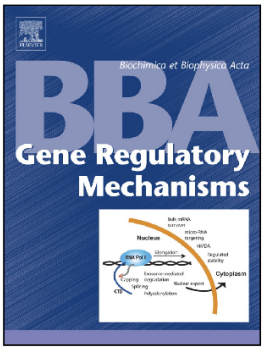
Journal Pre-proof

# The Gene Regulation Knowledge Commons

Martin Kuiper, Joseph Bonello, Jesualdo Tomás Fernández-Breis, Philipp Bucher, Matthias E. Futschik, Pascale Gaudet, Ivan V. Kulakovskiy, Luana Licata, Colin Logie, Ruth C. Lovering, Vsevolod J. Makeev, Sandra Orchard, Simona Panni, Livia Perfetto, David Sant, Stefan Schulz, Daniel R. Zerbino, Astrid Lægreid, The GRECO Consortium, Stefan Schulz, Christoph Bock, Stein Aerts, Klaas Vandepoele, Lejla Kapur-Pojskic, Naida Lojo-Kadrić, Ney Lemke, Vesselin Baev, Wyeth Wasserman, Jacques van Helden, Benoit Ballester, Juan Vaquerizas, Maria Gazouli, Dimitris Kardassis, Dávid Fazekas, Des Higgins, Livia Perfetto, Simona Panni, Luana Licata, Piero Carninci, Juris Viksna, Marcio Acencio, András Hartmann, Stephanie Kreis, Ernest Cachia, Joseph Bonello, Nikolai Pace, Julio Collado Vides, Kody Moodley, Michel Dumontier, Colin Logie, Sebastian Schmeier, Astrid Lægreid, Martin Kuiper, Steven Vercruysse, Anthony Mathelier, Matthias Futschik, Daniel Sobral, Filipe Castro, Pedro T. Monteiro, Marieta Costache, Gina Cecilia Pistol, Mihail Alexandru Gras, Sorina Dinescu, Ivan V. Kulakovskiy, Yulia Medvedeva, Vsevolod Makeev, Iva Pruner, Branislava Gemovic, Valentina Djordjevic, Damjana Rozman, Martin Krallinger, Alfonso Valencia, Miguel Vazquez, Ismael Navas-Delgado, José F. Aldana, José Manuel García-Nieto, María del Mar Roldán, Jesualdo Tomás Fernández Breis, Philipp Bucher, Fabio Rinaldi, Yavuz Oktay, Selcuk Sozer Tokdemir, Anton Popov, Sandra Orchard, Ruth Lovering, Chris Mungall, Paul Thomas, Karen Eibeck




PII: S1874-9399(21)00086-9

DOI: https://doi.org/10.1016/j.bbagrm.2021.194768

Reference: BBAGRM 194768

To appear in: *BBA - Gene Regulatory Mechanisms*

Received date: 20 April 2021

Revised date: 18 October 2021

Accepted date: 20 October 2021

Please cite this article as: M. Kuiper, J. Bonello, J.T. Fernández-Breis, et al., The Gene Regulation Knowledge Commons, *BBA - Gene Regulatory Mechanisms* (2021), https://doi.org/10.1016/j.bbagrm.2021.194768

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier B.V.

Journal Pre-proof

# The Gene Regulation Knowledge Commons

## The Action Area of GREEKC

**Authors:**
Martin Kuiper[1], Joseph Bonello[2], Jesualdo Tomás Fernández-Breis[3], Philipp Bucher[4], Matthias E. Futschik[5], Pascale Gaudet[6], Ivan V. Kulakovskiy[7], Luana Licata[8], Colin Logie[9], Ruth C Lovering[10], Vsevolod J Makeev[11], Sandra Orchard[12], Simona Panni[13], Livia Perfetto[14], David Sant[15], Stefan Schulz[16], Daniel R. Zerbino[17], Astrid Lægreid[18], The GRECO Consortium[19].

**Affiliations:**
[1]Systems biology group, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway.
[2]Faculty of Information & Communication Technology, University of Malta, Msida.
[3]Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, CP 30100, Murcia, Spain
[4]Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Amphipôle, 1015 Lausanne, Switzerland
[5]Systems Biology and Bioinformatics Laboratory (SysBioLab), Centre of Marine Sciences (CCMAR), University of Algarve, 8005-139 Faro, Portugal
[6]SIB Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1204 Geneva, Switzerland
[7]Institute of Protein Research, Russian Academy of Sciences, 142290, Institutskaya 4, Pushchino, Russia
[8]Department of Biology, University of Rome Tor Vergata, Rome, Italy.
[9]Department of Molecular Biology, Faculty of Science, Radboud University, PO box 9101, Nijmegen 6500HG, The Netherlands
[10]Functional Gene Annotation, Pre-clinical and Fundamental Science, Institute of Cardiovascular Science, University College London, 5 University Street, London, WC1E 6JF, UK
[11]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, Gubkina 3, Moscow, Russia
[12]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
[13]Department DIBEST, University of Calabria, Rende, Italy.
[14]Fondazione Human Technopole, Department of Biology, Via Cristina Belgioioso, 171, 20157 Milan, Italy.
[15]Department of Biomedical Informatics, University of Utah, 421 Wakara Way #140, Salt Lake City, UT 84108, United States
[16]Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerpl. 2, Graz, Austria
[18]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[18]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[19]The GRECO authors: see table at end of paper.

Journal Pre-proof

# Abstract

The COST Action *Gene Regulation Ensemble Effort for the Knowledge Commons* (GREEKC, CA15205, www.greekc.org) organised 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 parts of the knowledge cycle that is central to understanding gene regulatory mechanisms. The discussions between ontologists, curators, text miners, biologists, bioinformaticians, philosophers and computational scientists spawned a host of activities aimed to update and standardise existing knowledge management workflows, encourage new experimental approaches and thoroughly involve end-users in the process to design the Gene Regulation Knowledge Commons (GRKC). The GREEKC consortium describes its main achievements, contextualised in a state-of-the-art of current tools and resources that today represent the GRKC.


