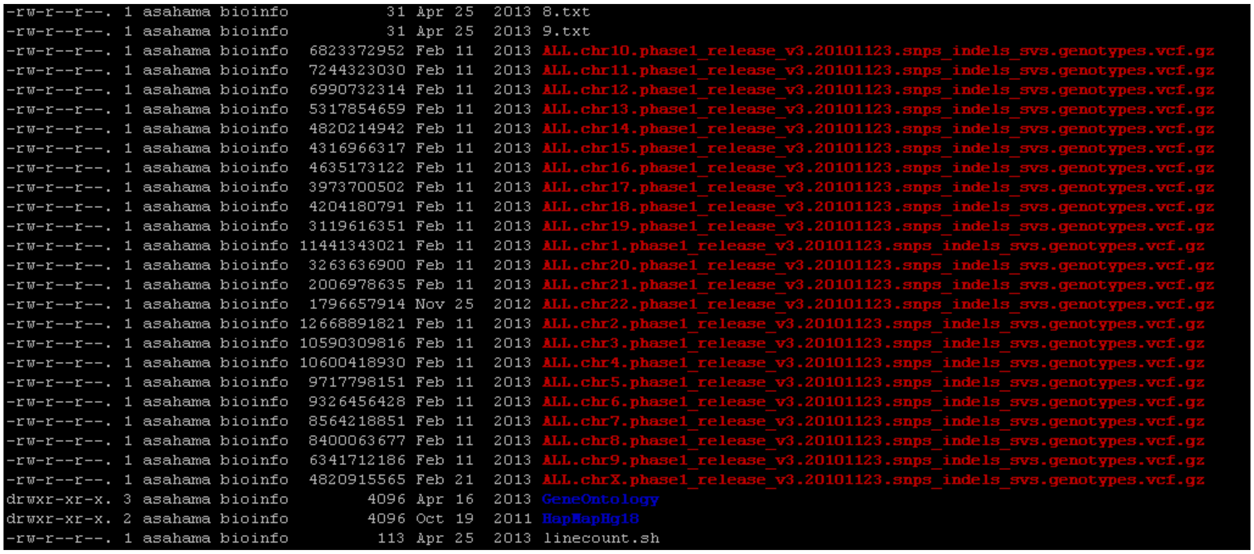
1.8 Appendix

1.8.1 Protocol for Assessing the Total Number of Genomic Differences in the 1000 Genomes Database

The 1000 Genomes data are saved in 22 (.gz) compressed files. Each file contains the complete data of 1,092 individuals of 1 chromosome.



```
-rw-r--r--. 1 asahama bioinfo          31 Apr 25  2013 8.txt
-rw-r--r--. 1 asahama bioinfo          31 Apr 25  2013 9.txt
-rw-r--r--. 1 asahama bioinfo  6823372952 Feb 11  2013 ALL.chr10.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  7244323030 Feb 11  2013 ALL.chr11.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  6990732314 Feb 11  2013 ALL.chr12.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  5317854659 Feb 11  2013 ALL.chr13.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  4820214942 Feb 11  2013 ALL.chr14.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  4316966317 Feb 11  2013 ALL.chr15.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  4635173122 Feb 11  2013 ALL.chr16.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  3973700502 Feb 11  2013 ALL.chr17.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  4204180791 Feb 11  2013 ALL.chr18.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  3119616351 Feb 11  2013 ALL.chr19.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo 11441343021 Feb 11  2013 ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  3263636900 Feb 11  2013 ALL.chr20.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  2006978635 Feb 11  2013 ALL.chr21.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  1796657914 Nov 25  2012 ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo 12668891821 Feb 11  2013 ALL.chr2.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo 10590309816 Feb 11  2013 ALL.chr3.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo 10600418930 Feb 11  2013 ALL.chr4.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  9717798151 Feb 11  2013 ALL.chr5.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  9326456428 Feb 11  2013 ALL.chr6.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  8564218851 Feb 11  2013 ALL.chr7.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  8400063677 Feb 11  2013 ALL.chr8.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  6341712186 Feb 11  2013 ALL.chr9.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
-rw-r--r--. 1 asahama bioinfo  4820915565 Feb 21  2013 ALL.chrX.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
drwxr-xr-x. 3 asahama bioinfo        4096 Apr 16  2013 GeneOntology
drwxr-xr-x. 2 asahama bioinfo        4096 Oct 19  2011 HapMapHg18
-rw-r--r--. 1 asahama bioinfo         113 Apr 25  2013 linecount.sh
```

- **First Program:**
  - ***IDs_seperator.pl:***

  This program is designed to skim genomic data of any one or two populations from the whole 22 chromosomes files. The population's skimmed data then become easier and faster to be processed.

  To run this program you will need to download 3 files "available in the website" into the same directory of the *IDs_seperator.pl* program.

  1. 20111108_1000genomes_samples2.txt

  This .txt file contains details and Identifier for all individuals that were actually sequences through the phase 1 of the 1K Genomes project. this fie should be in the same directory of the IDs_seperator program.

  2. head100_1000genome_testfile

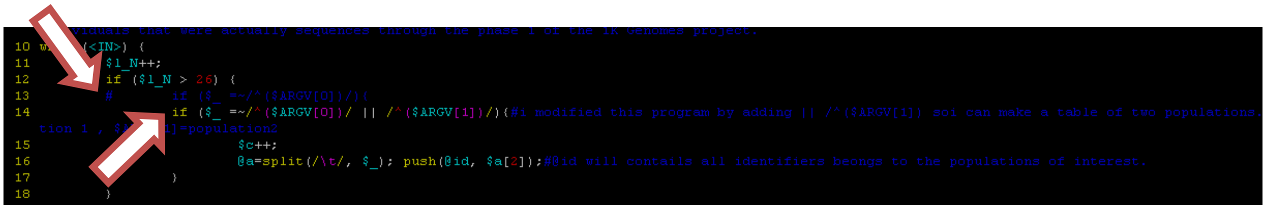
  This file contains the first 100 lines from the original table that has 1092 individual autosomes data.

  3. filenames

  A text file contains the names of the 22 chromosomes files.

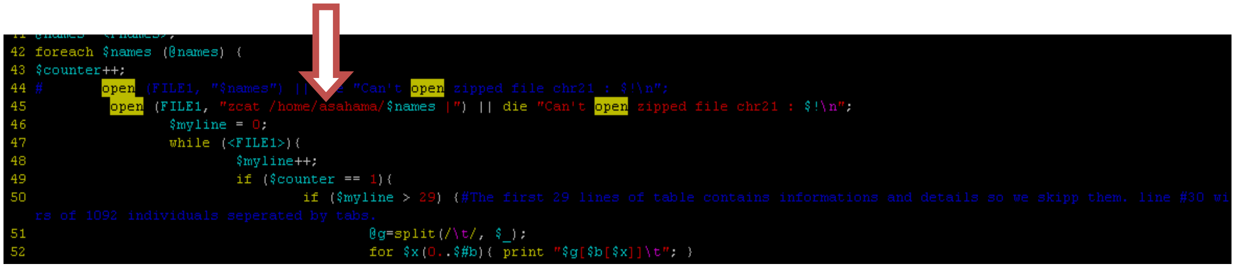
Use Line 14 and comment line 13 (If u would like to calculate genomic differences between two populations only)

Use Line 13 and comment line 14 (If u would like to calculate genomic differences in within one population only).



```
   viduals that were actually sequences through the phase 1 of the 1K Genomes project.
10 while (<IN>) {
11         $1_N++;
12         if ($1_N > 26) {
13         #       if ($_ =~/^($ARGV[0])/){
14                 if ($_ =~/^($ARGV[0])/ || /^($ARGV[1])/){#i modified this program by adding || /^($ARGV[1]) soi can make a table of two populations.
   tion 1 , $ARGV[1]=population2
15                         $c++;
16                         @a=split(/\t/, $_); push(@id, $a[2]);#@id will contails all identifiers beongs to the populations of interest.
17                 }
18         }
```

You will need to provide the path to the 22 chromosomes data in line 45.



```
42 foreach $names (@names) {
43 $counter++;
44 #       open (FILE1, "$names") || die "Can't open zipped file chr21 : $!\n";
45          open (FILE1, "zcat /home/asahama/$names |") || die "Can't open zipped file chr21 : $!\n";
46                  $myline = 0;
47                  while (<FILE1>){
48                          $myline++;
49                          if ($counter == 1){
50                                  if ($myline > 29) {#The first 29 lines of table contains informations and details so we skipp them. line #30 wi
   rs of 1092 individuals seperated by tabs.
51                                          @g=split(/\t/, $_);
52                                          for $x(0..$#b){ print "$g[$b[$x]]\t"; }
```

- ***Command line example: (one population):***
  - ***" one population":***
    nohup perl IDs_seperator.pl YRI > YRI_Genome_HOM_HET.txt &

    The created file will contain autosomal data for YRI individuals only.

  - ***" Two population" :***
    nohup perl IDs_seperator.pl YRI LWK> YRI_LWK_Genome_HOM_HET.txt &

    The created file will contain autosomal data for YRI and LWK individuals only.

