Biol Philos
DOI 10.1007/s10539-014-9472-9

What is a gene for?

Lindell Bromham

Received: 8 December 2014 / Accepted: 9 December 2014

Abstract The word "gene" means different things to different people, and can even be used in multiple ways by the same individual. In this review, I follow a particular thread running through Griffith and Stotz's "Genetics and Philosophy: an introduction", which is the way that methods of investigation influence the way we define the concept of "gene", from nineteen century breeding experiments to twenty-first century big data bioinformatics. These different views lead to a set of gene concepts, which only partially overlap each other, each of which picks up on a different part of gene behaviour, function or scientific utility. This plurality of concepts carries over to the use of the concept of "information" in biology, where the non-overlapping concepts can be connected to whether you view the genome as a blueprint for development, a response to environmental triggers, an engine of heritability, or a document of history.

Keywords Bioinformatics · Mutation · Heredity · Evolution · Genetics

Hence the talk of ''genes for any particular character" ought to be omitted, even in cases where no danger of confusion seems to exist. (Johannsen 1911).

Genetics is a slippery field. It changes so fast that it is nearly impossible to keep up. For example, in revising a textbook on molecular evolution, I have found that some things I had described as impossible or impractical in the first edition were not only possible but routine by the second edition, only 6 years later (Bromham 2008). In such a fast moving field, it is hard to hold onto any concept for long. It's like having your house rebuilt around you while you sleep, so you wake up in a different

L. Bromham (✉)
Macroevolution and Macroecology, Evolution Ecology and Genetics, Research School of Biology, Australian National University, Canberra, ACT 0200, Australia
e-mail: Lindell.Bromham@anu.edu.au

Published online: 17 December 2014

building than the one you went to bed in. It is hardly surprising that terminology has little hope of keeping up with progress.

"Genetics and philosophy: an introduction" by Paul Griffiths and Karola Stotz (Cambridge University Press 2013) captures the excitement and confusion of this fast-moving field by highlighting the way the definition of a gene has changed with the tools used to investigate it. The scope of this book is not so much genetics as a whole but the gene itself, and in particular, the thorny problem of how we are to define the word 'gene'. This book is primarily concerned with the gene as an actor in the construction of phenotype (morphology and behaviour), so discussions in the book revolve around the details of gene expression, development and environmental influence, with relatively little coverage of inheritance per se, and even less on genomic evolution. And, as the authors point out, this is resolutely not a book about ethics, medical genetics or human ancestry.

I should, at this point, warn the reader that I am neither a philosopher nor a geneticist, so should not be relied upon as an expert witness, but as an interested amateur. But the view of the interested amateur is of no small consequence here—the word 'gene' belongs to all of us, and understanding its meaning is critically important for informed discussion of many important ideas in medicine, biology and in wider society. None of us is unconnected to debates about what is a 'gene for'.

What does 'gene for' mean?

Someone I know came home from the doctor saying that she had been told she definitely had the gene for bowel cancer. Since no genetic screening had been performed, the doctor had presumably based this claim on two observations: the patient's uncle and grandfather had both died of bowel cancer, and she had some non-cancerous bowel polyps removed by endoscopy. So, really, the doctor's statement should be read as "I have identified possible risk factors that could increase the probability that you will get bowel cancer, when compared to the prevalence of the disease in the general population".

Actually, there is no 'gene for bowel cancer'. Less than one-third of colorectal cancers have an inherited basis, and the heritable factors associated with these cancers vary greatly in penetrance. Penetrance represents the number of people with a given allele that will express the phenotype of interest: a highly penetrant genetic disorder is one where carrying a particular allele (or alleles) gives you a very high chance of getting the disease. Of the cases of bowel cancer with some evidence of heritability, only one-sixth are due to highly penetrant genetic factors, such as Lynch syndrome which increases the frequency of a number of different cancers (Jasperson et al. 2010): someone with Lynch syndrome has an 80 % chance of having colon cancer during their lifetime. But the majority of bowel cancers associated with hereditary factors are due to 'low-penetrant susceptibility loci'; in other words, genome-wide scans have revealed one or more places in the genome where the particular genetic variant you carry appears to be non-randomly associated with a slightly increased risk of bowel cancer. These loci tend to have 'population attributable risks' in the low single figures, though there is some

evidence that the effect is cumulative over many loci (Bodmer and Bonilla 2008). These risk statistics are difficult to interpret: one common interpretation is that they represents the cases of the disease that might not exist if those risk factors were removed from the population, another interpretation is that the identified genetic risk factors contribute a proportion any one individual's risk of bowel cancer. So removing any one of these 'genes for bowel cancer' from the population could possibly reduce the incidence of bowel cancer by a couple of cases out of every hundred, or result in a small reduction in individual's lifetime risk of bowel cancer.

Does it matter that the doctor used the word 'gene for bowel cancer' when they meant 'family history of'? Should we worry that it will lead people to an incorrect assumption that the trait is unavoidable, or should we focus on the benefits of identifying risk for modifying behaviour? Griffiths and Stotz (G&S) point out that scientists can use several conceptions of the gene more or less simultaneously, yet most of the time there is no confusion because the context gives the meaning. Studies of patient interpretation of 'gene for' statements have similarly suggested that most patients can interpret these as implying risk factors, where outcomes are modulated by both environment and genetic factors (Bates et al. 2003).

Perhaps a greater problem for discourse about genetic inheritance is the lack of a commonly used term for a gene variant. Biologists use 'allele' (yet another term that can be traced to Wilhelm Johannsen's (1911) "Genotype concept of heredity"), but the term is not in common usage in everyday discussions. Without distinct terms for 'gene' (in the sense of a specific locus in the genome) and 'allele' (one particular variant found at that locus), it is more difficult to explain why some people carry risk factors for common traits. We all carry two copies of the *Huntington Disease* (*HD*) gene in every cell in our body (ignoring, for the sake of argument, anucleate blood cells and haploid gametes). People who have a copy of the *HD* gene that has more than 40 repeats of the nucleotides CAG at the beginning of the gene will get Huntington's Disease, a neurodegenerative disorder. So in most people the *Huntington Disease* gene does not give them Huntington's Disease, only those people unlucky enough to inherit an allele of the gene with more than 40 repeats will get the disease. The awkward naming system of genes, where they are known by the outcome of having rare alleles that prevent normal functioning, is a by-product of the history of gene discovery.

Gene discovery

> The "gene" is nothing but a very applicable little word, easily combined with others… As to the nature of the ''genes" it is as yet of no value to propose any hypothesis; but that the notion ''gene" covers a reality is evident from Mendelism. (Johannsen 1911)

The first chapter of "Genetics and Philosophy: an introduction" illustrates the way the concept of the gene has changed over time with the means of investigating heredity. Genes were originally described through the identification of mutations that had a consistent and heritable effect appearance or behaviour. Genes identified

through finding mutants that lacked normal gene function were named for what happens when they don't function, rather than what they do when they work. For example, ordinary *Drosophila* (fruitfly) have red eyes, but individuals in which the *white* gene (*w*) has been knocked out develop white eyes. So a working copy of the *white* gene is responsible for making eyes red.

