GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

For most human complex diseases and traits, SNPs identified by genome-wide association studies (GWAS) explain only a small fraction of the heritability. Here we report a user-friendly software tool called *g*enome-wide *c*omplex *t*rait *a*nalysis (GCTA), which was developed based on a method we recently developed to address the "missing heritability" problem. GCTA estimates the variance explained by all the SNPs on a chromosome or on the whole genome for a complex trait rather than testing the association of any particular SNP to the trait. We introduce GCTA's five main functions: data management, estimation of the genetic relationships from SNPs, mixed linear model analysis of variance explained by the SNPs, estimation of the linkage disequilibrium structure, and GWAS simulation. We focus on the function of estimating the variance explained by all the SNPs on the X chromosome and testing the hypotheses of dosage compensation. The GCTA software is a versatile tool to estimate and partition complex trait variation with large GWAS data sets.

Despite the great success of genome-wide association studies (GWAS), which have identified hundreds of SNPs conferring the genetic variation of human complex diseases and traits,[1] the genetic architecture of human complex traits still remains largely unexplained. For most traits, the associated SNPs from GWAS only explain a small fraction of the heritability.[2,3] There has not been any consensus on the explanation of the "missing heritability." Possible explanations include a large number of common variants with small effects, rare variants with large effects, and DNA structural variation.[2,4] We recently proposed a method of estimating the total amount of phenotypic variance captured by all SNPs on the current generation of commercial genotyping arrays and estimated that ~45% of the phenotypic variance for human height can be explained by all common SNPs.[5] Thus, most of the heritability for height is hiding rather than missing because of many SNPs with small effects.[5,6] In contrast to single-SNP association analysis, the basic concept behind our method is to fit the effects of all the SNPs as random effects by a mixed linear model (MLM),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \text{ with } \mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{W}\mathbf{W}'\sigma^2_{\mathrm{u}} + \mathbf{I}\sigma^2_{\varepsilon}, \quad \text{(Equation 1)}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypes with n being the sample size, $\boldsymbol{\beta}$ is a vector of fixed effects such as sex, age, and/or one or more eigenvectors from principal component analysis (PCA), $\mathbf{u}$ is a vector of SNP effects with $\mathbf{u} \sim N(0, \mathbf{I}\sigma^2_{\mathrm{u}})$, $\mathbf{I}$ is an $n \times n$ identity matrix, and $\boldsymbol{\varepsilon}$ is a vector of residual effects with $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_{\varepsilon})$. $\mathbf{W}$ is a standardized genotype matrix with the ij^{th} element $w_{ij} = (x_{ij} - 2p_i)/\sqrt{2p_i(1-p_i)}$, where x_{ij} is the number of copies of the reference allele for the i^{th} SNP of the j^{th} individual and p_i is the frequency of the reference allele. If we define $\mathbf{A} = \mathbf{W}\mathbf{W}'/N$ and define σ^2_{g} as the variance explained by all the SNPs, i.e., $\sigma^2_{\mathrm{g}} = N\sigma^2_{\mathrm{u}}$, with N being the number of SNPs, then Equation 1 will be equivalent to:[7–9]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \text{ with } \mathbf{V} = \mathbf{A}\sigma^2_{\mathrm{g}} + \mathbf{I}\sigma^2_{\varepsilon}, \quad \text{(Equation 2)}$$

where $\mathbf{g}$ is an $n \times 1$ vector of the total genetic effects of the individuals with $\mathbf{g} \sim N(0, \mathbf{A}\sigma^2_{\mathrm{g}})$, and $\mathbf{A}$ is interpreted as the genetic relationship matrix (GRM) between individuals. We can therefore estimate σ^2_{g} by the restricted maximum likelihood (REML) approach,[10] relying on the GRM estimated from all the SNPs. Here we report a versatile tool called *g*enome-wide *c*omplex *t*rait *a*nalysis (GCTA), which implements the method of estimating variance explained by all SNPs, and extend the method to partition the genetic variance onto each of the chromosomes and also to estimate the variance explained by the X chromosome and test for dosage compensation in females. We developed GCTA in five function domains: data management, estimation of the GRM from a set of SNPs, estimation of the variance explained by all the SNPs on a single chromosome or the whole genome, estimation of linkage disequilibrium (LD) structure, and simulation.