- **Second Program:**

- ***no_diff_deleter.pl:***

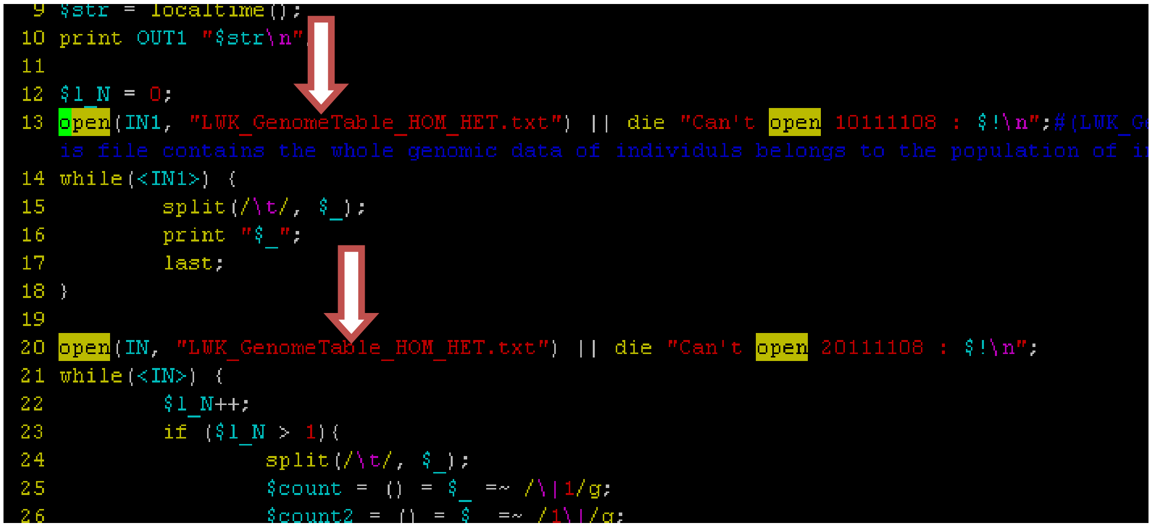
After creating a table of one or two population genomes by IDs_seperator.pl. The *no_diff_deleter.pl* program is designed to find and delete any 0|0 alleles that is shared by all of the individuals in that population (no difference between all the individual across all populations).

The result from this program will be a table that contains at least one mutant alleles at each position.

- ***Command line to execute this program:***

  nohup perl no_diff_deleter.pl > LWK_Genome_Het.txt & .

You will need to replace the file handle file name on lines 13 and 20 with the name of the table created by the *IDs_seperator.pl*.



```
 9 $str = localtime();
10 print OUT1 "$str\n";
11
12 $l_N = 0;
13 open(IN1, "LWK_GenomeTable_HOM_HET.txt") || die "Can't open 10111108 : $!\n";#(LWK_Ge
   is file contains the whole genomic data of individuls belongs to the population of i
14 while(<IN1>) {
15         split(/\t/, $_);
16         print "$_";
17         last;
18 }
19
20 open(IN, "LWK_GenomeTable_HOM_HET.txt") || die "Can't open 20111108 : $!\n";
21 while(<IN>) {
22         $l_N++;
23         if ($l_N > 1){
24                 split(/\t/, $_);
25                 $count = () = $_ =~ /\|1/g;
26                 $count2 = () = $_ =~ /1\|/g;
```

The file created by ***no_diff_deleter.pl*** will be used to calculate the total number of differences between each pair of individual using:

Intra_PopGenomeDif.pl
(Differences between pairs of the same population)

or

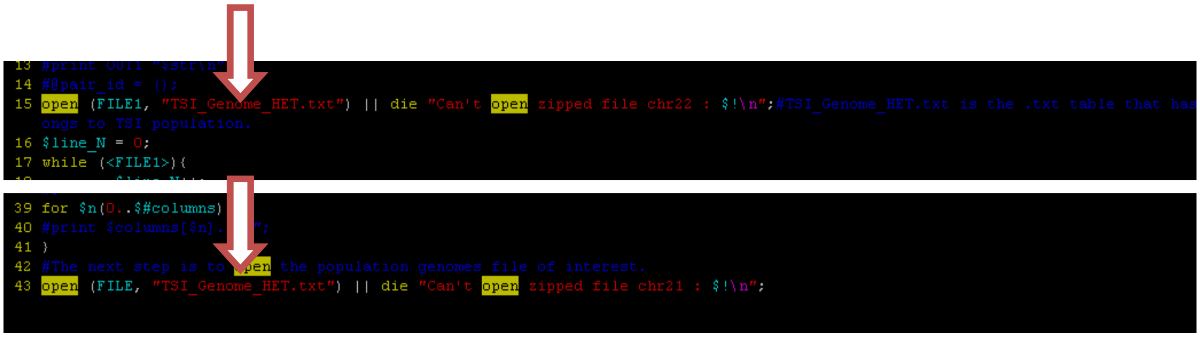
Inter_PopGenomeDif.pl
(Differences between pairs of different populations).

- **Third Program:**
  - ***Intra_PopGenomeDif.pl***

This program is created to calculate the total number of genomic variants differences between each pair of individuals within the same population.

To run this program:

Change the file name in lines 15 and 43 to the name of your file of your population of interest.



```
13 #print OUT1 "$str\n";
14 #@pair_id = ();
15 open (FILE1, "TSI_Genome_HET.txt") || die "Can't open zipped file chr22 : $!\n";#TSI_Genome_HET.txt is the .txt table that has
   ongs to TSI population.
16 $line_N = 0;
17 while (<FILE1>){
```

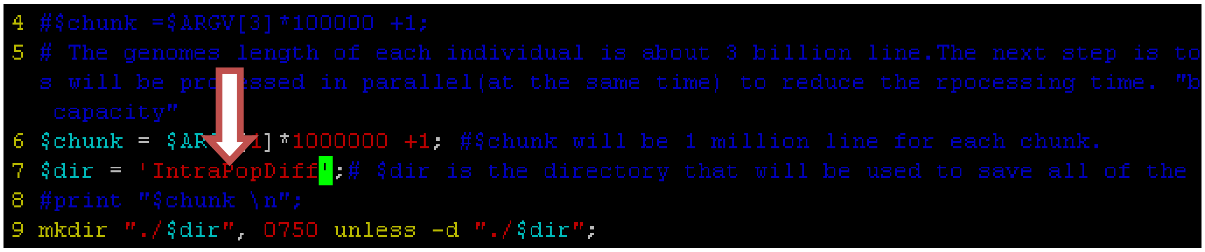
```
39 for $n(0..$#columns)
40 #print $columns[$n]."  ";
41 }
42 #The next step is to open the population genomes file of interest.
43 open (FILE, "TSI_Genome_HET.txt") || die "Can't open zipped file chr21 : $!\n";
```

This program is designed to process the genomic data per chunks of 1,000,000 lines. So you will need to find how many lines in your table (TSI_RareSNPs_Table.txt) in order to know how many times to run the program to process the entire data in that table.

For example: If your TSI_RareSNPs_Table.txt is 5,405,313 lines then you will need to run the program 6 times (1 time for each chunk) as below commands.

```
11  nohup perl Intra_PopGenomeDif.pl  TSI 0 &
12  nohup perl Intra_PopGenomeDif.pl  TSI 1 &
13  nohup perl Intra_PopGenomeDif.pl  TSI 2 &
14  nohup perl Intra_PopGenomeDif.pl  TSI 3 &
15  nohup perl Intra_PopGenomeDif.pl  TSI 4 &
16  nohup perl Intra_PopGenomeDif.pl  TSI 5 &
```

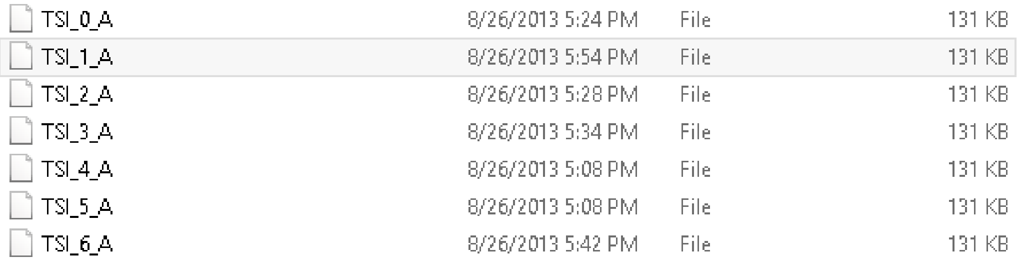
The results will be saved into a folder (IntraPopDiff) in the same directory of the program. To change the folder name in the program you will need to change the name in line 7



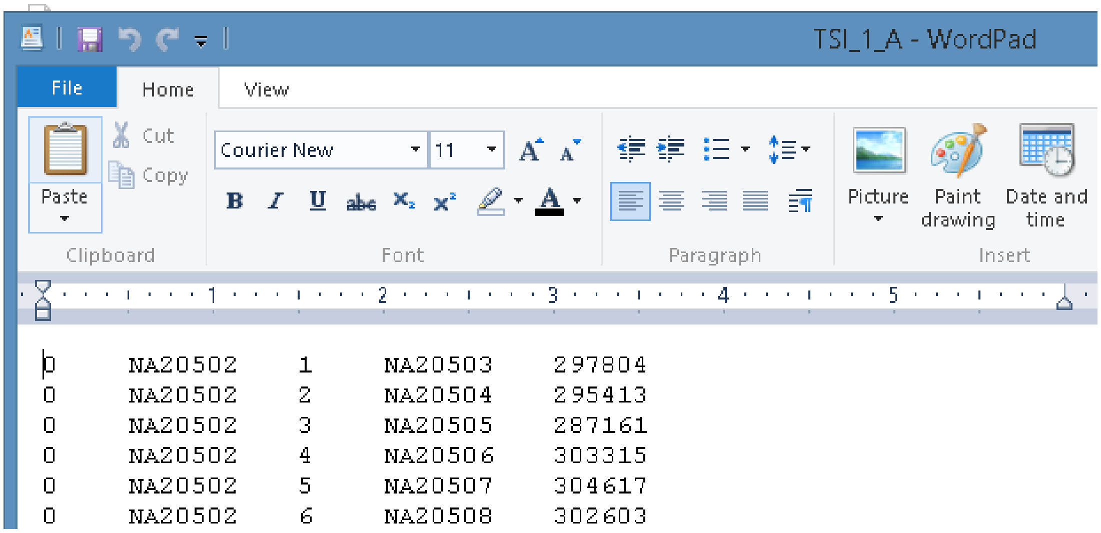
```
4 #$chunk =$ARGV[3]*100000 +1;
5 # The genomes length of each individual is about 3 billion line.The next step is to
  s will be processed in parallel(at the same time) to reduce the rpocessing time. "b
   capacity"
6 $chunk = $ARGV[1]*1000000 +1; #$chunk will be 1 million line for each chunk.
7 $dir = 'IntraPopDiff';# $dir is the directory that will be used to save all of the
8 #print "$chunk \n";
9 mkdir "./$dir", 0750 unless -d "./$dir";
```

Each chunk file will contain the total differences between each pair of individuals for the number of lines in that specific chunk.

