SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits

Osorio D Meirelles*,1, Jun Ding1, Toshiko Tanaka2, Serena Sanna3, Hsii-Te Yang1, Dawood B Dudekula1, Francesco Cucca3, Luigi Ferrucci2, Goncalo Abecasis4 and David Schlessinger1

Measurement error and biological variability generate distortions in quantitative phenotypic data. In longitudinal studies with repeated measurements, the multiple measurements provide a route to reduce noise and correspondingly increase the strength of signals in genome-wide association studies (GWAS). To optimize noise correction, we have developed Shrunken Average (SHAVE), an approach using a Bayesian Shrinkage estimator. This estimator uses regression toward the mean for every individual as a function of (1) their average across visits; (2) their number of visits; and (3) the correlation between visits. Computer simulations support an increase in power, with results very similar to those expected by the assumptions of the model. The method was applied to a real data set for 14 anthropomorphic traits in ~6000 individuals enrolled in the SardiNIA project, with up to three visits (measurements) for each participant. Results show that additional measurements have a large impact on the strength of GWAS signals, especially when participants have different number of visits, with SHAVE showing a clear increase in power relative to single visits. In addition, we have derived a relation to assess the improvement in power as a function of number of visits and correlation between visits. It can also be applied in the optimization of experimental designs or usage of measuring devices. SHAVE is fast and easy to run, written in R and freely available online.

European Journal of Human Genetics (2013) 21, 673–679; doi:10.1038/ejhg.2012.215; published online 24 October 2012

Keywords: genome-wide association study; bayesian; multiple measurements; biological variability; measurement error; random-intercept

INTRODUCTION

In contrast to Mendelian traits, for which the association with related penetrant mutations is patent, quantitative traits show a continuous range of values and smaller effect sizes of genetic variants. Thus, to identify genetic factors involved in quantitative traits, larger sample sizes and more refined statistical analyses are required.

Population studies with multiple visits and quantitative trait measurements a priori offer the possibility to increase power and determine the trajectory of trait values in relation to disease or other outcomes. Methods that take all the measurements into account can increase statistical power of the genome-wide association studies (GWAS) analyses that dominate current discovery efforts. Similar benefits of using multiple measurements have been shown in analyses of expression profiling on microarrays1–4 and more recently in studies of blood pressure.5 However, for most existing population cohorts, additional variability is introduced by different numbers of visits for individuals and by possible secular drift. To optimally model this type of data, we propose a shrinkage method that efficiently combines observations from different measurements, even when some visits are missing for some individuals.

The strength of shrinkage estimators compared with frequentist approaches has been clearly described in classical literature6,7 and more recently in GWAS and related areas such as imputation, fine mapping and meta-analysis.8 Our method implements an empirical Bayes algorithm, ‘Shrunken Average (SHAVE), using regression toward the mean for every individual as a function of their number of visits and the correlation between visits. We evaluated the performance of the method by simulations and confirmed the expectations in real data.

We used the SardiNIA cohort (http://sardinia.nia.nih.gov)9 consisting of >6000 individuals and a set of 14 traits that were measured up to three times in all individuals, at time ~3-year intervals. To evaluate the impact of the method, we selected top SNPs from single visits and meta-analysis studies, and compared the significance of the same SNPs for single visits, for the average across visits and for SHAVE. Variable but appreciable improvement in performance was found.

SHAVE is fast and easy to run, and can thus be added to approaches such as principal component and variance analysis. Finally, we suggest a way to estimate the cost-benefit of adding additional visits for GWAS signals and discuss the potential utility of SHAVE for other applications. The R code for SHAVE is available (http://sardinia.nia.nih.gov/Download/).

METHODS

We outline briefly how we test for the association of a single measurement of a trait with a given SNP, and then generalize for multiple replicates. Consider a given quantitative trait and a given SNP. Let \( y_{ij} \) \((i=1,...,n; j=1,...,k)\) denote individual i’s jth repeated measure of the trait, or his/her residual for that trait after adjusting for one or more variables (eg, sex and age). Let \( G \) denote the...
number of minor alleles of the given SNP for individual $i$. Let $G$ be the vector $(G_1, G_2, ..., G_n)$ containing the number of alleles ($0$, $1$, or $2$) for all individuals. Similarly, let $Y_1$ be the vector $(y_{11}, y_{21}, ..., y_{ni})$, containing the first measurement of the trait for all individuals. To test the association between $Y_1$ and $G$ using an additive model, $Y_1$ is regressed on $G$ such that $y_{ij} = \beta_i G_i + z_i \epsilon_i$ and an estimate for $\beta_i$ is obtained ($\beta_i^*$). We then divide $\beta_i^*$ by its standard error and obtain a $z$-statistic. Next for each $z$-statistic a corresponding $P$-value is obtained. This is done separately for every SNP and every trait.

