Title: Epigenetic this, epigenetic that: comparing two digital humanities methods for analyzing a slippery scientific term.

(Forthcoming in Synthese)

Stefan Linquist
Brady Fullerton
Akashdeep Grewal

Abstract

We compared two digital humanities methods in the analysis of a contested scientific term. "Epigenetics" is as enigmatic as it is popular. Some authors argue that its meaning has diluted over time as this term has come to describe a widening range of entities and mechanisms (Haig 2012). Others propose both a Waddingtonian "broad sense" and a mechanistic "narrow sense" definition to capture its various scientific uses (Stotz and Griffiths 2016). We evaluated these proposals by first replicating a recent analysis by Linquist and Fullerton (2021). We analyzed the 1100 most frequently cited abstracts on epigenetics across four disciplines: proximal biology, biomedicine, general biology, and evolution. Each abstract was coded for its heritability commitments (if any) and functional interpretation. A second study applied LDA topic modelling to the same corpus, thus providing a useful methodological comparison. The two methods converged on a discipline-relative ambiguity. Within such disciplines as biomedicine or molecular biology that focus on proximate mechanisms, "epigenetic(s)" refers to a range of molecular structures while specifying nothing in particular about their heritability. This proximal conception was primarily associated with the functions of gene regulation and disease. In contrast, a second relatively uncommon sense of "epigenetics" is restricted to a small proportion of evolutionary abstracts. It refers to many of the same molecular structures, but regards them as trans-generationally inherited and associated with adaptive phenotypic plasticity. This finding underscores the benefit of digital tools in complementing traditional conceptual analysis. Philosophers should be cautious not to conflate the relatively uncommon evolutionary sense of epigenetics with the more widely used proximal conception.

1. Introduction

To many philosophers, the sound of scientists arguing about definitions is a bit like the wail of an ambulance siren to a personal injury lawyer: a hopeful sign that soon there will be an opportunity to ply our trade. Philosophers of science are in the business of clarifying scientific concepts. Included in our portfolio of successful cases is a growing list of refined concepts including biological function (Millikan 2002; Garson 2016), adaptation (Brandon 2014; Godfrey-Smith 2001), group selection (Sober 1984; Okasha, 2006), the gene concept (Moss 2004; Stotz, Griffiths, and Knight 2004; Waters 1994), biodiversity (Takacs 1996; Sarkar 2002; Santana 2014), the innate/acquired distinction (Griffiths, Machery, and Linquist 2009), and the species concept (Wilkins 2009), among others. Not all of these philosophical projects–some of which are ongoing–resulted in a concise definition of the concept under analysis. Sometimes, progress involved discovering that a given term has multiple meanings that scientists themselves were insensitive to. In other cases, historical changes in meaning were revealed. The philosopher's job then is to explain the epistemic and ontological "ecological and evolutionary forces" that drive conceptual change (Stotz and Griffiths 2004). This kind of big-picture analysis is outside the purview of most working scientists and marks a way that philosophy can complement science. In yet other cases, philosophers see themselves as honest brokers hoping to resolve internal scientific disputes. Such cases are perhaps the most exciting as they stand to make theoretical contributions to the sciences themselves (Hull 1988).

This paper is a methodological investigation of two digital humanities approaches for the analysis of scientific concepts. Both methods take advantage of the growing availability of large, electronic datasets of scientific publications, but they do so in very different ways. The first, qualitative coding approach involves a human evaluator who classifies a sample of texts that

were extracted from a database. Each text is read and coded according to some preestablished set of criteria. As we illustrate below, this method can track differences in the uses of various concepts across disciplines or identify conceptual change. The second approach involves the use of a topic modelling technique called Latent Dirichlet Allocation (LDA). In this case, humans need not read all of the actual texts under analysis. Instead, a potentially very large corpus is analyzed by a Bayesian algorithm that ranks words according to their relative frequencies of co-occurrence (e.g. how often they cooccur in the same abstract) across the entire corpus. These ordered word-clusters or "topics" are then analyzed by human experts who interpret their theoretical significance. In what follows, we apply both strategies to a corpus of scientific abstracts on the general topic of epigenetics. The abstracts were classified into four disciplines and organized chronologically, allowing us to look for discipline-specific conceptions or conceptual changes over time. By applying both methods to an identical body of literature, we were able to assess their respective advantages and limitations.

This project is part of a larger investigation into the uses and abuses of the epigenetics concept across disciplines. The term epigenetics is as popular as it is enigmatic. Although biology is allegedly in the grip of an epigenetic revolution (Jablonka and Lamb 2005; Bonduriansky and Day 2020), many practicing scientists are unclear about what "epigenetic" means or whether it is used consistently across disciplines (Ptashne 2007; Häfner and Lund 2016; Deans and Maggert 2015). Our specific research questions are (1) whether the term "epigenetic(s)" is associated with specific heritability commitments and/or functional interpretations that vary across disciplines or time periods, and (2) whether the two aforementioned methods possess certain advantages or limitations when applied to this sort of task. Each of the following sections describes how we employed these methods to a corpus of

roughly 1100 scientific abstracts covering the past 25 years, and our findings. Before going into those details, however, it is helpful to consider why epigenetics is a suitable subject for this analysis.

2. A slippery Scientific Term

David Haig (2012) identified an interesting historical pattern in the frequency of scientific uses of "epigenetic(s)." After being coined by Waddington (1942), the term persisted at low frequency in journals for about fifty years. This is noteworthy in itself, given that most theoretical terms in science are "stillborn" (Haig 2012). At the start of the twenty-first century, however, the term exploded in popularity and its use continues to grow exponentially. Other researchers tracking this pattern describe differences among disciplines in the frequency and timing with which "epigenetic(s)" became popular (Linquist and Fullerton 2021). For instance, the proximal disciplines of developmental biology, molecular/cellular biology were early adopters of "epigenetic(s)", followed by biomedicine and general biology. Evolutionary biology stands in contrast to these disciplines, exhibiting a delayed and relatively muted adoption of this term. Both Haig (2012) and Linquist and Fullerton (2021) agree that a large and expanding proportion of biological articles published since 2000 mention "epigenetic(s)" in their titles and abstracts. This pattern is intriguing, given the amount of disagreement and confusion that exists today among scientists about its precise meaning.

Haig (2004, 2012) argues that part of the confusion surrounding the meaning of "epigenetic(s)" stems from its origin in two independent research traditions that eventually merged and gave rise to "hybrid recombinant offspring" (2012). The details of this historical trajectory are beyond the scope of this paper. However, a few noteworthy features of Haig's analysis provide insight into why this term is so slippery.

One of two early research traditions stemmed from Waddington's (1942, 1953, 1956) efforts to generate a mechanistic framework for explaining phenotypic development. More specifically, Waddington argued that an adequate understanding of development required the integration of two disciplines: the traditional field of epigenesis with the emerging discipline of genetics. Hence, "epigenetics" for Waddington was a contraction of these two (as he saw them) cognate disciplines (Haig 2004). At the same time, Waddington took a critical stance toward the genetics of his day. In particular, he saw a discipline that was overly focused on the effects of individual genes and insufficiently sensitive to interactions among multiple genes and between the genome and the ontogenetic environment. Haig notes that some authors, when discussing epigenetics, identify primarily with Waddington's opposition to an extreme form of gene reductionism. In contrast, other researchers continue to use this term in Waddington's core sense, to describe the mechanisms (whatever they might be) guiding phenotypic development. Importantly, although Waddington's understanding of epigenetics itself evolved over the course of his career (Nicoglou 2018), his early theoretical work "is notable for its lack of discussion about inheritance" (Haig 2012, p 14). This alleged tendency for "epigenetic(s)" to be associated with different ideas about heredity –or to lack any heritability commitment whatsoever –is one of the disputed issues surrounding this term.

The second epigenetic research tradition identified by Haig stems from David Nanney's (1958) work on epigenetic control systems. Nanney was interested not only in cellular differentiation but also the ability of developing cell lineages to remain stable in the face of changes to their surrounding context. Nanney reasoned that these phenomena could not be explained by genetics alone, since all cells in an organism have a common genome. He therefore posited a second system – the epigenetic system – that controls and stabilizes cellular

development. Hence, for Nanney, "epigenetics" referred to a system that was *above* or *in addition to* the genetic system. Haig adds that, "heredity was a potential property of epigenetic systems, but not a defining feature of such systems" (2010, p14).

One of the "hybrid offspring" produced by these two traditions involved a focus on methylation and its role in gene regulation. Methylation marks are small molecules that selectively bind to specific sequences of DNA, affecting the tendency for that DNA to be transcribed into RNA. The sequence-specificity of methylation makes it an effective system not only for regulating gene expression, but also for suppressing the activity of genetic parasites such as transposable elements (TEs) (Zhou et al 2020). Haig (2012) explains that for a brief period during the 1980s-1990s, the study of epigenetics became almost exclusively focussed on methylation. We draw three significant implications from this move. First, it marks a potential shift in the meaning of "epigenetic(s)" away from a purely functional definition to a more structural definition that pins this term to a particular entity. Second, during this period it became increasingly apparent that epigenetic marks could be transmitted in somatic cell linages. Hence, the question of epigenetic heritability became partly settled during this phase. The third implication concerns function. Methylation marks are widely recognized for their roles in gene regulation and (to a lesser extent) TE suppression. During the period when epigenetics was so focussed on methylation, discussions about epigenetic marks would have had a relatively clear functional interpretation.

As research in molecular biology has moved forward, a wider range of epigenetic marks (besides methylation) have been identified. These include a growing list of entities that range from histone modifications to noncoding RNA to transposable elements themselves. A unifying feature of this diverse range of epigenetic marks is that they interact with protein-coding genes

without affecting DNA sequence. However, this might be where their similarities end. The nature of their heritability and functions can vary widely. For instance, some types of mark (e.g. some noncoding RNAs) are transmitted in gametic cells from parent to offspring. A subset of these entities are thought to influence the early states of gene regulation in the embryo. This involves a very different sort of epigenetic inheritance than the transmission of a methylation mark from a somatic parent cell to its daughter. By the same token, although some marks have a role in gene regulation, it is misleading to apply this functional interpretation to all such marks across the board (Palazzo and Gregory 2014; Palazzo and Lee 2015; Graur et al. 2013; Doolittle 2013; Elliott, Linquist, and Gregory 2014).

Another approach to defining "epigenetics" has been defended by Karola Stotz and Paul Griffiths (2013, 2016). Although these authors are sensitive to the various historical twists and turns outlined by Haig and others, they claim to identify two stable definitions. Again, there is insufficient space to engage with their work in detail. However, we think that their position on epigenetic inheritance represents a popular gloss on the scientific literature, found also in the work of such influential authors as Susan Oyama (Oyama 2000), Evelyn Fox Keller (2014), Eva Jablonka and Marion Lamb (2014), and other scientific commentators. For reasons outlined below, this popular gloss arguably misrepresents the prevailing scientific views in molecular biology and biomedicine. Before exploring this discrepancy, it is important to note that Stotz and Griffiths' (2016) discussion of epigenetics is part of a wider argument against what they call the *sequence hypothesis* – i.e. the hypothesis that the causal specificity of a protein is contained entirely in the nucleotide sequence that encodes it. In their view, the fact that gene expression is partly controlled by epigenetic marks implies that the causes of protein synthesis are distributed across a range of both genic and non-genic sources. Although we have some concerns about

Griffiths' and Stotz's (2016) proposed definitions of epigenetics (see below), we do not take issue with their argument against the sequence hypothesis.

Stotz and Griffiths identify a "broad sense" of epigenetics that stems from Waddington, namely "the study of the causal mechanisms by which genotypes give rise to phenotypes" (2016, p 14). Interestingly, they make a point of adding that the broad sense definition is not only about the mechanisms of development, but also has implications for evolution. In particular, Stotz and Griffiths draw a connection between Waddington's model of development and the phenomenon of genetic accommodation, which some authors view as a challenge to the New Synthesis accounts of evolution (West Eberhard 2003). This analysis raises questions about the extent to which the broad-sense concept remains in contemporary use. Also, whether the meaning of this term has shifted away from its original developmental connotation to apply more specifically to the evolutionary effects of developmental plasticity.

Narrow-sense epigenetics is defined by Stotz and Griffiths as "the study of the mechanisms that determine which genome sequences will be expressed in the cell; the control of cell differentiation and of mitotically and sometimes meiotically heritable cell identity, that does not involve changes to the underlying genome sequence" (2016 p.3) – a definition that they plausibly associate with Nanney. However, Stotz and Griffiths further suggest that "Most molecular biologists today understand epigenetics in this narrow sense." We see at least three reasons for questioning this further claim that narrow sense epigenetics (as they define it) reflects contemporary usage in molecular biology. First, recall the move away from a purely functional definition toward the identification specific epigenetic structures. This move towards structure has involved a significant advance in the molecular understanding of epigenetics. Hence, it is potentially misleading to equate the modern molecular definition of epigenetic marks (which

include numerous structurally defined entities) with Nanney's purely functional definition. A second issue concerns the specific role of gene regulation. As we have noted, although some types of epigenetic mark are heavily implicated in regulation, other marks are less clearly tied to this functional role. To insist that gene regulation is contained in the contemporary definition of epigenetics is to presuppose a highly controversial empirical question. A third issue stems from Stotz and Griffiths' (2016) treatment of heritability commitments in their definition of narrow-sense epigenetics. While some types of epigenetic mark are widely recognized as mitotically inherited, this is not well established for all types of epigenetic mark (Deans and Maggert 2015). Indeed, even a smaller subset of epigenetic marks are thought to be meiotically inherited (sexually, from parent to offspring), and among those, there is very little evidence that meiotic inheritance could persists for an open-ended number of generations (ibid). A more sensitive definition of narrow sense epigenetics would not only distinguish epigenetic marks from their various potential functions (including an allowance for no function), but also allow for a range of heritability commitments (including the possibility of no heritability). The framework developed by Linquist and Fullerton (2021) and refined below (Tables 1 & 2) is a step in this direction.

In summarizing this section, it is interesting to ask why molecular biologists and biomedical researchers have retained "epigenetic" to describe a growing list of molecular structures that can vary so dramatically in their heritability and functions. Given the potential for equivocation, it would presumably make more sense to drop the epigenetic label altogether and speak instead of specific types of mark and their distinctive functional and heritability properties. Indeed, there is some indication that certain researchers are starting to move in this direction (Häfner and Lund 2016). Yet, the general label appears to be increasing in popularity (Linquist and Fullerton 2021). What then explains the appeal of such a confusing term? Haig suggests that

"The indefinite definition of epigenetics (together with the connotation of being 'above' or 'beyond' genetics) has meant that scientists from divergent disciplines, studying only loosely related phenomena, could all feel that they were engaged in epigenetic research near the cutting edge of modern biology" (2010, p.15). We take this to suggest a bandwagon effect where individual scientists are opting to promote the significance of their work at the potential expense of precision. Some authors suggest that this phenomenon is especially pronounced in the biomedical sciences, where competition for funding tends to be relatively fierce. For instance, Häfner and Lund suggest that it is no coincidence that epigenetics increased in popularity at around the same time that the failed promises of the Human Genome Project were becoming apparent:

> The knight in shining armor arrived in the form of epigenetics and suddenly the shortcomings of genetics metamorphosed into the sales pitch for epigenetics as genes do not hold all the keys to disease, epigenetics will. They were proclaimed the "biggest revolution in biology that is going to forever transform the way we understand genetics, environment, the way the two interact and what causes disease" in 2007. This sounds strangely familiar, doesn't it? (2016, p. 171)

Along similar lines, Jeungst et al (2014) are highly critical of a practice that they call *epigenetic risk messaging*. This involves an arguably flawed inference from a limited number of animal examples of maternal effects on fetal development to a range of warnings for human parents about potential risks that their behaviours might have for unborn offspring or grand-offspring. The aim of such messaging, according to Jeungst et al (2014) and others (Caufield 2018) is to draw attention to epigenetic research by stoking fear, ultimately with an aim to garner funding.

Our brief discussion of the tangled history of epigenetics suggests a few things to look out for. First, quantitative methods might enable us to determine whether Waddington's "broad sense" conception of epigenetics has been retained, and if so, in which disciplines. Second, these methods might enable us to evaluate the extent to which various heritability commitments have been associated with epigenetic marks. This information, in turn, could provide guidance on whether it makes sense to include a specific mode of inheritance as a part of the narrow-sense definition or, alternatively, whether such a definition should be neutral in its heritability commitment. A third potential benefit of digital humanities methods is to shed light on the kinds of functional interpretations that tend to be associated with epigenetic marks and whether there are differences in perspective across disciplines or over time. Finally, these methods can potentially determine the extent to which specific disciplines are speaking the same language about epigenetics – an important piece of information when considering the promise and commitments of the so-called epigenetic revolution.

3. Qualitative Coding of Epigenetics Abstracts (A Replication Study)

An analysis of scientific abstracts allowed us to replicate and expand upon the recent findings of Linquist and Fullerton (2021). They analyzed approximately 450 scientific abstracts with the term "epigenetic(s)" in the title, keywords, or associated metadata. Abstracts were organized into four "disciplines" that corresponded to Subject Categories supplied by the Web of Science search engine. Two of these disciplines (general biology and evolution) were stand-alone Subject Categories, the other two disciplines (proximal biology and biomedicine) were super-categories containing several thematically related Subject Categories. For instance, "proximal biology" contained abstracts from four Web of Science Subject Categories: cellular

biology, developmental biology, genetics/heredity, and biochemistry/molecular biology. In the original study, the top 25 most cited abstracts were downloaded for each five-year interval between 1995 and 2019. Every abstract was scrutinized for mentions of epigenetic marks and the authors classified these statements according to two, pre-established criteria. The first criterion was to establish a heritability commitment for each abstract. These commitments ranged from very minimal "bare marks" –which made no mention of any heritability—to "mitotically inherited" structures, to "limited meiotic" inheritance (restricted to three or fewer generations), to "open ended" or trans-generational inheritance. Every abstract was also assigned a functional interpretation. These included four potential roles for epigenetic structures: transposon suppression, gene regulation, contributing to disease, adaptive phenotypic plasticity. Using this classification schema, Linquist and Fullerton (2021) were able to identify historical trends and discipline-specific patterns in both the heritability commitments and functional interpretations assigned to epigenetic structures.

