This tutorial seeks to extend the original tutorial included in the parent QuicK-mer2 repository, in which sample NA12878 was downloaded and converted into a BED file with QuicK-mer2 numbers according to GrCH38. We now pass NA12878 through the Human Rarity Determination to find rare deletions and duplications, and all variants in the present sample.

Installation Steps Clone the Repository to your local machine using git clone https://github.com/antnguye/QuicK-mer2_human_rarity_determination (https://github.com/antnguye/QuicK-mer2_human_rarity_determination) Install the git file pip install dist/qm2_human_rarity_prj-0.1.0.tar.gz Follow the tutorial in the original QuicK-mer2 https://github.com/KiddLab/QuicK-mer2/blob/master/tutorial.md (https://github.com/KiddLab/QuicK-mer2/blob/master/tutorial.md) (optional) After those two steps, you can progress along the provided notebook below.

In [1]:
```
# Step 1
from qm2_human_rarity import compare_against_1000
```

In [2]:
```
# we use is_human = True here, to indicate that NA12878 was processed to Gr
filename = "../tutorial-sample-results/NA12878.qm2.CN.1k.bed"
output_table, output_dict = compare_against_1000.read_in_qm2(filename, is_h
```

In [3]:
```
output_table
```

Out[3]:
```
array([0.705958, 1.126735, 0.99964 , ..., 4.695913, 4.543899, 4.834303],
      dtype=float32)
```

In [4]:
```
output_dict
```

Out[4]:
```
{0: ('chr1', '0', '54484'),
 1: ('chr1', '54484', '60739'),
 2: ('chr1', '60739', '68902'),
 3: ('chr1', '68902', '82642'),
 4: ('chr1', '82642', '88348'),
 5: ('chr1', '88348', '108310'),
 6: ('chr1', '108310', '138016'),
 7: ('chr1', '138016', '184188'),
 8: ('chr1', '184188', '203334'),
 9: ('chr1', '203334', '277627'),
 10: ('chr1', '277627', '456910'),
 11: ('chr1', '456910', '591356'),
 12: ('chr1', '591356', '611444'),
 13: ('chr1', '611444', '630679'),
 14: ('chr1', '630679', '634342'),
 15: ('chr1', '634342', '662294'),
 16: ('chr1', '662294', '704251'),
 17: ('chr1', '704251', '754054'),
 18: ('chr1', '754054', '777156'),
```

In [5]:
```
NA12878_dups = compare_against_1000.find_dups(output_table, output_dict)
```

In [6]:
```
NA12878_dups
```

Out[6]:
```
[[[3.216503, 9665],
  [3.431669, 9666],
  [4.31106, 9667],
  [3.341199, 9668],
  [3.84454, 9669],
  [3.540264, 9670],
  [3.720474, 9671],
  [3.725578, 9672]],
 [[2.508482, 12174],
  [4.682363, 12175],
  [3.444574, 12176],
  [3.623551, 12177],
  [3.834896, 12178],
  [4.647865, 12179],
  [8.295877, 12180],
  [7.127389, 12181],
  [8.652528, 12182],
  [7.986285, 12183],
  [8.986205, 12184],
```

In [7]:
```
NA12878_dels = compare_against_1000.find_deletions(output_table, output_dic
NA12878_dels
```

Out[7]:
```
[[[0.705958, 0],
  [1.126735, 1],
  [0.99964, 2],
  [0.094265, 3],
  [0.088984, 4],
  [0.423472, 5]],
 [[0.733451, 1429], [0.743091, 1430], [1.362334, 1431]],
 [[0.745552, 2600], [0.013377, 2601], [0.529178, 2602]],
 [[1.211794, 2724], [1.423162, 2725], [1.301391, 2726]],
 [[1.287055, 9673], [1.458844, 9674], [1.210273, 9675]],
 [[1.2075, 18979],
  [1.028159, 18980],
  [1.166035, 18981],
  [1.074081, 18982],
  [1.134434, 18983],
  [1.012392, 18984]],
 [[1.488403, 18986],
  [1.03547, 18987],
  [1.093066, 18988],
```

In [9]:
```
normal_file_path = "../qm2_human_rarity/99_normalcy_range_tenk_genomes.npy"
normal_dups, normal_dels = compare_against_1000.compare_1000_genomes(output
```

In [15]:
```
for location, result in normal_dups.items():
    if result is True:
        print(location, result)
```

```
chr2:94507371-94516416 True
chr2:166640689-166644025 True
chr9:61318885-61406926 True
chr9:77153890-77162274 True
chr10:89035171-89042891 True
chr11:50117815-50165650 True
chr21:45348055-45351247 True
chr22:23710553-23720582 True
chrX:26792873-26803338 True
chrX:154548946-154567908 True
```

In [16]:
```
for location, result in normal_dels.items():
    if result is True:
        print(location, result)
```

```
chr2:154962837-154973458 True
chr3:177576353-177579658 True
chr4:181539613-181542788 True
chr5:1882412-1885451 True
chr5:105095630-105167989 True
chr6:132054168-132057615 True
chr7:1816693-1823698 True
chr8:40027371-40033327 True
chr9:650880-654148 True
chr12:183800-187447 True
chr12:268563-276065 True
chr18:32915395-32921570 True
```

In [ ]:
```
# Write output files
```

In [12]:
```
compare_against_1000.write_dups_and_dels(output_dict, NA12878_dups, NA12878
```

In [13]:
```
compare_against_1000.write_rarity("NA12878", normal_dups, normal_dels)
```

```
The above write steps will create two pairs of files:
NA12878_(duplications/deletions) contains each duplication and its median CN
NA12878.rare_(duplications/deletions) contains each duplication and a boolean
on whether or not if the variant is outside of 99% of the range contained in
the 1000 genomes (True if so)

Checking the normal_dups and normal_dels, we find a handful of rare
duplications and deletions, which are located in regions that NA12878 has a
copy number outside of 99% of the total human genome range in the 1000
Genomes. If this were a disease patient sample, we would recommend observing
these regions for evidence of gene presence or regulation.
```