frontiers in Genetics

METHODS
published: 07 April 2020
doi: 10.3389/fgene.2020.00282

Edited by:
Jianke Li,
Chinese Academy of Agricultural Sciences, China

Reviewed by:
Peipei Ma,
Shanghai Jiao Tong University, China
George R. Wiggans,
Agricultural Research Service (USDA), United States

***Correspondence:**
Yang Da
yda@umn.edu

†ORCID:
Dzianis Prakapenka
orcid.org/0000-0002-4592-7120
Chunkao Wang
orcid.org/0000-0002-3660-5736
Zuoxiang Liang
orcid.org/0000-0002-9058-9327
Cheng Bian
orcid.org/0000-0002-3638-5550
Cheng Tan
orcid.org/0000-0003-4190-2636
Yang Da
orcid.org/0000-0003-0119-7928

‡Present address:
Chunkao Wang,
Cobb-Vantress, Inc., Siloam Springs, AR, United States

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *27 November 2019*
Accepted: *09 March 2020*
Published: *07 April 2020*

Citation:
Prakapenka D, Wang C, Liang Z, Bian C, Tan C and Da Y (2020) GVCHAP: A Computing Pipeline for Genomic Prediction and Variance Component Estimation Using Haplotypes and SNP Markers. Front. Genet. 11:282. doi: 10.3389/fgene.2020.00282

GVCHAP: A Computing Pipeline for Genomic Prediction and Variance Component Estimation Using Haplotypes and SNP Markers

Dzianis Prakapenka[1†], Chunkao Wang[1†‡], Zuoxiang Liang[1†], Cheng Bian[1,2†], Cheng Tan[1,3†] and Yang Da[1*†]

[1] Department of Animal Science, University of Minnesota, Saint Paul, MN, United States, [2] State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, China, [3] National Engineering Research Center for Breeding Swine Industry, South China Agricultural University, Guangzhou, China

Haplotype prediction models open many possibilities to improve the accuracy of genomic selection but require more data processing and computing time than single-SNP prediction models. To facilitate haplotype analysis for genomic prediction and estimation using structural and functional genomic information, we developed a computing pipeline to implement haplotype analysis with capabilities for preparation of input data for haplotype analysis, genomic prediction and estimation using GVCHAP, and analysis of GVCHAP results. Data preparation includes utility programs for haplotype imputing; defining haplotype blocks by a fixed number of SNPs, a fixed distance in base pairs per block, or user defined block lengths based on structural or functional genomic information or a mixture of both types of information; and defining haplotype genotypes within each haplotype block. GVCHAP is the main program for genomic prediction and estimation, calculates GREML (genomic restricted maximum likelihood) estimates of variance components and heritabilities, and calculates GBLUP (genomic best linear unbiased prediction) for additive and dominance values of single SNPs as well as additive values of haplotypes with reliability estimates for training and validation populations. A two-step strategy and a method of multi-node processing are implemented to remove the computing bottleneck due to the creation of genomic relationship matrices for large samples. The analysis of GVCHAP results includes calculation of observed prediction accuracies from validation studies and preparation of input files for graphical visualization of heritability estimates of haplotype blocks as well as estimates of SNP effects and heritabilities. The entire pipeline provides an efficient and versatile computing tool for identifying the most accurate haplotype model among many candidate haplotype models utilizing structural and functional genomic information for genomic selection.

Keywords: genomic selection, haplotype, SNP, heritability, prediction accuracy

INTRODUCTION

Current methods using single nucleotide polymorphism (SNP) markers for genomic evaluation mostly use single-SNP prediction models. Haplotype analysis opens many possibilities of using structural and functional genomic information to improve the accuracy of genomic evaluation and gene discovery (Da, 2015). Compared to studies using single-SNP prediction models, only limited studies were available for using haplotype models in genomic evaluation (Calus et al., 2008; Villumsen et al., 2009; Mulder et al., 2010; Boichard et al., 2012; Cuyabano et al., 2015; Hess et al., 2017; Jónás et al., 2017; Jiang et al., 2018; Jan et al., 2019). Those studies achieved mixed results from little to substantial improvement in prediction accuracy due to the use of haplotypes relative to single-SNP models, and used haplotype blocking methods that are only a fraction of many possible haplotype models. However, investigating many haplotype models per trait is a computing challenge because haplotype models require considerably more data processing and computing resources than required by single-SNP models. Validation study is a commonly used approach to identify best haplotype models from a large number of candidate models but increases data processing work and computational difficulty. The computing pipeline in this article provides a full-featured computing tool for genomic prediction and variance component estimation using haplotypes with capability to minimize the data processing work, reduce computing difficulty, and conduct haplotype genomic prediction and estimation in an automated fashion.

METHODS

Mixed Model With SNP and Haplotype Effects

GVCHAP implements a multi-allelic haplotype mixed model that treats each haplotype block as a 'locus' and each haplotype within the haplotype block as an 'allele' (Da, 2015). The mixed model may include SNP additive and dominance effects and haplotype additive effects, with user flexibility to fit any or all of these three genomic effects. Haplotype dominance effects were coded but disabled in GVCHAP due to the large number of haplotype pairs that may exist in some haplotype blocks. The dominance effect of each haplotype pair requires all three genotypes to define, one heterozygous and two homozygous genotypes of the SNP, but one or both homozygous genotypes may be missing when many haplotypes with small frequencies exist.

Based on the quantitative genetics models of haplotypes and SNPs (Da et al., 2014; Da, 2015), the quantitative genetic model with SNP additive and dominance effects as well as haplotype additive effects is:

$$\begin{aligned}\mathbf{y} &= \mathbf{Xb} + \mathbf{Z}(\mathbf{W}_\alpha\boldsymbol{\alpha}_o + \mathbf{W}_\delta\boldsymbol{\delta}_o + \mathbf{W}_{\alpha h}\boldsymbol{\alpha}_{ho}) + \mathbf{e} \\ &= \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d} + \mathbf{a}_h) + \mathbf{e} \end{aligned} \quad (1)$$

where $\mathbf{Z}$ = N × n incidence matrix allocating phenotypic observations to each individual = identity matrix for one observation per individual (N = n), N = number of observations, n = number of individuals, $\boldsymbol{\alpha}_o$ = m × 1 column vector of SNP additive effects, m = number of SNPs, $\mathbf{W}_\alpha$ = n × m model matrix of $\boldsymbol{\alpha}_o$, $\boldsymbol{\delta}_o$ = m × 1 column vector for dominance effects of SNP genotypes, $\mathbf{W}_\delta$ = n × m model matrix of $\boldsymbol{\delta}_o$, $\boldsymbol{\alpha}_h = n_{\alpha h} \times 1$ column vector of haplotype additive effects, $n_{\alpha h}$ = number of haplotype additive effects, $\mathbf{W}_{\alpha h}$ = n × $n_{\alpha h}$ model matrix of $\boldsymbol{\alpha}_h$, $\mathbf{b}$ = c × 1 column vector of fixed effects such as herd-year-season in dairy cattle, c = number of fixed effects, $\mathbf{X}$ = N × c model matrix of $\mathbf{b}$, $\mathbf{a} = \mathbf{W}_\alpha\boldsymbol{\alpha}_o$ = SNP genomic additive values, $\mathbf{d} = \mathbf{W}_\delta\boldsymbol{\delta}_o$ = SNP genomic dominance values, $\mathbf{a}_h = \mathbf{W}_{\alpha h}\boldsymbol{\alpha}_{oh}$ = haplotype genomic additive values. The haplotype coding represented by $w_{\alpha h}^{ij,k}$ in $\mathbf{W}_{\alpha h}$is: $w_{\alpha h}^{ij,k} = 2p_k$ for i,j≠k (a_{ij} and α_{1k} do not share allele k), $w_{\alpha h}^{ij,k} = -(1 - 2p_k)$ for i≠j but i = k or j = k (a_{ij} and α_{1k} share allele k, i≠j), and $w_{\alpha h}^{ij,k} = -2(1 - p_k)$ for i = j = k (a_{ij} and α_{1k} share allele k, i = j), where a_{ij} = additive value of haplotype genotype with the i^{th} and j^{th} haplotypes, and α_{1k} = additive effect or the average effect of gene substitution as the difference between the allelic (haplotype) effects of the first and the k^{th} haplotypes (Da, 2015). SNP codings in $\mathbf{W}_\alpha$ and $\mathbf{W}_\delta$ are the same as defined by the single-SNP quantitative genetics model (Da et al., 2014). The first and second moments of Eq. 1 are $E(\mathbf{y}) = \mathbf{Xb}$, and

