GIFT: New method for the genetic analysis of small gene effects involving small sample sizes.

Cyril Rauch[*], Panagiota Kyratzi[#], Sarah Blott[*], Sian Bray[†], Jonathan Wattis[#]

[*]School of Veterinary Medicine and Science, University of Nottingham, College Road, Sutton Bonington, LE12 5RD, UK

[†]School of Life Sciences, University of Nottingham, University Park, NG7 2RD, UK

[#]School of Mathematical Sciences, University of Nottingham, University Park, NG7 2RD, UK

Corresponding authors: Dr Cyril Rauch. Tel.: +44 (0)115 9516451. Email: cyril.rauch@nottingham.ac.uk.

Key words: phenotype-genotype mapping, fields theory, complex traits, GWAS, GIFT

Abstract

Small gene effects involved in complex/omnigenic traits remain costly to analyse using current genome-wide association methods (GWAS) because of the number of individuals required to return meaningful association(s), a.k.a. 'study power'. Inspired by the fields theory in physics, we provide a different method called Genomic Informational Field Theory (GIFT). In contrast to GWAS, GIFT assumes that the phenotype is measured precisely enough and/or the number of individuals in the population is too small to permit the creation of categories. To extract information, GIFT uses the information contained in the cumulative sums difference of gene microstates between two configurations: (i) when the individuals are taken at random without information on phenotype values, and (ii) when individuals are ranked as a function of their phenotypic value. The difference in the cumulative sum of gene microstates is then attributed to the emergence of 'phenotypic fields'. We demonstrate that GIFT recovers GWAS, that is, Fisher's theory, when the phenotypic fields are linear (first order). However, unlike GWAS, GIFT demonstrates how the variance of microstate distribution density functions can also be involved in genotype-phenotype associations when field phenotypic fields are quadratic (second order). Using genotype-phenotype simulations based on Fisher's theory as a toy model, we illustrate the application of the method with a small sample size of 1000 individuals.

1. Introduction

Identifying the association between phenotypes and genotypes is the fundamental basis of genetic analyses. In the early days of genetic studies, beginning with Mendel's work at the end of the 19th century, genotypes were inferred by tracking the inheritance of phenotypes between individuals with known relationships (linkage analysis). In recent years, the development of molecular tools, culminating in high-density genotyping and whole genome sequencing, has enabled DNA variants to be directly identified and phenotypes to be associated with genotypes in large populations of unrelated individuals through association mapping. Genome-wide association studies (GWAS) have become the method of choice, largely replacing linkage analyses, because they are more powerful for mapping complex traits, that is, they can be used to detect smaller gene effects, and they provide a greater mapping precision as they depend on population-level linkage disequilibrium rather than close family relationships. For example, the 2021 NHGRI-EBI GWAS Catalogue currently lists 316,782 associations identified in 5149 publications describing GWAS results (1). Additionally, extensive data collection has been initiated through efforts such as the UK Biobank (2), Generation Scotland (3) and NIH All of Us research program (https://allofus.nih.gov/) with the expectation that large-scale GWAS will elucidate the basis of human health and disease and facilitate precision medicine.

Whilst genomic technologies used to generate data have rapidly advanced within the last 20 years, the statistical models used in GWAS to analyse the data are still predominantly based on Fisher's method published than 100 years ago (4,5). Using probability density functions (PDFs) and in particular the normal distribution, Fisher's method partitions genotypic values by performing a linear regression of the phenotype on marker allelic dosage (6). The regression coefficient estimates the average allele effect size, and the regression variance is the additive

genetic variance due to the locus (7). Whilst Fisher's method has been improved, for example using conditional probability linked to potential prior knowledge of genetic systems (Bayes' method) (8,9), the overall determination of genotype-phenotype mapping is still grounded on PDFs. However, the use of PDFs become problematic in the case of complex/omnigenic traits as they require large scale-study or equivalently, large sample size.

The results obtained by GWAS have demonstrated that some phenotypes, a.k.a. complex traits, are driven by a vast number of tiny-effect loci, namely a vast number of genes each with tiny-effect, and not by a handful of moderate-effect loci as initially thought. In turn, this has led to a re-conceptualisation of the genetic basis of complex traits from being polygenic (handful of loci/genes) to omnigenic (vast number of loci/genes) (10). Although the omnigenic paradigm is central to further our understanding of biology, there is a practical issue concerning the extraction of information to relate genotype to phenotype in this case. Indeed, tiny-effect loci necessitate a remarkably large population to extract information. Figure 1 exemplifies the limit of GWAS with a restricted sample size of 1000 individuals. This issue regarding the need for large sample sizes was present, but dismissed, in Fisher's seminal work (4) as he assumed an 'infinite population' from the start to use the normal distribution density function in the continuum limit. This assumption allowed him to provide a method able to extract, in theory, the genetic information required to map any genotype to phenotype.

Whilst one may assume an infinite population mathematically, in practice this comes at a huge cost. To give a 'real-life' example of the sample size needed to study complex traits the best is to turn to the phenotype 'height' in humans. The phenotype height in humans is a classical quantitative trait that has been studied for over a century as a model for investigating the genetic basis of complex traits (4,11) and whose measured heritability is well known (4,12,13). However, this phenotype has remained controversial (12) for a long time as current association

methods were not been able to fully recover the heritability measured (15,16). Whilst different reasons were put forward to explain this discrepancy including, for example, too restricted sample sizes, too stringent statistical tests or the involvement of the environment (6,17); this point seems to have been resolved only recently. Using a population containing a staggering 5.3 million individuals a recent study claims to have captured nearly all of the common single nucleotide polymorphisms (SNPs)-based heritability (18).

This important study confirms that the precision of current quantitative genetic methods to determine omnigenic traits comes at an astronomical cost in line with the assumptions used, namely the need for a staggeringly large (near infinite) population. In this context one may wonder whether such large-scale study will ever be replicated in any other species and in particular those near extinction where small sample sizes need to be considered. Alternatively, one may try to understand where the need for large sample sizes comes from and determine whether it is possible to extract information linking genotype to phenotype in a different way.

There is a very good reason as to why large populations will always be required for omnigenic phenotypes when GWAS is used. As mentioned above, the reason is rooted in the fact that GWAS is mostly based on frequentist probabilities a.k.a. probability density functions (PDFs). Indeed, GWAS is based on statistics and, by definition, statistics deals with the measurement of uncertainties (19). To draw inferences from the comparison of large datasets, a method that requires some understanding of its accuracy, including ways of measuring the uncertainty in data values, is needed. In this context, statistics is the science of collecting, analysing, and interpreting data, while PDFs defined through the notion of relative frequencies, is central to determining the validity of statistical inferences. In practice, the use of frequentist probabilities (or PDFs) and the resulting categorisation of data is justified when inaccuracy exists in experimental measurements. For example, measuring a continuous phenotype such as the

height of individuals with a ruler with centimetre graduations, that is, to the nearest centimetre, warrants the use of frequentist probability (or PDFs). In this case, a frequency table of phenotypic values can be defined through 1 cm-width bins or categories, from which the PDFs of the phenotype height and of the genotypes can be deduced to address the statistical inferences. However, the precision available for the inferences will always be, at best, given by the width of categories created and linked to the experimental precision achieved (1 cm in this case). Consequently, if instead of using a ruler with centimetre graduations one was using a ruler with millimetre graduations to increase the precision in inferences, a larger sample size would be required such as to match the new 1 mm-width of categories to reform the PDFs. The trading between the sample size and the precision achieved by GWAS is known as 'study power' and its *reason d'être* is linked to the fact that the entire field of probability, and therefore the PDFs, has been conceptualised mathematically to represent the fact that information on a system is limited. It is for this reason that the normal distribution was known before as the 'error function' or 'law of errors', where the term 'error' is defined experimentally (see rulers above). Accordingly, the creation of categories implies that a sort of 'imprecision' is necessary. Whilst the notion of imprecision can be genuine (see rulers above), the act of creating categories to use PDFs when precision in experimental measurements is available is not fully justified and can be seen as an act of 'wilful ignorance'. This is so, because information is lost by slotting different phenotypic values into the same category. To exemplify this point let us take an example, imagine a species close to extinction (very small population, say 50 individuals) and that it is possible to measure phenotypic values with very high precision, for example, using highly advanced imaging techniques or biosensing technologies (20). In this case, each measured individual could return a unique phenotypic value. Consequently, reforming categories to reform and use PDFs would mean embracing the relatively large width of categories created leading to the impression that a large imprecision is present, even so such

imprecision did not exist in the first place. One may argue that PDFs, such as the normal distribution in Fisher's theory, are not required since averages and variances can be mathematically calculated directly from data without the need to recreate PDFs. However, this argument is not valid as averages and variances (and any other moments) cannot be dissociated from PDFs, since they are the ontological parameters that define PDFs, and that PDFs are used to determine statistical inferences in the field of quantitative genetics. Consequently, thinking in term of averages and variances necessitate to conceptualise and hold valid probability density functions, i.e., categories, to describe any system.

Therefore, there is a need to formulate new methods using the full information generated through accurate and highly precise genotyping/phenotyping when sample sizes are small which does not require the categorisation of data. In fact, this problem is equivalent to finding a way to resolve genotype-phenotype mapping by assuming a finite-size population with phenotype values measured precisely enough to rule out the possibility that two phenotype values are in the same category. Taking this challenge as the starting point, a new and relatively simple method for extracting information for genotype-phenotype mapping can be defined. While this method is remarkably simple when explained in lay terms, its theoretical framework requires the introduction of a new concept called the 'phenotypic field'. Phenotypic fields can also be defined within the context of Fisher's theory.

The remainder of this paper is organised as follows. A theoretical framework and an illustration using simulated data are provided, with a particular focus on small-gene effects. Furthermore, it is demonstrated in Supplementary Materials that the method converges rapidly, ruling out the need for a large population size to extract precise genotype-phenotype information.

2. Heuristic presentation of GIFT: Introduction to the genetic path method

Let us consider a population of 'N' genotyped individuals, where the phenotype of interest has been measured precisely enough such that each individual has a unique phenotype value.

For diploid organisms, such as humans, and for a binary (bi-allelic, 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.

Concentrating on a particular genome position for all individuals. Noting 'N_{+1}', 'N_0' and 'N_{-1}' the number of genetic microstates '+1', '0' and '-1', respectively; the genetic microstate frequencies for this particular genome position are defined by, $N_{+1}/N = \omega_+^0$, $N_0/N = \omega_0^0$ and $N_{-1}/N = \omega_-^0$.

To develop the model, we consider two configurations, and to illustrate these two configurations, we assume the following setting: The individuals are horses in a yard where adjacent to this yard there are 'N' non nominative aligned paddocks numbered $i = 1 \ldots N$ (Figure 2A). The variable 'i' shall be referred as the (paddock) position.

In the first configuration, the horses are allocated randomly to paddocks; that is, there is no information regarding any phenotype values. As a result, when individuals are randomly selected their genetic microstates taken together is a disordered string of '+1', '0' and '-1'. The cumulative sum of the genetic microstates found is noted '$\theta_0(i)$' where 'i' is the paddock's number or, equivalently, the position in the string of microstates. Given the random allocation, the probabilities of finding '+1', '0' or '-1' as genetic microstate at any position 'i' are ω_+^0, ω_0^0 and ω_-^0, respectively. The resulting cumulative sum is then: $\theta_0(i) = (+1 \cdot \omega_+^0 + 0 \cdot \omega_0^0 - 1 \cdot \omega_-^0)i$. '$\theta_0(i)$' and is therefore a straight line. We call '$\theta_0(i)$' the 'default genetic path'(Figures 2B and 2C).

In the second configuration, after selecting the phenotype, the horses are ranked by their phenotypic value and sent to paddock accordingly. For example, if the phenotype is height, the

smallest horse is allocated to the first paddock, and the highest horse is allocated to the last one. The new cumulative sum of the microstates is then calculated at each position in the microstate string. If an association exists between the genome position and the phenotype, then one may expect a change in the configuration of the string of microstates, in turn impacting the new cumulative sum of microstates.

Thus, the only difference between the first and second configurations is the resulting shape of the cumulative sum of genetic microstates.

Noting $\omega_+(i)$, $\omega_0(i)$ and $\omega_-(i)$ the occurrence probabilities of the genetic microstates '+1', '0' and '-1' in the second configuration the new cumulative sum is then: $\theta(i) = \sum_1^i \left(+1 \cdot \omega_+(j) + 0 \cdot \omega_0(j) - 1 \cdot \omega_-(j)\right)$; where '$\theta(i)$' is defined as the 'phenotype-responding genetic path'.

