SAPIENZA
UNIVERSITÀ DI ROMA

Methods and Techniques

Vol. 6, No. 1 (2023)
ISSN: 2532-5876
Open access journal licensed under CC-BY
DOI: 10.13133/2532-5876/17811

On the Meaning of Averages in Genome-wide Association Studies: What Should Come Next?

Cyril Rauch,[a] Jonathan Wattis,[b] Sian Bray [c]*

[a] *University of Nottingham, School of Veterinary Medicine and Science, College Road, Sutton Bonington, LE12 5RD, UK*
[b] *University of Nottingham, School of Mathematical Sciences, University Park, NG7 5RD, UK*
[c] *University of Nottingham, School of Life Sciences, University Park, NG7 5RD, UK*

***Corresponding author:** Cyril Rauch, Email: cyril.rauch@nottingham.ac.uk

Abstract

Identifying the association between phenotypes and genotypes is the fundamental basis of genetic analyses. Although genomic technologies used to generate data have rapidly advanced within the last 20 years, the statistical models used in genome-wide associations studies (GWAS) to analyze these data are still predominantly based on the model developed by Fisher more than 100 years ago. The question is, does Fisher's theory need to be replaced or improved, and if so, what should come next? The theory developed by Fisher was inspired by the field of probability. To make use of probability not only did Fisher have to assume valid a number of questionable hypotheses, but he also had to conceptually frame genotype-phenotype associations in a specific way giving primordial importance to the notion of average. However, the "average" in probability results from the notions of "imprecision" or "ignorance". After reviewing the historical emergence and societal impact of probability as a method, it is clear what is needed now is a new method acknowledging precision in measurements. That is, a method that does not rely on categorizing or binning data.

Keywords: phenotype-genotype mapping, method of averages, GWAS, infinite population, normal distribution

Citation: Rauch, C, Wattis, J, & Bray, S 2023, "On the Meaning of Averages in Genome-wide Association Studies: What Should Come Next?'", *Organisms: Journal of Biological Sciences*, vol. 6, no. 1, pp. 7 –22. DOI: 10.13133/2532-5876/17811

Introduction

Genome-wide association studies (GWAS) based on statistics have had a huge impact on the field of genetics by providing a method to map genotypes (DNA variants) and continuous phenotypes, namely the observable characteristics of an organism varying in a continuous way. GWAS has in turn facilitated the understanding of biology, the development of new therapeutics in medicine and the improvement of agricultural species (Visscher *et al*. 2017). Statistical models describing the relationship between genotype and phenotype were first developed by R.A. Fisher more than 100 years ago and remain a cornerstone of genotype-phenotype mapping today (Fisher 1919; 1923). However, ongoing debates exist in this field related to the validity of Fisher's theory (Nelson, Pettersson, & Carlborg 2013; Visscher & Goddard 2019), in turn, raising questions regarding the current paradigm in quantitative genetics.

Controversies exist in sciences because a given theory is not supported by a subset of observations or is limited in its ability to provide information.

Controversy can act as an engine of progress; resulting in the generation of new developments that better describe or enable better description of reality. Such new ideas are not necessarily radical, *i.e.* do not negate the seminal idea, but come as a generalization of seminal concepts. In this context, the best and probably most notable example is the transition that occurred in physics between Newton and Einstein regarding the notions of space and time.

However, because "scientists" are also immersed in a social culture, new ideas rarely come out the blue, but result from a specific construction of knowledge that is, to some extent, biased by the society in which they exist. That is to say, to understand the true meaning of seminal ideas, it is also strongly advised to be cognizant that they are in part a product of their time.

Scientifically speaking, GWAS are situated at a junction between genetics, statistics, and probability. Genetics is a field of knowledge that has been studied in depth both scientifically, epistemologically and sociologically by many renowned authors (Boichard *et al.* 2016; Gayon 2016; Prunet & Meyerowitz 2016; Quintana-Murci 2016; Schacherer 2016; Weissenbach 2016), and there would be very little gain to add more to these works. Likewise, the history of statistics is a field that has been covered by many authors and in particular by S.M. Stigler in his remarkable book (Stigler 1990). On the contrary, the field of probability and its repercussion on GWAS and statistics, both scientifically, epistemologically, and sociologically is less well known. In fact, most graduate students in quantitative genetics who tend to be remarkably good at using statistics, would find very difficult to dissociate statistics from probability. Indeed, they will know and use the normal distribution or similar probability density functions to substantiate their inference(s) but only a few, if any, will wonder what the limits regarding the use of such distributions are and where they come from. Students are not to blame for this since the blending of statistics and probability virtually exists in all books dealing with population biology or quantitative genetics. For example, if one were to ask oneself "what is a phenotype?" and then look into books to get an answer, one will rapidly find that the notion of phenotype is represented exclusively as a probability density function. Why this is the case is linked to the rise of probability in the field of biology.

The bond between statistics and probability has permeated virtually all fields of biology to the point where the coupling of "statistics and probability" is now a biological reality, *i.e.* not a thought construction or a method anymore. An example of such widespread and subconscious use of probability concerns the notion "significance". From cell biology to population genetics, any result is deemed scientifically adequate, *i.e.* significant, provided that its p-value falls within agreed limits. Whilst this approach is mathematically sound, it also includes a number of assumptions without consideration of the restrictions they impose. In this context, it is important to recall that probability density functions originate through the notion of "imprecision" or "error". The normal distribution was known originally as the "error function" or "law of errors". The error function states that if an experiment can be repeated *identically to itself an infinite number of times in identical circumstances*, then, provided that the outcome of experiments are numerical data, the distribution of those data should follow the error function (the normal distribution). In essence, the error function: (i) justifies why experimental results are not identical, even though they arise from repeated and identical experiments, and (ii) tells us that the average is the numerical value of the thing that was meant to be measured.

Whilst one is free to use the field of probability to extract any result from measurements, the use of a probability density function such as the normal distribution imposes that the object studied must be conceptualized in a certain way. Perhaps one of the most important aspect as far as genotype-phenotype mapping and biology are concerned, involves the fact that all individuals are considered "identical" entities. To what extent two individuals in a population are identical is open to question. A nonetheless important aspect is the notion of "infinity". Interpolating a histogram representing the frequency of occurrence of categories of phenotype values using the formula of the normal distribution requires this notion of the continuum limit to be valid. However, to what extent the notion of "infinity" is granted in any field of sciences is rarely discussed. It is worth recalling that to justify his theory Fisher had to use the "infinite population" hypothesis, which we know, is unrealistic, if not impossible.

This opinion paper is not about questioning the entire field of probability, but to indicate the shortcomings when using probability as a tool to conceptualize the relationship between genotype and phenotype. This will show that the rise and societal importance of the notion of "average" in genotype-phenotype mapping came, historically, from the field of "biometry" in a dark period of our civilization marked by the predominance of eugenics theory; and that using probability as a mathematical field to substantiate any such premise was, simply, based on wrong assumptions. Although eugenics thoughts have been relegated to history, the predominance of the notion of "average" resulting from our initial belief in the normal distribution is still very present in our society. In fact, the rise of the normal distribution as well as its impact on our society has been defined as "biopolitics" and is now an entirely dedicated research field in sociology or philosophy (Rose 2001). Whilst conceptualizing the average is, in itself, not the real issue, it is its connection with something that ought to be normal, *i.e.* the normal distribution, that poses problem; as the tendency is to think that any value that is not average is linked to some randomness or error, *i.e.* is a nuisance.

We shall see that this reflection opens the way to different concepts to provide accurate information on genotype-phenotype mapping that are not based on the notion of "error".

