EvoKEN: evolutionary knowledge extraction in networks

Benjamin Linard, Ngoc Thanh Nguyen, Odile Lecompte, Olivier Poch, Julie D. Thompson

EvoKEN: evolutionary knowledge extraction in networks

Benjamin Linard*, Ngoc Hoan Nguyen, Odile Lecompte, Olivier Poch & Julie D. Thompson

Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Fédération de médecine translationelle de Strasbourg, France.

**Corresponding Author: benjamin.linard@gmail.com*

We introduce a multi-factorial, multi-level approach to build and explore evolutionary scenarios of complex protein networks. EvoKEN combines a unique formalism for integrating multiple types of data associated with network molecular components and knowledge extraction techniques for detecting cohesive/anomalous evolutionary processes. We analyzed known human pathway maps and identified perturbations or specializations at the local topology level that reveal important evolutionary and functional aspects of these cellular systems.

The dynamic molecular machinery underlying cellular systems is often represented by complex, hierarchical networks of interactions between the cell's constituents, such as proteins, DNA, RNA and small molecules. Ultimately, phenotypic traits and diseases can be described in terms of the complex intracellular and intercellular networks that link tissue and organ systems[1,2]. The structures of these networks, including metabolic, signaling or transcription regulatory networks, often share similar features even in distantly related species[3]. Understanding the evolution of these networks is therefore essential to

reconstruct the history of life, but also to better understand how the network structures correlate with the functioning of organisms at different granularity levels[4,5].

Application of evolutionary based methods in complex networks is challenging[6] and requires integration of multiple factors, such as gene spatial/temporal expression, protein sequence conservation, cellular localization signals, 3D structure, or binding/interaction sites[7]. In this context, we have developed an original formalism, called the evolutionary barcode or EvoluCode[8], to allow the integration of different parameters (e.g. genome context, protein organization, conservation patterns) in a common framework and to summarize the evolutionary history of a gene that leads to its current state in a given organism (Supplementary Fig. 1). EvoluCode thus facilitates the application of formal data mining and knowledge extraction techniques in evolutionary analyses. We previously used this approach to barcode all human protein-coding genes using 10 evolutionary data types from 17 vertebrate proteomes. Our systematic comparison of the human barcodes revealed protein function-evolution relationships that could not be observed by using only one or two biological parameters, for example using only sequence conservation[8].

Here, we introduce a unique protocol, called EvoKEN (Evolutionary Knowledge Extraction in Networks), that combines the EvoluCode formalism with knowledge extraction techniques, in order to study the evolution of genes in the context of their complex biological networks. We show how EvoKEN can be used at the pathway level to identify local topological motifs that have evolved cohesively and to highlight 'outlier' genes whose evolutionary history deviates from the local neighbors, suggesting different underlying evolutionary processes.

We then extend our work to investigate unusual evolutionary scenarios at the inter-pathway or 'cellular' level.

Our protocol can be applied to any biological system, where we define a system as a set of genes implicated in a common process or phenomenon (genetic information processing, signal transduction, metabolism, disease response, etc.) and mapped onto a molecular network. We demonstrate the utility of our approach by constructing and exploring evolutionary scenarios for the complete set of human pathway maps in the Kyoto Encyclopedia of Genes and Genomes (KEGG) knowledge base[9]. First, we mapped our EvoluCodes for the human proteome to the KEGG maps, thus producing pathway-level evolutionary maps (Fig. 1) for a total of 248 biological systems (available at lbgi.igbmc.fr/barcodes). We then applied a knowledge extraction algorithm (Online Methods) on each individual map in order to estimate its evolutionary cohesiveness and to identify genes with anomalous, 'outlier' barcodes that might reflect unusual evolutionary pressures within the system. Here, we used the Local Outlier Factor (LOF)[10], a powerful anomaly detection algorithm which is related to density-based clustering and is suitable for analyzing large-scale, multidimensional datasets where the underlying data distribution is unknown. The LOF method identified a total of 1147 outlier genes in 248 KEGG maps (lbgi.igbmc.fr/barcodes and Supplementary Fig. 2). The most cohesive pathways, i.e. those with the least outliers (Supplementary Fig. 3 and Supplementary Table 1), were typically involved in universal biological processes such as translation or cell growth/death, in line with previous observations[11].

To further investigate the biological significance of genes with anomalous evolutionary histories, we measured the correlations between the EvoKEN outliers and their local

topology in the corresponding networks. We focused on the metabolic pathways in KEGG, where the nodes in the networks represent metabolites (substrates, products and intermediates) that are linked by a reaction, associated with one or more genes/proteins. Within these networks, we manually defined 6 classes of local topological motifs based on 2 key node properties, redundancy and connectivity (Fig. 2a and Online Methods). The outlier genes from 20 metabolic pathways were then assigned to the different topology classes (Fig. 2b). We found that the cohesiveness of a gene in its network context depends on the local topological structure: for instance, the smallest proportion of outliers was found at the nodes involved in linear paths in the networks, particularly in non-redundant paths (class F). In contrast, more outliers were found at the start/end points of a pathway (class D), and at the interface between pathways, so called 'hubs' in the networks (class C). The correlation we observe between gene conservation and local network topology may be due to specific selection pressures, for instance on essential genes[12].

Having established the evolutionary cohesiveness of individual pathways, we then asked whether we could identify unusual evolutionary behavior at the cellular level. Individual pathways often function in a coordinated fashion and understanding the interactions or crosstalk between pathways is important for deciphering complex cellular processes, such as the appropriate physiological responses to internal or external stimuli. To investigate these high-level processes, we identified a set of genes involved in the crosstalk between 155 KEGG pathway maps, reflected by the fact that all the genes in the set were present in at least 3 maps. In this case, the evolutionary cohesiveness of a gene is context-dependent, i.e. a gene may be defined as cohesive in one of these pathways and as an outlier in another (Fig. 2c). Such cases of differential evolutionary conservation may indicate important events, such as gene duplications, rearrangements or losses and the subsequent gain or loss of

interactions in the network. For each pair of KEGG maps, we calculated the proportion of outlier genes observed in the overlapping set of genes shared between the two systems. We then constructed a global map of the relationships between the 155 maps, representing the evolutionary behavior of these pathways during vertebrate evolution (Fig. 2d and Supplementary Fig. 4). The exploration of this map provides a powerful and visual means of highlighting important events in the evolution of human biological systems.

Two examples are highlighted in Fig. 2d. First, the genes involved in both cell cycle and oocyte meiosis pathways are generally cohesive with the other genes in these pathways, but the crosstalk with the progesterone-mediated oocyte maturation pathway contains a higher proportion of outlier genes (Supplementary Table 2). In fact, cell cycle and oocyte meiosis pathways are conserved in most vertebrates, while the exact nature of oocyte maturation is more variable between species. A number of these functional specificities are highlighted by the EvoKEN outliers, such as the Myt1 gene coding for a cdc2-inhibitory kinase (PMYT1_HUMAN), which acts as a negative regulator of entry into mitosis during the cell cycle. Inspection of the Myt1 evolutionary barcode (Supplementary Fig. 5a) indicates a more divergent sequence family than is typical for this conserved pathway. This might be a result of the different functions of Myt1, which is implicated in control of entry into meiosis, either alone (as in *Xenopus*) or in concert with Wee1 (as in mouse oocytes)[13]. Other examples of outliers are provided in Supplementary Fig. 5.

The second example concerns the innate immune system, where pattern recognition receptors, such as Toll-like receptors (TLR), RIG-I-like receptors (RLR) or NOD-like receptors (NLR), recognize a wide variety of pathogens and endogenous molecules and trigger complex, overlapping intracellular signaling cascades. Outlier genes involved in the crosstalk

between these pathways are described in Supplementary Table 3. We highlight one example: the receptor interacting protein RIP1 (RIPK1_HUMAN), which plays a crucial role in the cellular response to TLR and RLR signals, switching between cell survival through RIP1 activation of NF-κB and cell death induced by caspase-8 cleavage of RIP1[14]. The RIP1 evolutionary barcode (Supplementary Fig. 6a) shows a typical sequence conservation in vertebrate evolution, but synteny is only observed in mammals and not in fishes for example where RIP1 plays a different role in TLR signaling[15]. Other examples of outliers are provided in Supplementary Fig. 6. Unraveling the evolutionary history of these pathways and their crosstalk will be important in understanding how the immune system functions and in developing effective therapeutic and vaccine strategies.

It is clear that more in-depth analysis, involving phylogenetic tree and ancestor reconstruction would be required to describe in detail the evolutionary events identified in these studies. The advantage of EvoKEN is that it provides an effective framework for investigating the evolution of large systems at different granularity levels from local network motifs to the cellular level, allowing the rapid identification of interesting patterns in a particular biological context. Hopefully, EvoKEN will contribute to the emerging field of evolutionary systems biology, with the goal of understanding and modeling the topological and dynamic properties of the complex networks that govern the behavior of the cell.

METHODS

Construction of EvoluCodes for the human proteome

The evolutionary barcodes (EvoluCodes) used in this study were constructed as described in [8]. Each protein-coding human gene is thus associated with one EvoluCode that is visualized as a 2D matrix. The columns of the matrix correspond to the studied organisms, which in this

work consist of 17 vertebrates with almost complete genomes from the Ensembl [16] database (version 51). The rows of the matrix correspond to different evolutionary parameters (Table 1) that were extracted from multiple alignments [17,18], synteny analysis and orthology data[19]. For each vertebrate organism, the most closely related homolog to the human reference gene was identified (based on percent residue identity) and 10 parameters were calculated.