As a result, the signature of a gene interacting with the phenotype when considering the two aforementioned genetic paths is the difference: '$\theta(i) - \theta_0(i)$', including a conservation relation since the two paths must meet when $i = N$, i.e., $\theta(N) = \theta_0(N)$, for they contain the same number of '+1', '0' and '-1' (Figures 2B and 2C).

Using physics terminology, one could say that '$\theta(i) - \theta_0(i) \neq 0$' because the phenotype acts as a 'field' to change the configuration of microstates in the string. More precisely, because no information is available on the phenotype in the first configuration, the shape of the cumulative sum as a straight line arises from the random ordering of genetic microstates, corresponding to the maximisation of the entropy of the set of genetic microstates. When an association exists between genotype and phenotype the signature is a cumulative sum with a curved shape and in this case, the phenotypic 'field' is involved to order the genetic microstates.

The core of the present study is to model '$\theta(i) - \theta_0(i)$' and demonstrate how genotype-phenotype associations can be inferred.

3. The genetic path model: mathematical principles underlying the method

3.1. Differential expression of the genetic path in the space of phenotype values

We assume that the phenotype values are measured precisely enough such that each individual has a unique phenotype value noted 'Ω_i'. As the population is composed of 'N' individuals one defines, $\hat{\imath} \stackrel{\text{def}}{=} i/N$ and '$\Omega_{\hat{\imath}}$' the new position and its corresponding phenotype value. Thus '$\Omega_{1/N}$' and 'Ω_1' are the smallest and largest phenotype values, respectively.

As a result, the cumulative sum of presence probability of genetic microstates as a function of '$\hat{\imath}$' can be written as: $\theta(\hat{\imath}) = \sum_{\hat{\jmath}=\frac{1}{N}}^{\hat{\jmath}=\hat{\imath}} \left(\omega_+(\hat{\jmath}) - \omega_-(\hat{\jmath})\right)$, where '$\omega_+(\hat{\jmath})$' and '$\omega_-(\hat{\jmath})$' are the presence probabilities of microstates '+1' and '-1' at the position '$\hat{\jmath}$'. Using the continuum limit,

$$\theta(\hat{\imath}) \sim \int_{\frac{1}{N}}^{\hat{\imath}} \left(\omega_+(\hat{\jmath}) - \omega_-(\hat{\jmath})\right) d\hat{\jmath} \qquad (1)$$

As the ranking of phenotype values was introduced to define '$\theta(\hat{\imath})$', the genetic path can also be expressed as a function of phenotype values under the form: $\hat{\theta}(\Omega_{\hat{\imath}}) = \sum_{\hat{\jmath}=1/N}^{\hat{\jmath}=\hat{\imath}} \left(\hat{\omega}_+(\Omega_{\hat{\jmath}}) - \hat{\omega}_-(\Omega_{\hat{\jmath}})\right)$ where the hat '$\hat{}$' is added to describe new functions linked to phenotypic values. Accordingly, by definition, $\hat{\theta}(\Omega_{\hat{\imath}}) - \hat{\theta}(\Omega_{\hat{\imath}-1}) = \hat{\omega}_+(\Omega_{\hat{\imath}}) - \hat{\omega}_-(\Omega_{\hat{\imath}})$. Assuming that the measured phenotype values are sufficiently close, the continuum limit can be used to determine $\hat{\theta}(\Omega_{\hat{\imath}}) - \hat{\theta}(\Omega_{\hat{\imath}-1}) \sim \frac{d\hat{\theta}}{d\Omega} \cdot \Delta(\Omega)$, where '$\Delta(\Omega)$' is the spacing between two consecutive individuals in the space of phenotype values. One deduces then: $\frac{d\hat{\theta}}{d\Omega} \sim \left(\hat{\omega}_+(\Omega) - \hat{\omega}_-(\Omega)\right) \frac{1}{\Delta(\Omega)}$. As a result, the genetic path can be expressed in the space of the phenotype values as:

$$\hat{\theta}(\Omega) \sim \int_{\Omega_{1/N}}^{\Omega} \left(\hat{\omega}_+(x) - \hat{\omega}_-(x)\right) \frac{1}{\Delta(x)} dx \qquad (2)$$

Where $\int_{\Omega_{1/N}}^{\Omega_1} \Delta(\Omega)\, d\Omega = \Omega_1 - \Omega_{1/N} = \Delta\Omega$.

As '$\Delta(\Omega)$' is the phenotypic space between individual, provided that the population is large and dense enough one can relate '$\Delta(\Omega)$' to the phenotype distribution density function, $P_\Omega(\Omega)$, under the form $1/\Delta(\Omega) \sim P_\Omega(\Omega)$ (SM1 in the Supplementary Materials).

The different elements that lead to the formation of a genetic path can now be addressed.

3.2. Entropy of genetic paths

Keeping the notations ω_+^0, ω_0^0 and ω_-^0, for the genetic microstate frequencies of a given genome position across the population of individuals, we aim to determine the expressions of $\omega_+(\hat{\imath})$, $\omega_0(\hat{\imath})$ and $\omega_-(\hat{\imath})$ given the information obtained upon ordering the genotypes as a function of phenotype values along the $\hat{\imath}$-axis.

The default genetic path 'θ_0' as a straight line is defined by the absence of information on phenotype values, which is similar to an absence of association between the genetic microstates and the phenotype values, leading to an apparent disordering of genetic microstates. One way to measure this disordering is by using the 'entropy' of the string of genetic microstates for the genome position considered. In this context, the entropy is given by calculating the number of possible combinations of placing $N_+ = N\omega_+^0$, $N_0 = N\omega_0^0$ and $N_- = N\omega_-^0$ genetic microstates over 'N' possible positions. Consequently, the entropy of the default genetic path is $S_0 = N!/N_+!\,N_0!\,N_-!$; and for 'N' large enough using Stirling's formula one deduces: $S_0/N \sim -\omega_+^0 \ln(\omega_+^0) - \omega_0^0 \ln(\omega_0^0) - \omega_-^0 \ln(\omega_-^0)$, that can be rewritten in the continuum limit as: $S_0 \sim -N \int_{1/N}^{1} \left(\omega_+^0 \ln(\omega_+^0) + \omega_0^0 \ln(\omega_0^0) + \omega_-^0 \ln(\omega_-^0)\right) d\hat{\imath}$. Note that, as the genetic microstate frequencies are constant in this case, the entropy can be rewritten using the phenotype values as

$$S_0 \sim -N \int_{\Omega_{1/N}}^{\Omega_1} \left(\omega_+^0 \ln(\omega_+^0) + \omega_0^0 \ln(\omega_0^0) + \omega_-^0 \ln(\omega_-^0)\right) \frac{1}{\Delta(\Omega)} d\Omega.$$

When information about phenotype values and their ranking is given and when the genome position considered is associated with the phenotype, 'S_0' is transformed to 'S' where:

$$S = -N \int_{\Omega_{1/N}}^{\Omega_1} \left(\widehat{\omega}_+(\Omega)\ln\left(\widehat{\omega}_+(\Omega)\right) + \widehat{\omega}_0(\Omega)\ln\left(\widehat{\omega}_0(\Omega)\right) + \widehat{\omega}_-(\Omega)\ln\left(\widehat{\omega}_-(\Omega)\right)\right)\frac{1}{\Delta(\Omega)}d\Omega$$

As a result, the entropy difference '$S - S_0$' when non-null provides information on whether the genome position is associated with phenotype values. Thus, '$S - S_0$' can be thought of as a 'transformation' in a physical/thermodynamic sense. That is, the difference in entropies must be balanced by a term that is linked to the association (or interaction) between the genetic microstates and the phenotype values.

3.3. Introduction of phenotypic fields as a source for genotype-phenotype associations

As '$S - S_0$' is linked to the information gained from knowing phenotypic values and ranking them, given the existence of three distinct genetic microstates, one can define three distinct functions a.k.a. phenotypic fields '$u_+(\Omega)$', '$u_0(\Omega)$' and '$u_-(\Omega)$' that are fundamentally related to changes in the phenotype-associated genetic path.

In this context, the entire genetic path can be defined with a function representing the sum of interactions between each of the genetic microstates and phenotypic fields:

$$E \sim N \int_{\Omega_{1/N}}^{\Omega_1} \left(u_+(\Omega)\widehat{\omega}_+(\Omega) + u_0(\Omega)\widehat{\omega}_0(\Omega) + u_-(\Omega)\widehat{\omega}_-(\Omega)\right)\frac{1}{\Delta(\Omega)}d\Omega$$

In this context, one may consider that the set of microstates changes the configuration because the fields are 'switch on'. This implies that for the genome positions that are not involved in the formation of the phenotype considered, the switch does not work, that is, the fields are null. In this context, one can consider the equivalence, $S - S_0 \sim E$.

As a result, the relationship to optimise is:

$$\Delta S - E = 0$$

Recalling the conservation of genetic microstates for the genome position considered: $\int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_+(\Omega)\frac{1}{\Delta(\Omega)}d\Omega = \omega_+^0$, $\int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_0(\Omega)\frac{1}{\Delta(\Omega)}d\Omega = \omega_0^0$ and $\int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_-(\Omega)\frac{1}{\Delta(\Omega)}d\Omega = \omega_-^0$ together with the conservation of probability, $\widehat{\omega}_+(\Omega) + \widehat{\omega}_0(\Omega) + \widehat{\omega}_-(\Omega) = 1$, Euler-Lagrange's method can then be used to determine the optimal configuration for $\widehat{\omega}_+(\Omega)$, $\widehat{\omega}_0(\Omega)$ and $\widehat{\omega}_-(\Omega)$ in a context where 's' is fixed and the phenotypic fields are imposed. By defining α_+, α_0 and α_-, the Lagrange multipliers for the conservation of genetic microstates, the relation to optimise with regard to the genetic microstate frequencies $\widehat{\omega}_+(\Omega)$, $\widehat{\omega}_0(\Omega)$ and $\widehat{\omega}_-(\Omega)$ is then:

$$\Delta S/N - E/N - \alpha_+\left(\omega_+^0 - \int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_+(\Omega)\frac{1}{\Delta(\Omega)}d\Omega\right) - \alpha_0\left(\omega_0^0 - \int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_0(\Omega)\frac{1}{\Delta(\Omega)}d\Omega\right) - \alpha_-\left(\omega_-^0 - \int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_-(\Omega)\frac{1}{\Delta(\Omega)}d\Omega\right) = 0$$

Using the conservation of genetic microstate frequencies, '$\widehat{\omega}_0(\Omega)$' can be replaced by '$1 - \widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega)$', and a variational calculus can be performed on the genetic microstate frequencies, leading to two conditions:

$$\delta\widehat{\omega}_+(\Omega)\left[\ln\left(\frac{\widehat{\omega}_+(\Omega)}{1-\widehat{\omega}_+(\Omega)-\widehat{\omega}_-(\Omega)}\right) - \delta u_+(\Omega) + (\alpha_+ - \alpha_0)\right] = 0$$

$$\delta\widehat{\omega}_-(\Omega)\left[\ln\left(\frac{\widehat{\omega}_-(\Omega)}{1-\widehat{\omega}_+(\Omega)-\widehat{\omega}_-(\Omega)}\right) - \delta u_-(\Omega) + (\alpha_- - \alpha_0)\right] = 0$$

Where $\delta\widehat{\omega}_+(\Omega)$ and $\delta\widehat{\omega}_-(\Omega)$ are small variations in the presence probabilities of microstates '+1' and '-1', and $\delta u_+(\Omega) \stackrel{\text{def}}{=} u_+(\Omega) - u_0(\Omega)$, $\delta u_-(\Omega) \stackrel{\text{def}}{=} u_-(\Omega) - u_0(\Omega)$. Finally, '$\widehat{\omega}_0(\Omega)$' can be deduced using $\widehat{\omega}_0(\Omega) = 1 - \widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega)$.

