Modularity map of the network of human cell differentiation

Viviane Galvão[1,2], José G. V. Miranda[2], Roberto F. S. Andrade[2],
José S. Andrade Jr.[3,4], Lazaros K. Gallos[5], Hernán A. Makse[3,5]

[1]*Departamento de Ciências Biológicas,*
Universidade Estadual de Feira de Santana,
44036-900, Feira de Santana, Bahia, Brazil

[2]*Instituto de Física, Universidade Federal da Bahia,*
40210-210, Salvador, Bahia, Brazil.

[3]*Departamento de Física, Universidade Federal do Ceará,*
Campus do Pici, 60455-760, Fortaleza, Ceará, Brazil.

[4] *IfB, HIF E12, ETH Honggerberg, 8093 Zurich, Switzerland*

[5]*Levich Institute and Physics Department,*
City College of New York, New York, New York 10031, USA

(Dated: February 14, 2022)

Abstract

Cell differentiation in multicellular organisms is a complex process whose mechanism can be understood by a reductionist approach, in which the individual processes that control the generation of different cell types are identified. Alternatively, a large scale approach in search of different organizational features of the growth stages promises to reveal its modular global structure with the goal of discovering previously unknown relations between cell types. Here we sort and analyze a large set of scattered data to construct the network of human cell differentiation (NHCD) based on cell types (nodes) and differentiation steps (links) from the fertilized egg to a crying baby. We discover a dynamical law of critical branching, which reveals a fractal regularity in the modular organization of the network, and allows us to observe the network at different scales. The emerging picture clearly identifies clusters of cell types following a hierarchical organization, ranging from sub-modules to super-modules of specialized tissues and organs on varying scales. This discovery will allow one to treat the development of a particular cell function in the context of the complex network of human development as a whole. Our results point to an integrated large-scale view of the network of cell types systematically revealing ties between previously unrelated domains in organ functions.

The cell differentiation process plays a crucial role in the prenatal development of multicellular organisms. Recent advances in the research on stem cell properties and embryonic development have uncovered several steps in the differentiation process [1–8]. Single and multiple sequences of cell differentiation have been identified through in-vivo observations of a particular embryo during early stages of development and through pathology studies of miscarriages during late stages of the process. While the identification of each cell differentiation step has been the subject of intense research, an integrated view of this complex process is still missing. Such a global view promises to reveal features associated with the large-scale modular organization of the cell types [5–7, 9–13] with the purpose of discovering new functional modules between cell types using novel theoretical network analysis for community detection [10–12]. In this letter, we take advantage of the current knowledge on the sequence of cell differentiation processes, which is spread over a vast specialized literature [1–6, 14–28] (see the Supplementary Information SI-Table I and references therein), to reveal and characterize the topological and dynamical features associated with the network of human cell differentiation (NHCD).

I. RESULTS

We construct the NHCD by systematically gathering the scattered information on the evolution of each cell type present in the embryo and fetus from a predecessor with a higher degree of differentiation potential into a more specialized type. The process of cell differentiation is then mapped onto a complex network which consists of 873 nodes connected through 977 edges. The nodes in the network represent distinct cell types reported in the literature [1–6, 14–28] and the edges represent the association between two cell types through a differentiation event.

The initial steps of the NHCD are shown in the inset of Fig. 1, while the resulting network structure is shown in the main panel of Fig. 1 (see also SI-Figs. 5a and 5b). The fertilized egg is followed by the ball stage, and the formation of the primary germ cell layers. Currently, it is known that until the ball stage, cell division is symmetric and produces further totipotent stem cells [1]. These cells then give rise to all the differentiated tissues of the organism as well as the extra-embryonic tissues (placenta, umbilical cord, etc.). Moreover, in the course of the entire process of organism formation, there is a monotonic decrease in the differentiation

potential (totipotent → pluripotent → multipotent → unipotent cells) accompanied with an increase in cell specialization.

Certain types of cells can be generated following more than one path from the fertilized egg. This process generates some closed loops of edges in the network. The NHCD comprises 529 branches of different lengths with each branch ending when the cell types do not undergo further differentiation. Note, however, that the most recent compilation of cell types in normal, healthy, human adults done in [6] reports only 407 cell types. Therefore, not all branch endpoints correspond to cell types in born humans. Thus, not all 873 cell types are present in a human being. Among those absent are the placenta cells that are generated from the fertilized egg during embryo development, as well as other somatic cell types that are important to control embryo and fetus development. The cell types that survive in a human are denoted by filled circles, while non-surviving ones are indicated by empty circles. The complete collected data is listed in the Supplementary Information, including an enumeration of cells and links between the cell types, their time of appearance in days after fecundation (T_a), and the reference to the publications reporting each link. To the best of our knowledge, the structure identified here provides the most complete schematic diagram of the human differentiation process to date.

It is visually apparent from Fig. 1 that the NHCD has a prominent modular structure. The continuous differentiation of cells into more specialized functions naturally leads to the formation of dense isolated clusters in the NHCD. As a first approach to understand this modular structure we cluster cell types in the network of Fig. 1 according to their known functions; different colors indicate 19 functional modules extracted from the literature ($C1 - C19$) (See SI-Table I and references therein. The largest communities were extracted from Refs. [1–6, 14–28]). There is, however, a certain degree of arbitrariness in this modular structure as the separation of the nodes into communities in our dataset is not unique. For instance, community $C12$, the neural lineage, could be divided into two sub-communities, representing the neural and the supporting (glial) cells [1–3, 5, 6, 14, 15]. On the other hand, the neural system module could be merged with the eye system module [1, 5, 6, 14] on a larger scale, since they have a common ancestral cell type. Therefore, a finer or coarser community structure can be extracted from the data when we look at the whole network at different scales of observation; a novel module-detection algorithm is needed to identify these communities in a systematic way.

Graph theoretical concepts allow us to unravel the scale dependence of the modular structure of the NHCD. Graph theory [12] defines the distance between two nodes (also called the chemical distance) as the number of links along the shortest path between the nodes in the network. We use this notion to propose a community detection algorithm that identifies modules of size ℓ composed of highly connected cell types. The algorithm finds the optimal tiling of the network with the smallest possible number of modules, N_B, of size ℓ [13] (each node is assigned to a module or box and all nodes in a module are at distance smaller than ℓ). This process results in an optimization problem which can be solved using the box-covering algorithm explained in Fig.2a, Materials and Methods Section III and reported in [29] as the Maximum Excluded Mass Burning algorithm (MEMB, the algorithm can be downloaded from `http://lev.ccny.cuny.edu/~hmakse/soft_data.html`). The requirement of minimal number of modules to cover the network (N_B) guarantees that the partition of the network is such that each module contains the largest possible number of nodes and links inside the module with the constraint that the modules cannot exceed size ℓ. This optimized tiling process gives rise to modules with the fewest number of links connecting to other modules implying that the degree of modularity, defined by [10–12, 30]

$$\mathcal{M}(\ell) \equiv \frac{1}{N_B} \sum_{i=1}^{N_B} \frac{L_i^{\rm in}}{L_i^{\rm out}}, \tag{1}$$

is maximized. Here $L_i^{\rm in}$ and $L_i^{\rm out}$ represent the number of links that start in a given module i and end either within or outside i, respectively. Large values of $\mathcal{M}$ ($L_i^{\rm out} \to 0$) correspond to a higher degree of modularity. The value of the modularity of the network $\mathcal{M}$ varies with ℓ, so that we can detect the dependence of modularity on different length scales, or equivalently how the modules themselves are organized into larger modules that enhance the degree of modularity.

For a given ℓ, we obtain the optimal coverage of the network with N_B modules (we use the MEMB algorithm [29] explained in Fig. 2a and Materials and Methods). Analysis of the modularity Eq. (1) in Fig. 3a reveals a monotonic increase of $\mathcal{M}(\ell)$ with a lack of a characteristic value of ℓ. Indeed, the data can be approximately fitted with a power-law functional form:

$$\mathcal{M}(\ell) \sim \ell^{d_M}, \tag{2}$$

which is detected through the modularity exponent d_M. We characterize the network using different snapshots in time and we find that $d_M \simeq 2.0$ is approximately constant over the

time evolution (Fig. 3a). This value reveals a considerable degree of modularity in the entire system (for comparison, a random network has $d_M = 0$ and a uniform lattice has $d_M = 1$ [30]), as evidenced by the network structure in Fig. 1. The lack of a characteristic length-scale in the modularity shown in Fig. 3a suggests that the modules appear at all length-scales, i.e. modules are organized within larger modules in a self-similar way, so that the inter-connections between those clusters repeat the basic modular character of the entire NHCD. Thus, the NHCD remains statistically invariant when observed at different scales. Varying the module size ℓ yields the scaling relation for the number of modules (Fig. 3b):

$$N_B(\ell) \sim \ell^{-d_B}, \tag{3}$$

where d_B represents the fractal dimension of the network [13]. We find that the fractal character of the modules is established at the early stages, yielding $d_B \simeq 1.4$ as early as 30 days (Fig. 3b). As the network evolves, the fractal dimension increases slightly and finally reaches $d_B \simeq 1.9$.