# Introduction

Understanding how complex biological systems operate is not possible without computational modeling of data and knowledge. In fact, the biological knowledge discovery process itself is becoming increasingly dependent on computational modeling and simulation, in systems biology approaches. The construction of computer models requires comprehensive knowledge of biological entities and their interactions, and abundant efforts are dedicated to providing such information in databases (Cook et al., 2020; Durinx et al., 2016; Sayers et al., 2021). Despite all this, joint undertakings that involve stakeholders across the different expert areas necessary to specify and design the necessary knowledge domains, formats, content, access and use (the knowledge life cycle) have been scant, explaining why many of the 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 genomic information governs cellular differentiation and function. The complex machinery that determines which genes are active requires an intricate and 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 defines the accessibility and proximity to their target genes. Real progress to comprehensively improve the construction and governance of knowledge repositories that provide detailed information about each of these types of gene regulatory actors and their causal interactions, needs multidisciplinary efforts engaging many different expert groups that should be open to collaborate and agree on common goals.

The COST Action *Gene Regulation Ensemble Effort for the Knowledge Commons* (GREEKC) was designed to improve the development of the Gene Regulation Knowledge Commons (GRKC). This GRKC is defined by the GREEKC consortium as: “The collection of freely accessible information resources, with data 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 worked toward improving the resources contributing to this GRKC by coordinating many of the efforts around production and consumption of ‘knowledge’ pertinent to this domain, following a responsible research and innovation (RRI) approach (Schomberg, 2013), i.e. engaging all stakeholders to optimise the deliverables of a scientific process, and to align scientific processes and outcomes to societal needs. We took this strategy as an iterative process of identifying and including stakeholders, starting with key players in the knowledge life cycle, and it needed constant

work of mutual adjustment of stakeholder concerns to ensure scientific and public trust. The strategy is to instill partakers to overcome competing interest and become mutually responsive and adaptive to each other concerns in light of the common goals that are being articulated in the process (Nydal et al., 2020).This RRI approach proved to be an extremely good fit with the main merit of COST Actions: bringing people together in one room who would not normally discuss or consult with each other.

GREEKC was the result of an initiative which started in 2013 and brought together a wide range of experts from the domain of gene regulatory mechanisms: The Gene Regulation Consortium (GRECO, www.theGRECO.org). Through a GRECO-inspired proposal we received funding from COST in 2016, allowing us to commence on a four-year journey using the different COST mechanisms (Workshops and Working Group meetings, Training Schools, Short Term Scientific Missions) to advance the coordinated building of the GRKC.

# GREEKC's field of operation and design

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 standardised formats in publicly available resources, if it is to become broadly available for computational integration and analysis (Holinski et al., 2020; International Society for Biocuration, 2018). 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 and able to identify and characterize specific biological entities and interactions described in papers or large-scale data repositories and to investigate their contents for experimental or other evidence that facilitates their use or supports particular claims about their biological function. These claims are described, or annotated, with the help of controlled vocabularies that provide standardised terms, descriptions and definitions for concepts that are relevant for a (sub)domain of biology. Domain Ontologies contain, in addition, machine-processable formal axioms and definitions of types of domain entities, hierarchically organized so that they can facilitate analysis at different levels (Guarino et al., 2009; Hastings, 2017; Jensen, 2009) thus constituting the building blocks for representing human knowledge (Schulz and Jansen, 2013). Describing biological entities and their relationships in specific contexts such as scientific experiments with the help of unambiguously defined ontology terms is performed in an annotation process that follows well-defined curation guidelines, so that different biocurators are able to interpret and annotate knowledge from a paper in identical ways. This is often supported by curation tools, which provide additional guidance as to the annotation details that need to be provided. There are many subdomains of biology that require such annotation efforts. GREEKC elected to focus on the area of gene regulatory mechanisms (Figure 1) whilst developing knowledge gathering and sharing principles that will have value across all biological domains.

The curation of information from scientific literature starts with the identification of content conveying curatable information. The identification of such content can be facilitated by text mining algorithms that flag papers, or segments thereof, for the presence of information that can be pre-marked for content appropriate for subsequent manual curation into a database. Alternatively, with appropriate post-processing, this information can be extracted directly for computational use. However, whereas the potential of text mining for assisting manual curation is well-established, its direct integration into curation workflows has not found wide

use yet. For those curation workflows that produce information for the GRKC, the breadth of annotations impacts their representation, storage in a database schema and subsequent sharing mechanisms. For instance, complex annotations need to meet well-defined curation guidelines and storage formats, and stored data require specific 'exchange languages' (e.g. XML-based or JSON-based) for downloads or web services that enable data transfer to the user.

The different elements of the gene regulation knowledge management life cycle as described above were converted into four challenges that were taken as a basis to assemble the four working groups of the GREEKC COST Action. These Working Groups (WGs) focused on:

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, addressed in five sub-working groups:

1. protein level
2. non-coding RNA level
3. nucleotide sequence recognition level (e.g. transcription factor binding sites)
4. genome level (DNA methylation status, histone modifications)
5. 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 the various regulatory molecular entities and their regulatory interactions;
WG4: The storing and sharing of annotations of gene regulatory interactions.