Using the conditions $\widehat{\omega}_+(\Omega) = \omega_+^0$, $\widehat{\omega}_0(\Omega) = \omega_0^0$ and $\widehat{\omega}_-(\Omega) = \omega_-^0$ when $s = 0$, we obtain

$$\widehat{\omega}_+(\Omega) = \frac{\omega_+^0 e^{-\delta u_+(\Omega)}}{\omega_0^0 + \omega_+^0 e^{-\delta u_+(\Omega)} + \omega_-^0 e^{-\delta u_-(\Omega)}} \quad (3)$$

$$\widehat{\omega}_-(\Omega) = \frac{\omega_-^0 e^{-\delta u_-(\Omega)}}{\omega_0^0 + \omega_+^0 e^{-\delta u_+(\Omega)} + \omega_-^0 e^{-\delta u_-(\Omega)}} \quad (4)$$

To make the asymmetries of the problem more apparent, the following are defined for genetic microstates:

$\Delta\omega_0 \stackrel{\text{def}}{=} \omega_+^0 - \omega_-^0$

$\omega_0 \stackrel{\text{def}}{=} \omega_+^0 + \omega_-^0 = 1 - \omega_0^0$

and for the phenotypic fields:

$2\bar{u}(\Omega) \stackrel{\text{def}}{=} \delta u_+(\Omega) + \delta u_-(\Omega) = u_+(\Omega) + u_-(\Omega) - s u_0(\Omega)$

$2\Delta u(\Omega) \stackrel{\text{def}}{=} \delta u_+(\Omega) - \delta u_-(\Omega) = u_+(\Omega) - u_-(\Omega)$

Then, the difference and sum of $\widehat{\omega}_+(\Omega)$ and $\widehat{\omega}_-(\Omega)$ can be rewritten as follow:

$$\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega) = \frac{\omega_0 e^{-\bar{u}(\Omega)}\left[\frac{\Delta\omega_0}{\omega_0}\mathrm{ch}\left(\Delta u(\Omega)\right) - s\ \left(\Delta u(\Omega)\right)\right]}{1-\omega_0+\omega_0 e^{-\bar{u}(\Omega)}\left[\mathrm{ch}\left(\Delta u(\Omega)\right) - \frac{\Delta\omega_0}{\omega_0}\mathrm{sh}\left(\Delta u(\Omega)\right)\right]}$$

$$\widehat{\omega}_+(\Omega) + \widehat{\omega}_-(\Omega) = \frac{\omega_0 e^{-\bar{u}(\Omega)}\left[\mathrm{ch}\left(\Delta u(\Omega)\right) - \frac{\Delta\omega_0}{\omega_0}\mathrm{h}\left(s\Delta u(\Omega)\right)\right]}{1-\omega_0+\omega_0 e^{-\bar{u}(\Omega)}\left[\mathrm{ch}\left(\Delta u(\Omega)\right) - \frac{\Delta\omega_0}{\omega_0}\mathrm{sh}\left(\Delta u(\Omega)\right)\right]}$$

Noting that: $-1 \leq \Delta\omega_0/\omega_0 \leq +1$, a new phenotype value is defined and noted 'Ω_0' by setting $\mathrm{th}\left(\Delta u(\Omega_0)\right) \stackrel{\text{def}}{=} \frac{\Delta\omega_0}{\omega_0}$. Then, the difference and sum of $\widehat{\omega}_+(\Omega)$ and $\widehat{\omega}_-(\Omega)$ can be rewritten as:

$$\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega) = \frac{\mathrm{sh}\left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)}{\alpha_0 e^{\bar{u}(\Omega)} + \mathrm{c}\ \left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)} \tag{5}$$

$$\widehat{\omega}_+(\Omega) + \widehat{\omega}_-(\Omega) = \frac{\mathrm{ch}\left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)}{\alpha_0 e^{\bar{u}(\Omega)} + \mathrm{ch}\left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)} \tag{6}$$

$$\alpha_0 \stackrel{\text{def}}{=} \frac{1-\omega_0}{\omega_0}\mathrm{ch}\left(s\Delta u(\Omega_0)\right) = \frac{1-\omega_0}{\omega_0}\frac{1}{\sqrt{1-\left(\mathrm{th}\left(s\Delta u(\Omega_0)\right)\right)^2}} = \frac{1-\omega_0}{\sqrt{{\omega_0}^2 - {\Delta\omega_0}^2}} = \frac{1}{2}\frac{\omega_0^0}{\sqrt{\omega_+^0\omega_-^0}}$$

The new variable 'Ω_0' is the phenotype value corresponding to the condition $\widehat{\omega}_+(\Omega_0) \sim \widehat{\omega}_-(\Omega_0)$.

The meaning of the constant 'α_0' can be related to the Hardy-Weinberg law. Hardy-Weinberg law based on random matting in a population provides a relationship between the genetic microstate frequencies under the form: $p^2 + 2pq + q^2 = 1$, where p^2 and q^2 are the genotype frequencies of genetic microstates '+1' and '-1', i.e. homozygote genotypes aa and AA, respectively; and $2pq$ the genotype frequency for genetic microstate '0', i.e. the heterozygote

genotype Aa. In our case, this corresponds to replacing p^2, q^2 and $2pq$ with, respectively, ω_+^0, ω_-^0 and ω_0^0. Consequently, the Hardy-Weinberg law imposes $\alpha_0 = 1$ with $\alpha_0 \neq 1$ corresponding to a deviation from the law. However, this term is expected to remain stable upon any changes of allele or genotype frequencies suggesting therefore that, genetically, any changes in '$\Delta\omega_0$' are to some extent compensated by corresponding changes in 'ω_0'.

We can now turn to the full expression of the genetic path difference in the space of the phenotype value:

3.4 Expression of the genetics paths difference and conservation of genetic microstate frequencies.

The phenotype-associated genetic path is simply the integration of Eq.5 over the phenotype values that is given, as seen above (Eq.2), by: $\int_{\Omega_{1/N}}^{\Omega}[\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega)]\frac{1}{\Delta(\Omega)}d\Omega$. The default genetic path is deduced from considering that the difference in the presence probabilities between the genetic microstates '+1' and '-1' is constant, i.e.: $\int_{\Omega_{1/N}}^{\Omega}\Delta\omega_0\frac{1}{\Delta(\Omega)}d\Omega$.

By rewriting '$\Delta\omega_0$' as '$\omega_0(\Delta\omega_0/\omega_0)$', where $\Delta\omega_0/\omega_0 = \mathrm{th}\big(\Delta u(\Omega_0)\big)$ and deducing 'ω_0' from $\alpha_0 = (1-\omega_0)\mathrm{ch}\big(\Delta u(\Omega_0)\big)/\omega_0$; it follows: $\Delta\omega_0 = \frac{\mathrm{sh}\big(\Delta u(\Omega_0)\big)}{\alpha_0+\mathrm{ch}\big(\Delta u(\Omega_0)\big)}$. As a result, the difference between the phenotype-associated and default genetic paths expressed as a function of phenotypic fields within the continuum limit is:

$$\Delta\hat{\theta}(\Omega) = \int_{\Omega_{1/N}}^{\Omega}\frac{\mathrm{sh}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}{\alpha_0 e^{\bar{u}(\Omega)}+\mathrm{ch}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}\frac{1}{\Delta(\Omega)}d\Omega - \frac{\mathrm{sh}\big(\Delta u(\Omega_0)\big)}{\alpha_0+\mathrm{ch}\big(\Delta u(\Omega_0)\big)}\int_{\Omega_{1/N}}^{\Omega}\frac{1}{\Delta(\Omega)}d\Omega \quad (7)$$

The conservation of genetic microstates needs to be added regardless of the genetic path taken, that is, $\Delta\hat{\theta}(\Omega_1) = 0$, expressed as:

$$\int_{\Omega_{1/N}}^{\Omega_1}\frac{\mathrm{sh}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}{\alpha_0 e^{\bar{u}(\Omega)}+\mathrm{ch}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}\frac{1}{\Delta(\Omega)}d\Omega = \frac{\mathrm{sh}\big(\Delta u(\Omega_0)\big)}{\alpha_0+c\ \big(\Delta u(\Omega_0)\big)} \quad (8)$$

Therefore, as 'α_0' is constant since a single genome position is considered, the genetic paths difference can be re-expressed integrally using two independent reduced phenotypic fields, i.e., $\Delta u(\Omega)$ and $\bar{u}(\Omega)$, and Eq.8 provides a coupling between those fields and '$\Delta(\Omega)$'. The advantage of using fields is the reduction of unknown parameters involved in the problem and the possibility of laying out genotype-phenotype associations based on fields' symmetry. For example, and as a minimalist model, one may wonder what sort of expression would take the fields if the reference field was null and, $s\delta u_+$ and $s\delta u_-$, were acting anti-symmetrically and linearly on the microstates '+1' and '-1'?

This minimalist model can be developed (see SM2 in the Supplementary Materials) and is similar to Fisher's seminal intuition (see below).

3.5. Coarse graining GIFT: Definition of Fisher's fields, implication for small gene effects and definition of variance fields.

3.5.1. Coarse graining GIFT

To derive a coarse-grained version of GIFT, that is, a genetic path difference for GWAS, we assume the existence of categories or bins and concentrate on the interval of phenotype values ranging between 'Ω' and '$\Omega + \delta\Omega$' defining one particular category or bin.

Based on frequentist probability, by noting by 'N' the total number of individuals in the population we can define by: $\delta N \sim N \cdot P_\Omega(\Omega)\delta\Omega$, the number of individuals in the phenotype category concerned namely with a phenotype value ranging between 'Ω' and '$\Omega + \delta\Omega$'.

Similarly, concentrating on a single genome potion, we can define by: $\delta N_+ \sim N_+^0 \cdot P_+(\Omega)\delta\Omega$, $\delta N_0 \sim N_0^0 \cdot P_0(\Omega)\delta\Omega$ and $\delta N_- \sim N_-^0 \cdot P_-(\Omega)\delta\Omega$ the respective number of '+1', '0' and '-1' genetic microstates for the phenotype values ranging between 'Ω' and '$\Omega + \delta\Omega$', where 'N_+^0',

'N_0^0' and 'N_-^0' correspond to the total number of '+1', '0' and '-1' microstates in the population with respective presence probability density functions '$P_+(\Omega)$', '$P_0(\Omega)$' and '$P_-(\Omega)$'.

The design of categories generates two conservation relationships: The first one concerns the total number of individuals and microstates, namely, that for a given genome position, the sum of all possible microstates is also the sum of all individuals. This relationship is written as follows:

$$N = N_+^0 + N_0^0 + N_-^0$$

The second conservation relationship is linked to the category considered. The number of individuals in the category concerned is also the sum of the microstates in this category:

$$\delta N = \delta N_+ + \delta N_0 + \delta N_-$$

Consequently, the conservation relation concerning the number of individuals and microstates in the concerned category can be rewritten as

$$\delta N/N = (N_+^0/N)(\delta N_+/N_+^0) + (N_0^0/N)(\delta N_0/N_0^0) + (N_-^0/N)(\delta N_-/N_-^0)$$

Using the probability density function of both microstates and the phenotype defined above, the following is deduced:

$$P_\Omega(\Omega) = \omega_+^0 P_+(\Omega) + \omega_0^0 P_0(\Omega) + \omega_-^0 P_-(\Omega) \qquad (9a)$$

From Eq.9 all the moments of microstate distributions are related to those of the phenotype distribution. Let us note by '$\langle\Omega\rangle$', and 'a_+', 'a_0', 'a_-' the average values of the phenotype and microstates '+1', '0', '-1' distribution functions, respectively; and by 'σ^2' and 'σ_+^2', 'σ_0^2', 'σ_-^2' the variances of the phenotype and microstates '+1', '0', '-1' distribution functions, respectively. From Eq.9a, one deduces the conservation relations for the first two moments in the form:

$$\langle\Omega\rangle = \omega_+^0 a_+ + \omega_0^0 a_0 + \omega_-^0 a_-$$

$$\sigma^2 = (\omega_+^0\sigma_+^2 + \omega_0^0\sigma_0^2 + \omega_-^0\sigma_-^2) + \omega_+^0(\langle\Omega\rangle - a_+)^2 + \omega_0^0(\langle\Omega\rangle - a_0)^2 + \omega_-^0(\langle\Omega\rangle - a_-)^2 \qquad (9b)$$

The relations provided by Eq.9b are valid by definition, namely, whatever probability density functions are involved. Whilst Fisher never formulated a conservation relation similar to Eq.9a, in his seminal paper (4) he used the notation 'α^2' to denote the genetic variance in the form:

$$\alpha^2 \stackrel{\text{def}}{=} \omega_+^0(\langle\Omega\rangle - a_+)^2 + \omega_0^0(\langle\Omega\rangle - a_0)^2 + \omega_-^0(\langle\Omega\rangle - a_-)^2$$

