DATABASE
The Journal of Biological Databases and Curation

Database, 2018, 1–15
doi: 10.1093/database/bay036
Database tool

Database tool

Signalling maps in cancer research: construction and data analysis

Maria Kondratova[1,2,3], Nicolas Sompairac[1,2,3], Emmanuel Barillot[1,2,3], Andrei Zinovyev[1,2,3] and Inna Kuperstein[1,2,3,*]

[1]Institut Curie, PSL Research University, F-75005 Paris, France, [2]INSERM, U900, F-75005 Paris, France and [3]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France

*Corresponding author: Tel: +33 (0) 1 56 24 69 87; Fax: +33 (0) 1 56 24 69 11; Email: inna.kuperstein@curie.fr

Citation details: Kondratova,M., Sompairac,N., Barillot,E. *et al.* Signalling maps in cancer research: construction and data analysis. *Database* (2018) Vol. 2018: article ID bay036; doi:10.1093/database/bay036

Received 3 November 2017; Revised 7 February 2018; Accepted 19 March 2018

Abstract

Generation and usage of high-quality molecular signalling network maps can be augmented by standardizing notations, establishing curation workflows and application of computational biology methods to exploit the knowledge contained in the maps. In this manuscript, we summarize the major aims and challenges of assembling information in the form of comprehensive maps of molecular interactions. Mainly, we share our experience gained while creating the Atlas of Cancer Signalling Network. In the step-by-step procedure, we describe the map construction process and suggest solutions for map complexity management by introducing a hierarchical modular map structure. In addition, we describe the NaviCell platform, a computational technology using Google Maps API to explore comprehensive molecular maps similar to geographical maps and explain the advantages of semantic zooming principles for map navigation. We also provide the outline to prepare signalling network maps for navigation using the NaviCell platform. Finally, several examples of cancer high-throughput data analysis and visualization in the context of comprehensive signalling maps are presented.

Introduction

Similar to geographical maps, the representation of biological knowledge as a diagram facilitates the study of complex processes in the living cell, in a visual and insightful way.

Goal of knowledge formalization

The representation of biological processes as comprehensive signalling network maps has three major goals: (i) to generate a resource containing a formalized summary of biological findings from many research groups, (ii) to provide a platform for sharing information and discussing biological mechanisms and (iii) to create an analytical tool useful for high-throughput data integration and analysis.

It can be helpful to systematically represent and formalize the molecular information distributed in thousands of scientific publications. An additional advantage of representing biological processes in a graphical form is to

capture the multiple cross-talks and interactions occurring between different cell processes (1).

Analysis and visualization of omics data in the context of signalling network maps can help to detect patterns in the data projected onto the molecular mechanisms there represented. For instance, identification of deregulated mechanisms and key players in human diseases have a direct clinical application (2, 3). Moreover, correlating the status of those deregulated mechanisms with patient survival helps for patient stratification according to their network-based signatures (4). Due to the complexity of mechanisms simultaneously involved in diseases, targeting combinations of molecular players is now the trend in treatment of complex diseases. The computational approaches using signalling maps allow testing multiple combinations *in silico*, considering large comprehensive signalling networks and omics data (5, 6). In addition, signalling networks can serve for modelling and prediction of cell fate decisions (7, 8) and suggestion of non-intuitive combinations of gene perturbations to explain phenotypes in health and disease (9).

To achieve these goals, the construction of a signalling map should become an accessible procedure that can be completed in a reasonable time. There are several solutions for biological knowledge formalization briefly described in this manuscript. We contribute to this global aim and formulate the main principles and steps of the established workflow for manual map construction. In addition, we suggest the biological network map navigation facilitated by Google Maps technology and provide examples of data analysis and visualization in this context.

Diagram types for molecular processes representation

Generally speaking, there are four main approaches (or diagram types) for representing molecular processes, each of them characterized by a certain depth of description: (i) *interaction diagram*, which shows simple binary relations between molecular entities; (ii) *activity-flow*, known as regulatory network or influence diagrams, representing the flow of information or influences of one entity on another; (iii) *entity relationship diagram*, depicting relations in which a given entity participates; and (iv) *process description (PD) diagram*, known in chemical kinetics as bipartite reaction network graphs (10).

Pathways and network maps approached for molecular processes representation

Using the aforementioned approaches of molecular processes representation, several pathway databases have emerged (11). They serve as biological knowledge information resource and as computational analytical tools for systems-based interpretation of data. A significant number of pathway databases has been also developed in the private domain, but the majority of pathway collections are free and open source (Supplementary Table S1).

The most common way of knowledge representation, used in the majority of these resources, is depicting separate processes referring to 'signal transduction pathways'. However, drawing individual pathways precludes clear representation of cross-regulations among biological processes. The alternative solution is to create seamless maps of biological mechanisms covering multiple cell processes simultaneously, similar to geographical maps. In order to apply this 'geography-inspired' approach to biological knowledge, it is necessary to address a number of challenges related to generation, maintenance and navigation of large signalling network maps. We discuss these challenges and suggest solutions based on our long-term experience manipulating biological maps.

Common standards for molecular processes representation

With the aim of creating a collection of exchangeable comprehensive signalling maps, common rules for drawing maps and standard graphical syntax should be developed and consistently applied. The current suggested solution in the field is the Systems Biology Graphical Notation (SBGN) syntax. This syntax is compatible with various pathway drawing and analytical tools, allowing to represent not only biochemical processes but also cell compartments and phenotypes (10). Furthermore, to increase cross-compatibility between pathway resources and analytical tools, several common formats for exchanging information on molecular interactions, such as BioPAX, SBML, PSI-MI and so on, have been suggested (7).

Tools for molecular processes representation

Since the generation of signalling maps is a long and laborious process, the choice of the appropriate drawing tool, best suited for the type of network, has to be thoughtfully done. There are several free and commercial tools to create biological network diagrams that differ in the process representation approach, syntax and requirement for the end users' technical skills, as SBGN-ED (12), visANT (13), CellDesigner (14) and so on (Supplementary Table S2).

Navigation platforms for comprehensive molecular network maps

Visualizing and exploring biological network diagrams became an important issue, because size and complexity of

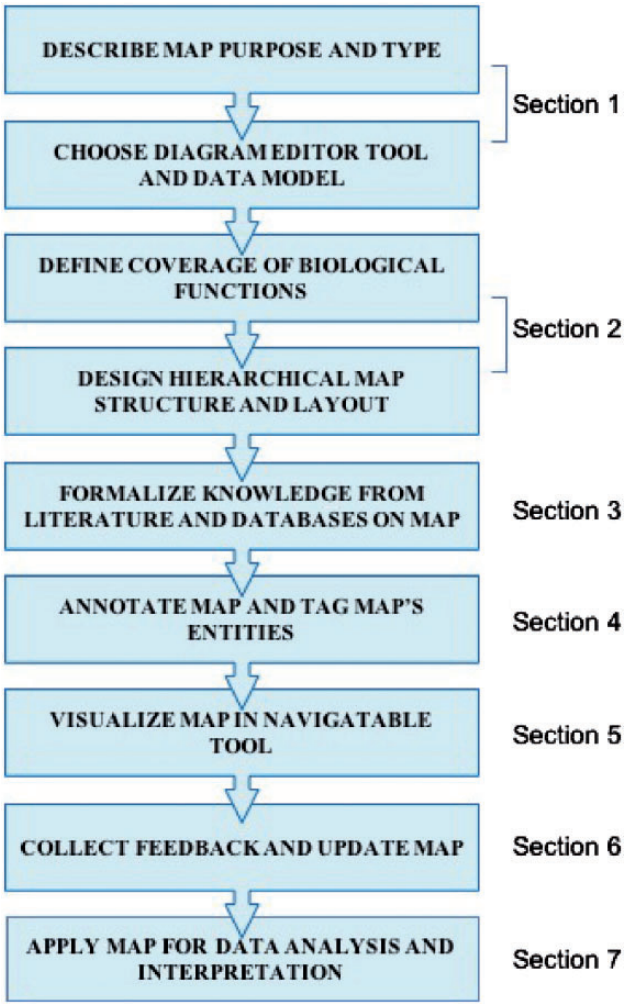

Figure 1. Map construction workflow scheme.The corresponding section in the texts is indicated.

molecular networks reach the dimensions of modern geographic maps. Therefore, several tools, such as CellPublisher (15), Pathways projector (16) MINERVA (17), NaviCell (18, 19) and so on, have adopted the navigation logic from Google Maps technology. These tools allow exploring big networks in a user-friendly manner thanks to such Google Maps features, as scrolling, zooming, markers and callouts (Supplementary Table S2).

