ASKAT Wrapper

Pablo Cingolani & Karim Oualkacha

Goals

1. Parallelize and speed up calculations.
2. Make software more “friendly” for future users (e.g. non-R experts)
   a. Formalize the install process: check for missing dependencies.
   b. Invoke program using only one OS command line, instead of several R commands.
3. Code improvements to facilitate future re-use or contributions.

What the wrapper does...

1. Make sure there are TPED and TFAM files are available (can create them from VCF)

2. Make sure that all ASKAT dependencies are installed.

3. Split input file in "blocks" of SNPs (kinship calculation). For each block:
   a. Calculate the kinship matrix
   b. Create a file in an intermediate 'askat' format

Call ASKAT R function (in parallel) using the pre-calculated kinship matrix:

i. One control thread is created
ii. Two additional threads are created to consume and parse STDOUT & STDERR
iii. Each R process splits the input into sub-blocks (typically consisting of 20 SNPs), prepares the input and TMP files
iv. Finally the R process calls Karim's ASKAT(...) function.

Results are show in STDOUT

c. Java Threads collect and parses STDOUT from R processes

4. Show summary

Command line options

```
$ java -jar Askat.jar
ASKAT algorithm by Karim Oualkacha
Askat wrapper version 0.1b (build 2012-05-10), by Pablo Cingolani

Usage: java -jar Askat.jar [options] genotype
Options:
      -b <num>       : Number of SNPs used for calculating the kinship matrix. Default: 100000
      -d             : Debug mode (implies verbose)
      -noDep         : Do not perform dependency check.
      -h             : Show this help and exit.
      -kin <type>    : Kinship estimation type. Options {chr, avg, all, block}. Default: CHROMOSOME
      -p <num>       : Number of parallel processes. Default: 8
      -pathBin <dir> : Path to binary programs (e.g. FastLmm). Default: './'.
      -pathR <dir>   : Path to R scripts (ASKAT scripts). Default './r/'.
      -sb <num>      : Number of SNPs used for calculating the ASKAT algorithm. Default: 20
      -v             : Be verbose.
```

Note: The number of default parallel processes is the number of CPU-Cores in the computer. So the default value may change in each computers.

Example

```
$ java -jar Askat.jar -v geno_cov
00:00:00.000        ASKAT algorithm by Karim Oualkacha
00:00:00.003        Askat wrapper version 0.1b (build 2012-05-10), by Pablo Cingolani

00:00:00.003        Checking dependencies.
00:00:00.003        Checking dependency: Program 'R'
00:00:00.259        OK
00:00:00.259        Checking dependency: Program 'Rscript'
00:00:00.411        OK
00:00:00.412        Checking dependency: Program 'fastlmmc'
00:00:00.513        OK
00:00:00.513        Checking dependency: R library 'GenABEL'
00:00:02.472        Checking dependency: R library 'CompQuadForm'
00:00:02.756        Checking dependency: R library 'nFactors'
00:00:03.228        Checking dependency: R library 'MASS'
00:00:03.522        Checking dependency: R library 'Runiversal'
00:00:03.806        All dependencies found.

00:00:03.807        Running algorithm.
00:00:03.832        Creating block 'geno_cov.block.1_0.tped'. Number of entries: 12
00:00:03.835        Running block: geno_cov.block.1_0.tped
00:00:03.836        Calculating kinship matrix for block: geno_cov.block.1_0
00:00:06.609        Starting block: geno_cov.block.1_0
00:00:06.610        Create batches.
                              File 'geno_cov.block.1_0.tped' has 12 lines.
                              Split up to 20 lines per batch.
00:00:06.619        Batch 1. Line 1. Creating batch : geno_cov.block.1_0.1.askat
ASKAT_RESUTS:       Block:    geno_cov.block.1_0.1.askat    Sub-Block:          1         12        p-value:  0.3340863 Q:        25191.44  Polygenic.VC:       0.001026282
          Env.VC:   2.433403  lambda:   24508.56  481.4885  441.8991  384.5358  318.7743  290.2661  122.1321  87.50353  65.80886  57.12162  40.80717
00:00:15.658        Finished block: geno_cov.block.1_0
00:00:15.658        Done.
```

Same example (less verbose)

```
$ java -jar Askat.jar -noDep geno_cov
ASKAT_RESUTS:  Block:    geno_cov.block.1_0.1.askat    Sub-Block:     1    12   p-value:
     0.3340863 Q:   25191.44  Polygenic.VC:  0.001026282    Env.VC:   2.433403  lambda:
     24508.56  481.4885  441.8991  384.5358  318.7743  290.2661  122.1321  87.50353
65.80886  57.12162  40.80717
```

R Code review

Performed mostly aesthetic and minor optimization changes. I made sure formulae remained intact throughout the process:

- Joined all functions into one file (askar.r)
- Functions accept file and dir names or are assigned to variable. Before some of them were hardcoded.
- Call to fastLmm should be in one function. It was split in two or more, thus creating several "modularity" issues.
- In VC.FaST.LMM you write to fastlmmOutFileName and immediately read the same file. Changed by those lines of code by using a faster command.
- In VC.FaST.LMM: Side effect deletes a file created outside the function. This is bad practice, we should change it.
- Mixed assignment operators: '=' and '<-' are both used. R programmers are quite 'orthodox' and usually prefer '<-'.
- Changed 'system("rm ...") by 'unlink(...)' which is system independent.
- Changed 'paste("...", sep="")' by just "...". It has the same effect.
- In VC.FaST.LMM, 'ans' is created just to return a list (removed).
- When writing pheno.txt and TFAM files, the separator was "   " (i.e.four spaces). Changed it to single space. Apparently it only made files larger and marginally slower to parse (when running this millons of times, everything counts).
- Added return statement to clarify the code.
- Show ASKAT results in one line (using ASKAT_RESULTS) to make it easier for other programs to parse this output.