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

Scientists agree that changes in the levels of gene expression are important for the cell differentiation process. 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. The evolution and intrinsic regulatory dynamics of differentiated multicellularity also lack a unified and falsifiable description. To fill this gap, 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-type/developmental-period samples. An analysis of genomic regions adjacent to transcription start sites generated for each cell-type/developmental-period 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 of differentiated multicellularity, which relies on the synergistic coupling across the extracellular space of two stochastically independent "self-organizing" systems constraining histone modification states at the same sites. This theory describes how the differentiated multicellular organism—understood as an intrinsic, higher-order, self-sufficient, self-repairing, self-replicating, and self-regulating constraint—emerges from proliferating undifferentiated cells. If it resists falsification, this theory will explain the intrinsic regulation of gene transcriptional changes during cell differentiation and the emergence of differentiated multicellular lineages throughout evolution.

Introduction

Cell differentiation, if seen as a motion picture in fast-forward, intuitively appears to be a teleological or "end-directed" process, its *telos* or "end" being the multicellular organism in its mature form. The first step for a scientific explanation of this apparent property was given 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].

The dynamics of the cell differentiation process have been associated with changes in chromatin states and concurrent heritable changes in gene expression that are uncorrelated to changes in the DNA sequence, and therefore 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]). Yet most key changes in gene expression during cell differentiation are not only independent from, but are even robust with respect to extrinsic variables. This indicates that cell differentiation is fundamentally an intrinsically regulated process, for which no falsifiable theory has emerged from the epigenetic framework. Due to our lack of understanding of the precise regulatory dynamics, this process has also been dubbed "The X-files of chromatin" [6].

To unravel these X-files, we have to look critically at (i) the regulation of cell differentiation as it is understood today, (ii) the non-genetic information capacity of primordial cells (zygotes, spores, or buds), and (iii) what is assumed to be pre-specified developmental information content in those primordial cells. 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 to chromatin higher-order structural changes [7, 8, 9], these epigenetic mechanisms not only propagate changes in gene expression at different loci as cells proliferate but, importantly, are also hypothesized to regulate the cell differentiation process intrinsically. This hypothesis is accepted as a well-established fact (as illustrated in [10]) even though the epigenetic mechanisms involved in cell differentiation have not been fully elucidated. Furthermore, this epigenetic regulation hypothesis leads to severe explanatory limitations and may even entail logical inconsistencies.

If one assumes that this hypothesis is true in its strictest sense, one accepts that gene self-regulation is a teleological property of cell differentiation. For example, one might assume that a certain *gene A* is an explanatory component of the general self-regulatory property once a researcher who modifies the expression levels of *gene A* in a given organism elucidates how these changes activate or repress the expression of a specific *gene B*, *gene C*, and *gene D* during differentiation. However, this assertion overlooks 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 system is truly regulating *gene B*, *gene C*, *gene D* AND *gene A*—and by extension, all genes during cell differentiation. Moreover, accounting for the regulation of transcriptional states at a gene locus by previous transcriptional states at other gene loci—in the same cell or any other—is only a non-explanatory regress (see Figure 1).

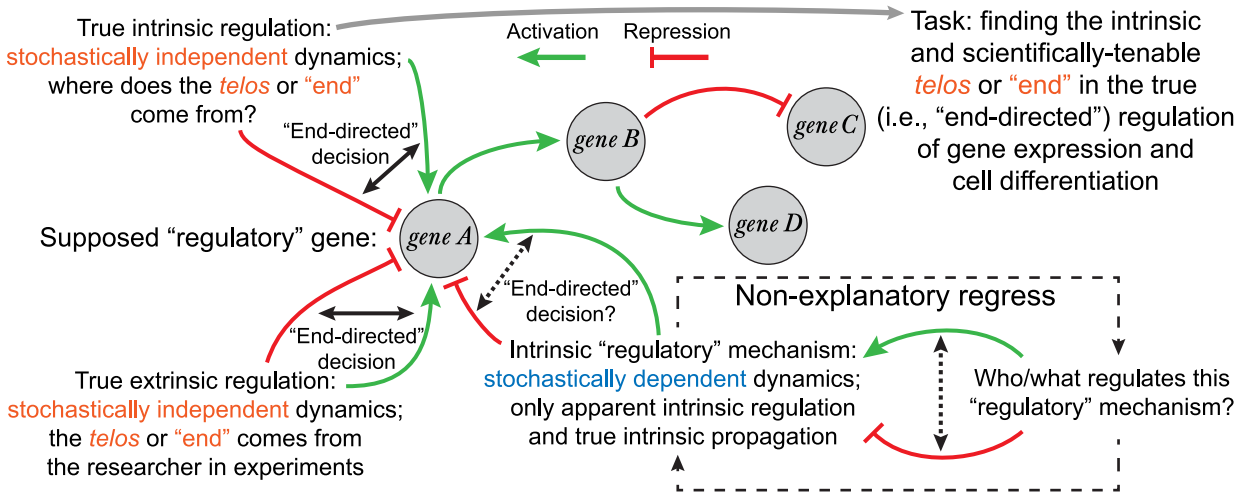

Figure 1: Some of the limitations of the epigenetic landscape
This framework either falls into a non-explanatory regress when attempting to account for the intrinsic regulation of changes in gene expression during cell differentiation or uses "regulation" simply as a placeholder for what is only the propagation of such changes, bracketing true intrinsic regulation from further inquiry.

If one assumes that the epigenetic regulation hypothesis is true in a loose sense, one has 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, an "epigenator"—a transient signal which probably originates in the environment of the cell—would trigger the epigenetic phenotype change after being transduced into the intracellular space [11]. However, if all "epigenators" in the developing organism are extrinsic to it, self-regulation is *ipso facto* unexplainable. Otherwise, the critical signaling property of an "epigenator" (i.e., what it refers to and how it does so) is left unexplained.

The question arises if it is possible that critical changes within a developing organism, and the intrinsic regulation of such changes, are completely different processes at the most fundamental level. Specifically, intrinsic regulation may not be a molecular mechanism correlating critical changes in gene expression within a developing organism but instead may involve particular *constraints* (understood as local thermodynamic boundary conditions that are level-of-scale specific) on those changes. Importantly, these particular constraints—imposed by the regulatory system we look for—are *stochastically independent* (see a formal definition in the Appendix) from the changes this system is supposed to regulate; otherwise the system is fundamentally just an additional mechanism propagating gene expression changes more or less extensively (depending, for example, on the presence of feedback loops) instead of *regulating* them. This explanatory limitation is inescapable: a nonlinear propagation of changes in gene expression only implies either a nonlinear dependence between those changes—describable by a dynamical systems model such as the epigenetic landscape—or chaotic behavior, not a *regulated* propagation of changes. Moreover, intrinsic regulation cannot be explained in terms of any mechanism, machine (e.g., autopoietic [12]), or any "self-organizing" system because all mechanisms, machines and "self-organizing" systems entail an explicit deterministic or stochastic dependence between all their component dynamics. Notably, however, the existence of "self-organizing" systems—a rather misleading term given there is no causally-efficacious *self* in such systems [13]—is a necessary condition for the intrinsic regulatory system of cell differentiation I propose here.

Regardless of the explanatory limitations inherent to the epigenetic landscape, it is generally believed that most, if not all, non-genetic information for later cell differentiation is "hardwired" in the primordial cells. If these cells indeed contain all this information, including that for intrinsic regulation [14, 15], the previously discussed explanatory gap could, in principle, be filled. Asymmetric early cleavage, shown to be able to resolve a few cell lineage commitments (into six founder cells) in the nematode *Caenorhabditis elegans* [16], supports this possibility at first glance, but 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 [17]. 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). However, this is a great underestimation because uncertainty remains with respect to at least two other variables, namely space and time. Therefore, the non-genetic information capacity necessary for the entire cell differentiation process far exceeds the few bits shown to be accounted for by epigenetic mechanisms. On the other hand, extrinsic constraints (e.g., diet-dependent hierarchy determination in eusocial insects [5], temperature-dependent sex determination in reptiles [18], or maternal regulation of offspring development [19]) do not account for all developmental decisions. These considerations highlight that certain *intrinsic* constraints must be identified to account for all the necessary non-genetic information in terms of capacity, which is measurable in units such as bits and content that must account for *how* each developmental decision is made.

The question also arises on how it is possible for an entire organism to develop from *any* totipotent cell, and for embryonic tissues to develop from *any* pluripotent stem cell, if the information for all cell fate decisions is contained in the primordial cell. The recently proposed "epigenetic disc" model for cell differentiation, under which the pluripotent state is only one among many metastable and directly interconvertible states [20], reflects the need to account for the significant dependence of developmental information on the cellular context.

Although David L. Nanney anticipated in 1958 that more important to development than the heritable material itself is the *process* by which heritable material may manifest different phenotypes [21], Waddington's epigenetic landscape prevailed and ever since developmental biology has built upon the premise that primordial cells are indeed complete blueprints of the mature organism (allowing for some limited degree of stochasticity [22, 23] and extrinsic influence as described previously). Thus, the epigenetic landscape framework is not only fundamentally unable to explain the intrinsic regulatory dynamics of cell differentiation, but has also lead research to ignore or reject the necessary *emergence* of developmental information content during ontogeny.

To shed light into "The X-files of chromatin," I designed and conducted a computational analysis of the combinatorial constraints on histone H3 post-translational modification states (to be referred to also as histone H3 crosstalk) because of their strong statistical relationship with transcriptional levels [24]. As data source, 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* cell-type, developmental-period, or developmental-time-point samples. The basis of the analysis was to define a numeric profile *ctalk_non_epi* (for "crosstalk that is non-epigenetic"), an n-tuple or ordered list of numerical values representing for any given sample the component of pairwise histone H3 crosstalk that is

stochastically independent from gene transcription in genomic regions adjacent to transcription start sites (Figure 2).

