mutations_distribution-chroms

August 26, 2018

1 Plotting the mutations density through the chromosomes

This notebook shows how to plot mutation density per chromosome alongside marks for centromeres and chromosome boundaries.

```
In [1]: %matplotlib inline

        import matplotlib.pyplot as plt
        import seaborn as sns

        sns.set()
```

We want to plot the distribution of the mutations along the chromosomes, so, we first read the positions of the mutations (read from a random sample of 100,000 mutations)

```
In [2]: from collections import defaultdict
        from ICGC_data_parser import SSM_Reader

        distribution = defaultdict(list)

        for record in SSM_Reader(filename='data/ssm_sample.vcf'):
            # Associate CHROMOSOME -> [MUTATION POSITIONS]
            distribution[record.CHROM].append(record.POS)
```

We want to add information of the positions of the centromeric regions and the chromosome boundaries. We read this from the table `data/chromosome-data.tsv`

```
In [3]: from collections import namedtuple

        # Create a custom class whose objects
        # hold information of a chromosome
        Chromosome = namedtuple('Chromosome', ['length',
                                               'centromere_start',
                                               'centromere_end'])
```

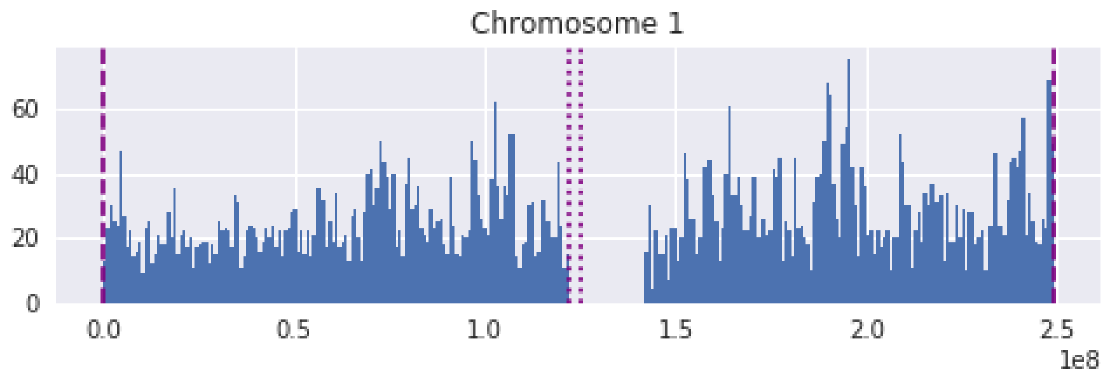
```
In [4]: import pandas as pd

        # Open the file with the information of the centromeric regions
        all_data = pd.read_table('data/chromosome-data.tsv',
                                 delimiter='\t')
        # Filter for human data
        human_data = all_data[ all_data['species'] == 'Homo sapiens' ]

        chromosomes = {}
        for _, record in human_data.iterrows():
            chrom = record['chromosome']
            length = record['chromosome length (bp)']
            c_start = record['centromeric region start']
            c_end = record['centromeric region end']

            chromosomes[chrom] = Chromosome(length, c_start, c_end)
```

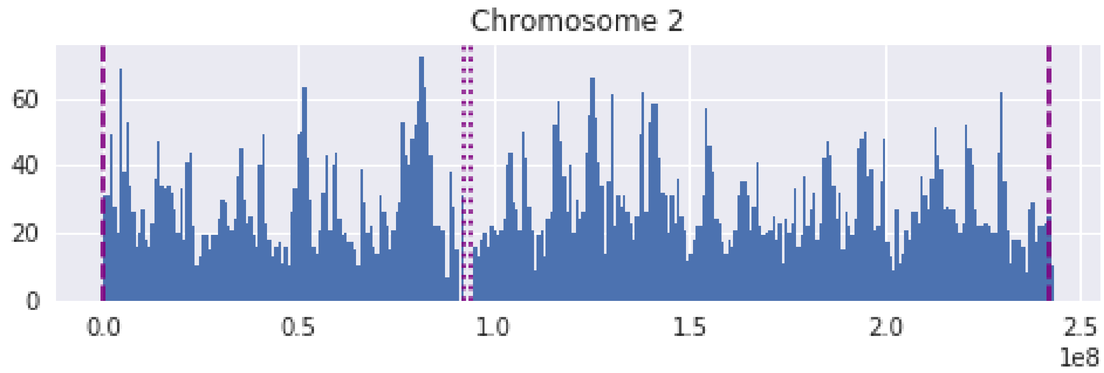
To ensure the chromosomes are plotted in the correct order, we provide a list that defines that order