To calculate the total number of genomic differences you will need to combine the chunks results by simply using Microsoft Excel.



```
TSI_0_A    8/26/2013 5:24 PM    File    131 KB
TSI_1_A    8/26/2013 5:54 PM    File    131 KB
TSI_2_A    8/26/2013 5:28 PM    File    131 KB
TSI_3_A    8/26/2013 5:34 PM    File    131 KB
TSI_4_A    8/26/2013 5:08 PM    File    131 KB
TSI_5_A    8/26/2013 5:08 PM    File    131 KB
TSI_6_A    8/26/2013 5:42 PM    File    131 KB
```



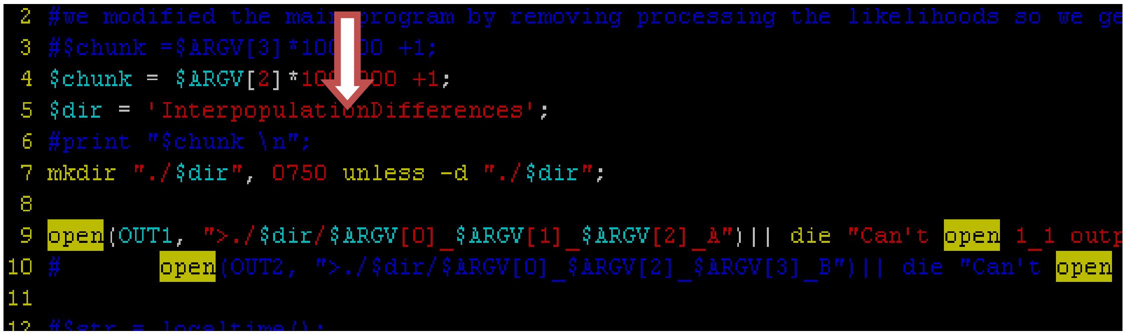
```
0       NA20502     1      NA20503      297804
0       NA20502     2      NA20504      295413
0       NA20502     3      NA20505      287161
0       NA20502     4      NA20506      303315
0       NA20502     5      NA20507      304617
0       NA20502     6      NA20508      302603
```

- **Fourth Program:**

– ***Inter_PopGenomeDif.pl***

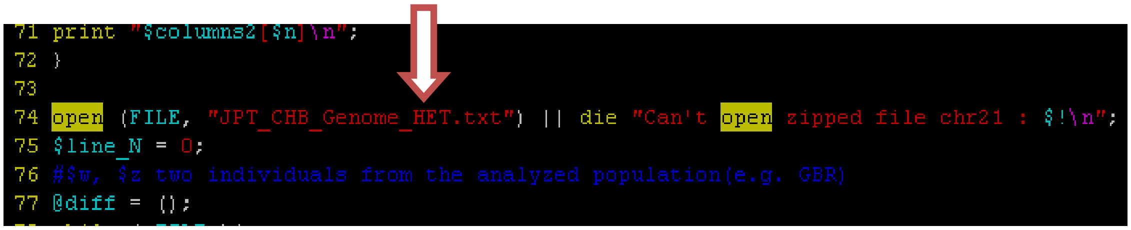
This program created to count the total number of genomic differences between each pair of individuals from two different populations.

All results from this file will be saved into a folder name " InterpopulationDifferences" and this folder name can be change through changing it in line 5.

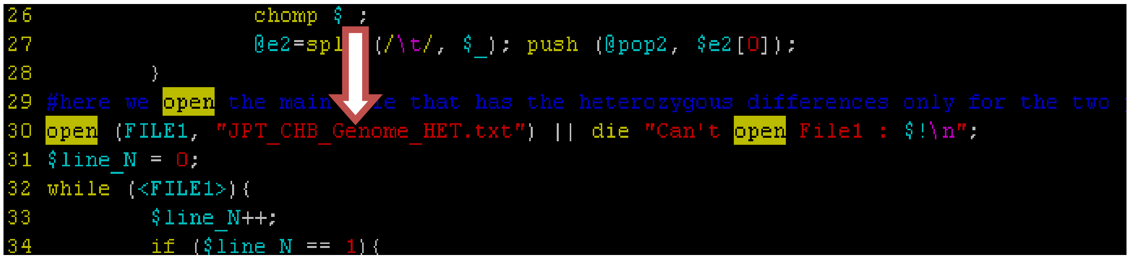


```
 2 #we modified the main program by removing processing the likelihoods so we ge
 3 #$chunk =$ARGV[3]*1000000 +1;
 4 $chunk = $ARGV[2]*1000000 +1;
 5 $dir = 'InterpopulationDifferences';
 6 #print "$chunk \n";
 7 mkdir "./$dir", 0750 unless -d "./$dir";
 8
 9 open(OUT1, ">./$dir/$ARGV[0]_$ARGV[1]_$ARGV[2]_A")|| die "Can't open 1_1 outp
10 #         open(OUT2, ">./$dir/$ARGV[0]_$ARGV[2]_$ARGV[3]_B")|| die "Can't open
11
12 #$str = localtime();
```

You will also need to change line 30 and line 74 to include the name of the table that have your populations of interest.



```
71 print "$columns2[$n]\n";
72 }
73
74 open (FILE, "JPT_CHB_Genome_HET.txt") || die "Can't open zipped file chr21 : $!\n";
75 $line_N = 0;
76 #$w, $z two individuals from the analyzed population(e.g. GBR)
77 @diff = ();
```



```
26                      chomp $_;
27                      @e2=spl (/\t/, $_); push (@pop2, $e2[0]);
28              }
29 #here we open the main e that has the heterozygous differences only for the two
30 open (FILE1, "JPT_CHB_Genome_HET.txt") || die "Can't open File1 : $!\n";
31 $line_N = 0;
32 while (<FILE1>){
33              $line_N++;
34              if ($line_N == 1){
```

To run this program you will need to have a genome table of two populations of your interest created by the *IDs_seperator.pl* and processed by the *no_diff_deleter.pl*.

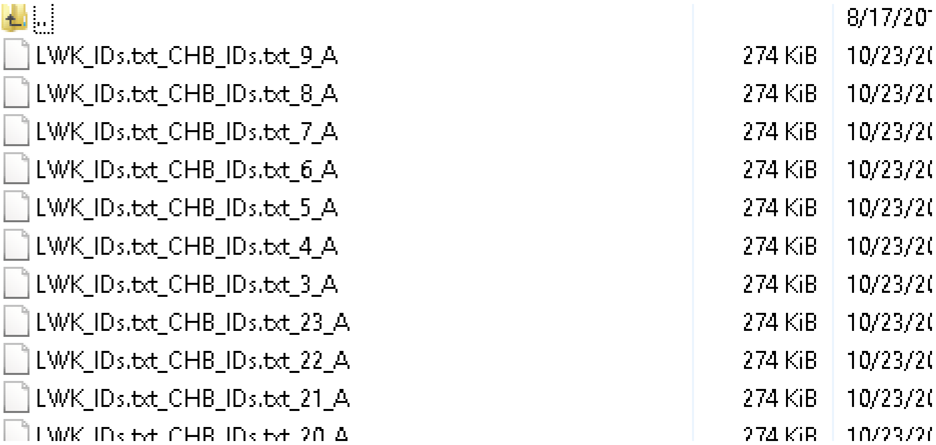
For example: "LWK_YRI_Genome_HOM_HET.txt"

You will also need to make two ".txt" files, each file will contain the identifiers of one population.

Let's assume we made LWK_IDs.txt and YRI_IDs.txt. The command line to run this program and process the table in chunks will be as the example below:

```
243 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 1 &
244 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 2 &
245 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 3 &
246 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 4 &
247 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 5 &
248 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 6 &
249 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 7 &
250 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 8 &
251 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 9 &
252 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 10 &
253 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 11 &
254 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 12 &
255 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 13 &
256 nohup perl Inter_PopGenomeDif.pl LWK_IDs.txt YRI_IDs.txt 14 &
```

To calculate the total number of genomic differences you will need to combine the chunks results by simply using Microsoft Excel.



