american physiological society
PHYSIOLOGICAL GENOMICS

Physiol Genomics 56: 791–806, 2024.
First published September 9, 2024; doi:10.1152/physiolgenomics.00049.2024

RESEARCH ARTICLE

Investigative power of genomic informational field theory relative to genome-wide association studies for genotype-phenotype mapping

Panagiota Kyratzi,[1,2] Oswald Matika,[3] Amey H. Brassington,[4] Connie E. Clare,[5] Juan Xu,[6] David A. Barrett,[7] Richard D. Emes,[8] Alan L. Archibald,[3] Andras Paldi,[2] Kevin D. Sinclair,[4] Jonathan Wattis,[9] and Cyril Rauch[1]

[1]*School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, United Kingdom;* [2]*École Pratique des Hautes Études, St-Antoine Research Center, PSL Research University, Inserm U938, Paris, France;* [3]*Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, Scotland, UK;* [4]*Agriculture and Horticulture Development Board, Middlemarch Business Park, East Coventry, United Kingdom;* [5]*School of Biosciences, University of Nottingham, Sutton Bonington, United Kingdom;* [6]*Shanghai Leadingtac Pharmaceutical Company Limited, Shanghai, People's Republic of China;* [7]*School of Pharmacy, Centre for Analytical Bioscience, University of Nottingham, Nottingham, United Kingdom;* [8]*Nottingham Trent University, Nottingham, United Kingdom; and* [9]*School of Mathematical Sciences, Centre for Mathematical Medicine and Biology, University of Nottingham, University Park, Nottingham, United Kingdom*

Abstract

Identifying associations between phenotype and genotype is the fundamental basis of genetic analyses. Inspired by frequentist probability and the work of R. A. Fisher, genome-wide association studies (GWAS) extract information using averages and variances from genotype-phenotype datasets. Averages and variances are legitimated upon creating distribution density functions obtained through the grouping of data into categories. However, as data from within a given category cannot be differentiated, the investigative power of such methodologies is limited. Genomic informational field theory (GIFT) is a method specifically designed to circumvent this issue. The way GIFT proceeds is opposite to that of GWAS. Although GWAS determines the extent to which genes are involved in phenotype formation (bottom-up approach), GIFT determines the degree to which the phenotype can select microstates (genes) for its subsistence (top-down approach). Doing so requires dealing with new genetic concepts, a.k.a. genetic paths, upon which significance levels for genotype-phenotype associations can be determined. By using different datasets obtained in *Ovis aries* related to bone growth (*dataset 1*) and to a series of linked metabolic and epigenetic pathways (*dataset 2*), we demonstrate that removing the informational barrier linked to categories enhances the investigative and discriminative powers of GIFT, namely that GIFT extracts more information than GWAS. We conclude by suggesting that GIFT is an adequate tool to study how phenotypic plasticity and genetic assimilation are linked.

NEW & NOTEWORTHY The genetic basis of complex traits remains challenging to investigate using classic genome-wide association studies (GWASs). Given the success of gene editing technologies, this point needs to be addressed urgently since there can only be useful editing technologies whether precise genotype-phenotype mapping information is available initially. Genomic informational field theory (GIFT) is a new mapping method designed to increase the investigative power of biological/medical datasets suggesting, in turn, the need to rethink the conceptual bases of quantitative genetics.

complex traits; GIFT; genotype-phenotype mapping studies; GWAS

INTRODUCTION

Identifying associations between phenotype and genotype is the fundamental basis of genetic analysis. The development of high-density genotyping and whole genome sequencing has enabled DNA variants to be directly identified, and genome-wide association studies (GWASs) have become the method of choice for mapping genotype to phenotype in large populations of unrelated individuals. GWASs have been used in many species, and especially in the study of human disease (1). By 2021, the National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog listed 316,782 associations identified in 5,149 publications describing GWAS results (2). In addition, extensive collection of data has been initiated through efforts such as the UK Biobank (3), Generation Scotland (4), and National Institutes of Health (NIH) All of Us research program (https://allofus.nih.gov/), in the expectation that large-

Correspondence: C. Rauch (cyril.rauch@nottingham.ac.uk).
Submitted 30 April 2024 / Revised 16 August 2024 / Accepted 3 September 2024

scale GWAS will elucidate the basis of human health and disease and facilitate precision medicine.

Although genomic technologies have advanced rapidly, statistical models used to analyze genetic data are still based on the models developed by Fisher more than 100 years ago (5, 6). GWASs essentially use the Fisher method of partitioning genotypic values by performing a linear regression of phenotype on marker allelic dosage (7). Regression coefficients estimate the average allele effect size, and the regression variance is the additive genetic variance due to the locus (8). However, an ongoing debate exists over whether the present analysis paradigm in quantitative genetics is at its limits for truly understanding complex traits, namely traits resulting from many genes, each with a very small effect size (9). As a result, one may wonder whether alternative statistical model(s) could be invented and used to determine genotype-phenotype mappings.

GWASs are fundamentally linked to frequentist probabilities that, defined through relative frequencies, determine the validity of statistical inferences. In practice, frequentist probabilities are generated through the grouping of data into bins or categories to generate a bar chart, which is then interpolated to create a distribution density function (DDF) in the continuum limit. The DDF is, in turn, used to determine statistical inferences including average, variance, p value, and so on. However, since the DDF approximates the bar chart (and not the converse), and because it is not possible to differentiate data from within any given group/category, the DDF is constructed mathematically on the implicit assumption that information is missing to differentiate data from within any given group/category.

The notion of "missing information" can be legitimate and defined experimentally. For example, measuring the phenotype human height with a ruler with centimeter graduations implies that any height can be measured to the nearest centimeter. Consequently, 1 cm-wide bins/categories need to be used to generate a frequency table of the range of phenotype values upon which the phenotype and genotype DDFs are defined. In this case, all the resulting statistical inferences are defined with a precision corresponding to the nearest centimeter. The "missing information" (i.e., that what cannot be measured by the ruler) corresponds then to subcentimetric scales (i.e., distances to the nearest millimeter for this example). In practice the "missing information" is therefore linked to the one of "imprecision" and deciding to provide more precise statistical inferences implies that the width of categories be reduced, which can only be achieved by increasing the sample size. It is not by chance that the "normal distribution" created by mathematicians and physicists was initially called the "law of errors," where the notion of error (misinformation) results from imprecisions in experimental measurements. As a result, GWAS is faced with a fundamental issue involving the extraction of precise information using a method that, conceptually, assumes that information is missing or that data are mis(in)formed.

In general, the problem concerning the "missing information" is never mentioned, since the DDF in the continuum limit is never considered as an approximation but as something that has its own reality. Namely, a DDF must exist independently of the data measured (i.e., data must fit the DDF and not the converse). The latter remark leads to an interesting conceptual territory where the notions of average and variance, and their usage, may be questioned. If one considers the normal distribution (or any other DDFs) is inherent to life and that data must fit it (them), then the moments of the distribution (e.g., average and variance) are also essential parameters to describe life, and the variance, often interpreted as noise in the data, is then a nuisance. If, on the contrary, data are the important thing, and that the DDF is considered solely as a tool to interpolate data based on missing information, then average and variance are parameters derived from a lack of information and are, as a result, poorly informative. The latter point should not come as a surprise, as reducing the huge diversity of populations to a handful of parameters (i.e., average and variance) is highly reductionist and likely to be poorly descriptive. Thus, although the notions of average and variance may help represent datasets, they are inventions nonetheless, i.e., thought constructions akin to the field of frequentist probability. Thus, using average and variance as a starting point to map genotype-phenotype (GWAS) is a matter of choice. Accordingly, different statistical methods can be suggested.

To avoid those conceptual and practical issues, a new method called genomic informational field theory (GIFT) has been designed and applied to simulated genotype-phenotype data in Wattis et al. (10) and Rauch et al. (11) and is reviewed in Rauch et al. (12). In short, to associate genotype to phenotype, GIFT does not presume that the only important information concerning the gene effects is found in averages or variances, nor does it presume that DDFs are central. On the contrary, GIFT starts with the prerequisite that phenotypic values, or phenotypic residuals after considering the environment/fixed effects, may be measured with sufficient precision to be unique in a population. Then, by avoiding the grouping of data into bins/categories, which would otherwise create an artificial imprecision, GIFT considers the entire information contained in the data (i.e., variance is not a nuisance anymore) by making use of the cumulative sum of microstates. Figure 1 provides the intuition underscoring GIFT as a method.

The current article extends our previous theoretic studies using simulated data to analyze for the first time two real datasets:

1) *Dataset 1* is derived from a study concerned with the genetic background of carcass composition in sheep (*Ovis aries*) (13). Using GWAS, this study demonstrated a strong association between chromosome 6 and the carcass composition trait "bone area at the ischium." We now apply GIFT to reanalyze this dataset to benchmark it against GWAS. Since GWAS previously identified a quantitative trait loci (QTL) in chromosome 6, our hypothesis was that GIFT would at least replicate GWAS results and identify additional putative QTLs.
2) *Dataset 2* comprises biochemical data arising from an ongoing study in sheep that seeks to identify risk allele variants in genes whose products direct a series of metabolic pathways, collectively referred to as one-carbon (1C) metabolism and associated epigenetic regulators. The gene array was designed to include all single-nucleotide polymorphisms (SNPs) linked to known biochemical

