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 field 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 is then attributed to the emergence of a phenotypic field. We demonstrate that GIFT recovers GWAS, that is, Fisher's theory, when the phenotypic field is 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 the phenotypic field is 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 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) (10–16). 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 (i.e., very small gene effects) 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 (4,17) and whose measured heritability is well known

92 (4,18,19). However, this phenotype has remained controversial (12) for a long time as current
93 association methods were not been able to fully recover the heritability measured (21,22).
94 Whilst different reasons were put forward to explain this discrepancy including, for example,
95 too restricted sample sizes, too stringent statistical tests or the involvement of the
96 environment (6,23); this point seems to have been resolved only recently. Using a population
97 containing a staggering 5.3 million individuals a recent study claims to have captured nearly
98 all of the common single nucleotide polymorphisms (SNPs)-based heritability (24).
99 This important study confirms that the precision of current quantitative genetic methods to
100 determine omnigenic traits comes at an astronomical cost in line with the assumptions used,
101 namely the need for a staggeringly large (near infinite) population. In this context one may
102 wonder whether such large-scale study will ever be replicated in any other species and in
103 particular those near extinction where small sample sizes need to be considered.
104 Alternatively, one may try to understand where the need for large sample sizes comes from
105 and determine whether it is possible to extract information linking genotype to phenotype in a
106 different way.
107
108 There is a very good reason as to why large populations will always be required for
109 omnigenic phenotypes when GWAS is used. As mentioned above, the reason is rooted in the
110 fact that GWAS is mostly based on frequentist probabilities a.k.a. probability density
111 functions (PDFs). Indeed, GWAS is based on statistics and, by definition, statistics deals with
112 the measurement of uncertainties (25). To draw inferences from the comparison of large
113 datasets, a method that requires some understanding of its accuracy, including ways of
114 measuring the uncertainty in data values, is needed. In this context, statistics is the science of
115 collecting, analysing, and interpreting data, while PDFs defined through the notion of relative
116 frequencies, is central to determining the validity of statistical inferences. In practice, the use

117 of frequentist probabilities (or PDFs) and the resulting categorisation of data is justified when
118 inaccuracy exists in experimental measurements. For example, measuring a continuous
119 phenotype such as the height of individuals with a ruler with centimetre graduations, that is,
120 to the nearest centimetre, warrants the use of frequentist probability (or PDFs). In this case, a
121 frequency table of phenotypic values can be defined through 1 cm-width bins or categories,
122 from which the PDFs of the phenotype height and of the genotypes can be deduced to address
123 the statistical inferences. However, the precision available for the inferences will always be,
124 at best, given by the width of categories created and linked to the experimental precision
125 achieved (1 cm in this case). Consequently, if instead of using a ruler with centimetre
126 graduations one was using a ruler with millimetre graduations to increase the precision in
127 inferences, a larger sample size would be required such as to match the new 1 mm-width of
128 categories to reform the PDFs. The trading between the sample size and the precision
129 achieved by GWAS is known as 'study power' and its *raison d'être* is linked to the fact that
130 the entire field of probability, and therefore the PDFs, has been conceptualised
131 mathematically to represent the fact that information on a system is limited. It is for this
132 reason that the normal distribution was known before as the 'error function' or 'law of
133 errors', where the term 'error' is defined experimentally (see rulers above). Accordingly, the
134 creation of categories implies that a sort of 'imprecision' is necessary.
135 Whilst the notion of imprecision can be genuine (see rulers above), the act of creating
136 categories to use PDFs when precision in experimental measurements is available is not fully
137 justified and can be seen as an act of 'wilful ignorance'. This is so because information is lost
138 by slotting different phenotypic values into the same category. To exemplify this point let us
139 take an example, imagine a species close to extinction (very small population, say 50
140 individuals) and that it is possible to measure phenotypic values with very high precision, for
141 example, using highly advanced imaging techniques or biosensing technologies (26). In this

142 case, each measured individual could return a unique phenotypic value. Consequently,
143 reforming categories to reform and use PDFs would mean embracing the relatively large
144 width of categories created leading to the impression that a large imprecision is present, even
145 so such imprecision did not exist in the first place. One may argue that PDFs, such as the
146 normal distribution in Fisher's theory, are not required since averages and variances can be
147 mathematically calculated directly from data without the need to recreate PDFs. However,
148 this argument is not valid as averages and variances (and any other moments) cannot be
149 dissociated from PDFs, since they are the ontological parameters that define PDFs, and that
150 PDFs are used to determine statistical inferences in the field of quantitative genetics.
151 Consequently, thinking in term of averages and variances necessitate to conceptualise and
152 hold valid probability density functions, i.e., categories, to describe any system.
153
154 Therefore, there is a need to formulate new methods using the full information generated
155 through accurate and highly precise genotyping/phenotyping when sample sizes are small
156 which does not require the categorisation of data. In fact, this problem is equivalent to
157 finding a way to resolve genotype-phenotype mapping by assuming a finite-size population
158 with phenotype values measured precisely enough to rule out the possibility that two
159 phenotype values are in the same category. Taking this challenge as the starting point, a new
160 and relatively simple method for extracting information for genotype-phenotype mapping can
161 be defined. While this method is remarkably simple when explained in lay terms, its
162 theoretical framework requires the introduction of a new concept called the 'phenotypic
163 field'. Phenotypic field can also be defined within the context of Fisher's theory.
164 The remainder of this paper is organised as follows. In the first part, an intuitive approach to
165 the method GIFT is presented in which one shall see that association between datasets (i.e.,
166 genotype and phenotype) can be analysed in specific way that do not involve the use of

167 means and variances necessarily, but phenotypic field instead. This is followed by a second
168 part stating and explaining the necessary ingredients from physics (entropy, energy and field)
169 and how they must be combined, to provide GIFT. Since GIFT is not too difficult to model,
170 we have relegated the theoretical development of GIFT in the Appendices. Finally (third part)
171 one will demonstrate how Fisher's seminal theory can be re-transcribed using GIFT. In
172 particular one shall see that Fisher's seminal intuition corresponds to the simplest form of
173 GIFT. Finally (fourth part), one will compare GIFT to GWAS using simulated genotype and
174 phenotype to demonstrate that GIFT outperforms GWAS.

175 **1. Position of the problem and heuristic presentation of GIFT a method**

176 The practical issue regarding genotype-phenotype mappings with current statistical methods
177 concerns the sample size needed to provide accurate/precise information when
178 complex/omnigenic traits are involved. As stated in the introduction, this issue stems from
179 the creation of categories historically linked to the notions of 'imprecision' or 'error' in
180 measurements. At the dawn of the 21st century we are getting more precise in our
181 measurements, and one may wonder what sort of scientific/mathematical tool we should be
182 using if one were able to attain any level of precision wanted in cases where the population
183 size studied is limited.

184 We recall here that for diploid organisms, such as humans, and for a binary (bi-allelic, A or a)
185 genetic marker, any microstate (genotype) can only take three values that we shall write as
186 '+1', '0' and '-1' corresponding to genotypes aa (homozygote), Aa (heterozygote) and AA
187 (homozygote), respectively.

188 One way to proceed to develop a method embracing precision is to start by looking at how
189 density distribution functions are transformed when precision in phenotypic measurements
190 increases. From Fig.2A, the conclusion is obvious, the bar charts are transformed into code
191 bars, where each bar originates from a particular phenotype value representing one individual
192 from the population studied. This result is expected since when the width of categories
193 decreases due to an increase in precision in measurements, there will be a point where there
194 can only be one individual per category. To extract information from the code bars
195 represented by the bottom right chart in Fig.2A, let us now wonder what it means to have
196 information on the phenotype as opposed to have none.

197 To answer this question the best thing is to further simplify the problem by considering the
198 coloured bars only and not their spacing. Imagine, therefore, that a set of individuals has been
199 genotyped and that those individuals are picked at random. That is, there is no information on

any phenotype. Imagine also that one decides to concentrate, for example, on the genome position 1'000'000 on chromosome 4 for all the individuals since this genome position happens to display a biallelic single nucleotide polymorphisms (SNPs) across the set of individuals.

Thus, upon calling randomly but sequentially individuals, the genotypic information obtained in due course can be represented as a random string of genotypes including '+1', '0' and '-1' microstates (representing homozygote-AA, heterozygote-Aa and homozygote-aa). An example of such random configuration is:

[-1, +1, 0, -1, -1, +1, +1, …, -1, +1, +1, 0, -1, 0, +1, …, 0, 0, -1, +1, 0, +1, -1]

Note that the order in which the individuals were called is linked to the position in the string. Let us now repeat the same experiment using the same individuals in a context where accurate information on a chosen phenotype is available. That is, we call the individuals as a function of the magnitude of their phenotype we consider. For example, if the phenotype is height, one starts by calling the smallest individual and all subsequent individuals through successive increments in their phenotype height. Note again that because each individual has a unique phenotype value there is no possibility for two individuals to be called at once.

