PHILOSOPHICAL
TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Jenkins D. 2024 How do stochastic processes and genetic threshold effects explain incomplete penetrance and inform causal disease mechanisms? *Phil. Trans. R. Soc. B* **379**: 20230045.
https://doi.org/10.1098/rstb.2023.0045

Received: 4 October 2023
Accepted: 16 January 2024

One contribution of 10 to a discussion meeting issue 'Causes and consequences of stochastic processes in development and disease'.

Subject Areas:
developmental biology, genetics

Keywords:
stochastic processes, genetic threshold effects, causal inference

Author for correspondence:
Dagan Jenkins
e-mail: d.jenkins@ucl.ac.uk

How do stochastic processes and genetic threshold effects explain incomplete penetrance and inform causal disease mechanisms?

Dagan Jenkins

Great Ormond Street Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK

DJ, 0000-0003-3293-2999

Incomplete penetrance is the rule rather than the exception in Mendelian disease. In syndromic monogenic disorders, phenotypic variability can be viewed as the combination of incomplete penetrance for each of multiple independent clinical features. Within genetically identical individuals, such as isogenic model organisms, stochastic variation at molecular and cellular levels is the primary cause of incomplete penetrance according to a genetic threshold model. By defining specific probability distributions of causal biological readouts and genetic liability values, stochasticity and incomplete penetrance provide information about threshold values in biological systems. Ascertainment of threshold values has been achieved by simultaneous scoring of relatively simple phenotypes and quantitation of molecular readouts at the level of single cells. However, this is much more challenging for complex morphological phenotypes using experimental and reductionist approaches alone, where cause and effect are separated temporally and across multiple biological modes and scales. Here I consider how causal inference, which integrates observational data with high confidence causal models, might be used to quantify the relative contribution of different sources of stochastic variation to phenotypic diversity. Collectively, these approaches could inform disease mechanisms, improve predictions of clinical outcomes and prioritize gene therapy targets across modes and scales of gene function.

This article is part of a discussion meeting issue 'Causes and consequences of stochastic processes in development and disease'.

1. Introduction

The identification of thousands of *bona fide* Mendelian disease genes represents one of the most important achievements in biomedical research, providing novel insights into disease mechanisms and promising to inform treatments and disease outcomes for patients [1–4]. However, the extensive phenotypic variability observed in patients with mutations in the same gene—even in pairs of monozygotic twins—reduces this predictive power and has proven to be a major limiting factor [5–8]. Given that most treatments represent the amelioration of severe forms of disease rather than being curative *per se*, it may be possible to identify new therapeutic targets by studying the mechanisms that underlie this variability and even to tailor medicines to individuals. Such work will also provide insight into fundamental genetic principles in the broad sense, or, as Waddington called it, *The Strategy of the Genes* [9]. This includes insight into the role of stochastic processes in biological systems, which is the focus of this Perspective.

THE ROYAL SOCIETY
PUBLISHING

Embryonic development and adult tissue homeostasis are exquisitely robust when viewed at macroscopic levels. As Wolpert discussed, it is bewildering that, for most people, their left and right arms remain almost identical in length and overall appearance throughout their lifetime, even though they have had no means of communicating with one another ever since the midline barrier was laid down in the early foetus [10]. This maintenance of mirror-image symmetry is also a feature of many other bilateral tissues. Yet when quantitative molecular and cellular assays are used to observe biological systems at microscopic levels they can be seen to be inherently noisy [11,12].

What then are the molecular and cellular sources of this noise? Stochastic processes can generate variation at the level of single cells, including intrinsic noise that results from non-statistical sampling of relatively small numbers of molecules. (For example, there are only two copies of each gene promoter located on an autosome.) This is thought to arise from processes such as transcriptional and translational bursting events [11–16] and can be ascertained by assessment of cell-to-cell variation.

By contrast extrinsic noise may arise from environmental fluctuations, and variability in the response to the environment may result from the disparity of cell state [17–20]. We might expect that extrinsic environmental factors such as temperature or diet are likely to be uniform across a tissue and generate inter-individual variation rather than noise at the single-cell level. Therefore, different sources of stochastic variation may have different signatures at the single molecule, cellular and tissue/inter-individual levels.

At the transcriptional level, the relative contributions of intrinsic versus extrinsic noise can be evaluated according to the correlation of pre-processed transcripts present at two identical promoters in single cells, such as the two alleles of an autosomal gene in an isogenic line [11,17,19]. Intrinsic versus extrinsic noise is indicated by the degree of allelic imbalance. Stochastic processes of this type have been demonstrated in a number of systems, and they have been shown to underlie incomplete penetrance, as will be outlined in detail below.

2. What is phenotypic variability?

Many Mendelian traits are discrete phenotypes that segregate with a defined mutation within a family. Variation in the severity of these phenotypes may involve individuals who carry the mutation but are unaffected, which is defined as incomplete penetrance. Alternatively, affected individuals may vary in severity, known as variable expressivity.

Most Mendelian diseases are syndromes which are compound phenotypes involving the combination of multiple independent features affecting a variety of organs or tissues [4,21–23]. While these diseases do appear to segregate according to Mendel's laws when considered as a whole, relatively few such disorders would actually exist without flexible systems of diagnosis [21–26]. For example, patients with Bardet-Biedl syndrome (BBS) exhibit a combination of six major and eight minor clinical features, and a clinical diagnosis of BBS is accepted only if a patient has four major criteria, or three major criteria and two or more secondary criteria [27,28]. For a series of genes causing brachydactylies, such as Robinow syndrome, heterozygous mutations cause isolated non-syndromic shortening of the digits while homozygous mutations in the same gene cause the same phenotype together with other syndromic features [29]. What seems to be the rule rather than the exception is that for these syndromes most individual clinical features are associated with incomplete penetrance when considered in isolation. For each independent phenotype mutations in Mendelian disease genes can be viewed as having three key features such that they are:

(1) rare susceptibility alleles, with;
(2) moderate to high penetrance, and;
(3) pleiotropic effects.

We may therefore consider that clinical variability in syndromic Mendelian disorders reflects incomplete penetrance combined across a number of independent clinical features. Incomplete penetrance may also explain variable expressivity where higher-level phenotypes involve multiple repeating structures or result from the combination of several constituent endophenotypes across biological scales. As a hypothetical example, a quantitative trait such as glomerular filtration rate may be the product of incomplete penetrance for discrete effects on each of approximately 10^6 nephrons within the kidney. Alternatively, a complex limb malformation might reflect incomplete penetrance for multiple independent developmental processes regulating digit number, identity and joint formation, which are temporally integrated [30].

