3.4 Quality of interactions of key protein sets

STRING, HIPPIE and MINT assign quality scores to each interaction and this is used to assess the confidence level of an analysis; HIPPIE and MINT calculate the confidence score based on accumulated experimental evidence of protein interactions (M.S. et al., submitted)(Ceol, et al., 2010). This stringent approach leads to scores below 0.5 for more than 75% of the interactions reported in these databases (Figure 3A). STRING calculates its confidence score based on the likelihood that two proteins have a functional association that is as specific as the association between an average pair of proteins present in the same KEGG pathway (Kanehisa, et al., 2010; Szklarczyk, et al., 2011). In addition, higher scores are assigned to associations supported by several sources of evidence. Consequently, intensively-studied interactions are more likely to be supported by higher confidence scores. Indeed, we find that more than 80% of the STRING interactions have scores above the acceptable cut-off of 400 (defined by the authors in the program web-site).

Next, we asked whether heavily studied proteins are correspondingly covered by good quality interactions in the PPI databases. To address this question, we selected the 10% most popular human genes/proteins from the literature (that is, 2921 genes/proteins), and ranked them by popularity based on the number of PubMed entries mentioning these genes (Supplementary Table 2); of those, 2,790 were present in HIPPIE, 2,460 in STRING and 1,653 in MINT database.

We performed pair-wise comparisons of the confidence levels of the interactions shared between databases, and that involve the 10% most intensively studied proteins (Figure 3B; Supplementary Table 3). We observed a lack of agreement between the scores calculated in the databases, i.e. several interactions reported as high confidence in one database are reported as low confidence interactions in the other. In the comparison between STRING and HIPPIE, ~70% of the interactions involving the 10% most studied proteins have a high confidence score in STRING but low confidence score in HIPPIE. On the other hand, we observed that 14% of shared interactions had a score above cut-off in both databases. An example is the interaction between TP53 and HMGB1 (Jayaraman, et al., 1998), with a score of 0.83 in HIPPIE and 932 in STRING.

As mentioned before, STRING and HIPPIE are derived databases, thus several interactions shared between them were originally reported in MINT. However, each database assign different scores to those interactions, resulting in no correspondence between the scores of different databases. Therefore, to search for tendencies or biases of each scoring scheme, we considered interactions involving at least one popular protein and with conflicting scores between the databases. With these interactions, we created four groups with distinct characteristics (Table 2) and evaluated a sample of 100 interactions (25 from each group), by manually searching experimental evidence supporting these interactions in the scientific literature (Supplementary Table 4).

We observed that a protein association had high confidence score only in STRING (and low scores in the other two databases), the experimental evidence supporting an association could not be readily identified, reflecting that the scoring scheme used by STRING - assigning a high score to proteins belonging to the same pathway - may be difficult to validate. In contrast, when either MINT or HIPPIE assigned high scores to an interaction, the supporting evidence could be confirmed in one or more publications; although HIPPIE has a very strict scoring scheme: occasionally more than one publication reported an interaction but it still received a low score. Lastly, as part of the iMEX curation guidelines (Orchard, et al., 2007), the scoring scheme used by MINT was very accurate: interactions with scores greater than 0.5 could be readily confirmed by manuscripts often containing the identity of both interacting partners in its title and specifically investigating that interaction.

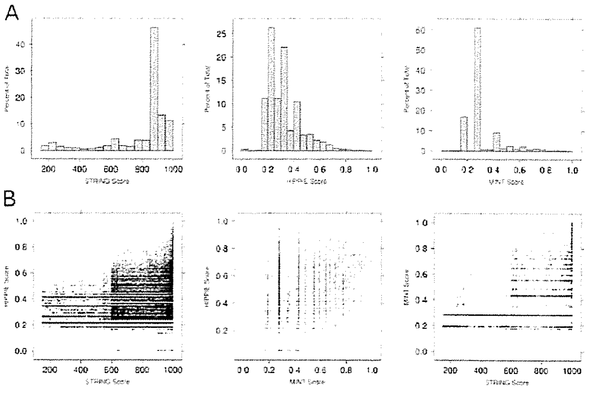


Fig. 3. (A) Three databases assign quality scores for protein interactions (HIPPIE, MINT) or functional associations (STRING). MINT and HIPPIE have a stringent quality score based on cumulative evidence from multiple sources and therefore the majority of its interactions are below 0.5. STRING on the other hand assigns a high score for proteins that are reported in pathway databases (Szklarczyk, et al., 2011). (B) Confidence scores of interactions that involve intensively studied proteins. We observed that in general there is no agreement between the database scores, with the exception that among the 31,229 interactions shared between STRING and HIPPIE, 4,539 have high confidence score in both databases. In addition, in both comparisons involving STRING, no proteins had low confidence score in MINT or HIPPIE and high confidence score in STRING.

Table 2. Groups of interactions

High-Score[1]	Low-Score[2]	Interactions[3]
STRING	HIPPIE	22,177
STRING	HIPPIE and MINT	2,225
STRING and HIPPIE	MINT	448
STRING and MINT	HIPPIE	353

[1] High-scores considered for STRING, MINT and HIPPIE were values greater than 400, 0.5 and 0.5. [2] Low-scores for STRING, MINT and HIPPIE were values lower than or equal to 400, 0.5 and 0.5. [3] All interactions included at least one popular protein.

Summarizing, we observed that although there are differences in the calculations of the quality score, interactions that are highly trustable are those that supported by different experimental systems (especially low-throughput methods), and are manually curated from literature. Ideally, interaction studies should be carried out in different experimental systems to overcome technique-specific bias (Braun, et al., 2009; Chen, et al., 2010; von Mering, et al., 2002).

3.5 Subnetworks based on organ- and cell type-specific expression data

Protein-protein interaction databases are used to address a wide range of questions that span different organisms, cell types, developmental stages, and/or phases of the cell cycle. To date, no public PPI database takes these issues into account, with the exception of the HPRD team, which in the long-term may also incorporate tissue-specific expression information. Some private companies, e.g. Ingenuity, provide tissue specific network construction, but as they limit the size of the PPI networks to be on the order of hundreds of nodes, these are not the most suitable tools for whole network studies. Here, we assessed how the incorporation of organ- and cell-type-specific expression data influence network analysis.

Using a gene expression dataset of 84 human organs and cell types (Su, et al., 2004; Wu, et al., 2009), we first selected all genes with moderate to high expression levels in each cell type (see Methods). Next, we evaluated the coverage of each database for the proteins expressed from these genes. STRING and HIPPIE cover about 60% of the organ/cell type-specific proteins, whereas the coverage reaches about 40-50% in the other databases (Supplementary Figure 4). It is also interesting to note that all databases have a relatively even coverage of all organs and cell types, although the number of genes expressed varies significantly between the different organs/cell types (Supplementary Figure 5). For example, ten times more genes are expressed in liver and heart as compared to the ovary; yet, the percent coverage in the PPI databases is comparable for these three organs.

To create organ/cell type-specific PPI networks, we then identified in the PPI database interactions for which both partners are expressed in the same organ/cell type (while eliminating interactions between proteins that are expressed in different organs/cell types). Each organ/cell type subnetwork was then built from the resulting dataset and we included 570 housekeeping proteins that are believed to be expressed in all tissues (Eisenberg and Levanon, 2003). As expected, the resulting organ/cell type-specific subnetworks possess significantly fewer interactions than the original PPI databases (between 1-25%) (Supplementary Figure 6). In addition, these subnetworks are considerably more fragmented than the parent networks, resulting in several smaller connected components (Supplementary Figure 7). We observed significant differences between the numbers of interactions for organ/cell type-specific subnetworks, which strongly correlated with the number of genes expressed in the respective organ/cell type (Supplementary Figure 8). For example, more than 6,000 different genes are expressed in BDCA dendritic cells, resulting in a subnetworks that retained 20% of the interactions found in the respective parental PPI databases. By contrast, fewer than 700 genes are expressed in ovary or skin, which reduced the specific subnetworks to just 0.4% of interactions reported in the parental networks (Supplementary Figure 6).

To assess the potential value of organ/cell type-specific subnetworks, we analyzed the interaction of cellular proteins with two medically relevant human viruses, hepatitis C virus (HCV) and human immunodeficiency virus (HIV). First, we obtained a list of 481 human proteins that interact with HCV proteins (de Chassey, et al., 2008) and compared these to the HIPPIE subnetwork created for liver. The HIPPIE database was chosen because it contains a relatively large number of interactions and covers most of the other databases; we focused on the liver subnetwork because of the relevance of this organ in HCV infection (Patrick, 1999).

From the original list of 481 HCV interactors, 98 proteins were present in the liver-specific subnetwork and they interacted with 394 different host proteins (Supplementary Table 5). Comparing the pathway membership of these 492 proteins (interactors and neighbors) with proteins specifically expressed in the liver as a background set, we observed appreciable enrichment in complement and coagulation cascades (p-value: 0.04), apoptosis (p-value: 2.94e-4), Chemokine signaling pathway (p-value: 0.0009) and focal adhesion (p-value: 1.03e-7). By contrast, when we used the complete HIPPIE database, 372 of 481 HCV interactors mapped to the database and were involved in 8,489 interactions with 3,317 different proteins. Using the same analysis that we used for the subnetwork analysis, the HCV interactors and their neighbors fell into many different categories, and no specific pathways or Gene Ontology categories, were significantly enriched, making it very difficult to identify critical pathways for the HCV pathogenesis. Hence, organ/cell type-specific subnetworks may aid in the identification of nodes that are critical in specific biological processes.

As a second example of subnetwork analysis, we studied the interaction of HIV with host cells. From the HIV-1 Human Protein Database (Ptak, et al., 2008), we obtained a dataset of 1,432 host proteins that interact with viral proteins. Next, we created subnetworks containing housekeeping genes and genes expressed in BDCA dendritic cells (DC), CD14+ monocytes, and CD4+ T-cells (all datasets were derived from the HIPPIE database). These datasets were chosen since these cell types play critical roles in HIV infections (Dragic, et al., 1996; McDonald, et al., 2003; Zhu, et al., 2002).

From the original list of 1,432 cellular proteins that interact with HIV proteins, 72 were exclusively found in the DC subnetwork and had 55 neighbors not present in the other two subnetworks. According to the pathway databases, these proteins are present in the systemic lupus erythematosus pathway (p-value: 0.001) and in the B-cell receptor signaling pathway (p-value: 0.01). By contrast, 65 cellular HIV interactors were restricted to the CD14+ monocyte subnetwork (interacting with 31 exclusive neighbors), and showed an enrichment for the apoptosis pathway (p-value: 0.08), Focal Adhesion (p-value: 0.007) and Fc gamma R-mediated phagocytosis (p-value: 0.04). Finally, 58 cellular HIV interactors (and 39 neighbors) were only detected in the CD4+ T-cell subnetwork, with an enrichment for T-cell receptor signaling (p-value: 6.8e-5) and primary immunodeficiency pathway (p-value: 0.05). These analyses demonstrate cell-type-specific interactions between HIV and cellular proteins that may be critical for the infection process. The complete list of cell-specific HIV interactors and neighbors is available in Supplementary Table 6 .

4 DISCUSSION

In this study, we compared six widely used public PPI databases for their basic characteristics, their neighborhood features, and their overlap with the other databases analyzed. In addition, we demonstrated that predictions could be significantly improved by the analysis of cell/tissue specific subnetworks, and by obtaining additional experimental verification for the interaction partners of the most intensively studied genes from literature.

The six databases compared here have different levels of coverage, in regard to both the number of proteins and the number of protein-protein interactions. Nonetheless, they assign similar topological positions to particular proteins within the network; hence, proteins with few or many interaction partners in one database are likely to have few or many interaction partners in the other databases analyzed. However, the identity of these interaction partners may differ between the databases, resulting in great uncertainty in model building. These differences reflect the differences in the algorithms, portion of literature curated by the different groups (Turinsky, et al., 2010), and the experimental techniques used to build the databases.

Many protein-protein interaction datasets are generated by expressing the two proteins of interest in one cell (for example, in the yeast two-hybrid system). In such in vitro assays, proteins may be co-expressed and interact, but in reality their expression may be dependent on cell type, different experimental stages, and/or during different phases of the cell cycle/organism development. As a result, the currently available PPI databases are believed to contain a significant percentage of false-positive entries (Deane, et al., 2002). To address this weakness, PPI databases could be combined with the increasing number of transcriptomics or proteomics datasets that assess the expression of genes or proteins in a specific organ, cell type, developmental or cell cycle stage. We here provide two examples that demonstrate the potential of this approach.

In one example, we show that the host cellular interaction partners of HCV proteins are not enriched for particular gene ontology categories or pathways in an analysis based on the entire HIPPIE database; in contrast, three KEGG pathways (apoptosis, focal adhesion, complement and coagulation cascades) are highly enriched when the HIPPIE database was analyzed in combination with a liver-specific gene expression dataset. Regulation of apoptosis may play a critical role in HCV infection to establish chronic or persistent infections (Bantel and Schulze-Osthoff, 2003). Activation of the complement and coagulation pathways has been described for HCV infections (Ueda, et al., 1993), and it was verified that hepatic inflammation can be reduced by administering CD55, a regulator of the complement pathway (Chang, et al., 2009). However, the significance of proteins involved in focal adhesion for HCV infections is currently not known, which may be addressed in further investigations. This example demonstrates how the generation of subnetworks may help in the prioritization of pathways for future studies.

In the second example, we show that each cell type subnetwork has exclusive proteins that interact with HIV. Among the exclusive proteins from each cell type are some representing critical processes studied and validated experimentally. Apoptosis induced by HIV proteins was reported to be a critical aspect of its pathogenicity (Castedo, et al., 2002; Rasola, et al., 2001; Zheng, et al., 2007). Cases of patients with concomitant systemic lupus erythematosus and HIV have been reported (Calza, et al., 2003; Gould and Tikly, 2004) and the interplay between autoimmune diseases and retroviruses is an active topic of research (Balada, et al., 2010). In addition, it was observed the association between HIV-infection and the down regulation of Fc-gammaR-mediated phagocytosis in HIV infected macrophages (Kedzierska, et al., 2002).

Some studies have generated subnetworks to address medical questions. In one example, subnetworks from normal and cancer cells have been established to identify protein-protein interactions that are characteristic of cancer development and could be targeted to 'rewire' these cells (Quayle, et al., 2007). In the context of a metabolic study, the creation of tissue-specific subnetworks helped to elucidate post-transcriptional regulation of genes from 10 different tissues that are involved in metabolic diseases (Shlomi, et al., 2008). Collectively, these and our own analyses demonstrate that cell/tissue specific subnetworks can be used to increment the biological relevance of PPI datasets.