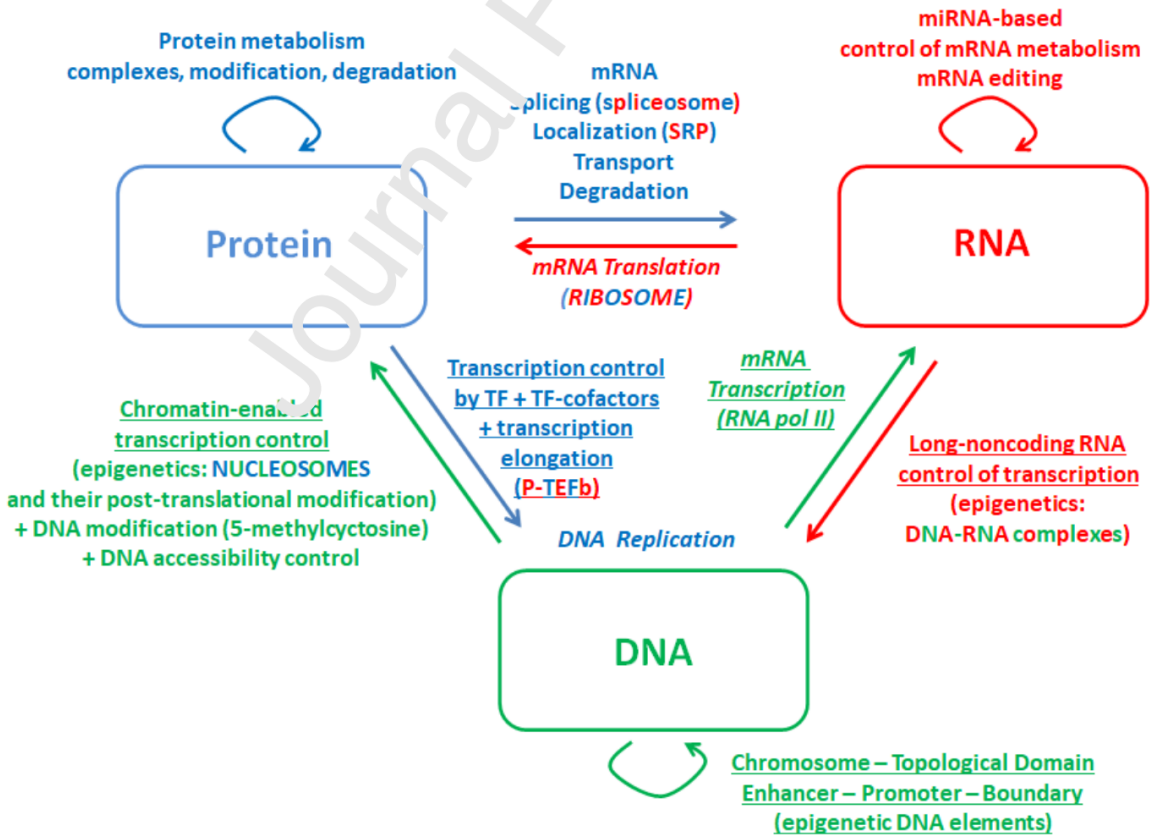


**Figure 1**: Schematic overview of gene regulation processes (courtesy of C. Logie). The regulation of gene expression involves the three genetically encoded polymer classes; DNA (green), RNA (red) and Protein (blue). Complexes involving different polymers, are described with mixed letter colors (**RNA-PROTEIN COMPLEXES, DNA-PROTEIN COMPLEXES, DNA-RNA COMPLEXES**); Underlined biological processes denote DNA-centric transcription control; Francis Crick's central dogma (Crick, 1958) is shown in italics. These various entities, complexes and processes represent the area of interest for the Gene Regulation Knowledge Commons.

# Ways of working and accomplishments

As a typical COST Action, the main mechanism to organize the scientific domain, stimulate discussions, strive for consensus and achieve progress (Kostelidou and Babiloni, 2010) was through the organization of Workshops, Training Schools and Short Term Scientific Missions (STSMs). Here we elaborate 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 main topics in GREEKC workshops since 2016, they were mostly intertwined with topics on ontologies and controlled vocabularies as well as text mining for gene regulation knowledge management and available annotation tools and their output formats. This means that much of what we achieved in GREEKC cannot be uniquely assigned to one particular Working Group but rather to the continuous joint efforts involving members from all groups.

**WG1: Ontologies**
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, (Ashburner et al., 2000; Gene Ontology Consortium, 2021) and the Sequence Ontology (SO) (Eilbeck et al., 2005) 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 the continued evolution of the knowledge life cycle and 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 (see 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 (see Gaudet et al., 2021, 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 DNA binding Transcription Factors) 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 SO. A restructuring of the SO has now been proposed (see 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 breath of protein-DNA interaction events and chromatin conformation states that have an impact on gene expression. Finally, efforts have been launched to follow up on the Gene Regulation Ontology (Beisswanger et al., 2008), 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) would pave the way for new knowledge bases able to semantically integrate all knowledge about gene regulatory events. The integration of gene regulatory mechanism data in the GRAO backbone would allow for complex queries addressing many aspects about regulatory context simultaneously, going well beyond the examples published for the Gene Expression Knowledge Base (Venkatesan et al., 2014).

**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, reliable information. GREEKC has addressed five different subdomains of biocuration in its workshops and in several areas notable progress was made (see below). Here we describe the main results to advance biocuration for the GRKC per WG2 subdomain.

The protein level:

GREEKC members have collaborated on the task of bringing together the knowledge that currently supports the classification of proteins as DNA binding Transcription Factors (dbTFs) (see 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 development of re-design of the GO transcription regulation molecular functions branch and an updated set of curation guidelines (see Gaudet et al., 2021, 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.

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 actions. In the last decade, strong efforts have been made to annotate both functional and physical RNA interactions in public repositories. While there are guidelines to use the Gene Ontology to capture the role of microRNAs in gene regulation (Huntley et al., 2016), no specific guidelines had been developed for the majority of other RNA roles, with the result that the knowledge extracted from one source is sometimes difficult to integrate or compare with other sources. During the COST meetings, valuable discussions among GREEKC members about entity identifiers, quality criteria for the reliability of the data, and necessary metadata, led to the definition of common standards for the annotation of microRNA-mRNA and microRNA-lncRNA interactions (Panni et al., 2020). MicroRNAs are the best-characterized regulatory RNAs, and their binding partners can be predicted using bioinformatic approaches that allow the interaction site to be mapped to the target gene. However, as each prediction tool provides different numbers of targets for each specific microRNA, the value of experimental confirmation of a microRNA-mRNA interaction should not be underestimated (Huntley et al., 2018). 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 and improved cooperation (Panni et al., 2020). In addition, a short term scientific mission has been supported by the COST programme that allowed one week of testing of the annotation guidelines. However, we have yet to do the same for functional interactions of the LncRNAs with genes and their role in transcriptional regulation.