Several methodological challenges could be raised with this approach. First, one might question the authors' competence as judges. Neither Linquist nor Fullerton have received extensive training in molecular biology. It is possible that they were unable to detect certain heritability or functional commitments that were relatively implicit in the abstracts. A replication of their methods by an evaluator with formal training in this discipline would help to corroborate their judgments. A second objection pertains to the sample of abstracts. The original study selected only the 25 most cited abstracts in each discipline for each of the five-year periods. The rationale for this choice was that, within the sciences generally, and in stark contrast to philosophy, citation rate is an indication of approval – i.e. articles that are seen as flawed tend to be ignored. Thus, Linquist and Fullerton assumed that the top 25 most cited papers reflect the

approval of researchers within each discipline. We do not contest this assumption. However, a larger sample of highly cited papers might provide greater confidence in the patterns that they observed.

3.1 Replication Methods

To address the first methodological objection of expertise, we employed an advanced student with training in both evolutionary and molecular biology (AG, the third author of this paper). He was provided with descriptions of each type of heritability commitment and functional interpretation, as described in the methods section of Linquist and Fullerton (2021). AG refined these categories slightly in ways that reflected his understanding of the subject matter (see Table 1 for heritability and Table 2 for function). A corpus of abstracts was downloaded using the Web of Science advanced search tools. To address the second methodological objection of sample size, we selected the top 50 most cited abstracts (instead of the previous top 25) for each five-year time period, within each of the four disciplines. In addition, we included an earlier time period (from 1990-1994) in the present study. This resulted in a corpus of 1097 abstracts spanning a 30 year period. Each abstract was evaluated by AG for its heritability and functional commitments. If an abstract contained either no discernable heritability commitment or functional interpretation, it was classified as not applicable ("N/A").

Table 1. Heritability commitments. The following four categories were used to classify heritability commitments associated with epigenetic marks for approximately 1100 scientific abstracts downloaded from the Web of Science using search filters "epigenetic*" in the Topic, which searches title, abstract, author keywords, and Keywords Plus.

Category	Description
Basic mark	The weakest level of heritability commitment, this was assigned to any abstract that mentioned an epigenetic modification (e.g., DNA methylation, histone modification, non-coding RNAs, etc.) but did not comment on heritability. Notably, the direct mention of an epigenetic modification was a sufficient but not necessary for applying this category. Authors often mentioned "epigenetic factors," "alterations," or "mechanisms" without specifying the specific nature of any modification – i.e. what kind of mark it was. They also sometimes mentioned the enzymes responsible for mediating epigenetic processes without mentioning any mark itself. These abstracts are also coded as having basic mark heritability as long as there was no discussion on the persistence of these epigenetic effects throughout mitosis or meiosis.
Mitotic	A stronger commitment than basic mark inheritance, this category was assigned to any abstract that discussed the inheritance of an epigenetic mark throughout mitotic cell division. There is no baseline for how many cell divisions a mark must persist for in order to be assigned mitotic heritability. As such, any abstract that discussed the proliferation of cells as part of an experimental procedure was classified mitotic heritability as long as the epigenetic modification was transmitted to daughter cells. This means that abstracts mentioning the use of cell culture, cell lines, immortalized cell lines, etc. were coded as having mitotic heritability as long as the epigenetic mark was found to persist throughout time. Notably, authors did not have to explicitly state that an epigenetic modification was mitotically inherited for this classification to apply. Processes such as X-inactivation, cellular memory (via maintenance methylation or other pathways), and clonal selection are all examples in which some epigenetic modification is mitotically inherited (e.g. Henikoff and Greally 2106, Mirang & Costello 2017).
Limited Meiotic	Applied to any abstract in which an epigenetic modification persisted throughout meiosis and/or is transmitted to progeny up to the F2 generation. Because this definition extends the heritability of epigenetic modifications to offspring, it is a considerably stronger commitment than mitotic heritability. This classification was also assigned to any abstract that discussed the transmission of an epigenetic mark to offspring without any comment on how many generations the mark is maintained for (e.g., "methylation is a heritable epigenetic modification.") As with mitotic heritability, it was not necessary for authors to explicitly mention the transmission of an epigenetic mark for this designation to be assigned. Processes like genomic imprinting, which involve epigenetic modifications in the germline, also constitute limited meiotic heritability (Tuccci et al 2019).
Open ended	The strongest heritability commitment, open-ended inheritance was reserved for abstracts that explicitly stated or clearly implied multi-generational transmission of an epigenetic mark. For example, abstracts that mention "transgenerational epigenetic inheritance" or the inheritance of epigenetic marks "across generations" go beyond the definition of limited meiotic heritability in that they indicate the possibility of transmission beyond one or two sexual generations. Thus, the demarcation between limited meiotic and open-ended heritability is not a matter of mechanism; rather, it is a matter of longevity (i.e., the duration in which an epigenetic modification/mark is able to persist).

Agreement between the original (Linquist and Fullerton 2021) and this replication study was calculated using a correlation coefficient. To make this calculation, we first summarized all of the assessments across time periods within each discipline. This provided the total number of abstracts for each discipline that had been classified into each of the four heritability or functional categories. The outputs were then transformed into percentages, in order to account for differences in corpus size. We then calculated a Pearson correlation for each of the 16 pairs of heritability assessments (original vs replicated) and again for the 16 pairs of functional interpretations (Figure 1).

To evaluate trends in our corpus, we generated a graphical presentation mirroring that of Linquist and Fullerton (2021). This was done according to the following procedure. First, the number of abstracts assigned to each of the five heritability and functional categories (four categories + N/A) were organized into five-year intervals and expressed as frequencies (Figures 2 &3). Notably, a small proportion of the abstracts were judged as having more than one heritability or functional commitment. In these cases, we followed the procedure employed by Linquist and Fullerton (2021) of selecting the "highest" level of classification. Specifically, for heritability: basic mark< mitotic < limited meiotic < open-ended; and for functionality: disease < TE suppression < gene regulation < adaptive phenotypic plasticity. Qualitative assessments of these graphs allowed us to identify differences among disciplines and temporal trends.

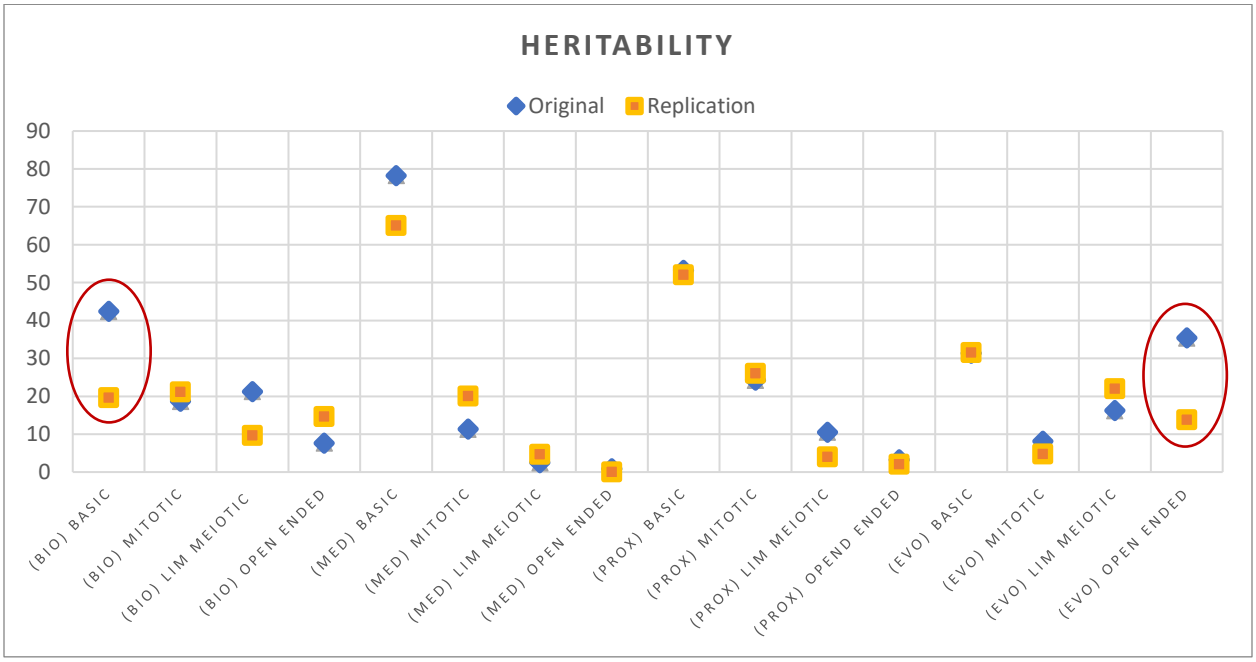
Table 2. Functional interpretations. The following four categories were used by AG to classify functional interpretations associated with epigenetic marks for approximately 1100 scientific abstracts downloaded from the Web of Science using epigenetic*" in the Topic, which searches title, abstract, author keywords, and Keywords Plus.

Category	Description
Disease related	A disease related role was assigned to any abstract in which an epigenetic mark/mechanism was implicated in some process related to disease. This included the promotion or suppression of disease, as well as any mention of epigenetic-based drugs being used to treat disease (e.g., the use of a histone deacetylase inhibitor in cancer). More broadly, this designation was also applied to any abstract that linked epigenetic modifications to deleterious effects in the organism without any specific mention of disease. For example, any abstract linking aberrant epigenetic modification with genomic instability was also classified as having a disease related role. However, it was not necessary for there to be discussion of epigenetic modifications for there to be a disease related functional ascription. The dysregulation of normal epigenetic processes causing disease (e.g., as occurs with genomic imprinting in Prader-Willi syndrome) is an example of how epigenetic mechanisms themselves may be considered disease related (e.g. Butler 2011)
Transposon suppression	Transposon suppression was assigned to any abstract which mentioned the suppression of transposable elements via the action of some epigenetic mark/modification. In addition, any abstract that discussed the activity of PIWI-interacting RNAs (piRNAs) was also classified as having a role in transposon suppression since piRNAs represent an epigenetic RNA-induced silencing mechanism that targets TEs (Deans and Maggert 2015).
Regulation	A regulatory functional role was assigned to any abstract which connected epigenetic modifications to the regulation of gene expression. This definition encompasses any comment on changes in transcriptional activity, gene activation/silencing, chromatin remodelling (e.g., the development of heterochromatin or euchromatin), etc. Although TEs are DNA sequences themselves, the silencing of TEs was not assigned a regulatory role unless the abstract separately mentioned the regulation of gene expression. It is important to note this demarcation in order to ensure that transposon suppression and regulation remain distinct functional categories. Taken more broadly, this functional role was also assigned to any abstract that implicated an epigenetic mark/mechanism in normal development or in the development of some specific phenotype, behaviour, or trait.
Phenotypic adaptation	An abstract was classified as positing phenotypic adaptation when some epigenetic mark/mechanism was described as either adapting an organism's phenotype to its environment or maintaining/stabilizing a beneficial phenotype. Abstracts that mentioned phenotypic plasticity were not automatically coded as having an adaptational role; rather, the most necessary prerequisite for this designation was for the authors to explicitly state that this plasticity was adaptive. Any mention of plasticity that was not said to be adaptive was assigned a regulatory role instead.

3.2 Results and Discussion

Heritability judgments were strongly correlated across the original study and our replication (Pearson's r = 0.892), suggesting that neither the judges' expertise nor sample size were biasing the original heritability results (Figure 1A). However, two categories showed a noteworthy discrepancy: the proportion of "basic marks" in general biology abstracts were approximately 20% higher in the original study. Likewise, the proportion of "open ended" heritability commitments were approximately 15% higher in the original study. Looking at the heritability results broken down over time, we can infer that these differences were partially influenced by the inclusion of an earlier period (1990-1994) in the replication study, where there was a larger proportion of "N/A" coded abstracts. This had the effect of diluting the overall number of abstracts for which a heritability commitment could be identified. In addition, it was our impression that AG was more reluctant to classify a given abstract as open-ended, defaulting more often to limited meiotic inheritance as a likely interpretation. Specifically, AG required that an abstract explicitly mention more than two generations to qualify as a case of open-ended inheritance (see Table 1). In contrast, the original study also classified an abstract as committed to open-ended inheritance if it discussed the evolutionary consequences of epigenetic inheritance. The latter judgment makes sense on the assumption that an epigenetic mark can only be of evolutionary significance if its effects persist for multiple (e.g. more than three) generations. To our knowledge this is a fairly widely held position, though not without its detractors (Stotz and Griffiths, 2016).

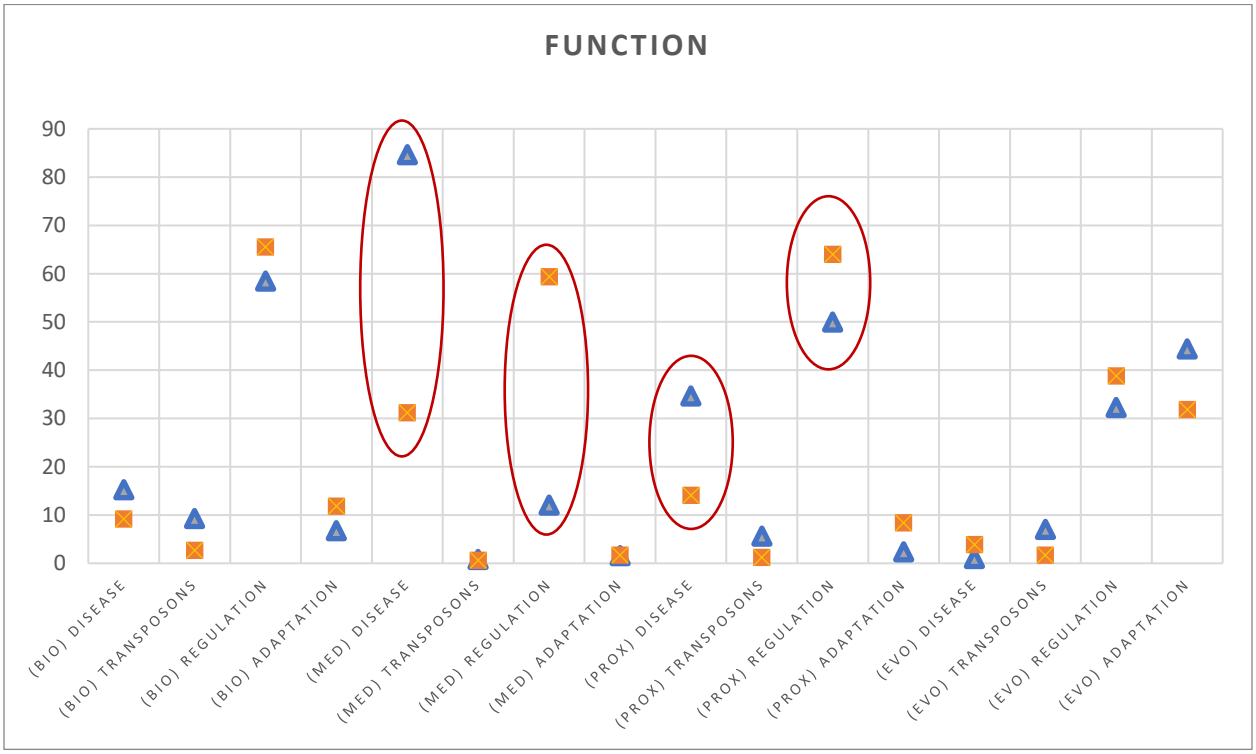
Figure 1. Original versus replicated heritability judgments. There was a strong correlation between the original study and the current replication that involved an independent evaluator and a larger sample size (Pearson's r= 0.89). Two points of divergence were the proportion of general biological abstracts committed to "basic marks" and the proportion of evolutionary abstracts that posited "open ended" functions (outlined in red).



Turning to our replication of Linquist and Fullerton's (2021) functional interpretations, the overall level of correlation was somewhat lower than for heritability judgments (Pearson's r= 0.66). However, this discrepancy was restricted to just four out of the sixteen categories (Figure 2). Within biomedicine, the original study classified over 80% of the abstracts as "disease" and less than 15% as "regulation." Our replication exhibits the inverse pattern, with roughly 60% of the biomedical abstracts classified as "regulation" and about 30% as "disease." The same inversion appears in the proximal biology category, but to a lesser degree. We offer two explanations for these inversions. First, AG possesses a greater familiarity with the specific molecular mechanisms described in biomedical abstracts. He was thus more often inclined to infer an underlying regulatory role for a given epigenetic mark, even if the focus of an abstract was primarily on its implications for disease. Second, as it was noted earlier, our replication

study included an earlier time period from 1990-1994. As it was discussed in the previous section, during this period the study of epigenetics was largely about methylation and this type of mark is associated with gene regulation.

Figure 2. **Original versus replicated functional interpretations.** There was a moderately strong correlation between the original study and a replication that involved an independent evaluator and a larger sample size (Pearson's r= 0.66). Most of the divergence was due to a much higher tendency in the replication study to judge biomedical and proximal biology abstracts as "regulatory" instead of merely "disease" (outlined in red).



Our replication study revealed five noteworthy trends resembling those described by Linquist and Fullerton (2021).

Trend 1: Biomedicine and proximal biology exhibited minimal heritability commitments.

In the original study, the most popular heritability commitment in both biomedicine and proximal biology (78% and 53% respectively) was that of bare marks– i.e. these abstracts described the effects of epigenetic marks (e.g. on disease or gene regulation) without alluding to

the inheritance of those marks. Our replication likewise found that proximal biology resembled biomedicine in its focus on bare marks as the primary heritability commitment (Figure 3). In other words, even in this larger sample of abstracts that were independently coded, we find that most discussions of epigenetic marks carry no commitment to heritability. In fact, the proportions of abstracts that referred just to bare marks (no explicit inheritance) increased over time across all four disciplines, especially biomedicine. The discipline of general biology provides a representative picture of this trend. In the earliest period (1990-1994), almost 50% of general biological abstracts were classified as "N/A" because they referred to no type of epigenetic mark whatsoever and so their heritability commitments could not be easily determined. However, from 2000-2004 onwards, general biology abstracts came to refer more explicitly to molecular structures and, in addition, exhibited a mixture of heritability commitments that favoured basic marks and limited meiotic inheritance over mitotic and open-ended inheritance. This pattern supports Haig's (2004, 2012) analysis that the term has come to refer to a range of specific structures in these disciplines while not carrying a strong association with any particular form of heritability.