Genes identified through breeding experiments are often named for their first studied phenotypic effect, so the name may seem incongruous with later research foci. For example, the regulatory gene *dsx* (*doublesex*) switches on different developmental programs to produce distinct morphs within a polymorphic butterfly population, such that individuals within the same population may mimic many different target species (Kunte et al. 2014). Why is a gene involved in butterfly wing patterns called *doublesex*? Because this gene was originally described as a factor controlling sex differentiation, which it does by being differentially spliced, so that males and females produce different transcripts from the same gene, triggering the developmental cascade that results in gender differentiation. *dsx* retains its role in sex differentiation in *Papilio* butterflies while taking on the additional role of determining the development of mimicry patterns.

Gene names carry with them their history of discovery. For example, the *hERG* gene codes for a potassium ion channel. The full name of this gene is *Human Ether-à-go-go-Related Gene*. Why? Because the gene was first described in *Drosophila* in the 1960s, where it was noted that mutant fruitflies had wobbly legs when treated with ether, which reminded the researchers of the dancing at the Whisky A Go-Go nightclub. So, in its name, your *hERG* gene carries with it redundant information, the imprint of a history nearly forgotten. The weight of history on gene names can also carry social baggage. Consider the *Drosophila* gene *fru* which is required for expressing normal mating behaviour (Demir and Dickson 2005). Loss of *fru* function can result in a lack of sex discrimination, so that male fruitflies will court other males. *fru* is an abbreviation of the original gene name "*fruity*" which was a contemporary reflection on this apparent homosexual activity. But as social attitudes changed, the discomfort caused by this name led to it being modified to "*fruitless*", ostensibly as a statement of the low probability of *fru*-mutants leaving any offspring.

The influence of experimental technique on conceptions of the gene was formally recognized by Seymour Benzer, who produced one of the first fine-scale genetic maps, working with bacteriophage genomes (Benzer 1957). Benzer described different forms of genes, and, since he was a physicist by training, he took inspiration from the division of the atom into the subatomic particles neutrons, electrons and protons (Holmes 2006). So he recognized mutons (the gene as the smallest unit of mutation, i.e. a nucleotide), recons (the gene as the smallest unit of recombination, also on the scale of nucleotides) and cistrons (the gene as a unit of biochemical function, typically the protein-coding sequence). In Benzer's scheme, we can see the transition from the atomic mendelian gene of the early geneticists to the continuous sequence-based molecular gene.

Richard Goldschmidt went further than this, and famously rejected the notion of a "corpuscular gene" as a clearly defined body inherited as an indivisible unit (Goldschmidt 1946). He felt that a geneticist could work quite happily considering

loci of mutations that could be recombined away from each other without ever asking what a gene actually was. In other words, we can do genetics on mutons and recons without connecting them to functioning cistrons. As G&S point out in Chapter 2, this view of the gene, while considered deeply heretical at the time, is not a million miles away from a modern perspective of the molecular gene represented by a nucleotide sequence that may not have clearly defined edges. The situation we find ourselves in now is not entirely dissimilar to the one facing the early geneticists, with the meaning of "gene" depending on how you look at the problem.

The mendelian gene

The early gene hunts were breeding experiments that were designed to identify mendelian genes, characterised by predictable patterns of inheritance which result from being independently sorted and sampled each generation. The easiest mendelian markers to detect are single-locus (due to one specific place in the genome) and highly penetrant (if you get the allele you show the phenotype). The great database of human heritable disease is called OMIM: Online Mendelian Inheritance in Man (a title that, by my count, contains two anachronisms, and mendelian isn't one of them). This database lists genetic markers associated with specific diseases, giving the co-ordinates for the location in the genome, a description of the allelic variants, and their effect on the biochemistry and health. In common biological parlance, a mendelian marker is one that follows predictable ratios of inheritance. Expected mendelian patterns of inheritance form the basis of many modern techniques in genetics (e.g. Cussens et al. 2013; Xu and Hu 2010).

A mendelian 'gene' is a pattern of inheritance: it does not need to be a 'gene' in the sense of being a cistron that makes a product or influences phenotype. DNA fingerprinting typically relies on determining the number of short nucleotide repeats at defined loci in the genome (referred to as microsatellites): the combination of alleles, each representing repeat number on one chromosome at a particular locus, is considered to be a near-unique identifier for individuals. These alleles are inherited as mendelian markers, so you get one microsatellite allele from each parent at each locus (with a small possibility of mutation to a different repeat number), which is why DNA fingerprinting can be used to identify genetic relatives. Although some microsatellites do influence phenotype (e.g. Bass et al. 2013), the ones selected for fingerprinting have no apparent role in determining phenotype.

Mendelian markers can be as small as a single SNP (single nucleotide polymorphism), or they can be thousands of bases in length. Human genomes contain large mendelian units that function like recons, stretches of DNA that are rarely divided by recombination thus contain a set of alleles that are consistently inherited together. Recombination can occur anywhere in the genome, including within genes, but in the human genome most recombination occurs at specific hotspots, creating haplotype blocks, typically tens of thousands of bases long. These blocks are remarkably useful, because any identifiable difference within the block acts as a "tag" that can be used to detect the presence of all the linked alleles in the haplotype without actually having to sequence them directly (Cardon and Abecasis

2003). The haplotype block is not strictly speaking a recon in Benzer's sense, because it is not the smallest unit of recombination, but it describes a similar phenomenon. Haplotype blocks can contain multiple genes (in the sense of separate functional cistrons). Because these haplotype blocks behave, most of the time, like good mendelian factors, they provide useful markers to aid genome-wide association studies. Using haplotype maps, researchers can identify genetic risk factors without actually knowing what the 'gene for' actually is.

Furthermore, something that acts as a cistron wont necessarily have a mendelian pattern of inheritance. On the one hand, *Huntington Disease* is like a classic mendelian factor, a single-gene disorder with a dominant effect on phenotype (one copy of the gene is sufficient to ensure expression of the Huntington's Disease phenotype). This is why *HD* was one of the first genes to be sequenced, because its inheritance could be traced over large pedigrees (Wexler et al. 2004). But in some ways it does not behave like a good mendelian gene at all. The age of onset of Huntington's Disease is strongly influenced by the number of repeats of the three-nucleotide CAG motif at the beginning of the gene: the more repeats, the earlier the onset. Repeat number is unstable, so individuals with high repeat number are likely to give rise to offspring with even higher repeat numbers, so that the age of onset can decrease with each generation, but usually only when inherited through the male line [this male-biased mutation might reflect genomic imprinting (Ridley et al. 1988), or may simply be a reflection of the greater chance for mutation in spermatogenesis (Wilson Sayres and Makova 2011)]. This phenomenon, where the severity and age of onset of a genetic disease worsens with each generation, has the somewhat unfortunate name of "genetic anticipation" (Friedman 2011), and is recognized in a number of other human diseases including Fragile X syndrome, one of the most common forms of inherited mental retardation.

