BBA - Gene Regulatory Mechanisms 1865 (2022) 194768

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

# The gene regulation knowledge commons: the action area of GREEKC

Martin Kuiper [a,*], Joseph Bonello [b], Jesualdo T. Fernández-Breis [c], Philipp Bucher [d], Matthias E. Futschik [e], Pascale Gaudet [f], Ivan V. Kulakovskiy [g], Luana Licata [h], Colin Logie [i], Ruth C. Lovering [j], Vsevolod J. Makeev [k], Sandra Orchard [l], Simona Panni [m], Livia Perfetto [n], David Sant [o], Stefan Schulz [p], Steven Vercruysse [a], Daniel R. Zerbino [q], Astrid Lægreid [r], The GRECO Consortium [1]

[a] Systems Biology Group, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway
[b] Faculty of Information & Communication Technology, University of Malta, Msida, Malta
[c] Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, CP 30100, Murcia, Spain
[d] Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Amphipôle, 1015 Lausanne, Switzerland
[e] Systems Biology and Bioinformatics Laboratory (SysBioLab), Centre of Marine Sciences (CCMAR), University of Algarve, 8005-139 Faro, Portugal
[f] SIB Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1204 Geneva, Switzerland
[g] Institute of Protein Research, Russian Academy of Sciences, 142290, Institutskaya 4, Pushchino, Russia
[h] Department of Biology, University of Rome Tor Vergata, Rome, Italy
[i] Department of Molecular Biology, Faculty of Science, Radboud University, PO box 9101, Nijmegen 6500HG, the Netherlands
[j] Functional Gene Annotation, Pre-clinical and Fundamental Science, Institute of Cardiovascular Science, University College London, 5 University Street, London WC1E 6JF, UK
[k] Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, Gubkina 3, Moscow, Russia
[l] European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[m] Department DIBEST, University of Calabria, Rende, Italy
[n] Fondazione Human Technopole, Department of Biology, Via Cristina Belgioioso, 171, 20157 Milan, Italy
[o] Department of Biomedical Informatics, University of Utah, 421 Wakara Way #140, Salt Lake City, UT 84108, United States
[p] Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerpl. 2, Graz, Austria
[q] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[r] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway

ARTICLE INFO

*Keywords:*
Knowledge Commons
Biocuration
Ontologies
Computational biology
Text mining
Web services

ABSTRACT

As computational modeling becomes more essential to analyze and understand biological regulatory mechanisms, governance of the many databases and knowledge bases that support this domain is crucial to guarantee reliability and interoperability of resources. To address this, the COST Action *Gene Regulation Ensemble Effort for the Knowledge Commons* (GREEKC, CA15205, www.greekc.org) organized nine workshops in a four-year period, starting September 2016. The workshops brought together a wide range of experts from all over the world working on various steps in the knowledge management process that focuses on understanding gene regulatory mechanisms. The discussions between ontologists, curators, text miners, biologists, bioinformaticians, philosophers and computational scientists spawned a host of activities aimed to standardize and update existing knowledge management workflows and involve end-users in the process of designing the Gene Regulation Knowledge Commons (GRKC). Here the GREEKC consortium describes its main achievements in improving this GRKC.

## 1. Introduction

Understanding how complex biological systems operate is not possible without computational modeling of data, information and knowledge. In fact, biological knowledge discovery itself is becoming increasingly dependent on computational modeling and simulation. The

* Corresponding author.
*E-mail address:* martin.kuiper@ntnu.no (M. Kuiper).
[1] The GRECO authors: see table at end of paper.

https://doi.org/10.1016/j.bbagrm.2021.194768
Received 20 April 2021; Received in revised form 18 October 2021; Accepted 20 October 2021
Available online 30 October 2021

construction of computer models requires comprehensive knowledge of biological entities and their interactions, and abundant efforts are dedicated to providing such information in databases [1–3]. Despite all this, multidisciplinary collaborations between stakeholders that represent the different expert areas necessary to specify and design the various knowledge domains, formats, content and access (the knowledge life cycle) have been scant, explaining why many of these valuable knowledge domains have remained only modestly interconnected.

The analysis of gene regulation mechanisms is of high importance to systems approaches because it is key to understanding how information in the genome governs cellular differentiation and function. The complex machinery that determines which genes are active requires a dynamic interplay between different types of transcription factors, the DNA regions where they engage in gene-specific transcription regulation, and the specific epigenetic context that affect the accessibility of these regions. Progress to comprehensively improve knowledge repositories that provide detailed information about each of these types of gene regulators and their causal interactions, needs input from expert groups that may not normally interact or collaborate.

The European Cooperation in Science and Technology (COST) Action *Gene Regulation Ensemble Effort for the Knowledge Commons* (GREEKC) is the result of an initiative which started in 2013: The Gene Regulation Consortium (GRECO, www.theGRECO.org). GRECO acquired funding from COST in 2016, allowing us to commence on a four-year journey using the different COST mechanisms (most importantly: Workshops and Working Group meetings, Training Schools and Short Term Scientific Missions). The main aim of GREEKC was to advance the coordinated building of the Gene Regulation Knowledge Commons (GRKC). This GRKC is defined by the GREEKC consortium as: "The collection of freely accessible gene regulation information resources, containing information that is well annotated with unambiguous descriptors according to quality criteria and standards that allow seamless integration and interoperability as well as automated computational access with third-party software".

From September 2016 to March 2021, GREEKC organized a series of workshops to discuss and assess efforts to produce and exploit 'knowledge' pertinent to this domain. In doing this, we followed a Responsible Research and Innovation (RRI) approach, which Schomberg [4] defined as engaging all stakeholders to optimize the deliverables of a scientific process, and to align scientific processes and outcomes to societal needs. The GREEKC consortium took this strategy as an iterative process of identifying and including stakeholders in discussions about, for instance, data curation or data sharing issues, starting with key players in the knowledge life cycle [5]. This RRI approach proved to be an extremely good fit with the main mechanism of COST Actions for facilitating discussions and establishing multidisciplinary partnerships to achieve scientific progress.

## 2. GREEKC field of operation and design

Gene regulatory mechanisms involve a complex interplay of many molecules and their causal relationships, some of which are described in Fig. 1. Different classes of biomolecules (Protein, RNA and DNA), acting often in multi-molecular complexes, are responsible for processes that enable transcription (e.g., accessibility of regulatory sequences at the DNA), drive transcription (DNA binding Transcription Factors (dbTF) and transcription cofactors (co-TF) in complexes bound to DNA) and support the use of transcripts for protein biosynthesis.

