Parallel Computing for Mixed Model Implementation of Genomic Prediction and Variance Component Estimation of Additive and Dominance Effects

Chunkao Wang[1], Dzianis Prakapenka[2], H. Birali Runesha[2], Yang Da[1]
[1]Department of Animal Science, University of Minnesota, St. Paul, MN 55108, USA
[2]Research Computing Center, The University of Chicago, Chicago, IL 60637, USA

ABSTRACT: We developed the GVCBLUP package using shared memory (SM) and Message Passing Interface (MPI) parallel computing for genomic prediction and variance component estimation using mixed model methods. The GREML_CE and GREML_QM programs in the package offer complementary computing advantages and identical results of GBLUP and GREML along with heritability estimates using a combination of EM-REML and AI-REML algorithms. GREML_CE was designed for large numbers of SNP markers and GREML_QM for large numbers of individuals. For the SM version, GREML_CE could analyze 50,000 individuals with 400K SNP markers and GREML_QM could analyze 100,000 individuals with 50K SNP markers. For the MPI version, GREML_CE was tested for 50,000 individuals with 1 million SNP markers and 100,000 individuals with 41K SNP markers.
Keywords: genomic selection, variance component, heritability, BLUP

Introduction

Genomic prediction using genome-wide single nucleotide polymorphism (SNP) has become a powerful approach to capture genetic effects dispersed over the genome for predicting an individual's genetic potential for a phenotype (Meuwissen et al., 2001; VanRaden et al., 2009). Genomic estimation of variance component using genome-wide SNP markers is a powerful tool for estimating the genetic contribution of the whole-genome to a phenotype and for addressing the missing heritability problem where a large number of causal variants explained only a small fraction of the phenotypic variation (Yang et al., 2011; Da et al., 2014). The purpose of this research is to develop parallel computing tools to implement two computationally complementary computing strategies for genomic prediction and variance component estimation of additive and dominance effects with a wide-range of flexibility and functionality.

Materials and Methods

Two complementary computing strategies. We implemented two sets of formulations with complementary computing advantages and identical results based on two equivalent mixed models, the CE set for large numbers of markers (m>q) and the QM set for large numbers of individuals (q>m) (Da and Wang, 2013; Da et al., 2014), where m = number of SNP markers, and q = number of individuals. Several methods for calculating genomic relationships were implemented.

Shared memory parallel computing. GVCBLUP was programmed in C++ language using Eigen and Intel Math Kernel libraries (MKL). Eigen is a C++ template library for linear algebra, supports large dense and sparse matrices and supplies easy-to-use coding expression for linear algebra, which was used for creation, transformation of matrices. Intel MKL provides BLAS and LAPACK linear algebra routines and is optimized for Intel processors with multiple cores by using shared memory parallel computing technology, which is used for dense matrix inversion.

MPI distributed shared memory implementation. A parallel version of the GVCBLUP (GREML-CE) has been implemented using MPI between compute nodes and shared memory parallelism within the node. The early MPI implementation is written in Fortran90 using 2D block cyclic distribution to scatter large matrices among given nodes. The ScaLAPACK library, a set of high-performance linear algebra routines for parallel distributed memory machines using MPI message-passing layer, BLACS communication subprograms, and BLAS routines, is used for generalized matrix inversion. The code was compiled with Intel MKL version of the library and leverages shared-memory parallelization for the cores in one node when allocated one MPI task per node. This approach decreases memory requirements per node to process larger data sets.

EM and AI-REML. A combination of EM type of REML (EM-REML) and AI-REML was implemented. AI-REML generally is much faster than EM but is not robust as EM. We required at least two iterations of EM-REML and the user may specify a larger number of EM-REML iterations before switching to AI-REML. The program automatically returns to EM-REML if AI-REML yields a negative estimate for any of the variance component estimates. This strategy is designed to guarantee estimates of variance components to be positive.