A workflow for construction of comprehensive signalling network maps

In this manuscript, we describe a set of good practices for building comprehensive signalling network maps. We provide the methodology that allows to overcome challenges associated with construction, navigation and exploration of large molecular interaction maps (Figure 1). We suggest an approach that is neither unique nor universal but provides verified practical solutions to comprehensive map construction and manipulation that successfully served to generate the ACSN resource (20) and also applied in other studies (9, 21).

Each section of the workflow starts with a challenge statement followed by a good practice solution, implemented in a typical example (Figure 1). The workflow covers the following steps: (i) defining the aim, coverage of biological knowledge and structure of the signalling map; (ii) literature mining, signalling map drawing and annotation in CellDesigner tool using SBGN-like syntax (14); (iii) map preparation in NaviCell format and navigation modes using Google Maps-based NaviCell tool (18); (iv) high-throughput data visualization on top of the signalling maps using NaviCell Web Service (19) and NaviCom (22).

The suggested workflow is in line with the Findable, Accessible, Interoperable, Re-usable (FAIR) principles for data management and sharing (23).The access to the detailed procedures is provided in the Supplementary material and available at https://github.com/sysbio-curie/NaviCell. The terms and definitions used through the paper are listed in Table 1.

The DNA repair map from ACSN resource is used as an example for map construction and annotation:

> https://acsn.curie.fr/navicell/maps/dnarepair/master/index.html (20).

Epithelial-mesenchymal transition (EMT) regulation map from NaviCell collection (9) is used for data visualization:

> https://navicell.curie.fr/pages/signalling_network_emt_regulation_description.html

General principles for map construction

Work organization

Signalling map construction requires an overview of broad scientific literature and a very meticulous work for correct representation of molecular processes in great detail. Therefore, several important decisions should be made prior to the map construction: (i) What is the purpose for constructing the map? (ii) What diagram type is suitable to properly represent the knowledge? (iii) What is the appropriate tool to build the map? (iv) What processes should be included in the map? (iv) How the map will look like? Once those questions are answered, and an agreement on the approach has been reached, the signalling diagram construction should follow fixed principles to ensure generation of a homogeneous and accurate map. An additional important step before constructing a map is to consult similar efforts in the field and evaluate the added value of a new map.

Table 1. Terms and definitions

Graphical standards and exchange formats terms and definitions	
SBGN	Systems Biology Graphical Notation (SBGN) is a standard graphical syntax for representation of biological processes and interactions. SBGN is compatible with multiple pathway drawing and analytical tools, http://sbgn.github.io/sbgn/
SBML	Systems Biology Markup Language (SBML) is a representation format, based on XML, for communicating and storing computational models of biological processes. It is a free and open standard language with widespread software support, http://sbml.org
Standard identifier (ID)	Community-accepted nomenclature for scientific naming of biomolecules as genes, proteins, chemicals, drugs and so on. The sources for standard IDs are repositories as UNIPROT, CHEB and HUGO, http://identifiers.org
Data and models exchange formats	Standard formats for data and models to facilitate networks and software intercompatibility. There are two major standard networks exchange formats, BIOPAX for complex networks and SIF for simple binary interactions. The CellDesigner xml format is commonly used exchange format compatible with multiple network analysis tools.
Signalling network map terms and definitions	
Map (in ACSN)	Diagram of detailed molecular interactions with meaningful layout reflecting a certain biological process, which is graphically represented in CellDesigner tool and converted to NaviCell format for exploration and curation.
Map layer (in ACSN)	Map area covering several molecular processes with similar functions.
Map module (in ACSN)	Part of the map representing a sequence of molecular interactions responsible for execution of a particular function.
Signal transduction pathway or sigalling cascade	Sequence of molecular interaction that transforms extracellular signals into an intercellular activity.
Map entity (in ACSN)	Component of the map graphically depicted using SBGN standards in CellDesigner tool.
Hierarchical structure of a map	System for signaling network map complexity reduction by dividing it into **hierarchically organized units** as layers meta-modules, modules that exist as independent submaps. The structure supports the **horizontal map navigation mode**.
Confidence score (ACSN)	Value representing the measure of accuracy of binary interactions on the map. There are two confidence scores in NaviCell maps. The **reference score** indicates number and 'weight' associated with the publications found in the annotation of a given reaction. The **functional proximity score** is computed based on the external network of PPI, HPRD.
NaviCell terms and definitions	
Semantic zoom	A mechanism providing several map views with different levels of details depiction achieved by gradual exclusion of details while zooming out. It simplifies navigation through large maps of molecular interactions by providing several levels of details, resembling navigation through geographical maps. Exploring the map from a detailed toward a top-level view is achieved by gradual exclusion and modification (simplification and abstraction) of details. One of the main principle of semantic zooming is in that every detail that is shown on the map at a current zoom level, should be readable. This feature supports the **vertical map navigation mode**.
Bird-eye view panel	Window containing top-level view of the map with indication on currently centered area; adapted from Google maps.
Zooming bar	Zoom control slider; adapted from Google maps.
Marker	Symbol indicating location of chosen objects on the map; adapted from Google maps.
Pop-up bubble	Small window that opens by clicking on marker. Contains short description and hyperlinks related to the marked entity.
Annotation post	Detailed map entity annotation created in CellDesigner by map manager. The annotation is converted to annotation post and displayed in the associated blog by NaviCell.

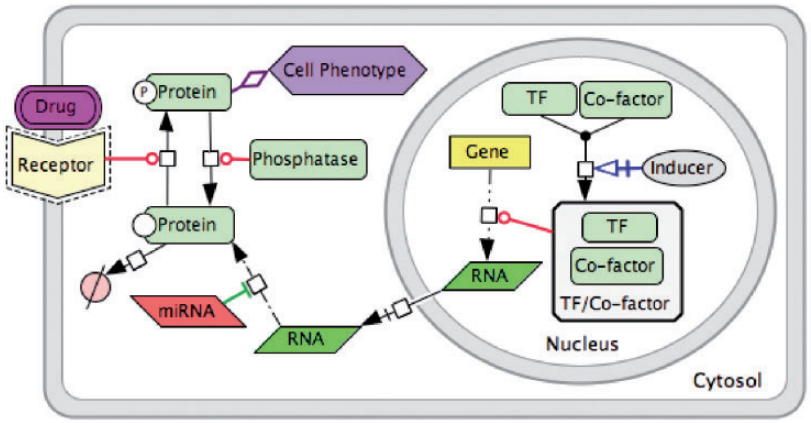

Figure 2. Data model for PD diagram using CellDesigner SBGN syntax.

Map purpose and type

Referring to our example (DNA repair map), the purpose was to understand how different types of DNA damage are repaired in the cell and how these various DNA repair mechanisms and the cell cycle coordinate. To address this challenge and gather together these molecular mechanisms, we decided to construct a comprehensive map of DNA repair and cell cycle signalling. The map can be applied to **detect the modes of DNA repair machinery rewiring** in different pathological situations such as cancer or genotoxic stress.

The mechanisms of DNA repair are well studied and information on involved molecules and regulation circuits is available. Thus, to preserve and accurately depict the processes in its whole complexity, we have chosen the **PD diagram type**, where the biochemical reactions can be explicitly depicted.

