aml035

December 24, 2019

1 Filter for cells that have over 200 reads mapping to mitochondrio, and at least one of the variants 3010G->A and 9698T>C. The paper says that 1,077 cells were found (see below), so trying to get the same number

2 Similar to the paper:

For the AML datasets previously generated by 10X Genomics (Zheng et al., 2017b), cells from two patients (AML027 and AML035) were analyzed for mitochondrial genotypes. Aligned and processed .bam files were downloaded from the 10X website (https:// support.10xgenomics.com/single-cell-gene-expression/datasets/) and further processed using custom Python scripts. Cell barco- des associated with at least 200 reads uniquely aligning to the mitochondrial genome were considered for downstream analysis. Barcodes were further filtered by requiring coverage by at least one read at two specific variants at mtDNA positions 3010 and 9698. We note that we did not observe a barcode that contained a read to support both alternate alleles (3010G > A and 9698T > C). We determined that 4 out of 1,077 cells were derived from the recipient (Figure 7M), a higher estimate than in the previously reported analysis performed with nuclear genome variants (reported exactly 0%) (Zheng et al., 2017b), though these four cells were not included in the published analysis as they did not pass the author's barcode/ transcriptome filters. We did not observe a well-covered set of variants separating the donor/ recipient pair in the AML027 dataset, and did not further analyze it for mutations but only for determining well-covered barcodes (Figures S7G and S7H

2.1 Steps:

1. Download the aml035 bam and index file.
2. Extract mitochondrion reads and create index file.
3. Get list of cell barcodes (corrected CB's and CR's) from the mitochondria file
4. Create text file of reads covering the two variants
5. Loop through those files and create a dictionary for the CB's that count how many times the alternative and reference are seen.
6. Create filter for having at least 200 basepairs

2.2 Load packages and set parameters

[1]:
```
import glob
import os
import pandas as pd
from tqdm import tqdm
import numpy as np
import pysam
import time
from collections import defaultdict
import pickle
from itertools import product
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

[2]:
```
from bam_barcodes_function import extract_barcode_info
```

[3]:
```
ORIG = "data/aml035_post_transplant_possorted_genome_bam.bam"
BAM = "data/aml035_post_transplant_possorted_genome_bam.MT.bam"
BARCODE_INFO = "data/barcode_data_aml035_post_transplant_possorted_genome_bam.
 ↪MT.p"

GENOME = "/data2/genome/human_GRCh38/cellranger/refdata-cellranger-GRCh38-3.0.0/
 ↪fasta/genome.fa"
NAME = 'aml_035'
```

[4]:
```
nucs = ["A", "T","C", "G"]
variants = ["3010", "9698"]
```

[5]:
```
if not os.path.exists("data"):
    os.mkdir("data")
```

2.3 1. Download files

[ ]:
```
cmd = "wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/
 ↪aml035_post_transplant/aml035_post_transplant_possorted_genome_bam.bam -O␣
 ↪{ORIG}"
print(cmd)
os.system(cmd)

cmd = "wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/
 ↪aml035_post_transplant/aml035_post_transplant_possorted_genome_bam_index.bam.
 ↪bai -O {ORIG}.bai"
```

```
os.system(cmd)
print(cmd)
```

2.4 2. Extract mitochondria reads

[6]:
```
cmd = f"samtools view {ORIG} -q 30 MT -b > {BAM}"
print(cmd)
os.system(cmd)

#index file
cmd = f"samtools index {BAM}"
print(cmd)
os.system(cmd)
```

```
samtools view data/aml035_post_transplant_possorted_genome_bam.bam -q 30 MT -b >
data/aml035_post_transplant_possorted_genome_bam.MT.bam
samtools index data/aml035_post_transplant_possorted_genome_bam.MT.bam
```

[6]:
```
0
```

2.5 3. Get barcode info from BAM files

[7]:
```
extract_barcode_info(BAM, BARCODE_INFO)
```

```
1475448it [00:17, 86474.29it/s]
```

[7]:
```
(defaultdict(int,
             {'AATGCGTGTTCTCA': 553,
              'ACGGATTGATCGAC': 7,
              'ACGTGATGTGCTTT': 14,
              'ACTCTCCTATTCCT': 783,
              'ATGGGTACTATGGC': 67,
              'ATTGGTCTACCGAT': 55,
              'ATTTGCACTCGCAA': 1118,
              'CATGTTACCTAAGC': 35,
              'CCAAGTGAGGTACT': 14,
              'CGCAGGACGGTAAA': 739,
              'GAGCTCCTGGAACG': 92,
              'GAGGTACTCCGATA': 453,
              'GAGTCTGAGATACC': 158,
              'GCGCATCTAACAGA': 12,
              'TAGGCTGATATCTC': 298,
              'TATGGGTGGATGAA': 476,
              'TCTAGTTGTGACTG': 60,
              'TCTCAAACCTAGCA': 515,
```

