SEEKIN: Sequence-based Estimation of Kinship and Inbreeding

version 1.00

Jinzhuang Dou, Chaolong Wang

Computational and Systems Biology

Genome Institute of Singapore

A*STAR, Singapore 138672, Singapore

September 14, 2016

The *SEEKIN* software is available at

https://github.com/jinzhuangdou/SEEKIN

1. Overview

1.1 What is *SEEKIN*

SEEKIN is a software program for estimating kinship and inbreeding coefficients for samples which are sequenced at low sequencing coverage (typically lower than 1x). The key features of this program include:

1) Account for the genotype uncertainties by leveraging the haplotype information from study or external samples. Considering the genotype calls maybe poor under low sequencing coverage, we first summary the sequencing reads for each sample into genotype likelihood (*GL*), and then use imputation/phasing software[1] to phase the genotypes with or without external reference panel. The kinship coefficients are then estimated based on the imputed genotypes.

2) Hand the genetic data from samples with population structure and admixture. To achieve this, we first apply PCA on genotyping data of a set of reference individuals to construct an ancestry map and then use *LASER* [2] to estimate the coordinates of study samples in reference ancestry space. The estimated PCs are used to predict individual-specific allele frequencies using linear regression model.

3) Analyze thousands of individuals in saving memory usage and computational time. We adopted a "single producer/consumer" design which have one "producer" queue to scan large VCF file and push every *L* SNPs into a data block, meanwhile, one "consumer" performs the computation for data block one by one in the multi-threading mode. Roughly, the maximal memory usage is O(*NL*), in which *N* is the sample size. This feature enables *SEEKIN* computationally efficient for large scale genetic datasets.

1.2 How to cite *SEEKIN*

The *SEEKIN* algorithm is described in ref [xxx]:

2 Download and installation

2.1 Installation

You can download the *SEEKIN* software at: https://github.com/jinzhuangdou/SEEKIN.The download package contains a standalone (i.e., statically linked) 64-bit Linux executable *seekin* (in the bin/), which has already been tested on Linux server. This static executable is recommended because it is well-optimized and no further installation is required. If you want to compile your own version of *SEEKIN* from the source code (in the src/), you need to ensure that the C/C++ compiler such as GNU gcc (Support for C++11 containers is required), as well as OpenBLAS and the Armadillo Linear Algebra Libraries installed in your computer. You will need to change the library paths in the Makefile accordingly. A simple Makefile is provided along in the source code (in the src/).

2.2 Help

If *SEEKIN* has been successfully installed into your directory, you can type the following command to get a list of help option.

./seekin –h

SEEKIN provides three modules 1) *modelAF* for calculating the PC-related regression coefficients of reference samples 2) *getAF* for estimating the individual allele frequencies of study samples; 3) *kinship* for estimating kinship coefficients for samples from either homogenous or admixture population. To get the detailed list of option for one module (for example *kinship*), you can type:

./seekin kinship –h

3 Running *SEEKIN*

Here we provide example usages of *SEEKIN* program based on the data provided in the folder named "example". In this folder, we have the Study.chr22.vcf.gz file that includes genotypes of chromosome 22 for 10 studied samples. The genotypes are called from shallow sequencing reads (~0.75x) and then phased using BeagleV4.0 with 1KG3 as the external reference panel. Another VCF format file is the SGVP_268.chr22.vcf.gz file that includes the genotypes at 20,773 SNP loci on chromosome 22 for 268 reference samples from the Singapore Genome Variation Project (SGVP) [3]. Except for the genotype files, we also have two PCA coordinate files in a plain text format: 1) SGVP_268.chr22.RefPC.coord file which contains PCA coordinates for the top 2 PCs of the reference individuals; 2) Study.chr22.ProPC.coord file which contains the top 2 PCs calculated by projecting the study samples on the reference panel using *LASER*[2]. Please refer to *LASER*/*TRACE* software for more information about the generation of these PCA coordinate files. These two PCA coordinate files plus the SGVP_268.chr22.vcf.gz are required for *SEEKIN* to estimate the individual allele frequencies if there is population structure/admixture for samples analyzed.

Kinship estimation using homogenous samples

The command line for running *SEEKIN* for homogenous samples is very simple. The following is one example:

```
./seekin kinship -i ./Study.chr22.vcf.gz  -r 0.3  -m 0.05   -d DS  -p homo  -n 2000  -t 3 \
                -w  1 -o Study.chr22.homo
```

It will generate result files with prefixes Study.chr22.homo specified by –o flag in the current directory. The detailed meanings of flags of "kinship" module are summarized below.

- -i Specifies the name of the SNP genotype input file of studied samples. SEEKIN only reads compressed (gzipped) VCF files. [no default value]
- -a Specifies the name of the individual allele frequency file of studied samples. SEEKIN only reads compressed (gzipped) VCF files. Note that this option cannot be used for homogenous estimation. [no default value]
- -r Remove sites with Rsq less than the "-r" value. [default 0.3]
- -m Remove sites with MAF less than the "-m" value. [default 0.05]
- -d Specifies the kinship estimation based on observed or imputed genotypes. It is "GT" if using observed genotypes and "DS" using imputed genotypes. If no DS information is available for a marker in the VCF file, the GT filed will be used even though the option –d is set to DS. If both GT and DS information available, we recommend using DS mode, because our model could account for genotype uncertainty effectively.

