Information topology of gene expression profile in dopaminergic neurons

Mónica TAPIA PACHECO[1,§], Pierre BAUDOT[1,§], Martial A. DUFOUR[1,2], Christine FORMISANO-TRÉZINY[1], Simone TEMPORAL[1], Manon LASSERRE[1], Jean GABERT[1,3], Kazuto KOBAYASHI[4] and Jean-Marc GOAILLARD[1,5]

[1] Unité de Neurobiologie des Canaux Ioniques et de la Synapse, INSERM UMR 1072, Aix Marseille Université, 13015 Marseille, FRANCE

[2] Current address: NYU Neuroscience Institute, New York University, New York, NY 10016, USA

[3] Department of Biochemistry & Molecular Biology, University Hospital Nord, Marseille, FRANCE

[4] Department of Molecular Genetics, Institute of Biomedical Sciences, Fukushima Medical University, Fukushima, 960-1295, JAPAN

[5] Corresponding author

§ These authors contributed equally to this work

Author contributions: M.T.P, P.B., C.F.T. and J.M.G. designed research. M.T.P., P.B., C.F.T., M.A.D., S.T., M.L., J.G., K.K. and J.M.G. performed research. M.T.P., P.B., C.F.T. and J.M.G. analyzed data. M.T.P., P.B., C.F.T. and J.M.G. wrote the manuscript.

Corresponding author: Jean-Marc GOAILLARD

UMR_S 1072, INSERM, Aix Marseille Université, Faculté de Médecine Secteur Nord, Marseille, FRANCE.

Email: jean-marc.goaillard@univ-amu.fr

SUMMARY PARAGRAPH

Extracting high-degree interactions and dependences between variables (pairs, triplets, … *k*-tuples) is a challenge posed by all omics approaches[1, 2]. Here we used multivariate mutual information (I_k) analysis[3] on single-cell retro-transcription quantitative PCR (sc-RTqPCR) data obtained from midbrain neurons to estimate the k-dimensional topology of their gene expression profiles. 41 mRNAs were quantified and statistical dependences in gene expression levels could be fully described for 21 genes: I_k analysis revealed a complex combinatorial structure including modules of pairs, triplets (up to 6-tuples) sharing strong positive, negative or zero I_k, corresponding to co-varying, clustering and independent sets of genes, respectively. Therefore, I_k analysis simultaneously identified heterogeneity (negative I_k) of the cell population under study and regulatory principles conserved across the population (homogeneity, positive I_k). Moreover, maximum information paths enabled to determine the size and stability of such transcriptional modules. I_k analysis represents a new topological and statistical method of data analysis.

MAIN TEXT

The recent evolution of single-cell transcriptomics has created much hope for our understanding of cell identity, cell development and gene regulation[4]. Using quantitative PCR or RNAseq, tens to thousands of mRNAs can be quantified from a single cell, generating particularly high-dimensional datasets (gene expression

profiles). Combined with clustering and dimensionality-reduction techniques, these approaches have been successfully used to identify and separate cell types in various tissues, including the brain[5]. Single-cell transcriptomics has also be used to shed light on the gene regulatory principles underlying the specific phenotype of different cell types[6, 7], frequently relying on pairwise analysis of gene expression levels to infer gene regulatory networks[6, 8]. However, the modular architecture of gene networks suggests that extracting higher-degree interactions between gene expression profiles may be necessary to understand gene regulation, and various approaches based on probability/information theory[8-10] or homology[11] have been proposed to tackle this issue.

Several transcriptomics studies have been performed on midbrain dopaminergic (DA) neurons[5, 12]: consistent with the heterogeneous vulnerability of this neuronal population in Parkinson's disease[13], qPCR and RNAseq performed at the single-cell level have revealed a significant diversity in gene expression profiles[5, 12]. In parallel, much work has also been performed to understand the gene regulatory networks and identify the regulatory factors underlying the emergence of the DA phenotype[14], with the therapeutical intent of producing functional DA neurons from induced pluripotent stem cells[15].

Here we implement multivariate mutual information (I_k) analysis on transcriptomics data from single midbrain DA neurons to simultaneously provide new insights about the molecular heterogeneity of this neuronal population and about the gene regulatory principles underlying its specific phenotype.

We performed sc-RTqPCR on acutely dissociated identified midbrain neurons using the microfluidic BioMark™ HD Fluidigm platform. TH-GFP mice were used to preferentially target putative DA neurons (identified by the presence of tyrosine

hydroxylase, TH-positive neurons, **Supplementary Figure 1a**). Electrophysiological recordings confirmed that acutely dissociated GFP and non-GFP neurons displayed the electrical properties expected for DA and non-dopaminergic (nDA) midbrain neurons[16, 17], respectively (**Supplementary Figure 2**). However, since TH presence alone has been shown to not be a reliable marker[18], DA and nDA phenotypes were refined based on the combined expression of *Th*/TH and *Slc6a3*/DAT (DA transporter) or lack thereof, allowing neurons collected from wild-type animals to be included (**Supplementary Figure 1b**). Based on *Th*-*Slc6a3* expression, 111 neurons were classified as DA and 37 as nDA.

We quantified the levels of expression of 41 genes (**Figure 1a**), including 19 related to ion channel function, 9 related to neurotransmitter definition, 5 related to neuronal activation and calcium binding, and 3 related to neuronal structure (**Supplementary Figure 1c**). As expected, DA metabolism and signaling-related genes such as *Th*/TH, *Slc6a3*/DAT, *Slc18a2*/VMAT2, *Drd2*/D2R were highly expressed in DA neurons only, while expression levels of *Slc17a6*/VGLUT2, *Gad1*/GAD67 and *Gad2*/GAD65 suggested that collected nDA neurons used mainly glutamate or GABA as neurotransmitters (**Figure 1a-b**, **Supplementary Figure 3**). While some ion channels showed similar expression profiles in DA and nDA neurons (*Cacna1c*/Cav1.2, *Cacna1g*/Cav3.1, *Hcn2*/HCN2, *Hcn4*/HCN4, *Kcna2*/Kv1.2, *Scn8a*/Nav1.6), others (*Kcnb1*/Kv2.1, *Kcnd3_2*/Kv4.3, *Kcnj6*/GIRK2, *Kcnn3*/SK3, *Scn2a1*/Nav1.2) displayed higher levels of expression in DA neurons (**Figure 1b**, **Supplementary Figure 3**). In addition, although a few genes displayed a fairly stable level of expression across DA neurons (*Th*/TH, *Slc6a3*/DAT, *Kcnd3_2*/Kv4.3, *Scn2a1*/Nav1.2), most genes displayed significant variability in their expression levels

(including dropout events) across cells (**Figure 1b**, **Supplementary Figure 3**), consistent with the already documented heterogeneity of midbrain DA neurons[5, 13, 14].