Table 1. Evolutionary parameters included in EvoluCodes of human proteome.

Parameter name	Description	Source
length	length of the vertebrate sequence	Multiple alignment
length_difference	difference in length between the human reference and vertebrate sequences	Multiple alignment
no_of_regions	number of conserved regions shared between the human reference and vertebrate sequences	Multiple alignment
sequence_identity	percent residue identity shared between the human reference and vertebrate sequences	Multiple alignment
no_of_domains	number of known protein domains (from the Pfam[20] database) in the vertebrate sequence	Multiple alignment
domain_conservation	parameter indicating domain structure conservation between the human reference and vertebrate sequences: unchanged domain structure/domain gains/domain losses/domain shuffling	Multiple alignment
hydrophilicity	average hydrophilicity of the vertebrate sequence	Multiple alignment
inparalog	number of human inparalogs with respect to the vertebrate species. This parameter represents the duplicability of a human gene compared to the other species	Ortholog/paralog database
co-ortholog	number of co-orthologs in the vertebrate species with respect to human. This parameter reflects gene duplications in the non human lineage	Ortholog/paralog database
Synteny	parameter indicating conservation of	Synteny database

	genome neighborhood: synteny on both sides of the gene / synteny either downstream or upstream of the gene / no synteny	

To facilitate visualization of the EvoluCode, a color is assigned to each matrix cell representing typical or atypical parameter values. To do this, the distribution of each parameter in each organism is first described by the sample percentiles, using the Emerson-Strenio formulas [21] implemented in the R software and color gradients are assigned to three intervals:

- Interval 1 represents values that are lower than what is generally observed for a specific parameter in a specific organism and is assigned a blue-to-green gradient
- Interval 2 represents values that correspond to what is generally observed for a specific parameter in a specific organism and is assigned a green color
- Interval 3 represents values that are higher than what is generally observed for a specific parameter in a specific organism and is assigned a green-to-red gradient.

By compiling several evolutionary parameters extracted from different biological levels, from residue data to phylum data, EvoluCodes incorporate an evolutionary systems biology point of view. Consequently, EvoluCodes can highlight important evolutionary events that could not be discovered using a single evolutionary parameter such as sequence conservation or domain composition. The complete set of 19778 human EvoluCodes can be visualized online at: lbgi.igbmc.fr/barcodes.

Analysis of human pathway data and definition of cohesive/outlier genes

We based our analysis on pathway data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) knowledge base. We analyzed 248 human pathways with the help of the KEGG SOAP server (http://www.kegg.jp/kegg/soap/). A total of 5849 EvoluCodes could be mapped to the genes in these pathways.

For each pathway, we then identified 'outlier' genes, i.e. genes with an unusual evolutionary history (EvoluCode) compared to the other genes in the pathway. We determined outliers using an anomaly detection algorithm called Local Outlier Factor (LOF) [22]. The basic concept of LOF is the local density, where locality is given by k nearest neighbors. By comparing the local density of an evolutionary barcode to the local densities of its neighbors, we identify regions of similar density, as well as barcodes that have a substantially lower density than their neighbors. These are considered to be outliers. The local density is estimated by the typical distance at which a barcode can be "reached" from its neighbors.

First, the 2D matrix representing an EvoluCode, consisting of 10 rows and 17 columns, is redimensioned to a 1D vector of length, n=170. Then, if A and B are 2 EvoluCodes in Euclidean n-space, with A = (a_1, a_2,...,a_n) and B = (b_1, b_2,...,b_n), the distance between A and B is:

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$

The Euclidean distance between the EvoluCode A and its k nearest neighbors is denoted kdist(A) and the set of k nearest neighbors is $N_k(A)$. The reachability distance is then calculated as:

$$reachability\_dist_k\ (A,B) = \max\{kdist(B), d(A,B)\}$$

The local reachability density (lrd) of EvoluCode A is defined as:

$$\text{lrd}_k(A) = 1 \Big/ \left(\frac{\sum_{B \in N_k(A)} reachability\_dist_k\ (A,B)}{|N_k(A)|}\right)$$

And the local reachability densities are then compared with those of the neighbors using:

$$\text{LOF}_k(A) = \left(\frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|}\right)$$

The LOF score thus represents the cohesiveness of the EvoluCode associated with each gene in the context of its pathway. The authors of the LOF algorithm consider that a score less than 1 indicates a clear inlier object, i.e. a cohesive barcode. Genes with a LOF score significantly greater than 1 are considered as outliers. However, the threshold determining a clear outlier depends on the dataset. Here, we defined the outlier threshold value as the upper quartile for the LOF scores of the EvoluCodes in the context of the 248 human pathways, which was 1.037.

Analysis of metabolic pathways

The KEGG database currently contains pathway data for 84 human KEGG metabolic pathways, where the nodes in the networks represent metabolites (substrates, intermediates and products). The edges between nodes represent reactions that are associated with one or more genes/proteins. For our experiment, we selected all pathways with more than 20 human genes, giving us an initial set of 20 pathways, containing a total of 875 different reactions, of which 671 reactions were associated with cohesive genes and 204

reactions with outlier genes. We defined 6 classes of local topological motifs within these pathways, based on node redundancy and connectivity (Table 2).

Table 2. Definition of local network topology classes.

Class	Redundancy	Connectivity	Description
A	yes	N/A	Alternative gene for same reaction step
B	yes	N/A	Alternative path for n>1 reactions
C	N/A	Inter-pathway	Pathway interface
D	N/A	1 intra-pathway	Start/end of pathway, single substrate/product
E	N/A	>2 intra-pathway	Multiple substrates and/or products
F	No	2 intra-pathway	Other: mostly linear paths, plus a small number of exceptions, such as unlinked genes

We then determined the topological localization of all biological reactions associated with the outlier genes. For each class, we calculated the proportion of cases where the reaction was associated with an outlier gene. In cases where a reaction was associated with more than one gene (protein complexes, genes with similar biochemical functions, etc.), we used the gene with the lowest LOF score. This choice reduced the number of reactions that are considered as outliers and only reactions with a clear outlier status were included for analysis.

Construction of the cellular level evolutionary map

We constructed a cellular level map, representing the evolutionary histories of the pathways in the KEGG database. For the 200 human KEGG pathways, we identified the genes shared

by each pair of pathways. We then focused our analysis on the pathway pairs sharing at least 3 genes, representing 155 KEGG pathways, which describe mainly cellular processes and signal transductions. In the evolutionary map, each node represents a specific KEGG pathway and the edge joining 2 nodes represents the genes shared by the two pathways. The node diameter is proportional to the number of genes implicated in the pathway. Each node is assigned a color representing the homogeneity of the EvoluCodes associated with the genes. The cohesiveness of the pathway evolution is estimated based on LOF value dispersion, using the IQR (interquartile range, IQR = $Q_3 - Q_1$). A low IQR indicates more cohesive barcodes associated with a given node.

Pathways with high cohesiveness are indicated by dark blue and pathways with low cohesiveness are light blue. The edge thickness is proportional to the number of genes shared by the 2 nodes, while the edge color indicates the proportion of shared genes identified as outliers in one or both linked pathways. A green edge links pathways that do not share any outlier genes. A red edge links pathways where all shared genes are outliers in at least one of the maps. Intermediate values are assigned a green to red color gradient.

SUPPLEMENTARY INFORMATION

Supplementary Figures 1-6 and Supplementary Tables 1-3 are are available in the second half of the document. For cytoscape files: contact Thompson(at)unistra(dot)fr

ACKNOWLEDGMENTS

This work was supported by the Decrypthon program, co-funded by Association Française contre les Myopathies (AFM, 14390-15392), IBM and Centre National de la Recherche Scientifique (CNRS). We acknowledge financial support from the ANR (EvolHHuPro: BLAN07-1-198915 and Puzzle-Fit: 09-PIRI-0018-02) and Institute funds from the CNRS, INSERM, and the Université de Strasbourg.

AUTHOR CONTRIBUTIONS

B.L., J.D.T. and O.P. conceived the study. B.L. and N.H.N. devised and implemented the algorithm and conducted experiments. B.L., N.H.N. O.L. J.D.T. and O.P. discussed the results and implications. O.L. and O.P. evaluated the biological relevance of the results. J.D.T. supervised the project. B.L. and J.D.T. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. T. F. Mackay, E. A. Stone, and J. F. Ayroles, *Nat Rev Genet* 10 (8), 565 (2009).
2. A. L. Barabasi, N. Gulbahce, and J. Loscalzo, *Nat Rev Genet* 12 (1), 56 (2011).
3. A. Schuler and E. Bornberg-Bauer, *Methods Mol Biol* 696, 273 (2011).
4. T. Yamada and P. Bork, *Nat Rev Mol Cell Biol* 10 (11), 791 (2009).
5. R. A. Pache and P. Aloy, *PLoS One* 7 (2), e31220 (2012).
6. G. Musso, A. Emili, and Z. Zhang, *Methods Mol Biol* 856, 363 (2012).
7. D. F. Veiga, B. Dutta, and G. Balazsi, *Mol Biosyst* 6 (3), 469 (2010).
8. B. Linard, N. H. Nguyen, F. Prosdocimi et al., *Evol Bioinform Online* 8, 61 (2012).
9. M. Kanehisa, S. Goto, Y. Sato et al., *Nucleic Acids Res* 40 (Database issue), D109 (2012).
10. J.H.M. Janssens, I. Flesch, and Postma. E.O., presented at the 8th International Conference on Machine Learning and Applications, Miami, USA, 2009.
11. L. Fokkens and B. Snel, *PLoS Comput Biol* 5 (1), e1000276 (2009).
12. E. V. Koonin and Y. I. Wolf, *Nat Rev Genet* 11 (7), 487 (2010).
13. M. Gaffre, A. Martoriati, N. Belhachemi et al., *Development* 138 (17), 3735 (2011).
14. N. Festjens, T. Vanden Berghe, S. Cornelis et al., *Cell Death Differ* 14 (3), 400 (2007).
15. A. Rebl, T. Goldammer, and H. M. Seyfert, *Vet Immunol Immunopathol* 134 (3-4), 139 (2010).
16. P. Flicek, M. R. Amode, D. Barrell et al., *Nucleic Acids Res* 40 (Database issue), D84 (2012).

