FIT-GWA: A new method for the genetic analysis of small gene effects, high precision in phenotype measurements and small sample sizes.

Cyril Rauch[*], Sarah Blott[*], Panagiota Kyratzi[#], 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, FIT-GWA

Abstract

Small gene effects involved in complex traits remains difficult to analyse using current genome-wide association methods (GWAS) due to the number of individuals required to return meaningful association(s), a.k.a. study power. Inspired by physics fields theory we provide a different method called Fields Informational Theory for Genome-Wide Associations (FIT-GWA). Contrary to GWAS, FIT-GWA that the phenotype is measured precisely enough and/or the number of individuals in the population is too small, to permit categories. To extract information FIT-GWA use the difference in the cumulated sum 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 phenotype value. Such difference can be accounted through the emergence of 'phenotypic fields'. We demonstrate that FIT-GWA recovers GWAS, i.e., Fisher's theory, when the phenotypic fields are linear. However, unlike GWAS, FIT-GWA permits to demonstrate how the variance of microstate distribution density functions are also involved in genotype-phenotype associations. Using genotype-phenotype simulations based on Fisher's theory we illustrate the application and power of the method with a small sample size of 1000 individuals.

1. Introduction

Identifying association between phenotype and genotype is the fundamental basis of genetic analysis. 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 phenotype between individuals with known relationship (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 phenotype to be associated with genotype 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, i.e., 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. GWAS have been employed in many species, and especially in the study of human disease (1). The 2021 NHGRI-EBI GWAS Catalog currently lists 316,782 associations identified in 5149 publications describing GWAS results (2). Additionally, extensive collection of data has been initiated through efforts such as the UK Biobank (3), Generation Scotland (4) and NIH All of Us research program (https://allofus.nih.gov/) in the expectation that large-scale GWAS will elucidate the basis of human health and disease and facilitate precision medicine.

While genomic technologies used to generate data have rapidly advanced within the last 20 years, the statistical models used to analyse the data are still based on the linear models derived by Fisher more than 100 years ago (5,6). GWAS essentially makes use of Fisher's method of partitioning genotypic values by performing a linear regression of the phenotype on marker allelic dosage (7). The regression coefficient estimates the average allele effect size, and the regression variance is the additive genetic variance due to the locus (8). However, ongoing

debate exists over whether the present analysis paradigm in quantitative genetics is at its limits for truly understanding complex traits.

For example, the phenotype height in humans is a classical quantitative trait that has been studied for well over a century as a model for investigating the genetic basis of complex traits (5,9) and whose measured heritability is well known to be around 80% (5,10,11). Yet, this phenotype remains controversial (12) as current association methods have not been able to recover fully the measured heritability (13,14). Whilst some explain this discrepancy as being associated with the restricted sample size used or the involvement of an ill-defined environment (7,15), one may wonder whether Fisher's method is accurate enough to detect small gene effects.

Another issue concerns the role of microstate variances. Whilst Fisher's method uses the distribution density function of microstates to detect differences between averages, it is still unclear as to how the variance of microstates are involved in genotype-phenotype associations, beyond their role in heredity (12).

Altogether, those results suggest that current genotype-phenotype mapping methods, based on Fisher's works, i.e., the method of averages, are potentially limited to capture the full potential of genotype-phenotype associations.

R.A. Fisher's work is based on statistics and by definition statistics deals with the measurement of uncertainties (16). To draw inferences from the comparison of data, a method is needed that requires some understanding about its accuracy, including ways of measuring the uncertainty in data values. In this context statistics is the science of collecting, analysing, and interpreting data; whilst probability, defined through relative frequencies, is central to determining the validity of statistical inferences. In practice, the use of frequentist probability, and the resulting binning or categorization of data, is justified when inaccuracy exists in experimental

measurement. For example, measuring a continuous phenotype such as the height of individuals with a ruler with centimetre graduations, i.e., to the nearest centimeter, warrants the use of frequentist probability. In this case, a frequency table of phenotype values can be defined through 1cm-width bins or categories, from which the probability density functions can be deduced to address the statistical inferences.

However, this method becomes problematic when the measurement of phenotype values can be carried out with very high precision, for example using highly advanced imaging techniques or biosensing technologies (17). In this case each individual measured could return a unique phenotype value. The phenotype values being unique how can 'randomness in the data' be defined, and frequentist probability used, to determine any inferences?

If this happens, only two options are possible. The first one is to artificially create categories and bin or group the data; the other is to increase the population size so as to ensure categories can be defined to match the precision required. Creating artificial imprecisions is clearly a step backwards in an era of advanced measurement technologies and increasing the number of individuals seems to be the way forward. However, even if the latter option was chosen, there is no guarantee that any population will be large enough to match the precision achieved by biosensing technologies, namely, to permit the creation of categories. This implies that the notion of 'precision' that can be achieved when frequentist probability is used, is always bounded by the population size, be it the world population. Figure 1 exemplifies the difficulty to extract genotype-phenotype association(s) with current methods when the sample size is restricted.

There is, therefore, a need to formulate new methods that make full use of the information generated through accurate and highly precise phenotyping which do not require categorization of the data. In fact, this problem is the equivalent of finding a way to resolve genotype-phenotype mapping by assuming a finite size population with phenotype values measured

precisely enough such as to rule out the possibility that two phenotype values are identical. Taking this challenge as the starting point a new and relatively simple method of extracting information for genotype-phenotype mapping can be defined. Whilst this method is remarkably simple when explained in lay terms, its theoretical framework requires the introduction of a new concept that we call the 'phenotypic field'. It will also be seen that phenotypic fields can be defined in the context of Fisher's theory.