As a first step in deciphering higher-degree relationships, we performed Pearson correlation analysis on the 33 most relevant genes (**Figure 1c-d, Supplementary Figure 4**). The patterns of correlations were clearly different for DA and nDA neurons, with more widespread correlations in DA neurons, as can be seen in the correlation maps (**Figure 1c**). This is only partly surprising as most of the genes were chosen because of their known expression in DA neurons, but it nonetheless demonstrates that specific signatures of second-degree linear relationships participate in the identity of the two populations under study (**Figure 1d**). While most of the cell type-specific correlations involved differentially expressed mRNAs, some similarly expressed genes displayed a stronger correlation in a specific population: *Kcnj6*/GIRK2 *vs Scn2a1*/Nav1.2 for instance in DA neurons, *Scn2a1*/Nav1.2 *vs Slc17a6*/VGLUT2 or *Hcn4*/HCN4 *vs Nefm*/NEF3 in nDA neurons (**Figure 1d**, **Supplementary Figure 4**). Several correlations were also present in both cell types (*Kcna2*/Kv1.2 *vs Nefm*/NEF3, *Hcn2*/HCN2 *vs Nefm*/NEF3). Interestingly, some of the strongest correlations found in DA neurons linked the group of genes involved in DA metabolism and signaling (*Th*/TH, *Slc6a3*/DAT, *Slc18a2*/VMAT2, *Drd2*/D2R) to a group of ion channel genes (*Kcnj6*/GIRK2, *Kcnd3_2*/Kv4.3, *Kcnn3*/SK3, *Scn2a1*/Nav1.2) (**Figure 1d**, **Supplementary Figure 4**), suggesting the existence of a large module of co-regulated genes. However, the size of such modules might only be accurately defined by methods capturing high-dimensional (beyond pairs) statistical dependences.

Various information theoretical approaches have been proposed to define gene regulatory modules based on the exploration of higher-degree relationships, notably

three-way interactions[8-10] (see also **Supplementary methods**). Here we present a method that combines in a single framework statistical and topological analysis of gene expression for systematic identification and quantification of such regulatory modules, based on the information cohomology developed by Baudot and Bennequin[3]. In this framework, joint-entropy (H_k) and multivariate mutual information (I_k) quantify the variability/randomness and the statistical dependences of the variables, respectively, while simultaneoustly estimating the topology of the dataset. We restricted the general setting defining information structures from the whole lattice of partitions of joint random variables to the simplicial sublattice of "set of subsets", thus computationally allowing an exhaustive estimation of H_k and I_k at all degrees k and for every k-tuple (for $k \leq n$=21, k being the degree/number of genes analyzed as a k-tuple, n being the total number of genes analyzed; **Figure 2a, Supplementary methods).** Information values obtained with this analysis provide a ranking of the lattices at each degree k (**Supplementary methods**). The H_k and I_k analysis therefore estimate the variability and statistical dependences at all degrees k, from 1 to n. I_k is defined as follows[3, 19, 20]:

$$I_k(X_1; ...; X_k) = \sum_{i=1}^{k} (-1)^{i-1} \sum_{I \subset [k]; \mathrm{card}(I)=i} H_i(X_I)$$

giving, for k=3,

$$\begin{aligned} I_3(X_1; X_2; X_3) = {} & H_1(X_1) + H_1(X_2) + H_1(X_3) \\ & - H_2(X_1, X_2) - H_2(X_1, X_3) - H_2(X_2, X_3) \\ & + H_3(X_1, X_2, X_3) \end{aligned}$$

, where X_I denotes the joint-variable corresponding to the subset I. I_k is equivalent to entropy for k = 1, has upper and lower limit values of $\log_2(N)$ and $-\log_2(N)$ bits (N being the number of bins or graining used to discretize the data; N=8 in the present case, **Supplementary Figure 5**), is always non-negative for $k < 3$, and can take negative values for $k \geq 3$ [19-21] (**Supplementary methods**). As an example, the

maxima and minima of I_3 for 3 binary variables are depicted in **Supplementary Figure 6**: while maxima (positive I_k) correspond to a fully redundant behavior (x_1, x_2 and x_3 are informationally equivalent), the minima (negative I_k) correspond to cases where variables are pairwise independent (I_2=0) but strictly tripletwise dependent (emergent behavior). In other terms, positive I_k captures co-variations and usual linear correlations as a subcase, zeros of I_k capture statistical independence, and negativity captures more complex relationships that cannot be detected on lower dimensional projections, such as degree-specific clustering patterns (also called synergy or frustation)[9, 21, 22] (**Supplementary methods**).

We applied I_k analysis to the gene expression levels measured in DA and nDA neurons for the 21 most relevant genes (**Figure 2**). The variability in expression of each gene X_i is quantified by the entropy $H_1(X_i)=I_1(X_i)$ (**Supplementary methods**). Consistent with the expression profiles depicted in **Figure 1b**, the smallest and largest values of I_1 were found for nDA neurons (**Figure 2b,d**). The genes sharing the strongest I_2 values (**Figure 2b**) significantly overlapped with those sharing strong Pearson correlations (**Figure 1d**), in particular for DA neurons (*Th*/TH, *Slc6a3*/DAT, *Slc18a2*/VMAT2, *Drd2*/D2R, *Kcnj6*/GIRK2, *Kcnd3_2*/Kv4.3, *Kcnn3*/SK3, *Scn2a1*/Nav1.2). Nevertheless the precise patterns of I_2-sharing genes were different, due to the fact that I_k also identifies non-linear dependences[23]. Interestingly, for $k \geq 3$, the modules of genes sharing the strongest positive I_3 and I_4 displayed dense overlap with those sharing the strongest I_2, while the groups of genes sharing the strongest negative I_3 and I_4 (*Cacna1g*/Cav3.1, *Calb1*/CB, *Drd2*/D2R, *Kcna2*/Kv1.2, *Kcnb1*/Kv2.1, *Kcnj11*/Kir6.2, *Nefm*/NEF3, *Slc17a6*/VGLUT2) had very little overlap with the strongly correlated (see **Figure 1d**) or strong I_2-sharing genes (**Figure 2b**), especially for DA neurons. I_k was also calculated for superior degrees (5 to 21), and

examples of the strongest positive and negative information modules are shown for I_5 and I_{10} in **Figure 2b**. Consistent with the theoretical examples presented in **Supplementary Figure 6**, strong negative I_4 was associated with clustering patterns of expression while strong positive I_4 corresponded to co-varying patterns of expression (**Figure 2c**). In general, the distribution of I_k at each degree was found to be very different between DA and nDA neurons, with a predominance of independence (0 values) and strong negative values in nDA compared to DA neurons (**Figure 2d**).