If the genome position 1'000'000 on chromosome 4 is involved in the formation of the phenotype, then one would expect a change in the configuration of the string of microstates based on the fact that homozygotes would be found at the extremities of the string and heterozygotes towards the middle (see Figure 2). An example of such a string would be, for example:

[+1, +1, +1, +1, 0, +1, +1, …, +1, 0, 0, 0, -1, 0, -1, …, -1, 0, -1, -1, -1, -1, -1]

Thus, the only thing that changes between the random and the phenotype-ordered configurations is the way the genetic microstates are allocated to positions in the string. However, as the genome position 1'000'000 on chromosome 4 is the only one that has been

considered, the two configurations contain the same number of '+1', '0' and '-1', since the same individuals were considered between the two configurations.

The *ansatz* is then to consider the cumulative sum of microstates as a function of the position in the string. Indeed, it is clear from the examples given above that if one starts by adding the microstates together, differences will be seen in the resulting cumulative sums. To give an example, let us consider the two strings above and note '$\theta_0(i)$' and '$\theta(i)$' the cumulative sums of microstates in the random and ordered configurations, respectively, where 'i' is the position in the string. Then adding the microstates starting from the left side of the strings one finds:

$\theta_0(1) = \mathbf{-1} = -1$ $\qquad$ $\theta(1) = \mathbf{+1} = +1$

$\theta_0(2) = \mathbf{-1} + \mathbf{1} = 0$ $\qquad$ $\theta(2) = \mathbf{+1} + \mathbf{1} = +2$

$\theta_0(3) = \mathbf{-1} + \mathbf{1} + \mathbf{0} = 0$ $\qquad$ $\theta(3) = \mathbf{+1} + \mathbf{1} + \mathbf{1} = +3$

….. $\qquad$ …..

As a result, the difference '$\theta(i) - \theta_0(i)$' is expected to be indicative of the importance of the phenotypic information and how gene microstates are related to the phenotype. The fact that the same individuals were considered in both configurations also impose a conservation relation under the form: $\theta(N) - \theta_0(N) = 0$. One shall call the cumulative sums: 'genetic paths' whose mathematical definition will be précised below. To conclude, it is the information on the phenotypic values that provides a change in the configuration of microstates and one can start developing the formulations of, $\theta_0(i)$ and $\theta(i)$.

Noting, N_{+1}, N_0 and N_{-1} the number of genetic microstates '+1', '0' and '-1', respectively. The genetic microstate frequencies for genome position 1'000'000 on chromosome 4 are defined by, $\mathrm{N}_{+1}/\mathrm{N} = \omega_+^0$, $\mathrm{N}_0/\mathrm{N} = \omega_0^0$ and $\mathrm{N}_{-1}/\mathrm{N} = \omega_-^0$.

When the positioning of the genetic microstates in the string is performed in a random fashion, 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: $\theta_0(i) = (+1 \cdot \omega_+^0 + 0 \cdot \omega_0^0 - 1 \cdot \omega_-^0)i$. Consequently, $\theta_0(i)$ is therefore a straight line. We shall call $\theta_0(i)$ the 'default genetic path'(Fig.2C).

In the second configuration the microstates are ordered. Noting $\omega_+(i)$, $\omega_0(i)$ and $\omega_-(i)$ the occurrence probabilities of the genetic microstates '+1', '0' and '-1' the 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' (Fig.2C).

As a result, the signature of a gene interacting with the phenotype when considering the two aforementioned genetic paths is the difference: $\theta(i) - \theta_0(i) = \sum_1^i \left[\left(\omega_+(j) - \omega_-(j)\right) - (\omega_+^0 - \omega_-^0)\right]$. One can then be a little bit more prescriptive by introducing the notion of phenotypic field.

2. A Physics-inspired model for GIFT: Notion of 'phenotypic field' and resulting difference between the phenotype-responding and default genetic paths

The difference, $\theta(i) - \theta_0(i)$, can be described using field theory. Indeed, as the only difference between the two configurations is the information linked to the phenotypic values, the phenotypic information can be thought as an external field impacting the configuration of microstates. To provide a physics-inspired definition of genotype-phenotype mapping, let us reconsider the random string above and assume that the set of individuals in the string are particles and that the different genetic microstates '+1', '0' and '-1' are their physical properties. One can then assume that it is those properties that interact with the field. Note that contrary to physics where a single field is defined, one needs in our case to define one field per microstate. Namely that the 'field' aforementioned is, in fact, composed of sub-fields. One shall note by $u_+(\Omega)$, $u_0(\Omega)$ and $u_-(\Omega)$ the phenotypic fields acting on the microstates '+1', '0' and '-1' respectively. Note that the variable Ω represents the phenotypic

values measured precisely. By assuming further that the particles cannot interact together and that, when they are not forced into a specific configuration by the fields, they can hop and exchange positions when the field is null (similar to a diffusion/thermal process), one can then model the string of microstates as a closed system. Fig.2C provides an idea of how the 'phenotypic field', when non null, impacts on the configuration of microstates. With those assumptions and using basic principles from statistical physics it is then possible to model the presence probability of microstates at any position in the string.

Thus, after re-expressing the genetic paths in the space of phenotypic values since the fields are function of the phenotypic values (Appendix 1), one can then construct functionals representing the entropy (Appendix 2) and the total interaction energy between the microstates and the subfields (Appendix 3). Finally, one can optimise a functional similar to the free energy to express how the fields are related to the asymmetry of states (Appendix 4). Consequently, one can demonstrate the familiar result concerning the presence probability of microstates expressed in the phenotypic space as,

$$\widehat{\omega}_{+}(\Omega) = \frac{\omega_{+}^{0} e^{-\delta u_{+}(\Omega)}}{\omega_{0}^{0} + \omega_{+}^{0} e^{-\delta u_{+}(\Omega)} + \omega_{-}^{0} e^{-\delta u_{-}(\Omega)}} \quad \text{(Eq.1)}$$

$$\widehat{\omega}_{0}(\Omega) = \frac{\omega_{0}^{0}}{\omega_{0}^{0} + \omega_{+}^{0} e^{-\delta u_{+}(\Omega)} + \omega_{-}^{0} e^{-\delta u_{-}(\Omega)}} \quad \text{(Eq.2)}$$

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

Where the hat '$\widehat{\ }$' is added to insist on the fact that the presence probabilities of microstates are expressed in the space of phenotypic values (and not positions) and, $\delta u_{+}(\Omega) \stackrel{\text{def}}{=} u_{+}(\Omega) - u_{0}(\Omega)$, $\delta u_{-}(\Omega) \stackrel{\text{def}}{=} u_{-}(\Omega) - u_{0}(\Omega)$. Eqs.1-3 are familiar to physicists when dealing with Boltzmann's weigh in statistical physics. Note that the default genetic path is defined when the fields are null. Noting, $\Delta\hat{\theta}(\Omega) \stackrel{\text{def}}{=} \hat{\theta}(\Omega) - \hat{\theta}_{0}(\Omega)$, the difference between the phenotype responding and default genetic paths expressed in the phenotypic space, $\Delta\hat{\theta}(\Omega)$ is therefore a function of the difference between Eq.1 and Eq3. One can then make the symmetries of the

problem more apparent by defining for the genetic microstates, $\Delta\omega_0 \stackrel{\text{def}}{=} \omega_+^0 - \omega_-^0$ and $\omega_0 \stackrel{\text{def}}{=} \omega_+^0 + \omega_-^0 = 1 - \omega_0^0$; and for the phenotypic fields, $2\bar{u}(\Omega) \stackrel{\text{def}}{=} \delta u_+(\Omega) + \delta u_-(\Omega) = u_+(\Omega) + u_-(\Omega) - u_0(\Omega)$ and $2\Delta u(\Omega) \stackrel{\text{def}}{=} \delta u_+(\Omega) - \delta u_-(\Omega) = u_+(\Omega) - u_-(\Omega)$. In this case, using hyperbolic functions one deduces (see Appendix 4 for development),

$$\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega) = \frac{\mathrm{sh}\big(\Delta u(\Omega_0) - \Delta u(\Omega)\big)}{\alpha_0 e^{\bar{u}(\Omega)} + \mathrm{ch}\big(\Delta u(\Omega_0) - \Delta u(\Omega)\big)} \qquad \text{(Eq.4)}$$

Where, $\mathrm{th}\big(\Delta u(\Omega_0)\big) \stackrel{\text{def}}{=} \frac{\Delta\omega_0}{\omega_0}$ and $\alpha_0 \stackrel{\text{def}}{=} \frac{1-\omega_0}{\omega_0}\mathrm{ch}\big(s\Delta u(\Omega_0)\big) = \frac{1}{2}\frac{\omega_0^0}{\sqrt{\omega_+^0\omega_-^0}}$. The new variable '$\Omega_0$' is the phenotype value corresponding to the condition $\widehat{\omega}_+(\Omega_0) \sim \widehat{\omega}_-(\Omega_0)$ and the meaning of the constant 'α_0' can be related to the Hardy-Weinberg law from population genetic. Hardy-Weinberg law based on random mating 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'. Finally, using Eq.4 one deduces the difference between the phenotype responding and default genetic paths expressed in the phenotypic space, $\Delta\hat{\theta}(\Omega)$, as (Appendix 5):

$$\Delta\hat{\theta}(\Omega) = \int_{\Omega_{1/N}}^{\Omega} \left[\frac{\mathrm{sh}\big(\Delta u(x_0) - \Delta u(x)\big)}{\alpha_0 e^{\bar{u}(x)} + \mathrm{ch}\big(\Delta u(x_0) - \Delta u(x)\big)} - \frac{\mathrm{sh}\big(\Delta u(x_0)\big)}{\alpha_0 + \mathrm{ch}\big(\Delta u(x_0)\big)}\right]\frac{1}{\Delta(x)}\,dx \qquad \text{(Eq.5)}$$

Where $\Omega_{1/N}$ is the smallest phenotypic value measured and, $\Delta(x)$, is the spacing between individuals in the code bar Fig.2A that can be related to the probability density function of the phenotype when the population measured is dense (see Appendix 1 and SM1 in the

Supplementary Materials). Finally, the conservation of genetic microstates needs to be added, that is, $\Delta\hat{\theta}(\Omega_1) = 0$, expressed as,

$$\int_{\Omega_{1/N}}^{\Omega_1} \frac{\mathrm{sh}(\Delta u(x_0)-\Delta u(x))}{\alpha_0 e^{\bar{u}(x)}+\mathrm{ch}(\Delta u(x_0)-\Delta u(x))} \frac{1}{\Delta(x)} dx = \frac{\mathrm{sh}(\Delta u(x_0))}{\alpha_0+c\ (\Delta u(x_0))} \qquad \text{(Eq.6)}$$

Where Ω_1 is the largest phenotypic value measured. 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.6 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 $u_0(\Omega)$ 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 concerning genotype-phenotype associations (see below).