The paper is structured as follows. A theoretical framework is provided with a particular focus on small gene effects. It is demonstrated that the method converges rapidly, ruling out the need for large population size to extract precise genotype-phenotype information. Finally, the application of the method is illustrated using simulated and real data.

2. Heuristic presentation of FIT-GWA: Introduction to the genetic path method

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

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

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

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

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

In the second configuration, after selecting on phenotype the horses are ranked by phenotype value. For example, if the phenotype is the height, the smallest horse is allocated to the first

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2. Entropy of genetic paths

Keeping the notations, ω_+^0, ω_0^0 and ω_-^0, for the genetic microstate frequencies of a given genome position across the population of individuals we aim to determine an expression for $\omega_+(\hat{\imath})$, $\omega_0(\hat{\imath})$ 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 that is similar to an absence of association between the genetic microstates and the phenotype values leading, as a result, 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. As a result, 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 phenotype values as:

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

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

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

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

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

As '$S - S_0$' is linked to the information gained on phenotype values and 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 the phenotypic fields:

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

There are two different but mathematically equivalent ways to consider the notion of phenotypic fields. The first way is to consider that the fields only appear due to the ranking of individuals. In this context one may consider that the set of microstates change 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, namely the fields are null. In this context one can consider the equivalence: $S - S_0 \sim E$.

The second way is to consider that the fields are always switched on for all genome positions, but that some 'work' is provided such as to transform a 'phenotype-responding genetic path' to a non-responding one. In this context the sum of interactions defining the default, i.e., non-responding, genetic path must be defined as: $E_0/N = \int_{\Omega_{1/N}}^{\Omega_1} (u_+(\Omega)\omega_+^0 + u_0(\Omega)\omega_0^0 + u_-(\Omega)\omega_-^0)\frac{1}{\Delta(\Omega)}d\Omega$. Then, a connection between '$S - S_0$' and '$E - E_0$' can be postulated through the existence of a genotype-phenotype association constant 's' such as to provide the equivalence: $S - S_0 = s(E - E_0)$; where 's' provides the 'magnitude' of the interaction or association between the genotype and the phenotype. It is this last consideration that we shall use throughout. As a result, the relation to optimize is:

$$\Delta S - s\Delta E = 0$$

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

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

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

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

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

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

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

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

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

In order to make the asymmetries of the problem more apparent the following are defined for the genetic microstates:

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

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

and for the phenotypic fields:

$$2s\bar{u}(\Omega) \stackrel{\text{def}}{=} s\delta u_+(\Omega)+s\delta u_-(\Omega)=su_+(\Omega)+su_-(\Omega)-2su_0(\Omega)$$

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

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

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

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

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

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

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

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

The new variable 'Ω_0' is then 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 Hardy-Weinberg Law. Hardy-Weinberg Law based on random matting in a population provides a relationship between the genetic microstate frequencies under the form: $p^2 + 2pq + q^2 = 1$, where p^2 and q^2 are the genotype frequencies of genetic microstates '+1' and '-1', i.e. homozygote genotypes aa and AA, respectively; and $2pq$ the genotype frequency for genetic microstate '0', i.e. the heterozygote genotype Aa. In our case this corresponds to replacing p^2, q^2 and $2pq$ by, respectively, ω_+^0, ω_-^0 and ω_0^0. As a result, Hardy-Weinberg Law imposes $\alpha_0 = 1$; and $\alpha_0 \neq 1$ 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 paths difference in the space of phenotype value

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

The phenotype-associated genetic path is simply the integration of Eq.5 over the phenotype values that is given, as seen above (Eq.2), by: $\int_{\Omega_{1/N}}^{\Omega}[\widehat{\omega}_+(\Omega) - \widehat{\omega}_-(\Omega)]\frac{1}{\Delta(\Omega)}d\Omega$. The default

genetic path is deduced from considering that the difference in the presence probabilities between the genetic microstates '+1' and '-1' is constant, i.e.: $\int_{\Omega_{1/N}}^{\Omega} \Delta\omega_0 \frac{1}{\Delta(\Omega)} d\Omega$.

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

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

To which the conservation of genetic microstates needs to be added whatever the genetic path taken, i.e. $\Delta\hat{\theta}(\Omega_1) = 0$, expressed as:

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

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

As it turns out this minimalist model can be developed (see SM2 in supplementary materials) and is similar to Fisher's theory (see below).

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

3.5.1. Coarse graining FIT-GWA

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

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

Similarly, concentrating on a single genome potion, we can define by: $\delta N_{+} \sim N_{+}^{0} \cdot P_{+}(\Omega)\delta\Omega$, $\delta N_{0} \sim N_{0}^{0} \cdot P_{0}(\Omega)\delta\Omega$ and $\delta N_{-} \sim N_{-}^{0} \cdot P_{-}(\Omega)\delta\Omega$ the respective number of '+1', '0' and '-1' genetic microstates for the phenotype values ranging between 'Ω' and '$\Omega + \delta\Omega$', where 'N_{+}^{0}', 'N_{0}^{0}' and 'N_{-}^{0}' correspond to the total number of '+1', '0' and '-1' microstates in the population with respective presence probability density functions '$P_{+}(\Omega)$', '$P_{0}(\Omega)$' and '$P_{-}(\Omega)$'.