Our analysis also revealed that current databases possess many interactions that are characterized by low confidence scores, a finding that is of particular concern for intensively studied proteins. While it is not feasible to verify all predicted interactions with different techniques, we suggest here focusing PPI evaluation efforts on the verification of low-confidence interactions of selected proteins widely used in research models but lacking high-confidence interactions. Towards this goal, we created a priority list of interactions that include highly investigated proteins such as TP53 (described earlier), MAPK1 (mitogen-activated protein kinase 1), BCL2 (B-cell CLL/lymphoma 2), or TNF (tumor necrosis factor F), among many others. Additional experimental data confirming or revealing new interactions of these 'key players' with their predicted cellular interaction partners will push PPI databases a step closer to becoming a reliable, daily-use tool for researchers, in the same way sequence analysis and protein structure databases already are.

ACKNOWLEDGEMENTS

We are grateful to the skillful and critical reading of the manuscript by Amie Eisfeld-Fenney and fruitful discussions with Takeshi Hase. We also acknowledge the Japanese Science and Technology Agency (JST), project ERATO Kawaoka.

REFERENCES

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks, *Nature*, **406**, 378-382.

Aranda, B., *et al.* (2010) The IntAct molecular interaction database in 2010, *Nucleic Acids Res*, **38**, D525-531.

Balada, E., Vilardell-Tarrés, M. and Ordi-Ros, J. (2010) Implication of Human Endogenous Retroviruses in the Development of Autoimmune Diseases, *International Reviews of Immunology*, **29**, 351-370.

Bantel, H. and Schulze-Osthoff, K. (2003) Apoptosis in hepatitis C virus infection, *Cell Death Differ*, **10**, S48-S58.

Bhattacharya, B., *et al.* (2004) Gene expression in human embryonic stem cell lines: unique molecular signature, *Blood*, **103**, 2956-2964.

Braaksma, M., *et al.* (2011) Metabolomics as a tool for target identification in strain improvement: the influence of phenotype definition, *Microbiology*, **157**, 147-159.

Braun, P., *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions, *Nat Methods*, **6**, 91-97.

Breitkreutz, B.J., *et al.* (2008) The BioGRID Interaction Database: 2008 update, *Nucleic Acids Res*, **36**, D637-640.

Calza, L., *et al.* (2003) Systemic and discoid lupus erythematosus in HIV-infected patients treated with highly active antiretroviral therapy, *Int J STD AIDS*, **14**, 356-359.

Castedo, M., *et al.* (2002) Sequential involvement of Cdk1, mTOR and p53 in apoptosis induced by the HIV-1 envelope, *EMBO J*, **21**, 4070-4080.

Ceol, A., *et al.* (2010) MINT, the molecular interaction database: 2009 update, *Nucleic Acids Res*, **38**, D532-539.

Chang, M.-L., *et al.* (2009) Hepatic inflammation mediated by hepatitis C virus core protein is ameliorated by blocking complement activation, *BMC Medical Genomics*, **2**, 51.

Chen, Y.C., *et al.* (2010) Exhaustive benchmarking of the yeast two-hybrid system, *Nat Methods*, **7**, 667-668; author reply 668.

Coulomb, S., *et al.* (2005) Gene essentiality and the topology of protein interaction networks, *Proc Biol Sci*, **272**, 1721-1725.

de Chassey, B., *et al.* (2008) Hepatitis C virus infection protein network, *Mol Syst Biol*, **4**, 230.

Deane, C.M., *et al.* (2002) Protein Interactions, *Molecular & Cellular Proteomics*, **1**, 349-356.

Dragic, T., *et al.* (1996) HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5, *Nature*, **381**, 667-673.

Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact, *Trends Genet*, **19**, 362-365.

Ewing, R.M., *et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry, *Mol Syst Biol*, **3**, 89.

Gentleman, R., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, **5**, R80.

Gould, T. and Tikly, M. (2004) Systemic lupus erythematosus in a patient with human immunodeficiency virus infection – challenges in diagnosis and management, *Clinical Rheumatology*, **23**, 166-169.

Han, J.D., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, **430**, 88-93.

Hase, T., *et al.* (2009) Structure of protein interaction networks and their implications on drug design, *PLoS Comput Biol*, **5**, e1000550.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protocols*, **4**, 44-57.

Jayaraman, L., *et al.* (1998) High mobility group protein-1 (HMG-1) is a unique activator of p53, *Genes & Development*, **12**, 462-472.

Jensen, L.J., *et al.* (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms, *Nucleic Acids Res*, **37**, D412-416.

Jeong, H., *et al.* (2001) Lethality and centrality in protein networks, *Nature*, **411**, 41-42.

Joy, M.P., *et al.* (2005) High-betweenness proteins in the yeast protein interaction network, *J Biomed Biotechnol*, **2005**, 96-103.

Kamburov, A., *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology, *Nucleic Acids Research*, **39**, D712-D717.

Kanehisa, M., *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Research*, **38**, D355-D360.

Kedzierska, K., *et al.* (2002) HIV-1 Down-Modulates γ Signaling Chain of FcγR in Human Macrophages: A Possible Mechanism for Inhibition of Phagocytosis, *The Journal of Immunology*, **168**, 2895-2903.

Krogan, N.J., *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature*, **440**, 637-643.

Mathivanan, S., *et al.* (2006) An evaluation of human protein-protein interaction data in the public domain, *BMC Bioinformatics*, **7 Suppl 5**, S19.

McDonald, D., *et al.* (2003) Recruitment of HIV and Its Receptors to Dendritic Cell-T Cell Junctions, *Science*, **300**, 1295-1297.

Orchard, S., *et al.* (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition, *Proteomics*, **7 Suppl 1**, 28-34.

Patil, A. and Nakamura, H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks, *FEBS Lett*, **580**, 2041-2045.

Patrick, M. (1999) Hepatitis C: the clinical spectrum of the disease, *Journal of hepatology*, **31**, 9-16.

Prasad, T.S., Kandasamy, K. and Pandey, A. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology, *Methods Mol Biol*, **577**, 67-79.

Ptak, R.G., *et al.* (2008) Cataloguing the HIV type 1 human protein interaction network, *AIDS Res Hum Retroviruses*, **24**, 1497-1502.

Quayle, A.P., Siddiqui, A.S. and Jones, S.J. (2007) Perturbation of interaction networks for application to cancer therapy, *Cancer Inform*, **5**, 45-65.

Ramirez, F., *et al.* (2007) Computational analysis of human protein interaction networks, *Proteomics*, **7**, 2541-2552.

Rasola, A., *et al.* (2001) Apoptosis enhancement by the HIV-1 Nef protein, *J Immunol*, **166**, 81-88.

Rual, J.F., *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, **437**, 1173-1178.

Shlomi, T., *et al.* (2008) Network-based prediction of human tissue-specific metabolism, *Nat Biotechnol*, **26**, 1003-1010.

Stumpf, M.P.H., *et al.* (2008) Estimating the size of the human interactome, *Proceedings of the National Academy of Sciences*, **105**, 6959-6964.

Su, A.I., *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A*, **101**, 6062-6067.

Szklarczyk, D., *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, *Nucleic Acids Res*, **39**, D561-568.

Turinsky, A.L., *et al.* (2010) Literature curation of protein interactions: measuring agreement across major public databases, *Database*, **2010**.

Ueda, K., *et al.* (1993) The association between hepatitis C virus infection and in vitro activation of the complement system, *Ann Clin Biochem*, **30 (Pt 6)**, 565-569.

Venkatesan, K., *et al.* (2009) An empirical framework for binary interactome mapping, *Nat Meth*, **6**, 83-90.

von Mering, C., *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic Acids Res*, **33**, D433-437.

von Mering, C., *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Wilhelm, B.T., *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, *Nature*, **453**, 1239-1243.

Wu, C., *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, *Genome Biol*, **10**, R130.

Yildirim, M.A., *et al.* (2007) Drug-target network, *Nat Biotechnol*, **25**, 1119-1126.

Zheng, L., *et al.* (2007) HIV Tat protein increases Bcl-2 expression in monocytes which inhibits monocyte apoptosis induced by tumor necrosis factor-alpha-related apoptosis-induced ligand, *Intervirology*, **50**, 224-228.

Zhu, T., *et al.* (2002) Evidence for Human Immunodeficiency Virus Type 1 Replication In Vivo in CD14+ Monocytes and Its Potential Role as a Source of Virus in Patients on Highly Active Antiretroviral Therapy, *J. Virol.*, **76**, 707-716.

Nature Reviews Genetics | AOP, published online 3 November 2011; doi:10.1038/nrg3096

STUDY DESIGNS

Software for systems biology: from tools to integrated platforms

Samik Ghosh[*], Yukiko Matsuoka[,‡], Yoshiyuki Asai[§], Kun-Yi Hsin[§] and Hiroaki Kitano[*,§,||]*

Abstract | Understanding complex biological systems requires extensive support from software tools. Such tools are needed at each step of a systems biology computational workflow, which typically consists of data handling, network inference, deep curation, dynamical simulation and model analysis. In addition, there are now efforts to develop integrated software platforms, so that tools that are used at different stages of the workflow and by different researchers can easily be used together. This Review describes the types of software tools that are required at different stages of systems biology research and the current options that are available for systems biology researchers. We also discuss the challenges and prospects for modelling the effects of genetic changes on physiology and the concept of an integrated platform.

Systems biology emerged in the mid-1990s with the aim of achieving a system-level understanding of living organisms and applying this knowledge in various fields, including medicine and biotechnology[1–4]. Early applications included modelling cell cycle dynamics[5–7], such as a computational model that explained the effects of over 120 knockout mutations on cell cycle dynamics in yeast[7]. Significant progress has also been made in the analysis of signalling pathways — for example, in understanding the dynamics of mitogen-activated protein kinase (MAPK) signalling[8] — and in cancer drug discovery applications, in which a reagent that was developed using model-based computational analysis is now in clinical trials[9,10].

System-level studies are often built on molecular and genetic findings and 'omics' studies, such as genomics, proteomics, and metabolomics. The main challenges in systems biology are the complexity of the systems, the vast quantities of data and the scattered pieces of knowledge; these all have to be integrated; therefore, systematic, computational tools are crucially important in systems biology. Software platforms have transformed industries — such as aviation, entertainment and electronics — by drastically improving productivity and by offering new capabilities[11]. Biological sciences are no different. In particular, the success of systems biology, and its application in areas such as systems drug design, requires sophisticated data handling, modelling, integrated computational analysis and knowledge integration. For example, the creation of computational models enables us to predict the behaviours of biological systems, thereby helping us to understand the underlying molecular mechanisms and to predict the impact of perturbations, such as drug treatments, on these biological systems.

Software tools and resources for systems biology need to be tailored to their intended applications in order to achieve the objectives of novel biological discoveries, drug design and answers to life-science research questions. A typical workflow for computational analysis is a cyclical process involving data acquisition, modelling and analysis. Prediction and explanation capabilities are associated with this cycle, and the integration and sharing of knowledge help to sustain these capabilities (FIG. 1).

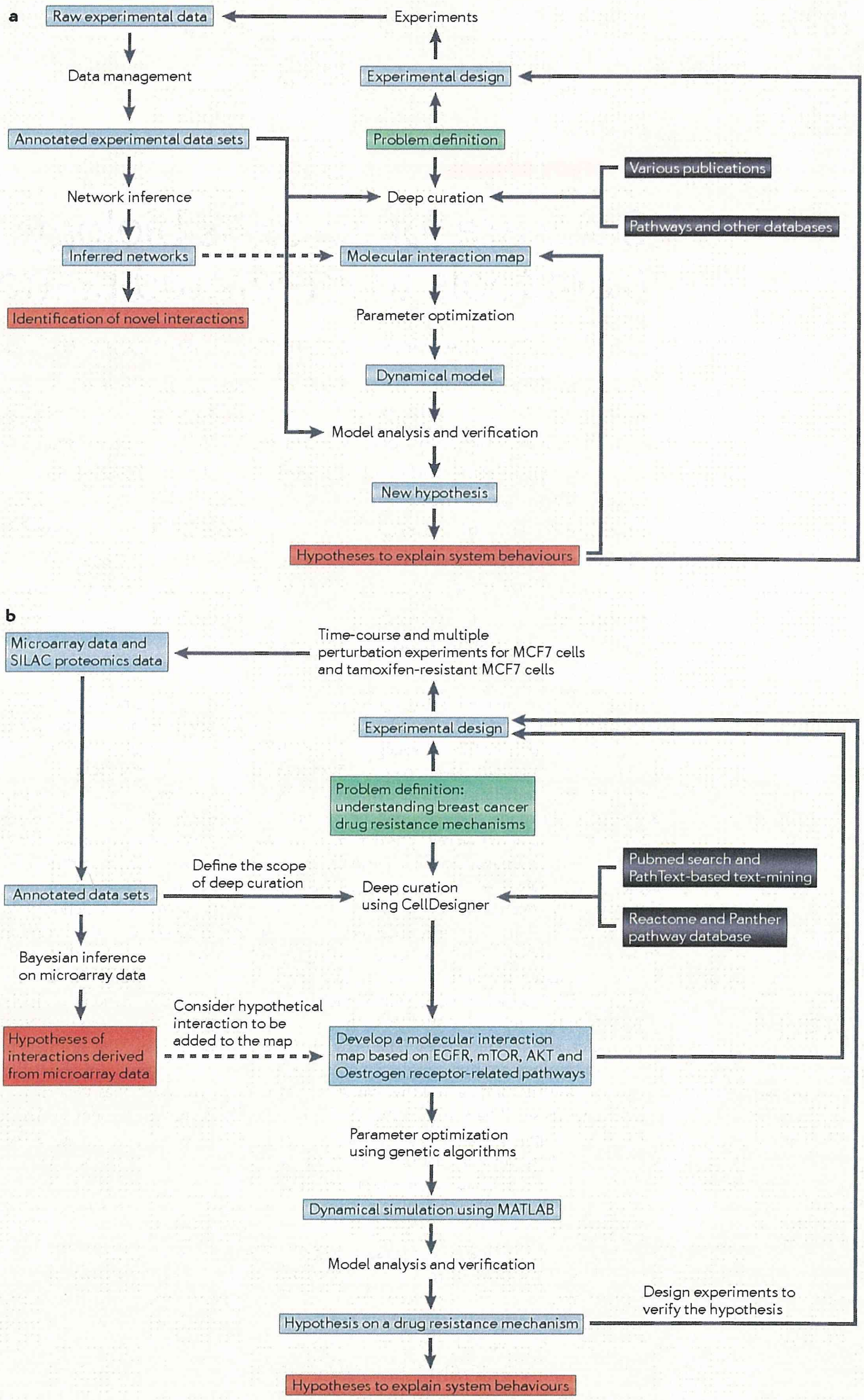
Here we describe the principles of each stage in this workflow and some examples of current tools. Links to the tools and resources mentioned in this Review are provided in Supplementary information S1 (table), along with information about their type and access policy. TABLE 1 provides a matrix to help users choose appropriate tools and resources. We provide a perspective on the current challenges facing systems biology software tools, and we describe our view that integrated software platforms will help to address future research problems in biology and medicine.