```
'AAGGCTACAACGAA': 334,
'ACCGCGGATGCTCC': 5,
'GGCTCACTCTCGAA': 1118,
'CCCTTACTCATTCT': 2,
'GAGGGCCGCTGTAG': 1,
'CTCCTACTGTTGTG': 14,
'GAGCATAGGAGGGT': 1,
'GCAGCCGATACTGG': 7,
'TACGACGAAGTGTC': 347,
'AATGTCCTGAACTC': 830,
'TAAGGCTGCTAAGC': 2111,
'ACGTCGCTTAGAAG': 4,
'ACTACGGACCTTGC': 910,
'AATAACACGCATAC': 26,
'GCACTGCTTTCGGA': 62,
'CCATATACACCATG': 131,
'CGACCGGATTTGTC': 311,
'ACGAACTGTCCAAG': 58,
'GGGGGGGGGGGGGG': 8655,
'CCAATGGAGGCATT': 298,
'AACGGTTGCCCGTT': 129,
'AATGTCCTCAACTA': 1,
'AGATATTGCGAGAG': 344,
'TGGAGACTGTAAAG': 1037,
'CAAGTTCTAACCGT': 27,
'GAAGCTTGTGTGGT': 404,
'GGCTCCCTCGCGAA': 3,
'AGCTGAACTAGCGT': 54,
'GAGTTGTGGTCACA': 47,
'AAATTCGAGACGAG': 251,
'ACGGCTCTGGACAG': 137,
'GGTGGAGAGACAGG': 983,
'GAGGGCCTGTGTAG': 1,
'CTCCACGAACCTCC': 3,
'GCCGTACTCCATGA': 1,
'ACGCTCACAACCTG': 39,
'TACGAGTGCTCTAT': 121,
'CCTTCACTCAGTCA': 133,
'CGCGATCTATCGAC': 8,
'GCTCCCTTTCACCC': 1,
'AAGAACGATTCATC': 57,
'CCCGCGGATGCGGC': 1,
'AATCGCAGAACAGA': 1,
'AATCGGTGAACCGT': 829,
'GGACTATGTATCTC': 144,
'TCAAGTCTGTATCG': 277,
'AAGGCTACAACGAC': 4,
```

```
'ACTCAGGAGGTATC': 125,
'AATGTAACAGCCCT': 1,
'GAGTCTGAGATCCC': 4,
'AATGGCTGCTAGTG': 429,
'GGTGGAGAGCCCGG': 6,
'TGAGGTACGGGCAA': 304,
'TAGAGCACCCATAG': 155,
'GACCGCCTCTGAAT': 1,
'CAATATGAATGTCG': 279,
'TGTGACGACAAAGA': 205,
'CACGGTTGCCCGTT': 1,
'AGAGCGGAAGTGTC': 159,
'AGGGCCACCCTGAA': 125,
'GCTTAACTTTGAGC': 20,
'CCCAAAGACGGGAA': 22,
'TATGGTCTAGACTC': 210,
'CAGCATGAGGTGTT': 12,
'CGCACACTCTCCAA': 1,
'TAGGACTGAACCAC': 148,
'CCCAAAGCCGGGAA': 1,
'TACCGCTGGAGCTT': 35,
'GAGGGCCTCTGTAG': 46,
'CTATTGACGTATCG': 175,
'GTTGGGGGGGGGCC': 1,
'AATGTAACACACTT': 3,
'GATACTCTTGAAGA': 896,
'TATGGTCTATAATC': 1,
'TGCAACGACAGTCA': 111,
'AGGGCCCCCCTGAA': 13,
'ACTTAGCTCAGAGG': 476,
'TGTCTAACTGGAAA': 32,
'CTCCACGACCCTCC': 1,
'AATGTAACAGCCAT': 351,
'GTAATAACCACCAA': 31,
'CAGAGGGCCAGCTC': 2,
'CAGAGGGACAGATC': 571,
'CACAGGGACAGATC': 5,
'CCCAGACTAGGAGC': 361,
'GGGGGGGGGGCCGG': 11,
'CTACTATGCCGCTT': 44,
'TATTTCCTAAGTAG': 61,
'AACAATACCTATTC': 546,
'GAGAAATGAGGGTG': 74,
'GAGAAAGGAGGGGG': 1,
'TTTCTACTCTGGAT': 72,
'GATTTGCTCGGTAT': 42,
'TGGTCAGACGACAT': 26,
```

```
'GCCCATACTTGAGC': 88,
'TACGCCACCTTACT': 17,
'CGATCTCGAGCGCG': 1,
'ACCTATTGCAACTG': 13,
'AATGTAACACACGT': 657,
'GCTACCTGTGCCAA': 762,
'TAGGACTGCACCCC': 1,
'GCAACCCTATTCTC': 27,
'GCCCTGCTTTCGGA': 1,
'GCCTGACTACTACG': 577,
'GGCGCATGGAAGGC': 366,
'TATTTCCTACCTGA': 198,
'AGATATTGCGACAG': 1,
'CTAACTACTTGTGG': 25,
'TCAACACTAGTCGT': 239,
'CCGGTAGGCCCCCG': 1,
'TATGGGGGGCTGCC': 1,
'GGCCATCATGTACC': 1,
'GGAACTACTGTCAG': 62,
'CACCACTGCAGCTA': 11,
'CAGGTATGAAAACG': 973,
'AACATGTGTTCCAG': 1,
'GTAACATGCGTTGC': 1,
'CCAATCCTGTGCAT': 1,
'GTAATATGCGTTGA': 25,
'TAGAGCACCCATAT': 3,
'TCAATAGAGAGCTT': 300,
'CTCGAAGACATTGG': 123,
'CCCATGTGTTGCAG': 345,
'TTCGCGGAGGTCGG': 1,
'TAGACCACCCATAG': 1,
'AATTCCTGCCTAAG': 255,
'TGATCCCTCTGACA': 5,
'TGATCACTCTGACA': 314,
'CATCCCGATGTTTC': 153,
'TATTTCCTTCCTGA': 2,
'TCAAAACTAGTCTT': 1,
'GAGTTGCTCGGGTG': 1,
'TATTAACTACCTGA': 1,
'ATTTGCACCCGAAT': 174,
'GCACGGTGAATCGC': 1571,
'TCATTAACCTGACA': 1,
'CGCACTACGTAGCT': 128,
'AATACTGAATTCGG': 4,
'TTGCATTGCCTTAT': 76,
'ATCTTTCTAAAAGC': 1,
'GCATGTGAAAGGCG': 124,
```