$$\begin{aligned}\mathrm{Var}(\mathbf{y}) = \mathbf{V} &= \mathbf{Z}(\sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}'_\alpha + \sigma^2_{\delta o}\mathbf{W}_\delta\mathbf{W}'_\delta + \sigma^2_{\alpha oh}\mathbf{W}_{\alpha h}\mathbf{W}'_{\alpha h})Z' \\ &+\sigma^2_e\mathbf{I}_N = \mathbf{Z}(\mathbf{G}_a + \mathbf{G}_d + \mathbf{G}_{ah})Z' + \sigma^2_e\mathbf{I}_N \end{aligned} \quad (2)$$

where $\sigma^2_{\alpha o}$ = SNP additive variance, $\sigma^2_{\delta o}$ = SNP dominance variance, $\sigma^2_{\alpha oh}$ = haplotype additive variance, σ^2_e = residual variance, $\mathbf{G}_a = \mathrm{var}(\mathbf{a}) = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}'_\alpha$, $\mathbf{G}_d = \mathrm{var}(\mathbf{d}) = \sigma^2_{\delta o}\mathbf{W}_\delta\mathbf{W}'_\delta$, and $\mathbf{G}_{ah} = \mathrm{var}(\mathbf{a}) = \sigma^2_{\alpha oh}\mathbf{W}_{\alpha h}\mathbf{W}_{\alpha h}{}'$.

Based on Eqs. 1, 2, the mixed model with genomic relationship matrices is a reparameterized and an equivalent model of Eqs. 1, 2, i.e.,

$$\begin{aligned}\mathbf{y} &= \mathbf{Xb} + \mathbf{Z}(\mathbf{T}_\alpha\boldsymbol{\alpha} + \mathbf{T}_\delta\boldsymbol{\delta} + \mathbf{T}_{\alpha h}\boldsymbol{\alpha}_h) + \mathbf{e} \\ &= \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d} + \mathbf{a}_h) + \mathbf{e} \end{aligned} \quad (3)$$

$$\mathrm{Var}(\mathbf{a}) = \sigma^2_\alpha\mathbf{A}_g = \sigma^2_\alpha\mathbf{T}_\alpha\mathbf{T}'_\alpha = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}'_\alpha = \mathbf{G}_a \quad (4)$$

$$\mathrm{Var}(\mathbf{d}) = \sigma^2_\delta\mathbf{D}_g = \sigma^2_\delta\mathbf{T}_\delta\mathbf{T}_\delta{}' = \sigma^2_{\delta o}\mathbf{W}_\delta\mathbf{W}'_\delta = \mathbf{G}_d \quad (5)$$

$$\mathrm{Var}(\mathbf{a}_h) = \sigma^2_{\alpha h}\mathbf{A}_{\alpha h} = \mathbf{T}_{\alpha h}\mathbf{T}'_{\alpha h} = \sigma^2_{\alpha oh}\mathbf{W}_{\alpha h}\mathbf{W}'_{\alpha h} = \mathbf{G}_{ah} \quad (6)$$

$$\begin{aligned}\mathrm{Var}(\mathbf{y}) = \mathbf{V} &= \mathbf{Z}(\sigma^2_\alpha\mathbf{A}_g + \sigma^2_\delta\mathbf{D}_g + \sigma^2_{\alpha h}\mathbf{A}_{gh})\mathbf{Z}' + \sigma^2_e\mathbf{I}_N \\ &= \mathbf{Z}(\mathbf{G}_a + \mathbf{G}_d + \mathbf{G}_{ah})\mathbf{Z}' + \sigma^2_e\mathbf{I}_N \end{aligned} \quad (7)$$

where $\mathbf{A}_g$ = SNP genomic additive relationship matrix, $\mathbf{D}_g$ = SNP genomic dominance relationship matrix, $\mathbf{A}_{gh}$ = haplotype genomic additive relationship matrix, $\mathbf{T}_\alpha = \mathbf{W}_\alpha/k_\alpha^{1/2}$, $\mathbf{T}_\delta = \mathbf{W}_\delta/k_\delta^{1/2}$, $\mathbf{T}_{\alpha h} = \mathbf{W}_{\alpha h}/k_{\alpha h}^{1/2}$, $\mathbf{a} = \mathbf{T}_\alpha\boldsymbol{\alpha} = \mathbf{W}_\alpha\boldsymbol{\alpha}_o$ = SNP genomic additive values, $\mathbf{d} = \mathbf{T}_\delta\boldsymbol{\delta} = \mathbf{W}_\delta\boldsymbol{\delta}_o$ = SNP genomic dominance values, $\mathbf{a}_h = \mathbf{T}_{\alpha h}\boldsymbol{\alpha}_h = \mathbf{W}_{\alpha h}\boldsymbol{\alpha}_{oh}$ = haplotype genomic additive

values, σ^2_α = SNP additive variance, σ^2_δ = SNP dominance variance, $\sigma^2_{\alpha h}$ = haplotype additive variance, σ^2_e = residual variance, **V** = phenotypic variance-covariance matrix, and

$$k_\alpha = tr(\mathbf{W}_\alpha \mathbf{W}'_\alpha)/n \quad (8)$$

$$k_\delta = tr(\mathbf{W}_\delta \mathbf{W}'_\delta)/n \quad (9)$$

$$k_{\alpha h} = tr(\mathbf{W}_{\alpha h} \mathbf{W}'_{\alpha h})/n \quad (10)$$

Each of k_α, k_δ and $k_{\alpha h}$ defined by Eqs. 8–10 is an average of the diagonal elements of $\mathbf{W}_j\mathbf{W}'_j$ (j = α, δ, αh). With this definition, variance components σ^2_α, σ^2_δ and $\sigma^2_{\alpha h}$ can be interpreted as the average of the corresponding variances of all individuals under the original quantitative genetics model of Eqs. 1, 2, i.e.,

$$\sigma^2_\alpha = tr(\mathbf{G}_a)/n = \sigma^2_{\alpha o} tr(\mathbf{W}_\alpha \mathbf{W}'_\alpha)/n = k_\alpha \sigma^2_{\alpha o} \quad (11)$$

$$\sigma^2_\delta = tr(\mathbf{G}_d)/n = \sigma^2_{\delta o} tr(\mathbf{W}_\delta \mathbf{W}'_\delta)/n = k_\delta \sigma^2_{\delta o} \quad (12)$$

$$\sigma^2_{\alpha h} = tr(\mathbf{G}_{ah})/n = \sigma^2_{\alpha oh} tr(\mathbf{W}_{\alpha h} \mathbf{W}'_{\alpha h})/n = k_{\alpha h} \sigma^2_{\alpha oh} \quad (13)$$

and the genomic relationship matrices in Eqs. 4–6 can be expressed as:

$$\mathbf{A}_g = \mathbf{T}_\alpha \mathbf{T}'_\alpha = \mathbf{W}_\alpha \mathbf{W}'_\alpha / k_\alpha \text{ (Hayes and Goddard, 2010)} \quad (14)$$

$$\mathbf{D}_g = \mathbf{T}_\delta \mathbf{T}_\delta{}' = \mathbf{W}_\delta \mathbf{W}'_\delta / k_\delta \quad (15)$$