Data management

The proper acquisition and handling of data is crucially important for both the generation and verification of hypotheses. The rapid development of high-throughput experimental techniques is transforming life-science research into 'big data' science[12], and although numerous data-management systems exist[13–16], the heterogeneity of

*The Systems Biology Institute, 5F Falcon Building, 5-6-9 Shirokanedai, Minato, Tokyo 108-0071, Japan.
‡JST ERATO Kawaoka Infection-induced Host Response Project, 4-6-1 Shirokanedai, Minato, Tokyo 108-8639, Japan.
§Okinawa Institute of Science and Technology, 1919-1, Tancha, Onna-son, Kunigami, Okinawa 904-0412, Japan.
||Sony Computer Science Laboratories, Inc., 3-14-13 Higashi-Gotanda, Shinagawa, Tokyo 141-0022, Japan.
Correspondence to S.G. and H.K.
e-mails: ghosh@sbi.jp; kitano@sbi.jp

doi:10.1038/nrg3096
Published online
3 November 2011


a
Raw experimental data
Experiments
Data management
Experimental design
Annotated experimental data sets
Problem definition
Various publications
Network inference
Deep curation
Pathways and other databases
Inferred networks
Molecular interaction map
Identification of novel interactions
Parameter optimization
Dynamical model
Model analysis and verification
New hypothesis
Hypotheses to explain system behaviours
b
Microarray data and SILAC proteomics data
Time-course and multiple perturbation experiments for MCF7 cells and tamoxifen-resistant MCF7 cells
Experimental design
Problem definition: understanding breast cancer drug resistance mechanisms
Pubmed search and PathText-based text-mining
Define the scope of deep curation
Annotated data sets
Deep curation using CellDesigner
Reactome and Panther pathway database
Bayesian inference on microarray data
Consider hypothetical interaction to be added to the map
Hypotheses of interactions derived from microarray data
Develop a molecular interaction map based on EGFR, mTOR, AKT and Oestrogen receptor-related pathways
Parameter optimization using genetic algorithms
Dynamical simulation using MATLAB
Model analysis and verification
Design experiments to verify the hypothesis
Hypothesis on a drug resistance mechanism
Hypotheses to explain system behaviours

◀ Figure 1 | **Workflow of computational tasks in systems biology.** A research cycle showing the computational modelling and analyses that are involved in the workflow. **a** | The workflow starts from the 'problem definition' of the research project (shown in the green box). One stream of the workflow starts with experimental design, followed by the execution of experiments, data management and network inference. A parallel stream of the workflow consists of deep curation, parameter optimization, dynamical model analysis and model verification using experimental data. Outputs are shown in red boxes. Discrepancies between simulation results from the computational model and experimental data indicates that some of the underlying hypotheses need to be modified; the simulation should then be tested again when these new hypotheses are incorporated into the model. Transformation of a network that is inferred from large-scale data into a precise, mechanism-based model is an important step. However, this step is not yet fully achievable in practice, as indicated by the dotted arrow in the figure. **b** | An example biological application of the workflow from part **a**; in this case, research aiming to understand mechanisms of drug resistance in breast cancer. After the definition of the problem, time-series, multiple perturbation experiments would be designed, followed by data annotation, data analysis and network inference. Results from the data analysis would be used to define the scope of deep curation. However, in some cases, a molecular interaction map would be created before the experiment is designed, so that the experiments could be designed based on existing knowledge. When moving from the molecular interaction map to dynamical simulation, often only a part of the deep-curation-based molecular interaction map would be used for dynamical modelling, by which possible hypotheses for drug resistance mechanisms could be generated. This is an iterative process involving both 'dry' and 'wet' research. EGFR, epidermal growth factor receptor; mTOR, mammalian target of rapamycin; SILAC, stable isotope labelling with amino acids in cell culture.

formats, identifiers and data schema pose serious challenges. In this context, data-management systems need standardized formats for data exchange, globally unique identifiers for data mapping[17] and common interfaces that allow the integration of disparate software tools in a computational workflow.

Data-management standards. The development of data representation and communication standards for systems biology and bioinformatics has become a distinct field of work[18]. Standards for data management have focused on three core aspects: minimum information, file formats and ontologies.

Minimum information is a checklist of required supporting information for data sets from different experiments. Examples include: Minimum Information About a Microarray Experiment (MIAME)[19], Minimum Information About a Proteomic Experiment (MIAPE)[20,21] and the Minimum Information for Biological and Biomedical Investigation (MIBBI) project[22]. An important element of these standardization efforts is the incorporation of metadata (that is, data about data), which has led to the definition of standards such as the International Organization for Standardization metadata registry (ISO–MDR) standard and the Dublin Core Metadata Initiative (DCMI) standard. Standards for file formats define how the minimum information should be stored. These formats are generally Extensible Markup Language (XML)-based, which facilitates automatic processing by computers. Organizations that have defined standards include the Microarray Gene Expression Data (MGED) Society, the Proteomics Standards Initiative (PSI) and the Metabolomics Standards Initiative (MSI).

Ontologies define the relationships and hierarchy between different terms and allow the unique, semantic annotation of data. Various specialized ontologies for biology are in development; for example, the Gene Ontology (GO) and the Systems Biology Ontology (SBO) (see Supplementary Information S1 (table) for a comprehensive list of biomedical ontologies).

Data-management and data-analysis tools. Current data-management systems can be broadly classified as spreadsheet-based or Web-based, or as laboratory information management systems (LIMS). Spreadsheet programs have historically been the most popular mode of data storage and communication in the life-science community, owing mainly to the ease of use and sharing; for example, template-based spreadsheets like MAGE-TAB (a spreadsheet-based, MIAME-supportive format for microarray data) and the Investigation–Study–Assay (ISA)-TAB formats. However, their integration with analysis tools and computational workflows requires custom-built interfaces that are not supported on all software platforms. In addition, a standardized practice for filling the spreadsheet is required.

More recently, online wiki-based document and project management has become a popular mode of exchange for different laboratories, and these formats now provide security and privacy options for data protection. Other alternatives are custom-built information systems for laboratory data storage and management, such as electronic lab notebooks (ELN). These are routinely deployed in large research laboratories. While providing various features and functionalities, they are usually associated with steep learning curves for users, which, together with the cost of deployment, creates a substantial barrier to the adoption of these systems across the scientific community.

A different option, which integrates data management and analysis, is the use of workflow-management systems (WMSs). These systems harness the power of the Web to integrate different tools and services in a computational pipeline. Systems like Konstanz Information Miner (KNIME), caGrid[23], Taverna[24], Bio-STEER[25] and Galaxy[26], allow the construction, execution and sharing of specialized workflows. A comprehensive catalogue of biological Web services is available at BioCatalogue. WMSs provide the first step in building a computational pipeline by enabling data exchange, data integration and inter-tool communication. However, most current systems are tailored for specific research workflows (for example, KNIME for bioinformatics tools and Galaxy for genomic data analysis), and they support only specific sets of tools and standards; this forces researchers to use several different WMSs for a holistic understanding of their biological system of interest.

There are emerging efforts that focus on data management, such as Sage Bionetworks and ELIXIR. Sage Bionetworks is currently focused on establishing a platform for data acquisition and curation. The future aim of this platform is for modelling, using an open collaborative approach for gathering expression profile and protein interaction data, with the specific aim of using these data for drug discovery. ELIXIR is a European effort that plans to build a biological data-management infrastructure.

Table 1 | **A resource matrix of software tools and data resources**

	Tools		Standards			Projects
	Software	Resources	Ontologies	File format	Minimum information	
Data and knowledge management	MAGE-TAB, ISA-TAB KNIME, caGrid, Taverna, Bio-STEER	BioCatalogue	SBO, OBO, NCBO	MGED (MAGE), PSI, MSI	MIAME, MIAPE, MIBBI, ISO MDR, DCMI	
Data-driven network inference	R, MATLAB, BANJO					DREAM Initiative, Sage Bionetworks
Deep curation	CellDesigner, EPE, Jdesigner, PathVISIO	KEGG, Reactome, Panther pathway database, BioModels.net, WikiPathways		SBML, SBGN, CellML, BioPAX, PSI-MI	MIRIAM	
***In silico* simulation**	COPASI, SBW, JSim, Neuron, GENESIS, MATLAB, ANSYS, FreeFEM, ePNK, ina, WoPeD, Petri nets, OpenCell, CellDesigner + COPASI, CellDesigner + SOSlib, PhysioDesigner (formerly *insilico*IDE)			SED-ML, SBRML, PNML, SBML	MIASE	
Model analysis	MATLAB, Auto, XPPAut, BUNKI, ManLab, ByoDyn, SenSB, COBRA, MetNetMaker, DBSolve Optimum, Kintecus, NetBuilder, BooleanNet, SimBoolNet					
Physiological modelling	JSim, PhysioDesigner (formerly *insilico*IDE), CellDesigner (cellular modelling), FLAME, OpenCell, Virtual Physiology (produced by cLabs), GENESIS, Neuron, Heart Simulator, AnyBody			CellML, SBML, NeuroML, MML		IUPS Physiome Project, Virtual Physiological Human, High-Definition Physiology
Molecular interaction modelling	AutoDock Vina, GOLD, eHiTS	RCSB PDB, ZINC, PubChem, PDBbind				

This table summarizes the tools and resources that correspond to each step in a systems biology workflow; please refer to FIG. 1 for an overview of the workflow and to Supplementary information S1 (table) for additional information and Weblinks to these resources.

Data-driven network inference

A specific kind of modelling from large-scale data, known as data-driven network-based modelling, has been developed over the last decade[27]. Data-driven network-based modelling approaches use computational algorithms to infer causal relationships among molecular entities (such as genes, transcription factors, proteins and metabolites) from high-throughput and time-course experimental data that has been collected under various perturbations. The models that result from this kind of modelling from large-scale data sets are variously known as inference networks, co-expression networks or association networks. Early studies focused on finding patterns in gene expression profiles to distinguish disease states from healthy states; for example, in breast cancer prognosis[28]. Further studies have integrated multi-dimensional data — including genome-scale DNA variation data[29–31], gene expression data[32–34], protein–protein interaction data, DNA–protein binding data and complex binding data — to construct probabilistic, causal gene networks[35–37]. The advent of next-generation sequencing technologies provides new opportunities to incorporate the knowledge of splicing variation and SNPs into network inference models.

Mutual information
A dimensionless quantity that measures the extent to which one random variable is informative about another variable. Zero mutual information between two random variables means that they are independent.

Approaches to network inference models. Network inference models have been predominantly based on Bayesian inference techniques; that is, computing the probability of a hypothesis (in this case, the relationship between two molecular entities) based on some kind of evidence or observations (known as priors). However, several alternative techniques have also been applied[38–45], including regression, correlation methods and mutual information approaches. Mutual information approaches compute the relationship between two genes or proteins based on mutual information (a quantity that measures the mutual dependence of two variables) to infer statistically significant associations between these variables[38,39].

The current focus of the research community is on the development of novel algorithms and techniques for reconstructing molecular interaction networks from large-scale experimental data sets. In this regard, standard tools and exchange formats are not yet well established, and most research groups develop their own implementation of network reconstruction algorithms. Common software tools for implementing network reconstruction algorithms include R, MATLAB and BANJO.

Standards in data-driven inference. One of the key challenges in network inference techniques is the problem of underdetermination[46], in which the number of possible inferred interactions far exceeds the number of independent measurements. The number of experiments and the systematic selection of perturbations and time points play an important part in the reliability of inferred networks. Also, there are no true benchmarking standards for biological data and networks, and most techniques currently have their accuracy evaluated using simulated data, which do not always capture the reality in biological systems. Recent efforts towards community-driven standardization and systematic, rigorous assessment have been initiated through Sage Bionetworks (see above), and the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative. The DREAM project attempts to evaluate and benchmark different algorithms that influence network inference. Analysis of DREAM results (from the DREAM2 and DREAM3 challenges) reveal that algorithms complement each other in a highly context-specific manner, and that a community-based, consensus-driven reverse-engineering approach can lead to high-quality network inference[46]. One of the explanations for why such a community-based approach performs better than the best algorithm in a pool of algorithms is the compensatory effects from multiple algorithms on the strength and weaknesses of each individual algorithm. This is an interesting observation and it is consistent with the proposed explanation for why IBM's DeepQA system (an open-domain, automatic question-answering computing system) was successful in a 'Jeopardy!' challenge[47], based on a US quiz show that requires participants to have a wide range of topical knowledge and to interpret nuances in subtle clues that are provided to them.

Deep curation

An alternative to data-driven network inference is the deep curation approach. The deep curation approach creates a detailed molecular interaction map by the large-scale integration of knowledge, such as information from publications, databases and high-throughput data[48–51]. Unlike the data-driven approach, in which hypotheses about interactions are generated automatically, the deep curation approach constructs the model manually or semi-manually, thus making it easier for researchers to add their own hypotheses into it. Users can explicitly add unknown interactions to deep curation pathways as 'hypotheses', but it would be helpful if these interactions were made distinct from the evidence-based interactions and if they also included a rationale to support the hypothesis. Although the data-driven approach, depending on observed data, might generate networks that represent inferred causality or the association of behaviours at the transcriptional or protein-protein interaction level, they do not provide mechanistic details nor confirm causality. By contrast, the deep curation approach can provide mechanistic details of each interaction because curators will look into the details of the reported molecular mechanisms and experiments in the literature and will read them critically. Precise and in-depth mechanistic-level models are essential not only for precise computer simulations and an understanding of biological mechanisms, but also for the proper evaluation of potential drug targets. In both basic research and drug discovery, a deep curation approach is essential when the priority is to understand the details of molecular mechanisms, rather than to identify novel molecules and novel interactions.

