A General Theory of Differentiated Multicellularity

Felipe A. Veloso[1]✉

[1]Faculty of Biological Sciences, Universidad Andrés Bello, Santiago, Chile

✉Correspondence: veloso.felipe.a@gmail.com

Abstract

It is widely accepted that changes in the levels of gene expression are important for the cell differentiation process. Furthermore, research in the field has customarily assumed that such changes regulate this process when they interconnect in space and time by means of complex epigenetic mechanisms. In fundamental terms, however, this assumed regulation refers only to the intricate propagation of changes in gene expression or else leads to logical inconsistencies. To tackle this problem, I analyzed publicly available high-throughput data of histone H3 post-translational modifications and mRNA abundance for different *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster* cell samples. An analysis of genomic regions adjacent to transcription start sites generated for each cell dataset a profile from pairwise partial correlations between histone modifications controlling for the respective mRNA levels. Here I report that these profiles, while explicitly uncorrelated to transcript abundance by construction, associate strongly with cell differentiation states. This association is not expected if cell differentiation is, in effect, regulated by epigenetic mechanisms. Based on these results, I propose a theory that relies on the synergistic coupling across the extracellular space of two mutually uncorrelated constraint-generating systems involving histone H3 modification states simultaneously. This theory describes how the differentiated multicellular organism—understood as an intrinsic, higher-order, self-sufficient, self-repairing, self-replicating, and self-regulating dynamical constraint—emerges from proliferating undifferentiated cells. If resists falsification, this theory will explain (i) the intrinsic regulation of gene transcriptional changes during cell differentiation and (ii) the emergence of differentiated multicellular lineages throughout evolution.

Introduction

The X-files of chromatin

Cell differentiation, if seen as a motion picture in fast-forward, intuitively appears to be a teleological process, its *telos*[1] being the multicellular organism in its mature form. The first step for a scientific explanation of this apparent property was given in 1957 when Conrad Waddington proposed his epigenetic landscape model. Influenced by earlier developments in dynamical systems theory [1], Waddington's model showed cell differentiation to be potentially predictable or at least potentially explainable without any teleological reference [2].

In terms of explainability, the dynamics of the cell differentiation process have been associated to changes in chromatin states and concurrent heritable changes in gene expression that are explicitly uncorrelated to changes in the DNA sequence (for this reason defined as epigenetic changes [3, 4]). In some cases these changes can be regulated extrinsically with respect to the developing organism, as observable in eusocial insects (e.g. a female honeybee larva develops into a worker or a queen depending on the royal jelly diet it is fed [5]). Nevertheless, most key changes in gene expression during cell differentiation are not only independent from, but are even robust with respect to extrinsic variables. This means cell differentiation is fundamentally an intrinsically regulated process, for which no falsifiable theory has emerged from the epigenetic framework since it was first advanced. Moreover, Peter Fraser has recently referred to this problem as "The X-files of chromatin" [6].

This research work was conceived and designed to, following Fraser's metaphor, declassify "The X-files of chromatin". In its initial phase, I conducted a computational analysis of the least relevant—for the epigenetic landscape—constraints on histone H3 post-translational modification states. Before outlining this analysis however, I must present here a case for the fundamental impossibility of explaining the cell differentiation self-regulatory dynamics under the framework pioneered by Waddington, however complex its underlying mechanisms may be (as also hinted by Fraser [6]). Only then will I be able to argue that these epigenetically irrelevant constraints on histone modification states are, in fact, key to a full understanding of differentiated multicellularity in terms of its self-regulation and evolution.

The conundrum of self-regulation

Avoiding non-explanatory teleological descriptions, modern science regards cell differentiation fundamentally as a dynamical system, where a fixed rule governs the transition between the realizable states of a complex network of molecular mechanisms. Ranging from low-order molecular interactions [7] to chromatin higher-order structural changes [8, 9], these epigenetic mechanisms not only propagate changes in gene expression in different loci as cells proliferate but, importantly, would also *regulate* intrinsically the cell differentiation process. Furthermore, and although the epigenetic mechanisms involved in cell differentiation are far from being

[1] τέλος is the Greek for "end", "goal", or "purpose".

completely elucidated, the hypothesis that cell differentiation is *regulated* by heritable changes in gene expression is routinely presented to the general public as a well-established scientific fact (as illustrated in [10]). However, this hypothesis—whether or not we accept it in its strict sense—leads to severe explanatory limitations and may even entail logical inconsistencies.

To assume the aforementioned hypothesis is true in its strict sense is to accept gene self-regulation as a scientifically tenable and explainable teleological property of cell differentiation (the "intuitive" *telos* here would be certain future transcriptional states to be timely achieved or maintained). To explore what this implies let us suppose, for simplicity without loss of generality, that a researcher modifies the expression levels of certain *gene A* in certain organism and then elucidates how those changes, during differentiation, activate or repress *gene B*, *gene C*, and *gene D*. At this point, we might regard the finding as evidence that *gene B*, *gene C*, and *gene D* are regulated by *gene A*. Consequently, we could also hold that *gene A* is an explanatory component of the general self-regulatory property. However, these assertions overlook that the researcher, not *gene A*, was the true regulator by purposefully imposing certain transcriptional states (on *gene A*, and by means of *gene A*, also *gene B*, *gene C*, and *gene D*). Yet, no human regulator is needed during the natural process, which raises the question of what is the system truly regulating *gene A*, *gene B*, *gene C*, *gene D*, and by extension, all genes during cell differentiation.

Moreover, explaining the regulation of transcriptional states in a gene locus by previous transcriptional states in other gene loci (in the same cell or any other) is only a non-explanatory regress. It takes the question about regulation, i.e. explaining a gene being at certain transcriptional states (and, importantly, at no other transcriptional states), to some other gene or genes, back in time. This regress inexorably leads—even in the simplest scenario—to the unexplained, timely regulation of one key gene (or more key genes, simultaneously) within undifferentiated cells.

On the other hand, to take the epigenetic-changes-regulate hypothesis in a loose sense is to use "self-regulation" only as a placeholder when referring to a certain class of molecular mechanisms *propagating* changes in gene expression. In this context we must note that a defining condition of any mechanism is that the dynamics it comprises are *explicitly correlated*. Thus, an epigenetic mechanism can be seen metaphorically as toppling dominoes (here the explicitly correlated dynamics are obvious). But as pointed out previously this mechanism, however numerous or intricately connected its correlated dynamics, says nothing about how the first domino tile (or any other whose fall is not attributable to the fall of other tiles) was toppled over. To fill this explanatory gap, it has been proposed that an "epigenator"—defined operationally as a transient signal which probably originates in the environment of the cell—triggers the epigenetic phenotype change after being transduced into the intracellular space [11]. Nonetheless, if all "epigenators" in the developing organism are extrinsic to it, self-regulation is *ipso facto* unexplainable. Otherwise if there is at least one intrinsic "epigenator" (e.g. a suggested "extracellular signal") its critical signaling property is left unexplained.

Importantly, these problems are inherent to *any* model based on Waddington's epigenetic landscape. This is because *the dynamics of any mechanism regulating changes in gene expression must be explicitly uncorrelated to such changes; otherwise this mechanism is fundamentally just an "additional set of arranged domino tiles" propagating gene expression changes more extensively instead of regulating them* (see **Figure 1A**). At this point the explanatory dead end becomes evident. Under the traditional approach in developmental biology no higher-order system within a living organism, however complex (e.g. displaying interlocked feedback loops or hypercyclic networks), can exert true intrinsic regulation because its dynamics are ultimately correlated to the lower-order dynamics it is supposed to regulate. Thus, the supposed *regulatory* power of changes in gene expression on the cell differentiation process is causally inefficacious in the most fundamental sense.

Progress comes from recognizing that the propagation of critical changes within a developing organism and the intrinsic regulation of such changes are entirely different processes. Specifically, intrinsic regulation is not a molecular mechanism—however complex—correlating the levels of critical variables within a developing organism but instead *constraints* on the realizable levels of said variables. Importantly, these constraints must be also *explicitly uncorrelated* to the regulated levels as argued previously.

Epigenetic information in theory and practice

Regardless of the explanatory limitations inherent to the epigenetic landscape, either all necessary information for the intrinsic regulation of cell differentiation is already contained in the zygote (or other primordial cell such as the spore) or it is not. This dichotomy may seem to be trivial but important implications follow it.

If the zygote (henceforth referring also to other primordial cells found in differentiated multicellular organisms) contains all necessary information for intrinsic regulation [12, 13], the previously discussed explanatory gap could, in principle, be filled. Asymmetric early cleavage, shown able to resolve a few cell lineage commitments (into six founder cells) in the nematode *Caenorhabditis elegans* [14], supports this possibility at first glance. Nevertheless, a closer look at the developmental complexity of this simple metazoan model organism suggests otherwise: the hermaphrodite *C. elegans* ontogeny yields 19 different cell types (excluding the germ line) in a total of 1,090 generated cells. From these two parameters alone, the required information capacity for the entire process can be estimated to be at least 983 bit (see details in the Appendix). Further, this is a great underestimation because uncertainty remains with respect to at least two more variables at least, namely space and time. Therefore, the non-genetic information capacity necessary for the entire process far exceeds the few bits of information that epigenetic mechanisms (even if asymmetric early cleavage is entirely explained by them) can account for. On the other hand, the explanatory power of extrinsic constraints (e.g. diet-dependent hierarchy determination in eusocial insects [5], temperature-dependent sex determination in reptiles [15], or maternal regulation of offspring development [16]) clearly does not account for all developmental decisions and in some cases it is not even necessary. These considerations highlight the remarkable explanatory power of certain *intrinsic* constraints—to be identified here—on developmental decisions in terms of information capacity.

Information not only requires a medium with capacity for its storage and transmission but also must have content, which resolves developmental decisions as cells proliferate. Here an additional problem appears: cell potency. An entire organism can develop (including extraembryonic tissues) from *any* totipotent cell, and all embryonic tissues can develop from *any* pluripotent stem cell. How is this possible if the information for all cell fate decisions is already contained in the zygote? The recently proposed—yet not explanatory—"epigenetic disc" model for cell differentiation, under which the pluripotent state is only one among many metastable and directly interconvertible states [17], reflects the necessity to account for the dependence of developmental information on cellular context.

With remarkable insight, David L. Nanney anticipated in 1958 explanatory pitfalls if the definition of epigenetics is limited to heritable changes. He further stated that "'cellular memory' is not an absolute attribute" [18]; or, in other words, that more important to development is the process by which heritable material may manifest different phenotypes than the heritable material itself. However, Waddington's epigenetic landscape prevailed and the field reinforced a "preinformationist" framework: although the zygote is not a complete miniature version of the mature organism (preformationism), it is indeed a complete blueprint of the mature organism (allowing for some degree of extrinsic influence as described previously and for stochasticity [19, 20]). If this is correct, we must also accept that in the mature human brain there is strictly less—since it is one among many outcomes of the developmental process—non-genetic and non-redundant information than in the human "developmental blueprint" (not surprisingly however, I failed to find a single research paper with such a proposition).

This *reductio ad absurdum* shows that the epigenetic landscape framework has forced research to ignore or reject the necessary *emergence* of not only some, but possibly most information content during cell differentiation. Specifically, if additional information content emerges during brain development, what would necessarily preclude information content from emerging in proliferating undifferentiated cells?

A proof-of-principle hypothesis

In the previous two subsections I argued that (i) explaining the self-regulatory dynamics of cell differentiation under the traditional epigenetic landscape approach is a fundamental impossibility, (ii) the constraints regulating the critical variables for cell differentiation must be explicitly uncorrelated to such variables, and (iii) any theory aiming to explain differentiated multicellularity must account for emergent developmental information, which is not structurally but dynamically embodied (that is, dependent on the extracellular context). Consequently, in this work I designed a computational analysis to search for constraints as defined in (ii) because their existence is, ultimately, the proof of principle for the theory referred to in (iii).

The specific objects of study were the combinatorial constraints on histone H3 post-translational modifications—also known simply as histone H3 crosstalk—because of their strong statistical relationship with transcriptional levels [21]. Notably, several high-throughput studies have underscored already the relevance of histone crosstalk by identifying highly significant pairwise relationships between post-translational modifications [22, 23, 24, 25].

Under these considerations, I defined the working hypothesis as follows: *for any given cell differentiation state and within genomic regions adjacent to transcription start sites, the component of pairwise histone H3 crosstalk that is explicitly uncorrelated to transcriptional levels associates with that differentiation state* (**Figure 1B, black dashed arrow**). Importantly, the null hypothesis (that is, no significant relationship exists between cell differentiation states and histone H3 crosstalk uncorrelated to mRNA levels) is further supported by the epigenetic landscape approach: if changes in mRNA levels not only associate with cell differentiation states [26, 27, 28] but also explain them completely, an additional non-epigenetic yet differentiation-associated level of constraints on histone H3 crosstalk is superfluous.

For the computational analysis I used publicly available tandem datasets of ChIP-seq (chromatin immunoprecipitation followed by high-throughput sequencing) on histone H3 modifications and RNA-seq (transcriptome high-throughput sequencing) on mRNA for *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster* (see Materials and Methods). The basis of the analysis was to define a numeric profile *ctalk_non_epi*, which represents the strength and sign of pairwise partial correlations between histone H3 modification states controlling for mRNA levels within genomic regions adjacent to RefSeq transcription start sites. In other words, *ctalk_non_epi* profiles represent the non-epigenetic component of pairwise histone H3 crosstalk (see decomposition as a sum of two covariances in **Figure 1B**) in genomic regions where the epigenetic component is significant.

The hypothesis testing rationale was to apply hierarchical clustering on the *ctalk_non_epi* profiles for different cell datasets in all three organisms, using nonparametric bootstrap resampling to assess cluster significance [29].

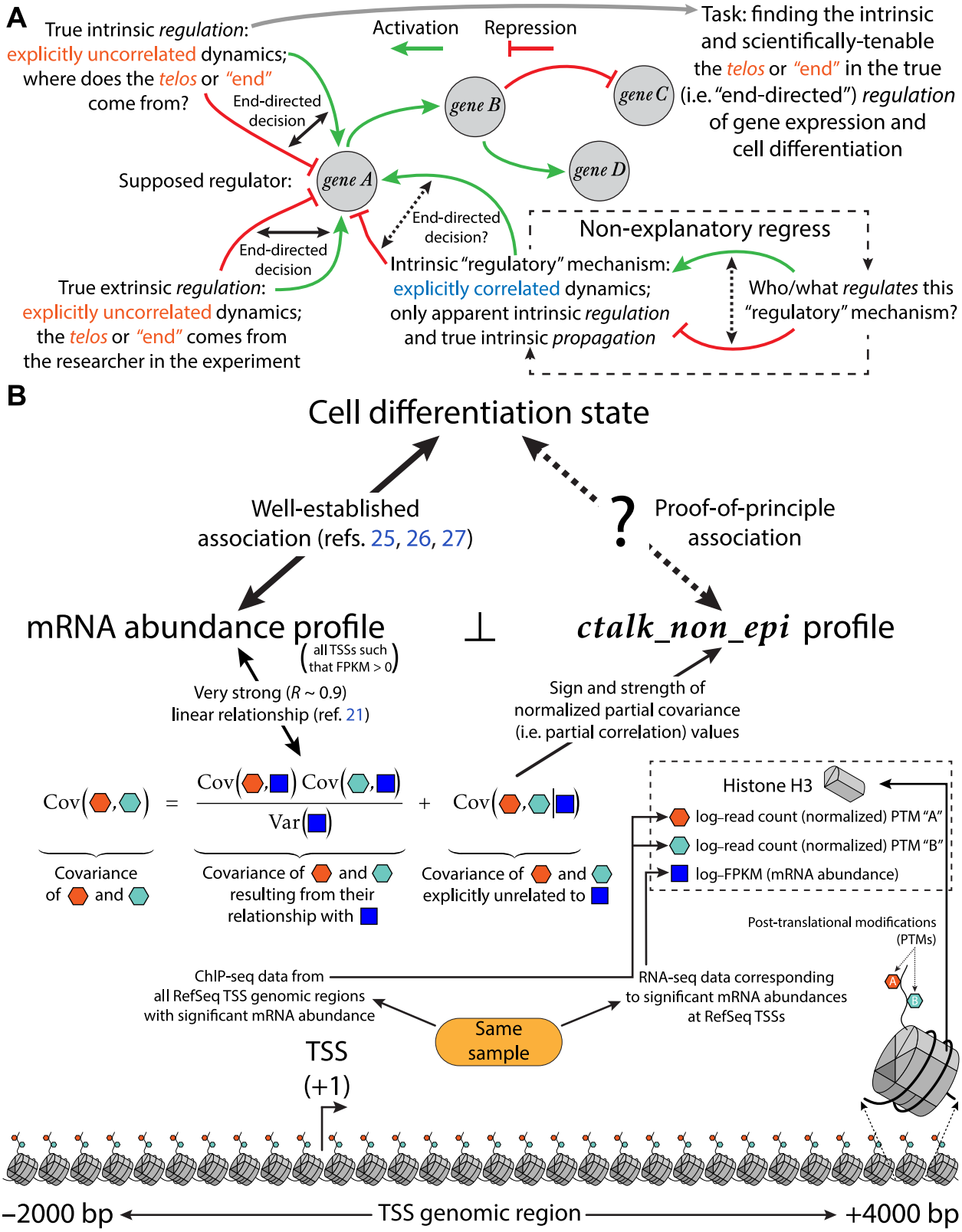

Figure 1: (A) Explanatory limitation of the epigenetic landscape approach in terms of the intrinsic regulation of gene expression. **(B)** Scheme of the proof-of-principle hypothesis described in the introduction and the computational analysis conducted for its testing (see details in Materials and Methods).

Results

If there is no significant association between *ctalk_non_epi* profiles and cell differentiation states (i.e. if the null hypothesis is true), after applying hierarchical clustering on those profiles the obtained clusters should be statistically insignificant, or else they should not associate with cell differentiation states. However, the results showed otherwise: in all analyses performed, *ctalk_non_epi* profiles fell into statistically significant clusters that associate with cell differentiation states in *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster*. Moreover, *ctalk_non_epi* profiles associated with cell differentiation states at least as clearly as mRNA abundance[2] profiles (as mentioned earlier, the relationship between transcriptional and cell differentiation states is known and well-established [26, 27, 28]). In sum, for all three organisms analyzed, the null hypothesis had to be consistently rejected in terms of a clear association between *ctalk_non_epi* profiles and cell differentiation states, thus providing proof of principle for the theory to be presented in this monograph.

ctalk_non_epi profiles of embryonic stem cells differ significantly from those of differentiated cell types in *Homo sapiens*

Using data for nine different histone H3 modifications (for details see Materials and Methods), I computed *ctalk_non_epi* profiles for six human cell types. From these, all profiles corresponding to differentiated cell types, namely HSMM (skeletal muscle myoblasts), HUVEC (umbilical vein endothelial cells), NHEK (epidermal keratinocytes), GM12878 (B-lymphoblastoids), and NHLF (lung fibroblasts) fell into the largest cluster. Such significance was expressed in the obtained *au* (approximately unbiased) and *bp* (bootstrap probability) significance scores, which were greater or equal than 95 (**Figure 2A, cluster #4**), indicating that the largest cluster was also statistically significant. One *ctalk_non_epi* profile was identified as the most dissimilar (i.e. excluded from the largest significant cluster) and corresponded to H1-hESC embryonic stem cells.

For comparison and positive control, mRNA abundance profiles for the six cell types were constructed from RNA-seq data (the same values that are controlled for in the computation of *ctalk_non_epi* profiles) and then hierarchically clustered. As observed for the *ctalk_non_epi* profiles, the mRNA abundance profile corresponding to H1-hESC (embryonic stem cells) was identified as significantly dissimilar, i.e. it resulted excluded from the largest significant cluster (**Figure 2B, cluster #3**), although in this case it was excluded along with the GM12878 B-lymphoblastoids profile.

These findings indicate that *ctalk_non_epi* profiles associate with cell differentiation states *in Homo sapiens* given the H1-hESC cells' profile was identified as significantly dissimilar to those of differentiated cell types (i.e. HSMM, HUVEC, NHEK, GM12878, and NHLF). Notably, in the cell types analyzed this association was clearer than the association observed between mRNA abundance profiles and cell differentiation states.

[2] Represented by $\log_2$-transformed FPKM values.

ctalk_non_epi profiles associate with cell differentiation states in *Mus musculus*

The analysis for mouse comprised five histone H3 modifications in five cell types. As in *Homo sapiens* the *ctalk_non_epi* profiles fell into significant clusters that associated with cell differentiation states. The five comprised cell type datasets were 8-weeks-adult heart, 8-weeks-adult liver, plus three datasets of E14 embryonic stem cells after zero, four, and six days of differentiation respectively. All three E14 *ctalk_non_epi* profiles fell into a significant cluster (**Figure 2C, cluster #2**) and within it, the profiles corresponding to latter time points (four and six days of differentiation) fell into another significant cluster (**Figure 2C, cluster #1**). Additionally, the liver *ctalk_non_epi* profile was found to be more similar to the profiles of the least differentiated states than the heart profile (**Figure 2C, cluster #3**).

Mouse mRNA abundance profiles also fell into significant clusters that associate with cell differentiation states as expected (**Figure 2D, clusters #1, #2 and #3**). As *ctalk_non_epi* profiles did, transcript abundance profiles resolved a significant difference between the earliest time point (zero days of differentiation) and latter time points (**Figure 2D, cluster #1**).

This second analysis showed that the association between *ctalk_non_epi* profiles and cell differentiation states is also observable in *Mus musculus*, since the profiles for E14 embryonic stem cells (0, 4, and 6 days of differentiation) were identified as significantly dissimilar to those of 8-weeks-adult heart and liver cells.

ctalk_non_epi profiles associate with developmental periods and time points in *Drosophila melanogaster*

ctalk_non_epi profiles were computed from data for six histone H3 modifications in nine periods/time points throughout *Drosophila melanogaster* development (0-4h, 4-8h, 8-12h, 12-16h, 16-20h and 20-24h embryos; L1 and L2 larval stages; pupae). As observed in human and mouse profiles, fruit fly *ctalk_non_epi* profiles fell into clusters that also associate strongly with the degree of cell differentiation (derivable from the degree of development). One significant cluster grouped *ctalk_non_epi* profiles of earlier developmental periods (**Figure 2E, cluster #5**) apart from later development profiles. Two more significant clusters grouped later time point *ctalk_non_epi* profiles (**Figure 2E, cluster #3**) and separated the L2 larvae profile (**Figure 2E, cluster #7**) from all other profiles.