Calculation of SNP heritabilities: Both GREML_CE or GREML_QM can output additive and dominance marker effects as well as additive and dominance marker heritbilities for every SNP marker. For SNP i, additive heritability or heritability in the narrow sense ($h^2_{\alpha i}$), dominance heritability ($h^2_{\delta i}$) and the total heritability or heritability in the broad sense (H^2_i) are:

$$h^2_{\alpha i} = \sigma^2_{\alpha i} / \sigma^2_y = (\hat{\alpha}^2_i / \sum_{i=1}^{m} \hat{\alpha}^2_i) h^2_{\alpha}$$

$$h^2_{\delta i} = \sigma^2_{\delta i} / \sigma^2_y = (\hat{\delta}^2_i / \sum_{i=1}^{m} \hat{\delta}^2_i) h^2_{\delta}$$

$$H^2_i = h^2_{\alpha i} + h^2_{\delta i}$$

where $\hat{\alpha}_i$ = GBLUP of additive effect of SNP i, $\hat{\delta}_i$ = GBLUP of additive effect of SNP i, $\sigma^2_y = \sigma^2_\alpha + \sigma^2_\delta + \sigma^2_e$ =

phenotypic variance, h^2_α = additive heritability of all SNP markers, and h^2_δ = dominance heritability of all SNP markers. The output file for the SNP effects and heritabilities was designed such that those results can be directly used as the input file for graphing and graphical viewing by SNPEVG2 (Wang et al., 2012).

Results and Discussion

Structure of GVCBLUP. The GVCBLUP package has three main programs, GREML_CE, GREML_QM and GCORRMX (Figure 1). This short article will focus on GREML_CE and GREML_QM.

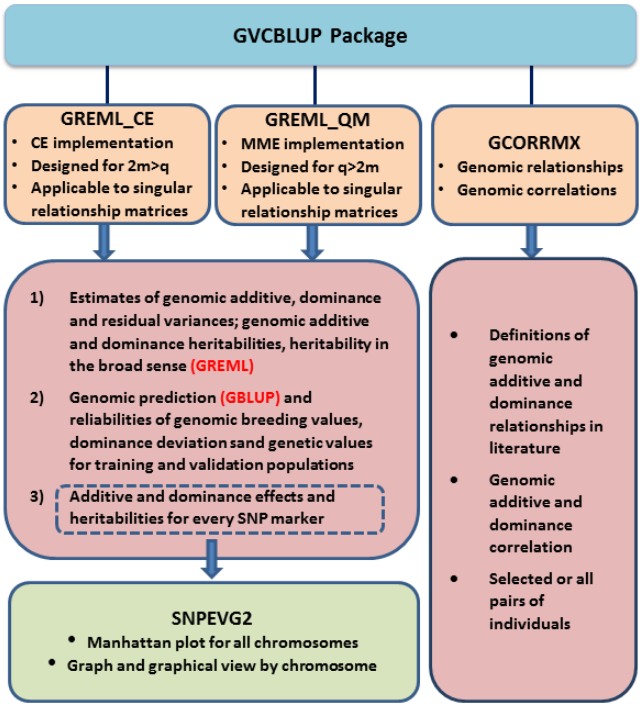

Figure 1. Structure of the GVCBLUP package.
(m = number of SNP markers, q = number of individuals)

GREML_CE and GREML_QM programs. The GREML_CE and GREML_QM programs calculate GREML estimates of additive, dominance and residual variances, additive and dominance heritabilities, as well as heritability in the broad sense, which is the summation of the additive and dominance heritabilities. GBLUP and reliability of breeding value, dominance deviation and genotypic value (summation of breeding value and dominance deviation) of each individual in the training or validation population are calculated at the end of variance component estimation. Assuming one observation per individuals, GREML_CE is more efficient than GREML_QM if 2m>q and is less efficient than GREML_QM if q>2m, where q = number of individuals and m = number of SNP markers. The example in Table 1 shows the complementary computing advantages of GREML_CE and GREML_QM. Both programs had identical results and required the same numbers of iterations. For 1000 individuals and 3000 SNP markers, GREML_CE required 5 seconds and GREML_QM required 69 seconds, whereas for 3000 individuals and 1000 SNP markers, GREML_CE required 32 seconds and GREML_QM required 6 seconds (Table 1).

