bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

There is wide scientific consensus on the relevance of changes in the levels of gene expression for the cell differentiation process. Furthermore, research in the field has customarily assumed that such changes regulate this process when they interconnect in space and time by means of complex epigenetic mechanisms. Nevertheless, this assumed regulatory power lacks a clear definition and may even lead to logical inconsistencies. To tackle this problem, I analyzed publicly available high-throughput data of histone H3 post-translational modifications and mRNA abundance for different *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster* cell samples. Comprising genomic regions adjacent to transcription start sites, this analysis generated for each cell dataset a profile from pairwise partial correlations between histone modifications controlling for the respective mRNA levels. Here I report that these profiles, while explicitly uncorrelated to transcript abundance by construction, associate strongly with cell differentiation states. This association is not to be expected if cell differentiation is, in effect, regulated by epigenetic changes of gene expression. Based on these results, I postulate in this paper a falsifiable theory of differentiated multicellularity. This theory describes how the differentiated multicellular organism—understood as an intrinsic, higher-order, self-sufficient, self-repairing, self-replicating, and self-regulating dynamical constraint—emerges from proliferating undifferentiated cells. If it survives falsification tests consistently this theory would explain in principle (i) the self-regulated gene transcriptional changes during ontogeny and (ii) the emergence of differentiated multicellular lineages throughout evolution.

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The X-files of chromatin

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

In practice however, system predictability has not been achieved yet, and research has rather focused on "reverse engineering" the ontogenetic process from experimental results. Still, this strategy has yielded remarkable results such as the induction of pluripotent stem cells (iPSCs) [3]. In terms of explainability, the dynamics of the cell differentiation process have been associated to changes in chromatin states and concurrent heritable changes in gene expression levels, which have been defined in turn as epigenetic changes [4, 5]). In some cases these changes can be generated extrinsically with respect to the developing organism, as clearly 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 [6]). Nevertheless, most changes of gene expression during cell differentiation are not only independent from, but are even robust with respect to extrinsic changes. This means that ontogeny is fundamentally an intrinsically regulated process, for which no falsifiable theory has emerged from the epigenetic framework since it was first advanced. Moreover, Peter Fraser has recently referred to this problem as "The X-files of chromatin" [7].

This research work was conceived and designed to, following Fraser's metaphor, declassify "The X-files of chromatin". In its initial phase, I conducted a computational analysis of the least relevant—for the epigenetic landscape—constraints on histone H3 post-translational modification states. Before outlining this analysis however, I must present here a case for the fundamental impossibility of explaining the cell differentiation self-regulatory dynamics under the framework pioneered by Waddington, however complex its underlying mechanisms may be (as also hinted by Fraser [7]). Only then will I be able to argue that these epigenetically irrelevant constraints on histone modification states are, in fact, key to a full understanding of differentiated multicellularity in terms of its self-regulation and evolution.

The conundrum of self-regulation

Avoiding non-explanatory teleological descriptions, modern science regards cell differentiation fundamentally as a dynamical system, where a fixed rule governs the transition between the realizable states of a complex network of molecular mechanisms. Ranging from low-order molecular interactions [8] to chromatin higher-order structural changes [9, 10], these mechanisms propagate changes of gene expression in different loci as cells proliferate. Both heritable

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

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and uncorrelated to changes in the DNA sequence, these changes (defined as epigenetic as mentioned previously) would in turn regulate cell differentiation. Furthermore, and although all epigenetic mechanisms involved in cell differentiation are far from being completely elucidated, the hypothesis that cell differentiation is regulated by epigenetic changes of gene expression is routinely presented to the general public as a well-established scientific fact (as illustrated in [11]). However, this hypothesis—whether or not we accept it in its strict sense—leads to severe explanatory limitations and may even entail logical inconsistencies.

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

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

On the other hand, to take the epigenetic-changes-regulate hypothesis in a loose sense is to use "intrinsic regulation" only as a placeholder when referring to a certain class of molecular mechanisms. If this is the case, we must note that any scientifically tenable mechanism requires that the changes it comprises are at least dynamically correlated. In this context, an epigenetic mechanism can be seen metaphorically as toppling dominoes (here the dynamically correlated events are obvious). But as pointed out previously, this mechanism, however numerous or intricately connected its correlated changes, says nothing about how the first domino tile (or any other whose fall is not attributable to the fall of other tiles) was toppled over. To fill this explanatory gap, it has been proposed that an "epigenator"—defined operationally as a transient signal which probably originates in the environment of the cell—triggers the epigenetic phenotype change after being transduced into the intracellular space [12]. Nonetheless, if all "epigenators" in the system are extrinsic to it, by definition intrinsic regulation cannot be explained. On the other hand, if there is at least one intrinsic "epigenator" in the system (e.g. a suggested "extracellular signal") its critical signaling property is left unexplained.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Importantly, these problems are inherent to *any* dynamical systems model intended to account for the self-regulatory dynamics of cell differentiation. This is because any system able to explain intrinsic "regulation" must be dynamically uncorrelated to the changes it "regulates"; otherwise the "regulator" is, fundamentally, just another domino tile that propagates changes regardless of its relative position. At this point the explanatory dead end becomes evident. Under the traditional approach in developmental biology no higher-order system within a living organism, however complex (e.g. displaying interlocked feedback loops or hypercyclic networks), exerts true intrinsic regulation because its dynamics are ultimately correlated to the lower-order dynamics it is supposed to regulate. Furthermore, in the epigenetic landscape any "intrinsic higher-order regulator" can be no more than an epiphenomenon: a causally inefficacious system—whether or not linear or predictable—resulting from molecular dynamics at the lowest level of scale.

Epigenetic information in theory and practice

Regardless of the explanatory limitations inherent to the traditional dynamical systems approach in developmental biology, either all necessary information for cell differentiation is already contained in the zygote or it is not. This dichotomy may seem to be trivial but important implications follow it.

If the zygote contains all necessary information [13, 14], the previously discussed explanatory gap could, in principle, be filled. Epigenetic imprinting, shown able to resolve a few early lineage commitments in *Caenorhabditis elegans* [15], supports this possibility at first glance. Nevertheless, a closer look at the complexity of this simple metazoan model suggests otherwise: *C. elegans* ontogeny yields 19 different cell types (excluding the germ line) in a total of 1,090 generated cells. From these two parameters alone, the required information capacity for the entire process can be estimated to be at least 983 bit (see details in Appendix). Further, this is a great underestimation since cell-fate uncertainty remains with respect two more variables at least, namely space and time. In effect, cell-fate decisions are made in specific regions within the organism and/or involve specific migration paths, and they are made in specific time points during differentiation. Therefore, the non-genetic information capacity necessary for the entire process far exceeds the few bits of information that epigenetic imprinting can account for.

Information not only requires a medium for its storage and transmission but also must have content which, in this context, resolves the fate of every cell: apoptosis before division, division without differentiation, or division with differentiation. Here an additional problem appears: stem cell potency. An entire organism can develop (including extraembryonic tissues) from *any* totipotent stem cell, and all embryonic tissues can develop from *any* pluripotent stem cell. How is this possible if cell fate decisions are already specified deterministically in the zygote? The recently proposed—yet not explanatory—"epigenetic disc" model for cell differentiation, under which the pluripotent state is only one among many metastable and directly interconvertible states, reflects the necessity to account for the context-dependent character of cell fate information [16].

With remarkable insight, in 1958 David L. Nanney anticipated explanatory pitfalls if the definition of epigenetics is limited to heritable changes. He further stated that "'cellular memory' is not an absolute attribute" [17]; or, in other words, that more important to development is

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the process by which heritable material may manifest different phenotypes than the heritable material itself. However, Waddington's epigenetic landscape prevailed and the field reinforced a "preinformationist" framework: although the zygote is not a complete miniature version of the mature organism (preformationism), it is indeed a complete blueprint of the mature organism (allowing for some degree of extrinsic control, as in eusocial insects [6] and stochastic gene expression [18]). If this is correct, we must also accept that in the mature human brain—indisputably, one among many products of the developmental process—there is strictly less non-genetic, non-redundant information than in the human zygote (not surprisingly however, I failed to find a single research paper with such a proposition).

This *reductio ad absurdum* shows that the traditional dynamical systems approach (i.e. the epigenetic landscape in developmental biology) has forced research to ignore or reject the necessary *emergence* of not only some, but possibly most information content during ontogeny. Specifically, if additional information content emerges during brain development, what would necessarily preclude information content from emerging in proliferating undifferentiated cells?

A proof-of-principle hypothesis

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

The specific objects of study were observed combinatorial constraints on histone H3 post-translational modifications (also known simply as histone H3 crosstalk). These modifications were chosen because of their well-established statistical association with transcriptional states [19]. Notably, several high-throughput studies have underscored already the relevance of histone crosstalk by identifying highly significant pairwise relationships between post-translational modifications [20, 21, 22, 23].

Under these considerations, I defined the working hypothesis as follows: *for any cell population in the same differentiation state and within genomic regions adjacent to transcription start sites, constraints on histone H3 crosstalk explicitly uncorrelated to mRNA levels* (i) *are statistically significant and* (ii) *associate with that differentiation state.* 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 dynamical systems approach: if heritable changes in mRNA levels explain completely cell differentiation states, an additional non-epigenetic yet differentiation-associated level of constraints on histone H3 crosstalk is superfluous.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

The hypothesis testing rationale was to apply unsupervised hierarchical clustering on the *ctalk_non_epi* profiles for different cell datasets in all three organisms, using non-parametric bootstrap resampling to assess cluster significance [24]. If the null hypothesis is true, the obtained clusters will be statistically insignificant, or else they will not associate with cell differentiation states.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

In all analyses performed, *ctalk_non_epi* profiles fell into statistically significant clusters that associate with cell differentiation states in *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster*. Moreover, the results in detail suggest that *ctalk_non_epi* profiles associate with cell differentiation states at least as strongly as do mRNA abundance[2] profiles (the relationship between transcriptional and cell differentiation states is known and well-established [25, 26, 27]). In summary, for all three organisms analyzed, the null hypothesis had to be consistently rejected, indicating that the proof of principle described in the Introduction was obtained.

The embryonic stem cells *ctalk_non_epi* profile differs significantly from those of differentiated cell types in *Homo sapiens*

Using data for nine different histone H3 modifications (for details see Materials and Methods), *ctalk_non_epi* profiles were computed 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 statistically significant cluster. Such significance was expressed in the obtained *au* (approximately unbiased) and *bp* (bootstrap probability) significance scores, which were greater or equal than 95 (**Figure 1A**, cluster #4). The *ctalk_non_epi* profile identified as dissimilar (i.e. excluded from the largest significant cluster) was the one corresponding to H1-hESC embryonic stem cells.

For comparison and positive control, mRNA abundance profiles for the six cell types were constructed from RNA-seq data (the same values that are controlled for in the computation of *ctalk_non_epi* profiles) and then hierarchically clustered. As expected, the transcriptional profile corresponding to H1-hESC (embryonic stem cells) was identified as significantly dissimilar, i.e. resulted excluded from the largest significant cluster (**Figure 1B**, cluster #3), although in this case it was excluded along with the GM12878 B-lymphoblastoids profile.

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

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

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

liver *ctalk_non_epi* profile was found to be more similar to the profiles of the least differentiated states than the heart profile (**Figure 1C**, cluster #3).

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

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

In the final analysis, *ctalk_non_epi* profiles were computed from data for six histone H3 modifications in nine periods/time points throughout *Drosophila melanogaster* development (0-4h, 4-8h, 8-12h, 12-16h, 16-20h and 20-24h embryos; L1 and L2 larval stages; pupae). As observed in human and mouse profiles, fruit fly *ctalk_non_epi* profiles fell into clusters that also associate strongly with the degree of cell differentiation (derivable from the degree of development). One significant cluster grouped *ctalk_non_epi* profiles of earlier developmental periods (**Figure 1E**, cluster #5) apart from later development profiles. Two more significant clusters grouped later time point *ctalk_non_epi* profiles (**Figure 1E**, cluster #3) and separated the L2 larvae profile (**Figure 1E**, cluster #7) from all other profiles.

General *ctalk_non_epi* cluster structure is not entirely consistent with developmental chronology as the pupae profile (**Figure 1E**, cluster #7) shows. It must be noted however that, unlike *Homo sapiens* and *Mus musculus* data where each *ctalk_non_epi* profile represented a specific or almost specific differentiation state, each *Drosophila melanogaster* data set was obtained by the authors from whole specimens (embryos, larvae and pupae). Especially for later development, this implies that each *ctalk_non_epi* profile has to be computed from more than one partially differentiated cell type at the same developmental period, thus limiting to a certain extent the power of the analysis. This caveat in fact highlights the overall *ctalk_non_epi* cluster consistence with developmental chronology, particularly when compared with that obtained from mRNA levels as will be detailed next.