The creation of artificial categories imposes two conservation relations. The first conservation relation concerns the total number of individuals and microstates, namely that for a given genome potion the sum of all possible microstates is also the sum of all individuals. This relation is written as:

$$N = N_{+}^{0} + N_{0}^{0} + N_{-}^{0}$$

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

$$\delta N = \delta N_{+} + \delta N_{0} + \delta N_{-}$$

As a result, the conservation relation concerning the number of individuals and microstates in the category concerned can be rewritten as:

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

By using the probability densities function of both microstates and phenotype defined above it is then deduced:

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

From Eq.9 all the moments of microstate distributions are related to the ones 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. One deduces, from Eq.9a, the conservation relations for the first two moments under the form:

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

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

The relations provided by Eq.9b are valid by definition, namely whatever probability density functions involved. Whilst Fisher never formulated a conservation relation similar to Eq.9a in his seminal paper (5), he used the notation 'α^2' to denote the genetic variance under the form:

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

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

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

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

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

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

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

$$\omega_+^0 - \omega_-^0$$

A direct mapping of fields can then be performed between GWAS and FIT-GWA (SM3 supplementary materials). As a result, one can define the coarse-grained versions of the genetic paths difference using the continuum limit under the form:

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

Eq.11 demonstrates that '$\Delta\hat{\theta}(\Omega)$' is sensitive to the probability density functions involved as a whole and not just to average values. That is to say that 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, can be envisaged.

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

Finally using Eq.7's notations one deduces:

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

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

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

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

3.5.2. Definition of Fisher's fields

In his seminal paper (5) Fisher hypothesized 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 that will depend on the variances of microstates (see 3.5.4., i.e., the definition of fields linked to the variance of microstates below).

Following Fisher's assumption, the probability of finding the microstates '+1', '0' and '-1' as a function of phenotype values are expressed as (5) (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 $(\delta a_-/\sigma)^2$ can be neglected, one deduces then:

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

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

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

In Eq.12a and Eq.12b the term: $\delta a_- - \delta a_+ = a_- - a_+ = 2a$ (see Fig.1A), is known as the 'gene effect' in GWAS. In Eq.12c the term: $2\delta a_0 - \delta a_+ - \delta a_- = 2a_0 - a_+ - a_- = d$ (see Fig.1A), is the dominance as defined in GWAS. As it turns out, Fisher considered: $d \sim 0$. Altogether these results demonstrate that Fisher's theory can be described by the minimalist model aforementioned (SM2 in supplementary materials). Using those fields, it is also possible to determine a generic solution to Eq.8 (see SM4 in supplementary materials).

3.5.3. Implication for small gene effects

Complex traits involve genes with very small effects that are difficult to characterise. The aim now is to determine the resulting genetic paths difference in this case, i.e., when $a \to 0$.

As probability density functions have been used, the integration interval can be altered using the convergence property of distributions. In this context, the conservation of genetic microstates (Eq.8) can be written using Fisher's fields as:

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

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

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

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

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

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

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

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

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

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

3.5.4. Fields linked to the variance of microstates

The involvement of variances of microstate distribution functions in genotype-phenotype associations is a highly debated matter, see (12) and references within.

As said above, the expression of the genetic paths difference considers the distribution density function as a whole including the role of variances. In this context, we saw that Eq.9b provides a relation between variances under the form of an ellipse. Whilst assuming a single variance

for all microstate and phenotype distributions, as Fisher did, is plausible; 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:

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

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

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

As a result, one can define the pseudo-gene effect and pseudo-dominance linked to the variances as:

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

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

To conclude, as the genetic paths difference 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 FIT-GWA using simulated data

It is our intention to illustrate how FIT-GWA can be applied using simulated data and assess, at least qualitatively, its sensitivity to extract information.

Data was simulated according to quantitative genetic models defined by Falconer and Mackay (1996) (18). A single bi-allelic quantitative trait locus (QTL) associated with a continuous phenotype was modelled, with additive allele effect, a, and allele frequencies, p and q, where

$p + q = 1$. 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 from Fisher's theory. Whilst 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 mid-way point between the means of the +1 and -1 microstates. For the purposes of the simulation dominance, d, was 0, i.e. 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}) is defined as (18): $\sigma^2_{QTL} = 2pq[a + d(q - p)]^2$. Each individual was given 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 then generated by adding a random environmental effect to the genotypic value of each individual. The environmental effect added was a random variate drawn from a normal distribution with mean 0 and variance: $1 - \sigma^2_{QTL}$. The phenotype was then rescaled to a value representing a realistic data set: Phenotype = (simulated phenotype x standard deviation of real data) + mean of real data. In this case, the real data set modelled was the Genotype-Tissue Expression (GTEx) project (19). The phenotype was height with mean = 68.17206 inches and standard deviation = 4.03007 inches. For each simulated replicate of N individuals the difference between the cumulative sums of microstates ordered by phenotype value and of genotypes in a randomised order with respect to phenotype was determined to create the genetic paths difference. The maximum value of this difference was identified and

its position and phenotypic value in the ordered string of microstates was recorded. Where the maximum value extended over several positions, the mean position and phenotypic value was taken. Finally, to ease representation the amplitude of the genetic paths were normalised by the population size ($N = 1000$ in this case).

Note that the standard deviation(s) arising from genotype-phenotype simulations were not taken into consideration in the analysis that will follow. Instead, we report the theoretical analysis of the convergence of the genetic paths difference method and its self-consistency, as well as its sensitivity to detect genotype-phenotype associations using simulations, in SM5 and SM6, respectively (see supplementary materials).

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=0.2 ($\Delta\omega_0 = -0.6$) 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 (20). The conclusion from Figure 3 is that 1000 individuals would not allow 80% power to be achieved unless the gene effect is large enough, i.e. $a/\sigma \geq 0.5$.