..		8/17/20
LWK_IDs.txt_CHB_IDs.txt_9_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_8_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_7_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_6_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_5_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_4_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_3_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_23_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_22_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_21_A	274 KiB	10/23/20
LWK_IDs.txt_CHB_IDs.txt_20_A	274 KiB	10/23/20

- **Fifth Program:**

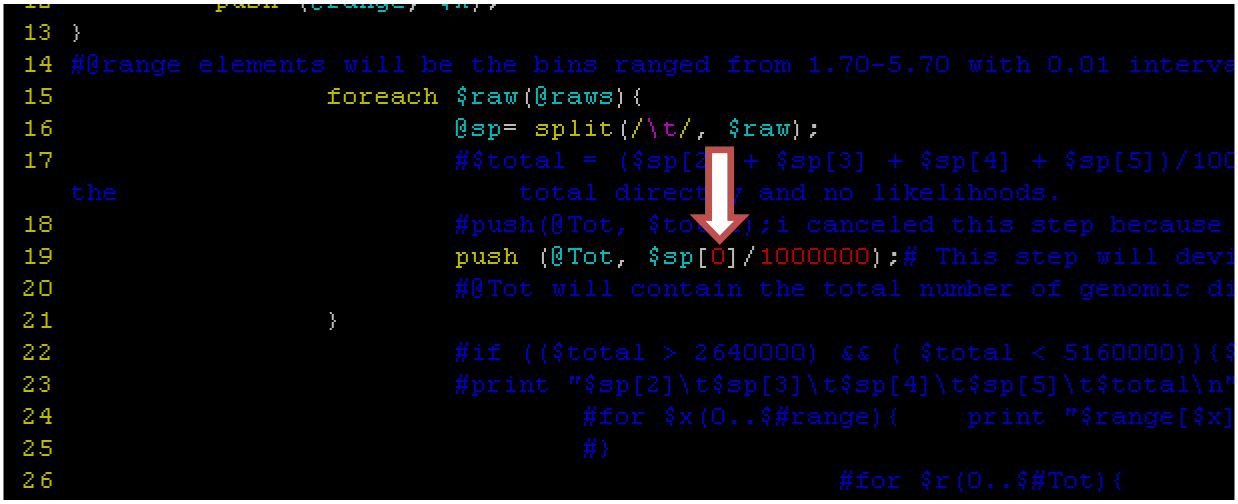
  - ***BinsMaker.pl***

  For easier demonstration of the total number of genomic differences between all pairs of individuals calculated by either *GenomeDiff_population_Chunks_HETonly.pl* or *InterpopulationDiff2individuals_HET.pl*

  We created this *BinsMaker.pl* program to group our results in bins.

  In order to run this program :

  - Change the column # in line 19 to refer to the column # that will have the totals of the differences in your result file.



```
12              push (@range, $x);
13 }
14 #@range elements will be the bins ranged from 1.70-5.70 with 0.01 interva
15                  foreach $raw(@raws){
16                          @sp= split(/\t/, $raw);
17                          #$total = ($sp[2] + $sp[3] + $sp[4] + $sp[5])/100
   the                          total directly and no likelihoods.
18                          #push(@Tot, $total);i canceled this step because
19                          push (@Tot, $sp[0]/1000000);# This step will devi
20                          #@Tot will contain the total number of genomic di
21                  }
22                          #if (($total > 2640000) && ( $total < 5160000)){$
23                          #print "$sp[2]\t$sp[3]\t$sp[4]\t$sp[5]\t$total\n"
24                                  #for $x(0..$#range){    print "$range[$x]
25                                  #}
26                                                  #for $r(0..$#Tot){
```

The command line to run this program is:

```
perl BinsMaker.pl LWK_YRI_InterPop.xls > LWK_YRI_InterPopBins.txt
```

The bins result can be then easily presented using Microsoft Excel chart.

1.8.2 Protocol for Assessing the vrGVs in the 1k Genome Database

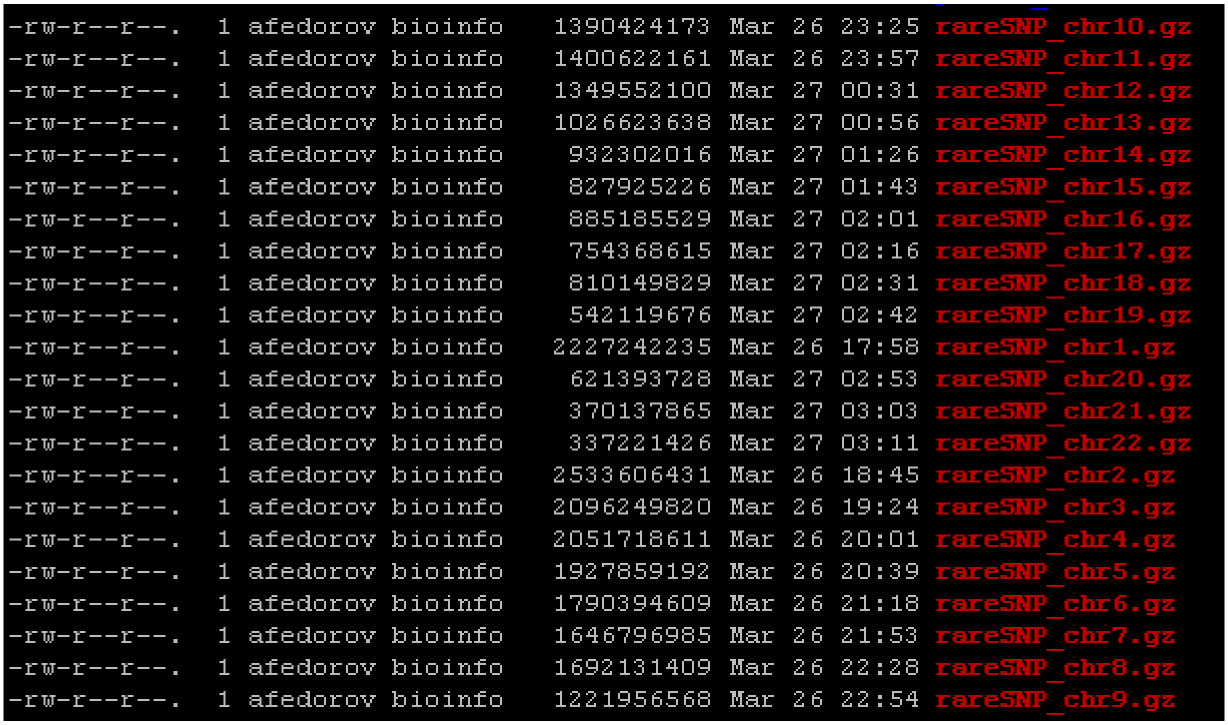
In order to assess vrGVs in the 1092 Individuals, a subset table was created for each chromosome file, this subset file contains only SNPs of 0.2% occurrence. This table was created using a Perl program *vrGVs.pl*.

```
#!/usr/local/perl

$str1 = localtime();
for $x (1..22) {
$names= '/home/asahama/ALL.chr'. $x .
'.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz';
$l_N = 0;
open (OUT, ">rareSNP_chr$x") || die "Can't open file rareSNP : $!\n";
open (IN, "zcat $names |") || die "Can't open zipped file chr21 :
$!\n";
while(<IN>) {
    $l_N++;
    if ($l_N < 31){print OUT $_;}
    if ($l_N > 30){
          $count = () = $_ =~ /\|1/g;
          $count2 = () = $_ =~ /1\|/g;
          if ($count + $count2 > 1 && $count + $count2 <5){
          #     print "$l_N \t $count \t $count2 \n";
                print OUT $_;
          }
    }
}
close(OUT);
}
$str2 = localtime();
print "$str1\t$str2\n";
```

In order to run this program you will need to specify the path for your 1092 individual autosomal genomes data (highlighted).

The subset tables will then be saved in a (.gz) compressed format files:



```
-rw-r--r--.  1 afedorov bioinfo   1390424173 Mar 26 23:25 rareSNP_chr10.gz
-rw-r--r--.  1 afedorov bioinfo   1400622161 Mar 26 23:57 rareSNP_chr11.gz
-rw-r--r--.  1 afedorov bioinfo   1349552100 Mar 27 00:31 rareSNP_chr12.gz
-rw-r--r--.  1 afedorov bioinfo   1026623638 Mar 27 00:56 rareSNP_chr13.gz
-rw-r--r--.  1 afedorov bioinfo    932302016 Mar 27 01:26 rareSNP_chr14.gz
-rw-r--r--.  1 afedorov bioinfo    827925226 Mar 27 01:43 rareSNP_chr15.gz
-rw-r--r--.  1 afedorov bioinfo    885185529 Mar 27 02:01 rareSNP_chr16.gz
-rw-r--r--.  1 afedorov bioinfo    754368615 Mar 27 02:16 rareSNP_chr17.gz
-rw-r--r--.  1 afedorov bioinfo    810149829 Mar 27 02:31 rareSNP_chr18.gz
-rw-r--r--.  1 afedorov bioinfo    542119676 Mar 27 02:42 rareSNP_chr19.gz
-rw-r--r--.  1 afedorov bioinfo   2227242235 Mar 26 17:58 rareSNP_chr1.gz
-rw-r--r--.  1 afedorov bioinfo    621393728 Mar 27 02:53 rareSNP_chr20.gz
-rw-r--r--.  1 afedorov bioinfo    370137865 Mar 27 03:03 rareSNP_chr21.gz
-rw-r--r--.  1 afedorov bioinfo    337221426 Mar 27 03:11 rareSNP_chr22.gz
-rw-r--r--.  1 afedorov bioinfo   2533606431 Mar 26 18:45 rareSNP_chr2.gz
-rw-r--r--.  1 afedorov bioinfo   2096249820 Mar 26 19:24 rareSNP_chr3.gz
-rw-r--r--.  1 afedorov bioinfo   2051718611 Mar 26 20:01 rareSNP_chr4.gz
-rw-r--r--.  1 afedorov bioinfo   1927859192 Mar 26 20:39 rareSNP_chr5.gz
-rw-r--r--.  1 afedorov bioinfo   1790394609 Mar 26 21:18 rareSNP_chr6.gz
-rw-r--r--.  1 afedorov bioinfo   1646796985 Mar 26 21:53 rareSNP_chr7.gz
-rw-r--r--.  1 afedorov bioinfo   1692131409 Mar 26 22:28 rareSNP_chr8.gz
-rw-r--r--.  1 afedorov bioinfo   1221956568 Mar 26 22:54 rareSNP_chr9.gz
```

- **First Program:**
  – *IDs_seperator_rareSNPs.pl:*
  This program is designed to skim genomic data of any one or two populations from the whole 22 vrGVs chromosomes files. The skimmed data then become easier and faster to be processed.