The significance of Eq. (2) is that the modules need to be interpreted at a given length-scale. Figure 2b shows an example of such hierarchical organization [Fig. 2c and SI-Fig. 6 show the full modular structure, while a list of detected modules appears in the SI Appendix]. Three types of communities of cell types are clearly identified in Fig. 2b as we change ℓ. *(i) The known functional modules:* The entire eye lineage [1, 5, 6, 14] is detected as a single module by the box-covering algorithm at $\ell = 11$, while the entire neural lineage [1–3, 5, 6, 14, 15] appears at $\ell = 15$. Finer and coarser novel modules are identified by the algorithm. *(ii) Sub-modules:* At $\ell = 11$ the neural lineage is split into the main neural and the supporting glial cell modules, while for $\ell = 7$ sub-modules are identified in the eye system. *(iii) Super-modules:* When we increase the length to $\ell = 19$, the eye and neural system form a single super-module. Thus, each cell type is connected to other types according to which groups of nodes of all sizes self-organize following a single principle. This property allows us to renormalize the network [13] by replacing each detected module by a single supernode to identify the network of modules as shown in Fig. 2c. Following the evolution and inter-dependence of these super-modules, as seen in Fig. 2c, identifies families of cell types at varying scales. This modularity map is useful in proposing future research ties between previously unrelated domains in organ functions.

The dynamics leading to such a structure can be unraveled by the study of the NHCD

as a growth process. The knowledge of the time of appearance of each cell type, T_a, makes it possible to follow the cumulative growth of the embryo and fetus in terms of the total number of cell types at time t, $N(t)$ as well as the number of cell types that eventually survive in the organism (Fig. 4a). As expected, surviving cells emerge in the later stages of the gestation period. However, the difference between the total and the surviving number of cell types indicates that generation of new types of non-surviving cells takes place even during the final gestation months.

The increase of the network size, $N(t)$, is initially approximately exponential and after $t_* = 40$ days changes into a slower growth (Fig. 4a). Only a small percentage of the nodes grow within a given time interval, so that the network activity is focused in a small number of them at a given time. The number of nodes that differentiate at a given time are shown in Fig. 4b. We observe an activity that increases monotonically up to around $t_* = 40$ days and then drops to lower values. The cross-over time $t_* = 40$ days observed in Figs. 4a and 4b separates two regimes of growth and approximately corresponds to the time below which most of the cells have a plastic characteristic (i.e., the capability to differentiate) and above which they start to become functional [1]. Interestingly, the two regimes observed in $N(t)$ merge into a single universal functional curve when we replot $N(\ell_{N1})$ as a function of the chemical distance to the fertilized egg, ℓ_{N1} (Fig. 4c). This result suggests that the topological distance in the network ℓ_{N1} is the natural variable to characterize the growth process in a universal form rather than the time. The dynamic of $N(\ell)$ follows a typical logistic (Verhulst) process of population growth where the rate of growth is restricted by environmental limitations:

$$\frac{dN}{d\ell} = rN[1 - \frac{N}{N_f}], \tag{4}$$

with solution,

$$N(\ell) = N_f \frac{\exp(r\ell)}{N_f + (\exp(r\ell) - 1)} \tag{5}$$

(see the fitting in Fig. 4c) where N_f is the final number of cell types and $r = 0.65$ is the growth rate of cell types.

Analysis of the network connectivity reveals that the average number of links per node in the final stages of the entire NHCD is $\langle k \rangle = 2.24$ (Fig. 4d). Even though $\langle k \rangle \approx 2$, there is a broad degree distribution (scale-free [12], $P(k) \sim k^{-\gamma}$, $\gamma \simeq 3.0$, SI-Fig. 7). This implies that there is always a small number of crucial cell types that differentiate much more than the

others, a fact that agrees with evidence on the existence of a few cells with large plasticity potential. As this potential is rapidly lost after 40 days, cell types change their development ability in favor of the organism life maintenance.

The fact that the average degree is close to 2 implies that the dynamical evolution of NHCD can be described by a critical branching process where every node has a certain probability of generating offsprings, in which case the critical condition for the branching to continue is $\langle k \rangle = 2$ [31]. This effectively means that each node needs to give at least one descendant in order for the network to keep growing. If $\langle k \rangle < 2$, the growth would stop early, while for $\langle k \rangle > 2$ the growth would be faster than exponential.

The network reaches the condition of criticality, $\langle k \rangle \approx 2$, at around $t_* = 40$ days (Fig. 4d) in conjunction with the transition from plasticity to functional behavior. After this, the average degree remains just above criticality to sustain a growth rate that guarantees the network survival. The majority of the nodes propagate the growth in a single line, but there are nodes which generate significantly more descendants to generate the diversity implied by the power-law distributions of degree and modularity.

II. DISCUSSION

In summary, we present the first large-scale study of the prenatal evolution of the human cell differentiation process from the fertilized egg to a developed human. The process of human cell differentiation can be mapped onto a complex network composed of cell types and differentiation steps. This mapping allows us to study the cell differentiation process with state of the art network theory for community detection with the goal of identifying hitherto unknown functional relations between cell types.

We discover a dynamical law of critical branching explaining the emergence of the network topology, which reveals a novel scale-invariant modular structure of the network of cell types. The self-similar modular features evidenced in Figs. 1, 2 and 3 are established early in the process and remain invariant during the evolution of the NHCD, although the network size changes significantly.

Using this law, we are able to observe the network at different scales. The emerging picture clearly identifies clusters of cell types, or modules, and their connectivity to other modules within its own and other functions. The resulting hierarchical organization consists

of sub-modules, known biological functions and super-modules of specialized tissues and organs emerging on varying scales. This discovery is useful in proposing future research ties between previously unrelated domains in organ functions in a systematic way. Furthermore, this information could be of importance in providing predictions of functional attributes to a number of identified modules of cell types in the NHCD.

III. MATERIALS AND METHODS

Module detection algorithm

The detection of modules or boxes in our work follows from the application of the box-covering algorithm [13, 29] at different length scales. The algorithm can be downloaded at `http://lev.ccny.cuny.edu/~hmakse/soft_data.html`. In box covering we assign every node to a module, by finding the minimum possible number of boxes, $N_B(\ell)$, that cover the network and whose diameter (defined as the maximum distance between any two nodes in this box) is smaller than ℓ. These boxes are characterized by the proximity between all their nodes, at a given length scale. Different values of the box diameter ℓ yield boxes of different size. These boxes are identified as modules which at a smaller scale ℓ may be separated, but merge into larger entities as we increase ℓ.

In this work we implement the Maximum Excluded Mass Burning (MEMB) algorithm from [29] for box covering. The algorithm uses the basic idea of box optimization, where we require that each box should cover the maximum possible number of nodes, and works as follows: For a given ℓ, we first locate the optimal 'central' nodes which will act as the origins for the boxes. This is done by first calculating the number of nodes (called the mass) within a diameter ℓ from each node. The node that yields the largest mass is marked as a center. Then we mark all the nodes in the box of this center node as 'tagged'. We repeat the process of calculating the mass of the boxes starting from all non-center nodes, and we identify a second center according to the largest remaining mass, while nodes in the corresponding box are 'tagged', and so on. When all nodes are either centers or 'tagged' we have identified the minimum number of centers that can cover the network at the given ℓ value. Starting from these centers as box origins, we then simultaneously burn the boxes from each origin until the entire network is covered, i.e. each node is assigned to one box (we call this process burning since it is similar to burning algorithms developed to investigate clustering statistics

in percolation theory [12]). In Fig. 2a of the main text we show how box-covering works for a simple network at two different ℓ values.

This algorithm is driven by the proximity between nodes and the maximization of the mass associated with each module center [13, 29]. Thus it detects boxes that maximize modularity, Eq. (1). In the case of MEMB we have the additional benefit of detecting modules at different scales, so that we can study the hierarchical character of modularity, i.e. modules of modules, and we can detect whether modularity is a feature of the network that remains scale-invariant.

The fractal dimension d_B of a complex network is an exponent that determines how the mass (equivalently: the number of nodes) around any given node scales with the length, which in networks corresponds to the shortest distance between two nodes. In order to numerically measure this exponent we optimally cover the network with boxes using the MEMB algorithm. A box is a set of nodes where all distances ℓ_{ij} between any two nodes i and j in this set are smaller than a given value of ℓ, the box size. Although there is a large number of coverings, for every value of ℓ we want to find the one which gives the smallest possible number of boxes, $N_B(\ell)$. Varying ℓ then yields the scaling relation Eq. (3). A finite fractal dimension reveals fundamental organizational principles of the underlying network, namely a self-similar structural character, where the network is built in a similar way even though we observe it at different length-scales. The boxes that are identified through this process correspond to the modules at varying scales.

[1] Sell, S (2004) *Stem Cells Handbook* (Humana Press, Totowa, NJ).

[2] Kirschstein R, Skirboll, LR (2001) *Stem Cells: Scientific Progress and Future Research Directions* (NIH, Bethesda).

[3] Freitas RA Jr (1999) *Nanomedicine, Volume I: Basic Capabilities* (Landes Bioscience, Georgetown, Texas).

[4] Alberts B *et al.* (2002) *Molecular Biology of the Cell* (Fourth ed., Garland Science, New York).

[5] Sadler, TW (2004) *Langman's Medical Embryology*, (9nd ed., Lippincott Williams and Wilkins, Baltimore).

[6] Vickaryous, MK, Hall, BK (2006) *Biol. Rev.* **81:**425–455.

[7] Valentine, JW (2003) in *Keywords and Concepts in Evolutionary Developmental Biology* (eds. B. K. Hall and W. M. Olson), pp. 35-53. (Harvard University Press, Cambridge).

[8] Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E (2005) *PLoS Comput Biol* **1**(5): e50.

[9] Ravasz E, Somera AL, Mongru, DA, Oltvai ZN, Barabási AL (2002) *Science* **297:**1551–1555.

[10] Newman MEJ, Girvan M (2004) *Phys. Rev. E* **69:**026113.

[11] Guimerà R, Amaral LAN (2005) *Nature* **433:**895–900.

[12] Caldarelli G, Vespignani A (eds) (2007) *Large scale structure and dynamics of complex networks.* (World Scientific, Singapore).

[13] Song C, Havlin S, Makse HA (2005) *Nature* **433:**392–395.