Map construction tool, graphical standard and data model

Signalling processes of DNA repair are represented as biochemical reactions in CellDesigner diagram editor based on the **standard** SBGN **syntax** (10), and the **Systems Biology Markup Language (SBML)**, for further computational modelling of the map (14). The **data model** used in our example is schematically depicted in Figure 2. It includes such molecular entities as proteins, genes, RNAs, antisense RNAs, simple molecules, ions, drugs, phenotypes and complexes. Edges on the map represent biochemical reactions or reaction regulations including post-translational modifications, translation, transcription, complex formation or dissociation, transport, degradation and so on. Reaction regulations are catalysis, inhibition, modulation, trigger and physical stimulation. It is also possible to depict cell compartments such as cytosol, nucleus, mitochondria and so on. See http://celldesigner.org for Cell Designer tool guide and http://www.sbgn.org for SBGN syntax explanation.

Map boundaries, layout and structure

Map boundaries and content

Given the limitations of graphical tools and difficulties in manipulating large molecular interaction maps, the challenge is to define map boundaries. The solution that we suggest is to assign **one biological function per map** (e.g. cell cycle, angiogenesis and immune response). This is challenging per se due to 'fuzziness' of borders between processes and overlaps between players and pathways. Therefore, biological function-driven maps can be assumed as components of a **global atlas** merged together via common players. To allow such combinable maps, common **standards for graphical representation** and **standard common identifiers** for naming the entities should be used through all maps. In addition, the community-based curation of maps is crucial for making more objective decisions regarding the map boundaries and content.

In our example, the boundaries and the content of the DNA repair map were defined in coordination with specialists in the corresponding fields. According to the seminal reviews and well-known databases, DNA repair machinery distinguishes 10 partially overlapping modes of repair, depending on the type of damage. The DNA damage types are depicted on the map as 'inputs' initiating the corresponding DNA repair mechanisms. In addition, the map covers four cell cycle phases and depicts regulatory circuits between cell cycle and DNA repair mechanisms coordinated by cell cycle checkpoints (24) (Figure 3A).

Map layout and hierarchical modular structure

The challenge of map dimensions and layout design should be addressed prior to the map drawing. This step is especially crucial in the case of collective map reconstruction where the final global map is assembled from a number of sub-maps generated by different map curators.

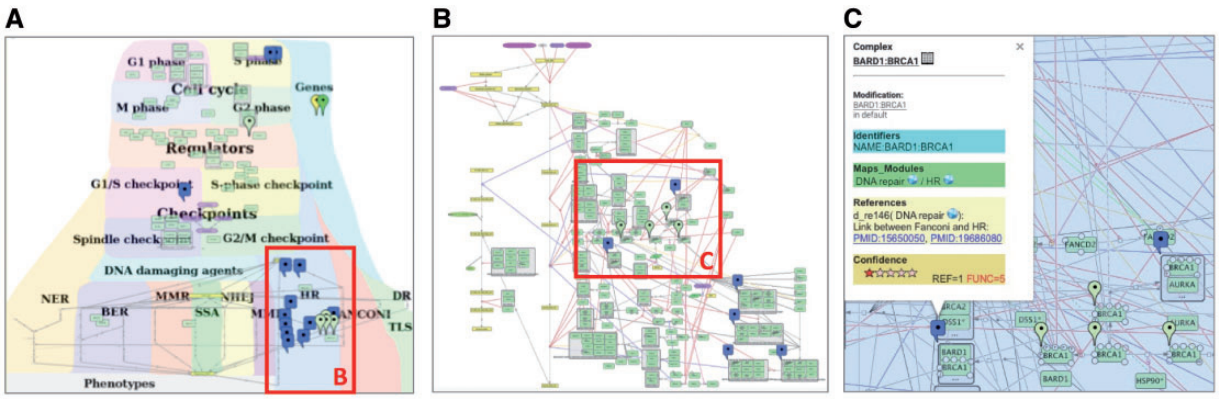

Figure 3. DNA repair map in NaviCell format. (**A**) Global layout of the map with maprker indicating BRCA1 protein distribution, (**B**) Individual module layout H HR, (**C**) Callout window with annotation of BARD: BRCA1 complex.

The aim of signalling map construction is not only to summarize molecular mechanisms but also to allocate the processes in a meaningful and biologically relevant way. Careful design of the **signalling map layout** helps for intuitive understanding of 'what is where' and 'what is close and what is distant'. In addition, a bird's eye view on the map can give a general impression about the map complexity.

There are at least three options of map layout design. (i) Representing **spatial localization of processes** in the context of the global cell architecture. Most of the signal transduction maps resemble an accepted view of **cell organization**, they include cellular compartments and signalling pathways are placed in the corresponding compartments in the map (see examples at https://acsn.curie.fr). (ii) Depicting **process propagation in time**, for instance, demonstrating propagation of the signalling through the four phases of cell cycle (https://acsn.curie.fr/navicell/maps/cell cycle/master/index.html). (iii) Placing processes together according to their **involvement in a particular biological function** as DNA repair, where each pathway is depicting one biological function. These pathways are allocated next to each other, creating together a DNA repair machinery signalling (Figure 3A).

The combination of layout types at the same map is also possible as in the case of the DNA repair map, combining the three aforementioned layout types (Figure 3A and https://acsn.curie.fr/navicell/maps/dnarepair/master/index.html).

The global layout of the map can be also used as guidance to split it into sub-maps of layers and functional modules and in generating a **hierarchical modular structure of the map** (see Table 1 for terms and definitions).

In our example, the **global map layout** of the DNA repair map has been designed to emphasize the **hierarchical organization** of three major layers in the map: the DNA repair machinery layer and its connection to the cell cycle layer via the checkpoints layer. The upper layer depicts the cell cycle, the middle layer represents cell cycle checkpoints, receiving signals from the cell cycle and coordinating the crosstalk between the cell cycle and the DNA repair machinery, represented in the lower layer. This hierarchical organization of the map in layers reflects the current understanding of signal propagation in the cell.

Further, the map can be divided into functional modules. Each functional module can exist as a part of the global map but also as an independent map. Exploring the module maps together within the global map can be supported by Google Maps-based map navigation (discussed in the Section 'Map generation and navigation in NaviCell'). The DNA map has a **modular structure** composed of 18 interconnected functional modules, corresponding to ten DNA repair mechanisms, four phases of the cell cycle and four checkpoints, all interconnected, with multiple regulatory circuits (Figure 3A).

Due to the complexity of comprehensive networks, members of the same functional module can be very distantly located in the map. For example, a molecule that participates in several functional modules can have a geographical location in one module, but its location with respect to other modules can be far from optimal. To cope with this problem, we suggest creating a separate map for each module with an **individual module map layout**. The size of the module map is smaller compared to the global comprehensive map, and its contents are limited to the processes participating in the given module. This allows creating a layout where all entities of a single module map are closely located. This manipulation results in a different layout of the module, comparing its design at the comprehensive map, however it does not change the backbone structure of the module. An example of a module map with optimized individual layout is the homologous recombination (HR) from the DNA repair map, shown in Figure 3B. For modular map

generation instructions see https://github.com/sysbio-curie/NaviCell.

Similarly, it is possible to generate **automatic layouts** for global, module maps or even for maps of an individual entity. The maps of an individual entity can represent 'life-cycle' reflecting reactions where the entity participates and the regulators of these reactions. This can be performed in Cytoscape software using the BiNoM plugin using the modularization and automatic layout functions (25). To facilitate the integration of separated signalling diagrams, there are at least two methods for map merging: (i) Merge Model plugin in CellDesigner (14) and (ii) BiNoM plugin on Cytoscape, which allows to reorganize, dissect and merge disconnected CellDesigner pathway diagrams (25). For the map merging procedure in BiNoM see https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf.

Data extraction and representation

Manual data mining

Retrieving knowledge and evidences from scientific papers, followed by a suitable graphical representation, requires a clear understanding of the biological processes and the experimental methodologies. Molecular interactions can be validated by various experimental methods. The most common and reliable experimental methods confirming different types of molecular interactions in the cell are summarized in Supplementary Table S3.