Within naturally breeding populations, such as humans with genetic diseases, incomplete penetrance and variable expressivity are attributed to modifiers owing to genetic and environmental heterogeneity amongst these individuals. Stochastic processes also contribute to clinical variability. In genetically identical experimental organisms in controlled environments, stochastic processes are the major source of variation. How stochasticity generates phenotypic variability is poorly understood, especially for complex morphological traits. As such, the 6000 or so Mendelian disease genes that have been identified [4], and the series of mutations found within them, provide us with many models to investigate the role of stochastic processes in disease susceptibility.

3. Incomplete penetrance and genetic threshold effects: stochastic variation informs threshold values for incompletely penetrant traits

The genetic threshold model was first proposed in 1934 by Wright [31,32], to explain the appearance of discrete phenotypes. It was invoked to explain preaxial polydactyly that was observed with different penetrance values in a number of strains of guinea pigs. There are several features of this model and historical aspects to consider. In its original form, the model was used to explain this discrete morphological trait in inbred (isogenic) lines of guinea pigs. As for polydactyly, a wide variety of phenotypes have subsequently been observed at background levels in isogenic lines of rodents, with strain-specific penetrance values [33–35]. In this early era of genetics and developmental biology (or embryology, as it was then known), Wright especially attributed this phenotypic variability to environmental variation. He considered that genetic background predisposed different strains to disease, and that these effects were modified by environmental factors. He was also open to the idea of stochasticity, although he did not use this term. Instead, he commented on, '…irregularities in development due to the intangible sort of causes to which the word chance is applied' [36, p. 328]. Wright particularly considered position effects within the uterus and differences in foetal blood supply as forms of environmental variation which are quite different from stochastic molecular processes.

The combination of all genetic and environmental modifiers impacting a particular phenotype in an individual is defined as the liability. It is important to realise that liability is an intangible construct that cannot be measured. It is not possible to define all genetic modifiers [37], not least because of statistical power considerations and small effect sizes [38,39], and the sum total of environmental factors are too complicated to realistically ascertain. By studying isogenic lines, Wright paved the way to remove genetic modifiers as a source of variation and greatly simplified the problem. This provided the basis for an experimentally tractable approach. We can now use gene-editing to introduce virtually any mutation in isogenic lines to study defined genetic variation in the context of a variety of fixed genetic backgrounds. We can therefore take essentially the same approach as Wright albeit in a molecularly targeted way.

A central difficulty for the theory of evolution was to understand the connection between particulate inheritance and quantitative phenotypic variation. In seminal work, Fisher demonstrated that quantitative traits could arise from particulate inheritance of multiple susceptibility alleles with small effect sizes across many different loci, i.e. oligogenic traits. In this model, the liability in populations of genetically heterogenous individuals in varied environments is taken to be normally distributed [40–42]. Later, and presumably taking inspiration from Wright's work, Carter, in his consideration of the inheritance of pyloric stenosis was the first to combine this concept with the threshold model such that individuals carrying a greater number of modifiers than a threshold liability value exhibit the trait [43,44]. Falconer subsequently did much to extend this model and understand its implications [45]. This is a model that permeates genetics because it helps us to think about the intersection between genes and environment, quantitative variation and discrete phenotypes, i.e. diseases. However, by being based on the concept of liability it is not particularly useful.

While Fisher's work was ground-breaking for evolutionary theory, its application in complex genetics placed the emphasis on heterogenous populations and oligogenic traits which forms the basis for the genetic threshold model typically seen in textbooks. This side-lined Wright's approach which had focused on isogenic models as experimentally tractable systems. Conscious of this, we can now build on Wright's model in the post-genomic era by considering the distribution of liability values for a *defined genotype* on a *specific genetic background* (with genotype refrerring to the Mendelian trait locus). In this situation, stochastic processes account for most of the variation in liability values. We must therefore *redefine liability to include naturally occurring molecular and cellular variation*. In this form, liability constitutes the functional effects of experimentally controlled genetic and environmental factors as well as stochastic variation in the form of tangible readouts such as gene product abundance (e.g. RNA and protein abundance) or activity (e.g. transcription factor binding to a promoter or a signalling output such as protein phosphorylation). The key is that such readouts can be measured, at least in principle. At the level of individuals, the liability value would fall below the threshold for unaffected individuals and above this threshold for those individuals that are affected. As will be discussed in the next section, these are not only readouts in the sense that is widely used in biomedical research, such as luciferase reporter assays or other bioreporters that are surrogate correlates of gene activity. They must be a node or combination of nodes on the causal path linking genotype to phenotype and capture the total stochastic variation within the wider network that impacts upon the phenotype (see below for a consideration of how this might be practically implemented using current technologies).

This leads us to several important features of the molecularly targeted genetic threshold model. Firstly, incomplete penetrance is the result of stochastic variation in liability values under experimentally controlled conditions for isogenic models—without this variation, all individuals would either be affected or unaffected. Incomplete penetrance means that the threshold liability value is located within the range of stochastic variation for a defined genotype in a particular strain. As we shall see in the next section, the penetrance value is exactly equal to the area under this curve that falls above the threshold liability value.

In the case of incomplete penetrance, stochastic variation therefore carries information about the threshold value in the form of its probability distribution function and the exact penetrance value for a particular trait.

4. Threshold effects demonstrated by simultaneous assessment of phenotype and readout in single cells

A genetic threshold is an exact value that relates phenotype penetrance to the distribution of liability values for a specified genotype (figure 1). Where the liability values for a series of genotypes on an isogenic background are described by the same type of statistical distribution (normal, Poisson, gamma, etc) with the same variance, differing only in their population means, the penetrance values across a series of genotypes will be described by the cumulative distribution function (CDF) of liability values (figure 1*c*). (Note: this is an idealized model for illustration. Empirically, mutations are typically found to be associated with greater stochastic variation. Furthermore, different mutations may be associated with different variances according to their mechanism of pathogenesis and different buffering mechanisms that may modulate their effects (splicing, nonsense-mediated mRNA decay, chaperones/protein folding etc.). Nonetheless, phenotype penetrance is given by the CDF for a particular genotype if the appropriate readout(s) is ascertained.)