(Da et al., 2014; Wang and Da, 2014)

$$\mathbf{A}_{gh} = \mathbf{T}_{\alpha h} \mathbf{T}'_{\alpha h} = \mathbf{W}_{\alpha h} \mathbf{W}'_{\alpha h} / k_{\alpha h} \text{ (Da, 2015)} \quad (16)$$

GVCHAP chose to implement the genomic relationship matrices using the average of the diagonal elements in $\mathbf{W}_j\mathbf{W}'_j$ (j = α, δ, αh) as the denominator, i.e., $k_j = tr(\mathbf{W}_j\mathbf{W}'_j)/n$ as the denominator (Eqs. 8–10), because this approach yields variance and heritability estimates in the study population that can be a random or an inbred population (Da, 2019). The genomic relationship matrix that uses the total SNP heterozygosity as the denominator (VanRaden, 2008) is not implemented by the current version of GVCHAP, unlike GVCBLUP that implements both methods with VanRaden's method as Definition I, and the Hayes-Goddard method as Definition II (Wang et al., 2014b). VanRaden's method preserves the properties of pedigree additive relationships, i.e., $a_{ii} = 1 + f$ and $a_{ij} = 2f_{ij}$, where a_{ij} = additive relationship between the i^{th} and j^{th} individuals, f_{ij} = coancestry coefficient between the i^{th} and j^{th} individuals, and f = inbreeding coefficient; but underestimates genetic variance components and heritability compared to the Hayes-Goddard method when inbreeding is present. The Hayes-Goddard method does not preserve the properties of pedigree additive relationships. Although these two methods for calculating genomic relationship matrices have different interpretations and generally have differences in estimates of variance components and heritabilities, these methods and the quantitative genetics model of Eqs. 1, 2 without using genomic relationships yield identical GBLUP and reliability (Da, 2019). The use of haplotype genomic relationships is in parallel to the use of SNP genomic relationships, but haplotype genomic relationships are not suitable for measuring relationships among individuals such as parent-offspring relationship due to recombination between SNPs within haplotype blocks (Da, 2015). The only practical application of haplotype genomic relationships is for multi-allelic markers such as microsatellite markers but such markers are virtually unused in current genetic research. This was another reason why only one method for defining genomic relationship matrix was implemented in GVCHAP.

Software Implementation

The GVCHAP program was developed based on the GREML_CE program in the GVCBLUP package for genomic prediction and variance component estimation using SNP markers (Wang et al., 2014b). As in GVCBLUP, GVCHAP is programmed in C++ language using Eigen and Intel Math Kernel library (MKL). Eigen is a C++ template library for linear algebra, supports large dense and sparse matrices. Intel MKL provides BLAS and LAPACK linear algebra routines and is optimized for Intel processors by using shared memory parallel computing technology. The multi-node processing (MNP) program is based on the input section of GVCHAP with modifications for processing the input SNP and haplotype files using multiple nodes. Nearly all the utility programs are written in Python. The GVCHAP implementation of the multi-allelic haplotype model was validated by a R-script that had the same results of EM-REML as GVCHAP for the same testing data. It was widely confirmed that EM-REML and AI-REML converge to the same estimates. The R-script and the testing data are included in the GVCHAP package.

RESULTS AND DISCUSSION

Structure of GVCHAP Computing Pipeline

The computing pipeline of GVCHAP (**Figure 1**) consists of three components: data preparation, GVCHAP analysis, and analysis of GVCHAP results. The complete list of computer programs in this pipeline and the technical details for using these programs are described in the GVCHAP User Manual (**Supplementary Material**). The following is an overview of this computing pipeline.

Computing Tools for Preparation of Input Files for GVCHAP Analysis

Five input files are required for running GVCHAP: SNP genotypes, haplotype genotypes, phenotypes, SNP map file, and parameter file. The SNP genotypic files, phenotypic file and map file are provided by the user and remain unchanged for GVCHAP analysis. The haplotype genotypic files can be tedious to prepare, particularly for studies evaluating many haplotype models using validation studies for each model. A set of utility programs prepares the haplotype genotypes and partially fill in the parameter file in an automated fashion. This process starts with converting the format of the user provided SNP genotypic file into the format for BEAGLE (Browning et al., 2018) or FINDHAP

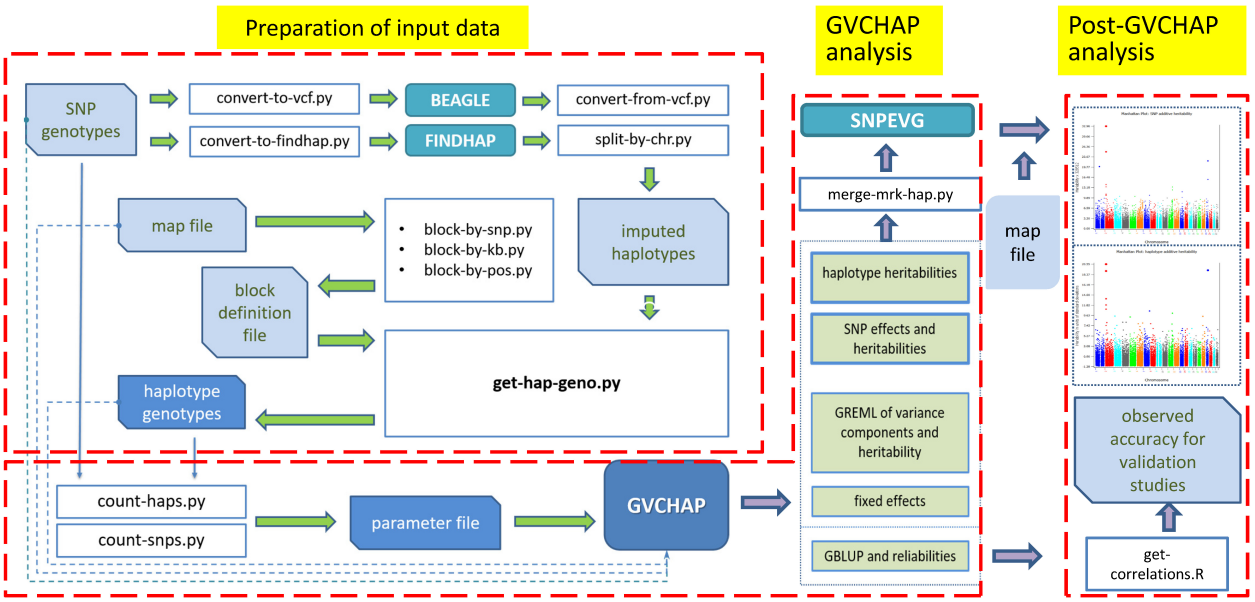

FIGURE 1 | Structure of GVCHAP computing pipeline. The pipeline consists of three components, preparation of input data for haplotype analysis, GVCHAP analysis of genomic prediction and estimation, and post-GVCHAP analysis.

(VanRaden et al., 2015), running BEAGLE or FINDHAP to produce imputed haplotypes, dividing the imputed haplotypes into haplotype blocks, and defining haplotype genotypes within each haplotype block (**Figure 2**).