The DNA level:
Whilst the dbTFs represent the protein side of the decoding of genome information, their specific binding sites in the genome uniquely interlink dbTF regulatory activity to specific genes. Because of their importance, these transcription factor binding sites have been extensively studied to characterize their nucleotide patterns (sequence motifs) and determine features that define binding specificity (Zambelli et al., 2013). A sequence motif recognized by a dbTF reflects the binding energy of a dbTF to a particular DNA segment (Rastogi et al., 2018), 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 (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015). However, the gold standard is still defined by position weight matrices (PWMs) which were suggested as early as 1982 (Stormo et al., 1982) and remain the most widespread and accepted way of describing dbTF binding specificity as a quantitative rather than a qualitative phenomenon (Berg and von Hippel, 1987). As a tool, PWMs are massively used to predict Transcription Factor Binding Sites (TFBS) in the genome and annotate regulatory sequence variants (Kulakovskiy and Makeev, 2013; Stormo and Zhao, 2010). 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 (Wingender, 2008), HOCOMOCO (Kulakovskiy et al., 2018), CIS-BP (Weirauch et al., 2014), and JASPAR (Fornes et al., 2020). Creating some common understanding for how these PWMs may 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 these PWM in practical analyses of new datasets (Annala et al., 2020).

The Genome level:
The sequence ontology (SO) is an essential source of terms that describe among others sequence concepts necessary to annotate regulatory sequences and transcription factor binding sites for a range of resources (e.g. Ensembl (Howe et al., 2021). SO was improved by the restructuring of terms related to cis-regulatory modules (CRMs), which are regulatory regions where transcription factor binding sites are clustered to regulate various aspects of transcription. CRMs contain 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 (Nanni et al., 2020). 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 (Nanni et al., 2020).

Level of interactions, regulatory complexes and network information flow:
The annotation of proteins in the GO database is based on well-established guidelines (Balakrishnan et al., 2013), but the underlying data model and output, the Gene Product Association Data (GPAD) file, does not fully take into account inter-relationships that more completely describe functional aspects relevant for a systems understanding of biological processes. 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 summarised by the limited vocabulary of the annotation extension (Huntley et al., 2014).

The GO-CAM data model (Thomas et al., 2019) 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 (see Juanes Cortés et al., 2021, this issue).

Transcription factors rarely bind as monomeric proteins but initially as homo-/heterodimers which then bind to co-factors to assemble the protein machinery required for transcription. GREEKC members (see Velthuijs et al., 2021, this issue) used data from the IMEx Consortium databases (Porras et al., 2020) and BioGRID (Oughtred et al., 2021) 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. A concomitant effort to manually curate transcription factor and coregulator complexes in the Complex Portal database (Meldal et al., 2015) has also been initiated as a result of the work of the GREEKC Action.

The PSI-MI standards that have been developed under the umbrella of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI) were the starting point (Hermjakob et al., 2004; Kerrien et al., 2007; Sivade Dumousseau et al., 2018) for discussions about future needs of the network modelling 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 signalling pathway to an up- or down-regulation of a set of genes. GREEKC was able to organise 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 ((Perfetto et al., 2019), 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 (Touré et al., 2020), 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 (Touré et al., 2021), to illustrate how a Visual Syntax Markup (VSM-box) data entry template engine (Vercruysse et al., 2020) can be used to support the presentation of an annotation standard in an organic way to a biocurator.

**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 organised 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 (see 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, 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, may omit it from analysis.

**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 standardisation 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 standardisation 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 should be given to open the possibility 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 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 (see Bonello et al, 2021, this issue). The SIGNOR database, for instance, abides by the FAIR principles and was an early adopter of the PSI-MI standards endorsed by IMEx. The SIGNOR curators led the activities supported by GREEKC to develop the HUPO-PSI MITAB2.8 (causalTAB) format that supports regulatory, causal relationships. However, this PSI-MI extension poses new demands for data exchange mechanisms, most notably the webservice PSICQUIC (Protein Standard Initiative Common Query Interface (del-Toro et al., 2013)), which, at the time of writing, is only able to serve queries for the PSI-MI 2.7 format. The GREEKC discussions

led to a Short Term Scientific Mission 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 (Cline et al., 2007), 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. Other web services that provide access to TF-TG interactions can be launched through Cytoscape Apps. The BioGateway App uses SPARQL queries (SPARQL Protocol and RDF Query Language (Prud'hommeaux E., Seaborne A., 2008)) to fetch regulatory information from the semantic web database BioGateway (Antezana et al., 2009), in the form of documented interactions between transcription factors and their target genes (see www.extri.org). Likewise, the OmniPath App (Ceccarelli et al., 2020) uses a REST type service to fetch TF-TG relationships from the dedicated transcription factor activity knowledge base DoRothEA (Garcia-Alonso et al., 2019).