Given the intangibility of liability, this term might reasonably be replaced with *change in gene function*, as discussed above. Gene function may correspond to RNA or protein abundance at lower levels of gene function, or any other higher-level effect of a mutation on the casual path to a phenotype, such as a dynamical signal transduction readout within a tissue. In the same way that the classical definition of liability is the sum total of genetic and environmental

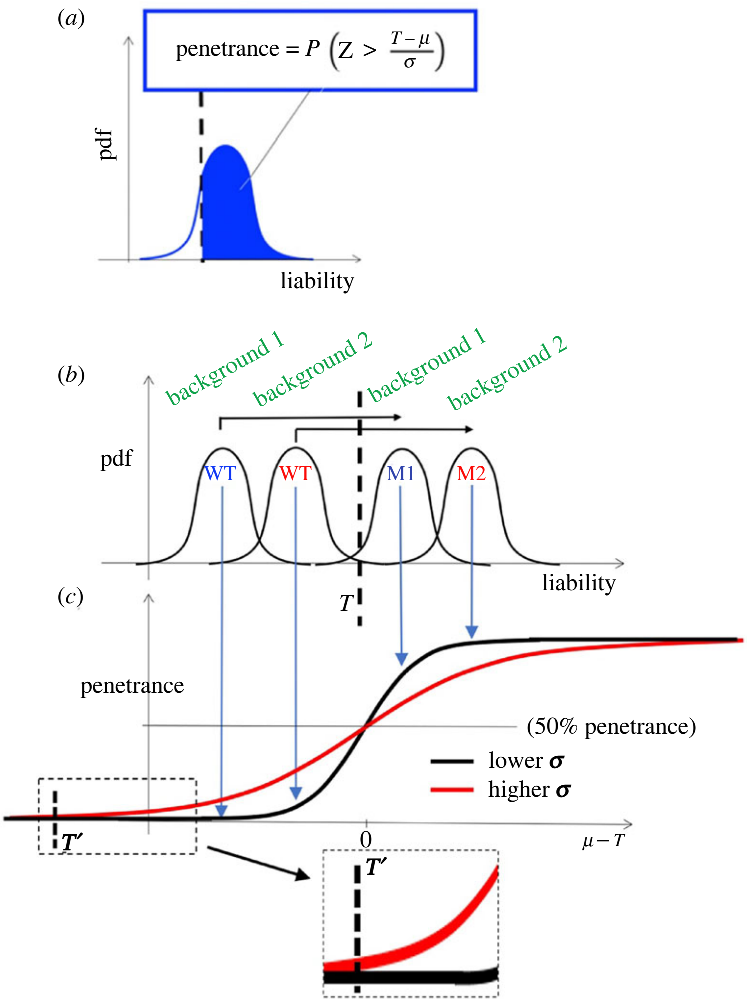

Figure 1. Genetic threshold model. (*a*) Basic model showing how the penetrance value for a trait is directly derived from the probability distribution function (pdf) of liability values in relation to a genetic threshold value—incomplete penetrance occurs when the threshold value falls within the pdf associated with a mutation. (*b*) Cartoon of pdfs for different mutations (M1, M2) on different genetic backgrounds showing that liability is the result of genetic background, mutational effect and stochastic variation. Different penetrance values for different genotypes are defined exactly by these pdfs in relation to the genetic threshold value (*T*, dashed line). (*c*) Disease penetrance relates to mean liability values for pdfs defined by genetic background, mutation and stochastic processes in (*b*) according to cumulative distribution functions (black line). The red line illustrates the cumulative distribution function for the same set of distributions in (*b*), but with higher levels of stochastic variation (note where the mean equals the threshold value, penetrance is 50% regardless of variance). The threshold value that can be ascertained by simultaneous ascertainment in single cells (*T′*) is a function of the genetic threshold value (*T*). Note that penetrance values increase above this threshold more rapidly where the variance is greater, reflecting the broader pdfs for genotypes below the threshold value (see text for further discussion).

risk factors for a particular phenotype, the penetrance value of a discrete phenotype would reflect the average change in gene function AND stochastic variation combined across *all* functional nodes within a network that converge on a phenotype. While liability, according to its classical definition, can be considered to be all genetic and environmental inputs into a system that define disease risk, the equivalent change in gene function is the highest-level readout(s) of biological function that causes a specific phenotype, and is the equivalent output derived from the input liability.

(We can see why we require such a high-level readout and its associated variation to capture all information about the genetic threshold value, and *vice versa*, as follows. A mutation that affects RNA abundance for instance will in turn influence protein abundance. As such, measurements of protein abundance will capture both the average mutational effect and associated noise at the RNA level as well as variation at the level of messenger RNA (mRNA) translation. In general, a higher-level readout will capture all of the mutational effects and variation at lower levels that feed into that particular node, as illustrated in figure 2. Therefore, only a readout at a level that is 'proximal' to a discrete phenotype of interest will accurately reflect the genetic threshold value.)

This high-level readout of gene function is the combined statistical distribution across all such nodes. In principle, this would allow for a parametric approach to calculate the genetic threshold value based on measurements of the relevant nodes within such a network. However, our limited understanding of the convoluted causal path linking functional readouts with phenotypes over time and biological scales precludes such an approach. The search for these predictive high-level readouts constitutes biomarker discovery, where a biomarker is a functional readout(s) that is either causally related to a phenotype or is correlated with a causal readout, and will be discussed further below.

In the absence of precise and detailed biological models such a parametric approach is not currently possible, and so the previous discussion of genetic threshold values was mainly illustrative. An alternative procedure that some

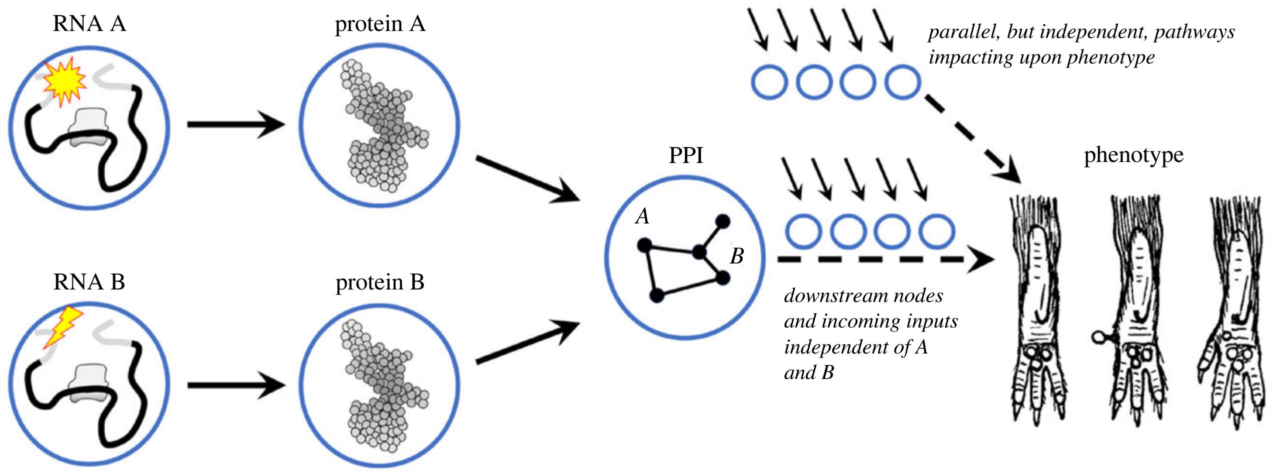
RNA A
protein A
RNA B
protein B
PPI
A
B
parallel, but independent, pathways impacting upon phenotype
phenotype
downstream nodes and incoming inputs independent of A and B