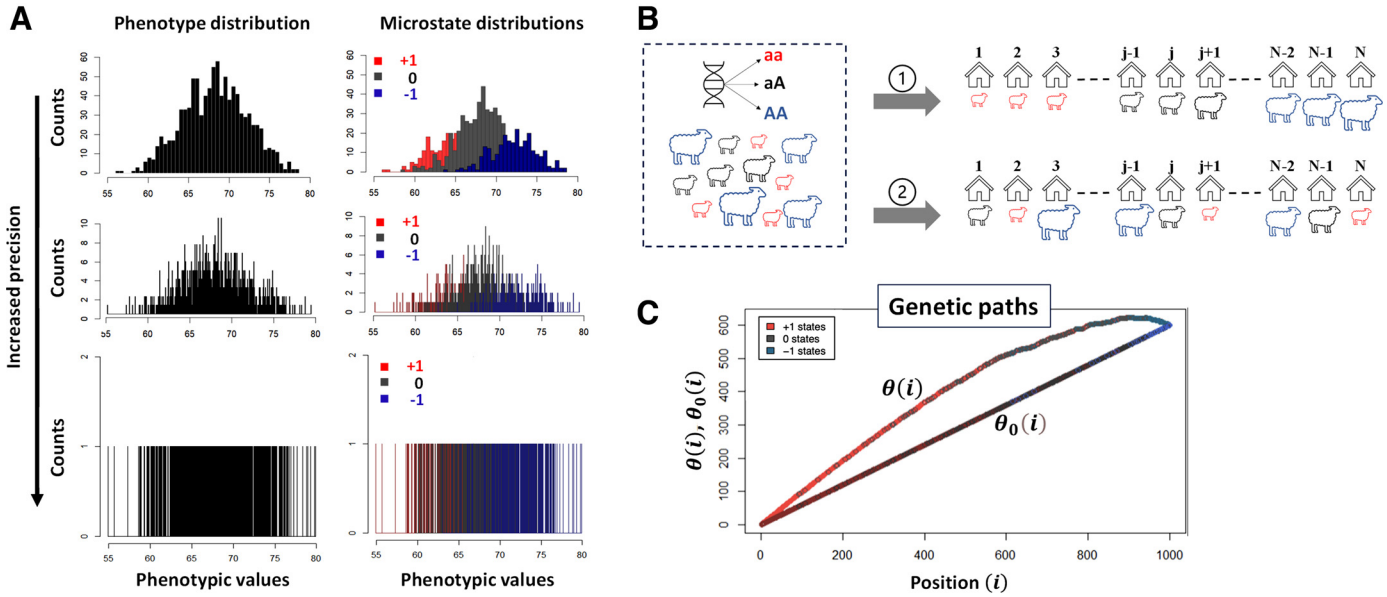


Figure 1. *A*: for diploid organisms and for a binary (biallelic, A or a) genetic marker, any microstate (genotype) can only take three values that we shall write as "+1," "0," and "−1" corresponding to genotypes aa, Aa, and AA, respectively. The genotypes are color coded to facilitate the representation of GIFT [+1: aa (red), 0: aA/Aa (black), and −1: AA (blue)]. GWASs rely on probability density functions formed through the grouping of data into bins/categories. The phenotype distribution density function (*A*, *top left*) is then decomposed onto the distribution density function of genetic microstates (*A*, *top right*) for every single-nucleotide polymorphism (SNP). Using an analysis of averages and variances, such decomposition determines whether the SNP studied is associated with the phenotype by comparing the average and variance of distributions. Repeating the same operation for every SNP in the genome permits to map genotype to phenotype. However, as more precise inferences can only come with and are only legitimized by a reduction in the width of categories, larger sample sizes are needed. To overcome this issue, one way to proceed is to deconstruct density functions and wonder what would happen if one were able to reduce the width of categories, that is increasing the precision in the measurement of the phenotype or equivalently getting access to the whole information of datasets, without changing the sample sizes (*A* from *top* to *bottom*). The mathematical object that emerges is a colored barcode, which is a list of microstates that can be analyzed precisely by GIFT. *B*: such barcode can be obtained simply at the practical level through field studies. Assume a flock of sheep has been genotyped and that their phenotype has been measured sufficiently precisely such as to exclude the possibility that any two phenotypic values are identical. In the figure, the magnitude of the phenotypic value for each sheep is characterized by the (unique) "size" of the sheep. The barcode is obtained by ranking animals as a function of the magnitude of their phenotypic values (configuration ① in *B*). The null hypothesis is obtained via the random ranking of sheep that is equivalent to a lack of information on phenotypic values (configuration ② in *B*). As GWAS works on phenotypic residual values after adjusting for fixed/environmental effects, a similar barcode can be generated considering the magnitude of residual phenotypic values. *C*: GIFT proceeds by plotting the cumulative sum of microstates as a function of the position in the list generating a curve called genetic path that is represented by θ(i) and is unique to the SNP considered. Although the curve θ(i) does not provide any significant information on its own, one may generate, for the same SNP, a curve (genetic path) corresponding to a sort of null hypothesis when ranking the phenotype does not bring any informational value. This is possible by scrambling (permutating) the string of microstates an infinite number of times. It is then possible to show that, in the asymptotic limit, the null hypothesis returns a straight line, noted θ_0(i), from which inferences may be suggested regarding potential association between the genotype and the phenotype by comparing θ_0(i) to θ(i). Note that the simulation shown in *A* adhering to Fisher seminal model is based on a constant sample size of 1,000, involving an arbitrary normally distributed phenotype with a mean of 68 and a variance of 4 units, respectively. Each microstate is normally distributed with a gene effect identical to the standard deviation of the phenotype but without dominance. The frequencies of the genotypes aa (red), Aa/aA (gray), and AA (blue) are 64%, 32% and 4%, respectively, and within Hardy-Weinberg ratio. GIFT, genomic informational field theory; GWAS, genome-wide association studies.

enzymes involved in these pathways. Given that *dataset 2* preselected genes for a targeted analysis of enzymes involved in these metabolic/epigenetic pathways, it can be considered more specific.

The present article initially introduces the reader to the way data may be used and analyzed differently using GIFT, contrasting to more conventional methods mostly based on an analysis of averages and variances. More specifically, in *part 1*, the null hypothesis defined by GIFT will be established. Using *dataset 1*, the concept of genetic path pertaining to GIFT will be introduced (*part 2*) out of which a *p* value for GIFT will be defined (*part 3*). Then, *dataset 1* (*part 4*) and *dataset 2* (*part 5*) will be analyzed by comparing the informational/investigative power of GIFT relative to GWAS using Manhattan plots before performing enrichment analyses.

MATERIALS AND METHODS

Biological Datasets

The first dataset (*dataset 1*) analyzed 600 pedigree-recorded Scottish Blackface lambs using CT scans to determine in vivo carcasses composition (13). The trait selected for the present study is the bone areas of the ischium (BAI), measured in mm^2 from cross-sectional computed tomography (CT) scans. The ischium is one of the three bones that make up the pelvis. It is located beneath the ilium and behind the pubis. The upper portion of the ischium forms a major part of the concave portion of the pelvis that forms the hip. The BAI crossed a genome-wide significance threshold on chromosome 6 (OAR6). The precorrected phenotype values were obtained by fitting fixed effects of the age of dam, year of birth, effect of management group (as sheep were from different farms), sex

(males or females), litter size (singles or twins), and as covariate the day of birth. Further information can be found in Matika et al. (13). Supplemental File S1 provides the raw data used (*dataset 1*).

The second dataset (*dataset 2*) was from previously unpublished data extracted from a large ongoing program of research to investigate genome regions (QTLs) that determine metabolic and epigenetic responses to nutritionally induced deficiencies in 1C metabolism (14, 15). For this study, sheep were used as an experimental model. All animal procedures relating to this study adhered to the Animals (Scientific Procedures) Act, 1986. Associated protocols complied with the ARRIVE guidelines and were approved by the University of Nottingham Animal Welfare and Ethical Review Body (AWERB) with Home-Office project licensed authority (30/3376; February 10, 2016). Supplemental File S2 provides the raw data used (*dataset 2*).

Dataset 2: Sheep Genome Resequencing, Custom Array Design, and SNP Profiling on Test Subjects

Twenty-four unrelated Texel ewes were sequenced to a depth of 30× in two pools at Edinburgh Genomics. DNA samples were prepared using Illumina's TruSeq PCR-free kits and sequenced on an Illumina HiSeq 2500 Rapid Mode (Serial No. D00125) with a read length of 150PE. Reads were trimmed to remove adapter sequences and low-quality bases using skewer with the commands (-Q 20, -q 3) (16) and mapped to the reference sheep genome assembly (Oar_v3.1) using bwa mem (options -M -t 4) (17). Following deduplication using Picard-tools v. 1.92, variants were called using the GATK pipeline (18), including realignment around known indels and recalibration of bases, as well as FreeBayes (–use-best-n-alleles 4 –pooled-discrete –min-alternate-count 4). Annotation of SNPs was performed using the Ensembl variant effect predictor (VEP) version Ensembl Tools Release 79 (19). A total of 15,347,831 variants were identified. Of these, ∼3 million were novel SNPs, and ∼12 million were already present in the Ensembl genome database. SNPs within annotated coding regions (VEP annotated "downstream gene variant" or "intron variant" removed) and those within 3Kb upstream of a gene were retained. SNPs with a minor allele frequency of greater than 0.5 were used to design an Illumina Infinium iSelect Custom Array consisting of 4,576 probes. This captured SNPs in 115 1C metabolism and related genes, 108 related epigenetic regulators, and 33 control SNPs (Supplemental File S1).

Liver samples were next collected postmortem from 360 male and female Texel lambs (6 to 11 mo of age) representing 11 farms dispersed regionally across the UK. Collections took place at regional abattoirs, and samples were immediately snap-frozen in liquid nitrogen and stored at −80°C until analyses. DNA was then extracted using AllPrep DNA/RNA Mini kit (QIAGEN, Manchester, UK). In brief, ∼20 mg of liver were mechanically disrupted using a TissueLyser (QIAGEN, Manchester, UK) in 600 RLT plus buffer containing β-mercaptoethanol. Tissue lysates were then used to extract RNA and DNA according to the manufacturer's instructions. The custom-designed array was then used to SNP-profile DNA from these Texel sheep. For this purpose, liver samples were collected postmortem from lambs (aged 6 to 11 mo) representing 11 farms dispersed regionally across the UK. Collections took place at regional abattoirs, and samples were immediately snap-frozen in liquid nitrogen and stored at −80°C until analyses. DNA was then extracted using AllPrep DNA/RNA Mini kit (QIAGEN, Manchester, UK). In brief, ∼20 mg of liver were mechanically disrupted using a TissueLyser (QIAGEN, Manchester, UK) in 600 RLT plus buffer containing β-mercaptoethanol. Tissue lysates were then used to extract RNA and DNA according to the manufacturer's instructions.

Dataset 2: Metabolic Profiling

For the purposes of the current study, the following seven liver metabolites were selected from a larger pool of 1C metabolites: *S*-adenosyl methionine (SAM), methylcobalamin (mB12), adenosylcobalamin (aB12), trimethylglycine (TMG), dimethylglycine (DMG), propionate (PPA), and methylmalonic acid (MMA). The first four metabolites were selected as representative intermediates of the methionine cycle, whereas the latter two are intermediates in the hepatic synthesis of succinate (15) (Fig. 2 and Supplemental File S1).

Hepatic concentrations of four metabolites (i.e., mB12, aB12, TMG, and DMG) were determined by hydrophilic interaction chromatography (HILIC) coupled with electrospray ionization tandem mass spectrometry (MS/MS) as reported previously (20). For the analysis of SAM (determined separately by HILIC), the standard was purchased from Sigma-Aldrich (Poole, Dorset, UK). Stock solutions of this standard were prepared in potassium phosphate extraction buffer (KH_2PO_4 and K_2HPO_4; 40 mmol/L) containing 0.1% L-ascorbic acid, 0.15% citric acid, and 0.1% MCE (adjusted to pH 7 with NaOH), each at a final concentration of 100 μmol/L. Also, for SAM, the mobile phase was modified from that used for the three other reported metabolites by adjusting the pH of the aqueous ammonium carbonate buffer solution from 3.5 to 9.1. Mass spectrometer parameters for SAM were as follows: retention time = 7.69 min; Q1mass = 399.1 amu; Q3 mass = 250.1 amu; declustering potential = 56; collision energy = 25; and collision cell exit potential = 16.

