Snp HeRitability Estimation Kit (SHREK) Development Log

Sam Shing Wan, Choi
choishingwan@gmail.com

Johnny Sheung Him, Kwan
shkwan@hku.hk

Henry Chi-Ming, Leung
cmleung2@cs.hku.hk

February 25, 2016

Contents

1 Assumption in implementation

To make life easier, I made a number of assumption when I implement the programme and they are listed as follow

1. The intervals within each individual bed files are non-overlapping
2. SNPs filtered out when reading the p-value file will not be included in the output no matter what (avoid storing unnecessary information)
3. We will remove *all* SNPs that are in perfect LD with each-other
   (a) Realistically, we can not remove *all* SNPs that are in perfect LD with each-other without first computing the LD matrix for the whole genome. However, that will incur a large speed penalty to the algorithm. To compromise, we remove SNPs that are in perfect LD with each-other *within the same window.* Take into consideration of the calculation of variance, we will need to plan forward and consider SNPs within 4 windows. That's definitely overkill but that avoid problem of crazy perfect LD interactions (to be honest, I would consider those as error, yet it is easier for me to just remove them)
4. We do not consider a SNP ambiguous unless we need to perform flipping
5. Only when the direction of effect is presented will we perform the detail variance estimation calculation. The equation for the variance calculation is: $\boldsymbol{R_{sq}^{-1}}\boldsymbol{n}(4\boldsymbol{R}\circ(\boldsymbol{z}\boldsymbol{z}^t)-2\boldsymbol{R_{sq}^{-1}})\boldsymbol{n}\boldsymbol{R_{sq}^{-1}}$ where $\boldsymbol{n}$ is a square diagonal matrix with diagonal as the sample size and $\boldsymbol{z}$ is the vector of z-statistics.

2 To Do

1. Compare the condition number of the LD matrix before and after the LD correction
2. Compare the performance of case control estimation with different sample size

3. For LD correction, might also worth testing how the density and total number of SNPs affect the estimation

4. For effect size estimation, don't use the complicated method

5. Compare the speed of the new and old implementation (remember that the complex variance calculation do takes more time to perform)

3 January 25, 2016

We are going to re-write SHREK starting from today. There are a number of main goal in this re-work

1. The whole parameter parsing

2. Better help messages

3. Allow filtering by imputation score (0-1)

4. Filtering of SNPs that have different reference/alternative allele with reference

5. Use Armadillo instead of EIGEN to improve speed

6. Filter *all* SNPs that are in perfect LD

7. Implement the advance variance calculation

8. Start documenting the codes

3.1 Finished Today

1. Finished the error message for different modes

2. Finished the parsing for binary trait

3. Finished the parsing for quantitative trait

4. Disabled risk prediction, should focus on the heritability estimation first

I have not perform debugging, there can be error

4 January 26, 2016

Continue on where I left yesterday Assumptions we made in the programme

1. p-value file has header
2. the numbers return from Command class are index (0 based)

I really want to write a more compacted code. Will try to restructure the command class

4.1 Finished Today

1. Finished refining the usage messages
2. Finished the basic parameter parsing (Still haven't finish the index based parameter though)

5 January 27, 2016

5.1 Aim

We want to at least finish the region parsing and SNP class update today

5.2 Finished Today

1. Completed the update of the Command class (compiled without problem)
2. Completed the update of the Region class (compiled without problem)
3. Completed the update of the Snp class (compiled without problem)

I also *haven't* test run the programme, so there might be additional syntax / logic error that are not picked up at the current stage.

6 January 28, 2016

6.1 Finished Today

1. Finished the checking of reference panel
2. Found a different algorithm for popcount, which should be platform independent
3. Installed armadillo and OpenBLAS
    (a) I have to add the OpenBLAS to the PATH and LD_LIBRARY_PATH for armadillo to detect it
4. Updated the makefile accordingly
    (a) However, I have not updated the decomposition class, therefore there will be problem when we try to compile the programme
5. Updated the r^2 calculation function such that it is using the correct sample size
    (a) The sample size when calculating the r^2 is defined as the number of samples containing *both* SNPs
    (b) Because of how complicated the genotype class is, I really don't want to modify it
    (c) The main bottleneck for the computation is in the decomposition anyway, so it should be fine

7 January 29, 2016

Need to at least finish the getSNP function. Must be careful with it as this is not only complicated, but also extremely important for the whole programme.