```
'GGCTGGGGGGGGTC': 1,
'GCCAAATGAGATGA': 243,
'TATGGTCTAGCCTC': 13,
'CTAATGCTCCCAAA': 138,
'TAAGGGCTTCATTC': 201,
'CCGGAGTGCTTGGA': 900,
'TGGAGCCTGTAAAG': 37,
'ATGTTAGACTTTAC': 247,
'CTAGGCCTTGTCGA': 64,
'CTAACGGACCCTAC': 313,
'CTAGAGACTGCTGA': 140,
'CCCGAACTCCAAGT': 357,
'GTATTGTTGTTGTT': 1,
'AGAGAACCACACAC': 1,
'AGATCGTGCTACGA': 1,
'CGCAAATGTTAGGC': 51,
'GGCACTCTTCCTTA': 465,
'CAGGTATGAACACG': 6,
'CCCTCACTATCGAA': 1,
'TGGAACTGCGTGAT': 243,
'ATAGTGCCCGCCTC': 1,
'AGAGTGCTTTTCAC': 25,
'AGCCACCTAGTCGT': 3197,
'AGGCCTCTGCTCCT': 296,
'CAGCATGATTTCAC': 100,
'CTAGTTACTATGCG': 227,
'GAAGCTACAGCGGA': 141,
'GCGGAGCTAGCTCA': 210,
'ACATTCTGAGCGTT': 683,
'TTCATCGACGTTGA': 161,
'TGGACTGATCCCAC': 99,
'CCCATTACGAACCT': 1,
'ACCCGTTGCAGATC': 68,
'CTCCTACTCGTACA': 420,
'GCGCCGGCGCCGCG': 1,
'ATCGCCACTCCTCG': 27,
'GGAGACACGAGACG': 1,
'TAAGCGTGTGGTTG': 76,
'AATGTACCACCCGT': 2,
'AGTCCAGATCTTAC': 177,
'TATGCGGAGATAAG': 238,
'TCATCAACCTCCCA': 12,
'AACGTGTGCTTATC': 40,
'CTCCTACTTCCGAA': 156,
'GGCTAATGTCGCCT': 179,
'GCTGATGACTTGGA': 159,
'GGAACTTGTAGCGT': 13,
'CAAGTTCTCACACA': 4,
'ATCCAGGCAGTAGA': 1,
'TCTGATACCAGTCA': 2,
'ACGCAATGGACGTT': 55,
'GAACGGGATCGATG': 30,
'CGTTCTGTACGCAA': 1,
```

2.5.1 Load barcode data

[8]:
```
[CR_read_number,CB_read_number,BC_read_number, barcodes, corrected_barcodes,␣
↪barcode_pairs] = pickle.load(open(BARCODE_INFO,"rb"))
```

[9]:
```
print('Number of CB (corrected) barcodes {}'.format(len(CB_read_number)))
print('Number of CR (uncorrected) barcodes {}'.format(len(CR_read_number)))
print('Number of BC (sample index) barcodes {}'.format(len(BC_read_number)))
BC_read_number
```

```
Number of CB (corrected) barcodes 44018
Number of CR (uncorrected) barcodes 172823
Number of BC (sample index) barcodes 12
```

[9]:
```
defaultdict(int,
            {'CGCAGGAG': 122082,
             'GCACCAGT': 123264,
             'ATACTGAG': 130876,
             'GATGCCTC': 147658,
             'GAATACTG': 100772,
             'TAGTACCA': 115213,
             'ATTGTTTC': 100875,
             'CGCGTGCA': 152216,
             'CGGAGACT': 145618,
             'TCCTATGA': 133150,
             'ATTCGTGC': 131056,
             'TCGACAAT': 72668})
```

2.6 4. Filter for reads with more than 200 bps

[10]:
```
count = 0
CB_filter = set()
for key in CB_read_number:
    if CB_read_number[key] >= 200:
        CB_filter.add(key)
        count += 1
print(count)
```

```
1426
```

[11]:
```
nucs = ["A", "T","C", "G"]
variants = ["3010", "9698"]

CB_df = pd.DataFrame(index=CB_filter, columns=["Number of␣
↪Reads"]+list(map(lambda x: "".join(x), product(variants,nucs))), dtype=int)
CB_df.loc[:,:] = 0
```

```
for i in CB_filter:
    CB_df.loc[i, "Number of Reads"] = CB_read_number[i]
CB_df
```

[11]:

	Number of Reads	3010A	3010T	3010C	3010G	9698A	9698T	9698C	9698G
ACCACCTGTTCCCG-3	256.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CATCGCTGTGGTAC-3	239.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CACTTAACTTATCC-1	208.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CTGGCACTTCTATC-3	335.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TATTGCTGTGCTGA-1	430.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
…	…	…	…	…	…	…	…	…	…
CTTAGACTTGAGCT-1	613.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GGACCTCTTAAGCC-2	211.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TTTCACGACTAGAC-1	351.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ACAAGAGACTCGCT-3	579.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GGAGTTACTGAAGA-3	215.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

[1426 rows x 9 columns]

2.7 Filter for reads with variants at 3010G and 9698T

2.7.1 fillmd compares to reference and puts = if the same. The variant at the position is added to the end of the file as V:letter , and this is used to filter for cells that do not have the reference

[12]:
```
cmd = f"""samtools view -b {BAM} MT:3010-3010 | samtools fillmd -e - {GENOME} |␣
↪grep -v "^@"| awk -v pos="3010" 'BEGIN {{OFS = FS = "\t" }} ;␣
↪{{n=split($10,a,"") ; if(a[(pos-$4)+1] != "=" ) print $0, "V:"␣
↪a[(pos-$4)+1]}}' > data/{NAME}_3010_reads.txt"""
cmd
os.system(cmd)
```

[12]: 0

[13]:
```
cmd = f"""samtools view -b {BAM} MT:9698-9698 | samtools fillmd -e - {GENOME} |␣
 ↪grep -v "^@"| awk -v pos="9698" 'BEGIN {{OFS = FS = "\t" }} ;␣
 ↪{{n=split($10,a,"") ; if(a[(pos-$4)+1] != "=" ) print $0, "V:"␣
 ↪a[(pos-$4)+1]}}' > data/{NAME}_9698_reads.txt"""
cmd
os.system(cmd)
```

[13]: 0

[27]:
```
nucs = ["A", "T","C", "G"]
variants = ["3010", "9698"]
ref_var = {"3010":"G", "9698":"C"}