The mRNA abundance profiles in *D. melanogaster* yielded a general cluster structure much less consistent with developmental chronology than the obtained from *ctalk_non_epi* profiles. For example, the profile for 0-4h embryos fell into the same significant cluster with the profiles for 16-20h and 20-24h embryos (**Figure 1F**, cluster #3). Additionally, the profile for 12-16h embryos fell into the same significant cluster with the profiles for L1 and L2 larvae (**Figure 1F**, cluster #5).

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

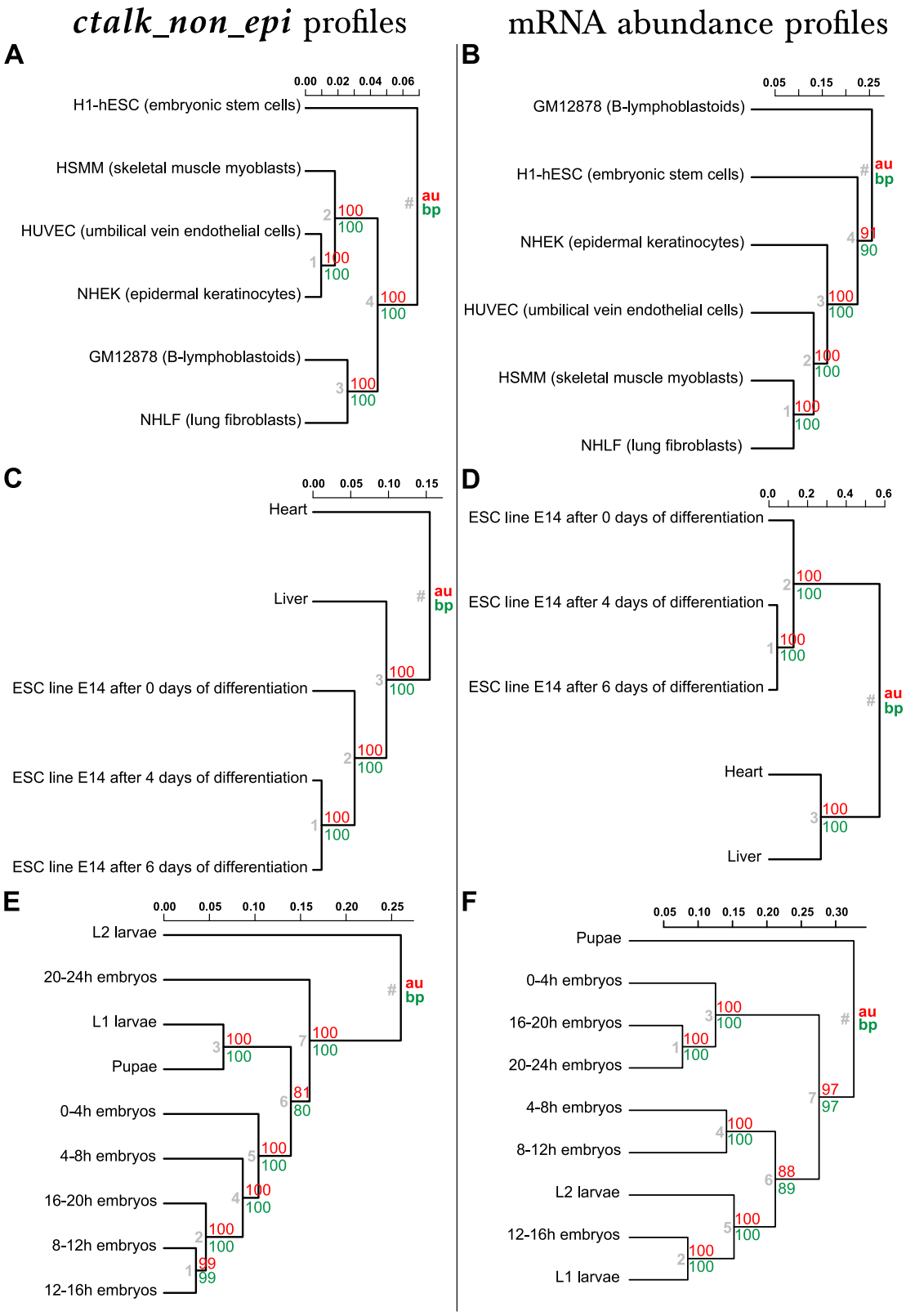

Figure 1: Unsupervised hierarchical clustering of *ctalk_non_epi* profiles and mRNA abundance profiles for *Homo sapiens* (**A**, **B**), *Mus musculus* (**C**, **D**), and *Drosophila melanogaster* (**E**, **F**). Metric: correlation ($1 - r$). Linkage method: "average" (also known as UPGMA). Significance scores [24]: **au** (approximately unbiased) and **bp** (bootstrap probability). Significant clusters were identified as those for which **au** and **bp** ≥ 95. Cluster numbers are in gray.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Beyond the obtained proof of principle

The most important aspect of the previously presented results is not the clear and statistically significant relationship between *ctalk_non_epi* profiles and cell differentiation states but instead the nature of the constraints represented by *ctalk_non_epi* profiles (provided such relationship exists). By definition, *ctalk_non_epi* profiles represent the strength and sign of pairwise partial correlations computed from observed histone modification states; the same observed states that previous research has shown able to predict transcriptional states with high accuracy ($R{\sim}0.9$) [19]. It follows directly from these considerations that, for all three analyzed organisms within regions adjacent to transcription start sites (henceforth TSSs), histone H3 modification states are subject to an additional type of constraints that are explicitly uncorrelated to mRNA levels and associated with cell differentiation states. In other words two systems, mutually uncorrelated and yet both associated to cell differentiation, ***simultaneously*** constrain histone H3 modification states.

Still, any theory of differentiated multicellularity developed on the basis of the critique of the traditional approach presented in the introduction and on the obtained proof of principle must address these six fundamental questions:

Q1 Since the constraints defining the proof of principle are explicitly uncorrelated to mRNA levels by definition, how do they come to be associated with cell differentiation states?

Q2 If they are indeed necessary for the intrinsic regulation of gene expression during cell differentiation, how is such regulation exerted?

Q3 Can these constraints be regarded as biologically meaningful information? If so, what is the content of this information?

Q4 Can they account for the remarkable and characteristic robustness of cell differentiation with respect to even moderate perturbations?

Q5 How do these constraints relate to the evolution of metazoans? Is this relationship extendable to the evolution of other differentiated multicellular lineages such as plants?

Q6 Are histone H3 modification states ultimately cause or effect of transcriptional states? (This last question is a rehash of a very important point raised previously by Peter Fraser and Wendy Bickmore [28].)

Q7 Why undifferentiated cells start to differentiate in the embryo at certain time point?

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Since Ernst Haeckel's "gastraea theory" [29], the explanatory accounts for the evolution of multicellularity that are regarded as the most solid are fundamentally divorced from those aiming to explain the dynamics of development such as the epigenetic landscape model. This is because Haeckel's hypothesis and the ones built upon it rely on the gradual specialization of same-species (or even different-species [30]) cell colonies or aggregations [31, 32, 33, 34, 35], whereas ontogeny and cell differentiation in particular start—in the development of every single multicellular organism—from a single cell or, in other words, "from the inside out". Although this divorce does not necessarily preclude that the "colonial" approach points in the right direction, it is also clear that a fundamental explanation for how a single cell came to embody this "dynamical reversal" of development with respect to its evolutionary origin is lacking and will be needed.

Notably, some alternative "non-colonial" and "non-epigenetic" hypotheses have been advanced aiming to explain the dynamics and informational requirements of cell-differentiation (which in turn could provide some hints on the evolution of multicellularity). One of them is the "darwinian cell differentiation" hypothesis by J. J. Kupiec, according to which gene expression instability and stochasticity, in the context of external metabolic substrate gradients, creates an intrinsic natural-selection-like mechanism able to drive the differentiation process [36]. Another "non-epigenetic" hypothesis, advanced by Andras Paldi, is that cell fate decisions are the result of the characteristic coupling of gene expression and metabolism: fates are determined by fluctuations in the nutrient/oxygen ratio, which are driven by the necessity to maintain the dissipative nature of the metabolic network, which in turn must be redox-neutral at all times [37].

At large, to my knowledge all explanatory accounts of the self-regulation of cell differentiation and of the evolution of multicellularity suffer at least one of the following problems: (i) failure to explain how structures or dynamics that supposedly account for the transition to multicellularity or to cell differentiation have fundamentally analogous counterparts in unicellular lineages or even prokaryotes, (ii) failure to account, at least in principle, for the information required in cell fate decisions or in the transition between strictly single-cell-related content to additional multicellular-individual-related content, (iii) failure to explain the reproducible and robust self-regulatory dynamics—apart from the propagatory—of gene expression during cell differentiation and, most importantly, (iv) unfalsifiability: this is why these accounts—importantly, including the epigenetic landscape—are widely regarded as hypotheses, models, or frameworks in spite of having been presented sometimes as theories by their authors.

In terms of overcoming these problems, it must be noted that Kupiec's hypothesis encompassed a variable that, I submit, is critical to the solution of the riddle: certain gradients in the extracellular space—not yet identified, but both fundamentally conceivable and experimentally verifiable—can be explicitly uncorrelated to gene expression profiles. It is possible that Kupiec did not consider this possibility because his attempt to explain cell differentiation relied only on random variation and selection, ruling out with this any explanatory role of emergent systems and properties.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In contrast to current hypotheses, the falsifiable theory to be postulated here regards the multicellular organism as a higher-order system that *emerges* from proliferating undifferentiated cells and *then* is subject to natural selection (as emerged the very first self-replicating and self-repairing system—ancestor of all known living organisms—beyond any reasonable doubt). Importantly, the theoretical development in this work is not based on the substrate-based[3] concept of irreducible emergence (fundamentally refuted by Jaegwon Kim [38, 39]) but instead converged (from the strict *explicitly-uncorrelated-dynamics* condition argued in the introduction) into what can be described as the constraint-based[4] concept of emergence for higher-order teleological systems, pioneered in a broader perspective by Terrence Deacon in 2011 [40].

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

[4]Understood as the dynamics explicitly *excluded* from realization in the system.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Preliminary theoretical definitions and notation

Before postulating the theory, I must introduce some new definitions and notation regarding molecular dynamics and spatial topology. A brief glossary sufficient for the theoretical formulation is provided below[5].

Context: $X_{(i;t)}$ is the ith 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 and will be operationally treated as sets.*

S_E **Extracellular space**: The entire space in an organism or cell population that is not occupied by the cells themselves at a given instant t. Positions in $S_E(t)$ will be specified in spherical coordinates, namely r (radial distance), θ (azimuthal angle), and ϕ (polar angle).

$C_W(X_{(i;t)})$ **Waddington's constraints**: The constraints associating certain subsets of the spatially-specified molecular nuclear phenotype of $X_{(i;t)}$ with the instantaneous transcription rates at the transcription start sites (TSSs), provided these Waddington's constraints $C_W(X_{(i;t)})$ are *explicitly uncorrelated* with the genomic sequence in dynamical terms.

$F_W(X_{(i;t)})$ **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 regions).

$F_W^{\rightarrow}(X_{(i;t)})$ **Waddington's extracellular propagators**: The largest subset of the entire spatially-specified and membrane-exchangeable (by facilitated diffusion) molecular phenotype of $X_{(i;t)}$ that excludes Waddington's embodiers $F_W(X_{(i;t)})$ but is capable of eliciting a change—intracellular signal transduction may be required—in those Waddington's embodiers $F_W(X_{(i;t)})$ after a certain time interval Δt.

$C_N(X_{(i;t)})$ **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 these Nanney's constraints $C_N(X_{(i;t)})$ are *explicitly uncorrelated* with the instantaneous transcription rates at the TSSs in dynamical terms.

$F_N(X_{(i;t)})$ **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 (e.g. histone H3 post-translational modifications in the TSS-adjacent regions, as shown in the Results).

[5]The complete list of formal definitions and notation can be found in the Appendix.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\boldsymbol{F}_{\boldsymbol{N}}^{\rightarrow}(\boldsymbol{X}_{(\boldsymbol{i};\boldsymbol{t})})$ **Nanney's extracellular propagators**: The subset of the entire spatially-specified and membrane-exchangeable (by facilitated diffusion) molecular phenotype of $\boldsymbol{X}_{(\boldsymbol{i};\boldsymbol{t})}$ that excludes Nanney's embodiers $\boldsymbol{F}_{\boldsymbol{N}}(\boldsymbol{X}_{(\boldsymbol{i};\boldsymbol{t})})$ but is capable of eliciting a change—intracellular signal transduction may be required—in those Nanney's embodiers $\boldsymbol{F}_{\boldsymbol{N}}(\boldsymbol{X}_{(\boldsymbol{i};\boldsymbol{t})})$ after a certain time interval $\Delta \boldsymbol{t}$.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A general theory of differentiated multicellularity

This theory mainly aims to explain how cell differentiation emerges in the ontogeny of extant multicellular lineages and how differentiated multicellular lineages emerged throughout evolution. To highlight the similarities of both phenomena at the most fundamental level, the theory will be postulated in parts described in parallel. Each part will be described in terms of the evolution of an ancestor eukaryotic species $\boldsymbol{U}$ towards differentiated multicellularity and in terms of the ontogenetic process starting from the zygote of a differentiated multicellular species $\boldsymbol{D}$. Importantly, and although its proof of principle was obtained from high-throughput metazoan data, this theoretical description makes no assumption whatsoever about a specific multicellular lineage. This is why it is referred to as a general theory here and also in the title.

