1 **Comparing genome-wide association study results from different measurements**

2 **of an underlying phenotype**

3

4 Authors:

5 Joseph L. Gage[1], Natalia de Leon[1]*, Murray Clayton[2]

6

7

8

9 Affiliations:

10 [1] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA

11 [2] Departments of Statistics and Plant Pathology, University of Wisconsin-Madison, Madison, WI
12    53706, USA

13 * Corresponding author: ndeleongatti@wisc.edu

14

15    **Abstract**

16          Increasing popularity of high-throughput phenotyping technologies, such as image-

17    based phenotyping, offer novel ways for quantifying plant growth and morphology.  These new

18    methods can be more or less accurate and precise than traditional, manual measurements.

19    Many large-scale phenotyping efforts are conducted to enable genome-wide association studies

20    (GWAS), but it is unclear exactly how alternative methods of phenotyping will affect GWAS

21    results.  In this study we simulate phenotypes that are controlled by the same set of causal loci

22    but have differing heritability, similar to two different measurements of the same morphological

23    character.  We then perform GWAS with the simulated traits and create receiver operating

24    characteristic (ROC) curves from the results.  The areas under the ROC curves (AUCs) provide a

25    metric that allows direct comparisons of GWAS results from different simulated traits.  We use

26    this framework to evaluate the effects of heritability and the number of causative loci on the

27    AUCs of simulated traits; we also test the differences between AUCs of traits with differing

28    heritability.  We find that both increasing the number of causative loci and decreasing the

29    heritability reduce a trait's AUC.  We also find that when two traits are controlled by a greater

30    number of causative loci, they are more likely to have significantly different AUCs as the

31    difference between their heritabilities increases.  These results provide a framework for deciding

32    between competing phenotyping strategies when the ultimate goal is to generate and use

33    phenotype-genotype associations from GWAS.

34

35    **Introduction**

36          As image-based methods for quantifying plant phenotypes grow in popularity, they

37    present the ability to measure phenotypes that previously could not be easily quantified as well

38    as an alternative way to measure phenotypes that previously had to be manually quantified. The

39    types of novel phenotypes enabled by image analysis include fractal dimension (Gage *et al.*

40    2017), principal component analysis of plant organ biomass (Miller *et al.* 2016) or shape

41    (Chitwood *et al.* 2014), and topological methods for quantifying branching patterns (Li *et al.*

42    2017).  Image-based phenotyping enables increases in resolution, such as in  Miller *et al.* (2016)

43    scanning maize ears on a flatbed scanner at 1,200 dots per inch; it enables advances in accuracy,

44    such as quantifying disease resistance (Bock *et al.* 2010); and it enables scalability and

45    throughput, by building multiple imaging devices (Durham Brooks *et al.* 2010) or using a

46    mobile imaging device (Men *et al.* 2012).  However, image-based phenotype measurements are

47    not always as accurate as high quality manually measured phenotypes, though the decrease in

48    accuracy may be paired with an increase in throughput or decrease in cost if high quality manual

49    phenotypes are time consuming or expensive to measure.  Though it is not strictly an image-

50    based method of phenotyping, one example of the tradeoff of accuracy for efficiency is the use of

51    near infrared reflectance spectroscopy (NIRS), which has been used for decades as a way to

52    predict chemical composition of silage feedstock without costly, expensive, and sometimes

53    hazardous wet lab assays (Park *et al.* 1998).

54       Potential tradeoffs between measurement accuracy and throughput need to be carefully

55    considered by scientists preparing for large-scale experiments.  In the fields of plant breeding

56    and plant genetics, genetic mapping experiments are one example of the type of study where

57    such considerations are crucial.  Genome-wide association studies (GWAS) involve measuring a

58    phenotype, usually in a replicated population of hundreds to thousands of genetically distinct

59    individuals, then scanning for statistical associations between individuals' phenotype and their

60    genotype at numerous genetic loci.  In such studies, the accuracy and precision with which a

61    phenotype is measured will have a direct impact on the ability to detect genetic associations by

62    GWAS.

63       At its core, GWAS involves testing for a difference in phenotype between individuals with

64    different genotypes at a particular single nucleotide polymorphism (SNP).  This process is

65   repeated separately for hundreds of thousands of SNPs across the genome. Ideally, SNPs within

66   or near genes that have some effect on the phenotype of interest will result in strong statistical

67   associations. As such, the precision and accuracy of phenotypic measurements influence the

68   ability to detect statistical differences between groups of individuals with different alleles. The

69   heritability of a phenotype, defined as the ratio of genotypic variance to phenotypic variance, is a

70   useful way to quantify the proportion of phenotypic variability that is attributable to genetic

71   differences between individuals. All other components held equal, heritability will increase as

72   precision of phenotypic measurement increases, due to decreasing phenotypic variability from

73   measurement error. Two methods of measuring the same 'true' phenotype with differing

74   precision will have different heritability, and thus different power to detect SNPs that are

75   statistically associated with the 'true' phenotype of interest. For the remainder of this study we

76   will refer to the 'true' phenotype of an individual as its character, and will refer to different

77   measurements of a character as traits.

78       In addition to heritability, another parameter that affects power in GWAS is the number

79   of loci that control a character. For two characters measured with the same heritability, one

80   controlled by fewer loci will have on average a larger proportion of variance explained by each

81   locus. Power to detect an association at a particular locus is positively related to the proportion

82   of phenotypic variance explained by that locus (Visscher 2008). Thus, phenotype-genotype

83   associations for characters controlled by a large number of small-effect loci tend to be more

84   difficult to detect.