CB_df = pd.DataFrame(index=CB_filter, columns=["Number of␣
 ↪Reads"]+list(map(lambda x: "".join(x), product(variants,nucs))), dtype=int)
CB_df.loc[:,:] = 0

for i in CB_filter:
    CB_df.loc[i, "Number of Reads"] = CB_read_number[i]
CB_df

rm_slash=False
for v in variants:
    print(f"data/{NAME}_{v}_reads.txt")
    with open(f"data/{NAME}_{v}_reads.txt", "r") as f:
        lines = list(map(lambda x: x.strip(), f.readlines()))
    #print(lines)
    for i in lines:
        if "CB:Z:" in i:
            if rm_slash:
                curr_bc = i.split("CB:Z:")[1].split("\t")[0].split("-")[0]
            else:
                curr_bc = i.split("CB:Z:")[1].split("\t")[0]

            ref_var
            if curr_bc in CB_df.index:
                if i[-1] == ":": #Then the reference should be it
                    CB_df.loc[curr_bc, v+ref_var[v]] += 1
                else:
                    CB_df.loc[curr_bc, v+i[-1]] += 1

CB_df
```

```
data/aml_035_3010_reads.txt
data/aml_035_9698_reads.txt
```

[27]:
```
                    Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
ACCACCTGTTCCCG-3              256.0    0.0    0.0    0.0    0.0    0.0    0.0
CATCGCTGTGGTAC-3              239.0    0.0    0.0    0.0    0.0    0.0    0.0
CACTTAACTTATCC-1              208.0    0.0    0.0    0.0    0.0    0.0    0.0
CTGGCACTTCTATC-3              335.0    0.0    0.0    0.0    0.0    0.0    0.0
TATTGCTGTGCTGA-1              430.0    0.0    0.0    0.0    0.0    0.0    0.0
…                               …      …      …      …      …      …      …
CTTAGACTTGAGCT-1              613.0    0.0    0.0    0.0    0.0    0.0    0.0
GGACCTCTTAAGCC-2              211.0    0.0    0.0    0.0    0.0    0.0    0.0
TTTCACGACTAGAC-1              351.0    0.0    0.0    0.0    0.0    0.0    0.0
ACAAGAGACTCGCT-3              579.0    0.0    0.0    0.0    0.0    0.0    0.0
GGAGTTACTGAAGA-3              215.0    0.0    0.0    0.0    0.0    0.0    0.0

                    9698C  9698G
ACCACCTGTTCCCG-3      7.0    0.0
CATCGCTGTGGTAC-3      7.0    0.0
CACTTAACTTATCC-1      6.0    0.0
CTGGCACTTCTATC-3      8.0    0.0
TATTGCTGTGCTGA-1     20.0    0.0
…                     …      …
CTTAGACTTGAGCT-1     17.0    0.0
GGACCTCTTAAGCC-2      4.0    0.0
TTTCACGACTAGAC-1     10.0    0.0
ACAAGAGACTCGCT-3     16.0    0.0
GGAGTTACTGAAGA-3      7.0    0.0

[1426 rows x 9 columns]
```

[26]:
```
#3010G->A and 9698T>C
```

2.8 Cells with different variants

[28]:
```
CB_df[(CB_df["9698T"] > 0) & (CB_df["9698C"] > 0)]
```

[28]:
```
                    Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
CATCATACCCAACA-3              464.0    0.0    0.0    0.0    0.0    0.0    1.0

                    9698C  9698G
CATCATACCCAACA-3     14.0    0.0
```

[29]:
```
CB_df[(CB_df["9698T"] > 0)]
```

[29]:
```
                    Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
CATCATACCCAACA-3              464.0    0.0    0.0    0.0    0.0    0.0    1.0

                    9698C  9698G
```

```
CATCATACCCAACA-3   14.0    0.0
```

[30]:
```
CB_df[(CB_df["9698C"] > 0)]
```

[30]:
```
                  Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
ACCACCTGTTCCCG-3            256.0    0.0    0.0    0.0    0.0    0.0    0.0
CATCGCTGTGGTAC-3            239.0    0.0    0.0    0.0    0.0    0.0    0.0
CACTTAACTTATCC-1            208.0    0.0    0.0    0.0    0.0    0.0    0.0
CTGGCACTTCTATC-3            335.0    0.0    0.0    0.0    0.0    0.0    0.0
TATTGCTGTGCTGA-1            430.0    0.0    0.0    0.0    0.0    0.0    0.0
…                             …      …      …      …      …      …      …
CTTAGACTTGAGCT-1            613.0    0.0    0.0    0.0    0.0    0.0    0.0
GGACCTCTTAAGCC-2            211.0    0.0    0.0    0.0    0.0    0.0    0.0
TTTCACGACTAGAC-1            351.0    0.0    0.0    0.0    0.0    0.0    0.0
ACAAGAGACTCGCT-3            579.0    0.0    0.0    0.0    0.0    0.0    0.0
GGAGTTACTGAAGA-3            215.0    0.0    0.0    0.0    0.0    0.0    0.0

                  9698C  9698G
ACCACCTGTTCCCG-3    7.0    0.0
CATCGCTGTGGTAC-3    7.0    0.0
CACTTAACTTATCC-1    6.0    0.0
CTGGCACTTCTATC-3    8.0    0.0
TATTGCTGTGCTGA-1   20.0    0.0
…                   …      …
CTTAGACTTGAGCT-1   17.0    0.0
GGACCTCTTAAGCC-2    4.0    0.0
TTTCACGACTAGAC-1   10.0    0.0
ACAAGAGACTCGCT-3   16.0    0.0
GGAGTTACTGAAGA-3    7.0    0.0