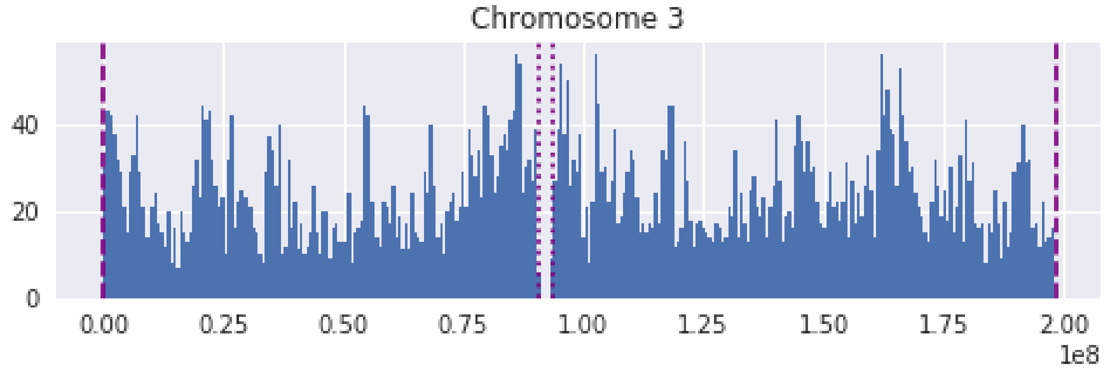
```
In [7]: chrom_names = [str(i+1) for i in range(22)] + ['X', 'Y', 'MT']
```

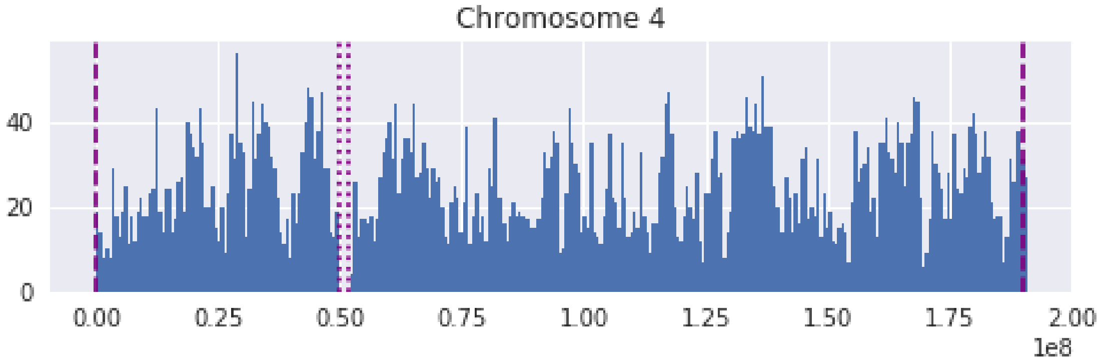
Finally, we can plot the mutations

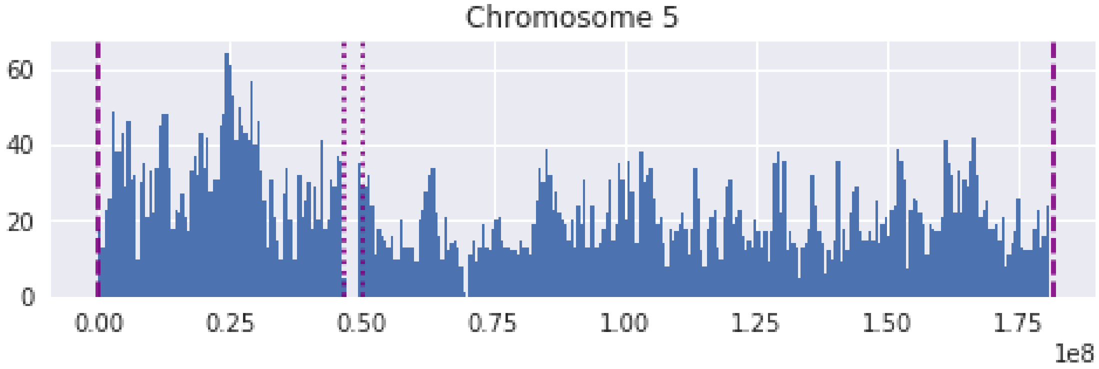
```
In [9]: for chrom in chrom_names:
            fig, ax = plt.subplots(figsize=(8, 2))

            # Main plot
            ax.hist(distribution[chrom], bins=300)
            ax.set(title=f'Chromosome {chrom}')

            if chrom in chromosomes:
                # Fetch data on chromosome
                # length and centromere positions
                chrom_data = chromosomes[chrom]

                # Chromosome boundaries
                ax.axvline(chrom_data.length, ls='--', color='purple')
                ax.axvline(0, ls='--', color='purple')

                # Chromosome centromeres
                ax.axvline(chrom_data.centromere_end, ls=':', color='purple')
                ax.axvline(chrom_data.centromere_start, ls=':', color='purple')

            plt.show()
```