Dividing haplotypes of each chromosome into haplotype blocks, to be referred to as haplotype blocking, is the first step for defining a specific haplotype model. Three utility programs for haplotype blocking allow three options for defining haplotype blocks: a fixed number of SNPs per block using 'block-by-snp.py,' or a fixed distance in kilo-bases per block using 'block-by-kb.py,' or a user provided haplotype blocking file that can have various block lengths using 'block-by-pos.py.' The user provided blocking file provides flexibility for using various types of structural and functional genomic information such as LD based blocking (LD = linkage disequilibrium) and gene based blocking. The parameter file contains controls for running GVCHAP, including the prediction model, the use of EM-REML and AI-REML, and information about the input and output files. Two utility programs (count-snps.py and count-haps.py) fill in the number of SNPs and the number of haplotype blocks for each SNP genotype file and each haplotype genotype file.

GVCHAP Analysis

The GVCHAP analysis produces GBLUP (genomic best linear unbiased prediction) and GREML (genomic restricted maximum likelihood estimation) for SNP effects and values as well as haplotype additive values with options to improve computing speed (**Figure 3**).

Seven Prediction Models

The prediction model can include SNP additive and dominance values, and haplotype additive values. For these values, GVCHAP offers seven models that include the full model of Eq. 3 and six variations of Eq. 3:

Model 1: SNP additive, dominance and haplotype additive values, $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d} + \mathbf{a}_h) + \mathbf{e}$;
Model 2: SNP and haplotype additive values, $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{a}_h) + \mathbf{e}$;
Model 3: SNP dominance values and haplotype additive values, $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}(\mathbf{d} + \mathbf{a}_h) + \mathbf{e}$;
Model 4: haplotype additive values only, $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}\mathbf{a}_h + \mathbf{e}$;
Model 5: SNP additive and dominance values, $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e}$;
Model 6: SNP additive values only, $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$;
Model 7: SNP dominance values only, $\mathbf{y} = \mathbf{Xb} + \mathbf{Zd} + \mathbf{e}$.

Models 1–4 contain haplotype additive values, and Models 5–7 are SNP models. The comparison between Models 1–4 and Models 5–7 for prediction accuracy provides an estimate whether haplotypes improve the prediction accuracy. For example, haplotypes improved the prediction accuracy if any of the haplotype models (Models 1–4) was more accurate than the three SNP models (Models 5–7) in validation studies. Similarly, validation studies can identify the most accurate model among the seven prediction models. Each of the seven models is configured by the starting values of the variance components in the parameter file through four parameters, var_snp_a, var_snp_d, var_snp_e, and var_hap_a for starting values of SNP additive variance, SNP dominance variance, residual variance, and haplotype additive variance respectively. These seven models and the main results of GBLUP and GREML for each model are summarized in **Table 1**.

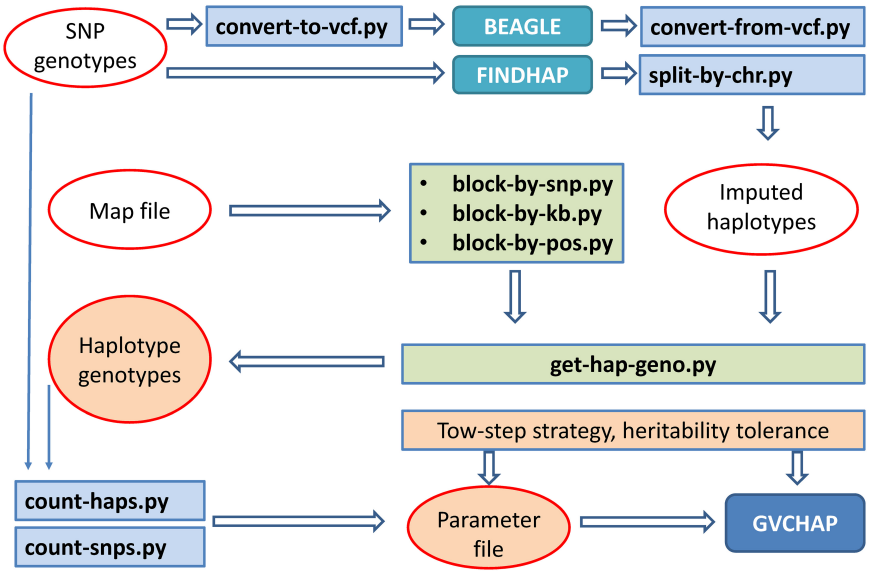

FIGURE 2 | Computing pipeline for preparing input files for GVCHAP analysis.

GREML Estimates of Variance Components and Heritabilities

GVCHAP calculates GREML estimates of variance components and hertiabilities for each of the genetic effects in the prediction model using a combination of EM-REML and AI-REML for iterative solutions adopted from GVCBLUP (Wang et al., 2014b). The program starts with a minimum of two EM-REML iterations, switches to AI-REML at iteration 3 by default, and switches back to EM-REML automatically when AI-REML fails. GREML calculates estimates of variance components with tolerance values, and heritability estimates with tolerance values and standard deviations.

GREML estimates may provide helpful information for choosing among the seven prediction models in **Table 1**. We recommend GREML with all types of genetic values in the prediction model (Model 1) as the initial GVCHAP analysis to determine whether any type of genetic values has a zero or near-zero heritability, and remove such genetic values from the prediction model. The inclusion of genetic values with negligible heritability may cause slow convergence for GREML and may only have negligible contribution to prediction accuracy. Therefore, removing such genetic values from the prediction model may significantly improve the computing speed with negligible change (positive or negative) to prediction accuracy. For example, the appropriate prediction model should be Model 2 if SNP dominance heritability is negligible, Model 3 if SNP additive heritability is negligible, or Model 4 if SNP additive and dominance hertiabilities are both negligible.

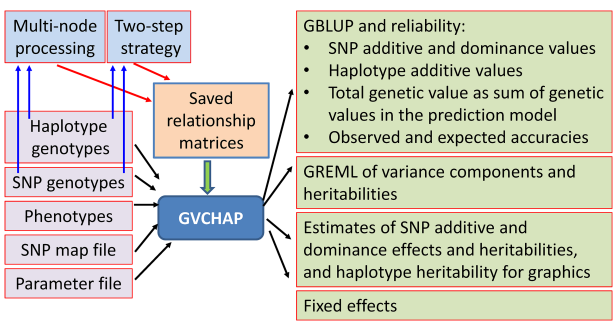

FIGURE 3 | Input and output files of GVCHAP analysis.