[default DS]

-p Specifies the population mode when estimating kinship. It is "homo" for homogenous estimation and "admix" for admixture estimation. Note that the option –a must be available when option –p is set to admix. [default homo]

-n Specifies the number of markers to include in each block for kinship calculation at one time. This option must be no more than the total number of markers in the input VCF file. [default 10,000]

-t Specifies the number of threads of execution. [default 1]

-w Specifies the weight scheme when combing genome-wide markers.

1: $w_m = 2\sqrt{\hat{p}_{im}(1-p_{im})\hat{p}_{jm}(1-\hat{p}_{jm})}\left(r_m^2\right)^2$; 2: $w_m = \left(r_m^2\right)^2$. [default 1]

-o Specifies the output file name prefix. The prefix may be an absolute or relative filename, but it cannot be a directory name.

Kinship estimation of samples with admixture

In this case, the individual allele frequency file is required. In this example, we use the SGVP as reference panel to model the PC-related linear regression coefficients. Users can use other datasets as reference panel (e.g., 1KG3[4], HapMap[5]). A typical command generating the PC-related regression coefficients using *modelAF* module implemented in *SEEKIN* is:

./seekin modelAF –i SGVP_268.chr22.vcf.gz –c SGVP_268.chr22.RefPC.coord -k 2 –o SGVP_268.chr22.beta

For different studies, there is no need for redundant calculation of the PC-related coefficients [6] if using the same set of reference individuals. Detailed meanings of flags of *modelAF* module are summarized below.

-i Specifies the name of the SNP genotype input file of reference samples. SEEKIN only reads compressed (gzipped) VCF files. [no default value]

-c Specifies the name of PCA coordinate file of reference samples. [no default value]

-k Specifies the number of PCs to compute. This number should be no more than the number of PCs in the input PCA coordinate file. [default 2].

-o Specifies the output file name. [no default value]

Using the SGVP268.beta file generated above, the following command can be used to estimate the individual allele frequencies of studies samples:

./seekin getAF –i Study.chr22.ProPC.coord -b SGVP_268.chr22.beta -k 2 –o Study.chr22.indvAF.vcf.

The detailed meanings of flags of *getAF* module are summarized below.

-i Specifies the PCA coordinate file of studies samples. [no default value]

-b Specifies the PC-related regression coefficient file of reference samples. [no default value]

-k Specifies the number of PCs to compute. This number should be no more than the

number of PCs in the input PCA coordinate file. [default 2].

-o Specifies the output file name. The output is the compressed VCF format. [no default value]

Finally, we can run the *kinship* module with the above Stdudy.chr22.indvAF.vcf.gz as the input, the command is:

```
./seekin kinship -i ./Study.chr22.vcf.gz  -a  ./Study.chr22.indvAF.vcf.gz  -r 0.3 \
        -m 0.05   -d DS  -p admix -n 2000  -t 3 -w 1  -o Study.chr22.admix
```

The command is similar with the homogenous case but requiring the estimated individual allele frequency file specified by the flag *–a* and using the "admix" mode specified by flag *–p*.

4 Input files

In this section, we will describe the basic input files that are taken by *SEEKIN* and describe how to prepare for these input files step by step.

4.1 Genotype file

SEEKIN assumes that the input genotype files are in the compressed VCF format, which should contain at least a *GT* (observed genotype) or *DS* (imputed genotype) field for each marker. If the DS format field is available, the INFO fields should also contain corresponding *DR2* (Dosage R-Squared) information. A preliminary data processes to generate *SEEKIN*-ready VCF files (including DS format field) staring from BAM files includes two steps: 1) Summary the low-sequencing coverage reads into the genotype likelihoods; 2) Obtain the posterior genotype probabilities using imputation/phasing software with genotype likelihoods as the input. For the first step, you can use the *mpileup* command in *samtools* to call the variants with the typical command as followings:

```
samtools mpileup –b bamlist  -l  target.bed  -f  hs37d5.fa  -r regionA  -q 30 –Q 20 –t DP \
                –v –o A.gl.vcf.gz
```

This will generate the file named "A.gl.vcf.gz" which contains the genotyped variants derived from region specified by *–l* flag for all samples. We recommend user to specify a region with "–r chrom:start-stop", which will improve the speed of data retrieving by leveraging the BAM file indexing. Details about other options of *samtools* is available in http://www.htslib.org/doc/samtools.html. The second step is to phase the genotypes with or without external reference panel using Beagle software with the command as following:

```
java -Xmx10g –jar beagle.22Apr16.1cf.jar  gl=A.gl.vcf.gz  ref=Ref.vcf.gz  chrom=regionA\
   impute=false out= A.gp.vcf.gz modelscale=2 nthreads=5 gprobs=true niterations=0
```

This command will generate a VCF file named A.gp.vcf.gz. The *gl* argument in the above command specifies the input file and the type of genotypes used to estimate the posterior genotype probabilities. The option "*gprobs=true*" specifies that the DS format field will be included in the results. Details of other arguments can be seen in https://faculty.washington.edu/browning/beagle/beagle_4.1_03May16.pdf. Preparation of the VCF file can be carried out easily for parallelizing analysis by chromosome segment with changing the *-r* option in *samtools* and *chrom* option in Beagle.

