COLLOQUIUM PAPER

Linear mixed model for heritability estimation that explicitly addresses environmental variation

David Heckerman[a,1], Deepti Gurdasani[b,c], Carl Kadie[d], Cristina Pomilla[b,c], Tommy Carstensen[b,c], Hilary Martin[b], Kenneth Ekoru[b,c], Rebecca N. Nsubuga[e], Gerald Ssenyomo[e], Anatoli Kamali[e], Pontiano Kaleebu[e], Christian Widmer[a], and Manjinder S. Sandhu[b,c]

[a]Microsoft Research, Los Angeles, CA 90024; [b]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom; [c]Department of Medicine, University of Cambridge, Cambridge CB2 0SP, United Kingdom; [d]Microsoft Research, Redmond, WA 98052; and [e]MRC/UVRI Uganda Research Unit on AIDS, Uganda

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 8, 2016 (received for review July 9, 2015)

The linear mixed model (LMM) is now routinely used to estimate heritability. Unfortunately, as we demonstrate, LMM estimates of heritability can be inflated when using a standard model. To help reduce this inflation, we used a more general LMM with two random effects—one based on genomic variants and one based on easily measured spatial location as a proxy for environmental effects. We investigated this approach with simulated data and with data from a Uganda cohort of 4,778 individuals for 34 phenotypes including anthropometric indices, blood factors, glycemic control, blood pressure, lipid tests, and liver function tests. For the genomic random effect, we used identity-by-descent estimates from accurately phased genome-wide data. For the environmental random effect, we constructed a covariance matrix based on a Gaussian radial basis function. Across the simulated and Ugandan data, narrow-sense heritability estimates were lower using the more general model. Thus, our approach addresses, in part, the issue of "missing heritability" in the sense that much of the heritability previously thought to be missing was fictional. Software is available at https://github.com/MicrosoftGenomics/FaST-LMM.

heritability estimation | linear mixed model | environment | Gaussian radial basis function | model misspecification

An important causal question comes from the age-old debate about nature versus nurture. For any phenotype such as height or intelligence quotient, how much of the phenotype is inherited and how much is determined by environment? This question was made precise by Fisher (1) and Wright (2) almost a century ago: Given observations of a phenotype from a population of individuals, what is the fraction of variance of the phenotype that is caused by inherited factors relative to the total variance of the phenotype due to both inherited and environmental factors? This fraction, termed "heritability," has been the subject of intense study across various phenotypes and populations since it was defined. Note that, in contrast to how some interpret the informal question around the nature-versus-nurture debate, heritability is not an absolute quantity but rather a quantity relative to a given population. For example, a phenotype in a population where environmental factors have large variation will have a smaller heritability than in an otherwise similar population where environmental factors have a small variation.

Over the years, many approaches have been developed to estimate heritability from data (3, 4). Here, we concentrate on an approach made possible by the recent ability to sequence genomes at a modest cost (5, 6). The approach uses a linear mixed model (LMM), a form of multivariate regression of the genomic and environmental factors on the phenotype, which we examine in detail in the next section.

In the standard LMM approach, the effects of environmental factors on the phenotype are modeled as noise. Specifically, the phenotype of each individual is assumed to be the sum of two random effects, one based on genomic factors and one based on environmental factors, where the latter is assumed to be mutually independent across individuals. As we shall see, this model for environmental effects can lead to inflated estimates of heritability.

To avoid this inflation, we could measure and model environmental effect explicitly (e.g., ref. 7). Unfortunately, in most circumstances there are many environmental factors to be measured. Furthermore, some environmental factors may be unrecognized and consequently are unmeasurable. In this work, we investigate the use of an easy-to-measure surrogate for environmental factors—namely, spatial location. We show how this surrogate can be incorporated into the LMM as an additional random effect. We investigate our more general model with simulated data and with data from a Ugandan cohort of about 5,000 individuals.

Results

Heritability Estimation. First, let us consider a standard approach for estimating heritability using an LMM (6, 8). The estimate is based on observations consisting of $\boldsymbol{y}$, an $N \times 1$ vector of phenotypes for the N individuals, and $\mathbf{X}$, an $N \times M$ matrix of causal genomic variants for the N individuals and M variants. Note that it is customary to normalize the causal variants so that each one has a mean of zero and an SD of one across individuals. Given these observations, we model $\boldsymbol{y}$ as a multivariate linear regression on $\mathbf{X}$:

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta}; \sigma_r^2 \mathbf{I}\right), \quad [1]$$

where $\boldsymbol{\mu}$ is an $N \times 1$ vector of offsets that can include the effects of covariates, $\boldsymbol{\beta}$ is the $M \times 1$ vector of linear weights relating the corresponding variants to the phenotype, $\mathbf{I}$ is the $N \times N$ identity matrix, and σ_r^2 is the residual variance of the multivariate normal distribution denoted by $\mathcal{N}(.;.)$. In addition, we assume that the elements of $\boldsymbol{\beta}$ are mutually independent, each having a normal distribution:

$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(0; \frac{\sigma_g^2}{M}\right), \; i = 1, \ldots, M. \quad [2]$$

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: D.H., D.G., C.K., C.W., and M.S.S. designed research; D.H., D.G., C.K., C.P., T.C., H.M., K.E., R.N.N., G.S., A.K., P.K., C.W., and M.S.S. performed research; D.H., C.K., and C.W. contributed new reagents/analytic tools; D.H., D.G., C.K., C.W., and M.S.S. analyzed data; and D.H., D.G., C.K., C.W., and M.S.S. wrote the paper.