85       Increased throughput of image-based phenotyping methods can make it possible to

86   collect measurements of more traits on more individuals than by manual measurement, which

87   makes image-based phenotyping an attractive way to generate phenotypic data for GWAS. We

88   can consider the manual measurement and the image-based measurement of an individual

89   character to be two traits with differing heritability but the same exact set of causative loci. As in

90    the NIRS example above, researchers may sometimes prefer a less precise method for

91    measuring a character because it is cheaper, faster, or otherwise more desirable.  It is unclear

92    just how much loss in precision (decrease in heritability) can occur before GWAS results begin

93    to suffer.  Part of the answer to this question lies in the goals and risk tolerance of the

94    researcher: if the goal of the experiment requires identification of few, strong signals then

95    perhaps lower power to detect associations will still be tolerable; if instead the goal is to identify

96    many small-effect loci, then even small reductions in heritability could negatively impact the

97    outcome of the study.  In this experiment, we use simulations to investigate the relationship

98    between trait heritability and the ability to detect genetic regions associated with a character.

99    We use receiver operating characteristic (ROC) curves to characterize detection of causative loci.

100   Previous studies have used ROC curves or similar visual aids to evaluate the efficacy of different

101   GWAS methods (e.g., Wang *et al.* 2014; Liu *et al.* 2016). In this study, however, we use ROC

102   curves to evaluate GWAS results for simulated traits, and test for differences between those ROC

103   curves using the area under the curve (AUC).  Using AUC to distill GWAS results to a single

104   statistic enables direct comparison of GWAS results from traits with differing simulation

105   parameters.  We use a test statistic for differences between two AUCs to construct a null

106   distribution from simulated traits with the same parameters, and use that distribution to predict

107   whether real traits measured manually and by image analysis have significantly different AUCs.

108   These results provide a framework for evaluating how differences in heritability between two

109   measures of a character can impact the efficacy of GWAS for identifying loci associated with the

110   character of interest.

111

112   **Results**

113   Manual and image-based phenotypes are highly correlated

114        Empirical phenotypic data used in this study comes from morphological measurements

115   of the male inflorescence of maize, known as the tassel.  Four tassel morphological characters

116   were measured manually and by image analysis in the Wisconsin Diversity panel, a population

117   of 942 diverse inbred maize lines (WiDiv-942).  The manually measured traits were tassel length

118   (TL), spike length (SL), branch number (BN), and tassel weight (TW), and their image-based

119   counterparts are referred to as TLp, SLp, BNp, and TWp, respectively.  TL, SL, and BN were

120   measured in replicated experiments over three years, while the other five traits were measured

121   in a replicated experiment in one year.  Best linear unbiased predictors (BLUPs) for manually

122   measured traits are highly correlated with BLUPs for the corresponding image-based traits, with

123   Pearson's correlation coefficients ranging from 0.81 to 0.9 (Figure 1) (Gage *et al.* 2018).

124   Estimated heritability for the traits ranges from 0.95 to 0.97 for manually measured traits and

125   from 0.79 to 0.86 for image-based traits (Table 1) (Gage *et al.* 2018).  TL, SL, and BN were

126   measured in three environments, and all image-based traits as well as TW were measured in a

127   single environment.  The difference in number of environments could inflate the heritability

128   estimates of the image-based traits, making it reasonable to conclude that the image-based

129   heritability estimates represent an upper bound for their true heritabilities.  As such, it is

130   reasonable to conclude that image-based measurements are less precise than the manual

131   measurements.

132

133   Power in GWAS of simulated traits varies with heritability and number of causal loci

134        To create a framework for comparing traits measured manually and by image analysis,

135   we first performed simulations to examine the impact of heritability and number of causative

136   loci (NCL) on GWAS results.  We first simulated a number of traits with different heritability

137    and NCL.  Phenotypes were simulated as being controlled by single nucleotide polymorphisms

138    (SNPs) from the WiDiv-942, which was genotyped at 529,018 SNPs discovered by RNA

139    sequencing.  Simulated phenotypes were controlled by varying NCL: 10, 100, or 1,000 randomly

140    selected SNPs were randomly assigned effect sizes drawn from a normal distribution.  For each

141    of the three values for NCL, traits were simulated with heritabilities ranging from 0.1 to 0.9 in

142    increments of 0.1. Each combination of NCL and heritability was simulated ten times.  GWAS

143    were performed on all simulated phenotypes, and empirical ROC curves were created with the

144    results from each GWAS.  ROC curves plot the proportion of true positives (true positive rate;

145    TPR) against the proportion of false positives (false positive rate; FPR), as the threshold for

146    labeling an observation as positive moves from stringent to more liberal.  The ROC curve for a

147    test with very good ability to identify true positives without too many false positives will rise

148    steeply from the origin and approach the point (0, 1), before flattening out and continuing on to

149    the point (1,1), producing an AUC close to 1.  A test that is no better than randomly guessing

150    which observations are positives will yield an ROC curve that follows a line with slope equal to

151    one from the origin to (1,1), producing an AUC of 0.5.

152        ROC curves are typically constructed by classifying a number of individuals as either

153    cases or controls, based on some continuous predictor variable.  The TPR and FPR are

154    calculated at different levels of the predictor variable to create the curve.  To create ROC curves

155    for GWAS results each SNP is treated as an individual, the true status of which is either

156    causative (case) or non-causative (control).  The continuous predictor variable is the $-\log_{10}$(p-

157    value) for each SNP from GWAS.

158        Our empirical results show that for any given NCL, simulated traits with higher

159    heritability generally had better ROC curves, as measured by AUC (Figure 2).  This was

160    expected, as greater heritability implies greater genetic variance relative to error, which makes it

161    easier to identify associations between phenotypic values and genotypic groups at causal loci.

162    However, higher heritability does not guarantee better ROC curves as there are ROC curves with

163    different heritability that intersect, particularly when the NCL is low (Figure 2).