17. F. Plewniak, L. Bianchetti, Y. Brelivet et al., *Nucleic Acids Res* 31 (13), 3829 (2003).
18. J. D. Thompson, A. Muller, A. Waterhouse et al., *BMC Bioinformatics* 7, 318 (2006).
19. B. Linard, J. D. Thompson, O. Poch et al., *BMC Bioinformatics* 12, 11 (2011).
20. M. Punta, P. C. Coggill, R. Y. Eberhardt et al., *Nucleic Acids Res* 40 (Database issue), D290 (2012).
21. J.D. Emerson and Strenio J., in *Understanding Robust and Exploratory Data Analysis*, edited by J.W. Tukey F. Mosteller, D.C. Hoaglin (Wiley, New York, 1983), pp. 58.
22. H.P. Kriegel, M.M. Breunig, R.T. Ng et al., presented at the ACM SIGMOD Int. Conf., 2000.

FIGURES

Figure 1 | Framework to construct and explore multi-level evolutionary networks.

Evolutionary barcodes, known as EvoluCodes, are assigned to individual genes and then mapped onto a known gene network, such as a KEGG pathway map. At the system level, the resulting evolutionary map provides a context for differentiating genes with 'cohesive' or 'outlier' (highlighted in red) evolutionary histories. At the cellular level, systems and inter-system crosstalk can be analyzed in terms of the cohesiveness of the underlying gene evolution.

Figure 2 | Characterization of outlier genes at the system and cellular levels. (a) Definition of 6 classes of local topological motifs in metabolic pathways, depending on the redundancy and connectivity of the reactions (and associated genes) in the network. (b) Identification of outlier genes and their distribution in the local topology classes. (c) The crosstalk between 2 systems is characterized by the proportion of shared outlier genes, indicated by a color gradient from green (all cohesive) to red (all outlier). (d) An integrated evolutionary map of selected human pathways showing the number and cohesiveness of the gene evolutionary histories, associated with individual pathways (nodes) and pathway crosstalk (edges).

Figure 1

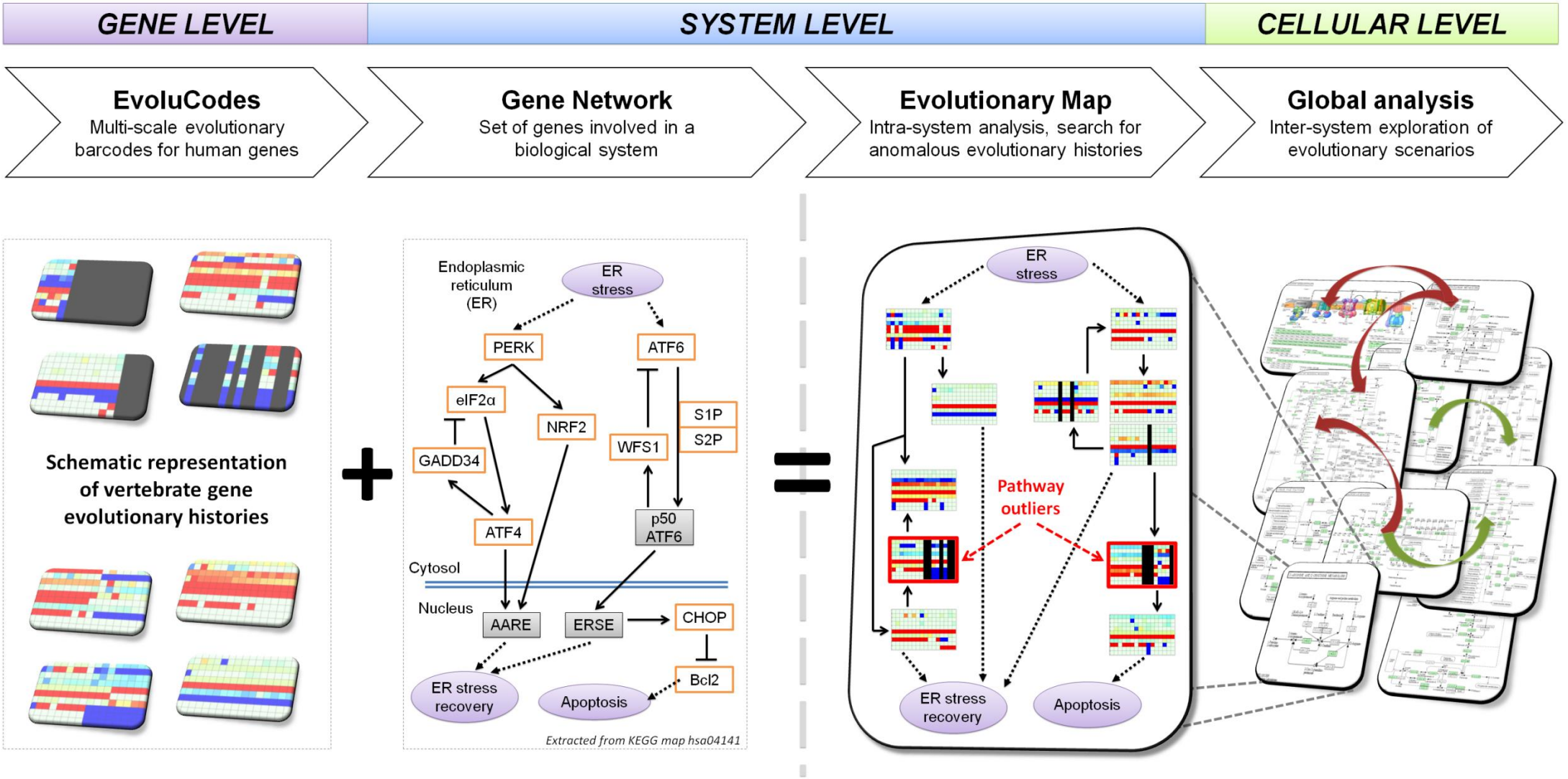
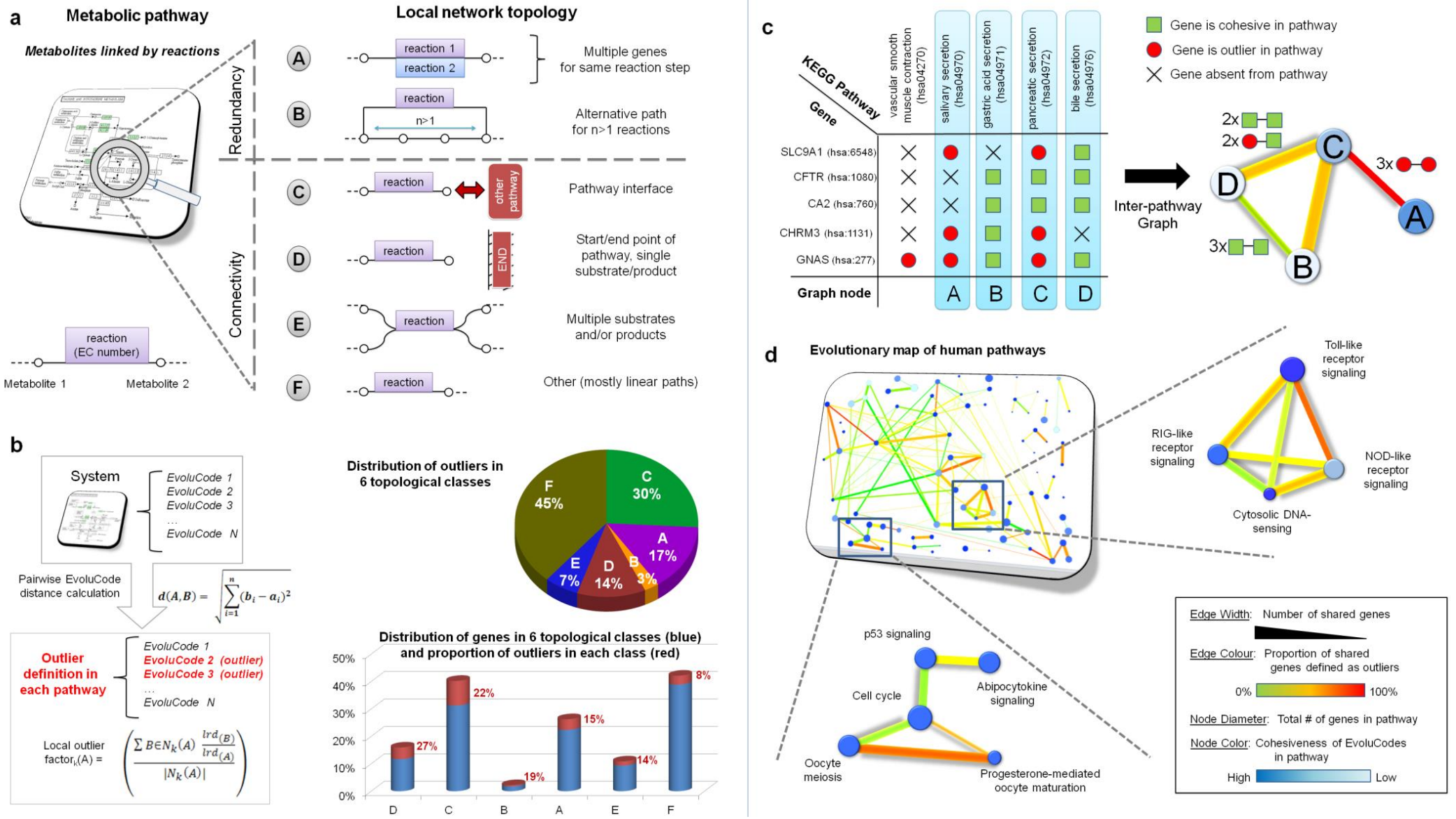
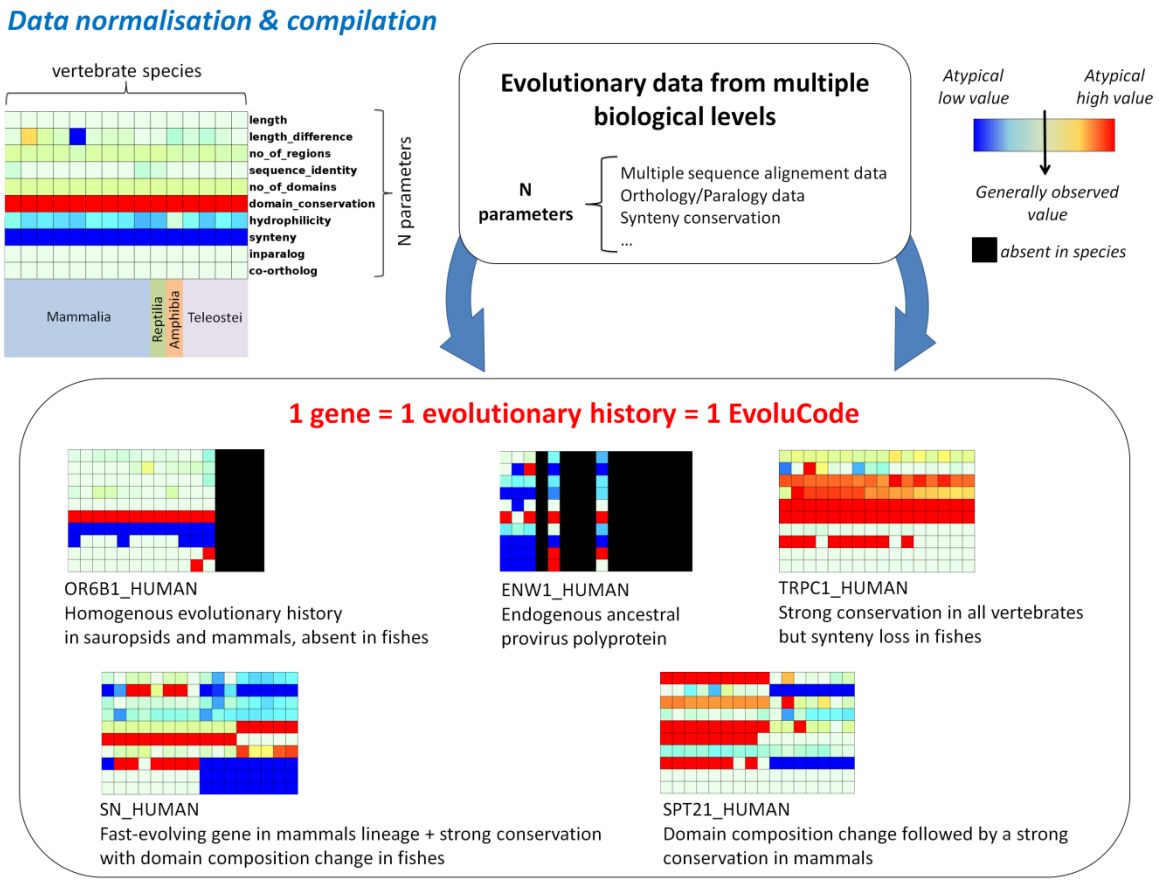

