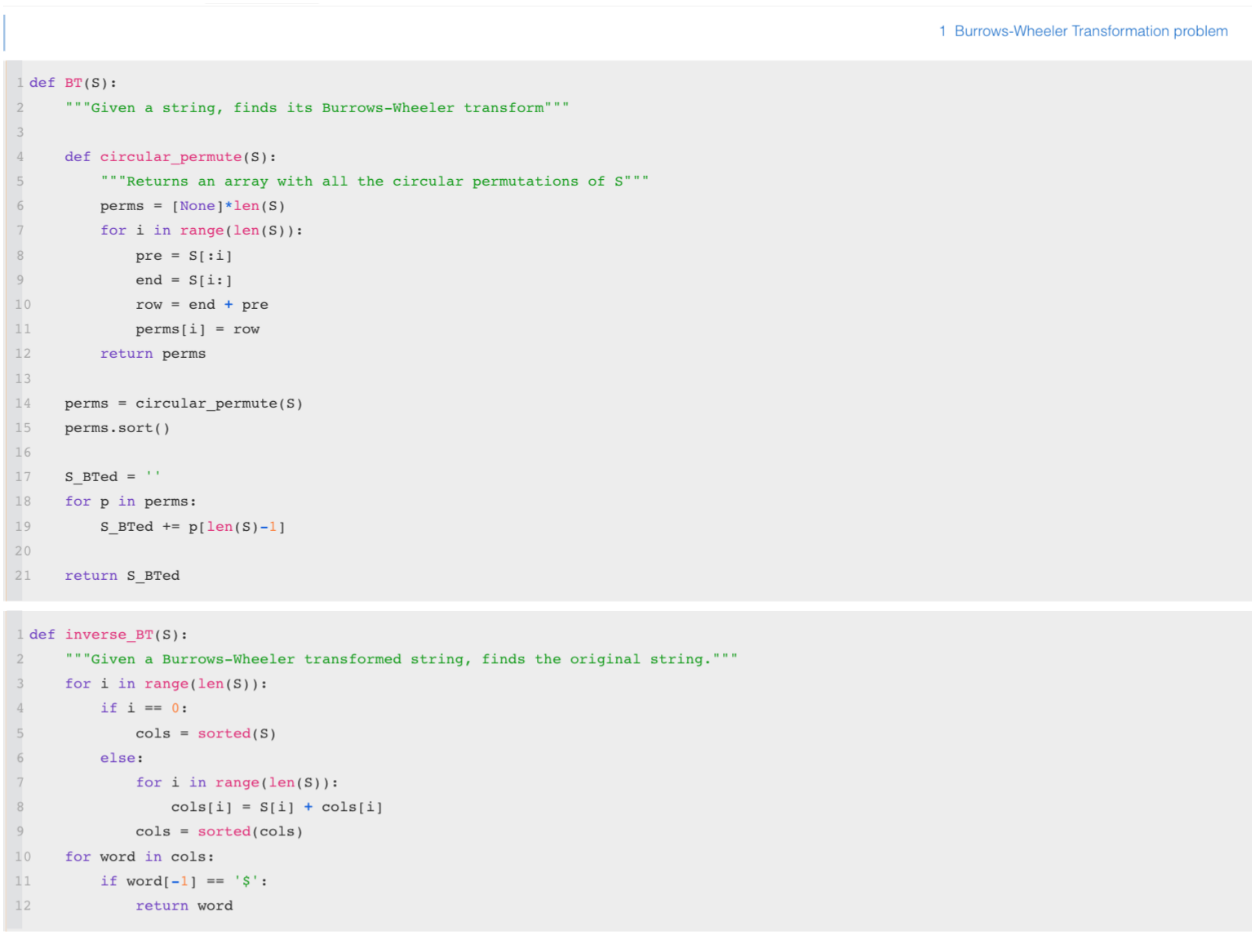
Figure 1. Code

Problem 3. Describe in detail a protocol for performing RNA-seq on human blood from a patient. How much blood will have to be drawn to build 3 separate RNA-seq libraries

This question is so poorly written as to be unanswerable. Using modern techniques, the absolute minimum quantity of blood required is probably less than 1mL, since we just need 3 single cells of whatever kind we are interested in.

A protocol:

- Find a human
- Insert needle
- Get blood
- Isolate single cells
- Perform Drop-Seq at the Single-Cell institute
- Pay a company to do analysis
- Write paper