Estimation of the Genetic Relationship from Genome-wide SNPs

One of the core functions of GCTA is to estimate the genetic relationships between individuals from the SNPs. From the definition above, the genetic relationship between individuals j and k can be estimated by the following equation:

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}. \quad \text{(Equation 3)}$$

We provide a function to iteratively exclude one individual of a pair whose relationship is greater than a specified cutoff value, e.g., 0.025, while retaining the maximum number of individuals in the data. For data collected from family or twin studies, we recommend that users estimate the genetic relationships with all of the autosomal SNPs and then use this option to exclude close relatives. The

[1]Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia; [2]Department of Food and Agricultural Systems, University of Melbourne, Parkville, Victoria 3010, Australia; [3]Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria 3086, Australia
*Correspondence: jian.yang@qimr.edu.au
DOI 10.1016/j.ajhg.2010.11.011.

reason for exclusion is that the objective of the analysis is to estimate genetic variation captured by all the SNPs, just as GWAS does for single SNPs. Including close relatives, such as parent-offspring pairs and siblings, would result in the estimate of genetic variance being driven by the phenotypic correlations for these pairs (just as in pedigree analysis), and this estimate could be a biased estimate of total genetic variance, for example because of common environmental effects. Even if the estimate is not biased, its interpretation is different from the estimate from "unrelated" individuals: a pedigree-based estimator captures the contribution from all causal variants (across the entire allele frequency spectrum), whereas our method captures the contribution from causal variants that are in LD with the genotyped SNPs.

As a by-product, we provide a function in GCTA to calculate the eigenvectors of the GRM, which is asymptotically equivalent to those from the PCA implemented in EIGENSTRAT[11] because the GRM (A_{jk}) defined in GCTA is approximately half of the covariance matrix (Ψ_{jk}) used in EIGENSTRAT. The only purpose of developing this function is to calculate eigenvectors and then include them in the model as covariates to capture variance due to population structure. More sophisticated analyses of the population structure can be found in programs such as EIGENSTRAT[11] and STRUCTURE.[12]

Estimation of the Variance Explained by Genome-wide SNPs by REML

The GRM estimated from the SNPs can be fitted subsequently in an MLM to estimate the variance explained by these SNPs via the REML method.[10] Previously, we included only one genetic factor in the model. Here we extend the model in a general form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \mathbf{g}_i + \boldsymbol{\varepsilon},$$

where $\mathbf{g}_i$ is a vector of random genetic effects, which could be the total genetic effects for the whole genome or for a single chromosome. In this model, the phenotypic variance (σ^2_{P}) is partitioned into the variance explained by each of the genetic factors and the residual variance,

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{A}_i \sigma^2_i + \mathbf{I}\sigma^2_{\varepsilon},$$

where σ^2_i is the variance of the i^{th} genetic factor with its corresponding GRM, $\mathbf{A}_i$.

In GCTA, we provide flexible options to specify different genetic models. For example:

(1) To estimate the variance explained by all autosomal SNPs, we can specify the model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon}$ with $\mathbf{V} = \mathbf{A}_{\mathrm{g}}\sigma^2_{\mathrm{g}} + \mathbf{I}\sigma^2_{\varepsilon}$, where $\mathbf{g}$ is an $n \times 1$ vector of the aggregate effects of all the autosomal SNPs for all of the individuals and $\mathbf{A}_{\mathrm{g}}$ is the GRM estimated from these SNPs. This model is the same as Equation 2.

(2) To estimate the variance of genotype-environment interaction effects (σ^2_{ge}), we can specify the model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{ge} + \boldsymbol{\varepsilon}$ with $\mathbf{V} = \mathbf{A}_{\mathrm{g}}\sigma^2_{\mathrm{g}} + \mathbf{A}_{\mathrm{ge}}\sigma^2_{\mathrm{ge}} + \mathbf{I}\sigma^2_{\varepsilon}$, where $\mathbf{ge}$ is a vector of genotype-environment interaction effects for all of the individuals with $\mathbf{A}_{\mathrm{ge}} = \mathbf{A}_{\mathrm{g}}$ for the pairs of individuals in the same environment and with $\mathbf{A}_{\mathrm{ge}} = \mathbf{0}$ for the pairs of individuals in different environments.

(3) To partition genetic variance onto each of the 22 autosomes, we can specify the model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{22} \mathbf{g}_i + \boldsymbol{\varepsilon}$ with $\mathbf{V} = \sum_{i=1}^{22} \mathbf{A}_i\sigma^2_i + \mathbf{I}\sigma^2_{\varepsilon}$, where $\mathbf{g}_i$ is a vector of genetic effects attributed to the i^{th} chromosome and $\mathbf{A}_i$ is the GRM estimated from the SNPs on the i^{th} chromosome.