Hepatic concentrations of PPA and MMA were determined by gas chromatography coupled to mass spectroscopic-detection (GC-MS). In brief, for PPA, 750 μL 5-sulfosalicylic acid (SSA, 0.04 mg/mL) was added to 150 mg frozen liver, homogenized for 2 min, and cooled on ice for 10 min. The sample was centrifuged for 15 min at 14,500 *g*, and 200 μL liver homogenate was transferred to a 2.5 mL screw-capped glass vial. To this, 20 μL internal standard (MBA, 400 μM), 3.5 μL HCl (37%), and 1 mL diethylether were added, vortexed for 2 min, and centrifuged for 10 min at 14,500 *g*. The upper layer (600 μL) was transferred to a screw-capped glass vial containing 3.5 μL 1-(*tert*-butyldimethylsilyl)imidazole (TMDMSIM, 97%), vortexed for 2 min, and heated at 60°C for 30 min. GC-MS analysis proceeded after cooling. The method used a DB-5MS column (J&W Scientific Agilent technology, 30 m × 0.25 mm; 0.25 μm film thickness). The carrier gas (He) was set at a constant flow rate of 1.3 mL/min. The injection volume was 5 μL for SCAN mode (for qualification) and SIM (selected ion monitoring) mode (for quantification), both using splitless mode. The injection port and MS selective detector interference temperatures

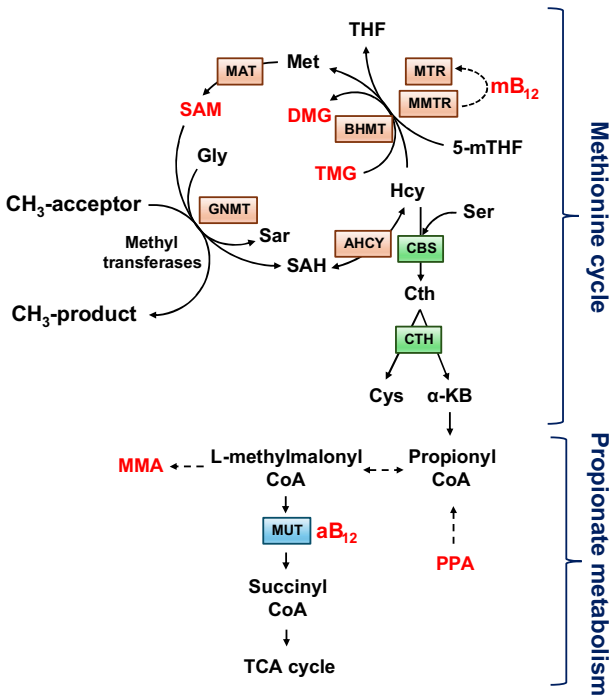


Figure 2. Linked methionine and propionate metabolism adapted from Clare et al. (15) where all metabolites studied for this study are in red. The methionine cycle facilitates the remethylation of homocysteine (Hcy) to methionine (Met) and ultimately *S*-adenosylmethionine (SAM) with methyl (CH_3) groups donated either from folate (5-mTHF) or betaine [trimethylglycine (TMG)], thus leading to the formation of dimethylglycine (DMG). Methylcobalamin (mB12) serves as a cofactor for the reduction of the inactive form of methionine synthase to its active state (MTR), which then transfers a methyl group from 5-mTHF to Hcy. The linked metabolism of propionate (PPA) to succinate (an intermediary metabolite in the tricarboxylic acid cycle) requires adenosylcobalamin (aB12), which serves as a cofactor for methylmalonyl-CoA-mutase (MUT), leading to the generation of succinyl-CoA and methylmalonic acid (MMA) in this pathway. Other intermediary metabolites and enzymes listed: α-KB, α-ketobutyrate; AHCY, adenosyl-homocysteinase; BHMT, betaine homocysteine methyltransferase; CBS, cystathionine β-synthase; Cth, cystathionine γ-lyase; Cth, cystathionine; Cys, cysteine; Gly, glycine; GNMT, glycine methyltransferase; MAT, methionine adenosyl-transferase; MMA, methylmalonic acid; SAH, *S*-adenosylhomocysteine; Sar, sarcosine; Ser, serine; THF, tetrahydrofolate.

were 260°C and 250°C, respectively. The chromatograph was programmed for an initial temperature of 40°C for 1 min, increased to 60°C at 70°C min^{-1}, then to 110°C at 15°C min^{-1}, and finally 250°C at 70°C min^{-1}. The MS was tuned regularly and operated in electron impact (EI) ionization mode with an ionization energy of 70 eV. SCAN mode measured at m/z: 30–300, and SIM ions were set at 159 (for MBA) and 131 (for PPA). The same method was used to produce a calibration curve for PPA using standards at concentrations ranging from 19.5 nmol/g to 5 μmol/g. The limit of detection was 19.5 nmol/g. CVs for low, medium, and high QCs were 10.4, 6.3, and 6.5%, and the interassay CV was 4.7%.

For MMA, 250 μL of 80% MeOH was added to 50 mg of frozen liver, homogenized for 2 min, and cooled on ice for 10 min. The sample was then centrifuged for 15 min at 14,500 *g*, and 200 μL liver homogenate was transferred to a 2.5 mL screw-capped glass vial. To this, 4 μL of internal standard [1 mM 4-chlorobutyric acid (CBA) in 1 mM HCl] followed by 250 μL 12% BF3-methanol were added, vortexed for 1 min, and heated at 95°C for 15 min. After cooling, 250 μL of cold distilled water and 250 μL of cold dichloromethane (CH_2Cl_2) were added to the vial, vortexed for 30 s, and centrifuged for 10 min at 14,500 *g*. The lower dichloromethane layer was transferred to a screw-capped glass auto-sampler vial with an insert for GC-MS analysis. The method used a DB-WAX column (cross-linked polyethylene glycol; J&W Scientific Agilent technology) (30 mm × 0.25 mm; 0.15 μm film thickness). The carrier gas (He) was set at a constant flow rate of 1.0 mL/min. The injection volume was 1 μL for SCAN mode (for qualification) and SIM mode (for quantification), both using splitless mode. The injection port and MS selective detector interference temperatures were 260°C and 280°C, respectively. The chromatograph was programmed for an initial temperature of 50°C for 2 min, increasing to 150°C at 8°C min^{-1}, then to 220°C at 100°C min^{-1}, and held for 5 min at the final temperature. MS was tuned regularly and operated in EI ionization mode with an ionization energy of 70 eV. The limit of detection was 0.75 nmol/g for both MMA and SA, and the inter-assay CVs were 8.4% for MMA and 11.0% for SA.

Dataset 2: Determination of GWAS for 1C Metabolites

Preliminary data analysis indicated the need to log-transform using the natural logarithm (Supplemental File S3) to approximate normality. The transformed data were then precorrected for the fixed effects of farm (F) and sex (S) in ASReml using the following model, $y_{ij} = \mu + F_i + S_j + E_{ij}$, where y_{ij} is the log-transformed phenotype, that is the log-transformed metabolite concentration studied; μ is the overall mean for the log-transformed metabolite concentration; F_i is the effect of the ith farm ($i = 1, \ldots, 11$); S_j is the effect of jth sex (male vs. female), and e_{ij} is the residual. The genotype dataset was filtered using PLINK (HWE p value threshold of 10^{-6}, call rate for genotypes of 10%, and a MAF of 5%), the number of independent SNPs was determined using BCFTOOLS (r^2 threshold = 0.1), and the GWAS Manhattan plots, linked to the determination of p_{GWAS}, were obtained using GEMMA. The same genotype and residual phenotypes as filtered by GWAS were used by GIFT.

Data Representation Using GIFT

Adjusted phenotypic data (i.e., residuals, from *dataset 1* and *dataset 2*) were used for this study. Regarding the representation of GIFT, upon selecting a SNP for all individuals, the different corresponding genotypes, aa, aA/Aa and AA, were assigned arbitrary values of +1, 0, and −1, respectively. With this convention, any barcode can be represented by a string of numbers from which a GIFT analysis can be inferred. More specifically, the assignment of values +1, 0, and −1 were done as a function of the base pairs as follows: AA = TT = +1, GG = CC = −1, and 0 otherwise. As shown schematically in Fig. 1, the residuals obtained were ranked by order of magnitude, and the cumulative sum of their corresponding genotypic values was performed to obtain the "genetic path" for the SNP considered. The genetic path of an SNP is noted θ(i) in the text (Fig. 1). The null hypothesis for GIFT, as well as the notion of significance when GIFT is used, will be introduced and fully explained in RESULTS.

RESULTS

Analysis of the Null Hypothesis $\theta_0(i)$ for GIFT

Although $\theta(i)$ is obtained using phenotypic information (configuration ① in Fig. 1 and *Data Representation Using GIFT* in MATERIALS AND METHODS), it is also possible to plot the cumulative sum of microstates when no phenotypic information is present that is equivalent to "scrambling" or permutating the string of microstates in Fig. 1*A*, which corresponds to configuration ② in Fig. 1*B*. Recall that since our focus is on a given SNP, the number of microstates, N_+, N_0, and N_-, are identical between the configurations ① and ②. This new cumulative sum noted $\theta_0(i)$ is expected to be a sort of null hypothesis solely dependent on the bulk microstate frequencies N_+/N, N_0/N, and N_-/N, where N_q $q \in \{+, 0, -\}$ is the number of microstates of type q. This is so because there is no further information that could inform on the positioning of microstates in their list when the scrambled state is considered. However, although $\theta(i)$ is unique since phenotypic information is used to generate it, $\theta_0(i)$ is not as each time the string of microstates from Fig. 1*A* is scrambled, a new $\theta_0(i)$ appears. Accordingly, one needs to consider the set of possible $\theta_0(i)$s generated bounded to the microstate frequencies N_+/N, N_0/N, and N_-/N.

Using a selection of theoretic SNPs defined by different microstate frequencies (Table 1), Fig. 3*A* illustrates the global shape resulting from simulating 1,000 $\theta_0(i)$s. The results demonstrate that the global shape of the $\theta_0(i)$s plotted as a function of the position in the string is ellipsoidal with short and long axes changing as a function of microstate frequencies involved, and where the different averages of $\theta_0(i)$s represented by black lines in Fig. 3*A* are straight lines with slopes linked to the difference, $\Delta N/N = (N_+ - N_-)$. The fact that the averages of $\theta_0(i)$s for a given set of microstates, N_+, N_0 and N_-, is always a straight line linked to microstate frequencies, N_+/N, N_0/N, and N_-/N, can be understood intuitively by the fact that scrambling or permutating an infinite number of times the string of microstates is equivalent to determining, for any position i, the presence probability, N_q/N, of each microstate in the string. Accordingly, for a given set of microstates, N_+, N_0, and N_-, the average of $\theta_0(i)$s, noted $\theta_0(i)$, is $\theta_0(i) = \frac{(N_+ - N_-)}{N} i$.

Table 1. *Theoretic SNPs used to capture the null hypothesis associated with GIFT upon 1,000 simulations of microstate permutation*

SNP Name	N_+	N_0	N_-	N
SNP1	25	25	515	565
SNP2	25	125	415	565
SNP3	25	225	315	565
SNP4	25	325	215	565
SNP5	25	425	115	565
SNP6	25	525	15	565

The difference between consecutive SNPs in the table is linked to the transfer of 100 microstates from the microstates "−1" to the microstate "0" leaving the number of microstates "+" invariant. By permutating the microstates "+" and "−" in the table, similar plots as those obtained in Fig. 3*A* could have been obtained; the only difference would have been the slopes of the average $\theta_0(i)$ changing sign. GIFT, genomic informational field theory; SNP, single-nucleotide polymorphism.

Further theoretic details can be found in Wattis et al. (10) and Rauch et al. (11). Using $\theta_0(i)$ as a reference for the null hypothesis, Fig. 3*B* shows the sur-imposition of the differences, $\Delta\theta_0(i) = \theta_0(i) - \langle\theta_0(i)\rangle$, obtained from simulations using SNPs from Table 1.

Finally, to assess the impact of the sample size (population size) on the null hypothesis, the initial size ($N = 565$, Table 1) was divided ($N = 280$) and multiplied ($N = 1{,}130$) by a factor ~2, while keeping constant the microstate frequencies N_+/N, N_0/N, and N_-/N from Table 1. The simulations in Fig. 3*A* show that the appearance of ellipsoids is affected when the sample size changes, becoming thinner as the population size increases. Plotting the standard deviation, $\sigma(i/N)$, as a function of the position once normalized by the sample size, $\sigma(i/N) = \sqrt{\left[\theta_0(i/N)^2 - \theta_0(i/N)^2\right]/N}$, resulting from the different simulations in Fig. 3*C* demonstrates that the standard deviation from GIFT is quadratic and independent of the sample size, as expected from a random allocation of different microstates in the string of positions.

At first sight and with this primary analysis, one could suggest that any genetic path departing from the cloud of genetic paths formed by the set of $\theta_0(i)$s upon the permutation of microstates (gray surface in Fig. 3*A* or black surface in Fig. 3*B*) would likely result in an association between the genotype and the phenotype. Although this assumption is true, it needs to be handed out carefully, as it is not exhaustive. Indeed, some genetic paths may be highly structured and of relatively small amplitude. Examples of genetic paths using real data from *dataset 1* will demonstrate this point.

