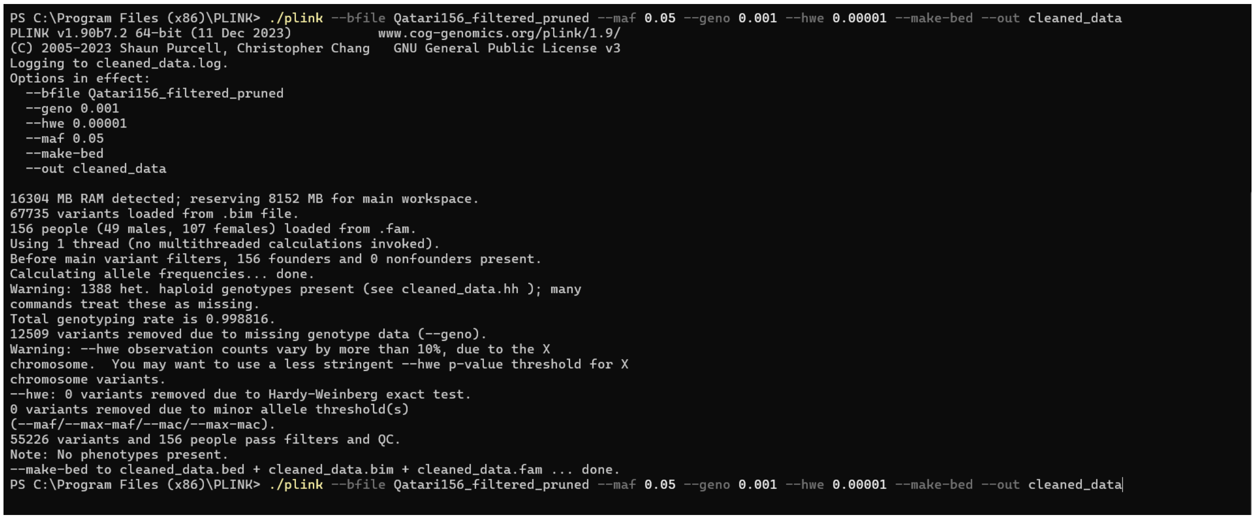
lab6

Abdullah Taman

2024-04-11

First Part: Data Preprocessing

plink –file Qatari156_filtered_pruned –maf 0.05 –geno 0.001 –hwe 0.00001 –make-bed –out cleaned_data 12509 variants removed due to missing genotype data (–geno). hwe: 0 variants removed due to Hardy-Weinberg exact test. 0 variants removed due to minor allele threshold(s)

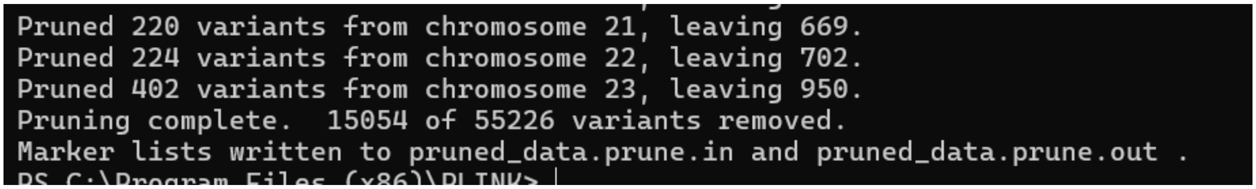


```
PS C:\Program Files (x86)\PLINK> ./plink --bfile Qatari156_filtered_pruned --maf 0.05 --geno 0.001 --hwe 0.00001 --make-bed --out cleaned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to cleaned_data.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.001
  --hwe 0.00001
  --maf 0.05
  --make-bed
  --out cleaned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see cleaned_data.hh ); many
commands treat these as missing.
Total genotyping rate is 0.998816.
12509 variants removed due to missing genotype data (--geno).
Warning: --hwe observation counts vary by more than 10%, due to the X
chromosome.  You may want to use a less stringent --hwe p-value threshold for X
chromosome variants.
--hwe: 0 variants removed due to Hardy-Weinberg exact test.
0 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to cleaned_data.bed + cleaned_data.bim + cleaned_data.fam ... done.
PS C:\Program Files (x86)\PLINK> ./plink --bfile Qatari156_filtered_pruned --maf 0.05 --geno 0.001 --hwe 0.00001 --make-bed --out cleaned_data
```

Figure 1: Caption for the image

./plink –bfile cleaned_data –indep-pairwise 100 5 0.1 –out pruned_data

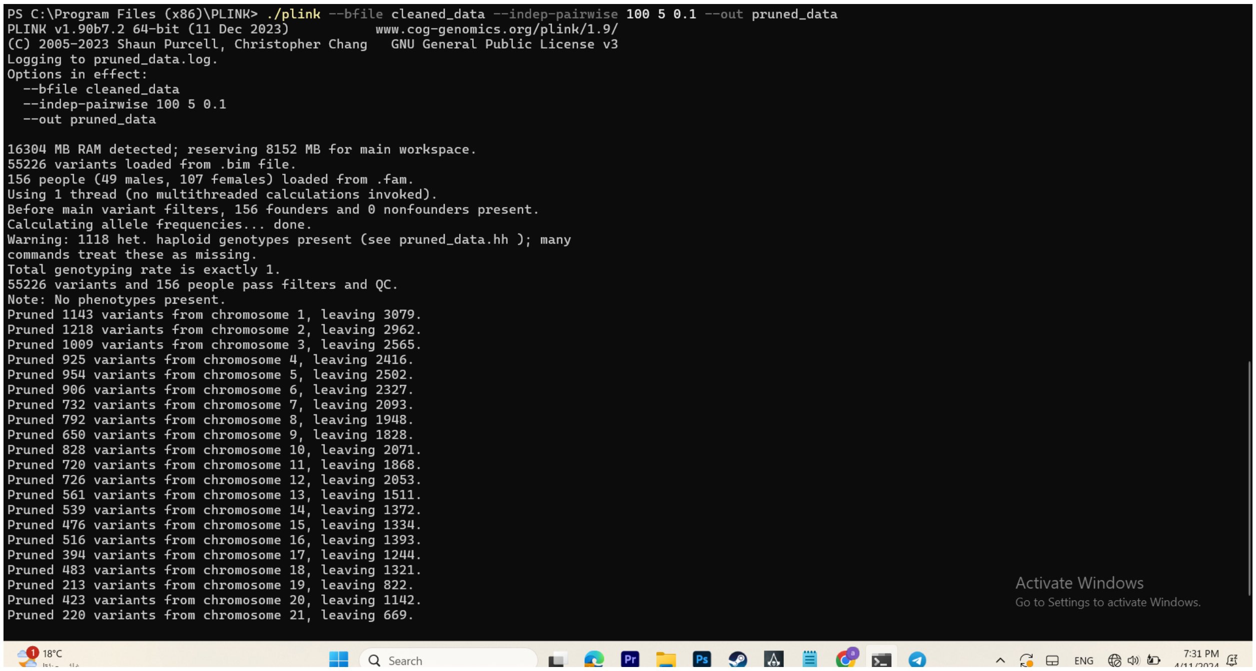
- Trying different Window Size (First Parameter) —————————
- At window size = 100



```
Pruned 220 variants from chromosome 21, leaving 669.
Pruned 224 variants from chromosome 22, leaving 702.
Pruned 402 variants from chromosome 23, leaving 950.
Pruning complete.  15054 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
PS C:\Program Files (x86)\PLINK>
```

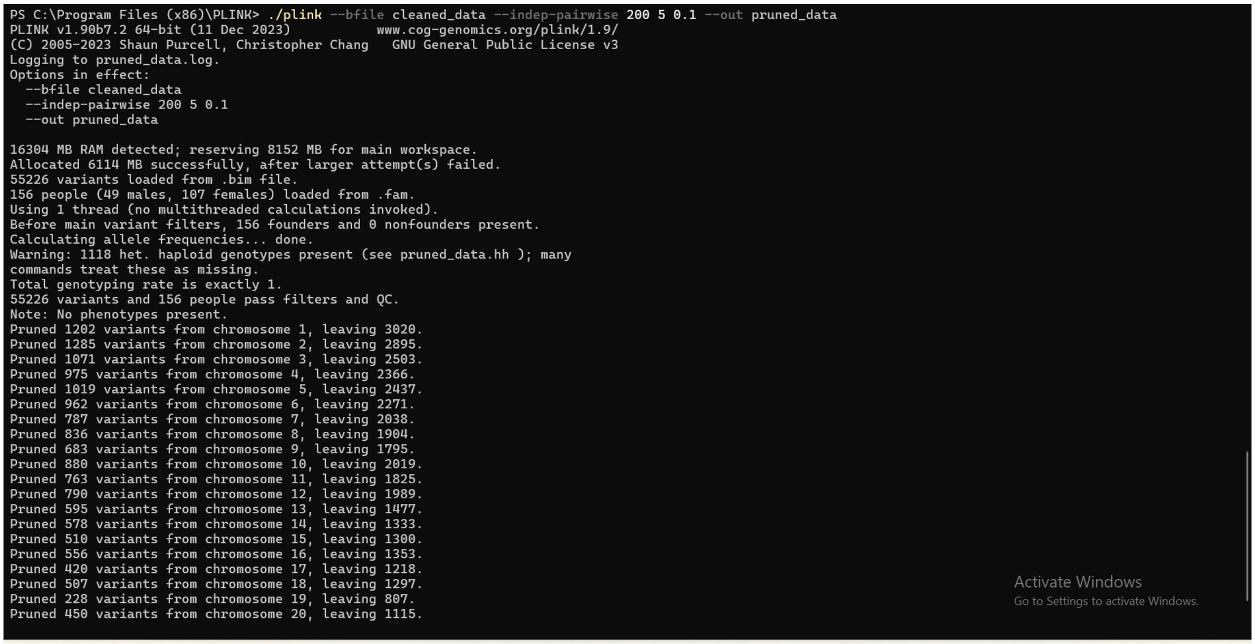
Figure 2: Caption for the image

15054 of 55226 variants removed



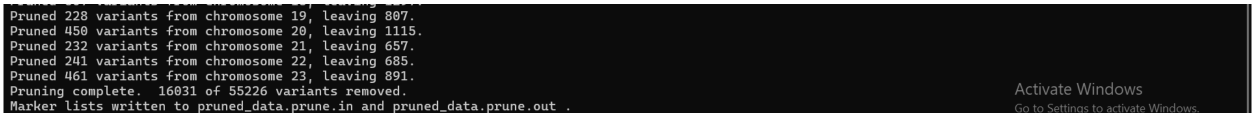
```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 100 5 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 5 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1143 variants from chromosome 1, leaving 3079.
Pruned 1218 variants from chromosome 2, leaving 2962.
Pruned 1009 variants from chromosome 3, leaving 2565.
Pruned 925 variants from chromosome 4, leaving 2416.
Pruned 954 variants from chromosome 5, leaving 2502.
Pruned 906 variants from chromosome 6, leaving 2327.
Pruned 732 variants from chromosome 7, leaving 2093.
Pruned 792 variants from chromosome 8, leaving 1948.
Pruned 650 variants from chromosome 9, leaving 1828.
Pruned 828 variants from chromosome 10, leaving 2071.
Pruned 720 variants from chromosome 11, leaving 1868.
Pruned 726 variants from chromosome 12, leaving 2053.
Pruned 561 variants from chromosome 13, leaving 1511.
Pruned 539 variants from chromosome 14, leaving 1372.
Pruned 476 variants from chromosome 15, leaving 1334.
Pruned 516 variants from chromosome 16, leaving 1393.
Pruned 394 variants from chromosome 17, leaving 1244.
Pruned 483 variants from chromosome 18, leaving 1321.
Pruned 213 variants from chromosome 19, leaving 822.
Pruned 423 variants from chromosome 20, leaving 1142.
Pruned 220 variants from chromosome 21, leaving 669.
```

Figure 3: Caption for the image



```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 200 5 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 200 5 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
Allocated 6114 MB successfully, after larger attempt(s) failed.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1202 variants from chromosome 1, leaving 3020.
Pruned 1285 variants from chromosome 2, leaving 2895.
Pruned 1071 variants from chromosome 3, leaving 2503.
Pruned 975 variants from chromosome 4, leaving 2366.
Pruned 1019 variants from chromosome 5, leaving 2437.
Pruned 962 variants from chromosome 6, leaving 2271.
Pruned 787 variants from chromosome 7, leaving 2038.
Pruned 836 variants from chromosome 8, leaving 1904.
Pruned 683 variants from chromosome 9, leaving 1795.
Pruned 880 variants from chromosome 10, leaving 2019.
Pruned 763 variants from chromosome 11, leaving 1825.
Pruned 790 variants from chromosome 12, leaving 1989.
Pruned 595 variants from chromosome 13, leaving 1477.
Pruned 578 variants from chromosome 14, leaving 1333.
Pruned 510 variants from chromosome 15, leaving 1300.
Pruned 556 variants from chromosome 16, leaving 1353.
Pruned 420 variants from chromosome 17, leaving 1218.
Pruned 507 variants from chromosome 18, leaving 1297.
Pruned 228 variants from chromosome 19, leaving 807.
Pruned 450 variants from chromosome 20, leaving 1115.
```

Figure 4: Caption for the image



```
Pruned 228 variants from chromosome 19, leaving 807.
Pruned 450 variants from chromosome 20, leaving 1115.
Pruned 232 variants from chromosome 21, leaving 657.
Pruned 241 variants from chromosome 22, leaving 685.
Pruned 461 variants from chromosome 23, leaving 891.
Pruning complete.  16031 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
```