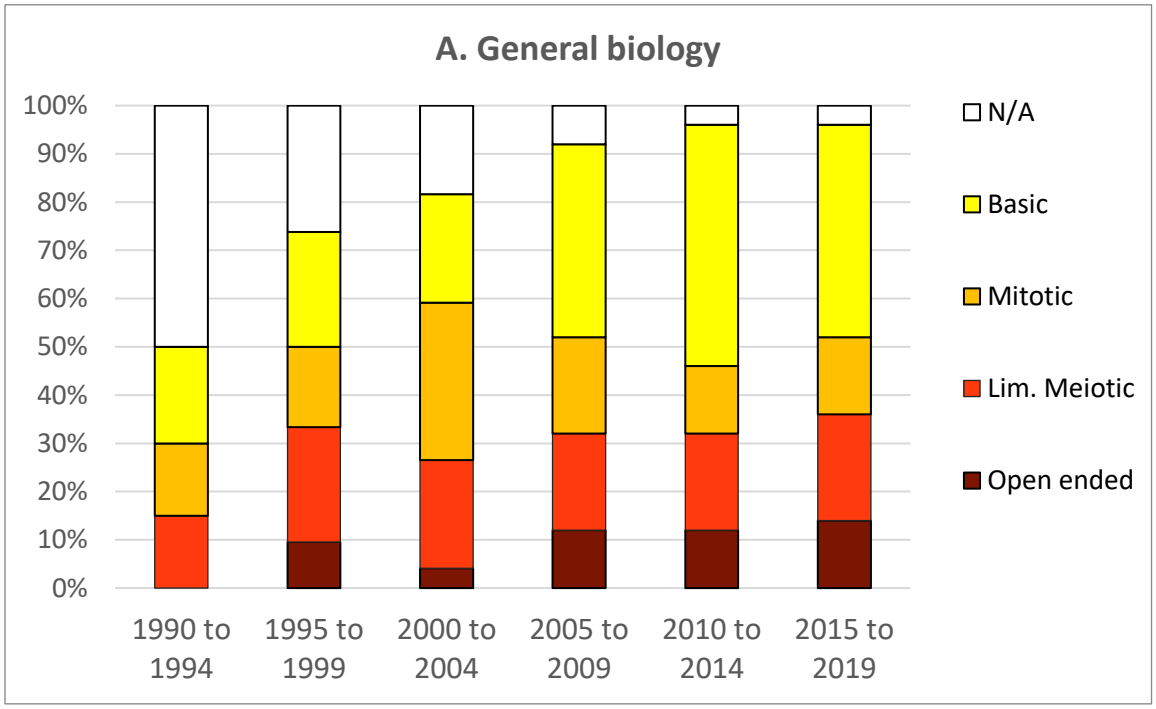
Open ended inheritance was mentioned in only < 1% of biomedical abstracts, all appearing in the earliest (1990-1994) period. Likewise, open-ended inheritance was mentioned in < 3% of proximal abstracts, smattered across the different time periods. General biology provides a point of useful comparison. In this discipline, open-ended inheritance was mentioned in 6% of the abstracts and, like biomedicine and proximal biology, showed no historical trend.

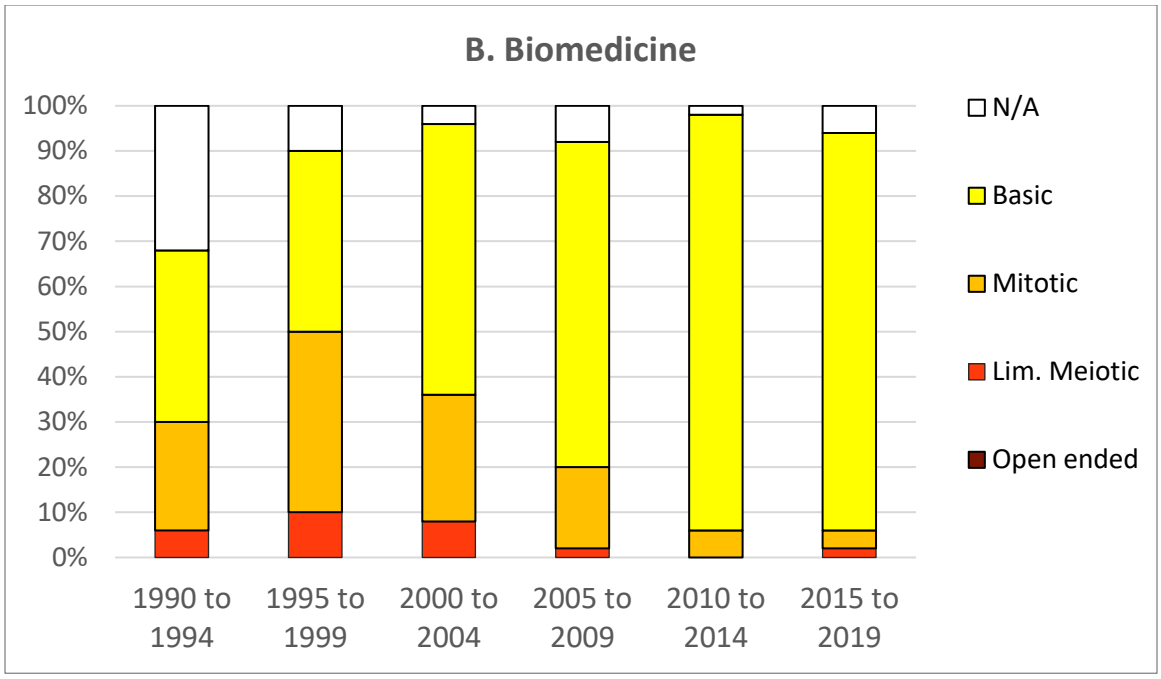
Trend 2: Evolution exhibited a relatively high frequency of open-ended inheritance.

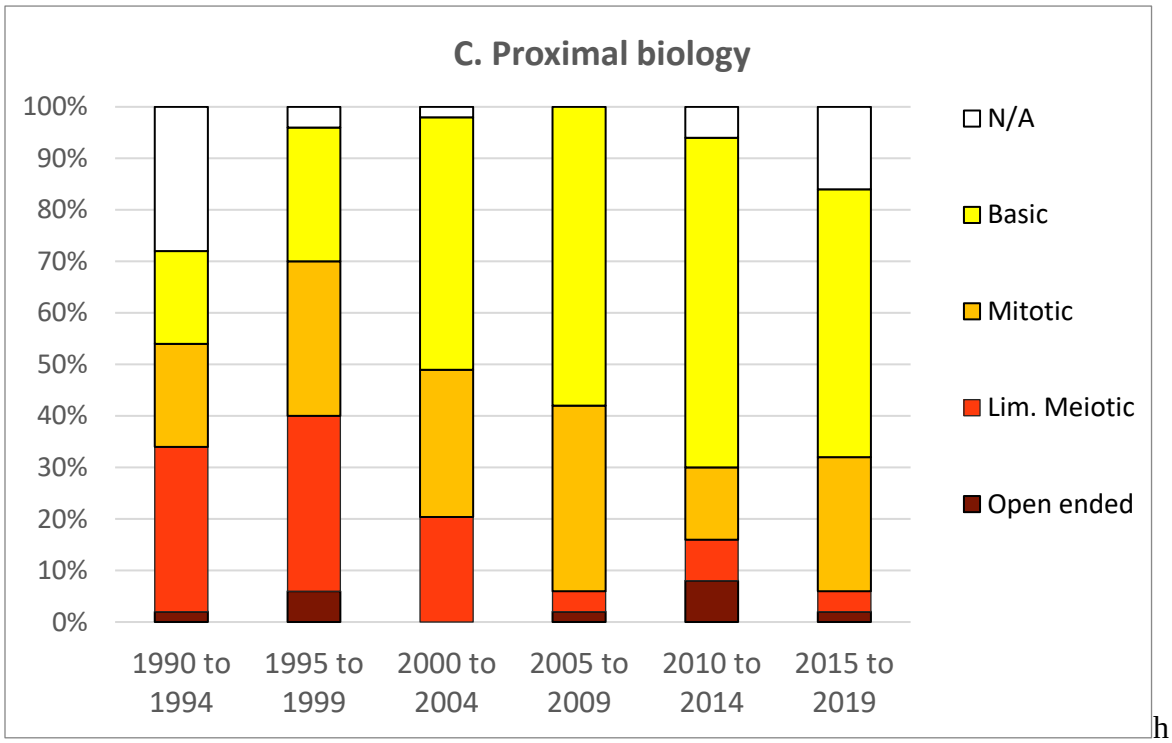
Linquist and Fullerton (2021) reported that open-ended inheritance was the most popular mode of inheritance discussed in evolutionary abstracts, accounting for 36% of the heritability

commitments summed across all time periods. In our sample (which included an earlier historical period and more abstracts) this pattern was somewhat less pronounced, with open-ended inheritance accounting for only 16% of the evolution abstracts. Nonetheless, this is a much higher frequency than in any of other the other three disciplines. It is also noteworthy that, like Linquist and Fullerton (2021), we observed an increase in the popularity of this functional interpretation over time (Figure 3D). Specifically, the proportion of evolution abstracts mentioning open-ended inheritance increased to over 20% of the abstracts after 2009. Although this is lower than the proportion reported by Linquist and Fullerton (2021), as it was noted earlier, the coder in this study (AG) adopted a slightly more conservative coding method.

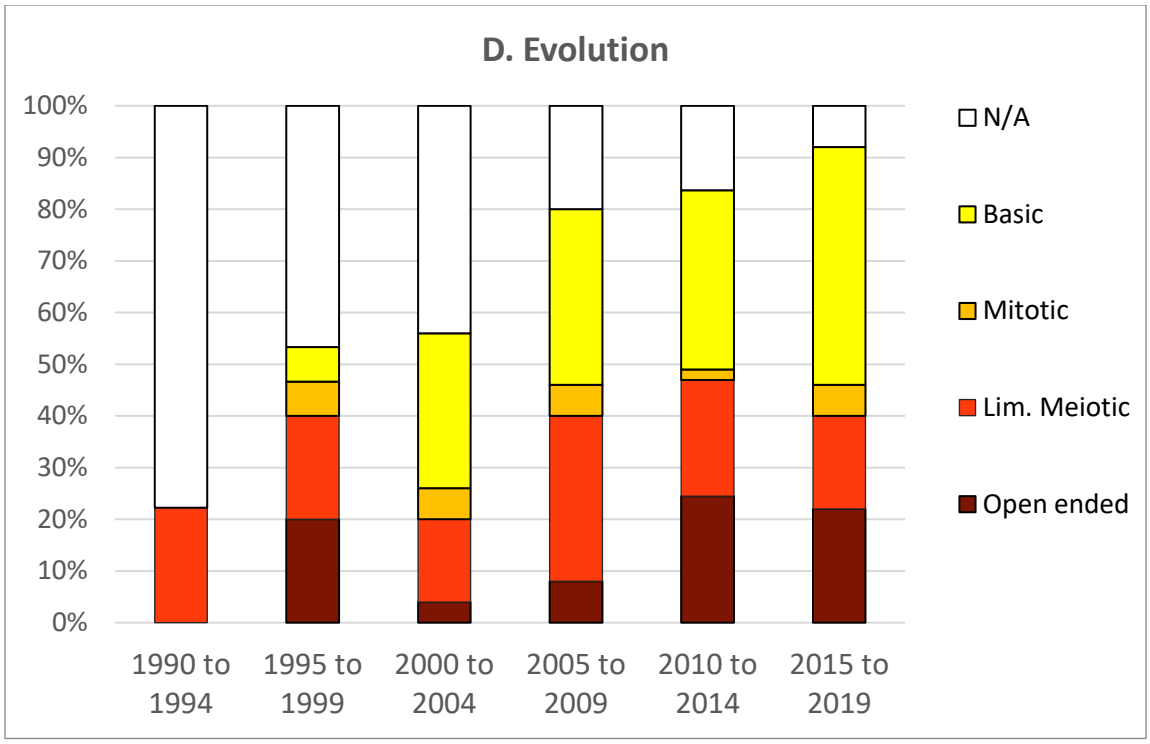
Figure 3. Heritability commitments. Heritability commitments attributed to epigenetic marks in a sample of the maximum 50 most cited abstracts in each of four disciplines over a thirty-year period, organized into five-year intervals.




B. Biomedicine
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1990 to 1994
1995 to 1999
2000 to 2004
2005 to 2009
2010 to 2014
2015 to 2019
N/A
Basic
Mitotic
Lim. Meiotic
Open ended


C. Proximal biology
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1990 to 1994
1995 to 1999
2000 to 2004
2005 to 2009
2010 to 2014
2015 to 2019
N/A
Basic
Mitotic
Lim. Meiotic
Open ended

h



Trend 3: Proximal and general biology exhibited a predominant focus on regulation.

Linquist and Fullerton (2021) reported that the most common functional role assigned to epigenetic marks in biomedicine was (predictably) disease related (85%). However, they observed that the disciplines of general and proximal biology focussed primarily on gene regulation (56% and 51% respectively). Within evolutionary biology, regulation was the most common functional role attributed to epigenetic marks from 1995-2004. However, this interpretation declined from 2005 onward. These trends were strongly supported by our replication (Figure 3). In addition, a large proportion of the biomedical abstracts also attributed regulation to epigenetic marks. Based on these results, it is fairly clear that among all disciplines that we studied – with the noteworthy the exception of evolution –if epigenetic marks were functionally interpreted, they were usually seen as playing a role in gene regulation.

Trend 4: Evolution exhibited growing interest in adaptive phenotypic plasticity.

According to Linquist and Fullerton (2021), adaptive phenotypic plasticity was a rare functional interpretation in all disciplines except evolution and, in this discipline, this interpretation increased dramatically after 2005. Specifically, this controversial functional interpretation jumped from roughly 12% of the evolution abstracts in 2000-2004 to over 50% in 2005-2009, and then climbed up to 60% in 2015-2019. Our replication data clearly confirm this trend (Figure 4D). In this larger sample of abstracts, we see an increase in the frequency of this functional interpretation that reaches a level of almost 60% in the most recent period. However, the overall proportion of evolutionary abstracts (summed across historical periods) mentioning adaptive plasticity was lower in this replication than in the original study. This could have been due either to expertise or to the larger sample. To drill down on this discrepancy, we conducted a further analysis in which the top 25 most highly cited evolutionary abstracts in our current sample was re-analyzed separately from the additional, next 25 most cited ones that we had included to increase sample size. This comparison revealed that adaptive plasticity was more than twice as likely to be mentioned in the top 25 most cited evolution abstracts. Hence, expanding our sample to include slightly less frequently cited abstracts had a diluting effect on the proportion of evolutionary abstracts that posited adaptive plasticity as a candidate function. In other words, it appears that this functional interpretation is somewhat restricted to the most highly cited evolution abstracts in which the term "epigenetic(s)" is used.

Trend 5: Decline in functionally indeterminate (N/A) abstracts over time.

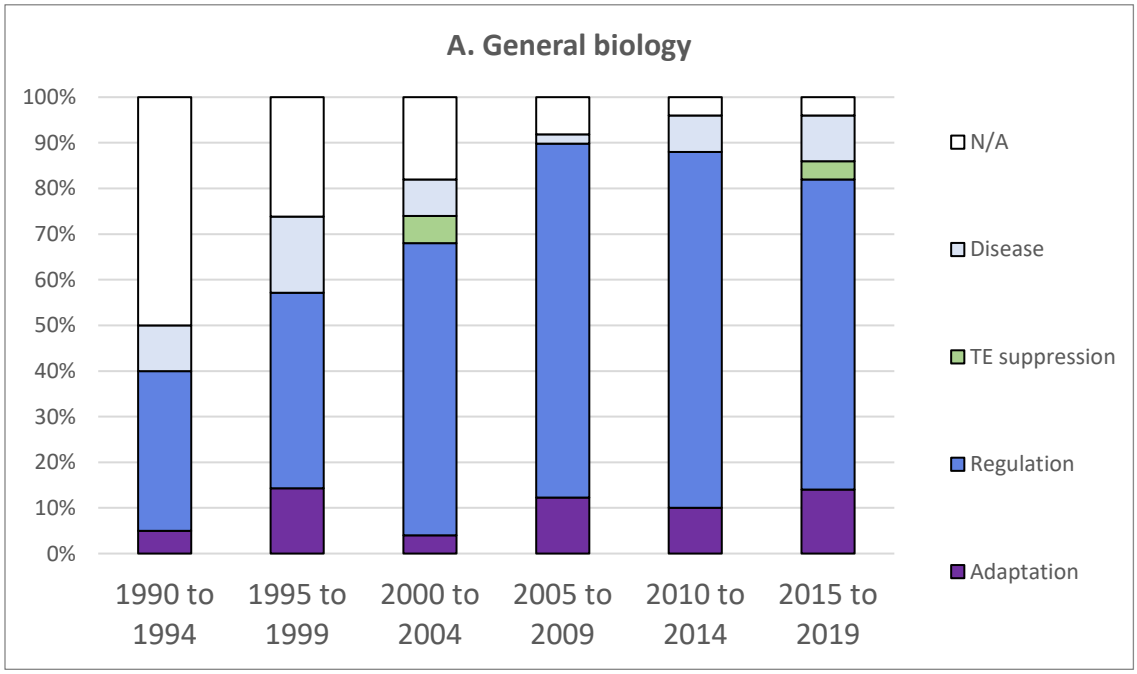
A fifth and novel trend was detected in our replication study. Within evolution there was a steady decline in the proportion of abstracts that were classified as functionally "N/A." For instance, in the 1990-1994 period almost 80% of the abstracts could not be functionally

