within_btwn_correlation

August 20, 2021

1 Population genetics simulation

1.1 Code

```
recomb_map = msprime.RecombinationMap.read_hapmap('./Simu3/genetic_map_GRCh37_chr22.txt')
ts = msprime.simulate(
    population_configurations = [
        msprime.PopulationConfiguration(sample_size=10000, growth_rate=0, initial_size=100000)
    ],
    demographic_events = [
        msprime.PopulationParametersChange(time=200, growth_rate=0.05)
    ],
    random_seed = 101,recombination_map=recomb_map, mutation_rate=2e-8
)
```

1.2 Input Summary

- Instead of fixed recombination rate, in this simulation, hapmap chromosome 22 recombination map is used
- Sample size = 10,000
- Initial population has a fixed growth rate but last 200 generation exponential growth with rate = 0.05
- Mutation rate = 2×10^{-8}
- Output is about 50 Mbp size

1.3 Output Summary

- Genome size $\approx$ 50 Mb
- Number of SNPs : 829,954
- Among the SNPs 827,702 has allele frequency < 0.005

1.4 Correlation within and between 1K region

```
array([[ 1.00000000e+00, -7.79879959e-04],
       [-7.79879959e-04,  1.00000000e+00]])
```

```
array([[1.        , 0.03597041],
       [0.03597041, 1.        ]])
```

We took a region in the begining of the genome of size 1,000 bp and another at the end of the genome. The correlation of mutational burden in these two regions are neglagible. Two adjacent regions of size 1K each in the begining of the genome has negligible positive correlation.

1.5 Correlations within and between 10K region

```
array([[1.00000000e+00, 9.31125926e-04],
       [9.31125926e-04, 1.00000000e+00]])

array([[1.00000000e+00, 9.31125926e-04],
       [9.31125926e-04, 1.00000000e+00]])
```

We see similar begavior when we took a 10kb region. There are negliginle or no correlation of mutational load between the regions. So far we included both the common and rare variants in the calculation. We wanted to see if the mutational burden shows any pattern when we only include the rare variants.

```
array([[ 1.        , -0.00423653],
       [-0.00423653,  1.        ]])

array([[1.        , 0.00250117],
       [0.00250117, 1.        ]])

array([[ 1.        , -0.00747924],
       [-0.00747924,  1.        ]])

array([[1.        , 0.00250117],
       [0.00250117, 1.        ]])
```

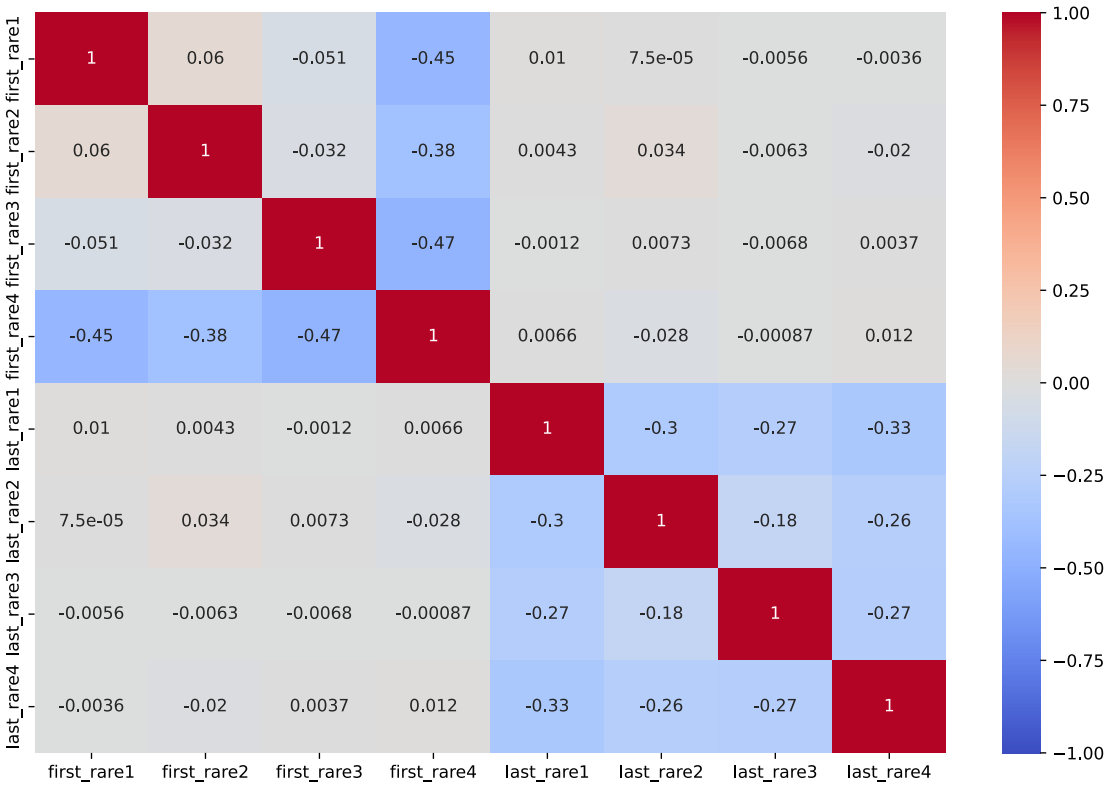
Even with taking 1K and 10K rare snps far apart and adjacent we don't see correlation of mutational burden among the sets. In our final check we will randomly select 10K rare variants form the first half and last half and see the correlations of mutaitonal burden.

```
array([[ 1.        , -0.00858777],
       [-0.00858777,  1.        ]])

array([[1.        , 0.00868325],
       [0.00868325, 1.        ]])
```