164        For any given heritability, the NCL plays an even larger role in the shape of the ROC

165    curve, with traits controlled by more loci having worse ROC curves than those with fewer loci

166    (Figure 2).  This result was also expected.  The effect sizes of individual loci become smaller as

167    NCL increases, making detection of true associations more difficult.

168

169    <u>Heritability and number of causal loci influence ability to detect differences between ROC curves</u>

170        The AUC can be interpreted as the probability that, given a randomly selected pair of one

171    causative and one non-causative SNP, the predictor variable [$-\log_{10}$(p-value)] for the causative

172    SNP will be greater than or equal to the predictor variable for the non-causative SNP (DeLong *et*

173    *al.* 1988).  AUCs provide a way to describe an ROC curve with a single value, and can be used for

174    testing differences between ROC curves. The empirical AUC of an ROC curve is equivalent to the

175    statistic generated by a Mann-Whitney test on the predictor scores of the cases and controls, or

176    in the context of this study, the causative and non-causative SNPs. Therefore, treating the AUCs

177    of two empirical ROC curves as Mann-Whitney statistics permits non-parametric testing of the

178    difference between two AUCs, the test statistic of which (Z) follows a standard normal

179    distribution (DeLong *et al.* 1988).  However, because we cannot be certain that the GWAS

180    results from this study satisfy the assumptions of a Mann-Whitney test or the assumptions for

181    testing the difference between AUCs, we still use the Z statistic but do not make the assumption

182    that the distribution of Z is normal.  We assume that ROC curves created from GWAS results on

183    traits with the same parameters (i.e., NCL and heritability) should not be significantly different

184    from each other.  Therefore, the empirical distribution of Z when the difference in heritability

185    (D) between two traits equals zero is the distribution of Z under the null hypothesis that the

186    AUCs of two ROC curves are the same.

187    We hypothesized that as D between two traits increased so would the Z score,

188    corresponding to a difference between AUCs of the two traits. We tested all 90 AUCs with the

189    same NCL (10 replications times nine levels of heritability) against each other, resulting in 4,005

190    test statistics for each NCL (comparisons were only made in one direction). The goal of this

191    study is to assess how differing heritability of two measurements of the same underlying

192    character affects GWAS results. Because manual and image-based measurements of a character

193    have equivalent underlying genetic structure, we limited our comparisons to AUCs of simulated

194    traits with the same NCL.

195    The Z scores for each pairwise test of two traits were regressed against D (Figure 3). As

196    expected, the Z values get more extreme as D gets larger – this is a reflection of higher

197    heritability traits generally having larger AUCs than lower heritability traits. Note that the tests

198    were always done in a consistent direction; therefore we mostly observed results with positive Z

199    scores. The relationship between D and Z is more extreme for simulated traits with greater

200    NCL. Practically, this indicates that within the assumptions of these simulations, heritability

201    plays a smaller role in the ability to detect GWAS associations when the trait is controlled by a

202    small NCL. ROC curves for traits with more complex genetic architectures, however, deteriorate

203    more quickly as heritability declines.

204

205    Alternative measurements of real phenotypes are not predicted to have differing AUCs

206    Having established an empirical relationship between Z and D for different NCL, we then

207    used the results from our simulations to predict whether there is a significant difference

208    between the AUCs of manual and image-based traits of a real character. We used the estimated

209    heritabilities of manual and image-based measurements of TL, SL, BN, and TW to predict

210    whether AUCs for the two measurement methods will be significantly different. We fit a

211    regression between Z scores and D values for each different NCL and used that regression to

212    obtain a predicted value of Z for each trait pair based on D estimates for manual measurements

213    and image-based measurements.  The estimates of D for real trait pairs were 0.1 (TW), 0.14

214    (BN), 0.16 (TL), and 0.17 (SL), with the manually measured trait always having higher

215    heritability than the image-based trait (Table 1).  The predicted values of Z for each trait ranged

216    from 0.15 to 0.26 when NCL=10, from 0.47 to 0.75 when NCL=100, and from 0.70 to 1.10 when

217    NCL=1,000 (Figure 3).  For each NCL we used the distribution of Z scores when D=0 and set the

218    2.5$^{th}$  and 97.5$^{th}$  percentiles as thresholds for significance to test the null hypothesis of  Z=0 at

219    α=0.05. The thresholds were (-1.26, 1.09) for NCL=10, (-1.49, 1.25) for NCL=100, and (-1.44,

220    1.60) for NCL=1,000.  Regardless of NCL, predictions of Z for all four tassel traits fall within the

221    thresholds for significance (Figure 3).  Therefore, under the assumptions made in these

222    simulations the manual and image-based measurements are expected to have AUCs that are not

223    significantly different from each other.

224

## Discussion

226        In this study, we use AUCs of ROC curves constructed from GWAS results of simulated

227    traits to test for significant differences in the ability to detect common genetic signal underlying

228    traits with differing heritability.  Our results show that as D increases, the test for differences

229    between the traits' AUCs becomes more significant.  Though there is a strong relationship

230    between D and Z, there is also substantial variability for Z scores at a given value of D.  We

231    predicted Z scores for real tassel morphological traits using the relationship between D and Z of

232    simulated traits.  Because each tassel trait was measured by manual and image-based methods,

233    we predicted Z using D from the estimated heritabilities of the two measurement methods.

234    Regardless of NCL, the predicted values of Z for real tassel traits were within the thresholds for

235    significance that were calculated from the null distribution of Z.  Based on these results, we

236  conclude that there is unlikely to be a significant difference between AUCs of measurements

237  made by different methods for any of the four tassel morphological phenotypes studied.