Conflict of interest statement: D.H., C.K., and C.W. were employees of Microsoft Research while performing this research.

This article is a PNAS Direct Submission.

Data deposition: The genomic data have been deposited at the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) (accession no. EGAS00001001558).

[1]To whom correspondence should be addressed. Email: heckerma@microsoft.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1510497113/-/DCSupplemental.

Plugging Eq. **2** into Eq. **1** and integrating out $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}; \sigma_g^2 \frac{1}{M}\mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_r^2 \mathbf{I}\right). \quad [3]$$

Model **3** is known as a linear mixed model with a random effect having the covariance matrix $\mathbf{K}_{\mathbf{causal}} = \frac{1}{M}\mathbf{X}\mathbf{X}^{\mathrm{T}}$ (9). It is also known as a Gaussian process with a linear covariance (or kernel) function (10, 11). Note that element i,j of $\mathbf{K}_{\mathbf{causal}}$ is the dot product $\sum_{k=1}^{M} \mathbf{X}_{ik}\mathbf{X}_{jk}$. The parameters of this model are typically fit by maximizing the restricted maximum likelihood (REML) of the data.

Narrow sense heritability, denoted h^2, is the fraction of the variance of $\boldsymbol{y}$ due to the genomic component. Given this model and the assumption that genomic variants are mutually independent, it follows that

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_r^2}. \quad [4]$$

Note that narrow-sense heritability accounts only for additive genomic effects. Genomic effects can also exhibit nonlinear interactions among each other (known as epistasis) and exhibit dominance, neither of which is captured in the model of Eq. **3**. The term "heritability" without the modifier "narrow sense" is typically reserved for the quantity that includes all genomic effects. Herein, for simplicity, we will concentrate on the estimation of narrow-sense heritability although, as we mention later, our approach can be extended to estimate more general quantities.

In practice, we do not know which genomic variants are causal, so we use an approximation for $\mathbf{K}_{\mathbf{causal}}$. One commonly used approximation—and one we will use in this work—is $\mathbf{K}_{\mathbf{IBD}}$, where element (i,j) is the fraction of the genome shared identical by descent (IBD) among individuals i and j (6). That is, we use the model

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}; \sigma_g^2 \mathbf{K}_{\mathbf{IBD}} + \sigma_r^2 \mathbf{I}\right).$$

As noted in the introduction, the standard LMM represents environmental effects as simple Gaussian noise. Here, let us consider a more general model for environmental effects based on the spatial location of individuals. Specifically, consider the addition of a random effect with covariance matrix $\mathbf{K}_{\mathbf{loc}}$:

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}; \sigma_g^2 \mathbf{K}_{\mathbf{IBD}} + \sigma_e^2 \mathbf{K}_{\mathbf{loc}} + \sigma_r^2 \mathbf{I}\right). \quad [5]$$

Assuming the genomic variants and spatial locations are mutually independent, we get a new estimate for narrow-sense heritability given by

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2 + \sigma_r^2}. \quad [6]$$

This model also allows us to estimate the fraction of variance of $\boldsymbol{y}$ due to the location component, denoted e^2:

$$e^2 = \frac{\sigma_e^2}{\sigma_g^2 + \sigma_e^2 + \sigma_r^2}. \quad [7]$$

In our analysis of the Ugandan cohort, we use $\mathbf{K}_{\mathbf{loc}}(i,j) = \exp\{-(d_{ij}/\alpha)^2\}$, where d_{ij} is the distance between individuals i and j, and α is a scaling parameter. Intuitively, the inclusion of $\mathbf{K}_{\mathbf{loc}}$ captures the notion that individuals physically closer to each other are more likely to be influenced by the same environmental factors and hence more likely to have similar phenotypes. Using other types of proximity—for example, social proximity—is also possible, but here we consider only physical proximity. The exponential form we use for $\mathbf{K}_{\mathbf{loc}}(i,j)$ is known as a Gaussian radial basis function and is often used in spatial analyses (10, 11). (We also tried the radial basis function $\exp\{-(d_{ij}/\alpha)\}$ but found it difficult to estimate α accurately.) The parameter α can be thought of as the spatial range of the environmental effect. The larger the value for α, the larger the range or extent of the effect. As in the standard case, we fit all parameters, now including σ_e^2 and α, with REML.

Recall that the standard LMM follows from modeling the phenotype as a regression on genomic variants. Similarly, Eq. **5** can be interpreted as the result of modeling the phenotype as a multivariate regression on both genomic variants and spatial location. In particular, Mercer's theorem (10, 11) states that, if $\mathbf{K}(z_i,z_j)$ is a continuous symmetric positive semidefinite function to $\mathbb{R}$ from z_i and z_j each in a compact Hausdorff space, then there exists a set of functions $\phi_k(z)$, $k = 1,\ldots,\infty$, such that $\mathbf{K}(z_i,z_j)$ is equal to the dot product $\sum_{k=1}^{\infty} \phi_k(z_i)\phi_k(z_j)$. Identifying z_i as the spatial location of individual i and $\mathbf{K}(z_i,z_j)$ as element i,j in $\mathbf{K}_{\mathbf{loc}}$, it follows that the inclusion of $\mathbf{K}_{\mathbf{loc}}$ in Eq. **5** is equivalent to conditioning on spatial features $\phi_k(z_i)$, $i = 1,\ldots,N$, $k = 1,\ldots,\infty$. We note that the Gaussian radial basis function is guaranteed to be positive semidefinite.