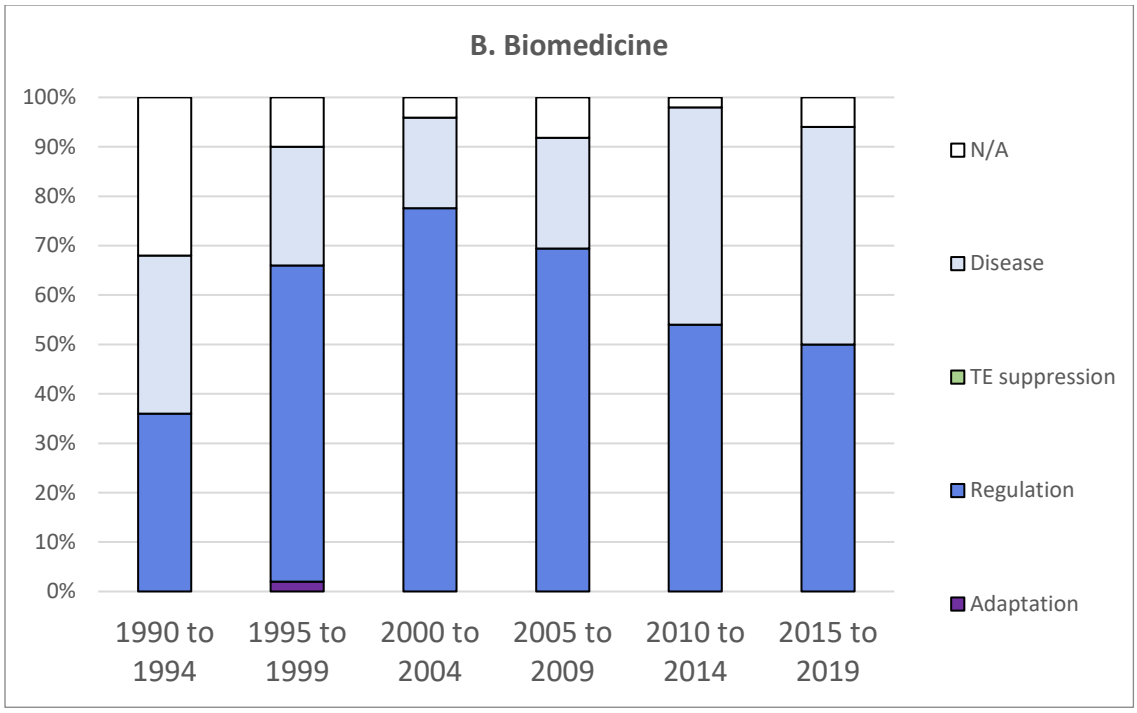
interpreted. Importantly, this does not necessarily mean that these abstracts were neutral on the topic of epigenetic function. Rather, earlier abstracts in evolution tended to discuss a much wider and somewhat motley array of subjects, many of which did not even mention any sort of epigenetic mark. If an abstract did not mention epigenetic marks even implicitly, then, accounting to our coding schema, a functional interpretation could not be assigned.

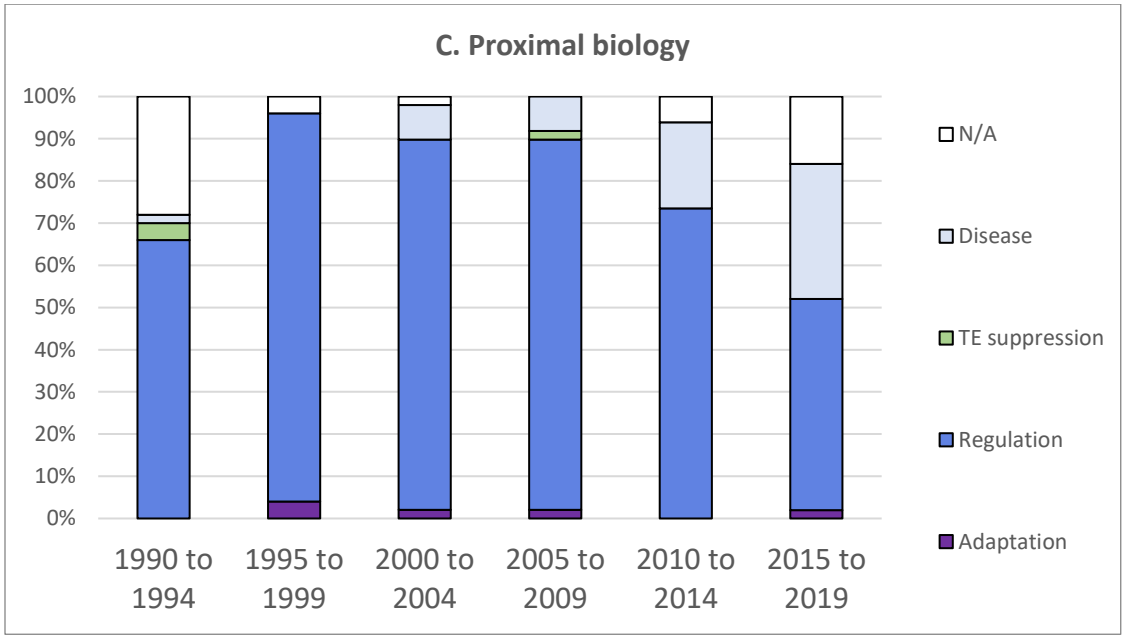
Putting all of this together, what does this qualitative coding study tell us about the meaning of "epigenetic(s)" across disciplines and over time? Generally speaking, our findings concur with Linquist and Fullerton (2021) that the disciplines of proximal biology and biomedicine are largely speaking the same language when it comes to epigenetics. In both disciplines, authors attributed minimal (if any) heritability to epigenetic marks and they were primarily interested in the functions of gene regulation and disease. These trends were mirrored in general biology, though to a lesser extent. In contrast, evolutionary biology stood out as a conceptual outlier. Evolutionary abstracts exhibited much higher levels of commitment to open-ended inheritance than other disciplines. Most pronounced however was the large and steadily increasing interest in adaptive phenotypic plasticity. On the one hand, these differences might not seem entirely surprising given that evolutionary biology tends to be interested in inter-generational phenomena. On the other hand, however, both phenomena have received only scant empirical support in eukaryotes (Sánchez-Tójar et al 2020). Other data reported by Linquist and Fullerton (2021) help to account for this apparent deviation from standard usage. In a separate study, they compared the total number of abstracts in the entire Web of Science database that mentioned "epigenetic(s)" across each of the four disciplines. This analysis of several hundred thousand abstracts did not involve reading individual abstracts, it looked only at relative frequencies in each discipline of papers that discuss epigenetics in some sense or other.

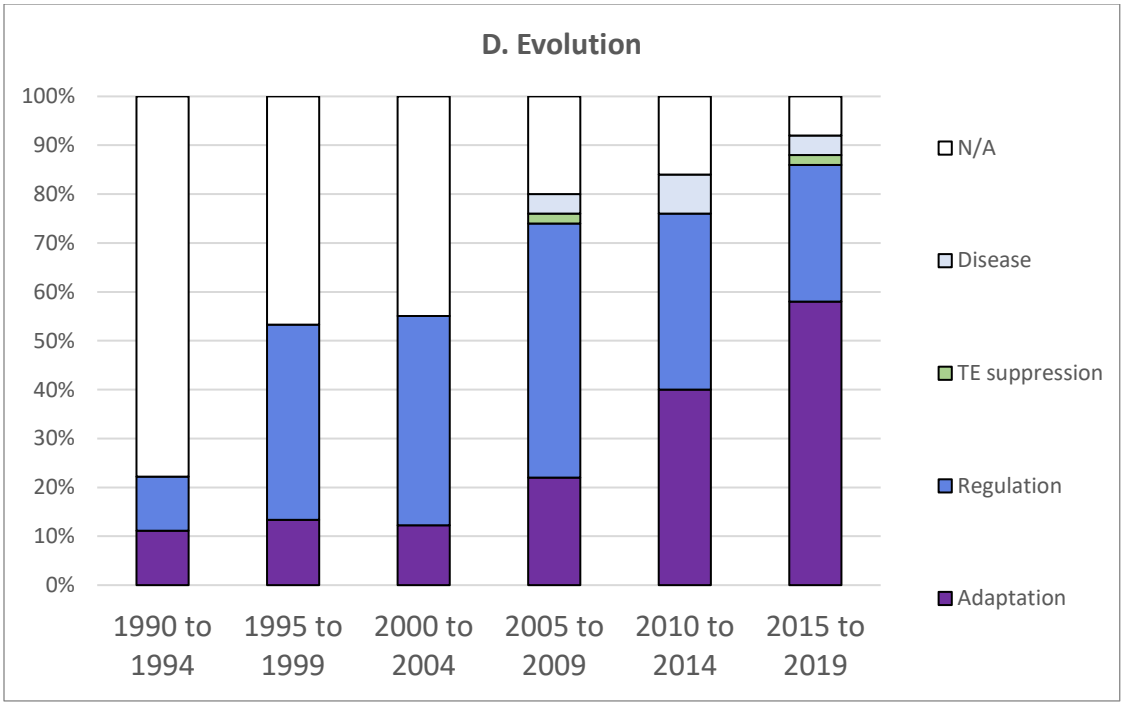
Nonetheless, their analysis revealed that “epigenetics” is much less frequently mentioned overall in evolutionary abstracts compared to the other three disciplines, controlling for differences in overall publication rates. For instance, whereas epigenetics appeared in > 12% of proximal biology abstracts and > 8% of biomedical articles, it appears in only about 3% of evolutionary abstracts. Hence, although a large proportion of evolution papers that mention epigenetics did involve open-ended inheritance, this is arguably not representative of the discipline, since only a small proportion of papers in evolutionary biology mention epigenetics in the first place. Our sample of highly cited evolutionary epigeneticists abstracts could therefore constitute a somewhat fringe group of scientists.

Figure 4. Functional interpretations. Functional interpretations attributed to epigenetic marks in a sample of the 50 most cited abstracts in each of four disciplines over a thirty-year period, organized into five-year intervals.




B. Biomedicine
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1990 to 1994
1995 to 1999
2000 to 2004
2005 to 2009
2010 to 2014
2015 to 2019
N/A
Disease
TE suppression
Regulation
Adaptation


C. Proximal biology
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1990 to 1994
1995 to 1999
2000 to 2004
2005 to 2009
2010 to 2014
2015 to 2019
N/A
Disease
TE suppression
Regulation
Adaptation



4. Topic modelling the same corpus of epigenetic abstracts

The previous replication study identified an ambiguity in "epigenetic(s)" associated with particular disciplines as well as some temporal dynamics in the heritability and functional interpretations within disciplines. A limitation of this approach, however, was that the coding criteria used to classify each abstract (Tables 1 & 2) were fairly course grained. This was evidenced, for instance, by the large proportion of "N/A" coded abstracts in the 1990-1994 period (Figure 3). Simply put, there might be more to the concept of epigenetics than its heritability and function – although these two factors have been central points of discussion in the philosophical and scientific literatures. To get a broader understanding of the uses of this term across disciplines, we generated a number of topic models using the same corpus as in the replication study. This method allows themes or "topics" to emerge out of a large body of

literature without the same sorts of biasing effects that can result from a pre-established coding system.

4.1 Topic Modelling Methods

Topic models were built using the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in the Python package Gensim (Rehurek & Sojka, 2011). Data visualization and exploration was completed using the pyLDAvis package (Sievert & Shirley, 2014). This dataset contained the same set of abstracts that were coded in the earlier, qualitative coding analysis. Data cleaning was relatively straightforward and involved only removing stop words as defined by the NLTK (Loper & Bird, 2002), and lemmatizing using SpaCy (Honnibal & Montani, 2017). Some further pre-processing could have been useful in grouping synonyms etc. but because of the relatively clean nature of abstracts, as opposed to full-text papers, we opted for the method that allowed us to work with the data quickly and flexibly. It should also be noted that during the construction of the corpus we allowed the inclusion of bigrams. This proved very useful in interpreting topics in certain situations. Gensim LDA models were completed for the corpus of every abstract for a given subject category (biomedicine, evolutionary biology, proximal biology, and general biology). The random state was set to 100, chunk size was set to 100, and passes were set to 10. While we had the computing power to complete more passes with larger chunk sizes, we found the results were not noticeably easier to interpret and the result was simply a more time-consuming process. This was important because we had to run the modelling process numerous times to find the ideal number of topics. In our first attempt at topic modelling, we built one set of topics for each subject category using the computed coherence score to find the most coherent number of topics from twenty-five or fewer topics. The cutoff at twenty-five topics was chosen because we found that larger numbers of topics became too esoteric and uninterpretable.

In order to make this study comparable to the qualitative coding replication, described earlier, it was important to capture a temporal dimension. However, separating abstracts into 50-article chunks would have generated corpuses too small to benefit from LDA modelling. We therefore divided each discipline into two corpuses of roughly equal size: an earlier set of abstracts published before 2005, and a later set published after 2004. This gave us two topic models for each of the four disciplines.

A general methodological question when employing topic models concerns the number of topics to generate. A persistent worry is that this is a significant source of bias, where individual researchers are free to select the level of topic-resolution that best suits their prior expectations. To mitigate this worry, we selected the number of topics according to the highest *coherence score* for each corpus. Coherence can be calculated easily with Gensim and is taken as a reliable indication of how interpretable a topic is by calculating the semantic similarity between the highest scoring terms in the topic. Twenty-five topics was chosen as the maximum cut-off because we found the process of coding and interpreting anything above this number of topics to be more laborious than it was worth.

Interpretation of topic models took place in two stages. First, each author of this paper independently assigned a label to every topic when possible and made notes on what they took it to be about. We then convened as a group to compare those notes and decide on labels and salient features for each topic. Only points of agreement among all three authors are reported below. For purposes of consistency in our analysis of the topics as well as our presentation of them (Figure 5), we relied on a relevance metric –or "lambda"– in pyLDAvis of 0.5, which we found to be the most interpretable. This practice draws on the work of Sievert and Shirley (2014)

who found in their own work that a lambda of 0.6 was optimal for the correct identification of topics (p. 67).

4.2 Topic Modelling Results and Discussion

Topics were analyzed in several ways. First and foremost was the assignment of a label to most topics. This label reflects our semantic interpretation a given topic's subject matter –the theme which, based on our interpretations of the relevant words and their orderings, we took each topic to be about. These assignments are influenced by a researcher's understanding of the relevant subject. Hence the labels that we attached to various topics might differ from those that would be assigned by others. Therefore, we encourage readers to look carefully at the word lists associated with each topic and to come up with their own interpretations (see Tables 3-6).

A second type of analysis relied on the pyLDAvis software package. This visualization software represents topics as circles within a two-dimensional landscape. Each circle represents one topic. The size of a given circle reflects its relative popularity in the overall sample of abstracts. For example, if a topic is represented by a large circle, then it appears in a greater proportion of the abstracts compared to a relatively small circle. This visualization software also represents the amount of topic overlap within a corpus. For example, if two circles overlap to a large extent, then the corresponding topics have a relatively high degree of word-order similarity compared to other circles exhibiting less overlap.

These two modes of analysis are complementary. The first, labelling phase tells us which topics feature prominently within a given sample of abstracts. The second, two-dimensional visualization reveals the overall organization of those topics – their relative popularity and thematic overlap. Taken together, we get a picture of the overall "topic landscape" for a given

sample of abstracts. Although this mode of topic model interpretation requires further validation, our analysis identified some interesting temporal and disciplinary trends. What follows is a discussion of each of the two topic models (pre 2005 and post 2004) for all four of the disciplines under investigation.

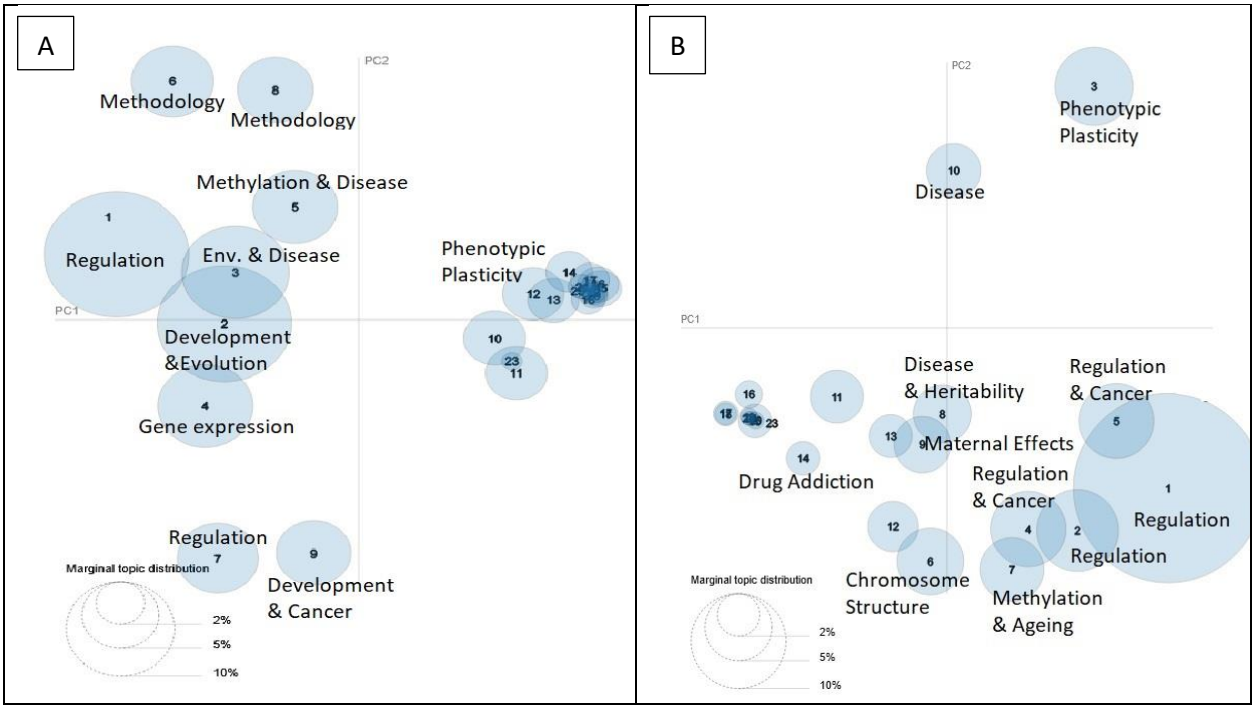
4.2.1 General Biology Topic Models

General biology provides a useful baseline for understanding the three more specialized disciplines. Recall that the number of topics in a given model was established using a coherence score, up to a limit of 25 topics. The fact that general biology exhibited a roughly consistent number of topics across the two time periods (pre-2005 = 23 topics, post-2005 = 25 topics) suggests that this discipline did not become more thematically differentiated in subject matter, at least, not at a level of resolution that we were able to detect.