General *ctalk_non_epi* cluster structure is not entirely consistent with developmental chronology as the pupae profile (**Figure 2E, cluster #7**) shows. It must be noted however that, unlike *Homo sapiens* and *Mus musculus* data where each *ctalk_non_epi* profile represented a specific or almost specific differentiation state, each *Drosophila melanogaster* data set was obtained by the authors from whole specimens (embryos, larvae and pupae). Especially for later development, this implies that each *ctalk_non_epi* profile has to be computed from more than one partially differentiated cell type at the same developmental period, thus limiting to a certain extent the power of the analysis. This caveat in fact highlights the overall *ctalk_non_epi* cluster consistence

with developmental chronology, particularly when compared with that obtained from mRNA levels as will be detailed next.

The mRNA abundance profiles in *D. melanogaster* yielded a general cluster structure much less consistent with developmental chronology than the obtained from *ctalk_non_epi* profiles. For example, the profile for 0-4h embryos fell into the same significant cluster with the profiles for 16-20h and 20-24h embryos (**Figure 2F, cluster #3**). Additionally, the profile for 12-16h embryos fell into the same significant cluster with the profiles for L1 and L2 larvae (**Figure 2F, cluster #5**).

These results indicate that the association between *ctalk_non_epi* profiles and cell differentiation states also holds in *Drosophila melanogaster* since the obtained cluster structure displayed a significant and clear association with developmental periods and time points, which is even more consistent than that observable for mRNA abundance.

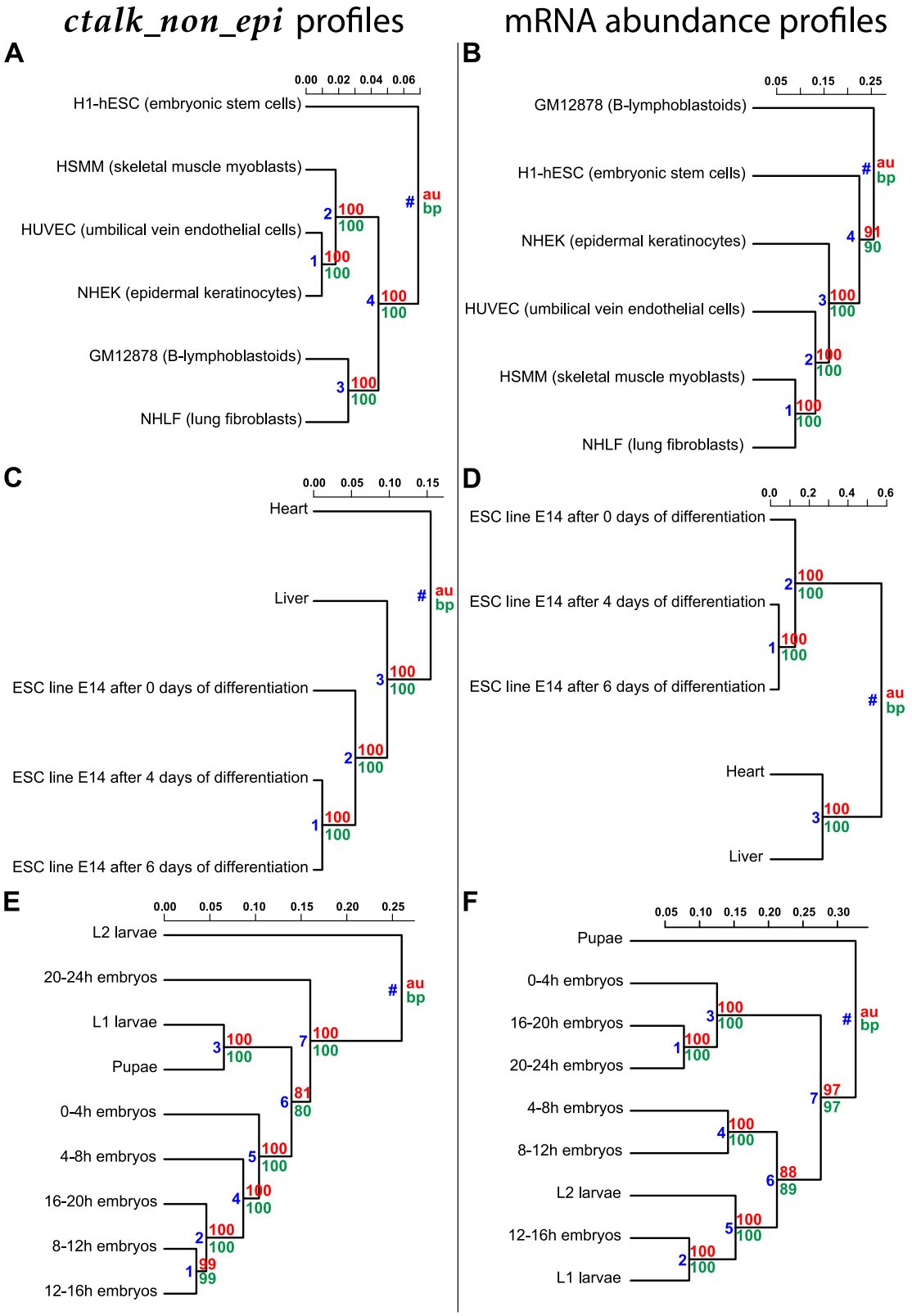

Figure 2: Unsupervised hierarchical clustering of *ctalk_non_epi* profiles and mRNA abundance profiles for *Homo sapiens* (**A**, **B**), *Mus musculus* (**C**, **D**), and *Drosophila melanogaster* (**E**, **F**). Metric: correlation $(1 - r)$. Linkage method: "average" (also known as UPGMA). Significance scores [29]: **au** (approximately unbiased) and **bp** (bootstrap probability). Significant clusters were identified as those for which **au** and **bp** ≥ 95. Cluster numbers are in blue.

Discussion

Beyond the obtained proof of principle

The most important aspect of my results is not the statistically significant relationship between *ctalk_non_epi* profiles and cell differentiation states but instead the nature of the constraints represented by *ctalk_non_epi* profiles (provided such relationship exists). By definition, *ctalk_non_epi* profiles represent the strength and sign of pairwise partial correlations (with mRNA abundance as control variable) computed from observed histone modification states; the same observed states that previous research has shown able to predict mRNA levels with high accuracy (R~0.9) [21]. It follows directly from these considerations that, for all three analyzed organisms within regions adjacent to transcription start sites (TSSs), histone H3 modification states are subject to an additional type of constraints that are explicitly uncorrelated to mRNA levels and associated with cell differentiation states. In other words two systems, mutually uncorrelated and yet both associated to cell differentiation states, ***simultaneously*** constrain histone H3 modification states.

Any resulting theory of differentiated multicellularity must address outstanding fundamental theoretical questions, including:

1) As the constraints defining the proof of principle are uncorrelated to mRNA levels, how are they associated with cell differentiation?

2) If the constraints are necessary for the intrinsic regulation of gene expression during cell differentiation, how is this regulation exerted?

3) Can the work performed by virtue of these constraints be regarded as biologically meaningful information? If so, what is the content of this information?

4) Can the constraints account for the robustness of cell differentiation under moderate perturbations?

5) How do the constraints relate to the evolution of metazoans? Is this relationship applicable to the evolution of other differentiated multicellular lineages such as plants?

6) Are histone H3 modification states cause or effect of transcriptional states? (This last question is a rehash of a very important point raised previously by Peter Fraser and Wendy Bickmore [30].)

7) Why do undifferentiated cells start to differentiate in the embryo at certain moment?

8) Why do cells stop differentiating? How does this relate to the termination of the ontogenetic process?

Table 1: Outstanding theoretical questions.

General theory of differentiated multicellularity

Based on the proof of principle obtained, I propose a general theory of differentiated multicellularity, which explains how gene expression is regulated during cell differentiation in extant multicellular lineages and how differentiated multicellular lineages emerged throughout evolution. This theory describes how two constraint-generating (also known as "self-organizing") systems elicit an emergent transition. The first system underpins the correlation between histone modification states and transcriptional states in the cell nucleus and the second system is a specific extracellular gradient generated by cell proliferation. At some moment each system starts to constrain each other synergistically, and the resulting emergent system is the differentiated multicellular organism as an individual, which must be understood as an intrinsic, higher-order constraint with logically consistent and scientifically-tenable teleological properties, in particular self-regulation. The theory explains describes how this multicellular individual is the true of gene expression during cell differentiation, and it is also falsifiable. Although its proof of principle was obtained from high-throughput metazoan data, this theoretical description makes no assumption about a specific multicellular lineage.

To highlight the similarities of molecular dynamics and spatial topology at the most fundamental level, the theory is presented in detail in ten parts described in parallel below. Each part is described in terms of the evolution of an ancestor eukaryotic species $\boldsymbol{U}$ towards differentiated multicellularity and in terms of the cell differentiation process starting from the zygote of a multicellular species $\boldsymbol{D}$. Definitions and notations are listed in **Table 2**.

Context	$X_{(i;t)}$ is the i^{th} cell of a given organism or cell population of the eukaryotic species X at a given instant t. In the same logic, *the following concepts must be understood in instantaneous terms.*
$S_E(X_{(1;t)},\ldots,X_{(n;t)})$	**Extracellular space**: The entire space in an organism or cell population that is not occupied by its n cells themselves at a given instant t. Positions in $S_E(t)$ will be specified in spherical coordinates, namely r (radial distance), θ (azimuthal angle), and ϕ (polar angle).
$C_W(X_{(i;t)})$	**Waddington's constraints**: The constraints associating certain subsets of the spatially-specified molecular nuclear phenotype of $X_{(i;t)}$ with the instantaneous transcription rates at the transcription start sites (TSSs), provided changes in these Waddington's constraints $C_W(X_{(i;t)})$ are *explicitly uncorrelated* with changes in the genomic sequence.
$F_W(X_{(i;t)})$	**Waddington's embodiers**: The largest subset of the spatially-specified molecular nuclear phenotype of $X_{(i;t)}$ for which the Waddington's constraints $C_W(X_{(i;t)})$ are significant (e.g. histone H3 post-translational modifications in the TSS-adjacent genomic regions).
$C_N(X_{(i;t)})$	**Nanney's constraints**: The constraints associating certain subsets of the spatially-specified molecular nuclear phenotype of $X_{(i;t)}$ with the Waddington's embodiers $F_W(X_{(i;t)})$, provided changes in these Nanney's constraints $C_N(X_{(i;t)})$ are *explicitly uncorrelated* with changes in the instantaneous transcription rates at the TSSs. In this work Nanney's constraints were represented by the *ctalk_non_epi* profiles.
$F_N(X_{(i;t)})$	**Nanney's embodiers**: The largest subset of the spatially-specified molecular nuclear phenotype of $X_{(i;t)}$ for which the Nanney's constraints $C_N(X_{(i;t)})$ are significant. Crucially, histone H3 post-translational modifications in the TSS-adjacent regions—as inferable from the Results—can be specifiable as Waddington's embodiers F_W and as Nanney's embodiers F_N *simultaneously.*
$F_N^{\rightarrow}(X_{(i;t)})$	**Nanney's extracellular propagators**: The subset of the entire spatially-specified molecular phenotype of $X_{(i;t)}$ that excludes Nanney's embodiers $F_N(X_{(i;t)})$ but is (i) secreted into the the extracellular space S_E and (ii) capable of eliciting a change (via facilitated diffusion/signal transduction) in Nanney's embodiers F_N within other cells after a certain time interval Δt.

Table 2: **Preliminary theoretical definitions and notation**.

Part I (Evolution) **The unicellular (or undifferentiated multicellular) ancestor**

- $\boldsymbol{U}_{(i;t_{U_0})}$ is the $\boldsymbol{i}^{th}$ cell in a population of the unicellular (or undifferentiated multicellular) species $\boldsymbol{U}$ (**Figure 3A, top**).
- $\boldsymbol{U}_{(i;t_{U_0})}$ displays Waddington's embodiers $\boldsymbol{F}_W(\boldsymbol{U}_{(i;t_{U_0})})$ (e.g. histone post-translational modifications able to elicit changes in transcriptional rates) but cell differentiation is not possible.
- Certain constraints exist on Waddington's embodiers $\boldsymbol{F}_W(\boldsymbol{U}_{(i;t_{U_0})})$ that are *explicitly uncorrelated* with transcriptional rates. In other words, significant Nanney's constraints $\boldsymbol{C}_N(\boldsymbol{U}_{(i;t_{U_0})})$ exist.
- However, the propagation (if any) of Nanney's constraints $\boldsymbol{C}_N$ is confined to $\boldsymbol{U}_{(i;t_{U_0})}$. In other words, Nanney's extracellular propagators $\boldsymbol{F}_N^{\rightarrow}$ do not exist in $\boldsymbol{U}_{(i;t_{U_0})}$.

Part I (Ontogeny) **The differentiated multicellular organism's zygote**

- $\boldsymbol{D}_{(1;t_{D_0})}$ is a zygote of the extant differentiated multicellular species $\boldsymbol{D}$ (**Figure 3A, bottom**).
- Like $\boldsymbol{U}_{(i;t_{D_0})}$, $\boldsymbol{D}_{(1;t_{D_0})}$ displays Waddington's embodiers $\boldsymbol{F}_W(\boldsymbol{D}_{(i;t_{D_0})})$ (e.g. histone post-translational modifications able to elicit changes in transcriptional rates) but cell differentiation is not observed *yet*.
- Certain constraints exist on Waddington's embodiers $\boldsymbol{F}_W(\boldsymbol{D}_{(1;t_{D_0})})$ that are *explicitly uncorrelated* with transcriptional rates. In other words, significant Nanney's constraints $\boldsymbol{C}_N(\boldsymbol{D}_{(1;t_{D_0})})$ exist.
- Unlike in $\boldsymbol{U}_{(i;t_{D_0})}$, the propagation of Nanney's constraints $\boldsymbol{C}_N$ is *not* confined to $\boldsymbol{D}_{(1;t_{D_0})}$. In other words, Nanney's extracellular propagators $\boldsymbol{F}_N^{\rightarrow}$ do exist in $\boldsymbol{D}_{(1;t_{D_0})}$.

Part II (Evolution) **Necessary novel alleles**

- At some time point $(\boldsymbol{t}_M - \boldsymbol{\Delta t}_M) > \boldsymbol{t}_{U_0}$ during evolution the genome of certain $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ cell suffers a change (**Figure 3A to 3B**) such that it now synthesizes a molecule specifiable as a Nanney's extracellular propagator $\boldsymbol{F}_N^{\rightarrow}$.
- As described in the preliminary definitions, this means a molecular substrate is synthesized that is membrane exchangeable and, once entering the cell, is also able to elicit a change in Nanney's embodiers $\boldsymbol{F}_N(\boldsymbol{U}_{(i;t_{U_0})})$ (e.g. histone post-translational modifications). Importantly, this change is *explicitly uncorrelated* with the transcriptional rates at the instant it is elicited.

- The genetic change implies that the genome now codes for all gene products necessary for the synthesis, facilitated diffusion/signal transduction of the novel Nanney's extracellular propagator(s) $\vec{F_N}$.
- Importantly, the novel alleles are a necessary but not sufficient condition for differentiated multicellularity (**Figure 3B**).

Part II (Ontogeny)

Already present necessary alleles

- At any instant $(t_D - \Delta t_D) > t_{D_0}$ the genome of any cell $D_{(i;t_D-\Delta t_D)}$ in the zygote's offspring is similar to the genome of the cell $U_{(k;t_M-\Delta t_M)}$ (see **Figure 3B, top**) in the sense that both genomes code for Nanney's extracellular propagators $\vec{F_N}$.
- Importantly, the alleles specified in the genome of the zygote $D_{(1;t_{D_0})}$—and in the genome of any cell in its offspring—are a necessary but not sufficient condition for cell differentiation (**Figure 3B**).

Part III (Evolution & Ontogeny)

Diffusion flux of Nanney's extracellular propagators and the geometry of the extracellular space S_E

- The existence of Nanney's extracellular propagators $\vec{F_N}$ allows to define a scalar field[3] Φ_N describing the concentration of $\vec{F_N}$ in the extracellular space S_E at any instant t.
- When the number of cells is small enough, diffusion flux is fast enough to overtake the spatial constraints imposed by the relatively simple geometry of S_E.
- Therefore, under these conditions the associated gradient[4] $\vec{\nabla}\Phi_N$ remains in magnitude—anywhere in S_E—under a certain critical value V_M for the offspring of the cell $U_{(k;t_M-\Delta t_M)}$ and under a critical value V_D for the offspring of the zygote $D_{(1;t_{D_0})}$ (**Figure 3B, bottom**).
- Importantly, the constraints represented by the gradient $\vec{\nabla}\Phi_N$ imply there is free energy available—whether or not there is cell differentiation—which, as will be described later, is in fact partially used as work in the emergence of new information content.

[3] A scalar field is a function associating a scalar (here concentration of Nanney's extracellular propagators $\vec{F_N}$) to every point in space.

[4] The gradient vector field $\vec{\nabla}$ of a scalar function (in this context, the scalar field Φ_N) is a vector operation that generalizes the concept of derivative represented by the differential operator—denoted by the ∇ (nabla) symbol and also called "del"—to more than one dimension.

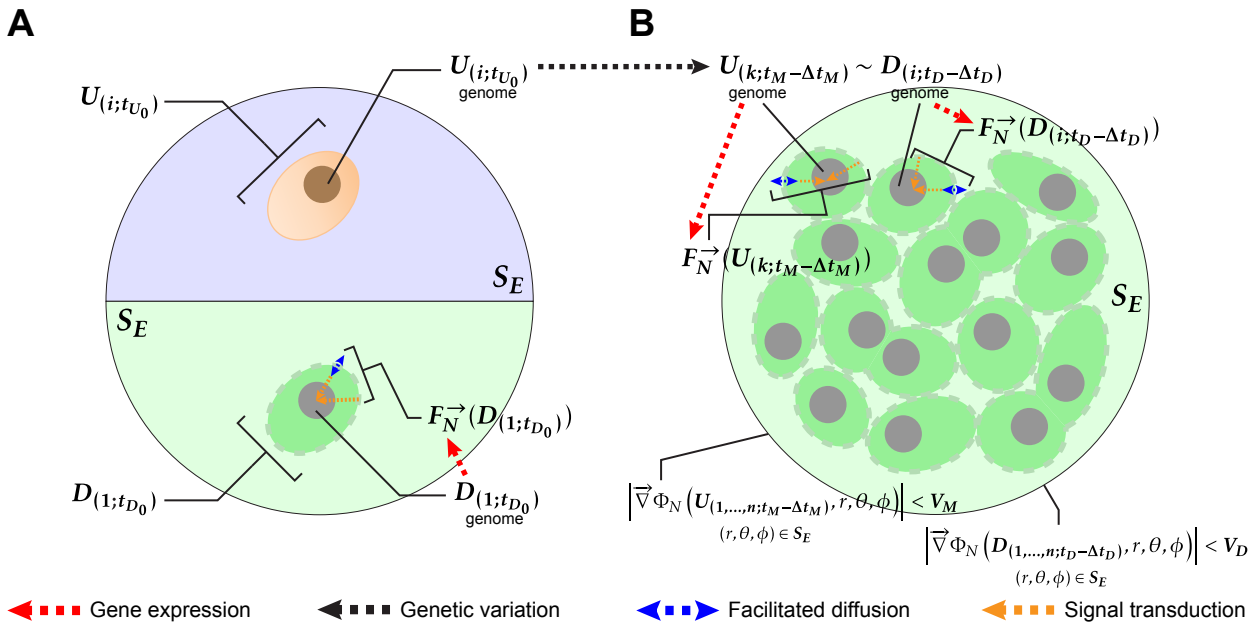

Figure 3: **A, (top)**: A cell of the unicellular and undifferentiated ancestor species U. **A (bottom)**: A zygote of the multicellular species D. **A (top) to B (top)**: The necessary genetic change for differentiated multicellularity occurs in the species U. **B (top)**: The similar and necessary alleles are now present in both species. **B (bottom)**: Cells proliferate but no significant $\vec{\nabla}\Phi_N$ gradients form yet in S_E and no differentiation is observed.

Part IV (Evolution)

The emergent transition to differentiated multicellularity

- At some instant t_M, later but relatively close to $(t_M - \Delta t_M)$, cell proliferation yields a significantly larger population. Now diffusion flux of Nanney's extracellular propagators $F_N^{\rightarrow}$ is no longer able to overtake the increasing spatial constraints in the extracellular space S_E.
- Under these conditions a significant gradient, in magnitude equal or greater—anywhere in S_E—than the critical value V_M forms, i.e. $\left|\vec{\nabla}\Phi_N\left(U_{(1;t_M)},\ldots,U_{(n;t_M)},r,\theta,\phi\right)\right| \geq V_M, (r,\theta,\phi) \in S_E$ (**Figure 4, bottom-left**).
- As consequence, Nanney's extracellular propagators $F_N^{\rightarrow}$ diffuse differentially into each cell, yielding unprecedented differential Nanney's constraints $\{C_N(U_{(1;t_M)}),\ldots,C_N(U_{(n;t_M)})\}$ in the cells' nuclei by virtue of no cell or gene product in particular but, importantly, of the constraints imposed by the entire proliferating cell population on the diffusion flux of $F_N^{\rightarrow}$ in S_E.
- These differential Nanney's constraints C_N in turn elicit differential changes in Waddington's embodiers $\{F_W(U_{(1;t_M)}),\ldots,F_W(U_{(n;t_M)})\}$ within the cells' nuclei (**Figure 4, top-left**), thus they now constrain the instantaneous

transcription rates in a differential and explicitly uncorrelated manner. This is how multicellular lineages, displaying self-*regulated* changes in gene expression during ontogeny, evolved.