[1411 rows x 9 columns]
```

[31]:
```
CB_df[(CB_df["9698C"] > 0) & (CB_df["3010G"] > 0)]
```

[31]:
```
                  Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
ATGAAACTCCAAGT-2            579.0    0.0    0.0    0.0    7.0    0.0    0.0
CTTAGGGACTCTAT-2            753.0    0.0    0.0    0.0    6.0    0.0    0.0
GGCCGAACTATCGG-1            287.0    0.0    0.0    0.0    3.0    0.0    0.0
TAGGTCGAAGGGTG-3           1434.0    0.0    0.0    0.0    3.0    0.0    0.0
CTAACGGAAGAAGT-2            398.0    0.0    0.0    0.0    3.0    0.0    0.0
TCCCAGACATCTTC-2            685.0    0.0    0.0    0.0    3.0    0.0    0.0
TGACCAGATGGTGT-2            201.0    0.0    0.0    0.0    1.0    0.0    0.0
ACCCAAGATGTCAG-3            725.0    0.0    0.0    0.0    1.0    0.0    0.0
TCGAGCCTGTTGAC-1            677.0    0.0    0.0    0.0    3.0    0.0    0.0
ACGTTTACACCATG-2            240.0    1.0    0.0    0.0    1.0    0.0    0.0
GATTCTTGGTTCTT-1           1554.0    0.0    0.0    0.0    3.0    0.0    0.0
ATTCGGGAAGTGTC-3            754.0    0.0    0.0    0.0    2.0    0.0    0.0
```

```
ATACCACTATGCCA-2           495.0    0.0    0.0    0.0    3.0    0.0    0.0
CATACTTGCTTAGG-3          1296.0    0.0    0.0    0.0    1.0    0.0    0.0

                    9698C  9698G
ATGAAACTCCAAGT-2      9.0    0.0
CTTAGGGACTCTAT-2     23.0    0.0
GGCCGAACTATCGG-1      9.0    0.0
TAGGTCGAAGGGTG-3     39.0    0.0
CTAACGGAAGAAGT-2      5.0    0.0
TCCCAGACATCTTC-2     12.0    0.0
TGACCAGATGGTGT-2      1.0    0.0
ACCCAAGATGTCAG-3     16.0    0.0
TCGAGCCTGTTGAC-1      4.0    0.0
ACGTTTACACCATG-2      6.0    0.0
GATTCTTGGTTCTT-1     31.0    0.0
ATTCGGGAAGTGTC-3     11.0    0.0
ATACCACTATGCCA-2      5.0    0.0
CATACTTGCTTAGG-3     33.0    0.0
```

[32]:
```
CB_df[(CB_df["9698C"] > 0) & (CB_df["3010A"] > 0)]
```

[32]:
```
                    Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
AGTACGTGAAGATG-2              234.0    1.0    0.0    0.0    0.0    0.0    0.0
GCGAAGGAGGTAGG-1              376.0    1.0    0.0    0.0    0.0    0.0    0.0
TAACAATGCTACCC-2             1237.0    2.0    0.0    0.0    0.0    0.0    0.0
GCCGAGTGTGCATG-2              999.0    2.0    0.0    0.0    0.0    0.0    0.0
AATGATACCTGTGA-3             1009.0    2.0    0.0    0.0    0.0    0.0    0.0
CTAGTTTGAGCGTT-2              269.0    1.0    0.0    0.0    0.0    0.0    0.0
GATATCCTCCAACA-2              226.0    4.0    0.0    0.0    0.0    0.0    0.0
GGGAACGATCGACA-2              283.0    1.0    0.0    0.0    0.0    0.0    0.0
CTCGAAGACCTACC-1              455.0    4.0    0.0    0.0    0.0    0.0    0.0
ACGTTTACACCATG-2              240.0    1.0    0.0    0.0    1.0    0.0    0.0
AGCCACCTAGTCGT-1             3630.0    1.0    0.0    0.0    0.0    0.0    0.0
TATAGCCTCTCAGA-1              332.0    2.0    0.0    0.0    0.0    0.0    0.0
GCGCGAACGATACC-2             1091.0    4.0    0.0    0.0    0.0    0.0    0.0
CGACCACTTAGAGA-2              449.0    3.0    0.0    0.0    0.0    0.0    0.0

                    9698C  9698G
AGTACGTGAAGATG-2      4.0    0.0
GCGAAGGAGGTAGG-1     21.0    0.0
TAACAATGCTACCC-2     22.0    0.0
GCCGAGTGTGCATG-2     15.0    0.0
AATGATACCTGTGA-3     25.0    0.0
CTAGTTTGAGCGTT-2      3.0    0.0
GATATCCTCCAACA-2      6.0    0.0
GGGAACGATCGACA-2      3.0    0.0
CTCGAAGACCTACC-1      5.0    0.0
```

```
ACGTTTACACCATG-2    6.0    0.0
AGCCACCTAGTCGT-1   52.0    0.0
TATAGCCTCTCAGA-1    3.0    0.0
GCGCGAACGATACC-2    2.0    0.0
CGACCACTTAGAGA-2    9.0    0.0
```

[33]:
```
CB_df[((CB_df["9698C"] == 0) & (CB_df["9698T"] > 0)) | ((CB_df["3010G"] == 0) &␣
↪(CB_df["3010A"] > 0)) |
    ((CB_df["9698C"] > 0) & (CB_df["9698T"] == 0)) | ((CB_df["3010G"] > 0) &␣
↪(CB_df["3010A"] == 0))]
```

[33]:
```
                  Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
ACCACCTGTTCCCG-3            256.0    0.0    0.0    0.0    0.0    0.0    0.0
CATCGCTGTGGTAC-3            239.0    0.0    0.0    0.0    0.0    0.0    0.0
CACTTAACTTATCC-1            208.0    0.0    0.0    0.0    0.0    0.0    0.0
CTGGCACTTCTATC-3            335.0    0.0    0.0    0.0    0.0    0.0    0.0
TATTGCTGTGCTGA-1            430.0    0.0    0.0    0.0    0.0    0.0    0.0
…                             …      …      …      …      …      …      …
CTTAGACTTGAGCT-1            613.0    0.0    0.0    0.0    0.0    0.0    0.0
GGACCTCTTAAGCC-2            211.0    0.0    0.0    0.0    0.0    0.0    0.0
TTTCACGACTAGAC-1            351.0    0.0    0.0    0.0    0.0    0.0    0.0
ACAAGAGACTCGCT-3            579.0    0.0    0.0    0.0    0.0    0.0    0.0
GGAGTTACTGAAGA-3            215.0    0.0    0.0    0.0    0.0    0.0    0.0