Looking at the 2-D visualization of topics in general biology for the pre-2005 model (Fig. 5A) reveals two general thematic clusters that are separated along the first principal component (x -axis). Some of the terms that are highly ranked across multiple topics in the first cluster (Topics 1-9), but rare or absent from the second cluster (Topics 10-23) include: development, chromatin, protein, mechanism, role, gene, and silence. In contrast, highly ranked terms appearing in the second cluster, but largely absent from the first, include: character, phenotypic, selection and variation. The first set of terms arguably refer to specific epigenetic marks and their mechanistic roles in development. The second cluster seems to be geared towards natural selection at the phenotypic level. Hence, we can infer that within general biology prior to 2005, these were the two predominant themes in popular discussion about epigenetics. One theme described molecular mechanisms, the other discussed phenotypic evolution.

Looking more closely at specific topics in the pre-2005 model, most of the larger topics received a consensus classification (Appendix 1). They were typically assigned labels that described the effects of epigenetic marks on either gene regulation or disease. Notably, none of them mention heredity. Although Topic 2 was labeled "development and evolution," none of the popular terms in this topic referred to heredity and the only explicit mention of evolutionary phenomena was "evolutionary theory." Interestingly, one of the smaller topics received the label "phenotypic plasticity and disease." Looking at the associated words, however, one sees little if any mention of heredity.

Figure 5. Topic models for General Biology corpus (A) 1990-2004 and (B) 2005– 2019. generated by LDA using pyLDAvis visualization at λ 0.5. See Appendix 1 for word lists.



Turning to the model of general biology abstracts published after 2004 (Figure 5B) we see two general clusters separated along the second principle component (y axis). A large cluster of mostly overlapping topics positioned in the lower half of the landscape is separated from two

satellite clusters at the other end. Common terms shared by the most of the first cluster included: function, activity, target, protein, chromatin, histone, methylation, mark, domain, DNA, and sequence. It seems reasonable that this large cluster describes epigenetic marks and their functions in gene regulation. One of the satellite topics (Topic 3) was labelled "adaptive phenotypic plasticity". The other satellite (Topic 10) was associated with disease.

At the level of individual topics, we again achieved consensus in our classifications for most topics. The dominant topic in general biology, post 2004 (Topic 1), described epigenetic marks involved in gene regulation – a consistently dominant theme from the earlier period. However, this more recent topic model reveals a much broader emphasis on biomedical issues including cancer (Topics 4 & 5), aging (Topic 7), disease heritability (Topics 8 & 14), maternal effects and disease (Topic 9) and drug addiction (Topic 14). Notably, the satellite topic of adaptive phenotypic plasticity (Topic 3) has grown in significance since the earlier period, now appearing in 6.8% of the abstracts. Also notable are topics on transposable elements and evolution (Topic 15) and genome structure (Topic 20).

To summarize our impression of the change in topic distributions in general biology across the two periods, we found: (1) a central focus on gene regulation across both periods, (2) a broader range of epigenetic marks and mechanisms besides just methylation in the later period, (3) slightly increased emphasis on phenotypic plasticity in the later period, and (4) greater emphasis on genome structure and transposable elements in the more recent period. Perhaps most noteworthy, however, was (5) relatively little mention of heredity across both periods.

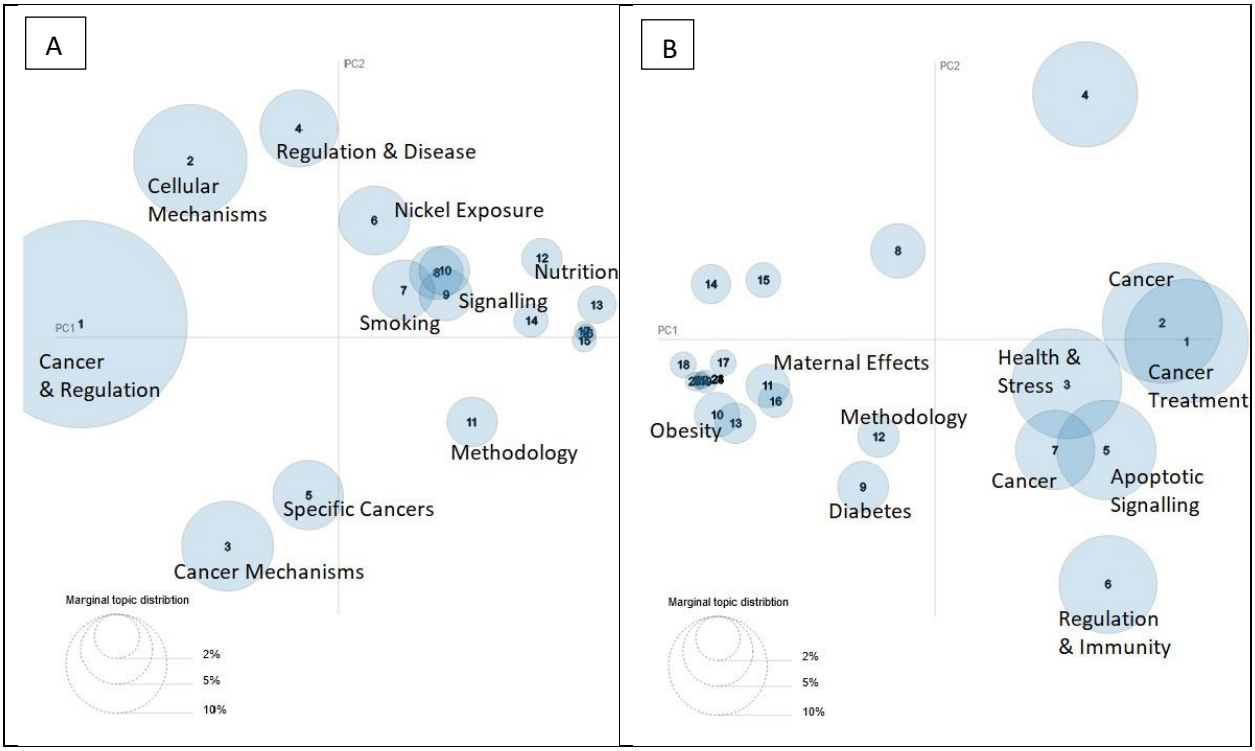
4.2.2 Biomedicine Topic Model

At the optimum levels of coherence, the earlier set of biomedical abstracts (pre-2005) generated a model of only 17 topics, compared to 24 topics in the later model (post-2004). Notably, the 2-D visualization of the earlier model consisted of a single dominant topic surrounded by several medium-sized satellites with little overlap (Fig 6A). In contrast, the more recent model has no single dominant topic, but instead a number of medium sized topics with a considerable amount of overlap (Fig 6B). Comparing just these 2-D visualizations, it would appear that the discipline of biomedicine has both become more differentiated (more topics) and more integrated (overlapping) in its subject matter. Of particular interest is the cluster of large topics (1-7, with the exception of 4). The dominant topic in this group (Topic 1) was clearly on methylation and its effects on cancer (Appendix 2). Most of the other large topics (2-11) also dealt with specific cancers, but often with less emphasis specifically on methylation. Only one topic (11) seemed to touch on methodological issues, and only two interpretable topics (14 and 15) were not predominantly cancer related.

Although post-2004 topics were still focused to a large extent on cancer (Topics 1,2,7,12, 13, 15, 16, 17), there was also a sizable number on other subjects (Table 4B). For instance, exposure and general health (Topic 3), apoptotic cell signalling (Topic 5) and diabetes (Topic 9).

Overall, we note the following changes in the biomedicine topics over time: (1) lower emphasis on methylation in later period, (2) Similar emphasis on cancer and its causes across both periods, (3) a greater range of topics in the more recent period. In addition, we found (4) relatively little mention of heredity in either period.

Figure 6. Topic models for Biomedicine corpus (A) 1990-2004 and (B) 2005 – 2019. generated by LDA using pyLDAvis visualization at λ 0.5. See Appendix 2 for word lists.



4.3 Proximal Biology Topic Model

At the peak level of coherence, the pre-2005 proximal biology model generated only 12 topics. In contrast, the post 2005 model generated 24 topics. As in the case of biomedicine, and in contrast to general biology, we took this to suggest a differentiation into a greater number of subjects. Comparing the landscape visualizations of each model, we see a pattern that is also reminiscent of biomedicine. Specifically, the earlier period was dominated by a large topic surrounded by medium-sized satellites and a cluster of small topics (Fig 7A). The later period exhibited a larger number of overlapping topics (Fig. 7B). If we divide the earlier model into two clusters along the first principal component, the cluster of large topics (1-3) share the following common terms that are not so frequent in the other cluster (Topics 7-12): mechanism,

methylation, heterochromatin, transcriptional, promoter, alternation, inactivation, and change. These terms clearly refer to epigenetic marks and their roles in gene regulation (Appendix 3).

The dominant topic in the pre-2005 model was the same as for the other two disciplines: the role of epigenetic marks in gene regulation (also Topic 6). Large satellite topics focussed on diseases (Topics 2, 3) including cancer (Topics 7,8,9), the process of epigenetic reprogramming (Topic 3), and limited meiotic inheritance (Topic 5), with a separate topic for the inheritance of disorders (Topic 7). Among the smaller satellites were a topic on the agouti mouse model organism (10), a topic about gross morphology (11) and plant genetics (12). It is interesting to note that heredity does show up as a topic in this earlier model of proximal biology abstracts in the context of parental effects, though we see no mention of trans-generational inheritance.

Topics in the post 2005 model were even more specific and well defined. Notably, not only was a biomedical topic (Topic 1) the most dominant in this period, it specifically emphasized such promotional terms as “drug, discovery, disease, therapy, future, health, promise.” Such language was infrequent in any other topic that we reviewed and it was surprising to see it as the dominant topic in post-2004 proximal biology. The role of epigenetic marks in gene regulation was the second most common topic in the more recent group of proximal abstracts. Although this topic was largely focused on methylation, other types of epigenetic mark featured prominently in other topics. In particular, there were several topics on noncoding RNA (7,8,11, and 13). This is consistent with what would be predicted on Haig’s historical analysis and consistent with what we observed in other disciplines. Namely, a broader range of molecular entities and mechanisms in the more recent period. In addition to the usual array of cancer related topics (3,4,15), there were specialized topics on clinical methods and

technologies (Topic 4, 12), maternal effects (Topic 5), immunity and aging (Topic 6) and methodological reflection (Topic 9).

Figure 7. Topic models for Proximal Biology corpus (A) 1990-2004 and (B) 2005 – 2019 generated by LDA using pyLDAvis visualization at λ 0.5. See Appendix 3 for word lists.

Comparing the two proximal biology models reveals the following: (1) greater diversity of more specific topics in the more recent period; (2) broader range of epigenetic marks in the later period, especially noncoding RNA; (3) prominent interest in maternal effects in both periods; (4) greater interest in biomedical applications in the later period, especially framed in popularizing language. Finally, (5) although there were relatively small topics associated with heredity in the earlier period, these do not persist into the later period. This could indicate reduced emphasis on heredity within the more recent abstracts.

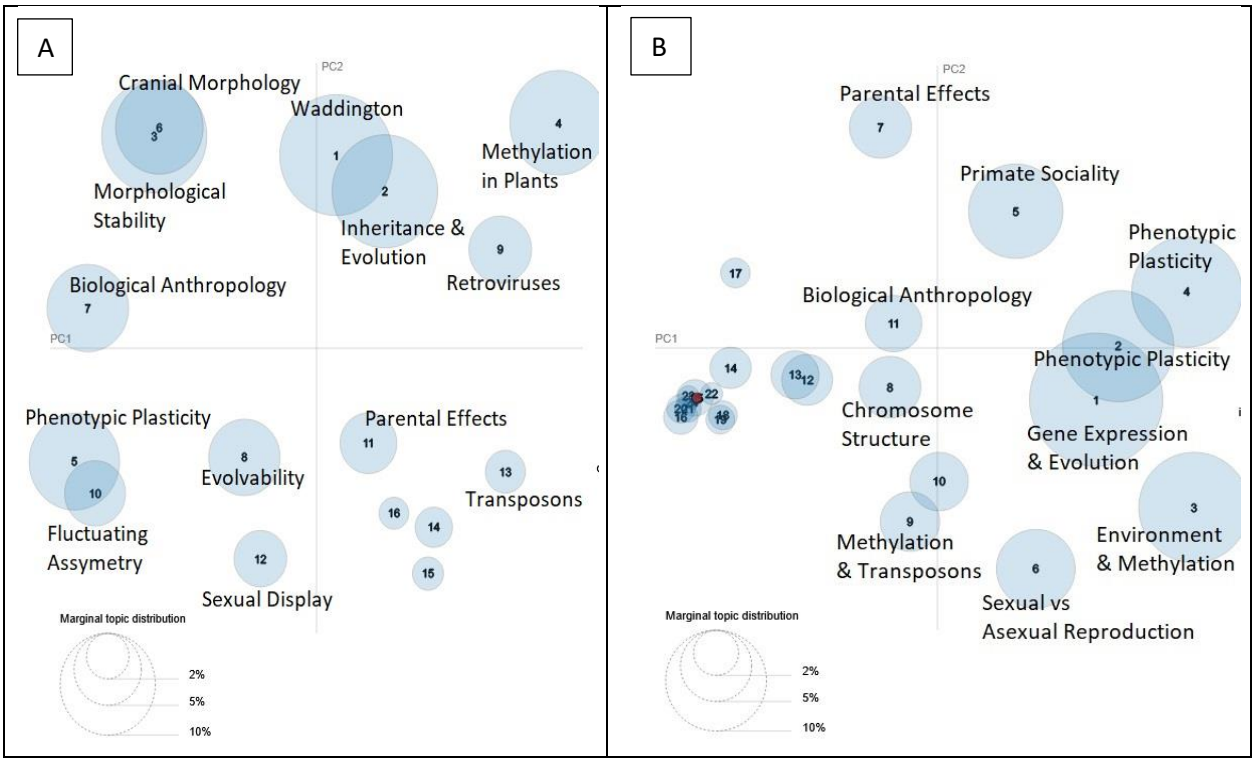
4.4 Evolution Topic Model

Abstracts in evolutionary biology also exhibited increased differentiation in their subject matter from pre-2005 to post 2005. The peak-coherence model for the earlier period generated 15 topics, whereas the peak-model for the latter period generated the maximum possible 25 topics. In the 2-D visualization, we see a fairly broad distribution in both models across the two principal components, with no noticeable shift in topic dominance or overlap (Figure 8A, 8B). However, differences across time periods are evident in the word lists. Several of the most popular topics in the earlier abstracts (Topics 1, 7, 8) contained Waddingtonian language including "canalized, canalization, landscape" (Appendix 4). Arguably, Topics 2 and 3 are more reflective of a Nanneyan picture of epigenetics, as they discuss inheritance and phenotypic stability – although, Waddington's conception of genetic assimilation features in Topic 2. Generally speaking, this particular topic model dovetails with Stotz and Giffiths' (2016) analysis of "epigenetics" as having both a broad sense Waddingtonian definition and a narrow sense Nannayan definition, both of which have evolutionary associations. Another notable feature of this topic model is a relatively strong emphasis on heredity, both in Topic 2 (labelled "epigenetic inheritance") and Topic 11 ("parental effects").

Things look quite different however in the second model of more recent evolution abstracts (Table 8B). Only one of these topics contains any Waddingtonian language (8), and it represents a mere 3.7% of the corpus. The most popular topic in this period (Topic 1) addresses gene regulation and evolution. However, the next three most prominent topics (2-4) are very clearly about adaptive phenotypic plasticity. Notably, Linquist and Fullerton also reported "a sharp and dramatic swing in the prevalent functional interpretation of epigenetic markers" in evolutionary biology that happened at around 2005" (2021, p. 150). This representation supports

their assessment. Also noteworthy is the appearance of parental effects (Topic 7) in this model – a form of limited meiotic, as opposed to open-ended inheritance.

Figure 8. Topic models for Evolution corpus (A) 1990-2004 and (B) 2005-2019 generated by LDA using pyLDAvis visualization at λ 0.5. See Appendix 4 for word lists.



Comparing the evolution models across the two periods reveals the following: (1) diversification of topics in the more recent abstracts, (2) predominant Waddingtonian and Nanneayan themes in the earlier period, (3) a shift away from Waddingtonian language in the more recent period, (4) increased emphasis on adaptive phenotypic plasticity in the more recent period, from just one fairly minor topic pre-2005 to multiple large topics post 2004.

5. Conclusions

What do these two digital humanities methods teach us about the meaning of "epigenetics?" Perhaps the most consistent finding on which both studies converged on a distinction between two senses of "epigenetic" that followed disciplinary boundaries. Within disciplines that focus on proximate mechanisms, such as molecular biology and biomedicine, the most highly cited epigenetics abstracts rarely discuss hereditary phenomena. This was revealed in the qualitative coding study by the prevalence of "bare marks" as the dominant heritability commitment in biomedicine and proximal biology. Likewise, topic models from these two disciplines contained little if any mention of hereditary terms. A second point of convergence is an emphasis on gene regulation and disease as the two dominant functional interpretations, not only within biomedicine and proximal biology, but also in general biology abstracts. Again, this pattern was clearest within the qualitative coding study. However, it was also quite noticeable in the topic models for biomedicine, proximal, and general biology where gene regulation and disease were dominant themes. From these two points of convergence we infer that there exists a fairly stable, proximal conception of epigenetics that can be stated as follows. The term "epigenetic" in the proximal sense refers to a growing list of structures associated with DNA that are involved in gene regulation and disease but which have no specific association with any particular pattern of inheritance, especially not trans-generationally.