Part IV (Ontogeny) **The emergent transition to cell differentiation**

- At some instant t_D, later but relatively close to $(t_D - \Delta t_D)$, embryonic growth yields certain number of undifferentiated cells. Now diffusion flux of Nanney's extracellular propagators is no longer able to overtake the increasing spatial constraints in the extracellular space S_E.
- Under these conditions a significant gradient forms, in magnitude equal or greater—anywhere in S_E—than the critical value V_D, i.e. $\left|\vec{\nabla}\Phi_N\left(D_{(1;t_D)},\ldots,D_{(n;t_D)},r,\theta,\phi\right)\right| \geq V_D, (r,\theta,\phi) \in S_E$ (**Figure 4, bottom-right**).
- As consequence, Nanney's extracellular propagators $\vec{F_N}$ diffuse differentially into each cell, yielding unprecedented differential Nanney's constraints $\{C_N(D_{(1;t_D)}),\ldots,C_N(D_{(n;t_D)})\}$ in the cells' nuclei by virtue of no cell or gene product but, importantly, of the constraints imposed by the entire growing embryo on the diffusion flux of Nanney's extracellular propagators in the extracellular space S_E (see **Table 1, question #7**).
- These differential Nanney's constraints C_N in turn elicit differential changes in Waddington's embodiers $\{F_W(D_{(1;t_D)}),\ldots,F_W(D_{(n;t_D)})\}$ within the cells' nuclei (**Figure 4, top-right**), thus they now constrain the instantaneous transcription rates in a differential and explicitly uncorrelated manner. This is how undifferentiated cells start to differentiate, displaying self-*regulated* changes in gene expression during ontogeny (see **Table 1, question #1**).

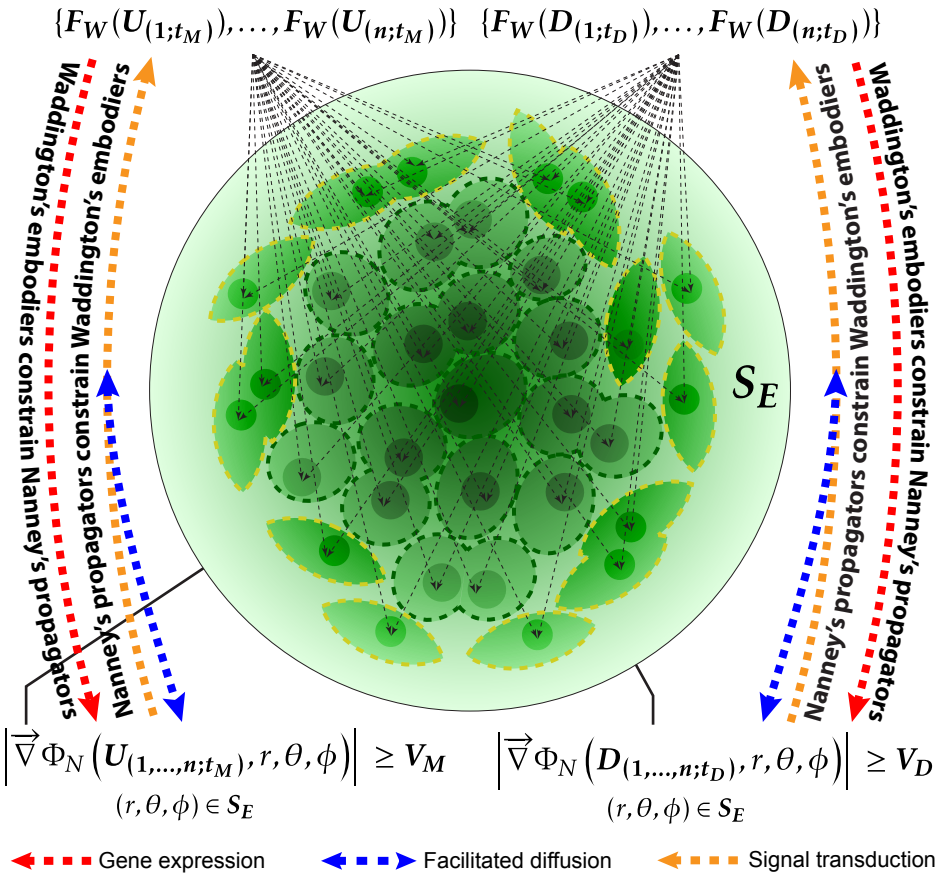

Figure 4: Emergent transition to differentiated multicellularity/cell differentiation. **Dashed arrows**: Intrinsic higher-order constraint emerges when significant gradients $\vec{\nabla}\Phi_N$ (bottom) couple the lower-order Nanney's constraints $\boldsymbol{C_N}$ and Waddington's constraints $\boldsymbol{C_W}$ synergistically across $\boldsymbol{S_E}$. **Top**: Waddington's embodiers $\boldsymbol{F_W}$ constrain—via gene expression (red dashed arrows)—the membrane exchange of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ driven by the gradients. **Bottom**: Gradients $\vec{\nabla}\Phi_N$ constrain in turn—via facilitated diffusion/signal transduction of $\boldsymbol{F_N^{\rightarrow}}$ (blue and orange dashed arrows)—Waddington's embodiers $\boldsymbol{F_W}$ and as a consequence constrain also transcription rates and gene expression levels.

Part V (Evolution)

What was the evolutionary breakthrough?

- Since the oldest undisputed differentiated multicellular organisms appear in the fossil record around 2.8 billion years after the first stromatolites [30], the necessary genetic change from the genome of the cell $\boldsymbol{U}_{(i;t_{U_0})}$ to the genome of the cell $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ can be safely regarded as a highly improbable step.
- Nevertheless, the major evolutionary breakthrough was not genetic but instead the unprecedented dynamical regime emerging from proliferating eukaryote cells at $\boldsymbol{t_M}$, or in more general terms at $\{\boldsymbol{t_{M_1}},\ldots,\boldsymbol{t_{M_n}}\}$ throughout evolution since extant differentiated multicellular organisms constitute a paraphyletic group [31, 32].
- This novel dynamical regime emerges as a higher-order constraint from the synergistic coupling of the lower-order Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$, able now to propagate through the extracellular space $\boldsymbol{S_E}$ (**Figure 4, dashed arrows**).
- Although dependent on the novel alleles in the genome of $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ to emerge given enough cell proliferation, this system is not a network of epigenetic mechanisms—however complex—but instead a particular example of the generic *teleodynamic system*, proposed by Terrence Deacon in his *theory of biological individuality by constraint coupling and preservation*[5] [33], which is presented to and shaped by natural selection at each instant. In this context, environmental constraints as oxygen availability [34] and even gravity (see Corollary #5) filter out specific emergent multicellular dynamics that are incompatible with them.
- In summary, the critical evolutionary novelty was the unprecedented multicellular individual or multicellular *self*, which can be described as an intrinsic, higher-order dynamical constraint that emerges spontaneously from a particular class of proliferating eukaryotic cells. Being a higher-order *constraint*, this multicellular *self* is causally-efficacious when regulating its intrinsic dynamics or its surroundings.

Part V (Ontogeny)

***Who* is regulating cell differentiation?**

- Contrary to what could be derived from Turing's approach [35], the theory hereby proposed does *not* regard the significant proliferation-generated extracellular gradient, i.e. $\left|\vec{\nabla}\Phi_N\right| \geq \boldsymbol{V_D}$ (anywhere in $\boldsymbol{S_E}$), as the fundamental regulator of the cell differentiation process.

[5] Although Deacon himself named his theory *emergent dynamics*, I am proposing here this longer but more descriptive name.

- Whereas differential Nanney's constraints $\{C_N(D_{(1;t_D)}),\ldots,C_N(D_{(n;t_D)})\}$ are *regulatory constraints* with respect to Waddington's embodiers $\{F_W(D_{(1;t_D)}),\ldots F_W(D_{(n;t_D)})\}$ as described in Part IV-Ontogeny (see **Figure 4, blue/orange dashed arrows**), the reciprocal proposition is also true. Namely, Waddington's constraints $\{C_W(D_{(1;t_D)}),\ldots,C_W(D_{(n;t_D)})\}$ are *explicitly uncorrelated* to Nanney's constraints, thus they are in turn *regulatory constraints* with respect to Nanney's extracellular propagators $\{\vec{F_N}(D_{(1;t_D)}),\ldots,\vec{F_N}(D_{(n;t_D)})\}$, e.g. changes in the expression of the protein channels, carriers or membrane receptors necessary for the facilitated diffusion/signal transduction of Nanney's extracellular propagators (see **Figure 4, red dashed arrows**).
- *Consequently, only if the explicitly uncorrelated Waddington's constraints C_W and Nanney's constraints C_N*[6] *become synergistically coupled* (**Figure 4, dashed arrows**) *across the extracellular space S_E true intrinsic regulation on the cell differentiation process is possible.*
- This implies in turn that both chromatin states and transcriptional states are simultaneously cause and effect with respect to each other (this regime, intuitively describable as "chicken-egg" dynamics, is the answer this theory provides to **question #6 in Table 1**).
- The true regulator of the cell differentiation process is then the developing multicellular organism itself. This is because the multicellular organism *is* the causally-efficacious, higher-order constraint emerging from and regulating *ipso facto* Nanney's constraints C_N and Waddington's constraints C_W (when coupled synergistically across the extracellular space S_E) in what would be otherwise a population or colony—however symbiotic—of unicellular eukaryotes (see **Table 1, question #2**).

Part VI (Evolution) **Unprecedented multicellular dynamics**

- Once the necessary alleles for differentiated multicellularity were present in some eukaryotic lineages, phenomena like mutation, gene duplication or alternative splicing—in the loci involved in the synthesis, facilitated diffusion or signal transduction of Nanney's extracellular propagators $\vec{F_N}$—made possible the emergence of a plethora of novel multicellular (*teleodynamic*) regimes.
- Moreover, the dependence of differentiated multicellularity on one or more coexisting $\vec{\nabla}\Phi_N$ gradients (i.e. constraints on diffusion flux) in S_E, which importantly depend on no cell in particular but on the entire cell population or embryo, yields an important implication in evolutionary terms. That is, since a higher-order constraint is taking over the regulation of changes

[6] Both emerge in turn from genetic (i.e. structurally embodied) constraints.

in gene expression within individual cells, it is predictable that said cells lose some cell-intrinsic systems that were critical at a time when eukaryotic life was only unicellular, even when compared with their prokaryotic counterparts[7].

- In this context a result obtained over a decade ago acquires relevance: in a genome-wide study it was found that that the number of transcription factor genes increases as a power law of the total number of protein coding genes, with an exponent greater than 1. In other words, the need for transcription-factor genetic information increases faster than the total amount of genetic information it is involved in regulating [36]. Remarkably, the eukaryotes analyzed—~10 genomes, most from differentiated multicellular organisms—were the group with the smallest (i.e. closest to linearity) power-law exponent. This means that the most complex organisms require proportionally *less* transcription-factor information. With data available today [37], a reproduction of the aforementioned analysis allowed in this work a robust confirmation: the power-law exponent for unicellular or undifferentiated multicellular eukaryotes is 1.33 ± 0.31 (37 genomes). For differentiated multicellular eukaryotes it is 1.11 ± 0.18 (67 genomes)[8]. The previously described loss of lower-order, cell-intrinsic regulatory systems in differentiated multicellular organisms—accounted for by the emergence of higher-order information content (see Part IX)—is consistent with the otherwise counterintuitive differences in power-law exponents.

Part VI (Ontogeny) **What does ontogeny recapitulate?**

- As the key to the evolution of any multicellular lineage displaying self-regulated changes in gene expression during cell differentiation, the proposed theory holds the emergent transition, spontaneous from cell proliferation shortly after Nanney's extracellular propagators $\vec{F_N}$ appeared.
- Therefore, this theoretical description rejects the hypothesis that metazoans—or, in general, any multicellular lineage displaying self-regulated cell differentiation—evolved from gradual specialization of single-cell colonies or aggregations [38, 39, 40, 32, 41, 42, 43, 44].
- Importantly, however, this rejection does not imply that precedent traits (e.g. cell-cell adhesion) were unimportant for the later fitness of differentiated multicellular organisms.

[7] T. Deacon generically described this as the offloading of teleodynamic constraints in lower-order systems—at the cost of losing teleodynamic properties—into the higher-order teleodynamic system emerging from them.

[8] The difference between the two estimates is statistically significant (assessed by 95%-confidence, nonparametric BCa bootstrapping).

- This theory does not reject Haeckel's famous assertion completely: in every extant multicellular lineage, this self-sufficient, self-repairing, self-replicating, and self-regulating system emerges over and over again from undifferentiated cells and presents itself to natural selection ever since its evolutionary debut. Therefore, at least in this single yet most fundamental sense, ontogeny does recapitulate phylogeny.

Part VII (Evolution & Ontogeny) **The role of epigenetic changes**

- Contrary to what the epigenetic landscape framework entails, under this theory the heritable changes in gene expression do not define let alone explain the intrinsic regulation of cell differentiation.
- The robustness, heritability, and number of cell divisions which any epigenetic change comprises are instead adaptations of the higher-order dynamical constraint emergent from proliferating individual cells (i.e. the multicellular organism).
- These adaptations have been shaped by natural selection after the emergence of each extant multicellular lineage and are in turn reproduced or replaced by novel adaptations in every successful ontogenetic process.

Part VIII (Evolution & Ontogeny) **Novel cell types, tissues and organs evolve and develop**

- Further genetic variation in the novel alleles in the genome of the cell $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ or the already present alleles in the genome of the $\boldsymbol{D}_{(1;t_{D_0})}$ (e.g. mutation, gene duplication, alternative splicing) imply than one or more than one $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ gradients emerge in $\boldsymbol{S_E}$ with cell proliferation.
- A cell type $\boldsymbol{T_j}$ will develop then in a region $\boldsymbol{S_{E_i}}$ of the extracellular space $\boldsymbol{S_E}$ when a relative uniformity of Nanney's extracellular propagators is reached, i.e. $\left(\left|\overrightarrow{\nabla}\Phi_{N_{1;T_j}}\right|,\ldots,\left|\overrightarrow{\nabla}\Phi_{N_{k;T_j}}\right|\right)<\left(\boldsymbol{V}_{N_{1;T_j}},\ldots,\boldsymbol{V}_{N_{k;T_j}}\right),(r,\theta,\phi)\in\boldsymbol{S_{E_i}}$, where $\left(\boldsymbol{V}_{N_{1;T_j}},\ldots,\boldsymbol{V}_{N_{k;T_j}}\right)$ are certain critical values (see a two-cell-type and two-gradient depiction in **Figure 5**).
- As highlighted earlier, cell differentiation is not *regulated* by these gradients themselves but by the higher-order constraint emergent from the synergistic coupling of Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ across $\boldsymbol{S_E}$.
- This constraint synergy can be exemplified as follows: gradients $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ can elicit changes in gene expression in a number of cells, which in turn may promote the dissipation of the gradients

(e.g. by generating a surrounding membrane that reduces dramatically the effective S_E size) or may limit further propagation of those gradients from S_E into the cells (e.g. by repressing the expression of genes involved in the facilitated diffusion/signal transduction of $F_N^{\rightarrow}$ in S_E).

- Thus, under this theory cell types, tissues, and organs evolved sequentially as "blobs" of relative $F_N^{\rightarrow}$ uniformity in regions $\{S_{E_1}, \ldots, S_{E_n}\}$ (i.e. regions of relatively small $\overrightarrow{\nabla}\Phi_N$ magnitude) within S_E displaying no particular shape or function—apart from not being incompatible with the multicellular organism's survival and reproduction—by virtue of genetic variation (involved in the embodiment and propagation of Nanney's constraints C_N) followed by cell proliferation.
- The $F_N^{\rightarrow}$-uniformity "blobs" emerged with no function in particular—apart from being compatible with the multicellular organism's survival and reproduction—by virtue of random genetic variation (involved in the embodiment and propagation of Nanney's constraints C_N) followed by cell proliferation.
- Then, these $F_N^{\rightarrow}$-uniformity "blobs" were shaped by natural selection from their initially random physiological and structural properties to specialized cell types, tissues, and organs. Such specialization evolves with respect to the emergent intrinsic higher-order constraint proposed here as the multicellular organism. The result of this emergence-selection process is observable in the dynamics characterizing the ontogeny of extant multicellular species (**Figure 6A**).

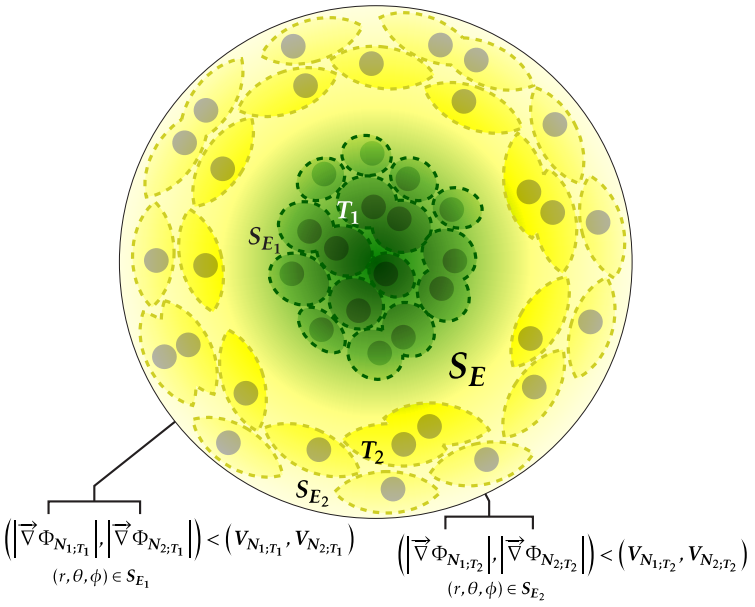

Figure 5: Two distinct cell types T_1 and T_2 develop respectively in regions S_{E_1} and S_{E_2} within S_E characterized by a relative small $\overrightarrow{\nabla}\Phi_N$ gradient magnitude, i.e. in extracellular regions of relative $F_N^{\rightarrow}$ uniformity.

Part IX (Evolution & Ontogeny)

Emergent *hologenic* information and multicellular self-repair

- As argued in the introduction, a significant amount of information content has to *emerge* to account for robust and reproducible cell fate decisions and for the self-regulated dynamics of cell differentiation in general.
- Under this theory, this content emerges when the significant gradient or gradients $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ form at some point from proliferating undifferentiated cells, entangling synergistically Nanney's constraints $\boldsymbol{C_N}$ and Waddington's constraints $\boldsymbol{C_W}$ across $\boldsymbol{S_E}$.
- Crucially, this information is *not* about any coding sequence and its relationship with cell-intrinsic and cell-environment dynamics (i.e. genetic information) *nor* about any heritable gene expression level/profile and its relationship with cell-intrinsic and cell-environment dynamics (i.e. epigenetic information).
- Instead, this information is *about the multicellular organism as a whole* understood as the emergent higher-order intrinsic constraint described previously and also about the environmental constraints under which this multicellular organism develops. For this reason I propose to call this emergent information *hologenic*[9] (see **Table 1, question #3**).
- No less importantly, at each instant the multicellular organism is not only interpreting hologenic information—by constraining its development into specific trajectories since it emerges—but also actively generating novel hologenic information (in other words displaying "chicken-egg" dynamics, similar to those described in Part V-Ontogeny).
- In the multicellular organism, the subset of the molecular phenotype that conveys hologenic information is not only the subset involved in the gradients $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ but the entire subset embodying or propagating Nanney's constraints $\boldsymbol{C_N}$.
- Additionally, since the gradients $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ conveying hologenic information depend on no cell in particular but on the spatial constraints imposed by the entire cell population or embryo, cell differentiation will be robust with respect to moderate perturbations such as some cell loss (see **Table 1, question #4**).

[9] ὅλος is the ancient Greek for "whole" or "entire".

Part X (Ontogeny) **Ontogeny ends and cell differentiation "terminates"**

- If under this theory cell differentiation emerges with the proliferation of (at the beginning, undifferentiated) cells, why should it terminate for any differentiation lineage? What is this "termination" in fundamental terms? These are no trivial questions. As an answer to the first, zero net proliferation begs the fundamental question; to the second, a "fully differentiated" cell state condition fails to explain the existence of adult stem cells. To address these issues three considerations are most important:

(i) For any cell or group of cells the molecules specifiable as Nanney's extracellular propagators $\vec{F_N}$ at any instant t may not be specifiable as such at some later instant $t + \Delta t$.

(ii) The emergent *telos* or "end" in this theory is the instantaneous, higher-order intrinsic constraint that emerges from proliferating undifferentiated cells (i.e. multicellular *self*); *not* the "intuitive" *telos* described in the introduction—such as the organism's mature form, a fully differentiated cell, or certain future transcriptional changes to achieve such states—which is logically inconsistent[10].

(iii) This causally-efficacious, higher-order constraint emerges from the synergistic coupling of lower-order Waddington's constraints C_W and Nanney's constraints C_N across the extracellular space S_E.

- Therefore, under this theory cell differentiation "terminates" in any given region S_{E_i} of the extracellular space if a stable or metastable equilibrium is reached where at least one of the two following conditions is true:

(a) The gradients $\{\vec{\nabla}\Phi_{N_1}, \ldots, \vec{\nabla}\Phi_{N_k}\}$ dissipate in S_{E_i} under certain critical values, i.e. $\left(\left|\vec{\nabla}\Phi_{N_1}\right|, \ldots, \left|\vec{\nabla}\Phi_{N_k}\right|\right) < \left(V_{D_1}, \ldots, V_{D_k}\right), (r, \theta, \phi) \in S_{E_i}$ (see **Table 1, question #8** and **Figure 6B, left**).