**SHAVE: the posterior expectation $\mu^*_i$**

Now consider the following random-intercept model:

$$y_{ij} | \mu_i \sim N(\mu_i, \sigma^2/w) \text{ and } \mu_i \sim N(0, \sigma^2)$$

where $w \geq 0$ and $\sigma^2 \geq 0$ are unknown parameters, which can be easily estimated from our dataset. Straightforward algebra, such as that presented in an elementary textbook on Bayesian statistics (e.g., Lee(10)) would give the posterior expectation of $\mu_i$ when $\sigma^2$ and $w$ are known. Thus, $\mu_i | y_1, y_2, \ldots, y_{ni}, w, \sigma^2 \sim N(\tilde{\mu}_i, \tilde{\sigma}_i^2)$, where $\tilde{\sigma}_i^2$ is the number of visits and $\tilde{\mu}_i$ is the average across the visits for individual $i$. For the proof, see Supplementary Materials Section 1. Let $\mu_i^*$ denote the posterior expectation $E[\mu_i | y_1, y_2, \ldots, y_{ni}, w, \sigma^2]$, then

$$\mu_i^* = (k_iw/(1 + k_iw))\tilde{\mu}_i$$

We call $\mu_i^*$ the SHAVE estimator for multiple visits. Note that not all individuals have the same number of visits, thus if an individual $i$ did not have visit $j$, the measurement $y_{ij}$ is set to missing, thus $\tilde{\mu}_i$ will be the average of non-missing values. The reason that $\tilde{\mu}_i$ varies between individuals is mainly due to missing data, which we are assuming is missing completely at random, that is, is unrelated to age, sex, and the measurement value itself.

The term $(k_iw/(1 + k_iw))$ will be referred to as the adjusting factor of the average, which is equal to one minus the shrinkage factor. We note that $\mu_i^*$ does not depend on $\sigma^2$ and is a function only of $k_i$, $w$ and $\tilde{\mu}_i$. Next, $\mu_i^*$ is regressed on $G$ such that $\mu_i^* = \beta G_i + a + \epsilon_i$ and the statistical significance of $\beta$ is calculated.

**Estimating $w$ and $\sigma^2$**

Let $n$ be the total number of individuals and $k_i$ be the number of visits for individual $i$. Given that equation 1 implies $\text{Var}(y_{ij}|\mu_i) = \sigma^2/w$ and $\text{Var}(y_i) = \sigma^2/w + \sigma^2$, both quantities can be respectively estimated by $s^2_{\text{within}} = \sum_{j=1}^{n} \sum_{i=1}^{k_i} (y_{ij} - \bar{y}_i)^2/(k_i - 1)$ and $s^2_{\text{total}} = \sum_{j=1}^{n} \sum_{i=1}^{k_i} (y_{ij} - \bar{y})^2/k_i$, and setting $\sigma^2/w = s^2_{\text{within}}$ and $\sigma^2/w + \sigma^2 = s^2_{\text{total}}$ yields the estimate of $w$. Although the estimate of $\sigma^2$ is not needed for $\mu_i^*$, $\sigma^2$ can be estimated by $s^2_{\text{total}} - s^2_{\text{within}}$. Thus, the weight estimate is given by:

$$w = (s^2_{\text{total}} - s^2_{\text{within}})/s^2_{\text{within}}$$

Therefore, in equation 2 the term $w$, which is unknown, should be replaced by its estimate. When all individuals have exactly two visits, $w$ is equal to $\rho/(1 - \rho)$, where $\rho$ is the sample correlation between the two visits, and $w$ also minimizes the least squares loss function $L_2 = \sum_{j=1}^{n} \sum_{i=1}^{k_i} [E[\mu_i | w, y_{ji}] - y_{ij}]^2$.

Another possibility is to use a more robust loss function $L_3 = \sum_{j=1}^{n} \sum_{i=1}^{k_i} |E[\mu_i | w, y_{ji}] - y_{ij}|$ and estimate $w$ that minimizes $L_3$. Although estimating $w$ by $L_2$ and $L_3$ will give different results, both weights were similar for most traits when using the SardiNIA data, and furthermore, their corresponding $z$-statistics for SHAVE were extremely similar for all traits. For more details see Supplementary Materials Section 2 and Table S2.

**Comparing different metrics – LOD ratio**

Comparisons were done in GWAS using three summary trait values from multiple visits: single visit, Average and SHAVE.