# Discussion and future challenges

We hope that the Gene Regulation Knowledge Commons (GRKC) may serve as an example of what can be achieved through a concerted effort to create an effective data infrastructure. The GRKC should fulfil the needs of two groups: on the one hand bench biologists to access detailed information on their 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. For both uses, the annotation needs to be enhanced by including information with 'richer' expressions 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. The achievement of such enhancements of the GRKC will require further innovations of curation approaches and tools, while making sure that these richer annotations remain fully interoperable. It should also involve all resources covering protein, RNA, chromatin, small molecule and DNA sequence information, as well as the molecular activities and biological processes in which these entities are involved. The curation tool Noctua (Thomas et al., 2019) and new experimental technologies like VSM (Vercruysse and Kuiper, 2020) provide a significant step in the direction of annotating biological systems rather than biological entities. These tools accommodate multiple entity, activation state and relation types, and provide for annotations based on multiple ontologies and supported by an elaborate set of evidence and biological context 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 work on the design of the Gene Regulation Application Ontology warrants follow-up efforts to produce for instance a prototype semantic knowledge base where GRKC information is integrated together with SO regulatory sequence concepts, the Complex portal and GO molecular function and biological process concepts to allow users to query for regulatory mechanism information that meets both location/sequence constraints, macromolecular assemblies and gene regulatory action constraints. This promises to supply users with information supported by a richer set of annotation details, and in formats that facilitates their integration and querying.

Users of the Knowledge Commons with information about gene regulatory mechanisms need rich and exhaustive information sources for the computational analysis of biological processes, for instance by way of the construction of comprehensive models of such processes. 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 (Krallinger et al., 2010) as well as by automated and manual curation, needs to gain more attention (Wei et al., 2019). 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 (Škunca et al., 2017). 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 while biocurators creating manual annotations often focus on including only true positives in their database content, this may come at the expense of discarding false negatives. Hence, information that is not included in a resource but upon closer inspection of additional or new data may in fact provide sufficient evidence to meet the database's inclusion criteria. Such information might be flagged by appropriate evidence codes, so that computational users may explore it either in a 'cautious' or 'greedy' mode (see Chatterjee et al., 2021, this issue). A typical example for this is information gathered from text mining, which has the potential to exhaustively explore all available text for valuable information, but which generates too many false positives to be acceptable for most curated databases.

Another case concerns the numerous ongoing efforts to computationally predict '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. More than 30 year long 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 profile. Yet, massive efforts in comparative studies of dbTF binding *in vitro* and *in vivo* in a variety of cell types gradually resulted in condensation of understanding of rules controlling recognition of particular DNA loci by protein factors in a particular cell type or biological condition. The main effort has been oriented to account for contributions of chromatin accessibility and dbTF affinity when predicting locus-specific DNA recognition, which would help to combine dbTF specificity assayed *in vitro* and data from chromatin accessibility profiling of the particular cell type. In the case of success, such bioinformatics 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 transcriptional 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 behaviour and transformation of cells existing in small quantities and/or short time frames, but having crucial impact on critical biological processes.

The development of advanced text mining and bioinformatics approaches to discover new knowledge in the literature or large-scale data, should not be taken as competing interests between manual, human-based curation and automated, computer-based curation: rather there seems to be plenty of potential synergy between the two approaches: with manual curation developing training sets and validating algorithm progress, and automated annotation step-wise improving itself with the help of manual annotators. Further dialogue between biocurators, text miners, bioinformaticians, computational biologists and

policymakers / granting agencies is needed to define productive working modes, information quality criteria and metadata that would satisfy all stakeholders.

Precision medicine is an emerging approach that aims to develop personalised therapies for individual patients, by taking into account patient-specific disease factors to increase the efficacy of drug treatment (Comte et al., 2020; Eduati et al., 2020). Precision medicine may be based on large scale omics data collections to obtain high-resolution molecular insight into health (Price et al., 2017), or on patient-specific mathematical models that serve as *in silico* patients, or 'digital twins' (Pappalardo et al., 2019). Obviously, this requires deep computational integration of a multitude of data types and prior knowledge. Important players in the domain of computational analysis are the builders and users of conceptual or mathematical models of biological processes. These computational and biomedical scientists 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 model. For instance, the Consortium for Logical Modelling Standards and Tools (CoLoMoTo (Naldi et al., 2015)) represents scientists engaged in constructing logical models and the Disease Maps consortium generates biological process information (Ostaszewski et al., 2019) 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 modelling world through these types of collaborations, possibly with the help of a future COST action, has the potential to further optimise curation and annotation processes for the Knowledge Commons.

# 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).

**References**

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

Ambrosini, G., Vorontsov, I., Penzar, D., Groux, R., Fornes, O., Nikolaeva, D.D., Ballester, B., Grau, J., Grosse, I., Makeev, V., Kulakovskiy, I., Bucher, P., 2020. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. Genome Biol. 21, 114. https://doi.org/10.1186/s13059-020-01996-3

Antezana, E., Blondé, W., Egaña, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., Kuiper, M., 2009. BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 10 Suppl 10, S11. https://doi.org/10.1186/1471-2105-10-S10-S11

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S.,

Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29. https://doi.org/10.1038/75556

Balakrishnan, R., Harris, M.A., Huntley, R., Van Auken, K., Cherry, J.M., 2013. A guide to best practices for Gene Ontology (GO) manual annotation. Database J. Biol. Databases Curation 2013, bat054. https://doi.org/10.1093/database/bat054

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

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

Ceccarelli, F., Turei, D., Gabor, A., Saez-Rodriguez, J., 2020. Bringing data from curated pathway resources to Cytoscape with OmniPath. Bioinforma. Oxf. Engl. 36, 2632–2633. https://doi.org/10.1093/bioinformatics/btz968

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A.R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B.R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G.J., Ideker, T., Bader, G.D., 2007. Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. 2, 2366–2382. https://doi.org/10.1038/nprot.2007.324

Comte, B., Baumbach, J., Benis, A., Basílio, J., Debeljak, N., Flebak, Å., Franken, C., Harel, N., He, F., Kuiper, M., Méndez Pérez, J.A., Pujos-Guillot, E., Režen, T., Rozman, D., Schmid, J.A., Scerri, J., Tieri, P., Van Steen, K., Vasudevan, S., Watterson, S., Schmidt, H.H.H.W., 2020. Network and Systems Medicine: Position Paper of the European Collaboration on Science and Technology Action on Open Multiscale Systems Medicine. Netw. Syst. Med. 3, 67–90. https://doi.org/10.1089/nsm.2020.0004

