Animal Science Papers and Reports vol. 39 (2021) no. 4, 303-320
Institute of Genetics and Animal Biotechnology
of the Polish Academy of Sciences, Jastrzębiec, Poland

Enabling technologies drive the diversification and re-interpretation of gene concept in animal science - a review

Łukasz Huminiecki

Institute of Genetics and Animal Biotechnology of the Polish Academy of Sciences,
Postępu 36A, Jastrzębiec, 05-552 Magdalenka, Poland

(Accepted November 6, 2021)

The concept of the gene was originally developed in the field of classical genetics, and can be traced back to Mendel. Mendel defined a paradigm of particle-based heredity that has been subsequently developed for 150 years. However, Mendel's proposals lacked details (reflecting contemporary technological limitations). In the 20th century, enabling experimental technologies frequently drove the development of gene concept in a bottom-up manner. This was the case for technologies such as the microscope, X-ray crystallography, recombinant DNA, Sanger sequencing, and automatic sequencers. In the first two decades of the 21st century, gene concept was further diversified by innovative genomic technologies (*e.g.* next generation sequencing or single-cell omics) and associated computational methods developed for analysis of resulting datasets. Presently, machine learning and deep learning transform genetics into a data science (with considerable impact on animal science). Old arguments about the assumptions and implications of the Mendelian paradigm re-emerge in the context of the annotation of genomes of farm animals, genomic prediction of commercial traits, meta-genomics of gut microbiomes, and other applications of genomics and computational methods in animal science.

KEYWORDS: gene concept / technology / experimentalism / molecular biology / genomics / bioinformatics / data science / animal science / animal genetics.

Introduction

As a means of introduction, I will start by underlying the long-term importance of the Mendelian paradigm of heredity. I also note that Mendel was inspired by Darwin

*Corresponding authors: l.huminiecki@igbzpan.pl

and that they were both breeders. Following the Introduction section, I will argue that the paradigm has been always transformed and diversified by new technologies of genetic experimentation. This trend for technological innovation in genetics culminated in genomics revolution. Presently, genetics is being transformed into a data science. This transformation is having considerable impact on animal science.

The paradigm of intra-cellular discrete and particulate differentiating elements mediating heredity (*differirenden Zellelemente* in German) was proposed by Gregor Mendel in his seminal 1866 paper [Mendel 1866]. Mendel was inspired by Darwin's evolutionary writings that were in the English tradition of natural history, but himself embraced a stricter experimental approach. Note that Darwin proposed that the species on Earth originate through the process of evolution, or what he liked to call descent with modification. He also postulated that the mechanism of evolution was based on natural selection, that is struggle for survival and existence in the face of the harsh conditions of life and the scarcity of resources. Somewhat paradoxically, many of the examples described in the treaty were that of artificial selection practiced by pigeon breeders (rose, pigeon, dog and horse breeders which were popular and socially acceptable hobbies among affluent English countryside gentry which were a part of Darwin's social circles.) Naturally, it is breeder's will to provide for and propagate their livestock that replaces natural selection under such conditions. Darwin underlined what was similar between breeding and the evolution in the wild, while choosing not to focus on differences. In fact, Darwin later added a complementary evolutionary mechanism in the form of sexual selection.

Mendel provided strong evidence in support of his hypothesis in the form of expertly designed experiments on plant hybridizations. Moreover, Mendel analysed the implications of the paradigm very well; his insights were mathematical, profound, and far-reaching. In effect, Mendel launched a new paradigm of heredity that was a scientific revolution [Portin 2015] similar in importance to the Copernican Revolution in astronomy [Kuhn 1962, Kuhn 1985]. Mendel's paradigm falsified a mechanism of inheritance assuming blending of hereditary material (understood as a mechanism of heredity rather than phenotypic outcomes of breeding [Fisher 1930, Porter 2014]), as well as several other theories that assumed inheritance of acquired characteristics. Although, new types of functional genomics data (*e.g.* microarrays, RNA-seq, epigenetic modifications) have diversified and stretched gene concept, fundamental paradigm shifts have not yet occurred.

Conceptual assumptions of the Mendelian paradigm

This section discusses theoretical assumptions about heredity that together make up the contents of Mendel's hypothesis. The following section discusses far-reaching implications of Mendel's assumptions.

The assumption of corpuscularity

The first assumption made by Mendel was that basic elements of heredity were not amorphous material that mixed or blend as implied in Darwin's theory of pangenesis [Holterhoff 2014], but were instead pairs of corpuscles – that is minute particles of solid matter. This idea resembles the ancient Greek idea of atoms as basic constituents of the material world. However, we do not know if Mendel also thought hereditary particles indivisible (the Greek word *atomos* means *indivisible*.) In terms of their physical behaviour, Mendel's hereditary elements behaved like solids, in contrast to liquids postulated by various blending theories. The differentiating elements could exist in alternative variants within the population, but such variants generally persisted over consecutive generations in the same stable form. A new variant could only rarely arise through a mutation. All this seems logically consistent. As liquids tend to mix, liquid-like genetic material would promptly eliminate variability. Mendel's scheme demanded that hereditary elements retain their distinct identity between generations and segregate independently preserving variability.

Of course, we tend to think today that eukaryotic genes are materially relatively short fragments of long deoxyribonucleic acid (DNA) macromolecules – chromosomes – suspended in colloid nucleoplasm. Mendel could not have had this knowledge because of technological limitations. However, the fragments of DNA molecules behave like particles for most practical purposes (although, many of the "particles" are linked together on chromosomes).

The assumption of pairing

The second assumption of Mendel's was that two parental hereditary elements paired in the offspring. In other words, an element derived from a paternal organism would conjugate with a corresponding maternal element. Again, this seems logically consistent. It is intuitive to assume two hereditary elements (that is a pair) in organisms with two sexes and a sexual mode of reproduction. Mendel also suggested the distinction between heterozygotes and homozygotes. Note that a hereditary substance implied by the blending theory could not be counted, as liquids are uncountable substances.

The assumption of the possibility of dominance

The third assumption of Mendel's was that varieties of the same kind of a hereditary element had higher strengths of promoting differentiation. Differentiation presumably started with cells, and was the process through which hereditary elements expressed their potential to encode traits. Dominant varieties could effectively mask the effects of other varieties of the same kind (that is those encoding the same trait) if paired with them within cells of the same organism.