[14] Paxinos G, Mai JK (2004) *The Human Nervous System* (2nd ed., Elsevier Academic Press).

[15] Temple S (2001) *Nature* **414:**112–117.

[16] Bianco P, Riminucci M, Gronthos S, Robey PG (2001) *Stem Cells* **19:**180–192.

[17] Chen JCJ, Goldhamer DJ (2003) *Reprod. Biol. Endocrinol.* **1:**101.

[18] Janeway CA, Travers P, Walport M, Shlomchik M (2001) *Immunobiology: the immune system in health and disease* (Fifth Edition, Garland Science).

[19] Anglani F, Forino M, Del Prete D, Tosetto E, Torregrossa R, D'Angelo A (2004) *J. Cell. Mol. Med.* **8:**474–487.

[20] Horster MF, Braun GS, Huber SM (1999) *Physiol. Rev.* **79:**1157–1191.

[21] Herrick SE, Mutsaers SE (2004) *Int. J. Biochem. Cell B.* **36:**621–642.

[22] Otto WR (2002) *J. Pathol.* **197:**527–535.

[23] Jessen KR, Mirsky R (2005) *Nat. Rev. Neurosci.* **6:**671–682.

[24] Nakashima M, Redid AH (2003) *Nat. Biotechnol.* **21:**1025–1032.

[25] Santagati F, Rijli FM (2003) *Nat. Rev. Neurosci.* **4:**806–818.

[26] Savage JJ, Yaden BC, Kiratipranon P, Rhodes SJ (2003) *Gene* **319:**1–19.

[27] Forge A, Wright T (2002) *Brit. Med. Bull.* **63:**5–24.

[28] Panteleyev A, Jahoda CAB, Christiano AM (2001) *J. Cell. Sci.* **114:**3419–3431.

[29] Song C, Gallos LK, Havlin S, Makse HA (2007) *J. Stat. Mech: Theory and Experiments* P03006.

[30] Gallos LK, Song C, Havlin S, Makse HA (2007) *Proc. Nat. Acad. Sci. USA*, **104:**7746–7751.

[31] Athreya KB, Ney PE (2004) *Branching processes* (Dover Publications).

Acknowledgments: We thank B. Brujić, B. Dubin-Thaler, H.D. Rozenfeld, S. Havlin, T. Rattei, M. Sigman and A. M. Andrade and J. M. M. Andrade for valuable discussions. This work was supported by National Science Foundation Grants SES-0624116 and EF-0827508. We thank the Brazilian agencies CNPq, CAPES, FAPESB and FUNCAP, and the EU FP7 Neuronano project (NMP4-SL-2008-214547) for financial support

Fig. 1. Complex network representation of the human cell differentiation process. The first steps of the NHCD construction are shown in the inset of this figure. These steps, known to also be present in the formation of the majority of multicellular organisms, include the first cleavage of a fertilized egg, which is subsequently followed by the ball stage and the formation of primary germ cell layers, namely, the ectoderm, mesoderm, and endoderm. The fertilized egg is a totipotent stem cell. The blastocyst, in turn, gives rise to both trophoblast and inner cell mass. These two cells further differentiate into other types of cells, and so on. Following the above process until the fetus is fully developed yields the complex network shown in this figure. Each node, plotted as a circle, corresponds to a cell type and the edges to a differentiation step. The entire network originates from the fertilized egg (denoted by a red square) and leads to the specialized cells of a developed human. Filled circles correspond to nodes that survive at the end of the development process, while empty circles correspond to non-surviving cell types. Nodes in communities of known functions from the literature are indicated by different colors, except for those cell types with no functional annotation (see SI-Table I for association to the known functions $C1$ to $C19$ extracted from the literature).

Fig. 2. Detection of modules and the network of modules at different scales. **a,** Demonstration of the box-covering algorithm for a schematic network, following the Maximum Excluded Mass Burning algorithm in [13, 29] (see SI-Section III for full details). We cover the network with the smallest possible number of boxes for a given ℓ value. This is done in a two-stage process: *(i)* We detect the smallest possible number of box origins (shown with cyan color) that provide the maximum number of nodes (mass) in each box, according to the following optimization algorithm: We calculate the mass associated with each node, and pick the first center as the node with largest mass and mark the nodes in this box as 'tagged'. We repeat the process from the remaining non-center nodes to identify a second center with the highest mass, and so on. *(ii)* We build the boxes through simultaneous burning from these center nodes, until the entire network is covered with boxes. For example, at $\ell = 3$ there are four boxes, where the maximum distance between any two nodes in a box is smaller than ℓ. Similarly, we can cover the same network with two boxes at $\ell = 6$. These two boxes are the result of merging two of the four boxes at $\ell = 3$. **b,** Detail of NHCD modules detected by the above box-covering algorithm for two particular functions. The algorithm detects a hierarchy of sub-modules, known functions and super-

modules of size ℓ plotted in different colors. We show the identified modules corresponding to $C12$-neural system and $C13$-eye system (full structure is in Fig. 2c and SI-Fig. 6), which first appear at $\ell = 15$ and $\ell = 11$, respectively. At other scales the box-covering algorithm detects new functional relations between cell types expressed in the obtained sub and super-modules. For instance, at $\ell = 11$ the neural lineage is further divided into two sub-modules, while at $\ell = 19$ the two functions merge into a super-module. **c,** The network of modules at different ℓ values, as detected through the box-covering algorithm. Every node corresponds to one of the three following types, in terms of increasing scale: *(i)* Sub-modules (small grey dots), which are fractions of a fully functional module, *(ii)* Known functional biological modules (colored circles), whose color corresponds to the functions $C1$-$C19$, and *(iii)* Super-modules (pie-charts), which represent the union of more than one known functional module, described by the colors of the pie-chart. The links that stem from known functional modules and super-modules are shown in red, and they progressively span the entire network as we increase ℓ.

Fig. 3. Modular properties of the NHCD. a, Degree of modularity of the network, $\mathcal{M}(\ell)$ at different times T_a (indicated in the figure) as a function of the scale of observation, ℓ. **b,** Number of boxes/modules, N_B, versus the size of the modules ℓ identified by the box-covering algorithm for different networks at time T_a.

Fig. 4. Growth properties of the NHCD. a, Number of cell types in the network, $N(t)$, as a function of time. We find precise information about the appearance time T_a for 782 among the 873 cell types. Those cells with missing appearance time have not been taken into account in this plot. Also shown are the time evolution of the number of surviving and non-surviving cells. **b,** Number of nodes whose degree increases at time t (red histogram) and number of new links appearing in the network (blue histogram) as a function of time. If all nodes were giving just one child then the two histograms would coincide. Inset: The average number of new links per node at a given time can be found by dividing the two histograms in the main plot. This plot shows how intense is the activity at that particular time. Despite the variation in activity, the new connections average around 1, which gives a critical branching ratio of $\langle k \rangle \simeq 2$. **c,** Number of cell types versus the chemical distance to the first node, ℓ_{N1}. This distance is only determined by the connections between the cell types, and is not influenced by the appearance time, so that we include all 873 cell types. **d,** Average degree $\langle k \rangle$ of the network as a function of time showing that the network achieves

the condition of critical branching process $\langle k \rangle \approx 2$ at around $t_* = 40$ days.