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

In Figures 4C, the profile of the phenotype distribution density function is recovered with an amplitude that decreases as 'a/σ' decreases. The red vertical line represents the average phenotype value. With the data simulated, Figure 4D demonstrates that a difference between genetic paths can be seen relatively clearly between the gene effects $a/\sigma \sim 0.1$ and $a/\sigma \sim 0.01$. One may then compare how perceptible are associations using the genetic paths difference by comparing Figures 1B&1C (method of averages) and Figures 5B&5D (method of genetic paths

difference) for identical allele frequencies and gene effects. The results suggest that FIT-GWA can be applied with 1000 individuals to return information regarding potential genotype-phenotype associations that would not be otherwise possible with current association studies.

Concentrating on different allele frequencies given by $p = 0.2$ ($\Delta\omega_0 = -0.6$) as an example. Figures 5A&5B are representations of '$\Delta\hat{\theta}(\Omega)$' and its log-scale transformation for log-scale values of 'a/σ'.

Differences are clearly visible between Figures 4C and 4D, and 5A and 5B since the phenotype values for which '$\Delta\hat{\theta}(\Omega)$'s' are maximal have been shifted from the average value of the phenotype (indicated by the vertical red dashed line). This is not surprising since the simulation only imposes a set of genetic variances without any constraint on the conservation of the average phenotype value.

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

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

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

similar to that of the average value of the phenotype, marked by the horizontal dashed line in Fig.5C.

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

Finally, as Eq.13 suggests that for small gene effects the genetic paths difference is proportional to the gene effect, a/σ, under the form: $\Delta\hat{\theta}(\Omega) \sim \omega_0 \frac{a}{\sigma} K e^{-\frac{1}{2\sigma^2}(\Omega - \langle\Omega\rangle)^2}$. It was decided to fit the log-values of '$\Delta\hat{\theta}(\Omega)$' in Figures 4D and 5B with quadratic curves of the form: $Ln[\Delta\hat{\theta}(\Omega)] \sim A\Omega^2 + B\Omega + C$; see Tables 2 and 3.

Then, as $\Delta\hat{\theta}(\Omega) \sim e^{\left(C - \frac{B^2}{2A}\right)} e^{A \cdot \left(\Omega + \frac{B}{2A}\right)^2}$ with such a fit, one expects by identification to Eq.13 that for small gene effects: $\bar{\Omega} \sim -\frac{B}{2A}$, $-\frac{1}{2\sigma^2} \sim A$ 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 '$\bar{\Omega}$' and 'σ' and setting $K = 1/\sqrt{2\pi}$, Figure 7 provides a comparison between the gene effect from the simulations, $\left(\frac{a}{\sigma}\right)_{sim}$, and the gene effect deduced from Eq.13, $\left(\frac{a}{\sigma}\right)_{theo}$.

Thus, recalling that the phenotype average and variance of the population modelled are, respectively, 68.17 inch and 16.24 inch^2; Table 2 and Table 3 demonstrate that fitting the genetic paths difference as a function of phenotype values with a quadratic curve recovers the magnitude of the average and variance of the phenotype used for the simulations for most log-

scale values of the gene effect. Furthermore, the amplitude of the genetic paths differences is also indicative of the gene effects involved.

5. Discussion

In his seminal paper (5), Fisher provided a synthesis between genetic inheritance of continuous trait and the Mendelian scheme of inheritance using statistics and probability. His theory has become a landmark in the field of genetics and heredity. Whilst statistics is a natural field to employ when dealing with large data sets, the interpretation of data as well as the inferences that can be drawn rely fundamentally on the probability of occurrence of observables.

Mathematically, the probability of one particular event is only fully defined once are known: (i) the set of all possible outcomes and, (ii) the process by which the event is generated. Without this information only conjectures can be made based on interpolating data. Thus, using a normal distribution as an underlying template for the occurrence of a measurable variable, as Fisher did for the phenotype and gene microstates, comes with limitations.

The first limitation, as mentioned in the introduction, is that the normal distribution expressed in the continuum limit is a probability density function that originates from the creation of categories. The act of creating categories implies losing information that, in turn, defines the notion of 'randomness in the data'. Whilst randomness in the data can be linked to genuine imprecision in measurements, there is no justification to use such a method when precision is potentially available.

The second limitation is linked to the first one in the sense that if randomness in the data is considered as a nuisance, then the only meaningful parameter is often the average. However, it is not because an average can be calculated from any data set that it is the most pertinent parameter. Inspired by statistical physics (21), Fisher used the central limit theorem (CLT) and the continuum limit as a central paradigm where average and variance are ontological

parameters. However, the CLT is valid only if the same 'experiment' can be repeated a very large (infinite) number of times under the same conditions. Notwithstanding the fact that individuals in a population are not the outcome of the same 'experiment', populations and genomes have also finite sizes, namely, populations are not infinite.

It is remarkable that Fisher never considered this limit (finite size of populations) and never thought about the consequences of creating categories to apply frequentist probabilities. Had he done so he would have deduced a relation similar to Eq.9a. Eq.9a and Eq.9b have far-reaching consequences since they imply that the variance of the phenotype can be decomposed formally in two terms only defined by the distribution density function of genes. Namely, one term linked to the genetic variance (noted 'α^2') and a second term representing the sum of variances of microstates weighted by their relative frequencies (noted '$\omega_+^0\sigma_+^2 + \omega_0^0\sigma_0^2 + \omega_-^0\sigma_-^2$', see Eq.9b). This second term defined by gene microstates is conceptually similar to what is known today as the variance linked to the 'environment' in GWAS. However, Eq.9a suggests that adding the notion of 'environment' is not required at this stage. The same consequence also applies to heredity whose definition, in a broad sense, is given by the ratio between the genetic and the phenotype variances. Finally, Eq.9a and Eq.9b demonstrate also that it is not possible to fully dissociate variances and averages and as a result any genotype-phenotype statistical associations must consider both, and not only averages. This point is visible when the fields linked to microstate variances are considered. Therefore, the paradigm used by Fisher is limited and the method of averages too restrictive.