  To run this program you will need to download 3 files "available in the website" into the same directory of the *IDs_seperator_rareSNPs.pl* program.

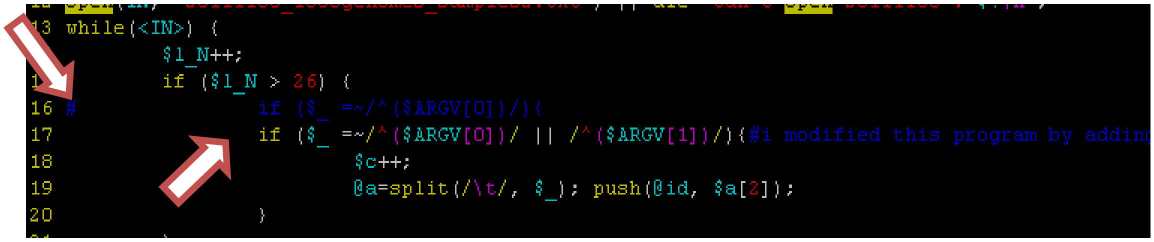
  1. 20111108_1000genomes_samples2.txt
  This .txt file contains details and Identifier for all individuals that were actually sequences through the phase 1 of the 1K Genomes project. this fie should be in the same directory of the *IDs_seperator_rareSNPs .pl* program.

  2. head100_1000genome_testfile
  This file contains the first 100 lines from the original table that has 1092 individual autosomes data.

  3. rareSNPs_Names.txt
  A text file contains the names of the 22 vrGVs chromosomes files.

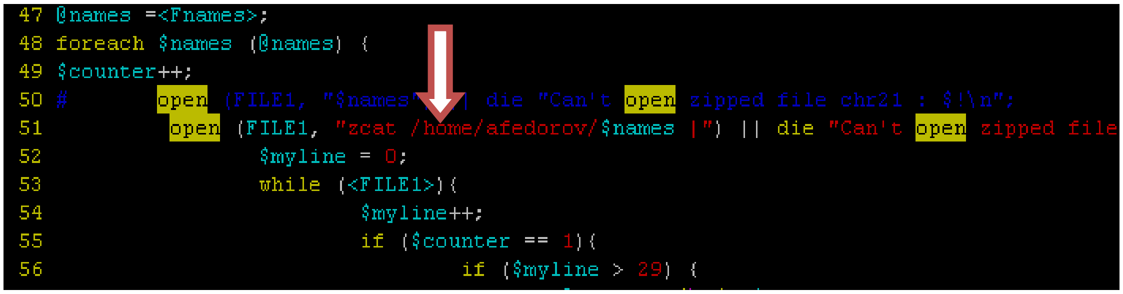
  Use Line 16 and comment line 17 (If u would like to calculate vrGVs between individuals from two different populations)

Use Line 17 and comment line 16 (If u would like to calculate vrGVs between individuals within the same population).



```
13 while(<IN>) {
                $1_N++;
1               if ($1_N > 26) {
16 #                    if ($_ =~/^($ARGV[0])/){
17                      if ($_ =~/^($ARGV[0])/ || /^($ARGV[1])/){#i modified this program by adding
18                              $c++;
19                              @a=split(/\t/, $_); push(@id, $a[2]);
20                      }
```

You will need to provide the path to the 22 chromosomes data in line 45.



```
47 @names =<Fnames>;
48 foreach $names (@names) {
49 $counter++;
50 #        open (FILE1, "$names") || die "Can't open zipped file chr21 : $!\n";
51            open (FILE1, "zcat /home/afedorov/$names |") || die "Can't open zipped file
52                    $myline = 0;
53                    while (<FILE1>){
54                            $myline++;
55                            if ($counter == 1){
56                                    if ($myline > 29) {
```

- ***Command line example: (one population):***
    - ***" one population":***
      nohup perl IDs_seperator_rareSNPs.pl ASW > ASW_RareSNPs_Table.txt &

      The created file will contain vrGVs data for ASW individuals only.

    - ***" Two population" :***
      nohup perl IDs_seperator_rareSNPs.pl LWK ASW > LWK_ASW_RareSNPs_Table.txt &

      The created file will contain vrGVs data for LWK and ASW individuals only.

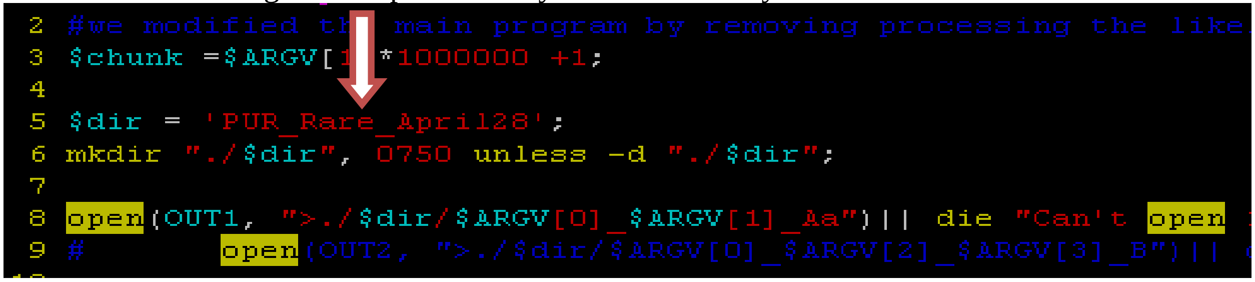
- **Second Program:**

  - ***Intra_PopGenomeDif_vrGVs.pl:***

  This program is designed to count the shared very rare SNPs between each pair of individuals with the same population.

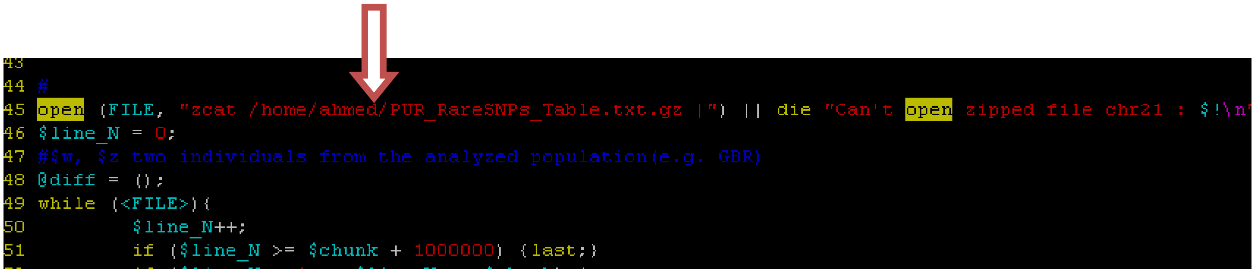
To run this program :

Change the output directory name in line to your desired name 5.

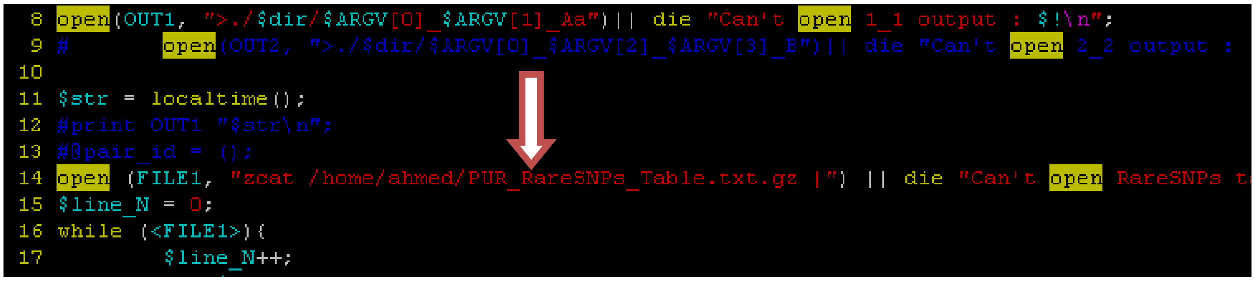


```
2 #we modified the main program by removing processing the like.
3 $chunk =$ARGV[1]*1000000 +1;
4
5 $dir = 'PUR_Rare_April28';
6 mkdir "./$dir", 0750 unless -d "./$dir";
7
8 open(OUT1, ">./$dir/$ARGV[0]_$ARGV[1]_Aa")|| die "Can't open
9 #         open(OUT2, ">./$dir/$ARGV[0]_$ARGV[2]_$ARGV[3]_B")||
```

Line 14 and line 45 should have the right directory path to the vrGVs table files.



```
43
44 #
45 open (FILE, "zcat /home/ahmed/PUR_RareSNPs_Table.txt.gz |") || die "Can't open zipped file chr21 : $!\n"
46 $line_N = 0;
47 #$w, $z two individuals from the analyzed population(e.g. GBR)
48 @diff = ();
49 while (<FILE>){
50         $line_N++;
51         if ($line_N >= $chunk + 1000000) {last;}
```



```
 8 open(OUT1, ">./$dir/$ARGV[0]_$ARGV[1]_Aa")|| die "Can't open 1_1 output : $!\n";
 9 #         open(OUT2, ">./$dir/$ARGV[0]_$ARGV[2]_$ARGV[3]_B")|| die "Can't open 2_2 output :
10
11 $str = localtime();
12 #print OUT1 "$str\n";
13 #@pair_id = ();
14 open (FILE1, "zcat /home/ahmed/PUR_RareSNPs_Table.txt.gz |") || die "Can't open RareSNPs ta
15 $line_N = 0;
16 while (<FILE1>){
17         $line_N++;
```

This program is designed to process the genomic data per chunks of 1,000,000 lines. So you will need to find how many lines in your table (PUR_rareSNPs.txt) in order to decide how many times to run the program to process the entire data in that table.