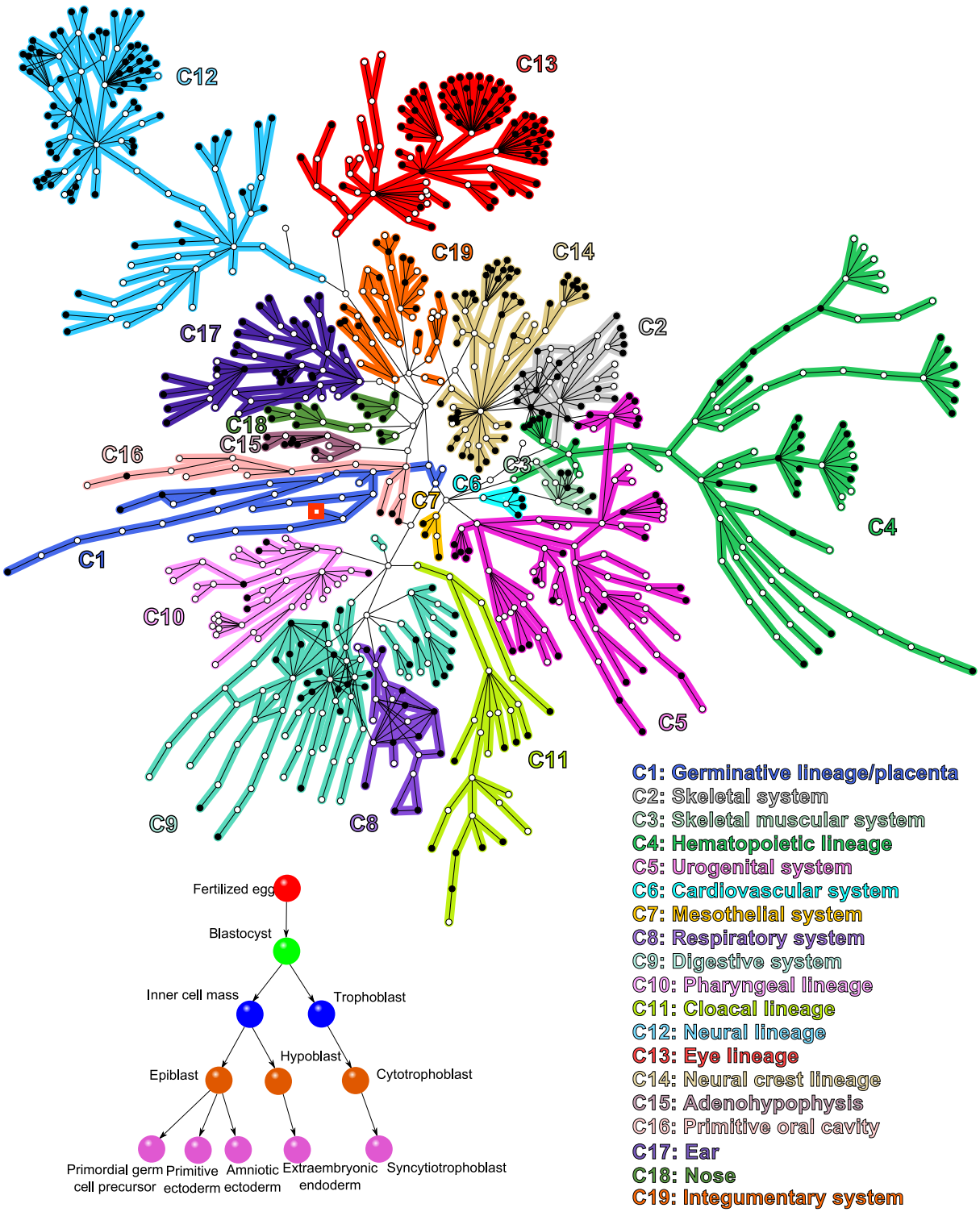

C12
C13
C19
C14
C17
C2
C18
C15
C16
C3
C6
C7
C4
C1
C10
C5
C11
C9
C8
Fertilized egg
Blastocyst
Inner cell mass
Trophoblast
Hypoblast
Epiblast
Cytotrophoblast
Primordial germ cell precursor
Primitive ectoderm
Amniotic ectoderm
Extraembryonic endoderm
Syncytiotrophoblast
C1: Germinative lineage/placenta
C2: Skeletal system
C3: Skeletal muscular system
C4: Hematopoietic lineage
C5: Urogenital system
C6: Cardiovascular system
C7: Mesothelial system
C8: Respiratory system
C9: Digestive system
C10: Pharyngeal lineage
C11: Cloacal lineage
C12: Neural lineage
C13: Eye lineage
C14: Neural crest lineage
C15: Adenohypophysis
C16: Primitive oral cavity
C17: Ear
C18: Nose
C19: Integumentary system

FIG. 1:

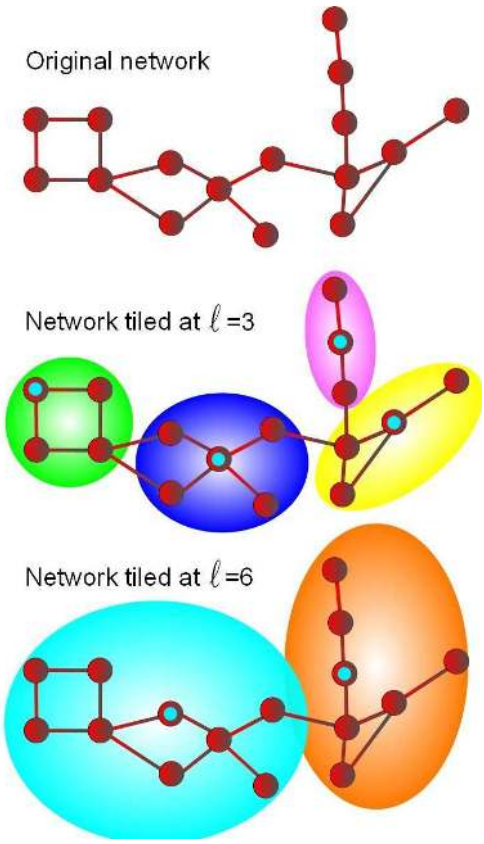

Original network
Network tiled at ℓ =3
Network tiled at ℓ =6

Fig. 2a

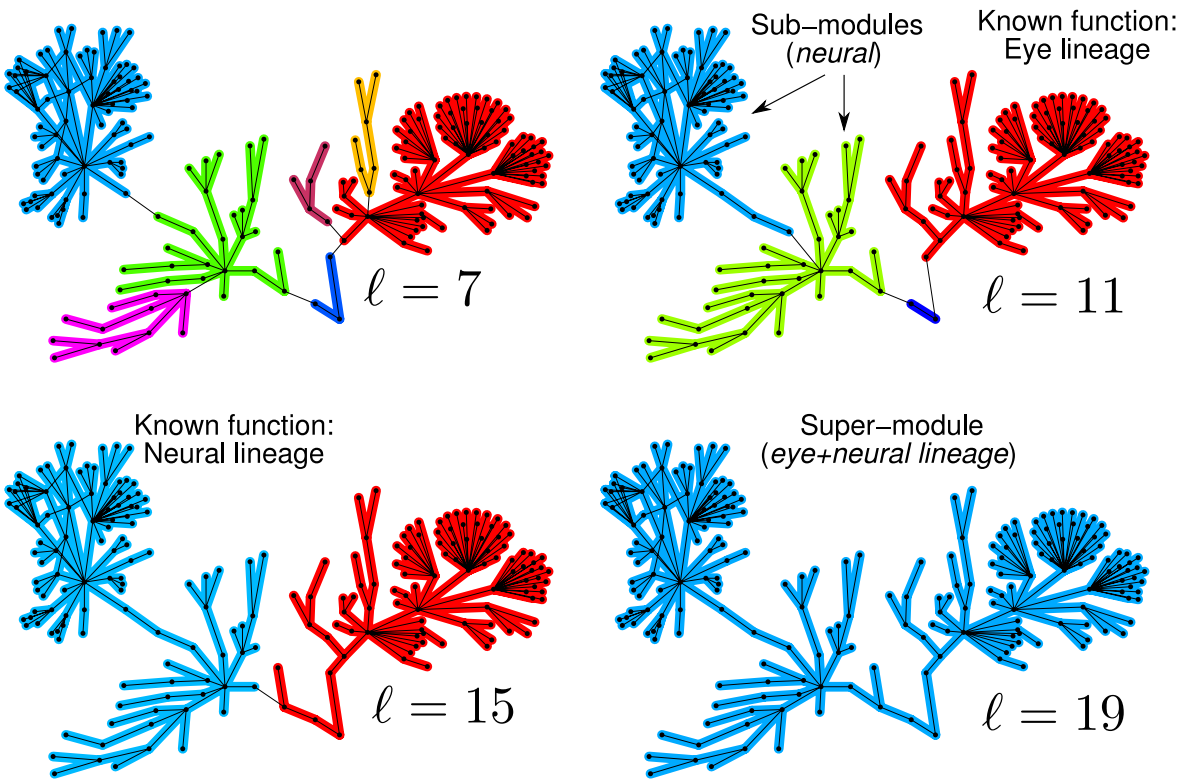

$\ell = 7$
Sub–modules
(*neural*)
Known function:
Eye lineage
$\ell = 11$
Known function:
Neural lineage
$\ell = 15$
Super–module
(*eye+neural lineage*)
$\ell = 19$