Let us now define the coarse-grained version of Eq.5 by noting '$\delta\widehat{\omega}_+(\Omega) - \delta\widehat{\omega}_-(\Omega)$' the difference in the presence probability of microstates '+1' and '-1' for the category of interest, it is then deduced:

$$\delta\widehat{\omega}_+(\Omega) - \delta\widehat{\omega}_-(\Omega) = \frac{\delta N_+}{\delta N} - \frac{\delta N_-}{\delta N} = \frac{\omega_+^0 P_+(\Omega) - \omega_-^0 P_-(\Omega)}{\omega_+^0 P_+(\Omega) + \omega_0^0 P_0(\Omega) + \omega_-^0 P_-(\Omega)} \qquad (10)$$

Also, it can be verified that the conservation of gene microstates holds by summing '$\delta N_+ - \delta N_-$' over all existing categories or bins, namely:

$$N_+ - N_- = \sum_{\text{bins}}(\delta N_+ - \delta N_-) = \sum_{\text{bins}}\big(\delta\widehat{\omega}_+(\Omega) - \delta\widehat{\omega}_-(\Omega)\big)\delta N$$

Using the definition $\delta N = NP_\Omega(\Omega)\delta\Omega$ together with Eq.9, we deduce using the continuum limit:

$$\frac{N_+ - N_-}{N} \sim \int_{\Omega_{1/N}}^{\Omega_1} \frac{\omega_+^0 P_+(\Omega) - \omega_-^0 P_-(\Omega)}{\omega_+^0 P_+(\Omega) + \omega_0^0 P_0(\Omega) + \omega_-^0 P_-(\Omega)} P_\Omega(\Omega)\delta\Omega = \int_{\Omega_{1/N}}^{\Omega_1} \big(\omega_+^0 P_+(\Omega) - \omega_-^0 P_-(\Omega)\big)\delta\Omega = \omega_+^0 - \omega_-^0$$

Direct mapping of fields can then be performed between GWAS and GIFT (SM3 in the Supplementary Materials). As a result, we can define the coarse-grained versions of the difference in the genetic paths using the continuum limit as:

$$\Delta\hat{\theta}(\Omega) \sim \int_{\Omega_{1/N}}^{\Omega} \left[\frac{\omega_+^0 P_+(\Omega) - \omega_-^0 P_-(\Omega)}{\omega_+^0 P_+(\Omega) + \omega_0^0 P_0(\Omega) + \omega_-^0 P_-(\Omega)} - (\omega_+^0 - \omega_-^0)\right] P_\Omega(\Omega)d\Omega \qquad (11)$$

Eq.11 demonstrates that '$\Delta\hat{\theta}(\Omega)$' is sensitive to the probability density functions involved as a whole and not just to the average values. In other words, the variance of microstates and their average values will impact on genotype-phenotype associations.

Note that in Eq.11 the integration interval is unchanged. However, the convergence in probability of distributions allows some freedom, for example changing the integration interval

from $[\Omega_{1/N};\Omega_1]$ to $[0;+\infty[$, or from $[\Omega_{1/N}-\Omega_m;\Omega_1-\Omega_m]$ to $]-\infty;+\infty[$, where 'Ω_m' is a median position.

The fields can then be defined setting: $-\delta u_+(\Omega) \stackrel{\text{def}}{=} \ln[P_+(\Omega)/P_0(\Omega)]$ and $-\delta u_-(\Omega) \stackrel{\text{def}}{=} \ln[P_-(\Omega)/P_0(\Omega)]$; and from those relations it is deduced: $\bar{u}(\Omega) = \ln\left[P_0(\Omega)/\sqrt{P_+(\Omega)\cdot P_-(\Omega)}\right]$ and $\Delta u(\Omega) = \frac{1}{2}\ln[P_-(\Omega)/P_+(\Omega)]$.

Finally using Eq.7's notations one deduces:

$$\Delta\hat{\theta}(\Omega) = \int_{\Omega_{1/N}}^{\Omega} \frac{\mathrm{sh}\left(\frac{1}{2}\ln\left(\frac{P_+(\Omega)}{P_-(\Omega)}\cdot\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right)\right)}{\frac{1}{2}\frac{\omega_0^0 P_0(\Omega)}{\sqrt{\omega_+^0 P_+(\Omega)\omega_-^0 P_-(\Omega)}}+c\left(\frac{1}{2}\ln\left(\frac{P_+(\Omega)}{P_-(\Omega)}\cdot\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right)\right)} P_\Omega(\Omega)d\Omega -$$

$$\frac{\mathrm{sh}\left(\frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right)\right)}{\frac{1}{2}\frac{\omega_0^0}{\sqrt{\omega_+^0\omega_-^0}}+c\left(\frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right)\right)}\int_{\Omega_{1/N}}^{\Omega} P_\Omega(\Omega)d\Omega$$

The significance of these fields can now be addressed. The field '$\bar{u}(\Omega)$' describes local deviations from the Hardy-Weinberg law, valid for each bin or category of phenotype values. For example, if a population was under no selection and random matting occurred, then the whole population would follow the Hardy-Weinberg equilibrium law, that is, $\frac{1}{2}\omega_0^0/\sqrt{\omega_+^0\omega_-^0}\sim 1$. However, selecting a particular bin or category of phenotype values would demonstrate a local deviation of this law, given by the term $e^{\bar{u}(\Omega)} = P_0(\Omega)/\sqrt{P_+(\Omega)P_-(\Omega)}$.

The signification of '$\Delta u(\Omega)$' can be addressed using Fisher's approach.

3.5.2. Definition of fields using Fisher's theory

In his seminal paper (5), Fisher hypothesised that in a context where the population is infinite to use the normal distribution, the genetic variance 'α^2' is much smaller than the phenotype variance and that the variances of microstate distribution density functions for each gene are similar to that of the variance of the phenotype. Whilst his hypothesis can be understood

intuitively when all distribution density functions nearly overlap, it can also be demonstrated using Eq.9b. Indeed, assuming $\alpha^2 \ll \sigma^2$ implies $\sigma^2 - \alpha^2 \sim \sigma^2$ and therefore $\sigma^2 \sim \omega_+^0\sigma_+^2 + \omega_0^0\sigma_0^2 + \omega_-^0\sigma_-^2$. As $\omega_+^0 + \omega_0^0 + \omega_-^0 = 1$, posing $\sigma_+^2 \sim \sigma_0^2 \sim \sigma_-^2 \sim \sigma^2$ as Fisher did, is one valid solution.

However, the relation $\sigma^2 \sim \omega_+^0\sigma_+^2 + \omega_0^0\sigma_0^2 + \omega_-^0\sigma_-^2$, is the equation of an ellipse, and an infinite number of solutions are, in theory, possible. Those solutions will dependent on the variances of microstates (see Section 3.5.4. below, that is, the definition of fields linked to the variance of microstates).

Following Fisher's assumption, the probability of finding the microstates '+1', '0' and '-1' as a function of phenotype values are expressed as (4) (Figure 1) : $P_+(\Omega) \sim \frac{K}{\sigma}\exp\left(-\frac{1}{2}\frac{(\Omega - a_+)^2}{\sigma^2}\right)$, $P_0(\Omega) \sim \frac{K}{\sigma}\exp\left(-\frac{1}{2}\frac{(\Omega - a_0)^2}{\sigma^2}\right)$ and $P_-(\Omega) \sim \frac{K}{\sigma}\exp\left(-\frac{1}{2}\frac{(\Omega - a_-)^2}{\sigma^2}\right)$. Where 'K' is the normalisation constant identical for all microstates distribution functions since they nearly overlap. Noting, $\delta x \stackrel{\text{def}}{=} x - \langle\Omega\rangle$, the difference between the variable 'x' and the average of phenotype values to simplify the notations, as the second orders defined by $(\delta a_+/\sigma)^2$, $(\delta a_0/\sigma)^2$ and $(\delta a_-/\sigma)^2$ are neglected in Fisher's context, one deduces then:

$$\Delta u(\Omega_0) = \frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) \sim \frac{1}{2}\left(\frac{\delta a_- - \delta a_+}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right) \tag{12a}$$

$$\Delta u(\Omega_0) - s\Delta u(\Omega) = \frac{1}{2}\ln\left(\frac{P_+(\Omega)}{P_-(\Omega)} \cdot \frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) = \frac{1}{2}\left(\frac{\delta a_- - \delta a_+}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma} - \frac{\delta\Omega}{\sigma}\right) \tag{12b}$$

$$\bar{u}(\Omega) = \ln\left(\frac{P_0(\Omega)}{\sqrt{P_+(\Omega)\cdot P_-(\Omega)}}\right) \sim \frac{1}{2}\left(\frac{2\delta a_0 - \delta a_+ - \delta a_-}{\sigma}\right)\frac{\delta\Omega}{\sigma} \tag{12c}$$

In Eq.12a and Eq.12b the term $\delta a_- - \delta a_+ = a_- - a_+ = 2a$ (see Fig.1A) is known as the 'gene effect' in GWAS. In Eq.12c, the term $2\delta a_0 - \delta a_+ - \delta a_- = 2a_0 - a_+ - a_- = d$ (see Fig.1A) is the dominance as defined in the GWAS. As it turns out, Fisher considered: $d \sim 0$.

Altogether, these results demonstrate that Fisher's theory can be described by phenotypic Fields and genetic paths. As it turns out Fisher's model corresponds to the minimalist model

aforementioned (SM2 in the Supplementary Materials). Using these fields, it is also possible to determine a generic solution to Eq.8 (see SM4 in the Supplementary Materials).

3.5.3. Implication for small gene effects

Complex traits involve genes with very small effects that are difficult to characterise. The aim is to determine the resulting difference in the genetic paths in this case, that is, when $a \to 0$. Because probability density functions are used, the integration interval can be altered using the convergence property of the distributions. In this context, the conservation of genetic microstates (Eq.8) can be written using Fisher's fields as:

$$\int_0^{+\infty} \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)}{\alpha_0 e^{\frac{1}{2}\left(\frac{d}{\sigma}\right)\frac{\delta\Omega}{\sigma}}+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)} P_\Omega(\Omega)d\Omega \sim \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}{\alpha_0+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}$$

Consider a phenotype distribution of the form, $P_\Omega(\Omega) \sim \frac{K}{\sigma} e^{-\frac{1}{2}\left(\frac{\Omega-\langle\Omega\rangle}{\sigma}\right)^2}$, and rescale the phenotype values in the integral using 'a/σ' as a scaling parameter as follows:

$$\int_0^{+\infty} \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)}{\alpha_0 e^{\frac{1d}{2a}\left(\frac{\delta\Omega}{\sigma}\right)}+\mathrm{ch}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)} K e^{-\frac{1}{2}\left(\frac{\delta\Omega}{a}\right)^2} d\left(\frac{\Omega}{a}\right) \sim \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}{\alpha_0+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}$$

By taking the limit $a \to 0$, the rescaled phenotype distribution becomes a Dirac distribution, dominating any convergences; thus, the left-hand side can be transformed as:

$$\lim_{a\to 0}\int_0^{+\infty} \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)}{\alpha_0 e^{\frac{1d}{2a}\left(\frac{\delta\Omega}{\sigma}\right)}+\mathrm{ch}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)} K e^{-\frac{1}{2}\left(\frac{\delta\Omega}{a}\right)^2} d\left(\frac{\Omega}{a}\right) \sim \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}{\alpha_0+\mathrm{ch}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}\right)\right)} \sim \lim_{a\to 0}\frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}{\alpha_0+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)} \sim 0$$

Therefore, small gene effects imply: $\delta\Omega_0/\sigma \ll 1$.

Recalling that $\Delta\omega_0/\omega_0 = \mathrm{th}\left(\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)$, one also deduces $\lim_{a\to 0}\mathrm{th}\left(\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right) \sim 0$, that is, small gene effects always involve common allele frequencies, namely $\Delta\omega_0/\omega_0 \to 0$.

Using this result, the genetic paths difference can then be developed when gene effects are small and by assuming that $P_\Omega(\Omega_{1/N}) \ll 1$ and that '$P_\Omega(\Omega)$' is normally distributed one obtains at the leading order:

$$\Delta\hat{\theta}(\Omega) \sim \frac{\omega_0}{2}\left(\frac{a}{\sigma}\right)\int_{\Omega_{1/N}}^{\Omega} -\left(\frac{\delta\Omega}{\sigma}\right) P_\Omega(\Omega)d\Omega + O(a^2) \sim \omega_0 \frac{a}{2} P_\Omega(\Omega) + O(a^2) \qquad (13)$$

Eq.13 shows that in the context of Fisher's theory, a small gene effect corresponds to an overlapping symmetry between the genetic microstates and the phenotype distribution, with an amplitude proportional to the gene effect.

3.5.4. Fields linked to the variance of microstates

The involvement of variances in microstate distribution functions in genotype-phenotype associations is a highly debated matter (see (14) and references within).