GBLUP, Reliability, and Expected Prediction Accuracy for Predicting Genetic Values

Converged GREML estimates of variance components are used for calculating GBLUP and reliability for each type of genetic values and the sum of SNP and haplotype values. In the output file for GBLUP and reliability, an individual is flagged with 'T' if the individual is in the training population with phenotypic observations, or 'V' if the individual has missing phenotypic value or is in the validation population with phenotypic value set as missing value. For the example of Model 1 with SNP additive and dominance values as well as haplotype additive values, GBLUP and reliability for each and the sum of these values are calculated. After sorting the output file by training (T) and validation (V) populations, the GBLUP estimates are:

$$\hat{\mathbf{a}} = (\hat{\mathbf{a}}_1', \hat{\mathbf{a}}_0') = \text{GBLUP of SNP additive values} \quad (17)$$

$$\hat{\mathbf{d}} = (\hat{\mathbf{d}}_1', \hat{\mathbf{d}}_0') = \text{GBLUP of SNP dominance values} \quad (18)$$

$$\hat{\mathbf{a}}_h = (\hat{\mathbf{a}}_{h1}', \hat{\mathbf{a}}_{h0}') = \text{GBLUP of haplotype additive values} \quad (19)$$

$$\hat{\mathbf{g}} = \hat{\mathbf{a}} + \hat{\mathbf{d}} + \hat{\mathbf{a}}_h = \text{GBLUP of total genotypic values} \quad (20)$$

TABLE 1 | Seven prediction models configured by parameters for starting values of variance components in the parameter file and the main results of GBLUP and GREML for each model.

	var_snp_a	var_snp_d	var_snp_e	var_hap_a
Model 1	var_snp_a NPV $\hat{\mathbf{a}}$, $\mathbf{R}^2_a$, $\hat{\sigma}^2_\alpha$, $\hat{h}^2_\alpha$	var_snp_d NPV $\hat{\mathbf{d}}$, $\mathbf{R}^2_d$, $\hat{\sigma}^2_\delta$, $\hat{h}^2_\delta$	var_snp_e NPV $\hat{\sigma}^2_e$	var_hap_a NPV $\hat{\mathbf{a}}_h$, $\mathbf{R}^2_{ah}$, $\hat{\sigma}^2_{\alpha h}$, $\hat{h}^2_{\alpha h}$, $\hat{\mathbf{g}}$, $\mathbf{R}^2_g$
Model 2	var_snp_a NPV $\hat{\mathbf{a}}$, $\mathbf{R}^2_a$, $\hat{\sigma}^2_\alpha$, $\hat{h}^2_\alpha$	#var_snp_d	var_snp_e NPV $\hat{\sigma}^2_e$	var_hap_a NPV $\hat{\mathbf{a}}_h$, $\mathbf{R}^2_{ah}$, $\hat{\sigma}^2_{\alpha h}$, $\hat{h}^2_{\alpha h}$, $\hat{\mathbf{g}}$, $\mathbf{R}^2_g$
Model 3	#var_snp_a	var_snp_d NPV $\hat{\mathbf{d}}$, $\mathbf{R}^2_d$, $\hat{\sigma}^2_\delta$, $\hat{h}^2_\delta$	var_snp_e NPV $\hat{\sigma}^2_e$	var_hap_a NPV $\hat{\mathbf{a}}_h$, $\mathbf{R}^2_{ah}$, $\hat{\sigma}^2_{\alpha h}$, $\hat{h}^2_{\alpha h}$, $\hat{\mathbf{g}}$, $\mathbf{R}^2_g$
Model 4	#var_snp_a	#var_snp_d	var_snp_e NPV $\hat{\sigma}^2_e$	var_hap_a NPV $\hat{\mathbf{a}}_h$, $\mathbf{R}^2_{ah}$, $\hat{\sigma}^2_{\alpha h}$, $\hat{h}^2_{\alpha h}$
Model 5	var_snp_a NPV $\hat{\mathbf{a}}$, $\mathbf{R}^2_a$, $\hat{\sigma}^2_\alpha$, $\hat{h}^2_\alpha$	var_snp_d NPV $\hat{\mathbf{d}}$, $\mathbf{R}^2_d$, $\hat{\sigma}^2_\delta$, $\hat{h}^2_\delta$, $\hat{\mathbf{g}}$, $\mathbf{R}^2_g$	var_snp_e NPV $\hat{\sigma}^2_e$	#var_hap_a
Model 6	var_snp_a NPV $\hat{\mathbf{a}}$, $\mathbf{R}^2_a$, $\hat{\sigma}^2_\alpha$, $\hat{h}^2_\alpha$	#var_snp_d	var_snp_e NPV $\hat{\sigma}^2_e$	#var_hap_a
Model 7	#var_snp_a	var_snp_d NPV $\hat{\mathbf{d}}$, $\mathbf{R}^2_d$, $\hat{\sigma}^2_\delta$, $\hat{h}^2_\delta$	var_snp_e NPV $\hat{\sigma}^2_e$	#var_hap_a

A non-zero positive value (NPV) after the parameter activates the genetic value in the prediction model, and a '#' sign in front of the parameter removes the genetic value from the prediction model. $\hat{\mathbf{a}}$ *= GBLUP of SNP additive value.* $\mathbf{R}^2_a$ *= reliability of* $\hat{\mathbf{a}}$*.* $\hat{\mathbf{d}}$ *= GBLUP of SNP dominance value.* $\mathbf{R}^2_d$ *= reliability of* $\hat{\mathbf{d}}$*.* $\hat{\mathbf{g}}$ *= GBLUP of total genotypic value,* $= \hat{\mathbf{a}} + \hat{\mathbf{d}} + \hat{\mathbf{a}}_h$ *for Model 1,* $= \hat{\mathbf{a}} + \hat{\mathbf{a}}_h$ *for Model 2,* $= \hat{\mathbf{d}} + \hat{\mathbf{a}}_h$ *for Model 3,* $= \hat{\mathbf{a}} + \hat{\mathbf{d}}$ *for Model 5.* $\mathbf{R}^2_g$ *= reliability of* $\hat{\mathbf{g}}$*.* $\hat{\sigma}^2_\alpha$ *= GREML estimate of SNP additive variance.* $\hat{\sigma}^2_\delta$ *= GREML estimate of SNP dominance variance.* $\hat{\sigma}^2_e$ *= GREML estimate of residual variance.* $\hat{\sigma}^2_{\alpha h}$ *= GREML estimate of haplotype additive variance.* $\hat{h}^2_\alpha$ *= GREML estimate of SNP additive heritability.* $\hat{h}^2_\delta$ *= GREML estimate of SNP dominance heritability.* $\hat{h}^2_{\alpha h}$ *= GREML estimate of haplotype additive heritability.*

where '∧' indicates estimated value, subscript '1' indicates training population, and subscript '0' indicates validation population or individuals with missing phenotypic observations. Each of the above GBLUP estimates is accompanied by its reliability. The square root of a reliability estimate is the correlation between GBLUP and the unobservable true genetic value being predicted by the GBLUP, and is the expected accuracy for predicting the unobservable true genetic valu'e, $\mathbf{a}$, $\mathbf{d}$, $\mathbf{a}_h$, or $g = a + d + \mathbf{a}_h$. In the absence of validation studies, reliability or the expected prediction accuracy is the measure of prediction accuracy for a type of genetic value, e.g., SNP additive or dominance value, or haplotype additive value, or the sum of all these genetic values. The reliability formula for $\hat{\mathbf{g}} = \hat{\mathbf{a}} + \hat{\mathbf{d}} + \hat{\mathbf{a}}_h$ of Model 1 is:

$$R^2_{gi} = \begin{pmatrix} \mathbf{G}_\alpha \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\alpha + \mathbf{G}_\delta \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\delta + \mathbf{G}_{\alpha h}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_{\alpha h} \\ +\mathbf{G}_\alpha \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\delta + \mathbf{G}_\delta \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\alpha + \mathbf{G}_\alpha \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_{\alpha h} \\ +\mathbf{G}_{\alpha h}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\alpha + \mathbf{G}_\delta \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_{\alpha h} + \mathbf{G}_{\alpha h}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\delta \end{pmatrix}_{ii} / \left(A^{ii}_g \sigma^2_\alpha + D^{ii}_g \sigma^2_\delta + A^{ii}_{gh} \sigma^2_{\alpha h} \right) \quad (21)$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$; A^{ii}_g, D^{ii}_g and A^{ii}_{gh} are the i^{th} diagonal elements of $\mathbf{A}_g$, $\mathbf{D}_g$ and $\mathbf{a}_{gh}$ respectively; and subscripts ii of the numerator indicates the i^{th} diagonal element of the numerator matrix. The reliability formula for any of Models 2–7 can be readily derived from Eq. 21, e.g., the reliability of Model 2 is obtained from Eq. 21 by deleting all terms involving 'δ.'