Examples of Genetic Path Using the Bone Area of the Ischium as Phenotype (*Dataset 1*)

The resulting average, $\theta_0(i)$, and variance, $\sigma(i)$, can be used to inform the null hypothesis of a particular SNP from "real" datasets. However, since there are as many different sets of $\theta_0(i)$s as number of SNPs, each SNP will return its own $\theta_0(i)$ (null hypothesis) upon scrambling. A comparison between SNPs using GIFT/genetic paths requires then to concentrate on the differences, $\Delta\theta_0(i) = \theta_0(i) - \langle\theta_0(i)\rangle$. In the remaining text, one shall rewrite $\langle\theta_0(i)\rangle$ as $\theta_0(i)$ to simplify notations.

Concentrating now on "real" dataset, the genetic paths were obtained further to ranking BAI residual values (*dataset 1*), using an incremental rank from small to large values. As an example, Fig. 4 shows the two genetic paths $\theta(i)$ and $\theta_0(i)$ for six SNPs, renamed SNP1-6 (see Table 2 for accurate genetic information), enabling us to appreciate the qualitative difference between the genetic paths. Although the null hypothesis, i.e., $\theta_0(i)$, resulting from the scrambling of phenotypic values many times, always returns a straight line with a different slope for each SNP, similar to what was seen in Fig. 3, the $\theta(i)$s for SNP1-6 in Fig. 4 have different shapes. To represent the set of $\theta(i)$s in relation to the different microstates involved, each datapoint of the $\theta(i)$s is color coded as in Fig. 1*C*.

Since $\theta_0(i)$ is linked to the difference between the genetic microstate frequencies of homozygotes, $\Delta N = N_+ - N_-$, in Fig. 4, we represent this difference by the angle α. Since $\tan(\alpha) = +N_+/N - N_-/N$, where N is the total number of positions ($i = 1, 2, \ldots, N$), $\theta_0(i)$ can be rewritten as $\theta_0(i) = \tan(\alpha)i$.

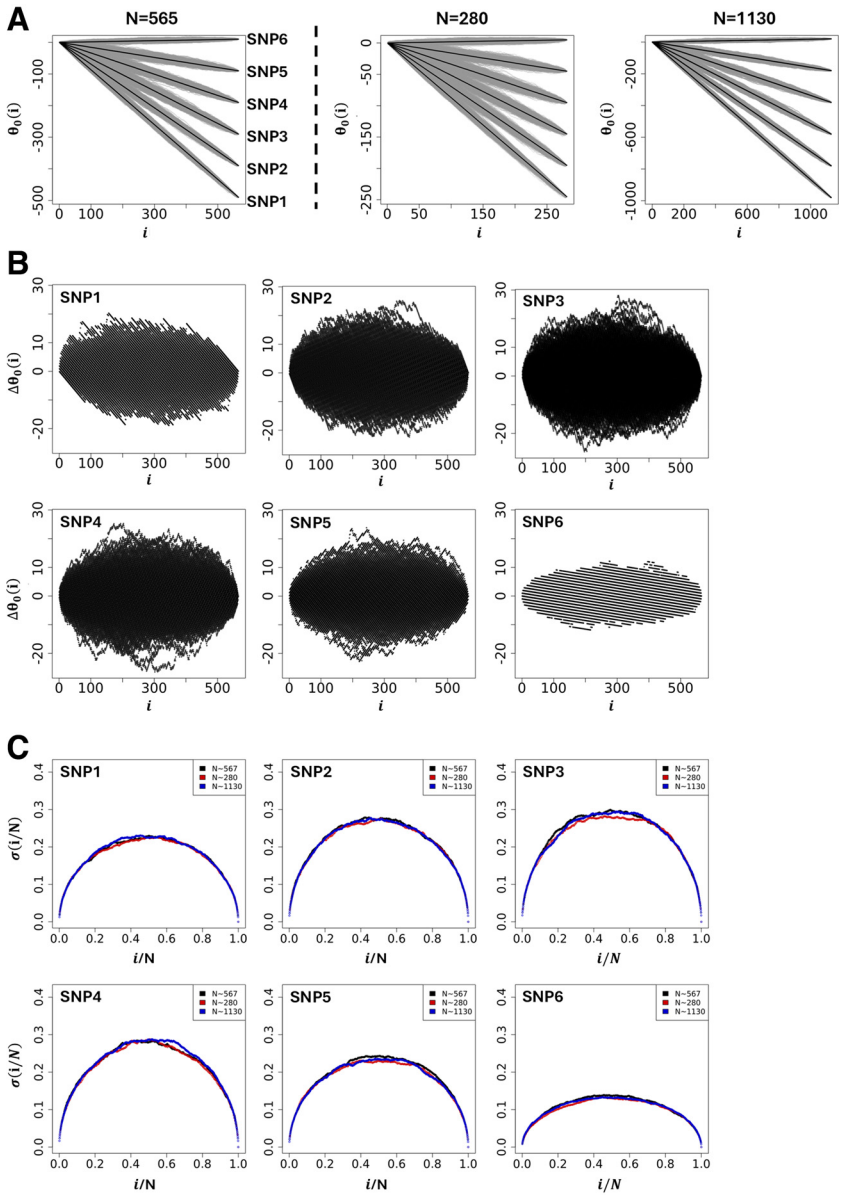


Figure 3. *A*, *left*: simulations of genetic paths corresponding to null hypotheses using GIFT as a method. The data used for the simulation are given in Table 1. *A*, *right*: simulations of genetic paths corresponding to null hypotheses when the sample size is divided or multiplied by a factor of two. *B*: representation of $\Delta\theta_0(i) = \theta_0(i) - \langle\theta_0(i)\rangle$ for the microstate data as given in Table 1. *C*: plots of the standard deviation normalized by the square root of the sample size and where the position is also normalized by the sample size. The code for the simulations is given in Supplemental File S4. GIFT, genomic informational field theory.

As any analysis must concentrate on the difference, $\Delta\theta(i) = \theta(i) - \theta_0(i)$, such as to cancel the apparent variability in the null hypothesis across SNPs, we represent the plots of the different $\Delta\theta(i)$ s obtained in the *right* panels of Fig. 4, *A–F*.

Figure 4, *A* and *B*, displays two distinct genetic paths that are globally similar. Although they have different number of microstates of each type (see Table 2), the $\Delta\theta(i)$s of SNP1 and SNP2 are characterized by their small amplitudes and the fact that they are erratic crossing several times the axis of position corresponding to the null hypothesis. In those cases, using the information contained in the phenotypic residuals, namely ranking the phenotypic residuals from small to large values, does not permit to fully differentiate $\theta(i)$ from $\theta_0(i)$. On the other hand, the *right* panels in Fig. 4, *C* and *D*, for SNP3 and SNP4 demonstrate, in a more noticeable way, a paraboloid shape for the $\Delta\theta(i)$s resulting from a segregation of microstates upon ordering the phenotypic residuals. The segregation of microstates $+1$ and -1 in opposite direction is reminiscent of Fisher theoretic works (Fig. 1). As it turns out, Fig. 4, *C* and *D*, shows some similarities with Fig. 1*C* based on a simulation inspired by Fisher's seminal works. Importantly the ΔN values of SNP1 and SNP4, while of opposite signs, are similar in absolute value, as those of SNP2 and SNP3, suggesting, in turn, the ΔN values do not impact on the ability to differentiate $\theta(i)$ from $\theta_0(i)$. Namely, that a segregation of microstates can also be inferred with relatively large and opposed ΔN values.

Envisaging the migration of microstates $+1$ and -1 in opposite directions, as initially postulated by Fisher, as the sole framework to associate genotype and phenotype is not always valid. This is demonstrated by SNP5 and SNP6 and the appearance of structured genetic paths displaying clear sigmoidal shapes for the $\Delta\theta(i)$s as shown in Fig. 4, *E* and *F*. Theoretically, this phenomenon can be understood and explained by the presence of nonlinear phenotypic fields, see Rauch et al. (11) and also reviewed in Rauch et al. (12), in turn breaking the symmetry postulated by Fisher, assuming the sole presence of linear phenotypic

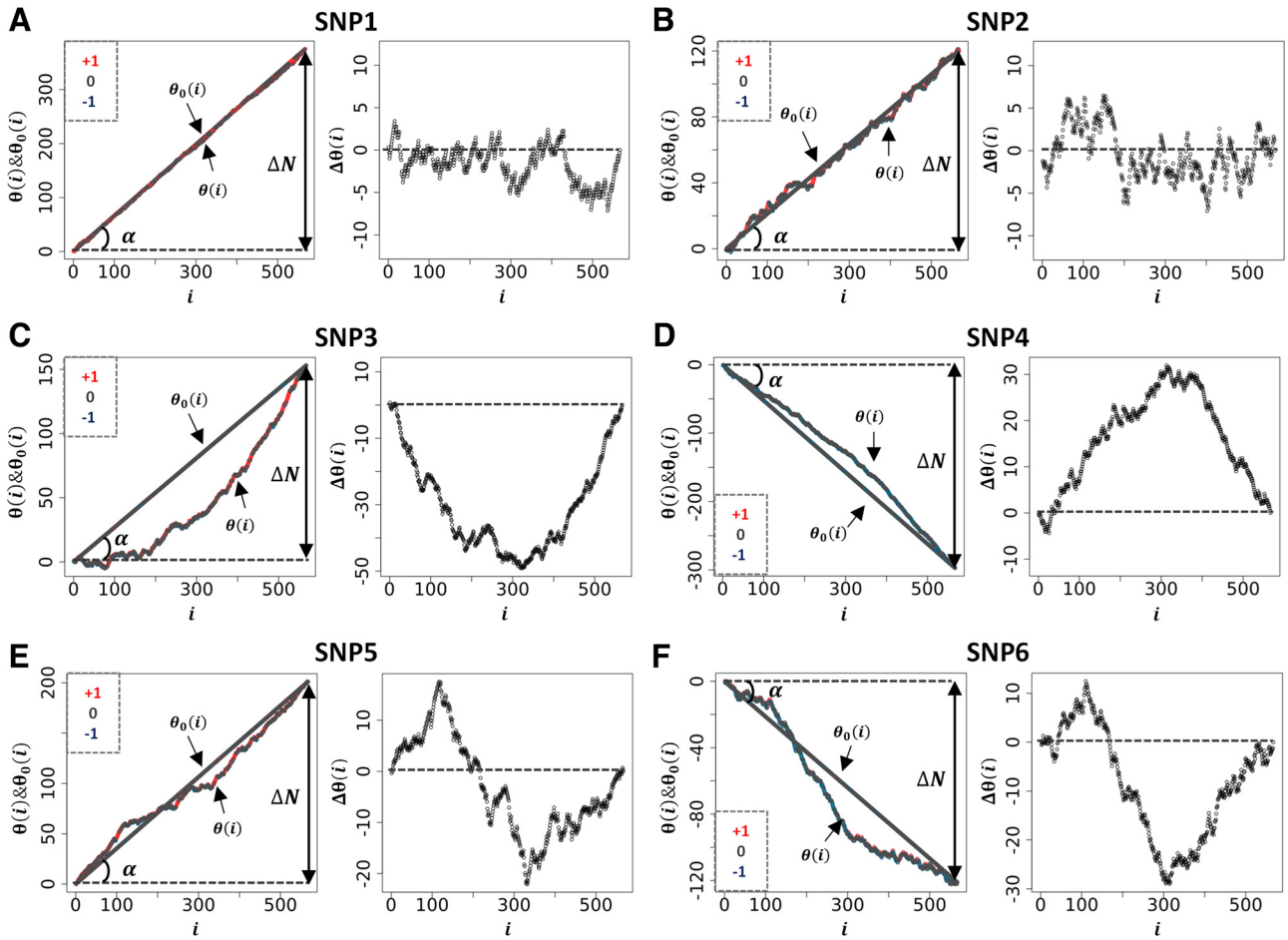


Figure 4. A sample of genetic paths selected from *dataset 1*. The details of the different SNPs displayed in *A–F* are given in Table 2. SNPs, single-nucleotide polymorphisms.

fields. This type of sigmoidal shapes is of interest since they inform on potential regulation mechanisms involving very probably "regulatory variants" (21). Indeed, the *right* panels in Fig. 4, *E* and *F*, can be envisioned as representing the genetic organization of two distinct subpopulations of phenotypic residual values, one above the dashed line and the other one underneath it. Taken separately, those two subpopulations draw curves like Fig. 4, *C* and *D*, or Fig. 1*C*. In this context, it is tempting to suggest that sigmoid genetic paths reveal a type of genotype-phenotype association that is inherently "scale dependent," namely function of the magnitude of phenotypic residuals. Because traditional GWAS concentrates on averages and variances, these sigmoid paths would be remarkably difficult to characterize with traditional methods. This is so because there is no clear antisymmetric segregation of microstates. As an example, using SNPs1-6 (from Fig. 4) we have plotted, in Fig. 5, the average values of phenotypic residuals for each microstate, and in Table 2 we provide the resulting gene/size effects and the dominances associated with those. Figure 5 and Table 2 demonstrate that sigmoid genetic paths (SNP5 and SNP6) are much less detectable with traditional methods, whereas paraboloid genetic paths (SNP3 and SNP4) are. Note that the numerical determination of "$\text{Log}_{10}(p_{\text{GIFT}})$" in Table 2, that is the significance for GIFT, is explained in the next part below.