                  9698C  9698G
ACCACCTGTTCCCG-3    7.0    0.0
CATCGCTGTGGTAC-3    7.0    0.0
CACTTAACTTATCC-1    6.0    0.0
CTGGCACTTCTATC-3    8.0    0.0
TATTGCTGTGCTGA-1   20.0    0.0
…                    …      …
CTTAGACTTGAGCT-1   17.0    0.0
GGACCTCTTAAGCC-2    4.0    0.0
TTTCACGACTAGAC-1   10.0    0.0
ACAAGAGACTCGCT-3   16.0    0.0
GGAGTTACTGAAGA-3    7.0    0.0

[1413 rows x 9 columns]
```

[34]:
```
CB_df[((CB_df["9698C"] > 0) | (CB_df["9698T"] > 0)) &  ((CB_df["3010G"] > 0) |␣
↪(CB_df["3010A"] > 0))]
```

[34]:
```
                  Number of Reads  3010A  3010T  3010C  3010G  9698A  9698T  \
ATGAAACTCCAAGT-2            579.0    0.0    0.0    0.0    7.0    0.0    0.0
AGTACGTGAAGATG-2            234.0    1.0    0.0    0.0    0.0    0.0    0.0
GCGAAGGAGGTAGG-1            376.0    1.0    0.0    0.0    0.0    0.0    0.0
TAACAATGCTACCC-2           1237.0    2.0    0.0    0.0    0.0    0.0    0.0
```

CTTAGGGACTCTAT-2	753.0	0.0	0.0	0.0	6.0	0.0	0.0
GCCGAGTGTGCATG-2	999.0	2.0	0.0	0.0	0.0	0.0	0.0
GGCCGAACTATCGG-1	287.0	0.0	0.0	0.0	3.0	0.0	0.0
TAGGTCGAAGGGTG-3	1434.0	0.0	0.0	0.0	3.0	0.0	0.0
AATGATACCTGTGA-3	1009.0	2.0	0.0	0.0	0.0	0.0	0.0
CTAACGGAAGAAGT-2	398.0	0.0	0.0	0.0	3.0	0.0	0.0
TCCCAGACATCTTC-2	685.0	0.0	0.0	0.0	3.0	0.0	0.0
TGACCAGATGGTGT-2	201.0	0.0	0.0	0.0	1.0	0.0	0.0
CTAGTTTGAGCGTT-2	269.0	1.0	0.0	0.0	0.0	0.0	0.0
ACCCAAGATGTCAG-3	725.0	0.0	0.0	0.0	1.0	0.0	0.0
GATATCCTCCAACA-2	226.0	4.0	0.0	0.0	0.0	0.0	0.0
GGGAACGATCGACA-2	283.0	1.0	0.0	0.0	0.0	0.0	0.0
CTCGAAGACCTACC-1	455.0	4.0	0.0	0.0	0.0	0.0	0.0
TCGAGCCTGTTGAC-1	677.0	0.0	0.0	0.0	3.0	0.0	0.0
ACGTTTACACCATG-2	240.0	1.0	0.0	0.0	1.0	0.0	0.0
GATTCTTGGTTCTT-1	1554.0	0.0	0.0	0.0	3.0	0.0	0.0
AGCCACCTAGTCGT-1	3630.0	1.0	0.0	0.0	0.0	0.0	0.0
ATTCGGGAAGTGTC-3	754.0	0.0	0.0	0.0	2.0	0.0	0.0
TATAGCCTCTCAGA-1	332.0	2.0	0.0	0.0	0.0	0.0	0.0
GCGCGAACGATACC-2	1091.0	4.0	0.0	0.0	0.0	0.0	0.0
CGACCACTTAGAGA-2	449.0	3.0	0.0	0.0	0.0	0.0	0.0
ATACCACTATGCCA-2	495.0	0.0	0.0	0.0	3.0	0.0	0.0
CATACTTGCTTAGG-3	1296.0	0.0	0.0	0.0	1.0	0.0	0.0

	9698C	9698G
ATGAAACTCCAAGT-2	9.0	0.0
AGTACGTGAAGATG-2	4.0	0.0
GCGAAGGAGGTAGG-1	21.0	0.0
TAACAATGCTACCC-2	22.0	0.0
CTTAGGGACTCTAT-2	23.0	0.0
GCCGAGTGTGCATG-2	15.0	0.0
GGCCGAACTATCGG-1	9.0	0.0
TAGGTCGAAGGGTG-3	39.0	0.0
AATGATACCTGTGA-3	25.0	0.0
CTAACGGAAGAAGT-2	5.0	0.0
TCCCAGACATCTTC-2	12.0	0.0
TGACCAGATGGTGT-2	1.0	0.0
CTAGTTTGAGCGTT-2	3.0	0.0
ACCCAAGATGTCAG-3	16.0	0.0
GATATCCTCCAACA-2	6.0	0.0
GGGAACGATCGACA-2	3.0	0.0
CTCGAAGACCTACC-1	5.0	0.0
TCGAGCCTGTTGAC-1	4.0	0.0
ACGTTTACACCATG-2	6.0	0.0
GATTCTTGGTTCTT-1	31.0	0.0
AGCCACCTAGTCGT-1	52.0	0.0
ATTCGGGAAGTGTC-3	11.0	0.0