I spent whole day in annotating the document of the class to make sure I don't overlook any problem.

8 February 1, 2016

Start to implement the getSNP function. A number of point to note

1. First get all the SNPs that passed all filtering (e.g. MAF)
2. Then for each window, calculate the LD
3. When encountered with perfect LD, we retain the *first* SNP
   (a) The summary statistics will be the mean of all SNPs in perfect LD
   (b) All SNPs except the first are *removed*
   (c) So they will not appear in subsequent analysis

9 February 2, 2016

Try to perform extensive check on codes that has been written

9.1 Finished Today

1. Finished debugging, everything up until the initialization of genotype-FileHandler is ok

10 February 3, 2016

Sat in a meeting with Pak today, now we have a few things to-do.

10.1 Finished Today

We have switched to use list instead of deque. The benefit of list is that it is a double linked list, therefore it is much easier for one to remove the middle part. Such property is ideal for the handling of perfect LD. However, it also means that we cannot access the elements of the list by index. The only way for us to access the elements is to use the iterators and the advance function from the iterator library. Although this makes the implementation more difficult, it should be able to help us to speed up our code even if we want to perform perfect LD removal.

1. Extensive testing of the previous classes
2. Now use iterators for the boundary, this is more dynamic and can work with the perfect LD removal without inducing much complication into the algorithm (however, the implementation itself can be complicated)

10.2 Problem identified today

So assuming that the SNPs A B C are occurring in the genome by alphabetical order, when we calculate the LD of A with B C, we do not observe any perfect LD. This will lead us to continue with the calculation of SNP B and C. However, if SNP B and SNP C is in perfect LD, we will need to go back to the LDs of SNP A to remove the column of SNP C. Not only is this procedure complicates the implementation, it will also significantly slow down the computation because of the constant resizing of the matrix.

To conclude, it might be better for us to use the original “sausage” approach, where we should also locate all the perfect LD location. Then we can remove the perfect LD afterwards.

11 February 4, 2016

Was spending the whole day in implementing the “sausage” approach. However, it was found that the calculated R^2 are terribly wrong. So now converting back to the old method of inclusion for the genotype reading. We will then check the implementation of the R^2 computation right after we have make sure the genotype reading is correct.

12 February 5, 2016

Try to continue on what we have been doing yesterday.

12.1 Finished Today

1. Finished debugging the genotype reading. By using the old inclusion method, we avoid a lot of complication and the programme now works accordingly. We also checked the bit arithmetic and decided to use

our original calculation which unlike what I previously thought, isn't platform specific.

12.2 Special Note

The NumberOfBit function in the usefulTools doesn't work in our case because our number is longer than uint32_t. I am rather uncertain how should we change the script accordingly because it involves extreme high level bit arithmetic. However, based on online sources, it seems like the __builtin_popcountll function will work as long as we compile the programme using gcc. So we will stick to that function for now.

13 February 11, 2016

Back from holiday, now taking the perfect LD seriously and start implementing the perfect LD removal algorithm. It is important to note that it is possible for the boundaries to be in perfect LD, therefore potentially changing the boundary. However, take into account of our current procedure, most of the time, only the last block need our attention, unless this is the beginning of the chromosome or decomposition region.

13.1 Finished Today

1. Implemented the slow but simple version of the perfect LD removal, need to test run it.
2. Solved some bugs, e.g. the genotype collection and LD construction bug (use the wrong index)
3. Found that there are still problems with the perfect LD removal. Specifically, when the SNP at the bound is in perfect LD with previous SNPs, the handling are still problematic. Need to change my test case in which I reduce the gap between the two SNPs (e.g. the $4->7$)

14 February 12, 2016

Finishing the first part of debugging. Basically, the LD construction seems to be ok for now. Also, I reused my old code for getting the mean of the test-statistics. Now, we store only the z-score in the Snp object. This reduces the number of information we store in Snp object but slightly increases the number of computation we have to perform.

14.1 Finished Today

1. Finished all the test run for linkage construction (have not test the cross chromosome yet)
2. Implemented the decomposition and variance calculation. (Using Tim's Equation) Will need to perform the simulation to see if the variance estimation works.
3. Updated the Snp class to accommodate the perfect LD computation.

14.2 Note

Unfortunately, I am not used to template usage in c++, therefore I will have to write two separate function for the decomposition of the matrix depending on whether if the sign is provided.