Our aim is now to demonstrate that the idea of genetic paths mediated by phenotypic fields already exists in Fisher theory. This can be shown by coarse graining the paths.

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

3.1. 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 position, 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,

$$\delta N/N = (N_{+}^{0}/N)(\delta N_{+}/N_{+}^{0}) + (N_{0}^{0}/N)(\delta N_{0}/N_{0}^{0}) + (N_{-}^{0}/N)(\delta N_{-}/N_{-}^{0}) \qquad \text{(Eq.7)}$$

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

$$P_{\Omega}(\Omega) = \omega_{+}^{0} P_{+}(\Omega) + \omega_{0}^{0} P_{0}(\Omega) + \omega_{-}^{0} P_{-}(\Omega) \qquad \text{(Eq.8)}$$

From Eq.8 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.8, one deduces the conservation relations for the first two moments in the form:

$$\langle\Omega\rangle = \omega_+^0 a_+ + \omega_0^0 a_0 + \omega_-^0 a_- \qquad \text{(Eq.9a)}$$

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

$$\text{(Eq.9b)}$$

The relations provided by Eqs.9 are valid by definition, namely, whatever probability density functions are involved. Whilst Fisher never formulated a conservation similar to the second from Eq.9b, 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.4 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:

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

Note that Eq.10 corresponds to a local relative difference, in the space of phenotypic values, of microstates '+1' and '-1'. 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 $\delta N = N P_\Omega(\Omega)\delta\Omega$ together with Eq.8, we deduce using the continuum limit, $\frac{N_+ - N_-}{N} \sim \int_{\Omega_{1/N}}^{\Omega_1} \frac{\omega_+^0 P_+(\Omega) - \omega_-^0 P_-(\Omega)}{\omega_+^0 P_+(\Omega) + \omega_0^0 P_0(\Omega) + \omega_-^0 P_-(\Omega)} P_\Omega(\Omega)\delta\Omega$ or equivalently, $\frac{N_+ - N_-}{N} \sim \int_{\Omega_{1/N}}^{\Omega_1} \big(\omega_+^0 P_+(\Omega) - \omega_-^0 P_-(\Omega)\big)\delta\Omega = \omega_+^0 - \omega_-^0$.

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:

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

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.5's notations one deduces:

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

$$\frac{\mathrm{sh}\left(\frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right)\right)}{\frac{1}{2}\frac{\omega_0^0}{\sqrt{\omega_+^0\omega_-^0}}+\mathrm{ch}\left(\frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right)\right)} \int_{\Omega_{1/N}}^{\Omega} P_\Omega(\Omega)\mathrm{d}\Omega \qquad \text{(Eq.12)}$$

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 mating 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.

3.2. 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 3.4. 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:

$$\Delta u(\Omega_0) = \frac{1}{2} \ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) \sim \frac{1}{2}\left(\frac{\delta a_- - \delta a_+}{\sigma}\right)\left(\frac{\delta \Omega_0}{\sigma}\right) \tag{13a}$$

$$\Delta u(\Omega_0) - s\Delta u(\Omega) = \frac{1}{2} \ln\left(\frac{P_+(\Omega)}{P_-(\Omega)} \cdot \frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) = \frac{1}{2}\left(\frac{\delta a_- - \delta a_+}{\sigma}\right)\left(\frac{\delta \Omega_0}{\sigma} - \frac{\delta \Omega}{\sigma}\right) \tag{13b}$$

$$\bar{u}(\Omega) = \ln\left(\frac{P_0(\Omega)}{\sqrt{P_+(\Omega) \cdot P_-(\Omega)}}\right) \sim \frac{1}{2}\left(\frac{2\delta\ _0 - \delta a_+ - \delta a_-}{\sigma}\right)\frac{\delta \Omega}{\sigma} \tag{13c}$$

In Eq.13a and Eq.13b 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).

3.3. 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 the gene effect, $a = (a_- - a_+)/2$ (see definition above), tends towards zero: $a \to 0$. 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.6) can be written using Fisher's fields as,

$$\int_0^{+\infty} \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)}{\alpha_0 e^{\frac{1}{2}\left(\frac{d}{\sigma}\right)\frac{\delta\Omega}{\sigma}}+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)} P_\Omega(\Omega)d\Omega \sim \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}{\alpha_0+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)} \qquad \text{(Eq.14)}$$

Consider, as Fisher did in his seminal paper (4), 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:

$$\int_0^{+\infty} \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)}{\alpha_0 e^{\frac{1d}{2a}\left(\frac{\delta\Omega}{\sigma}\right)}+\mathrm{ch}\left(\frac{1}{2}\left(\frac{\delta\Omega_0}{\sigma}-\frac{\delta\Omega}{\sigma}\right)\right)} K e^{-\frac{1}{2}\left(\frac{\delta\Omega}{a}\right)^2} d\left(\frac{\Omega}{a}\right) \sim \frac{\mathrm{sh}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)}{\alpha_0+\mathrm{ch}\left(\frac{1}{2}\left(\frac{a}{\sigma}\right)\left(\frac{\delta\Omega_0}{\sigma}\right)\right)} \qquad \text{(Eq.15)}$$

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:

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

(Eq.16)

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:

$$\Delta\hat{\theta}(\Omega)\sim\frac{\omega_0}{2}\left(\frac{a}{\sigma}\right)\int_{\Omega_{1/N}}^{\Omega}-\left(\frac{\delta\Omega}{\sigma}\right)P_\Omega(\Omega)d\Omega+O(a^2)\sim\omega_0\frac{a}{2}P_\Omega(\Omega)+O(a^2) \qquad \text{(Eq.17)}$$

Eq.17 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.

3.4. 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 (20) 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:

$$\Delta u(\Omega_0) = \frac{1}{2}\ln\left(\frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) = +\frac{1}{2}(\lambda_+^2 - \lambda_-^2)\left(\frac{\delta\Omega_0}{\sigma}\right)^2 + \ln\left(\frac{\lambda_+}{\lambda_-}\right) \quad (18a)$$

$$\Delta u(\Omega_0) - s\Delta u(\Omega) = \frac{1}{2}\ln\left(\frac{P_+(\Omega)}{P_-(\Omega)} \cdot \frac{P_-(\Omega_0)}{P_+(\Omega_0)}\right) = +\frac{1}{2}(\lambda_+^2 - \lambda_-^2)\left(\left(\frac{\delta\Omega_0}{\sigma}\right)^2 - \left(\frac{\delta\Omega}{\sigma}\right)^2\right) \quad (18b)$$

$$\bar{u}(\Omega) = \ln\left(\frac{P_0(\Omega)}{\sqrt{P_+(\Omega)\cdot P_-(\Omega)}}\right) \sim -\frac{1}{2}\left(\lambda_0^2 - \frac{1}{2}\lambda_+^2 - \frac{1}{2}\lambda_-^2\right)\left(\frac{\delta\Omega}{\sigma}\right)^2 + \ln\left(\frac{\lambda_+\lambda_-}{\lambda_0}\right) \quad (18c)$$

Consequently, pseudo-gene effect and pseudo-dominance linked to the variances of genetic microstates can be defined, respectively, as: $a' = \frac{1}{2}(\lambda_+^2 - \lambda_-^2)$ and $d' = \frac{1}{2}\left(\lambda_0^2 - \frac{1}{2}\lambda_+^2 - \frac{1}{2}\lambda_-^2\right)$.