GENE LEVEL
SYSTEM LEVEL
CELLULAR LEVEL
EvoluCodes
Multi-scale evolutionary barcodes for human genes
Gene Network
Set of genes involved in a biological system
Evolutionary Map
Intra-system analysis, search for anomalous evolutionary histories
Global analysis
Inter-system exploration of evolutionary scenarios
Schematic representation of vertebrate gene evolutionary histories
Endoplasmic reticulum (ER)
ER stress
PERK
ATF6
eIF2α
NRF2
S1P
S2P
WFS1
GADD34
ATF4
p50 ATF6
Cytosol
Nucleus
AARE
ERSE
CHOP
Bcl2
ER stress recovery
Apoptosis
Extracted from KEGG map hsa04141
ER stress
Pathway outliers
ER stress recovery
Apoptosis

Figure 2


a
Metabolic pathway
Metabolites linked by reactions
Local network topology
Redundancy
Connectivity
A
reaction 1
reaction 2
Multiple genes for same reaction step
B
reaction
n>1
Alternative path for n>1 reactions
C
reaction
other pathway
Pathway interface
D
reaction
END
Start/end point of pathway, single substrate/product
E
reaction
Multiple substrates and/or products
F
reaction
Other (mostly linear paths)
reaction (EC number)
Metabolite 1
Metabolite 2
b
System
EvoluCode 1
EvoluCode 2
EvoluCode 3
EvoluCode N
Pairwise EvoluCode distance calculation
Outlier definition in each pathway
EvoluCode 1
EvoluCode 2 (outlier)
EvoluCode 3 (outlier)
EvoluCode N
Local outlier factor_k(A) =
Distribution of outliers in 6 topological classes
F 45%
C 30%
A 17%
B 3%
D 14%
E 7%
Distribution of genes in 6 topological classes (blue) and proportion of outliers in each class (red)
50%
40%
30%
20%
10%
0%
27%
22%
19%
15%
14%
8%
D
C
B
A
E
F
c
KEGG Pathway
Gene
vascular smooth muscle contraction (hsa04270)
salivary secretion (hsa04970)
gastric acid secretion (hsa04971)
pancreatic secretion (hsa04972)
bile secretion (hsa04976)
SLC9A1 (hsa:6548)
CFTR (hsa:1080)
CA2 (hsa:760)
CHRM3 (hsa:1131)
GNAS (hsa:277)
Graph node
A
B
C
D
Gene is cohesive in pathway
Gene is outlier in pathway
Gene absent from pathway
Inter-pathway Graph
2x
2x
3x
3x
d
Evolutionary map of human pathways
Toll-like receptor signaling
RIG-like receptor signaling
NOD-like receptor signaling
Cytosolic DNA-sensing
p53 signaling
Cell cycle
Abipocytokine signaling
Oocyte meiosis
Progesterone-mediated oocyte maturation
Edge Width: Number of shared genes
Edge Colour: Proportion of shared genes defined as outliers
0%
100%
Node Diameter: Total # of genes in pathway
Node Color: Cohesiveness of EvoluCodes in pathway
High
Low