Part I (Evolution)

The unicellular and undifferentiated ancestor. Let $\boldsymbol{U}_{(i;t_{U_0})}$ be the $\boldsymbol{i}^{th}$ cell in a population of the species $\boldsymbol{U}$, which is the last unicellular eukaryotic ancestor species of the extant differentiated multicellular species $\boldsymbol{D}$. Here the spatially-specified phenotype $\boldsymbol{F}(\boldsymbol{U}_{(i;t_{U_0})})$ displays Waddington's embodiers (i.e. $\boldsymbol{F}_W(\boldsymbol{U}_{(i;t_{U_0})}) \neq \emptyset$, e.g. histone post-translational modifications) but cell differentiation is not possible. Also, significant constraints exist between the entire spatially-specified phenotype $\boldsymbol{F}(\boldsymbol{U}_{(i;t_{U_0})})$ and Waddington's propagators $\boldsymbol{F}_W(\boldsymbol{U}_{(i;t_{U_0})})$ regardless of $\boldsymbol{T}(\boldsymbol{U}_{(i;t_{U_0})})$ (i.e. significant Nanney's constraints $\boldsymbol{C}_N(\boldsymbol{U}_{(i;t_{U_0})})$ exist). However, Nanney's propagators (if any) are confined to $\boldsymbol{U}_{(i;t_{U_0})}$. In other words, here Nanney's extracellular propagators do not exist (i.e. $\boldsymbol{F}_{\overrightarrow{N}}(\boldsymbol{U}_{(i;t_{U_0})}) = \emptyset$; see **Figure 2A**, top)

Part I (Ontogeny)

The multicellular organism's zygote. Let $\boldsymbol{D}_{(1;t_{D_0})}$ be a zygote of the extant differentiated multicellular species $\boldsymbol{D}$. Like $\boldsymbol{F}(\boldsymbol{U}_{(i;t_{D_0})})$, the spatially-specified phenotype $\boldsymbol{F}(\boldsymbol{D}_{(1;t_{D_0})})$ displays Waddington's embodiers (i.e. $\boldsymbol{F}_W(\boldsymbol{D}_{(i;t_{D_0})}) \neq \emptyset$, e.g. histone post-translational modifications) but cell differentiation is not observed *yet*. Also, significant constraints exist between the entire spatially-specified phenotype $\boldsymbol{F}(\boldsymbol{D}_{(1;t_{D_0})})$ and Waddington's propagators $\boldsymbol{F}_W(\boldsymbol{D}_{(1;t_{D_0})})$ regardless of $\boldsymbol{T}(\boldsymbol{D}_{(1;t_{D_0})})$ (i.e. significant Nanney's constraints $\boldsymbol{C}_N(\boldsymbol{D}_{(1;t_{D_0})})$ exist). But unlike in $\boldsymbol{U}_{(i;t_{D_0})}$, Nanney's propagators are *not* confined to $\boldsymbol{D}_{(1;t_{D_0})}$. In other words, here Nanney's extracellular propagators do exist (i.e. $\boldsymbol{F}_{\overrightarrow{N}}(\boldsymbol{D}_{(1;t_{D_0})}) \neq \emptyset$; see **Figure 2A**, bottom).

Part II (Evolution)

The necessary novel alleles. At some instant $(\boldsymbol{t}_M - \Delta \boldsymbol{t}_M) > \boldsymbol{t}_{U_0}$ during evolution the genome $\boldsymbol{G}(\boldsymbol{U}_{(k;t_M-\Delta t_M)})$ of certain $\boldsymbol{k}^{\text{th}}$ cell of the species $\boldsymbol{U}$ changes such that at least one element of its associated phenotype is specifiable in the set of Nanney's extracellular propagators (i.e. $\boldsymbol{F}_{\overrightarrow{N}}(\boldsymbol{U}_{(k;t_M-\Delta t_M)}) \neq \emptyset$). As remarked in the previous subsection, this implies that $\boldsymbol{G}(\boldsymbol{U}_{(k;t_M-\Delta t_M)})$ accounts also for all other phenotypic gene products necessary for the facilitated diffusion of the molecule(s) specified in $\boldsymbol{F}_{\overrightarrow{N}}(\boldsymbol{U}_{(k;t_M-\Delta t_M)})$. Importantly, the novel alleles involved

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in the change $G(U_{(i;t_{U_0})}) \rightarrow G(U_{(k;t_M-\Delta t_M)})$ (**Figure 2A** to **2B**) are a necessary but not sufficient condition for differentiated multicellularity (**Figure 2B**).

Part II (Ontogeny)

The already present necessary alleles. At any instant $(t_D - \Delta t_D) > t_{D_0}$ preceding cell differentiation, the genome specified by $G(D_{(i;t_D-\Delta t_D)})$ (i.e. any daughter cell in the embryo) is similar to $G(U_{(k;t_M-\Delta t_M)})$ (see **Figure 2B**) in the sense that both genomes code for Nanney's extracellular propagators (i.e. the sets $F_N^{\rightarrow}(D_{(i;t_D-\Delta t_D)})$ and $F_N^{\rightarrow}(U_{(k;t_M-\Delta t_M)})$ are nonempty). Importantly, the alleles specified in the zygote's genome $G(D_{(1;t_{D_0})})$ and in $G(D_{(i;t_D-\Delta t_D)})$ (i.e. the genome of any of its daughter cells) are a necessary but not sufficient condition for cell differentiation.

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 $F_N^{\rightarrow}$ in any cell population $\{X_{(1;t)},\ldots,X_{(n;t)}\}$ (i.e. cells of the species X at any given instant t) implies that a scalar field[6] $\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi) \geq 0$ can represent the concentration of Nanney's extracellular propagators in $S_E(X_{(1;t)},\ldots,X_{(n;t)})$. 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 $S_E(X_{(1;t)},\ldots,X_{(n;t)})$. Under these conditions the associated gradient[7] $\vec{\nabla}\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi) = \left(\frac{\partial\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)}{\partial r}, \frac{1}{r}\frac{\partial\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)}{\partial\theta}, \frac{1}{r\sin\theta}\frac{\partial\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)}{\partial\phi}\right)$ remains in magnitude[8] under a certain critical value V_M in $S_E(X_{(1;t)},\ldots,X_{(n;t)})$ for the daughter cells of $U_{(k;t_M-\Delta t_M)}$ and under a critical value V_D for the differentiated multicellular species D. Importantly, the constraints represented in the gradient $\vec{\nabla}\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)$ imply there is free energy available—whether or not there is cell differentiation—which, as will be described later, is in fact partially utilized as work in the generation of new information content.

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

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

[8] Note that in spherical coordinates the magnitude of the gradient is simply the partial derivative of the scalar field $\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)$ (concentration of Nanney's extracellular propagators $F_W^{\rightarrow}$) with respect to the radial distance: $\left|\vec{\nabla}\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)\right| = \frac{\partial\Phi_N(X_{(1;t)},\ldots,X_{(n;t)},r,\theta,\phi)}{\partial r}$.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Part IV (Evolution) **The emergent transition to differentiated multicellularity**. At some later but relatively close instant t_M, cell proliferation yields a significantly larger population. Now diffusion flux of Nanney's extracellular propagators is no longer able to overtake the increasing spatial constraints in the extracellular space S_E. A significant gradient, in magnitude equal or greater—anywhere in S_E—than the critical value V_M forms then, i.e. $\left|\vec{\nabla}\Phi_N\left(U_{(1;t_M)},\ldots,U_{(n;t_M)},r,\theta,\phi\right)\right| \geq V_M, (r,\theta,\phi) \in S_E$. Therefore, Nanney's extracellular propagators $F_N^{\rightarrow}$ diffuse differentially into each cell, yielding unprecedented differential Nanney's constraints $\{C_N(U_{(1;t_M)}),\ldots,C_N(U_{(n;t_M)})\}$ in the cells' nuclei by virtue of no cell or gene product in particular but, importantly, of the constraints imposed by the entire proliferating cell population on the diffusion flux of $F_N^{\rightarrow}$ in S_E. These differential Nanney's constraints are in turn defined with respect to Waddington's embodiers $\{F_W(U_{(1;t_M)}),\ldots,F_W(U_{(n;t_M)})\}$, thus they now constrain the instantaneous transcription rates $\{T(U_{(1;t_M)}),\ldots,T(U_{(n;t_M)})\}$ in a differential and dynamically uncorrelated manner (**Figure 2C**). This is how multicellular lineages, displaying self-*regulated* changes of gene expression during ontogeny, evolved.

Part IV (Ontogeny) **The emergent transition to cell differentiation**. At some later but relatively close instant t_D, embryonic growth yields certain number of undifferentiated cells. Now diffusion flux of Nanney's extracellular propagators is no longer able to overtake the increasing spatial constraints in the extracellular space S_E. A significant gradient, in magnitude equal or greater—anywhere in S_E—than the critical value V_D forms then, i.e. $\left|\vec{\nabla}\Phi_N\left(D_{(1;t_D)},\ldots,D_{(n;t_D)},r,\theta,\phi\right)\right| \geq V_D, (r,\theta,\phi) \in S_E$ (see question Q7). Therefore, Nanney's extracellular propagators $F_N^{\rightarrow}$ diffuse differentially into each cell, yielding unprecedented differential Nanney's constraints $\{C_N(D_{(1;t_D)}),\ldots,C_N(D_{(n;t_D)})\}$ in the cells' nuclei by virtue of no cell or gene product but, importantly, of the constraints imposed by the entire growing embryo on the diffusion flux of Nanney's extracellular propagators in the extracellular space S_E. These differential Nanney's constraints are in turn defined with respect to Waddington's embodiers $\{F_W(D_{(1;t_D)}),\ldots,F_W(D_{(n;t_D)})\}$, thus they now constrain the instantaneous transcription rates $\{T(D_{(1;t_D)}),\ldots,T(D_{(n;t_D)})\}$ in a differential and dynamically uncorrelated manner (**Figure 2C**). This is how undifferentiated cells start to differentiate, displaying self-*regulated* changes of gene expression during ontogeny (see question Q1).

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 [41], the necessary microevolutionary represented by $G(U_{(i;t_{U_0})}) \rightarrow G(U_{(k;t_M-\Delta t_M)})$ can be safely regarded as a highly improbable step. Nevertheless, the major evolutionary breakthrough was not genetic but instead the unprecedented dynamical regime emerging

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [42, 33]. This novel dynamical regime emerges as a higher-order constraint[9] 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). Although dependent on the alleles in $G(U_{(k;t_M-\Delta t_M)})$ to emerge given enough cell proliferation, this system is not a network of epigenetic mechanisms—however complex—but instead a particular instantiation of a *teleodynamic system*, proposed by Terrence Deacon in his *theory of emergence by constraint coupling and preservation*[10] [40], which is presented to and shaped by natural selection at each instant. In this context, environmental constraints as oxygen availability [43] and even gravity (see Corollary #5) filter out specific emergent multicellular dynamics that are incompatible with them. In summary, the critical evolutionary novelty was the unprecedented multicellular *self*, which can be described as an intrinsic, higher-order, self-sustaining, self-repairing, self-replicating, and self-regulating dynamical constraint on individual eukaryotic cells.