238      This conclusion is highly dependent on the assumptions made when creating simulated

239  traits and performing subsequent GWAS.  Here, we assumed independent and randomly

240  positioned causal SNPs, when in fact quantitative traits can be controlled in part by numerous

241  causal variants clustered on the same locus (Allen *et al.* 2010).  Additionally, by choosing causal

242  SNPs directly from the WiDiv-942 genotypic data, we are assuming that the causal variants are

243  SNPs that are part of our genotypic data.  The 529,018 SNPs used in this study are a small

244  sample relative to the 60 million variants identified in the maize HapMap3 (Bukowski *et al.*

245  2015), and do not include other variant types that can affect quantitative traits such as

246  insertion/deletion and copy number variants.  By selecting causative SNPs from our genotyped

247  SNPs, we make the true associations easier to find by GWAS.  In reality, causative variants may

248  not be genotyped and therefore can only be identified by linkage disequilibrium with genotyped

249  SNPs.  For simplicity's sake, we drew the simulated effect sizes at each causal SNP from a

250  normal distribution.  Previous work by Hayes and Goddard (2001) has shown that quantitative

251  traits in livestock appear to follow a gamma distribution with a large number of very small effect

252  loci; they posit that there may be even more small effect loci than predicted by their

253  distributions.  This idea can be seen in its most extreme form in the omnigenic model proposed

254  by Boyle et al. (Boyle *et al.* 2017) which is based on Fisher's infinitesimal model (Fisher 1919).

255  By drawing our effect sizes from a normal distribution, we may be creating more large-effect

256  variants than is realistic, therefore increasing our ability to detect causal variants by GWAS.

257      Though our choices for location and effect size of causative SNPs may be increasing the

258  probability of detecting associations, we also assume that only identifying an exact chosen

259  causative SNP counts as a true positive.  In reality, identifying associations with SNPs that are

260  within the same gene as, or a small distance away from, the true causative SNP may be close

261   enough.  GWAS often serves as an initial sweep to find regions of interest for further study, and

262   associations that lead to fruitful downstream analysis may still be considered a 'success'.  This is

263   reflected in software that calculate power of GWAS by considering associations within a certain

264   distance of the causative variant to be true positives (Liu *et al.* 2016).  By only considering the

265   exact causative SNPs as true positives we make the true positives more difficult to identify,

266   partially counteracting the assumptions above that make the detection of associations easier.

267          By trying to predict the Z score for the difference between AUCs of real traits, we also

268   make some assumptions about the tassel morphological traits that we use.  Our estimates of

269   heritability are not exact; they are population- and experiment-specific.  Because the image-

270   based traits and TW were measured in one environment, whereas TL, SL, and BN were

271   measured in three, their estimates of heritability may have differing accuracy.  By predicting Z

272   scores for the real traits for NCL set to 10, 100, and 1,000, we were able to predict how Z scores

273   changed as NCL changed.  The true NCL for tassel morphological traits likely numbers in the

274   hundreds or higher, with an upper limit of the total number of expressed genes (tens of

275   thousands), as posited in the omnigenic model (Boyle *et al.* 2017).  If the true NCL is greater

276   than 1,000, the relationship between Z and D will be even steeper, meaning that the small values

277   of D for the real tassel morphological traits may in fact be responsible for significantly different

278   AUCs.

279          The use of AUC as a metric to quantify the success of GWAS is also accompanied by

280   assumptions about the goals of GWAS.  ROC curves, and thus the AUC, consider both power and

281   type I error, as measured by true and false positive rates.  Depending on the goals of the GWAS

282   study, power and type I error may not both be of equal importance.  For genomic prediction or

283   marker assisted selection, a high type I error rate is not particularly concerning as long as power

284   is high and trait prediction is accurate.  On the other hand, studies using GWAS to choose genes

285   for further molecular characterization have a large financial incentive to minimize type I error.

286    Using AUC to quantify the effectiveness of GWAS assumes that the entire ROC curve is of

287    interest.  When NCL is low this assumption may be true, but as NCL increases, it may be the

288    case that only the beginning of the ROC curve is of practical interest.  The simulated ROC curves

289    for NCL=1,000 (Figure 2c) are close to the 1:1 line that would be achieved by randomly selecting

290    SNPs as putatively causative.  It is unlikely that a researcher would want or expect to identify

291    every single causative locus when a trait is controlled by thousands of genes.  Instead, the

292    interest is often in large-effect loci that are likely to be identified by a stringent significance

293    threshold.  The ability to identify the most significant loci is characterized by the portion of the

294    ROC curve close to the origin.  Thus, for highly complex traits, the AUC of a partial ROC may be

295    more informative.

296    In this study, we use AUC of ROC curves to characterize and quantitatively compare

297    GWAS results from different traits.  Overall, our findings show an expected relationship between

298    NCL, heritability, and AUC of ROC curves.  Greater NCL and lower heritability both reduce the

299    AUC, while lower NCL and higher heritability can increase AUC.  Results suggest that there is no

300    significant difference between AUCs from GWAS using manual and image-based measurements

301    of typical maize tassel characters.  Creation of more nuanced simulation models and

302    consideration of partial ROC curves may enable improvement upon the results presented in this

303    study.  The results presented here provide a foundational framework that may facilitate

304    decision-making for researchers weighing the benefits of different phenotyping methods.