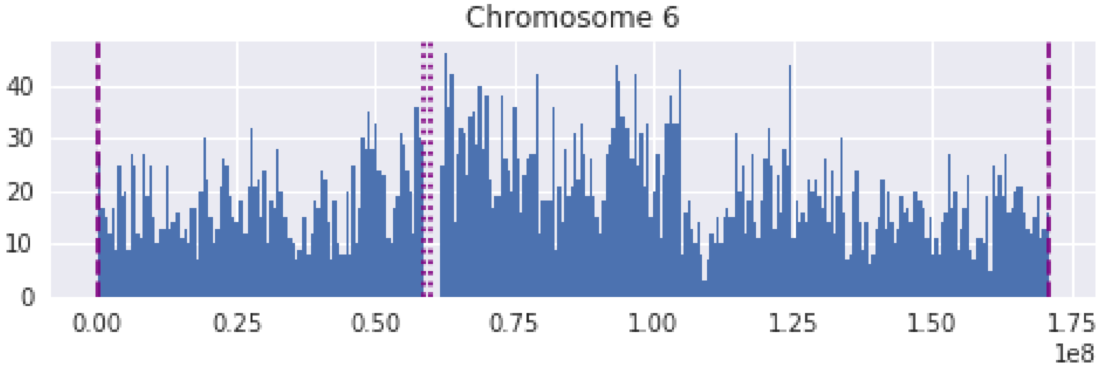

Chromosome 1
0
20
40
60
0.0
0.5
1.0
1.5
2.0
2.5
1e8

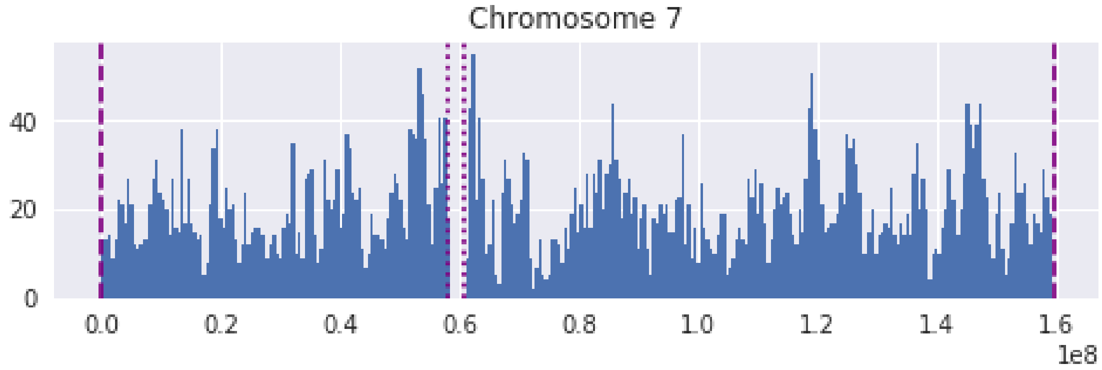

Chromosome 2
0
20
40
60
0.0
0.5
1.0
1.5
2.0
2.5
1e8

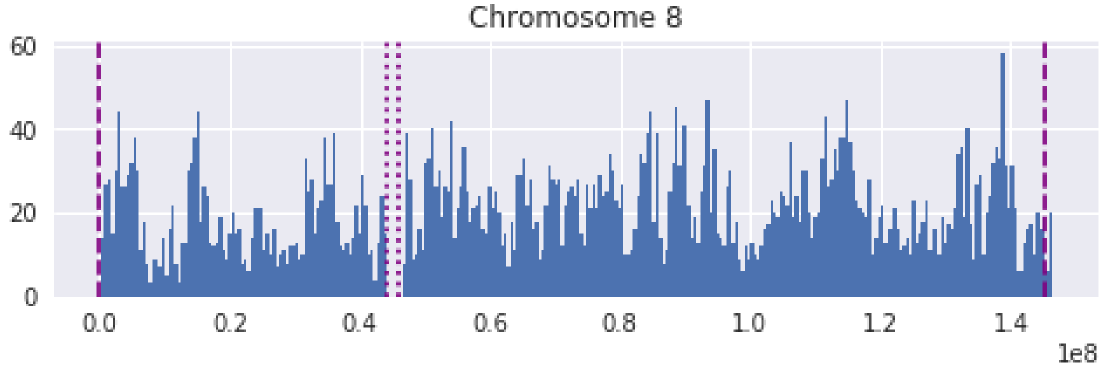

Chromosome 3
0
20
40
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e8

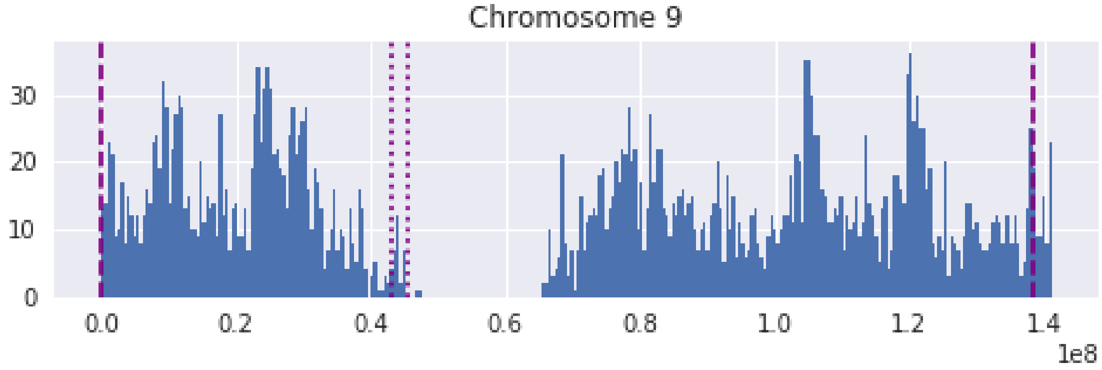