Note: To distinguish between the statistical term ‘average’ and the actual ‘Average’ among visits, we use the latter throughout this paper. To assess performance among different metrics, for a given trait and a given SNP, we run an association test for each visit, the Average and SHAVE, obtaining a corresponding slope and $z$-statistic and calculating the corresponding $z^2$. The LOD score, one of the outputs from the Merlin(11) software, is defined as $z^2\log(100)$ and was chosen as a performance measure because the LOD score (or equivalently $z^2$) is conveniently proportional to the sample size. For example, assume a true association between a trait and a specific SNP. If the sample size were equal to 2000 individuals with a corresponding $z^2$, then doubling the sample size to 4000 individuals would be expected to double $z^2$ as well. Next, we describe three common scenarios and provide an expected LOD ratio between different metrics, with $z_{AVG}$, $z_{SHAVE}$, and $z_{AVG}$ as the corresponding $z$-statistics for single visit, Average and SHAVE.

**Average vs single visit**

We start by considering that all individuals have the same number of visits (equation 4) – that is, from a balanced dataset – and we then account for a situation in which there are different numbers of visits for individuals (unbalanced dataset) (equation 5).

**In a balanced dataset,** $E[z_{AVG}^2]/E[z_i^2] \approx k(1 + w)/(kw + 1)$

**In an unbalanced dataset,** $E[z_{AVG}^2]/E[z_i^2] \approx (1 + w)/(w + E[1/k_i])$

Next we assume an unbalanced dataset for the LOD ratio between SHAVE vs Average (equation 6).

**SHAVE vs Average**

$$E[z_{SHAVE}^2]/E[z_{AVG}^2] \approx E[(1 + kw)/kw]E[kw/(1 + kw)]$$

Proofs for equations 4, 5 and 6 can be found in Supplementary Materials Section 3. At this stage we point out a salient fact that if every individual has the same number of visits, then by equation 2, SHAVE will be the Average multiplied by a constant factor $(kw/(1 + kw))$, which implies that $z_{SHAVE}$ is identical to $z_{AVG}$, also indicating that SHAVE will have the same power as the Average. This equality in power in balanced datasets between SHAVE and Average is also consistent with (equation 6), where replacing $k$ by $k_i$ results in a ratio equal to one.

**Simulation study**

Simulated unbalanced datasets were generated with 5000 individuals, with 2500 individuals with three visits and the remaining 2500 with a single visit. We conducted two types of simulations, one to estimate Type I error and the other to estimate power. In each type of simulation we compared SHAVE, Average and single visits.

**Simulation models**

As for all metrics described, $\sigma^2$ is independent from the $z$-statistics, we set $\sigma^2$ equal to one. Next, we describe two simulation models, where model 1 is used to measure Type I error and model 2 is used to measure power.

Model 1, for $\beta = 0$: $y_{ij} = \mu_i + \epsilon_i$, where $\mu_i \sim N(0, 1)$ and $\epsilon_i \sim N(0, 1/w)$ and $\epsilon_i$ is independent from $\mu_i$. In this model we can see that $\text{Var}(y_{ij}) = 1 + 1/w$. Model 2, for $\beta \neq 0$: $y_{ij} = \mu_i + \epsilon_i$, where $\mu_i = \beta_i - (1 - \beta)\text{Var}(G_i) - \beta(G_i - \overline{G})$, where $\beta_i \sim N(0, 1)$, $\epsilon_i \sim N(0, 1/w)$, $G_i$ is randomly generated based on the pre-defined allele frequency, $\overline{G}$ is the average number of alleles across all individuals, and $\epsilon_i$ is independent from $\mu_i$. Since our original random-intercept model assumed that $\text{Var}(y_{ij})$ does not depend on the genotype, the term $(1 - \beta)\text{Var}(G_i) - \beta(G_i - \overline{G})$ is introduced in order to have $\text{Var}(\beta_i - (1 - \beta)\text{Var}(G_i) - \beta(G_i - \overline{G})) = \text{Var}(\mu_i) = 1$, which implies that $\text{Var}(\beta_i) = 1 + 1/w$ in both models 1 and 2.

**Type 1 error simulations**

We set $z$ levels to $1 \times 10^{-5}$, $1 \times 10^{-6}$, $1 \times 10^{-7}$ and $5 \times 10^{-8}$. Ten billion simulations were performed to achieve an accurate Type I error estimation. Comparisons between correlations $w$ were set equal to 0.2 or 0.5, and minor allele frequency $P$ was set equal to 0.5. We then simulated $y_{ij}$ values for all three visits and all individuals according to model 1 using the ‘true’ weight $w = \rho/(1 - \rho)$. Next, we randomly set as missing 50% of the values for visits 2 and 3, and Average was then calculated for every individual based on non-missing values. Next, we estimated the sample weight $w$ using equation 3 and generated SHAVE. Finally, we simulated the vector $G$ based on the minor allele frequency
P (0.5). After performing the simple linear regression between each metric and G, P-values were obtained. Next, for each metric we measured Type I error as the proportion (over 10^10 simulations) of P-values smaller than each x level.