To make things even more complicated, a trait that does have a clear mendelian pattern of inheritance might not be the product of a single gene. Consider mimicry in butterflies, which requires the precise co-ordination of many different aspects of wing shape and patterning. Some mimetic butterflies are polymorphic, so that individuals within an interbreeding population can be highly dissimilar as they mimic entirely different target species. Since these mimicry morphs include differences in many aspects of wing shape, pattern and colouring, they must require the action of many different genes. But if all the genes controlling the different aspects of mimicry were each behaving like independent mendelian markers, they would get shuffled and reasserted each generation, and individuals would end up with the eyespots of one target species but the colouring of another, thus ruining the mimicry. It was assumed that butterflies avoided scrambling their mimicry patterns by joining all the relevant genes together to form a supergene, tightly linked together so that they were always inherited as a block. In this way, the alleles could be co-adapted, so that any given butterfly received a co-ordinated set of alleles that all fitted together to make a perfect mimic. A supergene combines many different cistrons into one mendelian factor by functioning as a recon, a block of alleles unlikely to be shuffled away from each other by recombination (Thompson and Jiggins 2014).

As it turns out, that is not how the polymorphic mimetic *Papilio polytes* butterflies have solved the multigene problem. Researchers used modern equivalents of the traditional breeding experiments to map the mimicry trait to a single stretch of the genome that contained the *doublesex* (*dsx*) gene, then showed that the mimetic phenotype correlated closely with the alleles of *dsx*. Just as *dsx* can switch development to a male or female program, it can also switch between different sets of mimicry traits. It seems that a butterfly inherits one version of the instruction "mimic this", which then triggers expression of all the right genes to produce the appropriate phenotype. The discovery that the regulatory gene *doublesex* co-ordinates the expression of the many genes involved in the mimicry phenotype allowed a single-locus explanation of how a multi-factorial trait behaves as a single mendelian factor (Kunte et al. 2014). This makes good biological sense: the supergene solution would always be vulnerable to rare recombination events that broke up the recon, but the regulatory solution ensures that the mendelian inheritance of a multi-gene trait is tied to a single cistron which produces a regulatory protein that acts as a switch for mimicry phenotype.

Uncovering the rich complexity of gene action has not dissolved the mendelian gene, but it has disassociated the mendelian gene from any particular form of DNA sequence. In casual conversation, when biologists speak of a "good mendelian gene" they mean a heritable factor with a clear pattern of inheritance suggesting a single locus that segregates predictably at meiosis and has a consistent effect on phenotype. But as with so many things, it is surprisingly difficult to put a line between a heritable factor that shows good mendelian inheritance and one that does not.

The *GNAS* complex locus in humans, which covers over 17,000 bases on the long arm of chromosome 20, can be considered a good mendelian gene because it has predictable patterns of inheritance. But its influence on phenotype is complicated by differential imprinting: a particular *GNAS* allele will cause a certain disease only when inherited from the mother, and other *GNAS* alleles cause disease only when only when inherited from the father (Weinstein et al. 2004). "Bowel cancer" as a phenotype is not a good mendelian trait, even though there are lots of genetically inherited risk factors each of which may independently behave as a mendelian factor. The *HD* (*Huntington Disease*) gene looks like a good mendelian gene with dominance and complete penetrance, but dynamic mutation of repeats combined with male-driven mutation creates patterns of inheritance more complex than classic mendelian genes.

A mendelian factor is an observation of hereditary patterns, which can be localised to a particular locus in the genome. It might co-incide with a gene that makes a product, or it might be a regulatory signal, or it might do nothing at all. A mendelian factor will be referred to as a 'gene for' when the phenotype tracks the marker. Many people will be comfortable calling a *dsx* allele a 'gene for' a particular mimicry pattern, even though it doesn't directly contribute to building that phenotype, but instead controls a developmental cascade that results in that phenotype. But being a good mendelian factor is not enough to be called a gene: a set of 15 CAG repeats at a non-coding microsatellite loci is not a 'gene for' anything, it's just there in the genome doing nothing but tracking ancestry. G&S emphasize that the 'molecular gene' (the DNA structure that produces a

biochemical product) did not replace the mendelian gene, nor did it simply provide the fine-scale structure for the mendelian gene. Instead, the molecular gene it is a different thing that sometimes overlaps with a mendelian gene, but sometimes does not. The molecular gene is essentially a biochemical description of the cistron (in Benzer's original sense of the word).

The postgenomic gene

While the mendelian gene may have a somewhat casual relationship with the classical molecular gene, the relationship between the molecular gene and what G&S refer to as the 'postgenomic gene' is much simpler. G&S define the postgenomic gene as the DNA sequences that have a linear correspondence to the gene product of interest, wherever these occur in the genome. This way you start with the product and work backwards to the DNA sequences that were needed for its construction. The 'product' is somewhat vague, as it will look different if we focus on the RNA transcript, a processed RNA molecule, or a translated protein. In some cases, the postgenomic gene maps nicely to the old paradigm of one gene, one product. For example, the protein Huntingtin is produced from the *HD* (*Huntington Disease*) gene, and the primary amino acid sequence in the protein bears a linear correspondence to the nucleotide sequence of the 67 exons (protein-coding sequences) in the *HD* gene (these exons make up <10 % of the length of the gene, most of which is intronic sequences that are removed from the transcript before translation). Using Benzer's terminology, *HD* is a cistron, although no one knows exactly what the gene product, Huntingtin, actually does.

But not all genes are as straightforward. As with most things in biology, the more we look the more complexity we find. G&S describe the increasingly complex view of genome function, such as split genes (where sequences transcribed from different parts of the genome are joined to allow construction of a single product) and alternative splicing (where different combinations of sequences from the same location in the genome are used to make different products). These mechanisms, when present, complicate the correspondence between the DNA sequence and the gene product.

Consider the *GNAS* complex locus in humans, which contains 17 different exons. Thanks to a complex set of promoter sequences, and regulated alternative splicing, this single locus can produce many different transcripts, including maternally, paternally, and biallelically expressed transcripts. Not only that, two completely different proteins (ALEX and XLαs) are produced from the same sequence by translating the same mRNA sequence for exon 1 in two different reading frames from alternative start codons (Nekrutenko et al. 2005). There is also evidence that *GNAS* locus can produce two different proteins, Nesp55 and Gαs, from the same bicistronic transcript (Wadhawan et al. 2008).

In this case, what is the 'gene'? If it is the DNA sequence that contains the nucleotide code that is translated, then we might say there is one gene (*GNAS*) with many products. If we define a gene by sequence elements, say by the presence of an active functional promoter sequence, then it gets a bit trickier because there are five

different promoters, that variously specify maternal, paternal and biallelic transcripts. If we take a product-based definition of the gene then we have many different genes co-occuring on the same DNA sequence, due to the inclusion or exclusion of different exons, and from reading exon 1 in two different frames. We would get a different answer again if we counted the number of different transcripts produced from the *GNAS* locus, since some transcripts produce more than one different protein.

But, while the growing catalogue of insanely complicated genes gives some idea of the bewildering complexity of gene action, it doesn't do much to change the actual molecular conception of the gene as the sequence that specifies a product. That gene might be uniquely coded in one DNA sequence, or it may represent a set of possible sequences read from a complex locus. But the po-ge-gene is a cistron, a sequence that provide linear determination of a gene product, albeit one that may be distributed across more than one locus. Each one of these distributed cistrons has a beginning, middle and end, containing the sequences that specify when it is to be expressed and what it should make.