In order to provide an exhaustive picture of the statistical dependences in both populations, we determined the information landscapes corresponding to the distribution of I_k values as a function of degree k (**Figure 3a**, **Supplementary Figure 7**). To help the reader understand this representation, two theoretical examples are given in **Supplementary Figure 7b**: for randomly equidistributed (independent) variables, $I_1 = \log_2(N)$, and $I_{2,\ldots,n} = 0$; while for strictly redundant variables (*e.g.* correlation of 1), $I_{1,\ldots,n} = \log_2(N)$. The information landscapes of DA and nDA neurons were found to be very different from these two theoretical examples and from each other: in particular, the landscape of nDA neurons mainly comprised strong negative and 0 I_k values for $k \geq 3$, suggesting that most k-tuples of genes are k-independent in these neurons. The prevalence of k-independence was found to be even stronger when the information landscape was computed for the 20 "less-relevant" genes in DA and nDA neurons (**Supplementary Figure 7c**). In contrast, the information landscape of DA neurons showed a predominance of negative I_k for $k < 5$ and predominance of positive I_k for $k \geq 5$ (**Figure 3a**). Therefore, this analysis revealed a complex combinatorial structure of gene expression profiles in DA and nDA neurons, mixing independent, synergistic and redundant k-tuples of genes for $k \geq$

3. In analogy with mean-field approximations, we also calculated the mean information for all degrees (**Figure 3a**, **Supplementary methods**). Due to the rather small number of cells analyzed and the inherent undersampling issue, the information landscapes computed here (especially the mean landscapes) should be intepreted with caution for $k > 6$ (DA) and $k > 5$ (nDA), even though maximal positive and negative I_k values are less sensitive to this limit (see **Supplementary methods**).

The I_k analysis presented in **Figure 2** revealed that modules of strong positive or negative I_k could persist across degrees, but did not allow us to estimate the size of these gene modules. In order to quantify the stability of information modules and determine their size, we estimated the information flow over paths in the lattice of random variables in DA neurons (**Figure 3b-c, Supplementary Figure 8**). For a given information path, the first derivative with respect to the degree k is given by the conditional mutual information with a minus sign (**Supplementary methods**):

$$X_i.I_{k-1}(X_1;...;\widehat{X_i};...;X_k) = I_{k-1}(X_1;...;\widehat{X_i};...;X_k) - I_k(X_1;...;X_k)$$

$$I(X_1;...;X_k) = I(X_1) - \sum_{i=2}^{k} X_i.H(X_1,...,X_{i-1})$$

, where ^ denotes the omission of X_i (the conditioning variable). $X_i.I_{k-1}$ stays positive (negative slope) if adding a variable X_i to the module increases the information while a negative $X_i.I_{k-1}$ (positive slope) indicates that adding a variable increases the uncertainty about the module. Therefore, reaching the first minima $X_i.I_{k-1} = 0$ indicates that adding a variable stops being informationally relevant, and allows to define the degree for which information modules become unstable. In other words, the degree of the first minima gives a definitive assessment of the size of a gene module.

We characterized the paths that maximized mutual information (most informative modules) or that minimize mutual information (sequence of variables that segregate the most the whole set of variables), and that stay stable (**Supplementary methods**). **Figure 3b** presents the 4 longest paths of maximal and minimal information, which correspond to stable modules of degree 6 and 4, respectively. We then built the scaffold composed of the 4 maximal and minimal information paths (**Figure 3c**). All the genes involved in defining DA metabolism and signaling were found in the scaffold of maximal paths (*Th*/TH, *Slc6a3*/DAT, *Slc18a2*/VMAT2, *Drd2*/D2R), together with three ion channel genes (*Kcnj6*/GIRK2, *Kcnd3_2*/Kv4.3, *Kcnn3*/SK3), in keeping with the pairs, triplets and quadruplets of positive I_k-sharing genes identified in **Figure 2b**. This finding brings new insights to our understanding of gene regulation in DA neurons. As shown in **Figure 2c**, the genes sharing strong positive I_k have co-varying profiles of expression, which is usually considered to indicate a co-regulation of expression[4, 8]. Therefore the positive information module determined using conditional mutual information (**Figure 3c**) should correspond to a group of genes co-targeted by the same regulatory factors. Several studies have demonstrated that the expression levels of *Th*/TH, *Slc6a3*/DAT, *Slc18a2*/VMAT2 and *Drd2*/D2R are indeed under the control of the same pair of transcription factors *Nurr1*/*Pitx3* [24] (**Supplementary Figure 9**). Our results are consistent with these observations, but moreover suggest that these four genes might be part of a larger transcriptional module ($\geq$ 7 genes) that also includes genes defining the electrical phenotype of DA neurons (*Kcnj6*/GIRK2, *Kcnd3_2*/Kv4.3, *Kcnn3*/SK3). This also means that defining the neurotransmitter identity and the electrical phenotype of these neurons might be the product of a single transcriptional program, involving at least the *Nurr1* and *Pitx3* transcription factors (**Supplementary Figure 9**). Alternatively,

this coupling between ion channel and DA metabolism genes might also reflect the documented activity-dependent regulation of DA-specific genes such as *Th*/TH, which has been shown to be sensitive to blockade of sodium (including Nav1.2) and potassium (including SK3) channel activity[25].

On the other hand, the minimal information paths identified the 8 genes that best segregate midbrain DA neurons (**Figure 3c**), supporting the already documented diversity of this neuronal population[12-14]. The presence of *Abcc8*/SUR1, *Cacna1g*/Cav3.1, *Calb1*/CB, *Gad2*/GAD65, *Kcnj11*/Kir6.2 and *Drd2*/D2R is perfectly consistent with several studies linking the expression of these genes to specific subpopulations of SNc and VTA neurons [13, 26-29] (**Supplementary Figure 9**). Importantly, our analysis reveals that other genes, in particular the potassium channels *Kcna2*/Kv1.2 and *Kcnb1*/Kv2.1 might be used as markers of midbrain DA neuron subpopulations.

In summary, we showed that the topology of a high-dimensional dataset defined by the independence, and the simple (redundant) and complex (synergistic) statistical dependences at all degrees can be estimated using multivariate mutual information analysis (I_k). Applied to sc-RTqPCR data, I_k analysis allowed us to simultaneously determine the size and identity of gene regulatory modules conserved across a cell population and the size and identity of gene modules underlying cell diversity (**Supplementary Figure 9**). Therefore, the specific complex combinatorial structure of genetic interactions (positive, negative, null) underlying the stability and diversity of a given cell type is described at once by the presented method. While applied here to transcriptomics data, I_k analysis could be applied to any type of high-dimensional data, within the limit of computational tractability.

ACKNOWLEDGEMENTS

This work was funded by the French National Research Agency (ANR JCJC grant ROBUSTEX to J.M.G.; supporting S.T.), the European Research Council (ERC consolidator grant 616827 *CanaloHmics* to J.M.G.; supporting M.T.P., P.B. and M.L.), and the French Ministry of Research (doctoral fellowship to M.A.D.). We would like to thank Pr. E. Marder for helpful discussions on the manuscript.