Chromosome 4
0
20
40
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e8

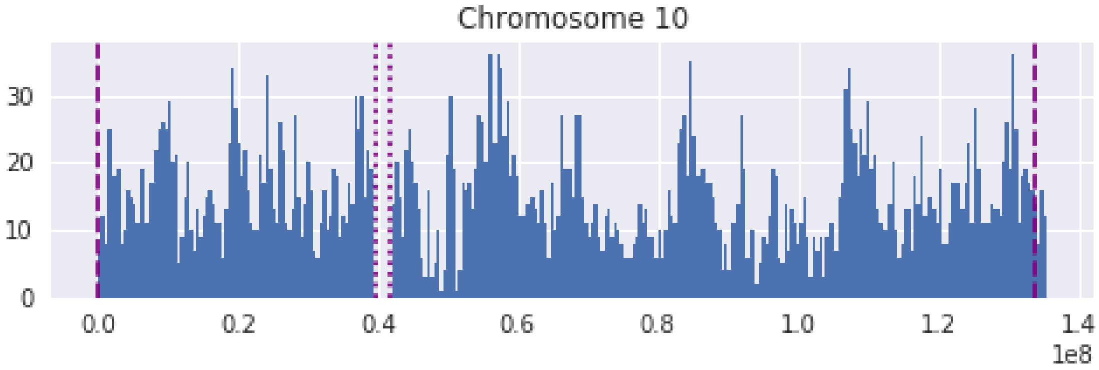

Chromosome 5
0
20
40
60
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e8

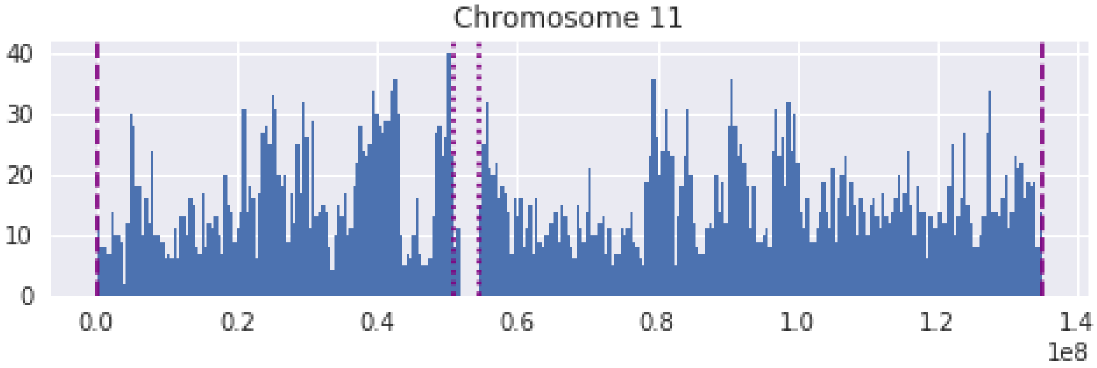

Chromosome 6
0
10
20
30
40
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e8