Power simulations
Simulations were conducted using x level of 5 × 10^-8, β values of 0.20, 0.25 and 0.30, minor allele frequencies P equal to 0.1 and 0.5, and correlations between visits ρ equal to 0.2 and 0.5. Simulated values were generated similarly to Type I error simulations with the main difference being that model 2 was used instead of model 1. One million simulations were performed for each combination of parameters (β, P and ρ), and as a result of each combination, we measured power as the proportion of P-values less than the 5 × 10^-8 cutoff, now considered the standard threshold to declare genome-wide significance findings.

Applying the method – SardiNIA dataset
The SardiNIA project was designed to investigate the genetics of quantitative traits in the Sardinian founder population. Over a 10-year period, from November 2001 to the present, residents of four towns in Sardinia, Italy, starting at age 14–95 years, were invited to participate to the study, and a total of 6320 individuals had up to three visits at ~3-year intervals. The total number of individuals in each visits one, two and three was 6177; 5670; and 1971, respectively, where each individual could be present or not in any of the visits. Individuals were characterized for >100 quantitative traits, and 14 traits were selected for this analysis (bilirubin, total cholesterol, γ-GT, glycemia, HDL, height, LDL, PR-interval, QT-interval, red blood cell counts (RBC), serum iron, transferrin, triglycerides and uric acid). Traits were selected based on previously reported meta-analysis studies (as of October 2011), which also showed top SNPs for visits 1 and 2 with P < 5 × 10^-8 from the SardiNIA dataset, where the same top SNPs had minor allele frequency >5% and were also SNPs were previously identified from the Hapmap project (SNPs with ‘rs’ as the first two characters). Genotype information was obtained from the Metabochip, a custom Illumina iSELECT genotyping array (http://www.sph.umich.edu/csg/kang/MetaboChip).

To minimize the effect of outliers, we applied an inverse normal transformation for every trait in each visit. Transformed traits were used as the dependent variable and modeled using linear regression, with age at the time of visit and sex as covariates for each separate trait and for each visit. As a result, each trait measurement version (a given trait for a given visit) generated standardized residuals (mean equal to zero and SD equal to one) as the output. Result, each trait measurement version (a given trait for a given visit) generated standardized residuals (mean equal to zero and SD equal to one) as the output. This standardization step is needed in order to assume that noise levels are the same for each visit. However, GWAS results without standardization were very similar (not shown).

Comparing performance of metrics
To measure the performance of metrics, the most significant SNP for each trait was selected based on three criteria: significance of the signal in visit 1, significance of the signal in visit 2, and significance in published meta-analyses. Next, for each SNP we ranked the P-values among metrics and then we obtained the average rank for each metric across all traits. As SNPs were selected based on reported meta-analysis (Table 3), but not all of those were present in the Metabochip, we used the SNAP algorithm to select a proxy SNP in the Metabochip that had the highest R^2 (≥0.80). As SardiNIA project is a family based study, to test for association while accounting for relatedness, we used a variance component method implemented in Merlin.

RESULTS
Simulation results – power and Type I error
Simulated Type I errors were very similar to expected (z), showing that Type I error is well controlled for all three metrics – single visit, Average of up to three visits and SHAVE of up to three visits (Supplementary Materials Table S1). We also noticed a clear increase in power for SHAVE relative to the Average and to a single visit (Table 1). With z level, minor allele frequency P and effect size β, respectively, set to 5 × 10^-8, 0.50 and 0.20, simulated power is shown as an increasing function of the correlation between visits (Figure 1 top). Similarly, with z, P and β set to 5 × 10^-8, 0.50 and 0.20, simulated power is shown as an increasing function of the effect size β (Figure 1 bottom). In addition, simulated and expected power was very similar for all three methods. A detailed description of the calculation of expected power can be found in Supplementary Materials Section 4.

SardiNIA dataset – performance by ranking
To compare metrics, we use the average rank across 14 traits (Tables 2–4), where lower average rank indicates higher overall significance. Data for all three visits in SardiNIA and selecting for the top SNP based on visit 1 (Table 1), the average ranks for visit 1, visit 2, and Average and SHAVE were 3.36, 3.50, 2.07 and 1.07. Similarly when selecting for the top SNP based on visit 2 (Table 3), corresponding average ranks were 3.64, 3.00, 2.21 and 1.14. When selecting for top Meta-Analysis SNP’s (Table 4), corresponding average ranks were 3.29, 3.57, 1.93 and 1.21. On the basis of these findings, Average was superior to any single visit in all three tables, with SHAVE having the best performance, (less significant than the Average only twice out of 42 cases (height and QT-interval in Table 4)). An alternative way to compare performance by ranking is shown in Supplementary Materials Table S3.