1. Statistics

At this point, it is important to recall what statistics is, at least for the sake of students. Statistics comes from the Latin *statisticium*, which refers to "the state of things" and is borne out from the need to order observations and represent those in the form of tables and graphs involving specific parameters summarizing the information contained in the data. Historically, collecting data outdates, by millennia, the field of probability and the reason is simple. Estimating the power of any chief of state, or similar, relies on good knowledge of characteristics related to population, military potential, wealth and so on. That is, governance and authority rely on data. Whilst Mesopotamians left traces of such activity in the form of tables of data made in clay dating back more than 6000 years (Droesbeke & Tassi 1990), the field of statistics as we know it today was reinvented through the rise of probability when scientists were trying to make sense of disparate data accumulated from scientific measurements. The need to determine as exactly as possible the "true" outcome or result of a set of scientific observations relied on understanding the notion of measurement errors, and it is the estimation of such errors that led to the collision between, and fusion of, the fields of statistics and probability.

To summarize, one can say that 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, analyzing, and interpreting data; whilst probability, defined through relative frequencies, is central to determining the validity of statistical inferences.

2. Probability

In early 20th century, the intertwined fields of statistics and probability had grown up to reach almost full maturity. Both fields arose through one of the greatest journeys of the human mind, trying to decipher the notion of evidence, *i.e.* what is provable, and provide this evidence in an interpretation to determine reality. Renowned authors in the field of probability agree that this field started with Jacob Bernoulli's (1654–1705) definition. Namely, that the probability of an event is the ratio of one outcome compared to all possible outcomes; defined by Bernoulli as,

> "that a particular thing will occur or not occur in the future as many times as it has been observed, in similar circumstances, to have occurred or not occurred in the past" (Stigler 1990, p. 65).

In Bernoulli's definition, the probability represents a degree of certainty that can only be described a posteriori using the frequency of occurrence of the "thing". Beyond characterizing a degree of certainty, this definition also encompasses indirectly a certain notion of "immanence" as the "thing" can be characterized by its reappearance. Indeed, the ratio of a specific outcome to all possible outcomes is "expected" to reoccur provided similar contexts are possible.

"Immanence" and "expectation" are interesting concepts when applied to sciences as they imply a certain degree of stability or invariance that may result from the presence of laws. However, a line should be drawn here between the notions of probability and scientific law, as a degree of certainty is by no way a proof or a demonstration. "Proof" or "demonstration" involve an articulation, *i.e.* a causality, between elements leading to the "particular thing" to be observed. Consequently, the "thing" is only secondary to this articulation, as it is this articulation that provides a conceptual understanding of its occurrence and notably its reason of being. Therefore, with such an articulation or causality leading to the "thing", the "thing" is necessarily defined as an evident *a priori* resulting from the scientific law.

A different way to phrase this is to say that averages and variances can always be defined in any population of data. The point, though, concerns their scientific meanings or pertinences. Pierre-Simon Laplace (1749–1827) gave the example of the Sun rising every morning and the time at which this occurs. Whilst regularity would be found in the data it would not inherently inform one of gravitational laws (Laplace 1995). That is to say that whilst a scientific law fits Jacob Bernoulli's definition of the probability, the converse is not necessarily true. Consequently, there is a vast conceptual difference between "empirical" and "mechanistic" sciences.

In his unfinished work *Ars Conjectandi* published eight years after his death, Jacob Bernoulli provided, thanks to his measure of probability, the weak law of large numbers (Todhunter 2014, pp. 56–77; Stigler 1990, pp. 63–98). This law was refined by Abraham de Moivre (1667–1754) (de Moivre 2013; Stigler 1990, pp. 63–98) providing a proof that if an observable is "expected" to occur with a defined degree of certainty, it must follow what we call today the Bernoulli distribution.

To avoid confusion a precision is required concerning the works by de Moivre and Gauss. De Moivre was interested in the probability of winning a game. When playing with cards for example, the entire set of outcomes can be determined as the set of cards is known and given from the start. This is different than trying to determine the "true" outcome from a set of data since the entire set of possible outcomes is unknown and given only as observational measurements. This point has led to some controversies as to who discovered the "normal distribution" first between C.F. Gauss (1777–1853) and Moivre as Gauss was interested in observational measurements (not games). The point however is that both manage to deduce the mathematical form of the Normal distribution in different ways.

In short, what Moivre demonstrated is a version of what, today, we call the central limit theorem. Moivre's theorem stipulates that if it is possible to make a very large number of independent measurements of the same "thing" in similar contexts, then a specific distribution of that "thing" would ensue. This was the first mathematical description of what would become the normal distribution with the "thing" being the expectation, *i.e.* average, with a variance inversely proportional to the number of measurements made. In essence, by doing a very large (infinite) number of measurements one would amplify and make visible the "thing" to be observed.

To Abraham de Moivre, this distribution demonstrated the intervention of God in which the "thing" was just awaiting to be discovered and measured, namely the "thing" had to have a fundamental meaning. His work was supposed

> "to cure a kind of superstition, which has been of long standing in the world, that there is … such a thing as Luck, good or bad" (Moivre 2013, p. 4 of the 1718 preface, 1st edition).

This way of thinking had to have a profound repercussion in different fields from biology to sociology. Indeed, this vision propelled the method of relative frequency, and therefore the normal distribution including its ontological parameters that are average and variance, as a reliable estimation of the *a priori* unknown probability. In short, the normal distribution had to happen since it provided the degree of certainty of the phenomenon observed. This, in turn, may explain why the notion of "infinite population" was used by Fisher as an attempt to promulgate scientific laws.

Whilst Thomas Bayes (1702–1761) and Pierre-Simon Laplace later demonstrated the weakness of the *a priori* argument as developed by de Moivre (see Appendix), the idea that the normal distribution was a fundamental trait of life was nonetheless accepted. The general acknowledgement of such trait of life was emphasized, for example, by Adolphe Jacques Quetelet (1796–1874) and his belief in the "average man" or the "social physics" (Porter 1985); or by Francis Galton's (1822–1911) narrative describing the "human ability" as a heritable trait (Galton 1886). As much as we know today that those sort of beliefs are strongly limited (wrong) since they exclude the socialization of individuals; it is important to recall that Quetelet and Galton were, during their times, trying to "improve" society and can be regarded as some sort of "sociologists"— missing an adequate term that could allude the notion of "past or outdated sociology"—that were the product of their times (Wright 2009).

To justify this statement, it is important here to recast the sociological impact that the field of probability has had on our society. Indeed, with the normal distribution being a fundamental trait of life, thinking or solving problems in term of probability by using the method of relative frequency was essential. In fact, with this method, it was, at least in theory, possible to forecast any event (*e.g.* being killed in the street; dying of a disease; being wrongly judged by a barrister, etc.) (Laplace 1995; Samueli & Boudenot 2009). The point to be emphasized here is that the field of probability has been used as a "scientific justification" of a "general biometry" whereby a set of people/individuals were, and still are, modelled as a "population". As an example, Quetelet believed that one ought to investigate the "social body" and not the "peculiarities distinguishing the individuals composing it" (Faerstein & Winkelstein 2012). This way of framing individuals at the end of the 18th century allowed a shift in judicial and social policies in which the "social body", that is the distribution density function of any population of measurements and its properties (averages and variances), formed the core of what needed to be understood and controlled (Rose 2001). Thus, the singular identity of individuals disappeared into the "social body", and the "social body" became then a tool to process the identification of individuals. It is therefore not surprising that during the same period the "judicial anthropometry" emerged, whereby arrested individuals were measured to construct database aiming at identifying potential criminals in the society (García Ferrari & Galeano 2016). Likewise, it is not surprising that how different phenotypes can be, they are represented by distribution density functions today.