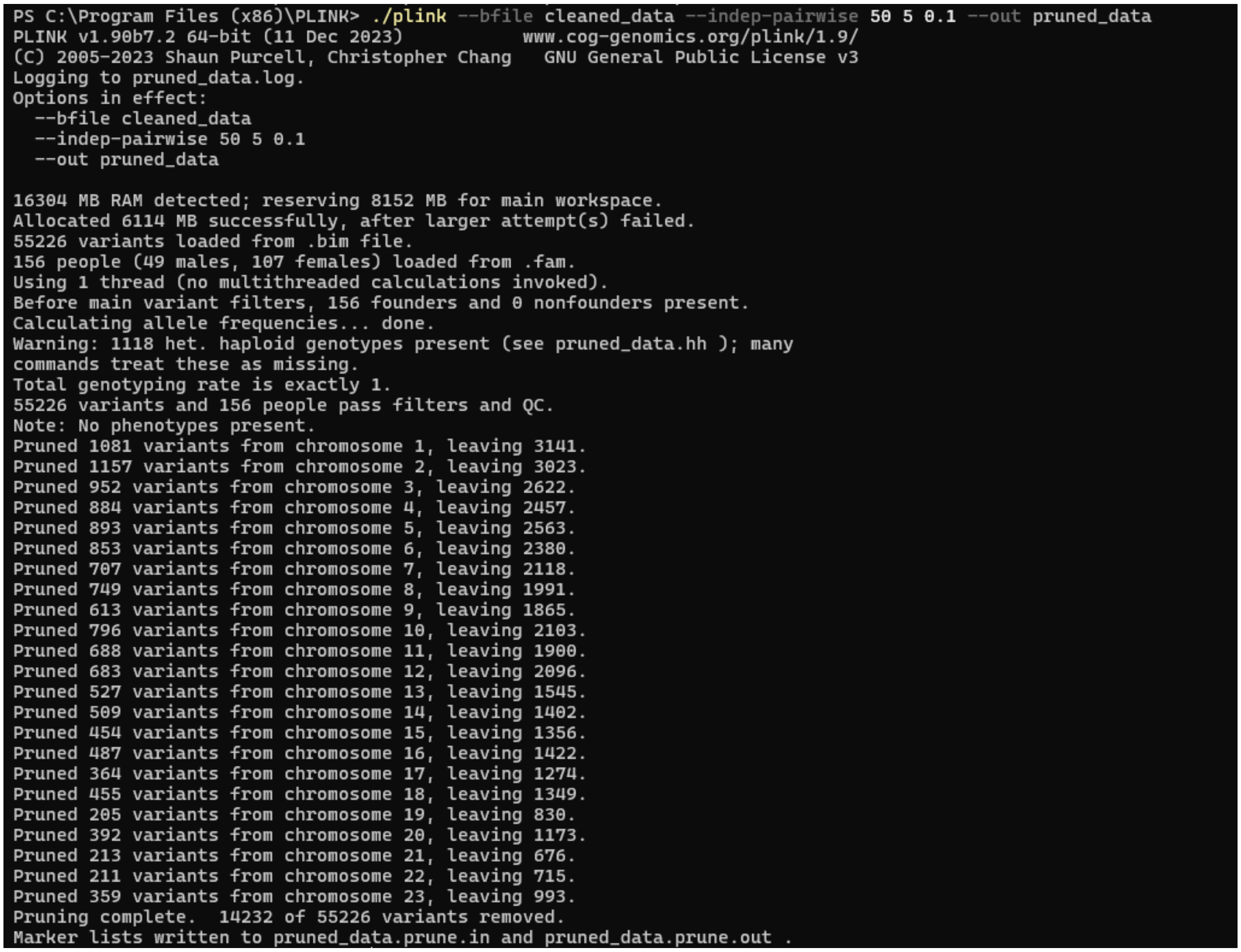
Figure 5: Caption for the image

- At window size = 200

16031 Variants were removed

- At window size = 50

14232 of 55226 variants removed



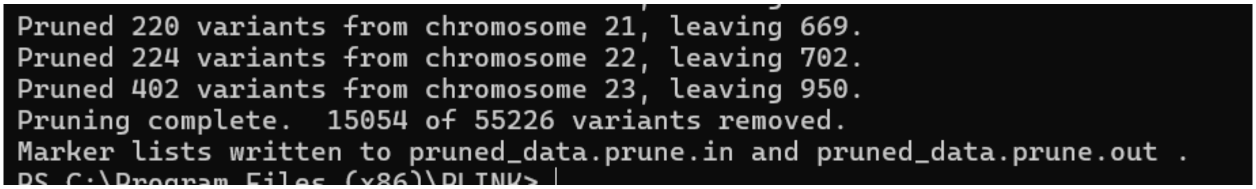
```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 50 5 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 50 5 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
Allocated 6114 MB successfully, after larger attempt(s) failed.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1081 variants from chromosome 1, leaving 3141.
Pruned 1157 variants from chromosome 2, leaving 3023.
Pruned 952 variants from chromosome 3, leaving 2622.
Pruned 884 variants from chromosome 4, leaving 2457.
Pruned 893 variants from chromosome 5, leaving 2563.
Pruned 853 variants from chromosome 6, leaving 2380.
Pruned 707 variants from chromosome 7, leaving 2118.
Pruned 749 variants from chromosome 8, leaving 1991.
Pruned 613 variants from chromosome 9, leaving 1865.
Pruned 796 variants from chromosome 10, leaving 2103.
Pruned 688 variants from chromosome 11, leaving 1900.
Pruned 683 variants from chromosome 12, leaving 2096.
Pruned 527 variants from chromosome 13, leaving 1545.
Pruned 509 variants from chromosome 14, leaving 1402.
Pruned 454 variants from chromosome 15, leaving 1356.
Pruned 487 variants from chromosome 16, leaving 1422.
Pruned 364 variants from chromosome 17, leaving 1274.
Pruned 455 variants from chromosome 18, leaving 1349.
Pruned 205 variants from chromosome 19, leaving 830.
Pruned 392 variants from chromosome 20, leaving 1173.
Pruned 213 variants from chromosome 21, leaving 676.
Pruned 211 variants from chromosome 22, leaving 715.
Pruned 359 variants from chromosome 23, leaving 993.
Pruning complete.  14232 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
```

Figure 6: Caption for the image

- Trying different Step Size (Second Parameter) ——————————
- At step size = 5

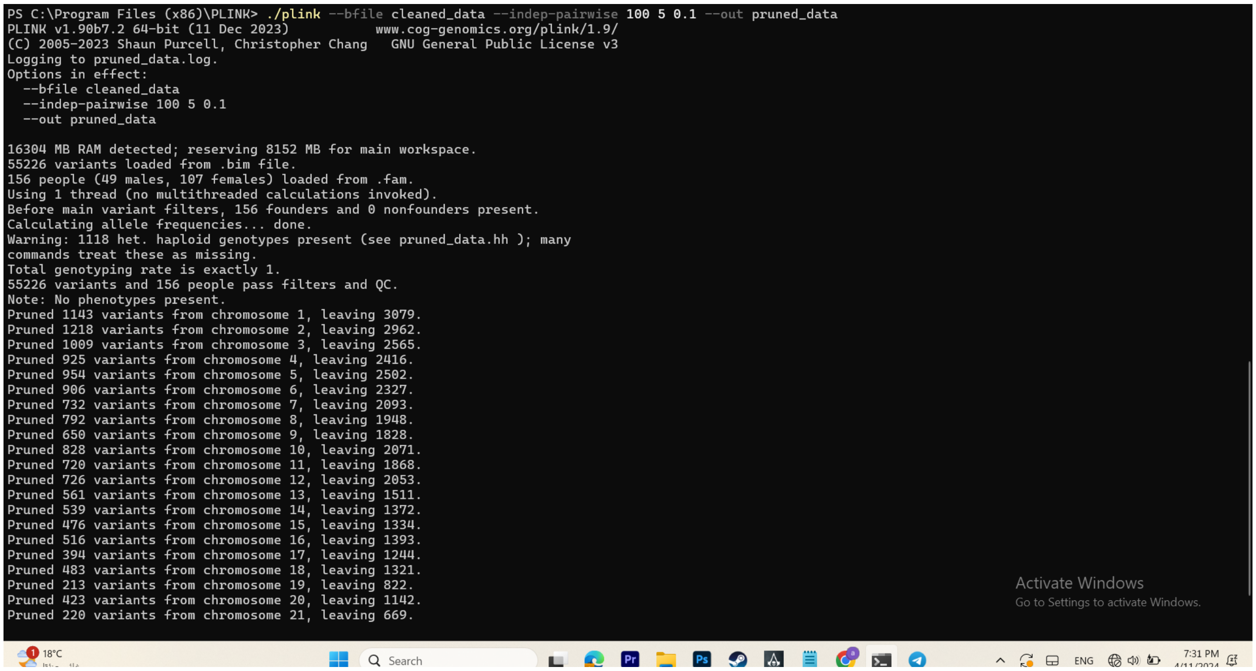
15054 of 55226 variants removed

- At step size = 20



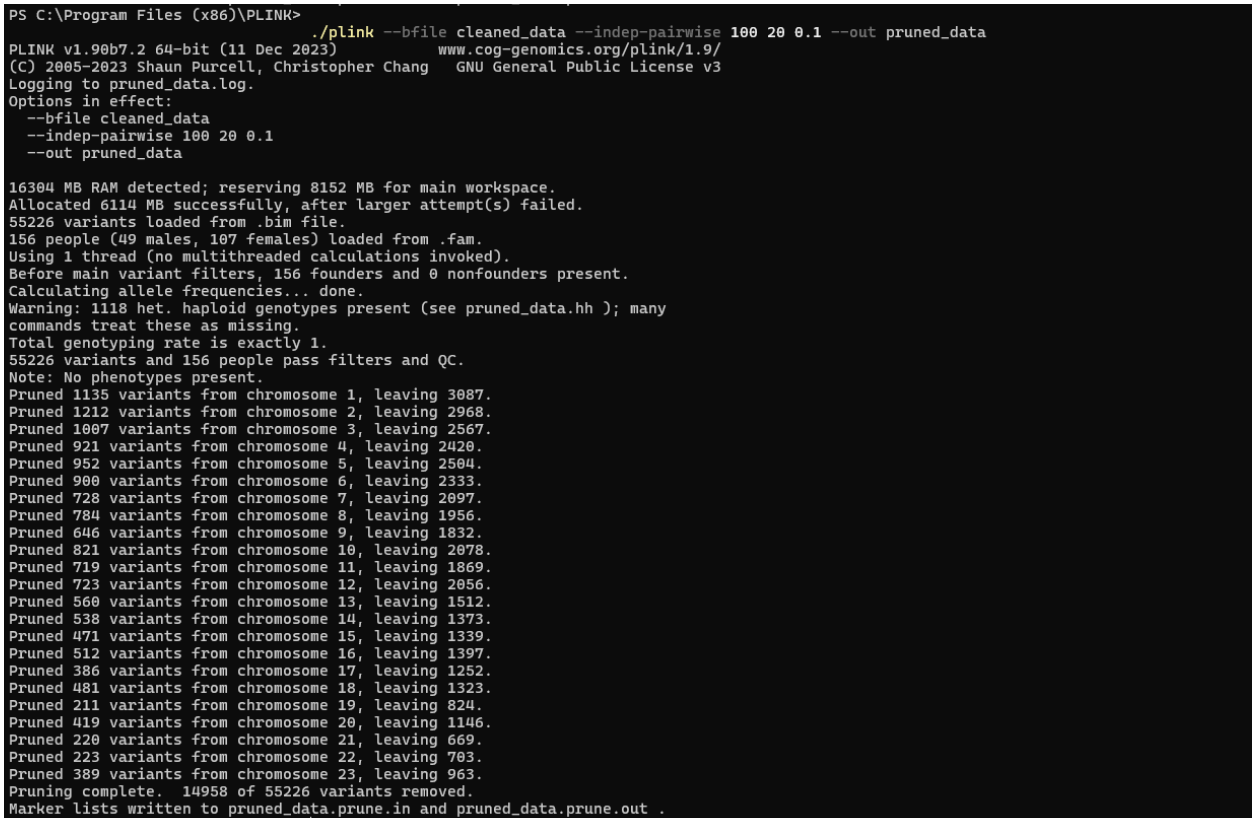
```
Pruned 220 variants from chromosome 21, leaving 669.
Pruned 224 variants from chromosome 22, leaving 702.
Pruned 402 variants from chromosome 23, leaving 950.
Pruning complete.  15054 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
PS C:\Program Files (x86)\PLINK>
```

Figure 7: Caption for the image



```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 100 5 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 5 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1143 variants from chromosome 1, leaving 3079.
Pruned 1218 variants from chromosome 2, leaving 2962.
Pruned 1009 variants from chromosome 3, leaving 2565.
Pruned 925 variants from chromosome 4, leaving 2416.
Pruned 954 variants from chromosome 5, leaving 2502.
Pruned 906 variants from chromosome 6, leaving 2327.
Pruned 732 variants from chromosome 7, leaving 2093.
Pruned 792 variants from chromosome 8, leaving 1948.
Pruned 650 variants from chromosome 9, leaving 1828.
Pruned 828 variants from chromosome 10, leaving 2071.
Pruned 720 variants from chromosome 11, leaving 1868.
Pruned 726 variants from chromosome 12, leaving 2053.
Pruned 561 variants from chromosome 13, leaving 1511.
Pruned 539 variants from chromosome 14, leaving 1372.
Pruned 476 variants from chromosome 15, leaving 1334.
Pruned 516 variants from chromosome 16, leaving 1393.
Pruned 394 variants from chromosome 17, leaving 1244.
Pruned 483 variants from chromosome 18, leaving 1321.
Pruned 213 variants from chromosome 19, leaving 822.
Pruned 423 variants from chromosome 20, leaving 1142.
Pruned 220 variants from chromosome 21, leaving 669.
```