In contrast, the evolutionary sense of "epigenetic" that emerged from our studies was markedly different in heritability commitments and functional associations. Qualitative coding revealed a persistent association with open-ended or transgenerational inheritance that has emerged since 2005. Interestingly, the earlier (pre 2005) topic model for evolutionary abstracts emphasized inheritance more strongly than the latter model, but this could have been due to a

greater divergence in topics over time. Taken together, our findings suggest that within the subset of evolutionary articles that discuss epigenetics it is widely assumed that epigenetic structures are transmitted over multiple generations. Even more striking perhaps is the commitment within evolutionary biology to adaptive phenotypic plasticity. Qualitative coding revealed a steady increase in this functional interpretation (Figure 4D). Likewise, phenotypic plasticity was the most common label for topics in the post-2005 topic model of evolutionary abstracts. (Table 6B). Hence, a rough definition of this evolutionary conception can be stated as follows. The term "epigenetic" refers to structures associated with DNA that are involved in the flexible adaptation of organismal phenotypes to the environment and which are often inherited trans-generationally.

Although other authors have described ambiguity in the meaning of epigenetic(s), to our knowledge, this study is the first to demonstrate empirically an association with specific disciplines. Our analysis nonetheless supports certain elements of Haig's (2004, 2007) analysis of this concept. Recall that Haig offered a dynamical interpretation of this concept which initially was associated with Waddingtonian themes, referred to a specific type of structure (methylation marks) during the 1990s, and which has come to refer to a wider range of structures in more recent decades. Haig also challenged the idea that epigenetics has ever been closely associated with any specific type of heredity, especially trans-generationally. Our topic models in particular reveal a growing list of different structures under the labels "epigenetic mechanisms." Likewise, as we have noted, heritability commitments are largely absent in most disciplines aside of evolution. Our analysis departs from Haig's in a few key respects. Most notable perhaps is our identification of an evolutionary sense of epigenetic that does carry explicit heritability commitments -in particular to trans generational inheritance. Our analysis also documents a

broader distribution of both heritability commitments and functional interpretations while tracking their change over time. In this respect our analysis is finer grained.

Our results lent more qualified support to Stotz and Griffiths (2016) analysis. We found little evidence for their broad sense or Waddingtonian definition of epigenetics, except in the earliest evolution abstracts. These showed up in the qualitative coding study as "N/A" coded abstracts, which were common in the 1990s but declined steadily over subsequent 5-year periods. Evidence for a broad sense conception also appeared in the form of Waddingtonian terminology, but only in the earlier of the two evolution topic models. We infer from these converging lines of evidence that by the 1990s, the broad sense definition was restricted to evolutionary biology and on its way out. Stotz and Griffiths (2016) also discussed a "narrow sense" definition that refers to specific molecular mechanisms thought to be heritable and quite likely involved in adaptive phenotypic plasticity. Importantly, they maintain that that the narrow sense definition is the received usage within the proximal sciences. Our results partially support the idea that there is a distinct conception of epigenetics that is prevalent in the proximal sciences. However, we find so support for the suggestion, emphasized by Stotz and Griffiths (2016), that the proximal conception of epigenetic has anything to do with transgenerational inheritance nor with adaptive phenotypic plasticity. On the contrary, it would appear that Stotz and Griffiths' narrow-sense definition corresponds most closely to what we have described as the evolutionary conception of epigenetic. As we have demonstrated, this conception is restricted to the evolutionary abstracts in our sample, which in turn represent a small fraction of evolution abstracts more generally (Linquist and Fullerton, 2021). Although the proximal conception of epigenetics is common in biomedicine, molecular biology and elsewhere, it would be a mistake to think that these disciplines are frequently talking about open-ended inheritance or adaptive

phenotypic plasticity. Conflating the relatively unpopular evolutionary conception of epigenetics with the much more widely used proximal conception could paint a seriously distorted picture of what the epigenetic revolution is all about. Perhaps the most important methodological lesson for philosophers concerns the danger of the disproportionate influence coming from a fairly narrow group of evolutionary researchers who are especially vocal about the prospects and promise of epigenetics.

Reflecting more generally on these digital humanities methods, we are encouraged by the extent to which the topic modelling corroborated many of the same trends identified by qualitative coding. For instance, the absence of heritability commitment and strong emphasis on gene regulation and disease within the proximal sciences came through as a clear signal in both studies. Likewise, the evolutionary emphasis on trans generational inheritance and especially adaptive plasticity was a striking outcome in both studies. It is noteworthy that the topic model revealed certain patterns that were not as noticeable with qualitative coding. Perhaps the most interesting of these can be described as the *biomedicalization* of general and proximal biology. By this term, we mean not only that discussions of epigenetics have focused more on disease-related topics, but also that these discussions have become more self-promotional in this regard. For instance, in general biological abstracts published before 2005 (Table 3A) there was only minimal reference to disease (Topic 3), and one other relatively small topic (9) on the subject of cancer. In contrast, general biological abstracts published after 2004 (Table 3B) contained two large topics on cancer (4 &5), as well as topics on aging (7), general diseases (8,14), drug addiction (14), and cognitive disorders (18). Similarly in proximal biology, prior to 2005 there was a substantial emphasis on disease (Topics 2, 4, 8). However, the dominant topic in post-2005 abstracts was a promotional topic touting the “positive” “future” benefits of epigenetic research

on disease and drug therapy. In addition, we identified 4 topics on cancer (3,8,14,15), two topics on aging (6,18) and one on general disease (13). This pattern underscores a criticism made by some molecular biologists that "epigenetics" has become a buzzword that functions more as a rhetorical tool for garnering funding and less as a descriptive term for specific molecular mechanisms (Deans and Maggert 2015; Häfner and Lund 2016). Indeed, the emphasis in more recent abstracts on such a wide range of potential human diseases (from obesity to cognitive dysfunction) is reminiscent of Jeungst et al's (2014) complaint against epigenetic risk messaging. An important question for future research is the extent to which "epigenetics" persists as a popular scientific term simply on account of its rhetorical value.

6. References

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bonduriansky, Russell, and Troy Day. 2020. *Extended Heredity: A New Understanding of Inheritance and Evolution*. Princeton University Press.

Brandon, Robert N. 2014. *Adaptation and Environment*. *Adaptation and Environment*. Princeton University Press. https://doi.org/10.1515/9781400860661.

Butler, Merlin G. 2011. Prader-Willi Syndrome: Obesity due to Genomic Imprinting. *Current Genomics* 12(3): 204-215.

Deans, Carrie, and Keith A Maggert. 2015. "What Do You Mean, 'Epigenetic'?" *Genetics* 199 (4): 887–96. https://doi.org/10.1534/genetics.114.173492.

Doolittle, W. Ford. 2013. "Is Junk DNA Bunk? A Critique of ENCODE." *Proceedings of the National Academy of Sciences* 110 (14): 5294–5300. https://doi.org/10.1073/pnas.1221376110.

Elliott, Tyler A., Stefan Linquist, and T. Ryan Gregory. 2014. "Conceptual and Empirical Challenges of Ascribing Functions to Transposable Elements." *The American Naturalist* 184 (1): 14–24. https://doi.org/10.1086/676588.

Garson, Justin. 2016. *A Critical Overview of Biological Functions*. SpringerBriefs in Philosophy. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32020-5.

Godfrey-Smith, Peter. 2001. "Three Kinds of Adaptationism." In *Adaptationism and Optimality*, edited by Steven Hecht Orzack and Elliott Sober, 1st ed., 335–57. Cambridge University Press. https://doi.org/10.1017/CBO9780511609084.012.

Graur, Dan, Yichen Zheng, Nicholas Price, Ricardo B.R. Azevedo, Rebecca A. Zufall, and Eran Elhaik. 2013. "On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE." *Genome Biology and Evolution* 5 (3): 578–90. https://doi.org/10.1093/gbe/evt028.

Griffiths, Paul E., and Eva M. Neumann-Held. 1999. "The Many Faces of the Gene." *BioScience* 49 (8): 656–62. https://doi.org/10.2307/1313441.

Griffiths, Paul, Edouard Machery, and Stefan Linquist. 2009. "The Vernacular Concept of Innateness." *Mind & Language* 24 (5): 605–30. https://doi.org/10.1111/j.1468-0017.2009.01376.x.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy : An Introduction*. Cambridge: Cambridge University Press.

Häfner, Sophia J., and Anders H. Lund. 2016. "Great Expectations – Epigenetics and the Meandering Path from Bench to Bedside." *Biomedical Journal* 39 (3): 166–76. https://doi.org/10.1016/j.bj.2016.01.008.

Haig, D. 2004. "The (Dual) Origin of Epigenetics." *Cold Spring Harbor Symposia on Quantitative Biology* 69 (January): 67–70. https://doi.org/10.1101/sqb.2004.69.67.

Haig, David. 2007. "Weismann Rules! OK? Epigenetics and the Lamarckian Temptation." *Biology & Philosophy* 22 (3): 415–28. https://doi.org/10.1007/s10539-006-9033-y.

———. 2012. "Commentary: The Epidemiology of Epigenetics." *International Journal of Epidemiology* 41 (1): 13–16. https://doi.org/10.1093/ije/dyr183.

Henikoff, Steven and John M. Greally. 2016. Epigenetics, cellular memory and gene regulation. *Current Biology* 26(14): 644-648.

Hull, David L. 2010. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. *Science as a Process*. University of Chicago Press. https://doi.org/10.7208/9780226360492.

Iwasaki, Yuka W., Mikiko C. Siomi., and Haruhiko Siomi. 2015. PIWI-Interacting RNA: Its Biogenesis and Functions. *Annual Review of Biochemistry* 84: 405-433.

Jablonka, Eva, and Marion J. Lamb. 1995. *Epigenetic Inheritance and Evolution: The Lamarckian Dimension*. Oxford University Press.

———. 2014. *Evolution in Four Dimensions, Revised Edition: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. MIT Press.

Keller, Evelyn Fox. 2014. "From Gene Action to Reactive Genomes." *The Journal of Physiology* 592 (11): 2423–29. https://doi.org/10.1113/jphysiol.2014.270991.

Kim, Mirang and Joseph Costello. 2017. DNA methylation: an epigenetic mark of cellular memory. *Experimental & Molecular Medicine* 49: e322.

Linquist, Stefan, and Brady Fullerton. 2021. "Transposon Dynamics and the Epigenetic Switch Hypothesis." *Theoretical Medicine and Bioethics* 42 (3): 137–54. https://doi.org/10.1007/s11017-021-09548-x.

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *ArXiv:Cs/0205028*. http://arxiv.org/abs/cs/0205028

Love, A. C. 2015. "Developmental Biology." In *The Stanford Encyclopedia of Philosophy*.

Millikan, Ruth. 2002. "Biofunctions: Two Paradigms." In *Functions: New Essays in the Philosophy of Psychology and Biology*, edited by Andre Ariew, Robert Cummins, and Mark Perlman. Clarendon Press.

Moss, Lenny. 2004. “What Genes Can’t Do - Google Books.” 2004. https://www.google.ca/books/edition/What_Genes_Can_t_Do/AGm7TgaKSR0C?hl=en&gbpv=1&dq=lenny+moss+genes&printsec=frontcover.

Nanney, D. L. 1958. “EPIGENETIC CONTROL SYSTEMS*.” *Proceedings of the National Academy of Sciences of the United States of America* 44 (7): 712–17.

Nicoglou, Antonine. 2018. “Waddington’s Epigenetics or the Pictorial Meetings of Development and Genetics.” *History and Philosophy of the Life Sciences* 40 (4): 61. https://doi.org/10.1007/s40656-018-0228-8.

Okasha, Samir. 2006. “Evolution and the Levels of Selection - Samir Okasha - Google Books.” 2006. https://books.google.ca/books?hl=en&lr=&id=aw9REAAAQBAJ&oi=fnd&pg=PR9&dq=evolution+and+the+levels+of+selection&ots=Dm69_NnUwd&sig=3O_GSUOqL2fxa5Z2MHbH2Syy8bQ&redir_esc=y#v=onepage&q=evolution%20and%20the%20levels%20of%20selection&f=false.

Oyama, Susan. 2000. “Causal Democracy and Causal Contributions in Developmental Systems Theory.” *Philosophy of Science* 67 (September): S332–47. https://doi.org/10.1086/392830.

Palazzo, Alexander F., and T. Ryan Gregory. 2014. “The Case for Junk DNA.” *PLOS Genetics* 10 (5): e1004351. https://doi.org/10.1371/journal.pgen.1004351.

Palazzo, Alexander F., and Eliza S. Lee. 2015. “Non-Coding RNA: What Is Functional and What Is Junk?” *Frontiers in Genetics* 0. https://doi.org/10.3389/fgene.2015.00002.

Ptashne, Mark. 2007. “On the Use of the Word ‘Epigenetic.’” *Current Biology* 17 (7): R233–36. https://doi.org/10.1016/j.cub.2007.02.030.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2), 2.

Rosenberg, Alex. 2006. “Is Epigenetic Inheritance a Counterexample to the Central Dogma?” *History and Philosophy of the Life Sciences* 28 (4): 549–65.

Rosenberg, Alexander. 2006. *Darwinian Reductionism: Or, How to Stop Worrying and Love Molecular Biology*. Chicago, IL: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/D/bo4094660.html.

Santana, Carlos. 2014. “Save the Planet: Eliminate Biodiversity.” *Biology & Philosophy* 29 (6): 761–80. https://doi.org/10.1007/s10539-014-9426-2.

Sarkar, Sahotra. 1998. *Genetics and Reductionism*. Cambridge University Press.

———. 2002. “Defining ‘Biodiversity’; Assessing Biodiversity.” *The Monist* 85 (1): 131–55. https://doi.org/10.5840/monist20028515.

Schaffner, Kenneth F. 1993. *Discovery and Explanation in Biology and Medicine*. University of Chicago Press.

Sievert, C., & Shirley, K. E. (2014). *LDAvis: A method for visualizing and interpreting topics*.

Sober, Elliot. 1984. *The Nature of Selection. Bradford*. MIT Press, Cambridge.

Stotz, K., and P. E. Griffiths. 2004. “Genes: Philosophical Analyses Put to the Test.” *History and Philosophy of the Life Sciences*. https://doi.org/10.1080/03919710412331341621.

Stotz, K., and P. E. Griffiths. 2016. "Epigenetics: ambiguities and implications." History and Philosophy of the Life Sciences 38(4): 22. doi: 10.1007/s40656-016-0121-2

Stotz, Karola, Paul E. Griffiths, and Rob Knight. 2004. “How Biologists Conceptualize Genes: An Empirical Study.” *Studies in History and Philosophy of Science Part C: Studies in*

History and Philosophy of Biological and Biomedical Sciences 35 (4): 647–73. https://doi.org/10.1016/j.shpsc.2004.09.005.

Takacs, David. 1996. *The Idea of Biodiversity: Philosophies of Paradise*. Johns Hopkins University Press.

Tucci, Valter., Anthony R. Isles., Gavin Kelsey., and Anne C. Ferguson-Smith. 2019. Genomic Imprinting and Physiological Processes in Mammals. *Cell* 176(5): 952-965.

Waters, C. Kenneth. 1994. "Genes Made Molecular." *Philosophy of Science* 61 (2): 163–85. https://doi.org/10.1086/289794.

Weber. 2004. "Philosophy of Experimental Biology - Marcel Weber - Google Books." 2004. https://books.google.ca/books?hl=en&lr=&id=1-9MSsqv5DwC&oi=fnd&pg=PP1&dq=Weber+experimental+biology&ots=STU3k3lxlA&sig=EEaOMJ5iNEy-3vfgOA08HYcGWVc&redir_esc=y#v=onepage&q=Weber%20experimental%20biology&f=false.

Wilkins, John S. 2009. *Species: A History of the Idea*. *Species*. University of California Press. https://doi.org/10.1525/9780520945074.

Zhou, Wanding, Gangning Liang, Peter L. Molly, Peter Jones. 2020. DNA methylation enables transposable element genome expansion. Proceedings of the National Academy of Science 117 (32): 19359-19366.

Appendix 1 General biology. Topic model of 50 most cited abstracts per 5 year interval published between (A) 1990-2004 and (B) 2005-2019, generated by LDA using pyLDAvis visualization at λ 0.5.