Given that the field of probability and its consequences, *i.e.* mean and variance, were "in the air" at that time, Fisher's theory, in which the notion of "average" is central was sociologically accepted by its contemporary society. Thus, the "infinite population" hypothesis that Fisher had to put forward to explain why genotype-phenotype can associate did not carry much doubt, how questionable it was.

More than 100 years later one can now try to think about those shortcomings.

3. Shortcomings of Genotype-Phenotype Mappings Using the Error Function

Whilst the notion of distribution opened the way to data analysis, the validity of the central limit theorem, *i.e.* the normal distribution, comes with some ties.

The first of which relates to the notions linked to the "thing" and "similar contexts", that is, the "thing" being measured as well as the context in which the "thing" is measured must be identical. The second tie resides in the utilization of "infinity", or the notion that a large number of experiments needs to be made for "God's will to be visible".

Those two ties are clear constraints concerning the use probabilities and as such are worth developing in the context of genotype-phenotype mapping since they will allow one to understand how the human mind has conceptually framed this field.

3.1. Identity and Probability

The first tie is a fundamental constraint as it reposes on a clear understanding of what identity is. One may say that one individual, say Paul, is identical to himself and that he defines his own context; but saying that Paul and Jacques can be considered as identical is a step that goes beyond any assumption defining a probability when biology is considered. Let us be precise. One can decide to measure Paul's height a large number of times and define a probability since the "thing" to measure and its "context" are always Paul's height and Paul, respectively. Accordingly, one would deduce Paul's average height and some standard deviation linked to some measurement errors. So, as much as a phenotype like height may be universally defined for human beings the identities between the individuals forming the population are different if two different individuals are measured, *i.e.* Paul is not Jacques. Naturally, nothing forbids us from determining the distribution of the phenotype height for each individual separately to reform the distribution of heights in a population. If so, one would then define the distribution of heights but not the distribution of individual heights since the identities of Paul and Jacques would be dissolved into the former distribution. This remark underlines that using the distribution of phenotypes in a population is equivalent to dissolving contexts, in this case, identities. Accordingly, with the distribution of heights one is left with the few moments of the distribution, *i.e.* average, variance, skewness and so on, providing a very short summary of the diversity and identity of individuals. Deciding to consider a phenotype distribution, despite the definition of probability, is then equivalent to consider Paul and Jacques as meaningless envelops of something more important that would spread across the population. Clearly, that "thing" awaiting to be observed or measured are genes (or Mendelian factors) and their effects, and the distribution of any phenotype would result from a condensate of independent genes without envelop/identity limiting them. The notion of "condensate" is historically important as R.A. Fisher was influenced by physics and most particularly in how statistical physics managed to relate the microscopic and macroscopic properties of ideal gases (Fisher 1923; Morrison 1997). One can then understand R.A. Fisher's method as a way to define each gene microstate across the population distribution as being a particular gas molecule with given property. The sum of genes including their properties would then define the moments of the phenotype distribution, *i.e.* average value and variance for example.

However, if GWAS is used to determine genotype-phenotype relationships, then there is an apparent problem when the "method of averages" as advocated by R.A. Fisher does not recover the average and variance of the phenotype. In this case, the notion of "environment" is added to complete the phenotype distribution. Despite the fact that the environment is in general ill-defined, it is added with the implicit intention to recover the phenotype distribution, *i.e.* to complete the faith in the normal distribution. What is puzzling with such intention is that one knows that for each phenotype measured, namely each individual, corresponds to genes in specific states and one may wonder whether dissolving individuals into a phenotype distribution, and assuming its universal relevance, does not lead to more complications.

Let us frame this in the context of frequentist probability as used in GWAS. We said earlier that the normal distribution was known as the "error" function. 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 centimeter 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 1 cm-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 (Macdonald, Hawkes, & Corrigan 2021). 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?

In general, the solution to this problem is to increase the population size to sample, such as to recreate bins or categories matching the available precision. How strange that, whilst precision is fully available in the first place, the method advocating phenotype categories, *i.e.* creating a sort of wilful ignorance regarding phenotype values, is still suggested. Again, this is linked the hundredth-plus years old faith in the normal distribution, *i.e.* the error function and the importance we ought to give to the notions of average and variance.

3.2. Infinity and Probability

Let us now explore the second tie, namely the "infinite population" hypothesis. This hypothesis is fascinating as its attempt is to reconcile genotype-phenotype mapping, *i.e.* GWAS, with the field of probability. It is mathematically true that if one were able to repeat the same experiment an infinite number of times one would be entitled to use the normal distribution in the continuum limit as an *a priori*, and use its full mathematical expression. One may then question to what extent is the notion of "infinity" granted in any field of knowledge. As an example, the normal distribution is inherent to some branches of physics and no one with a background in physics would question its usage and validity.

One may then argue that if physics is allowed to use the normal distribution and deduces average(s), then why would this be an issue for quantitative genetics?

The answer to this question lies in the very definition of what physics and biology try to address as sciences. If physics defines average(s), *i.e.* can conceptually consider the normal distribution a.k.a. error function leading to the Dirac distribution as an asymptotic limit when no other parameter (such as the thermal energy or the Planck constant) constrain this limit, it is because physics aims to uncover the intemporal Laws of Nature. It is this notion of intemporal Laws that underscores the notion of infinity or immanence or potential repeatability of experiments in physics. This warrants the use of the central limit theorem.

On the contrary, life is driven by evolution, *i.e.* changes in average(s). Thus, life's average(s) are not absolute but function of time and their history, *i.e.* are not immanent(s) but contingent(s). Time and history are fundamental conceptual parameters for understanding life.

To conclude, whilst the normal distribution can be a useful representation of data, the conclusions drawn, need to be mindful of the underlying conceptualization of the system studied as well as the scientific interpretations underscored. As a result, the normal distribution and its ontological parameters, *i.e.* average and variance need to be handled very carefully. Indeed, distribution density functions can always be derived for any process when data points, *i.e.* numbers, form the outcome of this process. That is to say that because distribution density functions in biology can always be derived, average and variance are not necessarily scientifically pertinent parameters.

3.3. What is the Option, What Comes Next Beyond the Binning or Categorization of Phenotype Values?

Controversy exists in the field of genotype-phenotype associations (Nelson, Pettersson, & Carlborg 2013). Attempts are being made to ameliorate inferences drawn by GWAS. For example, Bayesian models have been used to enhance any potential evidence of genotype-phenotype relationships (Beaumont & Rannala 2004). Whilst Bayesian and Fisher models are conceptually different since they envision the notion of probability and therefore, evidence, differently, they both rely on the concept of an *a priori* in different ways. For Fisher's model it is the importance of moments and in particular the notion of average and variance, namely the normal distribution; and for Bayes, the need to use *a priori* "information" whose exact formulation is either difficult to obtain (or unattainable in most cases). Whilst those two models are conceptually different, they both use the notion of probability in a specific way by defining probability density functions. However, using probability density functions is the central issue at hand.

Indeed, the notion of "imprecision" or "error" defines the concept of density that in turn, form the core of distribution density functions that lead to the definition of average and variance (other moments can be included if needed). However, binning or categorizing data to create density is equivalent to loosing information (wilful ignorance). At the dawn of the 21st century, we are getting more precise in our measurements, and one may wonder what sort of scientific/mathematical tool we should be using if one were able to attain any level of precision wanted. Whilst this sounds a bit idealistic and, perhaps, unattainable, it is worth recalling that not long ago physicists were able to measure remarkably small gravitational waves (Abbott *et al.* 2016) that were deemed out of reach a century ago.

The questions are then: how can genotype-phenotype mapping be possible without losing information? What method should we employ when there is no randomness in the data? Or said differently, how can we re-integrate the identity and diversity of individuals within genotype phenotype mapping such as to re-transform a population into a set of individuals?