The reactive genome

> In the future attention undoubtedly will be centered on the genome, and with greater appreciation of its significance as a highly sensitive organ of the cell, monitoring genomic activities and correcting common errors, sensing the unusual and unexpected events, and responding to them, often by restructuring the genome. (McClintock 1983)

As with nearly every aspect of biology, the more we find out about genes, the more complex they appear. To capture some of this complexity, G&S define the postgenomic gene as the set of all sequences that bear a linear correspondence to the gene product, even though those sequences could appear anywhere in the genome. This 'distributed specificity' need not fundamentally change our attitude towards genes: the molecular gene is still there, even if it is in pieces. But G&S suggest a more fundamental change to the conception of the gene. If the gene consists of all the sequences needed to make a gene product, then it can also be said to depend not only on the coding sequence but also the regulatory sequences that modulate gene expression. All genes require regulation so that they are expressed in the right place at the right time in the right amount. G&S emphasise that because factors other than coding sequences a specific difference to the linear order of elements in a gene product, these non-gene elements should therefore be regarded as having "Crick information" that contributes to the specificity of gene products: "the sequence of gene products, even if defined narrowly as the linear order of their elements, derives from regulatory mechanisms as well as the DNA sequence from which the produce is initially transcribed" (p97). The chains of regulatory action can be long and complex and take in many different loci in the genome. The mimicry genes in butterflies require the protein product of the *dsx* locus in order to be expressed, just

as *dsx* itself requires regulatory sequences that allow it to be expressed at the right time to specify both sex differentiation and mimicry traits.

For some traits, gene regulation must also be responsive to the environment in order to produce a functioning phenotype. 'Environment' includes not only conditions external to the organism, but the state of the cell and the behaviour of neighbouring cells. Because the environment modulates gene expression, altering phenotype, G&S reject the notion of the genome as the driver of development, the homunculus in charge of the vehicle. They base their argument on the concept of a reactive genome: as gene expression is modulated by environmental conditions, genes can be said to contribute to phenotype but do not wholly create it. Of course, this is not a new idea: Johannsen (1911) considered that the phenotype was generated by "reactions of the genotypical constituent", though the genotypic information remained constant, unchanged by the conditions that its parents were exposed to. But the reactive genome has gained momentum from the new genetics through the study of the complexities of gene regulation and expression. G&S suggest that the new genetics leads to a rejection of the claim that DNA sequences are the only important source of biological specificity of gene products.

G&S take evidence of the complexity of the interaction between the genotype and environment, and the complex web of regulatory factors that are required to regulate gene expression, as evidence in favour of the reactive genome. In this regard, they draw heavily on analysis of "non-coding" elements of the genome. For example, they state that the combinatorial complexity engendered by gene regulation "resolves the N-value paradox that the number of protein-coding genes of an organism doesn't seem to be correlated with its complexity". But it is too early to call resolution on this debate now. While it has been shown that the number of different transcripts exceeds the number of genes recognised on the basis of genomic sequence analysis alone, it has not yet been shown that the degree to which genes can be differentially regulated, or the number of different transcripts that can be produced, is causally linked to organismal complexity or lineage evolvability. The set of functionally relevant elements (inferred from studying the phenotype) is a subset of the set of actively transcribed elements (inferred from the transcriptome) which is a subset of all possible elements (inferred from sequence analysis), but what we don't know is the relative proportions of each these categories. Even if these were known, we make a conceptual leap if we assume that operational complexity—the number of working parts or connections in a system or the diversity of its products—must necessarily result in a better operational outcome (Parkinson and Osborn 1957).

Given the emphasis in earlier chapters on the way that modes of investigation have shaped notions of the gene, I was surprised that the 'new genetics' was not subject to greater critical scrutiny. Genome analysis has not opened a window through which we have a crystal clear view of reality. It has given us an absolutely enormous amount of data to play with, but that is quite a different thing. While some biologists feel that the analysis of 'non-coding' DNA paints a new picture of a complex and highly tuned genome, others feel that at this stage the story is more hype and inference than dispassionate observation (e.g. Doolittle 2013; Eddy 2013; Graur et al. 2013). The point here is not about who, in the end, will turn out to be

right, but that here again we have the concept of the gene seen through the filter of current technology and research environment. The post-post-genomic gene is a child of genomics and big data bioinformatics, and is young yet. Who knows what the gene will look like when these fields mature beyond their nascent excitement.

Several important points deserve greater emphasis in any discussion of the meaning of the new genetics for our understanding of the relationship between genotype and phenotype. The first is the potential for error in these analyses. Some error is technical, for example high-throughput sequencing techniques are error prone, and gene prediction and annotation methods vary greatly in their accuracy (Yandell and Ence 2012). These errors are likely to become corrected over time, but given the vast amount of data generated, even low levels of error can end up speaking with a loud voice. But some of the error may come from the genome itself. Genomes are wonderful, crazy, complex, messy things. Not everything they do is a Panglossian perfection (Lynch 2007). In other words, we need to consider that some of the error in the system is biological. Alternative splicing can produce multiple functionally relevant transcripts from the same DNA sequence, such as the ALEX and XLαs proteins derived from the *GNAS* locus. But this does not mean that all alternative transcripts are functionally relevant to phenotype. Splicing is error-prone, so some alternative transcripts are likely to arise from transcription errors, which may produce either isoforms that function as well as the canonical splice variant, or, perhaps much more likely, non-functional transcripts that cannot make working gene products (Pickrell et al. 2010; Sorek et al. 2004).

In any case, more complex genetic processing does not always make for a more sophisticated or reactive phenotype. Some single-celled organisms have insanely complicated systems of genome management and post-transcriptional processing, but it's not obvious that they are more complex, reactive, or evolvable than organisms with more boring genomes. Take, for example, the mitochondrial genome of the single-celled *Diplonema papillatum*. The mitochondrial genome is usually held in a single DNA molecule, neatly tied in a circle. But in *Diplonema papillatum*, the mitochondrial genome exists on lots of tiny linear fragments, each much smaller than a gene, so co-ordinated trans-splicing (joining different mRNA transcripts) is required in order to make functioning mitochondrial products such as cytochrome oxidase. Is this a sign of adaptive complexity, or an example of an organism surviving in spite of a truly ridiculous genomic arrangement? Might we not ask the same question of our own complex genome management? An increasing number of human diseases are now attributed to splicing errors, so even if complex machinery makes us more wonderful, it may do so at terrible cost.

Genes as information

> What is transmitted from one generation to another is not the form and substance of a pterodactyl or a mammoth, but primarily the capacity to synthesize particular proteins. The development of specific form is a

> consequence of this capacity, and the capacity itself depends on the self-replication properties of DNA (Maynard Smith 1958)

Much of the recent excitement in the new genetics has focussed around interpretations of the role of the fraction of the genome that does not seem to bear a linear correspondence to any gene product. This is sometimes referred to as 'non-coding DNA'. But the term 'non-coding DNA' is at least as vexatious as the word 'gene'. G&S state that the genome cannot be said to 'code for' the phenotype. They follow Godfrey-Smith (2000) in rejecting loose use of 'code for' to refer to any outcome attributable to genetic factors, and instead take a strictly functional view of coding that restricts the concept to concerning the way that some nucleotide sequences can be said to code for an amino acid sequence through the act of translation (the matching of specific amino acids to three-base codons in the messenger RNA). Classical protein-coding genes clearly 'code for' gene products: the nucleotide sequence in a functional protein-coding exon codes for the amino acid of the corresponding protein. Surely the same can be said for the nucleotide sequence of an RNA gene, where the nucleotides in the DNA sequence in the genome code for a complementary series of ribonucleotides in a functional RNA molecule. But after that things start to get blurry, and the blurring of definitions of 'function', 'coding' and 'information' are the source of some degree of angst and misunderstanding.