Part V (Ontogeny)

***Who* is regulating cell differentiation?** Contrary to what could be expected under the "top-down causation" framework (common to earlier formulations of causally efficacious emergent properties, and fundamentally refuted [38, 39] as mentioned previosusly), the theory hereby postulated does *not* regard the proliferation-generated extracellular gradient $\vec{\nabla}\Phi_N$ such that $\left|\vec{\nabla}\Phi_N\left(D_{(1;t_D)},\ldots,D_{(n;t_D)},r,\theta,\phi\right)\right| \geq V_D,(r,\theta,\phi) \in S_E$ as the fundamental regulator of the cell differentiation process. While differential Nanney's constraints $\{C_N(D_{(1;t_D)}),\ldots,C_N(D_{(n;t_D)})\}$ are *regulatory constraints* with respect to Waddington's embodiers $\{F_W(D_{(1;t_D)}),\ldots F_W(D_{(n;t_D)})\}$ (as described in Part IV-Ontogeny), the reciprocal proposition is also true: since Waddington's constraints $\{C_W(D_{(1;t_D)}),\ldots,C_W(D_{(n;t_D)})\}$ are dynamically uncorrelated to Nanney's constraints, they are in turn *regulatory constraints* with respect to Nanney's extracellular propagators $\{F_N^{\rightarrow}(D_{(1;t_D)}),\ldots,F_N^{\rightarrow}(D_{(n;t_D)})\}$ (e.g. changes in the expression of the protein pores or carriers necessary for the facilitated diffusion of Nanney's extracellular propagators). *If and only if the dynamically uncorrelated Waddington's constraints C_W and Nanney's constraints C_N*[11] *become synergistically coupled across the extracellular space S_E true intrinsic regulation on the cell differentiation process is possible.* This implies in turn that both chromatin states and transcriptional states are simultaneously cause and effect with respect to each other (this regime, intuitively describable as "chicken-egg" dynamics, is the answer this theory provides to question Q6). Ontogenic self-regulation is then exerted by the intrinsic higher-order constraint or *teleodynamic system* that

[9] Understood as the states explicitly excluded from being realized in the dynamics of the system.

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

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

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

emerges from proliferating cells. In other words, the differentiated multicellular organism *is* the causally efficacious, higher-order, coupled constraint emerging from and regulating *ipso facto* Nanney's constraints C_N and Waddington's constraints C_W in what would be otherwise a population or colony—however symbiotic—of individual eukaryotic cells (see question Q2).

Part VI (Evolution)

Unprecedented multicellular dynamics. Once the necessary microevolutionary change $G(U_{(i;t_{D_0})}) \rightarrow G(U_{(k;t_M-\Delta t_M)})$ took place in the species U phenomena like gene duplication or alternative splicing[12] made possible the appearance of a plethora of novel multicellular (*teleodynamic*) regimes and consequently novel cell types, tissues and organs. Moreover, the dependence of differentiated multicellularity on one or more coexisting $\vec{\nabla}\Phi_N$ gradients (i.e. constraints on diffusion flux) in S_E, which importantly depend on no cell in particular but on the entire cell population or embryo, yields at least two important implications in evolutionary terms. First, it explains in principle the remarkable robustness of differentiated multicellularity with respect to extrinsic perturbations (see question Q4). Second, 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 in a time when eukaryotic life was only unicellular, even when compared with their prokaryotic counterparts[13]. In this context a result obtained over a decade ago acquires relevance. In a genome-wide study comprising ~90 bacterial and ~10 eukaryote species, it was found that the number of genes involved in transcriptional change increases as a power law of the total number of genes [44], with an exponent of 1.87 ± 0.13 for bacteria. Remarkably, the corresponding exponent for eukaryotes was closer to one (i.e. to linearity): 1.26 ± 0.10. The previously described loss of lower-order, cell-intrinsic regulatory systems in differentiated multicellular organisms—by virtue of emergent $\vec{\nabla}\Phi_N$ gradients in S_E—is entirely consistent with this observation.

Part VI (Ontogeny)

What does ontogeny recapitulate? This theory holds the hereby proposed emergent transition, spontaneous from cell proliferation shortly after Nanney's extracellular propagators $F_N^{\rightarrow}$ appeared, as key to the evolution of any multicellular lineage displaying self-regulated changes of gene expression during cell differentiation. Therefore, it rejects in turn the hypothesis that metazoans—or, in general, any multicellular lineage displaying self-regulated cell differentiation—evolved from gradual specialization of single-cell colonies or aggregations [29, 31, 32, 33, 34, 35]. Importantly however, this is not to argue that potentially precedent traits (e.g. cell-cell adhesion) were necessarily unimportant for the later fitness of differentiated multicellular organisms. Neither is this to reject Haeckel's famous assertion completely: in every extant

[12] In the loci involved in the synthesis and/or facilitated diffusion of Nanney's extracellular propagators $F_N^{\rightarrow}$.

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

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

multicellular lineage this self-sufficient, self-repairing, self-replicating, and self-regulating system has emerged over and over again from undifferentiated cells and presented itself to natural selection ever since its evolutionary debut. Therefore, at least in this single yet most fundamental sense, ontogeny does recapitulate phylogeny.

Part VII (Evolution & Ontogeny)

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

Part VIII (Evolution & Ontogeny)

Novel cell types, tissues and organs evolve and develop. Further microevolutionary changes in the alleles specified in $G(U_{(k;t_M-\Delta t_M)})$ or already present in $G(D_{(1;t_{D_0})})$ (e.g. gene duplication, alternative splicing) imply than one or more than one $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ gradients emerge in S_E with cell proliferation. A cell type T_j will develop then in a region S_{E_i} of the extracellular space S_E when a relative uniformity of Nanney's extracellular propagators is reached, i.e. $\left(\left|\vec{\nabla}\Phi_{N_{1;T_j}}\right|,\ldots,\left|\vec{\nabla}\Phi_{N_{k;T_j}}\right|\right)\leq\left(V_{N_{1;T_j}},\ldots,V_{N_{k;T_j}}\right),(r,\theta,\phi)\in S_{E_i}$ (see a two-cell-type and two-gradient depiction in **Figure 2D**). As highlighted earlier, cell differentiation is not *regulated* by these gradients themselves but by the higher-order constraint emergent from their synergistic coupling with Waddington's constraints C_W within the cells. This constraint synergy can be exemplified as follows: gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ can elicit changes of gene expression in a number of cells, which in turn may promote the dissipation of the gradients (e.g. by generating a surrounding membrane that reduces dramatically the effective S_E size) or may limit further propagation of those gradients from S_E into the cells (e.g. by repressing the expression of protein pores or carriers involved in the facilitated diffusion of $F_{\vec{N}}$ in S_E). Thus, under this theory cell types, tissues, and organs evolved sequentially as "blobs" of relatively small magnitude $F_{\vec{N}}$ gradients in regions $\{S_{E_i},\ldots,S_{E_n}\}$ within S_E (as just described) displaying no particular shape. These "blobs" emerged with no function in particular—apart from not being incompatible with the multicellular organism's survival and reproduction—by virtue of random genetic variation (involved in the embodiment and propagation of Nanney's constraints C_N) followed by cell proliferation. Then, the "blobs" were shaped by natural selection from their initially random physiological and structural properties to specialized cell types, tissues, and organs (importantly, such specialization evolves with respect to the

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

emergent intrinsic higher-order constraint postulated here as the multicellular organism). The result of this evolutionary process is observable in the dynamics that emerge during the ontogeny of extant multicellular species (**Figure 2E**).

Part IX (Evolution & Ontogeny)

Emergent information content and multicellular self-repair. As argued in the introduction, a significant amount of information content has to *emerge* to account for robust and reproducible cell fate decisions and for the self-regulated dynamics of cell differentiation in general. Under this theory, this content emerges when the significant gradient or gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ form at some point from proliferating undifferentiated cells, entangling synergistically Nanney's constraints $\boldsymbol{C_N}$ and Waddington's constraints $\boldsymbol{C_W}$ across $\boldsymbol{S_E}$. Crucially, this information is *not* about any coding sequence and its relationship with cell-intrinsic and cell-environment dynamics (i.e. genetic information) *nor* about any heritable gene expression level/profile and its relationship with cell-intrinsic and cell-environment dynamics (i.e. epigenetic information). Instead, this information is *about the multicellular organism as a whole* (understood as the emergent higher-order intrinsic constraint described previously) and also about the environmental constraints under which this multicellular organism develops. For this reason I propose to call this emergent information *hologenic*[14] (see question Q3). No less importantly, at each instant the multicellular organism is not only interpreting hologenic information—by constraining its development into specific trajectories since it emerges—but also actively creating novel hologenic information (in other words displaying "chicken-egg" dynamics, similar to those described in Part V-Ontogeny). In the multicellular organism, the subset of the molecular phenotype that conveys hologenic information is not only the subset involved in the gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ but the entire subset embodying or propagating Nanney's constraints $\boldsymbol{C_N}$. Additionally, since the gradients $\{\vec{\nabla}\Phi_{N_1},\ldots,\vec{\nabla}\Phi_{N_k}\}$ depend on no cell in particular—not even on a sufficiently small group of cells—but on the whole cell population or embryo, cell differentiation will be robust with respect to moderate perturbations such as some cell loss (see question Q4).

Part X (Ontogeny)

Ontogeny ends and cell differentiation "terminates". If under this theory cell differentiation emerges with the proliferation of (at the beginning, undifferentiated) cells, why should it terminate for any differentiation lineage? What is this "termination" in fundamental terms? These are no trivial questions. As an answer to the first, zero net proliferation begs the fundamental question. To the second, a "fully differentiated" cell state condition fails to explain the existence of adult stem cells. To address these issues three considerations are most important: (i) for any cell or group of cells the molecules specifiable as Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ at any instant $\boldsymbol{t}$ may

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not be specifiable as such at some later instant[15] $t + \Delta t$, (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* a *telos* such as the organism's mature form, a fully differentiated cell, or certain future transcriptional changes to achieve such states (described as "intuitive" in the introduction), which are logically inconsistent[16] and unjustifiably homuncular and, (iii) this causally-efficacious, higher-order constraint emerges from the synergistic coupling of lower-order Waddington's constraints C_W and Nanney's constraints C_N across the extracellular space S_E. Therefore, under this theory, cell differentiation "terminates" (the quotes will be justified below) in any given region S_{E_i} of the extracellular space if a stable or metastable equilibrium is reached where (i) the gradients of Nanney's extracellular propagators dissipate in S_{E_i} under certain critical values, i.e. $\left(\left|\vec{\nabla}\Phi_{N_1}\right|,\ldots,\left|\vec{\nabla}\Phi_{N_k}\right|\right) < \left(V_{D_1},\ldots,V_{D_k}\right), (r,\theta,\phi) \in S_{E_i}$ (**Figure 2F**, left) and/or (ii) those gradients are unable to constrain Waddington's embodiers F_W in the cells' nuclei because the critical gene products (protein pores/carriers or intracellular transducers) are non-functional or not expressed, i.e. when the cells become "blind" to the gradients (**Figure 2F**, right). Condition (i) can be reached for example when development significantly changes the morphology—increasing the surface-to-volume ratio—of the cells. This is because such increase removes spatial constraints in S_E that facilitate the emergence/maintenance of the gradients. It is thus predictable under this theory a significant positive correlation between the degree of differentiation of a cell and its surface-to-volume ratio, once controlling for characteristic length (i.e. "unidimensional size") and also a negative significant correlation between stem cell potency/regenerative capacity and that ratio. On the other hand, condition (ii) can be reached when the cell differentiation process represses at some point the expression of the protein pores or carriers necessary for the facilitated diffusion of the *current* Nanney's extracellular propagators $F_{\vec{N}}$. Importantly, the stability of the equilibrium required in these conditions will depend on the cells' currently expressed phenotype, e.g. an adult multipotent or pluripotent stem cell—in stark contrast to a fully differentiated neuron—may differentiate if needed [45] and some differentiated cell may dedifferentiate given certain stimuli [46]. 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 (e.g. once a central nervous system develops/evolved).

[15] This exemplifies why the theoretical definitions and notation had to be developed in instantaneous terms.

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

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Part X (Evolution)

The evolutionarily-shaped *telos*. Whereas the causal power of the organism's mature form as ontogenetic *telos* is logically untenable and only apparent, the assumption that the zygote is a complete developmental blueprint containing all necessary information for the process—as argued in the introduction—is also untenable. In contrast, ontogeny is, under this theory, an emergent, evolutionarily-shaped and truly (instantaneously) 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, instant after instant, causal power on the ontogenetic process. Although the propagation of constraints within this process (e.g. propagated changes of gene expression) is decomposable into molecular interactions, its teleological causal power (e.g. self-regulation) is not. This is because its *telos* is a spontaneous, intrinsic higher-order *constraint* or "thermodynamic zero" emergent from lower-order constraints; it cannot be reduced or decomposed into molecular interactions—as the arithmetic zero cannot be divided and for the same fundamental reason—as T. Deacon first argued [40]. This is also why hologenic content (and in general any information content, as Deacon has argued as well) is thermodynamically *absent* or constrained: hologenic content is not in the molecular substrates conveying that content anymore than the content of this theory is in integrated circuits, computer displays, paper, or even in the complex neural interactions within the reader's brain. As described previously in less specific terms, what becomes constrained (i.e. thermodynamically "*absent*") in the dynamics of the multicellular organism is the content of hologenic information (see question Q3); the substrates propagating the critical constraints for this change can only then be identified as conveying hologenic information. Natural selection has thus shaped the content of hologenic information by shaping the genetic constraints it is ultimately emergent from, not any particular molecules or molecular interactions as media, which should be regarded in this context as means to the *telos*, as the etymology indirectly implies. Moreover, the necessary microevolutionary change $\boldsymbol{G}(\boldsymbol{U}_{(i;t_0)}) \rightarrow \boldsymbol{G}(\boldsymbol{U}_{(k;t_M-\Delta t_M)})$ (described in Part II-Evolution) could well have been significantly smaller—in terms of gene or protein sequence similarity—than the total changes undergone between $\boldsymbol{G}(\boldsymbol{U}_{(i;t_0)})$ and some of its own eukaryotic unicellular ancestors. In general, accounting for substantial differences in the phenotype and its properties[17] given comparatively small genetic changes is bound to be an intractable task if one or more teleodynamic transitions during evolution is/are involved yet ignored.