FIT-GWA is a method that circumvents this set of conceptual difficulties by concentrating on the information contained in genotype and phenotype in a different way. FIT-GWA can be used with Fisher's assumptions to recover key concepts from quantitative genetics including: (i) the Hardy-Weinberg coefficient locally, (ii) the Hardy-Weinberg coefficient at the

population level, (iii) the gene effect, (iv) the dominance and, (v) that small gene effects involve common allele frequencies (22). In addition, FIT-GWA has enabled us to define new parameters linked to the microstate variances, i.e., the pseudo-gene effect and pseudo-dominance, that will probably help resolving controversies (12). Therefore, taken as a whole the coarse-grained version of FIT-GWA generalizes Fisher's method.

Finally, applying FIT-GWA on simulated data using Fisher's assumption proves its sensitivity for extracting information on genotype-phenotype associations when sample sizes and gene effects are small. SM7 provides an example of application of FIT-GWA using real data (see supplementary materials).

6.Conclusion

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

Data accessibility. The datasets are available upon request.

Authors' contributions. Conceptualisation of FIT-GWA as a new method based on physics fields 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 FIT-GWA: C.R., S.B., P.K.

Competing interests. We declare we have no competing interests.

Funding. The work was supported by Physics of Life Network and MarsPetcare.

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

References

1. Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature [Internet]. 2007 Jun 7;447(7145):661–78. Available from: https://pubmed.ncbi.nlm.nih.gov/17554300
2. 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.
3. 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
4. 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.
5. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans R Soc Edinburgh. 1919;52(2):399–433.
6. 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
7. Visscher PM, Goddard ME. From R.A. Fisher's 1918 Paper to GWAS a Century Later. Genetics. 2019 Apr;211(4):1125–30.
8. Hivert V, Wray NR, Visscher PM. Gene action, genetic variation, and GWAS: A user-

friendly web tool. PLoS Genet. 2021 May;17(5):e1009548.

9. 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

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

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

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

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

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

15. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008 Apr;9(4):255–66.

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

17. 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.

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

19. 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

20. 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).

21. Morrison M. Physical Models and Biological Contexts. Philos Sci [Internet]. 1997 Aug 27;64:S315–24. Available from: http://www.jstor.org.nottingham.idm.oclc.org/stable/188413

22. 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

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 Fig.5D 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=0.1 p=0.5**	**a/σ=0.05 p=0.5**	**a/σ=0.01 p=0.5**	**a/σ =0.005 p=0.5**
A	-0.0371	-0.0321	-0.0312	-0.0311	-0.0361	-0.0435
B	5.0539	4.3799	4.2458	4.2425	4.9041	5.9241
C	-173.69	-151.48	-148.56	-149.5	-171.72	-208.35
$\mathbf{r^2}$	0.9960	0.9990	0.9983	0.9937	0.8805	0.8329
$\bar{\boldsymbol{\Omega}}$	**68.11**	**68.22**	**68.04**	**68.21**	**67.92**	**68.09**
σ	**13.48**	**15.58**	**16.02**	**16.08**	**13.85**	**11.49**

Table 3: Nonlinear fit results for Fig.6B 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=0.8	**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**
A	-0.0343	-0.0316	-0.0305	-0.0313	-0.0445	-0.0300
B	4.5867	4.2724	4.1448	4.2674	6.0789	4.0741
C	-155.49	-146.99	-144.95	-150.28	-214.24	-145.4
$\mathbf{r^2}$	0.9983	0.9990	0.9981	0.9960	0.9210	0.7554
$\bar{\boldsymbol{\Omega}}$	**66.86**	**67.60**	**67.95**	**68.17**	**68.30**	**67.90**
σ	**14.58**	**15.82**	**16.39**	**15.97**	**11.23**	**16.67**

Figure legends

Figure 1: Phenotype human height simulated using the real data from the Genotype-Tissue Expression (GTEx) project (19). **(A)** Phenotype histograms for the three genotypes (microstates), showing the +1 state in red, 0 state in dark grey and the -1 state in blue. The overall phenotype distribution is in dark grey. The mean genotypic (microstate) values are -a, d, and +a. **(B)** Example of limitations linked to current GWAS methods when the gene effect, a, changes by one order of magnitude from a=1 to a=0.1. In this context, the three populations of genetic microstates collapse, and their separation becomes very difficult to characterise unless the size of the population is increased to reduce the width of categories. In this example, it is almost impossible to distinguish between a=0.1 and a=0.01. Note that in this case the gene effect has been normalised by the phenotypic standard deviation of the population and there is no dominance (d=0). **(C)** Corresponding distributions of the different genetic microstates singled out based on the phenotype values they characterise. The method of averages, or method of moments, advocated by Fisher suggests plotting a straight line as best fit between the three distributions out of which the slope of the straight line is then indicative of genotype-phenotype association. However, the method of averages discards the information that is available in the spreading of data for each genotype. The method we suggest will make use of this information to describe genotype-phenotype association. Note that dominance can occur ($d \neq 0$) at which point and with the method of average a linear regression of the genotype means, weighted by their frequency, on the number of alleles needs to be performed to provide a new intercept and slope (see (7)).