Fig. 2b

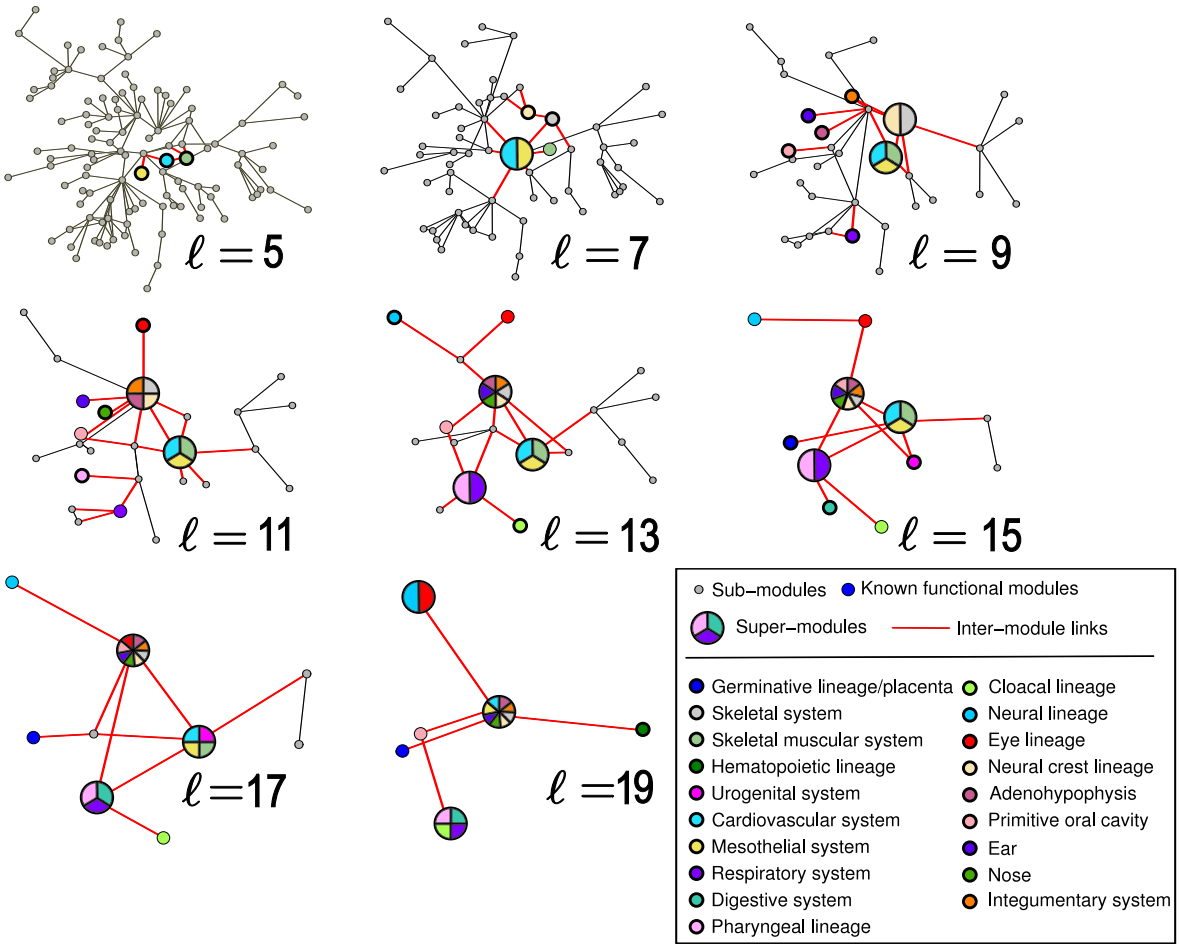

ℓ = 5
ℓ = 7
ℓ = 9
ℓ = 11
ℓ = 13
ℓ = 15
ℓ =17
ℓ =19
Sub−modules
Known functional modules
Super−modules
Inter−module links
Germinative lineage/placenta
Skeletal system
Skeletal muscular system
Hematopoietic lineage
Urogenital system
Cardiovascular system
Mesothelial system
Respiratory system
Digestive system
Pharyngeal lineage
Cloacal lineage
Neural lineage
Eye lineage
Neural crest lineage
Adenohypophysis
Primitive oral cavity
Ear
Nose
Integumentary system

Fig. 2c

FIG. 2:

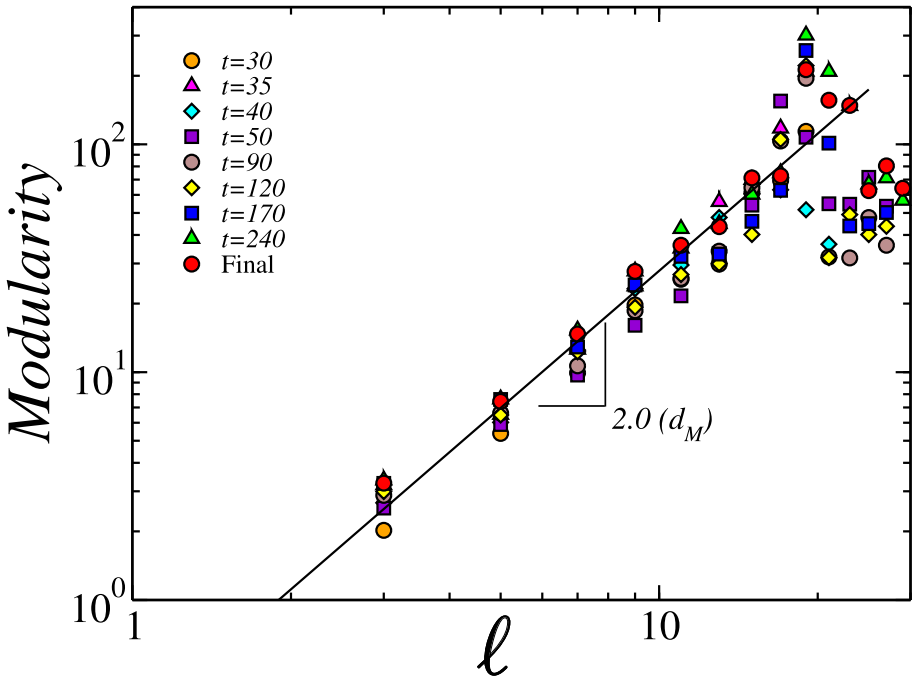


Fig. 3a

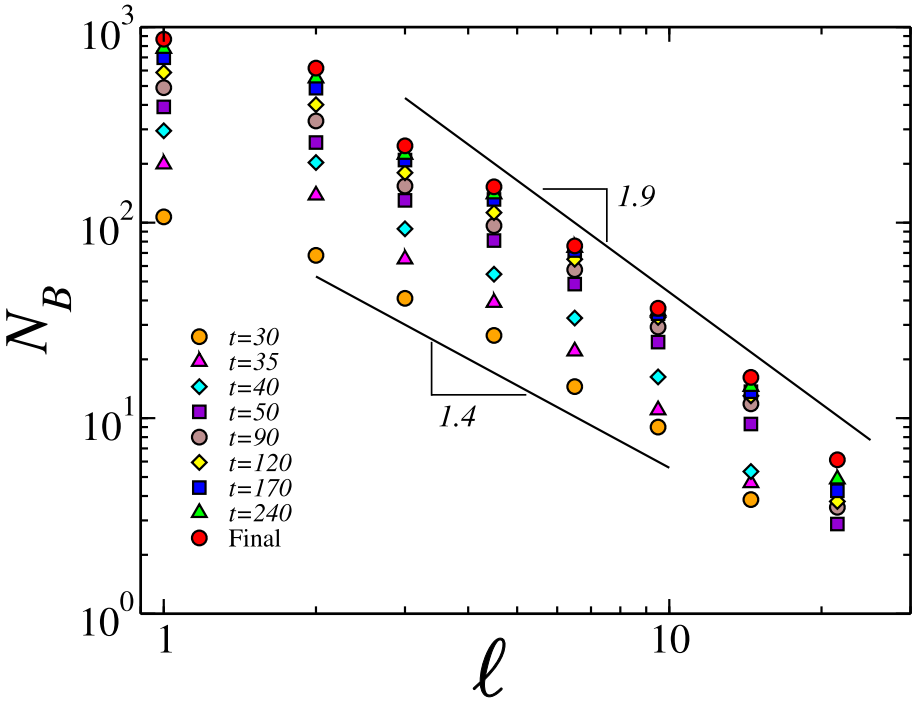


Fig. 3b

FIG. 3:

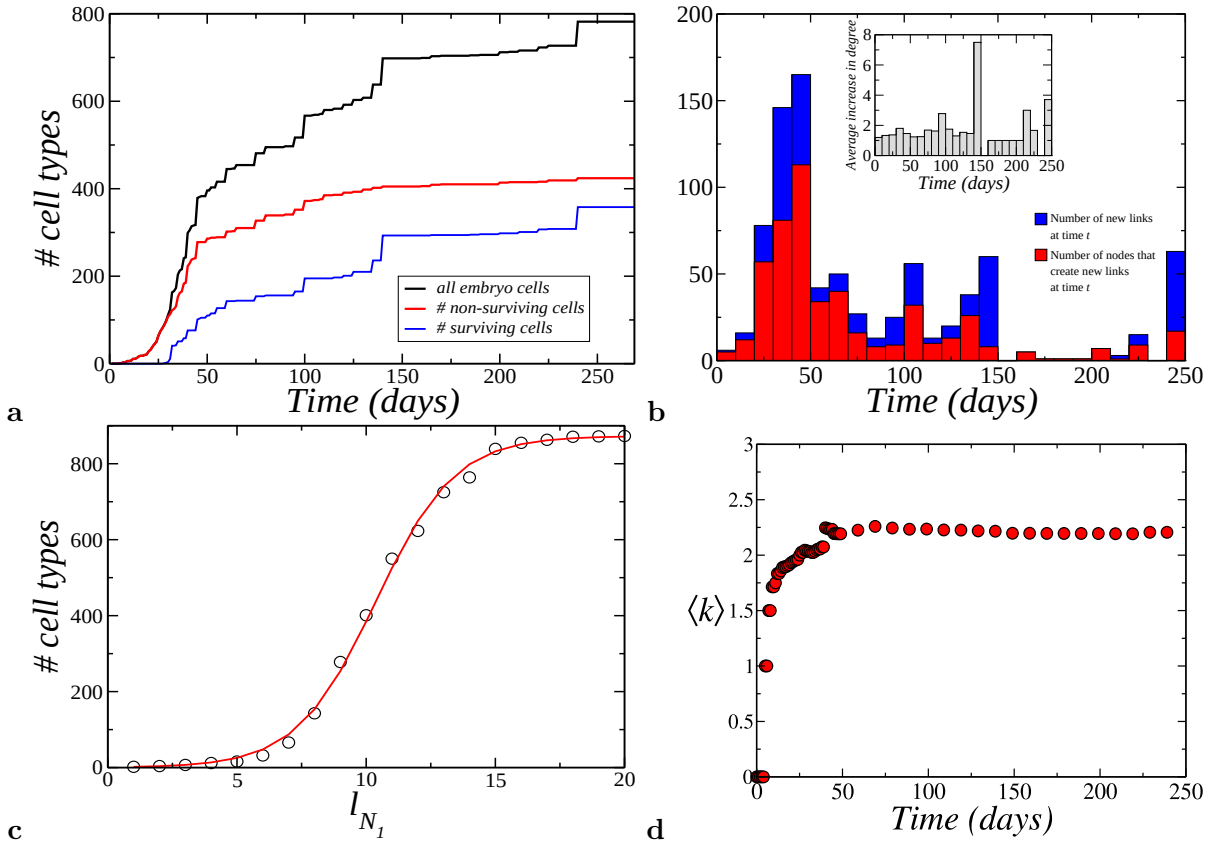

a
cell types
Time (days)
all embryo cells
non-surviving cells
surviving cells
b
Average increase in degree
Time (days)
Number of new links at time t
Number of nodes that create new links at time t
Time (days)
c
cell types
l_{N_1}
d
$\langle k \rangle$
Time (days)

FIG. 4:

SUPPLEMENTARY INFORMATION

Modularity map of the network of human cell differentiation

IV. ADDITIONAL INFORMATION FOR NHCD

FIGURE 5 provides an alternative representation of the NHCD. Different cell types appearing as bifurcation points of the network are stacked along the vertical axis. In SI-Fig. 5a the horizontal axis corresponds to the shortest path ℓ_{N1} calculated from node $N1$, the fertilized egg, to any given node, while in SI-Fig. 5b the same network is shown as a function of the node appearance time T_a. The links emerging from each cell type follow a bifurcation pattern ending up at the right side with $k-1$ branches, each one of them representing one of the more specialized cell types. Red nodes correspond to the surviving cell types, and they preferentially appear at later times, while the non-surviving cell types, the blue nodes, emerge during the early stages of the process. The color of the edges corresponds to one of the 19 functional groups identified in Table I as given in Fig. 1. Links that generate loops are plotted in red.