An important finding regarding the linking of the library. You must have the armadillo/usr/lib folder under your LD_LIBRARY_PATH. Also, you will need -larmadillo when you compile and it should come *after* the cpp file.

When performing the test code, it was found that the pinv function of armadillo is actually 100 times faster than my self-implemented pseudo inverse. Therefore, we should use their implementation instead.

What we have to do next is to store the results of the estimated variance and to generate the required output.

15 February 15, 2016

Start working on the variance capturing for the complex variance calculation. The most difficult part is to determine the denominator of the calculation. The best way might be to use median, which is less sensitive to outliers. However, the problem is, how to calculate the median without storing the whole matrix, which is memory intensive?

According to Tim, we only need the *rows* of the target SNPs, therefore we only need to capture each *rectangle* per run. Currently trying to list out all possible scenarios to avoid bugs.

16 February 16, 2016

Continue on what I was doing yesterday.

16.1 Finished Today

1. Finished implementing the whole programme, now need to start the debugging hell

17 February 17, 2016

DEPENDENCY HELL!!! AND BUGS, A LOT OF BUGS!!!

17.1 Today's work

1. Trying to debug the programme
2. Found that gdb doesn't work (too old)
3. Try to compile the gdb
4. Failed, so tried to compile latest gcc and see if that works
5. Need to make sure we recompile everything with the new gcc (otherwise, you will have tones of undefined referencing)
6. The variance is most likely to be wrong

18 February 18, 2016

Early simulation results suggested that the MSE of the new programme doubles that of the old one which is really bad. So now we need to check what's the problem of that.

Using the last simulated data, it was found that by not performing the LD correction, we can get a similar result as the new implementation. Therefore it is likely that the LD correction is accounting for the problem.

18.1 Note

Given the following plink simulation input, we would like to have an empirical variance of around 0.002729694

```
 10 snp 0.05 0.05 0.05 0.05 1 0.001 0
100 snp2 0.05 0.05 0.05 0.05 0.7 0.001 0
 890 null 0.05 0.05 0.05 0.05 0.8 0 0
```

18.2 Finished Today

1. The variance calculation has been fixed
2. It seems like the programme can produce sensible results for now

19 February 19,2016

The use of List is very smart in a way that it preserves the iterator, however, using the profiling, it was found that the List operations are taking up most time. Therefore, I will now have to change all the List usage back to deque, hopefully make things a bit faster

20 February 22, 2016

We have finished some basic debugging in the past few days. The heritability estimation now work as planned. However, the variance is still rather problematic. The original equation:

$$\mathrm{Var}(h^2) = \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \frac{4\boldsymbol{R} \circ \boldsymbol{z}\boldsymbol{z}^t - 2\boldsymbol{R}_{\boldsymbol{sq}}}{n^2} \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \tag{1}$$

Seems to return negative results. We need to keep working on it.

$$\begin{aligned}\mathrm{Var}(H) &= \boldsymbol{1^t R_{sq}^{-1}} \frac{4\boldsymbol{R} \circ \boldsymbol{\mu\mu^t} + 2\boldsymbol{R_{sq}}}{n^2} \boldsymbol{R_{sq}^{-1} 1} \\ &= \boldsymbol{1^t R_{sq}^{-1}} \frac{4\boldsymbol{R} \circ (\boldsymbol{zz^t} - \boldsymbol{R}) + 2\boldsymbol{R_{sq}}}{n^2} \boldsymbol{R_{sq}^{-1} 1} \\ &= \boldsymbol{R_{sq}}^{-1} \frac{4\boldsymbol{R} \circ \boldsymbol{zz}^t - 2\boldsymbol{R_{sq}}}{n^2} \boldsymbol{R_{sq}}^{-1} \end{aligned} \tag{2}$$

21 February 23, 2016

The main problem with our simulation is the speed required. From what we saw, the main bottle neck for the simulation is hapgen. So we might have to figure out a way to make it faster.

22 February 24, 2016

We are trying to tackle the problem of variance estimation. Seems like only when we use the sample genotypes can we correctly estimate the empirical variance.

23 February 25, 2016

Still trying to figure out the problem of the variance estimation. One thing we found is that our current simulation method actually violates the assumption of fixed X in Tim's formular of phenotype simulation. Therefore, we now try to use a large reference panel (e.g. 10000+) samples to calculate the $\mathrm{Var}(X\beta)$ and use this for the calculation of phenotype.

24 Useful Resources

Description	URL
Timing the functions in c++ programme	link