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) (27). 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 (27): $\sigma^2_{\mathrm{QTL}} = 2pq[a + d(q - p)]^2$. Each individual was assigned a genotypic value, depending on their microstate: $-a$ for the +1 microstate, 0 for the 0 microstate, and $+a$ 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 (28). 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 ($N = 1000$ 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 (29). Briefly, the power of a study is related to the concepts of Type I and Type II errors. A Type I Error (a.k.a. Alpha) is rejecting the null hypothesis in favor of a false alternative hypothesis, and a Type II Error (a.k.a. Beta) is failing to reject a false null hypothesis in favor of a true alternative hypothesis. The power of a study is then the probability of avoiding a Type II error. Mathematically, the power is defined by, 1 – Beta, where 0<Beta<1. If the power is close to 1, i.e., Beta~0, the hypothesis test is very good at detecting a false null hypothesis. Beta is commonly set at 0.2, to provide a power ~0.8 (or 80%). Powers lower than 0.8, while not impossible, would typically be considered too low for GWAS. The four primary factors affecting power are, the sample size, the significance level (or Alpha), the variance/variability in the measured response and the magnitude of the effect of the variable. Only the first variable can be altered in a study since all the others are fixed by the genes. To conclude, power is increased when

552 the sample size or effect sizes (gene effect) are increased. Accordingly, Figure 3
553 demonstrates that 1000 individuals would not allow 80% power to be achieved unless the
554 gene effect is sufficiently large, that is for $a/\sigma \geq 0.5$.

555 Using simulated data, we can now represent the genetic paths difference and its log
556 transformation for $\Delta\omega_0 = 0$, either in the space of positions (Figure 4A) or of phenotype
557 values (Figure 4B) for different log-scale values of the normalised gene effect a/σ.
558 As shown in Figure 4B, the profile of the phenotype distribution density function is recovered
559 with an amplitude that decreases as a/σ decreases. The red vertical dashed line in Figure 4D
560 represents the mean phenotypic value. Using the natural logarithm to transform $\Delta\hat{\theta}(\Omega)$
561 (Figure 4C) to $ln[\Delta\hat{\theta}(\Omega)]$ (Figure 4D) demonstrates that a difference between genetic paths
562 can be seen for small gene effects.
563 One may then compare how perceptible the associations are using the new method by
564 comparing Figures 1B and 1C (method of averages) and Figures 4D for identical allele
565 frequencies and similar gene effects. Recall that GWAS rely on determining difference in
566 averages (see Figure 1B or Figure 1C). However, the determination of a difference in the
567 microstate averages rely on a strong gene effect (Figure 1B) or a very large population (see
568 Figure 3) as otherwise the density functions of microstates collapse onto one. This is
569 particularly visible when one compares the right-hand and left-hand graphs in figure 1B or
570 Figure 1C. Thus, the results provided by Figure 4 suggest that GIFT can be applied to 1000
571 individuals to return information regarding potential genotype-phenotype associations that
572 would not otherwise be possible, or extremely difficult, with current association studies.
573 Concentrating on different allele frequencies given by $p = 0.8$ ($\Delta\omega_0 = 0.6$) as an example.
574 Figures 5A, and 5B are representations of $\Delta\hat{\theta}(\Omega)$ for log-scale values of a/σ using similar
575 transformations as done for $\Delta\omega_0 = 0$.

576 Differences are clearly visible between $\Delta\omega_0 = 0$ and $\Delta\omega_0 = 0.6$, since the phenotype values
577 for which $\Delta\hat{\theta}(\Omega)$'s are maximal have been shifted from the average value of the phenotype
578 (indicated by the vertical red dashed line). This is not surprising because the simulation only
579 imposed a set of genetic variances, without any constraint on the conservation of the average
580 phenotype value.
581 However, the shift of the phenotype value for which $\Delta\hat{\theta}(\Omega)$ is maximal is of interest. As
582 Eq.17 demonstrates that for small gene effects, the genetic path difference should be
583 proportional to the phenotype distribution, that is, the phenotype value for which $\Delta\hat{\theta}(\Omega)$ is
584 extreme should be the average value of the phenotype.
585 Thus, to obtain a better visualisation of the impact of the gene effect on the positioning of the
586 phenotype value $\Omega(\Delta\hat{\theta}_{\max})$ for which the genetic path difference is maximal, a set of
587 simulations were also performed based on allele frequencies defined by $p \in$
588 $\{0.5; 0.4; 0.3; 0.2; 0.1\}$ for log-ranging values of gene effects (Figure 5C). Note that $p \in$
589 $\{0.5; 0.6; 0.7; 0.8; 0.9\}$ can be deduced from the symmetry around the average value of the
590 phenotype.
591 Whilst the standard deviations obtained were not always negligible concerning $\Omega(\Delta\hat{\theta}_{\max})$,
592 typically between 0.5 and 1 phenotypic standard deviation for small gene effects; Figure 5A
593 demonstrated trends toward the average value of the phenotype with small gene effects.
594 Indeed, below the simulated gene effect of $a/\sigma \sim 10^{-1}$, the average value of $\Omega(\Delta\hat{\theta}_{\max})$ was
595 remarkably similar to that of the average value of the phenotype, marked by the horizontal
596 black dashed line.
597 To confirm this trend for small gene effects, that is, $a/\sigma \leq 0.1$, we varied the population size
598 from $N = 10^3$ to $N = 10^4$ to determine the presence of potential variations in $\Omega(\Delta\hat{\theta}_{\max})$
599 linked to the simulations. Results summarised in Figure 5B demonstrate that the only
600 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 5C and 5D pointing to different magnitude of the standard deviations). Namely, the initial symmetry of the phenotype distribution density function reappears, as expected (Eq.17).

Finally, Eq.17 suggests that for small gene effects $\Delta\hat{\theta}(\Omega)$ is proportional to the gene effect a/σ in the form, $\Delta\hat{\theta}(\Omega)\sim\omega_0\frac{a}{\sigma}Ke^{-\frac{1}{2\sigma^2}(\Omega-\langle\Omega\rangle)^2}$. As a consequence it was decided to fit the all the curves $\Delta\hat{\theta}(\Omega)$ in Figures 4D and 5B with quadratic equations of the form $\mathrm{Ln}\left[\Delta\hat{\theta}(\Omega)\right]\sim A\Omega^2+B\Omega+C$; 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.17 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 $K\sim1/\sqrt{2\pi}$, 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.17, $\left(\frac{a}{\sigma}\right)_{theo}$.

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,

624 the interpretation of data as well as the inferences that can be drawn from it rely
625 fundamentally on probability density functions. As the act of creating categories to work with
626 density function is acknowledging a sort of imprecision, our aim was to devise a different
627 method ruling out the need for category. This new theoretical method, inspired by physics
628 and named GIFT, uses the concept of phenotypic field, and concentrate on 'genetic path' to
629 extract information on genotype phenotype mappings. It is then important to discuss the
630 conceptual similarities and differences between GWAS and GIFT.
631 In term of conceptual similarities, we saw that the theory underscoring GIFT developed using
632 Fisher's assumptions recovers key concepts from quantitative genetics, including: (i) the
633 Hardy-Weinberg coefficient locally, (ii) the Hardy-Weinberg coefficient at the population
634 level, (iii) the gene effect, (iv) dominance, and (v) small gene effects involving common
635 allele frequencies (30). In this context, GIFT and GWAS are similar. Finally, applying GIFT
636 to simulated data based on Fisher's assumption proved its sensitivity for extracting
637 information on genotype-phenotype associations when sample sizes and gene effects were
638 small. The reason for not considering the dominance in the simulations is linked to the fact
639 that realistic genome-wide association studies have shown that with small effect sizes/small
640 gene effect (which is the main area of concern of the current paper), dominance effects are
641 often too small, and an additive model as suggested by Fisher works well enough (31).
642 In term of conceptual differences, three essential points can be discussed.
643 First of all, GIFT is more general that GWAS in the sense that the phenotypic fields can be
644 any, namely do not have to be linear. The prescription of a linear phenotypic field in Fisher's
645 context comes from the symmetry associated with using the normal distribution as a template
646 for any distribution density function, together with the assumption that the phenotypic and
647 microstate variances are identical (4). When the constraint on the variances is released, the
648 phenotypic field becomes quadratic involving the variances. In this context 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 (20).

Secondly, in term of genetics what has been achieved so far is rather at odds with traditional ways of thinking about the notion of gene. Indeed, by defining the difference in genetic paths, $\Delta\theta(\Omega)$, one can say that it is the phenotype, i.e., phenotypic field or information, that organises the configuration of genotypes and not the converse. In genetics the tradition is to think of genes as causing phenotypes. Here, a different way of thinking is suggested since it is the variation in phenotype values, resulting in our ability to generate a ranking process, that interacts with the microstates. Therefore, the phenotype is able to 'select' a set of genetic microstates. Recall that microstates 'respond' to, or interact with, the phenotypic field only if they are associated with the phenotype. Consequently, this model suggests considering a genotype-phenotype 'loop', a.k.a. self-consistency. That is to say that if genes cause phenotypes (traditional view) and that phenotypes select gene microstates (present view), then an equivalence exists between phenotype and genotype. Supplementary Materials contain more information and development concerning the convergence and self-consistency of GIFT (SM6).

Finally, Fisher's theory / GWAS has been built on considering the normal distribution. In general, 'real' density functions never come as normally distributed. Given that Fisher's theory gives biological meanings to average and variance only, to define the 'gene effect' and 'genetic/phenotypic variance' linked to heredity, respectively; there is no biological meaning to any other statistical/mathematical parameters describing real density functions, such as for example the 'skewness'. As GIFT uses curves, namely does not use average and variance as central parameters, this issue does not exist with GIFT. Said differently, GIFT frees GWAS from any preconceived idea of what statistics and probability applied to biology should be.