As mentioned above, the expression of the difference of the genetic paths considers the distribution density function as a whole, including the role of the microstate variances. In this context, we saw that Eq.9b provides a relation between variances in the form of an ellipse. Assuming a single variance for all microstate and phenotype distributions, as Fisher did, is plausible; but other solutions exist that would not violate Eq.9b.

In this context, let us imagine that the gene effect and dominance are nulls but that the distribution density function of microstates '+1', '0', '-1' and of the phenotype have distinct variances; written, respectively, as: $P_+(\Omega) \sim \frac{K}{\sigma_+}\exp\left(-\frac{1}{2}\frac{(\delta\Omega)^2}{\sigma_+{}^2}\right)$, $P_0(\Omega) \sim \frac{K}{\sigma_0}\exp\left(-\frac{1}{2}\frac{(\delta\Omega)^2}{\sigma_0{}^2}\right)$, $P_-(\Omega) \sim \frac{K}{\sigma_-}\exp\left(-\frac{1}{2}\frac{(\delta\Omega)^2}{\sigma_-{}^2}\right)$ and $P_\Omega(\Omega) \sim \frac{K}{\sigma}\exp\left(-\frac{1}{2}\frac{(\delta\Omega)^2}{\sigma^2}\right)$.

By noting $\lambda_+ = \sigma/\sigma_+$, $\lambda_0 = \sigma/\sigma_0$ and $\lambda_- = \sigma/\sigma_-$, the fields can be mapped under the form:

$$\Delta u(\Omega_0) = \frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) = +\frac{1}{2}(\lambda_+^2 - \lambda_-^2)\left(\frac{\delta\Omega_0}{\sigma}\right)^2 + \ln\left(\frac{\lambda_+}{\lambda_-}\right) \qquad (14a)$$

$$\Delta u(\Omega_0) - s\Delta u(\Omega) = \frac{1}{2}\ln\left(\frac{P_+(\Omega)}{P_-(\Omega)}\cdot\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) = +\frac{1}{2}(\lambda_+^2 - \lambda_-^2)\left(\left(\frac{\delta\Omega_0}{\sigma}\right)^2 - \left(\frac{\delta\Omega}{\sigma}\right)^2\right) \qquad (14b)$$

$$\bar{u}(\Omega) = \ln\left(\frac{P_0(\Omega)}{\sqrt{P_+(\Omega)\cdot P_-(\Omega)}}\right) \sim -\frac{1}{2}\left(\lambda_0^2 - \frac{1}{2}\lambda_+^2 - \frac{1}{2}\lambda_-^2\right)\left(\frac{\delta\Omega}{\sigma}\right)^2 + \ln\left(\frac{\lambda_+\lambda_-}{\lambda_0}\right) \quad (14c)$$

Consequently, pseudo-gene effect and pseudo-dominance linked to the variances of genetic microstates can be defined, respectively, as:

$$a' = \frac{1}{2}(\lambda_+^2 - \lambda_-^2)$$

$$d' = \frac{1}{2}\left(\lambda_0^2 - \frac{1}{2}\lambda_+^2 - \frac{1}{2}\lambda_-^2\right)$$

To conclude, as this new method does not only concentrate on average values, it captures more information as far as genotype-phenotype associations are involved.

4. Illustration of the application of GIFT using simulated data

We intend to illustrate how GIFT can be applied using simulated data and qualitatively assess its sensitivity to extract information.

4.1. Data simulations

The codes used are provided in SM5, see the Supplementary Materials.

Data was simulated according to quantitative genetic models defined by Falconer and Mackay (1996) (21). A single bi-allelic quantitative trait locus (QTL) associated with a continuous phenotype was modelled, with an additive allele effect, a, and allele frequencies, p and q, where $p + q = 1$. The simulation parameters were set as the number of individuals sampled, $N = 1000$; number of simulation replicates, $n = 1000$; allele frequency, p; additive allele effect, a, and dominance, d.; note that the number of simulation replicates allows one to determine the best outcomes. While the theory provided in this paper is general, the simulation of data will be restricted to individuals' genotypes allocated according to Hardy-Weinberg proportions. For N individuals, Np^2 had genotype AA (corresponding to microstate -1), $2pqN$ had genotype Aa (microstate 0), and Nq^2 had genotype aa (microstate +1). The allele effect, a, is defined as

half the difference between the +1 and -1 genotype (microstate) means, and d, is the position of the 0 genotype (microstate) mean (Figure 1). Dominance is measured as the deviation of the mean of microstate 0 from the midpoint between the means of the +1 and -1 microstates. For the purposes of the simulation dominance, d, was 0, that is, the mean of microstate 0 was mid-way between the mean of microstates +1 and -1.

The additive genetic variance due to the quantitative trait loci (σ^2_{QTL}) was defined as (21): $\sigma^2_{\mathrm{QTL}} = 2\mathrm{pq}[\mathrm{a} + \mathrm{d}(\mathrm{q} - \mathrm{p})]^2$. Each individual was assigned a genotypic value, depending on their microstate: $-\mathrm{a}$ for the +1 microstate, 0 for the 0 microstate, and $+\mathrm{a}$ for the -1 microstate. Individual phenotypes were generated by adding a random environmental effect to the genotypic value of each individual. The added environmental effect was a random variate drawn from a normal distribution with a mean of 0 and variance of $1 - \sigma^2_{\mathrm{QTL}}$. The phenotype was then rescaled to a value representing a realistic dataset: Phenotype = (simulated phenotype × standard deviation of real data) + mean of real data. In this case, the real dataset modelled was a summary of the Genotype-Tissue Expression (GTEx) project (22). In particular, the phenotype was height with a mean of 68.17206 inches and standard deviation of 4.03007 inches. For each simulated replicate of N individuals, the difference between the cumulative sums of microstates ordered by phenotype value and genotypes in a randomised order with respect to phenotype was determined to create the difference in the genetic paths difference. The maximum value of this difference was identified and its position and phenotypic value in the ordered string of microstates were recorded. Where the maximum value extended over several positions, the mean position and phenotypic value were recorded. Finally, to simplify representation, the amplitudes of the genetic path differences were normalised by population size ($\mathrm{N} = 1000$ in this case).

Note that the standard deviation(s) arising from genotype-phenotype simulations were not considered in the analysis that follows. Instead, we report a theoretical analysis of the

convergence of the genetic path difference method, and its self-consistency, as well as its sensitivity to detect genotype-phenotype associations using simulations, in SM6 and SM7, respectively, see Supplementary Materials.

4.2. Analysis of simulated results:

For information, Table 1 shows how genetic variance, gene effect, i.e. 'a/σ', and allele frequency are numerically related using GWAS method. Similarly Figures 3A and 3B represent, in the context of GWAS and for the allele frequencies p=0.5 ($\Delta\omega_0 = 0$) and p=0.8 ($\Delta\omega_0 = 0.6$) that will be used below as examples, the relationship between the power of the study, the gene effect and the sample size as described in (23). The conclusion from Figure 3 is that 1000 individuals would not allow 80% power to be achieved unless the gene effect is sufficiently large, that is, i.e. $a/\sigma \geq 0.5$.

Using simulated data, we can now represent the genetic paths difference and its log transformation for $\Delta\omega_0 = 0$, either in the space of positions (Figure 4A) or of phenotype values (Figure 4B) (, for different log-scale values of the normalised gene effect 'a/σ'.

As shown in Figure 4B, the profile of the phenotype distribution density function is recovered with an amplitude that decreases as 'a/σ' decreases. The red vertical dashed line represents mean phenotype value in either the space of position or of phenotype values. Using the natural logarithm to transform '$\Delta\hat{\theta}(\Omega)$' to '$ln[\Delta\hat{\theta}(\Omega)]$' (Figure 4C) demonstrates that a difference between genetic paths can be seen relatively for small gene effects.

One may then compare how perceptible the associations are using the new method by comparing Figures 1B and 1C (method of averages) and Figures 4C for identical allele frequencies and similar gene effects. The results suggest that GIFT can be applied to 1000 individuals to return information regarding potential genotype-phenotype associations that would not otherwise be possible, or extremely difficult, with current association studies.

Concentrating on different allele frequencies given by $p = 0.8$ ($\Delta\omega_0 = 0.6$) as an example. Figures 4D, 4E and 4F are representations of '$\Delta\hat{\theta}(\Omega)$' for log-scale values of 'a/σ' using similar transformations as done for $\Delta\omega_0 = 0$.

Differences are clearly visible between $\Delta\omega_0 = 0$ and $\Delta\omega_0 = 0.6$, since the phenotype values for which '$\Delta\hat{\theta}(\Omega)$'s' are maximal have been shifted from the average value of the phenotype (indicated by the vertical red dashed line). This is not surprising because the simulation only imposed a set of genetic variances, without any constraint on the conservation of the average phenotype value.

However, the shift of the phenotype value for which '$\Delta\hat{\theta}(\Omega)$' is maximal is of interest As Eq.13 demonstrates that for small gene effects, the genetic path difference should be proportional to the phenotype distribution, that is, the phenotype value for which '$\Delta\hat{\theta}(\Omega)$' is extreme should be the average value of the phenotype.

Thus, to obtain a better visualisation of the impact of the gene effect on the positioning of the phenotype value '$\Omega(\Delta\hat{\theta}_{max})$' for which the genetic path difference is maximal, a set of simulations were also performed based on allele frequencies defined by $p \in \{0.5; 0.4; 0.3; 0.2; 0.1\}$ for log-ranging values of gene effects (Figure 5C). Note that $p \in \{0.5; 0.6; 0.7; 0.8; 0.9\}$ can be deduced from the symmetry around the average value of the phenotype.

Whilst the standard deviations obtained were not always negligible concerning '$\Omega(\Delta\hat{\theta}_{max})$', typically between 0.5 and 1 phenotypic standard deviation for small gene effects; Figure 5A demonstrated trends toward the average value of the phenotype with small gene effects. Indeed, below the simulated gene effect of $a/\sigma \sim 10^{-1}$, the average value of '$\Omega(\Delta\hat{\theta}_{max})$' was remarkably similar to that of the average value of the phenotype, marked by the horizontal black dashed line.

To confirm this trend for small gene effects, that is, $a/\sigma \leq 0.1$, we varied the population size from $N = 10^3$ to $N = 10^4$ to determine the presence of potential variations in '$\Omega(\Delta\hat{\theta}_{max})$' linked to the simulations. Results summarised in Figure 5B demonstrate that the only difference was a reduction in the standard deviations obtained for '$\Omega(\Delta\hat{\theta}_{max})$' for the simulated gene effects comprised between 0.01 and 0.1 (see arrows Figures 5A and 5B pointing to different magnitude of the standard deviations). Namely, the initial symmetry of the phenotype distribution density function reappears, as expected (Eq.13).

Finally, as Eq.13 suggests that for small gene effects, '$\Delta\hat{\theta}(\Omega)$' is proportional to the gene effect a/σ, in the form $\Delta\hat{\theta}(\Omega) \sim \omega_0 \frac{a}{\sigma} K e^{-\frac{1}{2\sigma^2}(\Omega - \langle\Omega\rangle)^2}$. It was decided to fit the natural logarithm transformation of '$\Delta\hat{\theta}(\Omega)$' in Figures 4C and 4F with quadratic curves of the form $Ln[\Delta\hat{\theta}(\Omega)] \sim A\Omega^2 + B\Omega + C$; see Tables 2 and 3).

Then, as $\Delta\hat{\theta}(\Omega) \sim e^{\left(C - \frac{B^2}{2A}\right)} e^{A \cdot \left(\Omega + \frac{B}{2A}\right)^2}$ with such a fit, we expect by identification of Eq.13 that for small gene effects: $\overline{\Omega} \sim -\frac{B}{2A}$, $-\frac{1}{2\sigma^2} \sim A$ and $\frac{a}{\sigma} \sim \frac{1}{\omega_0 K} e^{\left(C - \frac{B^2}{2A}\right)}$. Tables 2 and 3 provide the estimations for both '$\overline{\Omega}$' and 'σ' and setting $K \sim 1/\sqrt{2\pi}$, Figure 6 provides a comparison between the gene effect from the simulations, $\left(\frac{a}{\sigma}\right)_{sim}$, and the gene effect deduced from Eq.13, $\left(\frac{a}{\sigma}\right)_{theo}$.