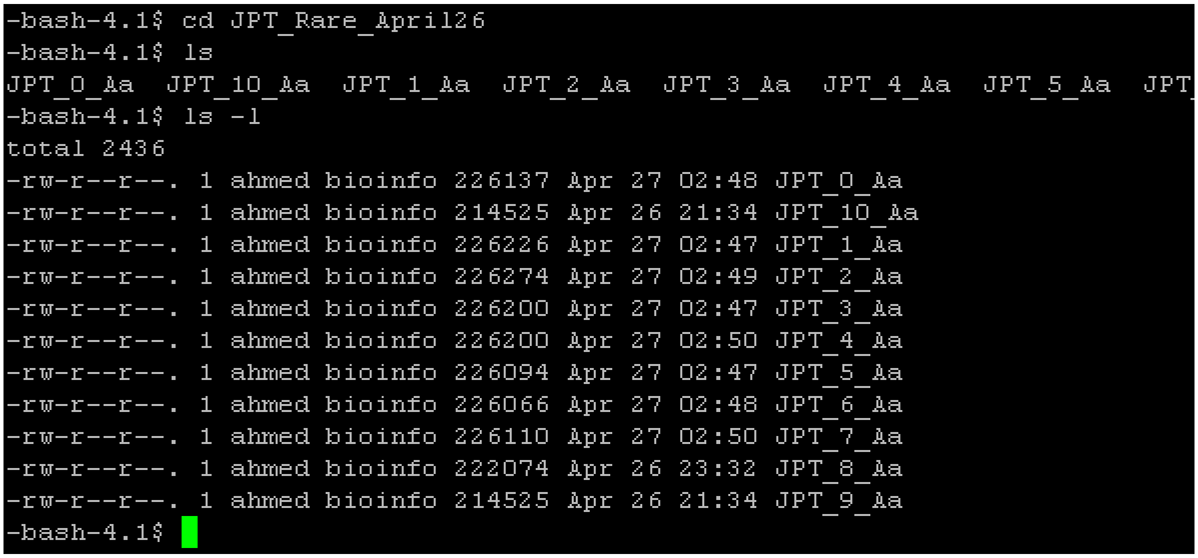
Unix command line (wc -l PUR_rareSNPs.txt) can easily count the number of lines in that file.

For example: If your PUR_rareSNPs.txt is 8,975,443 lines then you will need to run the program 9 times (1 time for each chunk) as below commands:

nohup perl Intra_PopGenomeDif_vrGVs.pl PUR 0 &
(Where the second argument variable is the chunk number)

The results will be saved into the output directory folder specified in line number 5.

Each chunk file will contain the total number of shared very rare SNPs (vrGVs) between each pair of individuals for the number of lines in that specific chunk.



```
-bash-4.1$ cd JPT_Rare_April26
-bash-4.1$ ls
JPT_0_Aa  JPT_10_Aa  JPT_1_Aa  JPT_2_Aa  JPT_3_Aa  JPT_4_Aa  JPT_5_Aa  JPT_
-bash-4.1$ ls -l
total 2436
-rw-r--r--. 1 ahmed bioinfo 226137 Apr 27 02:48 JPT_0_Aa
-rw-r--r--. 1 ahmed bioinfo 214525 Apr 26 21:34 JPT_10_Aa
-rw-r--r--. 1 ahmed bioinfo 226226 Apr 27 02:47 JPT_1_Aa
-rw-r--r--. 1 ahmed bioinfo 226274 Apr 27 02:49 JPT_2_Aa
-rw-r--r--. 1 ahmed bioinfo 226200 Apr 27 02:47 JPT_3_Aa
-rw-r--r--. 1 ahmed bioinfo 226200 Apr 27 02:50 JPT_4_Aa
-rw-r--r--. 1 ahmed bioinfo 226094 Apr 27 02:47 JPT_5_Aa
-rw-r--r--. 1 ahmed bioinfo 226066 Apr 27 02:48 JPT_6_Aa
-rw-r--r--. 1 ahmed bioinfo 226110 Apr 27 02:50 JPT_7_Aa
-rw-r--r--. 1 ahmed bioinfo 222074 Apr 26 23:32 JPT_8_Aa
-rw-r--r--. 1 ahmed bioinfo 214525 Apr 26 21:34 JPT_9_Aa
-bash-4.1$ 
```

A sample of a results file is showing in the table below:

	Individual 1		Individual 2					Rare SNPs shared
0	HG00403	1	HG00404	One_One:	13	Two_Two:	Two_One:	13
0	HG00403	2	HG00406	One_One:	97	Two_Two:	Two_One:	97
0	HG00403	3	HG00407	One_One:	9	Two_Two:	Two_One:	9
0	HG00403	4	HG00418	One_One:	9	Two_Two:	Two_One:	9
0	HG00403	5	HG00419	One_One:	29	Two_Two:	Two_One:	29
0	HG00403	6	HG00421	One_One:	12	Two_Two:	Two_One:	12
0	HG00403	7	HG00422	One_One:	7	Two_Two:	Two_One:	7
0	HG00403	8	HG00427	One_One:	20	Two_Two:	Two_One:	20
0	HG00403	9	HG00428	One_One:	8	Two_Two:	Two_One:	8
0	HG00403	10	HG00436	One_One:	4	Two_Two:	Two_One:	4
0	HG00403	11	HG00437	One_One:	5	Two_Two:	Two_One:	5
0	HG00403	12	HG00442	One_One:	11	Two_Two:	Two_One:	11
0	HG00403	13	HG00443	One_One:	15	Two_Two:	Two_One:	15
0	HG00403	14	HG00445	One_One:	3	Two_Two:	Two_One:	3
0	HG00403	15	HG00446	One_One:	4	Two_Two:	Two_One:	4
0	HG00403	16	HG00448	One_One:	8	Two_Two:	Two_One:	8

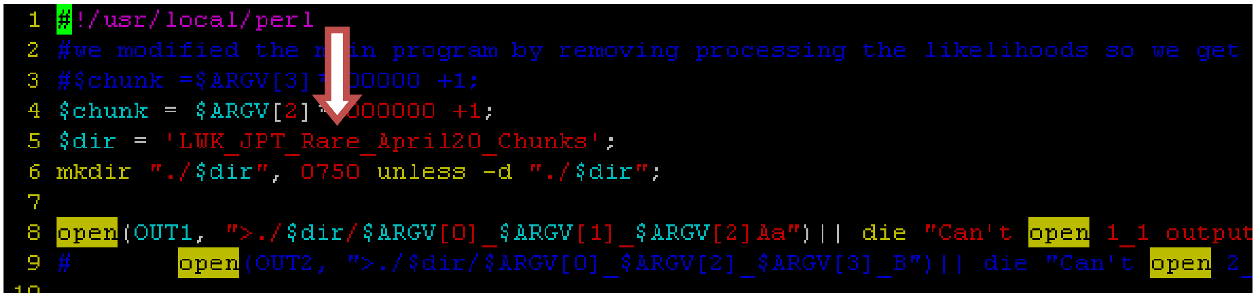
By simply using Microsoft Excel, Summing the number of rare SNPs shared for each specific pair of individuals from each chunk file to get the total number of shared rare SNPs.

- **Third Program:**
  – ***Inter_PopGenomeDif_vrGVs.pl***
  This program is created to calculate the number of rare genomic variants shared between each pair of individuals from two different populations.

  To run this program you should have (**20111108_1000genomes_samples2.txt**) in the same directory of the Inter_PopGenomeDif_vrGVs.pl program.