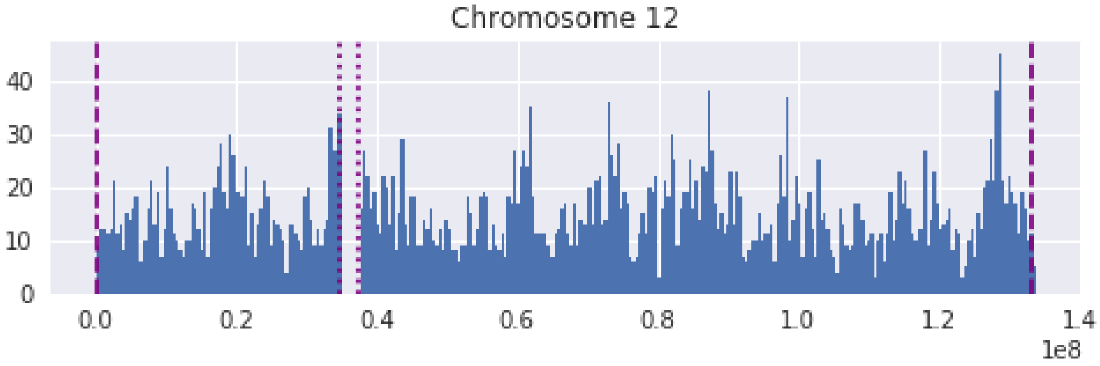

Chromosome 7
0
20
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e8

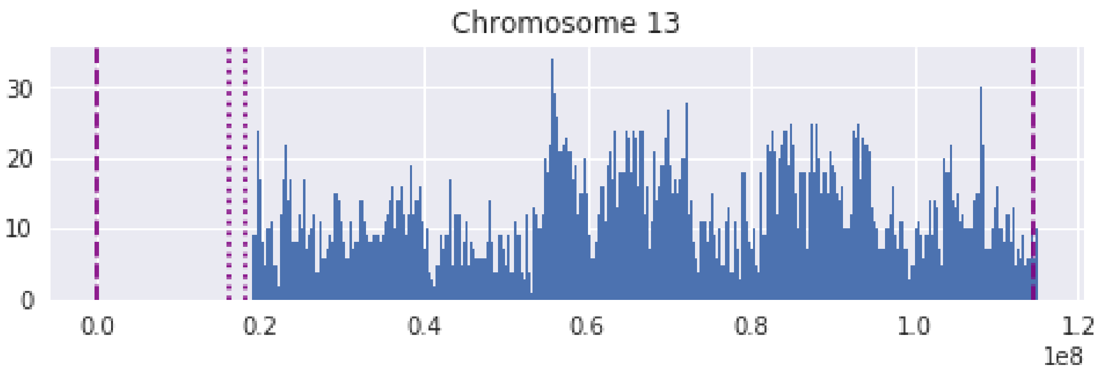

Chromosome 8
0
20
40
60
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e8


Chromosome 9
0
10
20
30
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e8

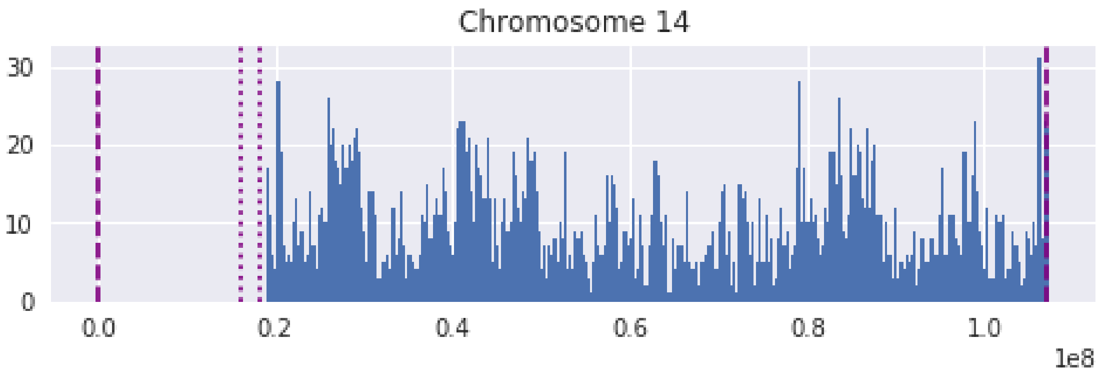

Chromosome 10
0
10
20
30