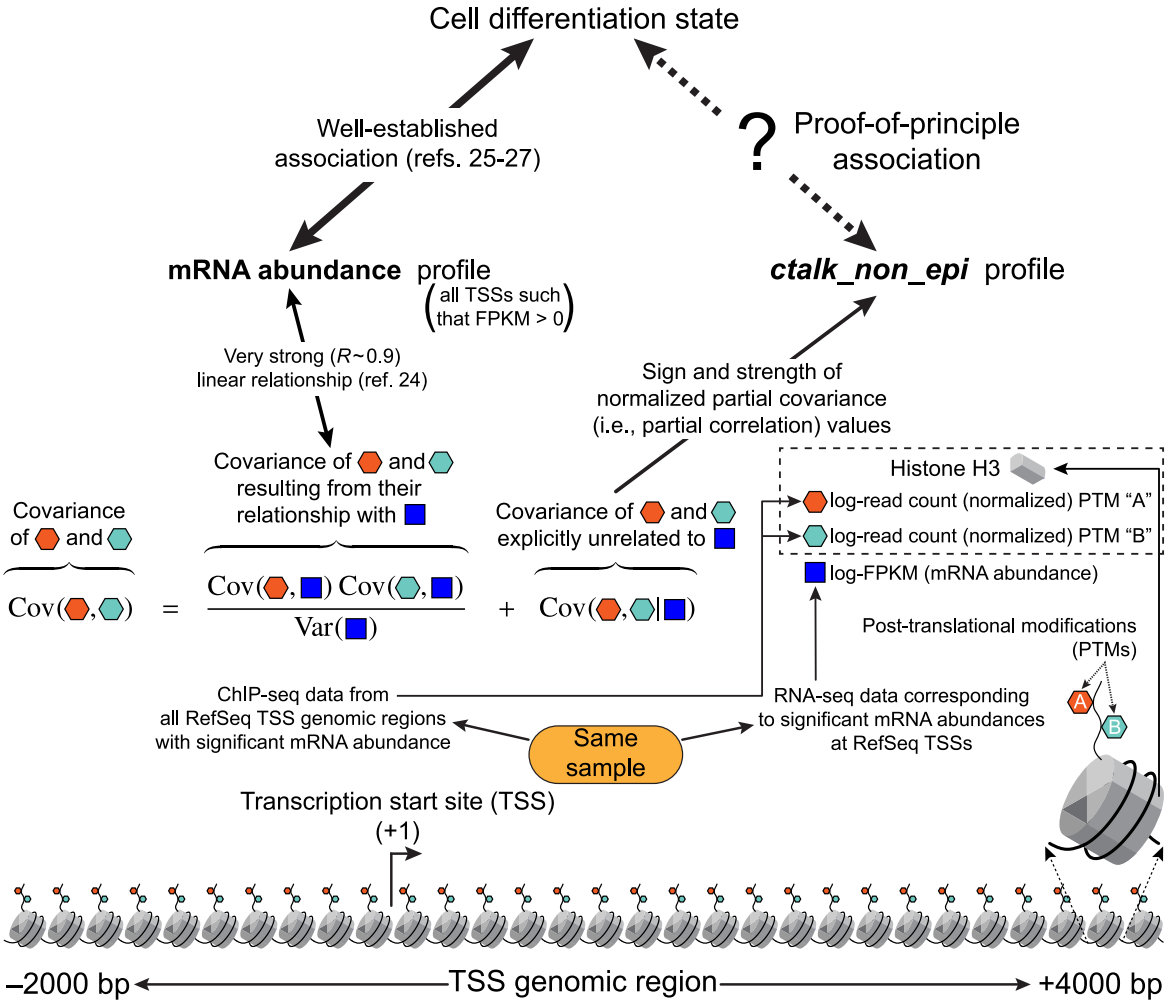

Figure 2: Scheme of the proof-of-principle hypothesis and the computational analysis for its testing.
ctalk_non_epi profiles represent constraints on histone H3 crosstalk that are stochastically independent from mRNA levels. The association between these *ctalk_non_epi* profiles and cell-differentiation states established the proof of principle for the theory proposed in this paper.

Results

Under the arguments presented in the introduction, the aim of the computational analysis was to test the following proof-of-principle, working hypothesis: *for a given cell differentiation state and within genomic regions adjacent to transcription start sites, the component of pairwise histone H3 crosstalk that is stochastically independent from transcriptional levels* (represented by the *ctalk_non_epi* profile) *associates with that differentiation state* (Figure 2, 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 [25, 26, 27] but also explain them completely, an additional non-epigenetic yet differentiation-associated type of constraint on histone H3 crosstalk is superfluous.

To test the proof-of-principle hypothesis, I applied hierarchical cluster analysis (HCA) on the *ctalk_non_epi* profiles for each organism analyzed. If there is no significant association between *ctalk_non_epi* profiles and cell differentiation states (i.e., if the null hypothesis is true), the obtained clusters should be statistically insignificant or else they should not associate with cell differentiation states. However, the results showed in all analyses performed that *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 strongly as mRNA abundance profiles. 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. This unambiguous result provides proof of principle for my theory, which requires differentiation-associated constraints on histone H3 crosstalk such that they are stochastically independent from mRNA levels.

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, 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. This cluster was also statistically significant as reflected in **au** (approximately unbiased) and **bp** (bootstrap probability) significance scores, which were greater than or equal to 95 (Fig. 3A, cluster #4), indicating that this cluster was also statistically significant.

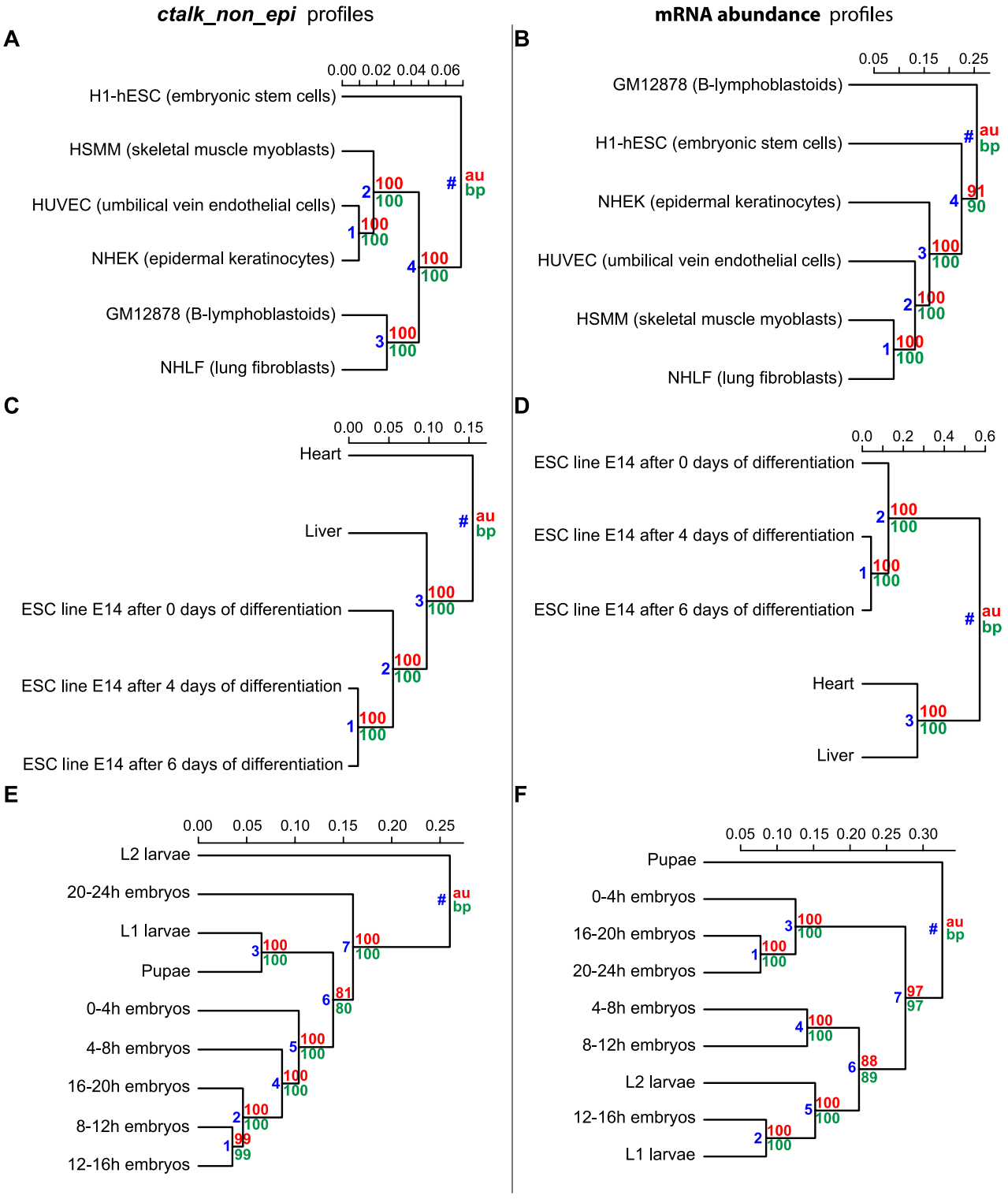

Figure 3: Hierarchical cluster analysis of *ctalk_non_epi* and mRNA abundance profiles
Organisms: *Homo sapiens* (A, B), *Mus musculus* (C, D), and *Drosophila melanogaster* (E, F).
Metric: correlation ($1 - r$). Linkage method: UPGMA. Significance scores: **au** (approximately unbiased) and **bp** (bootstrap probability) [28]. Significant clusters were identified as those for which **au** and **bp** ≥ 95. Cluster identification numbers are in blue.

The *ctalk_non_epi* profile corresponding to H1-hESC (embryonic stem cells) was identified as the most dissimilar with respect to the other profiles, which are all differentiated cell types. For comparison and positive control, mRNA abundance profiles for the six cell types were constructed from RNA-seq data and then hierarchically clustered. As observed for the *ctalk_non_epi* profiles, the mRNA abundance profile corresponding to H1-hESC cells was identified as significantly dissimilar, and therefore excluded from the largest significant cluster (Figure 3B, cluster #3) along with the GM12878 B-lymphoblastoids profile. These findings indicate that *ctalk_non_epi* profiles associate with cell differentiation states in *Homo sapiens*. Notably, in the cell types analyzed, this association was clearer than the association observed between mRNA abundance profiles and cell differentiation states.

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

The analysis for *Mus musculus* comprised five different histone H3 modifications in five cell types. The five cell type datasets analyzed 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. As in *Homo sapiens*, the *ctalk_non_epi* profiles for *Mus musculus* fell into significant clusters that associated with cell differentiation states. All three E14 *ctalk_non_epi* profiles were clustered into a significant, exclusive group (Figure 3C, cluster #2) and within it, the profiles corresponding to latter time points (four and six days of differentiation) fell into another significant cluster (Figure 3C, 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 3C, cluster #3).

Mouse mRNA abundance profiles also fell into significant clusters that associated with cell differentiation states (Figure 3D, clusters #1, #2 and #3). Like *ctalk_non_epi* profiles, mRNA abundance profiles resolved a significant difference between the earliest time point (zero days of differentiation) and latter time points (Figure 3D, cluster #1), indicating that the well-established association between transcriptional and cell differentiation states can be verified also from the data used for *Mus musculus*. Overall, this analysis showed that the association between *ctalk_non_epi* profiles and cell differentiation states is also observable in *Mus musculus*.

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

Similar to those from human and mouse data, *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 associated strongly with the degree of cell differentiation. One significant cluster grouped *ctalk_non_epi* profiles of earlier developmental periods (Figure 3E, cluster #5) apart from later development profiles. Two more significant clusters placed later time point *ctalk_non_epi* profiles (Figure 3E, cluster #3) and separated the L2 larvae profile (Figure 3E, cluster #7) from all other profiles.

General *ctalk_non_epi* cluster structure was not entirely consistent with developmental chronology as the pupae profile showed (Figure 3E, cluster #7). 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*** dataset was obtained from whole specimens (embryos, larvae and pupae). Especially for later developmental stages, 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 the power of the analysis.

The mRNA abundance profiles in ***D. melanogaster*** yielded a general cluster structure that was much less consistent with developmental chronology than the *ctalk_non_epi* profiles. For example, the profile for 0-4h embryos fell into the same significant cluster as the profiles for 16-20h and 20-24h embryos (Figure 3F, cluster #3). Additionally, the profile for 12-16h embryos fell into the same significant cluster as the profiles for L1 and L2 larvae (Figure 3F, cluster #5). Overall, these results indicate that the association between *ctalk_non_epi* profiles and cell differentiation states also holds in ***Drosophila melanogaster*** despite the limitations of the analysis imposed by the ChIP-seq/RNA-seq source data.

Beyond the obtained proof of principle

While the statistically significant association between *ctalk_non_epi* profiles and cell differentiation states is an immediate and critical result of the computational analysis, no less important is the nature of the constraints represented by *ctalk_non_epi* profiles. By definition, *ctalk_non_epi* profiles represent the strength and sign of pairwise partial correlations (with mRNA abundance as the 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 ($R \sim 0.9$) [24]. It follows directly from these considerations that, for all three analyzed organisms within regions adjacent to transcription start sites (TSSs), histone H3 crosstalk is subject to an additional type of constraints that are stochastically independent from mRNA levels and associated with cell differentiation states. In other words, two systems, ***stochastically independent*** and yet ***both associated to cell differentiation states***, constrain histone H3 crosstalk ***at the same sites.***

Discussion

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 certain moment these systems start 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 how this multicellular individual is the true regulator of gene expression during cell differentiation. The theory is also falsifiable (see problems with current hypotheses in the Appendix). Although its proof of principle was obtained from high-throughput metazoan data, the 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 primordial cell(s) of a differentiated multicellular species $\boldsymbol{D}$. Definitions and notation are listed in Table 1.