Such an enterprise means that the notion of average and variance must disappear from any association study, since it is the grouping of data into categories that generate those.

4. From the Method of Averages to a Method Based on Curves: Genomic Informational Field Theory (GIFT)

As stated, current genome-wide association studies rely on the consequences of using probability density functions in the continuum limit. That is, on the belief that the average and variance (and all the other moments of higher order if needed) are meaningful. However, other models can be suggested that do not require the grouping of data.

4.1. GIFT as a Method to Determine Genotype-Phenotype Mappings

GIFT is a method whose aim is to extract information from datasets without requiring the binning or categorization of data from which the notions of average and variance are ontological parameters when the method of relative frequencies is used. One way to position the problem is, therefore, to address how information can be extracted when phenotype and genotype are measured precisely enough such as to rule out the need of categories.

To answer this question, the best is to look at the impact of the notion of precision on data representations when one moves from imprecisions to precise measurements. Figure 1 demonstrates in the context of genome-wide association studies the impact of increasing the precision in phenotype measurements when a population has a finite size. The total number of individuals is 1000 in this case.

The conclusion is obvious: distribution density functions such as the normal distributions representing genotype and phenotype disappear. Instead, a set of code bars emerges. Those code bars are the individuals, *i.e.* people, forming the population. As a result, it is the structure of these code bars, namely their arrangement, that needs to be understood. Whilst, both color and spacing between the bars/individuals are important information since they are reminiscent from the use of the normal distributions to model genotype and phenotype initially, they are now two variables that were combined, or convolved, when the method of relative frequencies, *i.e.* normal distributions, were used.

To extract information from the code bars, let us now wonder what it means to have information on the phenotype as opposed to have none.

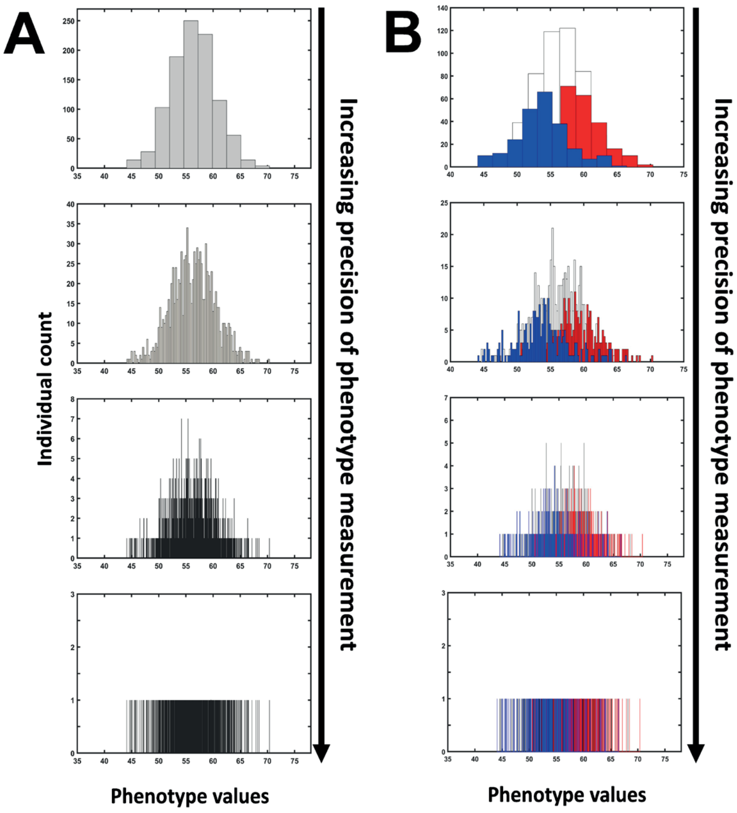

Figure 1: When applied to real data sets, current genome-wide association studies rely on probability distribution density functions (PDFs), namely the creation of frequency plots (method of relative frequencies) via the grouping of phenotype values into categories representing range of phenotype values (A, top-chart). The same method (PDFs) is then applied to genotypes (B, top-chart). 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 (blue), Aa (white) and AA (red), 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. 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”?

To answer this question the best thing is to further simplify the problem by considering the colored bars only and not their spacing. Imagine, therefore, that a set of individuals has been 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 1000000 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 therefore be represented as a random string of genotypes including "+1", "0" and "-1" microstates (representing homozyte-AA, heterozygote-Aa and homozygote-aa). An example of such random configuration is:

[0, +1, 0, -1, -1, +1,0, …, -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 1000000 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 1). An example of such a string would be:

[+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 1000000 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, then differences will be seen in the resulting cumulative sums. To give an example, let us consider the two strings above and note "$\theta_o(j)$" and "$\theta(j)$" the cumulative sums of microstates in the random and ordered configurations where "j" is the position in the string. Then adding the microstates starting from the left side of the strings one finds:

$\theta_o(j = 1) = 0 = 0$
$\theta_o(j = 2) = 0+1 = +1$
$\theta_o(j = 3) = 0+1+0 = +1$
.....

$\theta(j = 1) = +1 = +1$
$\theta(j = 2) = +1+1 = +2$
$\theta(j = 3) = +1+1+1 = +3$
....

As a result, the difference "$\theta(j)-\theta_o(j)$" is expected to be indicative of the importance of the phenotypic information. The fact that the same individuals were considered in both configurations also impose a conservation relation under the form: $\theta(N)-\theta_o(N) = 0$.

4.2. The notion of phenotypic fields

It is then possible to interpret the information on the phenotype as a field acting differently on microstates (Rauch *et al*. 2022; Wattis *et al*. 2022). The notion of phenotypic field is a natural concept since it is the information on the phenotype that promotes a migration of microstates, and as a result imposes a change between the two aforementioned configurations. To some extent, the different microstates "respond" differently to the phenotypic information and physics fields theory can be applied on this closed system. Closed system means that the individuals are the same between the two configurations.

Consider that there are “N_+”, “N_o” and “N_-” genetic microstates “+1”, “0” and “-1”, respectively, it follows that when the random configuration is considered, at any position in the string the probability of finding either “+1”, “0” or “-1”, is simply: $\omega_+^0 = N_+/N$, $\omega_+^0 = N_o/N$ and $\omega_o^0 = N_-/N$. That is to say that when no information on the phenotype is available the presence probability of microstates can be derived relatively simply. Accordingly, the cumulative sum of microstates in the random configuration, θ_o, is simply

$$\theta_o(j) = \sum_{x=1}^{J}(+1)\cdot\omega_+^0 + (0)\cdot\omega_o^0 + (-1)\cdot\omega_-^0 = \sum_{x=1}^{J}(\omega_+^0 - \omega_-^0)$$

One notes here that the difference “ω_+^0 - ω_-^0” can also be rewritten as

$$\omega_+^0 - \omega_-^0 = \frac{N_+ - N_-}{N} = \frac{N_+ - N_-}{N_+ + N_0 + N_-} = \frac{\frac{(N_+ - N_-)}{N}}{\frac{(N_+ + N_0 + N_-)}{N}} = \frac{\omega_+^0 - \omega_-^0}{\omega_+^0 + \omega_0^0 + \omega_-^0}$$