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

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

Figure 3: Power calculations for two allele frequencies as a function of the gene effect and sample size. The colour of the curves is associated with the sample size (one thousand individuals for black, ten thousand individuals for red, hundred thousand individuals for blue

and one million individuals for green). The figure illustrates that with 1000 individuals current genotype-phenotype methods are not powerful enough to detect small gene effects.

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

Figure 5: (A) Simulations of genetic paths difference in the space of phenotype values as a function of log-scale values of the gene effect normalised by the phenotypic standard deviation, noted 'a/σ', for $\Delta\omega_0 = -0.6$. **(B)** Log-scale representation of the genetic paths difference in the space of phenotype values as a function of 'a/σ', for $\Delta\omega_0 = -0.6$. **(C)** Representation of the shift in the phenotype value for which the genetic path difference is maximal as a function of the gene effect 'a/σ' for different values of allele frequencies. N=1000 represents the number of individuals in the simulation. **(D)** Representation of the shift in the phenotype value for which the genetic paths difference is maximal when the population is increased by a factor 10 (N=10,000) for the gene effects verifying $a/\sigma < 0.1$.

Figure 6: Comparison between simulated and theoretical gene effects.

Figure 1

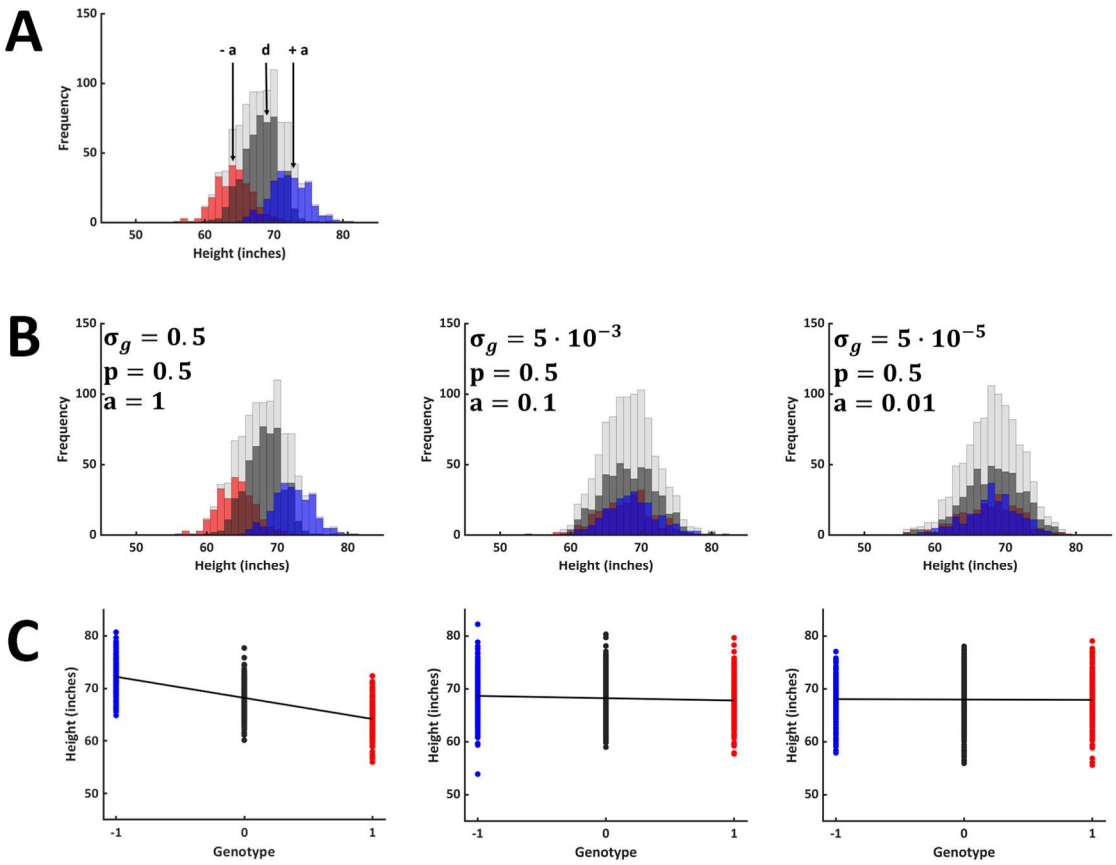
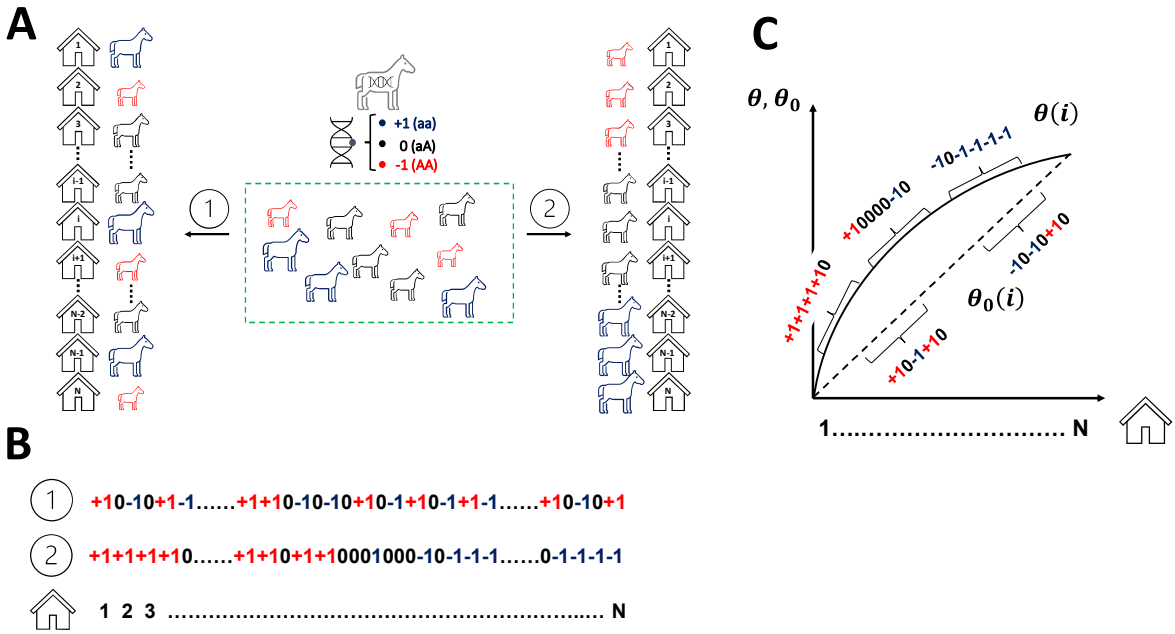