- Condition (a) can be reached for example when development significantly changes the morphology of the cells by increasing their surface-to-volume ratio, because such increase removes spatial constraints in S_E that facilitate the emergence/maintenance of the gradients.

- Thus, under this theory, one can predict a *significant positive correlation between the degree of differentiation of a cell and its surface-to-volume ratio* and also a *significant negative correlation between cell potency/regenerative capacity and that ratio.* These correlations must be assessed after controlling for cell characteristic length or "unidimensional size".

[10] Since such a *telos* entails the causal power of future events on events preceding them.

(b) The gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ are unable to constrain Waddington's embodiers $\boldsymbol{F_W}$ in the cells' nuclei because the critical gene products (protein channels/carriers or signal transducers) are non-functional or not expressed when the cells become "blind" to the gradients (see **Table 1, question #8** and **Figure 6B, right**).

- Condition (b) can be reached when the cell differentiation process represses at some point the expression of the protein channels or carriers necessary for the facilitated diffusion/signal transduction of the *current* Nanney's extracellular propagators $\boldsymbol{F}_{\vec{N}}$.
- Importantly, the stability of the equilibrium will depend on the cells' currently expressed phenotype. Thus, an adult multipotent or pluripotent stem cell may differentiate if needed [45] or some differentiated cells may dedifferentiate given certain stimuli [46] (metastable equilibrium), whereas a fully differentiated neuron does not (very stable equilibrium).
- These examples underscore that the *telos* of cell differentiation is not a "fully differentiated" state but, as this theory explains, the instantaneous, intrinsic higher-constraint which is the multicellular organism as a whole. Consequently, the "termination" of cell differentiation should be understood rather as an indefinite-as-long-as-functional stop, or even as apoptosis (see **Table 1, question #8**.
- The multicellular *telos* described will prevail in ontogeny (and did prevail in evolution) as long as an even higher-order *telos* does not emerge from it (e.g. once a central nervous system develops/evolved).

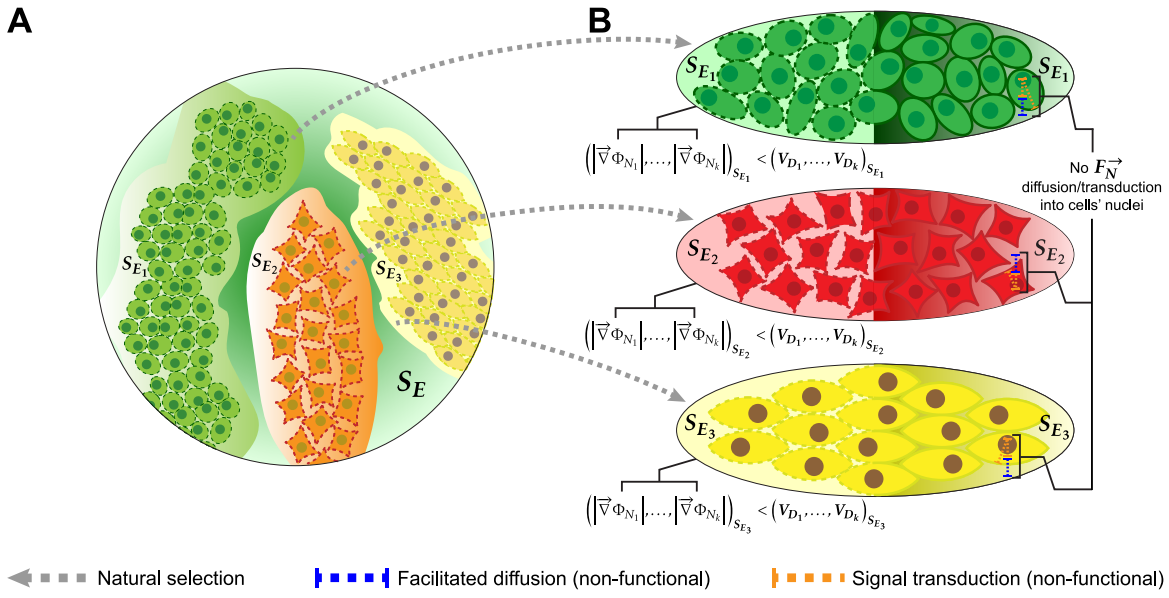

Figure 6: **A**: Cell types/tissues/organs evolve as emergent "blobs" of relatively small $\vec{\nabla}\Phi_N$ magnitude and then are shaped by natural selection **(E)**. **B**: Cell differentiation stops when the $\vec{\nabla}\Phi_N$ gradients dissipate **(left)**, or when they cannot diffuse/be transduced into the cells' nuclei **(right)**.

Part X (Evolution)

The evolutionarily-shaped multicellular *telos*

- Whereas the causal power of the organism's mature form as ontogenetic *telos* is logically inconsistent, the assumption that the zygote is a complete developmental blueprint containing all necessary information for the process, as argued in the introduction, is also untenable.
- In contrast, ontogeny is, under this theory, an emergent, evolutionarily-shaped and instantaneously-defined (i.e. logically consistent) teleological process. The reason why it intuitively appears to be "directed" to and by the organism's mature form is that the intrinsic higher-order constraint—the true (instantaneous) *telos* described previously—and the hologenic information content emerging along with it are exerting efficacious causal power on the ontogenetic process.
- Although the propagation of constraints within this process, such as propagated changes in gene expression, is decomposable into molecular interactions, its "end-directed" causal power in self-regulation *is not* because the *telos* or "end" (see **Figure 1, top-right**) is a spontaneous, intrinsic higher-order *constraint* or "dynamical analogue of zero" emergent from lower-order constraints. Therefore, this teleological causal power cannot be mechanistically reduced or decomposed into molecules and their interactions[11] as first argued by T. Deacon [33].
- Hologenic content is thermodynamically *"absent"* or constrained: hologenic content is not in the molecular substrates conveying that content anymore than the content of this theory is in integrated circuits, computer displays, paper, or even in the complex neural interactions within the reader's brain. *The realizable states that are constrained or made "absent" in the dynamics of the multicellular organism by the synergistic coupling of Waddington's constraints C_W and Nanney's constraints C_N across S_E is the content of hologenic information*; the substrates embodying and propagating the critical constraints for the coupling can only then be identified as conveying hologenic information.
- Evolution has thus selected the content of hologenic information by capturing the lower-order genetic information it is ultimately emergent from, not any particular molecules or molecular interactions as media, media that should be regarded in this context as means to the multicellular *telos* as the etymology implies. This explanation also implies a trade-off between cell independence and cell phenotypic complexity: the multicellular *telos* offloads regulatory work/constraints the cells were performing individually (as described in Part VI-Evolution) allowing them to use that free energy surplus in more complex and differentiated dynamics but also making them more dependent on that multicellular *telos*.

[11] Like its dynamical analogue the arithmetic zero cannot be divided, and for the same fundamental reason.

- In this context, the necessary genetic change from the genome of the cell $\boldsymbol{U}_{(i;t_{U_0})}$ to the genome of the cell $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ (described in Part II-Evolution) could well have been significantly smaller in terms of DNA or protein sequence than other genetic changes suffered by the eukaryotic ancestors of $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ while never leaving unicellularity or undifferentiated multicellularity. In general, accounting for substantial differences in the phenotype and its properties given comparatively small genetic changes is bound to be an intractable task if one or more teleodynamic transitions during evolution are involved but ignored.
- The description for the evolution of cell types, tissues and organs based on initial "blobs" of relative $\vec{\boldsymbol{F}_N}$ uniformity in $\boldsymbol{S_E}$ together with the predicted positive correlation between degree of cell differentiation and cell surface-to-volume ratio suggest an additional and more specific evolutionary implication.
- The high surface-to-volume ratio morphology needed for neuron function was only to be expected in the evolution of multicellularity, provided no rigid wall (of high relative fitness) impedes the tinkering with substantial increases of the cells' surface-to-volume ratio, as observable in plants.
- Together with the predicted negative correlation between cell potency and surface-to-volume ratio, this caveat suggests that if a multicellular lineage is constrained to display low cell surface-to-volume ratios, cell potency and regenerative capacity will be higher. These multicellular lineages can be expected to have a comparatively lower complexity but longer lifespan and more robustness to extrinsic damage (see **Table 1, question #5**).

The synergy in the coupling of Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ across $\boldsymbol{S_E}$ described in this theory does not preclude that cell differentiation may display phases dominated by proliferation and others dominated by differentiation itself: whereas significant gradients of Nanney's extracellular propagators $\vec{\boldsymbol{F}_N}$ in $\boldsymbol{S_E}$ emerge at some point given enough cell proliferation, it is also true that the exchange of such propagators between the cells and $\boldsymbol{S_E}$ is constrained by the dynamics of facilitated diffusion and/or ligand-receptor binding which, importantly, are saturable. Any representative simulation of cell differentiation according to this theory, however simple, will depend on an accurate modeling of the lower-order dynamical constraints it emerges from.

The proposed theory also encompasses coenocytic (also commonly called "syncytial") stages of development, where cell nuclei divide in absence of cytokinesis, such as in *Drosophila*). In such stages, Nanney's extracellular propagators have to be operationally redefined as Nanney's *extranuclear* propagators, while still maintaining their fundamental defining property: the ability to elicit a change in Nanney's embodiers $\boldsymbol{F_N}$ inside the nucleus.

Related to this theory, evidence has already been found for tissue migration across a migration-generated chemokine gradient in zebrafish [47, 48]. This finding demonstrates the feasibility of some of the dynamics proposed here, namely eukaryotic cells using certain free energy as work in their own intrinsic dynamics. Such free energy may be available in the spontaneous constraints on diffusion in S_E generated by cell migration/proliferation. The two linked processes—one spontaneous, the other non-spontaneous—exemplify a work cycle as proposed by Stuart Kauffman [49]. What remains to be verified is the synergistic coupling of—in this case two—lower-order constraint-generating (i.e. self-organizing) systems, as proposed by T. Deacon, into the intrinsic higher-order constraint or multicellular organism described here.

Two relevant simplifications or approximations were applied in my analysis: gene expression levels were represented theoretically by instantaneous transcription rates, which in turn where approximated by mRNA abundance in the analysis. These steps were justified since (i) the correlation between gene expression and mRNA abundance has been clearly established as positive and significant in spite of the limitations of the techniques available [50, 51], (ii) *ctalk_non_epi* profiles remain unchanged if gene expression can be accurately expressed as a linear transformation of mRNA abundance as the control variable and, (iii) the association between *ctalk_non_epi* profiles and cell differentiation states was robust with respect to these simplifications and approximations.

If the theory advanced here is ever tested and resists falsification attempts consistently, further research will be needed to identify the cell-and-instant-specific Nanney's extracellular propagators $\vec{F_N}$ at least for each multicellular model organism, and also to identify the implications (if any) of this theory on other developmental processes such as in aging or in diseases. Also, more theoretical development will be needed to quantify the capacity and classify the content of hologenic information that emerges along with cell differentiation.

On the other hand, the critique of the epigenetic landscape approach presented in the introduction (in terms of its assumed ability to explain the self-regulatory dynamics of cell differentiation) is completely independent from a potential falsification of the theory. Even in that case, I argue that if future research keeps on elucidating the mechanisms propagating changes in gene expression to an arbitrarily high level of detail while failing to recognize that the constraints that truly regulate changes must be explicitly uncorrelated yet coupled to the constraints that propagate those changes, advances in the fundamental understanding of the evolution and self-regulatory dynamics of differentiated multicellularity will not be significant.

Falsifiability

Popper's criterion of falsifiability will be met in this monograph by providing the three following experimentally-testable predictions:

1. Under the proposed theory, the gradient $\overrightarrow{\nabla}\Phi_N$ in the extracellular space $\boldsymbol{S_E}$ such that $\left|\overrightarrow{\nabla}\Phi_N\left(\boldsymbol{D_{(1;t_D)}},\ldots,\boldsymbol{D_{(n;t_D)}},r,\theta,\phi\right)\right| \geq \boldsymbol{V_D}, (r,\theta,\phi) \in \boldsymbol{S_E}$ is a necessary condition for the emergence of cell differentiation during ontogeny. It follows directly from this proposition that *if undifferentiated stem cells or their differentiating offspring are extracted continuously from a developing embryo at the same rate they are proliferating*, then at some instant $\boldsymbol{t_D + \Delta t}$ the significant gradient (if any) of Nanney's extracellular propagators in $\boldsymbol{S_E}$ will dissipate by virtue of the Second Law of thermodynamics, reaching everywhere values under the critical value, i.e. $\left|\overrightarrow{\nabla}\Phi_N\left(\boldsymbol{D_{(1;t_D+\Delta t)}},\ldots,\boldsymbol{D_{(n;t_D+\Delta t)}},r,\theta,\phi\right)\right| < \boldsymbol{V_D}, (r,\theta,\phi) \in \boldsymbol{S_E}$. Thus, as long as cells are extracted, *the undifferentiated cells will not differentiate or the once differentiating cells will enter an artificially-induced diapause or developmental arrest.* A proper experimental control will be needed for the effect of the cell extraction technique itself, in terms of applying the technique to the embryo but extracting no cells.

2. *A significant positive correlation will be observed between the overall cell-type-wise dissimilarity of Nanney's constraints $\boldsymbol{C_N}$ in an embryo and developmental time.* In practical terms, totipotent cells can be taken from early-stage embryos and divided into separate samples, and for each later developmental time point groups of cells can be taken (ideally according to distinguishable cell types or differentiated regions) from the embryos and treated as separate samples. Then, ChIP-seq on histone H3 modifications and RNA-seq on mRNA can be used to obtain the corresponding *ctalk_non_epi* profile, which represent Nanney's constraints $\boldsymbol{C_N}$ with histone H3 modifications (adjacent to TSSs) as embodiers, for each sample. If the extraction or sectioning technique is able to generate samples for ChIP-seq/RNA-seq with high cell-type specificity and the computational analysis fails to verify the predicted correlation, the theory proposed here should be regarded as falsified.

3. *If any molecule $\boldsymbol{M}$* (i) *is specifiable as a Nanney's extracellular propagator $\boldsymbol{F_N^{\rightarrow}}$* during a certain time interval for certain cells of a differentiated multicellular species (see Corollary #1) and (ii) *is also synthesized by a unicellular (or undifferentiated multicellular) eukaryote species $\boldsymbol{U}$*, then *experiments will fail to specify $\boldsymbol{M}$ as a Nanney's extracellular propagator $\boldsymbol{F_N^{\rightarrow}}$ for the species $\boldsymbol{U}$.*

Corollaries

Corollaries, hypotheses and predictions (not involving falsifiability) that can be derived from the proposed theory include:

1. **Nanney's extracellular propagators**. The strongest prediction that follows from the theory is *the existence of Nanney's extracellular propagators* $\boldsymbol{F_N^{\rightarrow}}$ *in any differentiated multicellular species*. Since these propagators are instantaneously defined, their identification should be in the form "molecule $\boldsymbol{M}$ is specifiable as a Nanney's extracellular propagator of the species $\boldsymbol{D}$ in the cell, cell population, or cell type $\boldsymbol{T_j}$ at the developmental time point $\boldsymbol{t}$ (or the differentiation state $\boldsymbol{s}$)". This will be verified if, for instance, an experiment shows that the *ctalk_non_epi* profiles in these $\boldsymbol{T_j}$ cell or cells vary significantly when exposed to differential concentrations of $\boldsymbol{M}$ in the extracellular medium. If this is the case, it is also predictable that $\boldsymbol{M}$ will be synthesized by the cells *in vivo* at a relatively constant rate (at least as long as $\boldsymbol{M}$ is specifiable as $\boldsymbol{F_N^{\rightarrow}}$). Importantly, there is no principle in this theory precluding a first messenger molecule $\boldsymbol{M}$ known to elicit transcriptional-rate changes (e.g. a well-known morphogen) from being also specifiable as a Nanney's extracellular propagator $\boldsymbol{F_W^{\rightarrow}}$. In other words, rather than the existence of a previously undescribed molecule, what will be verified is the ability of some secreted molecules to elicit changes in Nanney's constraints $\boldsymbol{C_N}$ in the cells' nuclei. One such example would be eliciting changes in histone H3 crosstalk in TSS-adjacent genomic regions irrespectively of what the transcriptional rates are. Note: although the existence of these Nanney's extracellular propagators is a very strong and verifiable prediction, it was not included in the falsification section because it is not falsifiable in a strict epistemological sense.

2. **Surface-to-volume ratio and the evolution and development of the extracellular matrix**. An important relationship between cell surface-to-volume ratio and the evolution of differentiated multicellularity was proposed earlier (Part X-Evolution), in particular between the neuron's high surface-to-volume ratio and the evolution of its function. Under the predicted relationship between regenerative capacity and surface-to-volume ratio (see Part X-Ontogeny) neuron-shaped cells are expected to be the most difficult to regenerate. This is the developmental price to pay for a higher-order, dynamically faster form of multicellular *self* that neurons make possible. On the other hand, glial cells (companions of neurons in the nervous tissue) have a smaller surface-to-volume ratio than neurons so they would support them by constraining to some extent the diffusion flux of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ in the neurons "effective" extracellular space[12]. Glial cells with the smallest surface-to-volume ratio are ependymal cells, which have been found able to serve as neural stem cells [52]. Because this analysis is based on constraints and not on their specific molecular embodiments, the logic of the neurons and glial cells example can be extended to the evolution and development of the extracellular matrix in general. That is, the extracellular matrix was not only shaped by natural selection making it provide the cells structural and biochemical support but also developmental support, understood as fine-tuned differential constraints to the diffusion flux of Nanney's extracellular propagators in $\boldsymbol{S_E}$. Moreover, the evolution of this developmental support

[12] Understood in this case as the neuroglia plus the neural extracellular matrix.

probably preceded the evolution of all other types of support, given the critical role of the $\boldsymbol{F_N^{\rightarrow}}$ gradients in the emergence and preservation of the multicellular *telos*.

3. **Natural developmental arrests or diapauses**. The account for natural diapauses in this theory follows directly from the description in Part X-Ontogeny. Such diapauses occur in arthropods [53] and some species of killifish (Cyprinodontiformes) [54]. Natural diapauses are a metastable equilibrium state characterized by (i) the dissipation of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ in $\boldsymbol{S_E}$ under certain critical values or (ii) the inability of these gradients to constrain Waddington's embodiers $\boldsymbol{F_W}$ in the cells' nuclei because the critical gene products for protein channels/carriers or signal transducers are non-functional or not expressed. For example, if in some organism the function of the gene products critical for the facilitated diffusion/signal transduction of the current $\boldsymbol{F_N^{\rightarrow}}$ is temperature-dependent, then at that time development will enter a diapause given certain thermal conditions and resume when those conditions are lost.

4. $\boldsymbol{F_N^{\rightarrow}}$ **gradients and tissue regeneration**. Whereas the scope of the theory is the dynamics of cell differentiation and the evolution of differentiated multicellularity, it may provide some hints about other developmental processes such as tissue regeneration after extrinsic damage. For instance, I hypothesize that an important constraint driving the regenerative response to wounds is the gradient $\left|\overrightarrow{\nabla}\Phi_N\left(\boldsymbol{D}_{(1;t_{\text{wound}})},\ldots,\boldsymbol{D}_{(n;t_{\text{wound}})},r,\theta,\phi\right)\right| \gg \left|\overrightarrow{\nabla}\Phi_N\left(\boldsymbol{D}_{(1;t_{\text{wound}}-\Delta t)},\ldots,\boldsymbol{D}_{(n;t_{\text{wound}}-\Delta t)},r,\theta,\phi\right)\right|, (r,\theta,\phi) \in \boldsymbol{S_E}$ generated by the wound itself. This drive occurs because a wound creates an immediate, significant gradient at its edges. Related evidence has been found already as extracellular H_2O_2 gradients mediating wound detection in zebrafish [55]. If relevant variables (such as $\boldsymbol{F_N^{\rightarrow}}$ diffusivity in the extracellular space $\boldsymbol{S_E}$, see Corollary #2) prevent this gradient from dissipating quickly, it should contribute to a developmental regenerative response as it dissipates gradually. If different tissues of the same multicellular individual are compared, a significant negative correlation should be observable between the regenerative capacity after injury in a tissue and the average cell surface-to-volume ratio in that tissue, once controlling for average cell characteristic length.

5. **Effects of microgravity on development**. In the last few decades a number of abnormal effects of microgravity on development-related phenomena have been described, including for mammal tissue culture [56], plant growth [57], human gene expression [58], cytoskeleton organization and general embryo development ([59] and references therein). A general explanation proposed for these effects is that microgravity introduces a significant degree of mechanical perturbation on critical structures for cells and tissues. These perturbed structures as a whole would be the "gravity sensors". Without dismissing these "gravity sensors" as relevant, I suggest that a key perturbation on development elicitable by microgravity is a significant alteration of the instantaneous $\boldsymbol{F_N^{\rightarrow}}$ distribution in the extracellular space $\boldsymbol{S_E}$. This could be explained in turn by changes in the diffusion dynamics as evidence for changes in the diffusion of miscible fluids suggest [60], and/or a significant density difference between the extracellular space $\boldsymbol{S_E}$ and the cells.

6. **Why plant seeds need water**. It is a well-known fact that plant seeds only need certain initial water intake to be released from dormancy and begin to germinate with no further extrinsic support. Whereas this specific requirement of water has been associated to embryo expansion and metabolic activation of the seeds [61, 62], I submit that it is also associated to the fundamental need for a proper medium in S_E where the critical $\overrightarrow{F_N}$ gradients can emerge. These gradients would be in turn required for the intrinsic regulation of the asymmetric divisions already shown critical for cell differentiation in plants [63].

Concluding remarks

Here, I showed that scientifically tenable, instantaneous teleology in nature can emerge only from lower-order systems that are *explicitly uncorrelated* with respect to each other in terms of their dynamics. Furthermore, the only way such requisite can be fulfilled is that an intrinsic higher-order *constraint* emerges from the synergistic coupling of lower-order *constraint* generating systems, as first argued by T. Deacon. Whereas this thermodynamically spontaneous, intrinsic constraint—the logically-consistent *telos*—is dependent on molecular substrates embodying it at any instant, these substrates can be added, replaced or even dispensed with at any instant as long as the *telos* is preserved. For all these reasons, the differentiated multicellular organism described in this theory is no mechanism or machine of any type as mechanisms and machines are definable by the explicit correlation between their component dynamics. Thus, the emergence of differentiated multicellularity throughout evolution and in every successful ontogenetic process has been—and still is—the emergence of unprecedented, constraint-based, thermodynamic *selves* in the natural world; *selves* which no mechanism or machine could ever be.