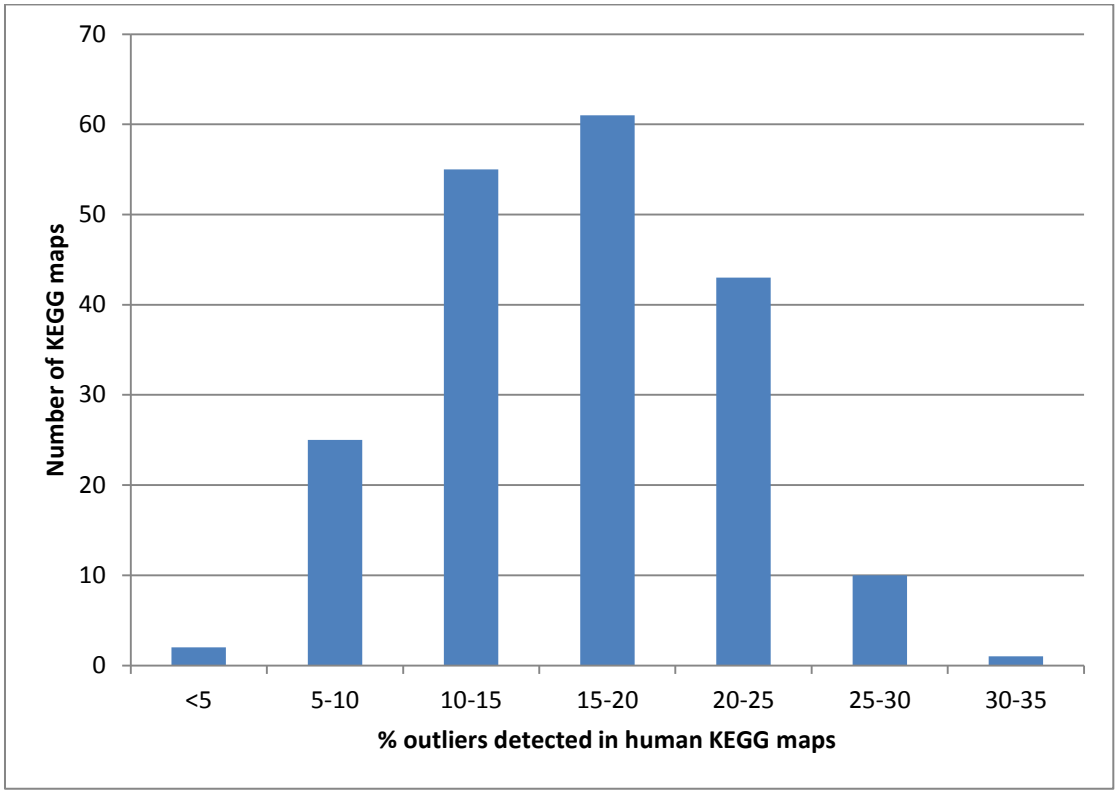
SUPPLEMENTARY DATA

Supplementary figure 1: Examples of EvoluCode evolutionary barcodes, representing 5 genes with different evolutionary histories.



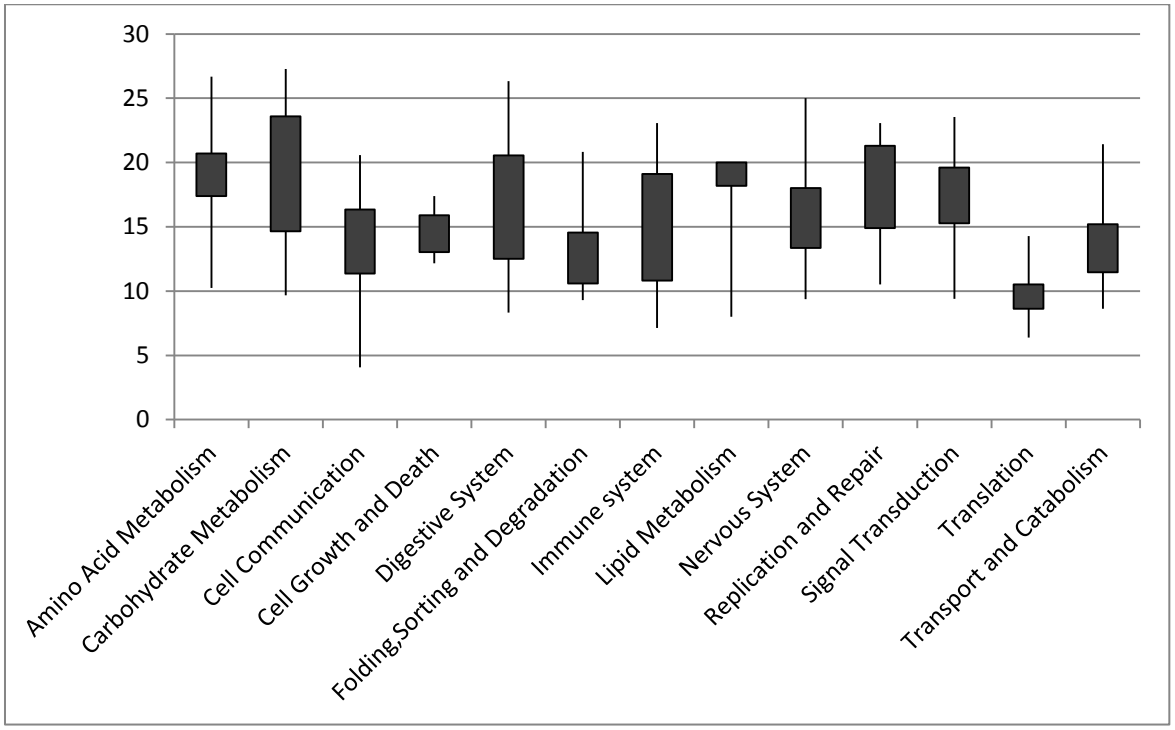
SUPPLEMENTARY DATA

Supplementary figure 2: Histogram of percentage of genes with anomalous, outlier EvoluCodes in 248 human metabolic pathways from the KEGG database.



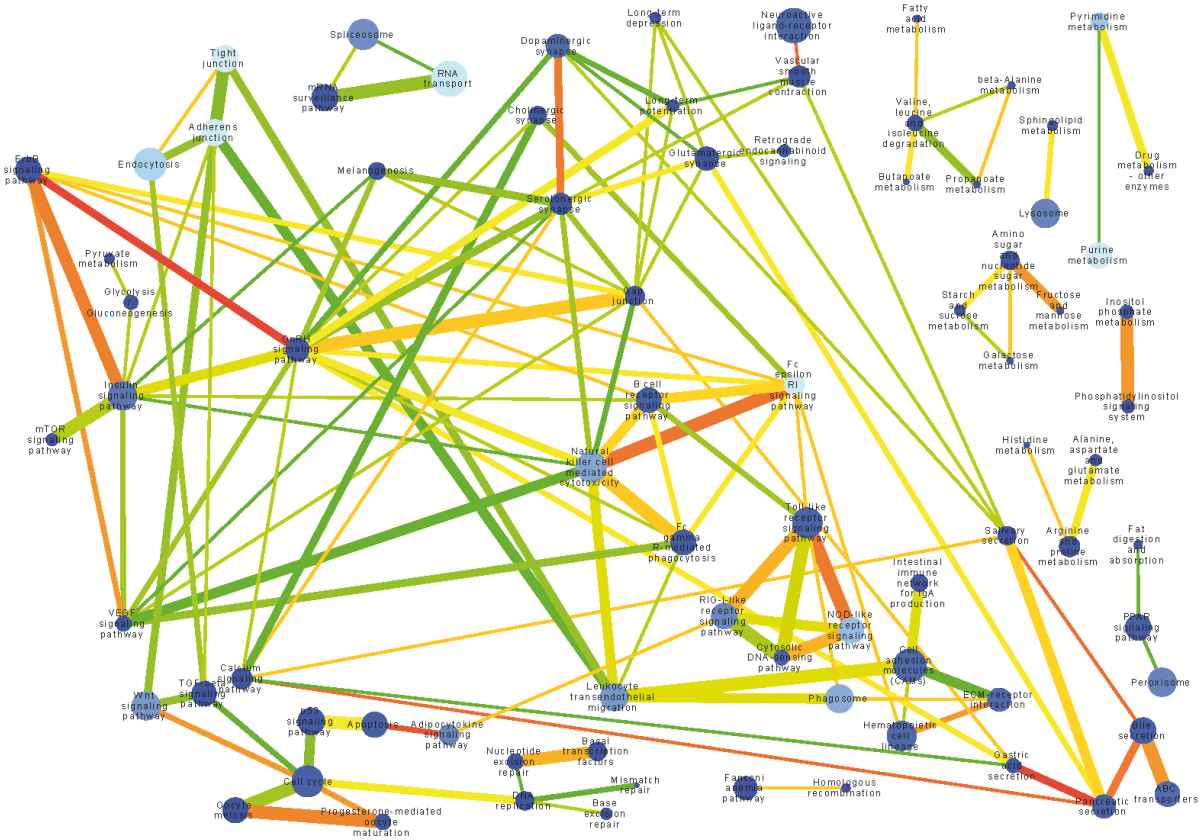
SUPPLEMENTARY DATA

Supplementary figure 3: Boxplots (minimum, maximum, lower and upper quartiles) of percentage of genes with anomalous, outlier EvoluCodes in 248 human metabolic pathways from KEGG, classified by functional groups. Pathways with a larger number of outlier genes have less cohesive evolutionary histories, e.g. some genes are more or less well conserved than the majority of genes in the pathway.



SUPPLEMENTARY DATA

Supplementary Figure 4: An integrated evolutionary map of 155 human pathways showing the number and cohesiveness of the gene evolutionary histories, associated with individual pathways (nodes) and pathway crosstalk (edges). Nodes and edges are colored as in Fig. 2 in main text. Map is visualized using Cytoscape.