In hindsight, 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. That is, the high surface-to-volume ratio morphology needed

[17] When great, these differences usually involve intrinsically teleological dynamics at a variety of levels, e.g. function, regulation, courtship, or planning.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for neuron function—and possibly neuron function itself—was only to be expected in the evolution of multicellularity and is only to be expected in multicellular-like life (if any) elsewhere in the Universe, provided no rigid wall (of high relative fitness) impedes the tinkering with substantial increases of the cells surface-to-volume ratio, as observable in plants. In turn this caveat—now together with the predicted negative correlation between stem cell potency and surface-to-volume ratio—suggests that if a multicellular lineage is constrained to always display low cell surface-to-volume ratios, stem cell potency and regenerative capacity will be higher. All other things being equal, these multicellular lineages should be characterized then by a comparatively lower complexity but also by longer lifespan and more robustness to extrinsic damage (see question Q5).

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

Importantly, this theory also encompasses coenocytic (also commonly called "syncytial") stages of development, where cell nuclei divide in absence of cytokinesis (observable in some invertebrates such as *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.

In terms of results indirectly related to this theory, it must be noted that evidence has already been found for tissue migration across a self-generated chemokine gradient in zebrafish [47, 48]. This finding demonstrates the feasibility of some of the dynamics proposed here, namely eukaryotic cells utilizing certain free energy (available in the spontaneous constraints on diffusion in S_E generated by cell migration/proliferation) as work in their own intrinsic dynamics. These two linked processes—one spontaneous, the other non-spontaneous—exemplify a work cycle as proposed by Stuart Kauffman [49]. What remains to be verified is the synergistic coupling of two (as in this theory) or more constraint systems, as proposed by T. Deacon, into the higher-order constraint or multicellular organism described here.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

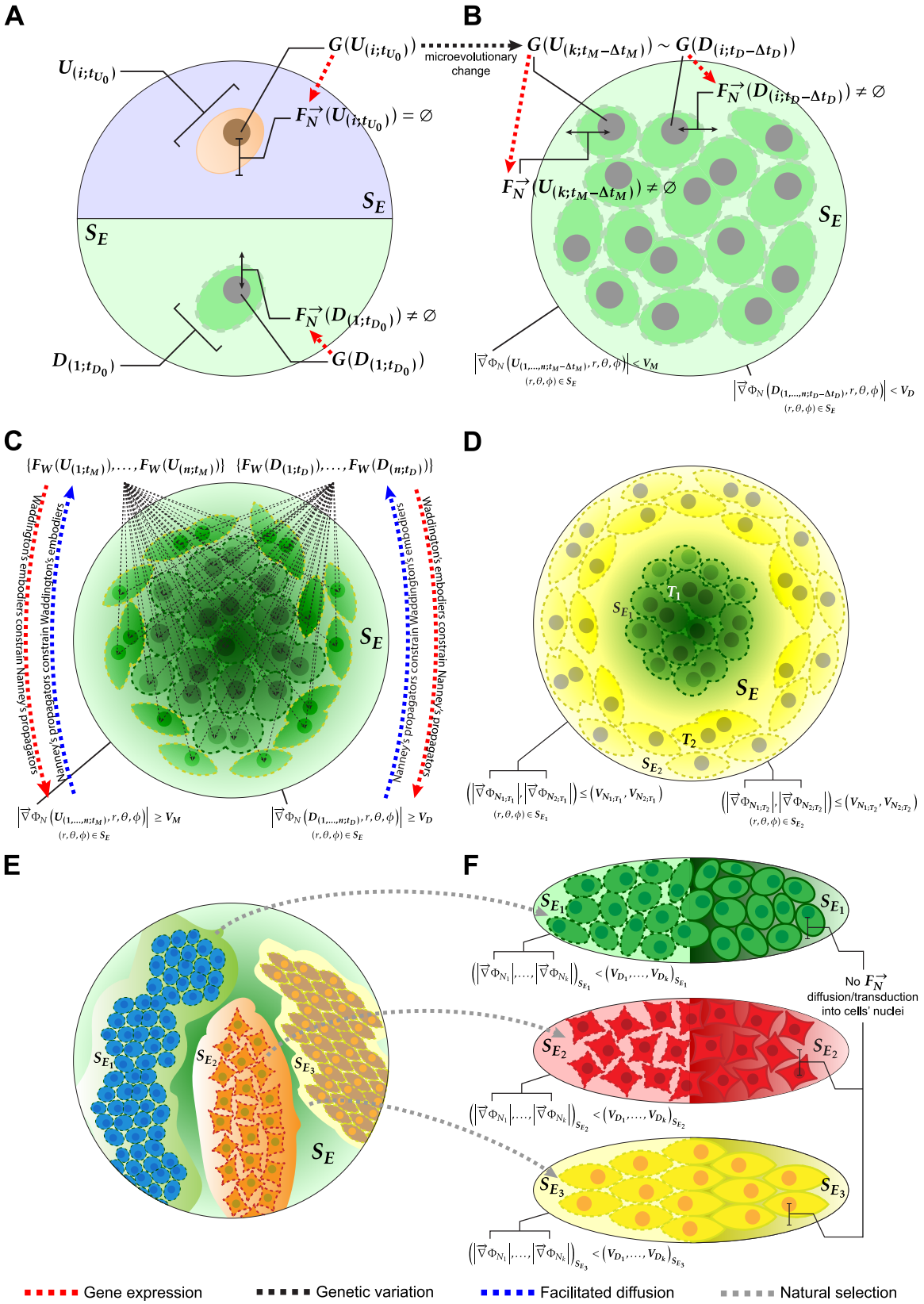

Figure 2: Main steps in the theory: 1. The unicellular and undifferentiated ancestor (**A**, top) and the zygote (**A**, bottom). 2. The necessary alleles are present and cells proliferate, but still no significant $\vec{\nabla}\Phi_N$ gradients form in S_E (**B**). 3. The multicellular *telos* (i.e. intrinsic higher-order constraint) emerges when significant $\vec{\nabla}\Phi_N$ gradients couple the lower-order C_N and C_W constraints synergistically across S_E (**C**). 4. Two cell types start to develop in differential regions with relative $\vec{\nabla}\Phi_N$ uniformity (**D**). 5. Cell types/tissues/organs evolve as emergent "blobs" of relatively small $\vec{\nabla}\Phi_N$ magnitude and then are shaped by natural selection (**E**). 6. Cell differentiation stops when $\vec{\nabla}\Phi_N$ gradients dissipate (**F**, left) or when they cannot diffuse/be transduced into the cells' nuclei (**F**, right).

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Falsifiability

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

1. Under the proposed theory, the gradient $\overrightarrow{\nabla}\Phi_N(\boldsymbol{D}_{(1;t)},\ldots,\boldsymbol{D}_{(n;t)},r,\theta,\phi)$ 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 or differentiating cells are extracted continuously from a developing embryo at the same rate they are proliferating*, then at some instant $\boldsymbol{t_D} + \boldsymbol{\Delta t}$ the gradient 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 (that is, applying it to the embryo but extracting no cells).

2. *There is a significant positive correlation between the cell-wise or cell-type-wise dissimilarity of Nanney's embodiers* $\boldsymbol{F_N}$ *in an embryo and developmental time, which will be observable given enough resolution in the experimental technique.* In practical terms, totipotent stem cells can be taken from an early-stage embryo and divided into subsamples, and embryos from later stages in the same species can be divided (e.g. by cryosection [50]) into subsamples. 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 $\boldsymbol{C_N}$ on histone H3 modifications (adjacent to TSSs) as embodiers—for each subsample. If the predicted correlation fails to be observed even using single-cell high-throughput sequencing methods [51], the theory postulated here should be regarded as falsified.

3. *If any molecular substrate* $\boldsymbol{M}$ (i) *is specifiable as a Nanney's extracellular propagator* during a certain time interval for certain cells of a differentiated multicellular species (see Corollary #1) and (ii) *is also synthesized by an unicellular eukaryote species* $\boldsymbol{U}$ *that is unable to differentiate* (e.g. the dinoflagellate *Lingulodinium polyedrum* [52]), then *experiments will fail to specify* $\boldsymbol{M}$ *as a Nanney's extracellular propagator for the species* $\boldsymbol{U}$.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Corollaries

Described next are some corollaries, hypotheses and predictions (not involving falsifiability) that can be derived from the theory.

1. **Nanney's extracellular propagators**. The strongest prediction that follows from the theory is the existence of Nanney's extracellular propagators, i.e. $\vec{F_N} \neq \emptyset$ for any differentiated multicellular species D. Since these propagators are instantaneously defined, their identification should be in the form "molecule M is specifiable as a Nanney's extracellular propagator of the species D in the cell, cell population, or cell type T_i at the developmental time point t (or the differentiation state s)". This will be verified if, for example, an experiment shows that the *ctalk_non_epi* profiles in these T_i cell or cells vary significantly when exposed to differential concentrations of M in the extracellular medium. If this is the case, it is also predictable that M will synthesized by the cells *in vivo* at a relatively constant rate (at least as long as M is specifiable as $\vec{F_N}$ for them). Importantly, there is no principle in this theory precluding a molecular substrate M from being specifiable as $\vec{F_N}$ and also as as Waddington's extracellular propagator $\vec{F_W}$[18]. Note: although the existence of Nanney's extracellular propagators is a very strong and verifiable prediction, it was not included in the previous subsection because it is not falsifiable in a strict epistemological sense.

2. **Surface-to-volume ratio and the evolution and development of the extracellular matrix**. It was proposed earlier (Part X-Evolution) an important relationship between cell surface-to-volume ratio and the evolution of differentiated multicellularity, in particular between the neuron's high surface-to-volume ratio and the evolution of its function. Importantly, 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 would have been the (developmental) price to pay for a higher-order, dynamically faster form of multicellular *self* (i.e. higher-order intrinsic constraint) that neurons—whose interconnectivity is underpinned by their high surface-to-volume ratio—make possible. On the other hand glial cells (companions of neurons in the nervous tissue) have a smaller surface-to-volume ratio than neurons so they would support them by constraining to some extent the diffusion flux of Nanney's extracellular propagators $\vec{F_N}$ in the neurons "effective" extracellular space[19]. Notably, the glial cells with the smallest surface-to-volume ratio are ependymal cells, which have been found able to serve as neural stem cells [53]. Since this analysis is based on constraints and not on specific material 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

[18] This dual specifiability is not unilkely, since the synergistic coupling of Waddington's constraints C_W and Nanney's constraints C_N across S_E requires that at least one type of molecular substrates is simultaneously specifiable as Waddington's embodiers F_W and Nanney's embodiers F_N.

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\vec{F_N}$ in S_E. Moreover, I submit that the evolution of this developmental support probably preceded the evolution of all other types of support, given the critical role of the $\vec{F_N}$ gradients in the emergence and preservation of the multicellular *telos*.

3. **Natural developmental arrests or diapauses**. The account for natural diapauses—observable in arthropods [54] and some species of killifish (Cyprinodontiformes) [55]—in this theory follows directly from the description in Part X-Ontogeny. That is, natural diapauses are a metastable equilibrium state characterized by (i) the dissipation of Nanney's extracellular propagators $\vec{F_N}$ in S_E under certain critical values (e.g. if some factor inhibits cell proliferation) or (ii) the inability of these gradients to constrain Waddington's embodiers F_W in the cells' nuclei because the critical gene products (protein pores/carriers or intracellular transducers) are non-functional or not expressed. For example, if in some organism the function of the protein pores/carriers critical for the facilitated diffusion of the current $\vec{F_N}$ is temperature dependent, then at that time development will enter a diapause given certain thermal conditions and resume when those conditions are lost.