Figure 2. Relationships between mutations, pathways, networks and phenotypes. In the centre, we focus on a specific node—in this case a protein-protein interaction (PPI) complex—into which linear pathways relating to the synthesis of two components, A and B, feed (left). A reductionistic approach can be taken whereby the average mutational effects and associated variance in RNA, protein and PPI abundance and function can be measured. Downstream of 'PPI' a series of steps eventually lead to a discrete phenotype. Although represented as linear, these components and functions form part of the wider network and the precise average values and stochastic variation associated with these components are influenced by these factors (arrows). By measuring 'PPI' we have no information about these values; however, the phenotype penetrance is influenced by all of these factors.

researchers have used is to simultaneously score the penetrance of a discrete trait and to quantify a biochemical readout(s) for simple phenotypes that are apparent at the level of single cells. This removes the need for temporal separation of each assay and can be done at scale to include large numbers of individual cells collected in bulk. This simultaneous ascertainment approach removes the requirement for a detailed biological model linking readout and phenotype over time.

An example of this is an analysis of threshold cyclin-dependent kinase activity in the regulation of cell division during mitosis and cell division in *Schizosaccharomyces pombe* [46,47]. Multiple cyclin-cyclin dependent kinase (CDK) complexes regulate cell cycle progression in fission yeast, and earlier work had suggested that quantitative changes in total CDK activity led to the orderly initiation of S phase and mitosis. By using genetic simplification of the CDK network [48] in mutants expressing only a single cyclin-CDK chimera in place of the four cyclin-CDK complexes usually present, Swaffer *et al.* [46] demonstrated a progressive increase in a range of phosphorylation substrates until the end of the cell cycle. Initiation of phosphorylation of several substrates at different stages of the cell cycle was related to different cyclin-CDK affinities, suggesting that a causal relationship between CDK activity thresholds, differential substrate phosphorylation and initiation of G1-to-S and G2-to-M transitions.

To ascertain these threshold values, Patterson *et al.* [47] developed a CDK activity biosensor which permitted the *in vivo* single-cell assessment of CDK activity and mitotic cell division. Fluorescence imaging allowed hundreds of individual cells to be simultaneously scored in bulk for cell division status, and their level of CDK activity to be quantified. By plotting CDK activity against the rate of division, a threshold of CDK could be seen directly. Below this threshold there was no cell division, but there was an exponential increase in the proportion of divided cells above this value. This is reminiscent of the exponential phase of a CDF that relates stochastic variation to phenotype penetrance (figure 1*c*).

A second study analysed threshold values within a gene regulatory network that specifies intestinal stem cell identities in *Caenorhabditis elegans* [13]. This network involves the maternal deposition of *skn-1* transcripts that induce expression of *end-1*, which in turn activates expression of *elt-2*. There is also a parallel path in the network whereby *skn-1* transcript regulates expression of *elt-2* via other transcription factors also activated by *skn-1* (figure 3*a*). Using single molecule fluorescence *in situ* hybridization (FISH), the authors were able to quantify processed transcript abundance for each of these components within intestinal cells, demonstrating a continuous distribution of stochastic variation in the expression levels of *end-1* in *skn-1* mutants, ranging from complete loss to normal expression at the level of single cells. By contrast, *end-1* expression was never zero in wild-types at the appropriate stage of development and demonstrated much less variability. Downstream of *end-1*, *elt-2* demonstrated a bimodal distribution of expression in individual cells either falling within the normal range or with no detectable expression (figure 3*b*).

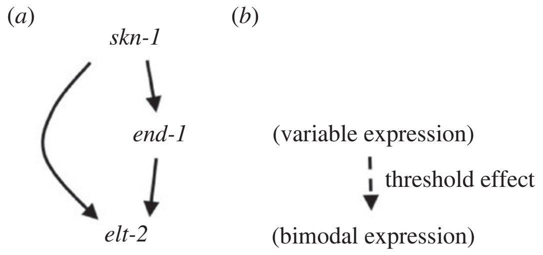

Figure 3. Simplified scheme of intestinal cell fate specification, adapted from [13].

To test the hypothesis that *elt-2* expression only occurs if stochastic variation in *end-1* transcript levels exceed a certain threshold value during a critical time-window, the levels of *end-1* and *elt-2* were simultaneously quantified in single cells. As for the CDK example outlined previously, this demonstrated a range of *end-1* values for which no *elt-2* expression was observed. A threshold value was defined as the maximum *end-1* value above which *elt-2* transcripts were non-zero in simultaneous quantification (figure 3*b*).

These studies offer additional layers of insight into threshold effects. In the CDK study, the same analysis was performed on both a wild-type and a CDK mutant strain, which suggested that the threshold value was the same for both. However, once the threshold had been reached, the phenotype penetrance (probability of cell division) increased more sharply with increasing CDK activity in the mutant versus wild-type strain. This is consistent with a CDF representing a greater level of stochastic variation in the mutant. In the intestinal specification study, the threshold value for two high penetrance *skn-1* alleles were the same, whereas for a weaker mutation, the threshold was lower. In this case it was suggested that *skn-1* may partially activate the parallel pathway, thereby changing the topology of the network and redefining the threshold for *end-1* activity. In this case, ascertainment of threshold values therefore had the added advantage of revealing a systems level property of the gene regulatory network.

This leads to two fundamental aspects of stochasticity. Firstly, while the average effects on gene function may differ for different genotypes, threshold values are properties of wider network topology: they are independent of genotype and remain constant for a given genetic background. In dynamical systems modelling, similar emergent properties of signalling pathways have been uncovered including distributed robustness and bifurcation points that determine cell fate decisions [49,50]. Second, stochastic noise is almost always greater in networks carrying a defined mutation (or any other specific perturbation) (figure 1*c*). This may relate to buffering mechanisms and localized network properties that confer robustness.

(This note addresses two points that were raised in review. It was stated in the previous paragraph that the threshold value is invariant for a given genetic background. This requires some justification as Wright does indeed assume constant threshold values across genetic backgrounds [31,32]. It is relevant for us to consider if this is correct. In Waddington's view [9], the epigenetic landscape constitutes the wider network consisting of the biophysical parameters that define the nodes and edges within a dynamical system (such as a large collection of ordinary differential equations). If the threshold value is a property of this wider network, rather than the mechanisms that modulate effects of a specific mutation (i.e. proximal to the mutation), then it would be invariant for a given genetic background in a controlled environment. It is also important to note that the threshold value is not 'picked' arbitrarily (i.e. scaled and shifted appropriately to reflect both the liability values and the penetrance of the mutation). Rather, it reflects network properties and is not peculiar to a particular mutation.)