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

Crick, F.H., 1958. On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163.

del-Toro, N., Dumousseau, M., Orchard, S., Jimenez, R.C., Galeota, E., Launay, G., Goll, J., Breuer, K., Ono, K., Salwinski, L., Hermjakob, H., 2013. A new reference implementation of the PSICQUIC web service. Nucleic Acids Res. 41, W601-606. https://doi.org/10.1093/nar/gkt392

Durinx, C., McEntyre, J., Appel, R., Apweiler, R., Barlow, M., Blomberg, N., Cook, C., Gasteiger, E., Kim, J.-H., Lopez, R., Nadalschi, N., Stockinger, H., Teixeira, D., Valencia, A., 2016. Identifying ELIXIR Core Data Resources. F1000Research 5. https://doi.org/10.12688/f1000research.9656.2

Eduati, F., Jaaks, P., Wappler, J., Cramer, T., Merten, C.A., Garnett, M.J., Saez-Rodriguez, J., 2020. Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. Mol. Syst. Biol. 16, e8664. https://doi.org/10.15252/msb.20188664

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

Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W.W., Mathelier, A., 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 48, D87–D92. https://doi.org/10.1093/nar/gkz1001

Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., Saez-Rodriguez, J., 2019. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 29, 1363–1375. https://doi.org/10.1101/gr.240663.118

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

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

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

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G.N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., Apweiler, R., 2004. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat. Biotechnol. 22, 177–183. https://doi.org/10.1038/nbt926

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

Howe, K.L., Achuthan, P., Allen, James, Allen, Jamie, Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., Garcia Giron, C., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O.G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Gonzalez Martinez, J., Marugán, J.C., Maurel, T., McMahon, A.C., Mohanan, S., Moore, B., Muffato, M., Oheh, D.N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M.P., Salam, A.I.A., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., De Silva, N., Flint, B., Frankish, A., Hunt, S.E., Ilsley, G.R., Langridge, N., Loveland, J.E., Martin, F.J., Mudge, J.M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S.J., Cunningham, F., Yates, A.D., Zerbino, D.R., Flicek, P., 2021. Ensembl 2021. Nucleic Acids Res. 49, D884–D891. https://doi.org/10.1093/nar/gkaa942

Huntley, R.P., Harris, M.A., Alam-Faruque, Y., Blake, J.A., Carbon, S., Dietze, H., Dimmer, E.C., Foulger, R.E., Hill, D.P., Khodiyar, V.K., Lock, A., Lomax, J., Lovering, R.C., Mutowo-Meullenet, P., Sawford, T., Van Auken, K., Wood, V., Mungall, C.J., 2014. A method for increasing expressivity of Gene Ontology annotations using a compositional approach. BMC Bioinformatics 15, 155. https://doi.org/10.1186/1471-2105-15-155

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

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

International Society for Biocuration, 2018. Biocuration: Distilling data into knowledge. PLoS Biol. 16, e2002846. https://doi.org/10.1371/journal.pbio.2002846

Jepsen, T.C., 2009. Just What Is an Ontology, Anyway? IT Prof. 11, 22–27. https://doi.org/10.1109/MITP.2009.105

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A.F., Vinod, N., Bader, G.D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J.J., Moore, S., Ceol, A., Chatr-Aryamontri, A., Oesterheld, M., Stümpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M.E., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., Hermjakob, H., 2007. Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. 5, 44. https://doi.org/10.1186/1741-7007-5-44

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

Krallinger, M., Leitner, F., Valencia, A., 2010. Analysis of biological processes and diseases using text mining approaches. Methods Mol. Biol. Clifton NJ 593, 341–382. https://doi.org/10.1007/978-1-60327-194-3_16

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

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., Kolpakov, F.A., Makeev, V.J., 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 46, D252–D259. https://doi.org/10.1093/nar/gkx1106

Lovering, R.C., Gaudet, P., Acencio, M.L., Ignatchenko, A., Jolma, A., Fornes, O., Kuiper, M., Kulakovskiy, I.V., Lægreid, A., Martin, M.J., Logie, C., 2020. A GO catalogue of human DNA-binding transcription factors. bioRxiv 2020.10.28.359232. https://doi.org/10.1101/2020.10.28.359232

Meldal, B.H.M., Forner-Martinez, O., Costanzo, M.C., Dana, J., Demeter, J., Dumousseau, M., Dwight, S.S., Gaulton, A., Licata, L., Melidoni, A.N., Ricard-Blum, S., Roechert, B., Skyzypek, M.S., Tiwari, M., Velankar, S., Wong, E.D., Hermjakob, H., Orchard, S., 2015. The complex portal--an encyclopaedia of macromolecular complexes. Nucleic Acids Res. 43, D479-484. https://doi.org/10.1093/nar/gku975

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

Nanni, L., Ceri, S., Logie, C., 2020. Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. Genome Biol. 21, 197. https://doi.org/10.1186/s13059-020-02108-x

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

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

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., Tyers, M., 2021. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. Publ. Protein Soc. 30, 187–200. https://doi.org/10.1002/pro.3978

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

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

Perfetto, L., Acencio, M.L., Bradley, G., Cesareni, G., Del Toro, N., Fazekas, D., Hermjakob, H., Korcsmaros, T., Kuiper, M., Lægreid, A., Lo Surdo, P., Lovering, R.C., Orchard, S., Porras, P., Thomas, P.D., Touré, V., Zobolas, J., Licata, L., 2019. CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination. Bioinforma. Oxf. Engl. 35, 3779–3785. https://doi.org/10.1093/bioinformatics/btz132

Porras, P., Barrera, E., Bridge, A., Del-Toro, N., Cesareni, G., Duesbury, M., Hermjakob, H., Iannuccelli, M., Jurisica, I., Kotlyar, M., Licata, L., Lovering, R.C., Lynn, D.J., Meldal, B., Nanduri, B., Paneerselvam, K., Panni, S., Pastrello, C., Pellegrini, M., Perfetto, L., Rahimzadeh, N., Ratan, P., Ricard-Blum, S., Salwinski, L., Shirodkar, G., Shrivastava, A.,

Orchard, S., 2020. Towards a unified open access dataset of molecular interactions. Nat. Commun. 11, 6144. https://doi.org/10.1038/s41467-020-19942-z

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

Prud'hommeaux E., Seaborne A., 2008. SPARQL query language for RDF. [WWW Document]. URL https://www.w3.org/TR/rdf-sparql-protocol/ (accessed 3.11.21).