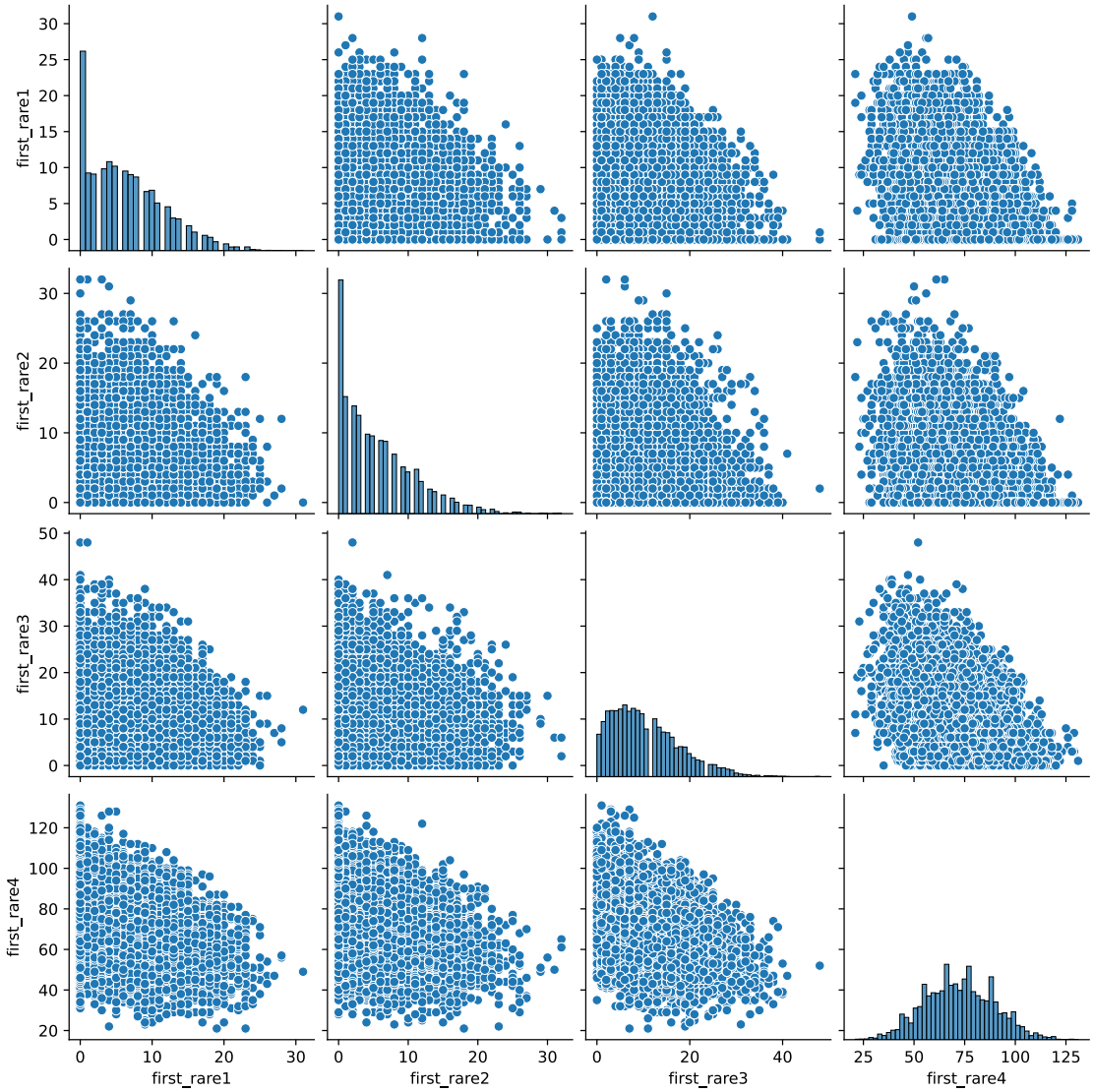
The last try was taking 10000 SNPs from first and last two third of the rare variants. This is done to check correlation of mutational load when there are overlap in region. In none of the cases explored above we saw a significant correlation in mutational load between two sets of SNPs. With a region the correlation of mutational load depends on how we stratify the SNPs, for example, grouping them by their LDscore.