GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher

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 genome-wide complex trait analysis (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, 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. 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. 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. Thus, most of the heritability for height is hiding rather than missing because of many SNPs with small effects. 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),

$$y = X \beta + Wu + \varepsilon$$ with \(var(y) = V = WW'\sigma^2_g + I_3\sigma^2_e.$$  

(Equation 1)

where \(y\) is an \(n \times 1\) vector of phenotypes with \(n\) being the sample size, \(\beta\) is a vector of fixed effects such as sex, age, and/or one or more eigenvectors from principal component analysis (PCA), \(u\) is a vector of SNP effects with \(u \sim N(0, I_3\sigma^2_g)\), \(I\) is an \(n \times n\) identity matrix, and \(\varepsilon\) is a vector of residual effects with \(\varepsilon \sim N(0, I_3\sigma^2_e)\). \(W\) is a standardized genotype matrix with the \(i^{th}\) 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 \(j^{th}\) SNP of the \(i^{th}\) individual and \(p_i\) is the frequency of the reference allele. If we define \(A = WW'N\) and define \(\sigma^2_g\) as the variance explained by all the SNPs, i.e., \(\sigma^2_g = N\sigma^2_gN\), with \(N\) being the number of SNPs, then Equation 1 will be equivalent to:

$$y = X \beta + g + \varepsilon$$ with \(V = A\sigma^2_g + I_3\sigma^2_e.$$  

(Equation 2)

where \(g\) is an \(n \times 1\) vector of the total genetic effects of the individuals with \(g \sim N(0, A\sigma^2_g)\), and \(A\) is interpreted as the genetic relationship matrix (GRM) between individuals. We can therefore estimate \(\sigma^2_g\) by the restricted maximum likelihood (REML) approach, relying on the GRM estimated from all the SNPs. Here we report a versatile tool called genome-wide complex trait analysis (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:

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

(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

1Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia; 2Department of Food and Agricultural Systems, University of Melbourne, Parkville, Victoria 3010, Australia; 3Biosciences Research Division, Department of Primary Industries, Bundool, Victoria 3086, Australia
*Correspondence: jian.yang@qimr.edu.au
DOI 10.1016/j.ajhg.2010.11.011. ©2011 by The American Society of Human Genetics. All rights reserved.
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, 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 because the GRM (Ajk) defined in GCTA is approximately half of the covariance matrix (Wjk) 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 and STRUCTURE.

### Estimation of the Variance Explained by Genomewide 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. Previously, we included only one genetic factor in the model. Here we extend the model in a general form as

\[
y = X\beta + \sum_{i=1}^{r} g_i + \epsilon,
\]

where \(g_i\) 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 (\(\sigma^2_p\)) is partitioned into the variance explained by each of the genetic factors and the residual variance,

\[
V = \sum_{i=1}^{r} A_i \sigma^2_{g_i} + \text{Residual},
\]

where \(\sigma^2_{g_i}\) is the variance of the \(i^{th}\) genetic factor with its corresponding GRM, \(A_i\).

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 \(y = X\beta + g + \epsilon\) with \(V = A_g \sigma^2_g + I \sigma^2_q\), where \(g\) is an \(n \times 1\) vector of the aggregate effects of all the autosomal SNPs for all of the individuals and \(A_g\) 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 (\(\sigma^2_{ge}\)), we can specify the model as \(y = X\beta + g + ge + \epsilon\) with \(V = A_g \sigma^2_g + A_{ge} \sigma^2_{ge} + I \sigma^2_q\), where \(ge\) is a vector of genotype-environment interaction effects for all of the individuals with \(A_{ge} = A_g\) for the pairs of individuals in the same environment and with \(A_{ge} = 0\) for the pairs of individuals in different environments.

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

GCTA implements the REML method via the average information (AI) algorithm. In the REML iteration process, the estimates of variance components from the \(t^{th}\) iteration are updated by \(\theta^{(t+1)} = \theta^{(t)} + (A_I^{(t)})^{-1} \partial L / \partial \theta\), where \(\theta\) is a vector of variance components (\(\sigma^2_1, \ldots, \sigma^2_q\)); \(L\) is the log likelihood function of the MLM (ignoring the constant), \(L = -1/2 \log |V| + \log |X'V^{-1}X + yPy|\) with \(P = V^{-1} X'V^{-1} X\). At the beginning of the iteration process, all of the components are initialized by an arbitrary value, i.e., \(\sigma^2_i = \sigma^2_q / (r + 1)\), which is subsequently updated by the expectation maximization (EM) algorithm, \(\sigma^2_i = (\bar{y}^2 PA_i PA_i y - yPA_i y + \bar{y}PA_i y) / n\). 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 \(L^{(t+1)} - L^{(t)} < 10^{-8}\), where \(L^{(t)}\) is the log likelihood of the \(t^{th}\) iteration. If the iteration process, any component that escapes from the parameter space (i.e., its estimate is negative) will be set to \(10^{-6} \times \sigma^2_q\). If a component keeps escaping from the parameter space, it will be constrained at \(10^{-6} \times \sigma^2_q\).