Table 2. *Determination of gene/size effect (a) and dominance (d) for SNP1-6 from dataset 1*

CHR	Name	Position	$-\text{Log}_{10}(p_{\text{GIFT}})$	$-\text{Log}_{10}(p_{\text{GWAS}})$	N_+	N_0	N_-	a*	d**
9	OAR9_58767921 (SNP1)	56039025	2.7895	0.2735	391	160	16	N/A	N/A
3	s02120 (SNP2)	213625709	2.8893	0.0018	198	291	78	N/A	N/A
6	OAR6_40855809 (SNP3)	36655091	**28.5105**	**9.8639**	229	262	76	96.85	−13.01
6	OAR6_38315830 (SNP4)	34256151	**20.7541**	**3.7366**	24	222	321	−70.02	−0.05
23	OAR23_35510473 (SNP5)	33556377	**19.7239**	0.2301	254	260	53	N/A	N/A
25	OAR25_30372586 (SNP6)	29046746	**18.5806**	1.0692	90	266	211	N/A	N/A

The level of significance for GIFT and GWAS is indicated as follows: normal font, not significant; and bold font, significant. a*: the gene/size effect is calculated considering the middistance between the average values of phenotypic residuals of microstates "−1" and "+1." d**: the dominance is calculated considering the difference between the gene/size effect (a) and the position of the average value of phenotypic residuals for the microstate "0."

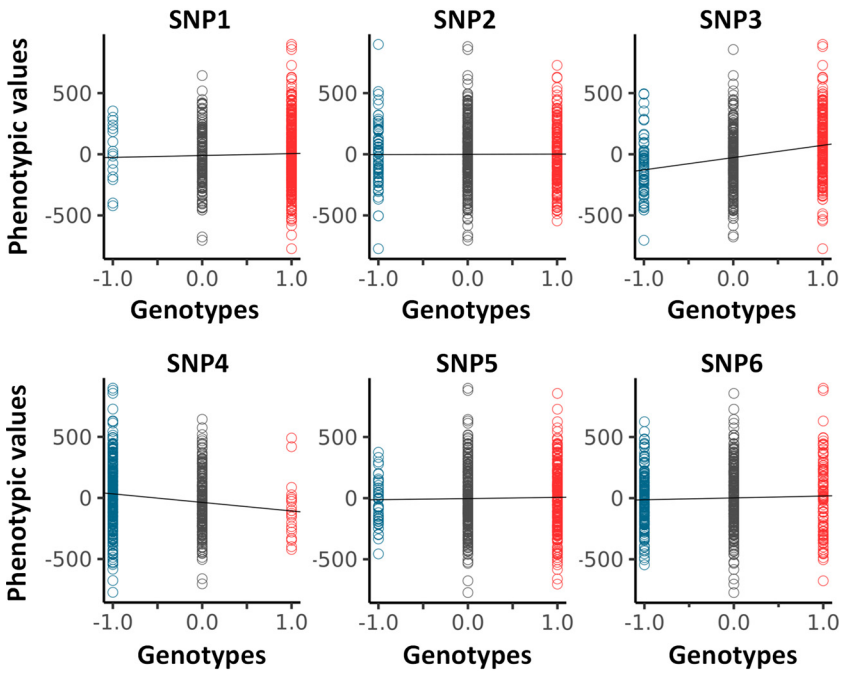


Figure 5. Analysis of averages (GWAS) for SNP1-6 (see Fig. 4 and Table 2). Values for the size/gene effects (a) and dominances (d) are given in Table 2. GWAS, genome-wide association studies.

To conclude, based on Fisher's theoretic works, the traditional GWAS method has been optimized to map SNPs that, using GIFT, would draw paraboloid genetic paths (see Fig. 1*C*). The potential novelty using GIFT resides in its ability to provide new information and detect relatively regular/structured sigmoid genetic paths that would otherwise not be detected by traditional methods.

p_{GIFT}: *p* Value for GIFT

GIFT and GWAS extract information on genotype-phenotype associations in totally different ways. Although GIFT concentrates on the significance of curves drawn using $\Delta\theta(i) = \theta(i) - \theta_0(i)$, GWAS focuses solely on the significance of difference of averages. However, to compare GIFT to GWAS, it is essential to determine a *p* value for GIFT that is exhaustive enough such as to also capture the information that GWAS provides. To this end, a *p* value was derived that concentrates on the maximal amplitude difference of genetic paths (see Fig. 6, *A* and *B*).

The *p* value for GIFT can be understood as follows. Since the number of possible paths is linked to the number of configuration possible resulting from lodging N_+, N_0, and N_- microstates into a list composed of $N = N_+ + N_0 + N_-$ components, the total number of possible paths is $N^0_{path} = \frac{N!}{N_+!N_0!N_-!}$. Let us now divide the genetic paths into regions Δi_1, Δi_2, and Δi_3 as shown in Fig. 6, *A* and *B*. As the number of microstates of each sort can be determined in each region using an adequate algorithm, the total number of possible genetic paths in this first, second, and third regions are, respectively, $N_1 = \frac{\Delta i_1!}{(n_+)_1!(n_0)_1!(n_-)_1!}$, $N_2 = \frac{\Delta i_2!}{(n_+)_2!(n_0)_2!(n_-)_2!}$ and $N_3 = \frac{\Delta i_3!}{(n_+)_3!(n_0)_3!(n_-)_3!}$, where $(n_q)_p$ is the number of microstate of type q in the pth region, $q \in \{+, 0, -\}$ and $p \in \{1, 2, 3\}$. Consequently, the probability of a genetic path in this context is, $\hat{p}_{GIFT} = N_1N_2N_3/N^0_{path}$. Using the null hypothesis simulations shown in Fig. 3, based on the theoretic SNPs given in Table 1, $\hat{p}_{GIFT}$ may be determined for each simulated genetic path. Its statistic plotted in Fig. 6*C* for each SNP demonstrates very little variations across SNPs or when the sample size changes by a factor of two. Based on this observation, confidence intervals were determined for all SNPs by averaging the $\hat{p}_{GIFT}$ values obtained. The upper and lower red dashed lines represent the 99% and 95% confidence intervals. To consider the false discovery rate (FDR) and adjust *p* values to remove type I errors, $\hat{p}_{GIFT}$ values in Fig. 6*C* were corrected using the Benjamini–Hochberg procedure leading to a new set of adjusted, i.e., reduced, *p* values, noted p_{GIFT} (see Fig. 6*D*), that may be used to determine the true significance of DNA variants (SNPs). Returning to Table 2, the numerical value of p_{GIFT} was determined for the genetic paths shown in Fig. 4, demonstrating that GIFT can extract information when sigmoid genetic paths are involved, whereas traditional GWAS is unable to do so.

Armed with p_{GIFT}, an analysis of datasets can now be performed.

Comparison Between GWAS and GIFT Considering the Bone Area of the Ischium as Phenotype (*Dataset 1*)

The first dataset (*dataset 1*) analyzed 567 pedigree-recorded Scottish Blackface lambs, concentrating on the bone areas of the ischium measured in mm^2 from cross-sectional CT scans (13). After adjusting the phenotypic values, the work demonstrated a clear involvement of chromosome 6 as shown in Fig. 7*A*. The genome-wide significant thresholds applied for GWAS in Fig. 7*A* correspond to Bonferroni corrections at 1% (upper red dashed line) and 5% (lower dashed red line), determined by using independent SNPs only. Formally a 1% (resp. 5%) Bonferroni correction is given by $-\log_{10}(0.01/N_{ind-SNPs})$ [resp. $-\log_{10}(0.05/N_{ind-SNPs})$], where $N_{ind-SNPs} = 10433$ is the number of independent SNPs. Using its own thresholds (Fig. 6*D*), GIFT was applied