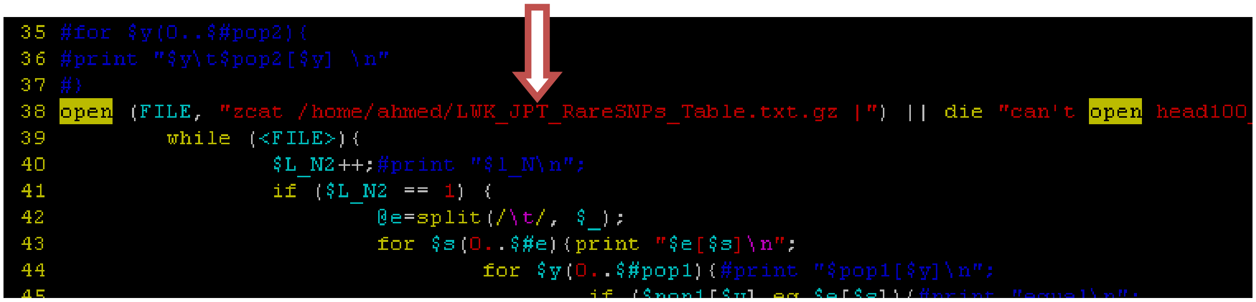
  The **20111108_1000genomes_samples2.txt** file contains details and Identifier for all individuals that were actually sequences through the phase 1 of the 1K Genomes project.

  Change the output directory name in line to your desired name 5.

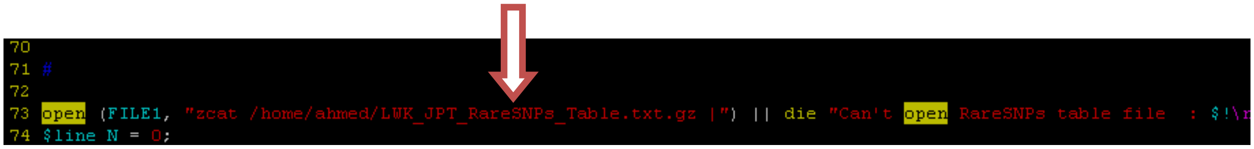


```
 1 #!/usr/local/perl
 2 #we modified the main program by removing processing the likelihoods so we get
 3 #$chunk =$ARGV[3]*100000 +1;
 4 $chunk = $ARGV[2] *1000000 +1;
 5 $dir = 'LWK_JPT_Rare_April20_Chunks';
 6 mkdir "./$dir", 0750 unless -d "./$dir";
 7
 8 open(OUT1, ">./$dir/$ARGV[0]_$ARGV[1]_$ARGV[2]Aa")|| die "Can't open 1_1 output
 9 #        open(OUT2, ">./$dir/$ARGV[0]_$ARGV[2]_$ARGV[3]_B")|| die "Can't open 2_
```

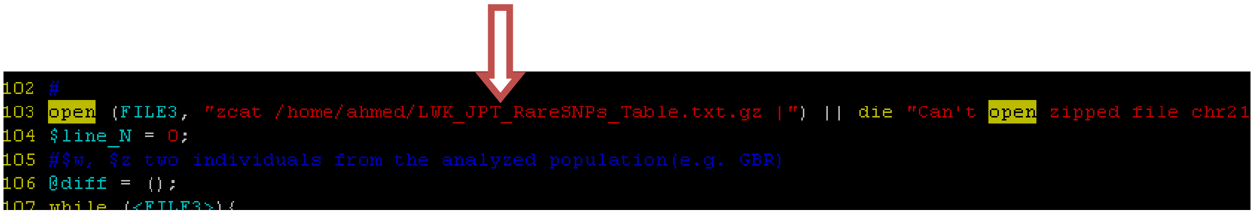
Lines 38, 73 and 103 should have the right directory to where the vrGVs tables are located at.



```
35 #for $y(0..$#pop2){
36 #print "$y\t$pop2[$y] \n"
37 #}
38 open (FILE, "zcat /home/ahmed/LWK_JPT_RareSNPs_Table.txt.gz |") || die "can't open head100_
39          while (<FILE>){
40                  $L_N2++;#print "$1_N\n";
41                  if ($L_N2 == 1) {
42                          @e=split(/\t/, $_);
43                          for $s(0..$#e){print "$e[$s]\n";
44                                  for $y(0..$#pop1){#print "$pop1[$y]\n";
45                                          if ($pop1[$y] eq $e[$s]){#print "equal\n";
```



```
70
71 #
72
73 open (FILE1, "zcat /home/ahmed/LWK_JPT_RareSNPs_Table.txt.gz |") || die "Can't open RareSNPs table file  : $!\n
74 $line N = 0;
```



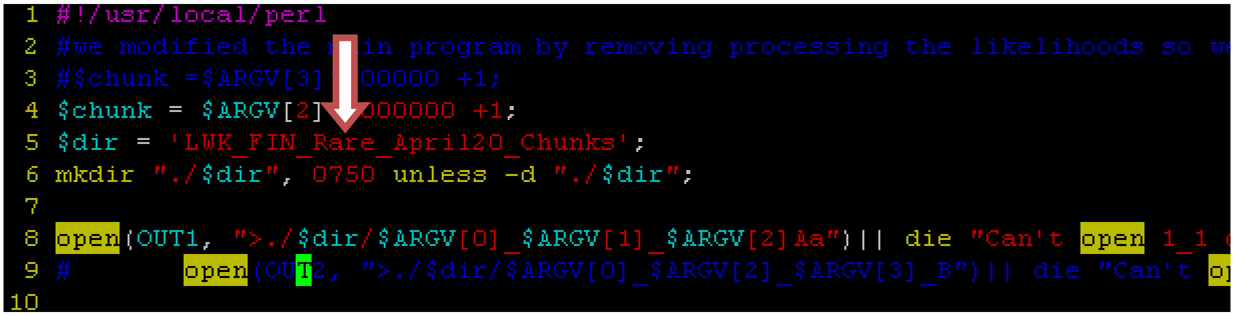
```
102 #
103 open (FILE3, "zcat /home/ahmed/LWK_JPT_RareSNPs_Table.txt.gz |") || die "Can't open zipped file chr21
104 $line_N = 0;
105 #$w, $z two individuals from the analyzed population(e.g. GBR)
106 @diff = ();
107 while (<FILE3>){
```

This program is designed to process the genomic data per chunks of 1,000,000 lines. So you will need to find how many lines in your table (LWK_FIN_RareSNPs_Table.txt.gz) in order to know how many times to run the program to process the entire data in that table.

For example: If your LWK_FIN_RareSNPs_Table.txt.gz is 8,975,443 lines then you will need to run the program 9 times (1 time for each chunk) as below commands.

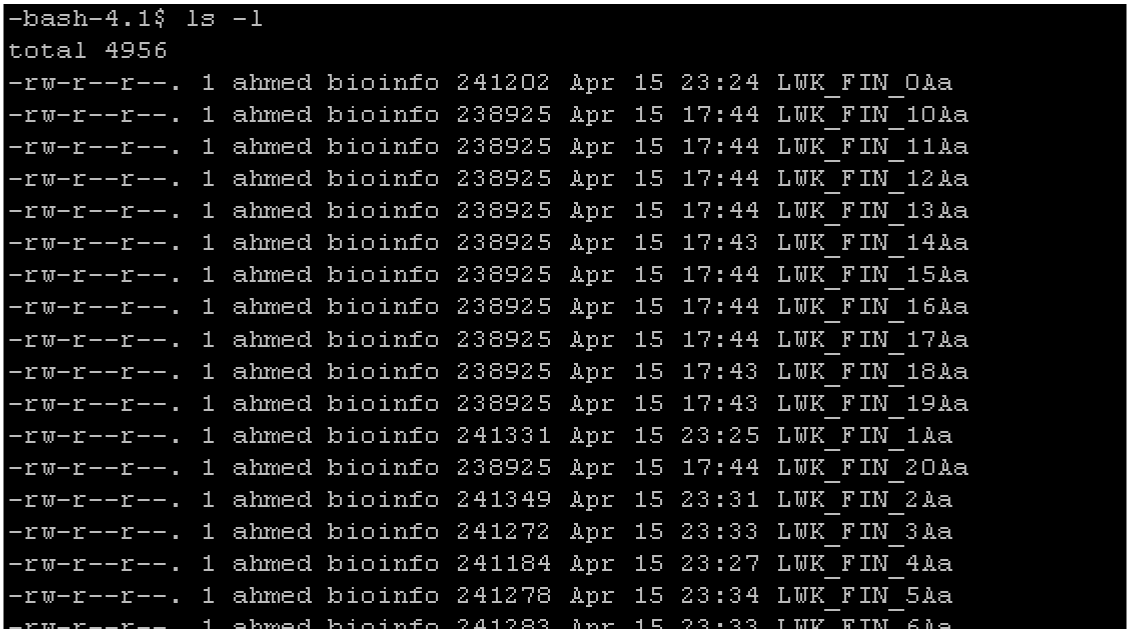
```
827  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 0 &
828  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 1 &
829  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 2 &
830  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 3 &
831  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 4 &
832  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 5 &
833  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 6 &
834  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 7 &
835  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 8 &
836  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 9 &
837  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 10 &
838  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 11 &
839  nohup  perl Intra_PopGenomeDif_vrGVs.pl LWK FIN 12 &
```

The results will be saved into a folder (LWK_FIN_Rare_April20_Chunks) in the same directory of the program. To change the folder name in the program you will need to change the name in line 5