4. $\vec{F_N}$ **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 (e.g. a cut in the skin) is the gradient $\left|\vec{\nabla}\Phi_N\left(D_{(1;t_{\text{wound}})},\ldots,D_{(n;t_{\text{wound}})},r,\theta,\phi\right)\right| \gg \left|\vec{\nabla}\Phi_N\left(D_{(1;t_{\text{wound}}-\Delta t)},\ldots,D_{(n;t_{\text{wound}}-\Delta t)},r,\theta,\phi\right)\right|, (r,\theta,\phi) \in S_E$ generated by the wound itself. This is because a cut creates an immediate, significant gradient at the wound edges (evidence has been already found for extracellular H_2O_2 gradients mediating wound detection in zebrafish [56]). If relevant variables (such as $\vec{F_N}$ 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—including mammal tissue culture [57], plant growth [58], human gene expression [59], cytoskeleton organization and general embryo development ([60] and references therein)—have been described. A general explanation proposed for these effects is that microgravity introduces a significant degree of mechanical perturbation on critical structures for cells and tissues which as a whole would be the "gravity sensors" [61]. Without dismissing these structural perturbations as relevant, I suggest that a key perturbation on development elicitable by microgravity is a significant alteration—with respect to standard gravity—of the instantaneous $\vec{F_N}$ 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 [62]) and/or a

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 further extrinsic support. Whereas this specific requirement of water has been associated to embryo expansion and metabolic activation of the seeds [63, 64], I submit that it is also associated to the fundamental need for a medium in S_E where the critical $\vec{F_N}$ gradients can emerge. This is because such gradients are in turn required for the intrinsic regulation of the asymmetric divisions already shown critical for cell differentiation in plants [65].

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Concluding remarks

The analysis conducted to search for the theoretical proof of principle in this work encompassed two relevant simplifications or approximations: gene expression levels were represented theoretically by instantaneous transcription rates, which in turn where approximated by mRNA abundance in the analysis. These steps were justified since (i) the correlation between gene expression and mRNA abundance has been clearly established as positive and significant in spite of the limitations of the techniques available [66, 67], (ii) if gene expression can be accurately expressed as a linear transformation of mRNA abundance as the control variable, the *ctalk_non_epi* profiles will remain unchanged (see details in Materials and Methods) and, (iii) the association between *ctalk_non_epi* profiles and cell differentiation states was robust with respect to these simplifications and approximations as shown in the Results.

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

On the other hand, I wish to underscore that the critique of the epigenetic landscape approach presented in the introduction (in terms of its supposed ability to explain the self-regulatory dynamics of cell differentiation) is completely independent from a potential falsification of the theory. Even that being the case, I argue that if future research keeps on elucidating the mechanisms propagating changes of gene expression to an arbitrarily high level of detail—while failing to recognize that the constraints that truly regulate changes[20] must be dynamically uncorrelated yet coupled to the constraints that propagate those changes—advances in the fundamental understanding of the evolution and self-regulatory dynamics of differentiated multicellularity will not be significant.

What underpins this view is that scientifically tenable (i.e. instantaneous) teleological dynamics in nature—unless we are still willing to talk about intrinsically teleological concepts like function, regulation, agency, courtship or planning in all fields of biology while holding they are fundamentally meaningless—must be dynamically uncorrelated to the lower-order dynamics they emerge from. Furthermore, the only way such requisite can be fulfilled is that an intrinsic higher-order constraint emerges from the synergistic coupling of lower-order constraints, as Terrence Deacon first proposed. Whereas these thermodynamically spontaneous, intrinsic constraints are dependent on molecular substrates embodying, propagating, and coupling them 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 (and any living system in general) is no mechanism or machine of any type (e.g. autopoietic [68])—interconnecting in this case a eukaryotic cell population—for mechanisms and machines fundamentally entail an *explicit correlation* between the dynamics within them.

[20] Whatever those constraints are if not the ones described in this theory.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Thus, the emergence of differentiated multicellularity throughout evolution and in every successful ontogenetic process has been—and still is—the emergence of unprecedented, constraint-based, thermodynamic *selves* in the natural world; *selves* which no machine or mechanism could ever be.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Materials and Methods

Data collection

The genomic coordinates of all annotated RefSeq TSSs for the hg19 (*Homo sapiens*), mm9 (*Mus musculus*), and dm3 (*Drosophila melanogaster*) assemblies were downloaded from the UCSC database. Publicly available tandem datafiles of ChIP-seq[21] on histone H3 modifications and RNA-seq[22] for each analyzed cell sample in each species were downloaded from the ENCODE, modENCODE or NCBI's SRA databases [69, 70, 71, 72, 73, 74, 75].

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

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

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

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

See Supplementary Information for the datafile lists in detail.

[21] Comprising 1x36bp, 1x50bp, and 1x75bp reads, depending on the data series (details available via GEO accession codes listed in Supplementary Information).

[22] Comprising 1x36bp, 1x100bp, and 2x75bp reads, depending on the data series (details available via GEO accession codes listed in Supplementary Information).

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

ChIP-seq read count processing

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

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

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

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

was applied (excluding failed-QC reads and duplicate reads by default) directly on the original BAM[23] file to generate the corresponding read count file in BED format.

RNA-seq data processing

The processed data were mRNA abundances in FPKM at RefSeq TSSs. When the original format was GTF (containing already FPKM values, as in the selected ENCODE RNA-seq datafiles for *H. sapiens*), those values were used directly in the analysis. When the original format was SAM, each datafile was pre-processed by first sorting it to generate then a BAM file using `samtools view -bS`. If otherwise the original format was BAM, mRNA levels at RefSeq TSSs were then calculated with FPKM as unit using *Cufflinks* [76] directly on the original file with the following options:

[23] The BAI file is required implicitly.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


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

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

Preparation of data input tables

For each of the three species, all $\mathrm{TSS_{def}}$—defined as those TSSs with measured mRNA abundance (i.e. FPKM > 0) in all cell types/cell samples—were determined. The number of $\mathrm{TSS_{def}}$ found for each species were $N_{\mathrm{TSS_{def}}}(\textit{Homo sapiens}) = 14{,}742$, $N_{\mathrm{TSS_{def}}}(\textit{Mus musculus}) = 16{,}021$, and $N_{\mathrm{TSS_{def}}}(\textit{Drosophila melanogaster}) = 11{,}632$. 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 $\mathrm{TSS_{def}}$ genomic coordinate: "−2000", "−1800", "−1600", "−1400", "−1200", "−1000", "−800", "−600", "−400", "−200", "0" ($\mathrm{TSS_{def}}$ 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 $\mathrm{TSS_{def}}$ 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[24]:

$$\begin{array}{cccccccccc} & \mathrm{H3[1]_-2000} & \dots & \mathrm{H3[}n_m\mathrm{]_-2000} & \cdots & \mathrm{H3[1]_3800} & \dots & \mathrm{H3[}n_m\mathrm{]_3{,}800} & \mathrm{FPKM} \\ 1 & & & & & & & & \\ \vdots & & & & & & & & \\ N_{\mathrm{TSS_{def}}} & & & & & & & & \end{array}$$

The tables were written then to these data files:

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

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

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

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Computation of *ctalk_non_epi* profiles

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

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

where the second summand $\mathrm{Cov}(X_j, Y_k|Z)$ is the partial covariance between X_j and Y_k given Z. It is easy to see that $\mathrm{Cov}(X_j, Y_k|Z)$ is a local approximation of Nanney's constraints $\boldsymbol{C_N}$ on histone H3 modifications, as anticipated in the preliminary theoretical definitions[25]. To make the *ctalk_non_epi* profiles comparable however, $\mathrm{Cov}\,(X_j, Y_k|Z)$ values have to be normalized[26] by the standard deviations of the residuals of X_j and Y_k with respect to Z. In other words, the partial correlation $\mathrm{Cor}(X_j, Y_k|Z)$ values were needed. Nevertheless, a correlation value does not have a straightforward interpretation, whereas its square—typically known as ***coefficient of determination***, ***effect size of the correlation***, or simply r^2—does: it represents the relative (i.e. fraction of) variance of one random variable explained by the other. For this reason, $\mathrm{Cor}(X_j, Y_k|Z)^2$ was used to represent the strength of the association, and then multiplied by the sign of the correlation to represent the direction of the association. Thus, after $\log_2$-transforming the X_j, Y_k and Z data, each pairwise combination of bin-specific histone H3 modifications $\{X_j, Y_k\}$ contributed with the value

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

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

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

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

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

The justification for this additional filtering was that some pairwise partial correlation values were expected a priori to be strong and significant, which was later confirmed. Namely, (i) those involving the same histone H3 modification in the same amino acid residue (e.g. $\mathrm{Cor}(\mathrm{H3K9ac_}{-200}, \mathrm{H3K9ac_}{-400}|\mathrm{FPKM}) > 0$; $\mathrm{Cor}(\mathrm{H3K4me3_}{-200}, \mathrm{H3K4me3_}{-200}|\mathrm{FPKM}) = 1$) (ii) those involving a different type of histone H3 modification in the same amino acid residue (e.g. $\mathrm{Cor}(\mathrm{H3K27ac_}{-800}, \mathrm{H3K27me3_}{-600}|\mathrm{FPKM}) < 0$), and (iii) those involving the same type of histone H3 modification in the same amino acid residue (e.g. $\mathrm{Cor}(\mathrm{H3K4me2_}{-400}, \mathrm{H3K4me3_}{-400}|\mathrm{FPKM}) > 0$) in part because ChIP-antibody cross reactivity has been shown able to introduce artifacts on the accurate assessment of some histone-crosstalk associations [20, 21]. For these reasons, in each species all pairwise combinations of histone H3 modifications involving the same amino acid residue were then identified as "trivial" and excluded from the *ctalk_non_epi* profiles construction. E.g., since for ***Mus musculus*** the comprised histone modifications were H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3 ($n_m = 5$), the pairwise combinations H3K4me1–H3K4me3 and H3K27ac–H3K27me3 were filtered out. Therefore, the length of the ***Mus musculus*** *ctalk_non_epi* profiles was $({}^5C_2 - 2) \times 900 = 7{,}200$.

Statistical significance assessment

The statistical significance of the partial correlation $\mathrm{Cor}(X_j, Y_k|Z)$ values, necessary for constructing the *ctalk_non_epi* profiles, was estimated using Fisher's z-transformation [77]. Under the null hypothesis $\mathrm{Cor}(X_j, Y_k|Z) = 0$ the statistic $z = \sqrt{N_{\mathrm{TSS_{def}}} - |Z| - 3}\ \frac{1}{2}\ln\left(\frac{1+\mathrm{Cor}(X_j,Y_k|Z)}{1-\mathrm{Cor}(X_j,Y_k|Z)}\right)$, where $N_{\mathrm{TSS_{def}}}$ is the sample size and $|Z| = 1$ (i.e. one control variable), follows asymptotically a $N(0,1)$ distribution. The p-values can be then computed easily using the $N(0,1)$ probability function.

Multiple comparisons correction of the p-values associated to each *ctalk_non_epi* profile was performed using the Benjamini-Yekutieli method [78]. The parameter used was the number of all possible[27] comparisons: ${}^{(n_m \times 30)}C_2$. From the resulting q-values associated to each *ctalk_non_epi* profile an empirical cumulative distribution was obtained, which in turn was used to compute a threshold t. The value of t was optimized to be the maximum value such that within the q-values smaller than t is expected less than 1 false-positive partial correlation. Consequently, if q-value$[i] \geq t$ then the associated partial correlation value was identified as not significant (i.e. zero) in the respective *ctalk_non_epi* profile.

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Unsupervised hierarchical clustering of *ctalk_non_epi* and mRNA abundance profiles

The goal of this step was to evaluate the significant *ctalk_non_epi*-profile clusters—if any—in the phenograms (i.e. "phenotypic similarity dendrograms") obtained from unsupervised hierarchical clustering analyses (unsupervised HCA). For each species, the analyses were conducted on (i) the *ctalk_non_epi* profiles of each cell type/sample (**Figure 1A, 1C, and 1E**) and (ii) the $\log_2$-transformed FPKM profiles (i.e mRNA abundance) of each cell type/sample (**Figure 1B, 1D, and 1F**). 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[28], hence it was the metric used here. On the other hand, the cluster-linkage method chosen was the "average" method or UPGMA (Unweighted Pair Group Method with Arithmetic Mean) in which the distance $D(A, B)$ between any clusters A and B is defined as

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

that is, the mean of all distances $d_r(pro[k], pro[l])$ such that $pro[k] \in A$ and $pro[l] \in B$ (this method was chosen because it has been shown to yield the highest cophenetic correlation values when using the "correlation distance" metric [79]). Cluster statistical significance was assessed as *au* (approximately unbiased) and *bp* (bootstrap probability) significance scores by non-parametric bootstrap resampling using the *Pvclust* [24] add-on package for the *R* software [80]. The number of bootstrap replicates in each analysis was 10,000.

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Suitability of FPKM as unit of mRNA abundance

Previous research has pinpointed that FPKM may not always be an adequate unit of transcript abundance in differential expression studies. It was shown that, if transcript size distribution varies significantly among the samples, FPKM/RPKM[29] will introduce biases. For this reason another abundance unit TPM (transcripts per million)—which is a linear transformation of the FPKM value for each sample—was proposed to overcome the limitation [81]. However, this issue was not a problem for this study.