673 Taken as a whole, the work presented here is a first step suggesting that GIFT can be
674 considered as a potential method for genotype-phenotype mappings. Supplementary Material
675 SM7 contains more information concerning the signal-to-noise ratio when GIFT is used and
676 SM8 (see Supplementary Materials) provides an initial illustration of the application of GIFT
677 using real data based on GWAS results.
678 However the authors agree with the fact that more work needs to be done to compare GIFT to
679 the vast literature concerning GWASes. For example, at present the model is quite simplistic
680 in the way that, by construction, it does not allow the easy incorporation of covariates. Future
681 works will relate covariates, such as age or sex, and for the case of human populations,
682 genetic principal components (to account for population structure). In addition, GIFT will be
683 compared against well-known statistical tests as used in GWAS (e.g., z-test/chi-square as
684 parametric test, or the Kolmogorov–Smirnov test as non-parametric).
685

686 **6.Conclusion**

687 A century ago, Fisher devised a statistical method to map genotypes and phenotypes, which
688 was essentially based on the measure of uncertainty. We present here a method taking as a
689 paradigm the fact that certainty can exist with the possibility to measure phenotype and
690 genotype with very high precision. In an associated paper, we present a theoretical
691 methodology based on Shannon's information enabling the significance of correlation using
692 real genotype-phenotype data to be quantified (32). To conclude, this new method (GIFT)
693 opens a new way to analyse genotype-phenotype mapping.

694

695 **Data accessibility.** Code for the simulations presented in the manuscript can be found in the
696 Supplementary Materials (see SM5). Genomic and phenotypic data analysed as part of this
697 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.

Appendices

Appendix 1: Differential expression of the genetic path in the space of phenotypic 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 $\Omega_{\hat{\imath}}$, as the new position and its corresponding phenotype value, respectively. 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, one deduces also, $\theta(\hat{\imath}) \sim \int_{\frac{1}{N}}^{\hat{\imath}} \left(\omega_+(\hat{\jmath}) - \omega_-(\hat{\jmath})\right) d\hat{\jmath}$.

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: $\hat{\theta}(\Omega_{\hat{\imath}}) = \sum_{\hat{\jmath}=1/N}^{\hat{\jmath}=\hat{\imath}} \left(\widehat{\omega}_+(\Omega_{\hat{\jmath}}) - \widehat{\omega}_-(\Omega_{\hat{\jmath}})\right)$ where the hat '$\widehat{\ }$' is added to describe new functions linked to phenotypic values. Accordingly, by definition, $\hat{\theta}(\Omega_{\hat{\imath}}) - \hat{\theta}(\Omega_{\hat{\imath}-1}) = \widehat{\omega}_+(\Omega_{\hat{\imath}}) - \widehat{\omega}_-(\Omega_{\hat{\imath}})$. 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: $\frac{d\hat{\theta}}{d\Omega} \sim \left(\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega)\right) \frac{1}{\Delta(\Omega)}$.

As a result, the genetic path can be expressed in the space of the phenotype values as:

$$\hat{\theta}(\Omega) \sim \int_{\Omega_{1/N}}^{\Omega} \left(\widehat{\omega}_+(x) - \widehat{\omega}_-(x)\right) \frac{1}{\Delta(x)} dx \qquad \text{(Eq.A1-1)}$$

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.

Appendix 2: Entropy of the string of microstates

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,

$$S_0 \sim -N \int_{\Omega_{1/N}}^{\Omega_1} \left(\omega_+^0 \ln(\omega_+^0) + \omega_0^0 \ln(\omega_0^0) + \omega_-^0 \ln(\omega_-^0)\right) \frac{1}{\Delta(\Omega)} d\Omega \qquad \text{(Eq.A2-1)}$$

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:

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

(Eq.A2-2)

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 the difference, $S - S_0$, 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.

Appendix 3: Interaction energy between microstates and subfields

As the difference, $S - S_0$, is linked to the information gained from knowing phenotypic values and ranking them (Appendix 2), 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 under the form:

$$E \sim N \int_{\Omega_{1/N}}^{\Omega_1} \left(u_+(\Omega) \widehat{\omega}_+(\Omega) + u_0(\Omega) \widehat{\omega}_0(\Omega) + u_-(\Omega) \widehat{\omega}_-(\Omega) \right) \frac{1}{\Delta(\Omega)} d\Omega \qquad \text{(Eq.A3-1)}$$

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: $\Delta S - E = 0$.

Appendix 4: Optimisation of $\Delta S - E$

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 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,

$$\Delta S/N - E/N - \alpha_+\left(\omega_+^0 - \int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_+(\Omega)\frac{1}{\Delta(\Omega)}d\Omega\right) - \alpha_0\left(\omega_0^0 - \int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_0(\Omega)\frac{1}{\Delta(\Omega)}d\Omega\right) - \alpha_-\left(\omega_-^0 - \int_{\Omega_{1/N}}^{\Omega_1} \widehat{\omega}_-(\Omega)\frac{1}{\Delta(\Omega)}d\Omega\right) = 0 \qquad \text{(Eq.A4-1)}$$

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:

$$\delta\widehat{\omega}_+(\Omega)\left[\ln\left(\frac{\widehat{\omega}_+(\Omega)}{1-\widehat{\omega}_+(\Omega)-\widehat{\omega}_-(\Omega)}\right) - \delta u_+(\Omega) + (\alpha_+ - \alpha_0)\right] = 0 \qquad \text{(Eq.A4-2)}$$

$$\delta\widehat{\omega}_-(\Omega)\left[\ln\left(\frac{\widehat{\omega}_-(\Omega)}{1-\widehat{\omega}_+(\Omega)-\widehat{\omega}_-(\Omega)}\right) - \delta u_-(\Omega) + (\alpha_- - \alpha_0)\right] = 0 \qquad \text{(Eq.A4-3)}$$

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 the fields are null, one obtains,

$$\widehat{\omega}_+(\Omega) = \frac{\omega_+^0 e^{-\delta u_+(\Omega)}}{\omega_0^0 + \omega_+^0 e^{-\delta u_+(\Omega)} + \omega_-^0 e^{-\delta u_-(\Omega)}} \qquad \text{(Eq.A4-4)}$$

$$\widehat{\omega}_-(\Omega) = \frac{\omega_-^0 e^{-\delta u_-(\Omega)}}{\omega_0^0 + \omega_+^0 e^{-\delta u_+(\Omega)} + \omega_-^0 e^{-\delta u_-(\Omega)}} \qquad \text{(Eq.A4-5)}$$

To make the asymmetries of the problem more apparent, the following are defined for genetic microstates: $\Delta\omega_0 \stackrel{\text{def}}{=} \omega_+^0 - \omega_-^0$ and $\omega_0 \stackrel{\text{def}}{=} \omega_+^0 + \omega_-^0 = 1 - \omega_0^0$; and for the phenotypic fields:

$2\bar{u}(\Omega) \stackrel{\text{def}}{=} \delta u_+(\Omega) + \delta u_-(\Omega) = u_+(\Omega) + u_-(\Omega) - u_0(\Omega)$ and $2\Delta u(\Omega) \stackrel{\text{def}}{=} \delta u_+(\Omega) - \delta u_-(\Omega) = u_+(\Omega) - u_-(\Omega)$. Then, the difference and sum of $\widehat{\omega}_+(\Omega)$ and $\widehat{\omega}_-(\Omega)$ can be rewritten as follow:

$$\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega) = \frac{\omega_0 e^{-\bar{u}(\Omega)}\left[\frac{\Delta\omega_0}{\omega_0}\mathrm{ch}\left(\Delta u(\Omega)\right) - \mathrm{sh}\left(\Delta u(\Omega)\right)\right]}{1-\omega_0+\omega_0 e^{-\bar{u}(\Omega)}\left[\mathrm{ch}\left(\Delta u(\Omega)\right) - \frac{\Delta\omega_0}{\omega_0}\mathrm{sh}\left(\Delta u(\Omega)\right)\right]} \quad \text{(Eq.A4-6)}$$

$$\widehat{\omega}_+(\Omega) + \widehat{\omega}_-(\Omega) = \frac{\omega_0 e^{-\bar{u}(\Omega)}\left[\mathrm{ch}\left(\Delta u(\Omega)\right) - \frac{\Delta\omega_0}{\omega_0}\mathrm{sh}\left(\Delta u(\Omega)\right)\right]}{1-\omega_0+\omega_0 e^{-\bar{u}(\Omega)}\left[\mathrm{ch}\left(\Delta u(\Omega)\right) - \frac{\Delta\omega_0}{\omega_0}\mathrm{sh}\left(\Delta u(\Omega)\right)\right]} \quad \text{(Eq.A4-7)}$$

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:

$$\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega) = \frac{\mathrm{sh}\left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)}{\alpha_0 e^{\bar{u}(\Omega)} + c\ \left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)} \quad \text{(Eq.A4-8)}$$

$$\widehat{\omega}_+(\Omega) + \widehat{\omega}_-(\Omega) = \frac{\mathrm{ch}\left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)}{\alpha_0 e^{\bar{u}(\Omega)} + \mathrm{ch}\left(\Delta u(\Omega_0) - \Delta u(\Omega)\right)} \quad \text{(Eq.A4-9)}$$

$$\alpha_0 \stackrel{\text{def}}{=} \frac{1-\omega_0}{\omega_0}\mathrm{ch}\left(s\Delta u(\Omega_0)\right) = \frac{1-\omega_0}{\omega_0}\frac{1}{\sqrt{1-\left(\mathrm{th}\left(s\Delta u(\Omega_0)\right)\right)^2}} = \frac{1-\omega_0}{\sqrt{{\omega_0}^2 - {\Delta\omega_0}^2}} = \frac{1}{2}\frac{\omega_0^0}{\sqrt{\omega_+^0\omega_-^0}} \quad \text{(Eq.A4-10)}$$

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 mating 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:

Appendix 5: Expression of the difference between the phenotype responding and default genetic paths expressed in the phenotypic space and conservation of genetic microstate frequencies.