GCTA implements the REML method via the average information (AI) algorithm.[13] In the REML iteration process, the estimates of variance components from the t^{th} iteration are updated by $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + (\mathbf{AI}^{(t)})^{-1}\partial L/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}^{(t)}}$, where $\boldsymbol{\theta}$ is a vector of variance components (σ^2_1, …, σ^2_r and σ^2_{ε}); L is the log likelihood function of the MLM (ignoring the constant), $L = -1/2(\log|\mathbf{V}| + \log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{P}\mathbf{y})$ with $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$; $\mathbf{AI}$ is the average of the observed and expected information matrices,

$$\mathbf{AI} = 1/2\begin{bmatrix} \mathbf{y}'\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{y} & \cdots & \mathbf{y}'\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{y} & \mathbf{y}'\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{P}\mathbf{y} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{y}'\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{y} & \cdots & \mathbf{y}'\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{y} & \mathbf{y}'\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{P}\mathbf{y} \\ \mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{y} & \cdots & \mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{y} & \mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{P}\mathbf{y} \end{bmatrix};$$

and $\partial L/\partial\boldsymbol{\theta}$ is a vector of first derivatives of the log likelihood function with respect to each variance component, $\partial L/\partial\boldsymbol{\theta} = -1/2\begin{bmatrix} tr(\mathbf{P}\mathbf{A}_1) - \mathbf{y}'\mathbf{P}\mathbf{A}_1\mathbf{P}\mathbf{y} \\ \vdots \\ tr(\mathbf{P}\mathbf{A}_r) - \mathbf{y}'\mathbf{P}\mathbf{A}_r\mathbf{P}\mathbf{y} \\ tr(\mathbf{P}) - \mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y} \end{bmatrix}$.[13] At the beginning of the iteration process, all of the components are initialized by an arbitrary value, i.e., $\sigma^{2(0)}_i = \sigma^2_{\mathrm{P}}/(r+1)$, which is subsequently updated by the expectation maximization (EM) algorithm, $\sigma^{2(1)}_i = [\sigma^{4(0)}_i\mathbf{y}'\mathbf{P}\mathbf{A}_i\mathbf{P}\mathbf{y} + \mathrm{tr}(\sigma^{2(0)}_i\mathbf{I} - \sigma^{4(0)}_i\mathbf{P}\mathbf{A}_i)]/n$. The EM algorithm is used as an initial step to determine the direction of the iteration updates because it is robust to poor starting values. After one EM iteration, GCTA switches to the AI algorithm for the remaining iterations until the iteration converges with the criteria of $L^{(t+1)} - L^{(t)} < 10^{-4}$, where $L^{(t)}$ is the log likelihood of the t^{th} iteration. In the iteration process, any component that escapes from the parameter space (i.e., its estimate is negative) will be set to $10^{-6} \times \sigma^2_{\mathrm{P}}$. If a component keeps escaping from the parameter space, it will be constrained at $10^{-6} \times \sigma^2_{\mathrm{P}}$.

From the REML analysis, GCTA has an option to provide the best linear unbiased prediction (BLUP) of the total genetic effect for all individuals. BLUP is widely used by plant and animal breeders to quantify the breeding value of individuals in artificial selection programs[14] and also by evolutionary geneticists.[15] Consider Equations 1 and 2, i.e., $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon}$. Because these two models are mathematically equivalent,[7–9] the BLUP of $\mathbf{g}$ can be transformed to the BLUP of $\mathbf{u}$ by $\widehat{\mathbf{u}} = \mathbf{W}'\mathbf{A}^{-1}\widehat{\mathbf{g}}/N$. Here the estimate of u_i corresponds to the coefficient w_{ij}, which is then rescaled for the original x_{ij} by

$\widehat{u}_i^* = \widehat{u}_i/\sqrt{2p_i(1-p_i)}$. We could obtain the BLUP of SNP effects in a discovery set by GCTA and predict genetic values of the individuals in a validation set ($\widehat{\mathbf{g}}_{\text{new}} = \mathbf{W}_{\text{new}}\widehat{\mathbf{u}}$). For example, GCTA could be used to predict SNP effects in a discovery set, and the SNP effects could be used in PLINK to predict whole-genome profiles via the scoring approach in a validation set. If the predictions are unbiased, then the regression slope of the observed phenotypes on the predicted genetic values is 1.[14] In that case, the genetic value calculated based on the BLUP of SNP effects is an unbiased predictor of the true genetic value in the validation set ($\mathbf{g}_{\text{new}}$), in the sense that $\mathrm{E}(\mathbf{g}_{\text{new}}|\widehat{\mathbf{g}}_{\text{new}}) = \widehat{\mathbf{g}}_{\text{new}}$.[16,17] Prediction analyses of human complex traits have demonstrated that many SNPs that do not pass the genome-wide significance level have substantial contribution to the prediction.[18,19] This option is therefore useful for the whole-genome prediction analysis with all of the SNPs, irrespective of their association p values.