SardiNIA dataset – performance by LOD ratios using top Meta-analysis SNPs
We performed two types of LOD ratios for every trait, the first between Average and single visit, the second between SHAVE and Average. To compare signals between Average and single visits, we selected a subset of individuals who had both visits 1 and 2 and compared their signals. We first obtained the z-statistics of the Average (z_{AVG}) and the z-statistics corresponding to visits 1 and 2 (z_1 and z_2). Next, we obtained the LOD ratio between Average (represented by the square of z_{AVG}) and a single visit (represented by the square of (z_1 + z_2)/2). Observed LOD ratios were all above one, indicating an increase in power using the Average vs a single visit (Figure 2). We note that traits with lower correlation between visits had the highest LOD ratios, and in the three traits with lowest correlation, transferrin, serum iron and QT-interval, LOD ratios were above 1.5. Similarly, traits with high correlation between visits, such as RBC and height, had LOD score ratios close to one. In general,

| Parameters | Simulated power | Expected power |
|-----------|----------------|---------------|
|           | P | β | P | Single | Average | SHAVE | Single | Average | SHAVE |
| 0.10 | 0.20 | 0.20 | 0.0028 | 0.0103 | 0.0185 | 0.0028 | 0.0103 | 0.0185 |
| 0.10 | 0.20 | 0.50 | 0.1137 | 0.2114 | 0.2406 | 0.1147 | 0.2129 | 0.2419 |
| 0.10 | 0.25 | 0.20 | 0.0179 | 0.0628 | 0.1066 | 0.0181 | 0.0629 | 0.1070 |
| 0.10 | 0.25 | 0.20 | 0.4437 | 0.6429 | 0.6872 | 0.4467 | 0.6456 | 0.6892 |
| 0.10 | 0.30 | 0.20 | 0.0772 | 0.2279 | 0.3447 | 0.0777 | 0.2283 | 0.3451 |
| 0.10 | 0.30 | 0.20 | 0.8262 | 0.9371 | 0.9531 | 0.8257 | 0.9391 | 0.9546 |
| 0.50 | 0.20 | 0.20 | 0.1640 | 0.4118 | 0.5648 | 0.1657 | 0.4131 | 0.5655 |
| 0.50 | 0.20 | 0.50 | 0.9495 | 0.9899 | 0.9935 | 0.9509 | 0.9902 | 0.9937 |
| 0.50 | 0.25 | 0.20 | 0.5581 | 0.8630 | 0.9430 | 0.5617 | 0.8634 | 0.9426 |
| 0.50 | 0.25 | 0.50 | 0.9997 | 1.0000 | 1.0000 | 0.9997 | 1.0000 | 1.0000 |
| 0.50 | 0.30 | 0.20 | 0.8987 | 0.9923 | 0.9987 | 0.9008 | 0.9922 | 0.9986 |
| 0.50 | 0.30 | 0.50 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Table 1: Simulated and expected power for alpha equal to 5 × 10^-8 and different levels of frequency P, slope β and correlation ρ

There were 1 million simulations for each combination of P, β and ρ. Expected power is estimated based on equations 5 and 6.
expected and observed LOD ratios were quite similar, suggesting that our observations match the expectations of the model.

To compare LOD ratios between SHA VE and Average, we generated a subset of the SardiNIA dataset in which all individuals had visit 1, and then, for the same individuals, we randomly selected 50% of them and included their second visits (setting the remaining visit 2 cases as 'missing values'). The main reason to look at this subset was that differences between SHA VE and Average are only appreciable in unbalanced datasets. Although the observed LOD ratios between SHA VE vs Average were modest when compared with Average vs single visit, the ratios were all greater than one, indicating a consistent increase in power of SHA VE relative to Average (Figure 3). Also, with the exception of transferrin, observed and expected LOD ratios were similar.

### Expected LOD ratios in a hypothetical dataset

To get a better estimate of the expected LOD ratio between Average and single visits, and between SHA VE and Average, we generated charts based on a hypothetical dataset with multiple visits (from 2 to 10 visits). When comparing Average vs single visit, we show the expected LOD ratio as a function of the number of visits and the correlation between visits (Figure 4). The expected LOD ratios decrease as the correlation between visits increases. Similarly, expected LOD ratios increase as the number of visits increases, and saturates as the number of visits $k$ becomes large, based on equation 4. When comparing SHA VE vs Average, we assumed a hypothetical dataset in which 50% of the individuals had a single visit and 50% of the individuals had $k$ visits (from 2 to 10) (Supplementary Materials Figure S1). Here LOD ratios are more modest when compared with Figure 4, but still show the same relation to number of visits and correlation.