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe48c56bf50>
```



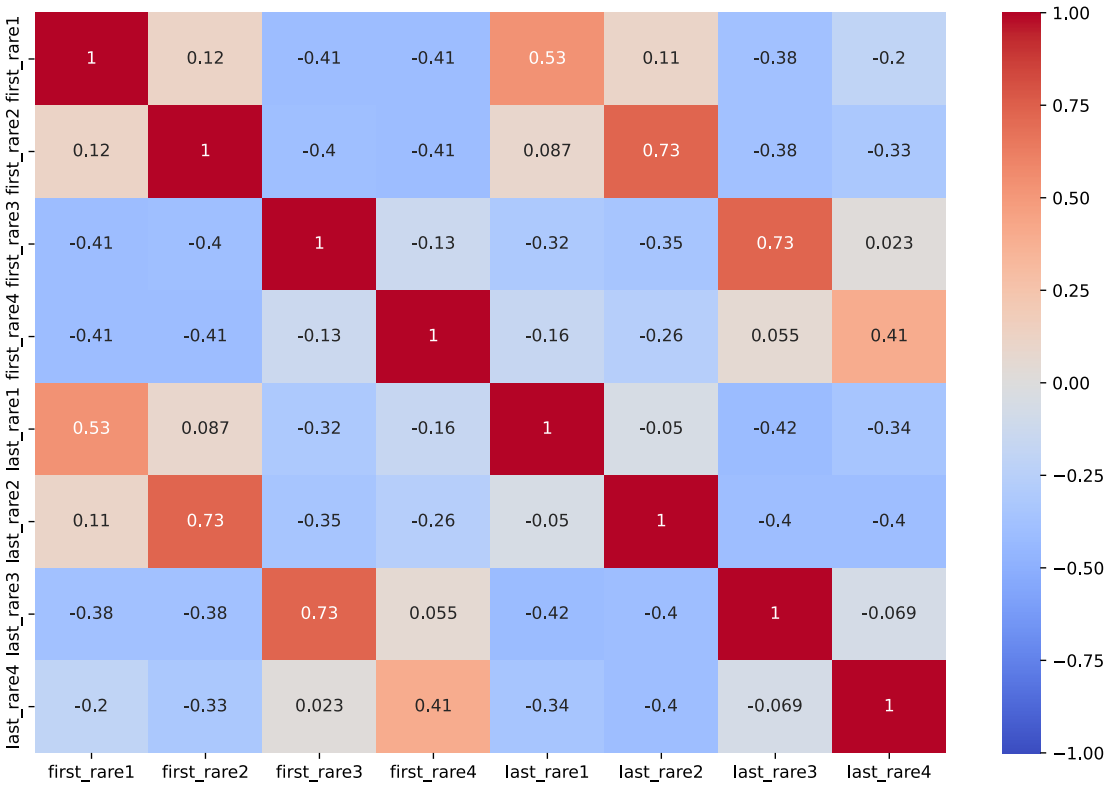
The above plot shows correlations of mutational burden in four rare bins within first and last one third of the variants. It shows that the mutational burden in rare 4 is negatively correlated with that in the other bins within the same set of SNPs. Between the sets of SNPs (first and last one thrid) there is no correlation.