The assumption of intracellularity

Finally, Mendel assumed that hereditary elements were located within cells, thus adapting cell theory to his purposes. Specifically, Mendel first wrote in his paper about generative cells (that is eggs and pollen) each cell of both the generative types having a single differentiating element [Mendel 1866]. Further in the text, Mendel also wrote about somatic cells of the embryo – each such cell bearing a pair of differentiating elements for every trait. One must also assume that all cells of one organism bear identical pairs of the elements for each trait.

Conceptual implications of Mendel's paradigm

Persistence and independent segregation of hereditary particles (versus blending)

Mendel's corpuscular theory implied the persistence of the identity of hereditary particles between generations. In other words, one should expect that variants would persist unchanged between generations despite being combined in each cell if hereditary elements were corpuscular and generally not modifiable. This is because solid hereditary particles would only combine transiently: independently segregating when germ cells were produced again. Today, we know that the molecular mechanism of separation is the anaphase stage of meiosis. Note that the corpuscular character of hereditary material also implied discrete rather than continuous variability of phenotypes. In contrast, fluid hereditary material of the blending theory would mix in the cells of the embryo – immediately loosing distinct identity and resulting in gradual variability between generations and between individuals.

The possibility of mutations

Hereditary varieties of cellular elements existed as variants corresponding to different macroscopic manifestations of a trait (*i.e.* as phenotypes). This is because new variants could be generated through mutations leading to variability in genotypes within the population of organisms of one species. What we call a gene corresponds to Mendel's single kind of a hereditary element, which included all its variants, *i.e.* alleles, existing in a population. In one of Mendel's examples, *smooth* and *wrinkled* were two such variants, which we today call alleles. Both these alleles were variants of the same element kind, *i.e.* of the gene encoding the shape of the seed.

The possibility of dominance

Mendel correctly inferred that both paternal and maternal alleles retained individual powers of promoting cellular differentiation despite being combined in each cell. The theory also assumed that allele differentiation powers could vary in force, implying the possibility of either dominant or recessive inheritance. Naturally, Mendel could not know anything about relevant mechanistic details, but we now know that dominance results from differential effects of alleles on the dynamics of gene regulatory networks, or metabolic and signalling pathways. The paradigm also implied that dominant inheritance could be identified depending on traits observed in consecutive generations of hybrids. For example, the existence of three phenotypes in first generation hybrids would suggest that the differentiating powers of a couple of parental varieties were similar. Because the phenotype of the heterozygote would differ from both homozygous phenotypes. In contrast, complete dominance would imply only a pair of hybrid phenotypes. As the heterozygote would have a phenotype identical to one of the two homozygotes.

Enabling experimental technologies

It is important to underline how severely Mendel was limited by the simplicity of experimental methods that were available to him. He knew exactly nothing of the physical form of the factors he postulated. For this reason, he did not specify what size a hereditary particle was, although logic implied it had to be much smaller than cells for the whole genome to fit inside a nucleus. Neither did Mendel suggest what chemical substance a gene was built with. Moreover, the distinction between a hereditary trait and a hereditary factor was not fully developed. Before Johannsen introduced his terms: phenotype and genotype, the distinction between a trait, such as Mendel's wrinkled-seed character of pea, and a hypothetical Mendelian factor underlying it was fuzzy at best. The practical interest of breeders was focused on visible traits, the theory of discrete Mendel factors was only reflected in how such visible traits were inherited.

As a consequence of simplicity of Mendel's methods, his paradigm had to be supplemented in the course of the 20th century as new experimental technologies became available. New types of data generated by innovative laboratory technologies both strengthened and supplemented the theory where it was at first incomplete. For example, many details were provided using experimental protocols of biochemistry, crystallography, molecular biology, recombinant nucleic acids, and genomics. See Table 1 for more examples. In the sections that follow, I will discuss in more detail the examples of the microscope and of genomics technologies.

The technology of microscope suggested that genes lie on chromosomes

As soon as chromosomes could be visually observed by cytologists using microscopes, the Mendelian paradigm was complemented by the chromosomal theory of heredity. In its earliest form, the chromosomal theory simply proposed that genes are somehow physically associated with chromosomes. In effect, the gene was conceptualized as a point location on a chromosome. However, the link was *ad hoc*: it was uncertain what the nature of the association precisely was. Neither was it known which biochemical material making up chromosomes was involved, *i.e.* whether these were nucleic acids, or perhaps proteins. Many biologists, at first, thought proteins more likely – a view that was partially vindicated by the discovery of prions, histone modifications / variants, chromatin-remodelling complexes, polycomb and trithorax proteins, as well as histone code [Allis D. 2015].

Microscope pioneers had great impact on genetics even if they initially started working in other fields. For example, a 19th century German biologist Theodor Boveri initially pioneered the use of the microscope to study the role of chromosomes in development [Scheer 2018]. Later, inspired by striking conceptual similarities between chromosomes and Mendelian factors, Boveri correctly inferred that the factors were associated with chromosomes. He reasoned that each species is characterized by a defined set of chromosomal particles, and that the whole set of such particles is necessary for proper embryonic development. Boveri also insightfully suggested that the role of chromosomes in heredity could be characterized by two functional principles further reminiscent of Mendel's *differentiating cellular elements*. The first principle was that chromosomes have unique individual hereditary characteristics (this thesis was known as the principle of individuality). The second principle was that chromosomes were continuous in identity between generations: this was known as the principle of continuity.

Boveri personally inspired a generation of practical experimental geneticists in America. For example, Walter Sutton was at first a technical inventor with great mechanical skills [Crow and Crow 2002]. During his doctorate, he was performing cytological research on chromosomes of grasshoppers abundant on prairies of surrounding rural Kansas. He identified a species with particularly large chromosomes, and demonstrated that a maternal and a paternal chromosome can pair, and segregate during meiosis. Such pairing and independent segregation was, of course, identical to the behaviour Mendel postulated for the corpuscular elements of heredity. Another example of a practical innovator was Thomas Hunt Morgan, who worked both in embryology and in genetics [Frezza and Capocci 2018]. Morgan developed the fruit fly as an advanced laboratory model, used the illuminated microscope to study the cell nucleus, and proposed a chromosomal theory of heredity. His lab used the fruit fly model to study patterns of inheritance in mutants, and to construct the first genetic maps. The explanatory power of the chromosomal theory of heredity developed by Morgan lied in the fact that deviations from standard Mendelian patterns of heredity could be, now, proven to result from the linkage of genes lying on the same chromosome.