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e8


Chromosome 11
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e8

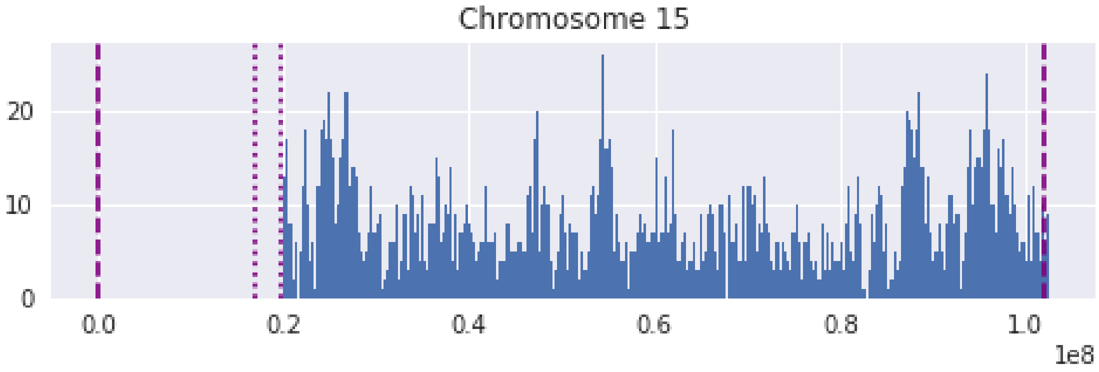

Chromosome 12
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e8



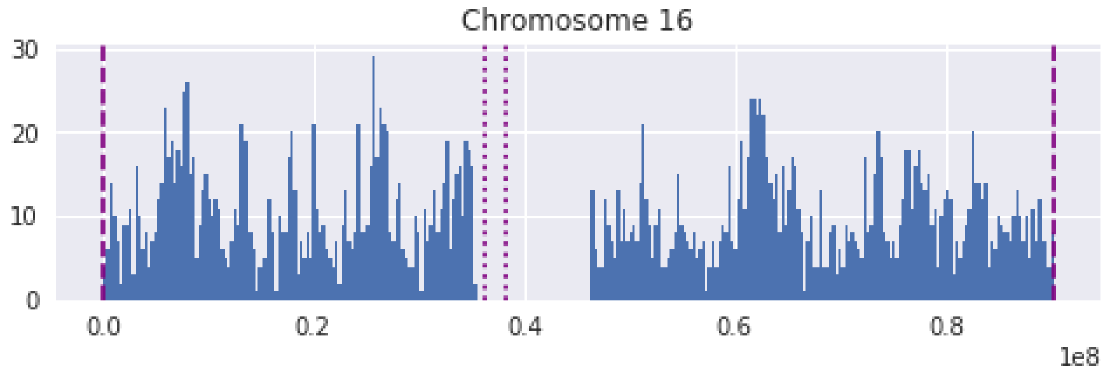
Chromosome 13

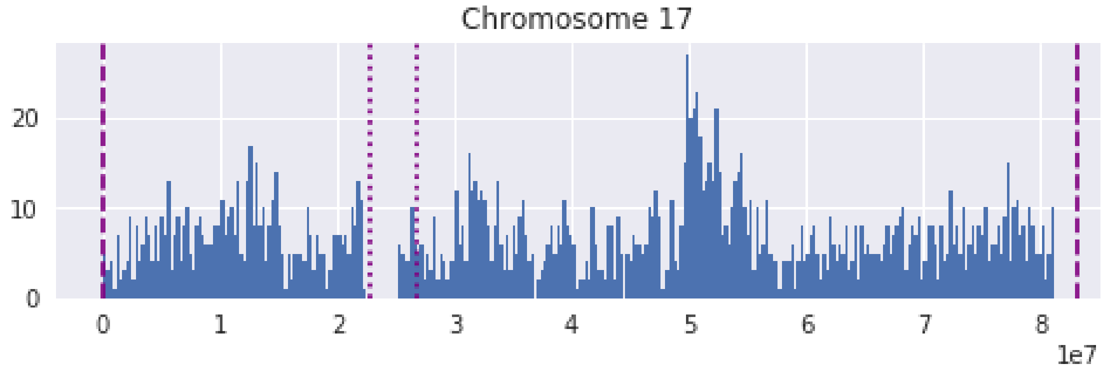


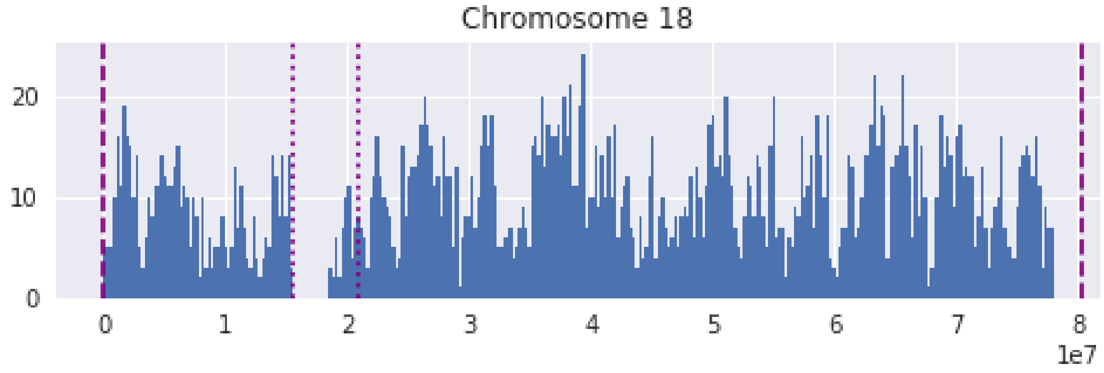
Chromosome 14