5. Could genetic threshold values be determined for complex (morphological) phenotypes?

These examples of simultaneous ascertainment demonstrate the possibility of determining threshold values for relatively simple cellular phenotypes. Could this be possible for complex phenotypes at the level of whole tissues occurring late in the process of development? It was stated above that a molecularly targeted genetic threshold model requires assessment of a combination of nodes that capture the total stochastic variation within the wider network that impacts upon the phenotype. A potential pushback on this idea is that this is an unknowable causal graph, and the strictly sufficient measurements (e.g. every protein molecule concentration, modification state, and location in every cell) are unrealistic to obtain without some strong casual assumptions. Is such a high-level readout that captures all relevant functional variation just as intractable as the classical definition of liability?

A great deal of work will be necessary to answer this question definitively, but it is illustrative to consider a practical scheme using single cell RNA-sequencing (scRNAseq) and spatial transcriptomic technologies. In his epigenetic landscape model, Waddington [9] proposed that only a limited set of discrete cellular and tissue-level fates can be arrived at over the course of developmental time in an individual animal (i.e. amongst groups of genetically identical cells). Each path within the landscape reflects its tolerance to stochastic variation and resistance to changes in cell fate. His idea also encapsulated tissue transplantations, thereby traversing all tissues within an embryo and reflecting the concepts of specification and determination in experimental embryology. Although less famous than his epigenetic landscape diagram, Waddington also considered a 'phase-space box diagram' consisting of three dimensions. One dimension was time, and the other two functioned to define clusters of cellular identities. These clusters overlapped and each identity mapped onto future clusters, representing specification and differentiation over time.

While this diagram was entirely theoretical and heuristic at the time, it bears remarkable similarity to what we are very familiar with from scRNAseq datasets where different cell types are revealed as clusters following dimensionality reduction. scRNAseq is necessarily destructive of the samples analysed and thereby strictly prohibiting temporal analyses (although trajectory analyses are possible given certain assumptions). However, it is possible to conceive of a simultaneous ascertainment approach, firstly using clustering to define discrete cellular phenotypes at the single cell level, and then using a form of principal component analysis to define quantitative stochastic variation that correlates with these cell types. In essence, this would serve to define both the high-level quantitative readout and associated variation that was discussed previously and suggests that this form of liability is ascertainable. In terms of higher-level phenotypes, more ambitious and exploratory models that relate these single cell phenotypes within a tissue to the penetrance of higher-level phenotypes would be required. Possible models include the number of each cell type (total/differentiated/proliferative) with or without weightings according to location or developmental time. Alternative models could be scrutinised by evaluating a variety of interventions using do-calculus.

While speculative, this illustrates how such a molecularly targeted liability model based on stochasticity might be possible.

6. Comparing experimental approaches to threshold determination with causal inference

The simultaneous ascertainment of thresholds described above is based on a direct comparison of the values for a particular readout of gene function in cells with or without a phenotype. Both 'affected' and 'unaffected' cells come from a mixed group that demonstrate stochastic variation in the values of the readout and incomplete penetrance for the trait. This approach is based on observational data alone. The stochasticity is useful in generating a distribution of values for a biological variable that would be difficult to achieve experimentally without artefact. It also captures natural biological variation.

Controlled experiment is the gold-standard for demonstrating causality. This is because correlations between parameters could arise from covariance of two readouts with a third confounding variable and not only through a causal relationship between them. In analysing heterogeneous human populations, it is not possible to control for all confounders, and in this case randomized control trials (RCTs; also invented by Fisher) are the gold-standard for establishing causality. Here, confounders are controlled for by *random assignment* of patients to different treatment groups. In analogy, stochasticity can also be used to achieve a form of random assignment to infer causal relationships.

In genetics, the stochastic nature of chromosomal segregation and crossing over provides a form of randomization by which causation can be inferred. Linkage analysis has been used to identify thousands of Mendelian disease genes [1–4,51,52]. The premise for this analytical approach is to first establish the mode of inheritance for a particular trait (e.g. dominant or recessive) and to genotype polymorphic genetic markers throughout the genome. Segregation and independent assortment of these variants occurs randomly with respect to disease status in regions of the genome that are not linked to the disease such that a particular autosomal allele has a 50% chance of passing from a parent to their offspring. As in RCTs, Mendelian inheritance serves as randomization with precisely defined ratios for the assignment of different regions of the genome to case (affected) and control (unaffected) groups. Deviation from this baseline value indicates linkage to a causative mutation. A special form of this causal analysis for quantitative traits is the transmission-disequilibrium test [53–56]. Another major area of causal inference is known as Mendelian randomization whereby genetic susceptibility variants for a specific phenotype are used as *instrumental variables*, and control for confounding of a risk factor under study [57,58]. Observational data with randomization through stochastic processes therefore allows causal mechanisms to be inferred.

We saw from the previous examples of threshold determination through simultaneous assessment of single cells that stochastic variation in biological networks can provide causal information. Specifically, for a mutant exhibiting a discrete phenotype, stochastic variation in a causative readout will exhibit a discontinuous correlation with this phenotype, i.e. there will be a correlation only above a threshold value relating to the truncated distribution in excess of the threshold (i.e. the line above blue shading in figure 1*a*). Furthermore, while the correlation coefficient may differ between strains with different mutations, the threshold value is constant for a particular isogenic strain where a high-level functional liability readout is quantified. Here, the randomization afforded by stochastic processes together with the concept of an invariant threshold value for a discrete trait and a uniform genetic background could provide a means to infer causal relationships in complex biological datasets. This concept of discontinuous correlation has parallels with an approach that is frequently used in the social sciences and econometrics known as regression discontinuity [59].

7. Structural equation models in causal inference and stochasticity

Causal inference (CI) is a statistical approach which allows sources of variation within a causal model to be quantified in relation to a variable outcome [60,61]. The causal model will have been derived from experimental evidence and/or will constitute a hypothesis relating to causation. As such, causal inference can only be applied relatively late within the discovery timeline of a research question, whereby extensive prior research has established high confidence causal relationships between various components. As will be described, it has several uses. CI can accommodate unknown confounders thereby allowing incomplete biological models to be analysed. It can quantify the relative contribution of different sources of variation to variable (phenotypic) outcomes in an unbiased way, and it can be used to infer these values even for variables that cannot be measured directly.

In CI a causal model is represented by a structural equation model (SEM) whereby nodes represent components within the system and edges are represented as arrows that depict causal relationships between these components. The purpose of CI is then to calculate weights for these edges that correspond to the proportion of variation in a particular outcome (such as a phenotype) that is accounted for by variation in a particular node. The weights for each edge are determined by calculating correlations between the values of different nodes (figure 4).