Estimation of the Variance Explained by the SNPs on the X Chromosome

The method of estimating the genetic relationship from the X chromosome is different to that for the autosomal SNPs, because males have only one X chromosome. We modified Equation 3 for the X chromosome as:

$$A_{jk}^{\mathrm{M}} = \sum_{i=1}^{N} \frac{\left(x_{ij}^{\mathrm{M}} - p_i\right)\left(x_{ik}^{\mathrm{M}} - p_i\right)}{p_i(1-p_i)} \text{ for a male-male pair,}$$

$$A_{jk}^{\mathrm{F}} = \sum_{i=1}^{N} \frac{\left(x_{ij}^{\mathrm{F}} - 2p_i\right)\left(x_{ik}^{\mathrm{F}} - 2p_i\right)}{2p_i(1-p_i)} \text{ for a female-female pair, and}$$

$$A_{jk}^{\mathrm{MF}} = \sum_{i=1}^{N} \frac{\left(x_{ij}^{\mathrm{M}} - p_i\right)\left(x_{ik}^{\mathrm{F}} - 2p_i\right)}{\sqrt{2}p_i(1-p_i)} \text{ for a male-female pair,}$$

where x_{ij}^{M} and x_{ij}^{F} are the number of copies of the reference allele for an X chromosome SNP for a male and a female, respectively.

Assuming the male-female genetic correlation to be 1, the X-linked phenotypic covariance between a pair of individuals is:[20]

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{M}}, y_k^{\mathrm{M}}\right) = \mathrm{E}\left(A_{jk}^{\mathrm{M}}\right)\sigma^2_{\mathrm{X(M)}} \text{ for a male-male pair,}$$

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{F}}, y_k^{\mathrm{F}}\right) = \mathrm{E}\left(A_{jk}^{\mathrm{F}}\right)\sigma^2_{\mathrm{X(F)}} \text{ for a female-female pair, and}$$

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{M}}, y_k^{\mathrm{F}}\right) = \mathrm{E}\left(A_{jk}^{\mathrm{MF}}\right)\sigma_{\mathrm{X(M)}}\sigma_{\mathrm{X(F)}} \text{ for a male-female pair,}$$

where $\sigma^2_{\mathrm{X(M)}}$ and $\sigma^2_{\mathrm{X(F)}}$ are the genetic variance attributed to the X chromosome for males and females, respectively.

The relative values of $\sigma^2_{\mathrm{X(M)}}$ and $\sigma^2_{\mathrm{X(F)}}$ depend on the assumption made regarding dosage compensation for X chromosome genes. There are two alleles per locus in females, but only one in males. If we assume that each allele has a similar effect on the trait (i.e., no dosage compensation), the genetic variance on the X chromosome for females is twice that for males: i.e., $\sigma^2_{\mathrm{X}} = \sigma^2_{\mathrm{X(F)}} = 2\sigma^2_{\mathrm{X(M)}}$. Thus,

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{M}}, y_k^{\mathrm{M}}\right) = 1/2\mathrm{E}\left(A_{jk}^{\mathrm{M}}\right)\sigma^2_{\mathrm{X}} \text{ for a male-male pair,}$$

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{F}}, y_k^{\mathrm{F}}\right) = \mathrm{E}\left(A_{jk}^{\mathrm{F}}\right)\sigma^2_{\mathrm{X}} \text{ for a female-female pair, and}$$

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{M}}, y_k^{\mathrm{F}}\right) = 1/\sqrt{2}\mathrm{E}\left(A_{jk}^{\mathrm{MF}}\right)\sigma^2_{\mathrm{X}} \text{ for a male-female pair.}$$

This can be implemented by redefining GRM for the X chromosome as $\mathbf{A}_{\mathrm{X}}^{\mathrm{ND}} = 1/2\mathbf{A}_{\mathrm{X}}$ for male-male pairs, $\mathbf{A}_{\mathrm{X}}^{\mathrm{ND}} = \mathbf{A}_{\mathrm{X}}$ for female-female pairs, and $\mathbf{A}_{\mathrm{X}}^{\mathrm{ND}} = 1/\sqrt{2}\mathbf{A}_{\mathrm{X}}$ for male-female pairs. If we assume that each allele in females has only half the effect of an allele in males (i.e., full dosage compensation), the X-linked genetic variance for females is half that for males: i.e., $\sigma^2_{\mathrm{X}} = \sigma^2_{\mathrm{X(F)}} = 1/2\sigma^2_{\mathrm{X(M)}}$. Thus,

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{M}}, y_k^{\mathrm{M}}\right) = 2\mathrm{E}\left(A_{jk}^{\mathrm{M}}\right)\sigma^2_{\mathrm{X}} \text{ for a male-male pair,}$$