```
TATAGCCTCTCAGA-1    3.0    0.0
GCGCGAACGATACC-2    2.0    0.0
CGACCACTTAGAGA-2    9.0    0.0
ATACCACTATGCCA-2    5.0    0.0
CATACTTGCTTAGG-3   33.0    0.0
```

[36]:
```
len(CB_df[((CB_df["9698C"] > 0) | (CB_df["9698T"] > 0)) &  ((CB_df["3010G"] >␣
↪0) | (CB_df["3010A"] > 0))])
```

[36]:
```
27
```

[37]:
```
CB_df[((CB_df["9698C"] > 0) | (CB_df["9698T"] > 0) | (CB_df["9698A"] > 0) |␣
↪(CB_df["9698G"] > 0)) &
     ((CB_df["3010C"] > 0) | (CB_df["3010T"] > 0) | (CB_df["3010A"] > 0) |␣
↪(CB_df["3010G"] > 0))]
```

[37]:

	Number of Reads	3010A	3010T	3010C	3010G	9698A	9698T	\
ATGAAACTCCAAGT-2	579.0	0.0	0.0	0.0	7.0	0.0	0.0	
ACCTCCGAGGAGGT-3	451.0	0.0	0.0	1.0	0.0	0.0	0.0	
AGTTCTACTCAAGC-2	736.0	0.0	0.0	2.0	0.0	0.0	0.0	
CCCGAACTCCAAGT-2	387.0	0.0	5.0	0.0	0.0	0.0	0.0	
TACGCCACGTAAAG-1	554.0	0.0	0.0	1.0	0.0	0.0	0.0	
AATGTCCTGTCTTT-1	509.0	0.0	0.0	1.0	0.0	0.0	0.0	
AGTACGTGAAGATG-2	234.0	1.0	0.0	0.0	0.0	0.0	0.0	
TAATGAACATTGGC-3	448.0	0.0	0.0	1.0	0.0	0.0	0.0	
GCGAAGGAGGTAGG-1	376.0	1.0	0.0	0.0	0.0	0.0	0.0	
ACCCAGCTAACGTC-1	260.0	0.0	1.0	0.0	0.0	0.0	0.0	
ACTACGGACCTTGC-3	1012.0	0.0	0.0	1.0	0.0	0.0	0.0	
TAACAATGCTACCC-2	1237.0	2.0	0.0	0.0	0.0	0.0	0.0	
CTTAGGGACTCTAT-2	753.0	0.0	0.0	0.0	6.0	0.0	0.0	
GCCGAGTGTGCATG-2	999.0	2.0	0.0	0.0	0.0	0.0	0.0	
ATTTGCACTCGCAA-2	1185.0	0.0	0.0	1.0	0.0	0.0	0.0	
GGCCGAACTATCGG-1	287.0	0.0	0.0	0.0	3.0	0.0	0.0	
ATCGAGTGAATCGC-2	661.0	0.0	3.0	0.0	0.0	0.0	0.0	
TAGGTCGAAGGGTG-3	1434.0	0.0	0.0	0.0	3.0	0.0	0.0	
CGCAGGACGGTAAA-1	797.0	0.0	0.0	1.0	0.0	0.0	0.0	
AATGATACCTGTGA-3	1009.0	2.0	0.0	0.0	0.0	0.0	0.0	
ATGTTCACACGCTA-1	635.0	0.0	0.0	1.0	0.0	0.0	0.0	
AGTCTACTTCAGAC-1	699.0	0.0	0.0	1.0	0.0	0.0	0.0	
CTAACGGAAGAAGT-2	398.0	0.0	0.0	0.0	3.0	0.0	0.0	
TTAGACCTGTAGCT-2	270.0	0.0	0.0	1.0	0.0	0.0	0.0	
GAAGGTCTGTATGC-2	610.0	0.0	1.0	0.0	0.0	0.0	0.0	
ATGTCACTTTGACG-1	369.0	0.0	0.0	1.0	0.0	0.0	0.0	
CTTGATTGTGTAGC-3	361.0	0.0	0.0	1.0	0.0	0.0	0.0	
TCCCAGACATCTTC-2	685.0	0.0	0.0	0.0	3.0	0.0	0.0	
TCAGCAGATTCATC-2	333.0	0.0	0.0	1.0	0.0	0.0	0.0	
AATTCCTGTAACGC-3	227.0	0.0	1.0	0.0	0.0	0.0	0.0	

TGACCAGATGGTGT-2	201.0	0.0	0.0	0.0	1.0	0.0	0.0
CTAGTTTGAGCGTT-2	269.0	1.0	0.0	0.0	0.0	0.0	0.0
GACGTCCTGAATAG-1	205.0	0.0	0.0	1.0	0.0	0.0	0.0
ACCCAAGATGTCAG-3	725.0	0.0	0.0	0.0	1.0	0.0	0.0
CTCGACACTGGAAA-1	1205.0	0.0	0.0	1.0	0.0	0.0	0.0
GATATCCTCCAACA-2	226.0	4.0	0.0	0.0	0.0	0.0	0.0
GGGAACGATCGACA-2	283.0	1.0	0.0	0.0	0.0	0.0	0.0
CTCGAAGACCTACC-1	455.0	4.0	0.0	0.0	0.0	0.0	0.0
AATGTCCTGAACTC-2	909.0	0.0	0.0	1.0	0.0	0.0	0.0
TCGAGCCTGTTGAC-1	677.0	0.0	0.0	0.0	3.0	0.0	0.0
CAGGTATGAAAACG-2	1035.0	0.0	0.0	1.0	0.0	0.0	0.0
AATCCTTGACGTGT-1	1221.0	0.0	0.0	1.0	0.0	0.0	0.0
ATATACGAGACACT-3	534.0	0.0	1.0	0.0	0.0	0.0	0.0
ACGTTTACACCATG-2	240.0	1.0	0.0	0.0	1.0	0.0	0.0
CCCAGTTGTTTGTC-3	592.0	0.0	0.0	1.0	0.0	0.0	0.0
GATTCTTGGTTCTT-1	1554.0	0.0	0.0	0.0	3.0	0.0	0.0
GACTGATGTATCTC-1	654.0	0.0	0.0	1.0	0.0	0.0	0.0
CACAATCTCTTCGC-3	1212.0	0.0	1.0	0.0	0.0	0.0	0.0
GGCCCAGAAAACGA-2	679.0	0.0	3.0	0.0	0.0	0.0	0.0
AGCCACCTAGTCGT-1	3630.0	1.0	0.0	0.0	0.0	0.0	0.0
ATTCGGGAAGTGTC-3	754.0	0.0	0.0	0.0	2.0	0.0	0.0
GCAACTGAGATACC-2	469.0	0.0	4.0	0.0	0.0	0.0	0.0
ATCCATACCTTATC-3	331.0	0.0	0.0	1.0	0.0	0.0	0.0
TATAGCCTCTCAGA-1	332.0	2.0	0.0	0.0	0.0	0.0	0.0
ACTGAGACCTAGAC-3	223.0	0.0	1.0	0.0	0.0	0.0	0.0
GGACTATGTCCTTA-1	427.0	0.0	1.0	0.0	0.0	0.0	0.0
GCGCGAACGATACC-2	1091.0	4.0	0.0	0.0	0.0	0.0	0.0
CGACCACTTAGAGA-2	449.0	3.0	0.0	0.0	0.0	0.0	0.0
ATACCACTATGCCA-2	495.0	0.0	0.0	0.0	3.0	0.0	0.0
CATACTTGCTTAGG-3	1296.0	0.0	0.0	0.0	1.0	0.0	0.0

	9698C	9698G
ATGAAACTCCAAGT-2	9.0	0.0
ACCTCCGAGGAGGT-3	11.0	0.0
AGTTCTACTCAAGC-2	19.0	0.0
CCCGAACTCCAAGT-2	7.0	0.0
TACGCCACGTAAAG-1	8.0	0.0
AATGTCCTGTCTTT-1	22.0	0.0
AGTACGTGAAGATG-2	4.0	0.0
TAATGAACATTGGC-3	24.0	0.0
GCGAAGGAGGTAGG-1	21.0	0.0
ACCCAGCTAACGTC-1	2.0	0.0
ACTACGGACCTTGC-3	22.0	0.0
TAACAATGCTACCC-2	22.0	0.0
CTTAGGGACTCTAT-2	23.0	0.0
GCCGAGTGTGCATG-2	15.0	0.0
ATTTGCACTCGCAA-2	10.0	0.0