4.2 PCA coordinate files

For kinship estimation of samples with population structure and admixture, user should prepare for two PCA coordinate files which contain the PCA coordinates of reference and study samples, respectively. These two files have the same format with tab-delimited. Each line represents one individual. The first two columns are population IDs and individual ID, and the following *k* columns represent the top *k* principal components (PCs). Please refer to LASER software on generating these PCA coordinate files. Below is an illustration of the format of coordinate file:

PopID	IndivID	PC1	PC2	PC3
CHS	IND1	10	9	-1
CHS	IND2	0	10	0
CHS	IND3	9	10	1

5 Output files

All output file will be saved in the current directory unless the path to a different directory given in the parameter value. We first describe the 5 output files from the *kinship* module which will be start with the prefix specified by the "-o" value (**5.1~5.3**).

5.1 _.log and terminal outputs

The terminal outputs are used to monitor and record the progress for each module when running *SEEKIN*. It starts with all parameter values used in the execution of *SEEKIN*, and report the progress of the program step by step. The log file is identical to the terminal outputs.

5.2 _.kin

This file provides the kinship estimation for all pairs of individuals. Each row in this file provides information for a pair of individuals. The first line is the header line. The first two columns correspond to the individual ID for the first and second individual of pair, respectively. The third column denotes the number of SNPs used for kinship estimation, and the fourth column represents the estimated kinship coefficient. One example is as following:

Ind1	Ind2	NSNP	Kinship
S1	S2	8592	0.0231
S1	S3	8592	0.0370
S1	S4	8592	0.0168

5.2 _.inbreed

This file provides the estimation of inbreeding coefficient estimation for each individual. Each row provides information for an individual. The columns are individual ID and *SEEKIN* inbreeding coefficient estimate, respectively. One example is as following:

Ind1	Inbreed_coef

ME-BBGMMRQ	0.0296
ME-HIUWPTI	0.0228
ME-5P732EC	-0.0338

5.3 _.index and _.matrix

The _.matrix file contains an $N \times N$ matrix (The variable N here means the sample size of study samples) of estimated kinship coefficients with the corresponding index of each individual shown in _.index file. For example, the kinship coefficient value given in row 2 and column 3 in the _.matrix file would correspond to the individuals in the _.index file who have indices of 2 and 3, respectively.

5.4 PC-related regression coefficient file of reference samples

Use the "*modelAF*" module in *SEEKIN*, we will generate the file which contains the PC-related regression coefficients for reference samples. The first line is a header line. Staring from the second row, each line represents information for one marker. From the first column to the fifth column are chromosome ID, genome position, reference allele, alternative reference allele, and allele frequencies of non-ref allele, respectively. And the remaining columns are the estimated coefficients for each PC. This file is also tab-delimited. An example is as following:

CHROM	POS	REF	ALT	AF	beta0	beta1
10	60969	C	A	0.48	0.96	-0.00
10	70969	G	A	0.41	0.96	-0.10

5.5 Individual allele frequencies file of studied samples

Use the "*getAF*" module in *SEEKIN*, we will generate the file which contains the individual allele frequencies for each sample. The generated file is the standard VCF in the compressed format with the first 4 header lines is as following:

```
##fileformat=VCFv4.2
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated Allele Frequencies of all samples">
##FORMAT=<ID=AF1,Number=A,Type=Float,Description="Estimated individual specific Allele Frequencies">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT ID1 ID2 ID3 ...
```

It is noted that we use AF1 field to specify the individual allele frequencies. Starting from the fifth line, each line denotes the individual allele frequency vector per marker. The SNP IDs and individual IDs are consistent with those from the input VCF file of study samples. The mean of individual allele frequencies is equal to the value specified by the AF field for each marker (i.e., the eighth column). An example is as following:

```
1 11008 . C G . . AF=0.0500 AF1 0.0534 0.0455 0.0536
1 12001 . A G . . AF=0.0200 AF1 0.0231 0.4451 0.1537
```

1	13102	.	C	T	.	.	AF=0.4500	AF1	0.2514	0.0123	0.0216
1	14052	.	C	G	.	.	AF=0.6500	AF1	0.0524	0.0252	0.9531

6 Version changes

Changes from previous versions of the *SEEKIN* software are noted here.

6.1 Version 1.0

-Initial release of the *SEEKIN* software.

7 References

1. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).
2. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* **46**, 409-15 (2014).
3. Teo, Y.Y. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res* **19**, 2154-62 (2009).
4. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
5. International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
6. Conomos, M.P., Reiner, A.P., Weir, B.S. & Thornton, T.A. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127-48 (2016).