Finally, let us consider nonlinear interactions between genomic and environmental components. We can model some of these interactions by introducing a third random effect to the LMM:

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}; \sigma_g^2 \mathbf{K}_{\mathbf{IBD}} + \sigma_e^2 \mathbf{K}_{\mathbf{loc}} + \sigma_{gxe}^2 \mathbf{K}_{\mathbf{GxE}} + \sigma_r^2 \mathbf{I}\right), \quad [8]$$

producing an estimate of the fraction of variance of $\boldsymbol{y}$ due to the interaction component given by

$$gxe^2 = \frac{\sigma_{gxe}^2}{\sigma_g^2 + \sigma_e^2 + \sigma_{gxe}^2 + \sigma_r^2}. \quad [9]$$

We use a particular form for $\mathbf{K}_{\mathbf{GxE}}$ where element i,j is the product of elements i,j from $\mathbf{K}_{\mathbf{causal}}$ and $\mathbf{K}_{\mathbf{loc}}$ (i.e., the Handamard product of $\mathbf{K}_{\mathbf{causal}}$ and $\mathbf{K}_{\mathbf{loc}}$). Using the nomenclature we have defined, it follows that element i,j of $\mathbf{K}_{\mathbf{GxE}}$ is given by $\sum_{k=1}^{M}\sum_{l=1}^{\infty} X_{ik}\phi_l(z_i)X_{jk}\phi_l(z_j)$. Consequently, inclusion of $\mathbf{K}_{\mathbf{GxE}}$ into the LMM is equivalent to conditioning on the features $\mathbf{X}_{ik}\,\phi_l(z_i)$, $i = 1,\ldots,N$, $k = 1,\ldots,M$, $l = 1,\ldots,\infty$. Product features such as these are often used to model nonlinear interactions (12). In our analysis of the Ugandan cohort, we use this instance of $\mathbf{K}_{\mathbf{GxE}}$, except we replace $\mathbf{K}_{\mathbf{causal}}$ with $\mathbf{K}_{\mathbf{IBD}}$ as an approximation.

We note that the standard model given by Eq. **3** is nested in the model given by Eq. **5**, which in turn is nested in the model given by Eq. **8**.

Heritability Analysis on Simulated Data. We first applied our approach to the analysis of simulated data. We generated from the Balding–Nichols model (13) with a 50:50 population ratio, a baseline minor allele frequency (MAF) sampled uniformly from [0.05, 0.5], and a value for Wright's F_{ST} equal to 0.1. We generated a spatial location for each individual by sampling randomly from one of two spherical Gaussian distributions with SD 625,000 and separation between Gaussian centers equal to $4 \times 625{,}000$. This procedure produced a distribution of spatial locations similar to the real data and satisfied an assumption underlying Eqs. **6** and **7** that genomic variants and spatial locations are independent. We next generated the phenotype using Eq. **5** with $\mathbf{K}_{\mathbf{IBD}}$ replaced with $\mathbf{K}_{\mathbf{causal}}$, and with $\sigma_g^2 = \sigma_e^2 = \sigma_r^2 = 1$, and $\alpha = 4 \times 625{,}000$. We generated 50 datasets and then, for each one, computed uncorrected and corrected heritability estimates, based on Eqs. **3** and **5**, respectively, with $\mathbf{K}_{\mathbf{IBD}}$ replaced with $\mathbf{K}_{\mathbf{causal}}$. For each dataset, we generated 1,000 causal SNPs for 5,000 individuals to mimic the real data.

The estimates of h^2 and e^2 based on the corrected model are unbiased, having mean ($\pm$ SE) of 0.33 $\pm$ 0.01 and 0.35 $\pm$ 0.02, respectively. In contrast, the estimates of h^2 based on the standard

or uncorrected model were inflated (0.42 ± 0.01). This inflation is not unexpected, because the "signal" produced by the spatial random effect needs to be accounted for by either the genomic random effect or the noise component, and there is no reason to expect that the noise component would account for all of it. Any leakage of the signal arising from the spatial random effect to the genomic random effect would yield inflated estimates of heritability. That is, model misspecification can lead to substantial bias in heritability estimates.

As mentioned, Eqs. **6** and **7** were derived under an assumption that genomic and spatial factors are independent. In practice, however, this assumption may not hold. To investigate the robustness of these estimates to nonindependence, we modified the above data-generation procedure to create a dependence between genomic and spatial variation. In particular, the spatial locations of all individuals from the same Balding–Nichols population were drawn from the same spherical Gaussian. Despite this relatively strong dependence, estimates of h^2 (corrected) and e^2 remained unbiased (0.33 ± 0.01 and 0.35 ± 0.02, respectively). Uncorrected estimates of h^2 were similarly inflated in the presence of dependence (0.46 ± 0.01).