A feature of SEMs that are amenable to CI is that they are directed acyclic graphs (DAGs). This means that there is a path of causation following a series of nodes to a specified outcome without forming a closed loop. This form of pathway analysis was first proposed by Wright in 1920 [36], where he applied it to breeding experiments that he had performed in various strains of guinea pigs that demonstrated differences in coat coloration, and later generalized in 1921 [62]. By formulating a scoring system to quantify this variation and normalizing the data, he was able to calculate the linear correlations between coat coloration in parents and their offspring, and between offspring within the same litter. He considered an SEM whereby fertilized zygotes gave rise to each of the parents' coat coloration phenotype as well as the parents' gametes (figure 4). Through the precise statistical ratios of Mendelian inheritance, the gametes from each parent combined to generate the zygotes that generated the subsequent generation. In both generations, the observed coat colour phenotypes of both the parents and the offspring were defined by genetic and environmental variables, as well as a third term which Wright coined '*development*' to account for the 'ontogenetic irregularities', i.e. stochastic processes.

By incorporating correlations between parents and siblings, and between siblings within the same litter into the SEM, it was possible to quantify the contribution of genetics, environment and development to variation in phenotypic outcomes. In this way this statistical approach was able to infer and quantify the contribution of variation in these otherwise intractable parameters to variation in coat colour. As a

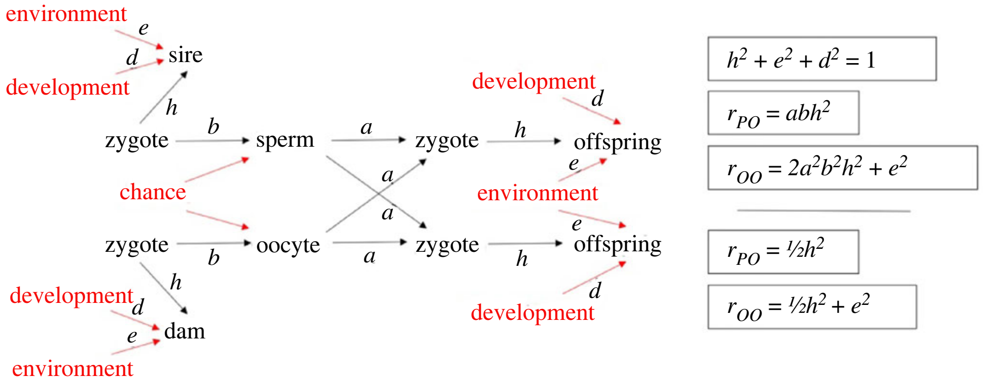

Figure 4. Structural equation model showing paths linking parental (sire, dam) phenotypes with offspring phenotypes. Nodes are as follows: tangible entities are in black, unmeasurable entities in red. Pathway coefficients are given in lower case (*a,b,e,d,h*). Simple mathematical relationships linking pathway coefficients with linear correlation coefficients between quantitative parent-offspring and offspring-offspring phenotypes are listed. With basic algebra, *h,e,d* can be derived although these would not otherwise be quantifiable. (Adapted from [36].) Note: Wright referred to '*chance*' as irregularities of development and also refers to '..ontogenetic irregularity..' in relation to the term *D*, hereby labelled *development*. *d* might reasonably represent weighting for stochastic processes, discussed in the text.

form of sanity check, Wright showed that environmental variation made the same contribution to phenotypic variability in both an isogenic line and an outbred strain of guinea pigs. By contrast, his calculation showed that genetic variation contributed significantly to phenotypic variability in the outbred strain but made no contribution to variability in the isogenic line.

Since Wright, CI was largely forgotten in the biomedical sciences, especially because of work by his contemporaries such as Galton/Pearson and Fisher who developed statistical methods for linear regression in genetics and RCTs, respectively. They established the mantra 'correlation not causation' (reviewed in [63]). They asserted that causal relationships could not be deduced from correlations within population-based datasets and could only be derived from RCTs, and this notion predominated. However, Wright had shown that correlations could reflect causation where a high confidence SEM is available.

CI based on SEMs has undergone extensive development. As stated by Meinshausen *et al*. [64, p. 7361], an SEM consists of: '(a) an underlying true causal influence diagram for random variables that are represented by nodes within the DAG, and; (b) a function that relates each variable to their parental variables and an error term. This so-called *'do-operator'* sets a particular variable to a deterministic value according to the SEM that relates the variable to its parental node, and can be applied to several variables simultaneously. This is a conditional probability of the kind – 'what is the probability of a specified outcome given that I assign a particular value to a specified variable (i.e. an intervention)'' [60,64,65].

What might the use of CI in understanding stochastic processes be? It might seem that, if a causal SEM is known, then there is no need for CI. However, CI would allow for a causal relationship to be defined in an SEM that subsequently turns out to have a weighting of zero. In other words, it may turn out that a parameter which exhibits some degree of variability and is thought to contribute variation to a phenotypic outcome might actually make no contribution at all. I began this *Perspective* by defining different sources of stochastic variation. Another use of CI could be to apply weights to these different sources of variation, thereby highlighting the importance of one or other form of stochastic variation associated with *a particular combination of mutation*, *genetic background and environment*. The complexities of the network that constitutes an SEM may unexpectedly render the system robust to variation in a particular node. In dynamical systems modelling, similar emergent properties of signalling pathways have been uncovered including distributed robustness and bifurcation points that determine cell fate decisions [49,50].

8. Conclusion and perspectives

In this *Perspective*, I have considered how stochastic processes could provide insight into mechanisms of pathogenesis. In isogenic model systems, stochastic variation is the major source of phenotypic variability and provides information about genetic threshold values where incomplete penetrance is observed. This is because there is an exact parametric relationship between phenotype penetrance and the probability distribution of stochastic variation. However, in order to make such a calculation, the sum total of average mutational effects and stochastic variation must be quantified across all nodes within a biological network that converge on a trait of interest. This has been achieved for relatively simple phenotypes by simultaneous quantification of a continuous biological readout and phenotype penetrance. However, this is unlikely to be possible for complex morphological traits where mechanisms of pathogenesis for such high-level traits are not well understood.

CI may provide an alternative means to harness this information that is afforded by stochastic processes. It may be possible to determine the degree to which different sources of stochastic variation within a network contribute to overall phenotypic variability. The main use of CI is to quantify the degree to which different sources of stochastic variation within a network contribute to overall phenotypic variability. This approach focuses on high-confidence causal relationships established through controlled experiments whereby SEMs can be drawn. An advantage of CI is that large gaps in the causal diagram can be tolerated by treating them as unknown confounders, and so a partial analysis of known causal mechanisms could be undertaken. By calculating correlations between different nodes within such a model, the extent to which variation in each node contributes to overall phenotypic variability can be estimated. It also allows inferences to be made for nodes that cannot be measured (but which are known) through these indirect calculations. In this way, unexpected and non-trivial weights can be given to known components in a causal framework.