Figure 8: Caption for the image



```
PS C:\Program Files (x86)\PLINK>
                                 ./plink --bfile cleaned_data --indep-pairwise 100 20 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 20 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
Allocated 6114 MB successfully, after larger attempt(s) failed.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1135 variants from chromosome 1, leaving 3087.
Pruned 1212 variants from chromosome 2, leaving 2968.
Pruned 1007 variants from chromosome 3, leaving 2567.
Pruned 921 variants from chromosome 4, leaving 2420.
Pruned 952 variants from chromosome 5, leaving 2504.
Pruned 900 variants from chromosome 6, leaving 2333.
Pruned 728 variants from chromosome 7, leaving 2097.
Pruned 784 variants from chromosome 8, leaving 1956.
Pruned 646 variants from chromosome 9, leaving 1832.
Pruned 821 variants from chromosome 10, leaving 2078.
Pruned 719 variants from chromosome 11, leaving 1869.
Pruned 723 variants from chromosome 12, leaving 2056.
Pruned 560 variants from chromosome 13, leaving 1512.
Pruned 538 variants from chromosome 14, leaving 1373.
Pruned 471 variants from chromosome 15, leaving 1339.
Pruned 512 variants from chromosome 16, leaving 1397.
Pruned 386 variants from chromosome 17, leaving 1252.
Pruned 481 variants from chromosome 18, leaving 1323.
Pruned 211 variants from chromosome 19, leaving 824.
Pruned 419 variants from chromosome 20, leaving 1146.
Pruned 220 variants from chromosome 21, leaving 669.
Pruned 223 variants from chromosome 22, leaving 703.
Pruned 389 variants from chromosome 23, leaving 963.
Pruning complete.  14958 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
```

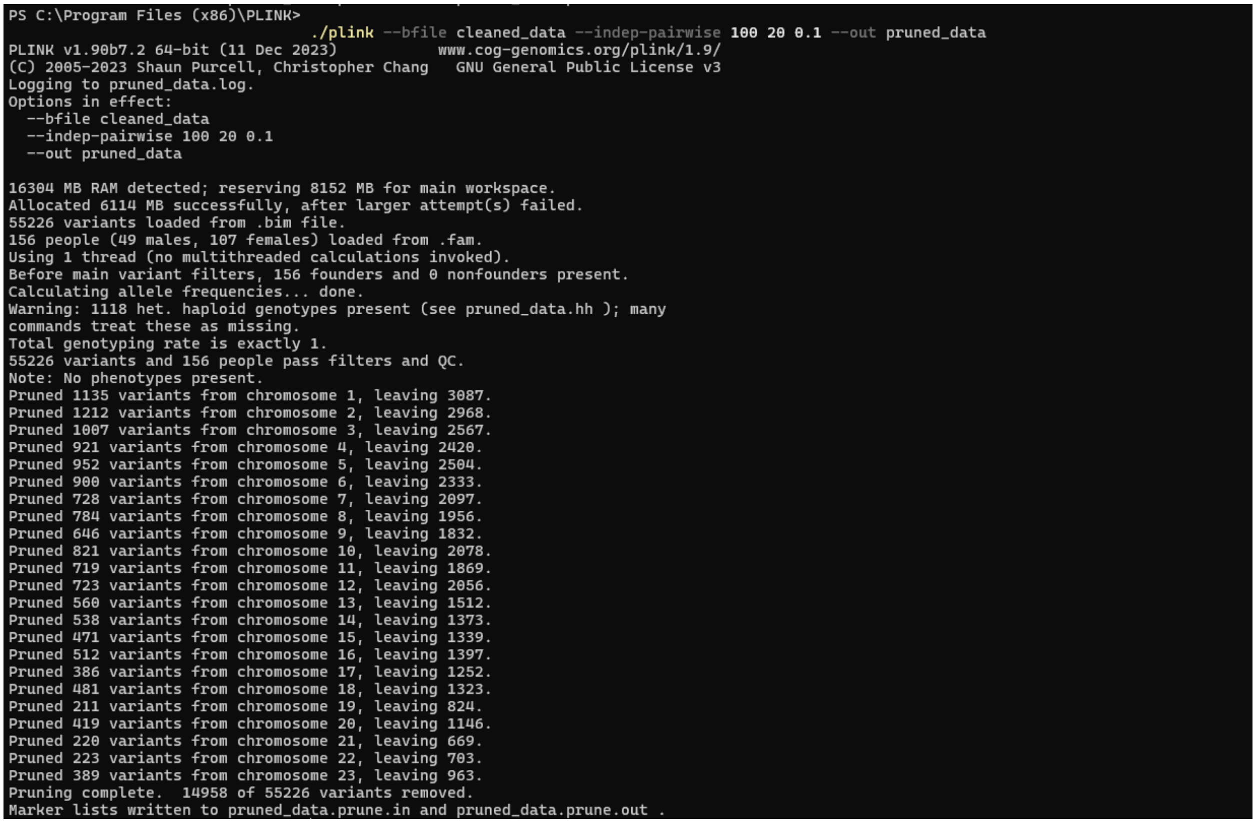
14958 of 55226 variants removed.

- At step size = 1

Pruning complete. 15066 of 55226 variants removed.

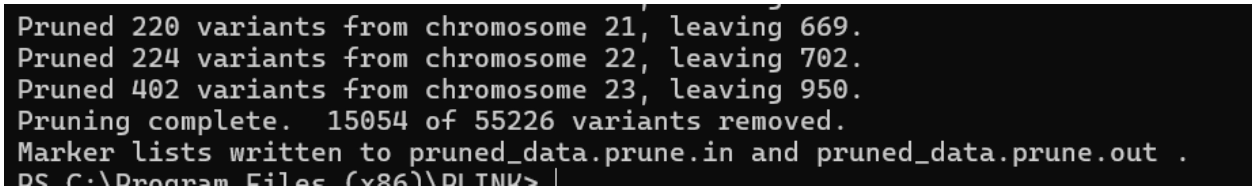
- Trying different LD (Third Parameter) —————————
- LD = .1

- LD = .9



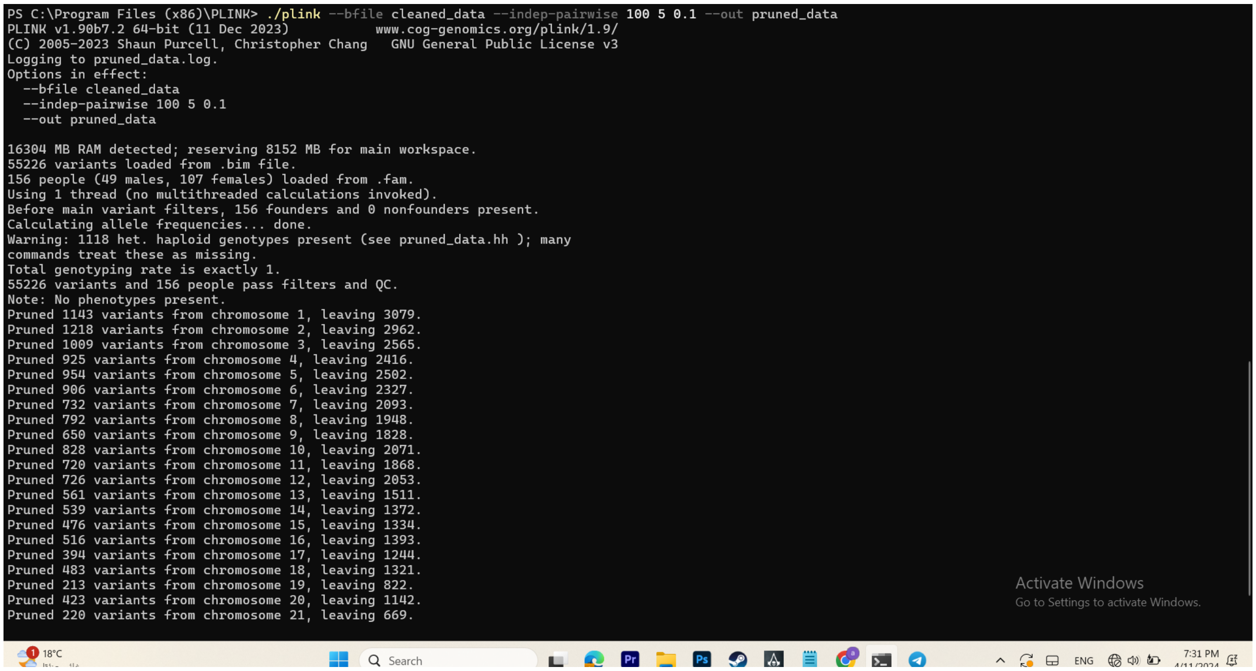
```
PS C:\Program Files (x86)\PLINK>
                                 ./plink --bfile cleaned_data --indep-pairwise 100 20 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 20 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
Allocated 6114 MB successfully, after larger attempt(s) failed.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1135 variants from chromosome 1, leaving 3087.
Pruned 1212 variants from chromosome 2, leaving 2968.
Pruned 1007 variants from chromosome 3, leaving 2567.
Pruned 921 variants from chromosome 4, leaving 2420.
Pruned 952 variants from chromosome 5, leaving 2504.
Pruned 900 variants from chromosome 6, leaving 2333.
Pruned 728 variants from chromosome 7, leaving 2097.
Pruned 784 variants from chromosome 8, leaving 1956.
Pruned 646 variants from chromosome 9, leaving 1832.
Pruned 821 variants from chromosome 10, leaving 2078.
Pruned 719 variants from chromosome 11, leaving 1869.
Pruned 723 variants from chromosome 12, leaving 2056.
Pruned 560 variants from chromosome 13, leaving 1512.
Pruned 538 variants from chromosome 14, leaving 1373.
Pruned 471 variants from chromosome 15, leaving 1339.
Pruned 512 variants from chromosome 16, leaving 1397.
Pruned 386 variants from chromosome 17, leaving 1252.
Pruned 481 variants from chromosome 18, leaving 1323.
Pruned 211 variants from chromosome 19, leaving 824.
Pruned 419 variants from chromosome 20, leaving 1146.
Pruned 220 variants from chromosome 21, leaving 669.
Pruned 223 variants from chromosome 22, leaving 703.
Pruned 389 variants from chromosome 23, leaving 963.
Pruning complete.  14958 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
```

Figure 9: Caption for the image



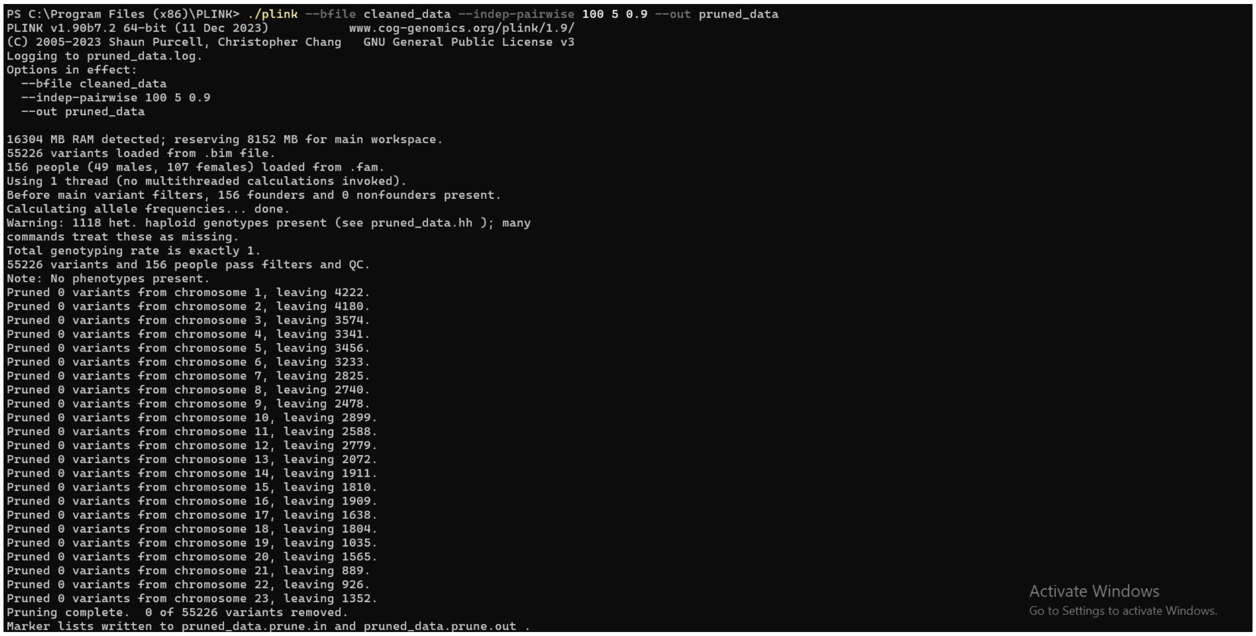
```
Pruned 220 variants from chromosome 21, leaving 669.
Pruned 224 variants from chromosome 22, leaving 702.
Pruned 402 variants from chromosome 23, leaving 950.
Pruning complete.  15054 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
PS C:\Program Files (x86)\PLINK>
```

Figure 10: Caption for the image



```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 100 5 0.1 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 5 0.1
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 1143 variants from chromosome 1, leaving 3079.
Pruned 1218 variants from chromosome 2, leaving 2962.
Pruned 1009 variants from chromosome 3, leaving 2565.
Pruned 925 variants from chromosome 4, leaving 2416.
Pruned 954 variants from chromosome 5, leaving 2502.
Pruned 906 variants from chromosome 6, leaving 2327.
Pruned 732 variants from chromosome 7, leaving 2093.
Pruned 792 variants from chromosome 8, leaving 1948.
Pruned 650 variants from chromosome 9, leaving 1828.
Pruned 828 variants from chromosome 10, leaving 2071.
Pruned 720 variants from chromosome 11, leaving 1868.
Pruned 726 variants from chromosome 12, leaving 2053.
Pruned 561 variants from chromosome 13, leaving 1511.
Pruned 539 variants from chromosome 14, leaving 1372.
Pruned 476 variants from chromosome 15, leaving 1334.
Pruned 516 variants from chromosome 16, leaving 1393.
Pruned 394 variants from chromosome 17, leaving 1244.
Pruned 483 variants from chromosome 18, leaving 1321.
Pruned 213 variants from chromosome 19, leaving 822.
Pruned 423 variants from chromosome 20, leaving 1142.
Pruned 220 variants from chromosome 21, leaving 669.
```