It would be ideal to combine deep curation and data-driven approaches, but this will require further work. For example, some of the interactions that have been inferred by data-driven approaches are likely to be confirmed by deep curation approaches, and some can be clearly rejected. The remaining inferred interactions can be prioritized for further studies, and resources can be focused on these hypotheses.

Resources, standards and software for deep curation. Deep curation requires an open-ended assembly of knowledge from diverse literature and data sources and is tailored for specific purposes. Therefore, if required, the scope of the model can span multiple pathways. A variety of pathway databases — such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[52], Reactome[53], Panther pathway database[54], Pathway Commons[55], BioCyc[56] — provide information that can be used to create an initial draft of the pathway model. There also are meta-databases, such as the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) and ConsensusPathDB, which integrate diverse knowledge resources and provide a broader context for pathway curation.

There are several machine-readable model-representation standards, which have been developed for different purposes; two widely used standards are the Systems Biology Markup Language (SBML)[57] and the Biological Pathways exchange (BioPAX)[58] format, both of which were designed to represent biomolecular networks from different perspectives. The Systems Biology Graphical Notation (SBGN)[59] was designed to standardize a human-readable pathway notation. This notation defines the graphical representation of networks so that users can interpret the diagrams consistently. Minimum Information Required in the Annotation of Models (MIRIAM)[60] defines the rules for model annotation. Workshops, such as the Computational Modelling in Biology Network (COMBINE) workshop, occur regularly and provide a forum for such standardization efforts. The establishment of standards enables data and models to be re-used across multiple software tools, promotes healthy competition among these tools and helps to build a pipeline of tools for efficient analysis.

Several tools and model databases are currently available to support deep curation efforts. CellDesigner[61] is one of the most widely used software tools[62] — it enables users to visually define a model of biological interactions and to comply with SBML and SBGN. A plug-in application programming interface (API) for CellDesigner enables users to develop various additional functionalities, including the conversion of models to other formats, such as BioPAX. Several other tools

Meta-database
A database for storing metadata, which was originally defined as 'data about data', such as tags and keywords. The database is used for integrating independent distributed databases.

provide graphical editing and visualization capabilities; for example, the Edinburgh Pathway Editor (EPE), JDesigner[63], PathVISIO[64] (which is for pathway curation) and Cytoscape[65] (which is a widely used tool for the visualization of molecular networks).

Challenges of deep curation. The quality of pathways in existing pathway databases is often compromised by fragmentation and inaccuracy because these databases cover a broad range of pathways and hence little time can be spent on curating each pathway. The current 'gold standard' is manually curated maps that have been carefully built by a small group of people who spend months studying a pathway, such that they would be familiar with almost every publication on that pathway[66]. Several such maps have been reported, including for the epidermal growth factor receptor (EGFR) pathway[49], the Toll-like receptor pathway[48], the mammalian target of rapamycin (mTOR) pathway[50], the yeast cell cycle[51] and the E2F pathway[67]. In addition, the community-based reconstruction of metabolic networks for several species has been accomplished through the systematic use of various omics databases and publications[68–71].

Another consideration is that pathways reflect a specific context, such as a tissue, a disease status or a species. Pathway databases do not always identify the tissues in which interactions have been identified, thus the context of interactions should be carefully noted during the curation process. In addition, tissue-specific proteomic and gene-expression data can be used to ascertain which parts of generic pathways actually exist in the tissue of interest. This is an important practice, especially when computational models are used to explain and predict cell-line-based drug-screening experiments[10]. An additional point to consider is that there can be crosstalk among pathways.

One of the main challenges of the deep curation approach is to keep the pathways up-to-date and to validate them. This is particularly important in view of the context-specificity of molecular maps. Several disease-specific maps have been curated — for example, for rheumatoid arthritis[72] and for cardiovascular pathways — but manually creating large-scale network maps from the literature is extremely labour-intensive and requires specific quality-control procedures. Also, it is challenging for curators to maintain the motivation to continuously update the map with new discoveries. There is a need to develop techniques that automate knowledge discovery, the aggregation of pathway components and the addition of context-specific control mechanisms to pathway maps. Automated literature mining has also been investigated extensively, but is not yet close to being ready to replace human curators. Pathway validation requires an expert knowledge of the underlying biology and the ability to transform literature evidence into pathway diagrams. Recruiting experts, assigning them to pathway curation and coordinating their efforts to build integrated pathways is a major challenge.

Another option is collaborative curation, and several approaches are being developed to enable community-driven pathway updates. An example is the Payao[73] system, which has been used to promote pathway development and annotation in large and geographically distributed teams. An alternative is the community-based development and refinement of pathways, as is used in WikiPathways[74]. However, insufficient participation from active users remains a challenge for such approaches, and it is not yet clear how the widespread engagement of the biological community can be incentivized.

In silico simulation models

Molecular interaction maps provide a static picture, but the dynamics of molecular interactions in time and space have a central role in the behaviour of cells and organisms. Dynamical simulations are mostly based on models that have been created by the deep curation approach, rather than by the data-driven approach. This is because deep curation captures causality, stoichiometry and mechanisms of interactions, which are mandatory in dynamical simulations. Here we provide a brief overview for readers who are unfamiliar with the subject; for further details we recommend reading reviews that are focused on simulation and analysis[62,75,76].

Simulations have an important role in the computational verification of biological models and the computational prediction of behaviours. After the initial model is created as a set of hypotheses, dynamical simulations examine whether the model behaves like the real biological system. When some observed behaviours are not reproduced by the model, this indicates that some hypotheses are inaccurate or missing, and alternative hypotheses should be incorporated into the model and verified. Thus, the proper identification of discrepancies between experimental results and model predictions is the key for successful computational research. Dynamical modelling of complex biological systems has been applied with varying degrees of success[10,77]. Ordinary differential equations (ODEs) have been used widely as a standard numerical method in many successful cases of biological modelling[5,6,9,10]. Dynamical models that capture the stochastic (random) behaviour of molecular interactions have successfully elucidated gene transcription and translation processes[78,79] or *Escherichia coli* fate decisions during phage infections[80]. Physiological models of systems also use partial differential equations (PDEs) and a different set of tools (see below). Other techniques, such as agent-based modelling[81], process algebra (for example, the Petri net[82] system) and rule-based modelling[83], have also been applied to study the behaviour of specific biological systems.

Reaction constants and other parameters are required for simulations, and the proper calibration of models remains a major bottleneck for biological systems. Researchers can consider using rate constants that have been measured using biochemical assays, but in many cases these differ from the rate constants within cells and have not been collected in a high-throughput manner. Thus, parameters must be measured *in vivo* or be estimated through parameter-optimization techniques that are supported by various simulation and model-analysis

Ordinary differential equations
(ODEs). A type of differential equation involving functions of one independent variable, such as time, and derivatives of the functions with respect to the variable.

Partial differential equations
A type of differential equation involving functions of several independent variables, such as time and spatial axes (that is, x, y and z), and partial derivatives of the functions with respect to those variables.

Agent-based modelling
A class of computational models that simulate the interaction of agents to study their effects on a system. Agents are autonomous, decision-making entities that have heterogeneous characteristics; examples of agents are molecules or cells.

Process algebra
A mathematical modelling language for describing the behaviour of distributed systems.

Rule-based modelling
When used in biochemical science, this term refers to a way to model molecules and proteins as objects that interact with each other. The interactions are described as rules that define how the objects transform their attributes and the relationships between the objects.

Box 1 | **Parameter optimization: stochastic search methods and gradient descent methods**

There are several methods to estimate parameters for models. The stochastic search approach generates a set of parameters randomly, but often following certain rules to make the search more efficient. Each parameter set is tested in the model to see whether it generates results that are consistent with the experimental results or other defined criteria. The best set is selected and parameter values are generated again, usually close in value to the selected set, to see if there are better parameter sets. Eventually, a parameter set that can be considered optimal will be found.

The gradient descent approach has a defined algorithm that tunes parameters. It depends on error gradients that can be calculated from the difference in error values between two parameter sets. The parameter value is chosen that is estimated to have a smaller error value. Such algorithms can quickly find the optimal parameters for simple problems in which there is only one optimal point and the parameter sets near this optimal point only gradually become suboptimal. However, it may only find a local optimal parameter set for highly nonlinear and multi-peak problems.

tools and reaction databases, such as the System for the Analysis of Biochemical Pathways — Reaction Kinetics (SABIO-RK)[84] database. Sophisticated parameter-estimation algorithms, and data to calibrate them, are essential. Algorithms for optimization include stochastic methods and gradient-descent methods (BOX 1). Nevertheless, there are limitations in the current technologies and resources for creating large-scale dynamical models; it may be more practical to select part of the pathways for precise dynamical modelling, rather than to try to use an entire pathway map that inevitability contains uncertain parameters.

Standards and tools for simulations. Several standardization efforts empower the modelling community. Examples include SBML[57], SBGN[59] and MIRIAM[60] for model representation and annotation. Minimum Information About a Simulation Experiment (MIASE)[85] is used to define the minimum set of information that is required to reproduce numerical simulations, and the Simulation Experiment Description Markup Language (SED-ML) is an XML-based specification for encoding configurations for simulations, for defining models to be used, for setting up numerical calculations and for formatting outputs. In addition, the Systems Biology Results Markup Language (SBRML)[86] is a complementary language to SBML that specifies the format of results of simulations carried out on models.

Based on these standards, a series of simulation tools and software has been developed, with tools such as MATLAB and the Complex Pathway Simulator (COPASI)[87] being widely used for model simulation and analysis. The Systems Biology Workbench (SBW) is a software platform that allows multiple applications — such as software packages for modelling, analysis or visualization — to communicate with each other; this aims to enhance model exchange and simulation efficacy. Several tools support process algebra and Petri net modelling. For example: ePNK, a modelling platform for Petri nets that is based on the Petri net Markup Language (PNML); Time Petri Net Analyser (TINA), a toolbox for the editing and analysis of Petri nets; and WoPeD, a tool for modelling, simulation and analyses of Petri nets that also supports PNML. BioModels.net provides a database portal for curated, validated dynamical models that can be used to kick-start a modelling effort by re-using well-known components.

Model analysis. The next step is to analyse models for insights into the intrinsic and dynamical nature of the system (FIG. 1). A conventional time-course simulation from a defined initial state gives an indication of how the system behaves under a specific condition; more in-depth insight is provided by systematic analyses of the system under different conditions. Different mathematical techniques have been developed to analyse the behaviour of complex biological models and are supported by specific software tools[88,89] (BOX 2).

Many model-analysis techniques focus on dynamical systems that are represented as set of ODEs (BOX 2), but alternative analyses have also been developed that are based on statistical network analysis[82]. In particular, Boolean network modelling of genetic regulatory networks has gained wide acceptance in the modelling community, based on pioneering work by Kauffman[90]. Several Boolean network simulators for biological systems have been developed, including NetBuilder, BooleanNet and SimBoolNet[91]. In addition, a series of tools is available for phase-space analysis and bifurcation analysis, such as XPPaut and BUNKI. We refer readers elsewhere for details of using these analysis approaches[5,76,88,92].

Multi-scale physiological modelling

The next level, in which there is an increasing interest, is to develop physiological models that are linked with underlying molecular networks and genetic polymorphisms. Developing these models is a substantial challenge, but such models should have important applications because genetic polymorphisms and the associated differences in network dynamics can influence many diseases. For example, mutations in the voltage-gated sodium channel SCN5A disrupt the flow of sodium ions into cardiac muscle cells, which affects heart electrophysiology and leads to clinical syndromes[93]. Understanding how genetic differences affect protein structure, ion channel function, molecular network dynamics and cellular behaviours (such as electrophysiology and cardiac events) would lead to a better understanding of diseases but requires well-integrated, multi-scale modelling.

Efforts are underway to achieve integrated multi-scale modelling that links molecules and genetics to physiology, especially for models of the heart [94], and large, community-driven projects have been

Phase-space analysis
A way to analyse the dynamics of a system in a space (the phase-space), in which each of the possible states of the system is represented as a unique point.

Bifurcation analysis
A way to analyse the qualitative changes in the dynamics of a system that are caused by varying one or several parameter values continuously.

Homeodynamics
A concept that views an organism as a dynamical system; this concept emerged after the concept of homeostasis. Biological systems can be considered as homeodynamic: they can lose stability and show diverse behaviours, such as bi-stability, periodicity and chaotic dynamics.

launched. The long-running International Union of Physiological Sciences (IUPS) Physiome Project aims to promote basic science and to provide a technological foundation for integrated physiological models. Two new initiatives that started in 2010 are the Virtual Physiological Human (VPH) project in Europe and the High-Definition Physiology (HD-Physiology) project in Japan. The HD-Physiology Project, funded by the Japanese government, is trying to develop a comprehensive platform for the virtual integration of models from the molecular to whole body levels. It focuses on developing a combined model of whole-heart electrophysiology that is interconnected with cellular-, pathway- and molecular-level models and a whole-body metabolism model (FIG. 2).

Box 2 | Model-analysis methods and tools

Several different mathematical techniques have been developed to analyse the behaviour of complex biological models[88,89]. Here we describe the basic principles of some of the options: sensitivity analysis, phase-space analysis and metabolic control analysis.

Sensitivity analysis

The sensitivity of a system against various parameter changes is one of the properties that affects the robustness and fragility of a system. Sensitivity analysis can reveal not only the stability of a system against various perturbations, but can also provide information about the controllability of a system.

Phase-space analysis

As living systems operate under conditions of cellular homeostasis and homeodynamics, it is highly informative to study complex biological models to discover possible steady state and dynamical behavioural tendencies. Bifurcation analysis (the analysis of a system of ordinary differential equations (ODEs) under parameter variation) and phase-plane analysis (for example, the analysis of null-clines and local stability) help to predict systems behaviour (such as equilibrium or oscillations) when parameters are perturbed. (For details, please consult dedicated textbooks and papers[5,76,88,92].)

Metabolic control analysis

Metabolic control analysis (MCA) is a powerful quantitative framework for understanding the relationship between the properties of a metabolic network (at steady state) that is characterized by its stoichiometric structure and component reactions. MCA has been widely applied for the analysis of cellular metabolism, particularly for the analysis of the regulation of cellular metabolism. An alternative to MCA is flux-balance analysis (FBA); this a constraint-based modelling technique that has been applied in metabolic engineering[108,109]. FBA does not require details of enzyme kinetics or metabolite concentrations. It aims to compute metabolic fluxes across a network that maximizes certain system properties (such as growth rates) under conditions of constraint. Notably, FBA has been shown to accurately predict the growth rates of *Escherichia coli* under different culture conditions[109].