This could help to disregard molecular and cellular mechanisms that seem intuitively to be causally related to a phenotype, potentially revealing novel mechanisms of buffering and robustness. It could also help to prioritize molecular targets for therapy. Genetic therapies are designed to target high-confidence causal mechanisms within the central dogma of molecular biology and many mutations affect multiple modes of gene function (RNA, protein etc). Weighting the relative contributions of different modes of gene function to phenotypic variability may therefore help to prioritize therapeutic targets. By making precise statistical statements for different causal relationships, CI may allow quantitative statements regarding the effectiveness of a particular therapy to be made. Similarly, it may permit the design of adjunct or combined therapies, targeted to a particular mode, where residual disease risk is present. In future, it may also be possible to make quantitative predictions about disease outcomes for individual patients undergoing a particular therapy. This would require disease models that represent a patient's total genetic constitution in which disease outcomes can be predicted (e.g. organoids, assembloids etc).

Data accessibility. This article has no additional data.

Declaration of AI use. I have not used AI-assisted technologies in creating this article.

Authors' contributions. D.J.: conceptualization, funding acquisition, writing—original draft.

Conflict of interest declaration. I declare I have no competing interests.

Funding. This work was funded by a Wellcome Trust Collaborative Award in Science (grant no. 210585/Z/18/Z) to D.J.

Acknowledgements. This research was also supported by the National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London.

References

1. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011 Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755. (doi:10.1038/nrg3031)
2. Bamshad MJ, Nickerson DA, Chong JX. 2019 Mendelian gene discovery: fast and furious with no end in sight. *Am. J. Hum. Genet.* **105**, 448–455. (doi:10.1016/j.ajhg.2019.07.011)
3. Baxter SM *et al.* 2022 Centers for Mendelian genomics: a decade of facilitating gene discovery. *Genet. Med.* **24**, 784–797. (doi:10.1016/j.gim.2021.12.005)
4. Zschocke J, Byers PH, Wilkie AOM. 2023 Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nat. Rev. Genet.* **24**, 442–463. (doi:10.1038/s41576-023-00574-0)
5. Zwijnenburg PJ, Meijers-Heijboer H, Boomsma DI. 2010 Identical but not the same: the value of discordant monozygotic twins in genetic research. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 1134–1149. (doi:10.1002/ajmg.b.31091)
6. Bruder CE *et al.* 2008 Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82**, 763–771. (doi:10.1016/j.ajhg.2007.12.011)
7. Czyz W, Morahan JM, Ebers GC, Ramagopalan SV. 2012 Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. *BMC Med.* **10**, 93. (doi:10.1186/1741-7015-10-93)
8. Wong AHC, Gottesman II, Petronis A. 2005 Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum. Mol. Genet.* **14**, R11–R18. (doi:10.1093/hmg/ddi116)
9. Waddington CH. 1957 *The strategy of the genes: a discussion of some aspects of theoretical biology.* London, UK: George Allen and Unwin.
10. Wolpert L. 2010 Arms and the man: the problem of symmetric growth. *PLoS Biol.* **8**, e1000477. (doi:10.1371/journal.pbio.1000477)
11. Raj A, van Oudenaarden A. 2008 Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226. (doi:10.1016/j.cell.2008.09.050)
12. Raj A, van Oudenaarden A. 2009 Single-molecule approaches to stochastic gene expression. *Annu. Rev. Biophys.* **38**, 255–270. (doi:10.1146/annurev.biophys.37.032807.125928)
13. Raj A, Rifkin SA, Andersen E, van Oudenaarden A. 2010 Variability in gene expression underliesincomplete penetrance. *Nature* **463**, 913–918. (doi:10.1038/nature08781)
14. Balázsi G, van Oudenaarden A, Collins JJ. 2011 Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910–925. (doi:10.1016/j.cell.2011.01.030)
15. Chubb JR, Liverpool TB. 2010 Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Curr. Opin Genet. Dev.* **20**, 478–484. (doi:10.1016/j.gde.2010.06.009)
16. Corrigan AM, Tunnacliffe E, Cannon D, Chubb JR. 2016 A continuum model of transcriptional bursting. *Elife* **5**, e13051. (doi:10.7554/eLife.13051)
17. Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002 Stochastic gene expression in a single cell. *Science* **297**, 1183–1186. (doi:10.1126/science.1070919)
18. Lin J, Amir A. 2021 Disentangling intrinsic and extrinsic gene expression noise in growing cells. *Phys. Rev. Lett.* **126**, 078101. (doi:10.1103/PhysRevLett.126.078101)
19. Fu AQ, Pachter L. 2016 Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. *Stat. Appl. Genet. Mol. Biol.* **15**, 447–471. (doi:10.1515/sagmb-2016-0002)
20. Vigilante A *et al.* 2019 Identifying extrinsic versus intrinsic drivers of variation in cell behavior in human iPSC lines from healthy donors. *Cell Rep.* **26**, 2078–2087.e3. (doi:10.1016/j.celrep.2019.01.094)
21. Kauffman MA, Calderon VS. 2023 The emergence of genotypic divergence and future precision medicine applications. *Handb. Clin. Neurol.* **192**, 87–99. (doi:10.1016/B978-0-323-85538-9.00013-4)
22. Thaxton C, Goldstein J, DiStefano M, Wallace K, Witmer PD, Haendel MA, Hamosh A, Rehm HL, Berg JS. 2022 Lumping versus splitting: how to approach defining a disease to enable accurate genomic curation. *Cell Genom.* **2**, 100131. (doi:10.1016/j.xgen.2022.100131)
23. Duque KR, Vizcarra JA, Hill EJ, Espay AJ. 2023 Disease-modifying vs symptomatic treatments: splitting over lumping. *Handb. Clin. Neurol.* **193**, 187–209. (doi:10.1016/B978-0-323-85555-6.00020-5)
24. Kingdom R, Wright CF. 2022 Incomplete penetrance and variable expressivity: from clinical studies to population cohorts. *Front Genet.* **13**, 920390. (doi:10.3389/fgene.2022.920390)
25. Miko I. 2008 Phenotype variability: penetrance and expressivity. *Nat. Educ.* **1**, 137.
26. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013 Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130. (doi:10.1007/s00439-013-1331-2)
27. Beales PL, Elcioglu N, Woolf AS, Parker D, Flinter FA. 1999 New criteria for improved diagnosis of Bardet – Biedl syndrome: results of a population survey. *J. Med. Genet.* **36**, 437–446. (doi:10.1136/jmg.36.6.437)
28. Slavotinek A, Beales PL. 2011 Clinical utility gene card for: Bardet – Biedl syndrome. *Eur. J. Hum. Genet.* **19**, 199. (doi:10.1038/ejhg.2010.199)