COMPETING INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Su, Y., Shi, Q. & Wei, W. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *Proteomics* **17** (2017).
2. Wang, Y.X. & Huang, H. Review on statistical methods for gene network reconstruction using expression data. *J Theor Biol* **362**, 53-61 (2014).
3. Baudot, P. & Bennequin, D. The Homological Nature of Entropy. *Entropy-Switz* **17**, 3253-3318 (2015).
4. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145 (2015).
5. Poulin, J.F. et al. Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling. *Cell Rep* **9**, 930-943 (2014).
6. Park, J. et al. Inputs drive cell phenotype variability. *Genome Res* **24**, 930-941 (2014).
7. Schulz, D.J., Goaillard, J.M. & Marder, E.E. Quantitative expression profiling of identified neurons reveals cell-specific constraints on highly variable levels of gene expression. *Proc Natl Acad Sci U S A* **104**, 13187-13191 (2007).
8. Villaverde, A.F., Ross, J. & Banga, J.R. Reverse engineering cellular networks with information theoretic methods. *Cells* **2**, 306-329 (2013).
9. Watkinson, J., Liang, K.C., Wang, X., Zheng, T. & Anastassiou, D. Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann N Y Acad Sci* **1158**, 302-313 (2009).
10. Margolin, A.A., Wang, K., Califano, A. & Nemenman, I. Multivariate dependence and genetic networks inference. *IET Syst Biol* **4**, 428-440 (2010).
11. Lockwood, S. & Krishnamoorthy, B. Topological features in cancer gene expression data. *Pac Symp Biocomput*, 108-119 (2015).
12. La Manno, G. et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580 e519 (2016).
13. Liss, B. & Roeper, J. Individual dopamine midbrain neurons: functional diversity and flexibility in health and disease. *Brain Res Rev* **58**, 314-321 (2008).
14. Bodea, G.O. & Blaess, S. Establishing diversity in the dopaminergic system. *FEBS Lett* **589**, 3773-3785 (2015).

15. Rivetti di Val Cervo, P. et al. Induction of functional dopamine neurons from human astrocytes in vitro and mouse astrocytes in a Parkinson's disease model. *Nat Biotechnol* **35**, 444-452 (2017).
16. Richards, C.D., Shiroyama, T. & Kitai, S.T. Electrophysiological and immunocytochemical characterization of GABA and dopamine neurons in the substantia nigra of the rat. *Neuroscience* **80**, 545-557 (1997).
17. Ungless, M.A., Magill, P.J. & Bolam, J.P. Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science* **303**, 2040-2042 (2004).
18. Lammel, S. et al. Diversity of transgenic mouse models for selective targeting of midbrain dopamine neurons. *Neuron* **85**, 429-438 (2015).
19. Hu, K.T. On the amount of information. *Theory of Probability and its Applications* **7**, 439-447 (1962).
20. Yeung, R.W. Information Theory and Network Coding. (Springer, 2008).
21. Matsuda, H. Information theoretic characterization of frustrated systems. *Physica A* **294**, 180-190 (2001).
22. Brenner, N., Strong, S.P., Koberle, R., Bialek, W. & de Ruyter van Steveninck, R.R. Synergy in a neural code. *Neural Comput* **12**, 1531-1552 (2000).
23. Reshef, D.N. et al. Detecting novel associations in large data sets. *Science* **334**, 1518-1524 (2011).
24. Jacobs, F.M. et al. Identification of Dlk1, Ptpru and Klhl1 as novel Nurr1 target genes in meso-diencephalic dopamine neurons. *Development* **136**, 2363-2373 (2009).
25. Aumann, T. & Horne, M. Activity-dependent regulation of the dopamine phenotype in substantia nigra neurons. *Journal of Neurochemistry* **121**, 497-515 (2012).
26. Evans, R.C., Zhu, M. & Khaliq, Z.M. Dopamine Inhibition Differentially Controls Excitability of Substantia Nigra Dopamine Neuron Subpopulations through T-Type Calcium Channels. *J Neurosci* **37**, 3704-3720 (2017).
27. Liss, B. et al. K-ATP channels promote the differential degeneration of dopaminergic midbrain neurons. *Nat Neurosci* **8**, 1742-1751 (2005).
28. Gonzalez-Hernandez, T., Barroso-Chinea, P., Acevedo, A., Salido, E. & Rodriguez, M. Colocalization of tyrosine hydroxylase and GAD65 mRNA in mesostriatal neurons. *Eur J Neurosci* **13**, 57-67 (2001).
29. Lammel, S. et al. Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron* **57**, 760-773 (2008).

MATERIAL AND METHODS

Acute midbrain slice preparation. Acute slices were prepared from P14–P23 TH-GFP mice (transgenic mice expressing GFP under the control of the tyrosine hydroxylase promoter) [30] of either sex. All experiments were performed according to the European and institutional guidelines for the care and use of laboratory animals (Council Directive 86/609/EEC and French National Research Council). Mice were

356 anesthetized with isoflurane (Piramidal Healthcare Uk) and decapitated. The brain
357 was immersed briefly in oxygenated ice-cold low calcium artificial cerebrospinal
358 fluid (aCSF) containing the following (in mM): 125 NaCl, 25 $NaHCO_3$, 2.5 KCl, 1.25
359 NaH_2PO_4, 0.5 $CaCl_2$, 4 $MgCl_2$, 25 glucose, pH 7.4, oxygenated with 95% O_2 / 5%
360 CO_2 gas. The cortices were removed and then coronal midbrain slices (250 µm) were
361 cut on a vibratome (Leica VT 1200S) in oxygenated ice-cold low calcium aCSF.
362 Following 30–45 min incubation in 32°C oxygenated low calcium aCSF, the slices
363 were incubated for at least 30 min in oxygenated aCSF (125 NaCl, 25 $NaHCO_3$, 2.5
364 KCl, 1.25 NaH_2PO_4, 2 $CaCl_2$, 2 $MgCl_2$ and 25 glucose, pH 7.4, oxygenated with 95%
365 O_2 / 5% CO_2 gas) at room temperature prior to electrophysiological recordings.
366 Picrotoxin (100 µM, Sigma Aldrich, St. Louis, MO) and Kynurenate (2 mM, Sigma
367 Aldrich) were bath-applied via continuous perfusion in aCSF to block inhibitory and
368 excitatory synaptic activity, respectively.
369 **Cell dissociation and collection**. Midbrain DA neurons were acutely dissociated
370 following a modified version of the methods described in references [31] and [32].
371 Regions containing the SNc, part of the VTA and SNr were excised from each
372 coronal midbrain slice. The tissue was submitted to papain digestion (2.5 mg/ml and
373 5mM L-cysteine) for 15-20 min in oxygenated low calcium HEPES aCSF (containing
374 10 mM HEPES, pH adjusted to 7.4 with NaOH) at 35-37º C and subsequently rinsed
375 in low-calcium HEPES aCSF supplemented with trypsin inhibitor and bovine serum
376 albumin (1mg/ml). Single cells were isolated by gentle trituration with fire-polished
377 Pasteur pipettes and plated on poly-L-Lysine-coated coverslips. Dissociated cells
378 were maintained in culture in low calcium HEPES aCSF at 37º in 5% CO_2 for at least
379 45 minutes. Coverslips were then placed in a cell chamber of a fluorescence
380 microscope and continuously perfused with HEPES-aCSF. Cells were collected by

aspiration into borosilicate glass pipettes mounted on a micromanipulator under visual control. Cell dissociation and collection were performed using RNA-protective technique and all solutions were prepared with RNase-free reagents when possible and filtered before use.