The phenotype-associated genetic path is simply the integration of (Eq.A4-8) over the phenotype values that is given, as seen above (Eq.A1-1 in Appendix 1), by: $\int_{\Omega_{1/N}}^{\Omega}[\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega)]\frac{1}{\Delta(\Omega)}\mathrm{d}\Omega$. 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.: $\int_{\Omega_{1/N}}^{\Omega}\Delta\omega_0\frac{1}{\Delta(\Omega)}\mathrm{d}\Omega$. 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:

$$\Delta\hat{\theta}(\Omega) = \int_{\Omega_{1/N}}^{\Omega}\frac{\mathrm{sh}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}{\alpha_0 e^{\bar{u}(\Omega)}+\mathrm{ch}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}\frac{1}{\Delta(\Omega)}\mathrm{d}\Omega - \frac{\mathrm{sh}\big(\Delta u(\Omega_0)\big)}{\alpha_0+\mathrm{ch}\big(\Delta u(\Omega_0)\big)}\int_{\Omega_{1/N}}^{\Omega}\frac{1}{\Delta(\Omega)}\mathrm{d}\Omega \qquad \text{(Eq.A5-1)}$$

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:

$$\int_{\Omega_{1/N}}^{\Omega_1}\frac{\mathrm{sh}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}{\alpha_0 e^{\bar{u}(\Omega)}+\mathrm{ch}\big(\Delta u(\Omega_0)-\Delta u(\Omega)\big)}\frac{1}{\Delta(\Omega)}\mathrm{d}\Omega = \frac{\mathrm{sh}\big(\Delta u(\Omega_0)\big)}{\alpha_0+\mathrm{c}\ \big(\Delta u(\Omega_0)\big)} \qquad \text{(Eq.A5-2)}$$

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. 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

9. Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet. 2004 Apr;5(4):251–61.

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. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res. 2018 Feb;122(3):433–43.

12. Herrmann M, Yampolsky LY. False and true positives in arthropod thermal adaptation candidate gene lists. Genetica. 2021 Jun;149(3):143–53.

13. Mähler N, Schiffthaler B, Robinson KM, Terebieniec BK, Vučak M, Mannapperuma C, et al. Leaf shape in Populus tremula is a complex, omnigenic trait. Ecol Evol. 2020 Nov;10(21):11922–40.

14. Zhang W, Reeves GR, Tautz D. Testing Implications of the Omnigenic Model for the Genetic Analysis of Loci Identified through Genome-wide Association. Curr Biol. 2021 Mar;31(5):1092-1098.e6.

15. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell. 2020 Sep;182(5):1214-1231.e11.

16. Mathieson I. The omnigenic model and polygenic prediction of complex traits. Am J Hum Genet. 2021 Sep;108(9):1558–63.

17. 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

899 18. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al.
900 Assumption-free estimation of heritability from genome-wide identity-by-descent
901 sharing between full siblings. PLoS Genet. 2006 Mar;2(3):e41.
902 19. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, et al.
903 Heritability of adult body height: a comparative study of twin cohorts in eight
904 countries. Twin Res. 2003 Oct;6(5):399–408.
905 20. Nelson RM, Pettersson ME, Carlborg Ö. A century after Fisher: time for a new
906 paradigm in quantitative genetics. Trends Genet. 2013 Dec;29(12):669–76.
907 21. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al.
908 Common SNPs explain a large proportion of the heritability for human height. Nat
909 Genet [Internet]. 2010/06/20. 2010 Jul;42(7):565–9. Available from:
910 https://pubmed.ncbi.nlm.nih.gov/20562875
911 22. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the
912 role of common variation in the genomic and biological architecture of adult human
913 height. Nat Genet. 2014 Nov;46(11):1173–86.
914 23. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and
915 misconceptions. Nat Rev Genet. 2008 Apr;9(4):255–66.
916 24. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A Saturated
917 Map of Common Genetic Variants Associated with Human Height from 5.4 Million
918 Individuals of Diverse Ancestries. bioRxiv [Internet]. 2022; Available from:
919 https://www.biorxiv.org/content/early/2022/01/10/2022.01.07.475305
920 25. Stigler SM. The History of Statistics: The Measurement of Uncertainty before 1900
921 [Internet]. Harvard University Press; 1990. Available from:
922 https://books.google.co.uk/books?id=nWn9DwAAQBAJ
923 26. 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.

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

28. 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

29. 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).

30. 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

31. 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

32. 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**
$\boldsymbol{\sigma}^2$	**15.72**	**17.12**	**17.73**	**16.67**

970 **Figure legends**

971

972 **Figure 1:** Phenotype human height simulated using the real data from the Genotype-Tissue
973 Expression (GTEx) project (28). We recall that for diploid organisms, such as humans, and
974 for a binary (bi-allelic, A or a) genetic marker, any microstate (genotype) can only take three
975 values that we shall write as '+1', '0' and '-1' corresponding to genotypes aa (red), Aa (grey)
976 and AA (blue), respectively. In the figures the letter, p, refers to the allele frequency and, σ_g,
977 to the genetic variance. Table 1 provides a relation between the two aforementioned
978 variables. **(A)** Phenotype histograms for the three genotypes (microstates), showing the '+1'
979 state in red, '0' state in dark grey and the '-1' state in blue. The overall phenotype
980 distribution is in light grey at the back. The mean genotypic (microstate) values are -a, d, and
981 +a. In genetics, a and d, are defined as the gene effect and the dominance, respectively. **(B)**
982 Example of limitations linked to current GWA methods when the gene effect, a, changes by
983 one order of magnitude from a=1 to a=0.1 whilst the respective number of microstates does
984 not. In this context, the three populations of genetic microstates collapse, and their separation
985 becomes very difficult to characterise unless the size of the population is increased to reduce
986 the width of categories. In this example, it is almost impossible to distinguish between a=0.1
987 and a=0.01. Note that in this case the gene effect has been normalised by the phenotypic
988 standard deviation of the population and there is no dominance (d=0). **(C)** Corresponding
989 distributions of the different genetic microstates singled out based on the phenotype values
990 they characterise. The method of averages, or method of moments, advocated by Fisher
991 suggests plotting a straight line as best fit between the three distributions out of which the
992 slope of the straight line is then indicative of genotype-phenotype association. However, the
993 method of averages discards the information that is available in the spreading of data for each
994 genotype. The method we suggest will make use of this information to describe genotype-
995 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)).