### DISCUSSION

Increasing the strength of a true genetic signal for a quantitative trait can provide overall benefits for GWAS studies, and we show here the extent to which measurements from multiple visits can contribute to that goal. In particular, when we compared the performance of SHA VE vs single visit and SHA VE vs Average using the SardiNIA dataset, some traits showed a large LOD ratio for their top SNPs, indicating that the same genome-wide significance can be achieved using a smaller sample with SHA VE. SHA VE increases power relative to the Average when the dataset is unbalanced (ie, individuals have different number of trait measurements). However, when a dataset is balanced, SHA VE and Average generate identical results. The increase in power for SHA VE was also supported by simulations, which showed both Type I error and power very close to that expected under the assumptions of the linear model.

Power increases with effect size (absolute value of the slope), number of visits and correlation between visits. Given the goal of maximizing the increase in power, when is SHA VE most useful? If power from a single visit is low — such that the top SNP is far from being genome-wide significant — then even the increase in power by SHA VE will not be sufficient for any SNP to achieve genome-wide significance. On the other hand, when a SNP shows marginal genome-wide significance in a single visit, the power boost from SHA VE may make a SNP genome-wide significant. Moreover, when a SNP is already genome-wide significant in a single visit, an increase in power by SHA VE will further improve genome-wide significance, providing additional confidence in the SNP effect.

A major assumption of the random-intercept model is that Var($y_{ij} | m_i$) = $s^2/w$ (a combination of biological variability and measurement error) is identical for each visit. This might not always be the case if better technology were used to measure a trait in a more recent visit (reducing measurement error), or if better protocols are used (reducing biological variability). However, SHA VE can easily be adapted to such datasets, and one potential improvement could be to estimate a different weight $w_j$ for each visit $j$. In such instances SHA VE and the Average will not be equivalent even in balanced datasets, with SHA VE expected to outperform the Average. Another key assumption is that the true variance (unknown) within each individual is constant. If this assumption is violated shrinkage
Table 2 Association results between 14 traits and their corresponding top SNPs, where top SNPs were selected based on visit 1 results of SardiNIA GWAS, and where z-statistics for Average and SHAVE are based on three visits

| TRAIT        | SNP        | z1   | z2   | zAVG | zSHAVE | Visit 1 | Visit 2 | Average | SHAVE |
|--------------|------------|------|------|------|--------|---------|---------|---------|-------|
| Bilirubin    | rs887829   | 27.33| 27.37| 31.44| 31.59  | 4       | 3       | 2       | 1     |
| Cholesterol  | rs4910742  | -6.25| -4.96| -5.88| -6.04  | 1       | 4       | 3       | 2     |
| γ-GT         | rs7310409  | -6.29| -6.52| -6.53| -6.69  | 4       | 3       | 2       | 1     |
| Glycemia     | rs853787   | -7.17| -7.50| -8.06| -8.24  | 4       | 3       | 2       | 1     |
| HDL          | rs247617   | 8.45 | 8.93 | 10.10| 10.19  | 4       | 3       | 2       | 1     |
| Height       | rs3132468  | 5.91 | 5.82 | 5.91 | 5.92   | 3       | 4       | 2       | 1     |
| LDL          | rs445925   | -5.89| -6.13| -6.74| -6.77  | 4       | 3       | 2       | 1     |
| PR-interval  | rs6800541  | 6.56 | 5.92 | 7.10 | 7.11   | 3       | 4       | 2       | 1     |
| QT-interval  | rs12036340 | 6.27 | 5.31 | 7.21 | 7.27   | 3       | 4       | 2       | 1     |
| RBC          | rs4910742  | 23.59| 22.19| 24.16| 24.26  | 3       | 4       | 2       | 1     |
| Serum iron   | rs4820268  | 8.77 | 7.52 | 10.43| 10.66  | 3       | 4       | 2       | 1     |
| Transferrin  | rs4854761  | 9.58 | 13.04| 14.96| 15.40  | 4       | 3       | 2       | 1     |
| Triglycerides| rs10401969 | -6.15| -4.72| -6.67| -6.76  | 3       | 4       | 2       | 1     |
| Uric acid    | rs13145758 | -11.84| -12.52| -13.80| -14.05 | 4       | 3       | 2       | 1     |