Depending on the purpose on the map, one can aim to represent biological processes with great precision including post-translational modifications, transport, complex association, degradation and so on. Phenotype nodes on the signalling maps normally serve to indicate signalling readouts or cell statuses or biological processes in general. This type of nodes can also serve for schematic representation of statements when the exact molecular mechanism is still unknown. Some details on cell signalling might be skipped or, on the contrary, represented rigorously, depending on the purpose of the map being drawn, and the opinion of the map creator. It is important to address the challenge of **persistent homogeneity** on how to present the information on the map. This is especially important for visualization and exploration of the map, because it will ensure a correct stepwise appearance of details at different zoom levels on the Google Maps-like map visualization platform (discussed in the Section 'Common exchange formats').

For efficient map construction, we suggest to perform a **systematic literature revision**. The hierarchical organization of the map, actually reflects a suggested approach for literature curation. First, we recommend to use **seminal review papers** in the field and the major **pathway databases** (Supplementary Table S1), in order to define the boundaries and general contents of the map. The information on canonical pathways retrieved from these sources, reflects the consensus view of the field and can serve as basis for drawing the backbone structure of the map. Thus, further details on molecular mechanisms can be compiled from the most recent studies and added to the map. The good practice is to comply with the requirement of supporting every interaction included in the map by at least **two independent investigations**.

In addition, the interpretation of scientific text and translation of written information into the correct and meaningful diagram is not always obvious. For consistency from text to diagram translation, we developed major rules for standardized interpretation of statements, see https://github.com/sysbio-curie/NaviCell. Figure 4 represents an example of scientific text translation into the PD diagram type.

Automatic data mining

The alternative approach to manual curation of scientific literature, is a wide range of automated or semi-automated text mining techniques (26). For example, the BioCreative initiative, an international community-wide effort for evaluating text mining and information extraction systems applied to the biological domain (http://www.biocreative.org). BioCreative gathers together multiple scientific teams and provides a list of text mining methods. In addition, it suggests common standards for text mining (27).

There are several automated text mining algorithms allowing to retrieve statements from scientific texts and to convert them into a meaningful graphical diagram. The advanced INDRA (Integrated Network and Dynamical Reasoning Assembler) approach allows building computational models directly from natural language using automated assembly of network (28). The HiPub approach translates PubMed and PMC texts to networks of knowledge (29). The ANDSystem: an Associative Network Discovery System for automated literature mining, distinguishes and reconstructs in a form of networks several types of interactions as genetic, protein, metabolic and so on. (30). Finally, the Information Hyperlinked over Proteins system, uses genes and proteins as hyperlinks between sentences and abstracts and converts the information from PubMed into one navigable resource for scientific literature investigation (31).

Among others, there is an approach of **formalized text-based intermediate language** as Biological Expression Language that structures the statements in a way that can be mechanistically converted into a meaningful biological network (32). Similarly, BiNoM, a Cytoscape plugin, provides a function that converts sentences formulated in the 'human' language, into a set of formal statements describing reactions that in turn can be automatically represented

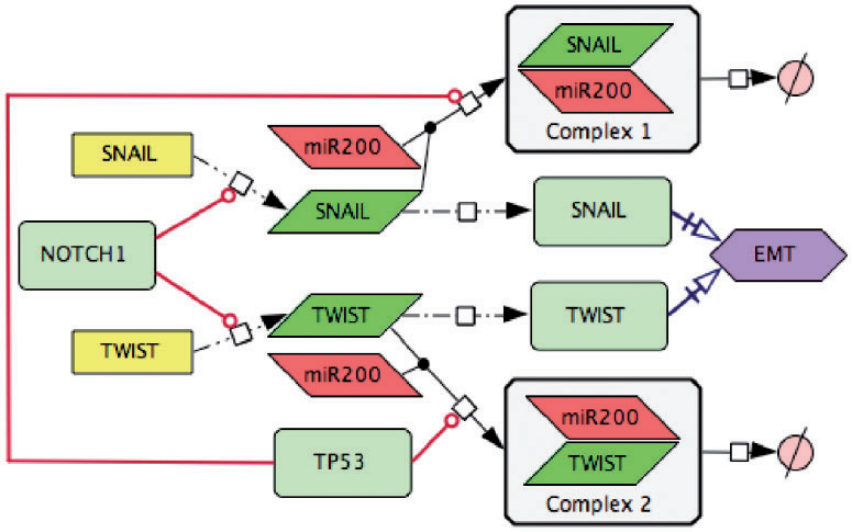

Figure 4. Transforming text to diagram: role of p53 and NOTCH in induction of EMT. The following statements were used for diagram construction: (1). Control of EMT program is performed by SNAIL and TWIST, the major transcription factors that can induce the executors EMT program (49). These transcription factors are under the control of several upstream mechanisms. (2) SNAIL and TWIST are inhibited by the p53 protein via a variety of microRNAs, including miR200 (50) (3) miR20 binds to the mRNAs of SNAIL and TWIST and triggers their degradation, this way preventing the translation of mRNAs into the corresponding proteins (51) (4) EMT program can be initiated due to excessive expression of NOTCH that directly activates transcription of SNAIL and TWIST (52).

as graphical diagrams. The statements can be prepared in any simple text-based editor and imported into Cytoscape through BiNoM and converted into the CellDesigner diagram as presented in Supplementary Figure S1. The syntax of this BiNoM Reaction Format language is depicted in (21, 25) and in the BiNoM manual https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf.

Map annotation

Map entities annotation

Consistent integration of **standard common identifiers (ID)** for map entities ensures compatibility of maps with other tools and facilitates data integration into the maps. A common entity annotation form is needed for efficient map exploration by users and for map cross-curation by specialists in the corresponding fields. To address the challenge of systematic and user-friendly entity annotations, we developed the **NaviCell annotation format**. The entity annotation is normally performed by the map creator during the map construction in CellDesigner (Figure 3C).

We suggest to structure the annotation panel in four sections: 'Identifiers', 'Maps_Modules', 'References' and 'Confidence'. We provide an example of BRCA1 entity annotation from the DNA repair map in Figure 3C and Supplementary Figure S2.

The **section 'Identifiers'** includes standard IDs and provides links to the corresponding pages in HGNC, UniProt, Entrez, SBO, GeneCards and cross-references in REACTOME (33), KEGG (34), Wiki Pathways (35) and so on. Metabolites and small compounds are annotated by corresponding IDs and linked to ChEBI (36), PubChem Compound (37) and KEGG Compound (38) databases.

The **section 'Maps_Modules'** includes links to modules of the map where the entity participates. Referring to our example, the DNA repair map is part of Atlas of Cancer Signalling Network (ACSN) resource, and multiple entities from the DNA repair map can participate in various maps in the resource. There is a challenge to demonstrate this multi-functionality of an entity and there is a need to provide links to all maps and modules of ACSN where an entity participates. To achieve this, we introduced a tagging system, included into the entity annotation. The tags with map and module names where the entity participates are systematically included into the annotation during the map construction in CellDesigner.

The **section 'References'** provides links to the corresponding publications from which the evidences on the entity or its reactions were collected. In addition, map creator's notes can be added in this section, in order to emphasize or to highlight important information related to the entity or its reactions. Each entity annotation is represented as a web page (Supplementary Figure S2) automatically generated when the NaviCell map is generated from the CellDesigner file (described in the Section '**Generation of NaviCell map using NaviCell factory**'). The extended description of annotation formats for each type of entities

in the map is provided in the documentation, see https://github.com/sysbio-curie/NaviCell.

Confidence scores

Despite that multiple network and pathway resources are believed to summarize the current knowledge on biological processes, there is always a remaining open question: to what extent the retrieved statements from the scientific literature and depicted in the diagrams, are indeed reflecting the biological reality? There is a permanent problem that map creators need to address: what processes should be included in the map diagrams? The challenge is how to systematically evaluate the experimental evidence from scientific publications and assign a confidence score. There is no consensus about the approach to evaluate the 'truth' or the 'degree of confidence' regarding to molecular mechanisms reported in the literature.