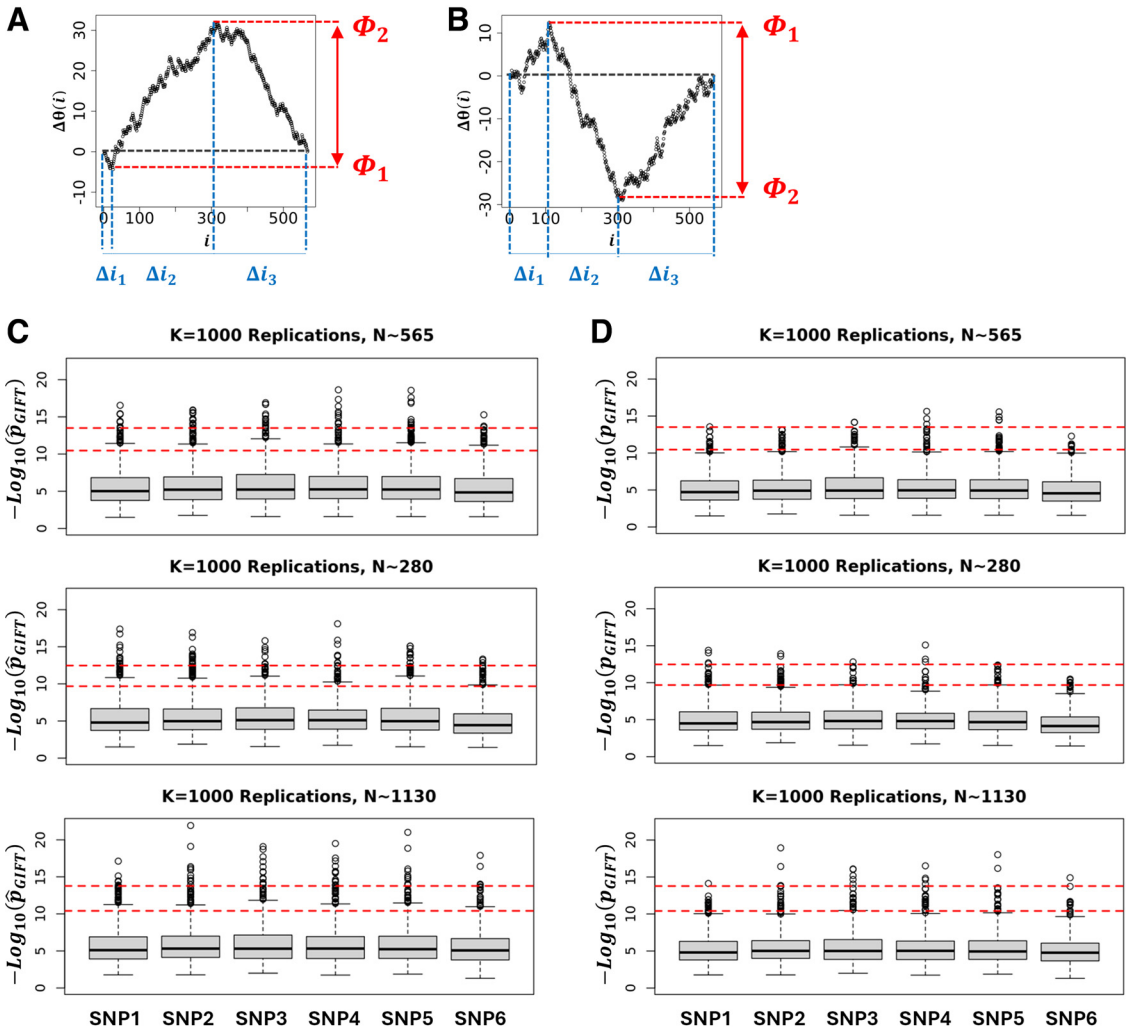


Figure 6. To provide a *p* value extracting genotype-phenotype associations in an exhaustive manner for both GWAS and GIFT, a method concentrating on the largest and smallest extreme values of the genetic path was focused upon. This method can be applied to paraboloid (GWAS-like or GIFT-like) (*A*) and sigmoid (GIFT-like) (*B*) genetic paths. The overall idea consists in determining how many paths N_1, N_2, and N_3 can be generated from the respective interval of positions Δi_1, Δi_2, and Δi_3 given that the constraints for the extrema are Φ_1 and Φ_2. Then a *p* value ($\hat{p}_{GIFT}$) can be determined as seen in the text. *C*: using simulations (K = 1,000 replicates), a statistic of $\hat{p}_{GIFT}$ for the null hypothesis can be generated using theoretic SNPs (Table 1). Simulations demonstrate that $\hat{p}_{GIFT}$ is relatively independent of the microstate's frequencies upon which a 99% (*top* dashed line) and 95% (*bottom* dashed line) interval confidences can be generated. *D*: $\hat{p}_{GIFT}$ values were adjusted to consider FDR using Benjamini–Hochberg procedure, leading to a new set of p_{GIFT} values. The code for the simulations is given in Supplemental File S4. FDR, false discovery rate; GIFT, genomic informational field theory; GWAS, genome-wide association studies; SNPs, single-nucleotide polymorphisms.

using the same set of phenotypic residuals. Figure 7, *A* and *B*, demonstrates the results obtained by GWAS and GIFT using Manhattan plots.

The significance threshold by GIFT was defined by a null hypothesis using theoretic SNPs. To demonstrate that the theoretic results obtained from Fig. 6*D* are transferrable to "real" SNPs (Fig. 7*B*), namely that the significant SNPs obtained in Fig. 7*B* have null hypotheses with similar properties like those shown in Fig. 6*D*, each significant SNP (Fig. 7*B*) had its genetic path randomly permutated a thousand times to determine the distribution of $-\log_{10}(p_{GIFT})$ values corresponding to their null hypothesis. Results show that the null hypotheses are remarkably similar across SNPs, and that the threshold determined using theoretic SNPs (Fig. 6*D*) holds when "real" SNPs are used (Supplemental File S5).

Overall, Fig. 7, *A* and *B*, demonstrates that there is an agreement between GWAS and GIFT that chromosome 6 is involved. However, differences exist that are shown through the involvement of several chromosomes when GIFT is used. Considering the thresholds involved, for GWAS, the phenotype studied may be considered as a sort of "single gene trait," whereas for GIFT, the phenotype looks very much like a "complex trait," involving more chromosomes than chromosome 6. Detailed information of all significant SNPs by GWAS or GIFT is given in Supplemental File S6.

Concentrating on chromosome 6 to address the overlap of information provided by GIFT and GWAS, a Venn diagram including highly significant SNPs only, namely SNPs beyond the *top* red dashed line in Fig. 7, *A* and *B*, was plotted. The Venn diagram (Fig. 7*C*) reveals that most SNPs

deemed significant by GWAS were also deemed significant by GIFT. Curiously, only one SNP seemed highly significant by GWAS but irrelevant for GIFT. As p_{GIFT} was designed to collect exhaustive information from GWAS, the SNP was identified (OAR6_40311379) and its genetic path, i.e., its $\Delta\theta(i)$, plotted (Fig. 7*D*, *left*) together with its GWAS representations (Fig. 7*D*, *right*). The genetic path, being erratic of relatively small amplitude and crossing several times the axis of positions, did not display any obvious "parabolic or sigmoidal" associations at first sight, in turn justifying its small p_{GIFT} value. The GWAS representation of OAR6_40311379, however, demonstrated the absence of microstate "−1" as well as a near overlap of microstates "0" and "+1" further demonstrated by the similarities between their boxplots, suggesting the occurrence of a false positive. To confirm this, a comparison of phenotypic means for the microstates "0" and "+1" was performed returning a *t* test value of 1.1485 (*p* value of 0.2512), confirming the presence of a false positive.

To assess the overlap of information between GWAS and GIFT, we plotted the first 100 more significant SNPs detected by GIFT and GWAS in Fig. 7*E*. Results confirm an overlap of GWAS and GIFT results for highly significant SNPs associated with the phenotypic residuals (see purple dots in Q_2 in Fig. 7*E*). Interestingly, two SNPs considered as significant by GWAS (two blue dots in Q_2) were not by GIFT. That is because the significance determined by GIFT for these dots were less than other SNPs detected by GIFT. As already stated earlier, many SNPs from other chromosomes were considered significant by GIFT that were not by GWAS (see red dots in Q_2). Finally, quadrant Q_2 in Fig. 7*E* confirms that OAR6_40311379, i.e., the false positive detected by GWAS, is a standalone SNP among the 100 SNPs. Finally, the biotype of significant SNPs on chromosome 6 for GIFT and GWAS is also presented in Fig. 7*F*.

The primary conclusion provided by Fig. 7, *A–F*, is that, when compared with GWAS, GIFT returns substantially more genetic information.

However, a central question concerns the genetic pertinence of the significant SNPs obtained by GIFT. As GIFT has been designed with the aim to increase the investigative power of biological datasets, we may assume that the significant SNPs obtained by GIFT once translated into gene names should underline some level of nonrandom gene-gene interactions. The latter point is particularly relevant since GIFT is expected to detect regulatory variants (cf. sigmoidal genetic paths). To assess this point, we performed an enrichment analysis based on gene names using the String database, which helps determine known and predicted protein-protein interactions. To apply String, the significant SNPs obtained using GWAS and GIFT were mapped to the reference sheep genome assembly from Ensembl (Oar_v3.1) to obtain the gene names. Using those gene names, String analyses were performed for GWAS and GIFT, using a minimum required interaction score of 0.4. Figure 7, *G* and *H*, shows the networks obtained. With enrichment *p* values for GWAS and GIFT of 0.176 and 0.00008, respectively, these results confirm that the set of genes determined by GIFT have more interactions among themselves than what would be expected for a random set of genes of the same size and degree distribution drawn from the genome. Namely that GIFT increases the investigative power of biological datasets.

At present, we do not know how the whole information provided by GIFT may inform on the putative biology of the phenotype studied (BAI). As it turns out, a full validation of the information provided by GIFT on *dataset 1* would require an in-depth mutational/deletion/insertion/gene-editing analyses in live animals, extending beyond the scope of this present article.

To demonstrate the relevance of the information provided by GIFT, we decided to challenge GIFT using a different dataset (*dataset 2*) concentrating on a complex trait related to 1C metabolism.

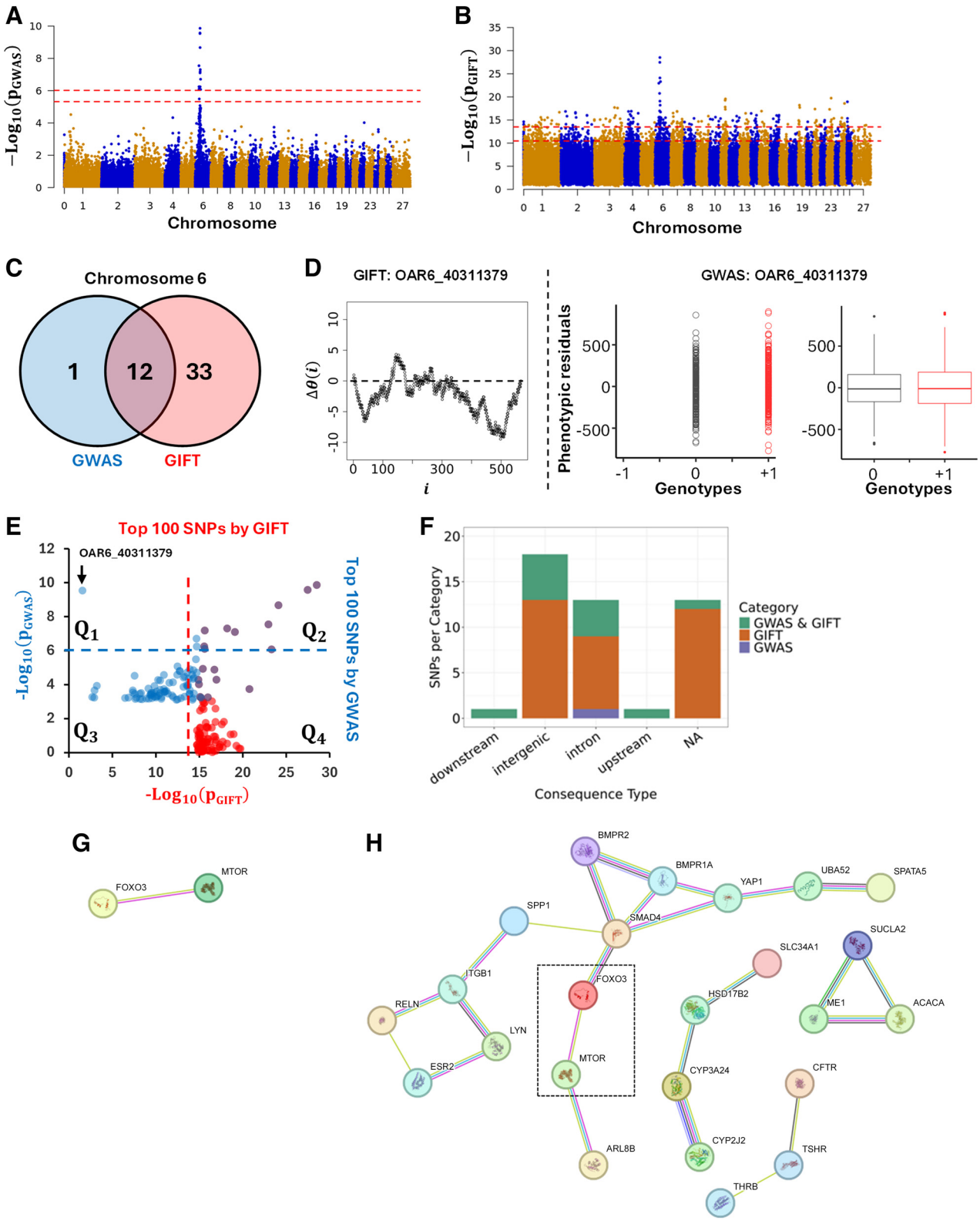
Comparison Between GWAS and GIFT Considering 1C Metabolites as Phenotype (*Dataset 2*)

Dataset 2 concerns biochemical data, which seeks to identify risk allele variants in genes whose products direct a specific series of metabolic pathways, known as one-carbon (1C) metabolism (Fig. 2). The significance of 1C metabolism is that it is a complex trait involving a series of interlinking metabolic pathways that provide 1C units (methyl groups) for the synthesis and methylation of biological molecules. After 1% and 5% Bonferroni corrections for GWAS and the Benjamini-Hochberg procedure applied to GIFT, the Manhattan plots were obtained (Fig. 8*A*). Note that the number of independent SNPs in this case is 624 (out of 3,923 SNPs from the gene array). Figure 8*A* demonstrates clearly that the informational power of GWAS is less than that of GIFT. Finally, in Fig. 8*B*, we provide the biotypes of the most significant SNPs shown by the upper red dashed lines obtained using GIFT. Detailed genetic information of the most significant SNPs obtained using GIFT is provided in Supplemental File S8.