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{F}}, y_k^{\mathrm{F}}\right) = \mathrm{E}\left(A_{jk}^{\mathrm{F}}\right)\sigma^2_{\mathrm{X}} \text{ for a female-female pair, and}$$

$$\mathrm{cov}_{\mathrm{X}}\left(y_j^{\mathrm{M}}, y_k^{\mathrm{F}}\right) = \sqrt{2}\mathrm{E}\left(A_{jk}^{\mathrm{MF}}\right)\sigma^2_{\mathrm{X}} \text{ for a male-female pair.}$$

Therefore, the raw $\mathbf{A}_{\mathrm{X}}$ matrix should be parameterized as $\mathbf{A}_{\mathrm{X}}^{\mathrm{FD}} = 2\mathbf{A}_{\mathrm{X}}$ for male-male pairs, $\mathbf{A}_{\mathrm{X}}^{\mathrm{FD}} = \mathbf{A}_{\mathrm{X}}$ for female-female pairs, and $\mathbf{A}_{\mathrm{X}}^{\mathrm{ND}} = \sqrt{2}\mathbf{A}_{\mathrm{X}}$ for male-female pairs. The third possibility is to assume equal genetic variance on the X chromosome for males and females, i.e., $\sigma^2_{\mathrm{X}} = \sigma^2_{\mathrm{X(F)}} = \sigma^2_{\mathrm{X(M)}}$, in which case the $\mathbf{A}_{\mathrm{X}}$ matrix is not redefined at all.

We can estimate σ^2_{X} by fitting the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}_{\mathrm{X}} + \mathbf{g} + \boldsymbol{\varepsilon}$, where $\mathbf{g}_{\mathrm{X}}$ is a vector of genetic effects attributable to the X chromosome, with $\mathrm{var}(\mathbf{g}_{\mathrm{X}}) = \mathbf{A}_{\mathrm{X}}^{\mathrm{ND}}\sigma^2_{\mathrm{X}}$ assuming no dosage compensation, $\mathrm{var}(\mathbf{g}_{\mathrm{X}}) = \mathbf{A}_{\mathrm{X}}^{\mathrm{FD}}\sigma^2_{\mathrm{X}}$ assuming full dosage compensation, and $\mathrm{var}(\mathbf{g}_{\mathrm{X}}) = \mathbf{A}_{\mathrm{X}}\sigma^2_{\mathrm{X}}$ assuming equal X-linked genetic variance for males and females. Test of dosage compensation can be achieved by comparing the likelihoods of model fitting under the three assumptions.

Estimation of the Variance Explained by Genome-wide SNPs for a Case-Control Study

The methodology described above is also applicable for case-control data, for which the estimate of variance explained by the SNPs corresponds to variation on the observed 0–1 scale. Under the assumption of a threshold-liability model for a disease, i.e., disease liability on the underlying scale follows standard normal distribution,[21] the estimate of variance explained by the SNPs on the

observed 0–1 scale can be transformed to that on the unobserved continuous liability scale by a linear transformation.[22] The relationship between additive genetic variance on the observed 0–1 and unobserved liability scales was proposed more than a half century ago,[23,24] and we recently extended this transformation to account for ascertainment bias in a case-control study, i.e., a much higher proportion of cases in the sample than in the general population (unpublished data). We provide options in GCTA to analyze a binary trait and to transform the estimate on the 0–1 scale to that on the liability scale with an adjustment for ascertainment bias. There is an important caveat in applying the methods described herein to case-control data. Any batch, plate, or other technical artifact that causes allele frequencies between case and control on average to be more different than that under the null hypothesis stating that the samples come from the same population will contribute to the estimation of spurious genetic variation, because cases will appear to be more related to other cases than to controls. Therefore, stringent quality control is essential when applying GCTA to case-control data. Quantitative traits are less likely to suffer from technical genotyping artifacts because they will generally not lead to spurious association between continuous phenotypes and genotypes.

Estimation of the Inbreeding Coefficient from Genome-wide SNPs

Apart from estimating the genetic relatedness between individuals, GCTA also has a function to estimate the inbreeding coefficient (F) from SNP data, i.e., the relationship between haplotypes within an individual. Two estimates have been used: one based on the variance of additive genetic values (diagonal of the SNP-derived GRM) and the other based on SNP homozygosity (implemented in PLINK).[25] Let $(1 - p_i)^2 + p_i(1 - p_i)F$, $2p_i(1 - p_i)(1 - F)$, and $p_i^2 + p_i(1 - p_i)F$ be the frequencies of the three genotypes of a SNP i and let $h_i = 2p_i(1 - p_i)$. The estimate based on the variance of additive genotype values is

$$\widehat{F}_i^{\mathrm{I}} = [x_i - E(x_i)]^2/h_i - 1 = (x_i - 2p_i)^2/h_i - 1 \text{ and } \mathrm{var}\left(\widehat{F}_i^{\mathrm{I}} \mid F\right) = (1 - h_i)/h_i + 7(1 - 2h_i)F/h_i - F^2,$$