Thus, recalling that the phenotype average and variance of the population modelled are, respectively, 68.17 inch and 16.24 inch^2; Table 2 and Table 3 demonstrate that fitting the genetic paths difference as a function of phenotype values with a quadratic curve recovers the magnitude of the average and variance of the phenotype used for the simulations for most log-scale values of the gene effect. Furthermore, the amplitude of '$\Delta\hat{\theta}(\Omega)$' is also indicative of the gene effects involved.

5. Discussion

In his seminal paper (4), Fisher provided a synthesis between the genetic inheritance of continuous traits and the Mendelian scheme of inheritance using statistics and probability. His theory has become a landmark in genetics and heredity and its conceptual framework is still used today. While statistics is a natural field to employ when dealing with large datasets, the interpretation of data as well as the inferences that can be drawn from it rely fundamentally on probability density functions.

As the act of creating categories is acknowledging a sort of imprecision, our aim was to devise a different method ruling out the need for category. This new theoretical method, inspired by physics and named GIFT, uses the concept of phenotypic field and concentrate on 'genetic path' to extract information on genotype phenotype mappings.

GIFT can be used with Fisher's assumptions and recover key concepts from quantitative genetics, including: (i) the Hardy-Weinberg coefficient locally, (ii) the Hardy-Weinberg coefficient at the population level, (iii) the gene effect, (iv) dominance, and (v) small gene effects involving common allele frequencies (24). In addition, GIFT has enabled us to define new parameters linked to microstate variances, that are, the pseudo-gene effect and the pseudo-dominance, which will probably help resolve controversies (14).

Finally, applying GIFT to simulated data based on Fisher's assumption proved its sensitivity for extracting information on genotype-phenotype associations when sample sizes and gene effects were small. The reason for not considering the dominance in the simulations is linked to the fact that realistic genome-wide association studies have shown that with small effect sizes/small gene effect (which is the main area of concern of the current paper), dominance effects are often too small, and an additive model as suggested by Fisher works well enough (25). SM8 (see Supplementary Materials) provides an initial illustration of the application of GIFT using real data based on GWAS results. In addition the Supplementary Materials contain

more information concerning the convergence and self-consistency of GIFT (SM6) and the signal-to-noise ratio when GIFT is used (SM7).

Taken as a whole, the work presented here is a first step suggesting that GIFT can be considered as a potential method for genotype-phenotype mappings. The authors agree with the fact that the simulations used to illustrate GIFT based on Fisher's seminal idea are, probably, the simplest possible and that more work needs to be done to compare GIFT to the vast literature concerning GWASes using real data. For example, at present the model is quite simplistic in the way that, by construction, it does not allow the easy incorporation of covariates. Future works will relate covariates, such as age or sex, and for the case of human populations, genetic principal components (to account for population structure). In addition, GIFT will be compared against well-known statistical tests as used in GWAS (e.g., z-test/chi-square as parametric test, or the Kolmogorov–Smirnov test as non-parametric).

6.Conclusion

A century ago, Fisher presented a statistical method to map genotypes and phenotypes, which was essentially based on the measure of uncertainty. We present here a method taking as a paradigm the fact that certainty can exist with the possibility to measure phenotype and genotype with very high precision. In an associated paper, we present a theoretical methodology based on Shannon's information enabling the significance of correlation using real genotype-phenotype data to be quantified (26). To conclude, this new method (GIFT) opens a new way to analyse genotype-phenotype mapping.

Data accessibility. Code for the simulations presented in the manuscript can be found in the Supplementary Materials (see SM5). Genomic and phenotypic data analysed as part of this study in Supplementary Materials (see SM8) is already published and freely available online.

The genomic data is part of the well-known Arabidopsis thaliana '1001 genomes' project (https://doi.org/10.1186/gb-2009-10-5-107) and can be found online here: https://1001genomes.org/data-center.html. The Phenotype data was published in The Plant Journal (DOI: 10.1111/tpj.15177) and is freely available online via Ion Explorer here: https://bitbucket.org/ADAC_UoN/dr000081-web-service-ionome- seed-and-leaf-map/. The code used for the real-data analysis in SM8 in the Supplementary Materials can be found on Dryad, DOI: https://doi.org/10.5061/dryad.vx0k6djtp.

Authors' contributions. Conceptualisation of GIFT as a new method based on physics field theory: C.R.; theoretical developments: C.R., J.W.; simulation of genotype and phenotype data based on Fisher's theory: C.R., Sa.Bl.; analysis of simulated or real data using GIFT: C.R., S.B., P.K.

Competing interests. We have no competing interests.

Funding. This work was supported by the Physics of Life Networks and The University of Nottingham.

Acknowledgements. We thank Professors Andras Paldi, James Leigh, Patricia Harris, and Darren Logan for fruitful discussions concerning the urgent need to improve genotype-phenotype association methods for small sample sizes.

References

1. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan;47(D1):D1005–12.
2. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med [Internet]. 2015 Mar 31;12(3):e1001779. Available from: https://doi.org/10.1371/journal.pmed.1001779
3. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. 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. 2013 Jun;42(3):689–700.
4. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans R Soc Edinburgh. 1919;52(2):399–433.
5. Fisher RA. XXI.—On the Dominance Ratio. Proc R Soc Edinburgh [Internet]. 2014/09/15. 1923;42:321–41. Available from: https://www.cambridge.org/core/article/xxion-the-dominance-ratio/2CDBE1D374EBE9B2774DF301BA98A584
6. Visscher PM, Goddard ME. From R.A. Fisher's 1918 Paper to GWAS a Century Later. Genetics. 2019 Apr;211(4):1125–30.
7. Hivert V, Wray NR, Visscher PM. Gene action, genetic variation, and GWAS: A user-friendly web tool. PLoS Genet. 2021 May;17(5):e1009548.
8. Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet. 2004 Apr;5(4):251–61.

9. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet [Internet]. 2009;10(10):681–90. Available from: https://doi.org/10.1038/nrg2615

10. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell [Internet]. 2017;169(7):1177–86. Available from: https://www.sciencedirect.com/science/article/pii/S0092867417306293

11. Galton F. Regression Towards Mediocrity in Hereditary Stature. J Anthropol Inst Gt Britain Irel [Internet]. 1886 Oct 22;15:246–63. Available from: http://www.jstor.org/stable/2841583

12. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2006 Mar;2(3):e41.

13. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 2003 Oct;6(5):399–408.

14. Nelson RM, Pettersson ME, Carlborg Ö. A century after Fisher: time for a new paradigm in quantitative genetics. Trends Genet. 2013 Dec;29(12):669–76.

15. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet [Internet]. 2010/06/20. 2010 Jul;42(7):565–9. Available from: https://pubmed.ncbi.nlm.nih.gov/20562875

16. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014 Nov;46(11):1173–86.

17. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and

misconceptions. Nat Rev Genet. 2008 Apr;9(4):255–66.

18. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries. bioRxiv [Internet]. 2022; Available from: https://www.biorxiv.org/content/early/2022/01/10/2022.01.07.475305

19. Stigler SM. The History of Statistics: The Measurement of Uncertainty before 1900 [Internet]. Harvard University Press; 1990. Available from: https://books.google.co.uk/books?id=nWn9DwAAQBAJ

20. Macdonald A, Hawkes LA, Corrigan DK. Recent advances in biomedical, biosensor and clinical measurement devices for use in humans and the potential application of these technologies for the study of physiology and disease in wild animals. Philos Trans R Soc London Ser B, Biol Sci. 2021 Aug;376(1831):20200228.

21. Falconer DS. Introduction to quantitative genetics. Harlow, England: Prentice Hall; 1996.

22. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet [Internet]. 2013;45(6):580–5. Available from: https://doi.org/10.1038/ng.2653

23. Delongchamp R, Faramawi MF, Feingold E, Chung D, Abouelenein S. The Association between SNPs and a Quantitative Trait: Power Calculation. Eur J Environ public Heal. 2018;2(2).

24. Park J-H, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci [Internet]. 2011;108(44):18026–31. Available from: https://www.pnas.org/content/108/44/18026

25. Sham PC, Cherny SS, Purcell S, Hewitt JK. Power of Linkage versus Association

Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data. Am J Hum Genet [Internet]. 2000;66(5):1616–30. Available from: https://www.sciencedirect.com/science/article/pii/S0002929707629929

26. Wattis J, Bray S, Kyratzi P, Rauch C. Analysis of Genotype-Phenotype Association using Fields and Information Theory. arXiv Prepr arXiv220211989. 2022;

Tables

Table 1: Genetic variances as a function of the gene effect and allele frequency

Gene effect in unit of phenottypic standard deviations (a/σ)	Allele Frequency (p)					Genetic variances
	0.1	0.2	0.3	0.4	0.5	
1	**0.18**	**0.32**	**0.42**	**0.48**	**0.5**	
0.8	**0.1152**	**0.2048**	**0.2688**	**0.3072**	**0.32**	
0.5	**0.045**	**0.08**	**0.105**	**0.12**	**0.125**	
0.3	**0.0162**	**0.0288**	**0.0378**	**0.0432**	**0.045**	
0.15	**0.00405**	**0.0072**	**0.00945**	**0.0108**	**0.01125**	
0.1	**0.0018**	**0.0032**	**0.0042**	**0.0048**	**0.005**	
0.05	**0.00045**	**0.0008**	**0.00105**	**0.0012**	**0.00125**	
0.01	**0.000018**	**0.000032**	**0.000042**	**0.000048**	**0.00005**	
0.005	**4.5E-06**	**0.000008**	**1.05E-05**	**0.000012**	**1.25E-05**	
0.001	**1.8E-07**	**3.2E-07**	**4.2E-07**	**4.8E-07**	**5E-07**	
0.0005	**4.5E-08**	**8E-08**	**1.05E-07**	**1.2E-07**	**1.25E-07**	

Table 2: Nonlinear fit results for Figure 4C using $\mathbf{Ln}\left[\Delta\widehat{\boldsymbol{\theta}}(\boldsymbol{\Omega})\right] = \mathbf{A}\boldsymbol{\Omega}^{\mathbf{2}} + \mathbf{B}\boldsymbol{\Omega} + \mathbf{C}$.

	a/σ=1 p=0.5	**a/σ=0.5 p=0.5**	**a/σ=0.05 p=0.5**	**a/σ =0.005 p=0.5**
A	-0.0371	-0.0321	-0.0311	-0.0435
B	5.0539	4.3799	4.2425	5.9241
C	-173.69	-151.48	-149.5	-208.35
$\mathbf{r^2}$	0.9960	0.9990	0.9937	0.8329
$\bar{\boldsymbol{\Omega}}$	**68.11**	**68.22**	**68.21**	**68.09**
$\boldsymbol{\sigma}^{\mathbf{2}}$	**13.48**	**15.58**	**16.08**	**11.49**

Table 3: Nonlinear fit results for Figure 4F using $\mathbf{Ln}\left[\Delta\widehat{\boldsymbol{\theta}}(\boldsymbol{\Omega})\right] = \mathbf{A}\boldsymbol{\Omega}^{\mathbf{2}} + \mathbf{B}\boldsymbol{\Omega} + \mathbf{C}$

	a/σ=1 **p=0.2**	**a/σ=0.5** **p=0.8**	**a/σ=0.05** **p=0.2**	**a/σ=0.005** **p=0.2**
A	-0.0318	-0.0292	-0.0282	-0.0300
B	4.2666	3.9583	3.8514	4.0741
C	-145.14	-136.71	-136.45	-145.40
$\mathbf{r^2}$	0.9135	0.8892	0.8004	0.7554
$\bar{\boldsymbol{\Omega}}$	**67.08**	**67.77**	**68.28**	**67.90**
$\boldsymbol{\sigma}^{\mathbf{2}}$	**15.72**	**17.12**	**17.73**	**16.67**

Figure legends

Figure 1: Phenotype human height simulated using the real data from the Genotype-Tissue Expression (GTEx) project (22). **(A)** Phenotype histograms for the three genotypes (microstates), showing the +1 state in red, 0 state in dark grey and the -1 state in blue. The overall phenotype distribution is in dark grey. The mean genotypic (microstate) values are -a, d, and +a. **(B)** Example of limitations linked to current GWAS methods when the gene effect, a, changes by one order of magnitude from a=1 to a=0.1. In this context, the three populations of genetic microstates collapse, and their separation becomes very difficult to characterise unless the size of the population is increased to reduce the width of categories. In this example, it is almost impossible to distinguish between a=0.1 and a=0.01. Note that in this case the gene effect has been normalised by the phenotypic standard deviation of the population and there is no dominance (d=0). **(C)** Corresponding distributions of the different genetic microstates singled out based on the phenotype values they characterise. The method of averages, or method of moments, advocated by Fisher suggests plotting a straight line as best fit between the three distributions out of which the slope of the straight line is then indicative of genotype-phenotype association. However, the method of averages discards the information that is available in the spreading of data for each genotype. The method we suggest will make use of this information to describe genotype-phenotype association. Note that dominance can occur ($d \neq 0$) at which point and with the method of average a linear regression of the genotype means, weighted by their frequency, on the number of alleles needs to be performed to provide a new intercept and slope (see (6)).