The research in this area has resulted in a wealth of information and knowledge, available in scientific publications and as large-scale

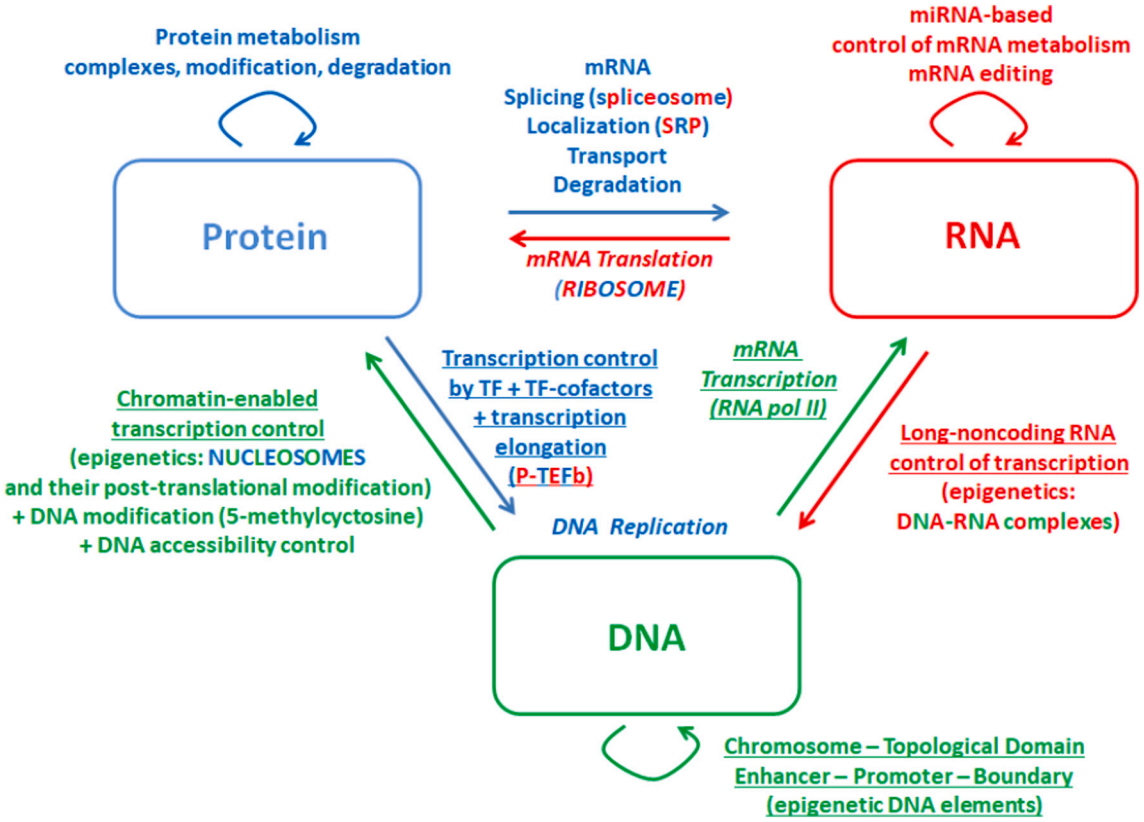


**Fig. 1.** Schematic overview of gene regulation processes. The regulation of gene expression involves the three genetically encoded polymer classes; DNA (green), RNA (red) and Protein (blue). Complexes involving different polymers, are described with mixed letter colors (RNA-PROTEIN COMPLEXES, DNA-PROTEIN COMPLEXES, DNA-RNA COMPLEXES); Underlined biological processes denote DNA-centric transcription control; Francis Crick's central dogma [66] is shown in italics. These various entities, complexes and processes represent the area of interest for the Gene Regulation Knowledge Commons.
(Courtesy of C. Logie).

datasets.

Yet, scientific results cannot be effectively shared for computational use through publications or data repositories alone. The information content of publications needs to be carefully checked, or curated, and archived in standardized formats in publicly available resources, if it is to become broadly available for computational integration and analysis [6,7]. Similarly, large-scale data must be curated and archived with proper metadata to provide well annotated resources for obtaining knowledge through computational processing and integration with other information sources.

Central to this value creation is the biocurator, who is typically an expert in a biology or bioinformatics domain. A trained biocurator is able to identify and characterize specific biological entities and interactions described in papers or large-scale data repositories and can investigate their contents for experimental or other evidence that supports particular claims about their biological function. These claims are described, or annotated, with the help of controlled vocabularies that provide standardized terms, descriptions and definitions for concepts that are relevant for a (sub)domain of biology. Domain Ontologies consist of machine-processable formal axioms and definitions of types of domain entities, hierarchically organized so that they can facilitate analysis at different levels [8–10] thus constituting the building blocks for representing human knowledge [11]. Describing biological entities and their relationships in specific contexts with the help of unambiguously defined ontology terms is performed in annotation workflows that follow well-defined curation guidelines, so that different biocurators are able to interpret and annotate knowledge from a paper in identical ways. Their work is supported by curation tools, which often provide additional guidance as to the annotation details that need to be provided. There are many subdomains of biology that require such annotation efforts. The focus of the GREEKC consortium has been the area of gene regulatory mechanisms (Fig. 1), but their efforts in developing knowledge gathering and sharing principles likely has value across all biological domains.

In addition to the curation of the various sources of information relevant for gene regulatory mechanisms, two other technology areas are also relevant to consider: text mining and data sharing. The curation of information from scientific literature starts with the identification of papers that have curatable information. Finding such content can be facilitated by text mining algorithms that identify and mark paper sections appropriate for subsequent manual curation. However, whereas the potential of text mining for assisting manual curation is well-established, its direct integration into curation workflows has not yet been widely adopted. For those curation workflows that produce information relevant to the GRKC, the breadth of annotation detail impacts their representation, storage in a database schema and subsequent sharing mechanisms. For instance, annotations need to meet well-defined curation guidelines and storage formats, and stored data require specific 'exchange languages' (e.g. based on XML or JSON) for downloads or web services.

Taken together, these different elements of the gene regulation knowledge management life cycle served to formulate four challenges that were addressed by four working groups of the GREEKC COST Action:

WG1: The development and maintenance of ontologies and controlled vocabularies;
WG2: The development of curation guidelines and workflows for the annotation of gene regulators at different levels:
a. protein level
b. non-coding RNA level
c. nucleotide sequence recognition level (e.g. transcription factor binding sites)
d. genome level (DNA methylation status, histone modifications)
e. level of interactions, regulatory complexes and network information flow;
WG3: The exploration of text mining to identify or extract information useful for annotation of gene regulators and to facilitate the identification of literature evidence that can be used to annotate regulatory molecular entities and their regulatory interactions;
WG4: The storing and sharing of annotations of gene regulatory interactions.

## 3. Ways of working and accomplishments

COST Actions can organize the scientific domain, stimulate discussions, strive for consensus and achieve progress [12] through the organization of Workshops, Training Schools and Short Term Scientific Missions (STSMs). This paper elaborates on the results of the workshops and some of the STSMs, as they have been most instrumental in generating new ideas and consensus about approaches to develop the structure and add content to the GRKC.

While biocuration and annotation efforts relevant for the GRKC have been the central topics of GREEKC workshops, many times discussions also involved the need for improving ontologies and controlled vocabularies as well as text mining for gene regulation knowledge management. This means that much of what we achieved in GREEKC cannot be uniquely assigned to one particular Working Group but rather to the joint efforts of all groups.

### 3.1. WG1: ontologies

Bio-ontologies form the semantic framework for the annotation of what we know and understand about the function of biological entities and their interrelationships. Both the Gene Ontology (GO), [13,14] and the Sequence Ontology (SO) [15] are central to the description of chromatin, gene, protein and RNA components involved in gene regulatory events.