Here, we provide several examples of confidence scores developed in well-established signalling networks resources. The scores in the BioGRID resource (39) (https://thebiogrid.org) are calculated taking into account if it is an original publication describing the interaction following the principle of CompPASS score system that uses unbiased metrics to assign confidence measurements to interactions from parallel nonreciprocal proteomic data sets (40). These scores are ranged from 0 to 1, with 1 being the most confident. The SIGNOR database (41) (SIGNOR http://signor.uniroma2.it) uses functional relevance of interactions provided in the form of a 'reliability score'. This score is defined by the probability that two partners in the interaction are cited together in the same paper. This co-citing probability is ranged from 0 to 1. In the RECON database (42) (https://vmh.uni.lu), each reaction in the network is associated with a confidence score ranging from 0 to 4. This score is based on the type of evidence identifying the reaction, assigning values for modelling (score 1), sequence (score 2), physiological (score 2), genetic (score 3) and for biochemical (score 4) studies. If multiple evidence types exist for the same interaction, a cumulative confidence score is assigned.

To evaluate the reliability of depicted molecular interactions in the ACSN resource, there are **two confidence scores** for protein complexes and reactions. The scores are automatically calculated during the conversion of the CellDesigner map to the NaviCell format and provided in the annotation's **section 'Confidence'**, in a form of a five-star diagram (Figure 3C and Supplementary Figure S2). Both scores are integers, ranging from 0 (undefined confidence) to 5 (high confidence).

The reference score, marked by 'REF' indicates both the number and the 'weight' associated with publications found in the annotation of a given reaction, with weight equal one point for an original publication and three points for a review article. For example, a reaction annotated by two reviews will obtain the highest reference score REF = 5. The functional proximity score, marked by 'FUNC' is computed based on the external network of protein–protein interactions (PPI), Human Protein Reference Database (HPRD) (43). The score reflects an average distance in the PPI graph between all proteins participating in the reaction as reactants, products or regulators. Therefore, if all the proteins participating in a reaction interact directly (functional distance 1) in the PPI network, then this reaction obtains the highest score (FUNC = 5). If in the reaction where phosphorylation of protein A is catalyzed by protein B, and A and B are separated in PPI graph by 2 (or 3, or 4) edges, then the reaction obtains the functional proximity score FUNC = 4 (or FUNC = 3, or FUNC = 2, respectively). If two reaction participants are not connected in the PPI network, then FUNC = 0. The functional proximity is computed using BiNoM, Cytoscape plugin (25).

Map generation and navigation in NaviCell

Generation of NaviCell map using NaviCell factory

CellDesigner maps, as the DNA repair map, annotated in the NaviCell format or not, can be converted into a NaviCell web-based front-end, which represents a set of HTML pages with embedded JavaScript code. It can be launched in a web browser locally or located on a web-server. The NaviCell factory is currently embedded in the BiNoM Cytoscape plugin but will be soon available as a stand-alone command line package. The detailed guide to use the NaviCell factory is provided at https://github.com/sysbio-curie/NaviCell.

Common exchange formats

The generated maps in CellDesigner and exposed in the NaviCell format, can also be provided in **common exchange formats,** as BioPAX and PNG formats. In addition, the modular composition of the maps can be provided in the form of GMT files. The description of map preparation in various formats using the BiNoM Cytoscape plugin is available in https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf.

Map navigation using NaviCell platform

The problem of navigation and exploration of modern digital interactive geographical maps is successfully addressed

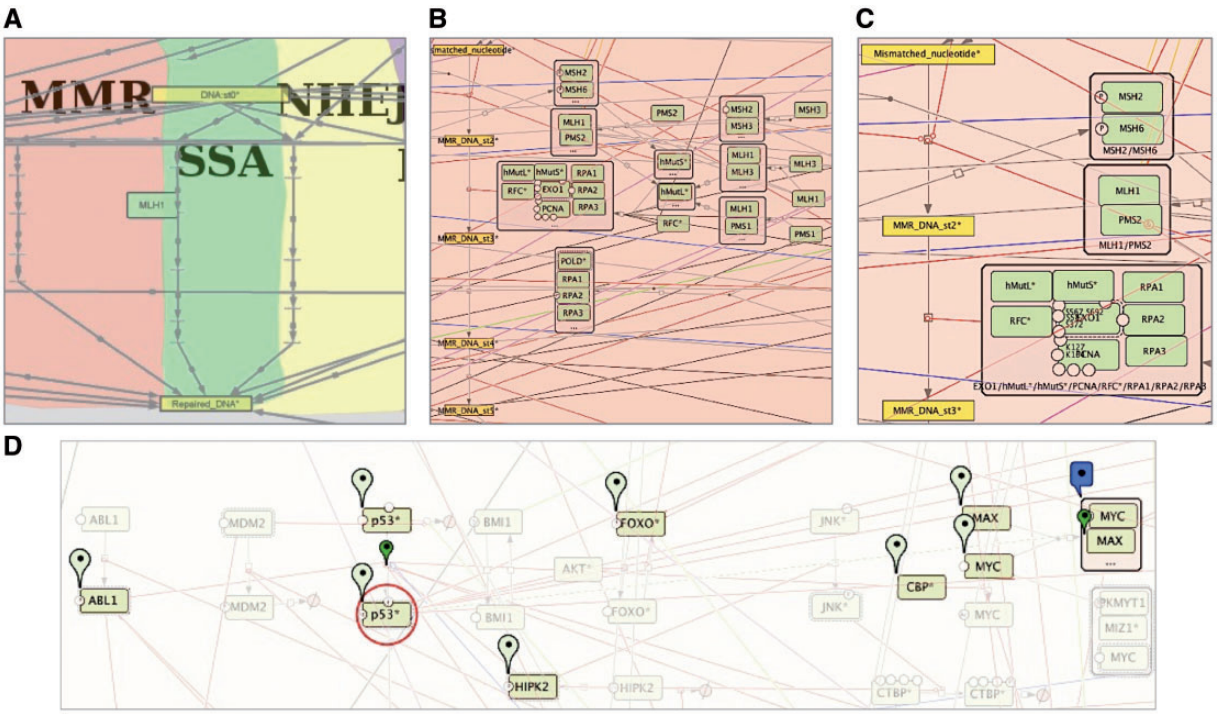

Figure 5. Semantic zooming and entity visualization on DNA repair map. (**A**) Canonical pathways view, (**B**) hide-details view, (**C**) detailed view, (**D**) highlighting p53 neighbours.

in Google Maps or similar systems. Large molecular interaction maps are comparable in size and complexity to geographical maps, therefore the navigability challenge can be addressed in a similar manner.

Comprehensive signalling maps, as the DNA repair map, contain a large number of nodes and edges, leading to a complex map navigation. To solve this problem, maps can be represented as clickable web pages with a user-friendly interface. We developed the NaviCell web-based environment, empowered by Google Maps engine for visualization and navigation of molecular maps (18). Scrolling, zooming, markers, callout windows as well as a zoom bar, are adopted from the Google Maps interface (Figure 3 and Supplementary Figure S3). The map in NaviCell is interactive and all map components are clickable. To find an entity of interest or querying for a single or multiple molecules is possible using the search window. Alternatively, the entity can be found in the selection panel or directly on the map.

Map navigation modes

The NaviCell platform is convenient for several map navigation modes. The so-called **horizontal navigation** is facilitated by the hierarchical modular structure of the maps, which is successfully exploited by NaviCell system. The modular structure of maps allows to transfer from the global map to the module maps. Thus, using our example of the DNA repair map, the involvement of BRCA1 protein across processes on the map can be appreciated (Figure 3A). However, in order to understand the details of the biochemical reactions where BRCA1 plays a role in a particular process as HR, the modular map is more suitable (Figure 3B).