Thus it is clear that the microscope – an advanced experimental technology able to powerfully enhance the human sense of vision – was from the beginning instrumental in making a connection between the chromosome and the gene. Subsequently, ever more detailed observations were enabled by improved microscopes as the technology advanced throughout the 20th century. For example, an improved microscope was introduced that deployed perfectly diffused light from an electric lamp to illuminate the sample. Phase-contrast microscope facilitated observations of translucent samples. Electron microscopy helped to overcome the limits of resolution correlated with the wavelength of visible light. Confocal microscopy and laser enabled observations of virtual slices through thick samples. Today, cryogenic electron microscopy enables direct visualization of individual RNA molecules [Gopal, Zhou *et al*. 2012] or DNA-enzyme complexes [Ilangovan, Kay *et al*. 2017].

The technologies of genomics modified our understanding of the material gene

Experimental protocols of molecular biology revolutionized genetics: defining the gene as a physicochemical entity – a coherent, sharply delimited, structural subunit on a linear chromosome – composed of DNA. This view was further developed by the technologies of genomics. For example, the Human Genome Project (HGP) led to the development of cheap high-throughput sequencing technologies and bioinformatics tools for their analyses. As a result, there was an avalanche of structural genomics data describing the anatomy of chromosomes, as well as functional data describing how genes are expressed and what their functions are. Crucially, the new types of genomic data not only provided details, but also diversified and stretched our understanding of the concept of the gene [Falk 1984, Portin 2002, Griffiths and Stotz 2006, Griffiths P. 2007, Portin 2015, Portin and Wilkins 2017].

First sequenced genomes raised doubts about what the fundamental unit of function. Is it an exon, a set of exons belonging to one gene, a transcript? Indeed, structural genomics provided many counter-examples to the traditional view of the gene as a coherent structural unit or an indivisible atom of heredity. Gene predictions prepared by Ensembl, which consisted of *ab initio* predictions of open reading frames on whole sequenced genomes, validated by homology to known proteins or transcripts [Birney, Clamp *et al.* 2004], suggested that many genes overlapped or even shared exons. Moreover, most genes had multiple exons with long intertwining introns. Therefore, it became clear that recombination, and consequently also gene conversion or duplication could affect only a portion of the gene by affecting some, but not all of its exons.

Further examples of the genomic complexity of the material gene were provided by international functional genomics consortia that followed on the success of the HGP. Such consortia were designed to functionally characterize genes in genomes of mammals and animal model species. In particular, project FANTOM – functional annotation of mammalian genomes – generated several catalogues of the transcriptional landscape in the human and mouse. The first such catalogue [Carninci, Kasukawa *et al.* 2005] was based on sequencing full-length transcripts and revealed the existence of ubiquitous antisense transcription [Katayama, Tomaru *et al.* 2005]. Full-length transcripts also confirmed the existence of many complex loci where genes overlapped [Engstrom, Suzuki *et al.* 2006]. Moreover, the structure of genes was not constant, with variable splicing and frequent aberrant transcripts [Frith, Wilming *et al.* 2006]. Indeed, full-length transcript sequencing demonstrated the existence of many pseudo-mRNAs [Frith, Wilming *et al.* 2006] that cannot encode fully functional proteins due to disrupted open reading frames or aberrant splicing. However, pseudo-mRNAs may have regulatory roles competing for mRNA-binding proteins regulating translation, or by encoding truncated dominant negative protein variants.

Presently, it is the technology of next generation sequencing (NGS) that generates most new genomics data, both for research and diagnostic use. Computational tools, protocols, and databases are indispensable to process, store, and access such NGS datasets. For example, the FANTOM consortium developed a new technology for single-molecule expression profiling called cap analysis of gene expression – CAGE [Kanamori-Katayama, Itoh *et al*. 2011]. FANTOM5 surveyed genome-wide transcriptional activity at a single-base resolution level using CAGE. The upgraded high-resolution catalogue showed that transcriptional activity is pervasive in mammalian genomes, and that many genes have distant enhancers [Andersson, Gebhard *et al*. 2014]. Moreover, most human genes have multiple transcriptional start sites that can drive contrasting patterns of gene expression [FANTOM5-Consortium 2014].

In parallel to FANTOM5, ENCODE [Dunham, Kundaje *et al*. 2012] and modENCODE [Gerstein, Lu *et al*. 2010] consortia characterized the patterns of DNA binding for dozens of transcription factors regulating gene expression in the human genome and in the worm. ENCODE used NGS sequencing following chromatin immunoprecipitation – ChIP-seq – to characterize gene regulatory sequences [Landt, Marinov *et al*. 2012]. Interestingly, ENCODE appear to confirm the suggestion of FANTOM5's data that most human genes have multiple transcriptional start sites. This is evident by frequent co-localization of CAGE and peaks.

Note that NGS technologies are rather varied, including data on gene structure and mutations: DNA-seq, regulation and expression: RNA-seq, epigenetic modification: methyl-seq, *etc*. DNA-seq is now routinely used to characterize gene variability in wild, or experimental populations, or in artificial breeding populations. DNA-seq was also used to sequence tumour genomes, and to identify polymorphisms associated with genetic diseases or quantitative traits.