Table 1: Theoretical definitions and notation

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 at a given instant t. Positions in S_E are 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 *stochastically independent* from 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 *stochastically independent* from changes in the instantaneous transcription rates at the TSSs. In this work Nanney's constraints are 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 specified as Waddington's embodiers F_W and *also* as Nanney's embodiers F_N.
$\vec{F_N}(X_{(i;t)})$	**Nanney's extracellular propagators**: the subset of the entire spatially-specified molecular phenotype of $X_{(i;t)}$ that excludes Nanney's embodiers $F_N(X_{(i;t)})$ and that is (i) secreted into the extracellular space S_E and (ii) capable of eliciting a significant change (via facilitated diffusion/signal transduction) in Nanney's embodiers F_N within other cells' nuclei after a certain time interval Δt.
$\vec{\nabla}\Phi_N(X_{(1;t)},\ldots,X_{(n;t)})$	**Gradient of Nanney's extracellular propagators**: the vector whose components are the partial derivatives of the concentration $\Phi_N(r,\theta,\phi)$ of Nanney's extracellular propagators $\vec{F_N}$ with respect to the coordinates (r,θ,ϕ) in the extracellular space S_E.

If indeed two systems, stochastically independent and yet both associated to cell differentiation states, constrain histone H3 crosstalk at the same sites, my theory regarding differentiated multicellularity must still address outstanding fundamental questions, including (i) how the association between *ctalk_non_epi* profiles and cell differentiation states is actually realized in the developing organism, (ii) how the intrinsic regulation of gene expression is exerted, and (iii) how the intrinsic regulatory system emerged throughout evolution and emerges within the ontogenetic process. These questions can be answered as follows:

Part I (Evolution): The unicellular (or undifferentiated multicellular) ancestor.

- $\boldsymbol{U}_{(i;t_{U_0})}$ is the i^{th} cell in a population of the unicellular (or undifferentiated multicellular) species $\boldsymbol{U}$ (Figure 4A, 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 *stochastically independent* from 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 $\boldsymbol{U}_{(i;t_{U_0})}$, i.e., Nanney's extracellular propagators $\vec{\boldsymbol{F}}_N$ do not exist in $\boldsymbol{U}_{(i;t_{U_0})}$.

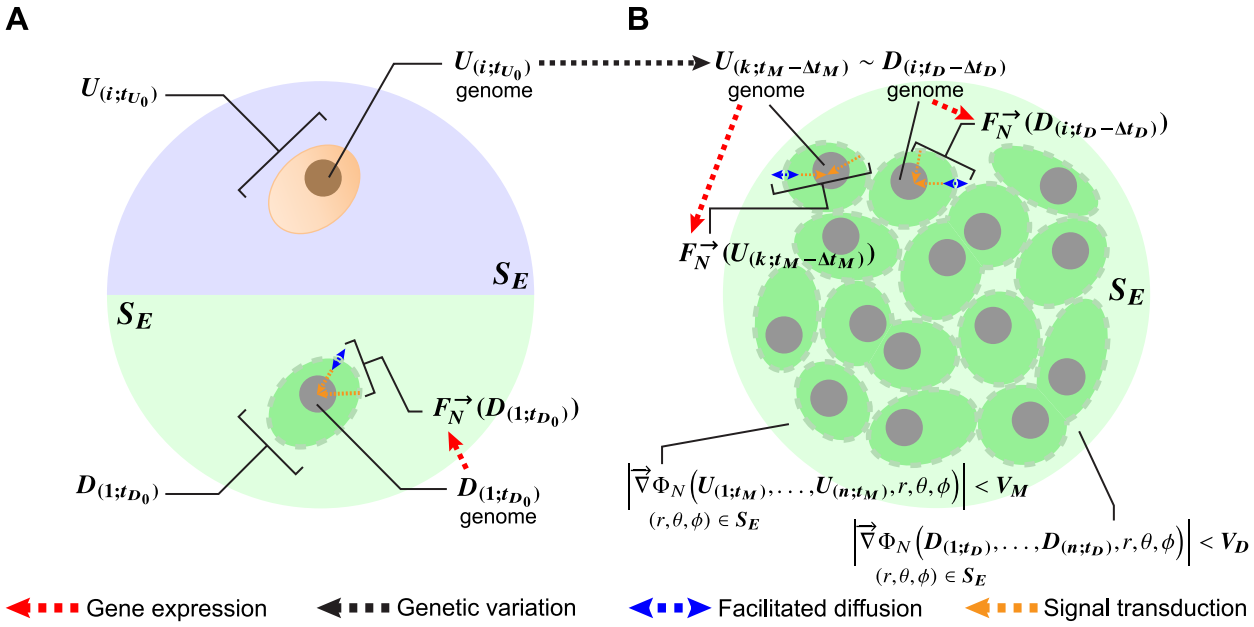

Figure 4: Necessary initial conditions for differentiated multicellularity
(A, top) A cell of the unicellular and undifferentiated ancestor species $\boldsymbol{U}$. (A, bottom) A primordial cell of the multicellular species $\boldsymbol{D}$. (A, top to B, top) The necessary genetic change for differentiated multicellularity occurs in the species $\boldsymbol{U}$. (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 I (Ontogeny): The differentiated multicellular organism's primordial cell

- $\boldsymbol{D}_{(1;t_{D_0})}$ is a primordial cell (as exemplified by zygotes, spores, or buds) of the extant differentiated multicellular species $\boldsymbol{D}$ (Figure 4A, 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 *stochastically independent* from 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 $\boldsymbol{D}_{(1;t_{D_0})}$. That is, 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 $(t_M - \Delta t_M) > t_{U_0}$ during evolution the genome of certain $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ cell suffers a change (Figure 4A to 4B) such that it now synthesizes a molecule specifiable as a Nanney's extracellular propagator $\boldsymbol{F}_N^{\rightarrow}$.
- A molecular substrate is synthesized that is membrane exchangeable and, once it enters the cell, it 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).
- Crucially, this change is *stochastically independent* from the current gene transcriptional rates when it is elicited.
- The genetic change implies that the genome now codes for all gene products necessary for the synthesis, facilitated diffusion, and/or signal transduction of the novel Nanney's extracellular propagator(s) $\boldsymbol{F}_N^{\rightarrow}$.
- Importantly, the novel alleles are a necessary but not sufficient condition for differentiated multicellularity (Figure 4B).

Part II (Ontogeny): Already present necessary alleles

- At any instant $(t_D - \Delta t_D) > t_{D_0}$ the genome of any cell $\boldsymbol{D}_{(i;t_D-\Delta t_D)}$ in the primordial cell's offspring is similar to the genome of the cell $\boldsymbol{U}_{(k;t_M-\Delta t_M)}$ (see Figure 4B, top) in that both genomes code for Nanney's extracellular propagators $\boldsymbol{F}_N^{\rightarrow}$.
- Importantly, the alleles specified in the genome of the primordial cell $\boldsymbol{D}_{(1;t_{D_0})}$—and in the genome of any cell in its offspring—are a necessary but not sufficient condition for cell differentiation (Figure 4B).

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 $F_N^{\rightarrow}$ allows us to define a scalar field Φ_N describing the concentration of $F_N^{\rightarrow}$ in the extracellular space S_E and its associated concentration gradient $\overrightarrow{\nabla}\Phi_N$ 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 $\overrightarrow{\nabla}\Phi_N$ 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 certain critical value V_D for the offspring of the primordial cell $D_{(1;t_{D_0})}$ (Figure 4B, bottom).
- The constraints represented by the gradient $\overrightarrow{\nabla}\Phi_N$ 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.

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 large population for which 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 forms, in magnitude equal to or greater than the critical value V_M—anywhere in S_E, i.e., $\left|\overrightarrow{\nabla}\Phi_N\left(U_{(1;t_M)},\ldots,U_{(n;t_M)},r,\theta,\phi\right)\right| \geq V_M, (r,\theta,\phi) \in S_E$ (Fig. 5, bottom-left).
- As a consequence, Nanney's extracellular propagators $F_N^{\rightarrow}$ diffuse differentially into each cell, yielding unprecedented differential changes in Nanney's embodiers $\{F_N(U_{(1;t_M)}),\ldots,F_N(U_{(n;t_M)})\}$ (i.e., histone post-translational modifications in TSS-adjacent genomic regions; see Fig. 5, center) in the cells' nuclei, not because of any cell or gene product in particular, but because of the constraints imposed by the entire proliferating cell population on the diffusion flux of $F_N^{\rightarrow}$ in S_E.
- Because Nanney's embodiers $\{F_N(U_{(1;t_M)}),\ldots,F_N(U_{(n;t_M)})\}$ are also specifiable as Waddington's embodiers $\{F_W(U_{(1;t_M)}),\ldots,F_W(U_{(n;t_M)})\}$ (as shown in the Results), these differential changes in turn elicit differential changes in the instantaneous transcription rates *irrespectively* of how gene transcriptional changes were propagating up to that instant. This part of the theory explains how—as a consequence—multicellular lineages evolved that display *self-regulated* changes in gene expression during ontogeny.

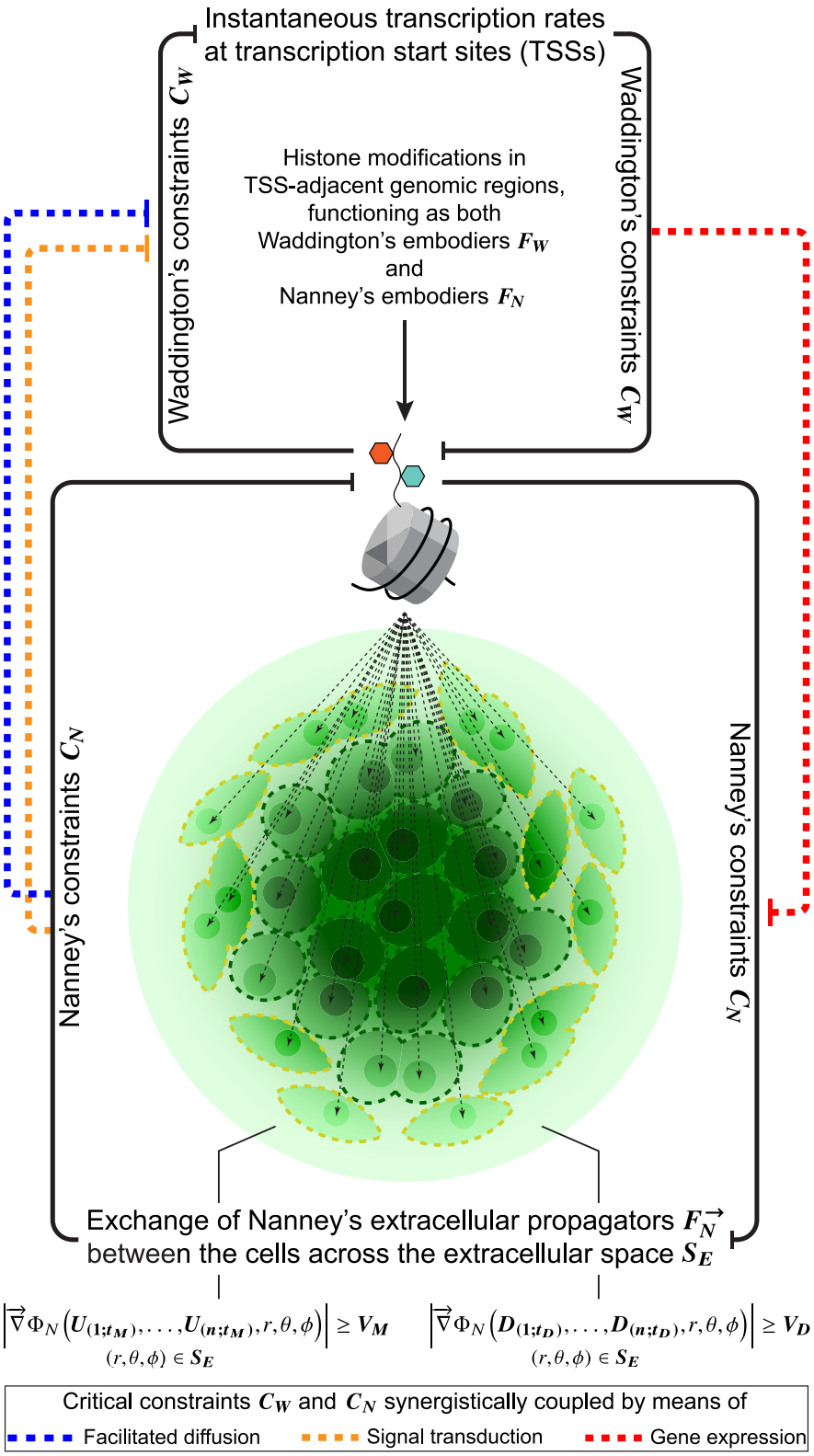

Figure 5: Emergent transition to differentiated multicellularity and to cell differentiation

Intrinsic higher-order constraint emerges when significant gradients $\vec{\nabla}\Phi_N$ couple the lower-order and stochastically independent Waddington's constraints C_W and Nanney's constraints C_N synergistically across S_E. Histone modification states in the TSS-adjacent genomic regions become thus constrained differentially by two stochastically independent "self-organizing" systems. This is how differentiated multicellular lineages emerged throughout evolution and how cell differentiation emerges during ontogeny, all displaying a truly *self-regulated* dynamical regime.