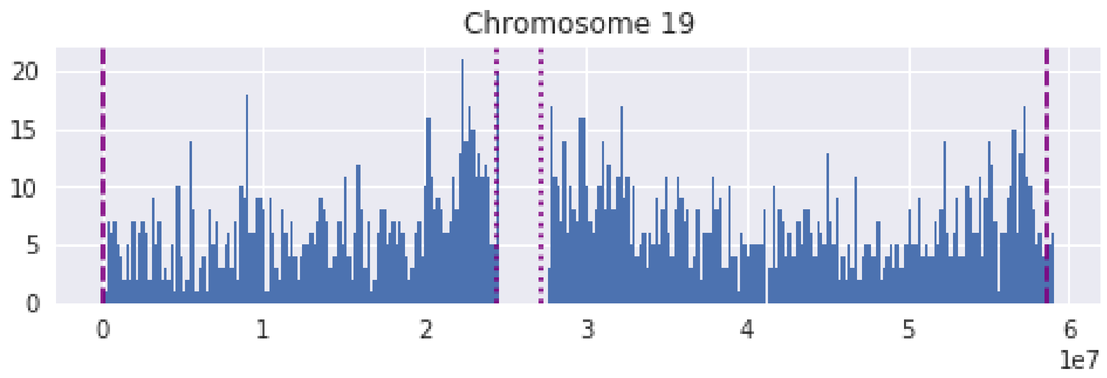


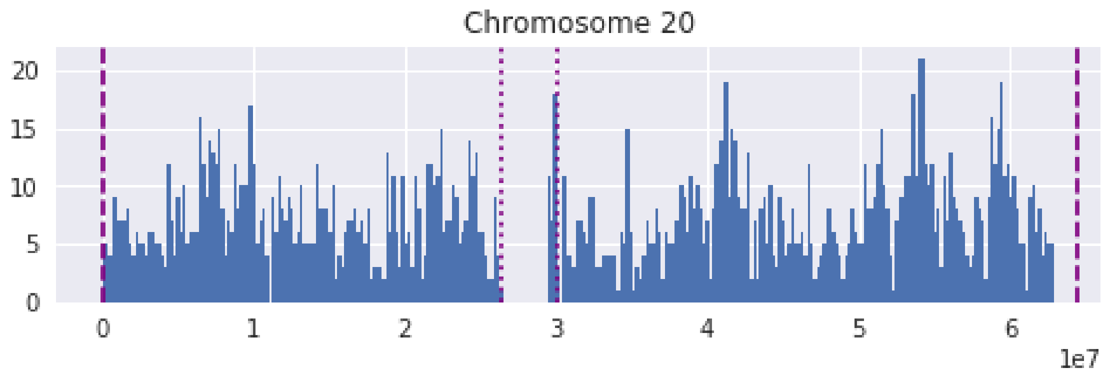
Chromosome 15

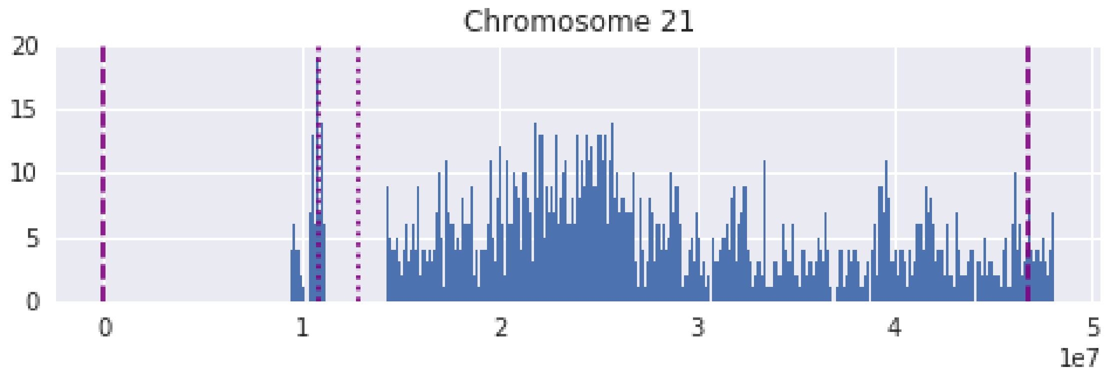

Chromosome 16
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1e8

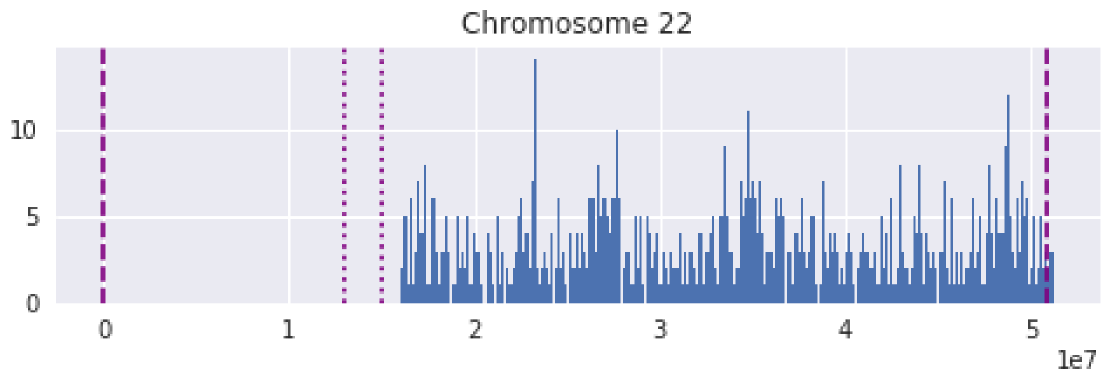

Chromosome 17
20
10
0
0
1
2
3
4
5
6
7
8
1e7

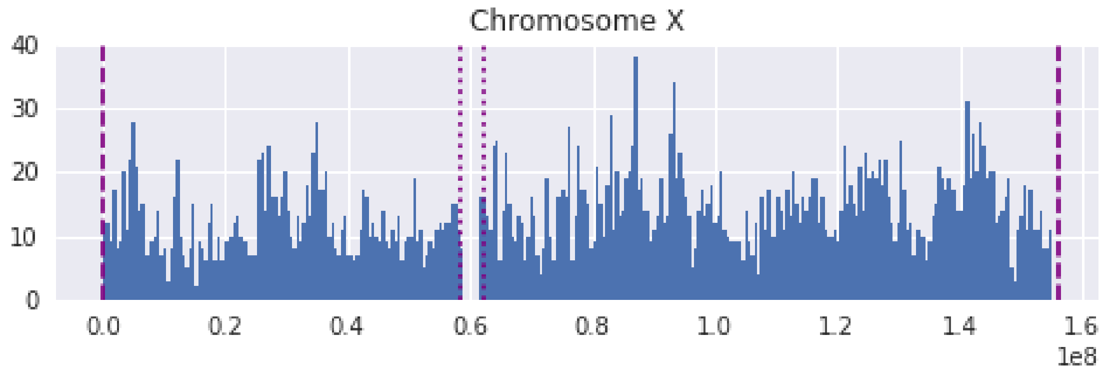

Chromosome 18
20
10
0
0
1
2
3
4
5
6
7
8
1e7

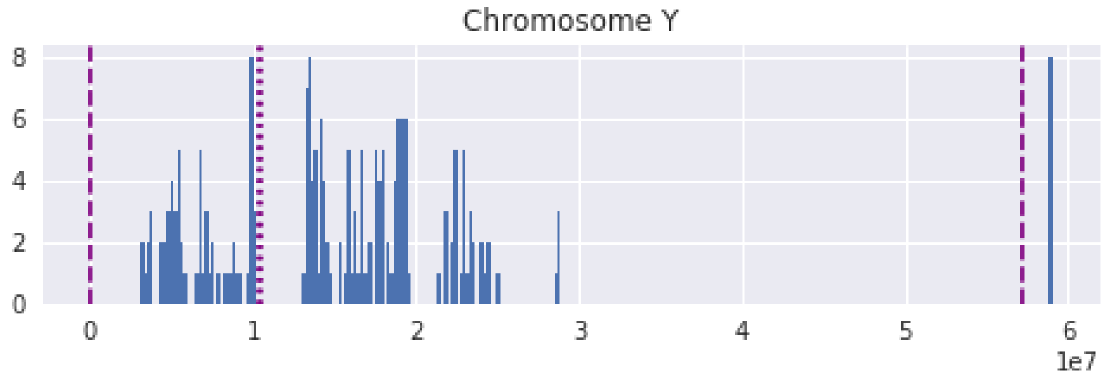