Average rank 3.36 3.50 2.07 1.07

Table 3 Association results between 14 traits and their corresponding top SNPs, where top SNPs were selected based on visit 2 results of SardiNIA GWAS, where z-statistics for Average and SHAVE are based on three visits

| TRAIT        | SNP        | z1   | z2   | zAVG | zSHAVE | Visit 1 | Visit 2 | Average | SHAVE |
|--------------|------------|------|------|------|--------|---------|---------|---------|-------|
| Bilirubin    | rs887829   | 27.33| 27.37| 31.44| 31.59  | 4       | 3       | 2       | 1     |
| Cholesterol  | rs6511720  | -4.60| -5.96| -5.49| -5.62  | 4       | 3       | 2       | 1     |
| γ-GT         | rs7310409  | -6.29| -6.52| -6.53| -6.69  | 4       | 3       | 2       | 1     |
| Glycemia     | rs853787   | -7.17| -7.50| -8.06| -8.24  | 4       | 3       | 2       | 1     |
| HDL          | rs247617   | 8.45 | 8.93 | 10.10| 10.19  | 4       | 3       | 2       | 1     |
| Height       | rs3132468  | 5.91 | 5.82 | 5.91 | 5.92   | 3       | 4       | 2       | 1     |
| LDL          | rs6511720  | -5.86| -6.93| -6.81| -6.97  | 4       | 3       | 2       | 1     |
| PR-interval  | rs67995970 | 6.18 | 5.94 | 6.80 | 6.86   | 3       | 4       | 2       | 1     |
| QT-interval  | rs12143842 | 6.27 | 5.47 | 7.30 | 7.33   | 3       | 4       | 2       | 1     |
| RBC          | rs4910742  | 23.39| 22.19| 24.16| 24.26  | 3       | 4       | 2       | 1     |
| Serum iron   | rs855791   | -8.03| -7.88| -10.24| -10.55 | 3       | 4       | 2       | 1     |
| Transferrin  | rs4854761  | 9.58 | 13.04| 14.96| 15.40  | 4       | 3       | 2       | 1     |
| Triglycerides| rs6999813  | -5.67| -7.63| -6.87| -6.97  | 4       | 3       | 2       | 1     |
| Uric acid    | rs13145758 | -11.84| -12.52| -13.80| -14.05 | 4       | 3       | 2       | 1     |

Average rank 3.64 3.00 2.21 1.14

The z-statistics are shown in order for visit 1 (z1), visit 2 (z2), Average (zAVG) and SHAVE (zSHAVE). In the next four columns their corresponding ranks within each trait are shown, where 1 is assigned to the most significant and 4 to the least. On the last row, we have the averages of the ranks for each metric.

Distortions result. In our model we assume that this true variance within individual i, denoted by $\eta_i^2$, is equal to $\sigma_i^2/w_i$. However, if $\eta_i^2$ is equal to $\sigma_i^2/w_i$, $w_i$ is the unknown weight for individual i, then if $\eta_i^2 > \sigma_i^2/w_i$, $w_i$ will be greater than his/her true weight $w_i$, leading to 'under-shrinkage', and similarly if $\eta_i^2 < \sigma_i^2/w_i$, then $w_i$ will be smaller than $w_i$ with 'over-shrinkage'. Thus, to minimize the effects of over shrinkage and under shrinkage, a potential improvement would be to estimate $w_i$ for each individual, were SHAVE will likely outperform the Average even in balanced datasets. Although preliminary results showed very small increase in power (Supplementary Materials Table S4), there is still potential for improvement in datasets with more visits, in which case the estimate of $w_i$ will be more precise.

Our method uses a two-step model where in the first step we estimate $w$ and use it to calculate SHAVE, and in the second step, SHAVE is regressed on G to obtain the estimate of $\beta$ and the corresponding z-statistic. One potential improvement would be to use a one-step model, in which $w$ and $\beta$ are estimated jointly. However, if the genetic variance of the top SNP of a trait, which is equal to $\beta^2 \text{Var}(G)$ is small relative to $\sigma^2$, the expected increase in power will be insignificant. Moreover, preliminary results comparing both one-step and two-step models were nearly identical (Supplementary Materials Table S5).

The derived relations of LOD score ratios, in which the simplest and most practical is $E[z^2_{\text{AVG}}]/E[z^2_{\text{SHAVE}}] \approx k(1 + w)/(kw + 1)$, can be applied in cost-benefit analysis for signal improvement in the usage of measuring devices and in experimental design. For example, suppose we are considering adding an additional visit for a trait, and that we have had some preliminary GWAS results for a given SNP. Under the assumption that the signal is true, by estimating the sample correlation between visits, one could estimate the potential increase in significance for that SNP if an additional visit were obtained. This can provide guidance in planning research. Moreover, one can
estimate the potential increase in significance for epidemiological studies and GWAS.