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 and also by evolutionary geneticists. Consider Equations 1 and 2, i.e., \(y = X\beta + Wu + \epsilon\) and \(y = X\beta + g + \epsilon\). Because these two models are mathematically equivalent, the BLUP of \(g\) can be transformed to the BLUP of \(u\) by \(u = W A^{-1} g / N\).

Here the estimate of \(u_i\) corresponds to the coefficient \(w_{ii}\), which is then rescaled for the original \(x_{0i}\) by
The raw $A_X$ matrix should be parameterized as $A_X^{ND} = 2A_X$ for male-male pairs, $A_X^{ND} = A_X$ for female-female pairs, and $A_X^{ND} = 1/\sqrt{2}A_X$ for female-male pairs. The third possibility is to assume equal genetic variance on the X chromosome for males and females, i.e., $\sigma_X^2 = \sigma_X^{ND} = \sigma_X^{MF}$. In which case the $A_X$ matrix is not redefined at all.

We can estimate $\sigma_X^2$ by fitting the model $y = X\beta + g_X + \mathbf{u} + \epsilon$, where $g_X$ is a vector of genetic effects attributable to the X chromosome, with $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ assuming no dosage compensation, $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ assuming full dosage compensation, and $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ 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, the estimate of variance explained by the SNPs on the 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: $\sigma_X^2 = \sigma_X^{ND} = 2\sigma_X^{MF}$. Thus,

$$\text{cov}_X(y_i^M, y_j^M) = 2E(A_{ik}^M)\sigma_X^2$$

for a male-male pair.

$$\text{cov}_X(y_i^F, y_j^F) = E(A_{ik}^F)\sigma_X^2$$

for a female-female pair, and

$$\text{cov}_X(y_i^M, y_j^F) = \sqrt{2}E(A_{ik}^{MF})\sigma_X^2$$

for a male-female pair.

Therefore, the raw $A_X$ matrix should be parameterized as $A_X^{ND} = 2A_X$ for male-male pairs, $A_X^{ND} = A_X$ for female-female pairs, and $A_X^{ND} = 1/\sqrt{2}A_X$ for female-male pairs. The third possibility is to assume equal genetic variance on the X chromosome for males and females, i.e., $\sigma_X^2 = \sigma_X^{ND} = \sigma_X^{MF}$. In which case the $A_X$ matrix is not redefined at all.

We can estimate $\sigma_X^2$ by fitting the model $y = X\beta + g_X + \mathbf{u} + \epsilon$, where $g_X$ is a vector of genetic effects attributable to the X chromosome, with $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ assuming no dosage compensation, $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ assuming full dosage compensation, and $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ 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 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:

$$A_X^M = \sum_{i=1}^N \frac{x_{ij}^M - p_i}{p_i(1-p_i)}$$

for a male-male pair,

$$A_X^F = \sum_{i=1}^N \frac{x_{ij}^F - 2p_i}{2p_i(1-p_i)}$$

for a female-female pair, and

$$A_X^{MF} = \sum_{i=1}^N \frac{x_{ij}^M - p_i}{\sqrt{2p_i(1-p_i)}}$$

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:

$$\text{cov}_X(y_i^M, y_j^M) = E(A_{ik}^M)\sigma_X^{MF}$$

for a male-male pair,

$$\text{cov}_X(y_i^F, y_j^F) = E(A_{ik}^F)\sigma_X^{MF}$$

for a female-female pair, and

$$\text{cov}_X(y_i^M, y_j^F) = E(A_{ik}^{MF})\sigma_X^{MF}$$

for a male-female pair,

where $\sigma_X^{MF}$ is the genetic variance attributable to the X chromosome for males and females, respectively.

The relative values of $\sigma_X^{MF}$ and $\sigma_X^{MF}$ 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: $\sigma_X^2 = \sigma_X^{MF} = 2\sigma_X^{MF}$. Thus,

$$\text{cov}_X(y_i^M, y_j^M) = 1/2E(A_{ik}^M)\sigma_X^2$$

for a male-male pair,

$$\text{cov}_X(y_i^F, y_j^F) = E(A_{ik}^F)\sigma_X^2$$

for a female-female pair, and

$$\text{cov}_X(y_i^M, y_j^F) = 1/\sqrt{2}E(A_{ik}^{MF})\sigma_X^2$$

for a male-female pair.

This can be implemented by redefining GRM for the X chromosome as $A_X^{ND} = 1/2A_X$ for male-male pairs, $A_X^{ND} = A_X$ for female-female pairs, and $A_X^{ND} = 1/\sqrt{2}A_X$ 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., $\sigma_X^2 = \sigma_X^{MF} = 1/2\sigma_X^{MF}$. Thus,

$$\text{cov}_X(y_i^M, y_j^M) = 2E(A_{ik}^M)\sigma_X^2$$

for a male-male pair,

$$\text{cov}_X(y_i^F, y_j^F) = E(A_{ik}^F)\sigma_X^2$$

for a female-female pair, and

$$\text{cov}_X(y_i^M, y_j^F) = \sqrt{2}E(A_{ik}^{MF})\sigma_X^2$$

for a male-female pair.