A
- a
d
+ a
Frequency
Height (inches)
B
$\sigma_g = 0.5$
$p = 0.5$
$a = 1$
$\sigma_g = 5 \cdot 10^{-3}$
$p = 0.5$
$a = 0.1$
$\sigma_g = 5 \cdot 10^{-5}$
$p = 0.5$
$a = 0.01$
C
Height (inches)
Genotype

Figure 2



B

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

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

1 2 3 ..... N

Figure 3

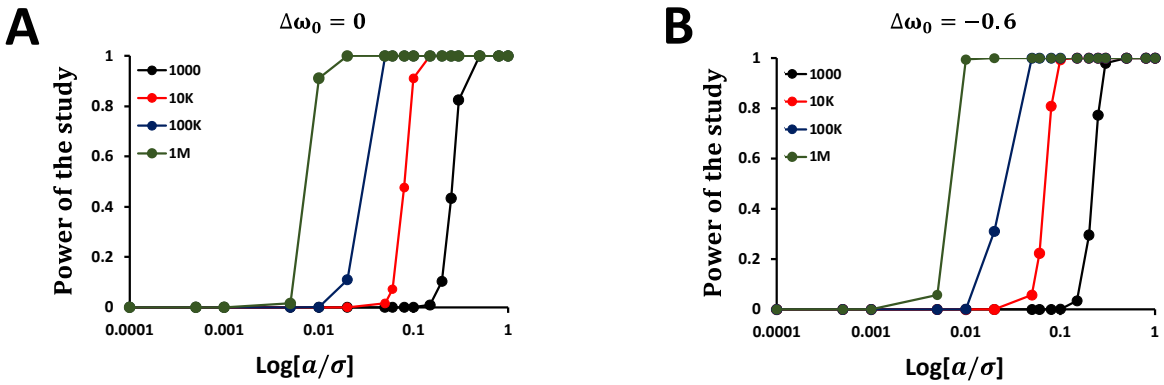

A
$\Delta\omega_0 = 0$
Power of the study
1000
10K
100K
1M
0
0.2
0.4
0.6
0.8
1
0.0001
0.001
0.01
0.1
1
Log[a/σ]
B
$\Delta\omega_0 = -0.6$
Power of the study
1000
10K
100K
1M
0
0.2
0.4
0.6
0.8
1
0.0001
0.001
0.01
0.1
1
Log[a/σ]