Sample code for these experiments in the form of an iPython notebook can be found in *SI Appendix*.

Heritability Analysis on the Ugandan Cohort. We next applied our approach to an analysis of a Ugandan cohort (*Methods*) across 34 phenotypes including anthropometric indices, blood factors, glycemic control, blood pressure, lipid tests, and liver function tests. A description of phenotypes can be found in Table 1. Not unexpectedly, heritability estimates varied widely, with corrected estimates ranging from 0.55 for mean platelet volume (MPV) to 0.10 for levels of alkaline phosphatase (Table 2).

Consistent with our studies on simulated data, uncorrected heritability estimates were inflated. The inflation was significant for 14 of the 34 phenotypes (Fig. 1). In addition, 23 phenotypes had a value for e^2 significantly greater than zero (Table 2). In general, we would expect that $e^2 > 0$ would be a necessary but not sufficient condition for a difference in corrected and uncorrected heritabilities. Consistent with this expectation, these 23 phenotypes are a superset of the 14. We note that for 11 of the phenotypes, e^2 was not significantly greater than zero. In these cases, the standard model, which is nested in our more general model, provided an adequate model for heritability.

For each phenotype, we were also able to determine the geographical range of the environmental effect (i.e., the optimized value of the scaling parameter α), which varied by more than three orders of magnitude across the phenotypes (Table 2).

Interestingly, corrections were most substantial for anthropometric indices, lipid tests, and measures of liver function. This pattern, which may or may not be real, is under investigation. We are also working to identify specific environmental effects responsible for the large heritability corrections. For the phenotype mean corpuscular hemoglobin concentration (MCHC), which had the largest correction, we ruled out several factors as substantial sources of environmental effects. In particular, elevation was excluded because the terrain of the study is essentially flat. Also, the heritability corrections were about the same for males and females, which we would not expect if iron was a contributor. In addition,

Table 1. A description of the phenotypes measured in the Ugandan cohort

Phenotype	Category	Description
BMI	Anthropometric index	Body mass index
Height	Anthropometric index	Height
HIP	Anthropometric index	Hip circumference
Waist	Anthropometric index	Waist circumference
Weight	Anthropometric index	Weight
WHR	Anthropometric index	Waist–hip ratio
Basophils	Blood factor	Basophil count
Eosinophils	Blood factor	Eosinophil count
Hematocrit	Blood factor	Hematocrit
Hemoglobin	Blood factor	Hemoglobin
Lymphocytes	Blood factor	Lymphocyte count
MCH	Blood factor	Mean corpuscular hemoglobin
MCHC	Blood factor	Mean corpuscular hemoglobin concentration
MCV	Blood factor	Mean corpuscular volume
Monocytes	Blood factor	Monocyte count
MPV	Blood factor	Mean platelet volume
Neutrophils	Blood factor	Neutrophil count
Platlets	Blood factor	Platelet count
RBC dstr width	Blood factor	Red blood cell distribution width
RBCs	Blood factor	Red blood cell count
WBC	Blood factor	White blood cell count
DBP	Blood pressure	Diastolic blood pressure
SBP	Blood pressure	Systolic blood pressure
HbA1c2	Glycemic control	HbA1c2
Cholesterol	Lipid test	Total cholesterol
HDL	Lipid test	High-density lipoprotein
LDL	Lipid test	Low-density lipoprotein
Triglycerides	Lipid test	Triglycerides
Alanine	Liver function	Alanine aminotransferase test
Albumin	Liver function	Serum albumin test
Alkaline	Liver function	Alkaline phosphatase test
Aspartate	Liver function	Aspartate aminotransferase test
Bilirubin	Liver function	Bilirubin
Gamma	Liver function	Gamma-glutamyl transpeptidase test