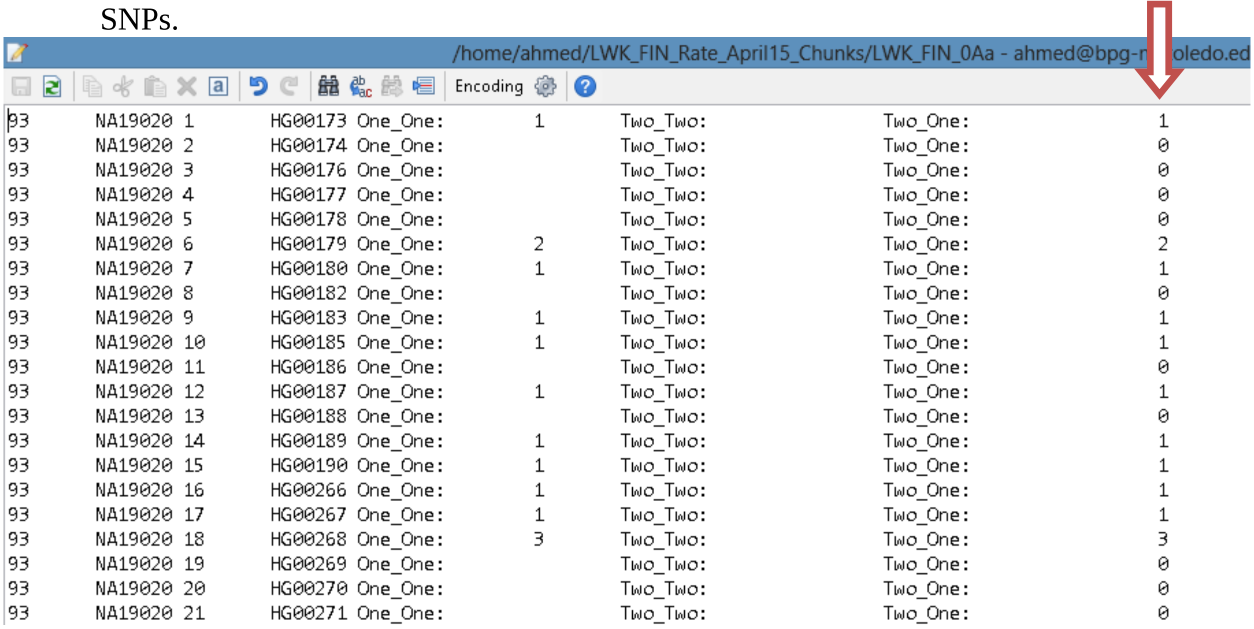
```
 1 #!/usr/local/perl
 2 #we modified the main program by removing processing the likelihoods so w
 3 #$chunk =$ARGV[3]*1000000 +1;
 4 $chunk = $ARGV[2]*1000000 +1;
 5 $dir = 'LWK_FIN_Rare_April20_Chunks';
 6 mkdir "./$dir", 0750 unless -d "./$dir";
 7
 8 open(OUT1, ">./$dir/$ARGV[0]_$ARGV[1]_$ARGV[2]Aa")|| die "Can't open 1_1
 9 #       open(OUT2, ">./$dir/$ARGV[0]_$ARGV[2]_$ARGV[3]_B")|| die "Can't op
10
```

Each chunk file will contain the total number of shared very rare SNPs(vrGVs) between each pair of individuals for the number of lines in that specific chunk.



```
-bash-4.1$ ls -l
total 4956
-rw-r--r--. 1 ahmed bioinfo 241202 Apr 15 23:24 LWK_FIN_0Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_10Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_11Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_12Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_13Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:43 LWK_FIN_14Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_15Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_16Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_17Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:43 LWK_FIN_18Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:43 LWK_FIN_19Aa
-rw-r--r--. 1 ahmed bioinfo 241331 Apr 15 23:25 LWK_FIN_1Aa
-rw-r--r--. 1 ahmed bioinfo 238925 Apr 15 17:44 LWK_FIN_20Aa
-rw-r--r--. 1 ahmed bioinfo 241349 Apr 15 23:31 LWK_FIN_2Aa
-rw-r--r--. 1 ahmed bioinfo 241272 Apr 15 23:33 LWK_FIN_3Aa
-rw-r--r--. 1 ahmed bioinfo 241184 Apr 15 23:27 LWK_FIN_4Aa
-rw-r--r--. 1 ahmed bioinfo 241278 Apr 15 23:34 LWK_FIN_5Aa
-rw-r--r--. 1 ahmed bioinfo 241283 Apr 15 23:33 LWK_FIN_6Aa
```

By simply using Microsoft Excel, Summing the number of rare SNPs shared for each specific pair of individuals from each chunk file to get the total number of shared rare SNPs.


/home/ahmed/LWK_FIN_Rate_April15_Chunks/LWK_FIN_0Aa - ahmed@bpg-r oledo.ed
Encoding

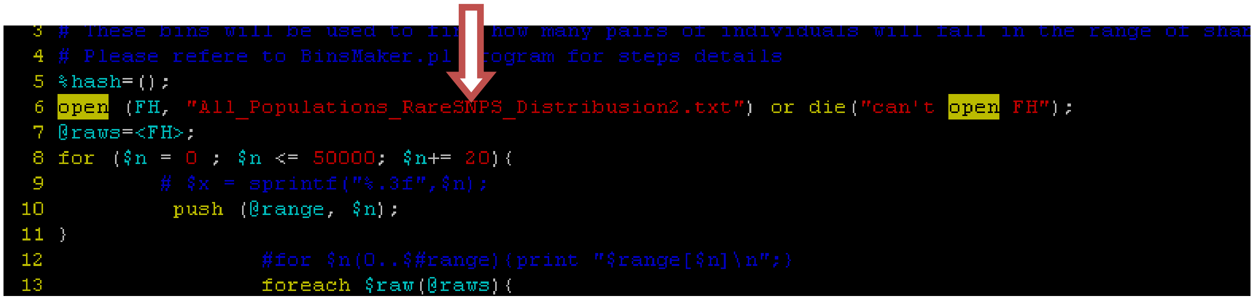
```
93      NA19020 1       HG00173 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 2       HG00174 One_One:                Two_Two:                Two_One:                0
93      NA19020 3       HG00176 One_One:                Two_Two:                Two_One:                0
93      NA19020 4       HG00177 One_One:                Two_Two:                Two_One:                0
93      NA19020 5       HG00178 One_One:                Two_Two:                Two_One:                0
93      NA19020 6       HG00179 One_One:        2       Two_Two:                Two_One:                2
93      NA19020 7       HG00180 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 8       HG00182 One_One:                Two_Two:                Two_One:                0
93      NA19020 9       HG00183 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 10      HG00185 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 11      HG00186 One_One:                Two_Two:                Two_One:                0
93      NA19020 12      HG00187 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 13      HG00188 One_One:                Two_Two:                Two_One:                0
93      NA19020 14      HG00189 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 15      HG00190 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 16      HG00266 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 17      HG00267 One_One:        1       Two_Two:                Two_One:                1
93      NA19020 18      HG00268 One_One:        3       Two_Two:                Two_One:                3
93      NA19020 19      HG00269 One_One:                Two_Two:                Two_One:                0
93      NA19020 20      HG00270 One_One:                Two_Two:                Two_One:                0
93      NA19020 21      HG00271 One_One:                Two_Two:                Two_One:                0
```

- **Fourth Program:**
  - ***Bins_shared_RareSNPs.pl***

For easier demonstration of the total number of shared very rare SNPs between all pairs of individuals calculated by either *Intra_PopGenomeDif_vrGVs.pl* or Inter_PopGenomeDif_vrGVs.pl

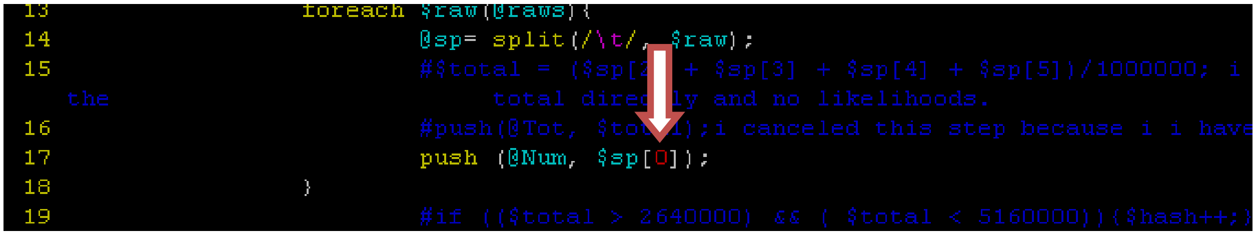
We created this *Bins_shared_RareSNPs.pl* program to group our results in bins.

In order to run this program:
Provide the result file directory in line number 6:



```
 3 # These bins will be used to find how many pairs of individuals will fall in the range of shar
 4 # Please refere to BinsMaker.pl program for steps details
 5 %hash=();
 6 open (FH, "All_Populations_RareSNPS_Distribusion2.txt") or die("can't open FH");
 7 @raws=<FH>;
 8 for ($n = 0 ; $n <= 50000; $n+= 20){
 9         # $x = sprintf("%.3f",$n);
10          push (@range, $n);
11 }
12                 #for $n(0..$#range){print "$range[$n]\n";}
13                 foreach $raw(@raws){
```

Change the column # in line 17 to refer to the column # that will have the totals of the Share Very Rare SNPs (vrGVs) in your result file.



```
13                  foreach $raw(@raws){
14                          @sp= split(/\t/, $raw);
15                          #$total = ($sp[2] + $sp[3] + $sp[4] + $sp[5])/1000000; i
   the                           total directly and no likelihoods.
16                          #push(@Tot, $total);i canceled this step because i i have
17                          push (@Num, $sp[0]);
18                  }
19                          #if (($total > 2640000) && ( $total < 5160000)){$hash++;}
```

The command line to run this program is:

```
perl Bins_shared_RareSNPs.pl > Pairwise_vrGVs_GBR_rawBins.xls
```

The bins result can be then easily presented using Microsoft Excel chart.