We can estimate $\sigma_X^2$ by fitting the model $y = X\beta + g_X + \mathbf{u} + \epsilon$, where $g_X$ is a vector of genetic effects attributable to the X chromosome, with $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ assuming no dosage compensation, $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ assuming full dosage compensation, and $\text{var}(g_X) = A_X^{ND} \sigma_X^2$ 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.
observed 0–1 scale can be transformed to that on the unobserved continuous liability scale by a linear transformation. The relationship between additive genetic variance on the observed 0–1 and unobserved liability scales was proposed more than a half century ago, 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 the observed 0–1 scale to that on the liability scale with an adjustment for ascertainment bias. There is an important caveat in 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.

### Estimating 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 (I) 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). Let (1 – p)² + p(1 – p)F + 2p(1 – p)(1 – F) and p² + p(1 – p)F be the frequencies of the three genotypes of a SNP and let h₁ = 2p(1 – p). The estimate based on the variance of additive genotype values is

\[ \hat{I}^\text{I} = \frac{\sum_i (x_i - E(x_i))^2}{h_1 - 1} = \frac{(x_i - 2p)^2}{h_1 - 1} \text{ and } \text{var}(\hat{I}^\text{I} | F) = \frac{(1 - h_1)}{h_1 + 7(1 - 2h_1)F/h_1 - F^2}, \]

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

\[ \hat{I}^\text{II} = \frac{O(# \text{ hom}) - E(# \text{ hom})}{1 - E(# \text{ hom})} = 1 - \frac{x_i(2 - x_i)}{h_1 - (1 - 2h_1)F/h_1 - F^2}, \]

where \( O(# \text{ hom}) \) and \( E(# \text{ hom}) \) are the observed and expected number of homozygous genotypes in the sample, respectively. Both estimators are unbiased estimates of I in the sense that \( E(\hat{I}^\text{I} | F) = E(\hat{I}^\text{II} | F) = F \), but their sampling variances are dependent on allele frequency, i.e., \( \text{var}(\hat{I}^\text{I}) = \frac{\text{var}(\hat{I}^\text{II})}{1 - h_1} = F \) if \( F = 0 \). In addition, the covariance between the two estimators is \( (3h_1 - 1) / h_1 + (1 - 2h_1)F / h_1 - F^2 \), so that the sampling covariance between the estimators is \( (3h_1 - 1) / h_1 \) and the sampling correlation is \( (3h_1 - 1) / (1 - h_1) \) when \( F = 0 \). We proposed an estimator based upon the correlation between uniting gametes:

\[ \hat{I}^\text{III} = \frac{[x_i^2 - (1 + 2p_i)x_i + 2p_i^2]}{h_1} \text{ and } \text{var}(\hat{I}^\text{III} | F) = 1 + 2(1 - 2h_1)F/h_1 - F^2. \]

\( \hat{I}^\text{III} \) is also an unbiased estimator of I in the sense that \( E(\hat{I}^\text{III} | F) = F \). If \( F = 0 \), \( \text{var}(\hat{I}^\text{III}) = 1 \) regardless of allele frequency, which is smaller than the sampling variance of \( \hat{I}^\text{I} \) and \( \hat{I}^\text{II} \), i.e., \( 1 \leq (1 - h_1) / h_1 \). When \( F < 1/3 \), \( \hat{I}^\text{III} \) also has a smaller variance than \( \hat{I}^\text{I} \) and \( \hat{I}^\text{II} \). In GCTA, we use \( 1 + \hat{I}^\text{III} \) rather than \( 1 + \hat{I}^\text{II} \) to calculate the diagonal of the GRM. For multiple SNPs, we average the estimates over all of the SNPs, i.e., \( \hat{I} = 1/N \sum_{i=1}^{N} \hat{I}_i \).

### Estimating LD Structure

In a standard GWAS, particularly with a large sample size, the mean (\( \lambda_{\text{mean}} \)) or median (\( \lambda_{\text{median}} \)) 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. An alternative explanation is that polygenic variation causes the observed inflated test statistic. To predict the genomic inflation factors, \( \lambda_{\text{mean}} \) and \( \lambda_{\text{median}} \) 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 \( \lambda_{\text{mean}} \) and \( \lambda_{\text{median}} \), 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 \( y = \mathbf{W}u + e \), 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 \( \sigma^2_y(1/h^2 - 1) \), respectively.
where $$\sigma^2_a$$ 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 PLINK25 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 MACH12 (*.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

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

where $$\beta = 1 - (c + 1 / N) / \text{var}(A_{jk})$$. 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.

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.3 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

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.