In summary, SHA VE takes advantage of multiple trait measurements to boost statistical power for GWAS of quantitative traits. Although, the specific weighting scheme used in this paper is a simple version that is easy to implement even in large-scale GWAS, there are many additional ways to improve the method. The method can also be adapted for more complicated scenarios with unique trait characteristics. For example, traits such as pulse wave velocity show a trait variance that increases with age, in which case weights can be estimated as a function of age. Other traits such as systolic and diastolic blood pressure show a trait variance that increases with the magnitude of the measurement, in which case weights can be estimated as a function of the trait. Such new weighting schemes could potentially further increase the statistical power of genetic studies of quantitative traits.

### Table 4: Association results between 14 traits and their corresponding top SNPs, where top SNPs were selected based on multi-study meta-analyses, and z-statistics for Average and SHA VE are based on three visits and results of SardiNIA GWAS

| TRAIT       | SNP      | z₁    | z₂    | zAVG   | zSHAVE | Visit 1 | Visit 2 | Average | SHA VE |
|-------------|----------|-------|-------|--------|--------|---------|---------|---------|--------|
| Bilirubin   | rs887829 | 27.33 | 27.37 | 31.44  | 31.59  | 4       | 3       | 2       | 1      |
| Cholesterol | rs467766*| −4.76 | −4.20 | −4.68  | −4.76  | 1       | 4       | 3       | 2      |
| γ-GT        | rs7310409| −6.29 | −6.52 | −6.53  | −6.69  | 4       | 3       | 2       | 1      |
| Glycemia    | rs10830963| 6.32  | 6.34  | 7.38   | 7.50   | 4       | 3       | 2       | 1      |
| HDL         | rs247617 | 8.45  | 8.93  | 10.10  | 10.19  | 4       | 3       | 2       | 1      |
| Height      | rs724016 | 5.02  | 4.25  | 5.26   | 5.24   | 3       | 4       | 1       | 2      |
| LDL         | rs467766*| −5.39 | −5.14 | −5.69  | −5.76  | 3       | 4       | 2       | 1      |
| PR-interval | rs6800541| 6.56  | 5.92  | 7.10   | 7.11   | 3       | 4       | 2       | 1      |
| QT-interval | rs7550692*| 5.89  | 4.36  | 6.66   | 6.55   | 3       | 4       | 1       | 2      |
| RBC         | rs4910742| 23.39 | 22.19 | 24.16  | 24.26  | 3       | 4       | 2       | 1      |
| Serum iron  | rs4820268| 8.77  | 7.52  | 10.43  | 10.66  | 3       | 4       | 2       | 1      |
| Transferrin | rs4854761*| 9.58  | 13.04 | 14.96  | 15.40  | 4       | 3       | 2       | 1      |
| Triglycerides | rs1260326| 5.12  | 4.75  | 5.92   | 5.95   | 3       | 4       | 2       | 1      |
| Uric acid   | rs9998811*| −11.58| −12.44| −13.61 | −13.87 | 4       | 3       | 2       | 1      |
| Average rank|          | 3.29  | 3.57  | 1.93   | 1.21   |         |         |         |        |

The z-statistics are shown in order for visit 1 (z₁), visit 2 (z₂), Average (zAVG) and SHA VE (zSHAVE). In the next four columns their corresponding ranks within each trait are shown, where 1 is assigned to the most significant and 4 to the least. On the last row, we have the averages of the ranks for each metric. *The original SNPs (which were replaced by proxy SNPs from the MetaboChip) are: cholesterol (total) rs629301 (R² = 1.00), LDL rs629301 (R² = 1.00), QT-interval rs2880058 (R² = 0.92), transferrin rs3811647 (R² = 0.96) and uric acid rs734553 (R² = 0.80).
Figure 4 Expected LOD ratio between Average and single visit for hypothetical datasets in which all individuals had \( k \) visits ranging from 2 to 10.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
This research was supported in part by the Intramural Research Program of the National Institute on Aging, NIH. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov). The SardiNIA (Progenia) team was supported by Contract NO1-AG-1-2109 from the NIA; the efforts of GRA were supported in part by contract 263-MA-410953 from the NIA to the University of Michigan and by research grant HG002651 and HL084729 from NIH (to GRA).

1. Astrand M, Mostad P, Rudemo M. Empirical Bayes models for multiple probe type microarrays at the probe level. BMC Bioinform 2008; 9: 156.
2. Meirelles O. Statistical Methods in Microarrays and High-Throughput Flow Cytometry, (PhD thesis). Albuquerque, NM: University of New Mexico, 2009.