Table 2. Results from an analysis of the Ugandan cohort

Phenotype	Category	h^2 uncorr	SE	h^2 corr	SE	P (diff = 0)	e^2	SE	P ($e^2 = 0$)	α	gxe^2	SE	P ($gxe^2 = 0$)	α_{gxe}
BMI	Anthropometric index	0.37	0.04	0.29	0.05	9.58E-05	0.074	0.017	0	56,000	0	0.000	—	56,000
Height	Anthropometric index	0.50	0.05	0.49	0.05	3.50E-02	0.007	0.003	0.0001	1,590,000	0	0.000	—	1,590,000
HIP	Anthropometric index	0.37	0.05	0.27	0.05	5.37E-06	0.076	0.016	0	56,000	0.100	0.090	0.0975	520,000
Waist	Anthropometric index	0.31	0.05	0.24	0.05	4.23E-04	0.065	0.017	0	56,000	0.029	0.090	0.3344	56,000
Weight	Anthropometric index	0.43	0.05	0.33	0.05	1.18E-05	0.071	0.016	0	67,000	0	0.000	—	67,000
WHR	Anthropometric index	0.14	0.05	0.12	0.05	1.83E-02	0.019	0.008	0	1,320,000	0.144	0.058	< 0.0001	1,320,000
Basophils	Blood factor	0.50	0.11	0.39	0.11	8.10E-04	0.062	0.014	0	430,000	0	0.100	0.4869	430,000
Eosinophils	Blood factor	0.39	0.11	0.36	0.11	2.39E-01	0.028	0.022	0.0266	118,000	0	0.000	—	118,000
Hematocrit	Blood factor	0.22	0.09	0.16	0.10	1.22E-01	0.032	0.020	0.0122	98,000	0.243	0.119	0.0001	171,000
Hemoglobin	Blood factor	0.20	0.09	0.17	0.09	2.20E-01	0.013	0.010	0.0148	298,000	0.210	0.106	0.0022	248,000
Lymphocytes	Blood factor	0.52	0.09	0.48	0.10	8.50E-02	0.024	0.015	0.0061	171,000	0.103	0.180	0.2294	171,000
MCH	Blood factor	0.53	0.11	0.48	0.12	3.13E-02	0.027	0.015	0.0006	360,000	0.225	0.231	0.106	298,000
MCHC	Blood factor	0.72	0.09	0.35	0.09	1.40E-13	0.220	0.029	0	248,000	0.029	0.122	0.3774	248,000
MCV	Blood factor	0.57	0.10	0.49	0.11	6.55E-03	0.038	0.016	0	430,000	0.363	0.181	0.009	430,000
Monocytes	Blood factor	0.43	0.10	0.39	0.10	1.38E-01	0.024	0.019	0.0086	360,000	0.091	0.243	0.288	360,000
MPV	Blood factor	0.57	0.09	0.55	0.09	2.61E-01	0.011	0.011	0.0411	520,000	0	0.000	—	520,000
Neutrophils	Blood factor	0.35	0.11	0.32	0.11	1.59E-01	0.015	0.010	0.0141	430,000	0	0.000	—	430,000
Platlets	Blood factor	0.48	0.09	0.45	0.10	4.08E-01	0.011	0.014	0.1111	171,000	0	0.000	—	171,000
RBC dstr width	Blood factor	0.33	0.09	0.25	0.08	0.003771	0.112	0.034	0	910,000	0.278	0.089	< 0.0001	760,000
RBCs	Blood factor	0.39	0.10	0.33	0.11	2.01E-01	0.064	0.048	0.0214	26,600	0.154	0.197	0.1285	56,000
WBC	Blood factor	0.44	0.10	0.43	0.10	5.76E-01	0.004	0.008	0.1736	910,000	0	0.020	—	910,000
DBP	Blood pressure	0.29	0.05	0.24	0.05	1.98E-04	0.034	0.008	0	760,000	0.070	0.075	0.141	760,000
SBP	Blood pressure	0.22	0.05	0.18	0.05	3.83E-03	0.024	0.010	0	1,320,000	0.059	0.082	0.2199	1,320,000
HbA1c2	HbA1c2	0.56	0.05	0.41	0.05	1.42E-16	0.101	0.013	0	360,000	0	0.000	—	360,000
Cholesterol	Lipid test	0.60	0.04	0.53	0.05	7.15E-07	0.052	0.010	0	206,000	0.032	0.095	0.3197	360,000
HDL	Lipid test	0.51	0.05	0.45	0.05	5.34E-07	0.041	0.009	0	430,000	0	0.000	—	430,000
LDL	Lipid test	0.60	0.04	0.54	0.05	3.81E-04	0.052	0.014	0	67,000	0	0.000	—	67,000
Triglycerides	Lipid test	0.27	0.05	0.25	0.05	4.18E-02	0.020	0.008	0.0002	142,000	0.125	0.088	0.0329	171,000
Alanine	Liver function	0.37	0.05	0.26	0.05	2.58E-05	0.122	0.025	0	530	0.001	0.082	0.4954	530
Albumin	Liver function	0.44	0.05	0.36	0.05	3.96E-07	0.047	0.009	0	360,000	0	0.000	—	360,000
Alkaline	Liver function	0.12	0.05	0.10	0.05	9.37E-02	0.010	0.003	0	760,000	0	0.000	—	760,000
Aspartate	Liver function	0.25	0.05	0.19	0.05	3.02E-04	0.061	0.015	0	46,000	0.001	0.064	0.4863	46,000
Bilirubin	Liver function	0.45	0.04	0.43	0.04	1.19E-02	0.019	0.008	0	520,000	0.042	0.102	0.2919	520,000
Gamma	Liver function	0.11	0.04	0.10	0.04	6.33E-01	0.006	0.013	0.2445	39,000	0	0.000	—	39,000

Uncorrected (uncorr) and corrected (corr) heritability estimates and estimates of e^2 and gxe^2 are shown along with their SEs. All SEs were computed from a 500-group jackknife. The P value testing the null that there is no difference between the uncorrected and corrected heritability estimates was a based on a two-sided test from a 500-group jackknife on the difference. The P values testing the null hypotheses that $\sigma^2_e = 0$ and $\sigma^2_{gxe} = 0$ were based on a one-sided test with 10,000 permutations (*Methods*). The values for α are in arbitrary units. The cells in green indicate statistical significance after Bonferroni correction. Columns 2–9 and 10–13 correspond to an analysis without and with the *gxe* variance component, respectively.

heritability correction remained high when MCHC was adjusted for primary occupation and alcohol consumption. On examining this spatial effect in more detail, we found that MCHC varied substantially from one village to the next. Previously, we have shown substantial variation in urbanicity indicators among these villages (14), consistent with this observation. We further explored the possibility of this spatial variation arising because villages were sampled in batches over time. We observed significant improvement in the model on inclusion of sampling date as a covariate. Nonetheless, even with this inclusion, the ratio of uncorrected to corrected h^2 remained almost the same. Our findings suggest that environmental factors influencing traits are complex, and understanding them will require further exploration in future studies with the relevant environmental phenotypic data.