G&S distinguish two senses in which information is commonly used in genetics: information about genes (in the sense of data to be stored, communicated or manipulated), and information in genes (in the sense that genes can be said to embody messages in the classic, information-theory sense). Both these senses are in common usage, but it is the second that G&S focus on, arguing that in most usages, the use of the word 'information' to describe the functioning of genetic systems is at best a sloppy metaphor. They conclude that the "only really substantial sense in which the genes carry information is their role in templating for gene products"—in other words, only the translation of nucleotide codons to amino acid sequence is truly a transfer of information. Here G&S target a particular idea of information as being directly concerned with biological specificity: a particular DNA sequence specifies a particular amino acid sequence through the mechanisms of transcription (DNA to RNA) and translation (RNA to protein). They are adamant that genes cannot be said to 'code for' any aspect of phenotype because the code written in DNA is a code for a series of amino acids only, not for phenotype as a whole, which may (sometimes) be affected by external factors.

It is surely no co-incidence that many of the cited defences of the use of the word 'information' in genetics are from biologists, while many of the critiques are from philosophers. As a non-philosopher, I'm afraid I have to admit to being somewhat bewildered by the discussion surrounding the sense, if any, in which the genome can be said to carry information. I can appreciate that biologists are failing to recognize the nuance in the concept of 'information' that might be much more salient to a philosopher. So I wonder: if biologists use 'information' to mean different things at different times, is this worse than using the word 'gene' to mean different things, as long as the meaning in any given statement is clear enough? Are there are cases

where confusion is wrought, or key ideas obscured, when biologists use 'information' in an unlicensed way?

Perhaps the most critical area where sorting out what we mean by 'information' may have significant impact on the way we talk about genes is in evaluating the role of non-genetic forms of inheritance. G&S use the example that antibodies in a mother's milk can be said to carry teleosemantic information, because they are designed to bring about a certain developmental change in her offspring. The antibodies are objects that by their physiochemical actions bring about a physiological effect, and these antibody molecules will eventually decay and not be passed on to the mother's grandchildren. The offspring does not receive the instructions how to make those antibodies from the milk, but it does receive those instructions in the genome, and it is those instructions that are passed down through the generations.

The same argument applies, though less cleanly, to epigenetic modification. Take, for example, methylation of DNA sequences, where a methyl group (CH_3) is added to particular nucleotides in a DNA sequence. It is referred to as "epigenetic" because the actual base sequence of the DNA is not altered, however methylation can function to turn genes on or off. In this sense, particular instances of methylation have clear teleosemantic effect on development. For example, because the *GNAS* locus is differentially methylated, certain alleles will give rise to different disease phenotypes depending on whether they have been inherited from the mother or the father. Another *GNAS* disease phenotype (with the mellifluous name of pseudo-hypoparathyroidism type 1B) arises from faulting epigenetic marking such that both alleles have been imprinted as paternal in origin (Weinstein et al. 2004). Methylation patterns are commonly re-established every generation, though in some cases they may be passed on to children or grandchildren. Like the antibodies in milk, it is the physical form of the methylation that is active in specifying phenotype, and the physical form is transient. Of course, the instructions for methylation can be inherited, which is why loci like *GNAS* are consistently and predictably methylated each generation. But the form of methylation that influences development belongs, by and large, to phenotype, not genotype.

Whether you consider epigenetic modification constitutes genetic information will depend on where you rest your gaze. If you are interested in the expression of the genotype in development, then epigenetic modification will be seen as information that contributes critically to the specification of phenotype, by modifying which genes are expressed and how. But if you are interested in the genome as the carrier of heritable information over many generations, then epigenetic modification itself does not appear to have much longevity over evolutionary timescales. Epigenetic modification alters the expression of the message without changing the message itself.

The medium and the message

However much the environment modifies development, such that we cannot say that the phenotype is the product of genotype alone but of many different factors acting together, there is something special about genotype that phenotype does not have.

The insight of Weismann's "continuity of the germplasm", Johannsen's "genotype-conception", and Crick's "central dogma" is that the gene is not the physical DNA molecule, but the message it contains (Crick 1970; Johannsen 1911; Weismann 1882). The embodiment of the genome in the cell may be in nucleotides, but it is still the genome when written in letters on a page, bands on a gel, or in the flashes of light generated by a pyrosequencing machine. In this sense, the online sequence repository GenBank contains the human genome in the same sense that one of my skin cells contains the human genome. The critical difference is that the DNA in my skin cell is both genotype (message) and phenotype (material) at the same time. That is because the physical genome has much more to it than the message written in nucleotides: its got all the histone modification, packaging, methylation and other epigenetic factors needed for the expression of the genotype to make the phenotype.

To highlight the distinction between genome as message and genome as material, consider DNA sequences that do not act solely as a template but also act directly as a binding site for other molecules which recognize the particular sequence of nucleotides as a kind of "docking station". We could communicate the message in signal lanterns or semaphore flags, so that someone else could write the genomic sequence (Fig. 1), but only when that sequence has the form of a nucleotide sequence can it act as a binding site for regulatory factors, through complementary base pairing. This means that DNA control elements that serve as recognition sequences are both genotype (a message that can be copied from one form to another) and phenotype (a structure that contributes to form and function). Like anything else in biology, the distinction between genotype and phenotype is bound to be blurry.

"Genetics and Philosophy: an introduction" is focussed on the role of the genotype in specifying phenotype, through the process of development. The role of the gene as the engine of heritability is given less attention, except when arguing for recognition of other non-genetic factors in heritability, such as environmental influences. As an evolutionary biologist, it is not surprising that my focus is on