Electrophysiology recordings, data acquisition and analysis. All recordings were performed as already described previously [33]. Picrotoxin and Kynurenate were present for all recordings to prevent contamination of the intrinsic activity by spontaneous glutamatergic and GABAergic synaptic activity. Statistical analysis (performed according to data distribution) included: unpaired *t* test, Mann Whitney, paired *t* test with a p value <0.05 being considered statistically significant. Statistics were performed utilizing SigmaPlot 10.0 (Jandel Scientific, UK) and Prism 6 (GraphPad Software, Inc., La Jolla, CA).

qPCR assays, specific retro-transcription and targeted amplification (RT-STA). Pre-designed TaqMan assays (TaqMan® Gene Expression Assays, Thermo Fisher Scientific) used in this study are listed in **Supplementary Table 1**. Assays were systematically selected to target the coding region and to cover all known splice variants. In the case of *Kcnd3* and *Kcnj6* genes, two different assays were used to detect all known splice variants. Excluding *Fos* (754 bp intron) and *Bdnf*, *Kcna2* and *Kcnj11* (both primers and probe within a single exon), assays spanning a large intron (>1000 bp) were chosen to avoid DNA amplification. *Gad1* primers and probe were designed according to Applied Biosystems criteria and MIQE recommendations [34]. TaqMan® assays were pooled (0.2x final concentration) and the preamplification step was validated using log serial dilutions of mouse brain total RNA (MBTR) [5, 6]. The following thermal profile was applied: 50ºC for 15 min, 95ºC for 2 min and 22 cycles of amplification [35] (95ºC for 15 s and 60ºC for 4 min) following Fluidigm

recommendations. For each assay, efficiency was estimated from the slope of the standard curve using the formula $E= (10^{(-1/slope)}-1) \times 100$. All assay efficiencies ($89.4 \leq E \leq 100.4$ %) are listed in S**upplementary Table 1.**

Single-cell RTqPCR, data processing and analysis. Individual GFP and non-GFP neurons were harvested directly into 5 µl of 2x Reaction buffer (CellsDirect™ One-Step qRTPCR, Lifetech) and kept at -80ºC until further processing. A reverse transcription followed by a specific targeted pre-amplification (RT-STA) was performed in the same tube (2.5 µl 0.2x assay pool; 0.5 µl SuperScript III) applying the same thermal profile described above. The pre-amplified products were treated with ExoSAPI (Affimetrix) and diluted 5-fold prior to analysis by qPCR using 96.96 Dynamic Arrays on a BioMark System (BioMark™ HD Fluidigm). Data were analyzed using Fluidigm Real-Time PCR Analysis software (Linear Baseline Correction Method and User detector Ct, Threshold Method). Two genes, *Kcnj6_c* and *Chat* were undetectable in all analyzed cells. Cells that had a Ct for *Hprt* above 21 were excluded from further analysis. After interplate calibration, all Ct values were converted into relative expression levels using the equation $Log_2Ex = Ct_{LOD} - Ct_{(Assay)}$ [36]. LOD (limit of detection) was set to Ct=25 by calculating the theoretical Ct value for 1 single molecule in the Biomark system from two custom-designed oligonucleotides: *Slc17a6* and *Penk*. All data pre-processing was performed in Microsoft Excel (Microsoft, Redmond, USA). Heatmap and correlation maps (Pearson correlation coefficient values excluding zero values, p value <0.5, n >5) were generated in the R environment (R Core Team 2016) using gplots, heatmap3, Hmics and corrplot packages. Gene expression scatter plots and frequency distribution plots were created in SigmaPlot 10.0 (Jandel Scientific) and Prism 6

(GraphPad Software, Inc, La Jolla, CA). Figures were prepared using Adobe Illustrator CS6.

Topological information data analysis

The present analysis is based on the information cohomology framework developed by Baudot and Bennequin[3] and relies on theorems establishing uniquely the usual entropy (H_k) and multivariate mutual information (I_k) as the first class of cohomology and coboundaries respectively with finite (non-asymptotic) methods (see **Supplementary Methods** for more detail).

Simplicial Information structures

The information functions are defined on the whole lattice of partitions of the probability simplex of atomic probabilities, providing the general random variable lattice of joint-variables. The application of this framework to data analysis is developed in the subcase of simplicial information homology, which consists in the exploration of the simplicial sublattice of "set of subsets" defined dually for joint and mutual (meet) monoid structures of random variables, and whose exploration follows binomial combinatorics with a complexity in $O(2^n)$. It allows an exhaustive estimation of the information structure, that is the joint-entropy H_k and the mutual information I_k, on all degrees k and for every k-tuple of variables (gene expression levels), defined respectively by the following equations:

$$\begin{aligned} H_k &= H(X_1, ..., X_k; P_{X_1,...,X_k}) \\ &= k \sum_{x_1,...,x_k \in [N_1 \times ... \times N_k]}^{N_1 \times ... \times N_k} p(x_1.....x_k) \ln p(x_1.....x_k) \end{aligned}$$

$$I_k(X_1; ...; X_k) = \sum_{i=1}^{k} (-1)^{i-1} \sum_{I \subset [k]; \mathrm{card}(I)=i} H_i(X_I)$$

for a probability joint-distribution $P_{X1,...,Xk}$ and joint-random variables (X_1,...,X_k) with alphabet [N_1...N_k] and k=-1/ln2, where *n* variables are mutually independent if and only if $\forall\ k \leq n$, I_k=0. Due to the combinatorial complexity, in the current study H_k and I_k values were computed for *n*=21 (for *n*=21, the total number of information elements to estimate is 2 097 152).

The distributions of I_k and H_k for every degree *k* (corresponding to *k*-tuples of variables) were represented as I_k and H_k landscapes (**Supplementary Figure 7**). The landscapes are representations of the simplicial information structures where each element of the lattice is represented as a function of its corresponding value of entropy or mutual information, and quantify the variability-ramdomness and statistical dependencies at all degrees *k*, respectivey, from 1 to *n*. Mean landscapes were calculated by averaging I_k and H_k for each degree *k* over the number of *k*-tuples. The mean information landscape quantifies the average behavior of the whole structure. The mean information landscape (or path) is given by:

$$\langle I_k \rangle = \frac{\sum_{T \subset [n]; card(T)=i} I_k(X_T; P)}{\binom{n}{k}}$$

Probability estimation

The probability estimation procedure is explained in **Supplementary Figure 5** for the simple case of two random variables (the expression levels of two genes). For each variable X_j, we consider the space in the intervals [min x_j, max x_j] and divide it into N_j boxes, *N* being the graining of the data. The empirical joint probability is estimated by box counting after a graining of the data space into N_1...N_k boxes (for *k*-tuple

probability estimation). In the current study, a graining of N_1=...=N_k=8 was chosen as it provided a correct description of the distribution of the expressions levels (see **Supplementary Figure 8** for the influence of changing the graining on the identification of gene modules).