Finally, we estimated the variance of the phenotypes due to interactions between genomic and environmental components, fitting the three random effects corresponding to $\mathbf{K_{IBD}}$, $\mathbf{K_{loc}}$, and $\mathbf{K_{GxE}}$ simultaneously. The variance σ^2_{gxe} for three phenotypes—hematocrit, red-blood-cell distribution width, and waist-to-hip ratio—was significantly greater than zero (Table 2).

Discussion

We have introduced an LMM approach that includes explicit representation of spatial location in estimates of narrow-sense heritability. Spatial location is presumably a surrogate for (some) environmental effects and, unlike many environmental variables, is easy to measure. On simulated data, we have shown that estimates of heritability based on the more general model seem to be unbiased, whereas the estimates based on the standard model are inflated for some phenotypes. Similarly, in an analysis of 34 phenotypes in a Ugandan cohort, we have found that the uncorrected estimates of heritability based on the standard model are inflated relative to corrected estimates based on the new model. Furthermore, on simulated data, we have shown that the degree of bias is not

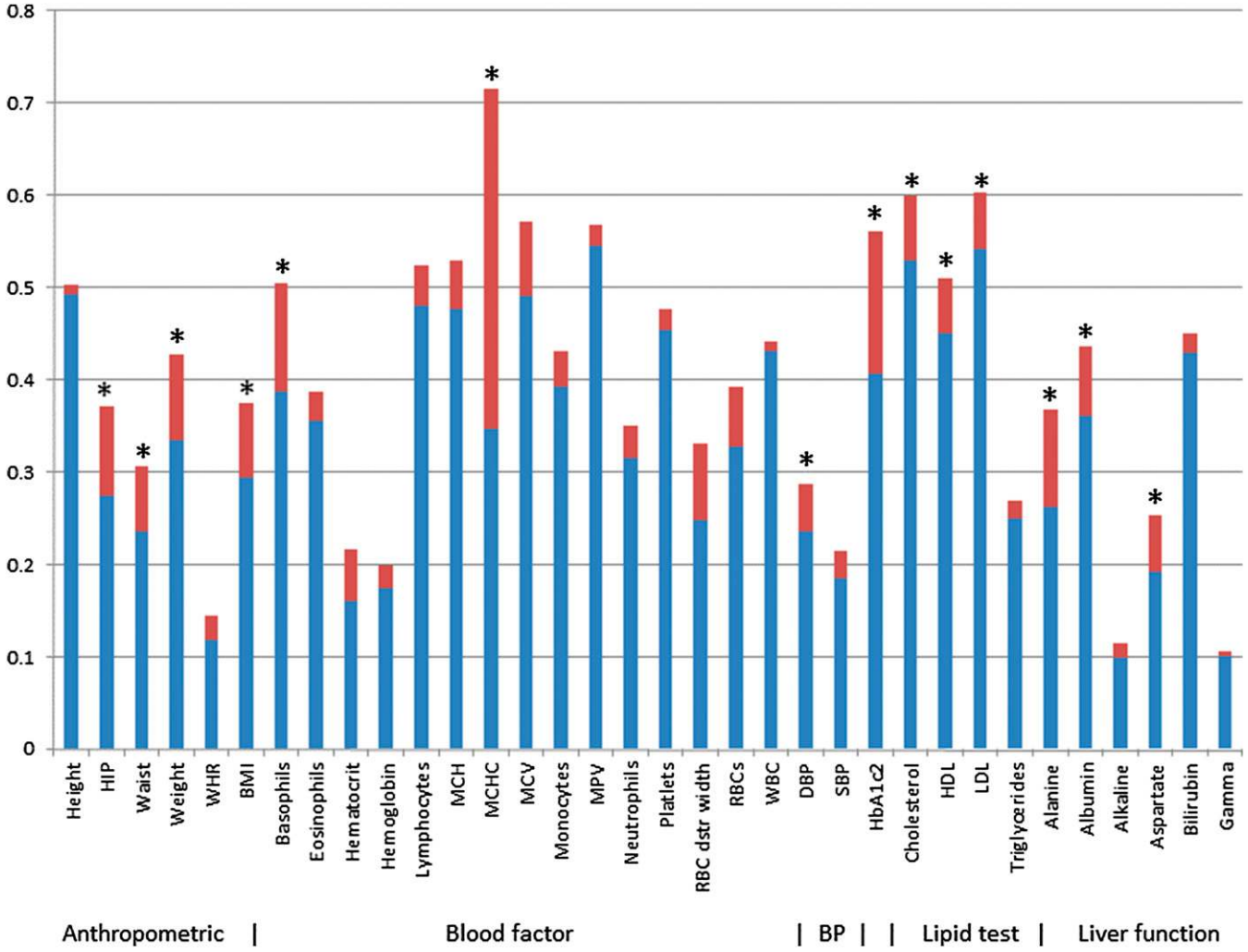

Fig. 1. Uncorrected and corrected estimates of narrow-sense heritability for phenotypes from the Ugandan cohort. The height of the blue and red bar combined corresponds to the uncorrected heritability estimate (based on Eq. **3**). The height of the blue bar corresponds to the corrected heritability estimate (based on Eq. **5**). Asterisks denote differences that are statistically significant after Bonferroni correction based on a two-sided test on the difference between uncorrected and corrected estimates from a 500-group jackknife.

substantially influenced by the absence or presence of dependence between genomic and environmental factors. Overall, we have demonstrated that estimates of heritability can depend on the nature of environmental variation.