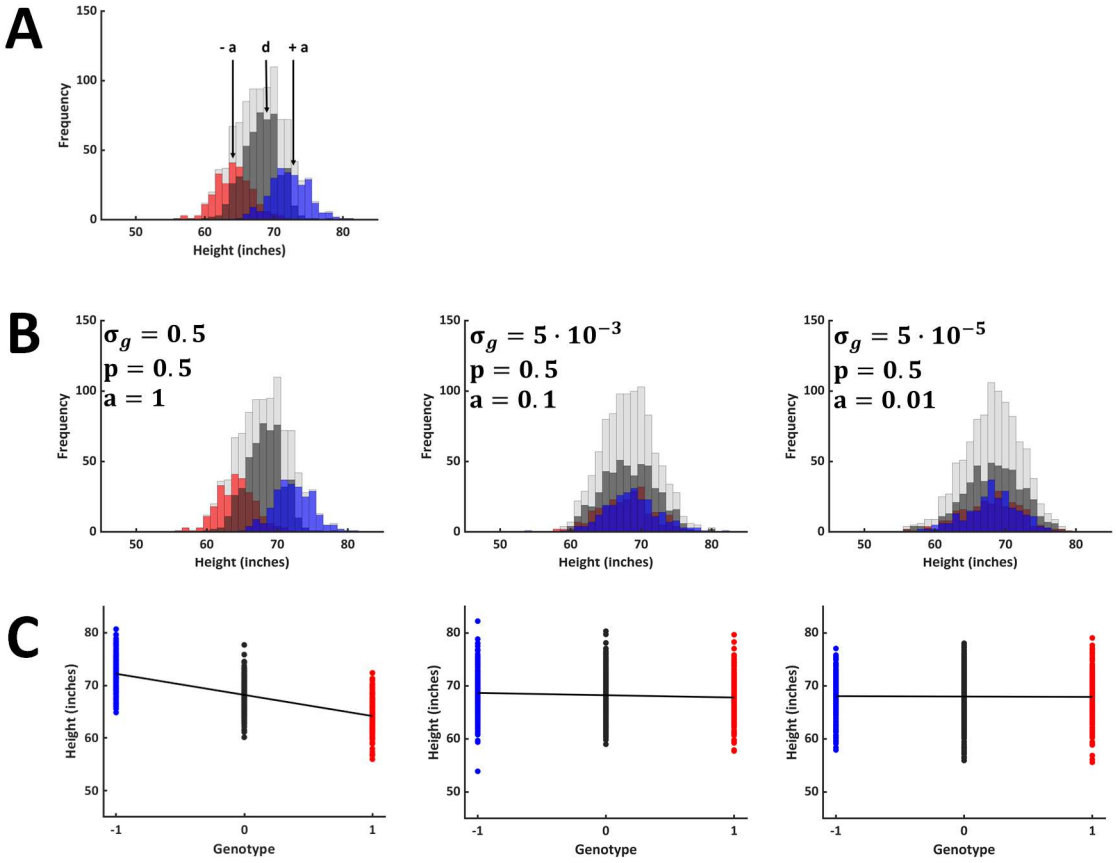
Figure 2: (A) & (B) When applied to real data sets, current genome-wide association studies rely on probability distribution density functions (PDFs) namely the creation of frequency plots (method of relative frequencies) via the grouping of phenotype values into categories representing range of phenotype values (A, top-chart). The same method (PDFs) is then applied to genotypes (B, top chart). We recall that 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 (red), Aa (grey) and AA (blue), respectively. The comparison of the two top charts in (A) and (B) demonstrates how genotype are associated with the phenotype, as in this case any phenotype category can be decomposed using the underlying microstate categories. However, grouping data into categories is legitimate so long that the width of the category is justified. The width of categories is justified provided that imprecision exists in phenotype measurements. For example, if height in human was the phenotype of interest measured with a ruler with inch graduations, namely measured to the nearest inch (scale of imprecision), then the width of categories would be 1 inch. However, a method based on the notion of imprecision has limited value when precision is available, and new methods are required in this case. Indeed, by increasing the precision in phenotype measurements it is possible to envision, in a near future, the possibility to deal with genotype and phenotype under the form of 'code bars' (A & B, bottom-charts) as opposed to PDFs . The question is then, how can information be extracted from those 'code bars'? **(C)** To analyse the 'code bar' in the bottom charts of (A&B), we rewrite it as a string of microstates. The first thing to note is that the way the microstates '+1', '0' and '-1' appear in the string (a.k.a. configuration/ordering), is linked to

1021 the fact that: (i) the genome position considered is associated with the phenotype and, (ii)
1022 phenotypic values are ranked as a function of their magnitude. Accordingly, if the genome
1023 position considered is not associated with the phenotype, the configuration would appear
1024 random (top string) as opposed to being ordered if a genotype-phenotype assoxaition exist
1025 (bottom string). **(D)** Thus, to determine as to whether a genome position is associated with
1026 the phenotype, the best way is to plot the cumulative sum of microstates and to compare it to
1027 the plot of its 'scrambled'/random configuration. When the string is ordered, a curve will
1028 emerge as shown by $\theta(i)$. On the contrary, when string is 'scrambled' a straight line is
1029 expected as shown by $\theta_0(i)$. The straight line resulting from scrambling of the string is
1030 expected as the presence probabilities of pulling '+1', '0' and '-1' from the scrambled string
1031 are constants in this case. The curve (ordered string) can be modelled by considering that the
1032 information about the phenotype acts like a 'field' to change the configuration of microstates
1033 from being disorderd to being ordered. The two curves intersect at the last position in the
1034 string because the numbers of '+1', '0' and '-1' in either string (ordered or scrambled) is
1035 conserved. The difference between the ordered and disordered strings provides a precise
1036 measure of the genetic influence on the phenotype. In Figure 4A we provide a plot of such a
1037 difference in the phenotypic space.
1038
1039 **Figure 3:** Power calculations for two allele frequencies as a function of the gene effect and
1040 sample size using Fisher's additive model when the dominance is null. The colour of the
1041 curves is associated with the sample size (one thousand individuals for black, ten thousand
1042 individuals for red, hundred thousand individuals for blue and one million individuals for
1043 green). The figure illustrates that with 1000 individuals current genotype-phenotype methods
1044 are not powerful enough to detect small gene effects.
1045

1046 **Figure 4: (A)** Simulation of genetic paths difference in the space of positions as a function of
1047 log-scale values of the gene effect normalised by the phenotypic standard deviation, noted
1048 'a/σ', for $\Delta\omega_0 = 0$. **(B)** Simulations of genetic paths difference in the space of phenotype
1049 values as a function of log-scale values of the gene effect normalised by the phenotypic
1050 standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0$. **(C)** Natural-logarithm transformation of the
1051 genetic paths difference in the space of phenotype values as a function of log-scale values of
1052 the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0$. **:**
1053 **(D)** Simulation of genetic paths difference in the space of positions as a function of log-scale
1054 values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for
1055 $\Delta\omega_0 = 0.6$. **(B)** Simulations of genetic paths difference in the space of phenotype values as a
1056 function of log-scale values of the gene effect normalised by the phenotypic standard
1057 deviation, noted 'a/σ', for $\Delta\omega_0 = 0.6$. **(C)** Natural-logarithm transformation of the genetic
1058 paths difference in the space of phenotype values as a function of log-scale values of the gene
1059 effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = 0.6$.
1060
1061 **Figure 5: (A)** Representation of the shift in the phenotype value for which the genetic path
1062 difference is maximal as a function of the gene effect 'a/σ' for different values of allele
1063 frequencies. N=1000 represents the number of individuals in the simulation. **(B)**
1064 Representation of the shift in the phenotype value for which the genetic paths difference is
1065 maximal when the population is increased by a factor 10 (N=10,000) for the gene effects
1066 verifying $a/\sigma < 0.1$.
1067
1068 **Figure 6:** Comparison between simulated and theoretical gene effects the black dots
1069 correspond to the cases where the gene effects simulated are identical to the theoretical gene
1070 effects.

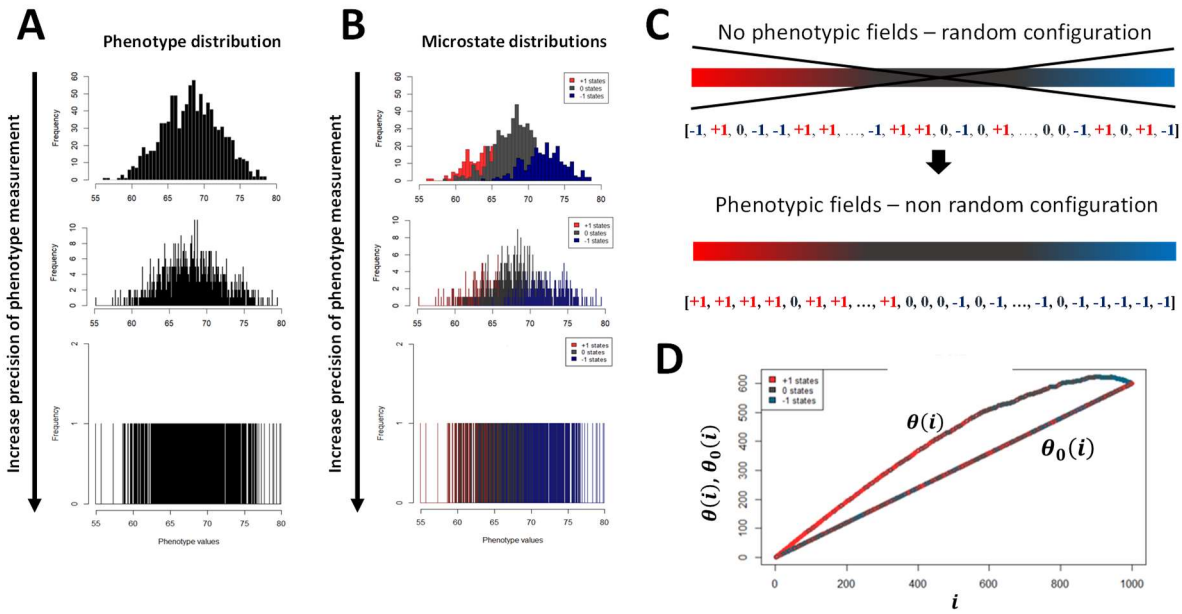
1071

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

1075 Figure 2

1076



1077

1078

Figure 3

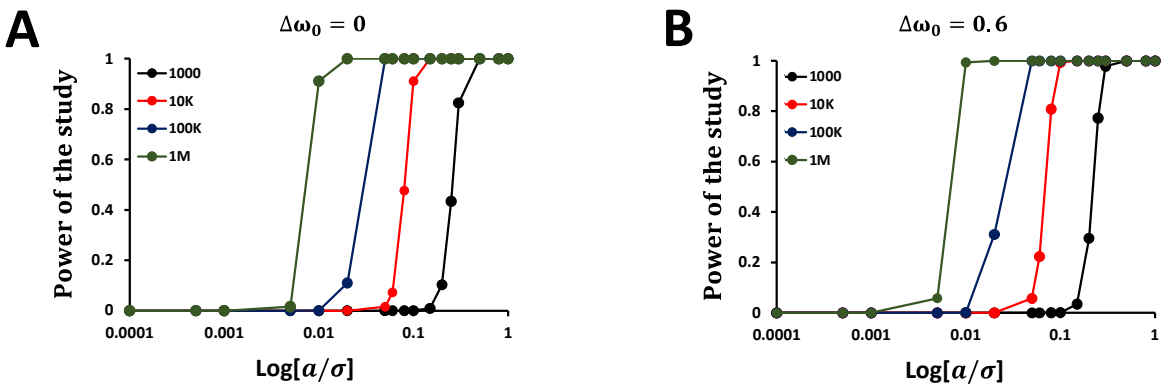

A
$\Delta\omega_0 = 0$
Power of the study
1000
10K
100K
1M
Log[a/σ]
B
$\Delta\omega_0 = 0.6$
Power of the study
1000
10K
100K
1M
Log[a/σ]