Figure 11: Caption for the image



```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 100 5 0.9 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 5 0.9
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 0 variants from chromosome 1, leaving 4222.
Pruned 0 variants from chromosome 2, leaving 4180.
Pruned 0 variants from chromosome 3, leaving 3574.
Pruned 0 variants from chromosome 4, leaving 3341.
Pruned 0 variants from chromosome 5, leaving 3456.
Pruned 0 variants from chromosome 6, leaving 3233.
Pruned 0 variants from chromosome 7, leaving 2825.
Pruned 0 variants from chromosome 8, leaving 2740.
Pruned 0 variants from chromosome 9, leaving 2478.
Pruned 0 variants from chromosome 10, leaving 2899.
Pruned 0 variants from chromosome 11, leaving 2588.
Pruned 0 variants from chromosome 12, leaving 2779.
Pruned 0 variants from chromosome 13, leaving 2072.
Pruned 0 variants from chromosome 14, leaving 1911.
Pruned 0 variants from chromosome 15, leaving 1810.
Pruned 0 variants from chromosome 16, leaving 1909.
Pruned 0 variants from chromosome 17, leaving 1638.
Pruned 0 variants from chromosome 18, leaving 1804.
Pruned 0 variants from chromosome 19, leaving 1035.
Pruned 0 variants from chromosome 20, leaving 1565.
Pruned 0 variants from chromosome 21, leaving 889.
Pruned 0 variants from chromosome 22, leaving 926.
Pruned 0 variants from chromosome 23, leaving 1352.
Pruning complete.  0 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
```

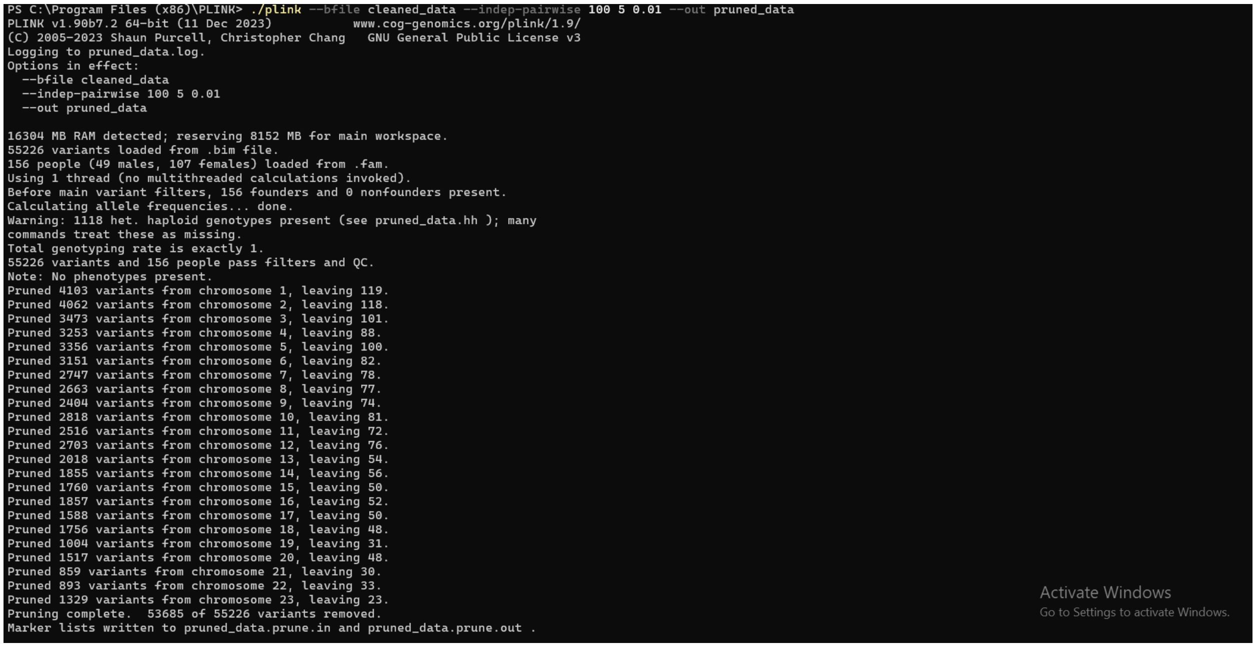
0 of 55226 variants removed.

- LD = .01

53685 of 55226 variants removed.

So when we: /////* Increase/Decrease window size -> more/less SNPs are removed Increase/Decrease step size -> less/more SNPs are removed Increase/Decrease LD -> less/more SNPs are removed

- We'll do the pca on the result of window size = 100, step size = 5, LD = .1 We had to recode the



```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --indep-pairwise 100 5 0.01 --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --indep-pairwise 100 5 0.01
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Pruned 4103 variants from chromosome 1, leaving 119.
Pruned 4062 variants from chromosome 2, leaving 118.
Pruned 3473 variants from chromosome 3, leaving 101.
Pruned 3253 variants from chromosome 4, leaving 88.
Pruned 3356 variants from chromosome 5, leaving 100.
Pruned 3151 variants from chromosome 6, leaving 82.
Pruned 2747 variants from chromosome 7, leaving 78.
Pruned 2663 variants from chromosome 8, leaving 77.
Pruned 2404 variants from chromosome 9, leaving 74.
Pruned 2818 variants from chromosome 10, leaving 81.
Pruned 2516 variants from chromosome 11, leaving 72.
Pruned 2703 variants from chromosome 12, leaving 76.
Pruned 2018 variants from chromosome 13, leaving 54.
Pruned 1855 variants from chromosome 14, leaving 56.
Pruned 1760 variants from chromosome 15, leaving 50.
Pruned 1857 variants from chromosome 16, leaving 52.
Pruned 1588 variants from chromosome 17, leaving 50.
Pruned 1756 variants from chromosome 18, leaving 48.
Pruned 1004 variants from chromosome 19, leaving 31.
Pruned 1517 variants from chromosome 20, leaving 48.
Pruned 859 variants from chromosome 21, leaving 30.
Pruned 893 variants from chromosome 22, leaving 33.
Pruned 1329 variants from chromosome 23, leaving 23.
Pruning complete.  53685 of 55226 variants removed.
Marker lists written to pruned_data.prune.in and pruned_data.prune.out .
```

Figure 12: Caption for the image

result of this operation to be able to find the data in ped and map format: ./plink –bfile cleaned_data –indep-pairwise 100 5 0.1 –out pruned_data –recode This is the result of the operation

Now we do this operation to make sure we got the right SNPs

PART II: Identify SNPs associated with population structure

I used this video in this part for guidance:

https://www.youtube.com/watch?v=vos6VeuNcaM&ab_channel=GenomicsBootCamp

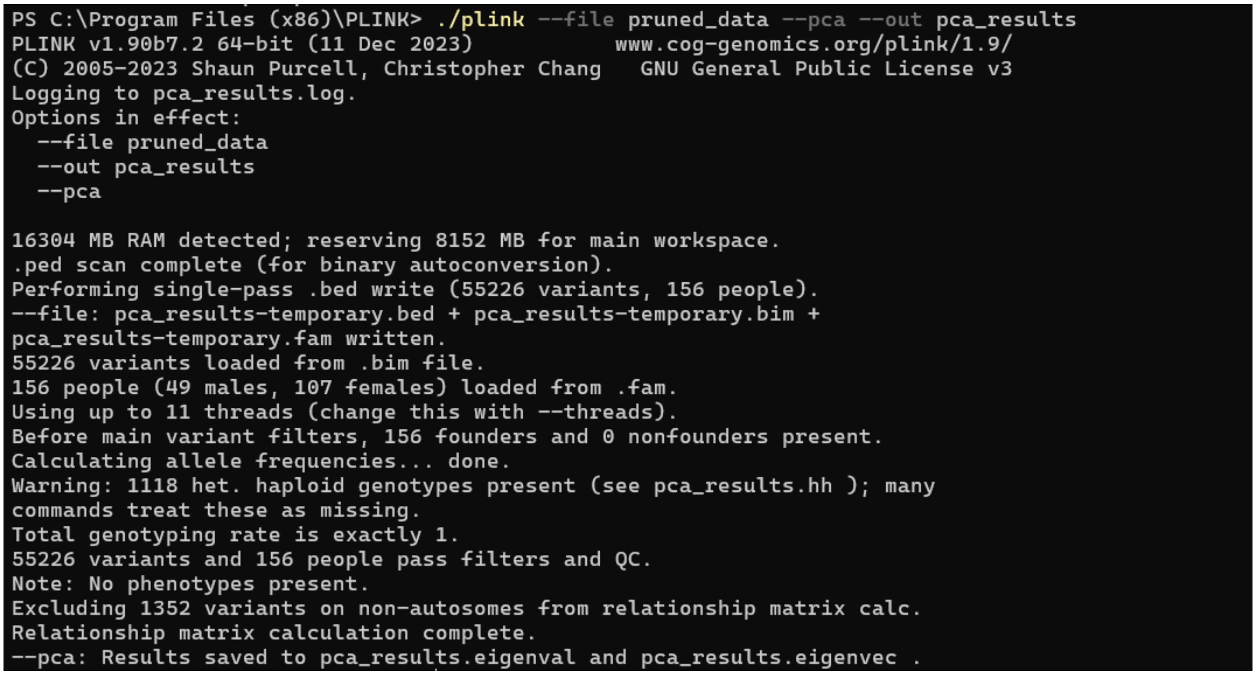
- First we run PCA on the cleaned_data ./plink –bfile cleaned_data –pca –out pca_results

This is the result:

- Second we read the .raw file C:/Program Files (x86)/PLINK/

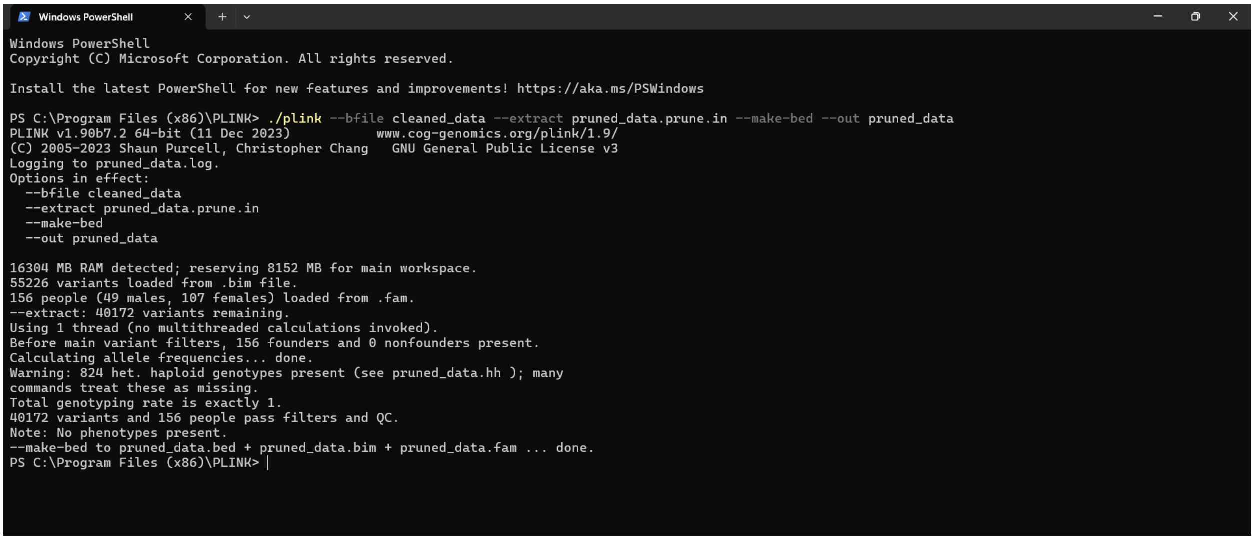
```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```



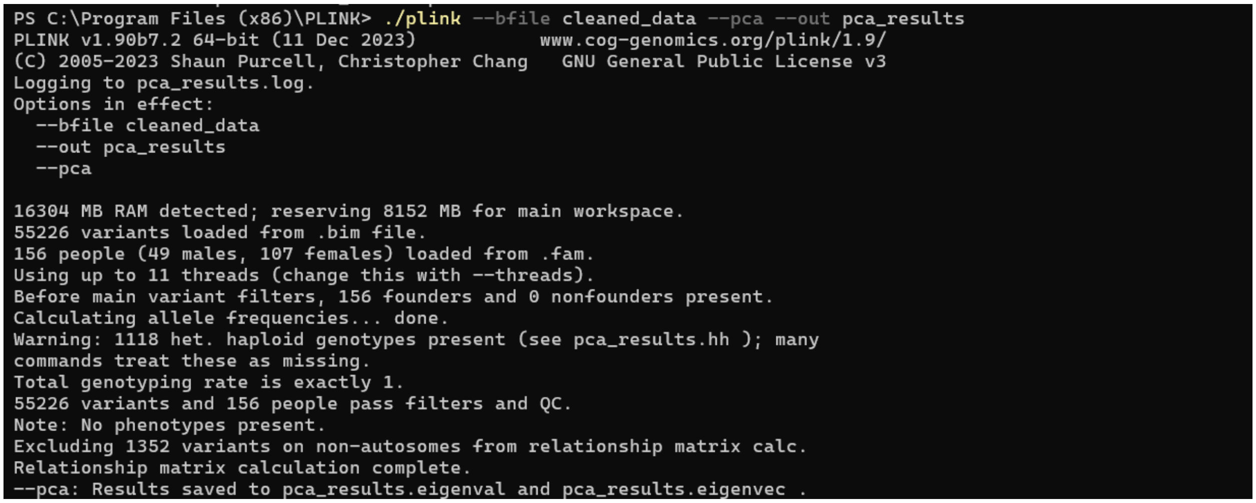
```
PS C:\Program Files (x86)\PLINK> ./plink --file pruned_data --pca --out pca_results
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pca_results.log.
Options in effect:
  --file pruned_data
  --out pca_results
  --pca