The so-called **vertical navigation** is facilitated by the semantic zooming feature of maps prepared in the NaviCell format. **Semantic zooming** (see Table 1 for terms and definitions) simplifies the navigation through the large maps of molecular interactions, showing a readable amount of details at each zoom level. This gradual detail appearance permits to explore the map contents from the top-level toward a detailed view. To prepare maps for this type of navigation, a map pruning is performed, eliminating hence non-essential information for each zoom level. Typically, it is recommended to generate three to four zoom levels, although the number of zoom levels is unlimited in NaviCell (Figure 5).

In our example of the DNA repair map, the top-level zoom displays the modules of the map, depicted with a coloured background shapes with the backbone structure of the pathways (Figure 5A). In the next zoom level, canonical cell signalling pathways are shown. These pathways are defined by intersecting the content of the map with the corresponding pathways in other databases (Figure 5B), at the last zoom all map elements are present (Figure 5C). For detailed instructions on map zoom levels creation, see https://github.com/sysbio-curie/NaviCell.

An additional way for map exploration is **highlighting individual entities** of the map and their **neighbours**. We developed a NaviCell function for selecting species of interest and its neighbours. This function allows step-wise enlarging of the neighbourhood coverage to understand the signalling propagation on the map, as shown for p53* molecule on the DNA repair map, in Figure 5D.

Map maintenance and curation

Biological knowledge about the majority of signalling pathways is not yet solid and grows continuously. Due to this, one of the major problems of signalling maps is their fast obsolescence. To address this challenge, permanent maintenance and map updating are needed. The community of users is the most reliable and trustable contributor to map maintenance, because specialists can support and update maps from their own research area. To enable such a community-based effort, efficient curation tools should be created. To our knowledge, there is only one community curation tool for comprehensive maps, the Payoa plugin of CellDesigner (44).

We suggest carrying out map curation in the context of NaviCell environment via a blog. The process of map curation and maintenance in NaviCell involves map managers that check the posts of the maps in the blog and latest scientific literature and update the maps accordingly. An automated procedure supports the map updating and archives older versions of posts, including comments, providing thus traceability of every map changes and all blog discussions (18).

Visualization of omics data in the context of signalling network maps

To make data visualization a straightforward and easy task, we developed a **built-in toolbox for visualization and analysis of high-throughput data** in the context of comprehensive signalling networks. The integrated NaviCell web-based toolbox allows importing and visualizing heterogeneous omics data on top of the maps and to perform simple functional data analysis. Standard common IDs in the annotation of proteins, genes (HUGO) and metabolites (CHEBI) allows the NaviCell data visualization functionality. NaviCell also computes aggregated values for sample groups and protein families. For visualization, the tool contains standard heatmaps, barplots and glyphs, as well as the map staining technique for displaying large-scale trends in the numerical values on top of the map. The combination of these flexible features provides an opportunity to adjust the modes of visualization to the type of data and to acquire the most meaningful picture (19). Extended documentation, a tutorial, a live example and a guide for data integration using NaviCell are provided at https://navicell.curie.fr/pages/nav_web_service.html.

To illustrate data visualization, the EMT regulation map from NaviCell collection (9) https://navicell.curie.fr/pages/signalling_network_emt_regulation_description.html was used to analyse omics data from ovarian cancer patients (45). The data files can be downloaded from https://navicell.curie.fr/pages/nav_web_service.html.

Expression and mutational profiles of proliferative and mesenchymal ovarian cancer classes were compared. Expression data from the two groups of patients were visualized on the EMT regulation map using the map staining technique in NaviCell, showing average expression of each gene across the samples. The mutation profiles for the same disease classes were visualized using glyphs. There is a clear difference in the expression pattern between the two classes, indicating a major activation of epithelial to mesenchymal transition (EMT) regulators in mesenchymal class (Figure 6B), characterized by invasive clinical outcome, opposite to the proliferative class (Figure 6A). A closer look at the EMT inducers ZEB1&2, SNAIL, CTNBB1, shows that they are largely mutated and demonstrate a low expression pattern in the proliferative, clinically, less invasive class (Figure 6C), in opposite to the invasive mesenchymal class (Figure 6D). The observation is consistent with the notion that in order to induce ETM and invasion, the cancer cell needs to be in low-proliferative status (9). Similar molecular portraits of cancer can be automatically generated using NaviCom, which connects cBioPortal and NaviCell allowing visualization of different high-throughput data types simultaneously on a network map in one click (22).

Maps generated using the procedure described earlier can be applied in different studies, not exclusively in NaviCell environment. For instance, the EMT regulation map (Figure 7A) was used to retrieve a minimal mechanistic model explaining the control of the EMT program. With this aim, hub players in each functional module of the map were identified (Figure 7B) and network complexity reduction was performed using path analysis function in BiNoM Cytoscape plugin (25) up to core regulators of EMT, apoptosis and proliferation, which were preserved through all levels of reduction (Figure 7C). The minimal model was used to predict genetic interactions leading to invasive phenotype in colon cancer (9).

Conclusions and perspectives

The knowledge about biological mechanisms is rapidly growing, demanding organization and systematic