SUPPLEMENTARY DATA

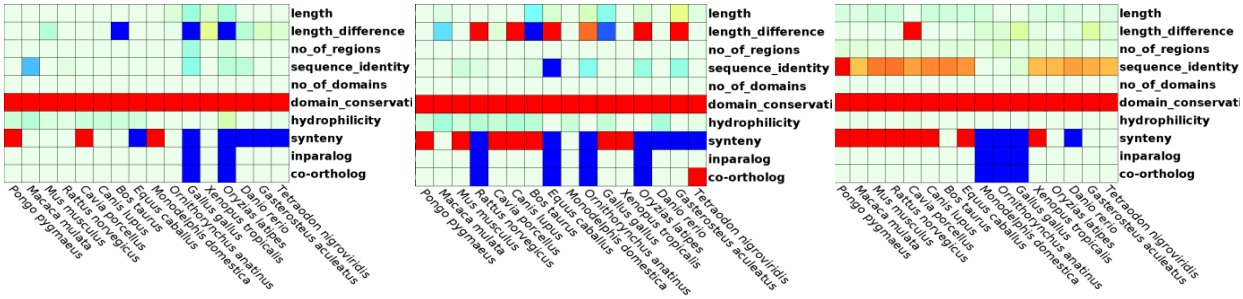
Supplementary Figure 5: Example outlier EvoluCodes.

a. Myt1 is an outlier in the cell cycle pathway. The Myt1 gene and its evolutionary barcode are described in the main text.

b. Mos is an outlier in the progesterone-mediated oocyte maturation pathway. The Mos gene is a key regulator of oocyte meiotic maturation, arresting the unfertilized oocyte at various meiotic stages depending on the species[1]. Inspection of the Mos evolutionary barcode shows that the protein sequences are generally conserved among vertebrates, however the patterns of synteny and paralogy are more divergent amongst species.

c. Cdk2 is an outlier in the progesterone-mediated oocyte maturation pathway. The cdk2 gene codes for a cyclin-dependent kinase that functions in the cell cycle in S phase progression[2]. It also plays an alternative role in the regulation of progesterone receptor (PR) signaling. PR and its coactivators are phosphoproteins. Cyclin A/Cdk2 phosphorylates several of the PR phosphorylation sites *in vitro* and there is evidence that it participates in PR phosphorylation *in vivo*[3]. Cdk2 is dispensable for the mitotic cell cycle, but it is crucial for the first meiotic division of male and female germ cells, and it has been suggested that Cdk2 might have evolved primarily as a meiotic kinase with a secondary role in the mitotic cell cycle [4]. Although the EvoluCode shows conservation in most vertebrates studied here, a perturbation is highlighted in *Monodelphis domestica* (opossum), *Ornithorhynchus anatinus* (platypus) and *Gallus gallus* (chicken) with lower sequence identity, loss of synteny and fewer inparalog/co-ortholog relationships.

a. Myt1 b. Mos c.Cdk2



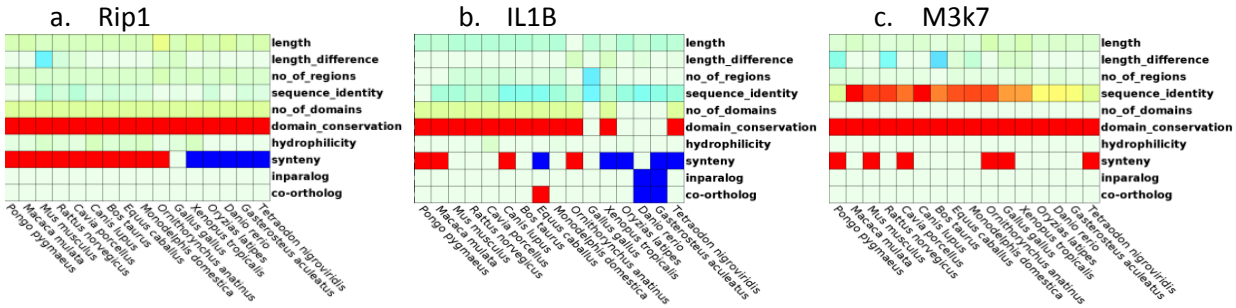
SUPPLEMENTARY DATA

Supplementary Figure 6: Example outlier EvoluCodes.

a. Rip1 is an outlier in the RIG-I-like receptor signaling pathway. The rip1 gene and its evolutionary barcode are described in the main text.

b. IL1B is an outlier in all 3 pathways where it is present. IL1B (interleukin-1 beta) is a cytokine that plays a crucial role in mediation and amplification of the innate immune response to bacterial pathogens. It is produced as an inactive precursor, termed pro-IL-1β, in response to molecular motifs carried by pathogens. After processing, the active IL-1β molecule is secreted. Given the similarity between the genomic organization of pro-IL-1a and pro-IL-1b genes, it has been hypothesized that pro-IL-1b may have arisen by a reverse transcriptase mediated duplication of the related alpha gene[5]. The EvoluCode of IL1B shows low conservation in terms of sequence identity, domain organization and synteny and is an outlier in the 3 innate immune system pathways where it is present.

c. M3K7 is an outlier in the Toll-like and NOD-like receptor signaling pathways. M3K7 (mitogen-activated protein kinase kinase kinase 7, also known as TAK1) is a serine/threonine kinase, which acts as an essential component of the MAPK signal transduction pathway. The EvoluCode associated with M3K7 shows unusually high sequence identity, but a complex synteny pattern.



SUPPLEMENTARY DATA

Supplementary table 1: Outliers detected in KEGG pathway maps. Only maps with at least 10 genes genes mapped to Uniprot proteins are shown, since outliers cannot be determined accurately for smaller sets of genes.

KEGG functional group	KEGG pathway	No. Genes	No. Outliers	% Outliers
Amino Acid Metabolism	hsa00280	29	4	13.8
Amino Acid Metabolism	hsa00270	20	3	15
Amino Acid Metabolism	hsa00250	23	4	17.4
Amino Acid Metabolism	hsa00380	23	4	17.4
Amino Acid Metabolism	hsa00330	36	7	19.4
Amino Acid Metabolism	hsa00310	15	3	20
Amino Acid Metabolism	hsa00260	29	6	20.7
Amino Acid Metabolism	hsa00350	23	5	21.7
Amino Acid Metabolism	hsa00340	15	4	26.7
Cancers	hsa05220	39	4	10.3
Cancers	hsa05214	28	3	10.7
Cancers	hsa05215	39	5	12.8
Cancers	hsa05212	38	5	13.2
Cancers	hsa05200	162	22	13.6
Cancers	hsa05202	146	20	13.7
Cancers	hsa05211	36	5	13.9
Cancers	hsa05216	13	2	15.4
Cancers	hsa05221	30	5	16.7
Cancers	hsa05222	36	6	16.7
Cancers	hsa05213	29	5	17.2
Cancers	hsa05217	17	3	17.6
Cancers	hsa05219	26	5	19.2
Cancers	hsa05210	36	7	19.4
Cancers	hsa05218	20	4	20
Cancers	hsa05223	32	7	21.9
Carbohydrate Metabolism	hsa00020	16	1	6.3
Carbohydrate Metabolism	hsa00500	22	2	9.1
Carbohydrate Metabolism	hsa00052	16	2	12.5
Carbohydrate Metabolism	hsa00562	26	4	15.4
Carbohydrate Metabolism	hsa00030	17	3	17.6
Carbohydrate Metabolism	hsa00640	16	3	18.8
Carbohydrate Metabolism	hsa00520	33	7	21.2
Carbohydrate Metabolism	hsa00010	27	6	22.2
Carbohydrate Metabolism	hsa00051	17	4	23.5

SUPPLEMENTARY DATA

Carbohydrate Metabolism	hsa00620	21	5	23.8
Carbohydrate Metabolism	hsa00650	15	4	26.7
Carbohydrate Metabolism	hsa00040	11	3	27.3
Cardiovascular Diseases	hsa05410	31	3	9.7
Cardiovascular Diseases	hsa05416	28	3	10.7
Cardiovascular Diseases	hsa05412	25	4	16
Cardiovascular Diseases	hsa05414	30	5	16.7
Cell Communication	hsa04520	49	2	4.1
Cell Communication	hsa04530	50	6	12
Cell Communication	hsa04510	59	9	15.3
Cell Communication	hsa04540	34	7	20.6
Cell Growth and Death	hsa04110	74	9	12.2
Cell Growth and Death	hsa04114	45	6	13.3
Cell Growth and Death	hsa04115	52	8	15.4
Cell Growth and Death	hsa04210	46	8	17.4
Cell Mobility	hsa04810	64	10	15.6
Ciculatory System	hsa04270	37	7	18.9
Ciculatory System	hsa04260	13	3	23.1
Development	hsa04380	68	5	7.4
Development	hsa04320	13	2	15.4
Development	hsa04360	60	11	18.3
Digestive System	hsa04976	48	4	8.3
Digestive System	hsa04970	33	3	9.1
Digestive System	hsa04974	22	3	13.6
Digestive System	hsa04975	19	3	15.8
Digestive System	hsa04972	40	7	17.5
Digestive System	hsa04978	30	6	20
Digestive System	hsa04971	27	6	22.2
Digestive System	hsa04973	19	5	26.3
Endocrine and Metabolic Diseases	hsa04930	20	2	10
Endocrine and Metabolic Diseases	hsa04940	19	3	15.8
Endocrine and Metabolic Diseases	hsa04950	25	4	16
Endocrine System	hsa03320	47	3	6.4
Endocrine System	hsa04916	31	4	12.9
Endocrine System	hsa04910	61	9	14.8
Endocrine System	hsa04914	29	5	17.2
Endocrine System	hsa04920	35	7	20
Endocrine System	hsa04912	40	8	20
Energy Metabolism	hsa00190	57	12	21.1
Environmental Adaptation	hsa04710	11	2	18.2