Figure 5b contains the same information as Fig. 5a but we plot each cell type according to its time of appearance rather than as a function of the chemical distance to $N1$, as in Fig. 5a. The white and yellow alternating vertical stripes divide the time axis in intervals in days. The branches have been extended so that each cell appears only in the corresponding interval. Colors and labels are the same as in Fig. 5a.

The catalog presented in Ref. [6] reports 407 distinct cell types in a healthy adult human body, all of which can be identified in our network representation. Most of them occupy the end points of the 529 branches in Figs. 1 and 5. Therefore, not all tree leaves (branch endpoints) correspond to cell types in born humans.

The average shortest path calculated from all cell types to $N1$ is $\langle \ell_{N1} \rangle = 10.93$, as expected from the large concentration of links in the interval $8 \leq \ell \leq 13$ (Fig. 5a).

FIGURE 6 shows the full modular structure of the NHCD as detected by the box-covering algorithm at different length scales. A detail of this process is represented in Fig. 2 in the main text. A list containing the nodes belonging to each module at a given ℓ is

contained in the file modules.txt.

FIGURE 7 shows the degree distribution of the NHCD for different times.

Table I lists the different known functional modules of the NHCD and the respective citations to the literature. The complete collected data is listed in the datafile: links.txt. This file includes the links between the cell types, their time of appearance in days after fecundation (T_a), and the known functional module they belong to. The references to the publications reporting each link appear in the table. Data on the structure of individual communities were obtained from the specialized literature. Some functional communities are: germ layer and extraembryonic tissue [1, 2, 4, 5], digestive system [1, 5], pharyngeal system [3, 5, 6], cloacal system [5], neural system [1, 6], eye [5, 6], primitive oral cavity [5], nose [5], skeletal system [3, 16], skeletal muscular system [1, 6, 17], hematopoietic system [18], urogenital system [19, 20], cardiovascular system [5, 6], mesothelium [21], respiratory [22], neural crest [5, 23–25], adenohypophysis [26], ear [6, 27], and integumentary system [5, 6, 28].

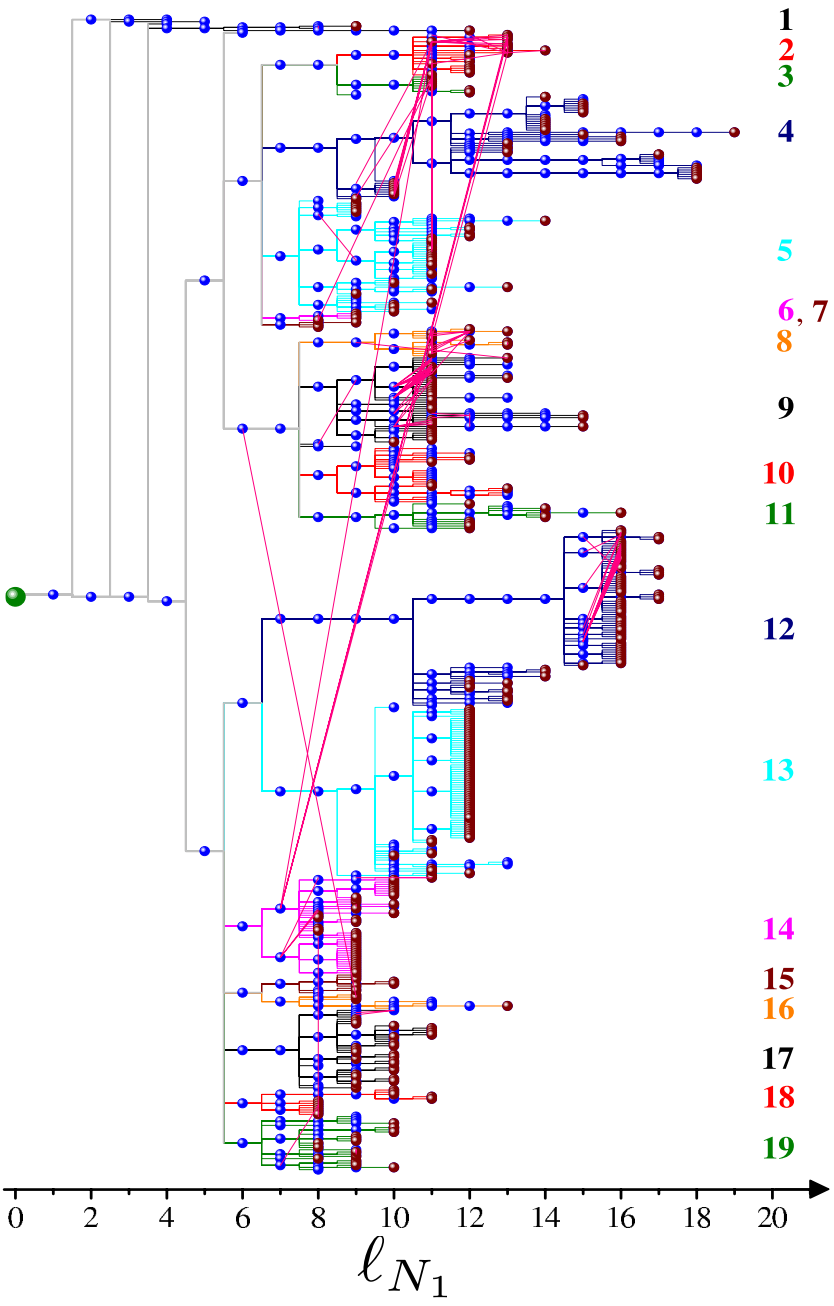

1
2
3
4
5
6, 7
8
9
10
11
12
13
14
15
16
17
18
19
0
2
4
6
8
10
12
14
16
18
20
ℓ_{N_1}