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

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

Schomberg, R. von, 2013. A Vision of Responsible Research and Innovation, in: Responsible Innovation. John Wiley & Sons, Ltd, pp. 51–74. https://doi.org/10.1002/9781118551424.ch3

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

Sivade Dumousseau, M., Alonso-López, D., Ammari, M., Bradley, G., Campbell, N.H., Ceol, A., Cesareni, G., Combe, C., De Las Rivas, J., Del Toro, N., Heimbach, J., Hermjakob, H., Jurisica, I., Koch, M., Licata, L., Lovering, R.C., Lynn, D.J., Meldal, B.H.M., Micklem, G., Panni, S., Porras, P., Ricard-Blum, S., Rosenheck, B., Salwinski, L., Shrivastava, A., Sullivan, J., Thierry-Mieg, N., Yehudi, Y., Van Roey, K., Orchard, S., 2018. Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. BMC Bioinformatics 19, 134. https://doi.org/10.1186/s12859-018-2118-1

Škunca, N., Roberts, R.J., Steffen, M., 2017. Evaluating Computational Gene Ontology Annotations. Methods Mol. Biol. Clifton NJ 1446, 97–109. https://doi.org/10.1007/978-1-4939-3743-1_8

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

Stormo, G.D., Zhao, Y., 2010. Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. 11, 751–760. https://doi.org/10.1038/nrg2845

Thomas, P.D., Hill, D.P., Mi, H., Osumi-Sutherland, D., Van Auken, K., Carbon, S., Balhoff, J.P., Albou, L.-P., Good, B., Gaudet, P., Lewis, S.E., Mungall, C.J., 2019. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat. Genet. 51, 1429–1433. https://doi.org/10.1038/s41588-019-0500-1

Touré, V., Vercruysse, S., Acencio, M.L., Lovering, R.C., Orchard, S., Bradley, G., Casals-Casas, C., Chaouiya, C., Del-Toro, N., Flobak, Å., Gaudet, P., Hermjakob, H., Hoyt, C.T., Licata, L., Lægreid, A., Mungall, C.J., Niknejad, A., Panni, S., Perfetto, L., Porras, P., Pratt, D., Saez-Rodriguez, J., Thieffry, D., Thomas, P.D., Türei, D., Kuiper, M., 2020. The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST). Bioinforma. Oxf. Engl. https://doi.org/10.1093/bioinformatics/btaa622

Touré, V., Zobolas, J., Kuiper, M., Vercruysse, S., 2021. CausalBuilder: bringing the MI2CAST causal interaction annotation standard to the curator. Database J. Biol. Databases Curation 2021. https://doi.org/10.1093/database/baaa107

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

Vercruysse, S., Kuiper, M., 2020. Intuitive Representation of Computable Knowledge. https://doi.org/10.20944/preprints202007.0486.v2

Vercruysse, S., Zobolas, J., Touré, V., Andersen, M.K., Kuiper, M., 2020. VSM-box: General-purpose Interface for Biocuration and Knowledge Representation. https://doi.org/10.20944/preprints202007.0557.v1

Wei, C.-H., Allot, A., Leaman, R., Lu, Z., 2019. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 47, W587–W593. https://doi.org/10.1093/nar/gkz389

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M.G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J.S., Govindarajan, S., Shaulsky, G., Walhout, A.J.M., Bouget, F.-Y., Ratsch, G., Larrondo, L.F., Ecker, J.R., Hughes, T.R., 2014. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158, 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009

Wingender, E., 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief. Bioinform. 9, 326–332. https://doi.org/10.1093/bib/bbn016

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

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

(Refs to be added while in press:)

Bonello et al., 2021, this issue:
Bonello, J, Cachia, E., Alfino, N. AutoFAIR - Automating FAIR Assessments Portal for Bioinformatics Tools.

Chatterjee et al., 2021, this issue:
Chatterjee A, Swierstra T, Kuiper M. The potential purity of polluted data: curator and user perspectives.

Gaudet et al., 2021, this issue:
Gaudet P, Logie C, Lovering RC, Kuiper M, Lægreid A, Thomas PD. Gene Ontology representation for transcription factor functions.

Juanes Cortés et al., 2021, this issue:
Juanes Cortés B, Vera-Ramos JA, Lovering RC, Gaudet P, Laegreid A, Logie C, Schulz S, del Mar Roldan-Garcia M, Kuiper M, Fernandez-Breis JT. Formalization of gene regulation knowledge using ontologies and Gene Ontology Causal Activity Models.

Lovering et al., 2021 this issue:
Lovering RC, Gaudet P, Acencio ML, Ignatchenko A, Jolma A, Fornes O, Kuiper M, Kulakovskiy IV, Lægreid A, Martin MJ, Logie C. A GO catalogue of human DNA-binding transcription factors.

Sant et al., 2021, this issue:
Sant WD, Sinclair M, Mungall CJ, Schulz S, Zerbino D, Lovering RC, Logie C, Eilbeck K. Sequence Ontology terminology for gene regulation.

Vazquez et al., 2021, this issue:
Vazquez M, Krallinger M, Leitner F, Kuiper M, Valencia A and Laegreid A. 2021. ExTRI: Extraction of Transcription Regulation Interactions from Literature. BBA details.