305

306    **Materials and Methods**

307    Plant populations: phenotyping and genotyping

308    The Wisconsin Diversity panel (WiDiv-942) is a set of 942 inbred maize lines that reach

309    grain physiological maturity in the upper Midwest region of the United States (Mazaheri *et al.*

310    2018, in press), and represents an expanded version of the 503 line diversity panel described by

311  Hirsch and colleagues (Hirsch *et al.* 2014). Phenotypic measurements of tassel morphology in

312  the WiDiv-942 were performed using both manual and image-based measurements. Manual

313  and image-based measurements are described in detail in (Gage *et al.* 2018). Briefly, manual

314  measurements included tassel length (TL), the distance (cm) from the lowest tassel branch to

315  the tip of the tassel spike; spike length (SL), the distance (cm) from the uppermost tassel branch

316  to the tip of the tassel spike; branch number (BN), the total number of primary tassel branches;

317  and tassel weight (TW), the weight (g) of the dried tassel biomass above and including the

318  lowest branch. Image-based measurements were made using the output of the tassel

319  phenotyping software TIPS (Gage *et al.* 2017). The data output by TIPS are image-based

320  measurements of tassel morphology and were used as explanatory variables in a partial least

321  squares regression model that performs image-based predictions of TL, SL, BN, and TW

322  (referred to as TLp, SLp, BNp, and TWp, respectively) (Gage *et al.* 2018). The WiDiv-942 was

323  grown using a replicated complete block design in three different environments: the University

324  of Wisconsin Arlington Agricultural Research Station in the summers of 2013 and 2014, and the

325  University of Wisconsin West Madison Agricultural Research Station in the summer of 2015.

326  TL, SL, and BN were measured in all three environments, while TW, TLp, SLp, BNp, and TWp

327  were measured only in 2015. Best linear unbiased predictors (BLUPs) for each inbred line for all

328  eight traits were calculated using random effects models to account for environment, genotype-

329  by-environment, and replication effects (Gage *et al.* 2018).

330  The WiDiv-942 was genotyped at 899,784 SNPs discovered by RNA sequencing

331  (Mazaheri *et al.* 2018, in press). SNP data contained 30% missing data, which were imputed

332  using fastPHASE (Scheet and Stephens 2006). After imputation, 0.3% of SNP calls were

333  missing, due to inability of the imputation program to call all missing SNPs. The remaining

334  missing data at any given SNP were imputed to the mean value for that SNP. SNPs with a minor

335  allele frequency of <0.02 were removed from the genotypic data, leaving 529,018 remaining

336  SNPs.

337

338    <u>Simulated phenotypes</u>

339    Heritability ($h^2$) is the ratio of genetic variability $\sigma_g^2$ in a population to overall phenotypic

340    variability $\sigma_p^2 = \sigma_g^2 + \sigma_\varepsilon^2$, where $\sigma_\varepsilon^2$ is all variance not attributed to differences between

341    genotypes. Thus, $h^2 = \frac{\sigma_g^2}{\sigma_p^2}$ is a measure of the strength of genetic signal in a particular population

342    for a particular phenotype. Two different measures of the same character, for instance, TL and

343    TLp, can be thought of as correlated traits with the same underlying genetic control but differing

344    heritabilities. The true value of the character for a particular individual cannot be measured

345    perfectly, but both the manual and image-based measurements of the character represent a

346    combination of true signal and some (differing) amount of measurement error.

347    To simulate traits with similar behavior, effect sizes were randomly drawn from a normal

348    distribution for a set of randomly chosen SNPs genotyped in the WiDiv-942, and the 'true'

349    phenotypic value was calculated for each of the 942 individuals. In a second step, noise was

350    added to each individual's true value in order to attain a desired heritability. We varied both the

351    number of causative loci (NCL), to simulate traits controlled by differing numbers of variants,

352    and the heritability of simulated traits. The NCL was set to 10, 100, and 1000, and $h^2$ ranged

353    from 0.1 to 0.9 in increments of 0.1. The causative loci were randomly selected a single time

354    from all SNPs genotyped in the WiDiv-942. The phenotype for an individual $i$ for a trait

355    controlled by NCL=$n$ SNPs and heritability $h^2$ is $y_i = \sum_{l=1}^{n} w_{il} a_l + \varepsilon_i$. The standardized

356    genotypic value, $w_{il} = \frac{g_{il} - 2p}{\sqrt{2p(1-p)}}$ is the genotype $g_{il}$ of individual $i$ at SNP $l$, expressed as 0, 1, or 2

357    copies of the major allele, centered by twice the major allele frequency, $p$, and divided by the

358    standard deviation of the SNP. The allelic effect $a_l$ is drawn from $N(0, 10)$, and $\varepsilon_i$ from $N(0, \sigma_\varepsilon^2)$

359    where $\sigma_\varepsilon^2 = \left(\frac{\sigma_g^2}{h^2}\right) - \sigma_g^2$. The variance for allelic effects was set arbitrarily, but could be any

360    reasonable number as the error variance is modified to ensure the desired heritability.

361     Genotypic variance $\sigma_g^2$ was calculated simply as the sample variance of the population's true

362     phenotypic values. For each combination of $n$ and $h^2$, phenotypes were simulated 10 times. The

363     difference between each simulation with the same set of parameters is simply the $\varepsilon_i$,

364     representing random draws from the same distribution each time. In total, 270 traits were

365     simulated: pairwise combinations of 3 levels of $n$ and 9 levels of $h^2$, each replicated 10 times.

366

367     Genome-wide association studies

368         Genome-wide association studies (GWAS) were conducted with the software GAPIT

369     (Lipka *et al.* 2012), implemented in R (R Core Team 2016). The same kinship matrix, calculated

370     from 10,000 randomly selected SNPs by the VanRaden method in GAPIT (VanRaden 2008;

371     Lipka *et al.* 2012), was used for all GWAS and no other covariates (e.g., principal component

372     scores) were included in the model. Compression was turned off by setting the group.to and

373     group.from parameters to 9999. GWAS was run separately for each of the 270 simulated traits.