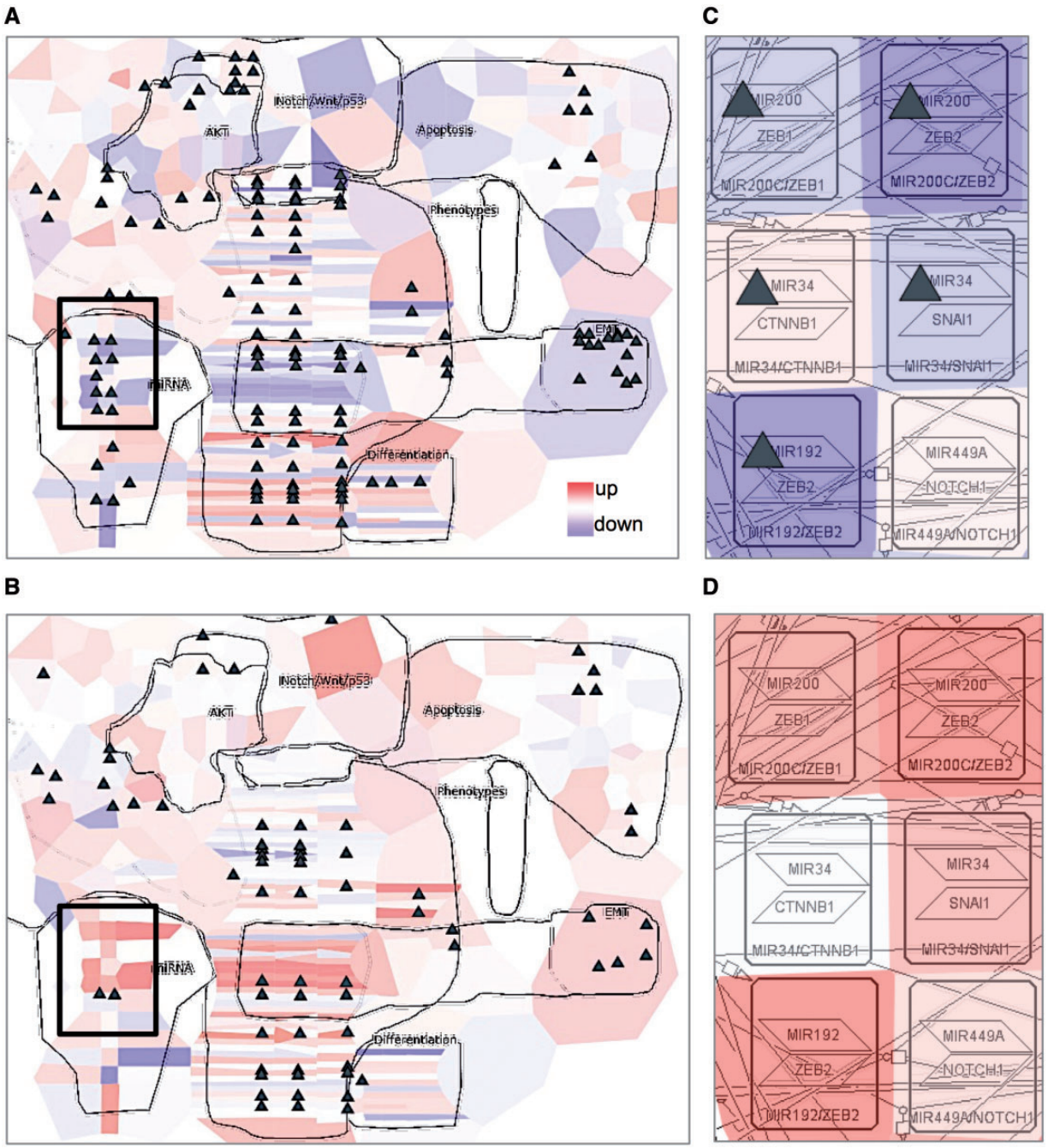

Figure 6. Visualization of cancer high-throughput data in the context of DNA repair map. Visualization of gene expression from ovarian cancer samples in a form of map staining and mutation profile in a form of glyph (triangle). (**A**) Proliferative and (**B**) mesenchymal classes of ovarian cancer. Zoom in on EMT regulators in (**C**) proliferative and (**D**) mesenchymal classes of ovarian cancer. Proliferative group, $n = 87$, mesenchymal group, $n = 96$.

representation of the information to create a global picture on cell molecular mechanisms (46). The current solution is to visualize cell signalling in a form of a diagram or a molecular map. The maps can vary in their size, complexity, contents and details' description. However, the majority of the map construction endeavours to address very similar challenges. Among others, the questions on how to improve the standardized representation of biological processes, provide intuitive map navigation tools and optimize maps update with the latest literature, are still not fully solved.

In this manuscript, we described the methodology developed following the long-standing experience with comprehensive generation and manipulation of maps. Using this approach, we created and currently maintain the Atlas of Cancer Signalling Network (ACSN) resource (20) and a collection of maps created in CellDesigner available at https://navicell.curie.fr/pages/maps.html.

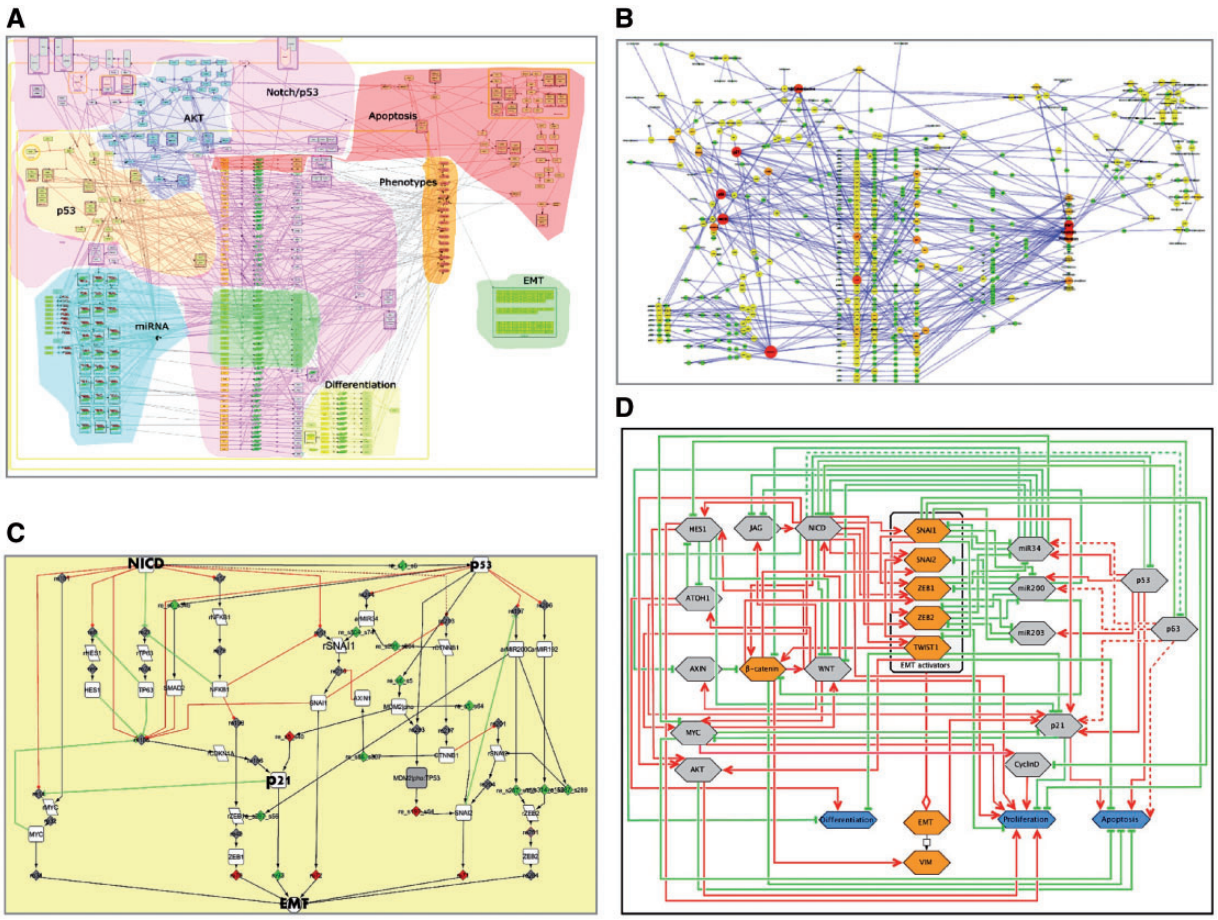

Figure 7. Retrieval of mechanistic model of EMT program control using EMT regulation map. (**A**) Comprehensive signalling map of EMT regulation, (**B**) hub player in functional modules of EMT map, (**C**) structural analysis and reduction of EMT map complexity; (**D**) scheme representing major mechanisms controlling EMT program.

We suggested a workflow for construction and annotation of signalling maps in CellDesigner, preparing hierarchical modular structure of maps and also generation of different map view levels, to allow semantic zooming-based exploration of maps in NaviCell. We introduced NaviCell, an environment for navigating large-scale maps of molecular interactions created in CellDesigner. NaviCell allows showing the content of the map in a convenient way, at several scales of complexity. It provides an opportunity to comment the map contents, thus facilitating its curation and maintenance. Finally, we showed how omics data can be visualized and interpreted in the context of the map.

Among many future challenges for the signalling network community, is the integration of similar efforts. Actually, the open-flow model of knowledge sharing and integration was suggested in the past by Kitano and colleagues (47). Nowadays, there are several examples of such efforts as Wiki Pathways (36) and Pathway Common (48). In addition, there is an emerging community activity, the Disease Maps Project, where mechanisms involved in various human diseases are brought together, allowing disease comorbidity and drug repositioning studies (http://disease-maps.org). Finally, the generation of comprehensive platforms for tools, data and knowledge sharing in systems biology and biomedical research, similar to GARUDA initiative (http://www.garuda-alliance.org), will facilitate the compatibility between tools and resources.

The map construction and annotation procedures that we expose in this manuscript aim to integrate community efforts, which are in line with FAIR principles (23). We are convinced that the application of standardized map construction procedures, providing access to the maps' content via web-based platforms, will increase the reuse of the existing maps, improving management and updating and will allow efficient generation of new high-quality signalling maps. Actually, the idea of exchangeable, re-usable maps or map modules lies in the basis of the Disease Community resource creation, where common mechanisms implicated in various diseases will be shared between disease-specific maps. Finally, standardly created maps provided in common formats will ensure their

compatibility with various systems biology tools, facilitating their applications for data analysis and modelling.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank L.C.M.G. for critical reading of the manuscript.