Figure 2: (A) Let us consider a single genome position that can only take three different genetic microstates. To represent the two configurations discussed let us consider a group of 'N' horses genotyped whose microstates (linked to different allele combinations) are color-coded. A

particular horse color corresponds to a particular genotype as indicated in the figure. In the first configuration (left arrow) the horses are sent randomly, i.e. without information related to their phenotype, in non-nominative paddocks numbered 1 to N. In the second configuration, right arrow, the same horses are sent to the same paddocks as a function of their phenotype value (height in this case). **(B)** In each configuration the genetic microstate of the horses is noted and reported to the paddock number. Given that each genetic microstate can only take either value '+1', '0' or '-1', two strings of numbers will emerge. However, as the same horses have been used for the two configurations the genetic microstates are conserved namely that the numbers of '+1', '0' and '-1' will remain the same. The only difference will concern their positioning in the string. If the genome position is related to the phenotype, then a migration of genetic microstates should ensue between the two configurations linked to the knowledge on the phenotype values. **(C)** In this context, a very simple way to represent whether the genome position is related to the phenotype is to calculate the cumulated sum of genetic microstates in either configuration. The random allocation, or absence of information on the phenotype, should always return a straight line. On the contrary, when the genome position 'interacts' with the phenotype and when the information on the phenotype is available, changes in the shape of the curve are expected due to the 'migration' of certain microstates in the string (B). One can then suggest that the cumulative sum 'responds' to the information available regarding the phenotype or, equivalently, that the phenotype acts as a field on the genetic microstates to generate the 'migration' of genetic microstates. The cumulative sum of genetic microstates is named 'genetic path' in the present work, and it is its variations that need to be quantified to determine the magnitude of the association between the phenotype and the genotype.

Figure 3: Power calculations for two allele frequencies as a function of the gene effect and sample size using Fisher's additive model when the dominance is null. The colour of the curves

is associated with the sample size (one thousand individuals for black, ten thousand individuals for red, hundred thousand individuals for blue and one million individuals for green). The figure illustrates that with 1000 individuals current genotype-phenotype methods are not powerful enough to detect small gene effects.

Figure 4: (A) Simulation of genetic paths difference in the space of positions as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0$. **(B)** Simulations of genetic paths difference in the space of phenotype values as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0$. **(C)** Natural-logarithm transformation of the genetic paths difference in the space of phenotype values as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0$. **:** **(D)** Simulation of genetic paths difference in the space of positions as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0.6$. **(B)** Simulations of genetic paths difference in the space of phenotype values as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0.6$. **(C)** Natural-logarithm transformation of the genetic paths difference in the space of phenotype values as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0.6$.

Figure 5: (A) Representation of the shift in the phenotype value for which the genetic path difference is maximal as a function of the gene effect 'a/σ' for different values of allele frequencies. N=1000 represents the number of individuals in the simulation. **(B)** Representation of the shift in the phenotype value for which the genetic paths difference is

maximal when the population is increased by a factor 10 (N=10,000) for the gene effects verifying $a/\sigma < 0.1$.

Figure 6: Comparison between simulated and theoretical gene effects the black dots correspond to the cases where the gene effects simulated are identical to the theoretical gene effects.

Figure 1

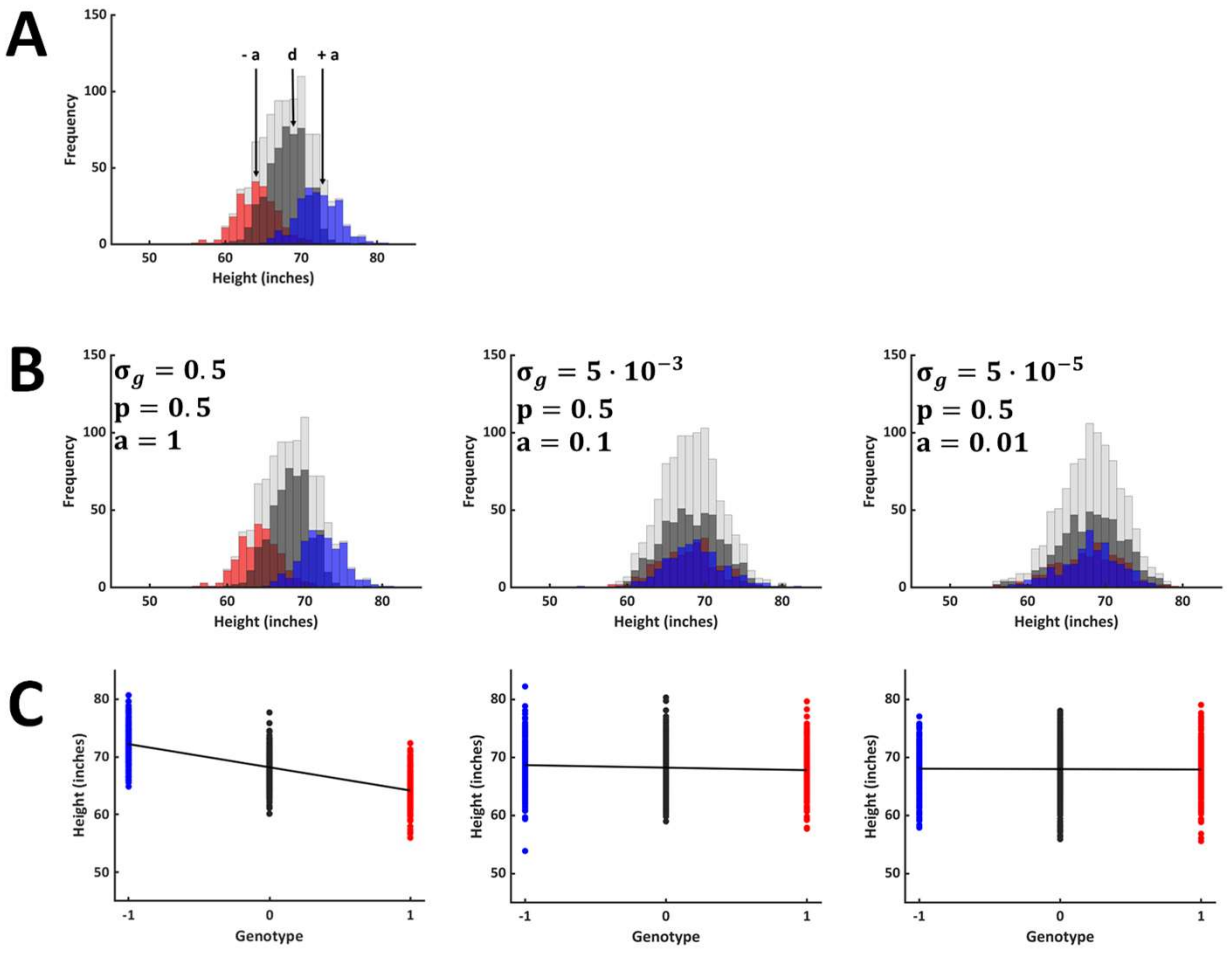
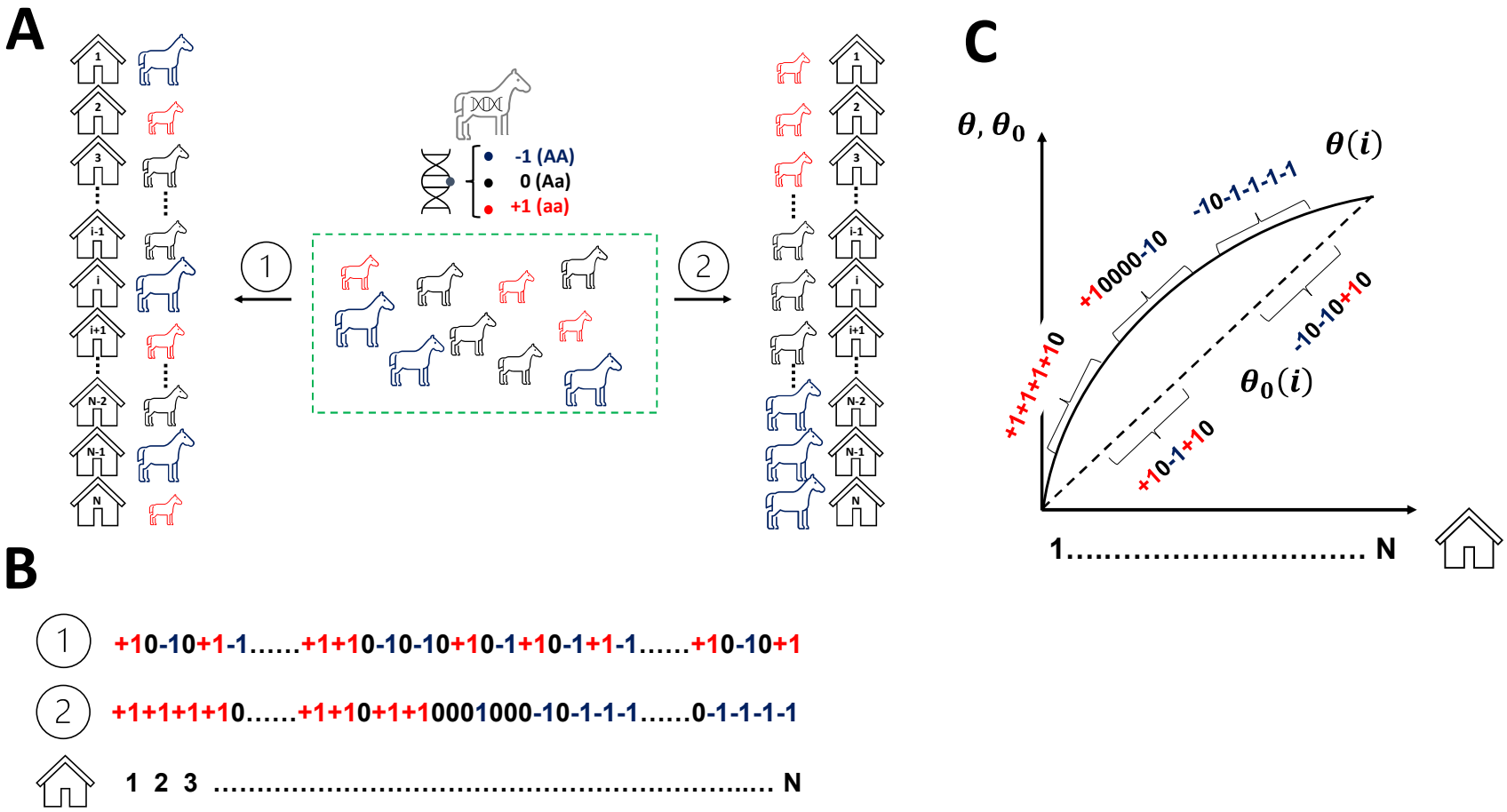

A
- a
d
+ a
Frequency
Height (inches)
B
$\sigma_g = 0.5$
$p = 0.5$
$a = 1$
$\sigma_g = 5 \cdot 10^{-3}$
$p = 0.5$
$a = 0.1$
$\sigma_g = 5 \cdot 10^{-5}$
$p = 0.5$
$a = 0.01$
C
Height (inches)
Genotype