```
<seaborn.axisgrid.PairGrid at 0x7f908dd26f10>
```



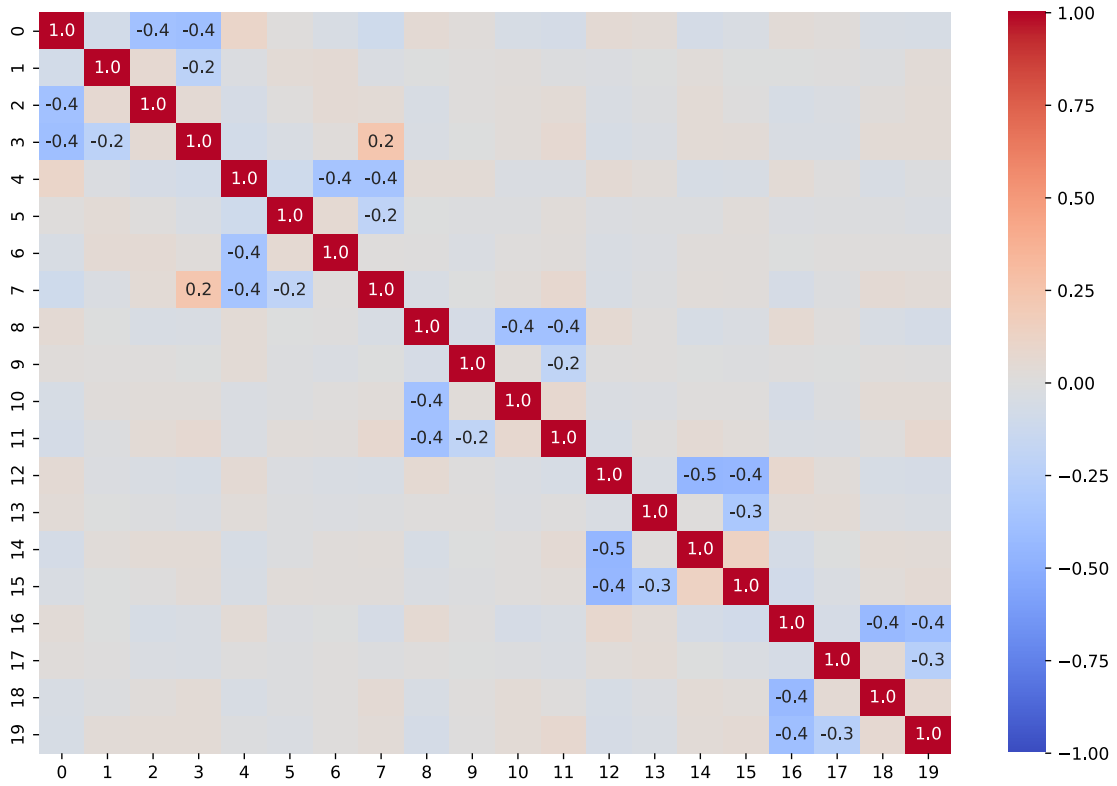
No abnormal behavior is seen in the scatter plot. The density plot shows, with increase of LD the mutational load goes from right skewed distribution to normality. We will now look at the correlations when the regions overlap. For that we took a set of variants from first two third and another set from last two third of the chromosome. In both sets the SNPs are binned by their LD score.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f913e199610>
```



2 Looking at the real data

I took a shuffled TOPMed sample and look into the correlation of mutational burden. In this analysis, five 5mb region from chromosome 21 is chosen and mutational burden is calculated on Rare 1 - 4 bins.



We see that there are negative correlations within a given region but no (or negligible) correlation between regions. The negative correlation means that if a person is a carrier of high LD rare variatns he is likely to have less number of low LD rare variants.