Materials and Methods

Data collection

The genomic coordinates of all annotated RefSeq TSSs for the hg19 (*Homo sapiens*), mm9 (*Mus musculus*), and dm3 (*Drosophila melanogaster*) assemblies were downloaded from the UCSC database. Publicly available tandem datafiles of ChIP-seq[13] on histone H3 modifications and RNA-seq[14] for each analyzed cell sample in each species were downloaded from the ENCODE, modENCODE or NCBI's SRA databases [64, 65, 66, 67, 68, 69, 70].

The criteria for selecting cell type/cell sample datasets in each species was (i) excluding those associated to abnormal karyotypes and (ii) among the remaining datasets, choosing the group that maximizes the number of specific histone H3 modifications shared. Under these criteria, the comprised cell type/sample datasets in this work were thus:

H. sapiens 6 cell types: HSMM (skeletal muscle myoblasts), HUVEC (umbilical vein endothelial cells), NHEK (epidermal keratinocytes), GM12878 (B-lymphoblastoids), NHLF (lung fibroblasts) and H1-hESC (embryonic stem cells).
9 histone H3 modifications: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, and H3K79me2.

M. musculus 5 cell types: 8-weeks-adult heart, 8-weeks-adult liver, E14-day0 (embryonic stem cells after zero days of differentiation), E14-day4 (embryonic stem cells after four days of differentiation), and E14-day6 (embryonic stem cells after six days of differentiation).
5 histone H3 modifications: H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3.

D. melanogaster 9 cell samples: 0-4h embryos, 4-8h embryos, 8-12h embryos, 12-16h embryos, 16-20h embryos, 20-24h embryos, L1 larvae, L2 larvae, and pupae.
6 histone H3 modifications: H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, and H3K27me3.

See Supplementary Information for the datafile lists in detail.

[13] Comprising 1×36 bp, 1×50 bp, and 1×75 bp reads, depending on the data series (details available via GEO accession codes listed in Supplementary Information).

[14] Comprising 1×36 bp, 1×100 bp, and 2×75 bp reads, depending on the data series (details available via GEO accession codes listed in Supplementary Information).

ChIP-seq read profiles and normalization

The first steps in the EFilter algorithm by Kumar *et al.*—which predicts mRNA levels in log-FPKM (fragments per transcript kilobase per million fragments mapped) with high accuracy (R~0.9) [21]—were used to generate ChIP-seq read signal profiles for the histone H3 modifications data. Namely, (i) dividing the genomic region from 2 kbp upstream to 4 kbp downstream of each TSS into 30 200-bp-long bins, in each of which ChIP-seq reads were later counted; (ii) dividing the read count signal for each bin by its corresponding control (Input/IgG) read density to minimize artifactual peaks; (iii) estimating this control read density within a 1-kbp window centered on each bin, if the 1-kbp window contained at least 20 reads. Otherwise, a 5-kbp window, or else a 10-kbp window was used if the control reads were less than 20. When the 10-kbp length was insufficient, a pseudo-count value of 20 reads per 10kbp was set as the control read density. This implies that the denominator (i.e. control read density) is at least 0.4 reads per bin. When replicates were available, the measure of central tendency used was the median of the replicate read count values.

ChIP-seq read count processing

When the original format was SRA, each datafile was pre-processed with standard tools in the pipeline

```
fastq-dump → bwa aln [genome.fa]→ bwa samse → samtools view -bS -F 4
→ samtools sort → samtools index
```

to generate its associated BAM and BAI files. Otherwise, the tool

```
bedtools multicov -bams [file.bam] -bed [bins_and_controlwindows.bed]
```

was applied (excluding failed-QC reads and duplicate reads by default) directly on the original BAM[15] file to generate the corresponding read count file in BED format.

RNA-seq data processing

The processed data were mRNA abundances in FPKM at RefSeq TSSs. When the original format was GTF (containing already FPKM values, as in the selected ENCODE RNA-seq datafiles for *H. sapiens*), those values were used directly in the analysis. When the original format was SAM, each datafile was pre-processed by first sorting it to generate then a BAM file using `samtools view -bS`. If otherwise the original format was BAM, mRNA levels at RefSeq TSSs were then calculated with FPKM as unit using *Cufflinks* [71] directly on the original file with the following options:

[15] The BAI file is required implicitly.

```
-GTF-guide <reference_annotation.(gtf/gff)>
-frag-bias-correct <genome.fa>
-multi-read-correct
```

When the same TSS (i.e. same genomic coordinate and strand) displayed more than one identified transcript in the *Cufflinks* output, the respective FPKM values were added. Also, when replicates were available the measure of central tendency used was the median of the replicate FPKM values.

Preparation of data input tables

For each of the three species, all TSS_{def}—defined as those TSSs with measured mRNA abundance (i.e. FPKM > 0) in all cell types/cell samples—were determined. The number of TSS_{def} found for each species were $N_{\text{TSS}_{\text{def}}}$(***Homo sapiens***) = 14,742; $N_{\text{TSS}_{\text{def}}}$(***Mus musculus***) = 16,021; and $N_{\text{TSS}_{\text{def}}}$(***Drosophila melanogaster***) = 11,632. Then, for each cell type/cell sample, 30 genomic bins were defined and denoted by the distance (in bp) between their 5′-end and their respective TSS_{def} genomic coordinate: “−2000”, “−1800”, “−1600”, “−1400”, “−1200”, “−1000”, “−800”, “−600”, “−400”, “−200”, “0” (TSS_{def} or ‘+1’), “200”, “400”, “600”, “800”, “1000”, “1200”, “1400”, “1600”, “1800”, “2000”, “2200”, “2400”, “2600”, “2800”, “3000”, “3200”, “3400”, “3600”, and “3800”. Then, for each cell type/cell sample, a ChIP-seq read signal was computed for all bins in all TSS_{def} genomic regions (e.g. in the “−2000” bin of the ***Homo sapiens*** TSS with RefSeq ID: NM_001127328, H3K27ac_−2000 = 4.68 in H1-hESC stem cells). Data input tables, with n_m being the number of histone H3 modifications comprised, were generated following this structure of rows and columns[16]:

	H3[1]_−2000	…	H3[n_m]_−2000	⋯	H3[1]_3800	…	H3[n_m]_3,800	FPKM
1								
⋮								
$N_{\text{TSS}_{\text{def}}}$								

The tables were written then to these data files:

H. sapiens: `Hs_Gm12878.dat, Hs_H1hesc.dat, Hs_Hsmm.dat, Hs_Huvec.dat, Hs_Nhek.dat, Hs_Nhlf.dat`

M. musculus: `Mm_Heart.dat, Mm_Liver.dat, Mm_E14-d0.dat, Mm_E14-d4.dat, Mm_E14-d6.dat`

D. melanogaster: `Dm_E0-4.dat, Dm_E4-8.dat, Dm_E8-12.dat, Dm_E12-16.dat, Dm_E16-20.dat, Dm_E20-24.dat, Dm_L1.dat, Dm_L2.dat, Dm_Pupae.dat`

[16] For reference, additional columns were appended in the generated .dat files after the FPKM column with the chromosome, position, strand and RefSeq ID of each TSS_{def}.

Computation of *ctalk_non_epi* profiles

If the variables X_j (representing the signal for histone H3 modification X in the genomic bin $j \in \{\text{“}-2000\text{”},\ldots,\text{“}3800\text{”}\}$), Y_k (representing the signal for histone H3 modification Y in the genomic bin $k \in \{\text{“}-2000\text{”},\ldots,\text{“}3800\text{”}\}$) and Z (representing FPKM values) are random variables, then the covariance of X_j and Y_k can be decomposed directly in terms of their linear relationship with Z as the sum

$$\mathrm{Cov}\left(X_j, Y_k\right) = \underbrace{\frac{\mathrm{Cov}\left(X_j, Z\right)\mathrm{Cov}\left(Y_k, Z\right)}{\mathrm{Var}\left(Z\right)}}_{\substack{\text{covariance of } X_j \text{ and } Y_k \\ \text{resulting from their} \\ \text{linear relationship with } Z}} + \underbrace{\mathrm{Cov}\left(X_j, Y_k \middle| Z\right)}_{\substack{\text{covariance of } X_j \text{ and } Y_k \\ \text{orthogonal to } Z}}, \tag{1}$$

where the second summand $\mathrm{Cov}(X_j, Y_k|Z)$ is the partial covariance between X_j and Y_k given Z. It is easy to see that $\mathrm{Cov}(X_j, Y_k|Z)$ is a local approximation of Nanney's constraints $\boldsymbol{C_N}$ on histone H3 modifications, as anticipated in the preliminary theoretical definitions[17]. To make the *ctalk_non_epi* profiles comparable however, $\mathrm{Cov}(X_j, Y_k|Z)$ values have to be normalized[18] by the standard deviations of the residuals of X_j and Y_k with respect to Z. In other words, the partial correlation $\mathrm{Cor}(X_j, Y_k|Z)$ values were needed. Nevertheless, a correlation value does not have a straightforward interpretation, whereas its square—typically known as *coefficient of determination*, *effect size of the correlation*, or simply r^2—does: it represents the relative (i.e. fraction of) variance of one random variable explained by the other. For this reason, $\mathrm{Cor}(X_j, Y_k|Z)^2$ was used to represent the strength of the association, and then multiplied by the sign of the correlation to represent the direction of the association. Thus, after $\log_2$-transforming the X_j, Y_k and Z data, each pairwise combination of bin-specific histone H3 modifications $\{X_j, Y_k\}$ contributed with the value

$$ctalk\_non\_epi(X_j, Y_k) = \underbrace{\mathrm{sgn}\left(\mathrm{Cor}(X_j, Y_k|Z)\right)}_{\substack{\text{partial correlation} \\ \text{sign} \in \{-1, 1\}}}\underbrace{\left(\mathrm{Cor}(X_j, Y_k|Z)\right)^2}_{\substack{\text{partial correlation} \\ \text{strength} \in [-1, 1]}}. \tag{2}$$

This implies that for each pairwise combination of histone H3 modifications $\{X, Y\}$, there are 30 (bins for X) $\times$ 30 (bins for Y) = 900 (bin-combination-specific *ctalk_non_epi* values). To increase the robustness of the analysis against the departures of the actual nucleosome distributions from the 30×200-bp bins model, the values were then sorted in descending order and placed in a 900-tuple.

[17] A straightforward corollary is that Waddington's constraints $\boldsymbol{C_W}$ can in turn be approximated locally by $\frac{\mathrm{Cov}\left(X_j, Z\right)\mathrm{Cov}\left(Y_k, Z\right)}{\mathrm{Var}\left(Z\right)}$.

[18] At the cost of losing the sum decomposition property, which was used here for explanatory purposes.

For a cell type/cell sample from a species with data for n_m histone H3 modifications, e.g. $n_m(\textit{Mus musculus}) = 5$, the length of the final *ctalk_non_epi* profile comprising all possible $\{X, Y\}$ combinations would be ${}^{n_m}C_2 \times 900$. However, a final data filtering was performed.

The justification for this additional filtering was that some pairwise partial correlation values were expected a priori to be strong and significant, which was later confirmed. Namely, (i) those involving the same histone H3 modification in the same amino acid residue (e.g. $\mathrm{Cor}(\text{H3K9ac\_}{-200}, \text{H3K9ac\_}{-400}|\mathrm{FPKM}) > 0$; $\mathrm{Cor}(\text{H3K4me3\_}{-200}, \text{H3K4me3\_}{-200}|\mathrm{FPKM}) = 1$), (ii) those involving a different type of histone H3 modification in the same amino acid residue (e.g. $\mathrm{Cor}(\text{H3K27ac\_}{-800}, \text{H3K27me3\_}{-600}|\mathrm{FPKM}) < 0$), and (iii) those involving the same type of histone H3 modification in the same amino acid residue (e.g. $\mathrm{Cor}(\text{H3K4me2\_}{-400}, \text{H3K4me3\_}{-400}|\mathrm{FPKM}) > 0$) in part because ChIP-antibody cross reactivity has been shown able to introduce artifacts on the accurate assessment of some histone-crosstalk associations [22, 23]. For these reasons, in each species all pairwise combinations of histone H3 modifications involving the same amino acid residue were then identified as "trivial" and excluded from the *ctalk_non_epi* profiles construction. E.g., since for ***Mus musculus*** the comprised histone modifications were H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3 ($n_m = 5$), the pairwise combinations H3K4me1–H3K4me3 and H3K27ac–H3K27me3 were filtered out. Therefore, the length of the ***Mus musculus*** *ctalk_non_epi* profiles was $({}^5C_2 - 2) \times 900 = 7{,}200$.

Statistical significance assessment

The statistical significance of the partial correlation $\mathrm{Cor}(X_j, Y_k|Z)$ values, necessary for constructing the *ctalk_non_epi* profiles, was estimated using Fisher's z-transformation [72]. Under the null hypothesis $\mathrm{Cor}(X_j, Y_k|Z) = 0$ the statistic $z = \sqrt{N_{\mathrm{TSS_{def}}} - |Z| - 3}\ \frac{1}{2} \ln\left(\frac{1+\mathrm{Cor}(X_j, Y_k|Z)}{1-\mathrm{Cor}(X_j, Y_k|Z)}\right)$, where $N_{\mathrm{TSS_{def}}}$ is the sample size and $|Z| = 1$ (i.e. one control variable), follows asymptotically a $N(0,1)$ distribution. The p-values can be then computed easily using the $N(0,1)$ probability function.

Multiple comparisons correction of the p-values associated to each *ctalk_non_epi* profile was performed using the Benjamini-Yekutieli method [73]. The parameter used was the number of all possible[19] comparisons: ${}^{(n_m \times 30)}C_2$. From the resulting q-values associated to each *ctalk_non_epi* profile an empirical cumulative distribution was obtained, which in turn was used to compute a threshold t. The value of t was optimized to be the maximum value such that within the q-values smaller than t is expected less than 1 false-positive partial correlation. Consequently, if q-value$[i] \geq t$ then the associated partial correlation value was identified as not significant (i.e. zero) in the respective *ctalk_non_epi* profile.

[19] Before excluding "trivial" pairwise combinations of histone H3 modifications, to further increase the conservativeness of the correction.

Unsupervised hierarchical clustering of *ctalk_non_epi* and mRNA abundance profiles

The goal of this step was to evaluate the significant *ctalk_non_epi*-profile clusters—if any—in the phenograms (i.e. "phenotypic similarity dendrograms") obtained from unsupervised hierarchical clustering analyses (unsupervised HCA). For each species, the analyses were conducted on (i) the *ctalk_non_epi* profiles of each cell type/sample (**Figure 2A, 2C, and 2E**) and (ii) the $\log_2$-transformed FPKM profiles (i.e mRNA abundance) of each cell type/sample (**Figure 2B, 2D, and 2F**). Important to the HCA technique is the choice of a metric (for determining the distance between any two profiles) and a cluster-linkage method (for determining the distance between any two clusters).

Different ChIP-seq antibodies display differential binding affinities (with respect to different epitopes or even the same epitope, depending on the manufacturer) that are intrinsic and irrespective to the biological phenomenon of interest. For this reason, comparing directly the strengths (i.e. magnitudes) in the *ctalk_non_epi* profiles (e.g. using Euclidean distance as metric) is to introduce significant biases in the analysis. In contrast, the "correlation distance" metric—customarily used for comparing gene expression profiles—defined between any two profiles $pro[i], pro[j]$ as

$$d_r(pro[i], pro[j]) = 1 - \mathrm{Cor}(pro[i], pro[j]) \tag{3}$$

compares instead the "shape" of the profiles[20], hence it was the metric used here. On the other hand, the cluster-linkage method chosen was the "average" method or UPGMA (Unweighted Pair Group Method with Arithmetic Mean) in which the distance $D(A, B)$ between any clusters A and B is defined as

$$D(A, B) = \frac{1}{|A||B|} \sum_{\substack{pro[k] \in A \\ pro[l] \in B}} d_r(pro[k], pro[l]), \tag{4}$$

that is, the mean of all distances $d_r(pro[k], pro[l])$ such that $pro[k] \in A$ and $pro[l] \in B$ (this method was chosen because it has been shown to yield the highest cophenetic correlation values when using the "correlation distance" metric [74]). Cluster statistical significance was assessed as *au* (approximately unbiased) and *bp* (bootstrap probability) significance scores by nonparametric bootstrap resampling using the *Pvclust* [29] add-on package for the *R* software [75]. The number of bootstrap replicates in each analysis was 10,000.

[20] As a consequence of what was highlighted previously, the "correlation distance" metric is also invariant under linear transformations of the profiles.

Suitability of FPKM as unit of mRNA abundance

Previous research has pinpointed that FPKM may not always be an adequate unit of transcript abundance in differential expression studies. It was shown that, if transcript size distribution varies significantly among the samples, FPKM/RPKM[21] will introduce biases. For this reason another abundance unit TPM (transcripts per million)—which is a linear transformation of the FPKM value for each sample—was proposed to overcome the limitation [76]. However, this issue was not a problem for this study.

This is because partial correlation, used to construct the *ctalk_non_epi* profiles later subject to HCA, is invariant under linear transformations of the control variable Z (i.e. $\mathrm{Cor}(X_j, Y_k|Z) = \mathrm{Cor}(X_j, Y_k|aZ+b)$ for any two scalars $\{a, b\}$). Importantly, this property also implies that *ctalk_non_epi* profiles are controlling not only for mRNA abundance but also for any other biological variable displaying a strong linear relationship with mRNA abundance (e.g. chromatin accessibility represented by DNase I hypersensitivity, as shown in [22]). Similarly, the unsupervised hierarchical clustering of mRNA abundance profiles is invariant under linear transformations of the profiles, since $\mathrm{Cor}(Z_i, Z_j) = \mathrm{Cor}(aZ_i + b, cZ_j + d)$ provided $ac > 0$.

[21] Reads per transcript kilobase per million fragments mapped.

Acknowledgements

I wish to thank the following people:

- John Tyler Dodge, horn soloist at the *Orquesta Filarmónica de Santiago*, for reviewing most of the English in the first complete draft of this monograph and his valuable questions, which pushed me to the limit of my abilities in the purpose of making all the content self-explanatory.
- Kenneth M. Weiss and José M. Neto for reviewing this manuscript and their valuable questions.
- Alejandro Maass for his interest in this work and his valuable questions.
- Miguel Allende, director of the FONDAP Center for Genome Regulation (see details in the institutional acknowledgements below).
- My anonymous colleagues who reviewed the grant proposal on behalf of FONDECYT.

Also, I wish to thank the following institutions:

- The National Fund for Scientific and Technological Development (FONDECYT, Chile) for the postdoctoral grant (see details in Funding).
- Universidad Andrés Bello and its Faculty of Biological Sciences for sponsoring my postdoctoral grant proposal to FONDECYT.
- The FONDAP Center for Genome Regulation (CGR, Chile) for generously granting me a workplace for more than a year and giving me the opportunity to share some preliminary results of this work with other colleagues at the CGR.
- The National Laboratory for High Performance Computing (NLHPC, Chile) for providing me with a free academic account, which helped me carry out efficiently most of the computational analyses described in this monograph.
- The *Math*^{omics} Lab (Chile), for kindly helping me with the setup of my NLHPC account.

Additional information

No institution (including the funder) or person other than the author had any role in study conception, design, publicly-available data collection, computational analysis, theory development, monograph writing, or the decision to submit this preprint to bioRxiv.

Copyright

Funding

Funder	Grant reference number	Author
National Fund for Scientific and Technological Development (FONDECYT)	3140328	Felipe A. Veloso

References

[1] Slack JMW (2002) Timeline: Conrad Hal Waddington: the last renaissance biologist? Nat Rev Genet 3: 889–895. doi: 10.1038/nrg933.

[2] Waddington CH (1957) The strategy of the genes: a discussion of some aspects of theoretical biology. London: Allen & Unwin.

[3] Wolffe AP (1999) Epigenetics: Regulation through repression. Science 286: 481–486. doi: 10.1126/science.286.5439.481.

[4] Bonasio R, Tu S, Reinberg D (2010) Molecular signals of epigenetic states. Science 330: 612-6. doi: 10.1126/science.1191078.

[5] Kamakura M (2011) Royalactin induces queen differentiation in honeybees. Nature 473: 478–483. doi: 10.1038/nature10093.

[6] Fraser P (2010). Defining epigenetics. Interviews by G. Riddihough. Science [Video podcast] 00:05:34-00:05:47. URL http://videolab.sciencemag.org/featured/650920373001/1.

[7] Orphanides G, Reinberg D (2002) A unified theory of gene expression. Cell 108: 439–451. doi: 10.1016/s0092-8674(02)00655-4.

[8] Li G, Reinberg D (2011) Chromatin higher-order structures and gene regulation. Current Opinion in Genetics & Development 21: 175–186. doi: 10.1016/j.gde.2011.01.022.

[9] Cope NF, Fraser P, Eskiw CH (2010) The yin and yang of chromatin spatial organization. Genome Biol 11: 204. doi: 10.1186/gb-2010-11-3-204.

[10] Ralston A, Shaw K (2008) Gene expression regulates cell differentiation. Nat Educ 1: 127. URL http://www.nature.com/scitable/topicpage/gene-expression-regulates-cell-differentiation-931.

[11] Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A (2009) An operational definition of epigenetics. Genes & Development 23: 781–783. doi: 10.1101/gad.1787609.

[12] Reinberg D (2010). Defining epigenetics. Interviews by G. Riddihough. Science [Video podcast] 00:01:25-00:01:35. URL http://videolab.sciencemag.org/featured/650920373001/1.