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 a certain number of undifferentiated cells for which 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., $\left|\overrightarrow{\nabla}\Phi_N\left(D_{(1;t_D)},\ldots,D_{(n;t_D)},r,\theta,\phi\right)\right| \geq V_D, (r,\theta,\phi) \in S_E$ (Figure 5, bottom-right).
- As a consequence, Nanney's extracellular propagators $F_N^{\rightarrow}$ diffuse differentially into each cell, yielding unprecedented differential changes in Nanney's embodiers $\{F_N(D_{(1;t_D)}),\ldots,F_N(D_{(n;t_D)})\}$ (i.e., histone post-translational modifications in TSS-adjacent genomic regions; see Figure 5, center) in the cells' nuclei by virtue of no cell or gene product in particular but because of the constraints imposed by the entire growing embryo on the diffusion flux of $F_N^{\rightarrow}$ in S_E.
- Because Nanney's embodiers $\{F_N(D_{(1;t_D)}),\ldots,F_N(D_{(n;t_D)})\}$ are also specifiable as Waddington's embodiers $\{F_W(D_{(1;t_D)}),\ldots,F_W(D_{(n;t_D)})\}$ (as shown in the Results), these differential changes in turn elicit differential changes in the instantaneous transcription rates *irrespectively* of how gene transcriptional changes were propagating up to that instant. This part of the theory explains as a consequence how undifferentiated cells start to differentiate, displaying *self-regulated* changes in gene expression during ontogeny.

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 [29], the necessary genetic change from the genome of the cell $U_{(i;t_{U_0})}$ to the genome of the cell $U_{(k;t_M-\Delta t_M)}$ can be safely regarded as a highly improbable step.
- The major evolutionary breakthrough was not genetic but instead the unprecedented dynamical regime emerging from proliferating eukaryote cells at t_M, or in more general terms at $\{t_{M_1},\ldots,t_{M_n}\}$ throughout evolution since extant differentiated multicellular organisms constitute a paraphyletic group [30, 31].
- This novel dynamical regime emerges as a higher-order constraint from the synergistic coupling of the lower-order Waddington's constraints C_W and Nanney's constraints C_N, able now to propagate through the extracellular space S_E (Figure 5, black lines).
- Although dependent on the novel alleles in the genome of $U_{(k;t_M-\Delta t_M)}$ to emerge given enough cell proliferation, this system is not a network of molecular mechanisms—however complex. Instead, it is a particular example of the generic *teleodynamic system*, proposed by Terrence Deacon in his *emergent dynamics* theory [32], which emerges when certain specific conditions are met and *then* is subject to and shaped by the interplay between its emergent properties and neo-Darwinian mechanisms (interplay henceforth referred to simply as evolution). In this context, environmental constraints such as oxygen availability [33] and

even gravity (see Corollary #5) filter out specific emergent multicellular dynamics that are incompatible with those constraints.

- In summary, the critical evolutionary novelty was the unprecedented multicellular individual or multicellular *self*, which can be described as an intrinsic, higher-order constraint that emerges spontaneously from a particular class of proliferating eukaryotic cells. Being an intrinsic, higher-order *constraint*, this multicellular *self* can be causally-efficacious when regulating its intrinsic dynamics or modifying its surroundings.

Part V (Ontogeny): *Who* is regulating cell differentiation?

- Contrary to what could be derived from Turing's hypothesis [34], the theory hereby proposed does *not* regard the significant proliferation-generated extracellular gradient, i.e., $\left|\overrightarrow{\nabla}\Phi_N\left(\boldsymbol{D}_{(1;t_D)},\ldots,\boldsymbol{D}_{(n;t_D)},r,\theta,\phi\right)\right| \geq \boldsymbol{V_D}$ (anywhere in $\boldsymbol{S_E}$), as the fundamental regulator of the cell differentiation process.
- Whereas differential Nanney's constraints $\{\boldsymbol{C_N}(\boldsymbol{D}_{(1;t_D)}),\ldots,\boldsymbol{C_N}(\boldsymbol{D}_{(n;t_D)})\}$ are *regulatory constraints* with respect to Waddington's embodiers $\{\boldsymbol{F_W}(\boldsymbol{D}_{(1;t_D)}),\ldots\boldsymbol{F_W}(\boldsymbol{D}_{(n;t_D)})\}$ as described in Part IV-Ontogeny (see Figure 5, blue/orange dashed lines), the reciprocal proposition is also true. Namely, Waddington's constraints $\{\boldsymbol{C_W}(\boldsymbol{D}_{(1;t_D)}),\ldots,\boldsymbol{C_W}(\boldsymbol{D}_{(n;t_D)})\}$ are *stochastically independent* from Nanney's constraints, thus Waddington's constraints $\boldsymbol{C_W}$ are in turn *regulatory constraints* with respect to Nanney's extracellular propagators $\{\overrightarrow{\boldsymbol{F_N}}(\boldsymbol{D}_{(1;t_D)}),\ldots,\overrightarrow{\boldsymbol{F_N}}(\boldsymbol{D}_{(n;t_D)})\}$, e.g., by modifying the expression of protein channels, carriers, membrane receptors, or intracellular transducers necessary for the facilitated diffusion/signal transduction of Nanney's extracellular propagators (Figure 5, red dashed lines).
- *Consequently, only if the stochastically independent Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ become synergistically coupled across the extracellular space $\boldsymbol{S_E}$, true intrinsic regulation on the cell differentiation process is possible* (Figure 5).
- This corollary implies in turn that both histone modification states in the TSS-adjacent regions and transcriptional states are reciprocally cause and effect with respect to each other, thus providing also a plausible fundamental account of the more general and so far unresolved causal relationship between nuclear organization and gene function (discussed in [35]). This causally circular dynamical regime—intuitively describable as "chicken-egg" dynamics—is characteristic of teleodynamic systems and teleodynamic systems only [36].
- The true regulator of the cell differentiation process is then the developing multicellular organism itself. This is because the individuated multicellular organism *is* the intrinsic and causally-efficacious higher-order *constraint* emerging from and regulating *ipso facto* Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ (when coupled synergistically coupling of across the extracellular space $\boldsymbol{S_E}$) in what would be otherwise an arbitrarily complex population or colony of unicellular eukaryotes.

Part VI (Evolution): Unprecedented multicellular dynamics

- Once the necessary alleles for differentiated multicellularity are present in some eukaryotic lineages (see Part II-Evolution), further variation due to phenomena like mutation, gene duplication or alternative splicing make possible the emergence of a plethora of novel (teleodynamic) multicellular 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 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 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 to their prokaryotic counterparts.
- 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 [37]. In other words, the need for transcription-factor genetic information increases faster than the total amount of genes or gene products it is involved in regulating. Intriguingly, the eukaryotes analyzed were the group with the smallest power-law exponent. This means that the most complex organisms require proportionally *less* transcription-factor information. With data available today [38], a reproduction I conducted of the aforementioned analysis allowed a robust confirmation: the power-law exponent for unicellular or undifferentiated multicellular eukaryotes is 1.33 ± 0.31 (based on 37 genomes, data not shown). For differentiated multicellular eukaryotes is 1.11 ± 0.18 (67 genomes). The loss of lower-order, cell-intrinsic regulatory systems in differentiated multicellular organisms described in the previous paragraph—in turn accounted for by the emergence of higher-order information content (see Part IX)—explains these 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 $\vec{F_N}$ began to be synthesized and exchanged through the cells' membrane.
- Therefore, the theoretical description presented here rejects the hypothesis that metazoans—or, in general, any multicellular lineage displaying self-regulated cell differentiation—evolved from gradual specialization and division of labor in single-cell colonies or aggregations [31, 39, 40, 41, 42, 43, 44].
- However, this rejection does not imply that precedent traits (e.g., cell-cell adhesion) were unimportant for the later fitness of differentiated multicellular organisms.

- Haeckel's famous assertion is not rejected completely because it contains some truth: 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 evolution ever since its phylogenetic debut. Therefore, 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 by themselves, the intrinsic regulation of cell differentiation.
- The robustness, heritability, and number of cell divisions, which any epigenetic change comprises, are instead *adaptations* of the intrinsic higher-order constraint emergent from proliferating individual cells (i.e., the multicellular organism).
- These adaptations have been shaped by evolution after the emergence of each extant multicellular lineage and are in turn reproduced, eliminated, 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 form 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., $\left(\left|\overrightarrow{\nabla}\Phi_{N_{1;T_j}}\right|,\ldots,\left|\overrightarrow{\nabla}\Phi_{N_{k;T_j}}\right|\right)^{\mathsf{T}} < \left(\boldsymbol{V}_{N_{1;T_j}},\ldots,\boldsymbol{V}_{N_{k;T_j}}\right)^{\mathsf{T}}, (r,\theta,\phi) \in \boldsymbol{S_{E_i}}$, 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 6).
- As highlighted earlier, cell differentiation is not *regulated* by these gradients themselves but by the intrinsic, 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 $\boldsymbol{S_E}$ size) or may limit further propagation of those gradients from $\boldsymbol{S_E}$ into the cells (e.g., by repressing the expression of genes involved in the facilitated diffusion/signal transduction of $\boldsymbol{F_N^{\rightarrow}}$ in $\boldsymbol{S_E}$).