Table 1. Computing time (seconds) using GREML_CE and GREML_QM for simulated datasets[1].

	q=1000 m=3000		q=3000 m=1000	
	CE	QM	CE	QM
Number of iteration	10	10	7	7
Total time	5	69	32	6

[1] Run on a personal computer (PC) with Intel Core i7-2600 (4 cores) of 3.40GHz and memory of 8.00GB.

Given q = 2m, the required memory storage of GREML_QM is approximately 1.5 times larger than GREML_CE, but GREML_QM is faster than GREML_CE due to the fact that GREML_CE requires twice as many matrix multiplication between large dense matrices. The shared memory parallel computing of GREML_CE and GREML_QM achieved excellent scalability on ItascaSB cluster in which each node contains two eight-core Sandy bridge E5-2670 processor chips (2.6 GHz) and 256 Gb memory and run *Linux* operating system.

Table 2. Capacity and speed of GREML_CE and GREML_QM for genomic estimation of additive, dominance and residual variances (tolerance = 10^{-8}) on ItascaSB supercomputer.

	CE	CE	QM	QM[1]
q	20,000	50,000	200,000	100,000
m	1 million	400K	10K	50K
Matrix creation	3.7 hrs	6.0 hrs	14.9 min	0.33 hrs
Total time	4.8 hrs	23.2 hrs	2 hrs	45.8 hrs
No. of iterations	12	13	20	20

[1]Not including computing time for GBLUP reliabilities.

The SNP input and the calculations of genomic relationships matrices ($\mathbf{A}_g$ and $\mathbf{D}_g$) required more computing time than per iteration of the estimation step. For shared memory parallel computing, GREML_CE was able to use 50,000 individuals with 400K SNP markers with total computing time of 23 hours for 13 iterations. For 20,000 individuals with one million SNP markers, GREML_CE only required 4.8 hours. GREML_QM was highly efficient for using low-density SNP markers, requiring only 2 hours for 200,000 individuals and 10K SNP markers. For 100,000 individuals with 50K SNP markers, GREML_QM required about 46 hours for 20 iterations (Table 2).

The implementation of the AI-REML algorithm for both GREML_CE and GREML_QM resulted in fast convergence rate, requiring between 12-20 iterations to converge with a strict tolerance level of 10^{-8}, compared to 295-458 iterations using EM-REML (Table 3). Although AI-REML was fast, extreme heritability levels (0 or 1) may cause failure of AI-REML. For eight of ten replications

with null heritability, AI-REML failed, but the variance components still can be estimated with EM type algorithm with large number of iterations. AI-REML was successful for all replications with heritability of 0.3 (Table 3).

Table 3. Comparison of iteration numbers of EM-REML and AI-REML (tolerance = 10^{-8}) using simulated data with different heritability levels.

Replicate	$h_\alpha^2 = 0.0, h_\delta^2 = 0.0$		$h_\alpha^2 = 0.3, h_\delta^2 = 0.3$	
	EM-REML	AI-REML	EM-REML	AI-REML
1	173	-[1]	322	9
2	481	18	458	10
3	138	-	295	10
4	1000	1000[1]	431	11

[1]AI-REML failed.

Preliminary results of the MPI implementation indicate that MPI is a promising computing solution to remove the computing bottleneck for large-scale genomic estimation requiring matrix inversions. For the two cases that the SM version cannot compute, the MPI version completed the computations easily on an Infiniband SandyBridge HPC Cluster (2.6Ghz-32G mem) with 16 cores per node. Each iteration required 9.5 minutes for 50,000 individuals with 1 million SNP markers using 60 nodes and required 98 minutes for 100,000 individuals with 41K SNP markers using 40 nodes with 16 cores per node (Table 4). The capability of the MPI version is expected to increase as the number of nodes increases.