Since the gene array was synthesized using SNPs from known genes involved in 1C metabolism, the relevance of string analyses (i.e., enrichment *p* values) would be minimal and of little interest.

Besides validating that GIFT may extract more information from genotype-phenotype datasets, it is worth underlying the biological importance and novelty of results obtained. 1C metabolism in sheep is comparable with that in humans. The significance of 1C metabolism is that it is a complex trait involving a series of interlinking metabolic pathways that provide 1C units (methyl groups) for the synthesis and methylation of chromatin among other molecules (15). *S*-adenosylmethionine (SAM) is a potent methyl donor within these cycles and serves as the principal substrate for methylation of DNA, associated proteins, and RNA. It was previously demonstrated in sheep, cattle, rodent, and human studies that disrupting these cycles during early pregnancy, by either dietary means (i.e., reducing dietary vitamin B12, folate, choline and/or methionine), or through exposure to environmental chemicals such as cigarette smoking, can lead to epigenetic dysregulation and impaired fetal development with long-term consequences for offspring cardiometabolic health (22–25). It was also advocated that interindividual and ethnic variability in epigenetic gene regulation arises because of SNPs within 1C genes, associated epigenetic

regulators, and differentially methylated target DNA sequences (15). However, information concerning the nature and extent of interactions between parental genotype, diet and EC exposure was, until now, limited to just a few 1C genes in humans (15). Consequently, data obtained by the current study provide new evidence concerning significant genetic variants in 1C metabolism and directly associated metabolic genes and epigenetic regulators that rely on SAM as the methyl donor, potentially applicable to the human species.



DISCUSSION

Although statistical association methods should not favor any biases when analyzing datasets, the way they are built mathematically is often indicative of a particular way of thinking. For example, with GWAS, the phenotype is decomposed onto more fundamental subdistributions characterized by the distribution of microstates (see Fig. 1*A*). This approach underlines a sort of bottom-up approach that, within a reductionist framework, defines genes as biological agents controlling the phenotype aligned with the "Neo-Darwinian synthesis." However, nothing prevents considering the opposite as far as statistical association methods are involved, and GIFT uses this degree of freedom. By using the full range of phenotypic information, GIFT transforms a random or disordered string of microstates (the straight line in the asymptotic limit seen in Fig. 1*C* or Fig. 3*A*) into an "ordered" configuration of microstates (see Fig. 1*C* or Fig. 4, *C–F*), in turn providing the signature of a genotype-phenotype association. Accordingly, since the phenotypic information controls the configuration of microstates, it is a top-down approach, which turns out to be remarkably sensitive. GIFT has been estimated to be ~1,000 more sensitive than GWAS (11).

There are three main reasons as to why GIFT is more sensitive. The first is that GIFT determines the significance of curves composed of an entire population of datapoints. As curves provide a greater level of significance than considering differences between microstate/phenotypic averages/variances as advocated by GWAS, GIFT is statistically more powerful. The second reason is that the null hypothesis for GIFT, namely $\theta_0(i)$, is contained in the definition of $\Delta\theta(i)$ and is therefore specific to the genome position, or SNP, studied. With GIFT, there are as many null hypotheses as SNPs. This contrasts with GWAS, defining a null hypothesis valid for all SNPs at the population level when the average of microstate distributions overlaps. Consequently, the discriminative power of GIFT is amplified. The third reason is that GIFT is simpler than GWAS. Indeed, based on R. A. Fisher's seminal work, GWAS is based on a complex theory that seeks to determine genotype-phenotype associations on one hand (*aim 1*) and the heritability of phenotypes/traits studied on the other (*aim 2*). To achieve those two aims, the GWAS approach relies on frequentist probability to determine the validity of statistical inferences giving the notions of average and variance fundamental meanings related to *aims 1* and *2*, respectively. However, because average and variance are antinomic, it is nearly impossible to have a clear picture of associations (size effects) since the noise (variance/heredity) blurs the average(s). On the other hand, by concentrating on genetic paths (curves), GIFT determines a global association. This does not mean that GIFT rules out the notions of size effect, dominance, and heritability; on the contrary, it encapsulates them under the generic notion of phenotypic field, i. e., size effect, dominance, and heritability can be rederived from the phenotypic field. The term "field" in the acronym GIFT is used to explain the disorder-order transition in the string of microstates using an analogy related to physics field theory; see Rauch et al. (11, 12) for more details.

Finally, it is important to reframe GIFT within current debates in the field of biology. With GIFT, it is the (information on the) phenotype that selects which SNP is required for its subsistence, and it is interesting to note that, at the conceptual level and as a top-down approach, GIFT has some familiarity with the notion of phenotypic plasticity. Phenotypic plasticity refers to the ability of phenotypes to respond to a change in the environment favoring a divergence from the ancestor phenotype. As the phenotype relies on traits (modules), the responsiveness to any new input(s) must involve a reorganization of the phenotype architecture by allowing phenotypic subcomponents (modular traits) to adapt the changes (26). Namely that genetic accommodation linked to a standing pool of genetic variations characterizing any trait is central to phenotypic plasticity that, through persistence, may genetically assimilate the new architecture (selection) (26, 27). In this context, the top-down method GIFT, which is essentially a phenotype-genotype (and not genotype-phenotype) association method, can pull out any standing genes awaiting to be used by phenotypes.

To conclude, we provide evidence that GIFT enhances the investigative power of biological datasets. In addition, we provide evidence also for the need to rethink the conceptual bases of genotype-phenotype association methods, such as use more information from the whole biodiversity of data.

DATA AVAILABILITY

Data including Supplemental Material are available using the following link: https://doi.org/10.1101/2024.04.16.589524.

SUPPLEMENTAL MATERIAL

Supplemental File S1 provides the raw data for *dataset 1*; Supplemental File S2 provides the raw data for *dataset 2*. Supplemental File S3 provides the statistical summary for the phenotypic adjustment before running GWAS on *dataset 2*; Supplemental File S4 provides the code to obtain Fig. 3 and Fig. 6C. Supplemental File S5 represents the permutation analysis of significant SNPs obtained by GIFT from *dataset 1*. Supplemental File S6 represents the list of significant SNPs obtained by GWAS and GIFT when applied on *dataset 1*; Supplemental File S7 provides the code to obtain Fig. 4, Fig. 5,

Figure 7. *A* and *B*: Manhattan plots based on *p* values obtained from GWAS (*A*) and GIFT (*B*) demonstrating significant differences between the methods concerning potential genotype-phenotype associations. Note that the presence of a chromosome "0" results from the fact that some SNPs identified by Matika et al. (13) were not allocated to specific chromosomes/genomic positions due to a lack of information at the time. A fathom chromosome (chromosome 0) was created to allocate those SNPs. *C*: Venn diagram representing the most significant SNPs by GWAS and GIFT. One SNP (OAR6_40311379) demonstrated a large *p* value for GWAS and a small *p* value for GIFT. A representation of its genetic path (*D*, *left*) did not underscore any "parabolic" or "sigmoidal" associations. As it turned out, this SNP was a false positive by GWAS, since the difference between the phenotypic means was not significant (*D*, *right*). *E*: the 100 most significant SNPs by GWAS and GIFT were extracted, and their *p* values were plotted against each other. The dashed lines represent the thresholds applied for GWAS (blue dashed line) and GIFT (red dashed line). The SNP OAR6_40311379 pointed by the black arrow is the single one standing out in Q_1 confirming its false positive status. *F*: biotypes of the most significant SNPs by GIFT and GWAS. *G*: string analysis was performed to determine gene networks using significant SNPs by GWAS. *H*: string analysis was performed to determine gene networks using significant SNPs by GIFT. Note that the dashed square underlines mTOR and FOXO3, which were determined by GWAS. The code for obtaining *B*, *C*, *D*, and *F* is given in Supplemental File S7. GIFT, genomic informational field theory; GWAS, genome-wide association studies; SNPs, single-nucleotide polymorphisms.