In evolutionary terms, whole genome sequencing and bioinformatics revealed a dynamic picture of genome architecture [Lynch 2007]: demonstrating the existence of thousands of pseudo-genes and gene families shaped by gene or genome duplication events [Torrents, Suyama *et al*. 2003, Coin and Durbin 2004]. Analyses of both individual gene families [Huminiecki, Goldovsky *et al*. 2009] and global patterns of gene duplication [Lynch and Conery 2000, Huminiecki and Heldin 2010, Perez-Bercoff, Makino *et al*. 2010, Huminiecki and Conant 2012] confirmed earlier theoretical speculations [Ohno 1970] that gene duplication is the most common recent source of the origins of new genes (this is certainly true in eukaryotic genomes). Accordingly, various molecular evolutionary models were proposed to explain why gene duplicates are retained and how they acquire new functions [Piatigorsky and Wistow 1991, Hughes 1994, Force, Lynch *et al*. 1999, Huminiecki and Wolfe 2004, Conant and Wolfe 2008, Des Marais and Rausher 2008, Hsiao and Vitkup 2008].

Development and critique of gene concept tended to follow on major technological advances

Let us underline repeating cycles of logical analysis and building of concepts in genetics, intertwined with empiricism driven by technological advances. For example, once microscopic observations made it clear that genes were associated with chromosomes, it was logical to ask of what physical nature this association was exactly. As chromosomes were known to be linear macromolecules, it was logical to argue either that (A) the gene is well localized, akin to a point or a well-demarcated interval on a line, or that it is (B) somehow diffusely encoded along the length of the chromosome.

View (A) was accepted by Morgan and most other contemporary figures in early genetical establishment. Indeed, the technology of genetic maps developed in Morgan's lab suggested that genes are arranged like pearls on a string. In this view, individual genes are functionally independent in agreement with Mendel's original assumptions (although, there may be pairwise linkage if two genes are located on the same chromosome). In contrast, view (B) is a dissenting view, first argued for by a cytologist-turned-geneticist: William Goldshmidt. In particular, Goldschmidt argued for the existence of reaction systems diffused along large parts or even the whole chromosome. In this view, genes could be organized into sorts of functional domains, perhaps with fuzzy boundaries.

Admittedly, molecular biology at first did not provide strong evidence in support of the theory of reaction systems postulated by Goldshmidt. However, his dissenting hypothesis is partially revived in the 21th century. Genomics suggest that distant chromosomal elements, both coding and regulatory, could form a hierarchy of elements [Fogle 1990] that together contributed to the functional expression of a single trait. Indeed, positional effects can be important for gene function. For example, chromosomal location (e.g., in a subtelomeric region) or gene neighbourhood (*e.g.*, clusters of neighbouring housekeeping [Lercher, Urrutia *et al*. 2002] or co-expressed genes [Weber and Hurst 2011]) could have impact on gene expression of genes located in the vicinity. Moreover, cell-specific expression domains, consisting of transcriptionally active chromatin delimited by Polycomb-binding regions, have been recently visualized using an innovative microscopy approach [Strack 2019] after the optical data was analyzed with a deep learning algorithm [Mateo, Murphy *et al*. 2019]. In some fungi, heritable epigenetic states of subtelomeric chromatin have been proven to play a role in adaptive stress response to stress, while chromosomal methylation patterns enable Darwinian evolution [Madhani 2021]. In plants, clusters of enzymes involved in biosynthesis of natural products have been described [Nutzmann, Huang *et al*. 2016].

Another wave of questioning of Mendel's assumptions followed on the discovery of DNA structure, and the genetic code encoded by the sequence of nucleobases. Was then the gene always a fundamental unit of function, heredity, and mutation? As a result of the molecular biology revolution, it became apparent that the gene could no longer be regarded as a mere point on a chromosome. This is after all a property of a DNA base pair. Indeed, technologies of recombinant DNA demonstrated beyond doubt that the gene is endowed with the property of length and included many base pairs. Similarly, the definition of a fundamental unit of mutation – a muton – had to be re-evaluated: a single base pair is a muton in case of point mutations. It is not my aim here to enumerate the basic facts of molecular genetics, which can be found in a number of excellent textbooks available. Of these textbooks, that by Watson et al. [Watson 2008] is particularly good with the processes on the interface of DNA and RNA biology and it keeps abreast of the developments in genome sequencing and systems biology.

Note that the concept of the material gene (encoded in DNA) emerged when basic empirical facts of molecular genetics turned out to be easy to generalize. For example, the genetic code turned out to be generally applicable across the domains of life. There also emerged a universal rule in that genetic information is used to synthesize proteins via the intermediate processes of transcription and translation. This rule, known as the fundamental dogma, is true for almost all species (except some specialized RNA viruses). The rule holds in unicellular, as well as multicellular organisms, both in development and in adulthood. In cases of the many fundamental processes of molecular genetics, the results obtained in simple model species (Lambda phage, the *E. coli* bacterium, or the fruit fly) could be directly generalized to animal genetics. However, many empirical findings in molecular genetics cannot be easily generalized to animal genetics, being applicable only to certain domains of the tree of life, or to a limited population of a single species. For example, early in the development of molecular genetics it became apparent that there are significant differences in gene structure or regulation between Prokaryotes and Eukaryotes [Watson 1988, Ptashne M 2002]. The rules concerning the structure of proximal promoters, transcribed as well as un-transcribed genic regions, as well as intron/exon junctions, seem to be more honoured in the breach than in the observance. There are many variants in animals, plants, bacteria, fungi and protozoans.

Towards a pragmatic approach to gene concept in animal science

My main aim was to highlight the great extent of the diversity of the molecular structures of animal genes that was revealed by enabling technologies of laboratory experimentation (such as techniques of molecular cloning, genomics, or systems genetics). Despite the fact that dictionary definitions normally strive for clarity and brevity, a classic dictionary of biology, now in its 14th edition updated to reflect progress in genomics [Lawrence 2008], signals the diversity of the concept of the gene as follows: *t*he basic unit of inheritance, by which hereditary characteristics are transmitted from parent to offspring. At the molecular level a single gene consists of a length of DNA (or in some viruses, RNA) which exerts its influence on the organism's form and function by encoding and directing the synthesis of a protein, or a tRNA, rRNA or other structural RNA. Each living cell carries a full complement of the genes typical of the species, borne in linear order on the chromosomes. Cells from diploid organisms carry two copies (alleles) of each gene.