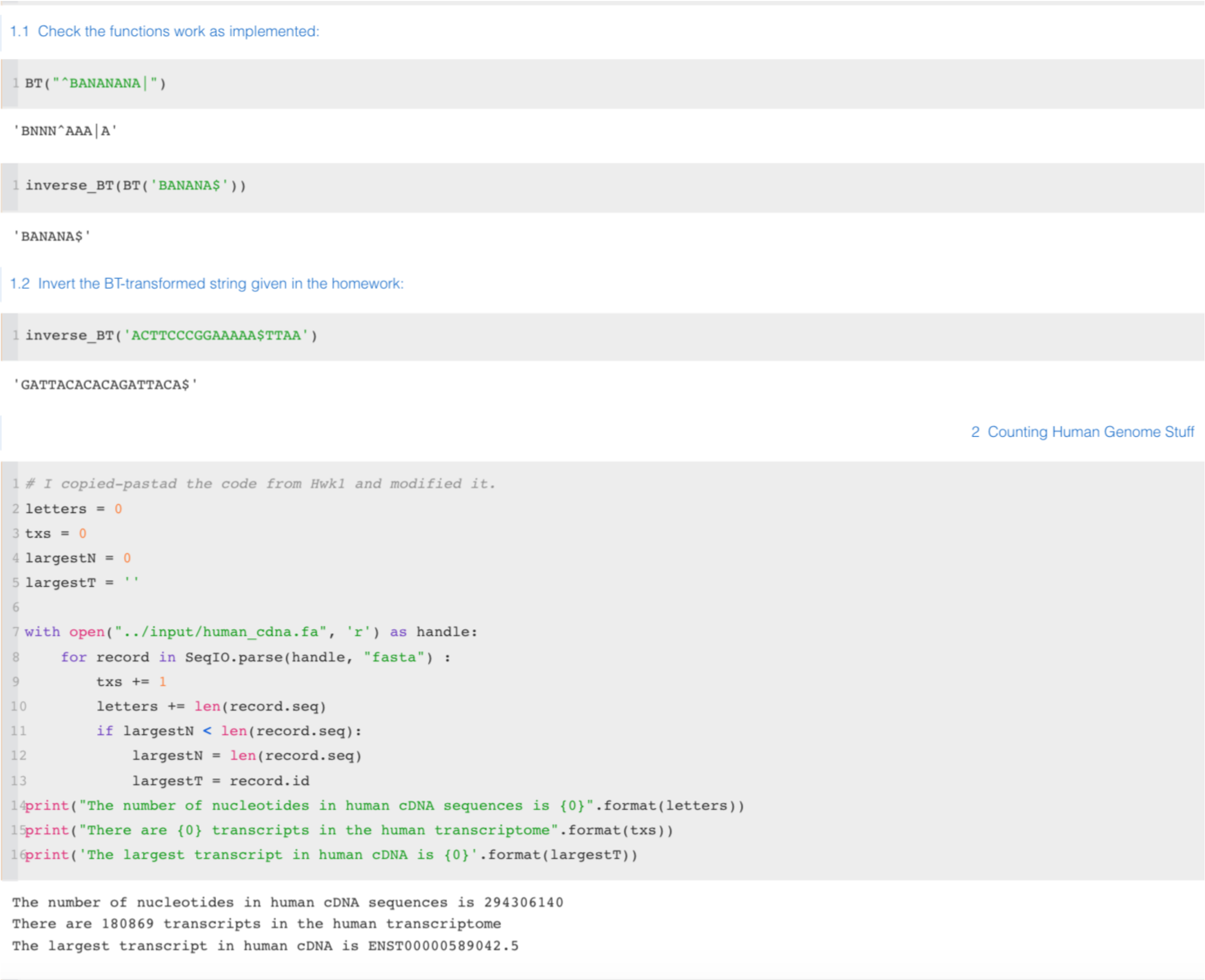


1 Burrows-Wheeler Transformation problem

```python
def BT(S):
    """Given a string, finds its Burrows-Wheeler transform"""

    def circular_permute(S):
        """Returns an array with all the circular permutations of S"""
        perms = [None]*len(S)
        for i in range(len(S)):
            pre = S[:i]
            end = S[i:]
            row = end + pre
            perms[i] = row
        return perms

    perms = circular_permute(S)
    perms.sort()

    S_BTed = ''
    for p in perms:
        S_BTed += p[len(S)-1]

    return S_BTed
```

```python
def inverse_BT(S):
    """Given a Burrows-Wheeler transformed string, finds the original string."""
    for i in range(len(S)):
        if i == 0:
            cols = sorted(S)
        else:
            for i in range(len(S)):
                cols[i] = S[i] + cols[i]
            cols = sorted(cols)
    for word in cols:
        if word[-1] == '$':
            return word
```

Figure 2. Code



1.1 Check the functions work as implemented:

```
BT("^BANANANA|")
```

```
'BNNN^AAA|A'
```

```
inverse_BT(BT('BANANA$'))
```

```
'BANANA$'
```

1.2 Invert the BT-transformed string given in the homework:

```
inverse_BT('ACTTCCCGGAAAAA$TTAA')
```

```
'GATTACACACAGATTACA$'
```

2 Counting Human Genome Stuff

```
# I copied-pastad the code from Hwk1 and modified it.
letters = 0
txs = 0
largestN = 0
largestT = ''

with open("../input/human_cdna.fa", 'r') as handle:
    for record in SeqIO.parse(handle, "fasta") :
        txs += 1
        letters += len(record.seq)
        if largestN < len(record.seq):
            largestN = len(record.seq)
            largestT = record.id
print("The number of nucleotides in human cDNA sequences is {0}".format(letters))
print("There are {0} transcripts in the human transcriptome".format(txs))
print('The largest transcript in human cDNA is {0}'.format(largestT))
```

```
The number of nucleotides in human cDNA sequences is 294306140
There are 180869 transcripts in the human transcriptome
The largest transcript in human cDNA is ENST00000589042.5
```

Figure 3. Code