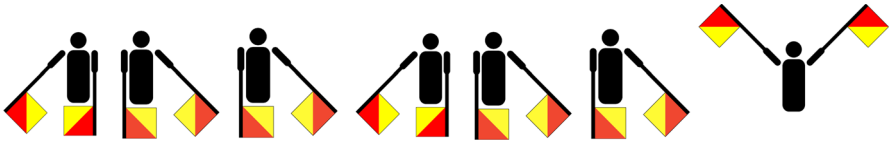

Fig. 1 A Shine-Dalgarno sequence, written in semaphore. Shine-Dalgarno sequences act as binding sites involved in the initiation of translation. This sequence fulfils the role of genotype, because it represents information that I could send it to a lab and, using an appropriate code, they could translate it exactly into nucleotides which they could put into a living bacterial genome where it would function perfectly well, and be inherited by all that bacterium's descendants. Note that this sequence can be translated using canonical semaphore to the RNA nucleotides AGGAGGU, but equally it could be translated to DNA nucleotides by reading the "flags-at-ten-to-two" symbol as thymine (T). We can also translate this sequence into the binding site of the 16s ribosomal RNA if we read the sequence from right to left, and translate the first character on the right (ten-to-two) as adenine (A), the second character (twenty–past-six) as cytosine (C) and the fourth character (half–past-eight) as uracil (U), thus reading ACCUCCU. The information in the sequence is independent of the actual symbols, and depends entirely on the code used to translate it. But this sequence would only function as phenotype when translated into ribonucleotides from left to right, because ribosomal RNA molecules do not bind to flags

heritability across vast numbers of generations. But patterns of inheritance do also have a practical impact on the role of the gene in formation of phenotype. For example, if a postgenomic gene has distributed specificity, in that critical sequences are spread over the genome, then it cannot be part of the same recon, so any allele in one part of the gene will be inherited independently from alleles in the other parts of the genes (Lemos et al. 2008). On the other hand, any loci that are inherited together in a haplotype block "recon" will have their alleles glued together and inherited as one. However these particular sequences act in development, their arrangement in the genome can influence the way they are inherited. This is why it is interesting to ask how *Papilio* butterflies have solved the problem of having polymorphic, polygenic phenotypes: how can they ensure reliable inheritance of mimicry phenotypes that are specified by many different genes?

This sense of genes as inheritance has been captured in the "transmission sense of information", which focuses on the transmission of information from one generation to the next (Bergstrom and Rosvall 2011). G&S describe this concept of information as being directed at answering a specific evolutionary question: why do "organisms have phenotypes that are well designed for environments of which they had no experience during their development" (p. 170). In this sense, the transmission concept focuses entirely on the adaptive part of the genome: it is described in terms of the role of selection in determining the information that must be passed from one generation to the next, as a signal of an appropriate way to develop in the environment likely to be encountered. This description of the transmission sense of information rests on genetic information having a "teleofunction" (Shea 2011): its purpose is to inform future generations.

But there is an awful lot of the genome that doesn't seem to do anything like shape the phenotype to suit the environment. Some of these non-adaptive sequences are apparently useless, some are probably harmful, and others are following their own adaptive agenda rather than that of the owner of the genome. These non-adaptive parts do not have teleofunction in the sense identified above, because they are not "for" telling future generations how to grow and behave (though of course transposable elements are serving their own purposes in this regard). But all DNA sequences, functional or not, carry historical information incidentally, as a by-product of descent with modification.

The non-adaptive parts of the genome may not be of a great deal of interest to someone primarily concerned with the role of the genotype in shaping phenotype. But they are very interesting to someone like me, who uses DNA primarily as a source of information on evolutionary past and processes. Compare the *HD* gene in a human and a pufferfish (*Fugu*). The 67 exons are recognizably similar and they form a similar protein which quite probably does much the same thing in a fugu as it does in a human (whatever that is). But the fugu *HD* gene is a tenth the size of the human version, because the introns are shorter (Baxendale et al. 1995). This follows the pattern found across the fugu genome of a significantly reduced amount of repetitive and non-functional DNA. It's a fair bet that we could trim some slack out of the introns of the human *HD* gene and still have it function perfectly well, as it does in the fugu. So the large amounts of intronic sequence in the human *HD* gene

are telling us something very interesting about the human genome, even though most of that sequence doesn't seem to code for anything.

G&S criticize Konrad Lorenz's distinction of "phylogenetic information" on the grounds that is has nothing unique to say about the way that genes are translated into phenotype: the same sequence would have the same effect on development, regardless of its history. But it is the historical information in sequences that I study in order to shine light on evolutionary patterns and processes: it matters to me if the *atp1* gene in a parasitic plant has actually been derived as a horizontal transfer from its host plant, rather than inherited from its ancestor, even if it functions just the same (Barkman et al. 2007; Nickrent et al. 2004). The gene as a carrier of information about history may sometimes overlap with the gene as a unit of function, but sometimes it will not. Your microsatellite alleles make no difference to your phenotype but their presence in the genome allows me to trace your ancestry. Perhaps this is 'gene' in the sense of 'selfish gene', because all that matters is that the sequence is transmitted across generations, it doesn't matter why or how or by whom (Sterelny and Kitcher 1988). So there is a type of information in the genome that doesn't determine aspects of phenotype, nor is it involved in transmission of useful instructions between generations, but is nonetheless the information that I make a living out of deciphering. I will leave it up to philosophers, or renegade physicists, to decide if it needs a separate name (heriton? pass-it-on?).

Are all forms of inheritance equal?

G&S consider that the difference between the heritability of genes and nongenetic factors, such as environments or genome modifications, is one of degree not kind. This may be true, but it seems to me that the degree of difference between genomic heritability on the one hand and environmental and "exogenetic" heritability on the other is so vast that there is blue sky between them. Some forms of epigenetic modification are systematically reset every generation: in such cases, epigenetic factors are important actors in development, and the genes specifying epigenetic marking are a critical part of the adaptation of the genome, but the marking itself is a part of phenotype and not passed through the genotype. But some cases of epigenetic modification have been shown to carry over several generations. For example, plants raised under conditions of water stress can modify the expression of genes by methylation to limit water loss, and this "hardening" can be passed on to their offspring through the epigenetic marking, but this inheritance is easily disrupted by change in environmental stress (Tricker et al. 2013). Perhaps evidence of inheritance over longer timespans will soon be revealed (Grossniklaus et al. 2013), but at this stage the evidence for a reach beyond the grandchildren is limited. By contrast there are some gene sequences that have been maintained, modified and inherited for billions of years (Bromham 2000). So even if there is no technical difference in the way we would describe both genetic and nongenetic factors as being heritable, in practise we end up with strikingly different consequences. Nongenetic inheritance is clearly important in determining phenotype, but may be

less so in stably transmitting variation across evolutionary timescales because it can be easily overridden or changed.

Similarly, the environmental niche is heritable and important to phenotype: we should never allow ourselves to ignore that most offspring inherit more from their parents than their genes, even those that never actually meet their mum and dad. But the continuity of environment can be altered much more easily and substantially than the genetic inheritance. I can take a couple of grey squirrels to Antarctica and feed them on peanut butter toast, and, as long as I meet their basic biological needs of warmth, food, shelter and water, they will continue to be grey squirrels and raise grey squirrel offspring. Which is not to say they will be unaffected: the baby squirrels may grow chubby on a junk food diet, they may even grow thicker fur coats against the cold, and they may have to adjust their behaviour to a different seasonal calendar. But they will continue to produce generations of nearly identical grey squirrels, despite the massive change in environment. Some species are probably more sensitive to environmental change (for example narrowly distributed alpine plants) and some less sensitive (think of the range of environments that a Norway rat can happily breed and persist in).