Model analysis is supported by many ODE solver systems (such as MATLAB), but more specialized tools are widely used in the community. Some examples are AUTO (a software package for bifurcation analysis) and XPPAut (a tool for solving ODEs that is capable of showing an orbit on the phase plane and that provides a user-friendly interface on AUTO). BUNKI and ManLab are MATLAB-based bifurcation analysis toolkits. Several tools support sensitivity analysis and parameter estimation; these include SBML-SAT, MATLAB SimBiology, ByoDyn and SensSB. SensSB is a MATLAB-based toolbox for the sensitivity analysis of systems biology models.

A related set of tools allows the study of metabolic networks. For example, DBSolve Optimum can be used for MCA computations and Kintecus is a software tool for simulating chemical kinetics, for MCA and for sensitivity analysis. These techniques fall into the category of constraint-based reconstruction and analysis (COBRA) methods, and several tools exist to support them. The COBRA Toolbox is a MATLAB-based toolbox that can be used to perform a variety of COBRA methods, including many FBA-based methods. MetNetMaker is a software tool that can create metabolic networks ready for FBA based on the KEGG LIGAND database.

Physiological modelling tools and standards. Currently there is no agreed standard for modelling physiological functions and for performing simulations at all levels of physiology. Indeed, more research is probably needed before these standards can be fully established. A hindrance to the development of standards in this field is the diversity of biological processes that operate at different spatiotemporal scales (such as in cells, tissues or organs); these processes require diverse modelling and numerical computation techniques[95]. CellML is a pioneering effort to define a markup language to describe mathematical models of physiology. Modelling languages are also available for specific fields, such as NeuroML[96] and NineML for describing models in computational neuroscience. Several tools that are based on these standards have been developed for physiological modelling (BOX 3). For example, the HD-Physiology project uses both CellDesigner (for cellular-level modelling) and PhysioDesigner, which is a software tool for modelling physiology from multicellular to whole-body levels. PhysioDesigner supports the *in silico* Markup Language (ISML)[97], which is an emerging standard XML-based language for multi-level physiological modelling, and is partially compatible with CellML and SBML. Both CellDesigner and PhysioDesigner can interface with other software platforms, and these tools are envisaged to be able to communicate with other tools through the Garuda platform (see below).

There also are publicly accessible resources that provide molecular structure and bioactivity data and that can be used for physiological modelling. These include RCSB PDB, ZINC, PubChem and PDBbind, the latter of which has had several of its commonly used programs comprehensively evaluated[98]. *In silico* simulation of protein–ligand interactions can be considered as an option for predicting the activity of small molecules, such as drugs[98,99]. This type of simulation can be performed using 'virtual docking' software, such as AutoDock Vina, GOLD or eHiTS.

Although integrating multiple levels of simulation has advantages, how this integration can be accomplished and how standards should be defined require further investigation. Some working standards are useful for clarifying the issues that need to be resolved and for outlining what can be achieved based on our current understanding; however, the introduction of obligatory standards may hamper the progress of the field.

An integrated software platform

Integrated software platforms have been a driving force of productivity, quality improvement and innovation in industries[11], and we can expect the same in systems biology. The concept is of an integrated software platform that enables users to access data and knowledge from any stage in the workflow, that allows the adaptation of the workflow to best fit the user's needs and that provides consistent user experiences and high levels of interoperability. All of these features can reduce the time costs that are associated with using independent and incompatible software. In principle, integrated platforms would significantly improve productivity and would reduce errors in the handling and analysis of complex data and models.

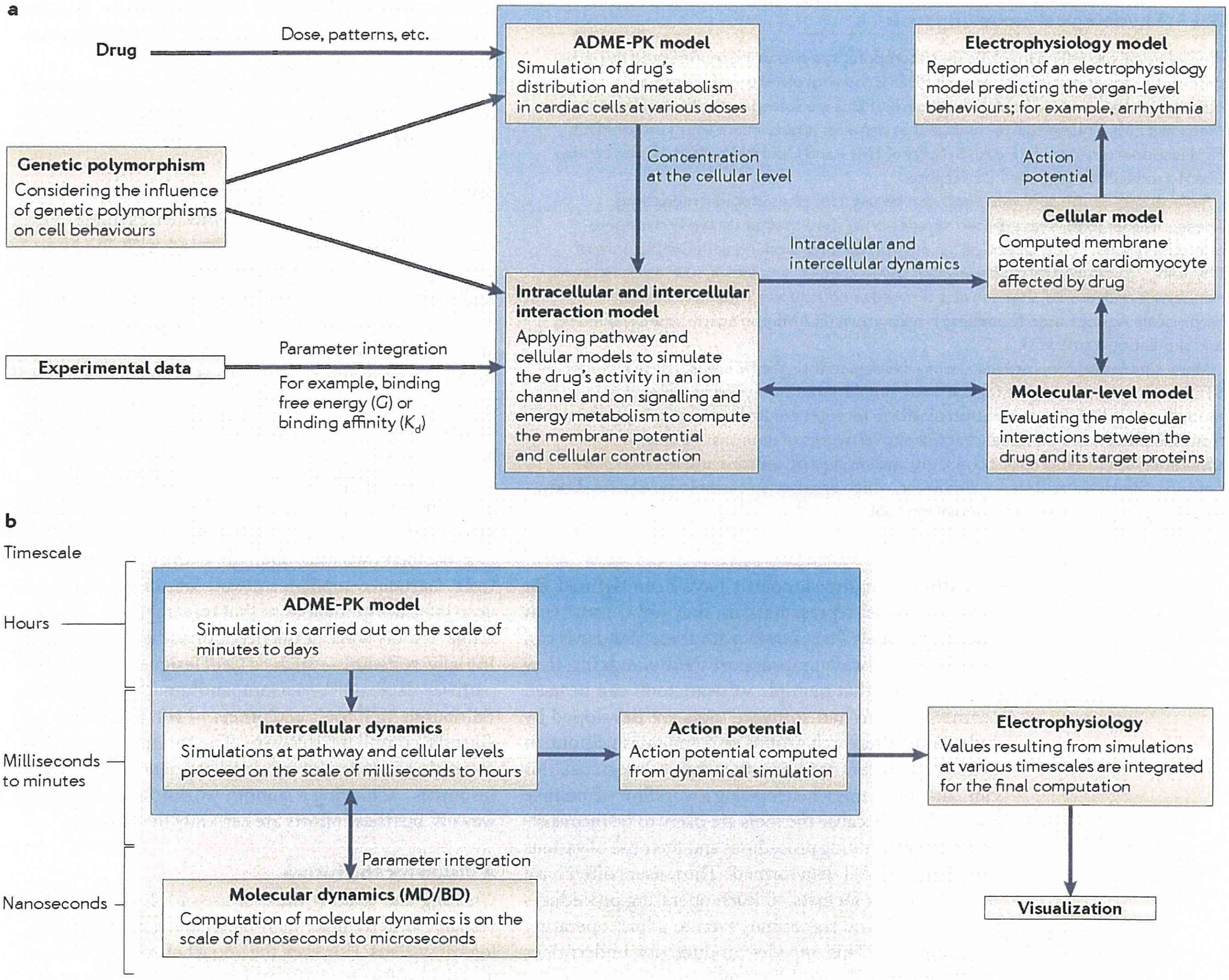

a
Drug
Dose, patterns, etc.
ADME-PK model
Simulation of drug's distribution and metabolism in cardiac cells at various doses
Electrophysiology model
Reproduction of an electrophysiology model predicting the organ-level behaviours; for example, arrhythmia
Genetic polymorphism
Considering the influence of genetic polymorphisms on cell behaviours
Concentration at the cellular level
Action potential
Cellular model
Computed membrane potential of cardiomyocyte affected by drug
Intracellular and intercellular dynamics
Intracellular and intercellular interaction model
Applying pathway and cellular models to simulate the drug's activity in an ion channel and on signalling and energy metabolism to compute the membrane potential and cellular contraction
Experimental data
Parameter integration
For example, binding free energy (G) or binding affinity (K_d)
Molecular-level model
Evaluating the molecular interactions between the drug and its target proteins
b
Timescale
Hours
Milliseconds to minutes
Nanoseconds
ADME-PK model
Simulation is carried out on the scale of minutes to days
Intercellular dynamics
Simulations at pathway and cellular levels proceed on the scale of milliseconds to hours
Action potential
Action potential computed from dynamical simulation
Electrophysiology
Values resulting from simulations at various timescales are integrated for the final computation
Parameter integration
Molecular dynamics (MD/BD)
Computation of molecular dynamics is on the scale of nanoseconds to microseconds
Visualization

Figure 2 | **An example application of the High-Definition Physiology Project. a** | A possible use of an integrated multi-scale model is to evaluate the effects of a drug on cardiac events. A simulation condition can be set that consists of a specific drug dose and its temporal pattern of administration. Absorption, distribution, metabolism and excretion pharmacokinetics (ADME-PK) models that are built based on various molecular properties can compute drug distribution and metabolism, so that a change in the drug dose that a cardiomyocyte is exposed to can be simulated. The molecular properties of the drug can also be calculated using *in silico* methods[110], such as quantitative structure–activity relationship (QSAR) modelling, and can be applied as a parametric component to a specific cell model. Pathway- and cellular-level models use the computed drug dose as an environmental factor in the simulation of ion channel activity, signalling and energy metabolism and then compute the membrane potential and cellular contraction. In some cases, genetic polymorphisms may change the behaviours of the cell. For novel protein structures of ion channels or other important molecules, *in silico* simulations of molecular interactions may be used to better estimate the interaction parameters that are not experimentally known. The computed membrane potential can be used to reproduce the organ-level electrophysiology of arrhythmia. **b** | Three different timescales have to be coupled for the simulations that are outlined in part **a**, and the methods that are relevant to each simulation are computationally intensive. ADME-PK are simulated on the scale from minutes to days. Cellular- and pathway-level simulations are mostly on the scale of milliseconds to hours. Molecular dynamics is computed on the scale of nanoseconds to microseconds. Owing to these large differences in timescales, loosely coupled, dynamically measured simulations and precomputed values are used for the final integrated computation. Inevitably, different numerical solution methods need to be used, but they must function coherently. For example, fluid dynamics of the blood in a heart can be described by partial differential equations (PDEs). An electrocardiogram that is derived from the cardiac electrical activity can also be computed using PDEs, but most of the intracellular signalling and the whole-body ADME-PK model will be calculated by ordinary differential equations (ODEs). Close linkage of ODEs and PDEs is crucial in such a model. In those cases in which the stochastic behaviour of molecules has a crucial role, stochastic computation may also need to be used. MD/BD, molecular dynamics or Brownian dynamics.

Constraint-based reconstruction and analysis (COBRA). A suite of methods to simulate, analyse and predict various phenotypes using genome-scale models. These methods are used particularly for metabolic networks.

Box 3 | Physiological modelling tools

Physiological modelling involves spatiotemporal systems being represented by partial differential equations (PDEs); solving PDEs is more involved than for ordinary differential equations (ODEs). In most cases, PDEs are solved by the Finite Element Method (FEM), a numerical technique that provides approximate solutions for PDEs. FEM is supported in tools like ANSYS, FreeFEM++ and OpenFEM. MATLAB can be also used to solve PDEs with the PDE Toolbox.

Several simulation software tools can be used for physiological modelling. These include: JSim, a Java-based simulation system that can solve first-degree, one-dimensional PDEs; OpenCell, an environment for working with CellML; Virtual Physiology (produced by cLabs), a set of tissue-specific simulators, such as for neurons, skin senses and muscle, that can also be used as educational tools; and Flexible Large-scale Agent-based Modelling Environment (FLAME), an agent-based modelling and simulation framework.

Many simulation tools are also being developed for specific areas. These include: STEPS, a stochastic engine for pathway simulation that is used in molecular modelling; GENESIS, a simulation platform for multiple levels of neural systems, from subcellular biochemical reactions to networks of neurons; Neuron, a simulation environment for computational models of neurons and networks of neurons; SimHeart and Heart Simulator, which are used in cardiology; and AnyBody, a full-body musculoskeletal simulation tool.

Although many standards have been defined for data and model representations, they only ensure that data and models that comply with these standards can be used by software that support these standards; they do not ensure that multiple software tools can be used seamlessly[100]. When software tools are developed by independent research groups or companies without an explicit agreement as to how they can be integrated, this can cause problems when forming a workflow of multiple tools. This is because the tools are likely to be inconsistent in their operating procedures and their use of various non-standardized data formats. Thus, users often have to convert data formats, to learn operating procedures for each tool, and sometimes even to adjust operating environments. This impedes productivity, undermines the flexibility of the workflow and is prone to errors.

As an example workflow, a researcher working on modelling an oncogenic MAPK pathway may wish to quickly access, by one click, the sequences of genes that are involved in this pathway to see the mutations that are associated with a specific subgroup of cancer patients. They might then search a protein structure database for the three-dimensional structures of proteins that are encoded by these mutated genes to see how the mutations might affect the three-dimensional structures. Next, they might explore possible docking interactions of candidate kinase inhibitors (using virtual docking simulations). Then, using advanced text-mining tools, the researcher could search for experimental and clinical articles that have reported possible effects of the compound and similar compounds on the cell line of interest or a cell line with similar mutations. Finally, the researcher could modify the original model to incorporate possible differences in networks owing to the mutations and could run dynamical simulations to see the effects on the cellular responses to specific compounds. Currently, this workflow requires multiple separate software tools, and there is no transfer of retrieved information among software tools. A successful, integrated software tool would enable such a workflow in a few clicks, so that users could concentrate on science rather than on software operation.

Recently, several initiatives have been launched to move towards software integration. The US Department of Energy is initiating the Systems Biology Knowledgebase project for building an integrated environment for data, knowledge and tools as part of their Genomes to Life programme. Another example is the Garuda Alliance, which was formed with the aim of creating a platform and a set of guidelines to achieve a highly productive and flexible software and data environment; that is, a one-stop service for systems biology and bioinformatics. The aim is to have a high level of interoperability among software in a language-agnostic manner, to provide consistent user experiences and to offer a broader accessibility of tools and resources. To achieve these objectives, the Garuda Core will provide defined and comprehensive APIs, a wide range of program and widget parts, and a series of design guidelines. Developers of tools will be able to use the provided APIs to make their own tools operational through the Garuda Core. Garuda-compliant software would need to adopt user-interface guidelines so that researchers could use a range of tools without the need for additional learning. Initially, software — such as CellDesigner, the Panther pathway database[101], bioCompendium, PathText[102], the Edinburgh SBSI tools and others — will be provided as Garuda-compliant software. The intention is to host increasing numbers of software and data or knowledge resources. Achieving a smooth workflow is still a long way off, but these efforts are certainly the first step.