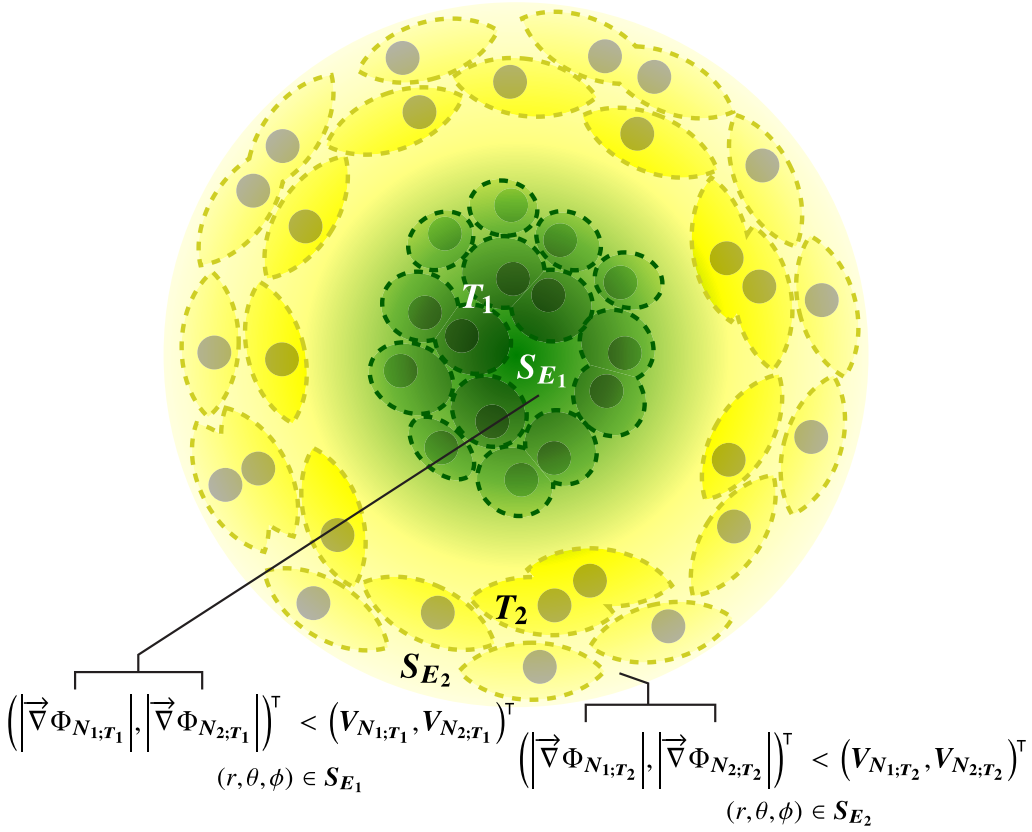

Figure 6: Novel cell types develop
Two distinct cell types T_1 and T_2 develop respectively in regions S_{E_1} and S_{E_2} within S_E characterized by a relative small $\vec{\nabla}\Phi_N$ gradient magnitude, i.e., in extracellular regions of relative $F_N^{\rightarrow}$ uniformity.

- 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 $\vec{\nabla}\Phi_N$ magnitude; see Figure 7A) within S_E displaying no particular shape or function—apart from being compatible 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.
- Then, these $F_N^{\rightarrow}$-uniformity "blobs" were shaped by evolution from their initially random physiological and structural properties to specialized cell types, tissues, and organs (Figure 7, gray dashed arrows). Such specialization evolved towards *serving* the emergent intrinsic higher-order constraint proposed here as being the multicellular organism itself. The result of this evolutionary process—which, importantly, includes emergent (teleodynamic) transitions—is observable in the dynamics characterizing the ontogeny of extant multicellular species.

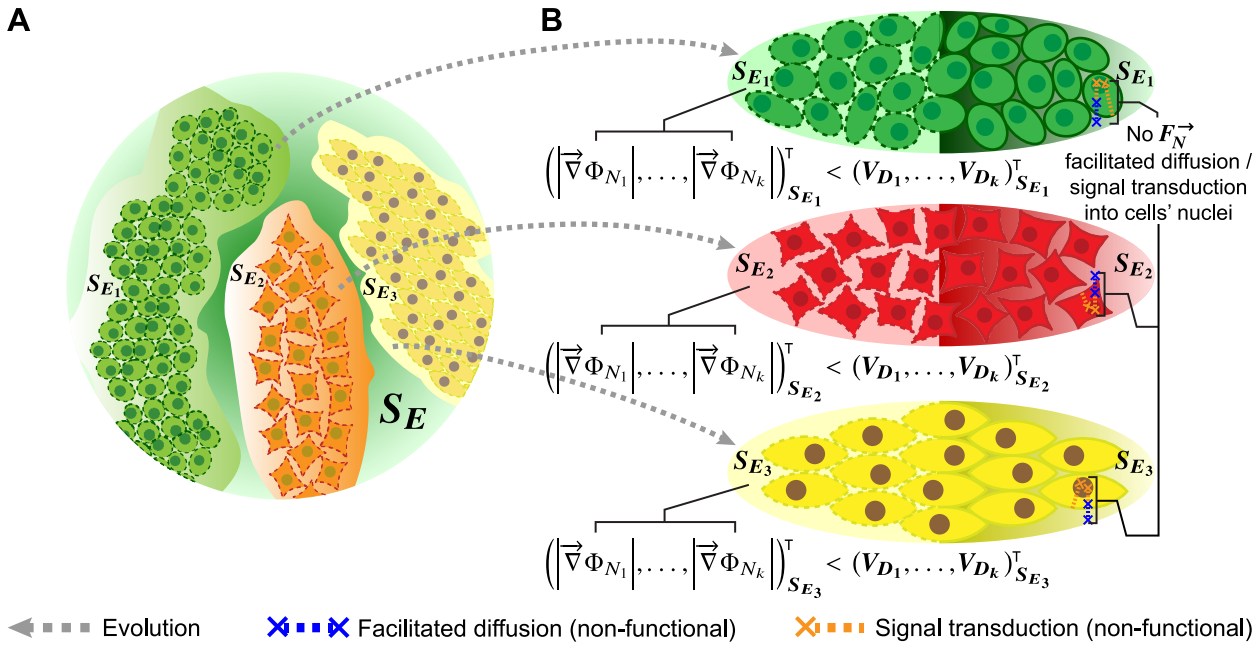

Figure 7: The evolution of types/tissues/organs and the "termination" of cell differentiation
(A) Cell types/tissues/organs evolve as emergent "blobs" of relatively small $\vec{\nabla}\Phi_N$ magnitude and then are shaped by evolution. (B) Cell differentiation stops when the $\vec{\nabla}\Phi_N$ gradients dissipate (left), or when they cannot diffuse/be transduced into the cells' nuclei (right).

Part IX (Evolution & Ontogeny): Emergent *hologenic* information and multicellular self-repair

- 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 $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ form at some point from proliferating undifferentiated cells, coupling synergistically Nanney's constraints C_N and Waddington's constraints C_W across 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* (the suffix -genic may denote "producing" or "produced by", which are both true under this theory as will be shown next).

- No less importantly, at each instant the multicellular organism is not only interpreting hologenic information—by constraining its development into specific trajectories since its emergence given the current $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ gradients in $\boldsymbol{S_E}$—but also actively generating novel hologenic information, e.g., when its very growth and the morphological changes in its differentiating cells modify the spatial constraints in $\boldsymbol{S_E}$ and, as a consequence, the $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ gradients. This causally circular dynamical regime is similar to that described in Part V-Ontogeny because it is underpinned by the same logic of constraint reciprocity (i.e., a teleodynamic relationship [32, 36]).
- Thus, in the most fundamental sense, *cell differentiation is an interpretive process*, not the replication or inheritance of any molecular "code", as David L. Nanney had indirectly anticipated [21]; its *defining interpreter of information*—endogenous such as hologenic information or exogenous such as that in royal jelly feeding [5]—*is the developing organism itself.*
- The subset of the molecular phenotype that conveys hologenic information is not only the subset involved in the gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ but the entire subset embodying or propagating Nanney's constraints $\boldsymbol{C_N}$.
- Hologenic information content is *"absent"* by virtue of intrinsic constraints: 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 otherwise realizable states that become constrained or made "absent" in the dynamics of the multicellular organism by the synergistic coupling of Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ across $\boldsymbol{S_E}$ 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.
- Additionally, since the gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ 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.

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" 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 $\boldsymbol{F_N^{\rightarrow}}$ at any instant $\boldsymbol{t}$ may not be specifiable as such at some later instant $\boldsymbol{t+\Delta t}$.

(ii) The emergent *telos* or "end" in this theory is the instantaneous, higher-order intrinsic constraint that emerges from proliferating undifferentiated cells (i.e., the 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—the "intuitive" *telos* is logically inconsistent, because such a *telos* entails the causal power of future events on events preceding them.

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

- Therefore, under this theory, cell differentiation "terminates" in any given region $\boldsymbol{S_{E_i}}$ 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 $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ dissipate in $\boldsymbol{S_{E_i}}$ under certain critical values, i.e., $\left(\left|\overrightarrow{\nabla}\Phi_{N_1}\right|,\ldots,\left|\overrightarrow{\nabla}\Phi_{N_k}\right|\right)^{\mathsf{T}} < \left(\boldsymbol{V_{D_1}},\ldots,\boldsymbol{V_{D_k}}\right)^{\mathsf{T}}, (r,\theta,\phi) \in \boldsymbol{S_{E_i}}$.

- 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 $\boldsymbol{S_{E_i}}$ that facilitate the formation/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*, once controlling for cell characteristic length.

(b) Extracellular gradients $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ are unable to constrain Waddington's embodiers $\boldsymbol{F_W}$ in the cells' nuclei because Nanney's extracellular propagators $\boldsymbol{F_{\overrightarrow{N}}}$ are non-functional or not expressed (i.e., the gradients dissipated), or the critical gene products for their facilitated diffusion/signal transduction are non-functional or not expressed.

- 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 $\boldsymbol{F_{\overrightarrow{N}}}$, i.e., the cells become "blind" to the gradients if they exist (Figure 7B, right).

- 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.

- 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.

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 primordial cell is a complete developmental blueprint containing all necessary information for the process 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, as first argued by T. Deacon [32]. Therefore, this teleological causal power cannot be mechanistically reduced or decomposed into molecules and their interactions.
- Evolution has shaped the content of hologenic information (from the initial "blobs" of relative $\boldsymbol{F}_{\vec{N}}$ uniformity in $\boldsymbol{S}_E$, see Figure 7A) by capturing the lower-order genetic constraints it is ultimately emergent from, not any particular molecules as media for their embodiment, media that should be regarded in this context as *means* to the multicellular *telos*. This explanation also implies a trade-off between cell independence and cell phenotypic complexity/diversity: the multicellular *telos* offloads regulatory work the cells were performing individually (as described in Part VI-Evolution), allowing them to use that free energy surplus for sustaining more complex and diverse phenotypes but also making them more dependent on the multicellular *telos* they serve.
- 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 $\boldsymbol{F}_{\vec{N}}$ 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 as a trait of highly differentiated cells in the evolution of multicellularity, provided no rigid wall impedes the tinkering with substantial increases of the cells' surface-to-volume ratio.

- Together with the predicted negative correlation between cell potency and cell 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 thus be expected to have a comparatively lower complexity but higher cell potency and robustness to extrinsic damage, as seen in the plant lineage: an adult plant can regenerate all its body parts and a cutting from it can develop into a whole new plant.

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 $\overrightarrow{\nabla}\Phi_N$ form in $\boldsymbol{S_E}$ at some point given enough cell proliferation, it is also true that the exchange of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ across $\boldsymbol{S_E}$ is constrained by the dynamics of facilitated diffusion and/or ligand-receptor binding, which are saturable. Any representative simulation of cell differentiation according to this theory, however simple, will depend on an accurate modeling of the lower-order constraints it emerges from.

The proposed theory also encompasses 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 changes in Nanney's embodiers $\boldsymbol{F_N}$ inside the cells' nuclei. Related to this theory, evidence has already been found for tissue migration across a migration-generated chemokine gradient in zebrafish [47, 48].

Two relevant simplifications or approximations were applied in my computational 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 because (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 [49, 50], (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 is robust with respect to these simplifications and approximations.

If the theory advanced here resists falsification attempts consistently, further research will be needed to identify the cell-and-instant-specific Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ 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 such as cancer (see Corollary #7). Also, more theoretical development will be needed to quantify the capacity and classify the content of hologenic information that emerges along with cell differentiation.

The critique of the epigenetic landscape approach, however, 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. To advance our fundamental understanding of the evolution and self-regulatory dynamics of differentiated multicellularity, future research needs to recognize that the propagation of changes in gene expression and the regulation of those changes must be processes stochastically independent from each other at certain critical parts of the nuclear phenotype.