Observed and Expected Accuracy for Predicting Phenotypic Values

For a validation study, this computing pipeline has a utility program (get-correlations.R) calculating the observed accuracy for predicting the phenotypic values using a type of genetic values (e.g., SNP additive values or haplotype additive values) or a sum of several types of genetic values (e.g., the sum of SNP dominance values and haplotype additive values) for each validation. This observed accuracy is calculated as the correlation between the predicted genetic values and the phenotypic values in the validation population (individuals flagged as 'V' in the output file) that were omitted when calculating GBLUP. The expected accuracy for predicting phenotypic values is the product between the expected accuracy for predicting genetic values and the square root of heritability (Legarra et al., 2008). For genomic prediction using a type of genetic values, e.g., SNP additive or dominance values, or haplotype additive values, or a sum of these genetic values, the observed and expected accuracies for predicting phenotypic values and the expected accuracies for predicting genetic values are:

$$\hat{R}_{0pi} = \mathrm{corr}(\hat{\mathbf{g}}'_{0i}, \mathbf{y}_0) \quad (22)$$

$$\hat{R}_{0p} = \sum_{i=1}^{k} \hat{R}_{0pi}/k \quad (23)$$

$$R_{0gi} = \sum_{j=1}^{n_{0i}} R_{0gij}/n_{0i} \quad (24)$$

$$R_{0g} = \sum_{i=1}^{k} R_{0gi}/k \quad (25)$$

$$R_{0p} = R_{0g}\sqrt{h^2} \quad (26)$$

where $\hat{R}_{0pi}$ = observed prediction accuracy for predicting the phenotypic value of the i^{th} validation population, $\hat{R}_{0p}$ = observed prediction accuracy for predicting the phenotypic value of k validation populations such as those from a k-fold validation

study, R_{0gij} = expected prediction accuracy for predicting the genotypic values of the j^{th} individual in the i^{th} validation population calculated as the square root of reliability from the GVCHAP output file for GBLUP and reliability based on Eq. 21,R_{0gi} = expected prediction accuracy for predicting the genotypic values of the i^{th} validation population, n_{0i} = number of individuals in the validation population, R_{0g} = expected prediction accuracy for predicting the genotypic values of k validation populations, R_{0p} = expected prediction accuracy for predicting the phenotypic values of k validation populations, and h^2 = heritability. The get-correlations.R program calculates Eqs. 22–25, but Eq. 26 needs to be calculated by the user using results in the GBLUP and GREML output files.

Haplotype and SNP Heritability Estimates for Graphic Visualization

GVCHAP calculates and saves SNP effects and heritability estimates, and saves haplotype heritability estimates in separate files for graphical visualization using SNPEVG2 (Wang et al., 2012) to identify SNPs and haplotype blocks with high heritability estimates. The graphical visualization of SNP effects and heritability estimates were available in GVCBLUP (Wang et al., 2014b), and the graphical visualization of haplotype heritability estimates is new to GVCHAP. A utility program (merge-mrk-hap.py) merges the two output files for haplotype heritabilities and SNP effects and heritabilities as a ".snpe" file that SNPEVG2 recognizes as an input file. Manhattan plots of and chromosome graphs of SNP and haplotype heritabilities can be produced by SNPEVG2, as shown by the example of **Figure 4**. The Manhattan plot and chromosome graphs of haplotype heritabilities may reveal chromosome regions with high heritability estimates that were not observed from the SNP plots and graphs.

Fixed Effects, and Genomic and Non-genetic Prediction of Phenotypic Values

GVCHAP calculates and saves fixed effects in a separate file and outputs the estimates of the fixed effects for all individuals. Two types of fixed effects can be included in the prediction model, classification variable for a fixed factor with a small number of levels such as male or female, and covariable for a fixed factor with many levels such as age. Selected SNPs can be fitted as fixed effects through declarations in the parameter file for column positions of those SNPs in the phenotype file, typically as covariables to minimize the use of degrees of freedom such that the residual degree of freedom is still sufficiently large.

Three potential applications involve the use of fixed effects. These include estimation of SNP contributions to prediction accuracy and genomic heritability (Tan et al., 2017), genomic prediction using fixed SNP effects for traits with large SNP effects (Spindel et al., 2015), and prediction of phenotypic values using genomic and non-genetic prediction (GNG) that combines genomic prediction and estimates of non-genetic fixed effects such as age and gender. For GNG, the current version of GVCHAP supports samples with one phenotypic observation per individual (N = n), and the general expression of the GNG values under this assumption is:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} + \hat{\mathbf{g}} \tag{27}$$

where $\hat{\mathbf{g}} = \hat{\mathbf{a}} + \hat{\mathbf{d}} + \hat{\mathbf{a}}_h$ for Model 1, $= \hat{\mathbf{a}} + \hat{\mathbf{a}}_h$ for Model 2, $= \hat{\mathbf{d}} + \hat{\mathbf{a}}_h$ for Model 3, $= \hat{\mathbf{a}}_h$ for Model 4, $= \hat{\mathbf{a}} + \hat{\mathbf{d}}$ for Model 5, $= \hat{\mathbf{a}}$ for Model 6, and $= \hat{\mathbf{d}}$ for Model 7.

Computing Speed of GVCHAP

The computing time of GVCHAP for calculating SNP additive and dominance genomic relationship matrices, haplotype additive relationship matrix, and each GREML iteration was evaluated using five samples and the Mesabi supercomputer at the University of Minnesota (**Table 2**). The results showed that the creation of the haplotype additive genomic matrix was the most time-consuming and memory-intensive computation, and could require 20 times as much time as creating the SNP additive and dominance genomic relationship matrices. For 15,098 individuals and 657,024 SNPs, the haplotype additive genomic relationship matrix required 22.11 h, whereas the SNP additive and dominance genomic relationship matrices together required 1.32 or 0.66 h per matrix. In contrast, each iteration required little time for any of the four samples, 19–82 s per iteration. To reduce the computing time for matrix creation, we developed two computing strategies, a strategy to remove repeated computations in validation studies and multivariate analysis, and a strategy of multi-node processing to eliminate the computing difficulty for creating the haplotype additive genomic relationship matrix. As the number of SNPs decreases, the computing speed for calculating the genomic matrices increases. For 7549 individuals with 82,128 SNPs, the calculation of the haplotype genomic additive relationship matrix required only 1.31 h, compared to 7.88 h when the number of SNPs became four times as many for the same number of individuals.