[13] Arnone MI, Davidson EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. Development 124: 1851-64.

[14] Maduro MF (2010) Cell fate specification in the c. elegans embryo. Dev Dyn 239: 1315-29. doi: 10.1002/dvdy.22233.

[15] Warner DA, Shine R (2008) The adaptive significance of temperature-dependent sex determination in a reptile. Nature 451: 566–568. doi: 10.1038/nature06519.

[16] Power ML, Schulkin J (2013) Maternal regulation of offspring development in mammals is an ancient adaptation tied to lactation. Applied & Translational Genomics 2: 55–63. doi: 10.1016/j.atg.2013.06.001.

[17] Ladewig J, Koch P, Brüstle O (2013) Leveling waddington: the emergence of direct programming and the loss of cell fate hierarchies. Nature Reviews Molecular Cell Biology 14: 225–236. doi: 10.1038/nrm3543.

[18] Nanney DL (1958) Epigenetic control systems. Proceedings of the National Academy of Sciences 44: 712–717. doi: 10.1073/pnas.44.7.712.

[19] Huang S (2012) The molecular and mathematical basis of waddington's epigenetic landscape: a framework for post-darwinian biology? Bioessays 34: 149-57. doi: 10.1002/bies.201100031.

[20] Losick R, Desplan C (2008) Stochasticity and cell fate. science 320: 65–68.

[21] Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, et al. (2013) Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotechnol 31: 615–622. doi: 10.1038/nbt.2596.

[22] Lasserre J, Chung HR, Vingron M (2013) Finding associations among histone modifications using sparse partial correlation networks. PLoS Comput Biol 9: e1003168. doi: 10.1371/journal.pcbi.1003168.

[23] Peach SE, Rudomin EL, Udeshi ND, Carr SA, Jaffe JD (2012) Quantitative assessment of chromatin immunoprecipitation grade antibodies directed against histone modifications reveals patterns of co-occurring marks on histone protein molecules. Mol Cell Proteomics 11: 128-37. doi: 10.1074/mcp.M111.015941.

[24] Schwammle V, Aspalter CM, Sidoli S, Jensen ON (2014) Large scale analysis of co-existing post-translational modifications in histone tails reveals global fine structure of cross-talk. Molecular & Cellular Proteomics 13: 1855–1865. doi: 10.1074/mcp.o113.036335.

[25] Zheng Y, Sweet SMM, Popovic R, Martinez-Garcia E, Tipton JD, et al. (2012) Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone h3. Proc Natl Acad Sci U S A 109: 13549-54. doi: 10.1073/pnas.1205707109.

[26] White KP (1999) Microarray analysis of *Drosophila* development during metamorphosis. Science 286: 2179–2184. doi: 10.1126/science.286.5447.2179.

[27] Cantera R, Ferreiro MJ, Aransay AM, Barrio R (2014) Global gene expression shift during the transition from early neural development to late neuronal differentiation in drosophila melanogaster. PLoS ONE 9: e97703. doi: 10.1371/journal.pone.0097703.

[28] Mody M, Cao Y, Cui Z, Tay KY, Shyong A, et al. (2001) Genome-wide gene expression profiles of the developing mouse hippocampus. Proceedings of the National Academy of Sciences 98: 8862–8867. doi: 10.1073/pnas.141244998.

[29] Suzuki R, Shimodaira H (2006) Pvclust: an r package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540–1542. doi: 10.1093/bioinformatics/btl117.

[30] Chen L, Xiao S, Pang K, Zhou C, Yuan X (2014) Cell differentiation and germ-soma separation in ediacaran animal embryo-like fossils. Nature doi: 10.1038/nature13766.

[31] Meyerowitz EM (2002) Plants compared to animals: The broadest comparative study of development. Science 295: 1482–1485. doi: 10.1126/science.1066609.

[32] Nielsen C (2008) Six major steps in animal evolution: are we derived sponge larvae? Evolution & Development 10: 241–257. doi: 10.1111/j.1525-142x.2008.00231.x.

[33] Deacon TW (2012) Incomplete nature: How mind emerged from matter. New York: W.W. Norton & Co., 1st edition.

[34] Donoghue PCJ, Antcliffe JB (2010) Early life: Origins of multicellularity. Nature 466: 41–42. doi: 10.1038/466041a.

[35] Turing AM (1952) The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences 237: 37–72.

[36] van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends in Genetics 19: 479–484. doi: 10.1016/s0168-9525(03)00203-8.

[37] Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2007) DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Research 36: D88–D92. doi: 10.1093/nar/gkm964.

[38] Haeckel E (1874) Die gastraea-theorie, die phylogenetische classification des thierreichs und die homologie der keimblätter. Jenaische Zeitschrift für Naturwissenschaft 8: 1-55.

[39] Metschnikoff E (1886) Embryologische Studien an Medusen: ein beitrag zur Genealogie der primitiv-Organ. A. Hölder.

[40] Kirk DL (2005) A twelve-step program for evolving multicellularity and a division of labor. Bioessays 27: 299–310. doi: 10.1002/bies.20197.

[41] Willensdorfer M (2009) On the evolution of differentiated multicellularity. Evolution 63: 306-23. doi: 10.1111/j.1558-5646.2008.00541.x.

[42] Mikhailov KV, Konstantinova AV, Nikitin MA, Troshin PV, Rusin LY, et al. (2009) The origin of metazoa: a transition from temporal to spatial cell differentiation. Bioessays 31: 758-68. doi: 10.1002/bies.200800214.

[43] Gavrilets S (2010) Rapid transition towards the division of labor via evolution of developmental plasticity. PLoS Computational Biology 6: e1000805. doi: 10.1371/journal.pcbi.1000805.

[44] Levin TC, Greaney AJ, Wetzel L, King N (2014) The rosetteless gene controls development in the choanoflagellate s. rosetta. eLife 3. doi: 10.7554/elife.04070.

[45] Young HE, Black AC (2003) Adult stem cells. Anat Rec 276A: 75–102. doi: 10.1002/ar.a.10134.

[46] Cai S, Fu X, Sheng Z (2007) Dedifferentiation: A new approach in stem cell research. BioScience 57: 655. doi: 10.1641/b570805.

[47] Donà E, Barry JD, Valentin G, Quirin C, Khmelinskii A, et al. (2013) Directional tissue migration through a self-generated chemokine gradient. Nature 503: 285-9. doi: 10.1038/nature12635.

[48] Venkiteswaran G, Lewellis SW, Wang J, Reynolds E, Nicholson C, et al. (2013) Generation and dynamics of an endogenous, self-generated signaling gradient across a migrating tissue. Cell 155: 674–687. doi: 10.1016/j.cell.2013.09.046.

[49] Kauffman S, Clayton P (2006) On emergence, agency, and organization. Biology & Philosophy 21: 501–521. doi: 10.1007/s10539-005-9003-9.

[50] Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mrna expression levels on a genomic scale. Genome Biol 4: 117.

[51] Ning K, Fermin D, Nesvizhskii AI (2012) Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-seq gene expression data. J Proteome Res 11: 2261–2271. doi: 10.1021/pr201052x.

[52] Meletis K, Barnabé-Heider F, Carlén M, Evergren E, Tomilin N, et al. (2008) Spinal cord injury reveals multilineage differentiation of ependymal cells. Plos Biol 6: e182. doi: 10.1371/journal.pbio.0060182.

[53] Sømme L (1982) Supercooling and winter survival in terrestrial arthropods. Comparative Biochemistry and Physiology Part A: Physiology 73: 519–543. doi: 10.1016/0300-9629(82)90260-2.

[54] Murphy WJ, Collier GE (1997) A molecular phylogeny for aplocheiloid fishes (atherinomorpha, cyprinodontiformes): the role of vicariance and the origins of annualism. Molecular Biology and Evolution 14: 790–799.

[55] Niethammer P, Grabher C, Look AT, Mitchison TJ (2009) A tissue-scale gradient of hydrogen peroxide mediates rapid wound detection in zebrafish. Nature 459: 996–999. doi: 10.1038/nature08119.

[56] Unsworth BR, Lelkes PI (1998) Growing tissues in microgravity. Nat Med 4: 901–907. doi: 10.1038/nm0898-901.

[57] Correll MJ, Pyle TP, Millar KDL, Sun Y, Yao J, et al. (2013) Transcriptome analyses of arabidopsis thaliana seedlings grown in space: implications for gravity-responsive genes. Planta 238: 519–533. doi: 10.1007/s00425-013-1909-x.

[58] Hammond T, Lewis F, Goodwin T, Linnehan R, Wolf D, et al. (1999) Gene expression in space. Nature Medicine 5: 359–359. doi: 10.1038/7331.

[59] Crawford-Young SJ (2006) Effects of microgravity on cell cytoskeleton and embryogenesis. The International Journal of Developmental Biology 50: 183–191. doi: 10.1387/ijdb.052077sc.

[60] Pojman JA, Bessonov N, Volpert V, Paley MS (2007) Miscible fluids in microgravity (MFMG): A zero-upmass investigation on the international space station. Microgravity Sci Technol 19: 33–41. doi: 10.1007/bf02870987.

[61] Rajjou L, Duval M, Gallardo K, Catusse J, Bally J, et al. (2012) Seed germination and vigor. Annu Rev Plant Biol 63: 507–533. doi: 10.1146/annurev-arplant-042811-105550.

[62] Finch-Savage WE, Leubner-Metzger G (2006) Seed dormancy and the control of germination. New Phytologist 171: 501–523. doi: 10.1111/j.1469-8137.2006.01787.x.

[63] Smet ID, Beeckman T (2011) Asymmetric cell division in land plants and algae: the driving force for differentiation. Nature Reviews Molecular Cell Biology 12: 177–188. doi: 10.1038/nrm3064.

[64] Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, et al. (2009) Unlocking the secrets of the genome. Nature 459: 927–930. doi: 10.1038/459927a.

[65] Ram O, Goren A, Amit I, Shoresh N, Yosef N, et al. (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell 147: 1628–1639. doi: 10.1016/j.cell.2011.09.057.

[66] Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, et al. (2011) A cis-regulatory map of the *Drosophila* genome. Nature 471: 527–531. doi: 10.1038/nature09990.

[67] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. doi: 10.1038/nature11247.

[68] Xiao S, Xie D, Cao X, Yu P, Xing X, et al. (2012) Comparative epigenomic annotation of regulatory DNA. Cell 149: 1381–1392. doi: 10.1016/j.cell.2012.04.029.

[69] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101–108. doi: 10.1038/nature11233.

[70] Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. (2012) An encyclopedia of mouse DNA elements (mouse ENCODE). Genome Biol 13: 418. doi: 10.1186/gb-2012-13-8-418.

[71] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515. doi: 10.1038/nbt.1621.

[72] Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika : 507–521.

[73] Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of statistics : 1165–1188.

[74] Saraçli S, Doğan N, Doğan I (2013) Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of Inequalities and Applications 2013: 203. doi: 10.1186/1029-242x-2013-203.

[75] R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[76] Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 131: 281–285. doi: 10.1007/s12064-012-0162-3.

[77] Altun Z, Hall D (2002). WormAtlas. URL http://www.wormatlas.org/.

[78] Hadzi J (1963) The evolution of the Metazoa. Macmillan.

[79] Tompkins N, Li N, Girabawe C, Heymann M, Ermentrout GB, et al. (2014) Testing turing's theory of morphogenesis in chemical cells. Proceedings of the National Academy of Sciences 111: 4397–4402.

[80] Kupiec JJ (1997) A darwinian theory for the origin of cellular differentiation. Molecular and General Genetics MGG 255: 201–208. doi: 10.1007/s004380050490.

[81] Paldi A (2012) What makes the cell differentiate? Prog Biophys Mol Biol 110: 41-3. doi: 10.1016/j.pbiomolbio.2012.04.003.

[82] Kim J (1999) Making sense of emergence. Philosophical Studies 95: 3–36. doi: 10.1023/a:1004563122154.

[83] Kim J (2006) Emergence: Core ideas and issues. Synthese 151: 547–559. doi: 10.1007/s11229-006-9025-0.

Appendix

Estimation of a lower bound for the necessary cell-fate information capacity in the hermaphrodite *Caenorhabditis elegans* ontogeny

Count	Nº
Cells generated	1,090
Deaths in the process	131
Final cells	959
Cell types developed	19

(Data source: WormAtlas website [77])

	Estimated as	Nº (approx.)
Total divisions	$2^{\log_2(\text{cells\_generated}+1)} - 1$	2,179
Cell-fate divisions	$2^{\log_2(\text{cell\_types}+1)} - 1$	37
Non-cell-fate divisions	total_divisions − (cell_fate_divisions + deaths)	2,011

	Estimated as	p	$-p \log_2 p$
Cell death	deaths / total_divisions	0.060	0.244
Non-cell-fate division	non_cell_fate_divisions / total_divisions	0.923	0.107
Cell-fate division	cell_fate_divisions / total_divisions	0.017	0.1
Uncertainty per division (Sum)			0.451

	Estimated as	(bit)
Uncertainty to resolve (total)	uncertainty_per_division × total_divisions	983

Note: germ line cells were excluded from the analysis.

Problems with current views on the self-regulation of cell differentiation and the evolution of multicellularity

Since Ernst Haeckel's "gastraea theory" [38], the most plausible models aimed to explain the evolution of differentiated multicellularity are fundamentally divorced from the epigenetic landscape model assumed to explain the self-regulatory dynamics underpinning differentiated multicellularity. This is because Haeckel's account and the models built upon it rely on the gradual specialization of same-species (or even different-species [78]) cell colonies or aggregations [39, 40, 32, 41, 42, 43, 44] while the developmental process starts from a single cell (zygote) or, in other words, "from the inside out". Because differentiated multicellularity is a single phenomenon whose evolution and self-regulation have been tackled by research under such divergent approaches, the resulting explanatory account is thus insufficiently substantiated as a whole, especially considering its lack of parsimony.

Other, "non-epigenetic" hypotheses have been advanced aiming to explain the dynamics and/or informational requirements of cell-differentiation (which in turn could provide some hints on the evolution of multicellularity). One of them holds that spontaneous intercellular reaction-diffusion patterns are responsible for morphogenesis, and for cell differentiation as a consequence [35]. Although this model has been tested in terms of chemical differentiation of synthetic "cells" [79], it does not explain the critical relationship in which real differentiating/differentiated cells *serve* the individuated multicellular organism as a whole. Another hypothesis suggests that gene expression instability and stochasticity, in the context of external metabolic substrate gradients, create an intrinsic natural-selection-like mechanism able to drive the differentiation process [80]. A third "non-epigenetic" hypothesis is that cell fate decisions are the result of the characteristic coupling of gene expression and metabolism [81].

All of these accounts, however, fail to explain how traits or dynamics that supposedly account for the transition to multicellularity or to cell differentiation have fundamentally analogous counterparts in undifferentiated multicellular or unicellular eukaryotic lineages. They also fail to account for the information required by developmental decisions for information and in the transition between strictly single-cell-related content to additional multicellular-individual-related content. Further, they fail to explain the reproducible and robust self-regulatory dynamics of gene expression during cell differentiation. These approaches also cannot describe the transition between a highly complex or symbiotic cell population/aggregation and a differentiated multicellular organism, and they lack the parsimony when encompassing both the evolution and self-regulation of differentiated multicellularity. Neither are they falsifiable.

In contrast to these current hypotheses, the falsifiable theory proposed here regards the multicellular organism as a higher-order system that emerges from proliferating undifferentiated cells and *then* is subject to natural selection. The theoretical development in this work is not based on the substrate-based concept of irreducible emergence[22] (fundamentally refuted by Jaegwon Kim [82, 83]) but instead converged from the strict *explicitly-uncorrelated-constraint-dynamics* condition argued in the introduction) into what can be described as the constraint-based4[23]

[22] Understood as molecules and their realizable interactions, which define the state space in a dynamical systems model such as the epigenetic landscape.

[23] Understood as the realizable states explicitly *excluded* from realization in the system's dynamics.

concept of emergence for higher-order teleological systems, pioneered in a broader perspective by Terrence Deacon in 2011 [33]. Importantly, this formulation of emergence does not build upon the traditional concepts of telos or "final cause" but instead redefines the telos as a thermodynamically spontaneous, intrinsic constraint whose causal power is exerted at the present instant.

Supplementary Information

Homo sapiens source data of ChIP-seq on histone H3 modifications (BAM/BAI files) [65]

For downloading, the URL must be constructed by adding the following prefix to each file listed:

`ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/`

Cell type	Antibody	GEO Accession	File URL suffix
GM12878	H3K27ac	GSM733771	wgEncodeBroadHistoneGm12878H3k27acStdAlnRep1.bam.bai
GM12878	H3K27ac	GSM733771	wgEncodeBroadHistoneGm12878H3k27acStdAlnRep1.bam
GM12878	H3K27ac	GSM733771	wgEncodeBroadHistoneGm12878H3k27acStdAlnRep2.bam.bai
GM12878	H3K27ac	GSM733771	wgEncodeBroadHistoneGm12878H3k27acStdAlnRep2.bam
GM12878	H3K27me3	GSM733758	wgEncodeBroadHistoneGm12878H3k27me3StdAlnRep1.bam.bai
GM12878	H3K27me3	GSM733758	wgEncodeBroadHistoneGm12878H3k27me3StdAlnRep1.bam
GM12878	H3K27me3	GSM733758	wgEncodeBroadHistoneGm12878H3k27me3StdAlnRep2.bam.bai
GM12878	H3K27me3	GSM733758	wgEncodeBroadHistoneGm12878H3k27me3StdAlnRep2.bam
GM12878	H3K27me3	GSM733758	wgEncodeBroadHistoneGm12878H3k27me3StdAlnRep3V2.bam.bai
GM12878	H3K27me3	GSM733758	wgEncodeBroadHistoneGm12878H3k27me3StdAlnRep3V2.bam
GM12878	H3K36me3	GSM733679	wgEncodeBroadHistoneGm12878H3k36me3StdAlnRep1.bam.bai
GM12878	H3K36me3	GSM733679	wgEncodeBroadHistoneGm12878H3k36me3StdAlnRep1.bam
GM12878	H3K36me3	GSM733679	wgEncodeBroadHistoneGm12878H3k36me3StdAlnRep2.bam.bai
GM12878	H3K36me3	GSM733679	wgEncodeBroadHistoneGm12878H3k36me3StdAlnRep2.bam
GM12878	H3K4me1	GSM733772	wgEncodeBroadHistoneGm12878H3k4me1StdAlnRep2.bam.bai
GM12878	H3K4me1	GSM733772	wgEncodeBroadHistoneGm12878H3k4me1StdAlnRep2.bam
GM12878	H3K4me1	GSM733772	wgEncodeBroadHistoneGm12878H3k04me1StdAlnRep1V2.bam.bai
GM12878	H3K4me1	GSM733772	wgEncodeBroadHistoneGm12878H3k04me1StdAlnRep1V2.bam
GM12878	H3K4me2	GSM733769	wgEncodeBroadHistoneGm12878H3k4me2StdAlnRep1.bam.bai
GM12878	H3K4me2	GSM733769	wgEncodeBroadHistoneGm12878H3k4me2StdAlnRep1.bam
GM12878	H3K4me2	GSM733769	wgEncodeBroadHistoneGm12878H3k4me2StdAlnRep2.bam.bai
GM12878	H3K4me2	GSM733769	wgEncodeBroadHistoneGm12878H3k4me2StdAlnRep2.bam
GM12878	H3K4me3	GSM733708	wgEncodeBroadHistoneGm12878H3k04me3StdAlnRep2V2.bam.bai
GM12878	H3K4me3	GSM733708	wgEncodeBroadHistoneGm12878H3k04me3StdAlnRep2V2.bam
GM12878	H3K4me3	GSM733708	wgEncodeBroadHistoneGm12878H3k4me3StdAlnRep1.bam.bai
GM12878	H3K4me3	GSM733708	wgEncodeBroadHistoneGm12878H3k4me3StdAlnRep1.bam
GM12878	H3K79me2	GSM733736	wgEncodeBroadHistoneGm12878H3k79me2StdAlnRep1.bam.bai
GM12878	H3K79me2	GSM733736	wgEncodeBroadHistoneGm12878H3k79me2StdAlnRep1.bam
GM12878	H3K79me2	GSM733736	wgEncodeBroadHistoneGm12878H3k79me2StdAlnRep2.bam.bai
GM12878	H3K79me2	GSM733736	wgEncodeBroadHistoneGm12878H3k79me2StdAlnRep2.bam
GM12878	H3K9ac	GSM733677	wgEncodeBroadHistoneGm12878H3k9acStdAlnRep1.bam.bai
GM12878	H3K9ac	GSM733677	wgEncodeBroadHistoneGm12878H3k9acStdAlnRep1.bam
GM12878	H3K9ac	GSM733677	wgEncodeBroadHistoneGm12878H3k9acStdAlnRep2.bam.bai
GM12878	H3K9ac	GSM733677	wgEncodeBroadHistoneGm12878H3k9acStdAlnRep2.bam
GM12878	H3K9me3	GSM733664	wgEncodeBroadHistoneGm12878H3k9me3StdAlnRep1.bam.bai
GM12878	H3K9me3	GSM733664	wgEncodeBroadHistoneGm12878H3k9me3StdAlnRep1.bam
GM12878	H3K9me3	GSM733664	wgEncodeBroadHistoneGm12878H3k9me3StdAlnRep2.bam.bai
GM12878	H3K9me3	GSM733664	wgEncodeBroadHistoneGm12878H3k9me3StdAlnRep2.bam
GM12878	H3K9me3	GSM733664	wgEncodeBroadHistoneGm12878H3k9me3StdAlnRep3.bam.bai
GM12878	H3K9me3	GSM733664	wgEncodeBroadHistoneGm12878H3k9me3StdAlnRep3.bam