Information paths

An information path IP_k or HP_k of degree k on I_k or a H_k landscape is defined as a sequence of elements of the lattice that begins at the leastest element of the lattice (the identity-constant "0"), travels along edges from element to element of increasing degree of the lattice and ends at the greatest element of the lattice of degree k. The first derivative of an IP_k path is minus the conditional mutual information. The ("non-Shannonian") information inequalities [19], e.g. the negativity of conditional mutual information that quantifies the instability of the mutual information along the path, are then equivalent to the existence of local minima on such paths (see **Supplementary methods**). The critical dimension of an IP_k path is the degree of its first minima. A positive information path is an information path from 0 to a given I_k corresponding to a given k-tuple of variables such that $I_k<I_{k-1}<...<I_1$. We call the marginal component of a path I_1 a self-information energy and the interacting compoenents functions I_k, $k > 1$, a free information energy. A maximal positive information path is a positive information path of maximal length: it ends at a minima of the free information energy function. In the current study, the length of maximal positive information paths was considered to indicate the size of a stable information module. The set of all these paths defines uniquely the minimum free information complex (see **Supplementary methods**). In simple terms, this complex is the homological formulation of the minimum energy principle with potentially many local and degenerate minima. The set of all paths of degree k is in one-to-one correspondence

with the symmetric group S_k and hence untractable computationally (complexity in $O(k!)$). In order to bypass this issue, we used a fast local algorithm that selects at each element of degree k of an IP path the positive information path with maximal or minimal I_{k+1} value or stops whenever $X_k.I_{k+1} \leq 0$ and rank those paths by their length.

Robustness of the method

To estimate the degree after which the sample size m becomes limiting and biases our estimations, the undersampling regime was quantified by the degree k_u beyond which a significant proportion (10%) of the H_k values get close to $\log_2(m)$. Using these criteria, with $\log_2(111)=6.79$ and $\log_2(37)=5.21$, the k_u obtained for DA neurons was 6 and 5 for nDA neurons, and I_k and H_k values beyond these degrees should be interpreted with caution (**Supplementary methods**). It must be noted however that this limit is calculated on the average H_k, whose value is mainly determined by non-relevant independent k-tuples. The biologically relevant statistical dependences correspond to extrema in the raw landscape (minimal H_k and maximal or minimal I_k) and therefore are less affected by this sampling problem. In order to evaluate the robustness of our results to sample size (m) and graining value (N), we calculated the maximal positive paths obtained for DA neurons for smaller samples (m = 28, 56, 84, taken fully arbitrarily among 111) and smaller (N=4, 6) or larger (N=10, 12) graining values (**Supplementary Figure 8**). The information paths of maximal length were found to be relatively robust to variations in N and m, even though, as expected, m=28 yielded significantly different paths. For most N and m combinations, the main genes identified in **Figure 3c** were also present in the maximal information paths, including in particular the DA metabolism/signaling genes and the two ion channel genes *Kcnd3*/Kv4.3 and *Kcnn3*/SK3. Concerning the statistical significance of the results, I_2 functions are Kullback-Leibler divergences [37] and estimate the divergence from 2-

independence. Their generalization to arbitrary degree k (I_k) can be interpreted as a statistical significance of a test, here against the null hypothesis of k-independence $I_k=0$. Our analysis, based on the ranking of the I_k for every k, considered only the 5 maximal (positive) and 5 minimal (negative) values of I_k, which are the 5 most significantly dependent positive and negative I_k-sharing k-tuples (for $k > 2$).

Computation and algorithm

The Information Topology open source program, written in Python, is available on Github depository. It allows to compute the information landscapes, paths, and minimum free energy complex, which encode and represent directly all the usual equalities, inequalities, and functions of information theory (as justified at length in **Supplementary methods**), and all the structures of the statistical dependences within a given set of empirical measures (up to the approximations, computational tractability and finite size biases, see previous sections). It can be run on a regular personal computer up to $k = n = 21$ random-variables in reasonable time (3 hours), and provide new tools for pattern detection, dimensionality reduction, ranking and clustering based on a unified homological and informational theory.

30. Sawamoto, K. et al. Visualization, direct isolation, and transplantation of midbrain dopaminergic neurons. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 6423-6428 (2001).
31. Guyon, A., Laurent, S., Paupardin-Tritsch, D., Rossier, J. & Eugene, D. Incremental conductance levels of GABAA receptors in dopaminergic neurones of the rat substantia nigra pars compacta. *J Physiol* **516 (Pt 3)**, 719-737 (1999).
32. Puopolo, M., Raviola, E. & Bean, B.P. Roles of subthreshold calcium current and sodium current in spontaneous firing of mouse midbrain dopamine neurons. *J Neurosci* **27**, 645-656 (2007).
33. Dufour, M.A., Woodhouse, A., Amendola, J. & Goaillard, J.M. Non-linear developmental trajectory of electrical phenotype in rat substantia nigra pars compacta dopaminergic neurons. *eLife* **3** (2014).
34. Bustin, S.A. et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**, 611-622 (2009).

552 35. Citri, A., Pang, Z.P., Sudhof, T.C., Wernig, M. & Malenka, R.C. Comprehensive qPCR
553 profiling of gene expression in single neuronal cells. *Nat Protoc* **7**, 118-127 (2012).
554 36. Stahlberg, A., Rusnakova, V., Forootan, A., Anderova, M. & Kubista, M. RT-qPCR
555 work-flow for single-cell data analysis. *Methods* **59**, 80-88 (2013).
556 37. Kullback, S. & Leibler, R. On information and sufficiency. *Annals of Mathematical*
557 *Statistics* **22**, 79-86 (1951).