The corrections were substantial, emphasizing the importance of explicitly modeling environmental effects in the estimation of heritability. Furthermore, the amount of inflation varied considerably across the 34 phenotypes, being the greatest for anthropometric indices, lipid tests, and measures of liver function. Presumably in this study, and perhaps in others, spatially related environmental factors affect these phenotypes more. A better understanding of differential bias across traits will require further exploration in cohorts with the relevant environmental phenotypic data.

Our approach has been applied only to the analysis of simulated data and the Ugandan cohort. Nonetheless, if the inflation seen here is typical, then this work offers a new interpretation of results of genome-wide association studies (GWAS). In particular, GWAS studies have so far revealed consistent "missing heritability," where the variability explained by SNPs identified as associated with a phenotype has been far less than the variability identified in heritability estimates (15). This work suggests that much of this missing heritability was not missing in the first place.

An important lesson from this work is that model misspecification can lead to substantial bias in heritability estimates. Because our use of the Gaussian radial basis function to quantify spatial similarity is itself likely to be misspecified, our corrected estimates of heritability on the Ugandan cohort may remain biased to some degree. Thus, we are investigating alternative similarity functions and methods to select the best one based on data.

We are also investigating modifications to the model beyond the form of the similarity function. One potential modification is based on the well-known fact that narrow-sense heritability estimates will be inflated when the data contains closely related individuals due to effects of dominance and epistasis (3, 4). This inflation could be mitigated by including variance components reflecting IBD2 (both alleles shared) (e.g., see ref. 16) and epistasis (e.g., ref. 17). In addition, one could include a variance component based on whether individuals are in the same household or a variance component based whether individuals are in the same village. As another example, one could include a variance component based on social connectivity, which is known to affect various phenotypes, including obesity (18).

Finally, in addition to heritability estimation, LMMs are also commonly used for identifying associations between genomic variants and phenotypes (e.g., genome-wide association studies) and for

prediction. The LMM models described in this work could be applied to these applications as well.

Methods

We collected data for 5,000 individuals from nine ethnolinguistic groups from the General Population Cohort (GPC), Uganda (19). The GPC is a population-based open cohort study established in 1989 by the Medical Research Council in collaboration with the Uganda Virus Research Institute (UVRI) to examine trends in prevalence and incidence of HIV infection and their determinants. Samples were collected from individuals during a survey from the study area located in southwestern Uganda in Kyamulibwa subcounty of Kalungu district, ~120 km from Entebbe town. The study area is divided into villages defined by administrative boundaries varying in size from 300 to 1,500 residents and includes families living within households. Data on health and lifestyle were collected using a standard individual questionnaire, blood samples obtained, and biophysical measurements taken, when necessary, as described previously (19). Spatial location was recorded in Global Positioning System coordinates. The measurements were translated and scaled to mitigate privacy concerns.

The GPC study was approved by the Uganda Virus Research Institute, Science and Ethics Committee (Ref. GC/127/10/10/25), the Uganda National Council for Science and Technology (Ref. HS 870), and the U.K. National Research Ethics Service, Research Ethics Committee (Ref. 11/H0305/5). Care was taken to obtain genuine informed consent from participants, including the use of reliable intermediaries as appropriate to ensure that the implications of participation were fully understood. Consent forms were translated from English into Luganda and checked for accuracy. The Lugandan translation was given to participants to read themselves, or was read out aloud to them by study staff. Participants could choose to consent to all, or just selected parts, of the survey. The informed consent of participants was obtained with a signature on the consent forms or a thumb print if the participant was unable to write. For participants aged 13–17 y, parental consent as well as child formal assent were collected. The immediate counter signature of a witness was then obtained. The APCDR committees are responsible for curation, storage, and sharing of the data under managed access. The genomic data have been deposited at the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) under accession number EGAS00001001558. Requests for access to phenotype data may be directed to data@apcdr.org.

We genotyped 5,000 samples from the Ugandan Survey on the Illumina HumanOmni 2.5M BeadChip array at the Wellcome Trust Sanger Institute. Sequenom quality control and gender checks were carried out before genotyping. A total of 2,314,174 autosomal and 55,208 X-chromosome markers were genotyped on the HumanOmni2.5–8 chip. Of these, 39,368 autosomal markers were excluded because they did not pass the quality thresholds for the SNP called proportion (<97%, 25,037 SNPs) and Hardy–Weinberg equilibrium (HWE) ($P < 10^{-8}$, 14,331 SNPs). HWE testing was only carried out on the founders for autosomes, and female unrelated individuals for the X chromosome defined by an IBD threshold <0.10 as estimated by PLINK. A total of 91 samples were dropped during sample quality control because they did not pass the quality thresholds for proportion of samples called (>97%) or heterozygosity (outliers: mean ± 3 SD), or the gender inferred from the X-chromosome data did not match the supplied gender. Three additional samples were dropped because of high relatedness (i.e., IBD >0.90). Principal component analysis was carried out on unrelated individuals projecting onto related individuals, for SNPs LD pruned at an r^2 threshold of 0.2, with a MAF threshold of >5%. No samples were identified as population/ancestry outliers based on this analysis.