To reiterate, historical arguments about the assumptions and implications of the Mendelian paradigm, which were outlined in introductory sections, re-emerge today in the context of the applications of genomics to animal science. Examples of the relevant applications of genomics include: (1) computational technologies such as algorithms for gene prediction and characterization used in the course of annotation of genes in genomes of farm animals[1]; (2) machine learning applications in genomic prediction of production traits of commercial value in breeding programs; (3) deep learning applied to functional annotation of genomes of farm animals; (4) Genome Wide Association Studies (GWAS) in farm animals; or (5) pipelines for the analysis of meta-genomes making up gut microbiomes and bacterial populations evolving antibiotic resistance. See Table 2 for an illustrative list of empirical observations that question classic assumptions about genes, including their independence, atomicity, corpuscularity, pairing, dominance, and intracellularity.

Genomic selection stands out as a key area in animal science in which machine learning has been successfully applied to the prediction of production traits effectively doubling the rate of genetic progress in breeding programs. Genomic selection was introduced on a massive scale and revolutionized the breeding industry. Machine learning facilitated prediction of traits of economic importance from tens of thousands of genomic SNPs, especially in dairy cattle [Wiggans, Cole *et al*. 2017], including in crossbreds [VanRaden, Tooker *et al*. 2020]. Note that deep learning in this application showed advantage over traditional machine learning in better capturing non-linear interactions between SNPs and in efficiently integrating different types of datasets [Montesinos-Lopez, Montesinos-Lopez *et al*. 2021]. The success of genomic selection supports a view on genetic information as diffused across the genome. This view on genes reminds one of Goldshmidt's reaction systems, rather than of atomistic and wholly independent genes.

GWAS studies in farm animals also became common in the last decade (studies in cattle, pigs, and chicken were reviewed by Sharma *et al*. in [Sharma, Lee *et al*. 2015]). GWAS suggest that pleiotropy[2] [Stearns 2010] is frequent in farm animals, especially for correlated traits (but can even occur in the case of uncorrelated production traits). In other words, a single gene can affect many different animal production traits to a different degree. It follows that genes have different levels of dominance in the context of different traits. For example, one study suggested that SNPs with a large effect on one trait can have small effects even on other uncorrelated traits [Xiang, MacLeod *et*

[1] Note that evolutionary characterization of predicted genes is typically a part of annotation of animal genomes. In the process, genes are grouped into gene families. Frequently, many evolutionarily related genes, or paralogs, are located within the same genome. Such paralogs can overlap in function, buffering against deleterious mutations.

[2] Pleiotropic genes have multiple effects which affect more than one phenotypic character Lawrence, E. (2008). Henderson's dictionary of biology. Harlow, England ; New York, Pearson Benjamin Cummings Prentice Hall.. Pleiotropy, redundancy, multi-gene traits, and functional overlap of paralogs call into question the one-gene-one-function rule of which the whole research program of reductionism (from genetics to molecular biology) tacitly relied. In some cases at least, the set of molecular domains relevant for some function could be dispersed across a wide chromosomal locus and include both coding and non-coding exons, as well as regulatory sequences.

al. 2017]. Pleiotropic gene functions frequently complicate breeding plans aiming to enhance a trait of economical importance due to inadvertent effects on other useful traits. This is because it may be difficult to select for a desired production trait without affecting other traits, sometimes in a fashion that is undesirable for the health of the animal or the plant, or otherwise detrimental from the point of view of economic productivity. However, a selection index can be used to manage such pleiotropic effects. For example, mastitis resistance correlates with milk production traits. Therefore, a breeding program for improved resistance to mastitis could simultaneously take into account milk production and udder health indices [Lund, Guldbrandtsen *et al*. 2008]. In another example, patterns of pleiotropy were characterized using multivariate analysis in the context of Australian selection index – an economic index of milk production traits [Bolormaa, Pryce *et al*. 2010].

Moreover, there are many paralogs overlapping in function: potentially also complicating breeding programs. Although, paralog redundancy is well known in molecular evolution, it has not yet received appropriate attention in animal science literature.

Note that human GWAS studies frequently found that SNPs tend to be only weakly associated with traits, and the majority fall in non-coding regions that cannot be easily linked with one gene. Regulatory SNPs affect regulatory sequences and change expression of a gene rather than the sequence of the encoded protein. Human disease studies also provided abundant evidence of pleiotropy [Sivakumaran, Agakov *et al*. 2011].

Admittedly, the counterexamples listed in Table 2 stretch and diversify the 19th century Mendelian paradigm, which assumed hereditary particles are always independent, atomistic, and have a fixed level of dominance. Thomas Fogle even argued that faced with the findings of molecular biology and genome projects, suggesting the diversity of gene's function and structure, we should abandon the idea of a unit. Instead, of thinking of the gene as a fundamentally integral unit of inheritance endowed with a singular function, and we should think of it as a set of molecular domains [Fogle 1990]. Fogle suggested, using metaphorical language, that the delicate bridge connecting Mendelian and molecular points of view in nearing collapse. Fogle gives several striking examples of gene overlap or dispersion, which do not work well with the concept of the gene as a unit. Note that a bioinformatician may interpret the gene not only as a set of domains, but possibly also as a hierarchy or a network of domains.

How should the paradigm be adapted in this confusing situation? The short and unexpectedly simple answer is: *pragmatically* – to maximize practical success in a given application. Indeed, Griffiths and Stotz, who are leading theoreticians of gene concept, proposed to use a family of closely related sub-concepts such as instrumental, nominal, and post-genomic genes [Griffiths and Stotz 2006, Griffiths P. 2007]. They also argued for a pragmatic approach to choosing between these and related gene sub-concepts [Griffiths P. 2013].