Fig. 5a

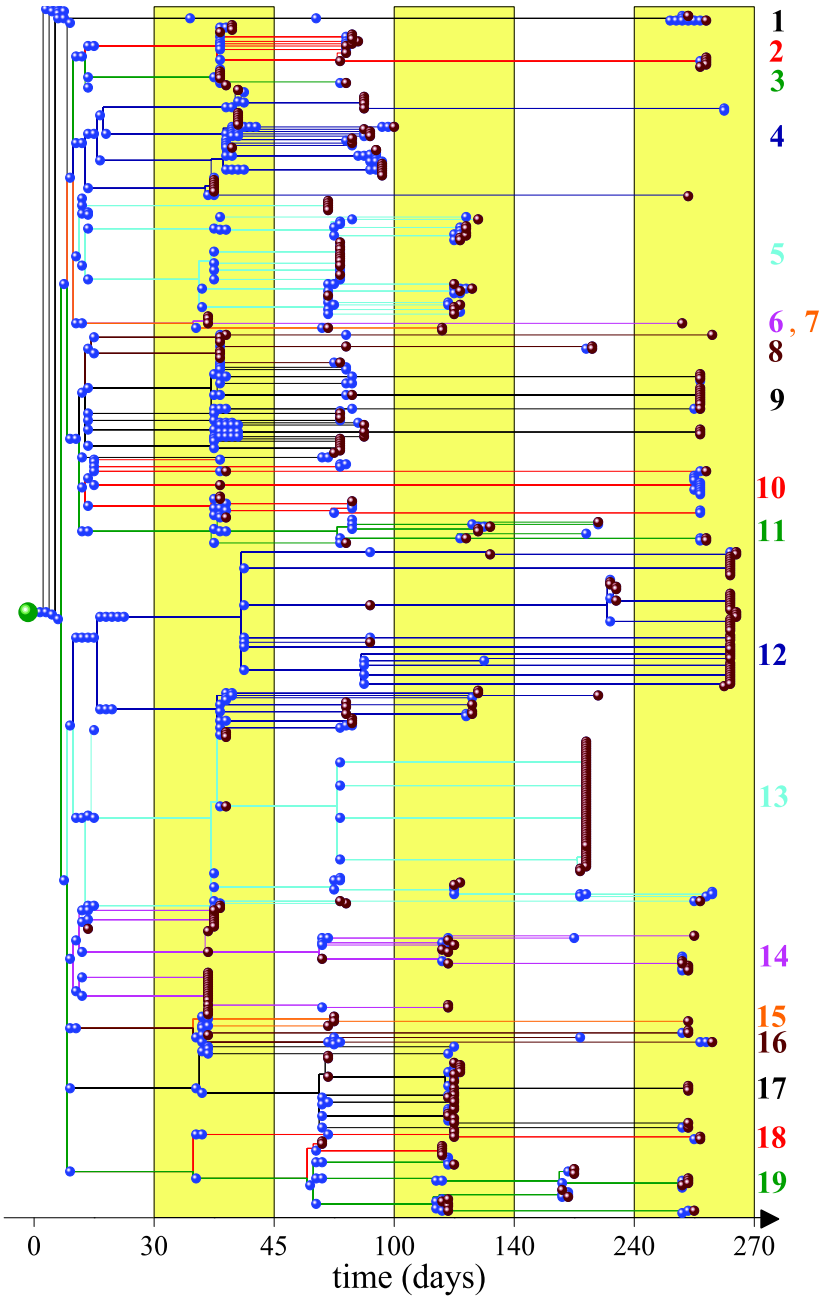
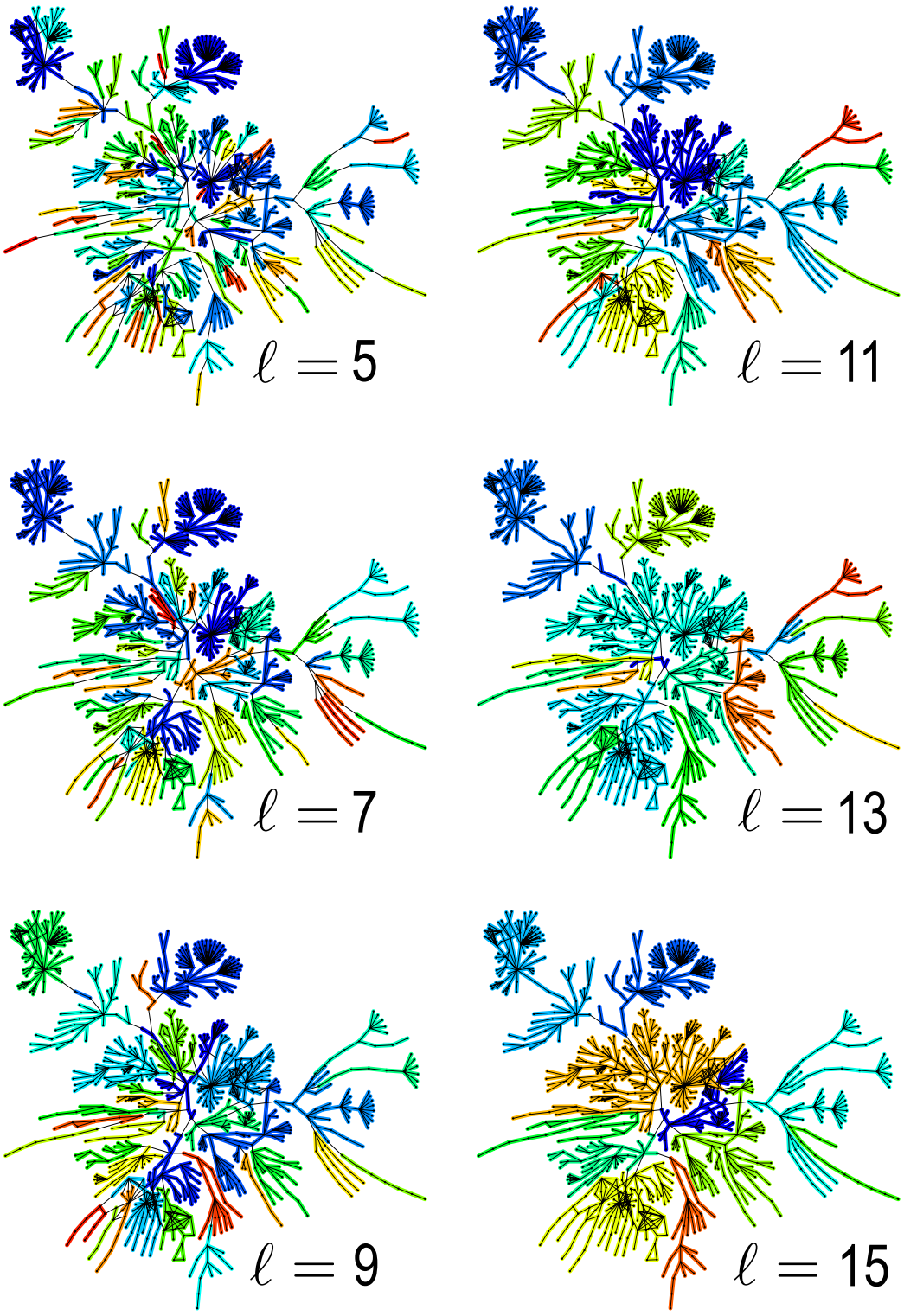


Fig. 5b

FIG. 5: **Alternative representation of the NHCD.** (**a**) The horizontal axis measures the shortest path from each cell type to the fertilized egg along the network, ℓ_{N_1}. (**b**) The horizontal axis denotes the appearance time of a given cell type, T_a. For simplicity we do not plot the links leading to loops. The colors of the branches denote the functional classes as in Fig. 1.


$\ell = 5$
$\ell = 11$
$\ell = 7$
$\ell = 13$
$\ell = 9$
$\ell = 15$

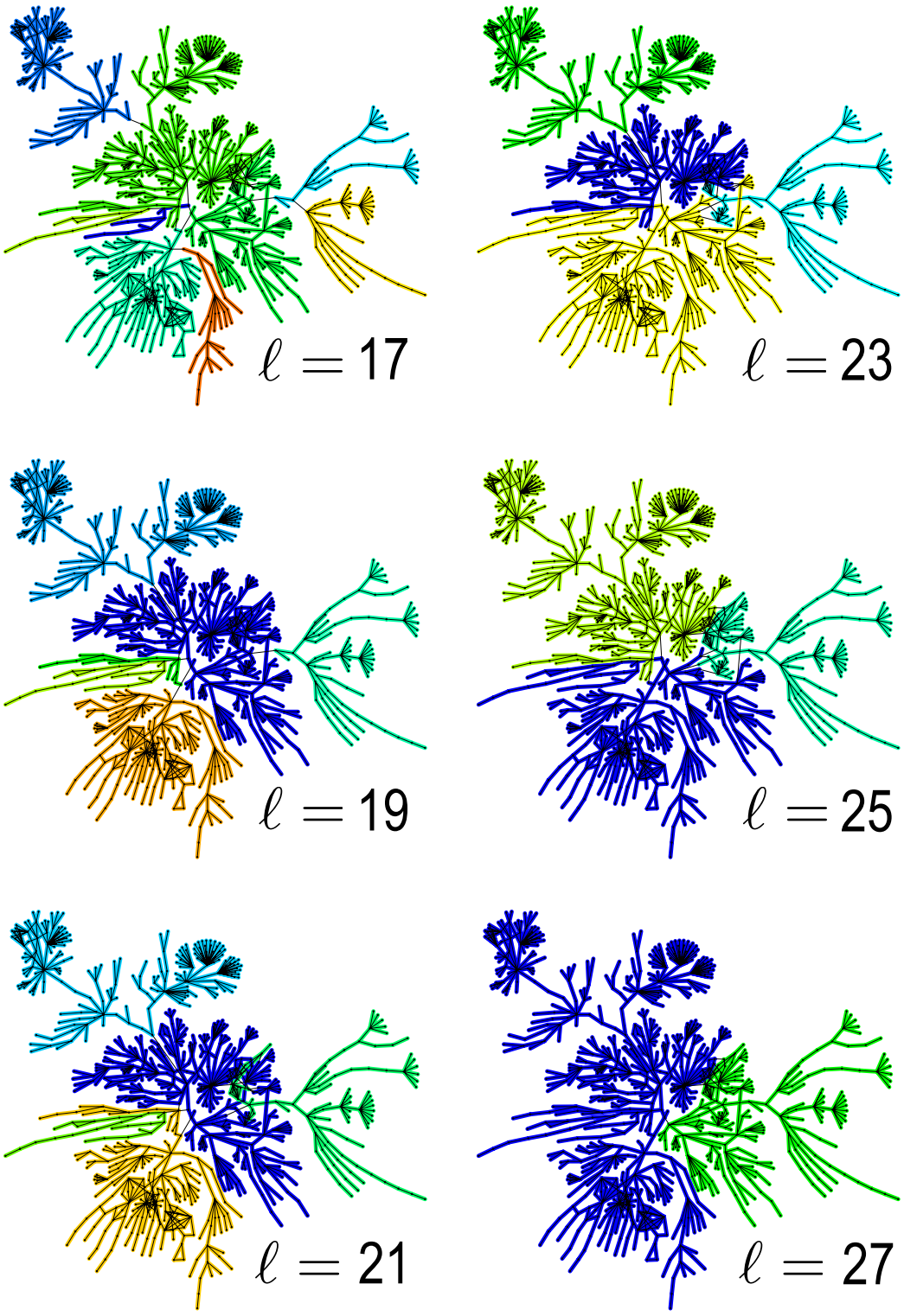


FIG. 6: **Full modular structure of the NHCD at the indicated length ℓ, as detected by the box-covering algorithm.** Each node is depicted with a different color indicating the module to which it belongs to.

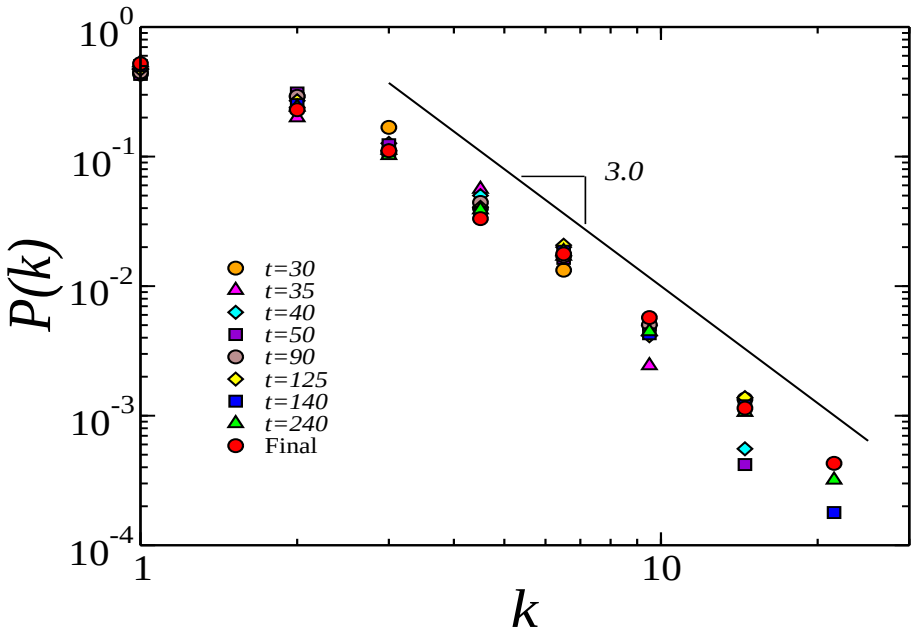


FIG. 7: Degree distribution $P(k)$ for the NHCD at different growth stages, from $t = 30$ days to $t = 240$ days.