Falsifiability

For the presented theory, Popper's criterion of falsifiability will be met 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 $\boldsymbol{D}$ (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 molecule $\boldsymbol{M}$ that is secreted into the extracellular space $\boldsymbol{S_E}$ and that activates or represses the expression of certain genes in other cells from being also specifiable as a Nanney's extracellular propagator $\boldsymbol{F_W^{\rightarrow}}$. In other words, more likely than the discovery of a previously undescribed molecule will be the verification of the ability of some known secreted molecules to elicit changes in Nanney's embodiers $\boldsymbol{F_N}$ 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 subsection because it is not falsifiable in a strict epistemological sense.

2. Cell 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 neurons by constraining to some extent the diffusion flux of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ in the extracellular space $\boldsymbol{S_E}$. This hypothesis is supported by the fact that the cells serving as neural stem cells are ependymal cells [51], which are precisely those of the smallest surface-to-volume ratio in the neuroglia. 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 probably preceded the evolution of all other types of support, given the critical role of the $\overrightarrow{\nabla}\Phi_N$ gradients in the emergence and preservation of the multicellular *telos*.

3. Natural developmental arrests or diapauses. The account for natural diapauses follows directly from the description in Part X-Ontogeny (such diapauses occur in arthropods [52] and some species of killifish [53]). Natural diapauses are under this theory a metastable equilibrium state characterized by (i) the dissipation of the extracellular gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ in S_E under certain critical values because the otherwise operating Nanney's extracellular propagators $F_N^{\rightarrow}$ are no longer expressed or functional, or (ii) the inability of these gradients to constrain Waddington's embodiers F_W in the cells' nuclei because the critical gene products for protein channels/carriers or signal transducers are not expressed or non-functional. For example, if at some developmental time point the expression/function/facilitated diffusion/signal transduction of the current Nanney's extracellular propagators $F_N^{\rightarrow}$ is temperature-dependent, then the developing organism will enter a diapause if certain thermal conditions are met and will exit the diapause later if those conditions are lost.

4. $\vec{\nabla}\Phi_N$ 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 $\left|\vec{\nabla}\Phi_N\left(\boldsymbol{D}_{(1;t_{\text{wound}})},\ldots,\boldsymbol{D}_{(n;t_{\text{wound}})},r,\theta,\phi\right)\right| \gg \left|\vec{\nabla}\Phi_N\left(\boldsymbol{D}_{(1;t_{\text{wound}}-\Delta t)},\ldots,\boldsymbol{D}_{(n;t_{\text{wound}}-\Delta t)},r,\theta,\phi\right)\right|, (r,\theta,\phi)\in S_E$ 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 [54]. If relevant variables (such as $F_N^{\rightarrow}$ diffusivity in the extracellular space 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 [55], plant growth [56], human gene expression [57], cytoskeleton organization and general embryo development ([58] 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 $F_N^{\rightarrow}$ distribution in the extracellular space 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 [59], and/or a significant density difference between the extracellular space 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 additional nutrient supply until they are able to photosynthesize. Whereas this specific requirement of water has been associated to embryo expansion and metabolic activation of the seeds [60, 61], I submit that it is also associated to the fundamental need for a proper medium in $\boldsymbol{S_E}$ where the critical gradients $\{\overrightarrow{\nabla}\Phi_{N_1},\ldots,\overrightarrow{\nabla}\Phi_{N_k}\}$ can form. These gradients would be in turn required for the intrinsic regulation of the asymmetric divisions already shown critical for cell differentiation in plants [62].

7. The (lost) multicellular *telos* of cancer cells. Previous research has shown that a normal cellular context can keep tissue-specific stem cells (TSSCs) with damaged DNA in check, preventing cancer onset for long time intervals ([63] and references therein). On the other hand, cancer cells proliferate faster than normal cells regardless of the needs of their host, which is frequently killed once cancer metastasizes. It can thus be hypothesized under this theory that a necessary condition for the onset of cancer is that the intrinsic, higher-order constraint emergent from the synergistic coupling of Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ across the extracellular space $\boldsymbol{S_E}$ has dissipated in cancer cells. As constraint is not intrinsically dependent on any molecular substrate for its embodiment but multiply realizable, its dissipation may be very difficult to account for with a single event such as a specific genetic mutation. Because cancer cells divide quickly, proliferation-generated gradients in $\boldsymbol{S_E}$ (in particular, $\overrightarrow{\nabla}\Phi_N$ gradients) will be observable with high probability. Thus, in cancer cells the dissipation of the synergistic coupling of Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$ could only be accounted for by (i) the otherwise functional Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ lacking their defining ability to elicit changes in Nanney's embodiers $\boldsymbol{F_N}$ within other cells' nuclei, (ii) no elements of the molecular nuclear phenotype being specified as Waddington's embodiers $\boldsymbol{F_W}$ and Nanney's embodiers $\boldsymbol{F_N}$ at the same time such as histone H3 modifications, preventing the coupling of Waddington's constraints $\boldsymbol{C_W}$ and Nanney's constraints $\boldsymbol{C_N}$, and/or (iii) Waddington's embodiers $\boldsymbol{F_N}$ not being able to constrain any longer—via gene expression and function—the facilitated diffusion/signal transduction of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ into the cells (e.g., the expression or function of critical protein channels, carriers, membrane receptors, or intracellular transducers is suppressed or impaired).

Concluding remarks

Here, I show that scientifically tenable teleology in nature can emerge only from local and level-of-scale-specific thermodynamic boundary conditions (i.e., constraints) that are *stochastically independent* with respect to each other at certain critical sites such as those for histone post-translational modifications in TSS-adjacent genomic regions. The only way such requisite of stochastic independence can be fulfilled intrinsically is if a higher-order constraint emerges from the synergistic coupling of lower-order constraint generating systems, an emergent transition first proposed 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, machine, or "self-organizing" system of any type as such systems entail an *explicit deterministic or stochastic dependence* 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 intrinsic constraints or *selves* in the natural world; *selves* whom no mechanism, machine, or "self-organizing" system could ever be.

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 (University of California, Santa Cruz) database [64]. Publicly available tandem datafiles of ChIP-seq (comprising 1×36 bp, 1×50 bp, and 1×75 bp reads, depending on the data series; details including GEO accession codes can be found in Supplementary Information) on histone H3 modifications and RNA-seq (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 Supporting Information) for each analyzed cell sample in each species were downloaded from the ENCODE, modENCODE or the SRA (Sequence Read Archives) database of the National Center for Biotechnology Information [65, 66, 67, 68, 69, 70, 71].

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 cell type/sample datasets included in this work for computing *ctalk_non_epi* and mRNA abundance profiles 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.

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) [24]—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 10 kbp 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 (Binary Sequence Alignment/Map) and BAI (BAM Index) 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 file (the BAI file is required implicitly) to generate the corresponding read count file in BED (Browser Extensible Data) format.