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
GM12878	Input	GSM733742	wgEncodeBroadHistoneGm12878ControlStdAlnRep1.bam.bai
GM12878	Input	GSM733742	wgEncodeBroadHistoneGm12878ControlStdAlnRep1.bam
GM12878	Input	GSM733742	wgEncodeBroadHistoneGm12878ControlStdAlnRep2.bam.bai
GM12878	Input	GSM733742	wgEncodeBroadHistoneGm12878ControlStdAlnRep2.bam
H1-hESC	H3K27ac	GSM733718	wgEncodeBroadHistoneH1hescH3k27acStdAlnRep1.bam.bai
H1-hESC	H3K27ac	GSM733718	wgEncodeBroadHistoneH1hescH3k27acStdAlnRep1.bam
H1-hESC	H3K27ac	GSM733718	wgEncodeBroadHistoneH1hescH3k27acStdAlnRep2.bam.bai
H1-hESC	H3K27ac	GSM733718	wgEncodeBroadHistoneH1hescH3k27acStdAlnRep2.bam
H1-hESC	H3K27me3	GSM733748	wgEncodeBroadHistoneH1hescH3k27me3StdAlnRep1.bam.bai
H1-hESC	H3K27me3	GSM733748	wgEncodeBroadHistoneH1hescH3k27me3StdAlnRep1.bam
H1-hESC	H3K27me3	GSM733748	wgEncodeBroadHistoneH1hescH3k27me3StdAlnRep2.bam.bai
H1-hESC	H3K27me3	GSM733748	wgEncodeBroadHistoneH1hescH3k27me3StdAlnRep2.bam
H1-hESC	H3K36me3	GSM733725	wgEncodeBroadHistoneH1hescH3k36me3StdAlnRep1.bam.bai
H1-hESC	H3K36me3	GSM733725	wgEncodeBroadHistoneH1hescH3k36me3StdAlnRep1.bam
H1-hESC	H3K36me3	GSM733725	wgEncodeBroadHistoneH1hescH3k36me3StdAlnRep2.bam.bai
H1-hESC	H3K36me3	GSM733725	wgEncodeBroadHistoneH1hescH3k36me3StdAlnRep2.bam
H1-hESC	H3K4me1	GSM733782	wgEncodeBroadHistoneH1hescH3k4me1StdAlnRep1.bam.bai
H1-hESC	H3K4me1	GSM733782	wgEncodeBroadHistoneH1hescH3k4me1StdAlnRep1.bam
H1-hESC	H3K4me1	GSM733782	wgEncodeBroadHistoneH1hescH3k4me1StdAlnRep2.bam.bai
H1-hESC	H3K4me1	GSM733782	wgEncodeBroadHistoneH1hescH3k4me1StdAlnRep2.bam
H1-hESC	H3K4me2	GSM733670	wgEncodeBroadHistoneH1hescH3k4me2StdAlnRep1.bam.bai
H1-hESC	H3K4me2	GSM733670	wgEncodeBroadHistoneH1hescH3k4me2StdAlnRep1.bam
H1-hESC	H3K4me2	GSM733670	wgEncodeBroadHistoneH1hescH3k4me2StdAlnRep2.bam.bai
H1-hESC	H3K4me2	GSM733670	wgEncodeBroadHistoneH1hescH3k4me2StdAlnRep2.bam
H1-hESC	H3K4me3	GSM733657	wgEncodeBroadHistoneH1hescH3k4me3StdAlnRep1.bam.bai
H1-hESC	H3K4me3	GSM733657	wgEncodeBroadHistoneH1hescH3k4me3StdAlnRep1.bam
H1-hESC	H3K4me3	GSM733657	wgEncodeBroadHistoneH1hescH3k4me3StdAlnRep2.bam.bai
H1-hESC	H3K4me3	GSM733657	wgEncodeBroadHistoneH1hescH3k4me3StdAlnRep2.bam
H1-hESC	H3K79me2	GSM1003547	wgEncodeBroadHistoneH1hescH3k79me2StdAlnRep1.bam.bai
H1-hESC	H3K79me2	GSM1003547	wgEncodeBroadHistoneH1hescH3k79me2StdAlnRep1.bam
H1-hESC	H3K79me2	GSM1003547	wgEncodeBroadHistoneH1hescH3k79me2StdAlnRep2.bam.bai
H1-hESC	H3K79me2	GSM1003547	wgEncodeBroadHistoneH1hescH3k79me2StdAlnRep2.bam
H1-hESC	H3K9ac	GSM733773	wgEncodeBroadHistoneH1hescH3k9acStdAlnRep1.bam.bai
H1-hESC	H3K9ac	GSM733773	wgEncodeBroadHistoneH1hescH3k9acStdAlnRep1.bam
H1-hESC	H3K9ac	GSM733773	wgEncodeBroadHistoneH1hescH3k9acStdAlnRep2.bam.bai
H1-hESC	H3K9ac	GSM733773	wgEncodeBroadHistoneH1hescH3k9acStdAlnRep2.bam
H1-hESC	H3K9me3	GSM1003585	wgEncodeBroadHistoneH1hescH3k09me3StdAlnRep1.bam.bai
H1-hESC	H3K9me3	GSM1003585	wgEncodeBroadHistoneH1hescH3k09me3StdAlnRep1.bam
H1-hESC	H3K9me3	GSM1003585	wgEncodeBroadHistoneH1hescH3k09me3StdAlnRep2.bam.bai
H1-hESC	H3K9me3	GSM1003585	wgEncodeBroadHistoneH1hescH3k09me3StdAlnRep2.bam
H1-hESC	Input	GSM733770	wgEncodeBroadHistoneH1hescControlStdAlnRep1.bam.bai
H1-hESC	Input	GSM733770	wgEncodeBroadHistoneH1hescControlStdAlnRep1.bam
H1-hESC	Input	GSM733770	wgEncodeBroadHistoneH1hescControlStdAlnRep2.bam.bai
H1-hESC	Input	GSM733770	wgEncodeBroadHistoneH1hescControlStdAlnRep2.bam
HSMM	H3K27ac	GSM733755	wgEncodeBroadHistoneHsmmH3k27acStdAlnRep1.bam.bai
HSMM	H3K27ac	GSM733755	wgEncodeBroadHistoneHsmmH3k27acStdAlnRep1.bam
HSMM	H3K27ac	GSM733755	wgEncodeBroadHistoneHsmmH3k27acStdAlnRep2.bam.bai
HSMM	H3K27ac	GSM733755	wgEncodeBroadHistoneHsmmH3k27acStdAlnRep2.bam
HSMM	H3K27me3	GSM733667	wgEncodeBroadHistoneHsmmH3k27me3StdAlnRep1.bam.bai
HSMM	H3K27me3	GSM733667	wgEncodeBroadHistoneHsmmH3k27me3StdAlnRep1.bam
HSMM	H3K27me3	GSM733667	wgEncodeBroadHistoneHsmmH3k27me3StdAlnRep2.bam.bai

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
HSMM	H3K27me3	GSM733667	wgEncodeBroadHistoneHsmmH3k27me3StdAlnRep2.bam
HSMM	H3K36me3	GSM733702	wgEncodeBroadHistoneHsmmH3k36me3StdAlnRep1.bam.bai
HSMM	H3K36me3	GSM733702	wgEncodeBroadHistoneHsmmH3k36me3StdAlnRep1.bam
HSMM	H3K36me3	GSM733702	wgEncodeBroadHistoneHsmmH3k36me3StdAlnRep2.bam.bai
HSMM	H3K36me3	GSM733702	wgEncodeBroadHistoneHsmmH3k36me3StdAlnRep2.bam
HSMM	H3K4me1	GSM733761	wgEncodeBroadHistoneHsmmH3k4me1StdAlnRep1.bam.bai
HSMM	H3K4me1	GSM733761	wgEncodeBroadHistoneHsmmH3k4me1StdAlnRep1.bam
HSMM	H3K4me1	GSM733761	wgEncodeBroadHistoneHsmmH3k4me1StdAlnRep2.bam.bai
HSMM	H3K4me1	GSM733761	wgEncodeBroadHistoneHsmmH3k4me1StdAlnRep2.bam
HSMM	H3K4me2	GSM733768	wgEncodeBroadHistoneHsmmH3k4me2StdAlnRep1.bam.bai
HSMM	H3K4me2	GSM733768	wgEncodeBroadHistoneHsmmH3k4me2StdAlnRep1.bam
HSMM	H3K4me2	GSM733768	wgEncodeBroadHistoneHsmmH3k4me2StdAlnRep2.bam.bai
HSMM	H3K4me2	GSM733768	wgEncodeBroadHistoneHsmmH3k4me2StdAlnRep2.bam
HSMM	H3K4me3	GSM733637	wgEncodeBroadHistoneHsmmH3k4me3StdAlnRep1.bam.bai
HSMM	H3K4me3	GSM733637	wgEncodeBroadHistoneHsmmH3k4me3StdAlnRep1.bam
HSMM	H3K4me3	GSM733637	wgEncodeBroadHistoneHsmmH3k4me3StdAlnRep2.bam.bai
HSMM	H3K4me3	GSM733637	wgEncodeBroadHistoneHsmmH3k4me3StdAlnRep2.bam
HSMM	H3K79me2	GSM733741	wgEncodeBroadHistoneHsmmH3k79me2StdAlnRep1.bam.bai
HSMM	H3K79me2	GSM733741	wgEncodeBroadHistoneHsmmH3k79me2StdAlnRep1.bam
HSMM	H3K79me2	GSM733741	wgEncodeBroadHistoneHsmmH3k79me2StdAlnRep2.bam.bai
HSMM	H3K79me2	GSM733741	wgEncodeBroadHistoneHsmmH3k79me2StdAlnRep2.bam
HSMM	H3K9ac	GSM733775	wgEncodeBroadHistoneHsmmH3k9acStdAlnRep1.bam.bai
HSMM	H3K9ac	GSM733775	wgEncodeBroadHistoneHsmmH3k9acStdAlnRep1.bam
HSMM	H3K9ac	GSM733775	wgEncodeBroadHistoneHsmmH3k9acStdAlnRep2.bam.bai
HSMM	H3K9ac	GSM733775	wgEncodeBroadHistoneHsmmH3k9acStdAlnRep2.bam
HSMM	H3K9me3	GSM733730	wgEncodeBroadHistoneHsmmH3k9me3StdAlnRep1.bam.bai
HSMM	H3K9me3	GSM733730	wgEncodeBroadHistoneHsmmH3k9me3StdAlnRep1.bam
HSMM	H3K9me3	GSM733730	wgEncodeBroadHistoneHsmmH3k9me3StdAlnRep2.bam.bai
HSMM	H3K9me3	GSM733730	wgEncodeBroadHistoneHsmmH3k9me3StdAlnRep2.bam
HSMM	Input	GSM733663	wgEncodeBroadHistoneHsmmControlStdAlnRep1.bam.bai
HSMM	Input	GSM733663	wgEncodeBroadHistoneHsmmControlStdAlnRep1.bam
HSMM	Input	GSM733663	wgEncodeBroadHistoneHsmmControlStdAlnRep2.bam.bai
HSMM	Input	GSM733663	wgEncodeBroadHistoneHsmmControlStdAlnRep2.bam
HUVEC	H3K27ac	GSM733691	wgEncodeBroadHistoneHuvecH3k27acStdAlnRep1.bam.bai
HUVEC	H3K27ac	GSM733691	wgEncodeBroadHistoneHuvecH3k27acStdAlnRep1.bam
HUVEC	H3K27ac	GSM733691	wgEncodeBroadHistoneHuvecH3k27acStdAlnRep2.bam.bai
HUVEC	H3K27ac	GSM733691	wgEncodeBroadHistoneHuvecH3k27acStdAlnRep2.bam
HUVEC	H3K27ac	GSM733691	wgEncodeBroadHistoneHuvecH3k27acStdAlnRep3.bam.bai
HUVEC	H3K27ac	GSM733691	wgEncodeBroadHistoneHuvecH3k27acStdAlnRep3.bam
HUVEC	H3K27me3	GSM733688	wgEncodeBroadHistoneHuvecH3k27me3StdAlnRep1.bam.bai
HUVEC	H3K27me3	GSM733688	wgEncodeBroadHistoneHuvecH3k27me3StdAlnRep1.bam
HUVEC	H3K27me3	GSM733688	wgEncodeBroadHistoneHuvecH3k27me3StdAlnRep2.bam.bai
HUVEC	H3K27me3	GSM733688	wgEncodeBroadHistoneHuvecH3k27me3StdAlnRep2.bam
HUVEC	H3K36me3	GSM733757	wgEncodeBroadHistoneHuvecH3k36me3StdAlnRep1.bam.bai
HUVEC	H3K36me3	GSM733757	wgEncodeBroadHistoneHuvecH3k36me3StdAlnRep1.bam
HUVEC	H3K36me3	GSM733757	wgEncodeBroadHistoneHuvecH3k36me3StdAlnRep2.bam.bai
HUVEC	H3K36me3	GSM733757	wgEncodeBroadHistoneHuvecH3k36me3StdAlnRep2.bam
HUVEC	H3K36me3	GSM733757	wgEncodeBroadHistoneHuvecH3k36me3StdAlnRep3.bam.bai
HUVEC	H3K36me3	GSM733757	wgEncodeBroadHistoneHuvecH3k36me3StdAlnRep3.bam
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep1.bam.bai
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep1.bam

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep2.bam.bai
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep2.bam
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep3.bam.bai
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep3.bam
HUVEC	H3K4me2	GSM733683	wgEncodeBroadHistoneHuvecH3k4me2StdAlnRep1.bam.bai
HUVEC	H3K4me2	GSM733683	wgEncodeBroadHistoneHuvecH3k4me2StdAlnRep1.bam
HUVEC	H3K4me2	GSM733683	wgEncodeBroadHistoneHuvecH3k4me2StdAlnRep2.bam.bai
HUVEC	H3K4me2	GSM733683	wgEncodeBroadHistoneHuvecH3k4me2StdAlnRep2.bam
HUVEC	H3K4me3	GSM733673	wgEncodeBroadHistoneHuvecH3k4me3StdAlnRep1.bam.bai
HUVEC	H3K4me3	GSM733673	wgEncodeBroadHistoneHuvecH3k4me3StdAlnRep1.bam
HUVEC	H3K4me3	GSM733673	wgEncodeBroadHistoneHuvecH3k4me3StdAlnRep2.bam.bai
HUVEC	H3K4me3	GSM733673	wgEncodeBroadHistoneHuvecH3k4me3StdAlnRep2.bam
HUVEC	H3K4me3	GSM733673	wgEncodeBroadHistoneHuvecH3k4me3StdAlnRep3.bam.bai
HUVEC	H3K4me3	GSM733673	wgEncodeBroadHistoneHuvecH3k4me3StdAlnRep3.bam
HUVEC	H3K79me2	GSM1003555	wgEncodeBroadHistoneHuvecH3k79me2AlnRep1.bam.bai
HUVEC	H3K79me2	GSM1003555	wgEncodeBroadHistoneHuvecH3k79me2AlnRep1.bam
HUVEC	H3K79me2	GSM1003555	wgEncodeBroadHistoneHuvecH3k79me2AlnRep2.bam.bai
HUVEC	H3K79me2	GSM1003555	wgEncodeBroadHistoneHuvecH3k79me2AlnRep2.bam
HUVEC	H3K9ac	GSM733735	wgEncodeBroadHistoneHuvecH3k9acStdAlnRep1.bam.bai
HUVEC	H3K9ac	GSM733735	wgEncodeBroadHistoneHuvecH3k9acStdAlnRep1.bam
HUVEC	H3K9ac	GSM733735	wgEncodeBroadHistoneHuvecH3k9acStdAlnRep2.bam.bai
HUVEC	H3K9ac	GSM733735	wgEncodeBroadHistoneHuvecH3k9acStdAlnRep2.bam
HUVEC	H3K9ac	GSM733735	wgEncodeBroadHistoneHuvecH3k9acStdAlnRep3.bam.bai
HUVEC	H3K9ac	GSM733735	wgEncodeBroadHistoneHuvecH3k9acStdAlnRep3.bam
HUVEC	H3K9me3	GSM1003517	wgEncodeBroadHistoneHuvecH3k09me3AlnRep1.bam.bai
HUVEC	H3K9me3	GSM1003517	wgEncodeBroadHistoneHuvecH3k09me3AlnRep1.bam
HUVEC	H3K9me3	GSM1003517	wgEncodeBroadHistoneHuvecH3k09me3AlnRep2.bam.bai
HUVEC	H3K9me3	GSM1003517	wgEncodeBroadHistoneHuvecH3k09me3AlnRep2.bam
HUVEC	Input	GSM733715	wgEncodeBroadHistoneHuvecControlStdAlnRep1.bam.bai
HUVEC	Input	GSM733715	wgEncodeBroadHistoneHuvecControlStdAlnRep1.bam
HUVEC	Input	GSM733715	wgEncodeBroadHistoneHuvecControlStdAlnRep2.bam.bai
HUVEC	Input	GSM733715	wgEncodeBroadHistoneHuvecControlStdAlnRep2.bam
HUVEC	Input	GSM733715	wgEncodeBroadHistoneHuvecControlStdAlnRep3.bam.bai
HUVEC	Input	GSM733715	wgEncodeBroadHistoneHuvecControlStdAlnRep3.bam
NHEK	H3K27ac	GSM733674	wgEncodeBroadHistoneNhekH3k27acStdAlnRep1.bam.bai
NHEK	H3K27ac	GSM733674	wgEncodeBroadHistoneNhekH3k27acStdAlnRep1.bam
NHEK	H3K27ac	GSM733674	wgEncodeBroadHistoneNhekH3k27acStdAlnRep2.bam.bai
NHEK	H3K27ac	GSM733674	wgEncodeBroadHistoneNhekH3k27acStdAlnRep2.bam
NHEK	H3K27ac	GSM733674	wgEncodeBroadHistoneNhekH3k27acStdAlnRep3.bam.bai
NHEK	H3K27ac	GSM733674	wgEncodeBroadHistoneNhekH3k27acStdAlnRep3.bam
NHEK	H3K27me3	GSM733701	wgEncodeBroadHistoneNhekH3k27me3StdAlnRep1.bam.bai
NHEK	H3K27me3	GSM733701	wgEncodeBroadHistoneNhekH3k27me3StdAlnRep1.bam
NHEK	H3K27me3	GSM733701	wgEncodeBroadHistoneNhekH3k27me3StdAlnRep2.bam.bai
NHEK	H3K27me3	GSM733701	wgEncodeBroadHistoneNhekH3k27me3StdAlnRep2.bam
NHEK	H3K27me3	GSM733701	wgEncodeBroadHistoneNhekH3k27me3StdAlnRep3.bam.bai
NHEK	H3K27me3	GSM733701	wgEncodeBroadHistoneNhekH3k27me3StdAlnRep3.bam
NHEK	H3K36me3	GSM733726	wgEncodeBroadHistoneNhekH3k36me3StdAlnRep1.bam.bai
NHEK	H3K36me3	GSM733726	wgEncodeBroadHistoneNhekH3k36me3StdAlnRep1.bam
NHEK	H3K36me3	GSM733726	wgEncodeBroadHistoneNhekH3k36me3StdAlnRep2.bam.bai
NHEK	H3K36me3	GSM733726	wgEncodeBroadHistoneNhekH3k36me3StdAlnRep2.bam
NHEK	H3K36me3	GSM733726	wgEncodeBroadHistoneNhekH3k36me3StdAlnRep3.bam.bai