TABLE I: Identification of the distinct biological functions of the cell types indicated in Fig. 1 and SI-Fig. 5. The third column lists the references used for building the NHCD.

Label	Biological function	Reference
1	Germinative Lineage and Placenta	Alberts et al., 2002; Kirschstein and Skirboll, 2001; Sadler, 2004; Sell, 2004
2	Skeletal System	Bianco et al., 2001; Freitas, 1999; Mochida, 2005; Sadler, 2004; Sell, 2004; Sanders et al., 1999; Towler and Gelberman, 2006; Vickaryous and Hall, 2006
3	Skeletal Muscular System	Chen and Goldhamer, 2002; Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
4	Hematopoietic Lineage	Alberts et al., 2002; Kirschstein and Skirboll, 2001; Janeway et al., 2001; Minasi et al., 2002; Paxinos and Mai, 2004; Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
5	Urogenital System	Anglani et al., 2004; Coulter, 2004; Horster et al., 1999; Lopez et al., 2001; Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
6	Cardiovascular System	Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
7	Mesothelial Lineage	Herrick and Mutsaers, 2004; Sadler, 2004
8	Respiratory System	Freitas, 1999; Otto, 2002; Sadler, 2004; Sell, 2004
9	Digestive System	Bardeesy and DePinho, 2002; Fausto, 2004; Freitas, 1999; Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
10	Pharyngeal Lineage	Blackburn and Manley, 2004; Freitas, 1999; Sadler, 2004; Vickaryous and Hall, 2006
11	Cloacal Lineage	Foster et al., 2002; Freitas, 1999; Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
12	Neural Lineage	Freitas, 1999; Kirschstein and Skirboll, 2001; Paxinos and Mai, 2004; Sadler, 2004; Sell, 2004; Temple, 2001; Vickaryous and Hall, 2006
13	Eye Lineage	Paxinos and Mai, 2004; Sadler, 2004; Sell, 2004; Vickaryous and Hall, 2006
14	Neural Crest Lineage	Jessen and Mirsky, 2005; Nakashima and Redid, 2003; Sadler, 2004; Santagati and Rijli, 2003; Sell, Szeder et al., 2003; Vickaryous and Hall, 2006
15	Adenohypophysis	Paxinos and Mai, 2004; Sadler, 2004; Savage et al., 2003; Vickaryous and Hall, 2006

16	Primitive Oral Cavity	Freitas, 1999; Nakashima and Redid, 2003; Sadler, 2004; Vickaryous and Hall, 2006
17	Ear	Forge and Wright, 2002; Freitas, 1999; Paxinos and Mai, 2004; Sadler, 2004; Vickaryous and Hall, 2006
18	Nose	Freitas, 1999; Sadler, 2004; Vickaryous and Hall, 2006
19	Integumentary System	Freitas, 1999; Hennighausen and Robinson, 2005; Panteleyev et al., 2001; Potten and Booth, 2002; Sadler, 2004; Stoeckelhuber et al., 2003; Vickaryous and Hall, 2006

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. *Molecular Biology of the Cell* (Fourth ed., Garland Science, New York, 2002).

2. Anglani, F., Forino, M., Del Prete, D., Tosetto, E., Torregrossa, R., & D'Angelo, A. In search of adult renal stem cells. *J. Cell. Mol. Med.* **8**, 474-487 (2004).

3. Bardeesy, N., & DePinho, R.A. Pancreatic cancer biology and genetics. *Nat. Rev. Cancer* **2**, 897-909 (2002).

4. Bianco, P., Riminucci, M., Gronthos, S., & Robey, P.G. Bone marrow stromal stem cells: nature, biology, and potential applications. *Stem Cells* **19**, 180-192 (2001).

5. Blackburn, C.C., & Manley, N.R. Developing a new paradigm for thymus organogenesis. *Nat. Rev. Immunol.* **4**, 278-289 (2004).

6. Chen, J.C.J., & Goldhamer, D.J. Skeletal muscle stem cells. *Reprod. Biol. Endocrinol.* **1**, 101 (2003).

7. Coulter, C.L. Functional biology of the primate fetal adrenal gland: advances in technology provide new insight. *Clin. Exp. Pharmacol.* **31**, 475-484 (2004).

8. Fausto, N. Liver regeneration and repair: hepatocytes, progenitor cells, and stem cells. *Hepatology* **39**, 1477-1487 (2004).

9. Freitas, R.A., Jr. *Nanomedicine, Volume I: Basic Capabilities* (Landes Bioscience, Georgetown, Texas, 1999).

10. Forge, A. & Wright, T. The molecular architecture of the inner ear. *Brit. Med. Bull.* **63**, 5-24 (2002).

11. Foster, C.S., Dodson, A., Karavana, V., Smith, P.H., & Ke, Y. Prostatic stem cells. *J. Pathol.* **197**, 551-565 (2002).

12. Hennighausen, L. & Robinson, G.W. Information networks in the mammary gland. *Nat. Rev. Mol. Cell Bio.* **6**, 715-725 (2005).

13. Herrick, S.E. & Mutsaers, S.E. Mesothelial progenitor cells and their potential in tissue engineering. *Int. J. Biochem. Cell B. 36*, 621-642 (2004).

14. Horster, M.F., Braun, G.S., & Huber, S.M. Embryonic renal epithelial: induction, nephrogenesis, and cell differentiation. *Physiol. Rev.* **79**, 1157-1191 (1999).

15. Kirschstein, R. & Skirboll, L. R. *Stem Cells: Scientific Progress and Future Research Directions* (NIH, Bethesda, 2001).

16. Janeway, C.A., Travers, P., Walport, M., & Shlomchik, M. *Immunobiology: the immune system in health and disease* (fifth ed., Garland Science, 2001).

17. Jessen, K.R. & Mirsky, R. The origin and development of glial cells in peripheral nerves. *Nat. Rev. Neurosci.* **6**, 671- 682 (2005).

18. Lopez, M.L.S.S, Pentz, E.S., Robert, B., Abrahamson, D.R., & Gomez, R.A. Embryonic origin and lineage of juxtaglomerular cells. *Am. J. Physiol. Renal Physiol.* **281**, 345-356 (2001).

19. Minasi MG, Riminucci M, De Angelis L, Borello U, Berarducci B, Innocenzi A, Caprioli A, Sirabella D, Baiocchi M, De Maria R, Boratto R, Jaffredo T, Broccoli V, Bianco P, & Cossu G. The meso-angioblast: a multipotent, self-renewing cell that originates from the dorsal aorta and differentiates into most mesodermal tissues. *Development* **129**, 2773-2783 (2002).

20. Mochida, J. New strategies for disc repair: novel preclinical trials. *J. Orthop. Sci.* **10**, 112-118 (2005).

21. Nakashima, M. & Redid, A.H. The application of bone morphogenetic proteins to dental tissue engineering. *Nat. Biotechnol.* **21**, 1025-1032 (2003).

22. Otto, W.R. Lung epithelial stem cells. *J. Pathol.* **197**, 527-535 (2002).

23. Panteleyev, A., Jahoda, C.A.B., & Christiano, A.M., Hair follicle predetermination. *J. Cell. Sci.* **114**, 3419-3431 (2001).

24. Paxinos, G. & Mai J.K. *The Human Nervous System* (second ed., Elsevier Academic Press, 2004).

25. Potten, C.S. & Booth, C. Keratinocyte Stem Cells: a Commentary. *J. Invest. Detmatol. 119*, 888-899 (2002).

26. Sadler, T.W. *Langman's Medical Embryology* (ninth ed., Lippincott Williams & Wilkins, Baltimore, 2004).

27. Santagati, F. & Rijli, F.M. Cranial neural crest and the building of the vertebrate head. *Nat. Rev. Neurosci.* **4**, 806-818 (2003).

28. Savage, J.J., Yaden, B.C. Kiratipranon, P., & Rhodes, S.J. Transcriptional control during mammalian anterior pituitary development. *Gene* **319**, 1-19 (2003).

29. Sell, S. *Stem Cells Handbook* (Humana Press, Totowa, NJ, 2004).

30. Stoeckelhuber, M., Stoeckelhuber, B.M., & Welsch, U. Human Glands of Moll: Histochemical and Ultrastructural Characterization of the Glands of Moll in the Human Eyelid. *J. Invest. Dermatol.* **121**, 28-36 (2003).

31. Szeder, V., Grim, M., Halata, Z., & Sieber-Bluma, M. Neural crest origin of mam-

malian Merkel cells. *Dev. Biol.* **253**, 258-263 (2003).

32. Temple, S. The development of neural stem cells. *Nature*, **414**, 112-117 (2001).

33. Towler, D.A. & Gelberman, R.H. The alchemy of tendon repair: a primer for the (S)mad scientist. *J. Clin. Invest.* **116**, 863-866 (2006).

34. Vickaryous, M.K. & Hall, B.K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev.* **81**, 425-455 (2006).