The development and maintenance of ontologies is intrinsically linked to established annotation processes and refinements thereof to keep up with evolving biological insights. Significant efforts have been made by GREEKC members to improve the annotation quality of the class of mammalian DNA binding transcription factors (dbTF) (Lovering et al., 2021, this issue), and, as a consequence, the GO molecular function subtree describing the regulation of gene expression by RNA polymerase II annotation has also undergone major restructuring (Gaudet et al., 2021, BBAGRM-D-21-00006, this issue). In addition, the SO has been critically reviewed. In several workshops, GREEKC members talked with the external experts responsible for constructing and using the SO, and arrived at a consensus on restructuring the part of SO that specifies the description of Gene Regulatory Elements within the genome. Since the original conception of the SO, the knowledge about the nature of gene regulation and the importance of the binding of proteins to regulatory control elements in the genome (most importantly the dbTFs) has advanced considerably and has revealed an abundance of transcription factor binding sites at multiple gene regulatory locations in the genome. In addition, the new notion of Topologically Associating Domains (TADs) was not yet supported by the SO and a restructuring has now been proposed (Sant et al., 2021, this issue) to align the definition and hierarchy of the SO regulatory element subtree with our current understanding of the full breadth of protein-DNA interaction events and chromatin conformation states that impact gene expression. Finally, efforts have been launched to follow up on the Gene Regulation Ontology [16], proposed as an application ontology for capturing broadly the entities and relationships that are essential for describing gene regulation at multiple levels (protein, RNA, small molecule, genome, DNA level and epigenetic level). The concept of the Gene Regulation Application Ontology (GRAO, https://github.com/greekc) provides an ontology framework for a knowledge base able to semantically integrate all available knowledge about gene regulatory events, allowing for complex queries addressing many aspects about regulatory context simultaneously, going well beyond the examples published for

the Gene Expression Knowledge Base [17].

### 3.2. WG2: curation guidelines

Biocuration involves a manual or computational assessment of the validity of a particular claim that may characterize a biological entity or relation, upon which this claim can be specified with the help of proper entity identifiers (IDs), ontology terms, evidence descriptions and provenance, e.g. the identifier of the publication based on which the biocurator made the assertion. It is the central process that generates knowledge base content that provides users with high quality information. GREEKC has addressed five different subdomains of biocuration in its workshops and in several areas notable progress was made:

#### 3.2.1. The protein level

GREEKC members have collaborated on the task of bringing together the knowledge that currently supports the classification of proteins as dbTFs (Lovering et al., 2021, this issue). The central role of these proteins in linking the cellular signaling machinery to the decoding of the regulatory genome has made them a prime focus of dedicated characterization and curation efforts over the years and the GREEKC review drove the re-design of the GO transcription regulation molecular functions branch and an updated set of curation guidelines (Gaudet et al., 2021, BBAGRM-D-21-00006, this issue). The updated GO transcription regulation branch also encompasses improvements in the GO structure and terms for co-transcription factors (coTFs) and general transcription factors (GTFs) and thus provides fertile ground for improved GO annotation of these protein entities with important roles in gene regulation.

#### 3.2.2. The RNA level

The gene regulatory network also includes RNA molecules that interact with proteins, with other RNAs or directly with genes to mediate their action. In the last decade, strong efforts have been launched to annotate both functional and physical RNA interactions in public repositories. While there were guidelines to use the Gene Ontology to capture the role of microRNAs in gene regulation [18], no specific guidelines had been developed for the majority of other RNA roles, with the result that knowledge extracted from one source is sometimes difficult to integrate or compare with other sources. Discussions among GREEKC members led to the definition of common standards for the annotation of microRNA-mRNA and microRNA-lncRNA interactions [19]. MicroRNAs are the best-characterized regulatory RNAs, and their binding partners can be predicted using bioinformatic approaches that map the interaction site to its target genes. However, as each prediction tool provides different sets of targets for each specific microRNA, the value of experimental confirmation of a microRNA-mRNA interaction should not be underestimated [20]. Meetings and round table discussions between members of the Working Groups 1 and 2 have led to recommendations for the annotation of interactions and ontologies focusing on microRNA regulatory mechanisms [19], and annotation guidelines have been tested through an STSM. However, we have yet to do the same for functional interactions of the lncRNAs with genes and their role in transcriptional regulation.

#### 3.2.3. The DNA level

Whilst the dbTFs represent the protein side of the decoding of genome information, their specific binding sites in the genome uniquely target dbTF regulatory activity to specific genes. Because of their importance, the transcription factor binding sites (TFBS) have been extensively studied to characterize their nucleotide patterns (sequence motifs) and determine features that define binding specificity [21]. A sequence motif recognized by a dbTF reflects the binding energy of a dbTF to a particular DNA segment [22], and there are many approaches to represent this relation in a computational model, from a basic consensus string to a 'black box' of advanced machine learning [23,24]. However, the gold standard is still defined by position weight matrices (PWMs) which were suggested as early as 1982 [25] and remain the most widespread and accepted way of describing dbTF binding specificity as a quantitative rather than a qualitative phenomenon [26]. PWMs are massively used to predict TFBS in the genome and annotate regulatory sequence variants [27,28]. Many TFBS motif discovery algorithms have been proposed over the years, and many experimental data sets were generated and analyzed, resulting in a multitude of motif collections, such as TRANSFAC [29], HOCOMOCO [30], CIS-BP [31], and JASPAR [32]. Creating a common understanding for how these PWMs should be used, represented, shared and interpreted was discussed in several workshops. As a result, a large-scale benchmarking was designed and carried out (aided by an STSM), resulting in a large set of publicly available performance measures that may improve the use of PWMs in practical analyses of new datasets [33].

#### 3.2.4. The genome level

The SO is an essential source of terms that, among others, describe sequence concepts necessary to annotate regulatory sequences and TFBS for a range of resources (e.g. Ensembl [34]). SO was improved by the restructuring of terms related to cis-regulatory modules (CRMs), which are regulatory regions where transcription factor binding sites are usually clustered to regulate various aspects of transcription. CRMs include enhancers, silencers, locus control regions, and insulators. A special type of CRM that was added to SO is the 'DNA_loop_anchor', which represents the ends of a DNA looping region. DNA looping is necessary to allow for areas of DNA that are separated by many kilobases to remain in close proximity within the cell, allowing CRMs to interact with distant genes [35]. Another set of updates to SO is the addition of terms related to topologically defined regions, which are areas where self-interaction of DNA occurs more frequently than expected by chance. An instance of self-interaction is a topologically associated domain, bordered by topologically associated domain boundaries. During interphase, DNA loop anchors are CCCTC-binding factor (CTCF) binding sites. Several studies have investigated CTCF binding to determine topologically defined regions [35].

#### 3.2.5. Level of interactions, regulatory complexes and network information flow

The annotation of proteins in the GO database is based on well-established guidelines [36], but the underlying data model and output, the Gene Product Association Data (GPAD) file, does not fully support all functional details about interactions between a protein and its interacting partners. One of the most significant shortcomings is caused by the limitation of the 'annotation extension' field in the tabular GPAD file. Target genes (TGs), and other protein interacting partners, bound by the transcription factor (dbTF) of interest, are captured in the annotation extension column but the result of transcription factor binding to a gene can only be summarized by the limited vocabulary of the annotation extension [37]. The GO-CAM data model [38] aims to remedy this, by allowing a biocurator to define linked annotations that use multiple ontologies to represent all aspects involved in biological functions involving multiple biological entities, essentially from a molecular function activity flow perspective. The GO-CAM approach has been discussed in several GREEKC workshops and its members have engaged in defining a set of templates in the Noctua curation tool that will guide a biocurator in the definition of new dbTF-TG interactions (Juanes Cortés et al., 2021, BBAGRM-D-21-00018, this issue).