16304 MB RAM detected; reserving 8152 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (55226 variants, 156 people).
--file: pca_results-temporary.bed + pca_results-temporary.bim +
pca_results-temporary.fam written.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using up to 11 threads (change this with --threads).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pca_results.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Excluding 1352 variants on non-autosomes from relationship matrix calc.
Relationship matrix calculation complete.
--pca: Results saved to pca_results.eigenval and pca_results.eigenvec .
```

Figure 13: Caption for the image



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --extract pruned_data.prune.in --make-bed --out pruned_data
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pruned_data.log.
Options in effect:
  --bfile cleaned_data
  --extract pruned_data.prune.in
  --make-bed
  --out pruned_data

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
--extract: 40172 variants remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 824 het. haploid genotypes present (see pruned_data.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
40172 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to pruned_data.bed + pruned_data.bim + pruned_data.fam ... done.
PS C:\Program Files (x86)\PLINK>
```

Figure 14: Caption for the image



```
PS C:\Program Files (x86)\PLINK> ./plink --bfile cleaned_data --pca --out pca_results
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to pca_results.log.
Options in effect:
  --bfile cleaned_data
  --out pca_results
  --pca

16304 MB RAM detected; reserving 8152 MB for main workspace.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using up to 11 threads (change this with --threads).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see pca_results.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
Excluding 1352 variants on non-autosomes from relationship matrix calc.
Relationship matrix calculation complete.
--pca: Results saved to pca_results.eigenval and pca_results.eigenvec .
```

Figure 15: Caption for the image

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(scatterplot3d)
```

```r
#install.packages("qqman")
library(qqman)
```

```
## Warning: package 'qqman' was built under R version 4.3.3
```

```
##
```

```
## For example usage please run: vignette('qqman')
```

```
##
```

```
## Citation appreciated but not required:
```

```
## Turner, (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Jour
```

```
##
```

```
##
## Attaching package: 'qqman'
```

```
## The following object is masked from 'package:lattice':
##
##     qq
```

```
genotype_data <- read.table("C:/Program Files (x86)/PLINK/recoded_data.raw", header = TRUE, sep = "")
```

```
genotype_data_filtered <- genotype_data[, c(1,7:ncol(genotype_data))]
```

```
t <- read.table("C:/Program Files (x86)/PLINK/recoded_data.raw", header = TRUE, sep = "")
```

```
eigval <- read.table("C:/Program Files (x86)/PLINK/pca_results.eigenval", header = FALSE, sep = "")
eigvec <- read.table("C:/Program Files (x86)/PLINK/pca_results.eigenvec", header = FALSE, sep = "")
```

```
df <- data.frame(x = genotype_data_filtered[, c(2: ncol(genotype_data_filtered))], y = eigvec[c(3: ncol
```

I'vw commented the following parts to be able to knit the file, as they may take 4~5 hours to finish

```
#num_tests <- (ncol(genotype_data_filtered) - 1) * 3
#result_df <- data.frame(SNP = character(num_tests), P = numeric(num_tests), PC = integer(num_tests), B
#ptr = 1

#for (i in 3:5) {
  #for (j in 2:ncol(genotype_data_filtered)) {
    #for (k in 3:5) {
      #for (l in 3:5) {
        #if (i != k & i != l & k != l && k < l) {
          # Construct formula
        #  formula <- reformulate(c(paste0("x.", names(genotype_data_filtered)[j]), paste0("y.V", k), p

          # Fit linear regression model
         # model <- lm(formula, data = df)

          # Extract p-value, beta coefficient, and standard error
        #  summary_coef <- summary(model)$coefficients
          #p_value <- summary_coef[2, "Pr(>|t|)"]
         # beta <- summary_coef[2, "Estimate"]
        #  se <- summary_coef[2, "Std. Error"]

          # Store results in result_df

          #          result_df[ptr, ] <- c(names(genotype_data_filtered)[j], p_value, i - 2, beta, se)
      #    ptr = ptr + 1
     #     print(ptr)
    #    }
   #   }
  #  }
```

```
 #}
#}


# Print the first few rows of the result data frame
#print(head(result_df))
```

```
map_data <- read.table("C:/Program Files (x86)/PLINK/map.map", header = FALSE, sep = "\t", col.names = c
```

```
#modified_result_df <- result_df
#modified_result_df$SNP <- substring(result_df$SNP, 1, nchar(result_df$SNP) - 2)
#merged_data <- merge(modified_result_df, map_data, by.x = "SNP", by.y = "SNP_ID", all.x = TRUE)
#merged_data$P <- as.numeric(merged_data$P)

#str(merged_data)
```

```
#pc1_data <- merged_data[merged_data$PC == 1, ]
#pc2_data <- merged_data[merged_data$PC == 2, ]
#pc3_data <- merged_data[merged_data$PC == 3, ]
```

```
#manhattan_data_1 <- pc1_data[, c("SNP", "Chromosome", "Physical_Position", "P")]
#colnames(manhattan_data_1) <- c("SNP", "CHR", "BP", "P")
#manhattan_data_1$P <- as.numeric(manhattan_data_1$P)
# Create the QQMan Manhattan plot
#manhattan(manhattan_data_1)
```

This was the result:

Now we'll sort the PC1 dataframe

```
#pc1_data_sorted <- pc1_data[order(pc1_data$P), ]
#pc1_data_top10 <- pc1_data_sorted[1:10, ]
#pc1_data_top10
#write.csv(pc1_data_top10, "j:/output_file.csv", row.names = TRUE)
```

This was the result

Now after searching in the dbSNP, using excel we constructed this table

Now we'll study the SNPs with the second PC, but for the next we only will draw the manhattan and show the top 10

```
#manhattan_data_2 <- pc2_data[, c("SNP", "Chromosome", "Physical_Position", "P")]
#colnames(manhattan_data_2) <- c("SNP", "CHR", "BP", "P")
#manhattan_data_2$P <- as.numeric(manhattan_data_2$P)
#manhattan(manhattan_data_2)
```

This was the result:

Again, we'llsort the resuls and get the top 10 SNPs and find the data about them

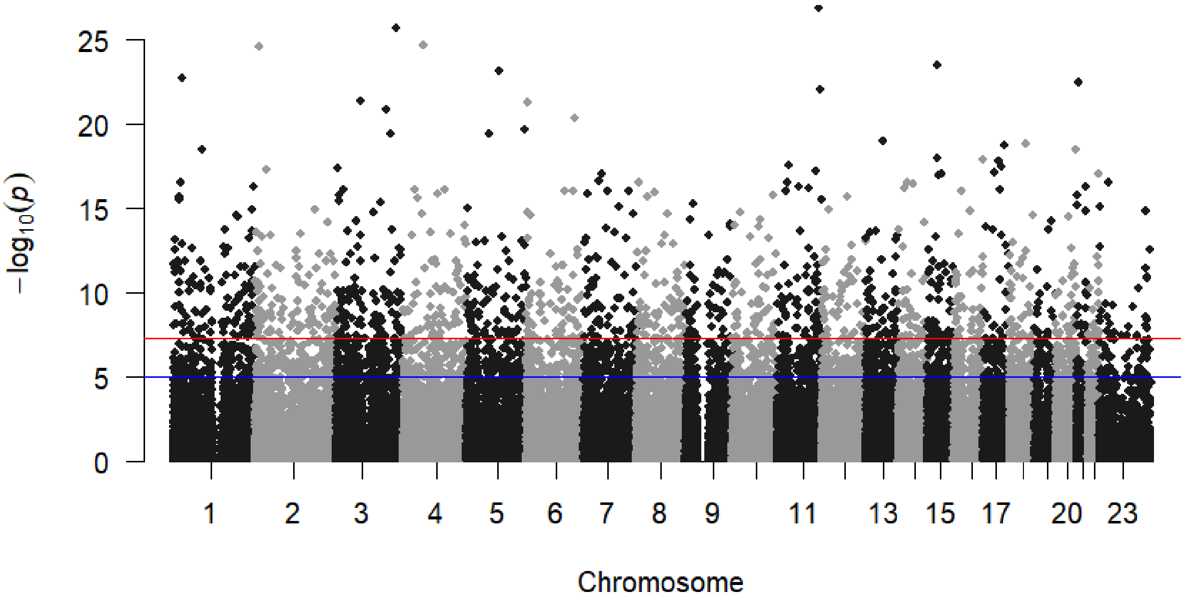
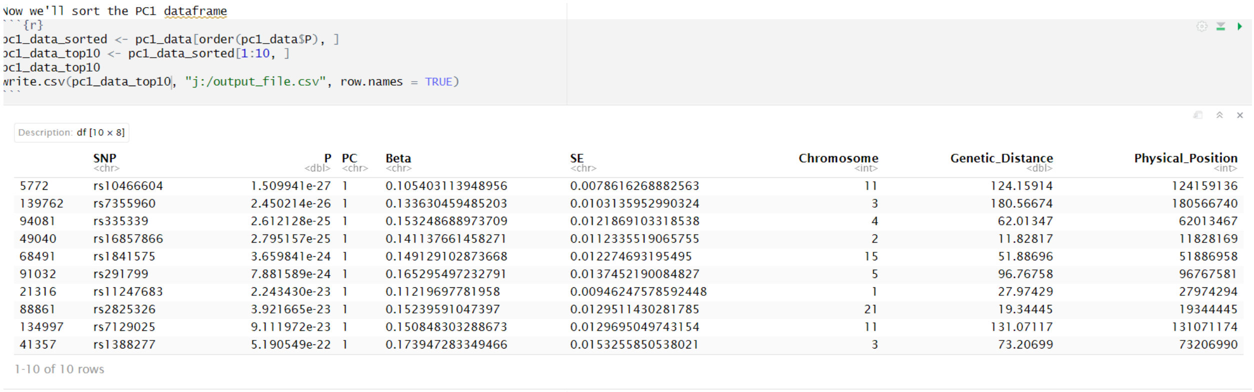


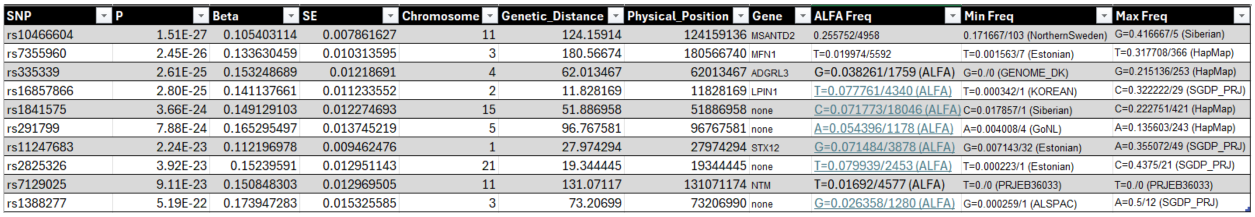
Figure 16: Caption for the image



~~~
Now we'll sort the PC1 dataframe
```{r}
pc1_data_sorted <- pc1_data[order(pc1_data$P), ]
pc1_data_top10 <- pc1_data_sorted[1:10, ]
pc1_data_top10
write.csv(pc1_data_top10, "j:/output_file.csv", row.names = TRUE)
```
~~~

```
Description: df [10 x 8]

        SNP                   P  PC   Beta                SE                      Chromosome  Genetic_Distance  Physical_Position
        <chr>             <dbl>  <chr> <chr>              <chr>                        <int>             <dbl>              <int>
5772    rs10466604  1.509941e-27  1   0.105403113948956  0.0078616268882563              11         124.15914          124159136
139762  rs7355960   2.450214e-26  1   0.133630459485203  0.0103135952990324               3         180.56674          180566740
94081   rs335339    2.612128e-25  1   0.153248688973709  0.0121869103318538               4          62.01347           62013467
49040   rs16857866  2.795157e-25  1   0.141137661458271  0.0112335519065755               2          11.82817           11828169
68491   rs1841575   3.659841e-24  1   0.149129102873668  0.012274693195495               15          51.88696           51886958
91032   rs291799    7.881589e-24  1   0.165295497232791  0.0137452190084827               5          96.76758           96767581
21316   rs11247683  2.243430e-23  1   0.11219697781958   0.00946247578592448              1          27.97429           27974294
88861   rs2825326   3.921665e-23  1   0.15239591047397   0.0129511430281785              21          19.34445           19344445
134997  rs7129025   9.111972e-23  1   0.150848303288673  0.0129695049743154              11         131.07117          131071174
41357   rs1388277   5.190549e-22  1   0.173947283349466  0.0153255850538021               3          73.20699           73206990