A vision for the future

Creating and making the best use of software and data resources will facilitate an in-depth understanding of biological systems. However, the impact of creating a widely accepted software platform may go far beyond productivity improvements in each research group because the platform could potentially connect research groups globally. Although international collaboration in scientific projects is common, determining how best to create a successful open collaboration is still a challenge. For example, creating and maintaining a comprehensive and in-depth model of a biological system is often beyond the scope of any single research group. Maintaining, updating and improving pathway databases — such as Reactome, KEGG, and the Panther pathway database — requires continuous funding. Also, such databases are not sufficiently in-depth for many complex pathways, especially when compared with interaction maps that have been developed by a few dedicated researchers who are focused on specific pathways[66].

Some alternative approaches have been proposed that use Web2.0 services, as Wikipedia does. There are several such attempts, including Wikipathways[74], Wikigenes[103] and Gene wiki[104]. However, many of these efforts are struggling[105]. One possible reason is the lack of incentive for scientists to contribute their knowledge and data. Why would somebody spend time to share their knowledge when such a contribution is not properly acknowledged[106]? Although there are discussions of schemes to systematically acknowledge such efforts, it is yet to be seen whether these schemes can change social dynamics and hence the motivations of potential contributors. There may be a great opportunity to enhance our scientific productivity when a 'network of intelligence' or 'wisdom of crowds'[107] approach can become a reality because everyone could gain from the ideas and experiences of others. However, we do not know yet how best to achieve this in reality. Web2.0 approaches are often suggested by computer science-based researchers because of the success of such approaches in their field. However, there are cultural differences in biological research, and overcoming these differences may be a substantial challenge and may also require the involvement of a broader range of experts, such as sociologists and psychologists.

Our vision is that the increased capability to navigate and relate various data and knowledge resources using integrated platforms would enable researchers to enjoy a higher level of productivity and a greater potential for innovation. Connecting genomics, molecular networks and physiology will provide us with a deeper understanding of how individual differences in the genome affect physiological processes through alterations in molecular networks. The current reality is that there are various software tools that can be used for a broad range of systems biology research, and these tools are being increasingly integrated owing to standardization and alliance efforts. Emerging comprehensive, consistent and community-wide software platforms enable us to promote systems biology research today, and also to think about what comes next.

1. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
2. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
3. Kitano, H. Perspectives on systems biology. *New Generation Computing* **18**, 199–216 (2000).
4. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
5. Tyson, J. J., Chen, K. & Novak, B. Network dynamics and cell physiology. *Nature Rev. Mol. Cell Biol.* **2**, 908–916 (2001).
6. Novak, B. & Tyson, J. J. Numerical analysis of a comprehensive model of M-phase control in *Xenopus* oocyte extracts and intact embryos. *J. Cell Sci.* **106**, 1153–1168 (1993).
7. Chen, K. C. *et al.* Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell* **15**, 3841–3862 (2004).
A pioneering study using computational modelling and analysis of the budding yeast cell cycle. The model computationally reproduced the phenotypes of various gene deletion mutants.
8. Aoki, K., Yamada, M., Kunida, K., Yasuda, S. & Matsuda, M. Processive phosphorylation of ERK MAP kinase in mammalian cells. *Proc. Natl Acad. Sci. USA* **108**, 12675–12680 (2011).
9. Schoeberl, B. *et al.* An ErbB3 antibody, MM-121, is active in cancers with ligand-dependent activation. *Cancer Res.* **70**, 2485–2494 (2010).
10. Schoeberl, B. *et al.* Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis. *Sci. Signal.* **2**, ra31 (2009).
11. Evans, D., Hagiu, A. & Schmalensee, R. *Invisible Engines: How Software Platforms Drive Innovation and Transform Industries.* (MIT Press, 2006).
An easy-to-read introduction to the concept of software platforms in industries.
12. Lee, T. L. Big data: open-source format needed to aid wiki collaboration. *Nature* **455**, 461 (2008).
13. Brown, F. Saving big pharma from drowning in the data pool. *Drug Discov. Today* **11**, 1043–1045 (2006).
14. Kröger, P. & Bry, F. A computational biology database digest: data, data analysis, and data management. *Distributed and Parallel Databases* **13**, 7–42 (2003).
15. Field, D., Tiwari, B. & Snape, J. Bioinformatics and data management support for environmental genomics. *PLoS Biol.* **3**, e297 (2005).
16. Keator, D. B. Management of information in distributed biomedical collaboratories. *Methods Mol. Biol.* **569**, 1–23 (2009).
17. Van Deun, K., Smilde, A. K., van der Werf, M. J., Kiers, H. A. & Van Mechelen, I. A structured overview of simultaneous component based data integration. *BMC Bioinformatics* **10**, 246 (2009).
18. Brazma, A., Krestyaninova, M. & Sarkans, U. Standards for systems biology. *Nature Rev. Genet.* **7**, 593–605 (2006).
19. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) —toward standards for microarray data. *Nature Genet.* **29**, 365–371, (2001).
20. Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nature Biotech.* **25**, 887–893 (2007).
21. Martens, L., Palazzi, L. M. & Hermjakob, H. Data standards and controlled vocabularies for proteomics. *Methods Mol. Biol.* **484**, 279–286 (2008).
22. Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotech.* **26**, 889–896 (2008).
23. Saltz, J. *et al.* caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics* **22**, 1910–1916 (2006).
24. Oinn, T. *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054 (2004).
25. Lee, S., Wang, T. D., Hashmic, N. & Cummings, M. P. Bio-STEER: A semantic Web workflow tool for Grid computing in the life sciences. *Future Generation Computer Systems* **23**, 497–509 (2007).
26. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
27. Schadt, E. E., Friend, S. H. & Shaywitz, D. A. A network view of disease and compound screening. *Nature Rev. Drug Discov.* **8**, 286–295 (2009).
28. van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
29. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
30. Dewan, A. *et al.* *HTRA1* promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992 (2006).
31. Yang, Z. *et al.* A variant of the *HTRA1* gene increases susceptibility to age-related macular degeneration. *Science* **314**, 992–993 (2006).
32. Chesler, E. J. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genet.* **37**, 233–242 (2005).
33. Monks, S. A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
34. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
35. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105**, 363–374 (2004).
36. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* **3**, e69 (2007).
37. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genet.* **40**, 854–861 (2008).
38. Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7** (Suppl. 1), S7 (2006).
39. Faith, J. J. *et al.* Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
40. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli. Nature Genet.* **31**, 64–68 (2002).
41. Alon, U. Network motifs: theory and experimental approaches. *Nature Rev. Genet.* **8**, 450–461 (2007).
42. Fadda, A. *et al.* Inferring the transcriptional network of *Bacillus subtilis*. *Mol. Biosyst.* **5**, 1840–1852 (2009).
43. Cho, B. K. *et al.* The transcription unit architecture of the *Escherichia coli* genome. *Nature Biotech.* **27**, 1043–1049 (2009).
44. Mendoza-Vargas, A. *et al.* Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PloS ONE* **4**, e7526 (2009).
45. Lemmens, K. *et al.* DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli. Genome Biol.* **10**, R27 (2009).
46. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nature Rev. Microbiol.* **8**, 717–729 (2010).
47. Ferrucci, D. *et al.* Building Watson: an overview of the DeepQA project. *AI Magazine* **31**, 3 (2010).
48. Oda, K. & Kitano, H. A comprehensive map of the toll-like receptor signaling network. *Mol. Syst. Biol.* **2**, 2006.0015 (2006).
49. Oda, K., Matsuoka, Y., Funahashi, A. & Kitano, H. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.* **1**, 2005.0010 (2005).
50. Caron, E. *et al.* A comprehensive map of the mTOR signaling network. *Mol. Syst. Biol.* **6**, 453 (2010).
51. Kaizu, K. *et al.* A comprehensive molecular interaction map of the budding yeast cell cycle. *Mol. Syst. Biol.* **6**, 415 (2010).
52. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
54. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284–D288 (2005).
55. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
56. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
57. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
An original paper on SBML that triggered various standardization efforts in systems biology.
58. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nature Biotech.* **28**, 935–942 (2010).
59. Le Novere, N. *et al.* The Systems Biology Graphical Notation. *Nature Biotech.* **27**, 735–741 (2009).
60. Le Novere, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotech.* **23**, 1509–1515 (2005).

61. Kitano, H., Funahashi, A., Matsuoka, Y. & Oda, K. Using process diagrams for the graphical representation of biological networks. *Nature Biotech.* **23**, 961–966 (2005).
62. Klipp, E., Liebermeister, W., Helbig, A., Kowald, A. & Schaber, J. Systems biology standards — the community speaks. *Nature Biotech.* **25**, 390–391 (2007).
63. Sauro, H. M. *et al.* Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* **7**, 355–372 (2003).
64. van Iersel, M. P. *et al.* Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* **9**, 399 (2008).
65. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
66. Bauer-Mehren, A., Furlong, L. I. & Sanz, F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* **5**, 290 (2009).
67. Calzone, L., Gelay, A., Zinovyev, A., Radvanyi, F. & Barillot, E. A comprehensive modular map of molecular interactions in RB/E2F pathway. *Mol. Syst. Biol.* **4**, 173 (2008).
68. Thiele, I. & Palsson, B. O. Reconstruction annotation jamborees: a community approach to systems biology. *Mol. Syst. Biol.* **6**, 361 (2010).
This paper discusses issues regarding community efforts to reconstruct comprehensive metabolic networks.
69. Thiele, I. & Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010).
70. Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. Reconstruction of biochemical networks in microorganisms. *Nature Rev. Microbiol.* **7**, 129–143 (2009).
A review on the current state-of-the-art in data-driven genome-wide network reconstruction.
71. Herrgard, M. J. *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotech.* **26**, 1155–1160 (2008).
72. Wu, G., Zhu, L., Dent, J. E. & Nardini, C. A comprehensive molecular interaction map for rheumatoid arthritis. *PLoS ONE* **5**, e10137 (2010).
73. Matsuoka, Y., Ghosh, S., Kikuchi, N. & Kitano, H. Payao: a community platform for SBML pathway model curation. *Bioinformatics* **26**, 1381–1383 (2010).
74. Pico, A. R. *et al.* WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184 (2008).
75. Wierling, C., Herwig, R. & Lehrach, H. Resources, standards and tools for systems biology. *Brief. Funct. Genomic. Proteomic.* **6**, 240–251 (2007).
76. Klipp, E. et al. *Systems Biology: A Textbook* (Wiley-VCH, 2009).
A text book with examples of modelling and computational analysis.
77. Lopez-Aviles, S., Kapuy, O., Novak, B. & Uhlmann, F. Irreversibility of mitotic exit is the consequence of systems-level feedback. *Nature* **459**, 592–595 (2009).
78. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA.* **94**, 814–819 (1997).
79. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
80. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648 (1998).
81. Emonet, T., Macal, C. M., North, M. J., Wickersham, C. E. & Cluzel, P. AgentCell: a digital single-cell assay for bacterial chemotaxis. *Bioinformatics* **21**, 2714–2721 (2005).
82. Hofestadt, R. & Thelen, S. Quantitative modeling of biochemical networks. *Stud. Health Technol. Inform.* **162**, 3–16 (2011).
83. Blinov, M. L., Faeder, J. R., Goldstein, B. & Hlavacek, W. S. BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics* **20**, 3289–3291 (2004).
84. Swainston, N. *et al.* Enzyme kinetics informatics: from instrument to browser. *FEBS J.* **277**, 3769–3779 (2010).
85. Waltemath, D. *et al.* Minimum Information About a Simulation Experiment (MIASE). *PLoS Comput. Biol.* **7**, e1001122 (2011).
86. Dada, J. O., Spasic, I., Paton, N. W. & Mendes, P. SBRML: a markup language for associating systems biology data with models. *Bioinformatics* **26**, 932–938 (2010).
87. Hoops, S. *et al.* COPASI — a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067–3074 (2006).
88. Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. *Systems Biology in Practice: Concepts, Implementation and Application* (John Wiley & Sons, 2005).
89. Haefner, J. W. *Modeling Biological Systems: Principles and Applications* (Kluwer Academic Pub, 1996).
90. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J.Theor. Biol.* **22**, 437–467 (1969).
91. Zheng, J. *et al.* SimBoolNet — a Cytoscape plugin for dynamic simulation of signaling networks. *Bioinformatics* **26**, 141–142 (2010).
92. Iglesias, P. & Ingaalls, B. *Control Theory and Systems Biology* (MIT Press, 2009).
An excellent collection of introductory articles on how control theory can be applied to systems biology analysis.
93. Chen, Q. *et al.* Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Nature* **392**, 293–296 (1998).
94. Noble, D. Modeling the heart — from genes to cells to the whole organ. *Science* **295**, 1678–1682 (2002).
95. Nomura, T. Towards integration of biological and physiological functions at multiple levels. *Front. Physiol.* **1**, 164 (2010).
96. Gleeson, P. *et al.* NeuroML: a language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS Comput. Biol.* **6**, e1000815 (2010).
97. Asai, Y. *et al.* Specifications of insilicoML 1.0: a multilevel biophysical model description language. *J. Physiol. Sci.* **58**, 447–458 (2008).
98. Plewczynski, D., La niewski, M., Augustyniak, R. & Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **32**, 742–755 (2011).
99. Englebienne, P. & Moitessier, N. Docking ligands into flexible and solvated macromolecules. 4: are popular scoring functions accurate for this class of proteins? *J. Chem. Inf. Model.* **49**, 1568–1580 (2009).
100. Swertz, M. A. & Jansen, R. C. Beyond standardization: dynamic software infrastructures for systems biology. *Nature Rev. Genet.* **8**, 235–243 (2007).
101. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* **563**, 123–140 (2009).
102. Kemper, B. *et al.* PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics* **26**, i374–i381 (2010).
103. Maier, H. *et al.* LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. *Nucleic Acids Res.* **33**, W779–W782 (2005).
104. Huss, J. W. *et al.* The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.* **38**, D633–D639 (2010).
105. Callaway, E. No rest for the bio-wikis. *Nature* **468**, 359–360 (2010).
106. Kitano, H., Ghosh, S. & Matsuoka, Y. Social engineering for virtual 'big science' in systems biology. *Nat. Chem. Biol.* **7**, 323–326 (2011).
This paper discusses social issues in community-driven efforts in systems biology.
107. Surowiecki, J. *The Wisdom of Crowds*. (Anchor, 2005).
108. Edwards, J. S. & Palsson, B. O. How will bioinformatics influence metabolic engineering? *Biotechnol. Bioeng.* **58**, 162–169 (1998).
109. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotech.* **19**, 125–130 (2001).
110. Smith, D. A. in *Metabolism, Pharmacokinetics and Toxicity of Functional Groups* 61–94 (Royal Society of Chemistry Publishing, 2010).