558

559

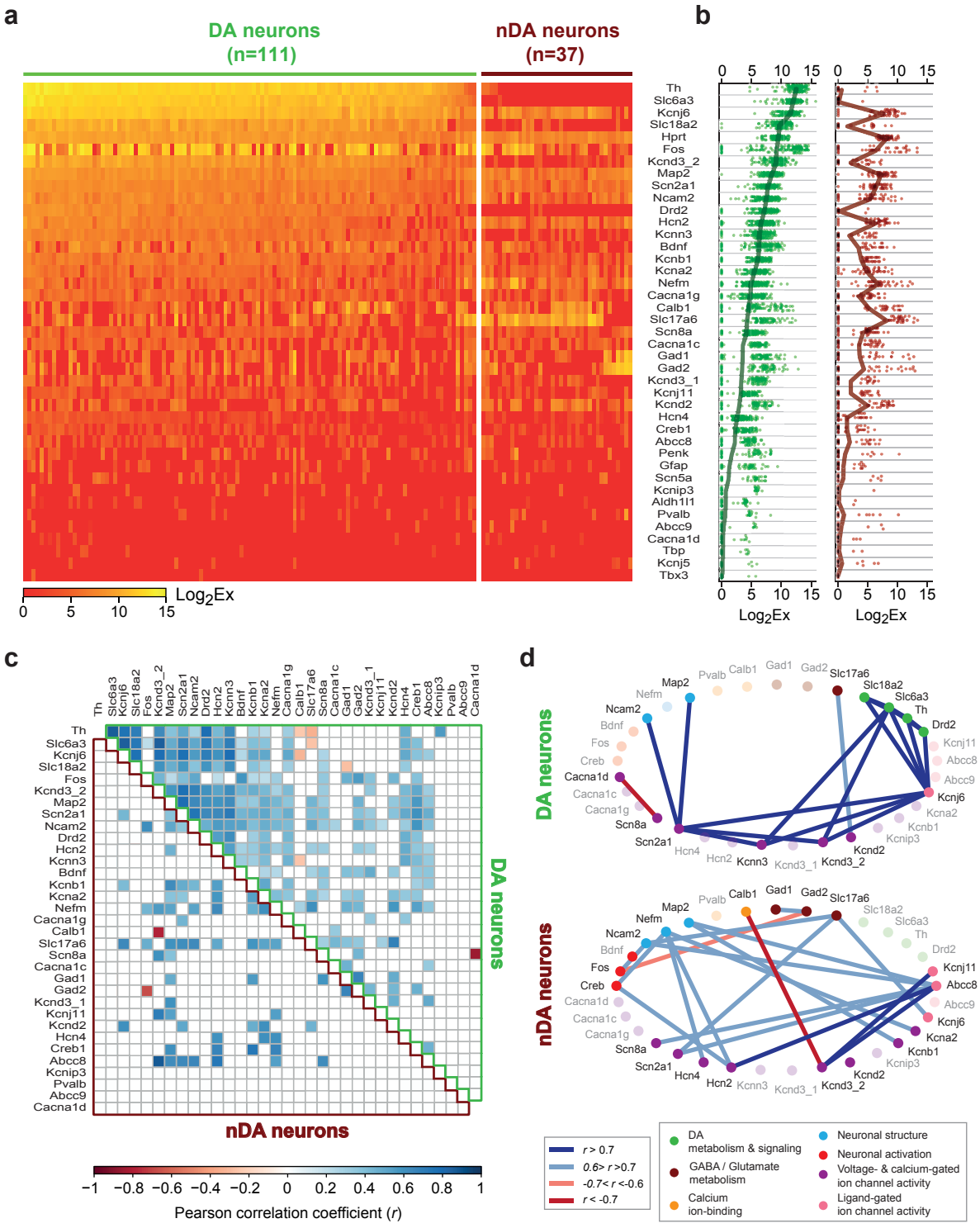

Figure 1. First and second order linear analysis reveals strong correlations in gene expression levels in midbrain DA and nDA neurons. a, heatmap representing the levels of expression of 41 genes in the collected 111 DA and 37 nDA neurons. Neurons are ordered based on *Th* and *Slc6a3* levels of expression, and genes are ordered based on their average level of expression in DA neurons (see **b**, left plot). **b**, levels of expression of the 41 genes presented in **a** in the DA population (left, green) and in the nDA population (right, dark red). The thick green and red lines represent the average expression levels while each dot corresponds to the expression level in one neuron. **c**, heatmap representing the significant correlations in expression levels for 33 genes in DA neurons (upper right triangle) and nDA neurons (lower left triangle) (Pearson correlation coefficient). Correlations were processed on non-zero values of expression, and only correlations with a p value < 0.05 and n > 5 are represented. Please note the difference in patterns of correlations between DA and nDA neurons. **d**, scaffold representations of the 20 most significant correlations in expression levels in DA (top) and nDA (bottom) neurons (r values > 0.6 or < -0.6, see color coding in the left box). mRNAs were ordered based on the known function of the corresponding proteins (see right box for the color coding of functions). The genes involved in the depicted correlations are highlighted (dark font, bright colors). Please note the strong connectivity between DA metabolism/signaling and ion channel genes in DA neurons.