For the second configuration, one can then deploy physics’ arsenal and it is then possible to write (Rauch *et al.* 2022; Wattis *et al.* 2022) the presence probabilities of microstates “+1”, “0” and “-1” at any position j = 1,…,N in the string as a function of the fields under the form

$$\omega_+(j) = \frac{\omega_+^0 e^{u_+(j)}}{\omega_+^0 e^{u_+(j)} + \omega_0^0 e^{u_0(j)} + \omega_-^0 e^{u_-(j)}}$$

$$\omega_0(j) = \frac{\omega_0^0 e^{u_0(j)}}{\omega_+^0 e^{u_+(j)} + \omega_0^0 e^{u_0(j)} + \omega_-^0 e^{u_-(j)}}$$

$$\omega_-(j) = \frac{\omega_-^0 e^{u_-(j)}}{\omega_+^0 e^{u_+(j)} + \omega_0^0 e^{u_0(j)} + \omega_-^0 e^{u_-(j)}}$$

Where “$u_+(j)$”, “$u_o(j)$” and “$u_-(j)$” are field functions to be defined representing the impact of the information on the phenotype on microstates “+1”, “0” and “-1”, respectively. The latter formulae are similar to “Laplace’s formula” (Box 1). When non-null, those fields guarantee a change in configurations. The second cumulative sum is then

$$\theta(j) = \sum_{x=1}^{J}(+1)\cdot\omega_+(x) + (0)\cdot\omega_o(x) + (-1)\cdot\omega_-(x) = \sum_{x=1}^{J}(\omega_+(x) - \omega_-(x))$$

As a result, the difference in the cumulative sums can be expressed as

$$\theta(j) - \theta_0(j) = \sum_{x=1}^{j}\left[\frac{\omega_+^0 e^{u_+(x)} - \omega_-^0 e^{u_-(x)}}{\omega_+^0 e^{u_+(x)} + \omega_0^0 e^{u_0(x)} + \omega_-^0 e^{u_-(x)}} - \frac{\omega_+^0 - \omega_-^0}{\omega_+^0 + \omega_0^0 + \omega_-^0}\right]$$

One deduces with this development that if the genome position 1000000 on chromosome 4 does not participate to the formation of the phenotype, *i.e.* when the fields are null, then one can set: $\theta(j)$-$\theta_o(j)$~0. That is, having no information on the phenotype is similar to an absence of genotype-phenotype association.

Finally, the conservation relation that is, $\theta(N)$-$\theta_o(N)$ = 0, is written as

$$\sum_{x=1}^{N}\left[\frac{\omega_+^0 e^{u_+(x)} - \omega_-^0 e^{u_-(x)}}{\omega_+^0 e^{u_+(x)} + \omega_0^0 e^{u_0(x)} + \omega_-^0 e^{u_-(x)}}\right] = N\frac{\omega_+^0 - \omega_-^0}{\omega_+^0 + \omega_0^0 + \omega_-^0}$$

4.3. Conceptual Consequences of GIFT: Genotype-Phenotype “Loop”

At the conceptual level, what has been done is intuitive and relatively simple. However, 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 “$\theta(j)$-$\theta_o(j)$” one can say that it is the phenotype, *i.e.* phenotypic field or information, that organizes 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, which 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 phenotype selects gene microstates (present view), then an equivalence exists between phenotype and genotype.

Stepping further in that direction one can also say that the difference “$\theta(j)$-$\theta_o(j)$” resulting from a change in microstates configuration is a decomposition of the phenotype in the genetic space.

Let us call “$\theta(j)$-$\theta_o(j)$” as the “genetic paths difference” of genome position 1000000 on chromosome 4, one way to capture the conceptual importance of this “loop” is to say that whilst a gene is “Darwinian”, the genetic paths difference is “Lamarckian” since the phenotype selects the set of microstates it needs to subsist. With GIFT those two visions (Darwin vs. Lamarck) are not mutually exclusive and as it turns out, Fisher’s theory does not disagree with this viewpoint either since GIFT can be transformed to “classic” GWAS provided categories are considered.

5. From GIFT to Fisher’s Theory by a Coarse-graining Process

GIFT is a method advocated when phenotype values are unique while traditional GWAS consider categories for the phenotype values. The correspondence between GWAS and GIFT can be determined provided artificial categories are created such as to lose information on the phenotype.

Let us consider the presence probability of the microstate “q” at the position “j” in the string, where “q” replaces the signs “+”, “o” or “-” to allow for succinct notations. This probability is formally written as

$\omega_q(j) = \omega_q^o e^{u_q(j)}$

Note that the denominator given by,

$\omega_+^o\, e^{u_+(j)}+\omega_o^o\, e^{u_o(j)}+\omega_-^o\, e^{u_-(j)}$,

is equal to one as by definition any position can either be a “+”, “o” or “-” microstate.

Consider now an interval of individual positions of width “Δj” centred around “j” and define by “ΔN_q” the number of microstates of type “q” in this interval. One can then determine the average number of microstates of type “q” in this interval under the form “$\Delta N_q/\Delta j$”. As it turns out, “$\Delta N_q/\Delta j$” is also the presence probability of microstate of type “q” in this interval.

Consequently, “$\Delta N_q/\Delta j$” can also be written as

$$\frac{\Delta N_q}{\Delta j} = \sum_{x=j-\Delta j/2}^{j+\Delta j/2} \omega_q(x) = \omega_q^0 e^{u_q(j)} \sum_{x=j-\Delta j/2}^{j+\Delta j/2} e^{u_q(x)-u_q(j)}$$

The discreet sum can be transformed into a continuous sum under the form:

$$\sum_{x=j-\Delta j/2}^{j+\Delta j/2} e^{u_q(x)-u_q(j)} \rightarrow \int_{j-\Delta j/2}^{j+\Delta j/2} e^{u_q(x)-u_q(j)}\, dx$$

Where “dx”, is defined as being the difference between two consecutive positions, that is the difference between the positions “x” and “x-1”.

As GWAS involves the phenotype values, the previous relation must be amended to provide the correspondence between GWAS and GIFT.

Noting “Ω_x” the phenotype value at the position “x”, one defines then the difference between two consecutive phenotype values as: $d\Omega_x = \Omega_x - \Omega_{x-1} \sim \lambda(\Omega_x)\, dx$. In this context “$\lambda(\Omega_x)$” is the rate of changes in phenotype values between two positions. Therefore, the difference between two positions “x” and “x-1”, that is “dx”, can be related to the difference of the two consecutive phenotype values at those positions under the form, $d\Omega_x/\lambda(\Omega_x) \sim dx$. Accordingly, the expression involving the integral can be transformed as follows

$$\omega_q^0 e^{u_q(j)} \int_{j-\Delta j/2}^{j+\Delta j/2} e^{u_q(x)-u_q(j)}\, dx$$
$$\rightarrow \omega_q^0 e^{\hat{u}_q(\Omega_j)} \int_{\Omega_j+\Omega_{\Delta j/2}}^{\Omega_j+\Omega_{\Delta j/2}} e^{\hat{u}_q(\Omega_x)-\hat{u}_q(\Omega_j)} \frac{d\Omega_x}{\lambda(\Omega_x)}$$

Where the hat on the field is added to inform that the field is now expressed in the space of phenotype values. Additionally, one can also drop the subscripts involving the position by re-writing “Ω_j” and “$\Omega_{\Delta j/2}$” as “Ω”and “$\Delta\Omega/2$”, respectively.

The two terms “ΔN_q” and “Δj” need also to be expressed in the space of phenotype values.

By definition, “ΔN_q” is the number of microstates of type “q” in the interval of phenotype values “$\Delta\Omega$”. Using probability density functions one can then rewrite, $\Delta Nq = N_q^o \cdot P_q(\Omega)\Delta\Omega$, where “$N_q^o$” is the total number of microstates of type “q” in the population for the genome position considered, and “$Pq(\Omega)$” is the probability density function of the microstate. Similarly, “Δj” is the number of individuals in the interval of phenotype values “$\Delta\Omega$”. Likewise, one can then rewrite, $\Delta j = N \cdot P(\Omega)\Delta\Omega$, where “$N$” is the total number of individuals in the population, and “$P(\Omega)$” is the probability density function of the phenotype.