where x_i is the number of copies of the reference allele for the i^{th} SNP. This is a special case of Equation 3 for a single SNP when $j = k$. The estimate based upon excess homozygosity is

$$\widehat{F}_i^{\mathrm{II}} = [\mathrm{O}(\#\mathrm{hom}) - E(\#\mathrm{hom})]/[1 - E(\#\mathrm{hom})] = 1 - x_i(2 - x_i)/h_i \text{ and } \mathrm{var}\left(\widehat{F}_i^{\mathrm{II}} \mid F\right) = (1 - h_i)/h_i - (1 - 2h_i)F/h_i - F^2,$$

where O(# hom) and E(# hom) are the observed and expected number of homozygous genotypes in the sample, respectively. Both estimators are unbiased estimates of F in the sense that $E(\widehat{F}_i^{\mathrm{I}}|F) = E(\widehat{F}_i^{\mathrm{II}}|F) = F$, but their sampling variances are dependent on allele frequency, i.e., $\mathrm{var}(\widehat{F}_i^{\mathrm{I}}) = \mathrm{var}(\widehat{F}_i^{\mathrm{II}}) = (1 - h_i) / h_i$ if $F = 0$. In addition, the covariance between the two estimators is $(3h_i - 1) / h_i + (1 - 2h_i)F / h_i - F^2$, so that the sampling covariance between the estimators is $(3h_i - 1) / h_i$ and the sampling correlation is $(3h_i - 1) / (1 - h_i)$ when $F = 0$. We proposed an estimator based upon the correlation between uniting gametes:[5]

$$\widehat{F}_i^{\mathrm{III}} = \left[x_i^2 - (1 + 2p_i)x_i + 2p_i^2\right]/h_i \text{ and } \mathrm{var}\left(\widehat{F}_i^{\mathrm{III}} \mid F\right) = 1 + 2(1 - 2h_i)F/h_i - F^2.$$

$\widehat{F}_i^{\mathrm{III}}$ is also an unbiased estimator of F in the sense that $E(\widehat{F}_i^{\mathrm{III}}|F) = F$. If $F = 0$, $\mathrm{var}(\widehat{F}_i^{\mathrm{III}}) = 1$ regardless of allele frequency, which is smaller than the sampling variance of $\widehat{F}_i^{\mathrm{I}}$ and $\widehat{F}_i^{\mathrm{II}}$, i.e., $1 \leq (1 - h_i) / h_i$. When $0 < F < 1/3$, $\widehat{F}_i^{\mathrm{III}}$ also has a smaller variance than $\widehat{F}_i^{\mathrm{I}}$ and $\widehat{F}_i^{\mathrm{II}}$. In GCTA, we use $1 + \widehat{F}_i^{\mathrm{III}}$ rather than $1 + \widehat{F}_i^{\mathrm{I}}$ to calculate the diagonal of the GRM. For multiple SNPs, we average the estimates over all of the SNPs, i.e., $\widehat{F} = 1/N\sum_{i=1}^{N}\widehat{F}_i$.

Estimating LD Structure

In a standard GWAS, particularly with a large sample size, the mean (λ_{mean}) or median ($\lambda_{\mathrm{median}}$) of the test statistics for single-SNP associations often deviates from its expected value under the null hypothesis of no association between any SNP and the phenotype, which is usually interpreted as the effect due to population stratification and/or cryptic relatedness.[11,26,27] An alternative explanation is that polygenic variation causes the observed inflated test statistic.[18] To predict the genomic inflation factors, λ_{mean} and $\lambda_{\mathrm{median}}$, from polygenic parameters such as the total amount of variance that is explained by all SNPs, we need to quantify the LD structure between SNPs and putative causal variants (unpublished data). GCTA provides a function to search for all the SNPs in LD with the "causal variants" (mimicked by a set of SNPs chosen by the user). Given a causal variant, we use simple regression to test for SNPs in LD with the causal variant within d Mb distance in either direction. PLINK has an option ("show targets") to select SNPs in LD with a set of target SNPs with LD r^2 larger than a user-specified cutoff value. This function is very useful to distinguish independent association signals but less suited to predict λ_{mean} and $\lambda_{\mathrm{median}}$, because the test statistics of the SNPs in modest LD with causal variants (SNPs at Mb distance with low r^2) will also be inflated to a certain extent, and these test statistics will contribute to the genomic inflation factors.