A. 1990- 2004

Topic # / label	**% corpus**	**Word list**
(1) The epigenetic regulation of genes	17.9%	Transcription, regulation, expression, involve, antisense, epigenetic, suggest, protein, mutant, mechanism, RNA, role, gene, silence, act, recently, implicate, require, chromosome, action, inheritance, maternal, plant, well, background, occur, sequence, state, highly, chromatin
(2) Development and evolution	15.3%	System, cell, group, drosophila, impact, biology, evolutionary, theory, activation, focus, conserve, physically, synthesize, mediate, genetic, review, developmental, yeast, influence, phase, functional, exist, study, exert, morphogenesis, future, critical, consistent, variant, homolog
(3) Environmental factors / disease / cell signalling / methylation	9.8%	Ecs, stress, exposure, decrease, degreesc, thermal, alteration, xist, even, weight, pre, respectively, temperature, congenital, survive, increase, disease, condition, methylation, affect,

		production, blood, later, reduce, compare, potentially, responsible, block, body
(4) Epigenetic effects on gene expression	7.7%	Post, transcriptional, link, rate, control, interesting, disrupt, originally, effect, mutation, bind, repression, expose, purpose, suppress, analyze, rapidly, gene, vivo, degree, level, usually, grow, dependence, specifically, silence, reversible, function, imprinted gene, emerge
(5) Effects of methylation on disease and evolution / cell differentiation	6.1%	Cancer, methylation, commitment, evolutionarily, year, demethylation, event, contribute, DNA, deficient, need, methylate, somatic, current, summarize, minority, precise, mechanistic, maintain, parental, biological, create, discuss, development, risk, normal, reprogramme, epigenetic, embryo, germline
(6) Methodological factors	5.7%	Initiation, genome, site, genomic, appear, replication, mammalian, DNA, human, study, frequency, recent, hypomethylation, principle, duplication, potentiate, suppressor, result, connect, interpretation, make, change, individual, simultaneously, important, region, understand, phenomenon, methylation, line
(7) Gene regulation	5.5%	Express, cytokine, gene, cell, raise, subject, possibility, receptor, damage, manner, terminal, fate, moreover, various, promote, report, class, apoptotic, allele, stimulation, repair, hematopoietic, stimulate, thus, induction, modulate, regulate, regard, variety, nore
(8) Methodological* (	4.9%	Hat, activity, modification, histone, acetylation, multiple, tail, element, code, provide, replicate, protein, hypothesis, finding, evidence, finally, determine, rather, long, addition, domain, regulation, require, particular, important, new, prion, rich, yeast, maintenance,
(9) Development and cancer	4.7%	Growth, tumor, cell, cycle, factor, gata, signal, represent, nervous, differentiation, differentiate, maturation, intrinsic, helper, secrete, distinguish, guide, respond, program, dependent, treatment, confer, extend, time, heritable, work, form, interaction, smad, transcriptional
(12) Phenotypic plasticity and disease	3.2%	Liver, surprisingly, oxidative, alpha, oncogene, cytogenetic, alone, age, induced, genotype, plasticity, peak, mouse, increase, myc, tumorigenesis, hematopoietic, adaptive, environmental, developmental, irradiate, character, incidence, phenotypic, adult, landscape, cluster, strong, apoptosis, genetic

(17) Chromosome structure	1.3%	Kinetochore, satellite, cenp, centromere, plate, centromeric, inner, nucleosome, purify, segregation, alpha, organization, molecular, assembly, classical, reflect, recognition, sequence, enrich, together, antisera, tagging, location, carboxy, ultrastructural, trilaminar, cse, neocentromere, substructure, dicentric

B. 2005 -2019

Topic label	**% corpus**	**Word list**
(1) Regulation of gene expression and development	39.9%	Gene, expression, mechanism, methylation, epigenetic, variation, DNA, change, genome, stress, control, plant, social, genetic, associate, show, remain, telomere, role, use, silence, identify, term, express, demonstrate, result, find, animal, individual
(2) Regulation and genome architecture	7.4%	acquire, transition, composition, feature, code, nuclear, aspect, network, resistance, poise, paradigm, RNA, transcription, neural, facilitate, genotype, range, temporal, less, address, pre, architecture, ctcf, dynamic, chromatin, drug, pluripotency, principle, dispensable, depletion
(3) Adaptive phenotypic plasticity	6.8%	Inheritance, effect, organism, epigenetic, exposure, environmental, environment, adaptive, evolutionary, plasticity, parent, developmental, offspre, consequence, adaptation, evolution, biology, consider, utero, expose, population, phenotype, great, condition, life, summarize, experience, intrauterine, understand, generation
(4) Epigenetic effects on regulation and cancer / TE suppression	6.3%	Cellular, post, tumor, protein, functional, module, source, RNA, pcg, turn, suppressor, method, bring, diversity, family, small, piwi, defect, oncogenic, illustrate, tumor, quality, elucidation, quantity, certain, absent, parallel, opportunity, spermatogenesis, need
(5) Regulation and cancer	6.2%	LncRNA, mediator, focus, impact, include, therapeutic, various, investigation, cancer, diverse, potential, inflammatory, research, review, nutrition, physiological, shape, unique, highlight, putative, macrophage, acetylation, exception, example, inhibitor, technology, mechanism, conservation, emerge, tissue
(6) Regulation and centromere structure	4.9%	Kinetochoer, DNA, methylation, centromere, structure, progression, accumulation, mark, produce, identify, promise, responsible, detect, microtubule, attachment, approximately, retain, surprisingly, enable cenh, RNA directe, specificity,

		question, spindle, summary, efficient, hypomethylation, last, typical, mitotic
(7) Methylation and ageing	4.5%	Age, enhancer, motif, bind, stem, cpg, aging, demethylase, suppression, DNA, hematopoietic, coordinate, site, sra, maturation, rich, loss, kyp, asymmetry, ring, methylate, methylcytosine, domain, activity, substrate, cg, concurrent, attenuation, shorten, carefully
(8) Methylation, disease, and heritability	3.7%	Tet, reprogramme, somatic, lie, cell, convert, climate, autism, repair, erasure, differentially, brain, advance, poorly, ber, apply, line, scale, ability, correspond, excision, capture, supply, restrict, inactivation, base, subsequently, disorder, germline, inhibition
(9) Maternal effects	3.5%	Obesity, microRNA, onset, blood, placental, vertebrate, fetal, respectively, assumption, help, schizophrenia, patient, previous, epigenomic, testis, biomarker, alter, cortisol, cord, impair, promoter, programming, analyze, nmis, unprecedented, donor, prominent, select, marrow
(10) Disease	3.3%	Disease, health, oxidative, pathway, chronic, risk, sex, flux, radiation, non, bystander, adverse, communication, dohad, intracellular, incidence, involvement, metabolic, diet, generation, plasma, little, irradiated, effect, late, life, increase, moreover, postnatal, purpose
(14) Drug addiction	1.2%	Memory, ino, nuclear periphery, recruitment, activation, reactivation, bulk, fosb, detailed, repress, localization, recruit, addiction, yeast, drug, promote, transcriptional, reward, abuse, modern, set, gal, previous, require, conserve, promoter, nucleoplasm, periphery, regulating, underway
(15) Transposons and evolution	1.2%	Recombination, rate, transposon, organization, computational, subtelomeric, divergence, allopolyploid, create, meiotic, suz, family, add, subtelomere, evolve, progress, eukaryotic, fast, poorly, duplication, contain, respectively, newly, evolution, metabolism, functional, polyploidy, frequent, mal, disaccharide
(18) Neurons and cognitive disease	0.5%	Neuron, neuronal, ehmt, reversible, orchestrate, fly, cognition, behavioral, learning, body, requirement, dimethylation, predominantly, kleefstra, cognitive, writer, corrupt, intellectual, mushroom, associative, adulthood, larval, disability, classic, seq, dendrite, courtship, neuroscience, memory, mutant
(19) Meiosis and sex chromosomes	0.3%	Meiosis, sex, continue, spermatid, persist, spermiogenesis, postmeiotic, rabl, msci, xy, prophase, postmeiotically, autosome, configuration, thereafter, hand, clarify,

		preinactivated, inactivation, chromosome, spermataogenesis, spawn, comprehensive, similarity, abundance, meiotic, compartment, body, discover, mature
(20) Centromere structure	0.3%	Strategy, cenp, neocentromere, centromere, amount, mass, paucity, complementary, randomly, segregate, fold, assemble, similar, centromeric, action, position, enrichment, contain, mitosis, prevent, inheritance, quantitative, represent, stochastic, mechanistic, implication, nucleosome, number, sufficient, segregation

Appendix 3 Biomedicine. Topic model of 50 most cited abstracts per 5 year interval published between (A) 1990-2004 and (B) 2005-2019 generated by LDA using pyLDAvis visualization at λ 0.5.

A. 1990- 2004

Topic label	**% corpus**	**Word list**
(1) Epigenetics in cancer and regulation	43.7%	Methylation, cancer, gene, DNA, tumor, promoter, epigenetic, associate, protein, mutation, role, human, silence, control, specific, development, study, tissue, tumor suppressor, mechanism, change, lung, include, find, use, hypomethylation, risk, occur, metastasis, group
(2) Cellular mechanisms / progression of cancer	12.7%	Cell, adhesion, metase, stem, epithelium, serum, system, increase, adult, reduce, mouse, concentration, epithelial, progression, prolong, tumorigenicity, emerge, induce, effect, contact, fibroblast, potential, induction, characteristic, impair, treatment, breast, activity, single, phenotype
(3) Epigenetic mechanisms of cancer	8.2%	Arf, igf, imprint, deletion, patient, tumor, age, cdkn, wild type, esophageal, normal, loi, mucosa, hypomethylation, functional, loh, chromosome, glioblastoma, homozygous, mdm, loss, product, apply, etv, gene, wilm, completely, sense, often, heterozygosity
(4) Epigenetics in disease treatment and gene regulation	6%	Chromatin, modification, histone, understand, differentiation, drug, great, resistance, become, repress, agent, new, acetylation, maintain, state, advance, melphalan, activity, cytosine, diverse, genome, myogenic, therapy, promise, way, mammalian, regulate, causal, impact, toxicity
(5) Epigenetics and specific cancers	4.9%	Cadherin, prostate, mgmt, cpg island, methylguanine, almost, line, neck, head, deacetylation, thyroid, remove, hpec, replication, prb, transferase, glutathione, similar, light, pps, function, pathway, invasion, senescence, karyotype, show, peroxisome, inactivation, relation, associated

(6) Effects of nickel exposure on cancer in murine models	4.8%	Dose, week, rearrangement, liver, chain, nutrient, statistical, rat, elsevi, compound, carcinogenic, exposure, fashion, nickel compound, tcr beta, focus, thereby, epigenetically, tumor, hepatocellular, investigation, strain, hypomethylate, mouse, rodent, nickel, hepatocyte, reversion, unilateral, gavage
(7) Relationship between smoking and cancer	3.7%	Passage, hematopoietic, smoke, area, revertant, property, leukemia, competence, cd, stability, receive, concordance, cell, germline, acute, environmental, yield, follow, gamma, msp, neo, stimulation, strong, imply, deoxycytidine, myelogenous, later, mesenchymal, nucleotide, fragment
(8) Study of cancer	2.8%	Variant, review, begin, bha, assumption, immune, respect, cervical, sufficient, attempt, discussed, icc, sscp, emphasis, initially, survive, concentrate, adenocarcinoma, heterogeneity, generation, elimination, question, program, possess, cytokine, widespread, fade, differentiate, metastatic, cell
(9) Disease, regulation, and cell signaling in the intestines	2.7%	Escape, reactivate, acf, cytokine, member, block, diagnose, phase, crypt, spread, demethylation, family, activation, consist, concomitant, transcription, express, signal, metaplastic, inflammatory, probably, commonly, percentage, central, primarily, elucidate, proliferative, large, restimulation, instruction
(10) Intestinal cancer	2.5%	Catenin, regard, concept, immune, defect, newly, anchorage, regression, duodenal, ohe, protocol, proceed, elucidation, nontransgenic, transgenic, system, number, intestinal, end, spontaneous, promotion, regulate, transformation, mediation, unconjugate, pge, soft, agar, hydroxylation, ckmgmt
(11) Methodology in cancer studies	2.5%	Nonmalignant, line, reaction, mesothelioma, plasma, autoimmune, trial, score, design, breast, secrete, genetically, reflectance, factor viii, month, vector, testing, apparent, culture, biopsy, lung, endpoint, conclusion, method, adjuvant, independent, transfecte, prognostic, recipient, variability
(12) Nutritional/molecular deficiencies and disease	1.6%	Folate, normally, guanine, deficient, mp, supply, deficiency, part, donor, incorporate, mode, prevention, hypoxanthine, resistant, fragmentation, passive, mutagenic, medium, viability, salvage, pool, aa, cho, ham, aprt, presently, arrest, effective, uv, leukemia
(13) Cellular mechanisms	1.4%	Transporter, glutamate, prior, free, self, history, available, capacity, post, amino, lateral, translational, excitatory, persist, central, subline, passage, selection, agonist, carrier, affinity, alanine, variant, transform, spontaneous, confluence, transformation, individual, threonine, regional
(14) Cellular structure	1.1%	Toxic, intracellular, junction, specificity, restore, gap junction, organism, gap, metabolic, couple, evolve, differentiated, translational, matrix, multi, bring, dysfunction, molecule, extracellular, production, chemical, adaptive, toxicology,

		synchronization, dysfunctional, connexin, metazoan, speculate, maladaptive, reproductive
(15) Heat shock response in yeast	0.6%	Yeast, heat stress, currently, transient, mild, life span, divide, life, mortality, longevity, elucidation, ra, extension, lack, nonlethal, resumption, ignore, curtail, afford, manipulation, mitochondrial, thermotolerance, hsp, recovery, hinder, altogether, cerevisiae, petite, responsive, saccharomyce

B. 2005 – 2019.

Topic label	**Percentage of corpus**	**Word list**
(1) Cancer treatment	15.6%	Cancer, tumor, cell, therapeutic, drug, heterogeneity, target, immune, therapy, discuss, lncRNA, progenitor, diagnostic, advance, resistance, anticancer, review, interaction, response, different, focus, strategy, stroma, transcriptional, agent, phenotypic, understand, include, hallmark, genetic
(2) Epigenetics marks in cancer	14.7%	Methylation, DNA, histone, epigenetic, cancer, gene, process, nature, various, role, biomarker, modification, discovery, translational, hdac, progression, consider, acetylation, hdac inhibitor, initiation, code, normal, involve, comprehensive, approve, repair, occur, enzyme, certain, cellular
(3) General health / effects of early exposure to stress	12.3%	Health, disease, body, chronic, system, evidence, also, sex, disorder, life, risk, exposure, prostate, death, source, increased, age, cellular, activation, rights reserve, challenge, early, stress, high, organ, change, defective, excessive, dysfunction, effect
(4) Undefined	11.3%	Pd, checkpoint, individual, patient, identify, treatment, base, mutation, genetic, often, blockade, diagnosis, time, sample, genomic mutant, analysis, biopsy, characterize, RNA, gene, somatic, use, people, evolution, animal, mutate, screen, datum, sequence
(5) Apoptotic signaling pathway	10.1%	Signal, cell, pathway, anti, apoptosis, canonical, bcl, infection, notch, shift, receptor, antibody, activate, overall, program, correlate, efficacy, overcome, exhaustion, resistance, efflux, respond, regulate, unique, regulator, functional, control, beta, molecule, finding
(6) Immune system and regulation of gene expression	9.8%	Cell, myc, innate, expression, deficient, cd, cytokine, brain, adaptive, subset, transcription factor, differentiation, foxp, culture, allergy, nfatc, dependent, growth, drive, transcription, promote, element, treg, proliferation, response, regulation, selective, helper, lineage, differentiate
(7) Epithelial to mesenchymal transition in cancer	6.4%	Emt, pancreatic, histone lysine, mesenchymal, report, stromal, aacr, isoform, line, metastasis, specificity, epithelial, reflect, plasticity, lysine, exhibit, integral, orchestrate, start, acquisition,

		set, see, first, carcinoma, ras, oscc, amino, established, bias, neoplastic
(8) Undefined	3%	Subtype, idh, metastatic, article, elucidate, combine, depend, copy, bcl, structurally, demethylase, embryogenesis, exemplify, derangement, transducer, defect, future, apply, domain, tet, protein, action, interference, hypothesis, share, identify, clinical, reversible, yy, cohort
(9) Impact of diabetes	2.6%	Country, ontogeny, complication, transporter, vascular, endothelial, propose, mitochondrial, iugr, diabetes, proportion, inducible, determinant, xenobiotic, persist, remodeling, hba, hyperglycemia, pdx, antioxidant, increase, follow, primarily, spike, diabetic, keap, function, reduce, nf kappa, cause
(10) Obesity epidemic	2.1%	Obesity, food, diet, public, physical, epidemic, account, composition, pool, able, address, weight, accept, policy, barrier, cycle, relevant, especially, utero, counter, abdominal, burden, build, assess, half, mass, cdks, fat, effort, even
(11) Maternal effects / intrauterine exposure	1.9%	Metabolic, maternal, behavior, ascorbate, undernutrition, pharmacological, electron, adversity, significant, offspring, offspre, theme, transport, difficult, gut, analogue, broad, induction, decitabine, shape, need, rationale, predict, clinical, deregulate, neoantigen, hpa, foetal, prodrug, donor
(12) Cancer, regulation, and methodology	1.7%	Divergent, crc, alternative, overlap, place, characterized, tgf beta, reversion, integrity, knockout, state, motility, tumor, extent, clonal, intra, feature, btg, dac, strongly, morphological, restore, silence, existence, observe, drive, suppressor, induction, substantial, driver
(13) Immunology, research, and cancer	1.7%	Macrophage, tissue, month, condition, test, mgmt promoter, pregnancy, compelling, hypertension, chain, exogenous, favorable, radiotherapy, interval, confidence, percent, monocyte, mortality, concept, diversity, methylate, temozolomide, rank, log, define, physiological, activation, derive, fat, hazard
(14) Epigenetics, cancer, and telomeres	1.6%	Uncover, advanced, less, glioblastoma, viral, secondary, telomere, acetyl, capability, hat, diffuse, oppose, transferase, definitive, prominent, absent, histologically, lifetime, harbor, precursor, person, telomerase, bone, differ, primary, people, directly, similar, largely, instability
(15) Longitudinal cancer research	1.2%	Decade, outline, option, subgroup, recurrent, origin, dismal, past, discussion, possibly, view, lastly, formation, marker, manner, progress, gbm, rise, direction, major, kill, prognosis, central, third, inherent, derive, research, mutation, drug, brain
(16) Leukemia and its treatment	1.2%	Mll, experience, deplete, leukemia, mgmt, haematopoietic, ongoing, regimen, alkylating agent, translocation, fusion, schedule, mixed, leukaemogenic, hox, lose, positively, poor, pseudosubstrate, chemoresistance, standard, nontoxic, dense, inevitably, note, mgmt promoter, useful, stem, glioma, unclear

(17) Dysregulation of hormones in cancer	0.7%	Aberrantly, dysregulate, androgen, ar, small molecule, neoantigen, antitumor, promising, repressive, et, durable, regression, erg, tmprss, management, prostate, recruitment, rapid, follow, disrupt, potentiating, fuse, rearrangement, massively, regulated, sequencing, iv, immunology, remission, widen

Appendix 3. Proximal Biology. Topic model of 50 most cited abstracts per 5 year interval published between 1990-2004 generated by LDA using pyLDAvis visualization at λ 0.5.