SUPPLEMENTARY DATA

Excretory system	hsa04961	21	3	14.3
Excretory system	hsa04960	18	3	16.7
Excretory system	hsa04962	16	3	18.8
Excretory system	hsa04964	12	3	25
Folding,Sorting and Degradation	hsa03050	43	4	9.3
Folding,Sorting and Degradation	hsa03018	52	5	9.6
Folding,Sorting and Degradation	hsa04120	96	13	13.5
Folding,Sorting and Degradation	hsa03060	21	3	14.3
Folding,Sorting and Degradation	hsa04141	82	12	14.6
Folding,Sorting and Degradation	hsa04130	24	5	20.8
Glycan Biosynthesis and Metabolism	hsa00601	13	1	7.7
Glycan Biosynthesis and Metabolism	hsa00510	32	3	9.4
Glycan Biosynthesis and Metabolism	hsa00563	24	3	12.5
Glycan Biosynthesis and Metabolism	hsa00534	14	2	14.3
Glycan Biosynthesis and Metabolism	hsa00514	15	4	26.7
Immune Diseases	hsa05323	42	4	9.5
Immune Diseases	hsa05330	16	3	18.8
Immune Diseases	hsa05320	20	4	20
Immune Diseases	hsa05340	34	7	20.6
Immune Diseases	hsa05322	29	6	20.7
Immune Diseases	hsa05332	13	3	23.1
Immune system	hsa04623	28	2	7.1
Immune system	hsa04622	47	4	8.5
Immune system	hsa04670	56	5	8.9
Immune system	hsa04610	57	6	10.5
Immune system	hsa04621	45	5	11.1
Immune system	hsa04666	44	5	11.4
Immune system	hsa04062	49	7	14.3
Immune system	hsa04620	67	10	14.9
Immune system	hsa04612	25	4	16
Immune system	hsa04672	31	5	16.1
Immune system	hsa04650	69	13	18.8
Immune system	hsa04660	62	12	19.4
Immune system	hsa04664	35	7	20
Immune system	hsa04662	43	9	20.9
Immune system	hsa04640	65	15	23.1
Infectious Diseases	hsa05160	62	5	8.1
Infectious Diseases	hsa05145	52	5	9.6
Infectious Diseases	hsa05168	91	9	9.9
Infectious Diseases	hsa05130	27	3	11.1

SUPPLEMENTARY DATA

Infectious Diseases	hsa05166	109	13	11.9
Infectious Diseases	hsa05164	96	12	12.5
Infectious Diseases	hsa05146	39	5	12.8
Infectious Diseases	hsa05140	39	5	12.8
Infectious Diseases	hsa05133	52	7	13.5
Infectious Diseases	hsa05134	43	6	14
Infectious Diseases	hsa05100	34	5	14.7
Infectious Diseases	hsa05144	34	5	14.7
Infectious Diseases	hsa05152	95	15	15.8
Infectious Diseases	hsa05162	73	12	16.4
Infectious Diseases	hsa05131	36	6	16.7
Infectious Diseases	hsa05150	30	5	16.7
Infectious Diseases	hsa05110	17	3	17.6
Infectious Diseases	hsa05120	33	6	18.2
Infectious Diseases	hsa05132	44	8	18.2
Infectious Diseases	hsa05142	55	10	18.2
Infectious Diseases	hsa05143	24	6	25
Lipid Metabolism	hsa00590	25	2	8
Lipid Metabolism	hsa00071	17	3	17.6
Lipid Metabolism	hsa00564	32	6	18.8
Lipid Metabolism	hsa00600	21	4	19
Lipid Metabolism	hsa00140	25	5	20
Lipid Metabolism	hsa00561	15	3	20
Lipid Metabolism	hsa00565	10	2	20
Membrane Transport	hsa02010	43	6	14
Metabolism of Cofactors and Vitamins	hsa00860	18	2	11.1
Metabolism of Cofactors and Vitamins	hsa00830	19	3	15.8
Metabolism of Cofactors and Vitamins	hsa00760	14	3	21.4
Metabolism of Other Amino Acids	hsa00480	19	3	15.8
Metabolism of Other Amino Acids	hsa00410	17	4	23.5
Nervous System	hsa04725	32	3	9.4
Nervous System	hsa04724	40	4	10
Nervous System	hsa04722	71	9	12.7
Nervous System	hsa04727	26	4	15.4
Nervous System	hsa04720	24	4	16.7
Nervous System	hsa04723	24	4	16.7
Nervous System	hsa04728	40	7	17.5
Nervous System	hsa04730	22	4	18.2

SUPPLEMENTARY DATA

Nervous System	hsa04726	37	7	18.9
Nervous System	hsa04721	16	4	25
Neurodegenerative Diseases	hsa05014	35	4	11.4
Neurodegenerative Diseases	hsa05016	47	6	12.8
Neurodegenerative Diseases	hsa05010	51	9	17.6
Neurodegenerative Diseases	hsa05012	26	5	19.2
Neurodegenerative Diseases	hsa05020	20	4	20
Nucleotide Metabolism	hsa00230	47	6	12.8
Nucleotide Metabolism	hsa00240	31	5	16.1
Replication and Repair	hsa03440	19	2	10.5
Replication and Repair	hsa03420	29	4	13.8
Replication and Repair	hsa03030	33	6	18.2
Replication and Repair	hsa03460	40	8	20
Replication and Repair	hsa03410	23	5	21.7
Replication and Repair	hsa03430	13	3	23.1
Sensory System	hsa04742	18	2	11.1
Sensory System	hsa04740	15	2	13.3
Sensory System	hsa04744	12	3	25
Signal Transduction	hsa04010	117	11	9.4
Signal Transduction	hsa04150	24	3	12.5
Signal Transduction	hsa04020	36	5	13.9
Signal Transduction	hsa04310	60	10	16.7
Signal Transduction	hsa04370	28	5	17.9
Signal Transduction	hsa04330	21	4	19
Signal Transduction	hsa04350	42	8	19
Signal Transduction	hsa04630	26	5	19.2
Signal Transduction	hsa04070	25	5	20
Signal Transduction	hsa04012	46	10	21.7
Signal Transduction	hsa04340	17	4	23.5
Signaling Molecules and Interaction	hsa04060	244	24	9.8
Signaling Molecules and Interaction	hsa04512	40	4	10
Signaling Molecules and Interaction	hsa04080	90	11	12.2
Signaling Molecules and Interaction	hsa04514	80	13	16.3
Substance Dependence	hsa05031	28	5	17.9
Substance Dependence	hsa05030	26	5	19.2
Transcription	hsa03040	70	9	12.9
Transcription	hsa03022	33	6	18.2
Transcription	hsa03020	28	6	21.4
Translation	hsa03008	53	2	3.8
Translation	hsa03013	94	6	6.4

SUPPLEMENTARY DATA

Translation	hsa00970	23	2	8.7
Translation	hsa03010	76	8	10.5
Translation	hsa03015	49	7	14.3
Transport and Catabolism	hsa04145	58	5	8.6
Transport and Catabolism	hsa04142	61	7	11.5
Transport and Catabolism	hsa04146	61	8	13.1
Transport and Catabolism	hsa04144	79	12	15.2
Transport and Catabolism	hsa04140	14	3	21.4
Xenobiotics Biodegradation and Metabolism	hsa00983	20	3	15
Xenobiotics Biodegradation and Metabolism	hsa00980	16	3	18.8
Xenobiotics Biodegradation and Metabolism	hsa00982	13	3	23.1

SUPPLEMENTARY DATA

Supplementary table 2. Cellular level analysis of KEGG pathways: cell cycle (hsa04110), oocyte meiosis (hsa04114) and progesterone-mediated oocyte maturation (hsa04914)*. **Cohesiveness of genes shared by at least 2 of the 3 pathways. Genes with cohesive barcodes for a given pathway are shown in green. Genes with outlier barcodes are highlighted in red. Genes shown in grey are not present in the pathway.**

KEGG Identifier	Uniprot Identifier	Cell cycle	Oocyte meiosis	Progesterone-mediated oocyte maturation
hsa:9088	PMYT1_HUMAN	1	0	0
hsa:1017	CDK2_HUMAN	0	0	1
hsa:699	BUB1_HUMAN	0	0	0
hsa:5347	PLK1_HUMAN	0	0	0
hsa:995	MPIP3_HUMAN	0	1	-1
hsa:9126	SMC3_HUMAN	0	0	-1
hsa:891	CCNB1_HUMAN	0	0	-1
hsa:9700	ESPL1_HUMAN	0	0	-1
hsa:4085	MD2L1_HUMAN	0	0	-1
hsa:991	CDC20_HUMAN	0	0	-1
hsa:898	CCNE1_HUMAN	0	0	-1
hsa:8454	CUL1_HUMAN	0	0	-1
hsa:64506	CPEB1_HUMAN	-1	1	0
hsa:3480	IGF1R_HUMAN	-1	1	0
hsa:3630	INS_HUMAN	-1	1	0
hsa:4342	CCNB2_HUMAN	-1	0	1
hsa:9133	MOS_HUMAN	-1	0	1
hsa:5604	MP2K1_HUMAN	-1	0	0
hsa:5241	PRGR_HUMAN	-1	0	0
hsa:993	MPIP1_HUMAN	0	-1	0
hsa:51343	FZR_HUMAN	0	-1	1

* The cell cycle and oocyte meiosis pathways are well conserved in most animals. In contrast, the exact nature of oocyte maturation varies in different species, since the females of some species produce thousands of eggs at a time, while in others, females produce relatively few mature eggs[6]. **The oocytes of most animal species arrest during meiotic prophase and complete meiosis in response to intercellular signaling in a process called meiotic maturation. Although the signals are different from species to species (e.g. steroid hormones in frogs and fishes, removal of a follicular inhibitor in mammals), they all activate signaling pathways that converge to the same target: the activation of the universal eukaryotic inducer of M-phase, MPF, a complex formed of the Cdk1 kinase, and Cyclin B[1].**