```
GGCCGAACTATCGG-1    9.0    0.0
ATCGAGTGAATCGC-2    9.0    0.0
TAGGTCGAAGGGTG-3   39.0    0.0
CGCAGGACGGTAAA-1   14.0    0.0
AATGATACCTGTGA-3   25.0    0.0
ATGTTCACACGCTA-1   28.0    0.0
AGTCTACTTCAGAC-1   26.0    0.0
CTAACGGAAGAAGT-2    5.0    0.0
TTAGACCTGTAGCT-2    4.0    0.0
GAAGGTCTGTATGC-2    9.0    0.0
ATGTCACTTTGACG-1   16.0    0.0
CTTGATTGTGTAGC-3    8.0    0.0
TCCCAGACATCTTC-2   12.0    0.0
TCAGCAGATTCATC-2    3.0    0.0
AATTCCTGTAACGC-3    1.0    0.0
TGACCAGATGGTGT-2    1.0    0.0
CTAGTTTGAGCGTT-2    3.0    0.0
GACGTCCTGAATAG-1    1.0    0.0
ACCCAAGATGTCAG-3   16.0    0.0
CTCGACACTGGAAA-1   22.0    0.0
GATATCCTCCAACA-2    6.0    0.0
GGGAACGATCGACA-2    3.0    0.0
CTCGAAGACCTACC-1    5.0    0.0
AATGTCCTGAACTC-2   14.0    0.0
TCGAGCCTGTTGAC-1    4.0    0.0
CAGGTATGAAAACG-2   10.0    0.0
AATCCTTGACGTGT-1   21.0    0.0
ATATACGAGACACT-3   16.0    0.0
ACGTTTACACCATG-2    6.0    0.0
CCCAGTTGTTTGTC-3    6.0    0.0
GATTCTTGGTTCTT-1   31.0    0.0
GACTGATGTATCTC-1   17.0    0.0
CACAATCTCTTCGC-3   25.0    0.0
GGCCCAGAAAACGA-2    9.0    0.0
AGCCACCTAGTCGT-1   52.0    0.0
ATTCGGGAAGTGTC-3   11.0    0.0
GCAACTGAGATACC-2    5.0    0.0
ATCCATACCTTATC-3    3.0    0.0
TATAGCCTCTCAGA-1    3.0    0.0
ACTGAGACCTAGAC-3    8.0    0.0
GGACTATGTCCTTA-1    2.0    0.0
GCGCGAACGATACC-2    2.0    0.0
CGACCACTTAGAGA-2    9.0    0.0
ATACCACTATGCCA-2    5.0    0.0
CATACTTGCTTAGG-3   33.0    0.0
```

[38]:
```
len(CB_df[((CB_df["9698C"] > 0) | (CB_df["9698T"] > 0) | (CB_df["9698A"] > 0) |␣
↪(CB_df["9698G"] > 0)) &
     ((CB_df["3010C"] > 0) | (CB_df["3010T"] > 0) | (CB_df["3010A"] > 0) |␣
↪(CB_df["3010G"] > 0))])
```

[38]:
```
60
```