But genetic inheritance cannot be so easily broken. Apart from the resampling of alleles each generation, the occasional new mutation, and the very small chance of picking up some horizontally acquired DNA, our Antarctic squirrels continue to receive basically the same genome despite the overwhelming change in environment, and that is why they can continue to produce baby squirrels. In all the talk of weird and wonderful genes, funky patterns of inheritance and reactive gene action, we must not lose sight of the fact that many genes have a much more predictable and pedestrian influence on phenotype. If you inherit a copy of the *HD* gene with 55 repeats, you will get Huntingdon's disease, regardless of your environment. A fly that only has mutant versions of the *white* gene will have white eyes. While some genetic factors provide relatively little information about phenotypic outcomes, like whether you will get bowel cancer, some sequences provide much more reliable information. For example eye colour can be predicted from a blood sample without knowing the environment the person grew up in (Walsh et al. 2011). The same can rarely be said, to anything like the same degree, for the environment: you can't often predict phenotype from information about the environment without also having the genetic information. You can't tell me someone's eye colour just by knowing where they grew up, unless I also give you information about their genetic inheritance (e.g. their parents' eye colour or ethnicity).

The distinction between genes as actors in development and genes as inherited sequences is an important distinction to make. For example, there are developmental genes that play key roles in the development of basic body plan, such as limbs in animals or petals in plants. Just because these genes are essential in the individual development of body plan does not mean they must have played a key role in the evolution of differences in body plan between lineages, which may be due to many more changes in multiple, downstream genes. Yet observation of the large changes in individual phenotype that can be achieved by manipulating single developmental genes have led to suggestions that changes in regulation have driven the evolution of

major differences in animal body plans, and these hypotheses have been used to argue for a new extended evolutionary synthesis (Carroll 2000; Pigliucci 2007).

Is the synthesis being extended?

The penultimate chapter, "The evolving genome", discusses the role of developmental plasticity in individual adaptation to environments, exogenetic inheritance and the origins of novelty, united under an umbrella of recent challenges to the modern synthesis view of evolution. It is not clear to me the extent to which the complexity of gene action revealed by modern techniques, and the role of environment in shaping phenotype, contributes to an extended synthesis. If 'extended' means a richer appreciation of the tangled web of interactions between genes, development and environment, then there are very few fields of biology that do not contribute to such an extension: the more we find out, the more we know, and the more fiendishly complicated everything looks. In this sense, developmental biology is no more privileged than ecology, physiology, biochemistry, behavioural studies and so on. There aren't many fields of biology where detailed investigation reveals that things are actually much simpler than we first supposed. If 'extended' means that knowledge of the complex interactions between genotype and phenotype cannot be accommodated in current evolutionary theory, then there are no concrete examples here that illustrate how this new knowledge does not fit into the established framework.

For example, claims that Darwin did not address the 'origin of the fittest' are no more substantial than arguments that he did not address the origin of new species (Mallet 2008, 2010). Darwin clearly set out his hypothesis for both the origin of new taxa and the origin of novelty, based on the accumulation of small heritable variations always present in all populations (Darwin 1859). The neo-Darwinians followed suit by demonstrating that the accumulation of many tiny changes over many generations could indeed produce large change over evolutionary time. You might dispute their hypothesis that both the origin of species and the origin of novelty can be explained through the accumulation of many small variants, but it is unfair to say that they did not put forward an explanation. Of course, we now know much more about evolutionary genetics than Darwin did. It is now clear that new species can arise by few genetic changes that engender genetic isolation, such as "magic" speciation traits that influence both mate choice and ecological preference (Servedio et al. 2011), speciation by hybridisation (Abbott et al. 2013), and "instant speciation" by changes in ploidy (Ainouche and Jenczewski 2010). There is no doubt that Darwin and his followers had an incomplete view of speciation which has since been enriched, but that is not the same thing as saying that these topics were ignored or that the basic theoretical framework developed has been shown to be inadequate.

Similarly, the importance of change in regulatory genes over time has enriched our understanding of the genetic basis of novelty (e.g. Brunetti et al. 2001; Kunte et al. 2014). So these fields of research do enlarge the explanatory scope of evolutionary biology because the more we find out, the more we know. But does this

mean that machinery of evolutionary theory that was built to operate on simpler information cannot also operate on more complex inputs? If we are talking about heritable traits that arise by mutation in one or few individuals then rise in frequency by some combination of selection and drift until they are carried by many or all members of a population, then surely we can accommodate those traits in the theoretical machinery of Darwinian evolution without the need for any dramatic rewiring. If not, then which part of this scenario are we rejecting?

Conclusion

> No one definition has as yet satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of a species (Darwin 1859)

Biologists have to learn to deal with the fact that few definitions are absolute, partly because we can't agree on the important aspects, but also because reality is a mess. Every naturalist knows there is such a thing as a species, yet the only thing we can be certain of is that no one definition will draw a line around all those things that naturalists might want to call a species. It may be that a species of bacteria cannot be usefully defined in the same way as a species of mammal.

Every geneticist knows what they mean by a gene, but do they mean the same thing as the person they are talking to? Given that G&S illustrate that what a gene looks like depend on who is looking at it, a gene may well need to be defined (if at all) with respect to an observer. What a bioinformatician calls a gene may or may not co-incide with what a biochemist or population biologist calls a gene. In some cases, all definitions of genes may hit the same thing, and in some cases they wont. Biology, alas, will not be fit into boxes. The final chapter of "Genetics and philosophy: an introduction" provides a balanced and clear account of how the concept of gene has varied over time, and can take different meanings in contemporary discussion. What remains to be seen is how these ideas will look in ten years time as the dust from the explosion of new genetics settles and we can survey the wreckage. What will be left of the gene, and what new ideas will take its place?

Acknowledgments Thanks to Brett Calcott, Kim Sterelny, Paul Griffiths, Karola Stotz and Emma Day for their feedback, and to Silo Kingston for providing an inspirational environment.

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, Butlin RK, Dieckmann U, Eroukhmanoff F, Grill A, Cahan SH, Hermansen JS, Hewitt G, Hudson AG, Jiggins C, Jones J, Keller B, Marczewski T, Mallet J, Martinez-Rodriguez P, Möst M, Mullen S, Nichols R, Nolte AW, Parisod C, Pfennig K, Rice AM, Ritchie MG, Seifert B, Smadja CM, Stelkens R, Szymura JM, Väinölä R, Wolf JBW, Zinner D (2013) Hybridization and speciation. J Evol Biol 26(2):229–246

Ainouche ML, Jenczewski E (2010) Focus on polyploidy. New Phytol 186(1):1–4

Barkman TJ, McNeal JR, Lim SH, Coat G, Croom HB, Young ND, dePamphilis CW (2007) Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. BMC Evol Biol 7:248

Bass C, Zimmer CT, Riveron JM, Wilding CS, Wondji CS, Kaussmann M, Field LM, Williamson MS, Nauen R (2013) Gene amplification and microsatellite polymorphism underlie a recent insect host shift. Proc Natl Acad Sci USA 110(48):19460–19465

Bates BR, Templeton A, Achter PJ, Harris TM, Condit CM (2003) What does "a gene for heart disease" mean? A focus group study of public understandings of genetic risk factors. Am J Med Genet Part A 119(2):156–161

Baxendale S, Abdulla S, Elgar G, Buck D, Berks M, Micklem G, Durbin R, Bates G, Brenner S, Beck S (1995) Comparative sequence analysis of the human and pufferfish Huntington's disease genes. Nature Genet 10(1):67–76

Benzer S (1957) The elementary units of heredity. In: McElroy WD, Glass B (eds) The chemical basis of heredity. Johns Hopkins University Press, Baltimore, pp 70–93