To generate the phased dataset, we first mapped pedigrees within our dataset based on relationships provided in the data. To detect any errors in these pedigrees, we ran KING (20) on each cohort and also used the results to identify any cryptic first-degree relationships that had not been mapped. We further removed pedigrees where age information was inconsistent with the pedigree specified. In addition to the quality control described, we also removed SNPs with a minor allele frequency in the founders less than 5%, or with more than 1% Mendelian errors. We set all remaining Mendelian errors to missing, as well as any genotypes flagged as unlikely by the detection algorithm Merlin (21). SNPs with more than 1% missingness were then removed. We phased this curated dataset of 1,340,101 SNPs using SHAPEIT2 (22), first phasing the samples ignoring family information, and then running a hidden Markov model on every parent–child duo. This procedure corrects phasing errors inconsistent with the pedigree structure, further improving phasing accuracy. We have previously shown this method produces highly accurate results in our cohort with negligible switch error rates (22). To construct $\mathbf{K_{IBD}}$ from these phased data, we used the method outlined in ref. 23.

Phenotypes were transformed before analysis. Residuals were obtained following regression of the trait on age, age squared, and sex. Residuals were then inverse-normally transformed for analysis. For HbA1c, regression was carried out on age, age squared, sex, and month of sample collection (as an indicator variable) to account for seasonal trends in HbA1C that have been described previously (24).

Heritability estimation was performed with the FaST-LMM toolset available at https://github.com/MicrosoftGenomics/FaST-LMM. To determine a P value for the null hypothesis $\sigma_e^2 = 0$, we performed a permutation test wherein the entries of $\mathbf{K_{loc}}$ were permuted by randomly shuffling the identifiers of the individuals. A P value for the null hypothesis $\sigma_{gxe}^2 = 0$ was determined similarly by permuting the entries of $\mathbf{K_{GxE}}$. In both cases, 10,000 permutations were used.

ACKNOWLEDGMENTS. We thank Christoph Lippert for discussions about Mercer's theorem, Noah Zaitlen for discussions about more general models for heritability estimation, Ashish Kapoor for discussions on how best to fit the scaling parameters for radial basis functions, and Johanna Riha for discussions regarding the sources of spatial variance for some phenotypes. We thank the African Partnership for Chronic Disease Research for providing a network to support this study as well as a repository for deposition of curated data. We also thank all study participants who contributed to this study and the National Institute of Health Research Cambridge Biomedical Research Centre for data collection and phenotype analysis. This work was funded by the Wellcome Trust, Wellcome Trust Sanger Institute Grant WT098051, Medical Research Council Grants G0901213-92157, G0801566, and MR/K013491/1, and the Medical Research Council/Uganda Virus Research Institute Uganda Research Unit on AIDS core funding.

1. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433.
2. Wright S (1920) The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proc Natl Acad Sci USA* 6(6):320–332.
3. Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics* (Longman, Harlow, UK).
4. Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA).
5. National Genome Human Research Institute (2015) National Genome Human Research Institute. Available at https://www.genome.gov/.
6. Zaitlen N, Kraft P (2012) Heritability in the genome-wide association era. *Hum Genet* 131(10):1655–1664.
7. Valdar W, et al. (2006) Genetic and environmental effects on complex traits in mice. *Genetics* 174(2):959–984.
8. Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569.
9. Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* 91(1):47–60.
10. Bernhard S, Smola AJ (2001) *Learning with Kernels* (MIT Press, Cambridge, MA).
11. Rasmussen CE, Williams CKI (2006) *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA).
12. Cordell HJ (2002) Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11(20):2463–2468.
13. Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96(1-2):3–12.
14. Riha J, et al. (2014) Urbanicity and lifestyle risk factors for cardiometabolic diseases in rural Uganda: A cross-sectional study. *PLoS Med* 11(7):e1001683.
15. Eichler EE, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11(6):446–450.
16. Zaitlen N, et al. (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* 9(5):e1003520.
17. Stern MP, et al. (1996) Evidence for linkage of regions on chromosomes 6 and 11 to plasma glucose concentrations in Mexican Americans. *Genome Res* 6(8):724–734.
18. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *N Engl J Med* 357(4):370–379.
19. Asiki G, et al.; GPC team (2013) The general population cohort in rural southwestern Uganda: A platform for communicable and non-communicable disease studies. *Int J Epidemiol* 42(1):129–141.
20. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.
21. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30(1):97–101.
22. O'Connell J, et al. (2014) A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* 10(4):e1004234.
23. Price AL, et al. (2011) Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* 7(2):e1001317.
24. Tseng CL, et al. (2005) Seasonal patterns in monthly hemoglobin A1c values. *Am J Epidemiol* 161(6):565–574.