374

375     Receiver operator characteristic curves

376         Receiver operator characteristic (ROC) curves are typically used to assess the ability of a

377     particular method for identifying the true, binary status (case or control) of an individual based

378     on various threshold levels of a continuous predictor variable. The *roc()* function in the R

379     package pROC (Robin *et al.* 2013) was used to construct ROC curves from GWAS results by

380     considering each SNP as an 'individual' and coding the randomly selected causative SNPs as

381     cases in the response, while all other SNPs were considered controls. The $-\log_{10}$(p-value) of

382     each SNP was used as the predictor variable. A separate empirical ROC curve was fitted to the

383     GWAS results for each of the simulated traits. A representative ROC curve was estimated from

384    all 10 replicates of each parameter combination by calculating the mean of the sensitivities and

385    specificities along the curve.

386         The area under the curve (AUC) for an empirical ROC curve is equivalent to a Mann-

387    Whitney statistic calculated on the predictor scores of the case and control individuals (DeLong

388    *et al.* 1988).  In addition, the AUC can be interpreted as the probability that for any pair of

389    randomly selected case and control individuals (or in this case, causative and non-causative

390    SNPs), the predictor value of the control individual will be less than or equal to the predictor

391    value of the case individual (DeLong *et al.* 1988).  In this context, that is the probability that the

392    $-\log_{10}$(p-value) of the non-causative SNP is less than or equal to the $-\log_{10}$(p-value) of the

393    causative SNP.  Because the empirical AUC of an ROC curve acts as a Mann-Whitney statistic,

394    the difference between AUCs of two ROC curves can be nonparametrically tested (DeLong *et al.*

395    1988), yielding a statistic we refer to as Z.  Because our Z scores were not standard-normally

396    distributed as expected (DeLong *et al.* 1988), we instead used the values of Z when the

397    difference in heritability (D) between traits was 0 as a null distribution for a given NCL.

398    Thresholds for significance at α=0.05 were calculated empirically from the 2.5[th] and 97.5[th]

399    percentiles of the null distribution of Z.

400         Z-scores were calculated for tests of the difference between all pairwise combinations of

401    ROC curves with the same NCL using the *roc.test()* function in pROC with the method argument

402    set to 'delong' (Robin *et al.* 2013).

403

404    <u>Estimating Z scores and confidence intervals for real phenotypes</u>

405         Because the relationship between Z and D for pairs of simulated traits was approximately

406    linear, we fit a simple linear regression of Z against D separately for each value of NCL.  Using

407    the differences between estimated heritability for each pair of real traits (TL and TLp; SL and

408    SLp; BN and BNp; TW and TWp), we predicted the value of Z for each character at different

409    NCL using the fitted regressions.  Using the thresholds for significance derived from the

410    empirical distribution of Z when D=0, we tested the null hypothesis Z=0 for each character at

411    each NCL.

412

413    Data Availability

414         Genotypic data are available from doi: 10.5061/dryad.n0m260p (paper still under

415    review; doi not publicly available as of submission).  Scripts for analysis can be found at

416    github.com/joegage/GWAS_AUC

417

418    **Acknowledgements**

419    Funding

424

425    Authors' contributions

426         All authors contributed ideas, and all authors were involved in writing, editing, and

427    approving the final manuscript.

428

429

430     **Figures**

431     Figure 1: Correlations between manual and image-based phenotypic values



432
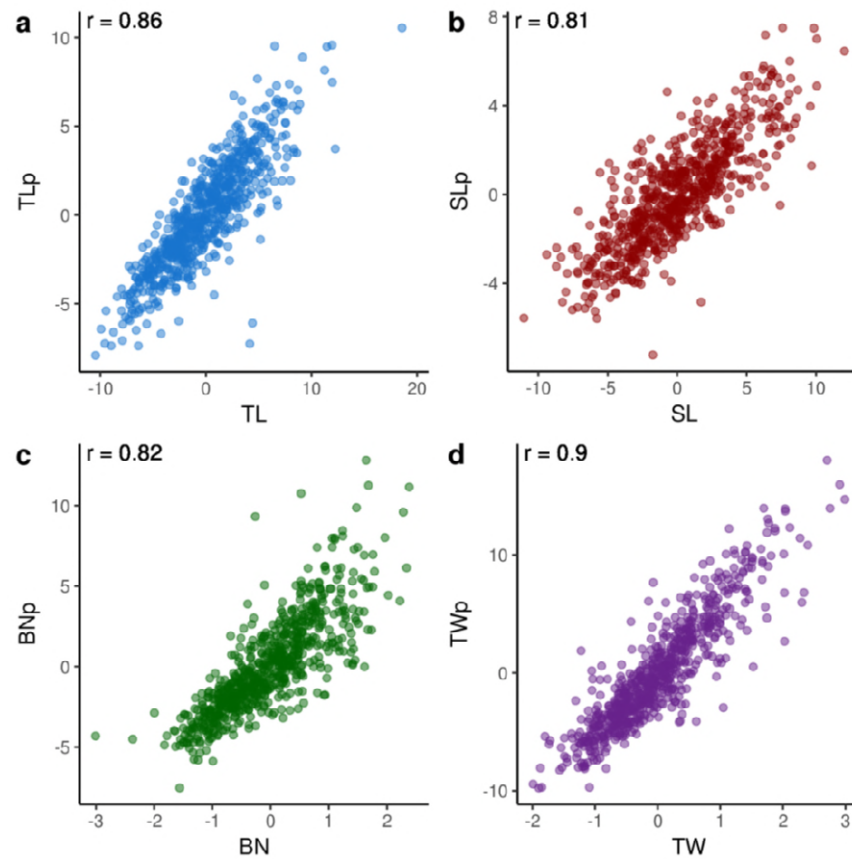
433          Scatter plots of best linear unbiased predictors (BLUPs) for manual vs image-based

434     measurements of tassel length (TL; a), spike length (SL; b), branch number (BN; c), and tassel

435     weight (TW; d).  Manually measured BLUPs are along the x-axis, while image-based

436     measurements are on the y-axis.  Values in the upper left corner of each plot are the Pearson

437     correlation coefficients for each trait.