1-10 of 10 rows
```

Figure 17: co



SNP	P	Beta	SE	Chromosome	Genetic_Distance	Physical_Position	Gene	ALFA Freq	Min Freq	Max Freq
rs10466604	1.51E-27	0.105403114	0.007861627	11	124.15914	124159136	MSANTD2	0.255752/4958	0.171667/103 (NorthernSweden)	G=0.416667/5 (Siberian)
rs7355960	2.45E-26	0.133630459	0.010313595	3	180.56674	180566740	MFN1	T=0.019974/5592	T=0.001563/7 (Estonian)	T=0.317708/366 (HapMap)
rs335339	2.61E-25	0.153248689	0.01218691	4	62.013467	62013467	ADGRL3	G=0.038261/1759 (ALFA)	G=0./0 (GENOME_DK)	G=0.215136/253 (HapMap)
rs16857866	2.80E-25	0.141137661	0.011233552	2	11.828169	11828169	LPIN1	T=0.077761/4340 (ALFA)	T=0.000342/1 (KOREAN)	C=0.322222/29 (SGDP_PRJ)
rs1841575	3.66E-24	0.149129103	0.012274693	15	51.886958	51886958	none	C=0.071773/18046 (ALFA)	C=0.017857/1 (Siberian)	C=0.222751/421 (HapMap)
rs291799	7.88E-24	0.165295497	0.013745219	5	96.767581	96767581	none	A=0.054396/1178 (ALFA)	A=0.004008/4 (GoNL)	A=0.135603/243 (HapMap)
rs11247683	2.24E-23	0.112196978	0.009462476	1	27.974294	27974294	STX12	G=0.071484/3878 (ALFA)	G=0.007143/32 (Estonian)	A=0.355072/49 (SGDP_PRJ)
rs2825326	3.92E-23	0.15239591	0.012951143	21	19.344445	19344445	none	T=0.079939/2453 (ALFA)	T=0.000223/1 (Estonian)	C=0.4375/21 (SGDP_PRJ)
rs7129025	9.11E-23	0.150848303	0.012969505	11	131.07117	131071174	NTM	T=0.01692/4577 (ALFA)	T=0./0 (PRJEB36033)	T=0./0 (PRJEB36033)
rs1388277	5.19E-22	0.173947283	0.015325585	3	73.20699	73206990	none	G=0.026358/1280 (ALFA)	G=0.000259/1 (ALSPAC)	A=0.5/12 (SGDP_PRJ)

Figure 18: c

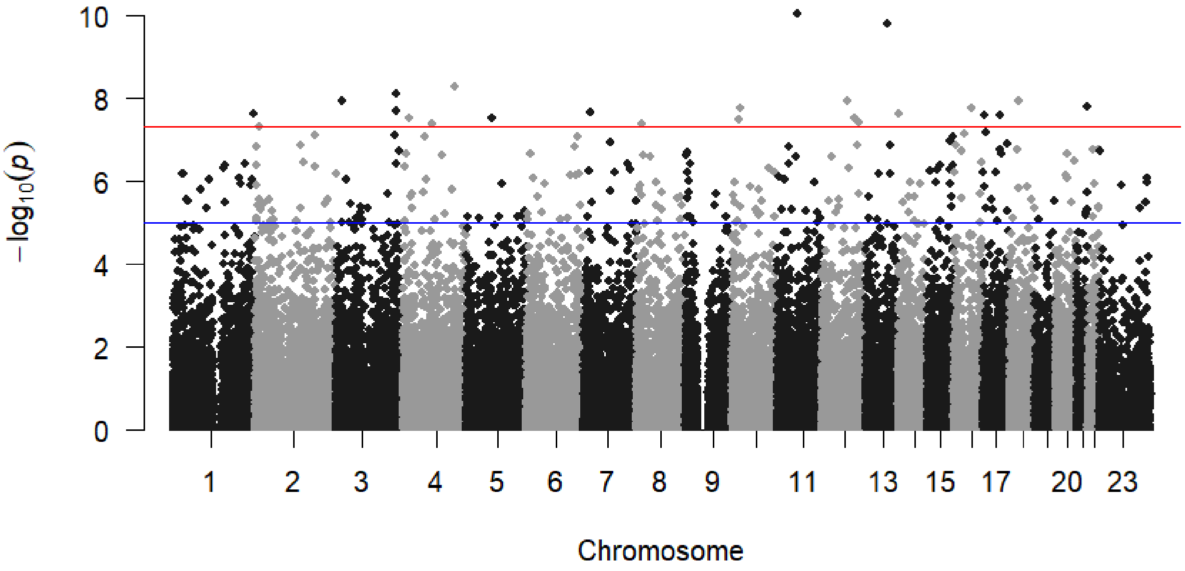
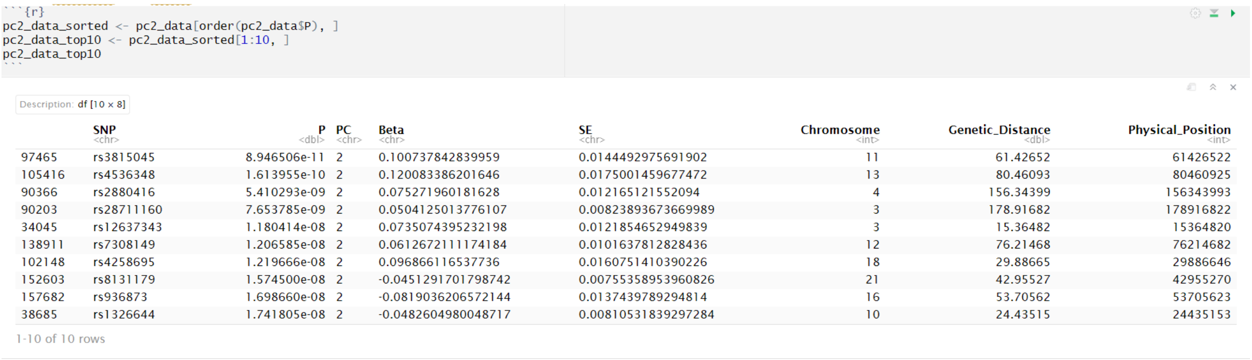


Figure 19: Caption for the image

```
#pc2_data_sorted <- pc2_data[order(pc2_data$P), ]
#pc2_data_top10 <- pc2_data_sorted[1:10, ]
#pc2_data_top10
#write.csv(pc2_data_top10, "j:/pc2_file.csv", row.names = TRUE)
```



~~~
```{r}
pc2_data_sorted <- pc2_data[order(pc2_data$P), ]
pc2_data_top10 <- pc2_data_sorted[1:10, ]
pc2_data_top10
```
~~~

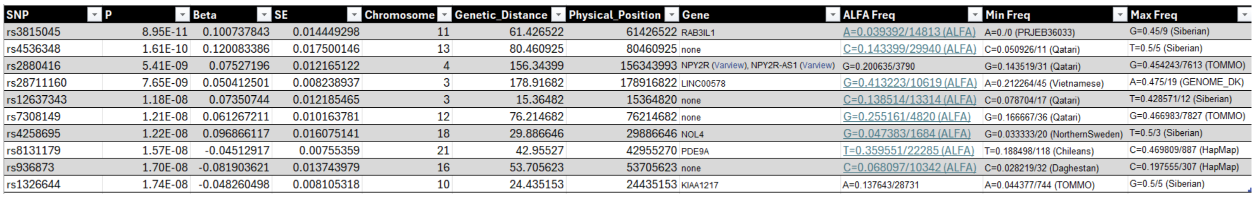
Description: df [10 x 8]

	SNP <chr>	P <dbl>	PC <chr>	Beta <chr>	SE <chr>	Chromosome <int>	Genetic_Distance <dbl>	Physical_Position <int>
97465	rs3815045	8.946506e-11	2	0.100737842839959	0.0144492975691902	11	61.42652	61426522
105416	rs4536348	1.613955e-10	2	0.120083386201646	0.0175001459677472	13	80.46093	80460925
90366	rs2880416	5.410293e-09	2	0.075271960181628	0.012165121552094	4	156.34399	156343993
90203	rs28711160	7.653785e-09	2	0.0504125013776107	0.00823893673669989	3	178.91682	178916822
34045	rs12637343	1.180414e-08	2	0.0735074395232198	0.0121854652949839	3	15.36482	15364820
138911	rs7308149	1.206585e-08	2	0.0612672111174184	0.0101637812828436	12	76.21468	76214682
102148	rs4258695	1.219666e-08	2	0.096866116537736	0.0160751410390226	18	29.88665	29886646
152603	rs8131179	1.574500e-08	2	-0.0451291701798742	0.00755358953960826	21	42.95527	42955270
157682	rs936873	1.698660e-08	2	-0.0819036206572144	0.0137439789294814	16	53.70562	53705623
38685	rs1326644	1.741805e-08	2	-0.0482604980048717	0.00810531839297284	10	24.43515	24435153

1-10 of 10 rows

Figure 20: co

```
#manhattan_data_3 <- pc3_data[, c("SNP", "Chromosome", "Physical_Position", "P")]
#colnames(manhattan_data_3) <- c("SNP", "CHR", "BP", "P")
#manhattan_data_3$P <- as.numeric(manhattan_data_3$P)
# Create the QQMan Manhattan plot
#manhattan(manhattan_data_3)
```



SNP	P	Beta	SE	Chromosome	Genetic_Distance	Physical_Position	Gene	ALFA Freq	Min Freq	Max Freq
rs3815045	8.95E-11	0.100737843	0.014449298	11	61.426522	61426522	RAB3IL1	A=0.039392/14813 (ALFA)	A=0./0 (PRJEB36033)	G=0.45/9 (Siberian)
rs4536348	1.61E-10	0.120083386	0.017500146	13	80.460925	80460925	none	C=0.143399/29940 (ALFA)	C=0.050926/11 (Qatari)	T=0.5/5 (Siberian)
rs2880416	5.41E-09	0.07527196	0.012165122	4	156.34399	156343993	NPY2R (Varview), NPY2R-AS1 (Varview)	G=0.200635/3790	G=0.143519/31 (Qatari)	G=0.454243/7613 (TOMMO)
rs28711160	7.65E-09	0.050412501	0.008238937	3	178.91682	178916822	LINC00578	G=0.413223/10619 (ALFA)	A=0.212264/45 (Vietnamese)	A=0.475/19 (GENOME_DK)
rs12637343	1.18E-08	0.07350744	0.012185465	3	15.36482	15364820	none	C=0.138514/13314 (ALFA)	C=0.078704/17 (Qatari)	T=0.428571/12 (Siberian)
rs7308149	1.21E-08	0.061267211	0.010163781	12	76.214682	76214682	none	G=0.255161/4820 (ALFA)	G=0.166667/36 (Qatari)	G=0.466983/7827 (TOMMO)
rs4258695	1.22E-08	0.096866117	0.016075141	18	29.886646	29886646	NOL4	G=0.047383/1684 (ALFA)	G=0.033333/20 (NorthernSweden)	T=0.5/3 (Siberian)
rs8131179	1.57E-08	-0.04512917	0.00755359	21	42.95527	42955270	PDE9A	T=0.359551/22285 (ALFA)	T=0.188498/118 (Chileans)	C=0.469809/887 (HapMap)
rs936873	1.70E-08	-0.081903621	0.013743979	16	53.705623	53705623	none	C=0.068097/10342 (ALFA)	C=0.028219/32 (Daghestan)	C=0.197555/307 (HapMap)
rs1326644	1.74E-08	-0.048260498	0.008105318	10	24.435153	24435153	KIAA1217	A=0.137643/28731	A=0.044377/744 (TOMMO)	G=0.5/5 (Siberian)

Figure 21: co

Here is the resulting one:

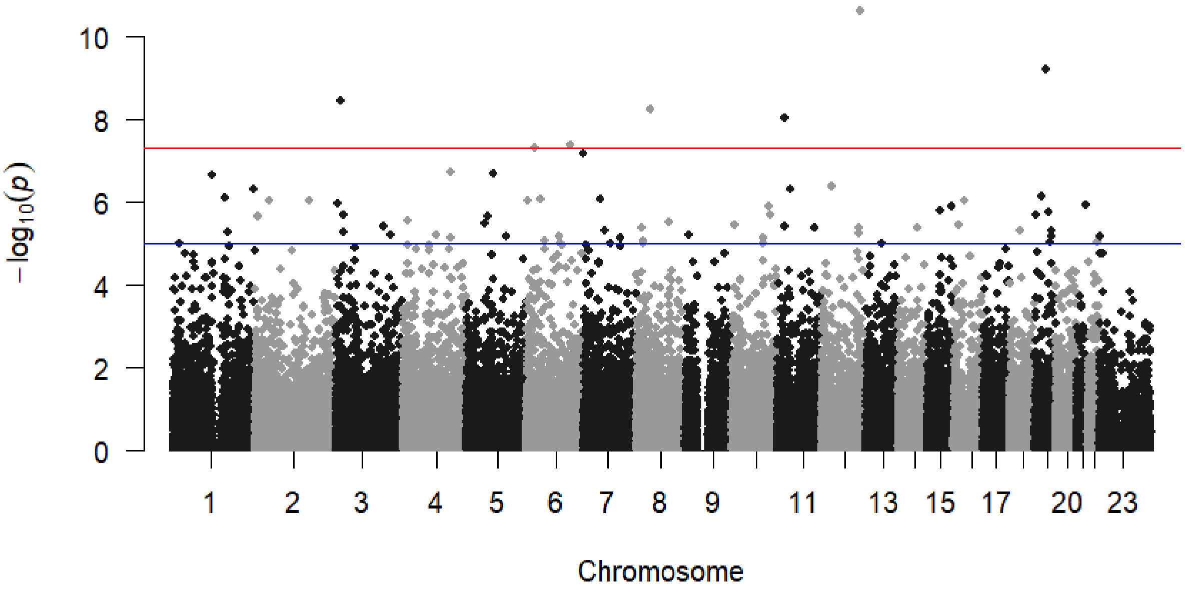


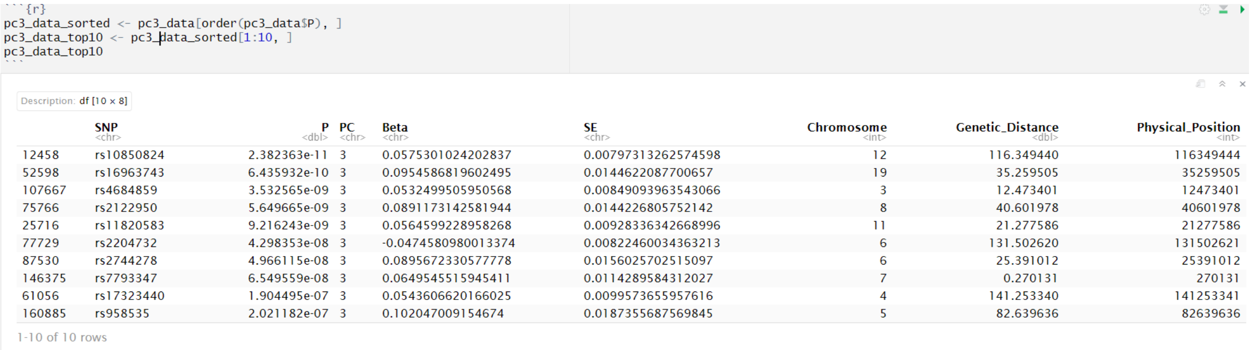
Figure 22: co

```
#pc3_data_sorted <- pc3_data[order(pc3_data$P), ]
#pc3_data_top10 <- pc3_data_sorted[1:10, ]
#pc3_data_top10
```

Here is the resulting table:

It’s enought to search for the previous 2 tables in dbSNP.

- Task 2.2 ---*-**- *We know from our previous knowledge about PCA that the new coordinates of the points are aquired by: D’ = Ut*Dt, where t is for transpose We know also that the U matrix is a d*r matrix. where d is the number of dimensions in the original space, and r is the new required dimensions of the reduced space. We want to map this to our problem, r should be 3 here, and if the PCA function of PLINK returns U not D’ we should have a matrix of 20 columns, and number of rows equal to number of SNPs which is ~56000. However, we see that the eigvec is consisting of 156 row, which is the number of samples, or n.