GWAS Simulation

We provided a function to simulate GWAS data based on the observed genotype data. For a quantitative trait, the phenotypes are simulated by the simple additive genetic model $\mathbf{y} = \mathbf{Wu} + \boldsymbol{\varepsilon}$, where the notation is the same as above. Given a set of SNPs assigned as causal variants, the effects of the causal variants are generated from a standard normal distribution, and the residual effects are generated from a normal distribution with mean of 0 and variance of $\sigma_g^2(1/h^2 - 1)$,

where σ_g^2 is the empirical variance of **Wu** and h^2 is the user specified heritability. For a case-control study, assuming a threshold-liability model, disease liabilities are simulated in the same way as that for the phenotypes of a quantitative trait. Any individual with disease liability exceeding a certain threshold T is assigned to be a case and a control otherwise, where T is the threshold of normal distribution truncating the proportion of K (disease prevalence). The only purpose of this function is to do a simple simulation based on the observed genotype data. More complicated simulation can be performed with programs such as ms,[28] GENOME,[29] FREGENE,[30] and HAPGEN.[31]

Data Management

We chose the PLINK[25] compact binary file format (*.bed, *.bim, and *.fam) as the input data format for GCTA because of its popularity in the genetics community and its efficiency of data storage. For the imputed dosage data, we use the output files of the imputation program MACH[32] (*.mldose.gz and *.mlinfo.gz) as the inputs for GCTA. For the convenience of analysis, we provide options to extract a subset of individuals and/or SNPs and to filter SNPs based on certain criteria, such as chromosome position, minor allele frequency (MAF), and imputation R^2 (for the imputed data). However, we do not provide functions for a thorough quality control (QC) of the data, such as Hardy-Weinberg equilibrium test and missingness, because these functions have been well developed in many other genetic analysis packages, e.g., PLINK, GenABEL,[33] and SNPTEST.[34] We assume that the data have been cleaned by a standard QC process before entering into GCTA.

Estimating Total Heritability

The method implemented in GCTA is to estimate the variance explained by chromosome- or genome-wide SNPs rather than the trait heritability. Estimating the heritability (i.e., variance explained by all the causal variants), however, relies on the genetic relationship at causal variants that is predicted with error by the genetic relationship derived from the SNPs as a result of imperfect tagging. We have previously established that the prediction error is $c + 1 / N$, with c depending on the distribution of the MAF of causal variants. We therefore developed a method based on simple regression to correct for the prediction error by

$$A_{jk}^{*} = \begin{cases} 1 + \beta(A_{jj} - 1), & j = k \\ \beta A_{jk}, & j \neq k, \end{cases}$$

where $\beta = 1 - (c + 1/N)/\mathrm{var}(A_{jk})$. The estimate of variance explained by all of the SNPs after such adjustment is an unbiased estimate of heritability only if the assumption about the MAF distribution of causal variants is correct.

Efficiency of GCTA Computing Algorithm

GCTA implements the REML method based on the variance-covariance matrix **V** and the projection matrix **P**. In some of the mixed model analysis packages, such as ASREML,[35] to avoid the inversion of the $n \times n$ **V** matrix, people usually use Gaussian elimination of the mixed model equations (MME) to obtain the **AI** matrix based on sparse matrix techniques. The SNP-derived GRM matrix, however, is typically dense, so the sparse matrix technique will bring an extra cost of memory and CPU time. Moreover, the dimension of MME depends on the number of random effects in the model, whereas the **V** matrix does not. For example, when fitting the 22 chromosomes simultaneously in the model, the dimension of MME is $22n \times 22n$ (ignoring the fixed effects), whereas the dimension of **V** matrix is still $n \times n$. We compared the computational efficiency of GCTA and ASREML. When the sample size is small, e.g., $n < 3000$, both GCTA and ASREML take a few minutes to run. When the sample size is large, e.g., $n > 10{,}000$, especially when fitting multiple GRMs, it takes days for ASREML to finish the analysis, whereas GCTA needs only a few hours.

System Requirements

We have released executable versions of GCTA for the three major operating systems: MS Windows, Linux/Unix, and Mac OS. We have also released the source codes so that users can compile them for some specific platforms. GCTA requires a large amount of memory when calculating the GRM or performing an REML analysis with multiple genetic components. For example, it requires ~4.8 GB memory to calculate the GRM for a data set with 3925 individuals genotyped by 294,831 SNPs, and it takes ~4 CPU hours (AMD Opteron 2.8 GHz) to finish the computation. We therefore recommend using the 64-bit version of GCTA for large memory support.