Acknowledgements

This work is, in part, supported by funding from the HD-Physiology Project of the Japan Society for the Promotion of Science (JSPS) to the Okinawa Institute of Science and Technology (OIST). Additional support is from a Canon Foundation Grant, the International Strategic Collaborative Research Program (BBSRC-JST) of the Japan Science and Technology Agency (JST), the Exploratory Research for Advanced Technology (ERATO) programme of JST to the Systems Biology Institute (SBI) and from a strategic cooperation partnership between the Luxembourg Centre for Systems Biomedicine and SBI.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

The Systems Biology Institute: http://www.sbi.jp
BioCatalogue: http://www.biocatalogue.org
BioModels.net: http://biomodels.net

SUPPLEMENTARY INFORMATION

See online article: S1 **(table)**

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Development 137, 1515-1522 (2010) doi:10.1242/dev.044545
© 2010. Published by The Company of Biologists Ltd

The oscillation of Notch activation, but not its boundary, is required for somite border formation and rostral-caudal patterning within a somite

Masayuki Oginuma[1,2,*], Yu Takahashi[3,*], Satoshi Kitajima[3], Makoto Kiso[2], Jun Kanno[3], Akatsuki Kimura[1,4] and Yumiko Saga[1,2,†]

SUMMARY

Notch signaling exerts multiple roles during different steps of mouse somitogenesis. We have previously shown that segmental boundaries are formed at the interface of the Notch activity boundary, suggesting the importance of the Notch on/off state for boundary formation. However, a recent study has shown that mouse embryos expressing Notch-intracellular domain (NICD) throughout the presomitic mesoderm (PSM) can still form more than ten somites, indicating that the NICD on/off state is dispensable for boundary formation. To clarify this discrepancy in our current study, we created a transgenic mouse lacking NICD boundaries in the anterior PSM but retaining Notch signal oscillation in the posterior PSM by manipulating the expression pattern of a Notch modulator, lunatic fringe. In this mouse, clearly segmented somites are continuously generated, indicating that the NICD on/off state is unnecessary for somite boundary formation. Surprisingly, this mouse also showed a normal rostral-caudal compartment within a somite, conferred by a normal Mesp2 expression pattern with a rostral-caudal gradient. To explore the establishment of normal Mesp2 expression, we performed computer simulations, which revealed that oscillating Notch signaling induces not only the periodic activation of *Mesp2* but also a rostral-caudal gradient of Mesp2 in the absence of striped Notch activity in the anterior PSM. In conclusion, we propose a novel function of Notch signaling, in which a progressive oscillating wave of Notch activity is translated into the rostral-caudal polarity of a somite by regulating *Mesp2* expression in the anterior PSM. This indicates that the initial somite pattern can be defined as a direct output of the segmentation clock.

KEY WORDS: Notch signaling, Hes7, Mesp2, Segmentation clock, Presomitic mesoderm, Lunatic fringe, Somitogenesis

INTRODUCTION

The periodicity of the segmented somites is established in the posterior presomitic mesoderm (PSM) via the function of a so-called molecular clock, which is based on complex gene regulatory networks under the control of three major signaling pathways: Notch, Fgf and Wnt (Dequeant et al., 2006; Dequeant and Pourquie, 2008). Among these pathways, Fgf and Wnt are implicated in the maintenance of immature PSM cells (Aulehla et al., 2003; Aulehla et al., 2008; Wahl et al., 2007; Delfini et al., 2005; Niwa et al., 2007), whereas Notch signaling might be directly involved in the generation of periodicity (Oginuma et al., 2008; Yasuhiko et al., 2006; Takahashi et al., 2000; Takahashi et al., 2003). In mice, Notch signal oscillations are produced by the suppressive function of the glycosyltransferase lunatic fringe (Lfng) as the levels of activated Notch1 (cleaved form of the Notch1 intracellular domain, referred to as cNICD hereafter) are upregulated in the *Lfng*-null mouse embryo (Morimoto et al., 2005). The expression of *Lfng* exhibits a biphasic pattern involving oscillation in the posterior PSM and a stabilized striped pattern in the anterior PSM (Aulehla and Johnson, 1999; McGrew et al., 1998; Morales et al., 2002; Cole et al., 2002). The oscillatory expression of *Lfng* is positively regulated by Notch signaling as it is greatly downregulated in *Dll1*-null mice, whereas it is negatively regulated by Hes7 as revealed by its upregulation in *Hes7*-null embryos (Barrantes et al., 1999; Bessho et al., 2003; Morales et al., 2002). The stabilized expression of *Lfng* is under the control of the Mesp2 transcription factor and stabilization does not occur in the absence of Mesp2 (Morimoto et al., 2005). In the absence of Lfng, no clear segmental border is defined and the rostral-caudal (R-C) compartmentalization within a somite is randomized (Zhang and Gridley, 1998; Evrard et al., 1998).

In the anterior PSM, the Mesp2 transcription factor plays an important role in the creation of a cNICD on/off state that corresponds to the future segmental boundary via the activation of Lfng transcription (Morimoto et al., 2005). This suggests that the Notch on/off state is important for boundary formation. However, a recent study has shown that mouse embryos expressing Notch activity throughout the PSM still show the ability to form more than ten somites, indicating that the Notch on/off state is dispensable for boundary formation (Feller et al., 2008). By contrast, however, other studies have reported that transgenic mice expressing *Lfng* only in the anterior PSM show normal segmental border formation after embryonic day 10.5 (E10.5), suggesting that the Notch on-off state generated in the anterior PSM is sufficient to create a somite boundary at least in the later stage embryos (Shifley et al., 2008; Stauber et al., 2009).

To resolve this discrepancy, we have, in our current study, generated a mouse that lacks the anterior striped Lfng expression pattern, but at the same time retains oscillating Lfng activity in the

[1]Department of Genetics, SOKENDAI, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. [2]Division of Mammalian Development, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan. [3]Division of Cellular and Molecular Toxicology, National Institute of Health Sciences, 1-18-1 Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan. [4]Cell Architecture Laboratory, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan.

*These authors contributed equally to this work
†Author for correspondence (ysaga@lab.nig.ac.jp)

Accepted 17 February 2010

posterior PSM. The resulting transgenic mouse shows no clear cNICD on/off state in the anterior PSM. Nevertheless, this mouse exhibits normal boundary formation, indicating that the cNICD boundary is dispensable for somite formation. In addition, our transgenic mouse shows normal R-C patterning within a somite. Further analyses by computer simulation have led us to conclude that Notch signaling oscillation functions as an output signal that is both required and sufficient to establish the *Mesp2* expression pattern needed for normal somitogenesis.

MATERIALS AND METHODS

Animals

The wild-type mice used in this study were the MCH strain (a closed colony established at CLEA, Japan). The *Lfng*-null (Evrard et al., 1998), *Mesp2*-null (*Mesp2$^{MCM/+}$*) (Takahashi et al., 2007) and *Mesp2-lacZ* (*Mesp2$^{lacZ/+}$*) (Takahashi et al., 2000) mouse lines are maintained in the animal facility of the National Institute of Genetics and National Institute of Health Sciences, Japan.

Gene targeting strategy to generate the *Mesp2^{Lfng}* allele

The knock-in strategy used to target the *Mesp2* locus is largely similar to our previously described method (Takahashi et al., 2000), except that *Lfng* cDNA was inserted. The *pgk-neo* cassette flanked by a *lox* sequence was removed by crossing with *CAG-Cre* mice (Sakai and Miyazaki, 1997).

Generation of the *Hes7-Lfng* transgenic mice

We used a 12 kb *Hes7* gene cassette comprising 5 kb of upstream sequence and all of the exons and introns, as this construct had previously been confirmed to be sufficient to reproduce the endogenous *Hes7* oscillation pattern when inserted in-frame at the translational start site (Kageyama et al., personal communications). We generated the construct *Lfng IRES-EGFP*, in which *IRES* (internal ribosomal entry site)-*EGFP* (enhanced GFP) was fused to the 3′ end of *Lfng* cDNA, and inserted this construct into the *Hes7*-translational initiation site. The resulting DNA was digested with restriction enzymes to remove vector sequences and gel purified. Transgenic mice were generated by microinjection of this construct into fertilized eggs, which were then transferred into the oviducts of pseudopregnant foster females.

In situ hybridization, immunohistochemistry, histology and skeletal preparations

The methods used for wholemount in situ hybridization, section in situ hybridization, immunohistochemistry, histology and skeletal preparation by Alcian Blue/Alizarin Red staining are described in our previous reports (Morimoto et al., 2005; Oginuma et al., 2008; Takahashi et al., 2000). The cNICD signal was detected by immunohistochemistry using anti-cleaved NICD (Val1744; 1:500; Cell Signaling Technology). Probes were prepared also as described previously: *Mesp2* exon-intron (Oginuma et al., 2008), *Mesp2* (Takahashi et al., 2000) and *Lfng* (Evrard et al., 1998). The *GFP* cRNA probe was prepared by PCR-amplification of *GFP* cDNA.

Computer simulation

Our computer simulation model is based on the previous mathematical description of a clock-and-wavefront model constructed by J. Lewis (Palmeirim et al., 1997). By using the basic oscillating function in the Lewis model, we modeled the activity of cNICD, *n*, at given time, *t*, and anteroposterior position, *x*, as:

$$n(x,t) = \left[1 - \cos\left\{2\pi \int_0^t \frac{1}{1+e^{(x+t)/2}}\,dt\right\}\right] / 2 .$$

For the control simulation with constant activity of cNICD, the cNICD activity, *n*, was set to 0.3. For the simulation with oscillating cNICD without wave, *n* was formulated as $n(x,t)=\{1-\cos(\pi t)\}/2$. The activity of Fgf8 is known to gradually decrease from posterior to anterior, and also according to the time elapsed. These features of Fgf8 fit well with the formulation of the clock cycling rate in the Lewis model and, thus, we calculated the activity of Fgf8, *f*, using the formula $f(x,t)=1/(1+e^{(x+t)/2})$.

We next added the regulation of Mesp2 and Tbx6 expression to the model. As cNICD and Fgf8 play positive and negative roles for *Mesp2* expression, respectively, we assumed that the increase of *Mesp2* expression occurs when the cNICD activity, *n*, exceeds that of Fgf8, *f*, with the amount dependent on *n–f*. Tbx6 (*b*) is also required for *Mesp2* expression. We thus modeled the *Mesp2* mRNA expression, *m*, and Mesp2 protein expression, *p*, as:

$$m(x,t+\Delta t) = m(x,t) + S_m \times \frac{[\{n(x,t)-f(x,t)\}/K_n]^{H_n}}{1+[\{n(x,t)-f(x,t)\}/K_n]^{H_n}}$$
$$\times \frac{\{b(x,t)/K_b\}^{H_b}}{1+\{b(x,t)/K_b\}^{H_b}} - D_m \times m(x,t) ,$$
$$p(x,t+\Delta t) = p(x,t) + S_p \times m(x,t-T) - D_p \times p(x,t) ,$$

with the initial condition $m(x,0)=0$, and $p(x,0)=0$. The degradation of Tbx6 is dependent on Mesp2 (Oginuma et al., 2008). We introduced a hypothetical molecule, *z*, that is expressed depending on Mesp2 and degrades Tbx6 by interacting with it. The expression of Tbx6 (*b*) and the Tbx6 degrading molecule (*z*) were modeled as:

$$z(x,t+\Delta t) = z(x,t) + S_z \times \frac{\{p(x,t-T)/K_p\}^{H_p}}{1+\{p(x,t-T)/K_p\}^{H_p}} - D_z \times z(x,t) ,$$
$$b(x,t+\Delta t) = b(x,t) - D_b \times \{b(x,t)\}^{Bb} \times \{z(x,t)\}^{Bz} ,$$

with the initial condition $z(x,0)=0$, and $b(x,0)=1.0$.

These formulas were implemented using C language. The activities of cNICD (*n*), Fgf8 (*f*), Mesp2 (*m* and *p*), Tbx6-regulator (*z*) and Tbx6 (*b*) were calculated over the ranges $-12.5 \leq x \leq -2.5$ and $0 \leq t \leq 20$. The calculations were conducted discretely with a single unit of *x* (Δx) of 1/10 and *t* (Δt) of 1/10. The parameter values we used are shown in Table S1 in the supplementary material. We also introduced time delay, $T=2\Delta t$, for protein expression (Lewis, 2003), which did not affect the results much.

RESULTS

Dissection of the *Lfng* expression pattern in the PSM

To examine the significance of the Notch on/off state during boundary formation, we focused on *Lfng* expression, which exhibits a biphasic pattern involving oscillation in the posterior PSM and a stabilized striped pattern in the anterior PSM (Aulehla and Johnson, 1999; Cole et al., 2002; McGrew et al., 1998; Morales et al., 2002). Each of these two patterns is implicated in the generation of the corresponding Notch activity profile via negative regulation. To induce only the oscillatory expression of *Lfng*, we utilized the *Hes7* transcriptional regulatory unit as the oscillation of *Lfng* and *Hes7* is regulated by similar factors, i.e. positively by Notch signaling and negatively by Hes7 protein. As shown in Fig. 1, these two transcripts show similar expression patterns in the oscillation phase. Both signals manifest a waved pattern within the Tbx6 expression domain from phase I to phase III (Fig. 1A-L). However, in phases I-II, *Hes7* expression is lost from the anterior domain (Fig. 1G-J), whereas that of *Lfng* persists for a longer period in the anterior PSM and forms a clear stripe (Fig. 1A-D,M,O). It should also be noted that the anterior *Lfng* expression domain was found to merge with that of the Mesp2 protein (Fig. 1N,P), the expression of which is restricted to the anterior PSM. This is not unexpected as *Lfng* expression is induced by Mesp2 in the anterior PSM and creates the Notch on/off state (Morimoto et al., 2005). Taken together, we concluded from these data that the *Lfng* expression pattern can be reproduced by two distinct regulatory systems – the *Hes7* promoter-enhancer and the *Mesp2* regulatory system – and this enabled us to further investigate the significance of Notch activities.