A. 1990- 2004

Topic #, label	**% corpus**	**Word list at λ 0.5 (pyLDAvis)**
(1) The epigenetic regulation of gene expression and genomic structure	31.1%	Chromatin, histone, methylation, gene, epigenetic, modification, bind, domain, site, state, protein, structure, complex, mechanism. DNA, heterochromatin, centromere, regulation, role, lys, activity, mecp, lysine, understand, specific, enzyme, nucleosome, provide, chromosome, transcriptional
(2) The influence of epigenetics on disease	17.6%	Cancer, tumor, loss, cell, gene, methylation, colorectal, associate, tumor suppressor, progression, aberrant, tumorigenesis, cellular, lung, cpg island, DNA, promoter, tumour, hypermethylate, beta, art, technology, repair, inactivation, inactivate, alteration, lesion, event, high, mgmt
(3) The epigenetic reprogramming of cells in the germline / development	12.6%	Cell, demethylation, reprogramme, germ, cd, imprint, stage, development, zygote, fertilization, insight, germline, occur, genomic, thymocyte, genome, first, stem, process, epigenetic, produce, mouse, pgcs, gene, normally, differentiate, runx, hepatocellular, hepatocyte, number
(4) The use of epigenetics in treating disease	10.3%	Suv, small RNA, sir, therapy, identifiy, target, protein, RNA, component, key, family, molecule, serve, specification, multiple, sequence, local, motif, promote, correspond, modifier, fate, predict, peptide, transient, significance, currently, possible, dependent, factor
(5) The heritability of epigenetic marks / parental effects / genomic analysis	9.7%	Igf, imprint, hypomethylation, sequence, fetal, express, paternal, ddm, transcript, maternal, tissue, environment, mutant, parental, methylation, allele, paternally, deletion, DNA, analyze, parent, region, observe, short, ref, repeat, self, progeny, snrpn, chromosome
(6) The relationship between epigenetic regulation and development	5.7%	Polycomb, history, respectively, group, embryo, synthesis, deposition, insulator, reporter, main, hold, exposure, great, procedure, homeotic, possibly, retinoblastoma, egg, extinction, maintenance, neural, design, endometrial, mlh, pcg, fifth, interesting, accelerate, occasionally, constitutive

(7) Inheritance of disorders	4.6%	Schizophrenia, twin, breast, prb, disorder, brca, specimen, effect, rtt, etiology, tnpa, hmr, contribution, heritability, selectively, benign, basic, autonomous, estimate, wild type, transgene, spm, chimera, regression, maternal, age, dz, discordant, probandwise, noninherite
(8) Cancer research using lab animals / mechanisms of blood clotting	3.3%	Cadmium, injection, human, testicular, rat, pai, dose, rate, rodent, prostate, occupational, metal, right, poorly, sa, cenp, plasma, factor viii, retroviral, induce, animal, carcinogenesis, secrete, reserve, mepai, hml, trx, satellite, effectively
(9) Epigenetics and cancer / bacteria	3%	Sirtuin, sirt, class, understanding, research, oncogene, area, oncogenesis, see, bacteria, protozoan, decitabine, phylogenetic, array, press, academic, prokaryotic, bacterial, prevention, extra, weak, sequential, comprehensive, year, effective, eukaryote, influence, elsevi, wide, epithelial
(10) Maternal effects and the agouti mouse	1.3%	Yellow, agouti, supplementation, nutrition, dietary, viable, donor, especially, in, deleterious, labile, ala, choline, beneficial, betaine, metastability, dam, lability, folic, metastable, presume, harbor, unintended, vitamin, offspring, cluster, mash, range, sibling, pseudoagouti
(11) Morphology	0.4%	Skull, individualized, ontogeny, decomposition, spline, facial, cranial, integration, cotton, complicate, familiar, heterotopy, heterochrony, spatiotemporal, integrated, weaning, geometric, sigmodon, adequately, precocial, thin, count, warp, priori, fulviventer, arbitrary, chew, whole, demanding, growth
(12) Plant Genetics	0.3%	Chs, sense, transgenote, cosuppression, antisense, synthase, readthrough, repetitiveness, chalcone, score, flower, phenotypically. pigmentation, construct, produce. transgene, cotton, complicate, familiar, heterotopy, heterochrony, spatiotemporal, integrated, weaning, geometric, sigmodon, adequately, precocial, thin, count

-

B: Proximal Biology (2005 -2019)

Topic label	**% corpus**	**Word list**
(1) Prospects/promise for biomedical research (1)	22.3%	Epigenetic, cellular, process, histone, role, cancer, mechanism, review, include, recent, drug, discover, biomarker, modification, biological, response, protein, many, therapy, involve, study, disease, year, influence, field, acetylation, target, discovery, question, regulation, health, future.
(2) Epigenetic regulation of gene expression	16.1%	Methylation, DNA, sequence, promoter, gene, level, inflammatory, genome, use, region, previously, context, critical, sex, cpgs, differentially, methylate, cpg, observe, activation, element, silence, loop, rapidly, probe, array, expression, interfere, potentially, resolution
(3) Epigenetics and cancer	12.7%	Cancer, cell, inhibitor, tumor, reprogramme, early, progression, selective, low, epigenetic, reveal, enhancer, progenitor, gene,

		mouse, part, abnormal, find, tissue, establish, stem, regulator, demethylation, bet, identification, germline, suggest, however, plasticity, development
(4) Clinical methods	11.8%	Molecular, clinical, pathway, prostate, demonstrate, tet, hematopoietic, method, alpha, include, decade, specific, activity, well, vivo, edit, study, develop, available, analysis, intervention, general, production, possible, apply, ability, investigation, give, induce, discuss
(5) Examination of maternal effects	6.9%	Maternal, treat, site, domain, newborn, mark, methylation, pattern, examine, cohort, status, methyl, division, success, specifically, past, bivalent, functionally, assess, discuss, specific, dinucleotide, mood, statistical, wide, presence, region, late, association, highly
(6) Immune system, ageing, and mortality	4.9%	Age, immune, chronological, factor, blood, individual, cd, estimate, life, rate, measure, foxp, death, lifespan, observation, enhance, treg, risk, concentration, naïve, hat, compelling, build, lead, cell, acceleration, response, contribute, suppression, predict
(7) Effect of ncRNAs on hematopoietic cells and genomic structure	3.9%	LncRNA, hsc, type, editing, independent, open, access, deep, nuclease, composition, lineage, correlate, location, memory, accessible, monocyte, hold, code, package, noncoding RNA, enable, local, readily, remodel, accomplish, distinguish, least, regulatory, generate, mitosis
(8) Noncoding RNA and effects on cancer and genome structure	3.8%	RNA, tumour, small, compartment, structure, dysregulation, uncover transcriptome, mRNA, subgroup, present, long, demethylase, organization, encode, scaffold, complexe, relevance, newly, defence, mode, disrupt, systematic, lesion, lncRNA, resource, polymerase, gbm, extensive, event
(9) Methodological reflection	2.6%	Application, tool, nature, set, researcher, query, subtype, mutate, datum, scna, link, interface, clinician, biomedical, dmrs, meeting, proteomic, potency, summary, short, guide, recurrent, explore, move, comparative, region, program, provide, interest, practical
(10) Reprogramming of somatic cells to induced pluripotent stem cells	2.5%	Stem, pluripotent, cell, es, differentiate, fibroblast, sox, rely, somatic, klf, adult, embryonic, disorder, ip, human, capable, ipsc, ips, state, surface, germline, origin, myc, generate, self, fold, transmission, induce, perpetual, lock
(11) Noncding RNAs and their function	2.5%	Diverse, accessibility, family, modulate, orchestrate, hox, recruit, strong, incRNA, guiding, allosteric, rapid, almost, interaction, substrate, diversity, regulatory, simultaneously, modifying, directly, landscape dpendent, analyse, ncRNA, conserve, hoxd, hotair, kilobase, remove, preimplantation
(12) Technologies	2.1%	Laboratory, next generation, testing, perform, variant, emphasis, consideration, example, sequence, chip, computational, describe, genetic, technology, technical, seq, experience, cardiovascular, pharmaceutical, canonical,

		mitochondrial, species, hallmark, datum, challenge, condition, comprise, pot, increase, therefore
(13) Noncoding RNAs and their role in disease and regulation	1.7%	MiRNA, dysregulate, oncogene, deletion, suppressor, ink, amplification, control, arf, biogenesis, degradation, interact, precursor, clear, turn, small, cnr, act, microRNA, become, fruit, epimutation, ripen, colorless, processing, cleavage, siRNA, damage, gene, guide
(14) Epigenetics and cancer	1.7%	Emt, transition, modifier, course, epithelial, mesenchymal, brd, diagnostic, representative, point, widespread, purine, modality, phenotypic, modulator, batch, link, peptide, address, recognize, inflammation, design, state, inactive, often, maintenance, plasticity, shift, inactivation, induction
(15) Cancer	1.1%	Heterogeneity, substantial, divergent, clonal, tumor, diagnosis, attribute, primary, source, receptor, metastatic, intra, morphological, right, reserve, elsevi, yet, experimental, clinical, extent, angiogenic, rare, startling, outgrowth, exception, furthermore, feature, time, clone, originate
(16) Epigenetic regulation in plants	0.9%	Plant, contact, dicer, lay, argonaute, biogenesis, territory, domain, small, chromosomal, silence, overlap, imprint, clustering, linearly, modeling, remote, realization, detailed, confine, drosophila, partition, systematically, conformation, hierarchically, extensively, unanticipated, draw, deepen, rdr
(17) Specification of germ layers	0.8%	Specification, robust, epiblast, pgc, competent, fungi, surround, easy, locate, appropriate, pgclcs, epilcs, esc, transmission, entity, pregastrulating, meticulously, episc, spermatogenic, ssea, reconstitution, gamete, properly, isolation, pgclc, multisteppe, surprise, irreversibly, integral, appropriately
(18) Ageing	0.5%	Annotation, prediction, biologically, accurately, predictor, healthy, number, data, set, clock, fourth, passage, freely, informative, meaningful, chimpanzee, cumulative, inversely, ontogenetically, applicable, surrogate, attributable, variance, core, multi, fact, greatly, predominant, third, year, site

Appendix 4. Evolutionary biology. Topic model of 50 most cited abstracts per 5 year interval published between (A) 1990-2004 and (B) 2005 – 2019 generated by LDA using pyLDAvis visualization at λ 0.5.

A. 1990- 2004

Topic label	% corpus	Word list
(1) Epigenetic mechanisms in evolution / Waddington	14.1%	Gene, level, sex, imprint, duplicate, chromosome, evolutionary, age, constraint, differential, extinction, flower, sexuality, developmental, silence, specific, mapping, duplication, conserve, polycomb, reduction, canalize, bnflc, removal, newly, study, canalization, mammal, methylation, expression
(2) Epigenetic inheritance as a complementary evolutionary system	12.3%	Cell, redundancy, memory, epigenetic, inheritance, DNA, system, state, clone, regulatory, particular, concept, heredity, architecture, gap, plasticity, genetic assimilation, mechanism, discuss, evolution, redundant, function, role, paper, transmit, gene, mutation, propose, fidelity, cytosine
(3) Variation and stability of morphological structures	12.2%	Lateral, epigenetic, insertion, femoral, stability, mal, meniscus, expect, joint, genetic, hominid, pattern, determine, maleness, carrier, phase, epiphysis, select, variation, covariance, interaction, expression, mutation, developmental, mitotic, offspring, development, tibial, lip, distal
(4) Effects of methylation on gene expression in plants	10.5%	Genome, gene, fot, arabidopsis, code, sequence, allopolyploid, polyploid, wheat, non, methylation, site, alteration, change, amplifiy, associate, novel, esterase, element, microarray, turn, duplication, pair, brassica, section, translocation, insecticide, resistant, chromosome, specie
(5) Traits that exhibit adaptive plasticity	9%	Tortoise, primitive, digit, phalangeal, hominin, afarensis, horn, head, forage, reconstruct, plastic, hypothesis, anatomy, behavior, character, size, trait, usually, adaptation, selection, datum, discrete, behaviour, ask, interested, null, individual, fish, specie, phenotypically
(6) Development of cranial morphology / Waddington	8.5%	Effect, shift, skull, integration, genotypic, maternal, approach, environmental, growth, observe, evolve, cranial, facial, individualized, developmental, factor, integrate, craniofacial, base, landscape, age, population, information, system, epigenetic, modern, dynamic, diverse, face, ontogeny
(7) Biological anthropology and sexual dimorphism	7.3%	Sexual dimorphism, sample, uterine, skeletal, canine, zygomatic, adenohypophysis, prehistoric, biodistance, australian, tayassuid, nhp, hypothalmic, population, size, process, relationship, heritable, extend, transfer, large, pattern, late, correlation, adult, significant, sex, mortuary, coastal, ad
(8) Evolvability of phenotypes / Waddington	5.6%	Aphid, asexual, parthenogenesis, trigger, morphological, evolvability, particulate, body, trade, lineage, paradigm, adaptive, postcranial, upper, evolution, overall, appearance, prion, read, translation, eventually, temporally, cycle, ecological, list, threshold, dependency, inform, confirmatory, archaic
(9) Retroviruses in plants	4.3%	Codon, eprvs, cpg, plant, cpt, cpa, epiallele, dinucleotide, contain, tabacum, substitution, confound, mammal, term, bat,

		epiallelic, tomentosiformis, nicotiana, sylvestris, eprv, affect, family, elevate, natural, rate, virus, fitness, impact, differ, represent
(10) Fluctuating asymmetry	4.1%	Neandertal, predator, marginal, fluctuating asymmetry, stress, symmetric, degree, development, bilaterally, good, side, trait, large, symbiose, ossification, exaggerated, prefer, otherwise, unable, cue, mate, area, concern, secondary, cranial, embryo, cranium, interaction, character, argument
(11) Parental effects	3.5%	Eg cell, copulation, line, germ, pef, latency, duration, allele, cell, somatic, resemble, chimera, ds, performance, fa, statistically, bristle, heritability, pfa, significant, embryonic, female, similar, hence, parental, igf, methylation, male, spartina, burst
(12) Sexual display	3.1%	Ornament, cortical, paternal, map, global, male, mite, sexual, ascribe, extravagant, secondary, elimination, demonstrate, preference, choice, flat, female, absolute, structure, relationship, size, cope, ocular, net, cortex, macaque, concentrate, approximate, anisotropic, uniform
(13) Transposons	1.8%	Selfish, repetitive, rie, narrative, functionality, DNA, component, remain, integrally, connection, couple, suppose, criticism, critique, untestable, overview, enigmatic, brief, specificity, elucidating, prokaryotic, monomer, story, selfishness, scheme, purpose, nearly, macro, substantial, replicator
(15) Parasites	1.1%	Parasitoid, virulence, host, resistance, defence, melanogaster, insect, parasitize, coevolutionary, dear, appraise, trail, coevolution, community, parasitism, elsewhere, hymenopterous, southern, survive, immunological, internal, mount, positively, critically, evidence, unlikely, explanation, central, attack, low

B. 2005 and after

Topic#, label	**% corpus**	**Word list**
(1) Gene expression and evolution	17.5%	Evolutionary, evolution, molecular, gene, provide, insect, question, methylate, include, process, genomic, new, te, novel, present, genome, future, require, facilitate, consider, gene expression, study, lack, together, diversity, major, toad, biology, rapid, advance
(2) Phenotypic plasticity / epigenetic switch	12.2%	Cue, plasticity, divergence, developmental, temperature, natural, integration, underline, whole, phenotype, development, challenge, source, scale, specie, match, adult, shape, respond, long, study, highlight, increase, different, great, trait, salt, divergent, phenotypic, differ
(3) Environmental influences on	12%	Methylation, DNA, climate, variation, epigenetic, age, significant, plant, response, find, occur, profile, site, estimate,

methylation / phenotypic responses		population, animal, role, association, cytosine, negative, strong, local, heritability, method, low, exhibit, treatment, level, datum, obtain
(4) Adaptive phenotypic plasticity	12%	Effect, environment, mutation, phenotypic, global, change, genetic, epigenetic, promoter, environmental, vertebrate, invertebrate, adaptation, life, plasticity, phenotype, result, population, selection, natural selection, model, generation, fitness, transmission, variation, influence, value, variant, contrast, condition
(5) Epigenetic effects on primate sociality	8.8%	Social, degree, generation, course, expression, relate, baboon, skeletal, control, mismatch, involve, differential, early, theory, human, outcome, other, measure, imprint, consistency, induce, day, link, behavioral, pattern, presence, inherit, line, growth, specify
(6) Relative advantages of sexual vs asexual modes of reproduction	6.1%	Habitat, asexual, reproductive, differentiation, apomictic, aflp, epigenetic, variation, reset, gradients, beneficial, length, directly, stage, genetically, lineage, interpret, genetic, slow, population, cycle, shift, correlation, end, variance, sexual, advantage, qualitatively, involvement, complete
(7) Parental effects	3.8%	Care, maternal, offspring, pass, hpa, trade, off, glucocorticoid, distribute, mother, paternal, follow, step, receptor, survival, last, offspre, modification, female, pituitary, hypothalamic, deprivation, biosynthesis, immune, isolate, elevate, transfer, information, modulate, short
(8) Chromosome features / Waddingtonian development	3.7%	Chromosome, environmentally, telomere, constraint, revolve, undergo, sex, locus, genome, especially, foster, newly, diverse, neocortex, host, assume, convergent, status, favor, stability, nonetheless, determine, region, canalization, additional, sessile, operate, parasitic, progress, compatible
(9) Chromosome features	3.5%	Centromere, satellite, within, predict, integrate, acession, non, recently, exploratory, order, select, DNA, humpback, centromeric, percentage, form, transcript, repeat, correlation, fit, site, cento, ant, part, chromatin, unit, muscle, bee, ccgg, inner
(10) Methylation, plants, and transposons	3.3%	Flower, amount, plant, demethylation, reduce, genotype, inheritance, quality, ecologically, inbreede, marker, dispersal, powerful, unknown, redundancy, nucleotide, arabidopsis, ltr, depression, idea, difference, experimental, rt, nectar, relatedness, noise, generality, heterogeneous, manipulate, drift
(11) Biological anthropology and diet	3.1%	Exposure, sensitivity, experimental, epigenome, laboratory, health, span, temporal, microwear, application, circumstance, pr afarensis, increasingly, scenario, food, concern, investigation, permanent, detailed, carry, explanatory, toxicant, consume, item, paleoecological, appear, robust, sufficient, susceptibility, limited