This is because partial correlation, used to construct the *ctalk_non_epi* profiles later subject to HCA, is invariant under linear transformations of the control variable Z (i.e. $\mathrm{Cor}(X_j, Y_k|Z) = \mathrm{Cor}(X_j, Y_k|aZ + b)$ for any two scalars $\{a, b\}$). Importantly, this property also implies that *ctalk_non_epi* profiles are controlling not only for mRNA abundance but also for any other biological variable displaying a strong linear relationship with mRNA abundance (e.g. chromatin accessibility represented by DNase I hypersensitivity, as shown in [20]). Similarly, the unsupervised hierarchical clustering of mRNA abundance profiles is invariant under linear transformations of the profiles, since $\mathrm{Cor}(Z_i, Z_j) = \mathrm{Cor}(aZ_i + b, cZ_j + d)$ provided $ac > 0$.

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

I wish to thank the following people:

- John Tyler Dodge, horn soloist at the *Orquesta Filarmónica de Santiago*, for reviewing most of the English in this paper and his valuable questions, which pushed me to the limit of my abilities in the purpose of making this paper self-explanatory.
- Miguel Allende, director of the FONDAP Center for Genome Regulation (see details in the institutional acknowledgements below).
- Alejandro Maass, professor at the Center for Mathematical Modeling (CMM), Universidad de Chile, for his special interest in this work and his interesting questions.
- My anonymous colleagues who reviewed the grant proposal on behalf of FONDECYT (see below).

Also, I wish to thank the following institutions:

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 preprint to bioRxiv.

Copyright

The copyright holder for this preprint is the author. It is made made available under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Funding

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

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

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

[3] Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663–676. doi: 10.1016/j.cell.2006.07.024.

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

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

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

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

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

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

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

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

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

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

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

[15] Rose LS, Kemphues KJ (1998) Early patterning of the *C. elegans* embryo. Annual Review of Genetics 32: 521–545. doi: 10.1146/annurev.genet.32.1.521.

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

[28] Fraser P, Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. Nature 447: 413–417. doi: 10.1038/nature05916.

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

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

[50] Combs PA, Eisen MB (2013) Sequencing mrna from cryo-sliced *Drosophila* embryos to determine genome-wide spatial patterns of gene expression. PLoS One 8: e71820. doi: 10.1371/journal.pone.0071820.

[51] Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, et al. (2013) Quantitative assessment of single-cell RNA-sequencing methods. Nat Meth 11: 41–46. doi: 10.1038/nmeth.2694.

[52] Roy S, Morse D (2012) A full suite of histone and histone modifying genes are transcribed in the dinoflagellate *Lingulodinium*. PLoS One 7: e34340. doi: 10.1371/journal.pone.0034340.

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

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

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

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

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

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

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

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

[61] Ingber D (1999) How cells (might) sense microgravity. The FASEB Journal 13: S3–S15.

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

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

[68] Varela FG, Maturana HR, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. Biosystems 5: 187–196.

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

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

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

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

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

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

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

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

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

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

[82] Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana: University of Illinois Press.

[83] Watanabe S (1960) Information theoretical analysis of multivariate correlation. IBM Journal of research and development 4: 66–82.

[84] Voss TC, Hager GL (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet 15: 69-81. doi: 10.1038/nrg3623.

[85] Altun Z, Hall D (2002). Wormatlas.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix

Formal theoretical definitions and notation

The following definitions and notation regard molecular dynamics and spatial topology. To avoid ambiguity, the definitions regarding molecular dynamics will be derived from instantaneously defined random variables—measurable only partially and approximately but easily imaginable from a fundamental point of view—using Shannon's information theory [82] measures[30]. For further explanatory convenience, these random variables will be seen sometimes as sets. Definitions regarding spatial topology will be derived from instantaneously-specified spherical coordinates given the spherical/circular symmetry that developing embryos and cell populations display in general. Additionally, a brief glossary will be provided with the most relevant notation and concepts, described there in less rigorous terms yet logically sufficient for the theoretical formulation. If desired, the reader may then skip to this glossary and return to the formal definitions at any point later.

Spatial topology

Let $\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)}$ be all cells in a given organism or cell population[31] of the eukaryotic species $\boldsymbol{X}$ at a given instant $\boldsymbol{t}$, spatially-specified in spherical coordinates. These spherical coordinates are r (radial distance), θ (azimuthal angle), and ϕ (polar angle). The origin of the coordinate system is the centroid of the cell population or embryo. Let $\boldsymbol{r}_{(n;t)}(\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)})$ be the radius of the entire cell population or embryo.

Definition *Overall space*: $\boldsymbol{S_O}(\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)})$

$\boldsymbol{S_O}(\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)}) \quad = \quad \left\{(r,\theta,\phi)\,\middle|\, r \leq 0 \leq \boldsymbol{r}_{(n;t)}(\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)}),\ 0 \leq \theta < 2\pi,\ 0 \leq \phi \leq \pi\right\}$; that is, the set of all points (r,θ,ϕ) within the radius $\boldsymbol{r}_{(n;t)}$.

Remark This is the space occupied by the entire cell population or embryo. It can be expressed as the sum of the space occupied by all individual cells in the population/embryo plus the associated extracellular space.

Definition *Cell-occupied space*: $\boldsymbol{S_C}(\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)})$

$\boldsymbol{S_C}(\boldsymbol{X}_{(1;t)},\ldots,\boldsymbol{X}_{(n;t)}) = \left\{(r,\theta,\phi)\,\middle|\,(r,\theta,\phi) \in \bigcup_{i=1}^{n} \boldsymbol{S_O}(\boldsymbol{X}_{(i;t)})\right\}$; that is, the union of the sets of points (r,θ,ϕ) spatially-specifying all $\boldsymbol{n}$ individual cells at a given instant $\boldsymbol{t}$.

[30] In a strict sense, molecular dynamics should be represented by multivariate random variables and as a consequence generalized measures (such as Watanabe's total correlation [83]) of statistical dependence would be needed. Univariate random variables were preferred here for simplicity in the notation.

[31] A single cell or zygote will be considered a cell population where $\boldsymbol{n} = 1$. All cells will be treated geometrically as spheres unless specified later. A dividing cell will be regarded as a single cell until division is complete.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Remark This is the space occupied by all n individual cells in the population/embryo at a given instant t.

Definition *Extracellular space*: $S_E(X_{(1;t)},\ldots,X_{(n;t)})$

$S_E(X_{(1;t)},\ldots,X_{(n;t)}) = S_O(X_{(1;t)},\ldots,X_{(n;t)}) \setminus S_C(X_{(1;t)},\ldots,X_{(n;t)})$; that is, the set difference of S_O and S_C, i.e. the set of points (r,θ,ϕ) such that $(r,\theta,\phi) \in S_T(X_{(1;t)},\ldots,X_{(n;t)})$ and $(r,\theta,\phi) \notin S_T(X_{(1;t)},\ldots,X_{(n;t)})$.

Remark Whereas this may seem an over-formalized definition of the widely-known concept of extracellular space, the theory postulated here shows its usefulness for describing some corollaries rigorously.

Molecular dynamics

Let $X_{(i;t)}$ be the ith cell[32] of a given organism or cell population of the eukaryotic species X at a given instant t, let $G(X_{(i;t)})$ be its genomic sequence, let $T(X_{(i;t)})$ be the instantaneous transcription rate at the transcription start site, let $F(X_{(i;t)})$ be its entire molecular phenotype spatially-specified with respect to the transcription start site (note that $F(X_{(i;t)})$ describes implicitly specific molecular abundances), let $F^{\circ}(X_{(i;t)})$ be the molecular phenotype of the nucleus of $X_{(i;t)}$, and let $F^{\star}(X_{(i;t)})$ be the cell's molecular phenotype that is membrane-exchangeable with the extracellular space S_E by facilitated diffusion (note that $F^{\circ}(X_{(i;t)}), F^{\star}(X_{(i;t)}) \subset F(X_{(i;t)})$). Importantly, the set $G(X_{(i;t)}) \cup T(X_{(i;t)}) \cup F(X_{(i;t)})$ describes an instantaneously realized state (of interest for this work), and it is not to be confused with a state space—the set of all realizable states—in dynamical systems theory. For notation simplicity, the argument $X_{(i;t)}$ will be implicit henceforth unless necessary.

Definition *Waddington's constraints* $C_W(X_{(i;t)}) = I(F;T|G)$; that is, the conditional mutual information between F° (spatially-specified nuclear phenotype) and T (instantaneous transcription rate at the TSS) given the value of G (genomic sequence).

Remark Waddington's constraints can be equivalently expressed in terms of Shannon's conditional entropies as $C_W(X_{(i;t)}) = H(F^{\circ}|G) - H(F^{\circ}|T,G)$.

Remark $C_W(X_{(i;t)})$ can be interpreted as a measure of the statistical dependence (i.e. constraint) between F° (spatially-specified nuclear phenotype) and T (transcription rates) for a given value of G.

Remark These constraints are determined by (i) the spatial coordinates in F° with respect to the TSS and (ii) the specific affinities of DNA with respect to the phenotypic elements specified in F° for any given value of G.

[32] A dividing cell will be regarded as a single cell until division is complete.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Definition *Waddington's embodiers* $F_W(X_{(i;t)})$ is the largest subset of $F^\circ(X_{(i;t)})$ such that $I(F_W;T|G) > 0$.

Remark In data analysis, the set $F_w(X_{(i;t)})$ is a subset of $F_W(X_{(i;t)})$ if sample $I(F_w;T|G)$ is significantly greater than zero.

Remark As mentioned earlier, the quantitative assessment of the hereby defined as Waddington's constraints has made possible to predict transcript abundance states from histone modification ChIP-seq enrichment profiles near TSSs with high accuracy (R~0.9) [19].

Example The previous remark implies the following: if T is approximated by transcript abundance and F_w is approximated by ChIP-seq enrichment, then histone H3 modifications can be specified as Waddington's embodiers F_W.

Remark If $X_{(j,t+\Delta t)}$ is a daughter cell of $X_{(i;t)}$ and if there are two sets $F_h(X_{(i;t)}) \subseteq F_W(X_{(i;t)})$ and $F_h(X_{(j,t+\Delta t)}) \subseteq F_W(X_{(j,t+\Delta t)})$ such that $I(F_h(X_{(i;t)});F_h(X_{(j,t+\Delta t)})) > 0$ (i.e. if Waddington's constraints are propagated in a heritable manner), then $(F_h(X_{i,t}) \cup F_h(X_{(j,t+\Delta t)}))$ specifies spatially a set of molecular substrates E that is customarily labeled as "epigenetic regulators". Whereas this is a trivial corollary in its strict sense, it highlights the substrate-centered character—as opposed to constraint-centered—of the traditional approach known as epigenetic landscape. Additionally, this corollary reinforces a point raised in the introduction: the misleading—and if taken in a strict sense, logically inconsistent—character of the "regulator" label on any molecular substrates satisfying the conditions that define the set E.

Definition *Waddington's extracellular propagators* $F_{\vec{W}}(X_{(i;t)})$ is the largest subset of $F^\star - F_W$ such that there is a minimal time interval Δt and a quantity $I_{\vec{W}(\Delta t)} > 0$ for which $I(F_W(X_{(i;t+\Delta t)});F_{\vec{W}}(X_{(i;t)})) = I_{W(t)} + I_{\vec{W}(\Delta t)} \implies I(F_W(X_{(i;t)});F_{\vec{W}}(X_{(i;t)})) = I_{W(t)}$; that is, the largest subset of $F^\star$ that is not a subset of Waddington's embodiers F_W at the instant t but elicits a significant change (measurable as the mutual information $I_{\vec{W}(\Delta t)}$) in Waddington's embodiers F_W observable after Δt.

Remark In data analysis, the set $F_{\vec{w}}(X_{(i;t)})$ is a subset of $F_{\vec{W}}(X_{(i;t)})$ if sample $I(F_W(X_{(i;t+\Delta t)});F_{\vec{w}}(X_{(i;t)})) - I(F_W(X_{(i;t)});F_{\vec{w}}(X_{(i;t)}))$ is significantly greater than zero and if the logical implication in the definition (i.e. causality) can be then inferred given the experimental design.

Remark Note that if $F_{\vec{W}}(X_{(i;t)}) \neq \emptyset$, then the alleles specified in G must account for all gene products in F (spatially-specified phenotype) necessary for the facilitated diffusion of the molecules specified in $F_{\vec{W}}(X_{(i;t)})$ (e.g. functional and sufficiently abundant protein pores or carriers, or intracellular transducers if needed).