438

439      Figure 2: Receiver operating characteristic curves for GWAS of simulated traits
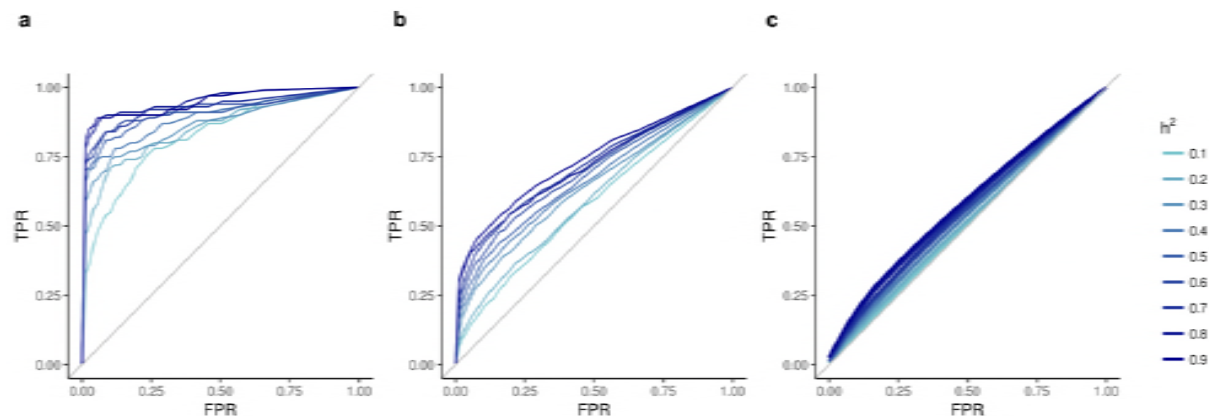


440

441          Receiver operating characteristic curves for GWAS results of traits controlled by 10 (a),

442    100 (b), or 1000 (c) causal loci.  For each number of causal loci, simulation of traits with

443    heritabilities ranging from 0.1 to 0.9 were replicated 10 times each.  Each curve represents the

444    average of the ten replications for each combination of causal loci and heritability.  TPR: true

445    positive rate; FPR: false positive rate; $h^2$: heritability.

446

447      Figure 3 Results of testing area under the curve for simulated phenotypes



448
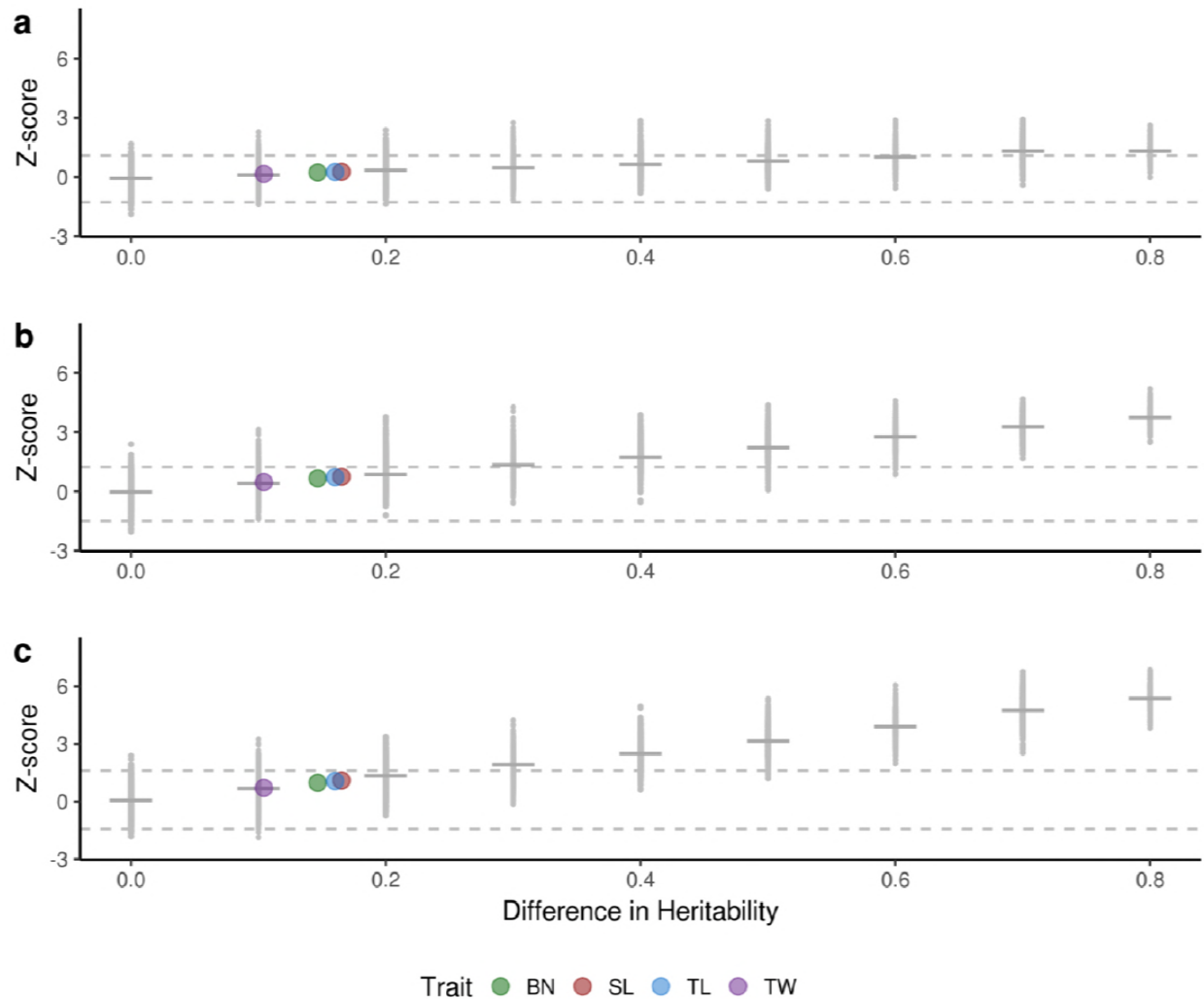
449      The Z score for testing the difference of two AUCs plotted against the absolute difference