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
NHEK	H3K36me3	GSM733726	wgEncodeBroadHistoneNhekH3k36me3StdAlnRep3.bam
NHEK	H3K4me1	GSM733698	wgEncodeBroadHistoneNhekH3k4me1StdAlnRep1.bam.bai
NHEK	H3K4me1	GSM733698	wgEncodeBroadHistoneNhekH3k4me1StdAlnRep1.bam
NHEK	H3K4me1	GSM733698	wgEncodeBroadHistoneNhekH3k4me1StdAlnRep2.bam.bai
NHEK	H3K4me1	GSM733698	wgEncodeBroadHistoneNhekH3k4me1StdAlnRep2.bam
NHEK	H3K4me1	GSM733698	wgEncodeBroadHistoneNhekH3k4me1StdAlnRep3.bam.bai
NHEK	H3K4me1	GSM733698	wgEncodeBroadHistoneNhekH3k4me1StdAlnRep3.bam
NHEK	H3K4me2	GSM733686	wgEncodeBroadHistoneNhekH3k4me2StdAlnRep1.bam.bai
NHEK	H3K4me2	GSM733686	wgEncodeBroadHistoneNhekH3k4me2StdAlnRep1.bam
NHEK	H3K4me2	GSM733686	wgEncodeBroadHistoneNhekH3k4me2StdAlnRep2.bam.bai
NHEK	H3K4me2	GSM733686	wgEncodeBroadHistoneNhekH3k4me2StdAlnRep2.bam
NHEK	H3K4me2	GSM733686	wgEncodeBroadHistoneNhekH3k4me2StdAlnRep3.bam.bai
NHEK	H3K4me2	GSM733686	wgEncodeBroadHistoneNhekH3k4me2StdAlnRep3.bam
NHEK	H3K4me3	GSM733720	wgEncodeBroadHistoneNhekH3k4me3StdAlnRep1.bam.bai
NHEK	H3K4me3	GSM733720	wgEncodeBroadHistoneNhekH3k4me3StdAlnRep1.bam
NHEK	H3K4me3	GSM733720	wgEncodeBroadHistoneNhekH3k4me3StdAlnRep2.bam.bai
NHEK	H3K4me3	GSM733720	wgEncodeBroadHistoneNhekH3k4me3StdAlnRep2.bam
NHEK	H3K4me3	GSM733720	wgEncodeBroadHistoneNhekH3k4me3StdAlnRep3.bam.bai
NHEK	H3K4me3	GSM733720	wgEncodeBroadHistoneNhekH3k4me3StdAlnRep3.bam
NHEK	H3K79me2	GSM1003527	wgEncodeBroadHistoneNhekH3k79me2AlnRep1.bam.bai
NHEK	H3K79me2	GSM1003527	wgEncodeBroadHistoneNhekH3k79me2AlnRep1.bam
NHEK	H3K79me2	GSM1003527	wgEncodeBroadHistoneNhekH3k79me2AlnRep2.bam.bai
NHEK	H3K79me2	GSM1003527	wgEncodeBroadHistoneNhekH3k79me2AlnRep2.bam
NHEK	H3K9ac	GSM733665	wgEncodeBroadHistoneNhekH3k9acStdAlnRep1.bam.bai
NHEK	H3K9ac	GSM733665	wgEncodeBroadHistoneNhekH3k9acStdAlnRep1.bam
NHEK	H3K9ac	GSM733665	wgEncodeBroadHistoneNhekH3k9acStdAlnRep2.bam.bai
NHEK	H3K9ac	GSM733665	wgEncodeBroadHistoneNhekH3k9acStdAlnRep2.bam
NHEK	H3K9ac	GSM733665	wgEncodeBroadHistoneNhekH3k9acStdAlnRep3.bam.bai
NHEK	H3K9ac	GSM733665	wgEncodeBroadHistoneNhekH3k9acStdAlnRep3.bam
NHEK	H3K9me3	GSM1003528	wgEncodeBroadHistoneNhekH3k09me3AlnRep1.bam.bai
NHEK	H3K9me3	GSM1003528	wgEncodeBroadHistoneNhekH3k09me3AlnRep1.bam
NHEK	H3K9me3	GSM1003528	wgEncodeBroadHistoneNhekH3k09me3AlnRep2.bam.bai
NHEK	H3K9me3	GSM1003528	wgEncodeBroadHistoneNhekH3k09me3AlnRep2.bam
NHEK	Input	GSM733740	wgEncodeBroadHistoneNhekControlStdAlnRep1.bam.bai
NHEK	Input	GSM733740	wgEncodeBroadHistoneNhekControlStdAlnRep1.bam
NHEK	Input	GSM733740	wgEncodeBroadHistoneNhekControlStdAlnRep2.bam.bai
NHEK	Input	GSM733740	wgEncodeBroadHistoneNhekControlStdAlnRep2.bam
NHLF	H3K27ac	GSM733646	wgEncodeBroadHistoneNhlfH3k27acStdAlnRep1.bam.bai
NHLF	H3K27ac	GSM733646	wgEncodeBroadHistoneNhlfH3k27acStdAlnRep1.bam
NHLF	H3K27ac	GSM733646	wgEncodeBroadHistoneNhlfH3k27acStdAlnRep2.bam.bai
NHLF	H3K27ac	GSM733646	wgEncodeBroadHistoneNhlfH3k27acStdAlnRep2.bam
NHLF	H3K27me3	GSM733764	wgEncodeBroadHistoneNhlfH3k27me3StdAlnRep1.bam.bai
NHLF	H3K27me3	GSM733764	wgEncodeBroadHistoneNhlfH3k27me3StdAlnRep1.bam
NHLF	H3K27me3	GSM733764	wgEncodeBroadHistoneNhlfH3k27me3StdAlnRep2.bam.bai
NHLF	H3K27me3	GSM733764	wgEncodeBroadHistoneNhlfH3k27me3StdAlnRep2.bam
NHLF	H3K36me3	GSM733699	wgEncodeBroadHistoneNhlfH3k36me3StdAlnRep1.bam.bai
NHLF	H3K36me3	GSM733699	wgEncodeBroadHistoneNhlfH3k36me3StdAlnRep1.bam
NHLF	H3K36me3	GSM733699	wgEncodeBroadHistoneNhlfH3k36me3StdAlnRep2.bam.bai
NHLF	H3K36me3	GSM733699	wgEncodeBroadHistoneNhlfH3k36me3StdAlnRep2.bam
NHLF	H3K4me1	GSM733649	wgEncodeBroadHistoneNhlfH3k4me1StdAlnRep1.bam.bai
NHLF	H3K4me1	GSM733649	wgEncodeBroadHistoneNhlfH3k4me1StdAlnRep1.bam

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
NHLF	H3K4me1	GSM733649	wgEncodeBroadHistoneNhlfH3k4me1StdAlnRep2.bam.bai
NHLF	H3K4me1	GSM733649	wgEncodeBroadHistoneNhlfH3k4me1StdAlnRep2.bam
NHLF	H3K4me2	GSM733781	wgEncodeBroadHistoneNhlfH3k4me2StdAlnRep1.bam.bai
NHLF	H3K4me2	GSM733781	wgEncodeBroadHistoneNhlfH3k4me2StdAlnRep1.bam
NHLF	H3K4me2	GSM733781	wgEncodeBroadHistoneNhlfH3k4me2StdAlnRep2.bam.bai
NHLF	H3K4me2	GSM733781	wgEncodeBroadHistoneNhlfH3k4me2StdAlnRep2.bam
NHLF	H3K4me3	GSM733723	wgEncodeBroadHistoneNhlfH3k4me3StdAlnRep1.bam.bai
NHLF	H3K4me3	GSM733723	wgEncodeBroadHistoneNhlfH3k4me3StdAlnRep1.bam
NHLF	H3K4me3	GSM733723	wgEncodeBroadHistoneNhlfH3k4me3StdAlnRep2.bam.bai
NHLF	H3K4me3	GSM733723	wgEncodeBroadHistoneNhlfH3k4me3StdAlnRep2.bam
NHLF	H3K79me2	GSM1003549	wgEncodeBroadHistoneNhlfH3k79me2AlnRep1.bam.bai
NHLF	H3K79me2	GSM1003549	wgEncodeBroadHistoneNhlfH3k79me2AlnRep1.bam
NHLF	H3K79me2	GSM1003549	wgEncodeBroadHistoneNhlfH3k79me2AlnRep2.bam.bai
NHLF	H3K79me2	GSM1003549	wgEncodeBroadHistoneNhlfH3k79me2AlnRep2.bam
NHLF	H3K9ac	GSM733652	wgEncodeBroadHistoneNhlfH3k9acStdAlnRep1.bam.bai
NHLF	H3K9ac	GSM733652	wgEncodeBroadHistoneNhlfH3k9acStdAlnRep1.bam
NHLF	H3K9ac	GSM733652	wgEncodeBroadHistoneNhlfH3k9acStdAlnRep2.bam.bai
NHLF	H3K9ac	GSM733652	wgEncodeBroadHistoneNhlfH3k9acStdAlnRep2.bam
NHLF	H3K9me3	GSM1003531	wgEncodeBroadHistoneNhlfH3k09me3AlnRep1.bam.bai
NHLF	H3K9me3	GSM1003531	wgEncodeBroadHistoneNhlfH3k09me3AlnRep1.bam
NHLF	H3K9me3	GSM1003531	wgEncodeBroadHistoneNhlfH3k09me3AlnRep2.bam.bai
NHLF	H3K9me3	GSM1003531	wgEncodeBroadHistoneNhlfH3k09me3AlnRep2.bam
NHLF	Input	GSM733731	wgEncodeBroadHistoneNhlfControlStdAlnRep1.bam.bai
NHLF	Input	GSM733731	wgEncodeBroadHistoneNhlfControlStdAlnRep1.bam
NHLF	Input	GSM733731	wgEncodeBroadHistoneNhlfControlStdAlnRep2.bam.bai
NHLF	Input	GSM733731	wgEncodeBroadHistoneNhlfControlStdAlnRep2.bam

Homo sapiens source data of RNA-seq transcript abundance in FPKM (GTF files) [69]

For downloading, the URL must be constructed by adding the following prefix to each file listed:

`ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/`

Cell type	GEO Accession	File URL suffix
GM12878	GSM958728	wgEncodeCaltechRnaSeqGm12878R2x75I1200TSSRep1V3.gtf.gz
GM12878	GSM958728	wgEncodeCaltechRnaSeqGm12878R2x75I1200TSSRep2V3.gtf.gz
H1-hESC	GSM958733	wgEncodeCaltechRnaSeqH1hescR2x75I1200TSSRep1V3.gtf.gz
H1-hESC	GSM958733	wgEncodeCaltechRnaSeqH1hescR2x75I1200TSSRep2V3.gtf.gz
H1-hESC	GSM958733	wgEncodeCaltechRnaSeqH1hescR2x75I1200TSSRep3V3.gtf.gz
H1-hESC	GSM958733	wgEncodeCaltechRnaSeqH1hescR2x75I1200TSSRep4V3.gtf.gz
HSMM	GSM958744	wgEncodeCaltechRnaSeqHsmmR2x75I1200TSSRep1V3.gtf.gz
HSMM	GSM958744	wgEncodeCaltechRnaSeqHsmmR2x75I1200TSSRep2V3.gtf.gz
HUVEC	GSM958734	wgEncodeCaltechRnaSeqHuvecR2x75I1200TSSRep1V3.gtf.gz
HUVEC	GSM958734	wgEncodeCaltechRnaSeqHuvecR2x75I1200TSSRep2V3.gtf.gz
NHEK	GSM958736	wgEncodeCaltechRnaSeqNhekR2x75I1200TSSRep1V3.gtf.gz

Continued on next page

Continued from previous page

Cell type	GEO Accession	File URL suffix
NHEK	GSM958736	wgEncodeCaltechRnaSeqNhekR2x75I1200TSSRep2V3.gtf.gz
NHLF	GSM958746	wgEncodeCaltechRnaSeqNhlfR2x75I1200TSSRep1V3.gtf.gz
NHLF	GSM958746	wgEncodeCaltechRnaSeqNhlfR2x75I1200TSSRep2V3.gtf.gz

Mus musculus source data of ChIP-seq on histone H3 modifications (SRA files) [70, 68]

For downloading, the URL must be constructed by adding the following prefix to each file listed:

`ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/`

Cell type	Antibody	Rep #	GEO Accession	File URL suffix
E14	IgG	1	GSM881345	SRR414/SRR414932/SRR414932.sra
E14-day0	H3K27ac	1	GSM881349	SRR414/SRR414936/SRR414936.sra
E14-day0	H3K27me3	1	GSM881350	SRR414/SRR414937/SRR414937.sra
E14-day0	H3K36me3	1	GSM881351	SRR414/SRR414938/SRR414938.sra
E14-day0	H3K4me1	1	GSM881352	SRR414/SRR414939/SRR414939.sra
E14-day0	H3K4me3	1	GSM881354	SRR414/SRR414941/SRR414941.sra
E14-day4	H3K27ac	1	GSM881357	SRR414/SRR414945/SRR414945.sra
E14-day4	H3K27me3	1	GSM881358	SRR414/SRR414946/SRR414946.sra
E14-day4	H3K36me3	1	GSM881359	SRR414/SRR414947/SRR414947.sra
E14-day4	H3K4me1	1	GSM881360	SRR414/SRR414948/SRR414948.sra
E14-day4	H3K4me3	1	GSM881362	SRR414/SRR414950/SRR414950.sra
E14-day6	H3K27ac	1	GSM881366	SRR414/SRR414955/SRR414955.sra
E14-day6	H3K27me3	1	GSM881367	SRR414/SRR414956/SRR414956.sra
E14-day6	H3K36me3	1	GSM881368	SRR414/SRR414957/SRR414957.sra
E14-day6	H3K4me1	1	GSM881369	SRR414/SRR414958/SRR414958.sra
E14-day6	H3K4me3	1	GSM881371	SRR414/SRR414960/SRR414960.sra
Heart (8 wks/o)	H3K27ac	1	GSM1000093	SRR566/SRR566827/SRR566827.sra
Heart (8 wks/o)	H3K27ac	2	GSM1000093	SRR566/SRR566828/SRR566828.sra
Heart (8 wks/o)	H3K27me3	1	GSM1000131	SRR566/SRR566903/SRR566903.sra
Heart (8 wks/o)	H3K27me3	2	GSM1000131	SRR566/SRR566904/SRR566904.sra
Heart (8 wks/o)	H3K36me3	1	GSM1000130	SRR566/SRR566901/SRR566901.sra
Heart (8 wks/o)	H3K36me3	2	GSM1000130	SRR566/SRR566902/SRR566902.sra
Heart (8 wks/o)	H3K4me1	1	GSM769025	SRR317/SRR317255/SRR317255.sra
Heart (8 wks/o)	H3K4me1	2	GSM769025	SRR317/SRR317256/SRR317256.sra
Heart (8 wks/o)	H3K4me3	1	GSM769017	SRR317/SRR317239/SRR317239.sra
Heart (8 wks/o)	H3K4me3	2	GSM769017	SRR317/SRR317240/SRR317240.sra
Heart (8 wks/o)	Input	1	GSM769032	SRR317/SRR317269/SRR317269.sra
Heart (8 wks/o)	Input	2	GSM769032	SRR317/SRR317270/SRR317270.sra
Liver (8 wks/o)	H3K27ac	1	GSM1000140	SRR566/SRR566921/SRR566921.sra
Liver (8 wks/o)	H3K27ac	2	GSM1000140	SRR566/SRR566922/SRR566922.sra
Liver (8 wks/o)	H3K27me3	1	GSM1000150	SRR566/SRR566941/SRR566941.sra
Liver (8 wks/o)	H3K27me3	2	GSM1000150	SRR566/SRR566942/SRR566942.sra
Liver (8 wks/o)	H3K36me3	1	GSM1000151	SRR566/SRR566943/SRR566943.sra
Liver (8 wks/o)	H3K36me3	2	GSM1000151	SRR566/SRR566944/SRR566944.sra

Continued on next page

Continued from previous page

Cell type	Antibody	Rep #	GEO Accession	File URL suffix
Liver (8 wks/o)	H3K4me1	1	GSM769015	SRR317/SRR317235/SRR317235.sra
Liver (8 wks/o)	H3K4me1	2	GSM769015	SRR317/SRR317236/SRR317236.sra
Liver (8 wks/o)	H3K4me3	1	GSM769014	SRR317/SRR317233/SRR317233.sra
Liver (8 wks/o)	H3K4me3	2	GSM769014	SRR317/SRR317234/SRR317234.sra
Liver (8 wks/o)	Input	1	GSM769034	SRR317/SRR317273/SRR317273.sra
Liver (8 wks/o)	Input	2	GSM769034	SRR317/SRR317274/SRR317274.sra

Mus musculus RNA-seq source data (BAM files) [70, 68]

For downloading, the URL must be constructed by adding one of the two following prefixes to each file listed:

1. `ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM881nnn/`
2. `ftp://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrRnaSeq/`

Cell type	Rep #	GEO Accession	File URL suffix
E14-day0	1	GSM881355	[*prefix_1*]GSM881355/suppl/GSM881355_E14_RNA.bam.gz
E14-day4	1	GSM881364	[*prefix_1*]GSM881364/suppl/GSM881364_E14_RNA_d4.bam.gz
E14-day6	1	GSM881373	[*prefix_1*]GSM881373/suppl/GSM881373_E14_RNA_d6.bam.gz
Heart (8 wks/o)	1	GSM929707	[*prefix_2*]wgEncodeLicrRnaSeqHeartCellPapMAdult8wksC57bl6AlnRep1.bam
Heart (8 wks/o)	2	GSM929707	[*prefix_2*]wgEncodeLicrRnaSeqHeartCellPapMAdult8wksC57bl6AlnRep2.bam
Liver (8 wks/o)	1	GSM929711	[*prefix_2*] wgEncodeLicrRnaSeqLiverCellPapMAdult8wksC57bl6AlnRep1.bam
Liver (8 wks/o)	2	GSM929711	[*prefix_2*]wgEncodeLicrRnaSeqLiverCellPapMAdult8wksC57bl6AlnRep2.bam

Drosophila melanogaster source data of ChIP-seq on histone H3 modifications (SRA files) [64, 66]

For downloading, the URL must be constructed by adding the following prefix to each file listed:

`ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR030/`

Developmental time point/period	Antibody	GEO Accession	File URL suffix
0-4h embryos	H3K27ac	GSM401407	SRR030295/SRR030295.sra
0-4h embryos	H3K27me3	GSM439448	SRR030360/SRR030360.sra
0-4h embryos	H3K4me1	GSM401409	SRR030297/SRR030297.sra
0-4h embryos	H3K4me3	GSM400656	SRR030269/SRR030269.sra
0-4h embryos	H3K9ac	GSM401408	SRR030296/SRR030296.sra
0-4h embryos	H3K9me3	GSM439457	SRR030369/SRR030369.sra
0-4h embryos	Input	GSM400657	SRR030270/SRR030270.sra

Continued on next page

Continued from previous page

Developmental time point/period	Antibody	GEO Accession	File URL suffix
4-8h embryos	H3K27ac	GSM401404	SRR030292/SRR030292.sra
4-8h embryos	H3K27me3	GSM439447	SRR030359/SRR030359.sra
4-8h embryos	H3K4me1	GSM401406	SRR030294/SRR030294.sra
4-8h embryos	H3K4me3	GSM400674	SRR030287/SRR030287.sra
4-8h embryos	H3K9ac	GSM401405	SRR030293/SRR030293.sra
4-8h embryos	H3K9me3	GSM439456	SRR030368/SRR030368.sra
4-8h embryos	Input	GSM400675	SRR030288/SRR030288.sra
8-12h embryos	H3K27ac	GSM432583	SRR030332/SRR030332.sra
8-12h embryos	H3K27me3	GSM439446	SRR030358/SRR030358.sra
8-12h embryos	H3K4me1	GSM432593	SRR030342/SRR030342.sra
8-12h embryos	H3K4me3	GSM432585	SRR030334/SRR030334.sra
8-12h embryos	H3K9ac	GSM432592	SRR030341/SRR030341.sra
8-12h embryos	H3K9me3	GSM439455	SRR030367/SRR030367.sra
8-12h embryos	Input	GSM432636	SRR030346/SRR030346.sra
12-16h embryos	H3K27ac	GSM432582	SRR030331/SRR030331.sra
12-16h embryos	H3K27me3	GSM439445	SRR030357/SRR030357.sra
12-16h embryos	H3K4me1	GSM432591	SRR030340/SRR030340.sra
12-16h embryos	H3K4me3	GSM432580	SRR030329/SRR030329.sra
12-16h embryos	H3K9ac	GSM439458	SRR030370/SRR030370.sra
12-16h embryos	H3K9me3	GSM439454	SRR030366/SRR030366.sra
12-16h embryos	Input	GSM432634	SRR030344/SRR030344.sra
16-20h embryos	H3K27ac	GSM401401	SRR030289/SRR030289.sra
16-20h embryos	H3K27me3	GSM439444	SRR030356/SRR030356.sra
16-20h embryos	H3K4me1	GSM401403	SRR030291/SRR030291.sra
16-20h embryos	H3K4me3	GSM400658	SRR030271/SRR030271.sra
16-20h embryos	H3K9ac	GSM401402	SRR030290/SRR030290.sra
16-20h embryos	H3K9me3	GSM439453	SRR030365/SRR030365.sra
16-20h embryos	Input	GSM400659	SRR030272/SRR030272.sra
20-24h embryos	H3K27ac	GSM401423	SRR030311/SRR030311.sra
20-24h embryos	H3K27me3	GSM439443	SRR030355/SRR030355.sra
20-24h embryos	H3K4me1	GSM439464	SRR030376/SRR030376.sra
20-24h embryos	H3K4me3	GSM400672	SRR030285/SRR030285.sra
20-24h embryos	H3K9ac	GSM401424	SRR030312/SRR030312.sra
20-24h embryos	H3K9me3	GSM439452	SRR030364/SRR030364.sra
20-24h embryos	Input	GSM400673	SRR030286/SRR030286.sra
L1 larvae	H3K27ac	GSM432581	SRR030330/SRR030330.sra
L1 larvae	H3K27me3	GSM439442	SRR030354/SRR030354.sra
L1 larvae	H3K4me1	GSM432588	SRR030337/SRR030337.sra
L1 larvae	H3K4me3	GSM400662	SRR030275/SRR030275.sra
L1 larvae	H3K9ac	GSM401422	SRR030310/SRR030310.sra
L1 larvae	H3K9me3	GSM439451	SRR030363/SRR030363.sra
L1 larvae	Input	GSM400663	SRR030276/SRR030276.sra
L2 larvae	H3K27ac	GSM401419	SRR030307/SRR030307.sra
L2 larvae	H3K27me3	GSM439441	SRR030353/SRR030353.sra
L2 larvae	H3K4me1	GSM401421	SRR030309/SRR030309.sra
L2 larvae	H3K4me3	GSM400668	SRR030281/SRR030281.sra
L2 larvae	H3K9ac	GSM401420	SRR030308/SRR030308.sra
L2 larvae	H3K9me3	GSM439450	SRR030362/SRR030362.sra
L2 larvae	Input	GSM400669	SRR030282/SRR030282.sra
Pupae	H3K27ac	GSM401413	SRR030301/SRR030301.sra
Pupae	H3K27me3	GSM439439	SRR030351/SRR030351.sra

Continued on next page

Continued from previous page

Developmental time point/period	Antibody	GEO Accession	File URL suffix
Pupae	H3K4mel	GSM401415	SRR030303/SRR030303.sra
Pupae	H3K4me3	GSM400664	SRR030277/SRR030277.sra
Pupae	H3K9ac	GSM401414	SRR030302/SRR030302.sra
Pupae	H3K9me3	GSM439449	SRR030361/SRR030361.sra
Pupae	Input	GSM400665	SRR030278/SRR030278.sra

Drosophila melanogaster RNA-seq source data (SAM files) [64, 66]

For downloading, the URL must be constructed by adding the following prefix to each file listed:

`ftp://data.modencode.org/all_files/dmel-signal-1/`

Developmental time point/period	GEO Accession	File URL suffix
0-4h embryos	GSM451806	2010_0-4_accepted_hits.sam.gz
4-8h embryos	GSM451809	2019_4-8_accepted_hits.sam.gz
8-12h embryos	GSM451808	2020_8-12_accepted_hits.sam.gz
12-16h embryos	GSM451803	2021_12-16_accepted_hits.sam.gz
16-20h embryos	GSM451807	2022_16-20_accepted_hits.sam.gz
20-24h embryos	GSM451810	2023_20-24_accepted_hits.sam.gz
L1 larvae	GSM451811	2024_L1_accepted_hits.sam.gz
L2 larvae	GSM453867	2025_L2_accepted_hits.sam.gz
Pupae	GSM451813	2030_Pupae_accepted_hits.sam.gz