SUPPLEMENTARY DATA

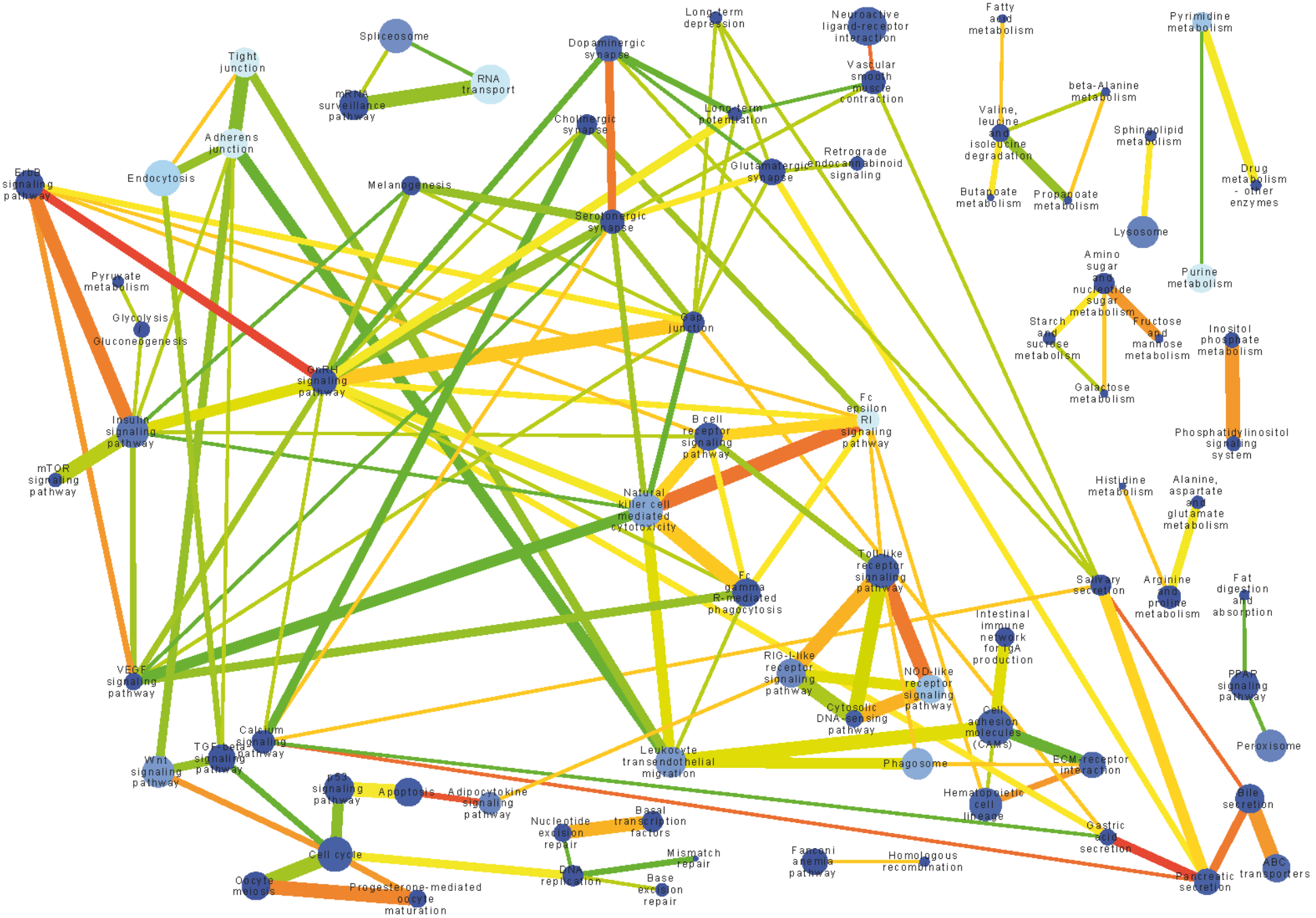
Supplementary table 3: Cellular level analysis of KEGG pathways: Toll-like receptor signaling (hsa4620), NOD-like receptor signaling (hsa4621), RIG-I-like receptor signaling (hsa4622) and cytosolic DNA sensing (hsa4623)*. Cohesiveness of genes shared by at least 2 of the 4 pathways. Genes with cohesive barcodes for a given pathway are shown in green. Genes with outlier barcodes are highlighted in red. Genes shown in grey are not present in the pathway.

KEGG Identifier	Uniprot Identifier	Toll-like	NOD-like	RIG-I-like	DNA sensors
hsa:1147	IKKA_HUMAN	1	0	0	0
hsa:8517	NEMO_HUMAN	0	0	1	-1
hsa:6885	M3K7_HUMAN	1	1	0	-1
hsa:3576	IL8_HUMAN	1	0	0	-1
hsa:7124	TNFA_HUMAN	1	0	0	-1
hsa:841	CASP8_HUMAN	0	0	0	-1
hsa:7189	TRAF6_HUMAN	0	0	0	-1
hsa:3551	IKKB_HUMAN	0	0	0	-1
hsa:3553	IL1B_HUMAN	1	1	-1	1
hsa:6352	CCL5_HUMAN	0	1	-1	0
hsa:5970	TF65_HUMAN	0	1	-1	0
hsa:3569	IL6_HUMAN	0	1	-1	0
hsa:4792	IKBA_HUMAN	0	0	-1	0
hsa:3665	IRF7_HUMAN	0	-1	0	1
hsa:8737	RIPK1_HUMAN	0	-1	1	0
hsa:3627	CXL10_HUMAN	0	-1	0	0
hsa:3661	IRF3_HUMAN,	0	-1	0	0
hsa:29110	TBK1_HUMAN	0	-1	0	0
hsa:9641	IKKE_HUMAN	0	-1	0	0
hsa:3456	IFNB_HUMAN	0	-1	0	0
hsa:10454	TAB1_HUMAN	0	0	-1	-1
hsa:8772	FADD_HUMAN	0	-1	1	-1
hsa:7187	TRAF3_HUMAN	1	-1	0	-1
hsa:6300	MK12_HUMAN	0	-1	0	-1
hsa:29108	ASC_HUMAN	-1	0	-1	0
hsa:834	CASP1_HUMAN	-1	0	-1	0
hsa:3606	IL18_HUMAN	-1	0	-1	0
hsa:340061	TM173_HUMAN	-1	-1	0	0
hsa:57506	MAVS_HUMAN	-1	-1	0	0
hsa:23586	DDX58_HUMAN	-1	-1	0	0

* The innate immune system relies on pattern recognition receptors (PRRs) that recognize different pathogens, such as viruses or bacteria, and that trigger intracellular signaling cascades ultimately culminating in the expression of proinflammatory molecules[7]. Toll-like

receptors are membrane-bound PRRs, located either at the cell surface where they mainly recognize bacterial products, or in intracellular compartments where they are involved in recognition of nucleic acids. Cytosolic PRRs, including RIG-I-like receptors and NOD-like receptors, mainly recognize intracellular RNA. Finally, cytoplasmic localization of DNA by cytosolic DNA sensors seems to be involved in mounting a response to both bacteria and DNA viruses. Three major signaling pathways responsible for mediating TLR-induced responses include nuclear factor kappa-B (NF-κB), mitogen-activated protein kinases (MAPKs), and IFN regulatory factors (IRFs). The RLR pathway involves two different signaling pathways, either NF-κB or IRFs. NOD protein signaling involves activation of NF-κB and MAPK. Regarding cytosolic DNA sensors, strong evidence suggests that these receptors signal to IRFs.

SUPPLEMENTARY DATA



SUPPLEMENTARY DATA

Supplementary References

1 Dupre, A., Haccard, O. & Jessus, C. Mos in the oocyte: how to use MAPK independently of growth factors and transcription to control meiotic divisions. *J Signal Transduct* **2011**, 350412 (2011).

2 Liu, J. & Kipreos, E. T. Evolution of cyclin-dependent kinases (CDKs) and CDK-activating kinases (CAKs): differential conservation of CAKs in yeast and metazoa. *Mol Biol Evol* **17**, 1061-1074 (2000).

3 Moore, N. L., Narayanan, R. & Weigel, N. L. Cyclin dependent kinase 2 and the regulation of human progesterone receptor activity. *Steroids* **72**, 202-209 (2007).

4 Malumbres, M. & Barbacid, M. Mammalian cyclin-dependent kinases. *Trends Biochem Sci* **30**, 630-641 (2005).

5 Clark, B. D., Collins, K. L., Gandy, M. S., Webb, A. C. & Auron, P. E. Genomic sequence for human prointerleukin 1 beta: possible evolution from a reverse transcribed prointerleukin 1 alpha gene. *Nucleic Acids Res* **14**, 7897-7914 (1986).

6 Vasudevan, S., Seli, E. & Steitz, J. A. Metazoan oocyte and early embryo development program: a progression through translation regulatory cascades. *Genes Dev* **20**, 138-146 (2006).

7 Mogensen, T. H. Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin Microbiol Rev* **22**, 240-273, (2009).