Nonadditive Genetic Variance

The analysis approach we have adapted is a logical extension of estimation methods based on pedigrees. It allows estimation of additive genetic variation that is captured by SNP arrays and is therefore informative with respect to the genetic architecture of complex traits. The estimate of variance captured by all of the SNPs obtained in GCTA is directly comparable to the heritability estimated from pedigree analysis in family and twin studies, as well as the variance explained by GWAS hits, so that missing and hiding heritability can be quantified.[5] Other sources of genetic variations such as dominance, gene-gene interaction, and gene-environment interaction are also important for complex trait variation but are less relevant to the "missing heritability" problem if the total heritability refers to the narrow-sense heritability, i.e., the proportion of phenotypic variance due to additive genetic variance. The current version of GCTA only provides functions to estimate and partition the variances of additive and additive-environment interaction effects. It is technically feasible to extend the analysis to include dominance and/or gene-gene interaction effects in the future. However, the power to detect the high-order genetic

variation will be limited, i.e., the sampling variance of estimated variance components will be very large. Future developments will also include options to do multivariate analyses, to read genotype or imputed probability data in different formats, and to implement other applications of whole-genome or chromosome segment approaches.

In summary, we have developed a versatile tool to estimate genetic relationships from genome-wide SNPs that can subsequently be used to estimate variance explained by SNPs via a mixed model approach. We provide flexible options to specify different genetic models to partition genetic variance onto each of the chromosomes. We developed methods to estimate genetic relationships from the SNPs on the X chromosome and to test the hypotheses of dosage compensation. GCTA is not limited to the analysis of data on human complex traits, but in this report we only use examples and specifications (e.g., the number of autosomes) for humans.

Acknowledgments

We thank Bruce Weir for discussions on the sampling variance of estimators of inbreeding coefficients. We thank Allan McRae and David Duffy for discussions and Anna Vinkhuyzen for software testing. We acknowledge funding from the Australian National Health and Medical Research Council (grants 389892 and 613672) and the Australian Research Council (grants DP0770096 and DP1093900).

Received: August 30, 2010
Revised: November 23, 2010
Accepted: November 29, 2010
Published online: December 16, 2010

Web Resources

The URLs for data presented herein are as follows:

Genome-wide Complex Trait Analysis (GCTA), http://gump.qimr.edu.au/gcta
MACH 1.0: A Markov Chain-based haplotyper, http://www.sph.umich.edu/csg/yli/mach
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink

References

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
3. Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18–21.
4. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
6. Gibson, G. (2010). Hints of hidden heritability in GWAS. Nat. Genet. *42*, 558–560.
7. Hayes, B.J., Visscher, P.M., and Goddard, M.E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. *91*, 47–60.
8. Strandén, I., and Garrick, D.J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. *92*, 2971–2975.
9. VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. *91*, 4414–4423.
10. Patterson, H.D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika *58*, 545–554.
11. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
12. Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.
13. Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average information REML: An efficient algorithm for variance parameters estimation in linear mixed models. Biometrics *51*, 1440–1450.
14. Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics *31*, 423–447.
15. Kruuk, L.E. (2004). Estimating genetic parameters in natural populations using the "animal model". Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 873–890.
16. Goddard, M.E., Wray, N.R., Verbyla, K., and Visscher, P.M. (2009). Estimating effects and making predictions from genome-wide marker data. Stat. Sci. *24*, 517–529.
17. de Los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: The promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.
18. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
19. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
20. Kent, J.W., Jr., Dyer, T.D., and Blangero, J. (2005). Estimating the additive genetic effect of the X chromosome. Genet. Epidemiol. *29*, 377–388.
21. Lynch, M., and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits (Sunderland, MA: Sinauer Associates).
22. Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. *29*, 51–76.

23. Dempster, E.R., and Lerner, I.M. (1950). Heritability of threshold characters. Genetics *35*, 212–236.
24. Robertson, A., and Lerner, I.M. (1949). The heritability of all-or-none traits; viability of poultry. Genetics *34*, 395–411.
25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
26. Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nat. Genet. *37*, 868–872.
27. Cardon, L.R., and Palmer, L.J. (2003). Population stratification and spurious allelic association. Lancet *361*, 598–604.
28. Hudson, R.R. (1990). Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology *7*, 1–44.
29. Liang, L., Zöllner, S., and Abecasis, G.R. (2007). GENOME: A rapid coalescent-based whole genome simulator. Bioinformatics *23*, 1565–1567.
30. Hoggart, C.J., Chadeau-Hyam, M., Clark, T.G., Lampariello, R., Whittaker, J.C., De Iorio, M., and Balding, D.J. (2007). Sequence-level population simulations over large genomic regions. Genetics *177*, 1725–1731.
31. Spencer, C.C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. *5*, e1000477.
32. Li, Y., and Abecasis, G.R. (2006). Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. Am. J. Hum. Genet. *S79*, 2290.
33. Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: An R library for genome-wide association analysis. Bioinformatics *23*, 1294–1296.
34. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
35. Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson, R. (2006). ASReml User Guide Release 2.0 (Hemel Hempstead, UK: VSN International).