Transcription factors often bind as homo-/heterodimers which then bind to co-factors to assemble the protein machinery required for transcription. GREEKC members (Velthuijs et al., 2021, this issue) used data from the IMEx Consortium databases [39] and BioGRID [40] to develop a pipeline to predict transcription factor coregulator complexes, which were subsequently validated using the CORUM (http://mips.helmholtz-muenchen.de/corum/) and hu.MAP (http://proteincomplexes.org) protein complex databases. Efforts to manually curate transcription factor and coregulator complexes in the Complex Portal database [41]

have also been inspired by the GREEKC Action.

The PSI-MI standards that have been developed under the umbrella of the Human Proteome Organization's Proteomics Standards Initiative (HUPO PSI) were the starting point [42–44] for discussions about future needs of the network modeling community. Although the existing data formats developed by this group were capable of describing TF-TG binding, the format was not designed to describe either the upstream dataflow from a cellular signaling pathway to an up- or down-regulation of a set of genes. GREEKC was able to organize several events together with the Proteomics Standards Initiative and ELIXIR to define an extension of HUPO-PSI MITAB2.7 that would cover the causality associated with (gene) regulatory interactions. The general importance of this type of interaction for the use in building conceptual and mathematical models of regulation networks called for a multidisciplinary agreement involving all relevant stakeholders (WG2 and WG4 members, many also active in the PSI-MI and ELIXIR community). This resulted in the definition of CausalTAB [45], which is also known as PSI MITAB2.8. The work on causal molecular interactions also exposed the need for a set of guidelines that describe the necessary and desirable contextual details that a user would need to find in order to be able to select and incorporate such causal statements in a model. These guidelines were created and are now published as the MI2CAST checklist [46], which has been endorsed globally by a broad group of biocurators, ontology developers, curation tool developers and users of molecular causal interaction statements. To the biocurator, the MI2CAST standard provides guidance in identifying contextual details that have to be minimally supplied in new annotations; to the curation tool developer it specifies the semantic resources and identifiers that should be chosen; to the user, the MI2CAST standard provides a summary of the contextual handles that are available for selecting proper data; and to the biological experimentalist, it defines the domain of study and reporting that will yield information most valuable for future computational integration and analysis. The MI2CAST standard has been implemented in the prototype curation tool causalBuilder [47], to illustrate how a Visual Syntax Markup (VSM-box) data entry template engine [48] can be used to support the presentation of an annotation standard in an organic way to a biocurator.

### 3.3. WG3: text mining for knowledge curation

The GREEKC consortium considered the value of text mining for aiding the curation workflow. These discussions have shown that the worlds of manual biocurators and text miners have many possible connections, but an active engagement where both sides benefit equally remains to be pursued. Text mining is an accepted method for triage, meaning the identification of e.g. a scientific paper that is likely to contain information that would satisfy a curation effort, implying it may contain the necessary information to warrant an annotation for a database. Conversely, curation is an accepted practice used to support text mining, both to assemble and prepare a text corpus that can be used for training of a text mining classifier, and for assessment of the quality of text mining results. But the results of manual curation (high-quality annotations of a limited subset of the available texts) and text mining (lower quality annotations of the widest possible range of texts) are unsatisfactory to the other expert group, which stands in the way of mutual efforts to marry the two without reservations. And to some extent the outcomes of both types of efforts may also serve different user communities: the high-quality curation resources serve the careful, cautious user, whereas the text mining result may serve the computational network analyst in settings where she is willing to accept that some of the information she is using may be of lower confidence than manually curated knowledge.

Several events have been organized by the text mining working group, but most notably the results from the collaboration between GREEKC members NTNU and BSC are worth mentioning. They have performed a text mining effort to specifically identify and retrieve gene regulatory interactions between a DNA binding transcription factor and a target gene (TF-TG relationships). The results (www.extri.org) were integrated and compared with several established curated resources with TF-TG relationships and indicate the sizable corpus of MedLine literature with information currently not represented in curated data resources (Vazquez et al., 2021, this issue). Moreover, they also indicate the potential gap of information pertaining to proteins currently not covered by functional studies reported in the literature, as about half of the putative dbTFs do not return any MedLine record of involvement in the regulation of a target gene. The ExTRI resource is available to the computational biologist through the BioGateway database and a Cytoscape app [49], and the potential problem of false positive records is mitigated by providing full provenance to the TRI sentence detected by text mining in its PubMed abstract, so that a user may check the validity and, if wrong, omit it from analysis.

### 3.4. WG4: databasing and sharing

The storing and sharing of curated information in databases provides the basis for dissemination of GRKC and thus has received particular attention in the GREEKC workshops. Among other issues, we were interested in the user perspective for GRKC and the standardization of information exchange. Regarding the former, we found that many commonly asked questions in gene regulation can be covered by a set of use cases (i.e. what are the known or predicted regulators of a gene?). For this reason, we have started to provide protocols for such use cases on the GREEKC website (https://www.greekc.org/use-cases). Regarding standardization and exchange of GRKC, the ELIXIR initiative has adopted criteria to assess the governance of knowledge bases and data repositories with the aim to identify Core Resources that comply with high governance and thus reliability standards. The identified Core Resources include several resources that contain information relevant for the GRKC, for instance GO, IntAct, UniProtKB and Ensembl. However, many additional valuable resources exist, making it imperative that careful consideration is given to make sure that their content is compliant with formats endorsed by ELIXIR Core resources and the FAIR principles. To assess the FAIRness of GRKC tools and datasets, a semi-automated tool was developed (Bonello et al., 2021, this issue) to score resources in terms of their compliance with the FAIR principles. Each principle is individually scored and a breakdown of the criteria is provided in a report generated by the scoring tool. The SIGNOR database, for instance, abides by the FAIR principles and was an early adopter of the PSI-MI standards endorsed by IMEx. The development of the CausalTAB / PSI.MITAB2.8 format described earlier poses new demands for data exchange mechanisms, most notably the webservice PSICQUIC (Protein Standard Initiative Common Query Interface [50]), which, at the time of writing, is only able to serve queries for the PSI-MI 2.7 format. The GREEKC discussions led to an STSM that resulted in a prototype PSICQUIC 1.0 webservice that has been implemented for communication with the SIGNOR database. Future work is needed to upgrade PSICQUIC web service functionality with common tools like Cytoscape [51], which supports the import of data through the Network from Public Databases / Universal Interaction Database Client. The MedLine extracted information on TF-TG interactions from the ExTRI text mining effort described above are available now through standard PSICQUIC web service (see http://www.ebi.ac.uk/Tools/webservices/psicquic/registry/registry?action=STATUS#, tfact2gene service). Other web services that provide access to TF-TG

interactions can be launched through Cytoscape Apps. The BioGateway App [49] uses SPARQL queries (SPARQL Protocol and RDF Query Language [52]) to fetch regulatory information from the semantic web database BioGateway [53], in the form of documented interactions between transcription factors and their target genes (see www.extri.org). Likewise, the OmniPath App [54] uses a REST type service to fetch TF-TG relationships from the dedicated transcription factor activity knowledge base DoRothEA [55].