Figure 2



B

1 +10-10+1-1.....+1+10-10-10+10-1+10-1+1-1.....+10-10+1

2 +1+1+1+10.....+1+10+1+10001000-10-1-1-1.....0-1-1-1-1

1 2 3 ..... N

Figure 3

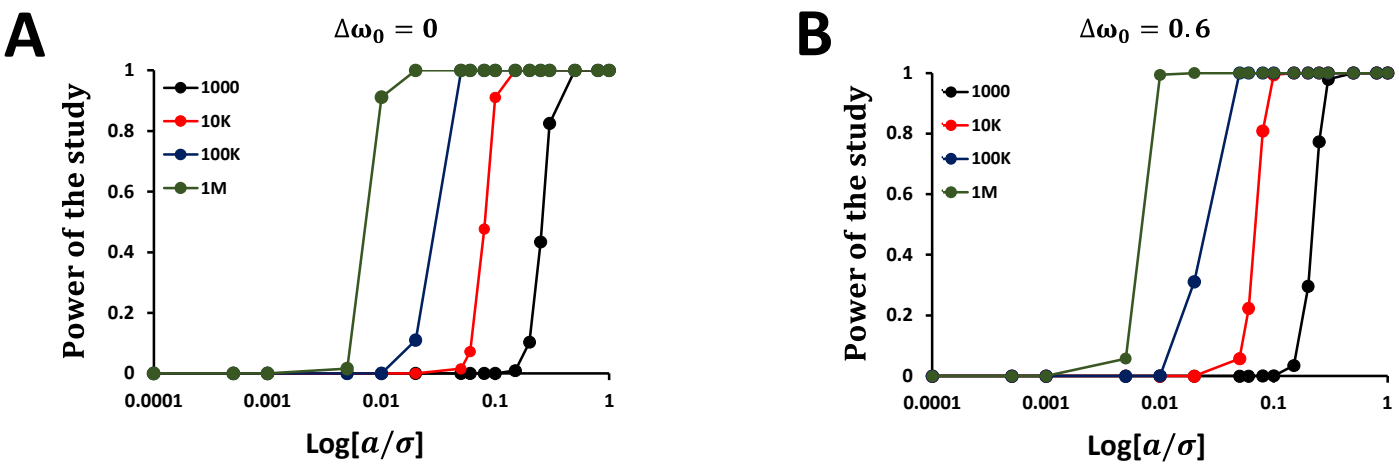

A
$\Delta\omega_0 = 0$
Power of the study
1
0.8
0.6
0.4
0.2
0
1000
10K
100K
1M
0.0001
0.001
0.01
0.1
1
Log[a/σ]
B
$\Delta\omega_0 = 0.6$
Power of the study
1
0.8
0.6
0.4
0.2
0
1000
10K
100K
1M
0.0001
0.001
0.01
0.1
1
Log[a/σ]

Figure 4

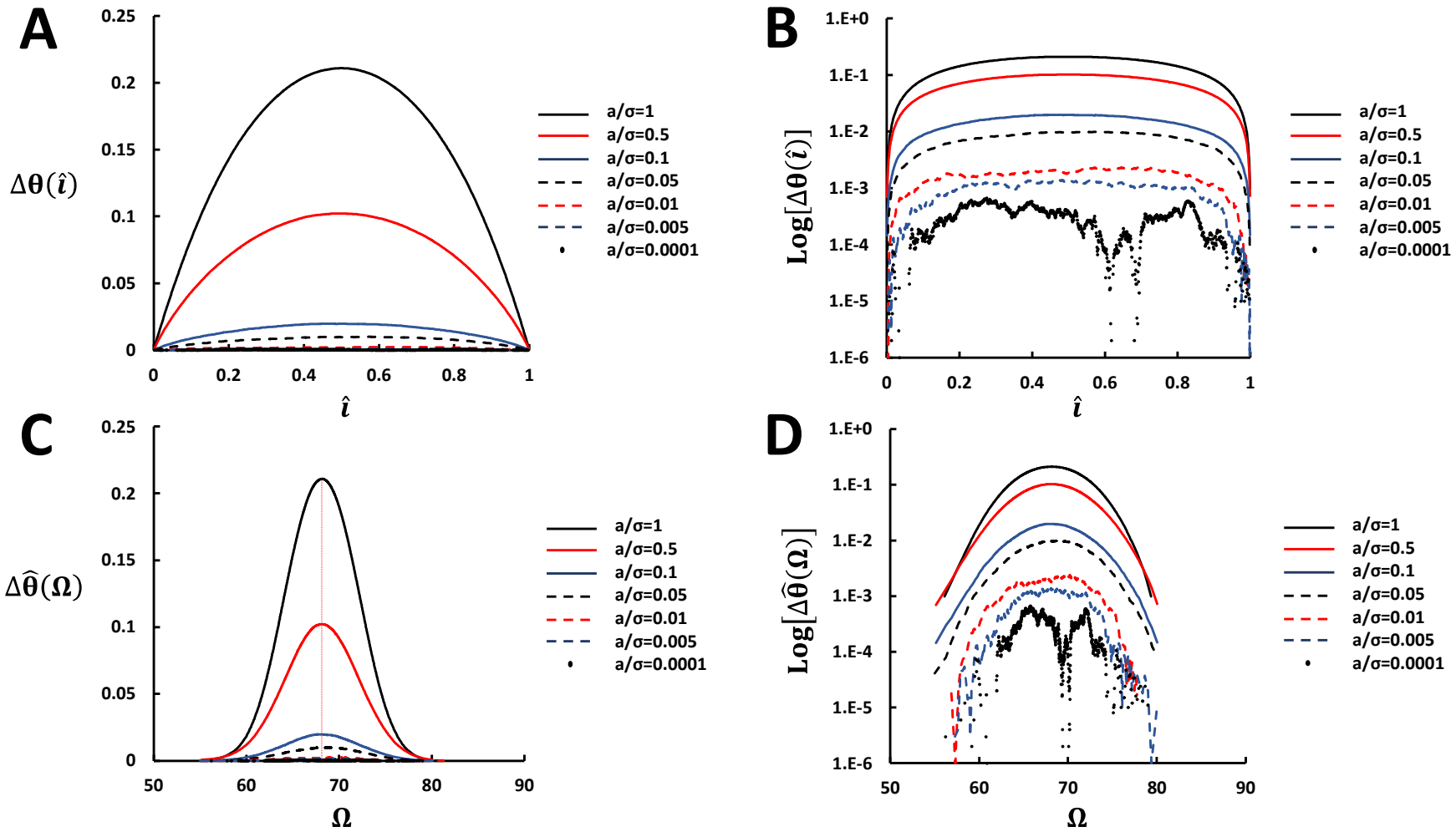

A
B
C
D
$\Delta\theta(\hat{\imath})$
$\mathrm{Log}[\Delta\theta(\hat{\imath})]$
$\Delta\hat{\theta}(\Omega)$
$\mathrm{Log}[\Delta\hat{\theta}(\Omega)]$
$\hat{\imath}$
Ω
a/σ=1
a/σ=0.5
a/σ=0.1
a/σ=0.05
a/σ=0.01
a/σ=0.005
a/σ=0.0001

Figure 5

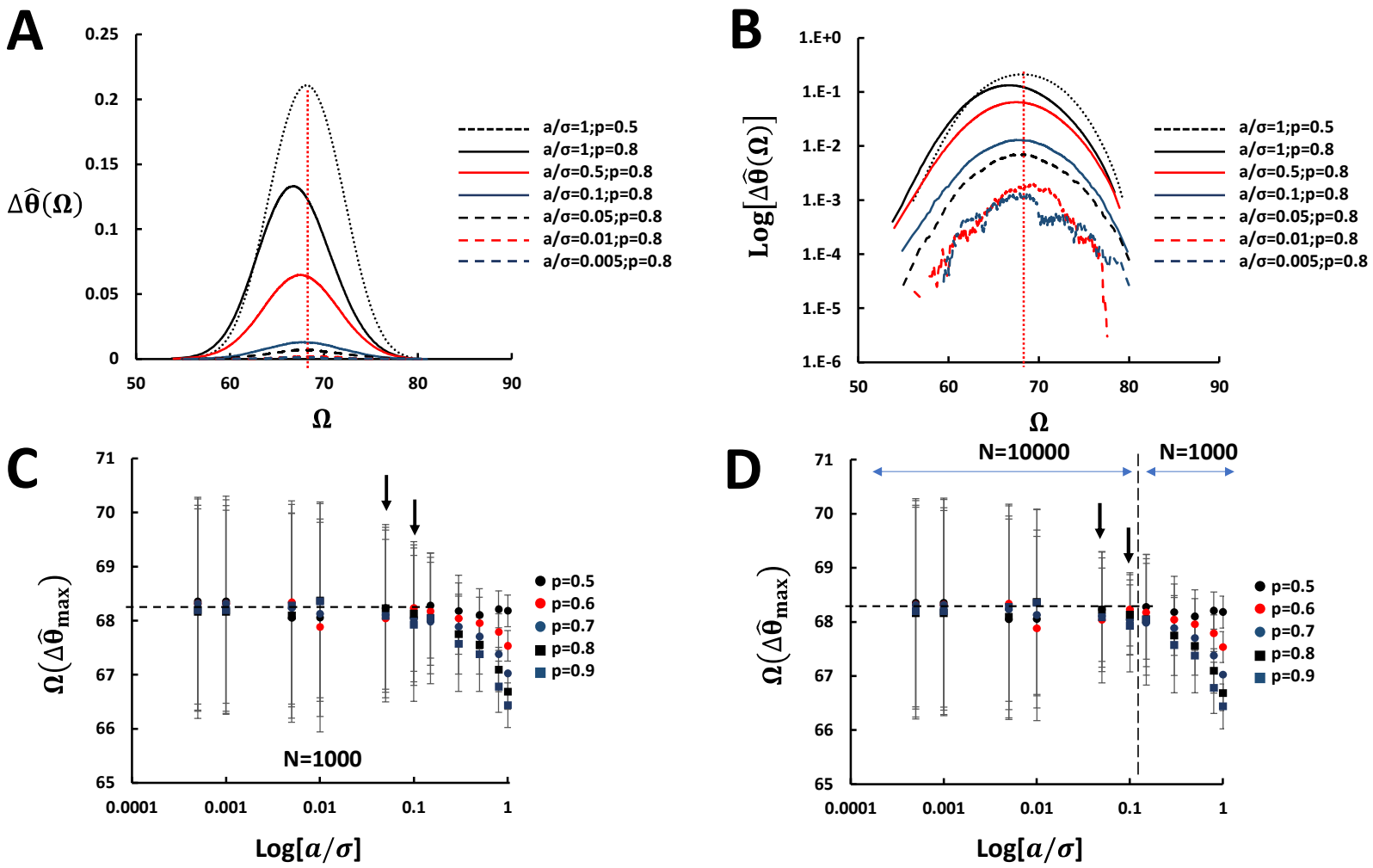
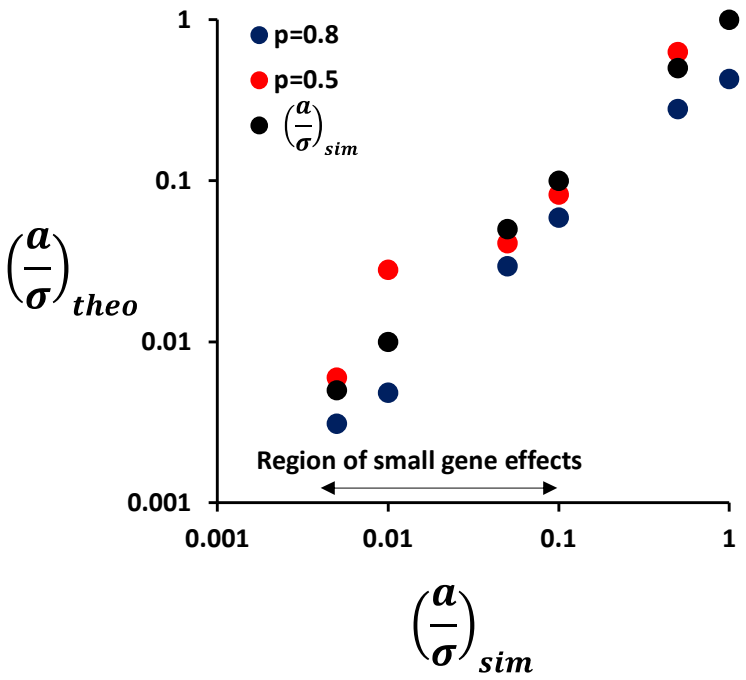

A
B
C
D
$\Delta\hat{\theta}(\Omega)$
$\mathrm{Log}[\Delta\hat{\theta}(\Omega)]$
Ω
a/σ=1;p=0.5
a/σ=1;p=0.8
a/σ=0.5;p=0.8
a/σ=0.1;p=0.8
a/σ=0.05;p=0.8
a/σ=0.01;p=0.8
a/σ=0.005;p=0.8
$\Omega(\Delta\hat{\theta}_{\max})$
p=0.5
p=0.6
p=0.7
p=0.8
p=0.9
N=1000
N=10000
$\mathrm{Log}[a/\sigma]$

Figure 6


p=0.8
p=0.5
$\left(\frac{a}{\sigma}\right)_{sim}$
1
0.1
0.01
0.001
$\left(\frac{a}{\sigma}\right)_{theo}$
Region of small gene effects
0.001
0.01
0.1
1
$\left(\frac{a}{\sigma}\right)_{sim}$