Consequently,

$$\frac{\Delta N_q}{\Delta j} = \frac{N_q^0 \cdot P_q(\Omega)\Delta\Omega}{N \cdot P(\Omega)\Delta\Omega} = \omega_q^0 \frac{P_q(\Omega)}{P(\Omega)}$$

And one deduces finally

$$\frac{P_q(\Omega)}{P(\Omega)} = e^{\hat{u}_q(\Omega)} \int_{\Omega-\Delta\Omega/2}^{\Omega+\Delta\Omega/2} e^{\hat{u}_q(y)-\hat{u}_q(\Omega)} \frac{dy}{\lambda(y)}$$

As a result, the field is a function of probability density functions taken as a whole, and not only a function of the average values. That is to say that the field contains information on all the moments of the probability density functions. With this formalism, the variance of microstate distribution density functions can be involved in “$\theta(j)$-θ_o (j)”, namely in genotype-phenotype associations.

To recover Fisher’s theory let us assume an infinitely dense population (infinite population). In this case the interval “ΔΩ” can tend toward zero and as a result, the field can be expected to be almost constant over the very small interval of phenotype values “ΔΩ”. One can then neglect the exponential in the integral since in this case $\hat{u}_q(y)$-$\hat{u}(\Omega)$~0. Furthermore, as by definition,

$$\int_{\Omega+\Delta\Omega/2}^{\Omega+\Delta\Omega/2} \frac{dy}{\lambda(y)} \sim 1$$

one obtains simply

$$\frac{P_q(\Omega)}{P(\Omega)} = e^{\hat{u}_q(\Omega)}$$

To express the field in Fisher context, consider now that the probability density functions of the microstate “q” and of the phenotype value are normally distributed, respectively written as,

$$P_q(\Omega) = \frac{K_q}{\sigma_q} exp\left(-\frac{1}{2}\left(\frac{\Omega - \langle\Omega\rangle_q}{\sigma_q}\right)^2\right)$$

$$\text{and } P(\Omega) = \frac{K}{\sigma} exp\left(-\frac{1}{2}\left(\frac{\Omega - \langle\Omega\rangle}{\sigma}\right)^2\right),$$

where Kq and K are normalization constants, “⟨┤⟩” denotes averages and “σ_q” and “σ” the variances.

In his seminal paper, (Fisher 1919), Fisher assumed also that the variance of microstates are similar to that of the phenotype, that is σ_q~σ. In this context one can defines Fisher’s field for the microstate of type “q” as

$$\hat{u}_q(\Omega) = \left(\frac{\langle\Omega\rangle - \langle\Omega\rangle_q}{\sigma}\right)\left(\frac{\Omega}{\sigma} - \frac{1}{2}\frac{\langle\Omega\rangle + \langle\Omega\rangle_q}{\sigma}\right) + ln\left(\frac{K_q}{K}\right)$$

That is to say that based on Fisher’s seminal idea the fields should be linear.

With this assumption, the gene effect, a = 1/2 [$\langle\Omega\rangle_+$-$\langle\Omega\rangle_-$], and the dominance, d = $\langle\Omega\rangle_o$-1/2 [$\langle\Omega\rangle_+$+$\langle\Omega\rangle_-$], correspond to derivative of the fields under the form

$$a = \frac{\sigma^2}{2}\frac{d}{d\Omega}[\hat{u}_-(\Omega) - \hat{u}_+(\Omega)]$$
$$d = \frac{\sigma^2}{2}\frac{d}{d\Omega}[\hat{u}_-(\Omega) + \hat{u}_+(\Omega) - 2\hat{u}_0(\Omega)]$$

6. Beyond Fisher

Assume now that σ_q≠σ, one deduces a more generic form for the field when normal distributions are employed,

$$\hat{u}_q(\Omega) = -\frac{1}{2}\left(\frac{\Omega - \langle\Omega\rangle_q}{\sigma_q}\right)^2 + \frac{1}{2}\left(\frac{\Omega - \langle\Omega\rangle}{\sigma}\right)^2 + ln\left(\frac{K_q}{K}\frac{\sigma}{\sigma_q}\right)$$

Thus, in the general case the fields are expected to be non-linear due to unequal variances. What the latter relation confirms also is that the variances as well as the averages are involved in genotype-phenotype associations.

Assume now that $\langle\Omega\rangle$~$\langle\Omega\rangle_q$. Traditional GWAS would conclude that the gene effect is null. However, in our case, provided that σ_q≠σ, the fields would be non-null still suggesting potential genotype-phenotype association. This suggests that considering averages only resulting in the notion of gene effect linked to averages difference is too restrictive.

.

7. Environment and Heredity

The difference given by “$\theta(j)$-$\theta_o(j)$” provides a way to determine genotype-phenotype association that depends only on a difference between two configurations involving microstates. That is there is no role given to the environment. In fact, the traditional notion of environment as defined in GWAS can be rederived considering the variance of microstates when the phenotypic field is considered.

In Fisher theory, the associations between genotype and phenotype are determined exclusively through the use of averages. In his seminal paper (Fisher 1919) and by considering one particular gene (Mendelian factor) involved in the formation of the phenotype, Fisher starts by defining two relations that relate the average value of microstate distribution density functions, $\langle\Omega\rangle_q$, to the average value of the phenotype, $\langle\Omega\rangle$, and to a new parameter called today the genetic variance, α^2, both expressed under the form

$$\langle\Omega\rangle = \sum_{q=+,0,-} \omega_q^0 \langle\Omega\rangle_q$$

$$\alpha^2 = \sum_{q=+,0,-} \omega_q^0 \left(\langle\Omega\rangle - \langle\Omega\rangle_q\right)^2$$

Accordingly, the environment is added to complete the phenotype distribution density function. More specifically, the effect of the environment is defined through a variance, σ_e^2, such that

$$\sigma^2 = \sigma_e^2 + \alpha^2$$

The variance linked to the environment can be derived explicitly. Let us recall the relation, $P_q(\Omega)/P(\Omega) = e^{\hat{u}_q(\Omega)}$, and rewrite it as, $\omega_q^o P_q(\Omega) = \omega_q^o e^{\hat{u}_q(\Omega)} P(\Omega)$. Summing the latter relation for each microstate, one deduces then

$$\sum_{q=+,0,-} \omega_q^0 P_q(\Omega) = P(\Omega) \sum_{q=+,0,-} \omega_q^0 e^{\hat{u}_q(\Omega)}$$

As $\sum_{q=+,0,-} \omega_q^o e^{\hat{u}_q(\Omega)} = 1$, it follows that the two first moments can be determined by

$$\sum_{q=+,0,-} \omega_q^0 \int (\Omega - \langle\Omega\rangle) P_q(\Omega) d\Omega = \int (\Omega - \langle\Omega\rangle) P(\Omega) d\Omega$$

$$\sum_{q=+,0,-} \omega_q^0 \int (\Omega - \langle\Omega\rangle)^2 P_q(\Omega) d\Omega = \int (\Omega - \langle\Omega\rangle)^2 P(\Omega) d\Omega$$

Where the integrals involve all possible phenotypic values. Those integrals can be rewritten also as

$$\sum_{q=+,0,-} \omega_q^0 \int \left(\Omega - \langle\Omega\rangle_q + \langle\Omega\rangle_q - \langle\Omega\rangle\right) P_q(\Omega) d\Omega = \int (\Omega - \langle\Omega\rangle) P(\Omega) d\Omega$$

$$\sum_{q=+,0,-} \omega_q^0 \int \left(\Omega - \langle\Omega\rangle_q + \langle\Omega\rangle_q - \langle\Omega\rangle\right)^2 P_q(\Omega) d\Omega = \int (\Omega - \langle\Omega\rangle)^2 P(\Omega) d\Omega$$

Owing to the fact that $\int(\Omega - \langle\Omega\rangle_q) P_q(\Omega) d\Omega = 0$, the first integral gives

$$\sum_{q=+,0,-} \omega_q^0 \left(\langle\Omega\rangle_q - \langle\Omega\rangle\right) = 0$$

As $\sum_{q=+,0,-} \omega_q^o = 1$, one deduces that the first integral provides indeed the first relation linking the averages as given by Fisher.