Remark Waddington's extracellular propagators $F_{\vec{W}}$ may need a certain set of Waddington's intracellular propagators (defined with respect to the set $F - (F^\star \cup F_W)$) to satisfy their own defining condition.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Definition *Nanney's constraints* $C_N(X_{(i;t)}) = I(F^\circ; F_W|T)$; that is, the conditional mutual information between F° (spatially-specified nuclear phenotype) and F_W (Waddington's propagators) given the value of T (instantaneous transcription rate at the TSS).

Remark Nanney's constraints can be equivalently expressed in terms of Shannon's conditional entropies as $C_N(X_{(i;t)}) = H(F^\circ|T) - H(F^\circ|F_W, T)$.

Remark $C_N(X_{(i;t)})$ can be interpreted as a measure of the statistical dependence (i.e. constraint) between F° (spatially-specified nuclear phenotype) and F_W (Waddington's propagators) for a given value of T.

Remark These constraints are determined by (i) the spatial coordinates in F° and Waddington's propagators F_W with respect to each other and (ii) the kinetic and structural constraints governing the interactions between Waddington's propagators in F_W and the entire nuclear phenotype in F° for any given value of T.

Definition *Nanney's embodiers* $F_N(X_{(i;t)})$ is the largest subset of F° such that $I(F_N; F_W|T) > 0$.

Remark In data analysis, the set $F_n(X_{(i;t)})$ is a subset of $F_N(X_{(i;t)})$ if sample $I(F_n; F_W|T)$ is significantly greater than zero.

Example The work cited previously [19], which demonstrated the high predictive power of histone modification profiles on transcript abundance, did so developing a trainable multivariate linear regression model. In the work presented in this paper, such linearity and demonstrated predictive power together made it possible to represent Nanney's constraints with the *ctalk_non_epi* profiles (see Materials and Methods). In turn, the high statistical significance of these *ctalk_non_epi* profiles (shown previously in the results) implies that histone H3 modifications can be specified as Nanney's embodiers F_N.

Remark At this point it is possible formalize what was highlighted in the beginning of this discussion: if $X_{(i;t)}$ is a human, mouse, or fruit fly cell and F_{H3} is its set of histone H3 modifications at transcription start sites, then $F_{\mathrm{H3}} \subseteq (F_W(X_{(i;t)}) \cap F_N(X_{(i;t)}))$. In other words, histone H3 modifications at the TSS are specifiable as Waddington's embodiers F_W and as Nanney's embodiers F_N *simultaneously*. This critical result will be generalized theoretically for cells of any differentiated multicellular organism.

Definition *Nanney's extracellular propagators* $F_N^{\rightarrow}(X_{(i;t)})$ is the largest subset of $F^{\star} - F_N$ such that there is a minimal time interval Δt and a quantity $I^{\rightarrow}_{N(\Delta t)} > 0$ for which $I(F_N(X_{(i;t+\Delta t)}); F_N^{\rightarrow}(X_{(i;t)})) = I_{N(t)} + I^{\rightarrow}_{N(\Delta t)} \implies I(F_N(X_{(i;t)}); F_N^{\rightarrow}(X_{(i;t)})) = I_{N(t)}$; that is, the largest subset of $F^{\star}$ that is not a subset of Nanney's embodiers F_N at the instant t but elicits a significant change (measurable as the mutual information $I^{\rightarrow}_{N(\Delta t)}$) in Nanney's embodiers F_N observable after Δt.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Remark In data analysis, the set $\boldsymbol{F_n^{\rightarrow}}$ is a subset of $\boldsymbol{F_n^{\rightarrow}}(\boldsymbol{X_{(i;t)}})$ if sample $\boldsymbol{I}(\boldsymbol{F_N}(\boldsymbol{X_{(i;t+\Delta t)}});\boldsymbol{F_n^{\rightarrow}}(\boldsymbol{X_{(i;t)}})) - \boldsymbol{I}(\boldsymbol{F_N}(\boldsymbol{X_{(i;t)}});\boldsymbol{F_n^{\rightarrow}}(\boldsymbol{X_{(i;t)}}))$ is significantly greater than zero and if the logical implication in the definition (i.e. causality) can be then inferred given the experimental design.

Remark Note that if $\boldsymbol{F_N^{\rightarrow}}(\boldsymbol{X_{(i;t)}}) \neq \emptyset$, then the alleles specified in $\boldsymbol{G}$ must account for all gene products in $\boldsymbol{F}$ (spatially-specified phenotype) necessary for the facilitated diffusion of the molecules specified in $\boldsymbol{F_N^{\rightarrow}}(\boldsymbol{X_{(i;t)}})$ (e.g. functional and sufficiently abundant protein pores or carriers, or intracellular transducers if needed).

Remark Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ may need a certain set of Nanney's intracellular propagators (defined with respect to the set $\boldsymbol{F} - (\boldsymbol{F}^{\star} \cup \boldsymbol{F_N})$) to satisfy their own defining condition.

Remark Whereas the existence of Nanney's extracellular propagators $\boldsymbol{F_W^{\rightarrow}}$ is indisputable [84], to my knowledge evidence for the existence of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ has not been searched for and thus, not surprisingly, is currently absent. However, the appearance of Nanney's extracellular propagators $\boldsymbol{F_N^{\rightarrow}}$ was a necessary condition for the evolution of differentiated multicellularity, as it is proposed in the theory.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [85])

	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.

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Information

Homo sapiens source data of ChIP-seq on histone H3 modifications (BAM/BAI 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/wgEncodeBroadHistone/`

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

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Homo sapiens source data of RNA-seq transcript abundance in FPKM (GTF files) [74]

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

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

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

Continued on next page

© 2015 Felipe A. Veloso CC BY

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Mus musculus source data of ChIP-seq on histone H3 modifications (SRA files) [75, 73]

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	H3K4mel	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	H3K4mel	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	H3K4mel	1	GSM881369	SRR414/SRR414958/SRR414958.sra
E14-day6	H3K4me3	1	GSM881371	SRR414/SRR414960/SRR414960.sra
Heart (8 wks/o)	H3K27ac	1	GSM1000093	SRR566/SRR566827/SRR566827.sra
Heart (8 wks/o)	H3K27ac	2	GSM1000093	SRR566/SRR566828/SRR566828.sra
Heart (8 wks/o)	H3K27me3	1	GSM1000131	SRR566/SRR566903/SRR566903.sra
Heart (8 wks/o)	H3K27me3	2	GSM1000131	SRR566/SRR566904/SRR566904.sra
Heart (8 wks/o)	H3K36me3	1	GSM1000130	SRR566/SRR566901/SRR566901.sra
Heart (8 wks/o)	H3K36me3	2	GSM1000130	SRR566/SRR566902/SRR566902.sra
Heart (8 wks/o)	H3K4mel	1	GSM769025	SRR317/SRR317255/SRR317255.sra
Heart (8 wks/o)	H3K4mel	2	GSM769025	SRR317/SRR317256/SRR317256.sra
Heart (8 wks/o)	H3K4me3	1	GSM769017	SRR317/SRR317239/SRR317239.sra
Heart (8 wks/o)	H3K4me3	2	GSM769017	SRR317/SRR317240/SRR317240.sra
Heart (8 wks/o)	Input	1	GSM769032	SRR317/SRR317269/SRR317269.sra
Heart (8 wks/o)	Input	2	GSM769032	SRR317/SRR317270/SRR317270.sra
Liver (8 wks/o)	H3K27ac	1	GSM1000140	SRR566/SRR566921/SRR566921.sra
Liver (8 wks/o)	H3K27ac	2	GSM1000140	SRR566/SRR566922/SRR566922.sra
Liver (8 wks/o)	H3K27me3	1	GSM1000150	SRR566/SRR566941/SRR566941.sra
Liver (8 wks/o)	H3K27me3	2	GSM1000150	SRR566/SRR566942/SRR566942.sra
Liver (8 wks/o)	H3K36me3	1	GSM1000151	SRR566/SRR566943/SRR566943.sra
Liver (8 wks/o)	H3K36me3	2	GSM1000151	SRR566/SRR566944/SRR566944.sra

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

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

Mus musculus source data of RNA-seq (BAM files) [75, 73]

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) [69, 71]

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

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

Developmental time point/period	Antibody	GEO Accession	File URL suffix
4-8h embryos	H3K27me3	GSM439447	SRR030359/SRR030359.sra
4-8h embryos	H3K4me1	GSM401406	SRR030294/SRR030294.sra
4-8h embryos	H3K4me3	GSM400674	SRR030287/SRR030287.sra
4-8h embryos	H3K9ac	GSM401405	SRR030293/SRR030293.sra
4-8h embryos	H3K9me3	GSM439456	SRR030368/SRR030368.sra
4-8h embryos	Input	GSM400675	SRR030288/SRR030288.sra
8-12h embryos	H3K27ac	GSM432583	SRR030332/SRR030332.sra
8-12h embryos	H3K27me3	GSM439446	SRR030358/SRR030358.sra
8-12h embryos	H3K4me1	GSM432593	SRR030342/SRR030342.sra
8-12h embryos	H3K4me3	GSM432585	SRR030334/SRR030334.sra
8-12h embryos	H3K9ac	GSM432592	SRR030341/SRR030341.sra
8-12h embryos	H3K9me3	GSM439455	SRR030367/SRR030367.sra
8-12h embryos	Input	GSM432636	SRR030346/SRR030346.sra
12-16h embryos	H3K27ac	GSM432582	SRR030331/SRR030331.sra
12-16h embryos	H3K27me3	GSM439445	SRR030357/SRR030357.sra
12-16h embryos	H3K4me1	GSM432591	SRR030340/SRR030340.sra
12-16h embryos	H3K4me3	GSM432580	SRR030329/SRR030329.sra
12-16h embryos	H3K9ac	GSM439458	SRR030370/SRR030370.sra
12-16h embryos	H3K9me3	GSM439454	SRR030366/SRR030366.sra
12-16h embryos	Input	GSM432634	SRR030344/SRR030344.sra
16-20h embryos	H3K27ac	GSM401401	SRR030289/SRR030289.sra
16-20h embryos	H3K27me3	GSM439444	SRR030356/SRR030356.sra
16-20h embryos	H3K4me1	GSM401403	SRR030291/SRR030291.sra
16-20h embryos	H3K4me3	GSM400658	SRR030271/SRR030271.sra
16-20h embryos	H3K9ac	GSM401402	SRR030290/SRR030290.sra
16-20h embryos	H3K9me3	GSM439453	SRR030365/SRR030365.sra
16-20h embryos	Input	GSM400659	SRR030272/SRR030272.sra
20-24h embryos	H3K27ac	GSM401423	SRR030311/SRR030311.sra
20-24h embryos	H3K27me3	GSM439443	SRR030355/SRR030355.sra
20-24h embryos	H3K4me1	GSM439464	SRR030376/SRR030376.sra
20-24h embryos	H3K4me3	GSM400672	SRR030285/SRR030285.sra
20-24h embryos	H3K9ac	GSM401424	SRR030312/SRR030312.sra
20-24h embryos	H3K9me3	GSM439452	SRR030364/SRR030364.sra
20-24h embryos	Input	GSM400673	SRR030286/SRR030286.sra
L1 larvae	H3K27ac	GSM432581	SRR030330/SRR030330.sra
L1 larvae	H3K27me3	GSM439442	SRR030354/SRR030354.sra
L1 larvae	H3K4me1	GSM432588	SRR030337/SRR030337.sra
L1 larvae	H3K4me3	GSM400662	SRR030275/SRR030275.sra
L1 larvae	H3K9ac	GSM401422	SRR030310/SRR030310.sra
L1 larvae	H3K9me3	GSM439451	SRR030363/SRR030363.sra
L1 larvae	Input	GSM400663	SRR030276/SRR030276.sra
L2 larvae	H3K27ac	GSM401419	SRR030307/SRR030307.sra
L2 larvae	H3K27me3	GSM439441	SRR030353/SRR030353.sra
L2 larvae	H3K4me1	GSM401421	SRR030309/SRR030309.sra
L2 larvae	H3K4me3	GSM400668	SRR030281/SRR030281.sra
L2 larvae	H3K9ac	GSM401420	SRR030308/SRR030308.sra
L2 larvae	H3K9me3	GSM439450	SRR030362/SRR030362.sra
L2 larvae	Input	GSM400669	SRR030282/SRR030282.sra
Pupae	H3K27ac	GSM401413	SRR030301/SRR030301.sra
Pupae	H3K27me3	GSM439439	SRR030351/SRR030351.sra
Pupae	H3K4me1	GSM401415	SRR030303/SRR030303.sra

Continued on next page

© 2015 Felipe A. Veloso

bioRxiv preprint doi: https://doi.org/10.1101/016840; this version posted March 27, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Continued from previous page

Developmental time point/period	Antibody	GEO Accession	File URL suffix
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 source data of RNA-seq (SAM files) [69, 71]

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

© 2015 Felipe A. Veloso