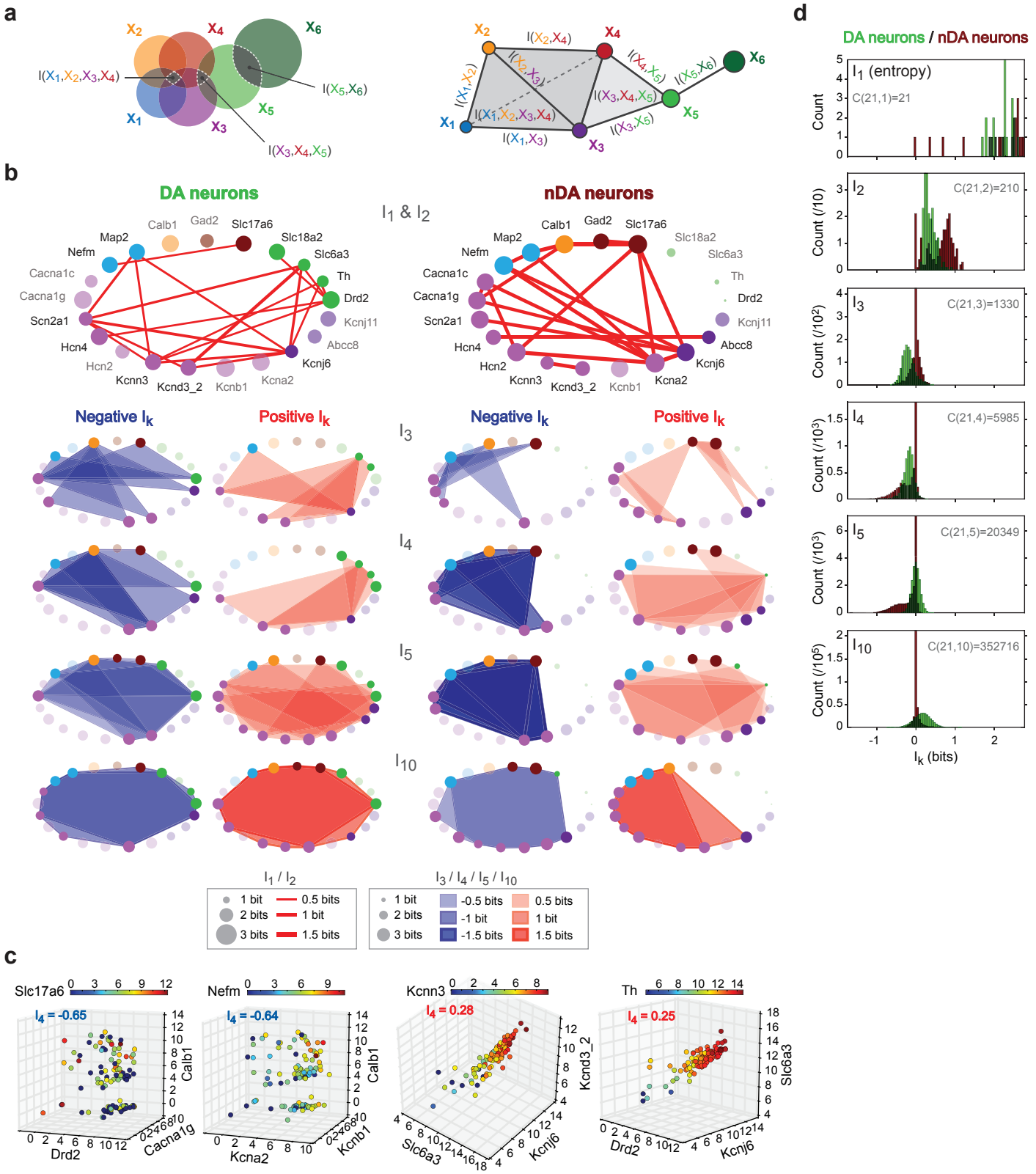

Figure 2. Mutual information analysis of gene expression levels reveals specific high-degree structures of transcriptomic profiles in midbrain DA and nDA neurons. a, left, Venn diagram illustrating a system of 6 random variables sharing mutual information at degree 2, 3 and 4. Right, same system represented on a simplicial complex. Each vertex represents a variable while the edges, faces and volumes represent joint-*n*-tuples of variables. **b**, scaffold representations of the most significant I_k values shared by pairs (I_2, top row), triplets (I_3, second row), quadruplets (I_4, third row), quintuplets (I_5, fourth row) and decuplets (I_{10}, fifth row) of genes in DA (left column) and nDA neurons (right column). Circle diameters are scaled according to entropy value (I_1). The red shapes (lines, triangles, quadrilaterals, etc) indicate positive I_k shared by genes while the blue shapes correspond to negative I_k. Only the most significant values of I_k are displayed on each scaffold (20 for I_2; 5 positive and 5 negative for I_3, I_4, I_5; and 2 positive and 2 negative for I_{10}). **c**, 4D-scatter plots representing the levels of expression of 2 quadruplets of genes sharing strong negative I_k (left-hand plots) and 2 quadruplets of genes sharing strong positive I_k (right-hand plots) in DA neurons. Please note that negative I_k is associated with a "clustering" or "heterogeneous" distribution of gene expression levels (left) while positive I_k is associated with a "co-varying" or "homogeneous" distribution of gene expression levels. **d**, histograms representing the distribution of I_k values for all the degrees presented in **b**. The total number of combinations C(*n*,*k*) for each degree (number of pairs for I_2; number of triplets for I_3, etc) is given in gray.

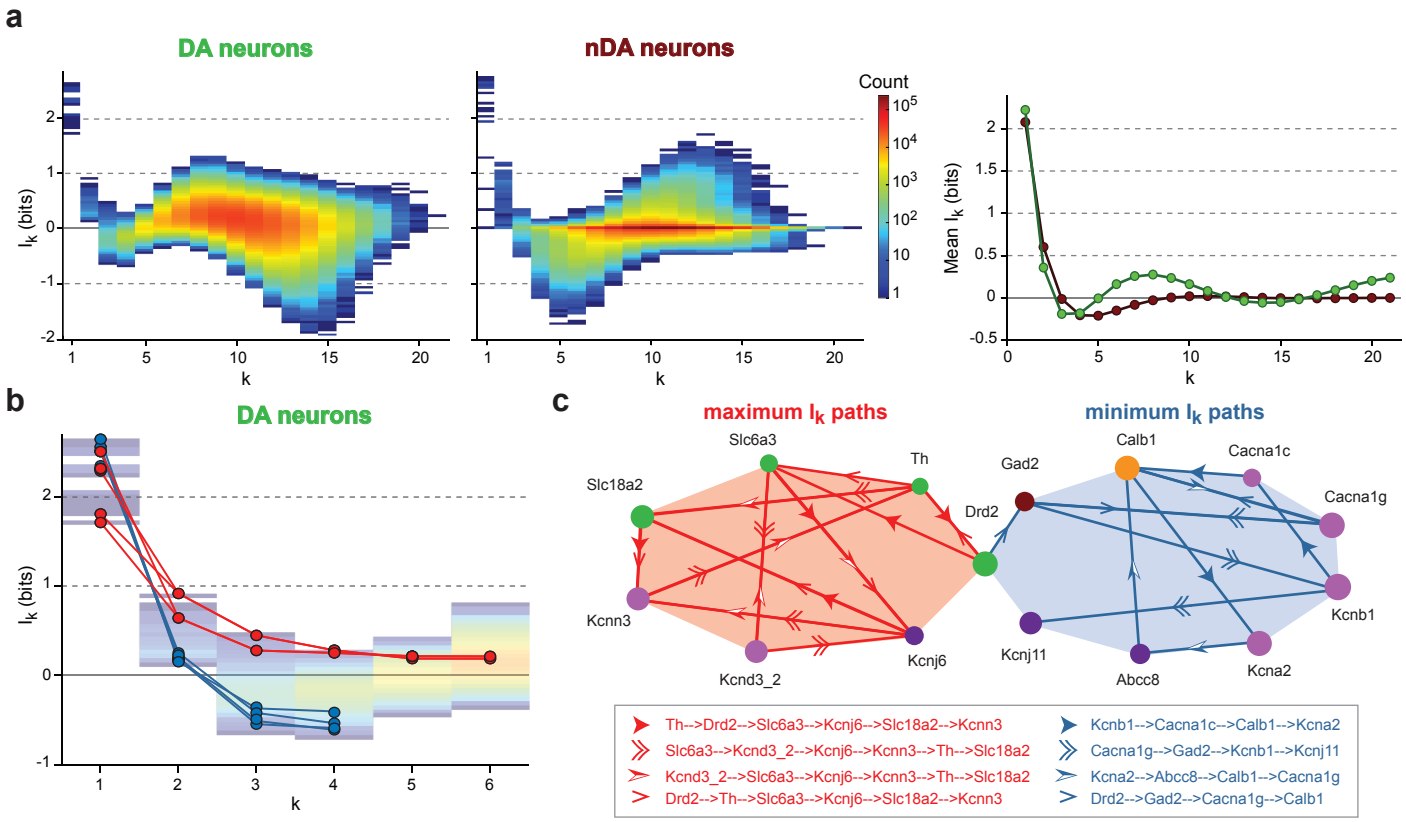

Figure 3. Conditional mutual information identifies stable modules of genes sharing strong statistical dependences. a, information landscapes representing the distribution of I_k values shared by the 21 genes presented in **Figure 2** as a function of degree *k* in DA (left) and nDA neurons (middle), and scatter plot representing mean I_k value (mean information landscape) as a function of degree for both populations (right). Color coding in the left and middle plots indicates the density of points for each I_k value. **b**, line and scatter plot illustrating the four maximum (red) and minimum (blue) information paths corresponding to stable information modules in DA neurons identified using conditional information computation. The paths have been superimposed to the total information landscape (transparent color coding) already shown in panel **a**. **c**, scaffold representation of the information modules corresponding to the maximum I_k paths (red) and minimum information paths delineated in panel **b**. Each path is identified with a specific arrowhead shape (see legend box).