Table 1. Enabling technologies drove the development of the concept of the gene

Empirical technology	Type of data generated	Corresponding conceptual advances
Basic microscopy	Cell morphology, embryology, observation of microorganisms	Cell theory. Ontogenesis
Excavation and identification of rocks and minerals for the mining industry	Description and relative dating of fossils	Geological uniformatism suggested gradual evolution of species (replacing supernatural explanations such as deluges)
Classification of genera, species and varieties, artificial selection, hybridization, horticulture, geographical surveys	Drawings and samples of biological specimen, maps, observations of climate, data on geographical distribution of species	Darwinian evolution implies heredity but the mechanism of heredity is uncertain. Darwin also proposed the idea of the tree of life (implying phylogenesis)
Experimental breeding; purebreds with well-defined traits	Ratios of traits in consecutive generations of hybrids	Mendelism implies the existence of factors (or cellular elements) that underlie heredity
Microscopy, sample illumination	Cytology, observations of chromosomes	Chromosomal theory of heredity
The fruit fly laboratory model and fly mutants	Genetic maps	Recombination, genetic linkage
Techniques of organic chemistry and biochemistry	Chemical structure of the compounds of carbon and nitrogen, measurements of acidity, identification of nucleobases	The discovery and characterization of nucleic acids
Biometrics, statistical methods for the analysis of biological samples, experimental design	Series of various measurements together with descriptive statistics, visualizations	Theories on genes in biological populations (population genetics)
X-ray crystallography	Diffraction patterns	The structure of DNA double helix is defined. (Pairing of the strands of DNA suggests a mechanism for copying.)
DNA and RNA isolation, enzymology, restriction enzymes, molecular cloning and other techniques of molecular biology	Reaction kinetics for biochemical reactions, relative length of nucleic acid fragments, temperatures of DNA denaturation	Replication, open reading frames, transcription, mutation, and recombination are described. Molecular biology leads to the development of molecular genetics
Sequencing of nucleic acids	Order of bases in DNA and RNA	Genetic code is determined; exons, introns, mobile elements, and repetitive sequences are described
Genomics and bioinformatics, microarrays, genome browsers	Physical maps, contigs, gene predictions, expression data, data integration	Details about the architecture and regulation of different classes of genes are revealed: transcription start sites, exons, introns, regulatory sequences, transcription factor binding sites, spatial and temporal expression patterns
Next generation sequencing, data science, machine learning, deep learning	Mutation screens, RNAseq, advanced classification of expression profiles, functional signatures for various biological processes	Gene is being redefined as a virtual gene – that is as concepts of genetic data science, objects of data integration, components of genetic regulatory networks, expression modules, etc

Table 2. Counterexamples to classic assumptions about genes prompt re-interpretation of the Mendelian paradigm

Genomic application	Genes are independent	Genes are indivisible (atomicity)	Hereditary particles do not blend (corpuscularity)	In diploids, there are exactly two alleles of each gene (pairing)	Alleles have differing but fixed capacity to promote a phenotype (dominance)	Genes are always located within cells (intracellularity)
Genome annotation, molecular diagnostics, and systems genetics	(1) Complex loci, gene overlap, nested genes, fuzzy gene boundaries, antisense transcription, functional overlap in gene families (2) Disease-associated SNPs are often difficult to link with individual genes[1]. (This is at least partially due to SNPs in diffuse or distant regulatory sequences) (3) Linkage disequilibrium[2]	(1) Recombination affecting individual exons[3] (2) Point mutations, small insertions and deletions. (The basic unit of mutation is typically a base pair, not a gene)	Loss of heterozygosity is common in cancer. (The mechanism may be either mutation, or epigenetic change, or recombination.)	Copy-number variants or somatic amplifications.	(1) Imprinting (2) Dynamic mutations (3) Hypothetical inheritance of epigenetic modifications (4) Epistasis[4] (5) Pleiotropy: genes are functionally linked in pathways with frequent cross talk, suggesting a networked[5] view of heredity	(1) Circulating tumor DNA (2) Viruses
Meta-genomics of gut microbiomes or antibiotic resistomes	Functionally coupled genes are frequently grouped in bacterial operons	The basic unit of mutation in bacteria or viruses is also a base pair	Bacteria are haploid, but plasmids are typically present in multiple copies. Plasmids can recombine, change copy number, or relative proportions of copies			Plasmids and bacteriophages are extrachromosomal mobile genetic elements that can be horizontally transferred

Empirical observations are generated through different genomic applications (see table rows). These are counter-examples to Mendel's assumptions about genes (*i.e.*, of their independence, atomicity, corpuscularity, pairing, dominance, and intracellularity).

[1] Similarly in genomic prediction, informative SNPs across the entire genomes are easier to interpret using machine learning (rather than to link with individual genes proven to be trait-associated by prior biochemical investigations).

[2] Occurs when alleles at two linked loci are non-randomly associated because of close physical proximity on a chromosome or selection

[3] This is because the domains of one gene can be "broken up" in the process of partial recombination during meiosis.

[4] Epistasis signifies masking or suppression of effects of a gene by another gene(s).

[5] The networked view of heredity can also explain why many quantitative traits are found during breeding programs to correlate with each other.

In conclusion, gene concept variants can and should be flexibly adapted to modern applications in animal science (realistically and depending on context). Overall, this means that practical rather than theoretical considerations should be prioritized. Indeed, animal scientists are justified in making pragmatic use of gene concept in the design and execution of breeding programs, or applied research projects. Context and project goals should decide how genes are represented in data mining projects and types of data need integrating. This is because alternative promoters, alternative or scrambled splice variants, epigenetic modifications, polymorphisms, epistasis, or pleiotropy may or may not have practical significance in specific applications.