450      in heritability (D) between the two traits.  Small gray dots represent the Z score from a single

451      pairwise test between simulated traits, while horizontal gray bars represent the median Z score

452      for a given D.  Larger colored dots represent the D estimates for real traits, plotted along the line

453      that best fits the Z scores of the simulated data. Dashed lines represent the thresholds for

454      significance at α=0.05 (i.e., 2.5[th] and 97.5[th] percentiles), calculated from the empirical

455      distribution of Z when D=0.  BN: branch number; SL: spike length; TL: tassel length; TW: tassel

456      weight.

457

## Tables

459   Table 1: Comparison of manual and image-based trait heritabilities

| Trait Name (unit) | Abbreviation | | Heritability | |
|---|---|---|---|---|
| | Manual | Image-Based | Manual | Image-Based |
| **Tassel Length (cm)** | TL | TLp | 0.95 | 0.79 |
| **Spike Length (cm)** | SL | SLp | 0.95 | 0.79 |
| **Branch Number (count)** | BN | BNp | 0.97 | 0.82 |
| **Tassel Weight (g)** | TW | TWp | 0.96 | 0.86 |

460        Heritabilities for four different tassel morphological traits, measured both manually and

461   using image-based methods.  TL, SL, and BN were measured in three environments, whereas

462   TW, TLp, SLp, BNp, and TWp were measured in one environment.

463

## References

Allen, H. L., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Bock, C. H., G. H. Poole, P. E. Parker, and T. R. Gottwald, 2010 Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. CRC. Crit. Rev. Plant Sci. 29: 59–107.

Boyle, E. A., Y. I. Li, and J. K. Pritchard, 2017 An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169: 1177–1186.

Bukowski, R., X. Guo, Y. Lu, C. Zou, B. He *et al.*, 2015 Construction of the third generation Zea mays haplotype map. bioRxiv 026963.

Chitwood, D. H., A. Ranjan, C. C. Martinez, L. R. Headland, T. Thiem *et al.*, 2014 A Modern Ampelography: A Genetic Basis for Leaf Shape and Venation Patterning in Grape. Plant Physiol. 164: 259–272.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson, 1988 Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44: 837.

Durham Brooks, T. L., N. D. Miller, and E. P. Spalding, 2010 Plasticity of Arabidopsis Root Gravitropism throughout a Multidimensional Condition Space Quantified by Automated Image Analysis. PLANT Physiol. 152: 206–216.

Fisher, R. A., 1919 The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinburgh 52: 399–433.

Gage, J. L., N. D. Miller, E. P. Spalding, S. M. Kaeppler, and N. de Leon, 2017 TIPS: a system for automated image-based phenotyping of maize tassels. Plant Methods 13: 21.

Gage, J. L., M. R. White, J. Edwards, S. Kaeppler, and N. de Leon, 2018 Selection signatures underlying dramatic male inflorescence transformation during modern hybrid maize breeding. bioRxiv.

Hayes, B., and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33: 209.

Hirsch, C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni *et al.*, 2014 Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26: 121–35.

Li, M., K. Duncan, C. N. Topp, and D. H. Chitwood, 2017 Persistent homology and the branching topologies of plants. Am. J. Bot. 104: 349–353.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li *et al.*, 2012 GAPIT: Genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Liu, X., M. Huang, B. Fan, E. S. Buckler, and Z. Zhang, 2016 Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome- Wide Association Studies. PLoS Genet. 12: 1–24.

502 Men, Y., Q. Yu, Z. Chen, J. Wang, Y. Huang *et al.*, 2012 A high-throughput imaging system to
503    quantitatively analyze the growth dynamics of plant seedlings. Integr. Biol. 4: 945.

504 Miller, N. D., N. J. Haase, J. Lee, S. M. Kaeppler, N. de Leon *et al.*, 2016 A robust, high-
505    throughput method for computing maize ear, cob, and kernel attributes automatically from
506    images. Plant J. n/a–n/a.

507 Park, R. ., R. . Agnew, F. . Gordon, and R. W. . Steen, 1998 The use of near infrared reflectance
508    spectroscopy (NIRS) on undried samples of grass silage to predict chemical composition
509    and digestibility parameters. Anim. Feed Sci. Technol. 72: 155–167.

510 R Core Team, 2016 R: A Language and Environment for Statistical Computing.

511 Robin, X., N. Turck, A. Hainard, F. Lisacek, J. Sanchez *et al.*, 2013 Package " pROC ." 2012-09-
512    10 09:34:56 1–71.

513 Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population
514    genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J.
515    Hum. Genet. 78: 629–44.

516 VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91:
517    4414–23.

518 Visscher, P. M., 2008 Sizing up human height variation. Nat. Genet. 40: 489–490.

519 Wang, Q., F. Tian, Y. Pan, E. S. Buckler, and Z. Zhang, 2014 A SUPER powerful method for
520    genome wide association study. PLoS One 9.:

521