Table 4. Preliminary testing results of the MPI version of GREML_CE for genomic estimation of additive, dominance and residual variances (tolerance = 10^{-8}).

q (number of individuals)	50,000	100,000
m (number of SNP markers)	1 million	41K
SNP input, matrix creation	82 min	58 min
Time per iteration	9.5 min	98 min
Number of nodes	60	40

GREML_CE and GREML_QM each has three output files for results of GREML, GBLUP, and SNP marker effects and heritabilities, in addition to screen displays. The GREML output files contain estimates and standard errors of variance components at each iteration, and the final estimates of variance components, heritabilities and their standard errors. The GBLUP output file contains GBLUP of breeding values and dominance deviations, genotypic values, as well as the corresponding reliabilities for both training and validation populations. These GBLUP results were calculated using the GREML estimates at the last iteration. Both GREML_CE and GREML_QM have a user option to output SNP additive and dominance marker effects and heritbilities for every SNP. The SNP effects and heritabilities can be readily graphed and displayed by SNPEVEG2 (Wang et al., 2012) including Manhattan plots and graphical figures by chromosome using the absolute GBLUP values (Figure 2A and 2B), or log10 of SNP heritability (Figure 2C and 2D).

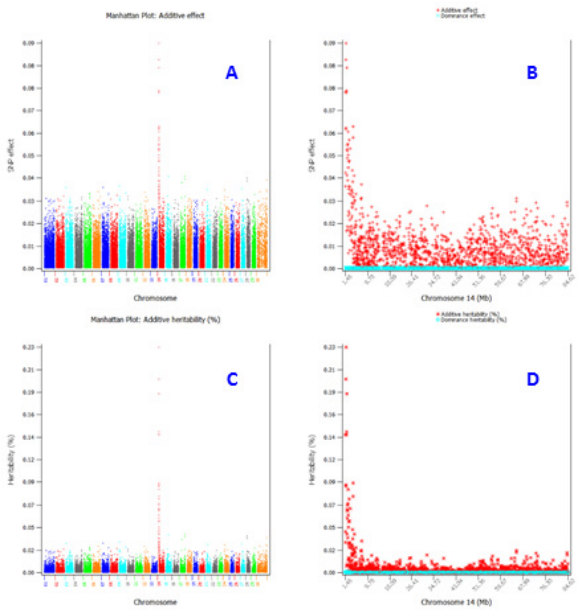

Figure 2. Graphical viewing of SNP additive and dominance effects and heritabilities. A: Manhattan plot using the absolute GBLUP of additive effects. **B**: Chromosome 14 graph using the absolute GBLUP of additive and dominance effects. **C**: Manhattan plot of SNP additive heritabilities. **D**: Chromosome 14 graph using SNP additive and dominance heritabilities. (Dominance GBLUP values were all near zero, consistent with the fact that the phenotypic values for fat percentage were PTA values that are additive effects. The highly significant chromosome 14 region is the DGAT1 region. The total additive heritability of SNP markers in the 1675278-4606904 Mb region of chromosome 14 including DGAT1 was 0.0248.)

Conclusion

The GVCBLUP package is a powerful and user friendly computing tool for assessing the type and magnitude of genetic effects affecting a phenotype by estimating whole-genome additive and dominance heritabilities of a phenotype using genome-wide SNP markers. MPI is a promising computing solution to address the memory and computing bottleneck for large-scale genomic estimation requiring matrix inversions.

Literature Cited

Meuwissen THE, Hayes BJ, Goddard ME. Genetics 2001, 157(4):1819-1829.

VanRaden PM, Van Tassell CP, Wiggans GR et al. J Dairy Sci 2009, 92:16-24.

Yang J, Lee SH, Goddard ME et al. Am J Hum Genet 2011, 88(1):76-82.

Da Y, Wang. Abstract #P1004, Plant and Animal Genome XXI Conference: January 12-16 2013; San Diego.

Da Y, Wang C, Wang S et al. PloS one 2014, 9(1):e87666.

Wang S, Dvorkin D, Da Y. BMC Bioinformatics 2012, 13:319.