~~~
```{r}
pc3_data_sorted <- pc3_data[order(pc3_data$P), ]
pc3_data_top10 <- pc3_data_sorted[1:10, ]
pc3_data_top10
```
~~~

Description: df [10 x 8]

	SNP <chr>	P <dbl>	PC <chr>	Beta <chr>	SE <chr>	Chromosome <int>	Genetic_Distance <dbl>	Physical_Position <int>
12458	rs10850824	2.382363e-11	3	0.0575301024202837	0.00797313262574598	12	116.349440	116349444
52598	rs16963743	6.435932e-10	3	0.0954586819602495	0.0144622087700657	19	35.259505	35259505
107667	rs4684859	3.532565e-09	3	0.0532499505950568	0.00849093963543066	3	12.473401	12473401
75766	rs2122950	5.649665e-09	3	0.0891173142581944	0.0144226805752142	8	40.601978	40601978
25716	rs11820583	9.216243e-09	3	0.0564599228958268	0.00928336342668996	11	21.277586	21277586
77729	rs2204732	4.298353e-08	3	-0.0474580980013374	0.00822460034363213	6	131.502620	131502621
87530	rs2744278	4.966115e-08	3	0.0895672330577778	0.0156025702515097	6	25.391012	25391012
146375	rs7793347	6.549559e-08	3	0.0649545515945411	0.0114289584312027	7	0.270131	270131
61056	rs17323440	1.904495e-07	3	0.0543606620166025	0.0099573655957616	4	141.253340	141253341
160885	rs958535	2.021182e-07	3	0.102047009154674	0.0187355687569845	5	82.639636	82639636

1-10 of 10 rows

Figure 23: co

```r
Xr <- t(eigvec[, 3:5])
Y <- t(genotype_data_filtered[, 2:ncol(genotype_data_filtered)])
nrow(Xr)
```

```
## [1] 3
```

```r
ncol(Xr)
```

```
## [1] 156
```

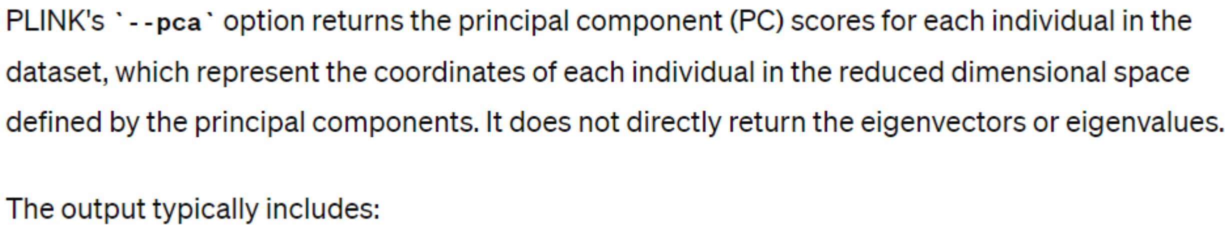
```r
nrow(Y)
```

```
## [1] 55226
```

```r
ncol(Y)
```

```
## [1] 156
```

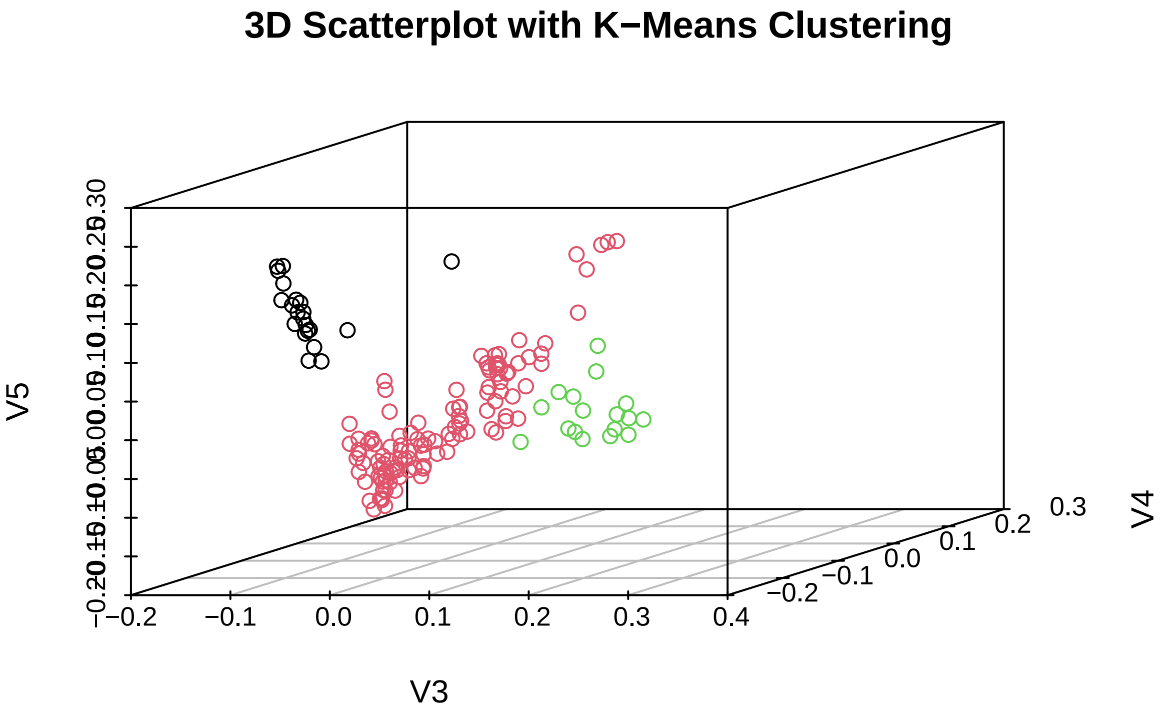
We deduce that PLINK returns the matrix D', dimensions of points in the new space of 3d. So we should apply the clustering on it directly. I asked chatgpt to make sure of this, it confirmed.



PLINK's `--pca` option returns the principal component (PC) scores for each individual in the dataset, which represent the coordinates of each individual in the reduced dimensional space defined by the principal components. It does not directly return the eigenvectors or eigenvalues.

The output typically includes:

```r
set.seed(123)
k <- 3
kmeans_result <- kmeans(eigvec[, 3:5], centers = k)

scatterplot3d(eigvec[, 3:5], color = kmeans_result$cluster, main = "3D Scatterplot with K-Means Cluster
```



```
new_df <- as.data.frame(eigvec[, 3:5])

# Create one-hot encoded cluster labels
cluster_labels <- matrix(0, nrow = nrow(new_df), ncol = k)
for (i in 1:k) {
  cluster_labels[kmeans_result$cluster == i, i] <- 1
}

# Add one-hot encoded cluster labels to the new dataframe
colnames(cluster_labels) <- paste0("cluster", 1:k)
new_df <- cbind(new_df, cluster_labels)
```

For saving time, the TA permitted us to do the work on 1000 - 10000 SNP instead of all of the data set, we'll work on 5000.

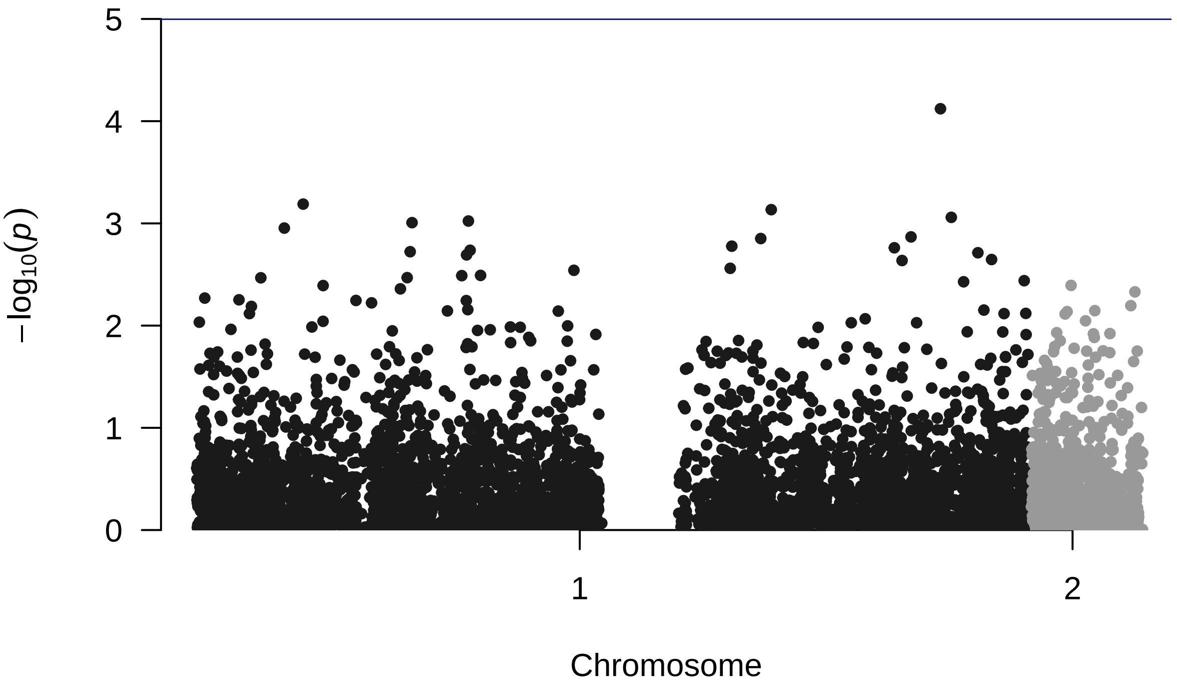
```
selected_columns <- genotype_data_filtered[, seq(2, ncol(genotype_data_filtered), by = 11)]
data <- cbind(new_df, selected_columns)
```

```
num_tests <- (ncol(data)) * 3
result_df1 <- data.frame(SNP = character(num_tests), P = numeric(num_tests), cluster = integer(num_test
ptr = 1

for (i in 1:k) {
  for (j in 7:ncol(data)) {
    formula <- reformulate(c(paste0(names(data)[j]), paste0("V", 3), paste0("V", 4), paste0("V", 5)), r
    model <- lm(formula, data = data)
    summary_coef <- summary(model)$coefficients
    p_value <- summary_coef[2, "Pr(>|t|)"]
    beta <- summary_coef[2, "Estimate"]
    se <- summary_coef[2, "Std. Error"]
    result_df1[ptr, ] <- c(names(genotype_data_filtered)[j], p_value, i, beta, se)
    ptr = ptr + 1
   }
}
```

```
modified_result_df1 <- result_df1
modified_result_df1$SNP <- substring(result_df1$SNP, 1, nchar(result_df1$SNP) - 2)
merged_data1 <- merge(modified_result_df1, map_data, by.x = "SNP", by.y = "SNP_ID", all.x = TRUE)
merged_data1 <- na.omit(merged_data1)
merged_data1$P <- as.numeric(merged_data1$P)
```

```
c1_data <- merged_data1[merged_data1$cluster == "1", ]
c2_data <- merged_data1[merged_data1$cluster == "2", ]
c3_data <- merged_data1[merged_data1$cluster == "3", ]
```