29. Patton MA, Afzal AR. 2002 Robinow syndrome. *J. Med. Genet.* **39**, 305–310. (doi:10.1136/jmg.39.5.305)
30. Zuniga A, Zeller R. 2020 Dynamic and self-regulatory interactions among gene regulatory networks control vertebrate limb bud morphogenesis. *Curr. Top. Dev. Biol.* **139**, 61–88. (doi:10.1016/bs.ctdb.2020.02.005)
31. Wright S. 1934 The results of crosses between inbred strains of guinea pigs, differing in number of digits. *Genetics* **19**, 537–551. (doi:10.1093/genetics/19.6.537)
32. Wright S. 1934 An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* **19**, 506–536. (doi:10.1093/genetics/19.6.506)
33. Casellas J. 2011 Inbred mouse strains and genetic stability: a review. *Animal* **5**, 1–7. (doi:10.1017/S1751731110001667)
34. Sittig LJ, Jeong C, Tixier E, Davis J, Barrios-Camacho CM, Palmer AA. 2014 Phenotypic instability between the near isogenic substrains BALB/cJ and BALB/cByJ. *Mamm. Genome* **25**, 564–572. (doi:10.1007/s00335-014-9531-1)
35. Simon MM *et al.* 2013 A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol.* **14**, R82. (doi:10.1186/gb-2013-14-7-r82)
36. Wright S. 1920 The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proc. Natl Acad. Sci. USA* **6**, 320–332. (doi:10.1073/pnas.6.6.320)
37. Génin E, Feingold J, Clerget-Darpoux F. 2008 Identifying modifier genes of monogenic disease: strategies and difficulties. *Hum. Genet.* **124**, 357–368. (doi:10.1007/s00439-008-0560-2)
38. Widmayer SJ, Evans KS, Zdraljevic S, Andersen EC. 2022 Evaluating the power and limitations of genome-wide association studies in *Caenorhabditis elegans*. *G3 (Bethesda)* **12**, jkac114. (doi:10.1093/g3journal/jkac114)
39. Chapman K, Ferreira T, Morris A, Asimit J, Zeggini E. 2011 Defining the power limits of genome-wide association scan meta-analyses. *Genet. Epidemiol.* **35**, 781–789. (doi:10.1002/gepi.20627)
40. Visscher PM, Goddard ME, From RA. 2019 Fisher's 1918 paper to GWAS a century later. *Genetics* **211**, 1125–1130. (doi:10.1534/genetics.118.301594)
41. Fisher RA. 1918 The correlation between relatives on the supposition of mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edin.* **52**, 399–433.
42. Norton B, Pearson ES. 1976 A note on the background to, and refereeing of, R. A. Fisher's 1918 paper 'On the correlation between relatives on the supposition of Mendelian inheritance'. *Notes Rec. R. Soc. Lond.* **31**, 151–162. (doi:10.1098/rsnr.1976.0005)
43. CARTER CO. 1961 The inheritance of congenital pyloric stenosis. *Br. Med. Bull.* **17**, 251–254. (doi:10.1093/oxfordjournals.bmb.a069918)
44. Carter CO. 1967 Clinical aspects of genetics, the genetics of common malformations and diseases. *Trans. Med. Soc. Lond.* **83**, 84–91.
45. Falconer DS. 1965 The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76. (doi:10.1111/j.1469-1809.1965.tb00500.x)
46. Swaffer MP, Jones AW, Flynn HR, Snijders AP, Nurse P. 2016 CDK Substrate phosphorylation and ordering the cell cycle. *Cell* **167**, 1750–1761.e16. (doi:10.1016/j.cell.2016.11.034)
47. Patterson JO, Basu S, Rees P, Nurse P. 2021 CDK control pathways integrate cell size and ploidy information to control cell division. *Elife* **10**, e64592. (doi:10.7554/eLife.64592)
48. Coudreuse D, Nurse P. 2010 Driving the cell cycle with a minimal CDK control network. *Nature* **468**, 1074–1079. (doi:10.1038/nature09543)
49. Exelby K, Herrera-Delgado E, Perez LG, Perez-Carrasco R, Sagner A, Metzis V, Sollich P, Briscoe J. 2021 Precision of tissue patterning is controlled by dynamical properties of gene regulatory networks. *Development* **148**, dev197566. (doi:10.1242/dev.197566)
50. Bénazet JD, Bischofberger M, Tiecke E, Gonçalves A, Martin JF, Zuniga A, Naef F, Zeller R. 2009 A self-regulatory system of interlinked signaling feedback loops controls mouse limb patterning. *Science* **323**, 1050–1053. (doi:10.1126/science.1168755)
51. Ott J. 1999 *Analysis of human genetic linkage*. Baltimore, MD: John Hopkins University Press.
52. Terwilliger JD, Ott J. 1994 *Handbook of human genetic linkage*. Baltimore, MD: John Hopkins University Press.
53. Spielman RS, McGinnis RE, Ewens WJ. 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516.
54. Allison DB. 1997 Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690. Erratum in: *Am. J. Hum. Genet*. 1997, **60**(6):1571.
55. Rabinowitz D. 1997 A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47**, 342–350. (doi:10.1159/000154433)
56. Bates S, Sesia M, Sabatti C, Candès E. 2020 Causal inference in genetic trio studies. *Proc. Natl Acad. Sci. USA* **117**, 24 117–24 126. (doi:10.1073/pnas.2007743117)
57. Sanderson E *et al.* 2022 Mendelian randomization. *Nat. Rev. Methods Primers* **2**, 6. (doi:10.1038/s43586-021-00092-5)
58. Burgess S, Small DS, Thompson SG. 2017 A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* **26**, 2333–2355. (doi:10.1177/0962280215597579)
59. Angrist JD, Pischke J-S. 2014 *'Mastering 'metrics: the path from cause to effect,' economics books*, 1st edn. Princeton, NJ: Princeton University Press.
60. Pearl J. 2000 *Causality*. New York, NY: Cambridge University Press.
61. Madelyn Glymour JP, Jewell NP. 2016 *Causal inference in statistics: a primer*. Chichester, UK: John Wiley & Sons Ltd.
62. Wright S. 1921 Correlation and causation. *J. Agricult. Res.* **20**, 557–585.
63. Pearl J, Mackenzie D. 2019 *The book of why*. Harlow, UK: Penguin Books.
64. Meinshausen N, Hauser A, Mooij JM, Peters J, Versteeg P, Bühlmann P. 2016 Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl Acad. Sci. USA* **113**, 7361–7368. (doi:10.1073/pnas.1510493113)
65. Squires C, Wang Y, Uhler C. 2020 *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR, 3-6 August 2020, online*, vol. 124, pp. 809-818. New York, NY: AUAI.