RNA-seq data processing

The processed data were mRNA abundances in FPKM at RefSeq TSSs. When the original file format was GTF (Gene transfer Format) 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 (Sequence Alignment/Map), 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* [72] directly on the original file with the following three options:

```
-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. When replicates were available the measure of central tendency used was the median of the replicate FPKM values.

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 $N_{\text{TSS}_{\text{def}}}(\textit{Homo sapiens}) = 14{,}742$; $N_{\text{TSS}_{\text{def}}}(\textit{Mus musculus}) = 16{,}021$; and $N_{\text{TSS}_{\text{def}}}(\textit{Drosophila melanogaster}) = 11{,}632$. 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:

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

The tables were then written to the following 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∎

Computation of *ctalk_non_epi* profiles

If the variables X_i (representing the signal for histone H3 modification X in the genomic bin $i \in \{\text{“}-2000\text{”},\ldots,\text{“}3800\text{”}\}$), Y_j (representing the signal for histone H3 modification Y in the genomic bin $j \in \{\text{“}-2000\text{”},\ldots,\text{“}3800\text{”}\}$) and Z (representing $\log_2$-transformed FPKM values) are random variables, then the covariance of X_i and Y_j can be decomposed in terms of their linear relationship with Z as follows:

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

where the second summand $\mathrm{Cov}(X_i, Y_j|Z)$ is the partial covariance between X_i and Y_j given Z. It is easy to see that $\mathrm{Cov}(X_i, Y_j|Z)$ is a local approximation of Nanney's constraints $\boldsymbol{C_N}$ on histone H3 modifications, as anticipated in the preliminary theoretical definitions (a straightforward corollary is that Waddington's constraints $\boldsymbol{C_W}$ can in turn be approximated by $\frac{\mathrm{Cov}(X_i,Z)\,\mathrm{Cov}(Y_j,Z)}{\mathrm{Var}(Z)}$). To make the *ctalk_non_epi* profiles comparable, however, $\mathrm{Cov}(X_i, Y_j|Z)$ values had to be normalized by the standard deviations of the residuals of X_i and Y_j with respect to Z. In other words, the partial correlation $\mathrm{Cor}(X_i, Y_j|Z)$ values were needed. Nevertheless, a correlation value does not have a straightforward interpretation, whereas its square—typically known as *coefficient of determination*, *strength 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, $\mathrm{Cor}(X_i, Y_j|Z)^2$ 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_i, Y_j and Z data, each pairwise combination of bin-specific histone H3 modifications $\{X_i, Y_j\}$ contributed with the value

$$ctalk\_non\_epi(X_i, Y_j) = \underbrace{\mathrm{sgn}\Big(\mathrm{Cor}(X_i, Y_j|Z)\Big)}_{\substack{\text{partial correlation} \\ \text{sign} \in \{-1, 1\}}} \underbrace{\Big(\mathrm{Cor}(X_i, Y_j|Z)\Big)^2}_{\substack{\text{partial correlation} \\ \text{strength} \in [0, 1]}}. \tag{2}$$

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.

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 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 [73, 74]. For these reasons, in each species all pairwise combinations of post-translational modifications involving the same amino acid residue in the H3 histone were then identified as "trivial" and excluded from the *ctalk_non_epi* profiles construction. E.g., for ***Mus musculus*** cell-type datasets the histone modifications comprised were H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3 (i.e., $n_m = 5$), then the combinations H3K4me1–H3K4me3 and H3K27ac–H3K27me3 were filtered out. Therefore, the length of the *ctalk_non_epi* profiles for ***Mus musculus*** was $({}^5C_2 - 2) \times 900 = 7{,}200$.

Statistical significance assessment

The statistical significance of the partial correlation $\mathrm{Cor}(X_i, Y_j|Z)$ values, necessary for constructing the *ctalk_non_epi* profiles, was estimated using Fisher's z-transformation [75]. Under the null hypothesis $\mathrm{Cor}(X_i, Y_j|Z) = 0$ the statistic $z = \sqrt{N_{\mathrm{TSS}_{\mathrm{def}}} - |Z| - 3}\ \frac{1}{2} \ln\left(\frac{1+\mathrm{Cor}(X_i, Y_j|Z)}{1-\mathrm{Cor}(X_i, Y_j|Z)}\right)$, where $N_{\mathrm{TSS}_{\mathrm{def}}}$ is the sample size and $|Z| = 1$ (i.e., one control variable), follows asymptotically a $N(0,1)$ distribution. The p-values can then be computed easily using the $N(0,1)$ probability function.

Multiple comparisons correction of the p-values associated with each *ctalk_non_epi* profile was performed using the Benjamini-Yekutieli method [76]. The parameter used was the number of all possible comparisons (i.e., ***before*** excluding "trivial" pairwise combinations of histone H3 modifications, to further increase the conservativeness of the correction): ${}^{(n_m \times 30)}C_2$. From the resulting q-values associated with 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 $q\text{-value}[i] \geq t$ then the associated partial correlation value was identified as not significant (i.e., equal to zero) in the respective *ctalk_non_epi* profile.

Hierarchical cluster analysis 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 hierarchical cluster analysis (HCA). For each species, HCA was performed on (i) the *ctalk_non_epi* profiles of each cell type/sample (Figure 3A, 3C, and 3E) and (ii) the $\log_2$-transformed FPKM profiles (i.e., mRNA abundance) of each cell type/sample (Figure 3B, 3D, and 3F). 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, hence it was the metric used here (as a consequence of what was highlighted previously, the "correlation distance" metric is also invariant under linear transformations of the profiles). On the other hand, the cluster-linkage method chosen was the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) or "average" method 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 [77]). Cluster statistical significance was assessed as *au* (approximately unbiased) and *bp* (bootstrap probability) significance scores by nonparametric bootstrap resampling using the *Pvclust* [28] add-on package for the *R* software [78]. The number of bootstrap replicates in each analysis was 10,000.

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 and RPKM (reads per kilobase of transcript per million reads mapped) may introduce significant 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 [79]. However, this issue was not a problem for this study.

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 and RPKM (reads per kilobase of transcript per million reads mapped) may introduce significant 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 [79]. However, this issue was not a problem for the study because partial correlation, used to construct the *ctalk_non_epi* profiles, is invariant under linear transformations of the control variable Z (i.e., $\mathrm{Cor}(X, Y|Z) = \mathrm{Cor}(X, Y|aZ+b)$ 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 [73]). Similarly, hierarchical clustering of mRNA abundance profiles is invariant under linear transformations of the profiles, because $\mathrm{Cor}(Z_i, Z_j) = \mathrm{Cor}(aZ_i+b, cZ_j+d)$ (provided $ac > 0$).

Acknowledgements

I wish to thank Angelika H. Hofmann for editing this paper into an English I could only dream of writing. I am especially grateful to John Tyler Dodge, horn soloist at the *Orquesta Filarmónica de Santiago*, for reviewing the English of the very first complete draft of this paper and his valuable questions, which pushed me to the limit of my abilities in the purpose of making the theoretical description self-explanatory. For reviewing this paper and their valuable comments I am indebted to Álvaro Glavic, Óscar M. Lazo, Inti Pedroso, Iván Sellés, and José Monserrat Neto. My thanks also extend to Alejandro Maass and Kenneth M. Weiss for their interest in this work and their valuable questions and to my anonymous colleagues who reviewed my grant proposal on behalf of FONDECYT.

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, paper writing, or the decision to submit this pre-print to *bioRxiv*.

Copyright

Funding

Funder	Grant reference number	Author
Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT, Chile)	3140328	Felipe A. Veloso

References

[1] Slack JMW. Timeline: Conrad Hal Waddington: the last Renaissance biologist? Nat Rev Genet. 2002 nov;3(11):889–895. Available from: http://dx.doi.org/10.1038/nrg933.

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

[3] Wolffe AP. Epigenetics: Regulation Through Repression. Science. 1999 oct;286(5439):481–486. Available from: http://dx.doi.org/10.1126/science.286.5439.481.

[4] Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. Science. 2010 Oct;330(6004):612–6. Available from: http://dx.doi.org/10.1126/science.1191078.

[5] Kamakura M. Royalactin induces queen differentiation in honeybees. Nature. 2011 apr;473(7348):478–483. Available from: http://dx.doi.org/10.1038/nature10093.

[6] Fraser P. Defining Epigenetics. Interviews by G. Riddihough; 2010. Science [Video podcast] 00:05:34-00:05:47. Available from: http://videolab.sciencemag.org/featured/650920373001/1.

[7] Orphanides G, Reinberg D. A Unified Theory of Gene Expression. Cell. 2002 feb;108(4):439–451. Available from: http://dx.doi.org/10.1016/s0092-8674(02)00655-4.

[8] Li G, Reinberg D. Chromatin higher-order structures and gene regulation. Current Opinion in Genetics & Development. 2011 apr;21(2):175–186. Available from: http://dx.doi.org/10.1016/j.gde.2011.01.022.

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

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

[11] Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes & Development. 2009 apr;23(7):781–783. Available from: http://dx.doi.org/10.1101/gad.1787609.

[12] Varela FG, Maturana HR, Uribe R. Autopoiesis: the organization of living systems, its characterization and a model. Biosystems. 1974;5(4):187–196. Available from: http://dx.doi.org/10.1016/0303-2647(74)90031-8.

[13] Abel DL, Trevors JT. Self-organization vs. self-ordering events in life-origin models. Physics of Life Reviews. 2006 dec;3(4):211–228. Available from: http://dx.doi.org/10.1016/j.plrev.2006.07.003.

[14] Reinberg D. Defining Epigenetics. Interviews by G. Riddihough; 2010. Science [Video podcast] 00:01:25-00:01:35. Available from: http://videolab.sciencemag.org/featured/650920373001/1.

[15] Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. Development. 1997 May;124(10):1851–64. Available from: http://dev.biologists.org/content/124/10/1851.long

[16] Maduro MF. Cell fate specification in the *C. elegans* embryo. Dev Dyn. 2010 May;239(5):1315–29. Available from: http://dx.doi.org/10.1002/dvdy.22233

[17] Altun Z, Hall D. WormAtlas. Altun ZF, Herndon LA, Crocker C, Lints R, Hall DH, (ed.s) 2002-2015. Available from: http://www.wormatlas.org/.

[18] Warner DA, Shine R. The adaptive significance of temperature-dependent sex determination in a reptile. Nature. 2008 jan;451(7178):566–568. Available from: http://dx.doi.org/10.1038/nature06519.

[19] Power ML, Schulkin J. Maternal regulation of offspring development in mammals is an ancient adaptation tied to lactation. Applied & Translational Genomics. 2013 dec;2:55–63. Available from: http://dx.doi.org/10.1016/j.atg.2013.06.001.

[20] Ladewig J, Koch P, Brüstle O. Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. Nature Reviews Molecular Cell Biology. 2013 mar;14(4):225–236. Available from: http://dx.doi.org/10.1038/nrm3543.

[21] Nanney DL. Epigenetic control systems. Proceedings of the National Academy of Sciences. 1958 jul;44(7):712–717. Available from: http://dx.doi.org/10.1073/pnas.44.7.712.

[22] Huang S. The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? Bioessays. 2012 Feb;34(2):149–57. Available from: http://dx.doi.org/10.1002/bies.201100031.

[23] Losick R, Desplan C. Stochasticity and cell fate. Science. 2008;320(5872):65–68. Available from: http://dx.doi.org/10.1126/science.1147888.

[24] Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotechnol. 2013 jun;31(7):615–622. Available from: http://dx.doi.org/10.1038/nbt.2596.

[25] White KP. Microarray Analysis of *Drosophila* Development During Metamorphosis. Science. 1999 dec;286(5447):2179–2184. Available from: http://dx.doi.org/10.1126/science.286.5447.2179.

[26] Cantera R, Ferreiro MJ, Aransay AM, Barrio R. Global Gene Expression Shift during the Transition from Early Neural Development to Late Neuronal Differentiation in *Drosophila melanogaster*. PLoS ONE. 2014 may;9(5):e97703. Available from: http://dx.doi.org/10.1371/journal.pone.0097703.

[27] Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, et al. Genome-wide gene expression profiles of the developing mouse hippocampus. Proceedings of the National Academy of Sciences. 2001 jul;98(15):8862–8867. Available from: http://dx.doi.org/10.1073/pnas.141244998.

[28] Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006 apr;22(12):1540–1542. Available from: http://dx.doi.org/10.1093/bioinformatics/btl117.

[29] Chen L, Xiao S, Pang K, Zhou C, Yuan X. Cell differentiation and germ–soma separation in Ediacaran animal embryo-like fossils. Nature. 2014 sep;516(7530):238–241. Available from: http://dx.doi.org/10.1038/nature13766.

[30] Meyerowitz EM. Plants Compared to Animals: The Broadest Comparative Study of Development. Science. 2002 feb;295(5559):1482–1485. Available from: http://dx.doi.org/10.1126/science.1066609.

[31] Nielsen C. Six major steps in animal evolution: are we derived sponge larvae? Evolution & Development. 2008 mar;10(2):241–257. Available from: http://dx.doi.org/10.1111/j.1525-142x.2008.00231.x.

[32] Deacon TW. Incomplete nature: How mind emerged from matter. 1st ed. New York: W.W. Norton & Co.; 2012.

[33] Donoghue PCJ, Antcliffe JB. Early life: Origins of multicellularity. Nature. 2010 jul;466(7302):41–42. Available from: http://dx.doi.org/10.1038/466041a.

[34] Turing AM. The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 1952;237(641):37–72. Available from: http://www.dna.caltech.edu/courses/cs191/paperscs191/turing.pdf

[35] Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007 may;447(7143):413–417. Available from: http://dx.doi.org/10.1038/nature05916.

[36] Deacon T, Koutroufinis S. Complexity and Dynamical Depth. Information. 2014 jul;5(3):404–423. Available from: http://dx.doi.org/10.3390/info5030404.

[37] van Nimwegen E. Scaling laws in the functional content of genomes. Trends in Genetics. 2003 sep;19(9):479–484. Available from: http://dx.doi.org/10.1016/s0168-9525(03)00203-8.

[38] Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Research. 2007 dec;36(Database):D88–D92. Available from: http://dx.doi.org/10.1093/nar/gkm964.

[39] Haeckel E. Die Gastraea-Theorie, die phylogenetische Classification des Thierreichs und die Homologie der Keimblätter. Jenaische Zeitschrift für Naturwissenschaft. 1874;8:1–55.