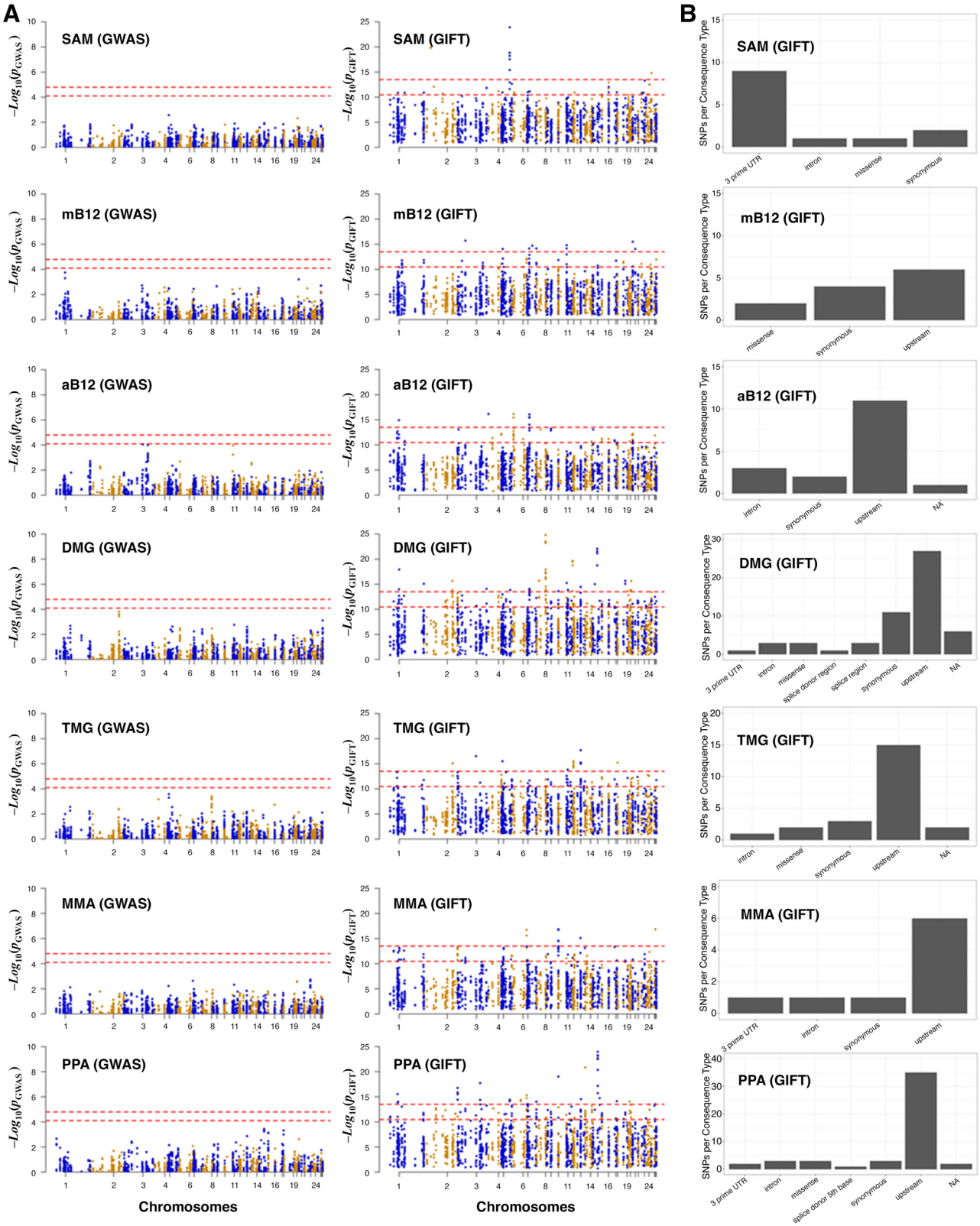


Figure 8. *A*: comparison of the information extracted by GWAS and GIFT shown using Manhattan plots for the metabolites presented in red in Fig. 2. We recall the acronyms *S*-adenosyl methionine (SAM), methylcobalamin (mB12), adenosylcobalamin (aB12), trimethylglycine (TMG), dimethylglycine (DMG), propionate (PPA), and methylmalonic acid (MMA). It should be noted that due to inherent difficulty linked to the measure of metabolite, the sample sizes were not similar across metabolites, that is, the values for *N* differ between the Manhattan plots (SAM: $N = 344$; mB12: $N = 183$; aB12: $N = 338$; DMG: $N = 338$; TMG: $N = 340$; MMA: $N = 348$; and PPA: $n = 345$). *B*: biotypes corresponding to the most significant SNPs for each metabolite determined by GIFT (a detailed list of information concerning those SNPs in given in Supplemental File S8). The code for the Manhattan plots and the determination of biotypes is given in Supplemental File S9. GIFT, genomic informational field theory; GWAS, genome-wide association studies.

and Fig. 7. Supplemental File S8 provides the list of significant SNPs by GIFT for *dataset 2*. Supplemental File S9 provides the code to obtain Fig. 8. Supplemental Material is available using the following link: https://doi.org/10.1101/2024.04.16.589524.

ACKNOWLEDGMENTS

The authors thank Dr. Barbara Bravi (Imperial College, London, UK), Dr. Wing-Yee Kwong (University of Nottingham, Sutton Bonington, UK), and Dr. Dongfang Li (University of Nottingham, Sutton Bonington, UK) for useful discussions and/or technical assistance.

GRANTS

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) Industrial Partnership Award with the Agriculture and Horticulture Development Board, Meat Promotion Wales and Agrisearch (BB/K017810/1 and BB/K017993/1), and National Institutes of Health (R01ES030374/ES/NIEHS NIH HHS/United States). P.K. is currently supported by a Doctoral Scholarship from the EPHE, Sorbonne University, in collaboration with the University of Nottingham. C.E.C. was in receipt of a BBSRC Doctoral Training Partnership scholarship (1796056), and A.H.B. was in receipt of a scholarship from The Perry Foundation.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

A.H.B., C.E.C., J.X., D.A.B., R.D.E., A.L.A., A.P., K.D.S., J.W., and C.R. conceived and designed research; P.K., A.H.B., C.E.C., and J.X. performed experiments; P.K., O.M., A.H.B., C.E.C., J.X., D.A.B., A.L.A., K.D.S., J.W., and C.R. analyzed data; A.H.B., C.E.C., J.X., D.A.B., A.P., K.D.S., J.W., and C.R., interpreted results of experiments; P.K. and C.R. prepared figures; C.R. drafted manuscript; A.P., K.D.S., J.W., and C.R. edited and revised manuscript; P.K., O.M., A.H.B., C.E.C., J.X., D.A.B., R.D.E., A.L.A., A.P., K.D.S., J.W., and C.R. approved final version of manuscript.

REFERENCES

1. **Wellcome Trust Case Control Consortium.** Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678, 2007. doi:10.1038/nature05911.
2. **Buniello A**, **MacArthur JAL**, **Cerezo M**, **Harris LW**, **Hayhurst J**, **Malangone C**, **McMahon A**, **Morales J**, **Mountjoy E**, **Sollis E**, **Suveges D**, **Vrousgou O**, **Whetzel PL**, **Amode R**, **Guillen JA**, **Riat HS**, **Trevanion SJ**, **Hall P**, **Junkins H**, **Flicek P**, **Burdett T**, **Hindorff LA**, **Cunningham F**, **Parkinson H.** The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47: D1005–D1012, 2019. doi:10.1093/nar/gky1120.
3. **Sudlow C**, **Gallacher J**, **Allen N**, **Beral V**, **Burton P**, **Danesh J**, **Downey P**, **Elliott P**, **Green J**, **Landray M**, **Liu B**, **Matthews P**, **Ong G**, **Pell J**, **Silman A**, **Young A**, **Sprosen T**, **Peakman T**, **Collins R.** UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12: e1001779, 2015. doi:10.1371/journal.pmed.1001779.
4. **Smith BH**, **Campbell A**, **Linksted P**, **Fitzpatrick B**, **Jackson C**, **Kerr SM**, **Deary IJ**, **Macintyre DJ**, **Campbell H**, **McGilchrist M**, **Hocking LJ**, **Wisely L**, **Ford I**, **Lindsay RS**, **Morton R**, **Palmer CNA**, **Dominiczak AF**, **Porteous DJ**, **Morris AD.** Cohort profile: generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* 42: 689–700, 2013. doi:10.1093/ije/dys084.
5. **Fisher RA.** XXI.—on the dominance ratio. *Proc R Soc Edinb* 42: 321–341, 1923. doi:10.1017/S0370164600023993.
6. **Fisher RA.** XV.—the correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52: 399–433, 1919. doi:10.1017/S0080456800012163.
7. **Visscher PM**, **Goddard ME.** From R.A. Fisher's 1918 paper to GWAS a century later. *Genetics* 211: 1125–1130, 2019. doi:10.1534/genetics.118.301594.
8. **Hivert V**, **Wray NR**, **Visscher PM.** Gene action, genetic variation, and GWAS: a user-friendly web tool. *PLoS Genet* 17: e1009548, 2021. doi:10.1371/journal.pgen.1009548.
9. **Nelson RM**, **Pettersson ME**, **Carlborg Ö.** A century after Fisher: time for a new paradigm in quantitative genetics. *Trends Genet* 29: 669–676, 2013. doi:10.1016/j.tig.2013.09.006.
10. **Wattis JAD**, **Bray SM**, **Kyratzi P**, **Rauch C.** Analysis of phenotype-genotype associations using genomic informational field theory (GIFT). *J Theor Biol* 548: 111198, 2022. doi:10.1016/j.jtbi.2022.111198.
11. **Rauch C**, **Kyratzi P**, **Blott S**, **Bray S**, **Wattis J.** GIFT: new method for the genetic analysis of small gene effects involving small sample sizes. *Phys Biol* 20: 016001, 2022. doi:10.1088/1478-3975/ac99b3.
12. **Rauch C**, **Wattis J**, **Bray S.** On the meaning of averages in genome-wide association studies: what should come next? *Organ J Biol Sci* 6: 7–22, 2023. doi:10.13133/2532-5876/17811.
13. **Matika O**, **Riggio V**, **Anselme-Moizan M**, **Law AS**, **Pong-Wong R**, **Archibald AL**, **Bishop SC.** Genome-wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. *Genet Sel Evol* 48: 11, 2016. doi:10.1186/s12711-016-0191-3.
14. **Clare CE**, **Pestinger V**, **Kwong WY**, **Tutt DAR**, **Xu J**, **Byrne HM**, **Barrett DA**, **Emes RD**, **Sinclair KD.** Interspecific variation in one-carbon metabolism within the ovarian follicle, oocyte, and preimplantation embryo: consequences for epigenetic programming of DNA methylation. *Int J Mol Sci* 22: 1838, 2021. doi:10.3390/ijms22041838.
15. **Clare CE**, **Brassington AH**, **Kwong WY**, **Sinclair KD.** One-carbon metabolism: linking nutritional biochemistry to epigenetic programming of long-term development. *Annu Rev Anim Biosci* 7: 263–287, 2019. doi:10.1146/annurev-animal-020518-115206.
16. **Jiang H**, **Lei R**, **Ding S-W**, **Zhu S.** Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15: 182, 2014. doi:10.1186/1471-2105-15-182.
17. **Li H**, **Durbin R.** Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760, 2009. doi:10.1093/bioinformatics/btp324.
18. **Van der Auwera GA**, **Carneiro MO**, **Hartl C**, **Poplin R**, **Del Angel G**, **Levy-Moonshine A**, **Jordan T**, **Shakir K**, **Roazen D**, **Thibault J**, **Banks E**, **Garimella KV**, **Altshuler D**, **Gabriel S**, **DePristo MA.** From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43: 11.10.1–11.10.33, 2013. doi:10.1002/0471250953.bi1110s43.
19. **McLaren W**, **Gil L**, **Hunt SE**, **Riat HS**, **Ritchie GRS**, **Thormann A**, **Flicek P**, **Cunningham F.** The Ensembl variant effect predictor. *Genome Biol* 17: 122, 2016. doi:10.1186/s13059-016-0974-4.
20. **Xu J**, **Clare CE**, **Brassington AH**, **Sinclair KD**, **Barrett DA.** Comprehensive and quantitative profiling of B vitamins and related compounds in the mammalian liver. *J Chromatogr B Analyt Technol Biomed Life Sci* 1136: 121884, 2020. doi:10.1016/j.jchromb.2019.121884.
21. **Boyle EA**, **Li YI**, **Pritchard JK.** An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169: 1177–1186, 2017. doi:10.1016/j.cell.2017.05.038.
22. **Maloney CA**, **Hay SM**, **Young LE**, **Sinclair KD**, **Rees WD.** A methyl-deficient diet fed to rat dams during the peri-conception period programs glucose homeostasis in adult male but not female offspring. *J Nutr* 141: 95–100, 2011. doi:10.3945/jn.109.119453.
23. **Sinclair KD**, **Allegrucci C**, **Singh R**, **Gardner DS**, **Sebastian S**, **Bispham J**, **Thurston A**, **Huntley JF**, **Rees WD**, **Maloney CA**, **Lea RG**, **Craigon J**, **McEvoy TG**, **Young LE.** DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. *Proc Natl Acad Sci USA* 104: 19351–19356, 2007. doi:10.1073/pnas.0707258104.
24. **Drake AJ**, **O'Shaughnessy PJ**, **Bhattacharya S**, **Monteiro A**, **Kerrigan D**, **Goetz S**, **Raab A**, **Rhind SM**, **Sinclair KD**, **Meharg AA**, **Feldmann J**, **Fowler PA.** In utero exposure to cigarette chemicals induces sex-specific disruption of one-carbon metabolism and DNA methylation in the human fetal liver. *BMC Med* 13: 18, 2015. doi:10.1186/s12916-014-0251-x.

25. **Rubini E**, **Snoek KM**, **Schoenmakers S**, **Willemsen SP**, **Sinclair KD**, **Rousian M**, **Steegers-Theunissen RPM.** First trimester maternal homocysteine and embryonic and fetal growth: the rotterdam periconception cohort. *Nutrients* 14: 1129, 2022. doi:10.3390/nu14061129.

26. **West-Eberhard MJ.** Developmental plasticity and the origin of species differences. *Proc Natl Acad Sci USA* 102, *Suppl 1*: 6543–6549, 2005. doi:10.1073/pnas.0501844102.

27. **Palmer AR.** Symmetry breaking and the evolution of development. *Science* 306: 828–833, 2004. doi:10.1126/science.1103707.