## 4. Discussion and future challenges

An overview of the published results of the GREEKC COST Action is shown in Table 1. In each of the areas of the Action, results have been published, either as part of this BBA-GRM special issue or elsewhere.

In the discussions about bottlenecks and solutions to enhance the GRKC, the needs of two groups were considered: on the one hand bench biologists who access detailed information on particular genes and proteins of interest and how they interact, and on the other hand computational biologists who need an abundance of computationally accessible and well-structured information resources. This requires that the content of the GRKC is both 'human readable' and browsable through a web interface, and available through an API or web service, for computational processing. Regardless of their use, annotations need to be enhanced by including information with 'richer' expression of the functions of molecular entities, the relations between entities, the 'emergent' effect of their interactions, as well as experimental evidence and biological context so as to underpin and enhance the use of this information in regulatory network building and computational analysis. To achieve this, further improvements and innovations of curation approaches and tools will be needed, so that the annotation process of not only biological entities, but of their systems interactions becomes and remains manageable. The curation tool Noctua [38], and new experimental technologies like VSM [56] provide significant steps in the direction of annotating biological systems rather than biological entities. These tools accommodate multiple entities, activation state and relation types, and provide for annotations based on multiple ontologies and supported by an elaborate set of evidence and contextual metadata. Although at the semantic level sufficient resources may be available to cover these domains individually, integrated resources are needed that interlink and support complex queries for obtaining regulatory information that spans the different levels. The design of the Gene Regulation Application Ontology has paved the way to produce a prototype semantic knowledge base where GRKC information is integrated together with SO regulatory sequence concepts, information from the Complex Portal and GO molecular function and biological process terms to allow users to query for regulatory mechanism information that meets both location/sequence constraints, macromolecular assemblies and gene regulatory action constraints.

Users will also need the Knowledge Commons to be as comprehensive as possible. Current literature curation efforts are too limited to cope with the increasing amount of information published on a daily basis. Therefore, the access of information generated by text mining [57] as well as by automated and manual curation [58], needs to gain more attention. Furthermore, improvements are needed in the associated metadata so that it is clear to the user what the quality and inclusion criteria are for a particular piece of information [59]. Demanding computational users will then be able to implement their own selection criteria for incorporating data into their analysis. In practice this can help ameliorate a well-known challenge in digital knowledge management, which is that in their annotation work, biocurators generally focus on including only cases with strong evidence ('true positives') in their database content, and discard cases with weak evidence (including possible false negatives). Information that is not included in a resource may, upon closer inspection of additional or new data, find support from sufficient evidence to meet the database's inclusion criteria. Such information might be flagged by appropriate evidence codes, so that users may apply their own filters when exploring it either in a 'cautious' or 'greedy' mode (Chatterjee et al., this issue).

While modern sequencing technologies provide great power at low cost to detect transcriptional activity (e.g. by RNA-seq or Ribo-seq), or TF binding (e.g. by ChIP-seq) on a genome-wide scale, no experimental technology exists that comprehensively captures TF activity across the genome. Therefore, an area where further coordinated work is essential concerns the computational prediction of 'active' binding sites of transcription factors (including those of homo- and heterodimers) combining evidence from multiple experimental, often large-scale data sources, to infer transcription factor-target gene interactions. For more than 30 years, efforts of decoding a "regulatory code of transcription factors" have been undermined by the notorious ability of transcription factors to recognize quite dissimilar DNA sequences depending on the availability of different protein partners for complex formation and local and overall chromatin accessibility profiles. Yet, massive efforts in comparative studies of dbTF binding in vitro and in vivo in a variety of cell types are gradually providing an understanding of rules controlling recognition of particular DNA loci by dbTFs in a particular cell type or biological condition. Main bioinformatics efforts try to account for contributions of chromatin accessibility and dbTF affinity when predicting locus-specific DNA recognition, which may help to combine dbTF specificity assayed in vitro and data from chromatin accessibility profiling of the particular cell type. If successful, such bioinformatics

**Table 1**
Major results achieved by GREEKC. Progress in the four areas of the GREEKC COST Action is published in this special issue (BBAGRM), or elsewhere.

| Action area | Result | Reference |
|---|---|---|
| WG1: Ontologies | GO: Updates of GO Transcription Factor branch | Gene Ontology Consortium [14] |
| | SO: Update of Gene Regulatory Element branch | Gaudet et al., this issue |
| | GRAO: Gene Regulation Application Ontology | Sant et al., this issue |
| | | https://github.com/GREEKC |
| WG2: Curation | dbTF to GO Catalogue | Lovering et al. this issue |
| | coTFs from predicted complexes | Velthuijs et al. this issue |
| | ncRNA curation | Panni et al. [19] |
| | TFBS PWM benchmarking | Ambrosini et al. [33] |
| | CTCF binding to topologically defined regions | Nanni et al. [35] |
| | MI2CAST curation guidelines | Touré et al. [46] |
| | The causalBuilder curation tool | Touré et al. [47] |
| | GO-CAM: TF-TG curation templates | Juanes Cortés et al., this issue |
| WG3: Text mining | ExTRI TF-TG text mining corpus | Vazquez et al., this issue |
| WG4: Data sharing | CausalTAB - PSI-MITAB 2.8 | Perfetto et al. [45] |
| | FAIR assessment GRKC | Bonello et al., this issue |
| | Purity and curation | Chatterjee et al., this issue |
| | PSICQUIC and SPARQL sharing of ExTRI data | Holmås et al. [49] |

strategies would save the researchers from exhaustive assessment of the active regulome of DNA binding transcription factors substituting it with reliable prediction of dbTF binding profiles at single base resolution and further pinpoint dbTF target genes. This is especially important for hard-to-get or transient cell types, and thus vital in the context of developmental biology or in studying the transcription response of different cells to particular physiological, environmental or stress conditions. Fortunately, future prospects to tackle such challenges are brightened by emerging opportunities to obtain single cell data relevant for gene regulation, such as transcriptomics, transcription factor binding and chromatin states and topologies. With support from comprehensive and well documented prior knowledge resources, such data might allow the researcher to unveil cell state-specific gene regulatory (sub)networks, which control behavior and transformation of cells existing in small quantities and/or short time frames but having a crucial impact on critical biological processes.

Precision medicine is an emerging approach that aims to develop personalized therapies for individual patients, by taking into account patient-specific disease factors to increase the efficacy of drug treatment [60,61]. Precision medicine may be based on large scale omics data collections to obtain high-resolution molecular insight into health [62], or on patient-specific mathematical models that serve as in silico patients, or 'digital twins' [63]. The builders and users of these patient-specific models are often involved in curation themselves, to make models complete and to audit literature in order to verify database information against contextual details of the processes that they are modeling. For instance, the Consortium for Logical Modeling Standards and Tools (CoLoMoTo [64]) represents scientists engaged in constructing logical models and the Disease Maps consortium generates biological process information [65] to support the analysis of many diseases. It is noteworthy that despite the large efforts in building resources that describe regulatory information that involves molecular components, be it genes or proteins, additional efforts are still needed to obtain the information to construct process diagrams or mathematical models that capture what we know about gene regulatory mechanisms adequately checked to have validity in a specific biological setting or context. Having an integration of the curation world with the modeling world through these types of collaborations, possibly with the help of a future COST action, has the potential to further optimize curation and annotation processes for the Knowledge Commons.