REFERENCES

1. ALLIS D., C.M.L., JENUWEIN T., REINBERG D., 2015 – Epigenetics. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.
2. ANDERSSON R.C. GEBHARD C., MIGUEL-ESCALADA I., HOOF I., BORNHOLDT J., BOYD M., CHEN Y., ZHAO X., SCHMIDL C., SUZUKI T. *et al.,* 2014 – An atlas of active enhancers across human cell types and tissues. ***Nature*** 507(7493), 455-461.
3. BIRNEY E., CLAMP M., DURBIN R., 2004 – GeneWise and Genomewise. ***Genome Research*** 14(5), 988-995.
4. BOLORMAA, S.J.E. PRYCE J.E., HAYES B.J., GODDARD M.E., 2010 – Multivariate analysis of a genome-wide association study in dairy cattle. ***Journal of Dairy Science*** 93(8), 3818-3833.
5. CARNINCI P.T. KASUKAWA T., KATAYAMA S., GOUGH J., FRITH M.C., MAEDA, N., OYAMA R., RAVASI T., LENHARD B., WELLS C. *et al.*, 2005 – The transcriptional landscape of the mammalian genome. ***Science*** 309(5740), 1559-1563.
6. COIN L., DURBIN R., 2004 – Improved techniques for the identification of pseudogenes. ***Bioinformatics*** 20 Suppl 1, I94-I100.
7. CONANT G.C., WOLFE K.H., 2008 – Turning a hobby into a job: how duplicated genes find new functions. ***Nature reviews. Genetics*** 9(12), 938-950.
8. CROW E.W., CROW J.F., 2002 – 100 years ago: Walter Sutton and the chromosome theory of heredity. ***Genetics*** 160(1), 1-4.
9. DES MARAIS D.L., RAUSHERM.D., 2008 – Escape from adaptive conflict after duplication in an anthocyanin pathway gene. ***Nature*** 454(7205), 762-765.
10. DUNHAM I.A., KUNDAJE A., ALDRED S.F., COLLINS P.J., DAVIS C.A., DOYLE F., EPSTEIN C.B., FRIETZE S., HARROW J., KAUL R. *et al.,* 2012 – An integrated encyclopedia of DNA elements in the human genome. ***Nature*** 489(7414), 57-74.
11. ENGSTROM P.G., SUZUKI H., NINOMIYA N., AKALIN A., SESSA L., LAVORGNA G., BROZZI A., LUZI L., TAN S.L., YANG L. *et al.,* 2006 – Complex Loci in human and mouse genomes. ***PLoS Genetics*** 2(4), e47.
12. FALK R., 1984 – The gene in search of an identity. ***Human Genetics*** 68(3), 195-204.
13. FANTOM5-Consortium 2014 – A promoter-level mammalian expression atlas. ***Nature*** 507(7493), 462-470.
14. FISHER R.A., 1930 – The genetical theory of natural selection. Oxford, at the Clarendon Press.
15. FOGLE T., 1990 – Are Genes Units of Inheritance. ***Biology & Philosophy*** 5(3), 349-371.
16. FORCE A., LYNCH M., PICKETT F.B., AMORES A., YAN Y.L., POSTLETHWAIT J. *et al.,* 1999 – Preservation of duplicate genes by complementary, degenerative mutations. ***Genetics*** 151(4), 1531-1545.

17. FREZZA G., CAPOCCI M., 2018 – Thomas Hunt Morgan and the invisible gene: the right tool for the job. ***History and Philosophy of the Life Sciences*** 40(2), 31.
18. FRITH M.C., WILMING L.G., FORREST A., KAWAJI H., TAN S.L., WAHLESTEDT C., BAJIC V.B., KAI C., KAWAI J., CARNINCI P. *et al.,* 2006 – Pseudo-messenger RNA: phantoms of the transcriptome. ***PLoS Genetics*** 2(4), e23.
19. GERSTEIN M.B., LU Z.J., VAN NOSTRAND E.L., CHENG C., ARSHINOFF B.I., LIU T., YIP K.Y., ROBILOTTO R., RECHTSTEINER A., IKEGAMI K., *et al.* 2010 – Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. ***Science*** 330(6012), 1775-1787.
20. GOPAL A., ZHOU Z.H., KNOBLER C.M., GELBART W.M., 2012 – Visualizing large RNA molecules in solution. ***RNA*** 18(2), 284-299.
21. GRIFFITHS P.S., 2007 – Gene. The Cambridge companion to the philosophy of biology. D. L. Hull and M. Ruse. Cambridge; New York, Cambridge University Press: xxvii, 513 p.
22. GRIFFITHS P.S.K., 2013 – Genetics and Philosophy: An Introduction, Cambridge University Press.
23. GRIFFITHS P.E., STOTZ K., 2006 – Genes in the postgenomic era. ***Theoretical Medicine and Bioethics*** 27(6), 499-521.
24. HOLTERHOFF K., 2014 – The history and reception of Charles Darwin's hypothesis of pangenesis. ***Journal of the History of Biology*** 47(4), 661-695.
25. HSIAO T.L., VITKUP D., 2008 – Role of duplicate genes in robustness against deleterious human mutations. ***PLoS Genetics*** 4(3), e1000014.
26. HUGHES A.L., 1994 – The evolution of functionally novel proteins after gene duplication. ***Proceedings of the Royal Society of London. Series B: Biological Sciences*** 256(1346), 119-124.
27. HUMINIECKI L., CONANT G.C., 2012 – Polyploidy and the evolution of complex traits. ***International Journal of Evolutionary Biology*** 292068.
28. HUMINIECKI L., GOLDOVSKY L., FREILICH S., MOUSTAKAS A., OUZOUNIS C., HELDIN C.H., 2009 – Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom. ***BMC Evolutionary Biology*** 9, 28.
29. HUMINIECKI L., HELDIN C.H., 2010 – 2R and remodeling of vertebrate signal transduction engine. ***BMC Biology*** 8, 146.
30. HUMINIECKI L., WOLFE K.H., 2004 – Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. ***Genome Research*** 14(10A), 1870-1879.
31. ILANGOVAN A., KAY C.W.M., ROIER S., EL MKAMI H., SALVADORI E., ZECHNER E.L., ZANETTI G., WAKSMAN G., 2017 – Cryo-EM Structure of a Relaxase Reveals the Molecular Basis of DNA Unwinding during Bacterial Conjugation. ***Cell*** 169(4), 708-721 e712.
32. KANAMORI-KATAYAMA M., ITOH M., KAWAJI H., LASSMANN T., KATAYAMA S., KOJIMA M., BERTIN N., KAIHO A., NINOMIYA N., DAUB C.O. *et al.,* 2011 – Unamplified cap analysis of gene expression on a single-molecule sequencer. ***Genome Research*** 21(7), 1150-1159.
33. KATAYAMA S., TOMARU Y., KASUKAWA T., WAKI K., NAKANISHI M., NAKAMURA M., NISHIDA H., YAP C.C., SUZUKI M., KAWAI J. *et al.,* 2005 – Antisense transcription in the mammalian transcriptome. ***Science*** 309(5740), 1564-1566.
34. KUHN T.S., 1962 – The structure of scientific revolutions. Chicago, University of Chicago Press.
35. KUHN T.S., 1985 – The Copernican Revolution-Planetary Astronomy in the Development of Western Thought. Cambridge, Mississippi, Harvard University Press.
36. LANDT S.G., MARINOV G. K. , KUNDAJE A., KHERADPOUR P., PAULI F., BATZOGLOU S., BERNSTEIN B.E., BICKEL P., BROWN J.B., CAYTING P. *et al.,* 2012 – ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. ***Genome Research*** 22(9): 1813-1831.
37. LAWRENCE E., 2008 – Henderson's dictionary of biology. Harlow, England; New York, Pearson Benjamin Cummings Prentice Hall.