Funding

This work has been supported by the COLOSYS grant ANR-15-CMED-0001-04, provided by the Agence Nationale de la Recherche under the frame of ERACoSysMed-1, the ERA-Net for Systems Medicine in clinical research and medical practice and by INSERM Plan Cancer N° BIO2014-08 COMET grant under ITMO Cancer BioSys program. This work received support from MASTODON program by CNRS (project APLIGOOGLE).

Conflict of interest. None declared.

References

1. Kuperstein,I., Grieco,L., Cohen,D.P.A. *et al.* (2015) The shortest path is not the one you know: application of biological network resources in precision oncology research. *Mutagenesis*, **30**, 191–204.
2. Krogan,N.J., Lippman,S., Agard,D.A. *et al.* (2015) The cancer cell map initiative: defining the hallmark networks of cancer. *Mol. Cell*, **58**, 690–698.
3. Wang,J., Zuo,Y., Man,Y. *et al.* (2015) Pathway and network approaches for identification of cancer signature markers from omics data. *J. Cancer*, **6**, 54–65.
4. Dorel,M., Barillot,E., Zinovyev,A. *et al.* (2015) Network-based approaches for drug response prediction and targeted therapy development in cancer. *Biochem. Biophys. Res. Commun.*, **464**, 386–391.
5. Garg,A., Mohanram,K., Di Cara,A. *et al.* (2013) Efficient computation of minimal perturbation sets in gene regulatory networks. *Front. Physiol.*, **4**, 361.
6. Csermely,P., Korcsmáros,T., Kiss,H.J.M. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.*, **138**, 333–408.
7. Cohen,D., Kuperstein,I., Barillot,E. *et al.* (2013) From a biological hypothesis to the construction of a mathematical model. *Methods Mol. Biol.*, **1021**, 107–125.
8. Barillot,E., Calzone,L., Hupe,P. *et al.* (2012) *Computational Systems Biology of Cancer*. CRC Press, Boca Raton, FL.
9. Chanrion,M., Kuperstein,I., Barrière,C. *et al.* (2014) Concomitant Notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut. *Nat. Commun/*, **5**, 5005.
10. Le Novère,N., Hucka,M., Mi,H. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.
11. Chowdhury,S. and Sarkar,R.R. (2015) Comparison of human cell signaling pathway databases–evolution, drawbacks and challenges. *Database*, **2015**, 1–25.
12. Czauderna,T., Klukas,C. and Schreiber,F. (2010) Editing, validating and translating of SBGN maps. *Bioinformatics*, **26**, 2340–2341.
13. Hu,Z., Mellor,J., Wu,J. *et al.* (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, **5**, 17.
14. Kitano,H., Funahashi,A., Matsuoka,Y. *et al.* (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.*, **23**, 961–966.
15. Flórez,L.A., Lammers,C.R., Michna,R. *et al.* (2010) CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling and regulatory pathways. *Bioinformatics*, **26**, 2997–2999.
16. Kono,N., Arakawa,K., Ogawa,R. *et al.* (2009) Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS One*, **4**, e7710.
17. Gawron,P., Ostaszewski,M., Satagopam,V. *et al.* (2016) MINERVA-a platform for visualization and curation of molecular interaction networks. *NPJ Syst. Biol. Appl.*, **2**, 16020.
18. Kuperstein,I., Cohen,D.P.A., Pook,S. *et al.* (2013) NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Syst. Biol.*, **7**, 100.
19. Bonnet,E., Viara,E., Kuperstein,I. *et al.* (2015) NaviCell web service for network-based data visualization. *Nucleic Acids Res.*, **43**, W560–W565.
20. Kuperstein,I., Bonnet,E., Nguyen,H.-A. *et al.* (2015) Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, **4**, e160.
21. Calzone,L., Gelay,A., Zinovyev,A. *et al.* (2008) A comprehensive modular map of molecular interactions in RB/E2F pathway. *Mol. Syst. Biol.*, **4**, 173.
22. Dorel,M., Viara,E., Barillot,E. *et al.* (2017) NaviCom: a web application to create interactive molecular network portraits using multi-level omics data. *Database*, **2017**, 1–11.
23. Wilkinson,M.D., Dumontier,M., Aalbersberg. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
24. Tian,H., Gao,Z., Li,H. *et al.* (2015) DNA damage response–a double-edged sword in cancer prevention and cancer therapy. *Cancer Lett.*, **358**, 8–16.
25. Bonnet,E., Calzone,L., Rovera,D. *et al.* (2013) BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Syst. Biol.*, **7**, 18.
26. Rzhetsky,A., Seringhaus,M. and Gerstein,M. (2008) Seeking a new biology through text mining. *Cell*, **134**, 9–13.
27. Arighi,C.N., Wu,C.H., Cohen,K.B. *et al.* (2014) BioCreative-IV virtual issue. *Database*, **2014**, 1–6.
28. Gyori,B.M., Bachman,J.A., Subramanian,K. *et al.* (2017) From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.*, **13**, 954.
29. Lee,K., Shin,W., Kim,B. *et al.* (2016) HiPub: translating PubMed and PMC texts to networks for knowledge discovery. *Bioinformatics*, **32**, 2886–2888.
30. Ivanisenko,V.A., Saik,O.V., Ivanisenko,N.V. *et al.* (2015) ANDSystem: an Associative Network Discovery System for

automated literature mining in the field of biology. *BMC Syst. Biol.*, **9**, S2.
31. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
32. Szostak,J., Ansari,S., Madan,S. *et al.* (2015) Construction of biological networks from unstructured information based on a semi-automated curation workflow. *Database*, **2015**, 1–14.
33. Croft,D., Mundo,A.F., Haw,R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
34. Kanehisa,M., Goto,S., Sato,Y. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
35. Kelder,T., van Iersel,M.P., Hanspers,K. *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
36. Degtyarenko,K., de Matos,P., Ennis,M. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
37. Li,Q., Cheng,T., Wang,Y. *et al.* (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today*, **15**, 1052–1057.
38. Kanehisa,M. (2013) Molecular network analysis of diseases and drugs in KEGG. *Methods Mol. Biol.*, **939**, 263–275.
39. Chatr-Aryamontri,A., Breitkreutz,B.-J., Oughtred,R. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
40. Sowa,M.E., Bennett,E.J., Gygi,S.P. *et al.* (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell*, **138**, 389–403.
41. Perfetto,L., Briganti,L., Calderone,A. *et al.* (2016) SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.*, **44**, D548–D554.
42. Thiele,I. and Palsson,B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.
43. Keshava Prasad,T.S., Goel,R., Kandasamy,K. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
44. Matsuoka,Y., Ghosh,S., Kikuchi,N. *et al.* (2010) Payao: a community platform for SBML pathway model curation. *Bioinformatics*, **26**, 1381–1383.
45. Bell,D., Berchuck,A., Birrer,M. *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
46. Ghosh,S., Matsuoka,Y. and Kitano,H. (2010) Connecting the dots: role of standardization and technology sharing in biological simulation. *Drug Discov. Today*, **15**, 1024–1031.
47. Kitano,H., Ghosh,S. and Matsuoka,Y. (2011) Social engineering for virtual "big science" in systems biology. *Nat. Chem. Biol.*, **7**, 323–326.
48. Cerami,E.G., Gross,B.E., Demir,E. *et al.* (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
49. Li,J. and Zhou,B.P. (2011) Activation of β-catenin and Akt pathways by Twist are critical for the maintenance of EMT associated cancer stem cell-like characters. *BMC Cancer*, **11**, 49.
50. Kim,T., Veronese,A., Pichiorri,F. *et al.* (2011) p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. *J. Exp. Med.*, **208**, 875–883.
51. Schubert,J. and Brabletz,T. (2011) p53 Spreads out further: suppression of EMT and stemness by activating miR-200c expression. *Cell Res.*, **21**, 705–707.
52. Sahlgren,C., Gustafsson,M.V., Jin,S. *et al.* (2008) Notch signaling mediates hypoxia-induced tumor cell migration and invasion. *Proc. Natl. Acad. Sci. U.S.A*, **105**, 6392–6397.