## Declaration of competing interest

The authors declare no conflict of interest.

## Acknowledgments

This publication is based upon work from COST Action CA15205: GREEKC, supported by COST (European Cooperation in Science and Technology). RCL has been supported by Alzheimer's Research UK grant (ARUK-NAS2017A-1) and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. IVK was supported by RSF grant 20-74-10075. MEF was supported by national Portuguese funding through a developmental grant (IF/00881/2013) by the FCT (Fundação para a Ciência e Tecnologia).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2021.194768.

## References

[1] C.E. Cook, O. Stroe, G. Cochrane, E. Birney, R. Apweiler, The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences, Nucleic Acids Res. 48 (2020) D17–D23, https://doi.org/10.1093/nar/gkz1033.

[2] C. Durinx, J. McEntyre, R. Appel, R. Apweiler, M. Barlow, N. Blomberg, C. Cook, E. Gasteiger, J.-H. Kim, R. Lopez, N. Redaschi, H. Stockinger, D. Teixeira, A. Valencia, Identifying ELIXIR core data resources, F1000Research 5 (2016), https://doi.org/10.12688/f1000research.9656.2.

[3] E.W. Sayers, J. Beck, E.E. Bolton, D. Bourexis, J.R. Brister, K. Canese, D.C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T. L. Madden, N. O'Leary, L. Phan, S.H. Rangwala, V.A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B.W. Trawick, K.D. Pruitt, S.T. Sherry, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 49 (2021) D10–D17, https://doi.org/10.1093/nar/gkaa892.

[4] R. von Schomberg, A vision of responsible research and innovation, in: Responsible Innovation, John Wiley & Sons. Ltd., 2013, pp. 51–74, https://doi.org/10.1002/9781118551424.ch3.

[5] R. Nydal, G. Bennett, M. Kuiper, A. Lægreid, Silencing trust: confidence and familiarity in re-engineering knowledge infrastructures, Med. Health Care Philos. 23 (2020) 471–484, https://doi.org/10.1007/s11019-020-09957-0.

[6] A. Holinski, M.L. Burke, S.L. Morgan, P. McQuilton, P.M. Palagi, Biocuration - mapping resources and needs, F1000Research 9 (2020), https://doi.org/10.12688/f1000research.25413.2.

[7] International Society for Biocuration, Biocuration: distilling data into knowledge, PLoS Biol. 16 (2018), e2002846, https://doi.org/10.1371/journal.pbio.2002846.

[8] N. Guarino, D. Oberle, S. Staab, What is an ontology? in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems Springer, Berlin, Heidelberg, 2009, pp. 1–17, https://doi.org/10.1007/978-3-540-92673-3_0.

[9] J. Hastings, Primer on ontologies, Methods Mol. Biol. Clifton NJ 1446 (2017) 3–13, https://doi.org/10.1007/978-1-4939-3743-1_1.

[10] T.C. Jepsen, Just what is an ontology, anyway? IT Prof. 11 (2009) 22–27, https://doi.org/10.1109/MITP.2009.105.

[11] S. Schulz, L. Jansen, Formal ontologies in biomedical knowledge representation, Yearb. Med. Inform. 8 (2013) 132–146.

[12] K. Kostelidou, F. Babiloni, Why bother with a COST action? The benefits of networking in science, Nonlinear Biomed. Phys. 4 (Suppl. 1) (2010) S12, https://doi.org/10.1186/1753-4631-4-S1-S12.

[13] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.

[14] Gene Ontology Consortium, The Gene Ontology resource: enriching a GOld mine, Nucleic Acids Res. 49 (2021) D325–D334, https://doi.org/10.1093/nar/gkaa1113.

[15] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, The Sequence Ontology: a tool for the unification of genome annotations, Genome Biol. 6 (2005) R44, https://doi.org/10.1186/gb-2005-6-5-r44.

[16] E. Beisswanger, V. Lee, J.-J. Kim, D. Rebholz-Schuhmann, A. Splendiani, O. Dameron, S. Schulz, U. Hahn, Gene Regulation Ontology (GRO): design principles and use cases, Stud. Health Technol. Inform. 136 (2008) 9–14.

[17] A. Venkatesan, S. Tripathi, A. Sanz de Galdeano, W. Blondé, A. Lægreid, V. Mironov, M. Kuiper, Finding gene regulatory network candidates using the gene expression knowledge base, BMC Bioinformatics 15 (2014) 386, https://doi.org/10.1186/s12859-014-0386-y.

[18] R.P. Huntley, D. Sitnikov, M. Orlic-Milacic, R. Balakrishnan, P. D'Eustachio, M. E. Gillespie, D. Howe, A.Z. Kalea, L. Maegdefessel, D. Osumi-Sutherland, V. Petri, J. R. Smith, K. Van Auken, V. Wood, A. Zampetaki, M. Mayr, R.C. Lovering, Guidelines for the functional annotation of microRNAs using the Gene Ontology, RNA N. Y. N 22 (2016) 667–676, https://doi.org/10.1261/rna.055301.115.

[19] S. Panni, R.C. Lovering, P. Porras, S. Orchard, Non-coding RNA regulatory networks, Biochim. Biophys. Acta Gene Regul. Mech. 2020 (1863) 194417, https://doi.org/10.1016/j.bbagrm.2019.194417.

[20] R.P. Huntley, B. Kramarz, T. Sawford, Z. Umrao, A. Kalea, V. Acquaah, M.J. Martin, M. Mayr, R.C. Lovering, Expanding the horizons of microRNA bioinformatics, RNA N. Y. N 24 (2018) 1005–1017, https://doi.org/10.1261/rna.065565.118.

[21] F. Zambelli, G. Pesole, G. Pavesi, Motif discovery and transcription factor binding sites before and after the next-generation sequencing era, Brief. Bioinform. 14 (2013) 225–237, https://doi.org/10.1093/bib/bbs016.

[22] C. Rastogi, H.T. Rube, J.F. Kribelbauer, J. Crocker, R.E. Loker, G.D. Martini, O. Laptenko, W.A. Freed-Pastor, C. Prives, D.L. Stern, R.S. Mann, H.J. Bussemaker, Accurate and sensitive quantification of protein-DNA binding affinity, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E3692–E3701, https://doi.org/10.1073/pnas.1714376115.

[23] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (2015) 831–838, https://doi.org/10.1038/nbt.3300.

[24] J. Zhou, O.G. Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model, Nat. Methods 12 (2015) 931–934, https://doi.org/10.1038/nmeth.3547.

[25] G.D. Stormo, T.D. Schneider, L. Gold, A. Ehrenfeucht, Use of the "perceptron" algorithm to distinguish translational initiation sites in E. coli, Nucleic Acids Res. 10 (1982) 2997–3011, https://doi.org/10.1093/nar/10.9.2997.