By developing the quadratic term in the second integral and owing to the fact that, $\int(\Omega - \langle\Omega\rangle_q)^2 P_q(\Omega) d\Omega = \sigma_q^2$, one deduces

$$\sum_{q=+,0,-} \omega_q^0 \sigma_q^2 + \sum_{q=+,0,-} \omega_q^0 \left(\langle\Omega\rangle_q - \langle\Omega\rangle\right)^2 = \sigma^2$$

As by definition $\alpha^2 = \sum_{q=+,0,-} \omega_q^o (\langle\Omega\rangle_q - \langle\Omega\rangle)^2$, the environment is therefore linked to the variance of microstates under the form

$$\sum_{q=+,0,-} \omega_q^0 \sigma_q^2 = \sigma^2 - \alpha^2 = \sigma_e^2$$

To conclude, with GIFT the definition of the environment in genotype-phenotype associations results from the variance of microstates. However, a theory entirely focused on averages to determine associations and considering the variances as mere fluctuations would have missed the importance of the variance of microstates in the associations themselves. This is why the environment is often considered as an "intruder" in GWAS but always present, and why heredity linked to the variances and defined as the ratio between the genetic variance and the phenotypic variance is often used to determine genotype-phenotype associations.

Conclusion

The field of probability is borne out from our desire to provide a foundation to the notion of "evidence". The method of relative frequencies is fundamentally based on the notions of "imprecision", "uncertainty", "error" or "ignorance". Whilst there are some advantages to using frequentist probabilities to work with derived parameters such as, for example, the average or the variance when the conditions underlying the existence of probability are met; it is paramount to realize that the "average" and the "variance" result from the acknowledgement that a void exist in our knowledge. Because those two parameters have had a life on their own sociologically, mostly through diverse analogies such as for example the definition of the "social body", they appear legitimate to us. However, there are no good reasons to think always in term of "average" or "variance" or both. One can still feel ripples of such analogies in the 21st century. For example, the Body Mass Index (BMI) was invented by Quetelet (Faerstein & Winkelstein 2012) and is used to underscore health/obesity based on a distribution density function. One may then wonder about the universality of considering this distribution density function when rugby or American football players who won the six-nation tournament or the super bowl are considered, who would probably offset any BMI limits. The problem is that deciding to consider those players separately would split the "social body" demonstrating the overall futility of considering probability density functions as universal identifying of population. Again, it is the individuals/people that form a population, not the opposite way around.

Aside from considering "population", the problem culminates when, in addition, one tries to force a population into the field of probability as a number of assumptions need to be made that are not always realistic.

The method suggested (GIFT) tries to remove our reliance on the notion of average by considering the shortcoming of frequentist probability and creating a new mathematical object. This new mathematical object, called the genetic paths difference, takes for granted that no obvious void is present in our knowledge because precision (in phenotypic measurements) can exist. The advantage of using this model is that it does not contradict Fisher but, instead, generalizes it by giving a role also to the variance of microstates. Indeed, specific fields can be derived using Fisher's assumptions. The potential role of the variance of microstates in genotype-phenotype associations is, currently, a highly debated matter (Nelson, Pettersson, & Carlborg 2013). The model exposed herein will probably help in this matter.

Perhaps the most important point with this model (GIFT) is that, as opposed to using a population to determine genotype-phenotype associations, the reintegration of individuals into genome-wide association studies permits us to think about the self-consistency of genetics that is the "loop" that exists (and must exist) between phenotype and genotype. This can provide a basis to comprehend the notion of epigenetics and in particular the notion of phenotype plasticity in evolution and in genome-wide association studies, whereby phenotype alterations can happen without affecting the DNA composition (Fusco & Minelli 2010; Sommer 2020).

Data accessibility. The code used to generate the figure 1 can be found in the supplementary materials of Rauch, C *et al.* 2022, *bioRxiv*, p. 2022.02.25.479563. doi: 10.1101/2022.02.25.479563.

Authors' contributions. Conceptualization and writing: C.R., J.W. and S.B.

Competing interests. We have no competing interests.

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

Acknowledgements. We thank Professors Patricia Harris and James Leigh for fruitful discussions concerning the urgent need to improve genotype-phenotype association methods.

Appendix: "Probability" (Abraham de Moivre) vs. "Conditional Probability" (Thomas Bayes) vs. "Generalized Probability" (Pierre-Simon Laplace)

The theory of probability is to define a mathematical framework to model random events. There are different ways to define, as well as interpret, a probability epistemologically. Defined by Bernoulli and developed by de Moivre the most common definition is when the frequency of events can be defined, also known as frequentist probability. In this case, the probability of a particular event can be defined objectively. However, the use of the normal distribution formula defined in the continuum limit implies the possibility to repeat independently an infinite number of times the same experiment. There is thus an empirical problem since it is not possible to clearly define "infinite number of times”. This in turn means that the probability can only be defined subjectively when dataset is limited. This subjective approach was developed by T. Bayes and is known as “conditional probability”. Bayes managed to provide an expression for the resulting probability of a hypothesis upon the addition of some evidence to the antecedent body of knowledge. In this case, Bayes showed that the posterior probability varies directly as the prior or antecedent probability. That is to say that if the evidence is what is expected, it casts little credit upon any particular hypothesis. Consequently, trying to promulgate Bayes’ method as an objective one is, practically speaking, impossible since there is nothing trivial in determining a meaningful antecedent probability out of the blue. In Bayes case, the only solution to generate an objective probability is by knowing all antecedent probabilities. This viewpoint was developed by Laplace. Laplace understood that the field of probability can be used as a measure of our “ignorance” concerning a process only in two different cases. Assume an event determined by different causes. One can then determine the probability of the event knowing the causes or, the probability of the causes knowing the event. To demonstrate this point, assume that three possible causes, noted “+1”, “0” and “-1”, generate an event and note by $P(+1)$, $P(0)$ and $P(-1)$ the probability of these causes. Then $P(+1)$, $P(0)$ and $P(-1)$ can be rewritten, respectively, as $P(+1) = N_{+1}/N$, $P(0) = N_0/N$ and $P(-1) = N_{-1}/N$, where “N_{+1}”, “N_0”, “N_{-1}” are the number of times the causes “+1”, “0” or “-1” were observed/measured, and “N” is the total number of observations or measurements made. If among those “N” observations or measurements made the number of times the event “E” was observed/measured is “N_E”, then the probability of the event “E” occurring is, $P(E) = N_E/N$. One can also determine the probability that the event “E” occurs as a result of the cause “+1”, noted $P(E/+1)$. In this case, $P(E/+1) = (N_E)_{+1}/N_{+1}$, where “$(N_E)_{+1}$” and “$N_{+1}$” are, respectively, the number of times the event “E” and the cause “+1” were simultaneously observed or measured. Note that $(N_E)_{+1}$ is a subset of the total number of events “N_E” since they are only determined by “+1”. Consequently, since only three causes can determine the event “E” one can write, $(N_E)_{+1}+(N_E)_0+(N_E)_{-1} = N_E$, and as a result, $N_E = P(E/+1)N_{+1}+P(E/0)N_0+P(E/-1)N_{-1}$. Dividing the latter relation by “N” one finds, $P(E) = P(E/+1)P(+1)+P(E/0)P(0)+P(E/-1)P(-1)$. One can then determine the probability that the event observed is caused by “+1” by using the ratio $(N_E)_{+1}/N_E = P(E/+1) N_{+1}/N_E$. By multiplying and dividing the right-hand side by “N” one deduces finally, $(N_E)_{+1}/N_E = P(E/+1)P(+1)/[P(E/+1)P(+1)+P(E/0)P(0)+P(E/-1)P(-1)]$. The ratio $(N_E)_{+1}/N_E$ is the probability that “+1” caused the event and as a result this ratio can be re-noted $P(+1/E)$. One deduces then the formula wrongly attributed to Bayes since Laplace derived it in 1776:

$$P(+1/E) = P(E/+1)P(+1)/[P(E/+1)P(+1)+P(E/0)P(0)+P(E/-1)P(-1)]$$

This type of formula is the one used in this manuscript and derived in (Rauch *et al.* 2022; Wattis *et al.* 2022). The important property of this relation is that the notion of “density” disappears since the right-hand side is a ratio of probabilities. Note also that if an event is always observed/measured then $N_E = N$ and the denominator is equal to one leading to: $P(+1/E) = P(E/+1)P(+1)$. The later relation is the true Bayes’ formula.

References

Beaumont, MA, & Rannala, B 2004, “The Bayesian revolution in genetics”, *Nature Reviews. Genetics*, vol. 5, no. 4, pp. 251–261. doi: 10.1038/nrg1318.

Boichard, D *et al.* 2016, “Genomic selection in domestic animals: Principles, applications and perspectives”, *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 274–277. doi: https://doi.org/10.1016/j.crvi.2016.04.007.

Abbott, BP *et al.* 2016, “Observation of gravitational waves from a binary black hole merger”, *Physical Review Letters*, vol. 116, no. 6, p. 61102. doi: 10.1103/PhysRevLett.116.061102.

Faerstein, E, & Winkelstein, WJ 2012, “Adolphe Quetelet: Statistician and more”, *Epidemiology*, vol. 23, no. 5. Available from: https://journals.lww.com/epidem/Fulltext/2012/09000/Adolphe_Quetelet__Statistician_and_More.18.aspx. [20 October 2022].

Falconer, DS 1996, *Introduction to quantitative genetics*. Harlow: Prentice Hall.

García Ferrari, M, & Galeano, D 2016, “Police, anthropometry, and fingerprinting: The transnational history of identification systems from Rio de la Plata to Brazil”, *História, Ciências, Saúde - Manguinhos*, Dec. 23, Suppl. 1, pp. 171–194. doi: 10.1590/S0104-59702016000500010.

Fisher, RA 1919, “XV.—The correlation between relatives on the supposition of Mendelian inheritance.’, *Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433. doi: 10.1017/S0080456800012163.

Fisher, RA 1923, “XXI.—On the dominance ratio”, *Proceedings of the Royal Society of Edinburgh*, vol. 42, pp. 321–341. doi: 10.1017/S0370164600023993.

Fusco, G, & Minelli, A 2010, “Phenotypic plasticity in development and evolution: Facts and concepts. Introduction”, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 365, no. 1540, pp. 547–556. doi: 10.1098/rstb.2009.0267.

Galton, F 1886, “Regression towards mediocrity in hereditary stature”, *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 15, pp. 246–263. doi: 10.2307/2841583.

Gayon, J 2016, “From Mendel to epigenetics: History of genetics”, *Comptes Rendus Biologies*, vol. 339, no. 7–8, pp. 225–230. doi: 10.1016/j.crvi.2016.05.009.

Droesbeke, JJ & Tassi, P 1990, *Histoire de la statistique*. Paris: Presses Universitaires de France.

Laplace, P-S 1995, *Théorie analytique des probabilités (Volume 1)*. Paris: Editions J. Gabay.

Lonsdale, J *et al.* 2013, “The genotype-tissue expression (GTEx) project”, *Nature Genetics*, vol. 45, no. 6, pp. 580–585. doi: 10.1038/ng.2653.

Macdonald, A, Hawkes, LA, & Corrigan, DK 2021, “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”, *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 376, no. 1831, p. 20200228. doi: 10.1098/rstb.2020.0228.

de Moivre, A 2013, *The doctrine of chances: Or, a method of calculating the probability of events in play*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139833783.

Morrison, M 1997, “Physical models and biological contexts”, *Philosophy of Science*, vol. 64, pp. S315–S324. Available from: www.jstor.org.nottingham.idm.oclc.org/stable/188413.

Nelson, RM, Pettersson, ME, & Carlborg, Ö 2013, “A century after Fisher: Time for a new paradigm in quantitative genetics”, *Trends in Genetics: TIG*, vol. 29, no. 12, pp. 669–676. doi: 10.1016/j.tig.2013.09.006.

Porter, TM 1985, “The mathematics of society: Variation and error in Quetelet’s statistics”, *The British Journal for the History of Science,* vol. 18, no. 1, pp. 51–69. Available from: www.jstor.org.nottingham.idm.oclc.org/stable/4026265.

Prunet, N, & Meyerowitz, EM 2016, “Genetics and plant development”, *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 240–246. doi: https://doi.org/10.1016/j.crvi.2016.05.003.

Quintana-Murci, L 2016, “Genetic and epigenetic variation of human populations: An adaptive tale”, *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 278–283. doi: https://doi.org/10.1016/j.crvi.2016.04.005.

Rauch, C *et al.* 2022, “FIT-GWA: A new method for the genetic analysis of small gene effects, high precision in phenotype measurements and small sample sizes”, *bioRxiv*, p. 2022.02.25.479563. doi: 10.1101/2022.02.25.479563.

Rose, N 2001, “The politics of life itself”, *Theory, Culture & Society*, vol. 18, no. 6, pp. 1–30. doi: 10.1177/02632760122052020.

Samueli J-J, & Boudenot J-C 2009, *Une histoire des probabilites des origines a 1900*. Paris: Ellipses.

Schacherer, J 2016, “Beyond the simplicity of Mendelian inheritance”, *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 284–288. doi: https://doi.org/10.1016/j.crvi.2016.04.006.

Sommer, RJ 2020, “Phenotypic plasticity: From theory and genetics to current and future challenges”, *Genetics*, vol. 215, no. 1, pp. 1–13. doi: 10.1534/genetics.120.303163.

Stigler, SM 1990, *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Todhunter, I 2014, *A history of the mathematical theory of probability: From the time of Pascal to that of Laplace*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139923576.

Visscher, PM *et al.* 2017, “10 Years of GWAS discovery: Biology, function, and translation”, *American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22. doi: 10.1016/j.ajhg.2017.06.005.

Visscher, PM, & Goddard, ME 2019, “From R.A. Fisher’s 1918 paper to GWAS a century later”, *Genetics*, vol. 211, no. 4, pp. 1125–1130. doi: 10.1534/genetics.118.301594.

Wattis, J *et al.* 2022, “Analysis of genotype-phenotype association using fields and information theory’, *arXiv* preprint arXiv:2202.11989.

Weissenbach, J 2016, “The rise of genomics”, *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 231–239. doi: https://doi.org/10.1016/j.crvi.2016.05.002.

Wright, JD 2009, “The founding fathers of sociology: Francis Galton, Adolphe Quetelet, and Charles Booth: Or what do people you probably never heard of have to do with the foundations of sociology?”, *Journal of Applied Social Science*, vol. 3, no. 2, pp. 63–72. doi: 10.1177/193672440900300206.