Two-Step Strategy to Remove Repeated Computing for Genomic Relationship Matrices

To remove repeated calculations to create genomic relationship matrices in validation studies and multivariate analysis, we designed a two-step strategy that saves genomic relationship matrices from the first run as binary files and loads the saved genomic relationship matrices for the remaining runs. For the example of a 10-fold validation, genomic relationship matrices are calculated only for the first of the 10 validation runs, and the remaining 9 validation runs load the stored genomic relationship matrices calculated and saved by the first run. This strategy virtually eliminates computing time for creating genomic relationship matrices when running GVCHAP. Our experience showed loading the saved genomic relationship matrices was nearly instantaneous and required only negligible computing times. With this strategy, days of computing time could be saved even for the smallest sample with 7549 individuals and 328,512 SNPs in our test runs (**Table 2**). For a study investigating many candidate haplotype models for multiple traits using k-fold

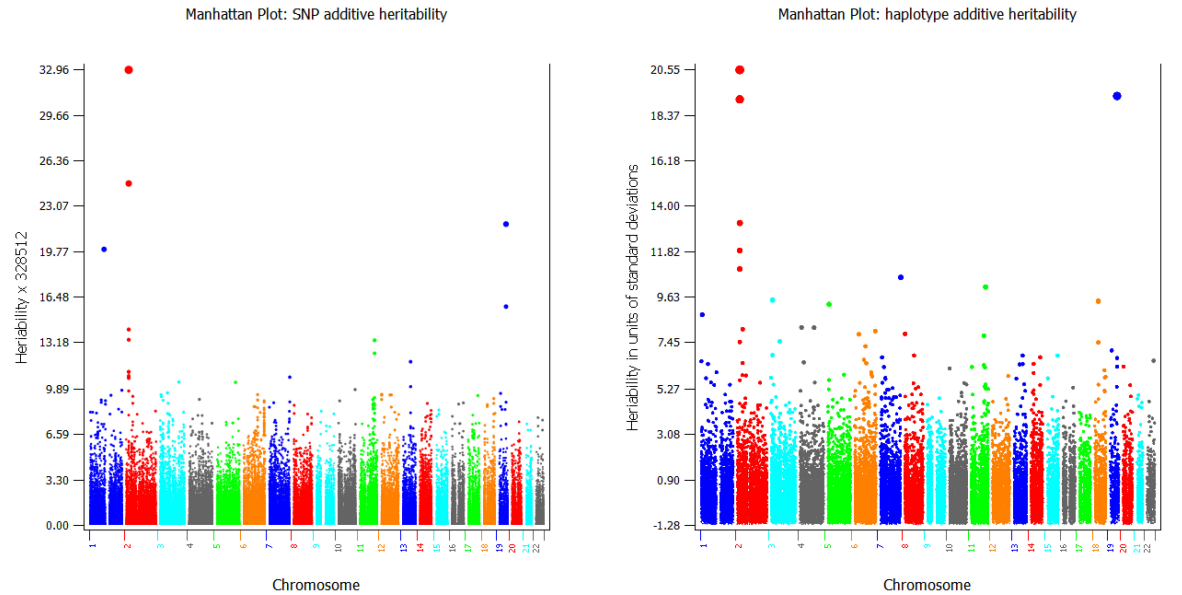

FIGURE 4 | Examples of graphical visualization of SNP heritabilities **(left)** and haplotype heritabilities **(right).**

TABLE 2 | Computing time of GVCHAP using the Mesabi supercomputer.

	GVCHAP using one node with 62 Gb memory				Multi-node processing (MNP)
Number of individuals (n)	7549	7549	15,098	15,098	37,745[a]
Number of SNPs (m)	82,128	328,512	328,512	657,024	328,512
SNP $\mathbf{A}_g$ and $\mathbf{D}_g$	0.12 h	0.71 h	0.70 h	1.32 h	2.67 h
Haplotype $\mathbf{A}_{gh}$	1.31 h	7.88 h	14.31 h	22.11 h	≈3 h[b]
Time per iteration	19 s[c]	37 s[c]	82 s[c]	62 s[c]	22–29 min[c] (one node)

[a]Haplotype genotypes of this dataset could not be processed by a single node with 62 Gb memory limit and was used as an example using the MNP method. The haplotype genotypes were divided into 75 files each with 200 blocks for most of the 75 files that were processed using 75 different nodes. [b]Each $\boldsymbol{W}^{i}_{\alpha h}\boldsymbol{W}^{i'}_{\alpha h}$ file required 0.831 ± 0.248 h to process, and the summation of the 75 $\boldsymbol{W}^{i}_{\alpha h}\boldsymbol{W}^{i'}_{\alpha h}$ matrices and the calculation of $\boldsymbol{A}_{gh} = \boldsymbol{W}_{\alpha h}\boldsymbol{W}'_{\alpha h}/k_{\alpha h}$ took about 1.66 h to complete. Adding 0.831 and 1.66 h indicates the calculation of $\boldsymbol{A}_{gh}$ approximately required 3 h. [c]The computing time per iteration is affected by the number of jobs running on the Mesabi supercomputer.

validations (e.g., 10-fold), days even months of computing time could be saved using this two-step strategy alone (**Table 3**).

Multi-Node Processing (MNP) for Genomic Relationship Matrices of Large Samples

The long computing time required by GVCHAP to create a haplotype additive genomic relationship matrix using a single node (**Table 2**) was due to the successive processing of haplotypes by chromosome. This successive processing results in the waiting of the next chromosome to be processed until the current chromosome finished its processing, which involves reading the haplotype genotype data, calculation of allele (haplotype) frequencies, and matrix multiplication and addition. Moreover, the memory limit of a single node sets a limit for the sample size and number of haplotypes that can be processed. To remove the limitation of using successive processing of the haplotype genotypes and a single node, we developed a multi-node processing (MNP) approach that divides the haplotype genotype files into s small files. One node processes each of the s files and all the s small files are processed simultaneously using different nodes. Although SNP additive and dominance genomic relationship matrices only required minor computing time relative to the time required by haplotypes (**Table 2**), the MNP approach is also implemented for SNP genomic relationships. The MNP approach is based on the result that the numerator matrix of each relationship matrix of Eqs. 14–16 can be expressed as a sum of the numerator matrices of all s small files, i.e.,

$$\mathbf{A}_g = \mathbf{W}_\alpha \mathbf{W}'_\alpha / k_\alpha = \left(\sum_{i=1}^{s} \mathbf{W}^i_\alpha \mathbf{W}^{i'}_\alpha \right) / k_\alpha \quad (28)$$

$$\mathbf{D}_g = \mathbf{W}_\delta \mathbf{W}'_\delta / k_\delta = \left(\sum_{i=1}^{s} \mathbf{W}^i_\delta \mathbf{W}^{i'}_\delta \right) / k_\delta \quad (29)$$

$$\mathbf{A}_{gh} = \mathbf{W}_{\alpha h} \mathbf{W}'_{\alpha h} / k_{\alpha h} = \left(\sum_{i=1}^{s} \mathbf{W}^i_{\alpha h} \mathbf{W}^{i'}_{\alpha h} \right) / k_{\alpha h} \quad (30)$$

where k_α, k_δ, and $k_{\alpha h}$ are defined by Eqs. 8–10. Different $\mathbf{W}^i_\alpha \mathbf{W}^{i'}_\alpha$ matrices in Eq. 28, $\mathbf{W}^i_\delta \mathbf{W}^i_\delta$ in Eq. 29, and $\mathbf{W}^i_{\alpha h} \mathbf{W}^i_{\alpha h}$ in Eq. 30 are processed by different nodes and are saved separately as a binary file. These matrices are then added together to create each genomic relationship matrix using Eqs. 28–30. Each genomic relationship matrix for all haplotypes or SNPs is saved as a binary file. The GVCHAP analysis loads these saved matrices with negligible computing time and the main computing time required for the GVCHAP analysis becomes that for iterations. The MNP results in the same savings in computing time due to processing of the genomic relationship matrices as the two-step strategy (**Table 3**) and can be considered as a multi-node version of the two-step strategy. The unique advantage of the MNP approach is the removal of hardware limitation of a single node and the ability to use multiple nodes simultaneously, making what is undoable using a single node doable using multiple nodes. For the example of 35,745 individuals and 328,512 SNPs, a single node with 62 Gb memory could not complete the haplotype genomic relationship matrix, and this matrix could be created in about 3 h using MNP (**Table 2**).