Bergstrom CT, Rosvall M (2011) The transmission sense of information. Biol Philos 26(2):159–176

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nature Genet 40(6):695–701

Bromham L (2000) Conservation and mutability in molecular evolution. Trends Ecol Evol 15:355

Bromham L (2008) Reading the story in DNA: a beginner's guide to molecular evolution. Oxford University Press, Oxford

Brunetti CR, Selegue JE, Monteiro A, French V, Brakefield PM, Carroll SB (2001) The generation and diversification of butterfly eyespot color patterns. Curr Biol 11:1578–1585

Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trend Genet 19(3):135–140

Carroll RC (2000) Towards a new evolutionary synthesis. Trends Ecol Evol 15:27–32

Crick FHC (1970) Central dogma of molecular biology. Nature 227:561–563

Cussens J, Bartlett M, Jones EM, Sheehan NA (2013) Maximum likelihood pedigree reconstruction using integer linear programming. Genet Epidemiol 37(1):69–83

Darwin C (1859) The origin of species by means of natural selection: or the preservation of favoured races in the struggle for life. John Murray, London

Demir E, Dickson BJ (2005) Fruitless splicing specifies male courtship behavior in *Drosophila*. Cell 121(5):785–794

Doolittle WF (2013) Is junk DNA bunk? a critique of ENCODE. Proc Natl Acad Sci USA 110(14):5294–5300

Eddy SR (2013) The ENCODE project: missteps overshadowing a success. Curr Biol 23(7):R259–R261

Friedman JE (2011) Anticipation in hereditary disease: the history of a biomedical concept. Hum Genet 130(6):705–714

Godfrey-Smith P (2000) On the theoretical role of "genetic coding". Philos Sci 67:26–44

Goldschmidt RB (1946) Position effect and the theory of the corpuscular gene. Cell Mol Life Sci 2(7):250–256

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E (2013) On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5(3):578–590

Grossniklaus U, Kelly WG, Ferguson-Smith AC, Pembrey M, Lindquist S (2013) Transgenerational epigenetic inheritance: how important is it? Nature Rev Genet 14(3):228–235

Holmes FL (2006) Reconceiving the gene: Seymour Benzer's adventures in phage genetics. Yale University Press, New Haven

Jasperson KW, Tuohy TM, Neklason DW, Burt RW (2010) Hereditary and familial colon cancer. Gastroenterology 138(6):2044–2058

Johannsen W (1911) The genotype conception of heredity. Am Nat 45(531):129–159

Kunte K, Zhang W, Tenger-Trolander A, Palmer D, Martin A, Reed R, Mullen S, Kronforst M (2014) Doublesex is a mimicry supergene. Nature 507(7491):229–232

Lemos B, Landry CR, Fontanillas P, Renn SCP, Kulathinal R, Brown KM, Hartl DL (2008) Evolution of genomic expression. In: Pagel MD, Pomiankowski A (eds) Evolutionary genomics and proteomics. Sinauer Associates, Sunderland

Lynch M (2007) The origins of genome architecture. Sinauer Associates, Sunderland

Mallet J (2008) Mayr's view of Darwin: was Darwin wrong about speciation. Biol J Linn Soc 95:3–16

Mallet J (2010) Why was Darwin's view of species rejected by twentieth century biologists? Biol Philos 25(4):497–527

Maynard Smith J (1958) The theory of evolution. Penguin, London

McClintock B (1983) The significance of responses of the genome to challenge. In: Federoff N, Botstein D (eds) Nobel prize lecture: reprinted in the dynamic genome: Barbara McClintock's ideas in the century of genetics. 1992. Cold Spring Laboratory Press, New York

Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD (2005) Oscillating evolution of a mammalian locus with overlapping reading frames: an XLas/ALEX relay. PLoS Genet 1(2):e18

Nickrent D, Blarer A, Qiu Y-L, Vidal-Russell R, Anderson F (2004) Phylogenetic inference in Rafflesiales: the influence of rate heterogeneity and horizontal gene transfer. BMC Evol Biol 4:1471–2148

Parkinson CN, Osborn RC (1957) Parkinson's law, and other studies in administration. Houghton Mifflin, Boston

Pickrell JK, Pai AA, Gilad Y, Pritchard JK (2010) Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet 6(12):e1001236

Pigliucci M (2007) Do we need an extended evolutionary synthesis? Evolution 62:2743–2749

Ridley R, Frith C, Crow T, Conneally P (1988) Anticipation in Huntington's disease is inherited through the male line but may originate in the female. J Med Genet 25(9):589–595

Servedio MR, Doorn G, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: "magic" but not rare? Trends Ecol Evol 26(8):389–397

Shea N (2011) What's transmitted? inherited information. Biol Philos 26(2):183–189

Sorek R, Shamir R, Ast G (2004) How prevalent is functional alternative splicing in the human genome? Trends Genet 20(2):68–71

Sterelny K, Kitcher P (1988) The return of the gene. J Philos 85:339–361

Thompson M, Jiggins C (2014) Supergenes and their role in evolution. Heredity 113:1–8

Tricker PJ, López CMR, Gibbings G, Hadley P, Wilkinson MJ (2013) Transgenerational, dynamic methylation of stomata genes in response to low relative humidity. Int J Mol Sci 14(4):6674–6689

Wadhawan S, Dickins B, Nekrutenko A (2008) Wheels within wheels: clues to the evolution of the Gnas and Gnal loci. Mol Biol Evol 25(12):2745–2757

Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M (2011) IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet 5(3):170–180

Weinstein LS, Liu J, Sakamoto A, Xie T, Chen M (2004) Minireview: GNAS: normal and abnormal functions. Endocrinology 145(12):5459–5464

Weismann A (1882) Studies in the theory of descent. AMS Press, New York

Wexler NS, Lorimer J, Porter J, Gomez F, Moskowitz C, Shackell E, Marder K, Penchaszadeh G, Roberts SA, Gayan J, Brocklebank D, Cherny SS, Cardon LR, Gray J, Dlouhy SR, Wiktorski S, Hodes ME, Conneally PM, Penney JB, Gusella J, Cha JH, Irizarry M, Rosas D, Hersch S, Hollingsworth Z, MacDonald M, Young AB, Andresen JM, Housman DE, De Young MM, Bonilla E, Stillings T, Negrette A, Snodgrass SR, Martinez-Jaurrieta MD, Ramos-Arroyo MA, Bickham J, Ramos JS, Marshall F, Shoulson I, Rey GJ, Feigin A, Arnheim N, Acevedo-Cruz A, Acosta L, Alvir J, Fischbeck K, Thompson LM, Young A, Dure L, O'Brien CJ, Paulsen J, Brickman A, Krch D, Peery S, Hogarth P, Higgins DS Jr, Landwehrmeyer B (2004) Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc Natl Acad Sci USA 101(10):3498–3503

Wilson Sayres MA, Makova KD (2011) Genome analyses substantiate male mutation bias in many species. Bioessays 33(12):938–945

Xu S, Hu Z (2010) Mapping quantitative trait loci using distorted markers. Int Journal Plant Genomics 2009

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nature Rev Genet 13(5):329–342