The cNICD on/off state is not required for somite boundary formation

To further elucidate the functional significance of the oscillatory cNICD in the posterior PSM and that of the cNICD on/off state in the anterior PSM, we generated a transgenic mouse line by inserting *Lfng* cDNA flanked with *IRES-EGFP* under the control of the Hes7 promoter (see Fig. S1A in the supplementary material). As expected, the expression pattern of this transgene, examined by in situ hybridization using *EGFP* as a probe, was found to be very similar to that of endogenous *Hes7* and *Lfng* except for the lack of anterior striped expression (see Fig. S1B-D in the supplementary material). We then introduced this transgene into the *Lfng*-null genetic background to establish the *Hes7>Lfng/Lfng*$^{-/-}$ mouse line and examined the expression pattern of exogenous *Lfng* and cNICD expression in the absence of endogenous *Lfng* expression (i.e. an *Lfng*-null background). In wild-type embryos, *Lfng* and cNICD showed biphasic patterns, these being oscillation in the posterior PSM and stabilization in the anterior PSM, whereas cNICD oscillation was barely detectable and a constant level of cNICD could be observed through the entire PSM in the absence of *Lfng*, as reported previously (Morimoto et al., 2005). In the *Hes7>Lfng/Lfng*$^{-/-}$ embryo, however, we observed the recovery of cNICD oscillation in the posterior PSM, which overlapped with *Lfng* expression (Fig. 2A-F), clearly indicating that the *Lfng* transgene was functionally active in these embryos. In addition, we previously showed that cNICD and Mesp2 generate a clear boundary in the anterior PSM, which demarcates the presumptive segmental border in phase-II embryos (Morimoto et al., 2005) (Fig. 2G-I). In the absence of *Lfng*, this clear border between cNICD and Mesp2 was not generated and a merged pattern was instead observed

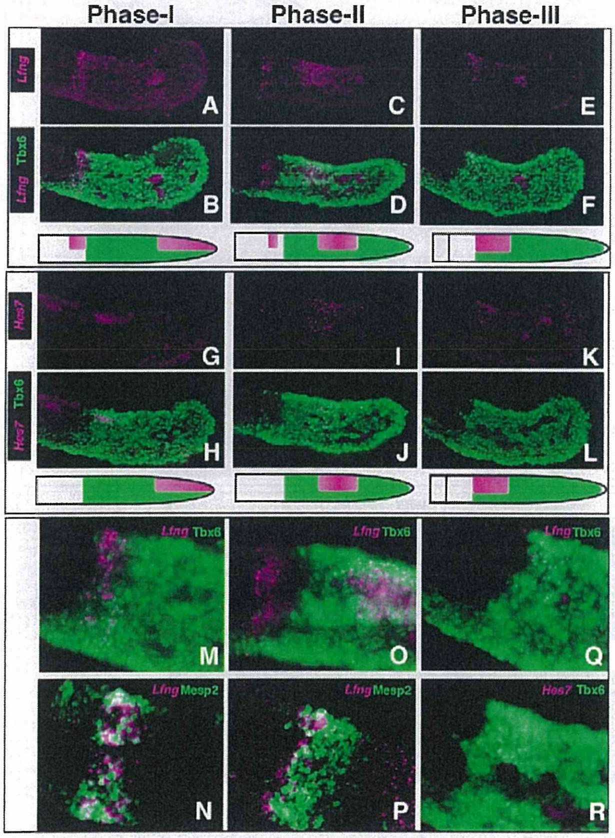


Fig. 1. Comparison of the *Hes7* and *Lfng* expression patterns. In situ hybridization analysis of the spatiotemporal changes in the *Lfng* (**A-F**) and *Hes7* (**G-L**) transcription patterns during somitogenesis by double staining for the Tbx6 protein as the reference point. The stained sections shown in the vertical rows are derived from a single embryo. The phase was defined by the location of the *Hes7* and *Lfng* transcripts and the waves of oscillating *Hes7* and *Lfng* were initiated at the posterior PSM (Phase I). The oscillating wave then moves to the intermediate PSM (Phase II) and reaches the anterior PSM (Phase III). (**M,O,Q,R**) Magnified images of B, D, F and L, respectively. Phase I and Phase II sections were also subjected to double staining for *Lfng* mRNA and Mesp2 (**N,P**).

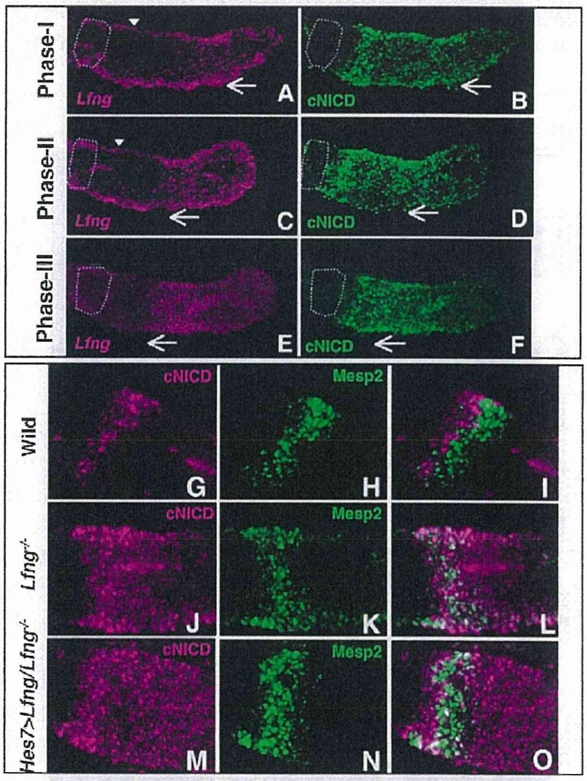


Fig. 2. *Hes7>Lfng/Lfng*$^{-/-}$ mice show cNICD oscillation in the posterior PSM but do not form a cNICD boundary in the anterior PSM. (**A-F**) The patterns of *Lfng* mRNA (A,C,E) and cNICD (B,D,F) expression were revealed in each channel by double staining of these signals using single embryos of *Hes7>Lfng/Lfng*$^{-/-}$ mice at three different phases, I-III, respectively. *Lfng* expression shows a traveling wave (arrow) but no stabilized stripe (arrowheads, A,C). The first somite is indicated by a white dotted line. The wave of oscillating cNICD is initiated at the posterior PSM (B; Phase I; n=3), moves to the intermediate PSM (D; Phase II; n=4) and eventually reaches the anterior PSM (F; Phase III; n=3). (**G-O**) The relationship between cNICD and Mesp2 in Phase II was compared among wild-type (G-I), *Lfng*$^{-/-}$ (J-L) and *Hes7>Lfng/Lfng*$^{-/-}$ (M-O) embryos by double staining. Single channels for cNICD (G,J,M) and Mesp2 (H,K,N), and merged images of both (I,L,O), are shown. In the wild-type embryos, cNICD and Mesp2 generate a clear boundary (I). *Lfng*$^{-/-}$ and *Hes7>Lfng/Lfng*$^{-/-}$ mice, however, do not show a clear segregation between cNICD and Mesp2 (L,O).

(Fig. 2J-L). In the *Hes7>Lfng/Lfng*$^{-/-}$ embryo, as expected by the lack of Lfng expression in the anterior PSM, we did not detect segregation between the cNICD and Mesp2 domains (Fig. 2M-O). *Lfng*$^{-/-}$ embryos did not show clear somite boundaries, although incomplete somites did appear to be formed (see Fig. S2 in the supplementary material), as also suggested previously (Evrard et al., 1998; Zhang and Gridley, 1998). Very surprisingly, however, *Hes7>Lfng/Lfng*$^{-/-}$ embryos showed clearly segmented somites (Fig. 3A-C). This strongly indicates that the oscillatory expression of cNICD mediated via oscillating Lfng is sufficient to provide the conditions for normal somitogenesis to occur and that the cNICD boundary in the anterior PSM is not required for this process.

Recently, we and others have suggested that the Mesp2 downstream events, such as the activation of ephrin-EphA4 signaling and the formation of a Tbx6 protein boundary, were more important for segmental border formation (Watanabe et al., 2009; Oginuma et al., 2008; Nakajima et al., 2006). In *Lfng*$^{-/-}$ embryos, the expression of *EphA4* and the Tbx6 protein boundary were found to be diffuse or randomized (Fig. 3E,J-L), whereas in *Hes7>Lfng/Lfng*$^{-/-}$ embryos, these expression patterns appeared to be normal (Fig. 3F,M-O), i.e. similar to those in wild-type embryos (Fig. 3D,G-I). Taken together, our current findings show that the cNICD boundary is dispensable, but that the Mesp2 boundary might be required, for the creation of the segmental border through the regulation of downstream genes.

R-C polarity is completely recovered in *Hes7>Lfng/Lfng*$^{-/-}$ embryos

We next further examined the morphological features of the *Hes7>Lfng/Lfng*$^{-/-}$ embryo. Surprisingly, these transgenic embryos showed a completely normal skeletal system, with segmented vertebra and ribs (Fig. 4A-C). Furthermore, the expression pattern of *Uncx4.1*, a caudal marker of R-C polarity (Fig. 4D), was fully recovered in the *Hes7>Lfng/Lfng*$^{-/-}$ embryo (Fig. 4F), which contrasts with the randomized pattern we observed in the *Lfng*$^{-/-}$ embryo (Fig. 4E). These results suggest that the cNICD boundary in

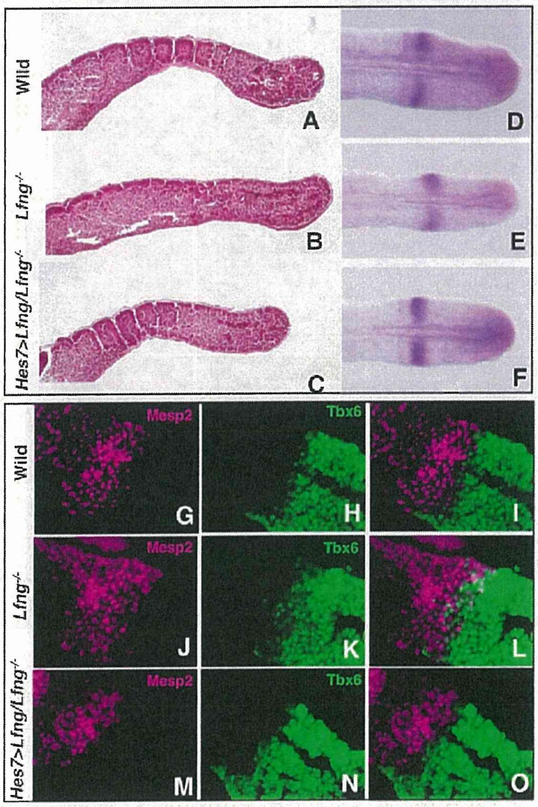


Fig. 3. Exogenous *Lfng* controlled by the *Hes7* promoter completely rescues the boundary formation defect in the *Lfng*$^{-/-}$ mice. The segmental morphologies (**A-C**), the *EphA4* expression pattern (**D-F**) and the relationship between Mesp2 and Tbx6 in Phase II (**G-O**) were compared among wild-type (A,D,G-I), *Lfng*$^{-/-}$ (B,E,J-L) and *Hes7>Lfng/Lfng*$^{-/-}$ (C,F,M-O) using E11.5 embryonic tail regions. Single channels for Mesp2 (G,J,M) and Tbx6 (H,K,N), and merged images of both (I,L,O), are shown. Expression of the *EphA4* and Tbx6 protein boundary forms a clear border in the wild-type (D, *n*=7; G-I, *n*=4) and *Hes7>Lfng/Lfng*$^{-/-}$ embryos (F, *n*=4; M-O, *n*=4), but this is diffuse or randomized in the *Lfng*$^{-/-}$ embryos (E, *n*=4; J-L, *n*=3).

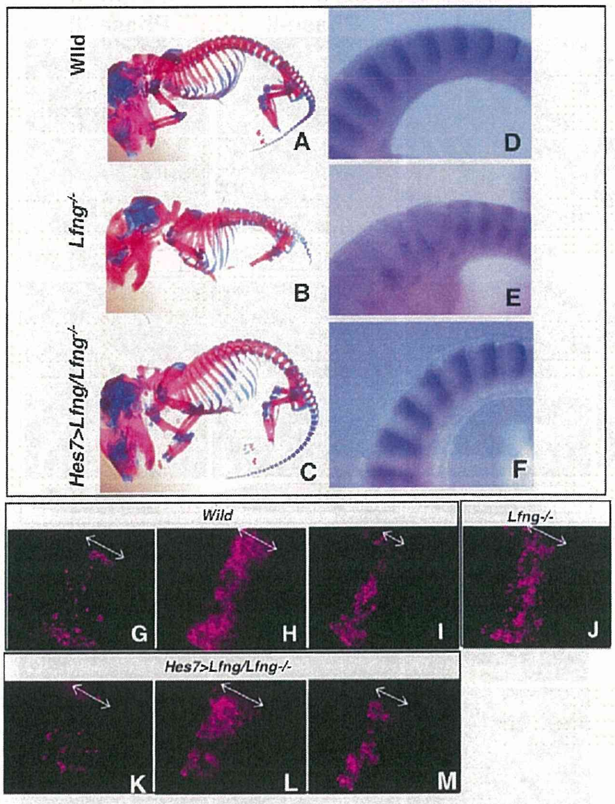


Fig. 4. Exogenous *Lfng* under the control of the *Hes7* promoter completely rescues the R-C patterning defect in the somites of *Lfng*$^{-/-}$ mice. A comparison of the segmental morphologies of skeletal preparations of E17.5 embryos (**A-C**) and the expression pattern of *Uncx4.1*, indicative of R-C patterning within a somite (**D-F**). *Hes7>Lfng/Lfng*$^{-/-}$ mice show a normal skeleton (C, *n*=4) and expression pattern of *Uncx4.1* (F, *n*=3), whereas *Lfng*$^{-/-}$ mice show randomized pattern of skeleton (B) and *Uncx4.1* expression (E). (**G-M**) *Mesp2* transcription states revealed by high resolution in situ hybridization analysis of wild-type embryos for transcriptional initiation (G, *n*=3), active state (H, *n*=5) and rostral localization (I, *n*=3), and *Lfng*-null (J, *n*=11) and *Hes7>Lfng/Lfng*$^{-/-}$ embryos for transcriptional initiation (K, *n*=2), active state (L, *n*=3) and rostral localization (M, *n*=3). Double arrows indicate the length of the *Mesp2* transcription domains.