Heritability Tolerance to Reduce the Number of Iterations

With the removal of the computing bottleneck for processing haplotype genotypes by the two-step strategy and MNP method, the computing bottleneck becomes iterative solutions for GREML. The computing time per iteration increased from less than 1.5 min for 15,098 individuals to 22–29 min for 37,745 individuals (**Table 2**). The current version of GVCHAP uses shared memory parallel computing that can utilize all cores within a single node, but cannot use multiple nodes simultaneously. When AI-REML is used, GREML iterations generally converge fast, but EM-REML is used automatically when AI-REML fails and EM-REML may require many iterations to converge. To reduce the computing time required by EM-REML, GVCHAP implements two types of tolerance levels, heritability tolerance and variance component tolerance. Heritability tolerance can be substantially less stringent than variance component tolerance, e.g., 10^{-6} for heritability tolerance and 10^{-8} for variance component tolerance. Since heritability estimates are rarely reported with more than three decimal points, the 10^{-6} tolerance for heritability could significantly reduce the number of EM-REML iterations. With two types of tolerance levels, iteration stops when any of the two tolerance levels is reached.

TABLE 3 | Approximate saving of computing time of GVCHAP due to the two-step strategy or multi-node processing (MNP) for a 10-fold validation study relative to the use of a single node of the Mesabi supercomputer without the two-step strategy of MNP.

Number of individuals (n)	7549	15,098	15,098
Number of SNPs (m)	328,512	328,512	657,024
Saving in computing time			
10-fold validation model/trait	3.22 days	6 days	9 days
10-fold validation 10 models per trait	32.2 days	60 days	90 days

Challenge of Large Numbers of Individuals

The number of individuals is the limiting factor within each iteration because of matrix inversions and the need to store genomic relationship matrices for GREML. As the number of individuals increases, the sizes of the genomic relationship matrices and hence the required memory to store those matrices increase, and the computing time may increase rapidly. The choice of statistical model has a major impact on the memory requirement. Model 1 with haplotype additive effects and SNP additive and dominance effects requires the most memory, whereas Model 4 with haplotype additive effects requires the least memory among all models with haplotype effects. Therefore, Model 4 is the computationally competitive choice among Models 1–4 if Models 1–3 do not have substantially better prediction accuracy than Model 4. As shown in **Table 2**, increasing the number of individuals from 15,098 to 37,745 increased the computing time to 22–29 min from 82 s per iteration due to the larger genomic relationship matrices, 37,745 × 37,745 for 37,745 individuals versus 15,098 × 15,098 for 15,098 individuals. For this case, the matrix size increased 2.5 times and required 6.25 times as much memory to store each genomic relationship matrix. The ultimate computing solution to increase the GVCHAP capability for large numbers of individuals would be the use of distributed memory parallel computing, or parallel computing with massage passing interface (MPI) for using multiple nodes to reduce computing time and to use a large amount of memory. We have developed a MPI version of GVCBLUP (Wang et al., 2014a), and a MPI version of GVCHAP may be developed in the future.

CONCLUSION

The GVCHAP program provides a capability for GBLUP and GREML to identify optimal prediction models using haplotypes based on structural and functional genomic information for genomic selection. The utility programs for data preparation and summary analysis of GVCHAP results eliminate most of the tedious and time-consuming data work. The entire pipeline provides an efficient and versatile computing tool to investigate candidate haplotype models utilizing structural and functional genomic information for genomic selection.

SOFTWARE AVAILABILITY

The GVCHAP computing pipeline is available at http://animalgene.umn.edu.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YD conceived the study. DP was the author of the current version of GVCHAP and the utility programs. CW was the author of the initial version of GVCHAP. ZL was the author of the MNP program. CB and CT provided extensive evaluation that improved GVCHAP and the utility programs. ZL and DP evaluated computing time required by GVCHAP. DP, ZL, and YD prepared the user manual and prepared the manuscript.

FUNDING

This research was supported by grant no. 2018-67015-28128 from the USDA National Institute of Food and Agriculture, and by project MIN-16-124 of the Agricultural Experiment Station at the University of Minnesota. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The Minnesota Supercomputer Institute at the University of Minnesota provided supercomputer computing time and storage for this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00282/full#supplementary-material

REFERENCES

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M., Boscher, M., et al. (2012). Genomic selection in French dairy cattle. *Anim. Prod. Sci.* 52, 115–120.

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Calus, M., De Roos, A., and Veerkamp, R. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178, 553–561. doi: 10.1534/genetics.107.080838

Cuyabano, B. C., Su, G., and Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* 47, 1–11. doi: 10.1186/s12711-015-0143-3

Da, Y. (2015). Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genet.* 16:144. doi: 10.1186/s12863-015-0301-1

Da, Y. (2019). *Mixed Model Methods for Genetic Analysis. Classnotes for AnSc 8141*. Saint Paul: Department of Animal Science, University of Minnesota. Available online at: https://animalgene.umn.edu/sites/animalgene.umn.edu/files/ansc8141_2019.pdf (accessed March 17, 2020).

Da, Y., Wang, C., Wang, S., and Hu, G. (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* 9:e87666. doi: 10.1371/journal.pone.0087666

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53, 876–883. doi: 10.1139/G10-076

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* 49:54. doi: 10.1186/s12711-017-0329-y

Jan, H. U., Guan, M., Yao, M., Liu, W., Wei, D., Abbadi, A., et al. (2019). Genome-wide haplotype analysis improves trait predictions in *Brassica napus* hybrids. *Plant Sci.* 283, 157–164. doi: 10.1016/j.plantsci.2019.02.007

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3: Genes, Genomes, Genetics* 8, 1687–1699. doi: 10.1534/g3.117.300548

Jónás, D., Ducrocq, V., and Croiseau, P. (2017). The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle. *J. Dairy Sci.* 100, 2905–2908. doi: 10.3168/jds.2016-11798

Legarra, A., Robert-Granié, C., Manfredi, E., and Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics* 180, 611–618. doi: 10.1534/genetics.108.088575

Mulder, H. A., Calus, M. P., and Veerkamp, R. F. (2010). Research prediction of haplotypes for ungenotyped animals and its effect on marker-assisted breeding value estimation. *Genet. Sel. Evol.* 42:10. doi: 10.1186/1297-9686-42-10

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., et al. (2015). Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11:e1004982. doi: 10.1371/journal.pgen.1004982

Tan, C., Wu, Z., Ren, J., Huang, Z., Liu, D., He, X., et al. (2017). Genome-wide association study and accuracy of genomic prediction for teat number in Duroc pigs using genotyping-by-sequencing. *Genet. Sel. Evol.* 49:35. doi: 10.1186/s12711-017-0311-8

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

VanRaden, P. M., Sun, C., and O'Connell, J. R. (2015). Fast imputation using medium or low-coverage sequence data. *BMC Genet*. 16:82. doi: 10.1186/s12863-015-0243-7

Villumsen, T., Janss, L., and Lund, M. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126, 3–13. doi: 10.1111/j.1439-0388.2008.00747.x

Wang, C., and Da, Y. (2014). Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. *PLoS One* 9:e114484. doi: 10.1371/journal.pone.0114484

Wang, C., Prakapenka, D., Runesha, H. B., and Da, Y. (2014a). "Parallel computing for mixed model implementation of genomic prediction and variance component estimation of additive and dominance effects," in *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production* (Vancouver, BC).

Wang, C., Prakapenka, D., Wang, S., Pulugurta, S., Runesha, H. B., and Da, Y. (2014b). GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. *BMC Bioinformatics* 15:270. doi: 10.1186/1471-2105-15-270

Wang, S., Dvorkin, D., and Da, Y. (2012). SNPEVG: a graphical tool for GWAS graphing with mouse clicks. *BMC Bioinformatics* 13:319. doi: 10.1186/1471-2105-13-319

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.