Velthuijs et al., 2021, this issue:

Velthuijs N, Integration of transcription co-regulator complexes with sequence-specific DNA-binding factor interactomes.

Bucher et al., 2021, this issue.
Rozman et al., 2021, this issue.
Collado-Vides et al., 2021, this issue.
Wasserman et al., 2021, this issue.

**Extended author list:**

| GRECO member | Institution | Country |
|---|---|---|
| Stefan Schulz | Medical University of Graz | Austria |
| Christoph Bock | Research Center for Molecular Medicine | Austria |
| Stein Aerts | Flanders Interuniversity Institute of Biotechnology | Belgium |
| Klaas Vandepoele | Ghent University | Belgium |
| Lejla Kapur-Pojskic | Institute for Genetic Engineering and Biotechnology | Bosnia and Herzegovina |
| Naida Lojo-Kadrić | Institute for Genetic Engineering and Biotechnology | Bosnia and Herzegovina |
| Ney Lemke | São Paulo State University | Brazil |
| Vesselin Baev | University of Plovdiv | Bulgaria |
| Wyeth Wasserman | University of British Columbia | Canada |
| Jacques van Helden | Aix-Marseille Université | France |
| Benoit Ballester | INSERM Institute of Health and Medical Research | France |
| Juan Vaquerizas | Max Planck Institute for Molecular Biomedicine | Germany |
| Maria Gazouli | University of Athens | Greece |
| Dimitris Kardassis | University of Crete | Greece |

Journal Pre-proof

| | | |
|---|---|---|
| Dávid Fazekas | Eötvös Loránd University | Hungary |
| Des Higgins | University College Dublin | Ireland |
| Livia Perfetto | Fondazione Human Technopole | Italy |
| Simona Panni | University of Calabria | Italy |
| Luana Licata | University of Rome Tor Vergata | Italy |
| Piero Carninci | RIKEN | Japan |
| Juris Viksna | Institute of Mathematics and Computer Science | Latvia |
| Marcio Acencio | University of Luxembourg | Luxembourg |
| András Hartmann | University of Luxembourg | Luxembourg |
| Stephanie Kreis | University of Luxembourg | Luxembourg |
| Ernest Cachia | University of Malta | Malta |
| Joseph Bonello | University of Malta | Malta |
| Nikolai Pace | University of Malta | Malta |
| Julio Collado Vides | National Autonomous University of Mexico | Mexico |
| Kody Moodley | Maastricht University | Netherlands |
| Michel Dumontier | Maastricht University | Netherlands |
| Colin Logie | Radboud University Nijmegen | Netherlands |
| Sebastian Schmeier | Massey University | New Zealand |
| Astrid Lægreid | Norwegian University of Science and Technology | Norway |
| Martin Kuiper | Norwegian University of Science and Technology | Norway |

Journal Pre-proof

| | | |
|---|---|---|
| Steven Vercruysse | Norwegian University of Science and Technology | Norway |
| Anthony Mathelier | University of Oslo | Norway |
| Matthias Futschik | Centro de Ciências do Mar | Portugal |
| Daniel Sobral | Instituto Gulbenkian de Ciencia | Portugal |
| Filipe Castro | Interdisciplinary Centre of Marine and Environmental Research | Portugal |
| Pedro T. Monteiro | University of Lisbon | Portugal |
| Marieta Costache | Department of Biochemistry and Molecular Biology, University of Bucharest | Romania |
| Gina Cecilia Pistol | National Research and Development Institute for Animal Biology and Nutrition | Romania |
| Mihail Alexandru Gras | National Research and Development Institute for Animal Biology and Nutrition | Romania |
| Sorina Dinescu | University of Bucharest | Romania |
| Ivan V. Kulakovskiy | Institute of Protein Research | Russian Federation |
| Yulia Medvedeva | Research Center of Biotechnology RAS | Russian Federation |
| Vsevolod Makeev | Vavilov Institute of General Genetics | Russian Federation |
| Iva Pruner | Institute of Molecular Genetics and Genetic Engineering | Serbia |
| Branislava Gemovic | University of Belgrade | Serbia |
| Valentina Djordjevic | University of Belgrade | Serbia |
| Damjana Rozman | University of Ljubljana | Slovenia |
| Martin Krallinger | Barcelona Supercomputing Center | Spain |

Journal Pre-proof

| | | |
|---|---|---|
| Alfonso Valencia | Barcelona Supercomputing Center | Spain |
| Miguel Vazquez | Barcelona Supercomputing Center | Spain |
| Ismael Navas-Delgado | University of Málaga | Spain |
| José F. Aldana | University of Málaga | Spain |
| José Manuel García-Nieto | University of Málaga | Spain |
| María del Mar Roldán | University of Málaga | Spain |
| Jesualdo Tomás Fernández Breis | University of Murcia | Spain |
| Philipp Bucher | Swiss Institute of Bioinformatics | Switzerland |
| Fabio Rinaldi | University of Zurich | Switzerland |
| Yavuz Oktay | Dokuz Eylul University | Turkey |
| Selcuk Sozer Tokdemir | Istanbul University | Turkey |
| Anton Popov | National Technical University of Ukraine | Ukraine |
| Sandra Orchard | European Molecular Biology Laboratory | United Kingdom |
| Ruth Lovering | University College London | United Kingdom |
| Chris Mungall | Lawrence Berkeley Lab | USA |
| Paul Thomas | University of Southern California | USA |
| Karen Eibeck | University of Utah | USA |

Journal Pre-proof

Journal Pre-proof

Highlights

- The COST Action GREEKC produced the results of a BBA-GRM-21 special issue.
- GREEKC worked to enhance knowledge management in the gene regulation domain.
- Improvements of the Gene Regulation Knowledge Commons are presented.

Journal Pre-proof