[40] Kirk DL. A twelve-step program for evolving multicellularity and a division of labor. Bioessays. 2005;27(3):299–310. Available from: http://dx.doi.org/10.1002/bies.20197.

[41] Willensdorfer M. On the evolution of differentiated multicellularity. Evolution. 2009 Feb;63(2):306–23. Available from: http://dx.doi.org/10.1111/j.1558-5646.2008.00541.x

[42] Mikhailov KV, Konstantinova AV, Nikitin MA, Troshin PV, Rusin LY, Lyubetsky VA, et al. The origin of Metazoa: a transition from temporal to spatial cell differentiation. Bioessays. 2009 Jul;31(7):758–68. Available from: http://dx.doi.org/10.1002/bies.200800214

[43] Gavrilets S. Rapid Transition towards the Division of Labor via Evolution of Developmental Plasticity. PLoS Computational Biology. 2010 jun;6(6):e1000805. Available from: http://dx.doi.org/10.1371/journal.pcbi.1000805.

[44] Levin TC, Greaney AJ, Wetzel L, King N. The rosetteless gene controls development in the choanoflagellate *S. rosetta*. eLife. 2014 oct;3. Available from: http://dx.doi.org/10.7554/elife.04070.

[45] Young HE, Black AC. Adult stem cells. Anat Rec. 2003;276A(1):75–102. Available from: http://dx.doi.org/10.1002/ar.a.10134.

[46] Cai S, Fu X, Sheng Z. Dedifferentiation: A New Approach in Stem Cell Research. BioScience. 2007;57(8):655. Available from: http://dx.doi.org/10.1641/b570805.

[47] Donà E, Barry JD, Valentin G, Quirin C, Khmelinskii A, Kunze A, et al. Directional tissue migration through a self-generated chemokine gradient. Nature. 2013 Nov;503(7475):285–9. Available from: http://dx.doi.org/10.1038/nature12635

[48] Venkiteswaran G, Lewellis SW, Wang J, Reynolds E, Nicholson C, Knaut H. Generation and Dynamics of an Endogenous, Self-Generated Signaling Gradient across a Migrating Tissue. Cell. 2013 oct;155(3):674–687. Available from: http://dx.doi.org/10.1016/j.cell.2013.09.046.

[49] Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003;4(9):117. Available from: http://genomebiology.com/2003/4/9/117

[50] Ning K, Fermin D, Nesvizhskii AI. Comparative Analysis of Different Label-Free Mass Spectrometry Based Protein Abundance Estimates and Their Correlation with RNA-Seq Gene Expression Data. J Proteome Res. 2012 apr;11(4):2261–2271. Available from: http://dx.doi.org/10.1021/pr201052x.

[51] Meletis K, Barnabé-Heider F, Carlén M, Evergren E, Tomilin N, Shupliakov O, et al. Spinal Cord Injury Reveals Multilineage Differentiation of Ependymal Cells. Plos Biol. 2008;6(7):e182. Available from: http://dx.doi.org/10.1371/journal.pbio.0060182.

[52] Sømme L. Supercooling and winter survival in terrestrial arthropods. Comparative Biochemistry and Physiology Part A: Physiology. 1982 jan;73(4):519–543. Available from: http://dx.doi.org/10.1016/0300-9629(82)90260-2.

[53] Murphy WJ, Collier GE. A molecular phylogeny for aplocheiloid fishes (Atherinomorpha, Cyprinodontiformes): the role of vicariance and the origins of annualism. Molecular Biology and Evolution. 1997;14(8):790–799. Available from: http://dx.doi.org/10.1093/oxfordjournals.molbev.a025819

[54] Niethammer P, Grabher C, Look AT, Mitchison TJ. A tissue-scale gradient of hydrogen peroxide mediates rapid wound detection in zebrafish. Nature. 2009 jun;459(7249):996–999. Available from: http://dx.doi.org/10.1038/nature08119.

[55] Unsworth BR, Lelkes PI. Growing tissues in microgravity. Nat Med. 1998 aug;4(8):901–907. Available from: http://dx.doi.org/10.1038/nm0898-901.

[56] Correll MJ, Pyle TP, Millar KDL, Sun Y, Yao J, Edelmann RE, et al. Transcriptome analyses of Arabidopsis thaliana seedlings grown in space: implications for gravity-responsive genes. Planta. 2013 jun;238(3):519–533. Available from: http://dx.doi.org/10.1007/s00425-013-1909-x.

[57] Hammond TG, Lewis FC, Goodwin TJ, Linnehan RM, Wolf DA, Hire KP, et al. Gene expression in space. Nature Medicine. 1999 apr;5(4):359–359. Available from: http://dx.doi.org/10.1038/7331.

[58] Crawford-Young SJ. Effects of microgravity on cell cytoskeleton and embryogenesis. The International Journal of Developmental Biology. 2006;50(2-3):183–191. Available from: http://dx.doi.org/10.1387/ijdb.052077sc.

[59] Pojman JA, Bessonov N, Volpert V, Paley MS. Miscible Fluids in Microgravity (MFMG): A zero-upmass investigation on the International Space Station. Microgravity Sci Technol. 2007 dec;19(1):33–41. Available from: http://dx.doi.org/10.1007/bf02870987.

[60] Rajjou L, Duval M, Gallardo K, Catusse J, Bally J, Job C, et al. Seed Germination and Vigor. Annu Rev Plant Biol. 2012 jun;63(1):507–533. Available from: http://dx.doi.org/10.1146/annurev-arplant-042811-105550.

[61] Finch-Savage WE, Leubner-Metzger G. Seed dormancy and the control of germination. New Phytologist. 2006 jul;171(3):501–523. Available from: http://dx.doi.org/10.1111/j.1469-8137.2006.01787.x.

[62] Smet ID, Beeckman T. Asymmetric cell division in land plants and algae: the driving force for differentiation. Nature Reviews Molecular Cell Biology. 2011 mar;12(3):177–188. Available from: http://dx.doi.org/10.1038/nrm3064.

[63] Bissell MJ, LaBarge MA. Context, tissue plasticity, and cancer. Cancer Cell. 2005 jan;7(1):17–23. Available from: http://dx.doi.org/10.1016/j.ccr.2004.12.013.

[64] Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Research. 2004 jan;32(90001):493D–496. Available from: http://dx.doi.org/10.1093/nar/gkh103.

[65] Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, et al. Unlocking the secrets of the genome. Nature. 2009 jun;459(7249):927–930. Available from: http://dx.doi.org/10.1038/459927a.

[66] Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, et al. Combinatorial Patterning of Chromatin Regulators Uncovered by Genome-wide Location Analysis in Human Cells. Cell. 2011 dec;147(7):1628–1639. Available from: http://dx.doi.org/10.1016/j.cell.2011.09.057.

[67] Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map of the *Drosophila* genome. Nature. 2011 mar;471(7339):527–531. Available from: http://dx.doi.org/10.1038/nature09990.

[68] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 sep;489(7414):57–74. Available from: http://dx.doi.org/10.1038/nature11247.

[69] Xiao S, Xie D, Cao X, Yu P, Xing X, Chen CC, et al. Comparative Epigenomic Annotation of Regulatory DNA. Cell. 2012 jun;149(6):1381–1392. Available from: http://dx.doi.org/10.1016/j.cell.2012.04.029.

[70] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 sep;489(7414):101–108. Available from: http://dx.doi.org/10.1038/nature11233.

[71] Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. 2012;13(8):418. Available from: http://dx.doi.org/10.1186/gb-2012-13-8-418.

[72] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 may;28(5):511–515. Available from: http://dx.doi.org/10.1038/nbt.1621.

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

[74] Peach SE, Rudomin EL, Udeshi ND, Carr SA, Jaffe JD. Quantitative assessment of chromatin immunoprecipitation grade antibodies directed against histone modifications reveals patterns of co-occurring marks on histone protein molecules. Mol Cell Proteomics. 2012 May;11(5):128–37. Available from: http://dx.doi.org/10.1074/mcp.m111.015941.

[75] Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;p. 507–521. Available from: http://dx.doi.org/10.2307/2331838

[76] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of statistics. 2001;p. 1165–1188. Available from: http://www.jstor.org/stable/2674075.

[77] Saraçli S, Doğan N, Doğan I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of Inequalities and Applications. 2013;2013(1):203. Available from: http://dx.doi.org/10.1186/1029-242x-2013-203.

[78] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available from: http://www.R-project.org/.

[79] Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012 aug;131(4):281–285. Available from: http://dx.doi.org/10.1007/s12064-012-0162-3.

[80] Hadzi J. The evolution of the Metazoa. Macmillan; 1963.

[81] Tompkins N, Li N, Girabawe C, Heymann M, Ermentrout GB, Epstein IR, et al. Testing Turing's theory of morphogenesis in chemical cells. Proceedings of the National Academy of Sciences. 2014;111(12):4397–4402. Available from: http://dx.doi.org/10.1073/pnas.1322005111.

[82] Kupiec JJ. A Darwinian theory for the origin of cellular differentiation. Molecular and General Genetics MGG. 1997 jun;255(2):201–208. Available from: http://dx.doi.org/10.1007/s004380050490.

[83] Paldi A. What makes the cell differentiate? Prog Biophys Mol Biol. 2012 Sep;110(1):41–3. Available from: http://dx.doi.org/10.1016/j.pbiomolbio.2012.04.003.

[84] Kim J. Making Sense of Emergence. Philosophical Studies. 1999;95(1/2):3–36. Available from: http://dx.doi.org/10.1023/a:1004563122154.

[85] Kim J. Emergence: Core ideas and issues. Synthese. 2006 aug;151(3):547–559. Available from: http://dx.doi.org/10.1007/s11229-006-9025-0.

[86] Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948 oct;27(4):623–656. Available from: http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Appendix

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

Since Ernst Haeckel's "gastraea theory" [39], 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 [80]) cell colonies or aggregations [40, 31, 41, 42, 43, 44] while the developmental process starts from one or a few primordial cells (zygotes, spores, or buds) 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.

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 [34]. Although this model has been tested in terms of chemical differentiation of synthetic "cells" [81], 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 [82]. A third "non-epigenetic" hypothesis is that cell fate decisions are the result of the characteristic coupling of gene expression and metabolism [83].

All of these accounts, however, fail to (i) 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, and/or (ii) 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, and/or (iii) explain the reproducible and robust self-regulatory dynamics of gene expression during cell differentiation. These approaches also do not describe in an objective and unambiguous way the transition or difference between a highly complex or symbiotic cell population/aggregation and a differentiated multicellular individual, and they lack 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 (fundamentally refuted by Jaegwon Kim [84, 85]) but instead converged from the strict *stochastically-independent-dynamics* condition argued in the introduction into what can be described as the constraint-based concept of emergence of unprecedented, higher-order teleological systems, pioneered in a broader

perspective by Terrence Deacon in 2011 [32]. Importantly, this formulation of emergence does not build upon 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.

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 [17])

	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.

Stochastically independent dynamics

Let $X_1,\ldots,X_n$ be n discrete random variables representing certain dynamics of n thermodynamic systems respectively.

If $H(X_i)$ is the Shannon entropy of X_i, then by joint Shannon entropy subadditivity [86], it is always true that

$$H(X_1,\ldots,X_n) \leq H(X_1)+\ldots+H(X_n).$$

These n dynamics will be *stochastically independent* if and only if

$$H(X_1,\ldots,X_n) = H(X_1)+\ldots+H(X_n)$$

or, in other words, if these n dynamics are explicitly uncorrelated.

If the relationships are linear, the condition of *stochastically independent* dynamics for two systems can be expressed as:

$$\mathrm{Cov}(X_i,X_j) = 0$$

In general, the covariance of two random variables $\{X_i,X_j\}$ can be decomposed in terms of their linear relationship with a third random variable X_k as follows:

$$\mathrm{Cov}(X_i,X_j) = \underbrace{\frac{\mathrm{Cov}(X_i,X_k)\,\mathrm{Cov}(X_j,X_k)}{\mathrm{Var}(X_k)}}_{\substack{\text{covariance of } X_i \text{ and } X_j \\ \text{resulting from their} \\ \text{linear relationship with } X_k}} + \underbrace{\mathrm{Cov}(X_i,X_j|X_k).}_{\substack{\text{covariance of } X_i \text{ and } X_j \\ \text{orthogonal to } X_k}} \tag{5}$$

Importantly, the dynamics of the two systems that respectively account for each summand above are *stochastically independent* (see also Figure 2 and Methods).

$$\mathrm{Cov}(X_i,Y_j) = \underbrace{\frac{\mathrm{Cov}(X_i,Z)\,\mathrm{Cov}(Y_j,Z)}{\mathrm{Var}(Z)}}_{\substack{\text{covariance of } X_i \text{ and } Y_j \\ \text{resulting from their} \\ \text{linear relationship with } Z}} + \underbrace{\mathrm{Cov}(X_i,Y_j|Z).}_{\substack{\text{covariance of } X_i \text{ and } Y_j \\ \text{orthogonal to } Z}} \tag{6}$$

Supplementary Information

Homo sapiens source data of ChIP-seq on histone H3 modifications (BAM/BAI files) [66]

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

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
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
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

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
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
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

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
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	H3K4mel	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep1.bam.bai

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
HUVEC	H3K4me1	GSM733690	wgEncodeBroadHistoneHuvecH3k4me1StdAlnRep1.bam
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

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
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
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

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
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
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

Continued on next page

Continued from previous page

Cell type	Antibody	GEO Accession	File URL suffix
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) [70]

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

Continued on next page

Continued from previous page

Cell type	GEO Accession	File URL suffix
HUVEC	GSM958734	wgEncodeCaltechRnaSeqHuvecR2x75I1200TSSRep2V3.gtf.gz
NHEK	GSM958736	wgEncodeCaltechRnaSeqNhekR2x75I1200TSSRep1V3.gtf.gz
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) [71, 69]

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

Continued on next page

Continued from previous page

Cell type	Antibody	Rep #	GEO Accession	File URL suffix
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
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) [71, 69]

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) [65, 67]

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
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

Continued on next page

Continued from previous page

Developmental time point/period	Antibody	GEO Accession	File URL suffix
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
Pupae	H3K4me1	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) [65, 67]

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`