[26] O.G. Berg, P.H. von Hippel, Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters, J. Mol. Biol. 193 (1987) 723–750, https://doi.org/10.1016/0022-2836(87)90354-8.

[27] I.V. Kulakovskiy, V.J. Makeev, DNA sequence motif: a jack of all trades for ChIP-Seq data, Adv. Protein Chem. Struct. Biol. 91 (2013) 135–171, https://doi.org/10.1016/B978-0-12-411637-5.00005-6.

[28] G.D. Stormo, Y. Zhao, Determining the specificity of protein-DNA interactions, Nat. Rev. Genet. 11 (2010) 751–760, https://doi.org/10.1038/nrg2845.
[29] E. Wingender, The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, Brief. Bioinform. 9 (2008) 326–332, https://doi.org/10.1093/bib/bbn016.
[30] I.V. Kulakovskiy, I.E. Vorontsov, I.S. Yevshin, R.N. Sharipov, A.D. Fedorova, E. I. Rumynskiy, Y.A. Medvedeva, A. Magana-Mora, V.B. Bajic, D.A. Papatsenko, F. A. Kolpakov, V.J. Makeev, HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis, Nucleic Acids Res. 46 (2018) D252–D259, https://doi.org/10.1093/nar/gkx1106.
[31] M.T. Weirauch, A. Yang, M. Albu, A.G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S.A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M.G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J.S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A.J.M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J.R. Ecker, T.R. Hughes, Determination and inference of eukaryotic transcription factor sequence specificity, Cell 158 (2014) 1431–1443, https://doi.org/10.1016/j.cell.2014.08.009.
[32] O. Fornes, J.A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B.P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, A. Mathelier, JASPAR 2020: update of the open-access database of transcription factor binding profiles, Nucleic Acids Res. 48 (2020) D87–D92, https://doi.org/10.1093/nar/gkz1001.
[33] G. Ambrosini, I. Vorontsov, D. Penzar, R. Groux, O. Fornes, D.D. Nikolaeva, B. Ballester, J. Grau, I. Grosse, V. Makeev, I. Kulakovskiy, P. Bucher, Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study, Genome Biol. 21 (2020) 114, https://doi.org/10.1186/s13059-020-01996-3.
[34] K.L. Howe, P. Achuthan, James Allen, Jamie Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, A.G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. Garcia Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O.G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. Gonzalez Martinez, J.C. Marugán, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D.N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M.P. Sakthivel, A.I.A. Salam, B.M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. De Silva, B. Flint, A. Frankish, S.E. Hunt, G.R. IIsley, N. Langridge, J. E. Loveland, F.J. Martin, J.M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S.J. Trevanion, F. Cunningham, A.D. Yates, D.R. Zerbino, P. Flicek, Ensembl 2021, Nucleic Acids Res. 49 (2021) D884–D891, https://doi.org/10.1093/nar/gkaa942.
[35] L. Nanni, S. Ceri, C. Logie, Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries, Genome Biol. 21 (2020) 197, https://doi.org/10.1186/s13059-020-02108-x.
[36] R. Balakrishnan, M.A. Harris, R. Huntley, K. Van Auken, J.M. Cherry, A guide to best practices for Gene Ontology (GO) manual annotation, Database J. Biol. Databases Curation 2013 (2013), bat054, https://doi.org/10.1093/database/bat054.
[37] R.P. Huntley, M.A. Harris, Y. Alam-Faruque, J.A. Blake, S. Carbon, H. Dietze, E. C. Dimmer, R.E. Foulger, D.P. Hill, V.K. Khodiyar, A. Lock, J. Lomax, R. C. Lovering, P. Mutowo-Meullenet, T. Sawford, K. Van Auken, V. Wood, C. J. Mungall, A method for increasing expressivity of Gene Ontology annotations using a compositional approach, BMC Bioinformatics 15 (2014) 155, https://doi.org/10.1186/1471-2105-15-155.
[38] P.D. Thomas, D.P. Hill, H. Mi, D. Osumi-Sutherland, K. Van Auken, S. Carbon, J. P. Balhoff, L.-P. Albou, B. Good, P. Gaudet, S.E. Lewis, C.J. Mungall, Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems, Nat. Genet. 51 (2019) 1429–1433, https://doi.org/10.1038/s41588-019-0500-1.
[39] P. Porras, E. Barrera, A. Bridge, N. Del-Toro, G. Cesareni, M. Duesbury, H. Hermjakob, M. Iannuccelli, I. Jurisica, M. Kotlyar, L. Licata, R.C. Lovering, D. J. Lynn, B. Meldal, B. Nanduri, K. Paneerselvam, S. Panni, C. Pastrello, M. Pellegrini, L. Perfetto, N. Rahimzadeh, P. Ratan, S. Ricard-Blum, L. Salwinski, G. Shirodkar, A. Shrivastava, S. Orchard, Towards a unified open access dataset of molecular interactions, Nat. Commun. 11 (2020) 6144, https://doi.org/10.1038/s41467-020-19942-z.
[40] R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, M. Tyers, The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions, Protein Sci. Publ. Protein Soc. 30 (2021) 187–200, https://doi.org/10.1002/pro.3978.
[41] B.H.M. Meldal, O. Forner-Martinez, M.C. Costanzo, J. Dana, J. Demeter, M. Dumousseau, S.S. Dwight, A. Gaulton, L. Licata, A.N. Melidoni, S. Ricard-Blum, B. Roechert, M.S. Skyzypek, M. Tiwari, S. Velankar, E.D. Wong, H. Hermjakob, S. Orchard, The complex portal—an encyclopaedia of macromolecular complexes, Nucleic Acids Res. 43 (2015) D479–D484, https://doi.org/10.1093/nar/gku975.
[42] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S.G.N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, R. Apweiler, The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data, Nat. Biotechnol. 22 (2004) 177–183, https://doi.org/10.1038/nbt926.
[43] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, B. Aranda, A.F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J.J. Salama, S. Moore, A. Ceol, A. Chatr-Aryamontri, M. Oesterheld, V. Stümpflen, L. Salwinski, J. Nerothin, E. Cerami, M.E. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler, H. Hermjakob, Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions, BMC Biol. 5 (2007) 44, https://doi.org/10.1186/1741-7007-5-44.
[44] M. Sivade Dumousseau, D. Alonso-López, M. Ammari, G. Bradley, N.H. Campbell, A. Ceol, G. Cesareni, C. Combe, J. De Las Rivas, N. Del-Toro, J. Heimbach, H. Hermjakob, I. Jurisica, M. Koch, L. Licata, R.C. Lovering, D.J. Lynn, B.H. M. Meldal, G. Micklem, S. Panni, P. Porras, S. Ricard-Blum, B. Roechert, L. Salwinski, A. Shrivastava, J. Sullivan, N. Thierry-Mieg, Y. Yehudi, K. Van Roey, S. Orchard, Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions, BMC Bioinformatics 19 (2018) 134, https://doi.org/10.1186/s12859-018-2118-1.
[45] L. Perfetto, M.L. Acencio, G. Bradley, G. Cesareni, N. Del Toro, D. Fazekas, H. Hermjakob, T. Korcsmaros, M. Kuiper, A. Lægreid, P. Lo Surdo, R.C. Lovering, S. Orchard, P. Porras, P.D. Thomas, V. Touré, J. Zobolas, L. Licata, CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination, Bioinforma. Oxf. Engl. 35 (2019) 3779–3785, https://doi.org/10.1093/bioinformatics/btz132.
[46] V. Touré, S. Vercruysse, M.L. Acencio, R.C. Lovering, S. Orchard, G. Bradley, C. Casals-Casas, C. Chaouiya, N. Del-Toro, Å. Flobak, P. Gaudet, H. Hermjakob, C. T. Hoyt, L. Licata, A. Lægreid, C.J. Mungall, A. Niknejad, S. Panni, L. Perfetto, P. Porras, D. Pratt, J. Saez-Rodriguez, D. Thieffry, P.D. Thomas, D. Türei, M. Kuiper, The minimum information about a molecular interaction causal statement (MI2CAST), Bioinforma. Oxf. Engl. (2020), https://doi.org/10.1093/bioinformatics/btaa622.
[47] V. Touré, J. Zobolas, M. Kuiper, S. Vercruysse, CausalBuilder: bringing the MI2CAST causal interaction annotation standard to the curator, Database J. Biol. Databases Curation 2021 (2021), https://doi.org/10.1093/database/baaa107.
[48] S. Vercruysse, J. Zobolas, V. Touré, M.K. Andersen, M. Kuiper, VSM-box: General-purpose Interface for Biocuration and Knowledge Representation, 2020, https://doi.org/10.20944/preprints202007.0557.v1.
[49] S. Holmås, R.R. Puig, M.L. Acencio, V. Mironov, M. Kuiper, The Cytoscape BioGateway App: explorative network building from the BioGateway triple store, Bioinforma. Oxf. Engl. (2019), https://doi.org/10.1093/bioinformatics/btz835.
[50] N. del-Toro, M. Dumousseau, S. Orchard, R.C. Jimenez, E. Galeota, G. Launay, J. Goll, K. Breuer, K. Ono, L. Salwinski, H. Hermjakob, A new reference implementation of the PSICQUIC web service, Nucleic Acids Res. 41 (2013) W601–W606, https://doi.org/10.1093/nar/gkt392.
[51] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A.R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G.J. Warner, T. Ideker, G.D. Bader, Integration of biological networks and gene expression data using Cytoscape, Nat. Protoc. 2 (2007) 2366–2382, https://doi.org/10.1038/nprot.2007.324.
[52] E. Prud'hommeaux, A. Seaborne, SPARQL query language for RDF. [WWW document], URL, https://www.w3.org/TR/rdf-sparql-protocol/, 2008 (accessed 3.11.21).
[53] E. Antezana, W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, M. Kuiper, BioGateway: a semantic systems biology tool for the life sciences, BMC Bioinformatics 10 (Suppl. 10) (2009) S11, https://doi.org/10.1186/1471-2105-10-S10-S11.
[54] F. Ceccarelli, D. Turei, A. Gabor, J. Saez-Rodriguez, Bringing data from curated pathway resources to Cytoscape with OmniPath, Bioinforma. Oxf. Engl. 36 (2020) 2632–2633, https://doi.org/10.1093/bioinformatics/btz968.
[55] L. Garcia-Alonso, C.H. Holland, M.M. Ibrahim, D. Turei, J. Saez-Rodriguez, Benchmark and integration of resources for the estimation of human transcription factor activities, Genome Res. 29 (2019) 1363–1375, https://doi.org/10.1101/gr.240663.118.
[56] S. Vercruysse, M. Kuiper, Intuitive Representation of Computable Knowledge, 2020, https://doi.org/10.20944/preprints202007.0486.v2.
[57] M. Krallinger, F. Leitner, A. Valencia, Analysis of biological processes and diseases using text mining approaches, Methods Mol. Biol. Clifton NJ 593 (2010) 341–382, https://doi.org/10.1007/978-1-60327-194-3_16.
[58] C.-H. Wei, A. Allot, R. Leaman, Z. Lu, PubTator central: automated concept annotation for biomedical full text articles, Nucleic Acids Res. 47 (2019) W587–W593, https://doi.org/10.1093/nar/gkz389.
[59] N. Škunca, R.J. Roberts, M. Steffen, Evaluating computational gene ontology annotations, Methods Mol. Biol. Clifton NJ 1446 (2017) 97–109, https://doi.org/10.1007/978-1-4939-3743-1_8.
[60] B. Comte, J. Baumbach, A. Benis, J. Basílio, N. Debeljak, Å. Flobak, C. Franken, N. Harel, F. He, M. Kuiper, J.A. Méndez Pérez, E. Pujos-Guillot, T. Režen, D. Rozman, J.A. Schmid, J. Scerri, P. Tieri, K. Van Steen, S. Vasudevan, S. Watterson, H.H.H.W. Schmidt, Network and systems medicine: position paper of the European collaboration on science and technology action on open multiscale systems medicine, Netw. Syst. Med. 3 (2020) 67–90, https://doi.org/10.1089/nsm.2020.0004.
[61] F. Eduati, P. Jaaks, J. Wappler, T. Cramer, C.A. Merten, M.J. Garnett, J. Saez-Rodriguez, Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies, Mol. Syst. Biol. 16 (2020) e8664, https://doi.org/10.15252/msb.20188664.