```
manhattan_data_1 <- c1_data[, c("SNP", "Chromosome", "Physical_Position", "P")]
colnames(manhattan_data_1) <- c("SNP", "CHR", "BP", "P")
manhattan_data_1$P <- as.numeric(manhattan_data_1$P)
# Create the QQMan Manhattan plot
manhattan(manhattan_data_1)
```

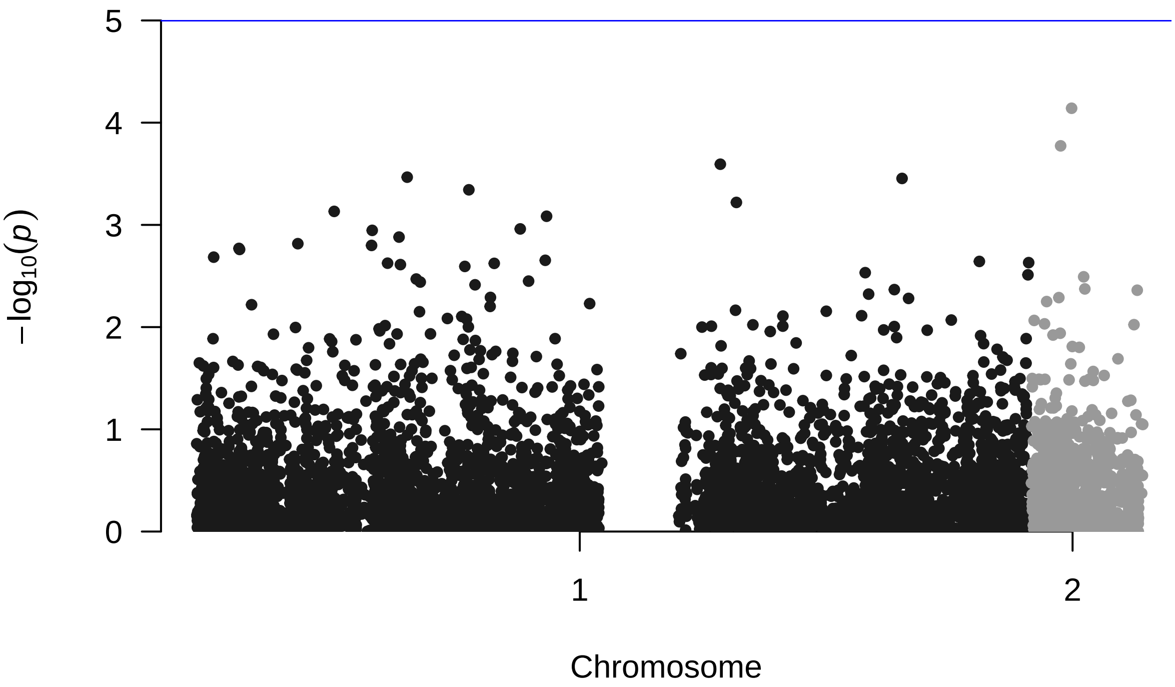


No SNP is significantly associated with first cluster

```
c1_data_sorted <- c1_data[order(c1_data$P), ]
c1_data_top10 <- c1_data_sorted[1:10, ]
c1_data_top10
```

```
##              SNP            P cluster                Beta                SE
## 6122  rs17162892 7.560011e-05       1  0.0866962622210513 0.0213019793355854
## 14282   rs804429 6.475575e-04       1 -0.0621534366554175 0.0178417985831305
## 5341  rs16844658 7.340055e-04       1 -0.0707296246838893 0.0205180832739289
## 9572   rs3913657 8.727141e-04       1  0.0597698425679151 0.0175978156668449
## 12244  rs6671683 9.477386e-04       1 -0.0467057475360225 0.0138508099293416
## 11299   rs555146 9.836841e-04       1  0.0487687651396717 0.0145101371826581
## 10858  rs4908343 1.111680e-03       1   0.051925594743484 0.0156189330744395
## 5743  rs17020918 1.355751e-03       1 -0.0517753429763544 0.0158591723478737
## 14923   rs961404 1.406868e-03       1 -0.0641849722296378 0.0197282958483413
## 2578  rs11587040 1.671335e-03       1  -0.038814668408714  0.012126850922408
##       Chromosome Genetic_Distance Physical_Position
## 6122           1        220.42402         220424015
## 14282          1         33.33810          33338100
## 5341           1        170.74873         170748734
## 9572           1        223.59646         223596463
## 12244          1         81.84146          81841463
## 11299          1         65.29999          65299995
## 10858          1         27.80429          27804285
## 5743           1        211.77346         211773462
## 14923          1        167.67247         167672474
## 2578           1        159.16641         159166407
```

```
manhattan_data_2 <- c2_data[, c("SNP", "Chromosome", "Physical_Position", "P")]
colnames(manhattan_data_2) <- c("SNP", "CHR", "BP", "P")
manhattan_data_2$P <- as.numeric(manhattan_data_2$P)
manhattan(manhattan_data_2)
```

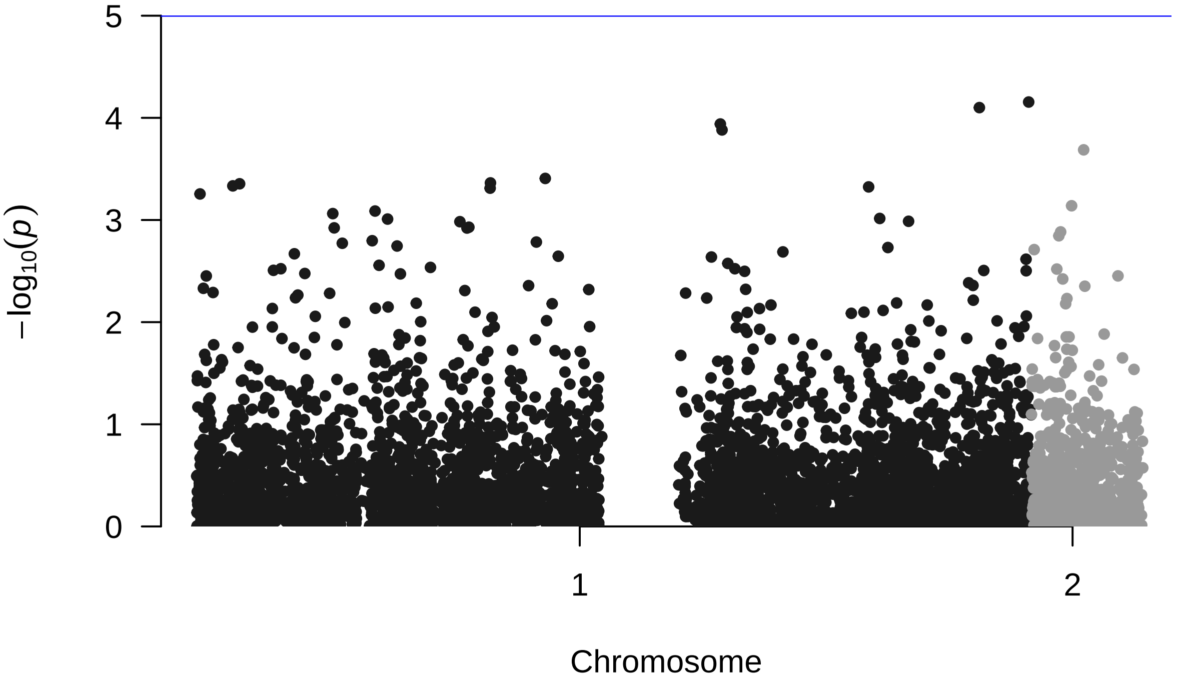


No one again is significanlty associated with second cluster

```
c2_data_sorted <- c2_data[order(c2_data$P), ]
c2_data_top10 <- c2_data_sorted[1:10, ]
c2_data_top10
```

```
##              SNP            P cluster                Beta                 SE
## 5498  rs16857866 7.227758e-05       2  -0.123323034152977  0.030215251367673
## 9157    rs345903 1.685911e-04       2  0.0769376203809053 0.0199398773382473
## 7701   rs2244798 2.548312e-04       2  -0.178407619417091  0.047622427672023
## 7748   rs2269252 3.413882e-04       2   0.105855789863394 0.0288797511105179
## 14651   rs924569 3.511197e-04       2 -0.0717764529648769 0.0196241648420027
## 3377  rs12090448 4.537164e-04       2   0.117295293031882 0.0327143030517655
## 13516  rs7525955 6.027103e-04       2 -0.0694915618341067 0.0198303133491292
## 12133  rs6665972 7.377420e-04       2  0.0721657692523048 0.0209436945324718
## 2932  rs12021686 8.228888e-04       2 -0.0915899554837591 0.0268298510324016
## 3209  rs12060538 1.094881e-03       2  0.0785154506940166 0.0235846841048672
##       Chromosome Genetic_Distance Physical_Position
## 5498           2         11.82817          11828169
## 9157           2          8.60861           8608610
## 7701           1        155.79459         155794587
## 7748           1         63.87065          63870653
## 14651          1        209.14624         209146241
## 3377           1         81.97615          81976150
## 13516          1        160.51446         160514456
## 12133          1         42.44611          42446110
## 2932           1        104.79172        104791721
## 3209           1         97.05897         97058969
```

```
manhattan_data_3 <- c3_data[, c("SNP", "Chromosome", "Physical_Position", "P")]
colnames(manhattan_data_3) <- c("SNP", "CHR", "BP", "P")
manhattan_data_3$P <- as.numeric(manhattan_data_3$P)
manhattan(manhattan_data_3)
```



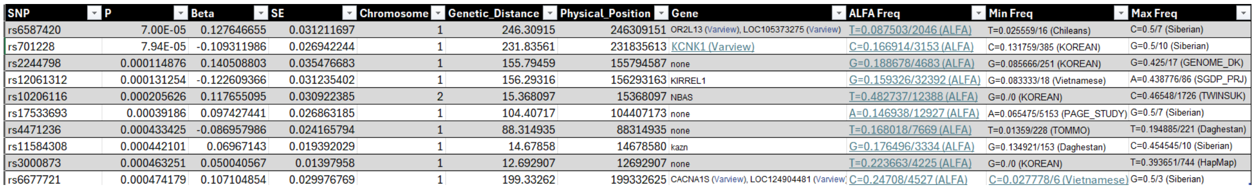
One only SNP is significantly associated to the third cluster we shall search for it.

```
c3_data_sorted <- c3_data[order(c3_data$P), ]
c3_data_top10 <- c3_data_sorted[1:10, ]
c3_data_top10
```

```
##              SNP            P cluster                Beta                 SE
## 11879  rs6587420 7.001009e-05       3    0.12764665504888 0.0312116970971629
## 13110   rs701228 7.936884e-05       3   -0.10931198581659 0.0269422442730761
## 7699   rs2244798 1.148761e-04       3   0.140508802695761 0.0354766831124132
## 3211  rs12061312 1.312542e-04       3  -0.122609365933987 0.0312354015702992
## 225   rs10206116 2.056257e-04       3   0.117655094639041 0.0309223852742068
## 6571  rs17533693 3.918598e-04       3  0.0974274413365295 0.0268631853975606
## 10070  rs4471236 4.334250e-04       3 -0.0869579861890864 0.0241657943419837
## 2536  rs11584308 4.421008e-04       3  0.0696714304688626 0.0193920290827495
## 8888   rs3000873 4.632510e-04       3  0.0500405671516608 0.0139795795819277
## 12343  rs6677721 4.741785e-04       3   0.107104853530145 0.0299767693480647
##       Chromosome Genetic_Distance Physical_Position
## 11879          1        246.30915         246309151
## 13110          1        231.83561         231835613
## 7699           1        155.79459         155794587
## 3211           1        156.29316         156293163
## 225            2         15.36810          15368097
## 6571           1        104.40717         104407173
## 10070          1         88.31494          88314935
## 2536           1         14.67858          14678580
## 8888           1         12.69291          12692907
## 12343          1        199.33262         199332625
```

NO significant SNP in relation to any cluster. However we'll do the research in the dbSNP becqause it's required.

After researching, here is the result.



SNP	P	Beta	SE	Chromosome	Genetic_Distance	Physical_Position	Gene	ALFA Freq	Min Freq	Max Freq
rs6587420	7.00E-05	0.127646655	0.031211697	1	246.30915	246309151	OR2L13 (Varview), LOC105373275 (Varview)	T=0.087503/2046 (ALFA)	T=0.025559/16 (Chileans)	C=0.5/7 (Siberian)
rs701228	7.94E-05	-0.109311986	0.026942244	1	231.83561	231835613	KCNK1 (Varview)	C=0.166914/3153 (ALFA)	C=0.131759/385 (KOREAN)	G=0.5/10 (Siberian)
rs2244798	0.000114876	0.140508803	0.035476683	1	155.79459	155794587	none	G=0.188678/4683 (ALFA)	G=0.085666/251 (KOREAN)	G=0.425/17 (GENOME_DK)
rs12061312	0.000131254	-0.122609366	0.031235402	1	156.29316	156293163	KIRREL1	G=0.159326/32392 (ALFA)	G=0.083333/18 (Vietnamese)	A=0.438776/86 (SGDP_PRJ)
rs10206116	0.000205626	0.117655095	0.030922385	2	15.368097	15368097	NBAS	T=0.482737/12388 (ALFA)	G=0./0 (KOREAN)	C=0.46548/1726 (TWINSUK)
rs17533693	0.00039186	0.097427441	0.026863185	1	104.40717	104407173	none	A=0.146938/12927 (ALFA)	A=0.065475/5153 (PAGE_STUDY)	G=0.5/7 (Siberian)
rs4471236	0.000433425	-0.086957986	0.024165794	1	88.314935	88314935	none	T=0.168018/7669 (ALFA)	T=0.01359/228 (TOMMO)	T=0.194885/221 (Daghestan)
rs11584308	0.000442101	0.06967143	0.019392029	1	14.67858	14678580	kazn	G=0.176496/3334 (ALFA)	G=0.134921/153 (Daghestan)	C=0.454545/10 (Siberian)
rs3000873	0.000463251	0.050040567	0.01397958	1	12.692907	12692907	none	T=0.223663/4225 (ALFA)	G=0./0 (KOREAN)	T=0.393651/744 (HapMap)
rs6677721	0.000474179	0.107104854	0.029976769	1	199.33262	199332625	CACNA1S (Varview), LOC124904481 (Varview	C=0.24708/4527 (ALFA)	C=0.027778/6 (Vietnamese)	G=0.5/3 (Siberian)

Figure 24: co

THANK YOU!