Figure 4

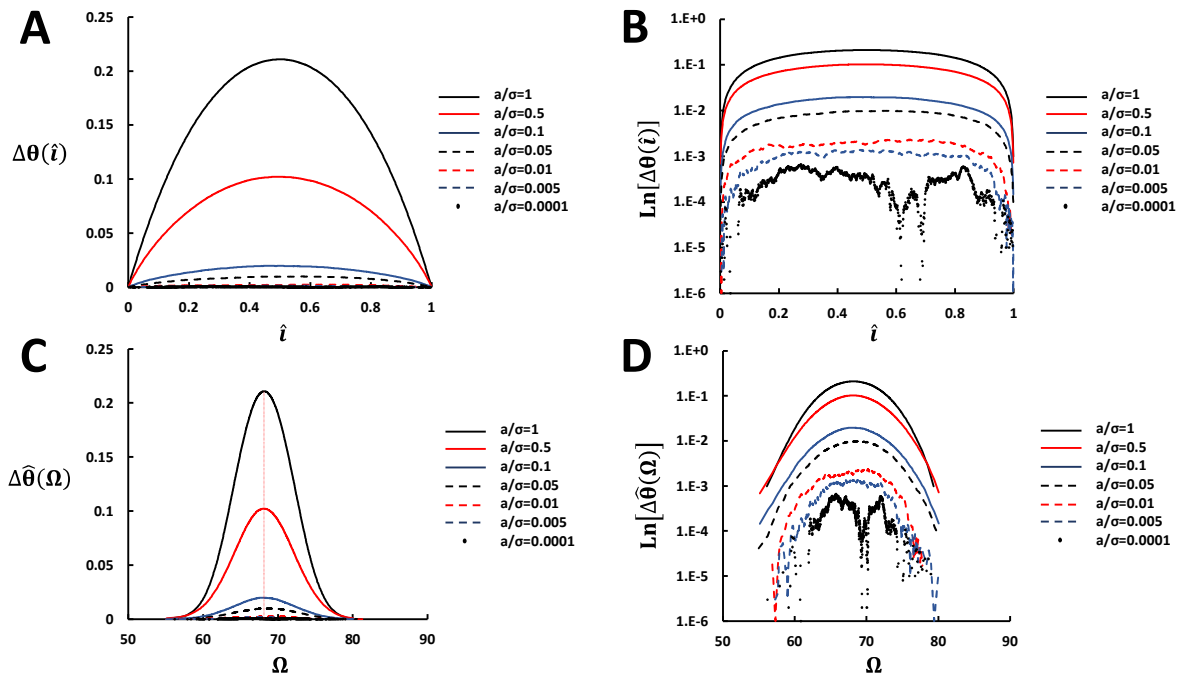

A
B
C
D
Δθ(î)
Ln[Δθ(î)]
Δθ̂(Ω)
Ln[Δθ̂(Ω)]
î
Ω
a/σ=1
a/σ=0.5
a/σ=0.1
a/σ=0.05
a/σ=0.01
a/σ=0.005
a/σ=0.0001

Figure 5

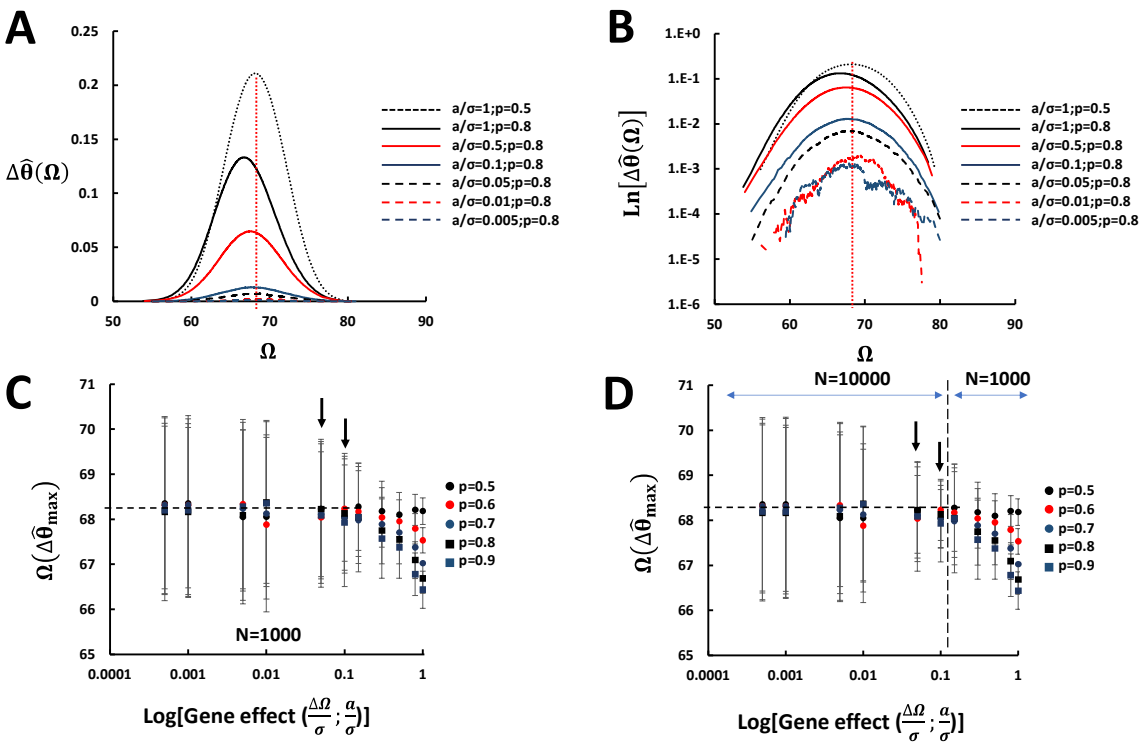
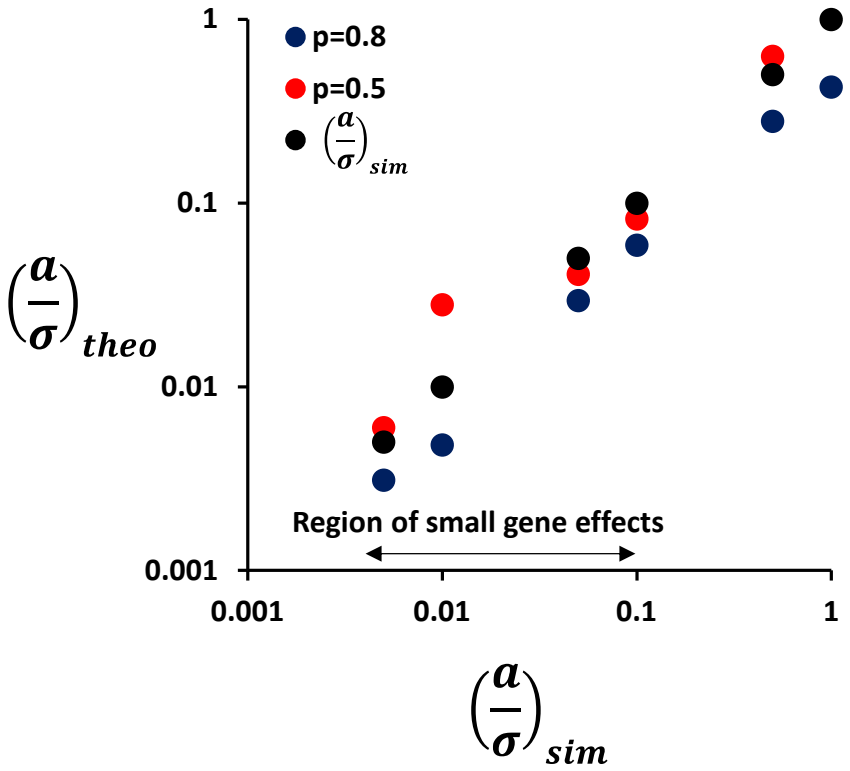

A
B
C
D
a/σ=1;p=0.5
a/σ=1;p=0.8
a/σ=0.5;p=0.8
a/σ=0.1;p=0.8
a/σ=0.05;p=0.8
a/σ=0.01;p=0.8
a/σ=0.005;p=0.8
Ω
N=1000
N=10000
N=1000
p=0.5
p=0.6
p=0.7
p=0.8
p=0.9
Log[Gene effect (ΔΩ/σ ; a/σ)]

Figure 6


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