Chromosome 19
0
5
10
15
20
0
1
2
3
4
5
6
1e7

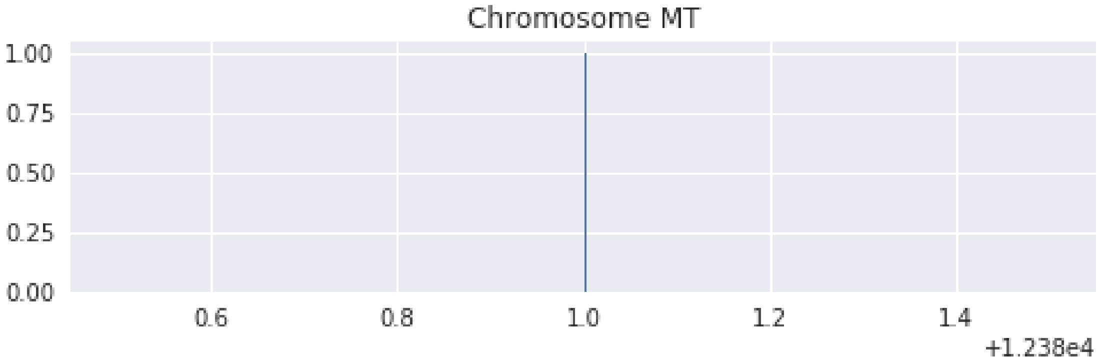

Chromosome 20
0
5
10
15
20
0
1
2
3
4
5
6
1e7


Chromosome 21
0
5
10
15
20
0
1
2
3
4
5
1e7


Chromosome 22
0
5
10
0
1
2
3
4
5
1e7


Chromosome X
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e8


Chromosome Y
0
2
4
6
8
0
1
2
3
4
5
6
1e7


Chromosome MT
1.00
0.75
0.50
0.25
0.00
0.6
0.8
1.0
1.2
1.4
+1.238e4