Figure 4

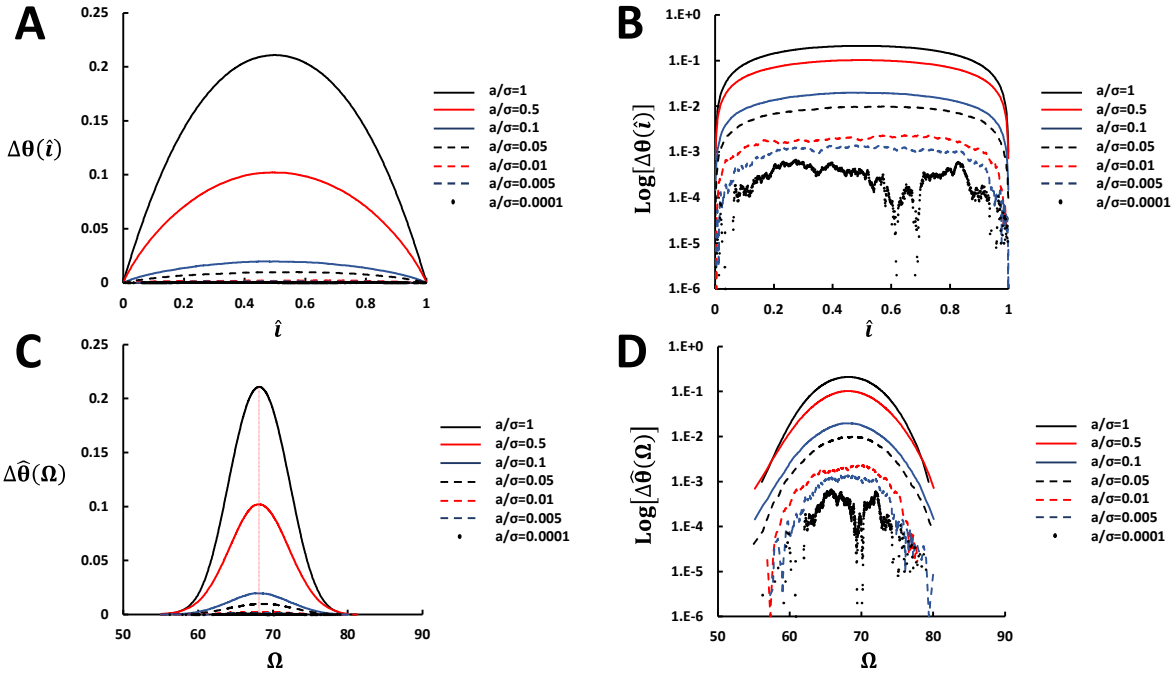

A
B
C
D
$\Delta\theta(\hat{\imath})$
$\mathrm{Log}[\Delta\theta(\hat{\imath})]$
$\Delta\hat{\theta}(\Omega)$
$\mathrm{Log}[\Delta\hat{\theta}(\Omega)]$
$\hat{\imath}$
Ω
a/σ=1
a/σ=0.5
a/σ=0.1
a/σ=0.05
a/σ=0.01
a/σ=0.005
a/σ=0.0001

Figure 5

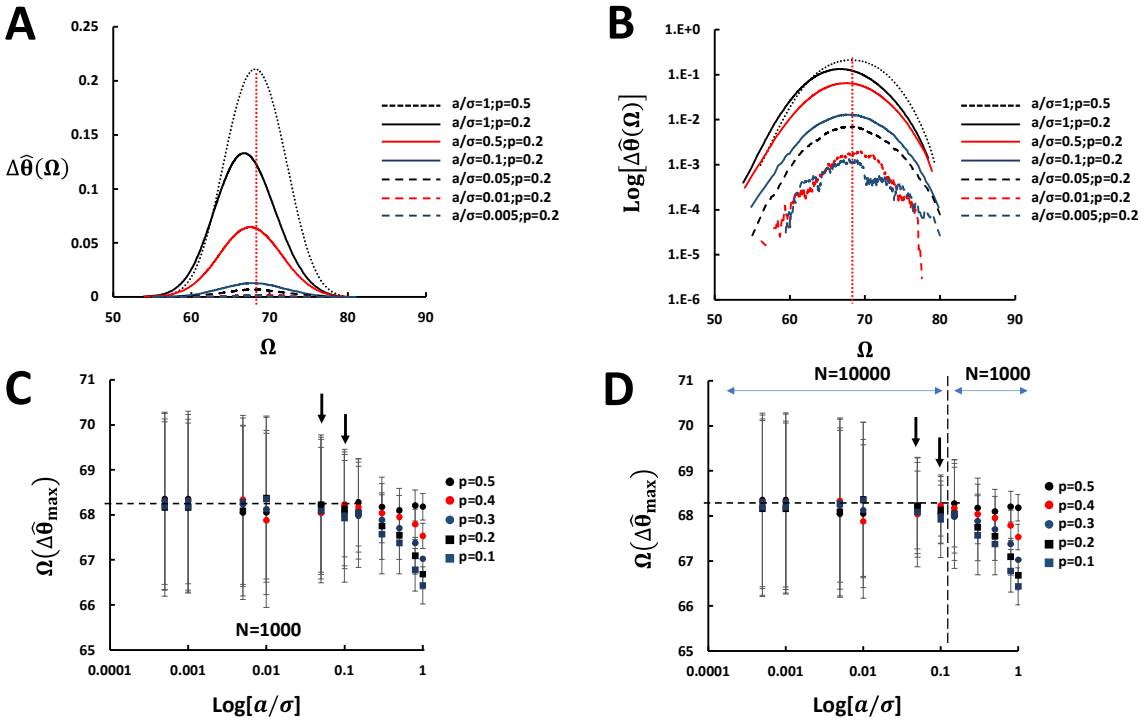
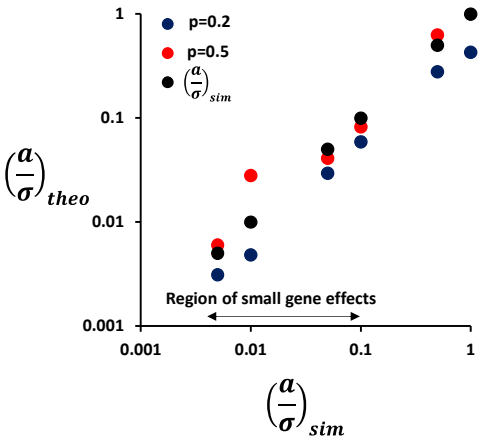

A
B
C
D
$\Delta\hat{\theta}(\Omega)$
$\mathrm{Log}[\Delta\hat{\theta}(\Omega)]$
Ω
a/σ=1;p=0.5
a/σ=1;p=0.2
a/σ=0.5;p=0.2
a/σ=0.1;p=0.2
a/σ=0.05;p=0.2
a/σ=0.01;p=0.2
a/σ=0.005;p=0.2
$\Omega(\Delta\hat{\theta}_{max})$
N=1000
N=10000
p=0.5
p=0.4
p=0.3
p=0.2
p=0.1
$\mathrm{Log}[a/\sigma]$

Figure 6


p=0.2
p=0.5
$\left(\frac{a}{\sigma}\right)_{sim}$
1
0.1
0.01
0.001
$\left(\frac{a}{\sigma}\right)_{theo}$
Region of small gene effects
0.001
0.01
0.1
1
$\left(\frac{a}{\sigma}\right)_{sim}$