[62] N.D. Price, A.T. Magis, J.C. Earls, G. Glusman, R. Levy, C. Lausted, D.T. McDonald, U. Kusebauch, C.L. Moss, Y. Zhou, S. Qin, R.L. Moritz, K. Brogaard, G.S. Omenn, J. C. Lovejoy, L. Hood, A wellness study of 108 individuals using personal, dense, dynamic data clouds, Nat. Biotechnol. 35 (2017) 747–756, https://doi.org/10.1038/nbt.3870.

[63] F. Pappalardo, G. Russo, F.M. Tshinanu, M. Viceconti, In silico clinical trials: concepts and early adoptions, Brief. Bioinform. 20 (2019) 1699–1708, https://doi.org/10.1093/bib/bby043.

[64] A. Naldi, P.T. Monteiro, C. Müssel, Consortium for Logical Models and Tools, H. A. Kestler, D. Thieffry, I. Xenarios, J. Saez-Rodriguez, T. Helikar, C. Chaouiya, Cooperative development of logical modelling standards and tools with CoLoMoTo, Bioinforma. Oxf. Engl. 31 (2015) 1154–1159, https://doi.org/10.1093/bioinformatics/btv013.

[65] M. Ostaszewski, S. Gebel, I. Kuperstein, A. Mazein, A. Zinovyev, U. Dogrusoz, J. Hasenauer, R.M.T. Fleming, N. Le Novère, P. Gawron, T. Ligon, A. Niarakis, D. Nickerson, D. Weindl, R. Balling, E. Barillot, C. Auffray, R. Schneider, Community-driven roadmap for integrated disease maps, Brief. Bioinform. 20 (2019) 659–670, https://doi.org/10.1093/bib/bby024.

[66] F.H.C. Crick, On protein synthesis, 1957. Manuscript. Cold Spring Harbor Laboratory Archives, SB/11/5/4, http://libgallery.cshl.edu/items/show/52220.