38. LERCHER M.J., URRUTIA A.O., HURST L.D., 2002 – Clustering of housekeeping genes provides a unified model of gene order in the human genome. ***Nature Genetics*** 31(2), 180-183.
39. LUND, M.S., GULDBRANDTSEN B., BUITENHUIS A.J., THOMSEN B., BENDIXEN C., 2008 – Detection of Quantitative Trait Loci in Danish Holstein Cattle Affecting Clinical Mastitis, Somatic Cell Score, Udder Conformation Traits, and Assessment of Associated Effects on Milk Yield. ***Journal of Dairy Science*** 91(10), 4028-4036.
40. LYNCH M., 2007 – The Origins of Genome Architecture, Sinauer Associates.
41. LYNCH M., CONERY J. S. 2000 – The evolutionary fate and consequences of duplicate genes. ***Science*** 290(5494), 1151-1155.
42. MADHANI H.D., 2021 – Unbelievable but True: Epigenetics and Chromatin in Fungi. ***Trends in Genetics*** 37(1), 12-20.
43. MATEO L.J., MURPHY S.E., HAFNER A., CINQUINI I.S., WALKER C.A., BOETTIGER A.N. *et al.,* 2019 – Visualizing DNA folding and RNA in embryos at single-cell resolution. ***Nature*** 568(7750), 49-54.
44. MENDEL J.G., 1866 – Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn(IV), 3–47.
45. MONTESINOS-LOPEZ, O.A., MONTESINOS-LOPEZ A., PEREZ-RODRIGUEZ P., BARRON-LOPEZ J. A., MARTINI J.W.R., FAJARDO-FLORES S.B., GAYTAN-LUGO L.S., SANTANA-MANCILLA P.C., CROSSA J. *et al.*, 2021 – A review of deep learning applications for genomic selection. ***BMC Genomics*** 22(1), 19.
46. NUTZMANN H.W., HUANG A., OSBOURN A., 2016 – Plant metabolic clusters - from genetics to genomics. ***New Phytologist*** 211(3), 771-789.
47. OHNO S., 1970 – Evolution by gene duplication. Berlin, Springer-Verlag.
48. PEREZ-BERCOFF A., MAKINO T., MCLYSAGHT A., 2010 – Duplicability of self-interacting human genes. ***BMC Evolutionary Biology*** 10, 160.
49. PIATIGORSKY J., WISTOW G., 1991 – The recruitment of crystallins: new functions precede gene duplication. ***Science*** 252(5010), 1078-1079.
50. PORTER T.M., 2014 – The curious case of blending inheritance. ***Stud Hist Philos Biol Biomed Sci Part C*** 46, 125-132.
51. PORTIN P., 2002 – Historical development of the concept of the gene. ***The Journal of Medicine and Philosophy*** 27(3): 257-286.
52. PORTIN P., 2015 – The Development of Genetics in the Light of Thomas Kuhn's Theory of Scientific Revolutions. ***Recent Advances in DNA & Gene Sequences*** 9(1), 14-25.
53. PORTIN P., WILKINS A., 2017 – The Evolving Definition of the Term "Gene". ***Genetics*** 205(4), 1353-1364.
54. PTASHNE M.G.A., 2002 – Genes and Signals, Cold Spring Harbor Laboratory Press.
55. SCHEER U., 2018 – Boveri's research at the Zoological Station Naples: Rediscovery of his original microscope slides at the University of Wurzburg. ***Marine Genomics*** 40, 1-8.
56. SHARMA A., LEE, J. S., DANG C.G., SUDRAJAD P., KIM H.C., YEON S.H., KANG H.S., LEE S.H., 2015 – Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. ***Asian-Australasian Journal of Animal Sciences*** 28(10), 1371-1379.
57. SIVAKUMARAN S., AGAKOV F., THEODORATOU E., PRENDERGAST J.G., ZGAGA L., MANOLIO T., RUDAN I., MCKEIGUE P., WILSON J.F., CAMPBELL H., 2011 – Abundant pleiotropy in human complex diseases and traits. ***American Journal of Human Genetics*** 89(5), 607-618.
58. STEARNS F.W., 2010 – One hundred years of pleiotropy: a retrospective. ***Genetics*** 186(3), 767-773.

59. STRACK R., 2019 – Imaging chromatin and RNA in embryos. ***Nat Methods*** 16(5), 361.
60. TORRENTS D., SUYAMA M., ZDOBNOV E., BORK P., 2003 – A genome-wide survey of human pseudogenes. ***Genome Research*** 13(12), 2559-2567.
61. VANRADEN P.M., TOOKER M.E., CHUD T.C.S., NORMAN H.D., MEGONIGAL J.H., JR., HAAGEN I. W., WIGGANS G.R., 2020 – Genomic predictions for crossbred dairy cattle. ***Journal of Dairy Science*** 103(2), 1620-1631.
62. WATSON J.D., 1988 – Molecular biology of the gene. Menlo Park, Calif., Benjamin/Cummings Pub. Co.
63. WATSON J.D. 2008 – Molecular biology of the gene. San Francisco, Calif., Pearson/Benjamin Cummings.
64. WEBER C.C., HURST L.D., 2011 – Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. ***Genome Biology*** 12(3), R23.
65. WIGGANS G.R., COLE J.B., HUBBARD S.M., SONSTEGARD T.S. *et al.*, 2017 – Genomic Selection in Dairy Cattle: The USDA Experience. ***Annu Rev Anim Biosci*** 5, 309-327.
66. XIANG R., MACLEOD I.M., BOLORMAA S., GODDARD M.E., 2017 – Genome-wide comparative analyses of correlated and uncorrelated phenotypes identify major pleiotropic variants in dairy cattle. ***Scientific Reports*** 7(1), 9248.